A STUDY OF EFFICIENCY OF VARIANCE ESTIMATION IN MULTI-STAGE CLUSTER SAMPLING WITH OBSERVATIONAL STUDIES

Md Abdul Qudus Sheikh[1]
Research Scholar Department Of Mathematics, Faculty of School of Teaching department, Dr. A. P. J. Abdul Kalam University Indore, MP, India
Dr. Anjna Rajoria [2]
Research Scholar Department Of Mathematics, Faculty of School of Teaching department, Dr. A. P. J. Abdul Kalam University Indore, MP, India
[1] maqudussheikh@gmail.com,[2] anjna_rajoria@rediffmail.com

ABSTRACT

An estimator (a statistic) cannot be better than the parameter it estimates. Multi-stage sampling thrives because of its cost-effectiveness and not because its estimator is better. This study emphasis the need to develop higher-order estimators of multi-stage sampling and their use for large-scale sample survey. In this study, the researchers developed estimators of population total and variance of four-stage cluster sampling and illustrated their use with numerical example. Researchers should not sacrifice the development and application of higher-order estimators of multi-stage sampling that are robust statistical techniques on the altar of the complexity of design nature of large population. To reap the actual fruit of sampling that culminates in the high level of predication and estimation accuracy in sample survey, higher-order multi-stage estimators should be used in large-scale sample surveys.

KEYWORDS: Estimators, multi-stage sampling, population total, population variance, cost-effectiveness, simple random sampling

INTRODUCTION

Although a lot of estimation procedures of population parameters have been in existence, more still need to be done in order to make sampling schemes and their results depict the real life situations which they attempt to explain. Meteorological predictions of weather conditions have high level of acceptability, because almost all predictions of meteorologists at local, national and international levels are successful. It would not be odd if sample survey estimates yield even more generally acceptable results. The level of success and acceptance of estimation is dependent on the level of efficiency and adequacy of estimator. Multistage sampling is a sampling scheme which deepens the real essence of sampling. Its cost-effectiveness attests to this fact. The scheme has attracted the attention of many renowned scholars who have developed estimation procedures for it. Some of the procedures include those of Cochran (1977), Kelton (1983), Okafor (2002) and Nafiu et al (2012) that derived a generalized form of multistage cluster sampling design based on lower-order stage of the Nafius (2012) procedure. Most of the procedures aforementioned focus on lower – order multistage, such as two or three-stage and their generalizations are often based on the lower – order stage. This limits the ability of other researchers who might not be in a position to develop, derive or extend the generalized procedure for use in large scale sampling. In this study, the researchers are of the view that higher – order multistage sampling procedures should be developed and used to improve the reliability of sample surveys, especially in large scale surveys. Moreso, every stage of a multistage sampling is a sampling technique itself, Lukman (2012). In order to make available higher – order multistage sampling procedure, the researchers' aim in this study is to develop estimators of the parameters of four-stage cluster sampling based on Okafor (2002) procedure. The objectives of the study are to (i) develop estimators of the population total and variance of four-stage cluster sampling, (ii) apply the developed estimators to the estimation of population total of farmers in a vicinity within Nasarawa and Keffi local government areas of Nasarawa State and (iii) determine the standard error and coefficient of variation of the estimated population total. Wolter, (2007) noted that naturally, representative samples of large populations often have complex design features because of their cost efficiency and size, among other reasons. For purposes of estimating sampling variances based on complex multi-stage sample design involving cluster sampling, sampling error codes are provided by survey organizations in public use survey files, in some developed countries of the world. The codes, where they exist, are used for "ultimate cluster selection" for individuals, Rendtel and Harms (2009). In some developed countries and developing countries like Nigeria where those codes do not exist, the normal procedure for estimating variance holds. This procedure is explained in the next section of this study. It is however necessary to note that if the first stage units (fsu's) were actually sampled without replacement in multi-stage design, which is typical of this study, standard estimates of population totals are estimated and Kerry and Bland, (2006) asserted that the variance estimator, in this case, only have a slight positive bias with the selection of ultimate cluster. Lavallee (2010) argue that since unbiased estimators for estimating totals in multi-stage cluster samples of various stages are primarily driven by the variance estimated totals within a stage, at least two first stage units are needed within a first stage sampling cluster for design based variance estimation purposes. It is a common mathematical knowledge that variance estimators for sample totals based on complex designs featuring without replacement selection of first stage units at the sampling stage, Canette (2010) affirm, are function of finite population correlations (FPC's) which account for the proportion of the finite population that was not included in a sample selected without replacement and joint probabilities of selection for sampled units. Carle (2009) and Rabe-Hesheth and Skrondal (2006) stress that generally, large sample sizes lead to FPC.

Notations for Sample Selection

The following notations are used in this study

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

n = the number of clusters selected from N fsu's by simple random sampling without replacement (srswor)

Mi = the number of clusters in the second stage unit (ssu)

mi= the number of clusters selected by srswor from Mi(i = 1,2,....... n) ssu's

Mij = the number of clusters in the third stage unit (tsu)

mij = the number of clusters selected by srswor from Mij (j = 1,2,......, mi) tsu's Finally
Mijk = the number of clusters (elements) in the fourth stage unit (4th su)
mijk = the number of clusters (elements) selected by srswor from Mijk (K = 1,2, ...mij) 4thsu's in each of the selected tsu's
Yijkl = the value of the characteristic, y, from lth4th su in the kthtsu of the jthssu in the ith fsu.
Y = the population total

Sample Selection in Multi-Stage Sampling

(Equal probability sampling with first stage units of unequal sizes)

From a population of N first stage units (fsu's), n fsu's are selected by simple random sampling without replacement (srswor). Next, a srswor of size mi second stage units (ssu's) is selected from Mi (i=1,2,......,n) within each of the selected fsu's. Then, a srswor of size mij third stage units (tsu's) is selected from Mij (j = 1,2.......mi) within each of the selected ssu's. A srswor of size mijk fourth stage units (4thsu's) is lastly selected from Mijk (K = 1,2,.....mij) within each of the selected tsu's

Now, we assume that the value of the characteristic, y, of interest, yijkl, in view of the notations defined above, is yijkl= the value of the characteristic, y, from lth 4thsu in the kth tsu of the jth ssu in the ith fsu.

List of Population Means and Totals of Interest

$$\overline{Y}_{ijk\,.} = \frac{1}{M_{ijk}} \sum_{1=1}^{Mijk} Y_{ijkl} = \frac{Y_{ijk\,.}}{M_{ijk}} =$$ the population mean of the kth tsu in

the jth ssu of the ith fsu.

$Y_{ijk\,.}$ = the population total of the kth tsu in the jth ssu of the ith fsu

$$\overline{Y}_{ij..} = \frac{1}{M_{ij}} \sum_{K=1}^{Mij} Y_{ijk.} = \frac{Y_{i\,j..}}{M_{ij}} =$$ the population mean of the jth ssu in the ith fsu

$Y_{ij..}$ = the population total of the jth ssu in the ith fsu

$$\overline{Y}_{i...} = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij..} = \frac{Y_{i...}}{M_i} = \quad =$$ the population mean of the ith fsu

$Y_{i..}$ = the population total of the ith fsu

List of Sample Means and Totals of Interest

$\bar{y}_{ijk\cdot} = \frac{1}{m_{ijk}} \sum_{i=1}^{mijk} y_{i\,jkl} = \frac{y_{ijk.}}{m_{ijk}}$ = the sample mean of the kth tsu in the jth ssu of the ith fsu

$y_{ijk.}$ = the sample total of the kth tsu in the jth ssu of the ith fsu

$\bar{y}_{ij\cdot} = \frac{1}{m_{ij}} \sum_{k=1}^{mij} y_{i\,jk.} = \frac{y_{ij..}}{m_{ij}}$ = the sample mean of the jth ssu in the ith fsu

y_{ij} = The sample total of the jth ssu in the ith fsu

$\bar{y}_{i...} = \frac{1}{m_i} \sum_{j=1}^{mi} y_{i\,j..} = \frac{y_{i...}}{m_i}$ = the sample mean of the ith fsu

$y_{i...}$ = the sample total of the ith fsu

Estimation of Population Total

We start from the inclusion probability as follows:

The inclusion probability of Horvitz-Thompson Estimator in the jth ssu within the ith fsu according to Okafor (2002) and Wolter (2007) is given by

$$\pi_{(ij)} = (\frac{n}{N})(\frac{m_i}{M_i}) \qquad (1)$$

Where

$$\frac{n}{N} = f_1 \quad , \qquad (2)$$

is the sampling fraction of fsu,

$$\frac{m_i}{M_i} = f_2 \quad , \qquad (3)$$

is the sampling fraction of ssu

In view of equation (1), we add that

the inclusion probability of the kth tsu within the jth ssu in the ith fsu is

$$\pi_{(ijk)} = (\frac{n}{N})\,(\frac{m_i}{M_i})\,(\frac{m_{ij}}{M_{ij}}) \qquad (4)$$

Where

$$\frac{m_{ij}}{M_{ij}} = f_3 \quad , \qquad (5)$$

is the sampling fraction of the tsu
Similarly, we include that;
the inclusion probability of the lth 4th su within the kth tsu in the jth ssu of the ith fsu is

$$\pi_{(ijk)} = (\frac{n}{N})\,(\frac{m_i}{M_i})\,(\frac{m_{ij}}{M_{ij}})(\frac{m_{ijk}}{M_{ijk}}) \qquad (6)$$

Where

$$\frac{m_{ijk}}{M_{ijk}} = f_4 \qquad , \qquad (7)$$

is the sampling fraction of the 4thsu.

$$\frac{N}{n} = \frac{1}{f_1};\ \frac{M_i}{m_i} = \frac{1}{f_2};\ \frac{M_{ij}}{m_{ij}} = \frac{1}{f_3} \qquad (8)$$

are the raising factors of fsu's, ssu's, tsu's, respectively. Using the inclusion probability, in this study, as shown in (1) from which (4) and (6) were derived, we further state that the unbiased estimator of the population total, y, of four-stage cluster sampling, , ∧ Y given by

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{mi}\frac{M_{ij}}{m_{ij}}\sum_{k=1}^{mij} \frac{M_{ijk}}{m_{ijk}}\sum_{i=1}^{mijk} Y_{ijkl} = \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{mi}\frac{M_{ij}}{m_{ij}}\sum_{k=1}^{mij} M_{ijk}\,\bar{y}_{ijk} \quad (9)$$

Where,

$$= \frac{M_{ijk}}{m_{ijk}} = \frac{1}{f_4} \qquad , \qquad (10)$$

is the raising factor of the 4th su

Notations for variance estimation

E4 = the conditional expectation over the selection of the 4th su's selected from the tsu's that was obtained from the ssu's in each fsu.
V4 = its corresponding variance
E3 = the conditional expectation over the selection of tsu's from the ssu's in each fsu
V3 = its corresponding variance
E2 = the conditional expectation over the selection of the ssu's from each fsu
V2 = its corresponding variance
E1 = the unconditional expectation over all possible samples of n fsu
V1 = its corresponding variance

In the same vein, we state that the variance of the estimator of the population total of four-stage cluster sampling, $\hat{Y}$, which is $V\left(\hat{Y}\right)$ is derived by the model

$$V\left(\hat{Y}\right) = E_1E_2E_3V_4\left(\hat{Y}\right) + E_1E_2V_3E_4\left(\hat{Y}\right) + E_1V_2E_3E_4\left(\hat{Y}\right) + V_1E_2E_3E_4\left(\hat{Y}\right) \qquad (11)$$

Applying the general theorem for obtaining the sampling variance of the sample mean $\left(\bar{y}\right)$ in element sampling which states that

$$V\left(\bar{y}\right) = \frac{\sigma^2}{n} \quad \textit{for srswr} \qquad (12)$$

Or $V\left(\bar{y}\right) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$ *for srswor* (13)

Where $\sigma^2 = \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 / N =$ the population (element) variance, (14)

using the analysis of variance approach which gives the population variance as

$$S^2 = \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 /(N-1), \qquad (15)$$

and by substituting equation (15) in (14); then, (14) in (13), we obtain that the sampling variance of the fsu is

$$V^1\left(\hat{y}\right) = N^2\frac{1-f_1}{n}S_1^2 \qquad (16)$$

Similarly, the sampling variance of the ssu and the tsu are obtained. The sampling variance of the 4thsu is therefore derived as

$$V^{1111}\left(\hat{y}\right) = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{M_{ij}}{m_{ij}}\sum_{k=1}^{m_{ij}} M_{ijk}\frac{1-f_{4ijk}}{m_{ijk}}S_{4ijk}^2 \qquad (17)$$

Hence, we further state that the unbiased sample estimator of the sampling variance, $\hat{V}\left(\hat{Y}\right)$ of a four-stage cluster sampling is given by

$$\hat{V}\left(\hat{y}\right) = N^2 \frac{1-f_1}{n} s_1^2 + \frac{N}{n} \sum_{i=1}^{n} M^2{}_i \frac{1-f_{2i}}{m_i} s_{2i}^2 + \frac{N}{n} \sum_{i=1} \frac{M_i}{m_i} \sum_{j=1}^{mi} M_i{}^2 \frac{1-f_{3ij}}{m_{ij}} s_{3ij}^2$$

$$+ \frac{N}{n} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{mi} \frac{M_{ij}}{m_{ij}} \sum_{k=1}^{mij} M^2{}_{ijk} \frac{1-f_{4ij}}{m_{ijk}} s_{4ijk}^2 \qquad (18)$$

Where

$$s_1^2 = \sum_{j=1}^{n} \left(\hat{y}_{i\dots} - \frac{\wedge}{y_{\dots}} \right)^2 /(n-1) \qquad , \qquad (19)$$

is the sample variability among cluster totals in the fsu

$$s_{2i}^2 = \sum_{j=1}^{mi} \left(\hat{y}_{ij..} - \frac{\wedge}{y_{i\dots}} \right)^2 /(m_i - 1) \qquad (20)$$

$$s_{3ij}^2 = \sum_{k=1}^{mij} \left(\hat{y}_{ijk.} - \frac{\wedge}{y_{ij..}} \right)^2 /(m_{ij} - 1) \qquad (21)$$

Equations (20), (21) and (22) are the sample variabilities among cluster totals in the 2nd , 3 rd and 4th stage units, respectively. f1, f2, f3, and f4 are as defined in equations 2, 3, 5 and 7, respectively.

CONCLUSION

The benefits of sampling are much more appreciated in multi-stage cluster sampling. Its elegance is more obvious in higher-order multistage sampling which is typical of largescale surveys. The uses of lower-order multi-stage estimators for the analysis of data arising from large population, more often than not, end up in misleading or unacceptable results. The researchers, in this study, used the estimators of two-stage cluster sampling to illustrate the use of lower-order multi-stage estimators while the derived four-stage estimator was used to illustrate the use of higher-order multi-stage estimator. Researchers should be encouraged to develop and use higher-order multi-stage estimators for the analysis of large-scale surveys to increase the level of accuracy and acceptability of sample survey results. Encouragement, in this regard, should be in the form of sponsorship of survey, commercialization and patent (where necessary).

REFERENCES

[1]. Lavallee, P. (2010): Cross Section Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. Survey Methodology, 32, 165-176

[2]. Lukman, A. N. (2012): Comparison of One-Stage, Two-Stage and Three-Stage Estimators Using Finite Population. The Pacific Journal of Science and Technology, Volume 13, Number 2, November 2012.

[3]. Nafiu, L. A, Oshungade, I. O. and Adewara, A. A. (2012): Generalization (Unequal Probability Class) of Multi-Stage Cluster Sampling Using Finite Population. International Journal of Engineering and Applied Sciences.

[4]. Okafor, F. C. (2002): Sample Survey Theory with Applications. Afro-Orbis Publications Ltd 33 Catering Rest House Road, P. O. Box 3160, University of Nigeria, Nsukka, Nigeria.

[5]. Rabe-Hesheth, S. and Skrondal, A. (2006): Multilevel Modelling of Complex Survey Data. Journal of the Royal Statistical Society-Series A, 169, 805-827

[6]. Rendtel, W. and Harms, T. (2009): Weighting and Calibration for Household Panels in Methodology of Longitudinal Surveys. Ed. P. Lynn, Chichester: John Wiley and Sons, 265-286

[7]. Wolter, K. M. (2007): Introduction to Variance Estimation. Second Edition Springer-Verlag

[8]. Ram Kumar, Gunja Varshney , Tourism Crisis Evaluation Using Fuzzy Artificial Neural network, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011

[9]. Ram Kumar, Jasvinder Pal Singh, Gaurav Srivastava, “A Survey Paper on Altered Fingerprint Identification & Classification” International Journal of Electronics Communication and Computer Engineering Volume 3, Issue 5, ISSN (Online): 2249–071X, ISSN (Print): 2278– 4209

[10]. Kumar, R., Singh, J.P., Srivastava, G. (2014). Altered Fingerprint Identification and Classification Using SP Detection and Fuzzy Classification. In: , et al. Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Advances in Intelligent Systems and Computing, vol 236. Springer, New Delhi. https://doi.org/10.1007/978-81-322-1602-5_139

[11]. A. S. Rajawat and A. R. Upadhyay, "Web Personalization Model Using Modified S3VM Algorithm For developing Recommendation Process," 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170701.

[12]. A. Singh Rajawat and S. Jain, "Fusion Deep Learning Based on Back Propagation Neural Network for Personalization," 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-7, doi: 10.1109/IDEA49133.2020.9170693.

[13]. Chetan Chauhan, Ravindra Gupta and Kshitij Pathak. Article: Survey of Methods of Solving TSP along with its Implementation using Dynamic Programming Approach. International Journal of Computer Applications 52(4):12-19, August 2012.