International Research Journal of Social Sciences ISSN 2319–3565
Vol. **7(5),** 34-38, May **(2018)** *Int. Res. J. Social Sci.*

Review Paper

A review on survey designs adopted by statistics Botswana

Raghunath Arnab*, Thapelo Sediadie and Vijai Kumar Dwivedi
University of Botswana, Private Bag: UB 00705, Gaborone, Botswana
arnabr@mopipi.ub.bw

Available online at: www.isca.in, www.isca.me
Received 25th January 2018, revised 27th April 2018, accepted 10th May 2018

Abstract

Statistics Botswana, Statistics South Africa and several other statistical organizations conduct nation wide surveys on regular basis using complex survey designs. The designs comprises of statification, clustering and unequal probability of selection of samples. In such sampling designs, the first-stage units are selected with unequal probability sampling design and the second-stage units are selected with systematic sampling procedures. It has been pointed out that the unbiased variance estimation for certain population characteristics such as the population mean and proportion for such sampling designs do not exist. Apart from the variance estimation, several standard statistical tests such as goodness of fit, independency and homogenity are not valid for such designs. Standard software packages SPSS, BMDP and SAS provides very often erroneous results. In this paper we have reviewed various methods of variance estimation, independency and testing hypothesis for complex survey designs adapted by the Statistics Botswana.

Keywords: Complex survey design, design effect, inclusion probability proportional to size, Jackknife method, normed size measure, random group method, variance estimation.

Introduction

Statistics Botswana (former Central Statistics Office (CSO)) has a parastatel status in the Ministry of Finance and Development Planning. It is the apex body in the official statistical system of the country. Most official statistical operations fall within the mandate of this office. Statistics Botswana's primary function is to provide Government Ministries and Departments, NGO's and members of the public in general, with information for monitoring, evaluation and formulation of development plans and programmes. It performs this function through collection, processing and analysis of data, reporting and dissemination of results through publications, workshop and seminars. For the republic of South Africa (RSA) such responsibilities are bestowed on the Statistics South Africa.

Statistics Botswana conducts large scale nation wide surveys on a regular basis using stratified multi-stage sampling designs recommended by UNDP. Such surveys include: Household Income Expenditure Surveys (HIES 2004), Botswana Aids impact surveys (BAIS, 2004, 2008), Botswana Family Heath Survey (BFHS 2007), amongst others. The districts of Botswana are considered as strata and each of the 27 districts is divided into a number of enumeration areas (EA). The EA's are identical to the population and housing Census of Botswana, 2001. The (i) Cities & Towns, (ii) Urban villages and (iii) Rural Districts were treated as separate strata. EA's of a district were treated as first-stage units (fsu's) while households within EA's were treated as second-stage or ultimate units as the statistical information is collected from the sampled households. The EAs are selected by PPS systematic sampling scheme where inclusion probabilities are exactly proportional to the household size of the respective EA. Similar sampling designs are applied by Statistics South Africa for national household surveys[1].

Variance estimation is important for estimating efficiencies of the survey estimates, interval estimation and hypothesis testing relating to the population parameters. It is also required in determining the optimal sample size and the evaluation of cost of a survey[2]. For a multi-stage sampling design where the first-stage units are selected using without replacement (WOR) sampling and the systematic or any other sampling procedure is used for second and subsequent stages of sampling, unbiased variance estimates of the population parameters is not possible[3].

In this paper, we will review various methods for variance estimation and hypothesis testing proposed by various researchers based on Statistics Botswana survey designs. We also study the existence of suitable alternative sampling designs which Statistics Botswana could be used for deriving more reliable inferences.

Sampling Designs

The entire country Botswana is divided into H strata (district). The h -th stratum consists of N_h EAs (first-stage units) and the i-th EA of the h -th stratum consists of M_{hi} households (second-stage units). Thus the total number of the households in Botswana is $M = \sum_h \sum_i M_{hi}$ .From each of the stratum e.g. h -

th stratum, a sample s_h of size $n_h\ (= nN_h / N, N = \sum_h N_h)$ EAs is selected by PPS systematic sampling procedure (described below) with $p_{hi}\ (= M_{hi} / M)$ as normed size measure for i-th EA of the hth stratum $i = 1,..,M_{hi}; h = 1,..,H$. If the i-th EA of the h-th stratum is included in the sample s_h, a sub-sample s_{hi} of size m households is selected from the M_{hi} households of i-th EA by linear systematic sampling scheme (assuming M_{hi} / m an integer).

PPS Systematic sampling scheme: The PPS systematic sampling scheme was introduced in early 1950[4,1]. Let $T_{hi} = n_h \sum_{j=1}^{hi} p_{hj}$ for $i = 1,..., M_h$ and $T_{h0} = 0$. In this method, for the h-th-stratum, we select a random sample (called random start) d_h from a uniform distribution with the range (0, 1). This random start d_h selects a sample from the h-th statum with those units whose index, " j " , satisfies $T_{hj-1} \le d_h + k < T_{hj}$ for $k = 0,1,..,n_h - 1$ [5].

The sampling scheme ensures that the inclusion probability of the i[th] EA of the h[th] stratum is $\pi_{i|h} = n_h p_{hi}$ where n_h is the sample size and $p_{hi}\ \left(p_{hi} > 0, \sum_{i=1}^{M_h} p_{hi} = 1 \right)$ is the normed size measure of the ith EA of the hth stratum. The sampling scheme is very simple to implement and it is applicable to any value of n_h as long as $\pi_{i|h} = n_h p_{hi} \le 1$ However, the main drawback of this procedure is that the expressions for the second order inclusion probabilities $\pi_{ij|h}$ are highly complex.

Assuming that the labels of the units are attached at random, the following approximate expression of $\pi_{ij|h}$ to the order $O(N_h^{-4})$ when p_{hi} is of $O(N_h^{-1})$, n_h is relatively small to N_h and N_h is moderately large was given by [6].

$$\pi_{ij|h} = n(n-1) p_{hi} p_{hj} \begin{bmatrix} 1 + \{(p_{hi} + p_{hj}) - \sum_j p_{hj}^2\} + \{2(p_{hi}^2 + p_{hj}^2) - 2\sum_j p_{hj}^3 \\ + 2 p_{hi} p_{hj} - 3(p_{hi} + p_{hj}) \sum_j p_{hj}^2 + 3(\sum_j p_{hj}^2)^2\} \end{bmatrix} \quad (1)$$

Estimation of the population mean per unit: Let y_{hij} and q_{hij} be the value of the variable understudy (indicators) y and family size q for the jth household of the ith EA of the hth stratum. Then the population total of y and q are:

$$Y = \sum_{h=1}^{H} Y_h \text{ and } Q = \sum_{h=1}^{H} Q_h$$

where $Y_h = \sum_{i=1}^{N_h} Y_{hi}$, $Y_{hi} = \sum_{j=1}^{M_{hi}} y_{hij}$, $Q_h = \sum_{i=1}^{N_h} Q_{hi}$ and $Q_{hi} = \sum_{j=1}^{M_{hi}} q_{hij}$.

The estimator for the population mean per individual $\bar{Y} = Y / Q$ is given by

$$\hat{\bar{Y}} = \hat{Y} / \hat{Q} \quad (2)$$

where $\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h$, $\hat{Y}_h = \sum_{i \in s_h} \frac{\hat{Y}_{hi}}{\pi_{i|h}}$, $\hat{Y}_{hi} = M_{hi} \bar{y}_{hi}$, $\hat{Q} = \sum_{h=1}^{H} \hat{Q}_h$, $\hat{Q}_h = \sum_{i \in s_h} \frac{\hat{Q}_{hi}}{\pi_{i|h}}$, $\hat{Q}_{hi} = M_{hi} \bar{q}_{hi}$; $\bar{y}_{hi}$ and $\bar{q}_{hi}$ are the sample means of y and q -values of the systematic sample s_{hi} .

The estimator (2) is a ratio estimator and it is unbiased for $\bar{Y}$ for large sample sizes[7]. An approximate mean-square error or variance is given by

$$V\left(\hat{\bar{Y}}\right) \cong \frac{1}{Q^2} \sum_{h=1}^{H} V\left(\hat{D}_h\right)$$

$$= \frac{1}{Q^2} \left[\sum_{h=1}^{H} \left\{ \frac{1}{2} \sum_{i \neq}^{N_h} \sum_{j=1}^{N_h} \left(\pi_{i|h}\pi_{j|h} - \pi_{ij|h}\right) \left(\frac{D_{hi}}{\pi_{i|h}} - \frac{D_{hj}}{\pi_{j|h}} \right)^2 + \sum_{i=1}^{N_h} \sigma_{dhi}^2 \right\} \right] \quad (3)$$

where $\hat{D}_h = \hat{Y}_h - \bar{Y}\hat{Q}_h$, $D_h = Y_h - \bar{Y}Q_h$, $D_{hi} = Y_{hi} - \bar{Y}Q_{hi}$ and $\sigma_{dhi}^2 = V\left(\hat{Y}_{hi} - \bar{Y}\hat{Q}_{hi}\right) = M_{hi}^2\ V\left(\bar{y}_{hi} - \bar{Y}\bar{q}_{hi}\right)$.

In case Q is known e.g. from the latest census, and $\hat{\sigma}_{dhi}^2$, an unbiased estimator of σ_{dhi}^2 is available, one could estimate $V\left(\hat{\bar{Y}}\right)$ unbiasedly as

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{Q^2} \left[\frac{1}{2} \sum_{i \neq} \sum_{j \in s_h} \left(\frac{\pi_{i|h}\pi_{j|h} - \pi_{ij|h}}{\pi_{ij|h}} \right) \left(\frac{\hat{D}_{hi}}{\pi_{i|h}} - \frac{\hat{D}_{hj}}{\pi_{j|h}} \right)^2 + \sum_{i \in s} \frac{\hat{\sigma}_{dhi}^2}{\pi_{i|h}} \right] \quad (4)$$

where $\hat{D}_{hi} = \hat{Y}_{hi} - \hat{\bar{Y}}\hat{Q}_{hi}$.

An unbiased variance estimator $\hat{V}\left(\hat{\bar{Y}}\right)$ given in (4) cannot be obtained since unbiased estimator $\hat{\sigma}_{dhi}^2$ doesnot exist since it is based on a systematic sampling scheme[3].

To avoid such complexity, the following traditional variance estimator is proposed, treating the initial sample s_h as PPSWR (probability proportional to size with replacement) method.

$$\hat{V}_{pps} = \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)} \left(\frac{\hat{D}_{hi}}{\pi_{i|h}} - \hat{D}_h \right)^2 \quad (5)$$

The estimator (5) overestimates the variance[8]. Various researchers proposed alternative variance estimators and these are reviewed as follows.

Methods of Variance Estimation

Sediadie: Statistics Botswana method of estimating the standard error for selected indicators and introduced Jackknife re-sampling technique using WesVar statistical software was appraised[9]. The analysis was based on the 2002/3 Household Income and Expenditure Survey data set. Statistics Botswana method estimated the standard error for selected proportions as 0.0000 for each of the three strata (Cities//towns Urban villages and Rural areas). The standard errors for the proportions based on Jackknife statistical method ranged from 0.0008 to 0.0021 at stratum level. At national level, the standard error estimates based on Statistics Botswana method were not zeros but rather larger compared to Jackknife estimates. The lower and upper bounds of the confidence intervals for each indicator were similar at stratum level and wider at national level for Statistics Botswana method. The interval width ranged from 0.00041 to 0.763 for the Statistics Botswana method and from 0.00029 to 0.0439 for the Jackknife. Jacknife method revealed realiable and meaningful estimates of standard errors and confidence intervals compared to Statistics Botswana method.

Arnab et al.: Somenew variance estimators for the above mentioned sampling designs were proposed and compared with the traditional methods used by Statistics Botswana, Botswana, the Random group (RG) and the Jackknife (JK) methods[2]. For comparison of the variance estimators simulation studies were employed using HIES 02/03 survey data for selected six indicators (study variables). Treating Gaborone district as population, a sample of n enumeration areas (EA's) was selected by the PPS systematic sampling procedure. From each of the selected EA's, households were selected by linear systematic sampling method. The proposed variance estimators for the mean per unit $\bar{Y}$ are given as follows:

Traditional method: The conventional estimator of $V\left(\bar{Y}\right)$ is given by

$$\hat{V}_{pps} = \frac{1}{Q^2} \frac{1}{n(n-1)} \frac{1}{2} \sum_{i \neq} \sum_{j \in s} \left(\frac{\hat{Y}_i}{p_i} - \frac{\hat{Y}_j}{p_j} \right)^2 \quad (6)$$

where $\hat{Y}_i = M_i \bar{y}_i$ and $\bar{y}_i$ is the sample means of y (variable under study) for the systematic sample s_i selected from i-th EA.

Use of Wolter method: The following seven consistent estimators for $V\left(\hat{\bar{Y}}\right)$ was proposed as follows:

$$V_k\left(\hat{\bar{Y}}\right) = \frac{1}{Q^2} \left[\frac{1}{2} \sum_{i \neq} \sum_{j \in s} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{\hat{D}_i}{\pi_i} - \frac{\hat{D}_j}{\pi_j} \right)^2 + \sum_{i \in s} \frac{\hat{\sigma}_{di}^2(k)}{\pi_i} \right], \; k = 1,..,7 \quad (7)$$

where π_i and π_{ij} are the inclusion probabilities of the i-th, and i-th and j-th $(i \neq j)$ EA's; Q is the total population of Gaborone district, $\hat{D}_i = M_i\left(\bar{y}_i - \bar{q}_i\right)$, M_i = population size of the ith EA, $\bar{q}_i$ is the sample mean q (household size) -values of the systematic sample s_i selected from the ith EA. $\hat{\sigma}_{di}^2(k)$ is the k-th estimate of the variance of $\hat{D}_i = M_i\left(\bar{y}_i - \bar{q}_i\right)$ [8]. For the detailed expressions of $\hat{\sigma}_{di}^2(k)$ the readers are referred to [2].

Random group method: The sample s of n EA's is divided into G groups at random, of size $n_\alpha = n/G$ each assuming n/G is an integer. Let $\hat{\bar{Y}}_\alpha$ be an estimator of $\bar{Y}$, computed form the α th group s_α.

The proposed estimator of $V\left(\hat{\bar{Y}}\right)$ for the random group method is given by

$$\hat{V}_{RG} = \frac{1}{G(G-1)} \sum_{\alpha=1}^{G} \left(\hat{\bar{Y}}_\alpha - \hat{\bar{Y}}_G \right)^2 \quad (8)$$

where $\hat{\bar{Y}}_G = \frac{1}{G} \sum_{\alpha=1}^{G} \hat{\bar{Y}}_\alpha$

Jackknife method

Let $\hat{\bar{Y}}_{s-j} = \frac{\sum_{i \in s-j} \frac{\hat{Y}_i}{\pi_i}}{\sum_{i \in s-j} \frac{\hat{Q}_i}{\pi_i}} = \frac{\sum_{i \in s} \frac{\hat{Y}_i}{\pi_{i|}} - \frac{\hat{Y}_j}{\pi_j}}{\sum_{i \in s} \frac{\hat{Q}_i}{\pi_i} - \frac{\hat{Q}_j}{\pi_j}}$

with $\hat{Q}_i = M_i \bar{q}_i$ be an estimator of $\bar{Y}$ obtained by deleting jth EA from the sample s.

The Jackknife estimator for $V\left(\hat{\bar{Y}}\right)$ is given by

$$\hat{V}_J = \frac{1}{n(n-1)} \sum_{j=1}^{n} \left(\theta_{(j)} - \bar{\theta} \right)^2 \quad (9)$$

where

$\theta_{(j)} = n\bar{Y} - (n-1)\bar{Y}_{s-j}$ and $\bar{\theta} = \frac{1}{n} \sum_{j=1}^{n} \theta_{(j)}$

The results of the simulation studies indicate that all the variance estimators considered in this study overestimate the variance in general. The variance estimators proposed by Statistics Botswana, conventional Jackknife and Random group variance estimators always overestimate the variance. However, the proposed methods occasionally overestimate the variance. The traditional estimator used by Statistics Botswanais found to have substantially very high bias compared to the Jackknife and Random group methods. The remaining methods yield remarkably lower bias. Except the Random group method, all the proposed variance estimators inclusive Jackknife and Random group methods yield substantial gains in efficiency with respect to the estimator proposed by CSO[10, 11, 12].

Arnab and Arcos: Although the proposed variance estimators $\hat{V}_J$ performs better than the traditional estimator $\hat{V}_{pps}$, its main drawback is that as it needs compute the second order inclusion probabilities mentioned in expression (1) which are very much complex[5]. Further, the Jackknife variance estimation $\hat{V}_J$ overestimates the variance $V\left(\hat{\hat{Y}}\right)$ and the bias is independent of the group size and the second-stage variances σ_i^2's[5]. Therefore an alternative variance estimator and of the total $\hat{Y}$ by adjusting the bias was proposed[5]. The adjusted variance estimator is of the form:

$$\hat{V}^*_{adJ} = \left[\hat{V}_J - \left\{\sum_{i\in s}\frac{\hat{Y}_i^2 - \hat{\sigma}_i^2}{\pi_i} - \frac{2\hat{Y}}{n}\sum_{i\in s}\hat{Y}_i + \hat{Y}^2\left(\frac{1}{n}\sum_{i\in s_h}\frac{1}{p_i}\right)\right\}\right] \Big/ \left\{1 - \left(\frac{1}{n}\sum_{i\in s}\frac{1}{p_i}\right)\right\} \quad (10)$$

The performance of the proposed adjusted variance estimator V^*_{adJ} with respect to $\hat{V}_J$ using HIES 2002/03 data was compared[5]. The data was based on only one stratum (Gaborone district) with 13 EA's. The simulation studies revealed that the proposed adjusted Jackknife estimator possesses the lower bias and higher efficiency than the original Jackknife estimator in almost all the situations. The reduction in biases of the proposed estimator and relative efficiencies vary together. The percentage reduction of biases vary from -1.72% to 26.92% while efficiencies vary from 95.75% to 126.50%.

Alternative Designs

A few sampling designs alternative to the sampling design proposed by Statistics Botswana, as proposed[13]. The performances of the estimators were tested with the live data collected by Statistics Botswana, for HIES 1993/94 survey with six indicators. Gaborone district was considered as the population. A sample of households from Gaborone district was selected by multi-stage sampling designs. The following six sampling strategies were considered:

Strategy I: PPS Systematic (first-stage) and Systematic (second-stage)

Strategy II: PPS Systematic (first-stage) and Simple random sampling without replacement (SRSWOR) (second-stage)

Strategy III: Rao-Hartley-Cochran[6] (first-stage) and Systematic (second-stage)

Strategy IV: Rao-Hartley-Cochran[6] (first-stage) and SRSWOR (second-stage)

Strategy V: SRSWOR (first-stage) and Systematic (second-stage)

Strategy VI: SRSWOR (first-stage) and SRSWOR (second-stage)

The simulation studies revealed that none of the Statistics Botswana, design and the proposed alternatives is found applicable for all the six indicators considered. The investigation further reveals that the design effect which is the ratio of the variance of an estimator assuming the design is SRSWOR to the variance of the original estimator varies between 0.85 to 1.31 for all indicators and strategies. The design effect set by Statistics Botswanawas 2.0, which was too high compared to its actual value which may affect efficiency and cost of the survey. The authors recommended Strategy IV or Strategy VI for the estimation of variance and interval estimation.

Categorical Data Analysis

It is well known that the traditional Pearsonian chi-square tests for goodness of fit, independence and homogeneity are large sample size tests where the samples are selected by simple random sampling with replacement (SRSWR) method. SPSS, BMDP and SAS statistical packages provides chi-square test statistics assuming sample size is large and SRSWR sampling used for selection of the sample. The results (p-values) provided by the above packages are often erroneous for complex survey designs as it is based on sampling designs involving stratification, clustering and unequal sampling designs. On the basis of the empirical studies, it has been shown that that the chi-square tests for complex survey designs may provide type 1 error high as 40% for the target type I error 5%[14]. A modified chi-square test for goodness of fit, independence and homogeneity for the complex survey designs based on BAIS III data was provided[14, 7]. It was shown that the modified chi-square test statistics often provides lower values of the test statistics compared to the traditional chi-square test statistics[7].

Conclusion

Variance estimation plays important role in analyzing survey data such as determination of accuracy of the estimates and interval estimation. For determination of the optimal sample size, design effect and hypothesis testing. Pearson chi-square test for categorical data analysis is essential for goodness of fit, independence and homogeneity tests.

This paper reviewed some methodological aspects of the existing sampling designs adopted by Statistics Botswana (former Central Statistics Office (CSO)) for national household surveys with some recommendations in particular of variance estimation. Furthermore the paper highlighted investigations on Pearson chi square test for categorical data analysis. Some comparison of various approaches have been illustrated with real data from household surveys to come up with better methods and estimates. All the investigations mentioned in this paper are limited to small sample sizes only. Hence there is absolute need of further research using large sample size to find suitable alternative sampling strategies which may provides efficient estimates, unbiased variance estimation as well as appropriate methods of testing hypotheses.

References

1. Goodman R. and Kish L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association,* 45, 350-372.

2. Arnab R., Zewotir T. and North D. (2015). Variance estimation from complex survey designs: A case study of Household Income and Expenditure Survey Design 2002/03, Botswana. *Communications in Statistics-Theory and Methods,* 44, 63-79.

3. Chaudhuri A. and Arnab R. (1982). On Unbiased Variance-Estimation with Various Multi-Stage Sampling Strategies. *Sankhyā,* 44, 92-101.

4. Madow W.G. (1949). On the Theory of Systematic Sampling, II. *The Annals of Mathematical Statistics*, 20, 333-354.

5. Arnab R. and Arcos A. (2016). Jackknife variance estimation from complex survey designs. *Hacettepe Journal of Mathematics and Statistics* (in press).

6. Hartley H.O. and Rao J.N.K. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics*, 33, 350-374.

7. Arnab R., Malambo T. and North D. (2014). Chi-square tests for complex survey designs: A case study of Botswana Aids Impact survey III. *Pakistan Journal of Statistics*, 30, 373-384.

8. Wolter K.M. (1985). Introduction to Variance Estimation. Sprinkler–Verlag, New York.

9. Sediadie T. (2008). Variance estimation techniques in complex survey sampling design Case study: Botswana Household Income and Expenditure Survey (HIES) 2002/03. unpublished MA statistics thesis, University of Botswana.

10. Central Statistical Office (2004). Household Income and Expenditure Survey 2002/03., Republic of Botswana.

11. Central Statistics Central Statistical Office (2005). The Botswana Aids Impact Survey II 2004. Republic of Botswana.

12. Central Statistical Office (2009). The Botswana Family Health Survey 2007. Republic of Botswana.

13. Arnab R. and North D. (2012). An appraisal of Household Income and Expenditure Survey design. *Pakistan Journal of Statistics*, 28, 423-436.

14. Rao J.N.K and Scott A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association,* 76, 221-230.

15. Rao J.N.K., Hartley H.O. and Cochran W.G. (1962). A simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society,* B., 24, 482-491.