TITLE: Jackknife variance estimation from complex survey designs

AUTHORS: Raghunath ARNAB,Antonio ARCOS

PAGES: 909-919

Jackknife variance estimation from complex survey designs

Raghunath Arnab*† and Antonio Arcos‡

Abstract

Large scale surveys very often involve multi-stage sampling design, where the first-stage units are selected with varying probability sampling without replacement method and the second and subsequent stages units are selected with varying or equal probability sampling schemes. It is well known (vide Chaudhuri and Arnab (1982)) that for such sampling designs it impossible to find an unbiased estimator of the variance of the estimator of the population total (or mean) as a homogeneous quadratic function of the estimators of the totals (means) of second-stage units without estimating variances of the estimators of the totals (means) of the second and sub-sequent stages of sampling. Wolter (1985) has shown that the Jackknife estimators of the population total based on unequal probability sampling overestimates the variance. In this paper we have proposed an alternative Jackknife estimator after reduction of bias from the original Jackknife estimator. The performances of the proposed Jackknife estimator and the original estimator are compared through simulation studies using Household Income and Expenditure Survey (HIES) 2002/03 data collected by CSO, Botswana. The simulation studies reveal that the proposed estimator fares better than the original Jackknife estimator in terms of relative bias and mean-square error.

———————

*University of Botswana and University of KwaZulu-Natal, Email: arnabr@mopipi.ub.bw
†Corresponding Author.
‡University of Granada, Spain, Email: arcos@ugr.es

1. Introduction

A sampling design other than the simple random sampling (SRS) is known as a complex sampling design. Complex designs involve clustering, stratification and varying probability sampling (VPS) among others. Most real life surveys are complex surveys and for such surveys we often need to estimate several parametric functions such as the population mean, population ratio of the total of two characteristics, population coefficient of variation, population regression coefficient and population correlation coefficient.

For example, Household Income and Expenditure Survey 2002/03 (HIES 2002/03) conducted by the Central Statistics Office (CSO), Botswana, used enumeration areas (EA's) as first-stage units (fsu's) and households in an EA as second-stage units (ssu's). The EA's are selected by inclusion probability proportional to size (IPPS) sampling design using PPS systematic sampling procedure (Goodman and Kish (1950)) taking the number of households in an enumeration area as measure of size variable while the households in the selected EA's by systematic sampling procedure. The same survey design was used by CSO for Botswana Aids Impact Surveys (BAIS) 2004, 2008, 2012 amongst others.

Variance estimation is essential for estimating the precision of survey estimates, calculation of confidence intervals, determination of optimum sample sizes and for testing of hypotheses, amongst others. In particular, finding the optimum sample size is the key factor in the determination of the cost of a survey and the subsequent precision of estimates.

In a multistage sampling, if the fsu's are selected by without replacement sampling procedure, the variance of the population total (or mean) cannot be estimated unbiasedly as a homogeneous quadratic function of the estimates of the ssu totals only. It requires unbiased estimates of variances of the estimators of the second and sub-sequent sages (vide Chaudhury and Arnab (1982)). For example if the fsu's are selected by IPPS sampling design and ssu's are selected by simple random sampling procedure, then the unbiased estimator of the variance of the estimator of the population total (mean) cannot be estimated unbiasedly as a quadratic function of the sample means of ssu's of the selected fsu's only. It should also involve sample variances of the selected ssu's totals (means) of the second-stage units. To avoid the complexity of the unbiased variance estimation, conventional approximate variance estimators such as Random group (RG), Jackknife (JK), Balanced repeated replications (BRR), Bootstrap (BT) methods are proposed (vide Wolter, 1985). It is well known that for a multi-stage sampling, the RG and JK methods very often overestimate the variance (Vide Wolter, 1985). Singh et al. (1998, 1999, 2011) among others proposed alternative methods of variance estimation for complex survey designs. Arnab et al. (2012, 2015) proposed methods of unbiased estimation of the variances of the population totals for multi-stage sampling designs. Their variance estimators involve unbiased estimators of the variances of the first-stage units as well as unbiased estimators (or approximate estimators) of the variances of the estimators of the second-stage units. The performances of their proposed variance estimators were compared using simulation studies based on the HIES 02/03 data with six indicators (study variables). The simulation studies revealed that variance estimators proposed by Arnab et al. (2015) yielded better estimators in respect to bias and mean square errors than the conventional Jackknife and Random group methods.

Although, the method of variance estimation involving unbiased variance estimation of the first-stage units produces better variance estimators, it requires computation of second order inclusion probabilities which are very complex and difficult to compute for general IPPS sampling designs. The PPS systematic sampling design proposed by Goodman and Kish (1950) is very easy to execute and imposes the least restriction on

p_i (viz. $p_i \leq 1/n$) where p_i ($\sum p_i = 1$) is the normed size measure for the ith unit and n is the sample size. The expression for the second order inclusion probabilities for such sampling scheme was obtained by Hartley and Rao (1962) assuming the units are labelled at random. Such random labelling of units is not possible in practice since the units adjacent to each other are normally labelled by adjacent numbers. So, in this present paper, we have considered Sampford's (1967) IPPS sampling design which is described in Section 2.

Asok and Sukhatme (1976) showed that the variances of the Horvitz-Thompson (1952) estimator

$$(1.1) \qquad \hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

(with $\pi_i = np_i$ = inclusion probability of the ith unit) for a finite population total Y based on a sample s of size n selected by the Goodman-Kish (1950) or Sampford (1967)'s sampling scheme, correct to $O(N^{-1})$ are the same and they are exactly equal to

$$(1.2) \qquad V(\hat{Y}) = \frac{1}{n} \left[\sum_{i=1}^{N} p_i z_i^2 - (n-1) \sum_{i=1}^{N} p_i^2 z_i^2 \right]$$

where $z_i = \dfrac{y_i}{p_i} - Y$.

Expression (1.2) indicates that the Horvitz-Thomson estimator based on the PPS systematic as well as Sampford's procedures possess a uniformly smaller variance than that of the Hansen-Hurwitz estimator based on a PPSWR sampling design of the same sample size n. Furthermore, when the variance is considered to be $O(N^{-2})$, the Horvitz-Thomson estimator based on Sampford sampling (V_{SAM}) has a uniformly smaller variance than that of the PPS systematic sampling procedure V_{GK} and their difference

$$(1.3) \qquad V_{GK} - V_{SAM} = (n-1) \left(\sum_{i=1}^{N} p_i^2 z_i \right)^2$$

is non-negative and increases with the sample size n.

In the present paper, we have proposed an alternative variance estimation formula for multi-stage sampling design where the first-stage units are selected by Sampord's (1967) IPPS sampling design. The proposed variance formula is obtained by removing the bias of the conventional Jackknife variance estimator. The adjusted Jackknife variance estimator is obtained after reduction of bias from the original Jackknife estimator and it is free from second order inclusion probabilities. The performance of the proposed variance estimator is tested by simulation studies using HIES (2003/2004) data with six indicators (study variables). The simulation studies reveal that the proposed variance formula brings enormous gain in efficiency with respect to bias and mean square error. We have also proved that the bias of the conventional Jackknife estimator does not depend on the group size m and the variances of the estimators of the second-stage units σ_i^2's.

2. Variance Estimation from Multi-stage Sampling

Consider a finite population $U = (U_1, \ldots, U_i, \ldots, U_N)$ of N first-stage units (fsu's). The ith fsu U_i consists of M_i second-stage units (ssu's). Let y_{ij} be the value of variable of interest y for the jth ssu of the ith fsu and

$$(2.1) \qquad Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} Y_i$$

be the population total where $Y_i = \sum_{j=1}^{M_i} y_{ij}$ is the ith fus total.

From the population U, a sample s of n fsu's is selected with probability $p(s)$ using Sampford's (1967) IPPS sampling design (described below) with normed size p_i measure attached to the i th unit so that the inclusion probability for the ith unit becomes $\pi_i = np_i$. If the ith fsu is selected in the sample s, a sub-sample s_i of size m_i (pre-determined number) ssu's is selected from it by using some suitable sampling procedure. Each of the selected fsu's are sub-sampled independently.

The Sampford's (1967) sampling design is described as follows:

On the first draw the ith unit is selected with probability $p_i(1) = p_i$. Then the remaining $(n-1)$ fsu's are drawn with replacement from the entire population with probability proportional to $\lambda_i = p_i/(1-np_i)$ attached to the i th unit i.e. the probability of selecting ith unit at kth draw is $p_i(k) = \lambda / \sum_{j=1}^{N} \lambda_j$, $k = 2, \ldots, n$. The selected units are accepted as a sample if all the n units happen to be different, otherwise the entire selection is discarded and this process is repeated unless a set of n distinct units is obtained. Sampford (1967) has shown that the inclusion probability for the selection of ith unit is $\pi_i = np_i$ and $\Delta_{ij} = \pi_i\pi_j - \pi_{ij} \geq 0$. The expression for the second order inclusion probabilities is not simple. However, approximate expression of π_{ij} correct to $O(N^{-4})$, derived by Asok and Sukhatme (1976) is given for $n \geq 3$ as follows:

$$\pi_{ij} = n(n-1)p_ip_j \left[1 + \left(p_i + p_j - \sum_j p_j^2 \right) + \left\{ 2(p_i^2 + p_j^2) - 2\sum_j p_j^3 \right.\right.$$

$$(2.2) \qquad \left.\left. -(n-2)p_ip_j + (n-3)(p_i + p_j)\sum_j p_j^2 - (n-3)(\sum_j p_j^2)^2 \right\} \right]$$

The Horvitz-Thompson (1952) estimator for the population total Y is

$$(2.3) \qquad \hat{Y} = \sum_{i \in s} \frac{\hat{Y}_i}{\pi_i}$$

where $\hat{Y}_i$ is an unbiased estimator of Y_i and $\sum_{i \in s}$ denotes the sum over the distinct units in s.

The variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = \frac{1}{2}\sum_{i \neq}^{N}\sum_{j=1}^{N} \Delta_{ij} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 + \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i}$$

$$(2.4) \qquad = V_{\pi ps} + \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i}$$

where $\sigma_i^2 = V(\hat{Y}_i)$ and

$$(2.5) \qquad V_{\pi ps} = \frac{1}{2}\sum_{i \neq}^{N}\sum_{j=1}^{N} \Delta_{ij} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

An exact unbiased estimator of (2.4) was proposed by Chaudhuri and Arnab (1982) as

$$(2.6) \qquad \hat{V}(\hat{Y}) = \frac{1}{2}\sum_{i \neq}\sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j} \right)^2 + \sum_{i \in s} \frac{\hat{\sigma}_i^2}{\pi_i}$$

where $\hat{\sigma}_i^2$ is an unbiased estimator of σ_i^2.

3. Jackknife Variance Estimation

To use the exact variance estimator $\hat{V}(\hat{Y})$ in practice becomes difficult because of the computation of π_{ij} as well as $\hat{\sigma}_i^2$ for $i, j \in s$. To avoid such tedious computation, one may use the Jackknife variance estimator. The Jackknife method of variance estimation for varying probability sampling was proposed by Wolter (1985) and Särndal et al. (1992). Following Wolter and Särndal et al., in the proposed Jackknife method the first-stage sample s of size n is partitioned at random into k groups each so that the j th group $\tilde{s}_j$ consists of $m = n/k$ (assuming integer) units. The Jackknife estimator of the variance $V(\hat{Y})$ is

$$(3.1) \qquad V_J = \frac{k-1}{k} \sum_{j=1}^{k} \left(\hat{Y}_{-j} - \hat{Y}_\bullet \right)^2$$

where $\hat{Y}_{-j} = \dfrac{n}{n-m} \left(\sum_{i \in s} \dfrac{\hat{Y}_i}{\pi_i} - \sum_{k \in \tilde{s}_j} \dfrac{\hat{Y}_k}{\pi_k} \right)$ is an estimator of Y obtained after deleting

the j th group $\tilde{s}_j$ from the sample s and $\hat{Y}_\bullet = \dfrac{1}{k} \sum_{j=1}^{k} \hat{Y}_{-j}$.

3.1. Theorem.
(i) The bias of V_J is

$$B(V_J) = \frac{n}{n-1} (V_{pps} - V_{\pi ps})$$

where $V_{pps} = \dfrac{1}{n} \sum_{i=1}^{N} p_i \left(\dfrac{Y_i}{p_i} - Y \right)^2$ and $V_{\pi ps}$ is defined in (2.5).

(ii) The estimator V_J overestimates the variance $V(\hat{Y})$ and also independent of the group size m (i.e. k) and the second-stage variances σ_i^2's.

Proof. (i)

$$E(V_J) = \frac{k-1}{k} E\left[\sum_{j=1}^{k} E\left\{\left(\hat{Y}_{-j} - \hat{Y}\right)^2 \mid s\right\}\right]$$

$$= \frac{k-1}{k} E\left[\sum_{j=1}^{k} V\left\{\hat{Y}_{-j} \mid s\right\}\right]$$

$$= \frac{k-1}{k} n^2 \left(\frac{1}{n-m} - \frac{1}{n}\right) \frac{k}{n-1} E\left[\sum_{j \in s} \left(\frac{\hat{Y}_j}{\pi_j}\right)^2 - \frac{\hat{Y}^2}{n}\right]$$

$$= \frac{n}{n-1} E\left[\sum_{j \in s} \left(\frac{\hat{Y}_j}{\pi_j}\right)^2 - \frac{\hat{Y}^2}{n}\right]$$

$$= \frac{n}{n-1} E\left[\sum_{j \in s} \frac{Y_j^2 + \sigma_j^2}{\pi_j^2} - \frac{V_{\pi ps} + Y^2}{n}\right]$$

$$= \frac{n}{n-1} \left[\sum_{i=1}^{N} \frac{Y_i^2 + \sigma_i^2}{\pi_i} - \frac{\frac{1}{2}\sum_{i \neq}^{N}\sum_{j=1}^{N} \Delta_{ij}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i} + Y^2}{n}\right]$$

$$= V(\hat{Y}) + \frac{n}{n-1}\left[\sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} - \frac{Y^2}{n} - V_{\pi ps}\right]$$

$$(3.2) \qquad = V(\hat{Y}) + \frac{n}{n-1}(V_{pps} - V_{\pi ps})$$

Hence the bias of the Jackknife estimator is

$$(3.3) \qquad B(V_J) = \frac{n}{n-1}(V_{pps} - V_{\pi ps})$$

(ii) From expressions (1.3) and (3.3), we note that the magnitude of bias $B(V_J)$ is positive and independent of m (i.e. k) and σ_i^2's. $\square$

4. Proposed Variance Estimator

Following Asok and Sukhatme (1976), we can approximate the variance $V_{\pi ps}$ up to order $O(N^{-2})$ as

$$(4.1) \qquad V_{\pi ps} \simeq \frac{1}{n}\left[\sum_{i=1}^{N} p_i \left(\frac{Y_i}{p_i} - Y\right)^2 - (n-1)\sum_{i=1}^{N} p_i^2 \left(\frac{Y_i}{p_i} - Y\right)^2\right]$$

Now substituting (4.1), in the expression of bias (3.3), we find an approximate expression of bias as

$$B(V_J) \simeq \sum_{i=1}^{N} p_i^2 \left(\frac{Y_i}{p_i} - Y\right)^2$$

$$(4.2) \qquad = \sum_{i=1}^{N} Y_i^2 - 2Y\sum_{i=1}^{N} Y_i p_i + Y^2 \sum_{i=1}^{N} p_i^2$$

Let V_{adJ} be an improved adjusted estimator of the variance of $V(\hat{Y})$. Then an approximate of unbiased estimator of $B(V_J)$ as

$$(4.3) \qquad \hat{B}(V_J) = \sum_{i \in s} \frac{\hat{Y}_i^2 - \hat{\sigma}_i^2}{\pi_i} - \frac{2\hat{Y}}{n} \sum_{i \in s} \hat{Y}_i + (\hat{Y}^2 - V_{adJ}) \left(\sum_{i=1}^{N} p_i^2 \right)$$

The proposed adjusted Jackknife variance estimator of $V(\hat{Y})$ is obtained as a solution of the following equation

$$
\begin{aligned}
V_{adJ} &= V_J - \hat{B}(V_J) \\
(4.4) \qquad &= V_J - \left[\sum_{i \in s} \frac{\hat{Y}_i^2 - \hat{\sigma}_i^2}{\pi_i} - \frac{2\hat{Y}}{n} \sum_{i \in s} \hat{Y}_i + (\hat{Y}^2 - V_{adJ}) \left(\sum_{i=1}^{N} p_i^2 \right) \right]
\end{aligned}
$$

The equation (4.4) yields

$$
\begin{aligned}
V_{adJ} &= \left[V_J - \left\{ \sum_{i \in s} \frac{\hat{Y}_i^2 - \hat{\sigma}_i^2}{\pi_i} - \frac{2\hat{Y}}{n} \sum_{i \in s} \hat{Y}_i + \hat{Y}^2 \left(\sum_{i=1}^{N} p_i^2 \right) \right\} \right] / \\
(4.5) \qquad &\quad \left\{ 1 - \left(\sum_{i=1}^{N} p_i^2 \right) \right\}
\end{aligned}
$$

Now replacing $\left(\sum_{i=1}^{N} p_i^2 \right)$ by it's unbiased estimate $\sum_{i \in s} \frac{p_i^2}{\pi_i} = \frac{1}{n} \sum_{i \in s} \frac{1}{p_i}$, the expression (4.5), we find an improved jackknife estimator (vide Särndal et al.(1992)) as

$$
\begin{aligned}
V_{adJ}^* &= \left[V_J - \left\{ \sum_{i \in s} \frac{\hat{Y}_i^2 - \hat{\sigma}_i^2}{\pi_i} - \frac{2\hat{Y}}{n} \sum_{i \in s} \hat{Y}_i + \hat{Y}^2 \left(\frac{1}{n} \sum_{i \in s} \frac{1}{p_i} \right) \right\} \right] / \\
(4.6) \qquad &\quad \left\{ 1 - \left(\frac{1}{n} \sum_{i \in s} \frac{1}{p_i} \right) \right\}
\end{aligned}
$$

5. Relative Efficiency

Here we compare the performance of the proposed variance estimator V_{adJ}^* with respect to V_J through simulation studies. For the simulation study, we consider the stratum "Gaborone district" of HIES 2002/03 survey as population. The Gaborone district comprises 13 ($= N$) enumeration areas (EA's) of 120, 132, 140, 120, 120, 112, 64, 72, 96, 120, 100, 90 and 80 households respectively. From the population (Gaborone district) a sample s of EA's is selected by Sampford's (1967) IPPS sampling scheme taking number of households M_i of the ith EA as measure of size variable. The inclusion probability of the ith EA is $\pi_i = np_i$ where $p_i = M_i/M$, $M = \sum_{i \in U} M_i = 1366$. If the ith fsu (EA) U_i is selected in the sample s, a sub-sample s_i of size $m_i = \gamma M_i$ ($\gamma = 0.50, 0.33, 0.10$) households (second-stage units) is selected from it by SRSWOR method. Here m_i's are pre-determined numbers and the subsamples s_i's are selected independently from each of the fsu's (first-stage units). The variance $\sigma_i^2(V(\hat{Y}_i))$ is unbiasedly estimated by $\hat{\sigma}_i^2 = M_i^2 (\frac{1}{m_i} - \frac{1}{M_i}) \frac{1}{m_i - 1} \sum_{j \in s_i} (y_{ij} - \bar{y}_i)^2$, $\bar{y}_i = \frac{1}{m_i} \sum_{æ \in s_i} y_{ij}$.

From the selected households information relating to six different indicators viz. Total consumption, Cash earnings, School meals, Gross Income, Earned Income and Income Tax are collected.

From the selected sample estimates of population total $\hat{Y}$, Jackknife estimator V_J (with $k = n$) and the proposed adjusted Jackknife estimator V_{adJ}^* are obtained. The selection of sample and estimation of the variance estimators from each the selected sample (iteration) are repeated $R = 100,000$ times. Let $\hat{Y}[r]$, $V_J[r]$ and $\hat{V}_{adJ}(1)[r]$ be the value of $\hat{Y}$, V_J and V_{adJ}^* based on the r th iteration, $r = 1, \ldots, R$.

The relative bias of $\hat{Y}$ is obtained as

$$RB(\hat{Y}) = \left[\frac{1}{R} \sum_{r=1}^{R} \hat{Y}[r] - Y \right] / Y$$

The relative biases and mean square errors (MSE) of V_J and V_{adJ}^* are obtained as follows:

Relative bias of $V_J = RB(V_J) = B(V_J)/\overline{V} = \left[\frac{1}{R} \sum_{r=1}^{R} V_J[r] - \overline{V} \right] / \overline{V}$,

Relative bias of $V_{adJ}^* = RB(V_{adJ}^*) = B(V_{adJ}^*)/\overline{V} = \left[\frac{1}{R} \sum_{r=1}^{R} V_{adJ}^*[r] - \overline{V} \right] / \overline{V}$,

MSE of $V_J = M(V_J) = \frac{1}{R} \sum_{r=1}^{R} (V_J[r] - \overline{V})^2$ and

MSE of $V_{adJ}^* = M(V_{adJ}^*) = \frac{1}{R} \sum_{r=1}^{R} (V_{adJ}^*[r] - \overline{V})^2$,

where the true variance $\overline{V}$ is obtained 100,000 simulated samples. The efficiency of the proposed variance estimator V_{adJ}^* with respect to V_J is given by

$$RE = \frac{M(V_J)}{M(V_{adJ}^*)}$$

Table 1

n	γ	Average sample size	$RB(\hat{Y})\%$	$cv(\hat{Y})$	RBR%	PRE%
		Total consumption				
3	0.50	164.536	-0.195	0.172	9.849	111.29
	0.33	110.055	-0.413	0.190	7.392	108.74
	0.10	32.827	-0.047	0.278	-0.915	100.76
4	0.50	219.562	-0.483	0.192	21.033	125.62
	0.33	146.606	0.881	0.203	18.834	120.66
	0.10	43.878	0.121	0.273	3.596	105.90
		Cash earnings				
3	0.50	164.536	-0.171	0.178	9.718	111.35
	0.33	110.055	-0.377	0.196	7.331	108.71
	0.10	32.827	-0.027	0.285	-0.869	100.84
4	0.50	219.562	0.285	0.192	22.114	125.79
	0.33	146.606	1.657	0.204	19.804	120.88
	0.10	43.878	0.864	0.276	4.114	106.15
		School meals				
3	0.50	164.536	0.001	0.169	9.314	111.06
	0.33	110.055	-0.150	0.187	7.255	108.82
	0.10	32.827	0.329	0.266	0.4	103.64
4	0.50	219.562	-0.468	0.175	22.601	126.50
	0.33	146.606	0.890	0.184	19.936	123.24
	0.10	43.878	-0.164	0.247	5.496	110.38
		Gross Income				
3	0.50	164.536	-0.304	0.169	9.588	110.98
	0.33	110.055	-0.569	0.188	7.05	108.19
	0.10	32.827	-0.053	0.278	-1.721	99.15
4	0.50	219.562	-0.509	0.195	19.774	124.58
	0.33	146.606	0.880	0.206	17.457	119.32
	0.10	43.878	0.103	0.278	2.454	103.63
		Earned Income				
3	0.50	164.536	0.108	0.187	9.454	107.63
	0.33	110.055	0.085	0.214	6.54	103.53
	0.10	32.827	1.318	0.323	-1.311	95.98
4	0.50	219.562	-0.239	0.183	26.922	117.71
	0.33	146.606	1.037	0.197	21.583	113.12
	0.10	43.878	0.806	0.294	3.223	99.76
		Income Tax				
3	0.50	164.536	0.077	0.182	9.432	107.36
	0.33	110.055	0.053	0.208	6.473	103.89
	0.10	32.827	1.368	0.313	-1.328	95.75
4	0.50	219.562	-0.301	0.186	25.786	117.67
	0.33	146.606	1.040	0.198	20.616	113.11
	0.10	43.878	0.744	0.289	3.115	99.73

Table 1 gives the average sample size, Percentage relative bias $RB(\hat{Y})\%(= RB(\hat{Y}) \times 100)$ and Coefficient of variation (cv) of $\hat{Y}$, Percentage relative bias reduction $RBR\%(= \{RB(V_J) - RB(V_{adJ}^*)\} \times 100)$ and percentage relative efficiency $PRE = RE \times 100$ of the adjusted Jackknife estimator over the conventional Jackknife estimator. The Table 1 shows 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%. For a given fsu size (n) both the reduction of bias and efficiency increase with the increase of ssu size (m_i) while for a given ssu size (m_i), both the

reduction of biases and efficiencies increase with the fsu size (n). However, there is virtually no reduction of biases and increase of efficiencies when both fsu and ssu sizes are small for some of the indicators e.g. Gross Income $(n = 3, \gamma = 0.1)$, Earned Income $(n = 3, \gamma = 0.1; n = 4, \gamma = 0.1)$ and Income Tax $(n = 3, \gamma = 0.1; n = 4, \gamma = 0.1)$.

6. Conclusions

Multi-stage sampling designs are used extensively in real life surveys. The expression of the exact unbiased variance estimators of the population total (or mean) is complex. So, Jackknife variance estimators are used for computing standard errors of the estimators. Standard errors are used for determination of confidence intervals, testing of hypothesis and determination of the optimal sample size. The jackknife estimators generally overestimate the bias and hence results provide inappropriate inferences. It is proved in this paper that the bias of the Jackknife variance estimators is independent of the group size m and second-stage variances (σ_i^2) of the sampling designs used. An alternative Jackknife variance estimator has been proposed in this paper by eliminating bias of the usual Jackknife estimator. The performances of the proposed Jackknife variance estimators are compared with the existing one using HIES 2002/data. The study reveals that the proposed adjusted variance estimators perform better than the original Jackknife estimators in terms of reduction of bias and enhancing relative efficiencies in almost all the situations. The present study is based on a limited sample size due to small resources. However, a similar study with large sample size will certainly provide more conclusive performance of the proposed variance estimator.

Acknowledgement

References

[1] Asok.C. and Sukhatme, B.V. *On Sampford's procedure of unequal probability sampling without replacement*, J. Amer. Statist. Assoc., **71**, 912-918, 1976.

[2] Arnab, R. and North, D. *An appraisal of household income and expenditure survey design*, Pak. J. Statist., **28**, 423-436, 2012.

[3] Arnab, R., Zewotir, T. and North, D. *Variance estimation from complex survey designs: A case study of household income and expenditure survey design 2002/03, Botswana*, Communi. Statist. Theory. Methods., **44**, 63-79, 2015.

[4] CSO (2005). 2004 Botswana AIDS Impact Survey II

[5] CSO (2009). 2008 Botswana AIDS Impact Survey III

[6] CSO (2016). 2012 Botswana AIDS Impact Survey VI (in print)

[7] Chaudhuri, A. and Arnab, R. *On unbiased variance estimation with various multi-stage sampling strategies*, Sankhyā B, **44**, 92-101, 1982.

[8] CSO Household Income and Expenditure Survey 2002/03, Republic of Botswana, 2004.

[9] Goodman, R. and Kish, L. *Controlled selection-a technique in probability sampling*, J. Amer. Statist. Assoc., **45**, 350-372, 1950.

[10] Hartley, H.O and Rao, J.N.K. *Sampling with unequal probabilities and without replacement*, Ann. Math. Statist., **33**, 350-374, 1962.

[11] Horvitz, D.G. and Thompson, D.J. *A generalization of sampling without replacement from a finite universe*, J. Amer. Statist. Assoc., **47**, 663-685, 1952.

[12] Sampford, M.R. *On sampling without replacement with unequal probability selection*, Biometrika, **67**, 639-650, 1967.

[13] Särndal, C.E., Swensson, B. and Wretman, J. Model Assisted Survry Sampling, Springer-Verlag, New York, 1992.

[14] Singh, H.P., Tailor, R., Singh, S. and Kim, J. M. *Estimation of population variance in successive sampling*, Quality & Quantity, **45**, 477- 494, 2011.

[15] Singh, S., Horn, S. , and Yu, F. *Estimation of variance of the general regression estimator: Higher level calibration approach*, Survey Methodology, **24**, 41-50, 1998.

[16] Singh, S., Horn, S., Choudhuri, S. and Yu, F. *Calibration of the estimators of variance*, Austral. and New Zealand J. Statist. **41**, 199-212, 1999.

[17] Wolter, K. *Introduction to variance estimation*, Springer-Verlag, New York, 1985.