Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
MAT2012-0229

Variance Estimation for Complex Survey Designs

Raghunath Arnab
Department of Statistics
University of Botswana, Gaborone
Botswana
Honorary Research Fellow
University of Kwa-Zulu Natal
South Africa

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209
Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

Printed in Athens, Greece by the Athens Institute for Education and Research.

ISSN **2241-2891**
19/09/2012

An Introduction to
ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. The papers published in the series have not been refereed and are published as they were submitted by the author. The series serves two purposes. First, we want to disseminate the information as fast as possible. Second, by doing so, the authors can receive comments useful to revise their papers before they are considered for publication in one of ATINER's books, following our standard procedures of a blind review.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Arnab, R. (2012) "**Variance Estimation for Complex Survey Designs**" Athens: ATINER'S Conference Paper Series, No: MAT2012-0229.

Variance Estimation for Complex Survey Designs

Raghunath Arnab
Department of Statistics
University of Botswana, Gaborone
Botswana
Honorary Research Fellow
University of Kwa-Zulu Natal
South Africa

Abstract

The Central Statistical Office (CSO), Botswana and Statistics South Africa conduct large scale nationwide surveys on a regular basis. The survey design involves selection of first stage units (fsu's) with inclusion probability proportional to size (IPPS) sampling design introduced by Goodman and Kish (1950) and the second stage units (ssu's) by systematic sampling scheme. Chaudhuri and Arnab (1982) proved that for the sampling design used by CSO, the variance of the population total or mean cannot be estimated unbiasedly. Unbiased variance estimation is required for estimating precision of the survey estimates, confidence interval, optimum sample size and testing of hypothesis amongst others. The optimum sample size is the key factor of determination of cost of a survey and precisions of estimates. The formula used by CSO for the estimation of variance is not appropriate. In this present paper, we have proposed a few alternative methods of variance estimation in systematic and complex survey designs.

Key words: Variance estimation, IPPS sampling, Systematic sampling

Contact Information of Corresponding author:

Introduction

The Central Statistical Office (CSO), Botswana and Statistics South Africa conduct large scale nationwide surveys on a regular basis. Both the organizations use the same survey design recommended by the UNDP. The survey design involves selection of first stage units (fsu) with inclusion probability proportional to size (IPPS) sampling design introduced by Goodman and Kish (1950) and the second stage units (ssu's) by systematic sampling scheme. CSO employed
the same sampling design for "Household Income and Expenditure Surveys (HIES, 2002/03)" and "BAIS II Survey (2004)". Chaudhuri and Arnab (1982) proved that for the sampling design used by CSO, the variance of the population total or mean cannot be estimated unbiasedly. Unbiased variance estimation is required for estimating precision of the survey estimates, confidence interval, optimum sample size and testing of hypothesis amongst others. The optimum sample size is the key factor of determination of cost of a survey and precisions of estimates. The formula used by CSO for the estimation of variance is not appropriate. In this paper, we have proposed a few alternative methods of estimation of variance other than the traditional approximate methods viz. Random grouping (RG), Jackknife (JK), Balanced Repeated Replication (BRR) and Bootstrap method (BT).

Proposed Method

Consider a finite population $\wp$ consisting H strata. The h th stratum $\wp_h =$ $(U_{h1},..,U_{hi},..,U_{hM_h})$ consists of M_h first-stage units (EA's). The ith first-stage unit (fsu) of the hth stratum U_{hi} consists of M_{hi} second stage units (households) $U_{hi1},..,U_{hij},..,U_{hiM_{hi}}$. The jth household of the ith enumeration area U_{hij} consists of M_{hij} individuals (household size). The total number of individuals in the population is $M = \sum_{h=1}^{H}\sum_{i=1}^{M_h}\sum_{j=1}^{M_{hi}} M_{hij}$. The quantity M, is generally unknown since M_{hij}'s are known only for the sampled households. Let y_{hijk} be the value of variable under study y for the unit U_{hijk}, kth member of the household U_{hij}. The expression of the population mean is given by

$$\bar{Y} = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}}\sum_{k=1}^{M_{hij}} y_{hijk}/M \qquad (1)$$

BAIS II & HIES Sampling Design

Under this sampling design, from the stratum h, a sample s_h of size m_h fsu (EA's) is selected by following Goodman and Kish (1950, GK) sampling design with inclusion probability $\pi_{hi}(=m_h p_{hi})$ for the hi th and $\pi_{hi,hj}$ for hi and hj th unit ($hi \neq hj$) of the stratum h. Here p_{hi} is the normed size measure of the ith unit of the hth stratum. If the unit U_{hi} (EA's) is selected in the sample s_h, a sub-sample s_{hi} of size m_{hi} ssu (households) is selected from U_{hi} by a systematic sampling procedure. Finally, all the eligible members of the selected households were interviewed.

Estimation of mean and variance

An unbiased estimator for the population mean $\bar{Y}$ is given by

$$\hat{\bar{Y}} = \sum_{h=1}^{H} \hat{Y}_h / \hat{M}$$

(2)

where

$\hat{Y}_h = \sum_{i\in s_h} \frac{\hat{Y}_{hi}}{\pi_{hi}}$ with $\hat{Y}_{hi} = M_{hi}\bar{y}_{hi}$, $\bar{y}_{hi} = \sum_{j\in s_{hi}} Y_{hij} / m_{hi}$, $Y_{hij} = \sum_{k=1}^{M_{hij}} y_{hijk}$,

$\hat{M} = \sum_{h=1}^{H} \hat{Q}_h$

$\hat{Q}_h = \sum_{i\in s_h} \frac{\hat{Q}_{hi}}{\pi_{hi}}$, $\hat{Q}_{hi} = M_{hi}\bar{q}_{hi}$, $\bar{q}_{hi} = \sum_{j\in s_{hi}} Q_{hij} / m_{hi}$ and $Q_{hij} = M_{hij}$

Here $\hat{\bar{Y}}$ given in (2) is a ratio estimator and it is approximately unbiased for $\bar{Y}$ at least for large sample. Now writing $Q_{hi} = \sum_{j=1}^{M_{hi}} Q_{hij}$, $Q_h = \sum_{i=1}^{M_h} Q_{hi}$, $D_{hij} = Y_{hij} - \pi_A(y)Q_{hij}$, $D_{hi} = Y_{hi} - \pi_A(y)Q_{hi}$, $\hat{D}_{hi} = M_{hi}\bar{d}_{hi}$, $\bar{d}_{hi} = \sum_{i\in s_{hi}} D_{hij} / m_{hi}$ and $\hat{D}_h = \hat{Y}_h - \pi_A(y)\hat{Q}_h$ an approximate expression for the mean square error of $\hat{\bar{Y}}$ is obtained as

$$MSE(\hat{\bar{Y}}) \cong \frac{1}{M^2}\sum_{h=1}^{H} Var(\hat{D}_h)$$

$$= \frac{1}{M^2}\sum_{h=1}^{H}[Var\{\sum_{i\in s_h} \frac{E(\hat{D}_{hi} \mid s_{hi})}{\pi_{hi}}\} + E\{\sum_{i\in s_h} \frac{V(\hat{D}_{hi} \mid s_{hi})}{\pi_{hi}^2}\}]$$

$$= \frac{1}{M^2}\sum_{h=1}^{H}[\{Var(\sum_{i\in s_h}\frac{D_{hi}}{\pi_{hi}})\}+E\{\sum_{i\in s_h}\frac{\delta_{hi}^2(sys)}{\pi_{hi}^2}\}]$$

where

$\delta_{hi}^2(sys)=V(\hat{D}_{hi}\mid s_h)=M_{hi}^2\sigma_{hi}^2(sys)$ and $\sigma_{hi}^2(sys)$ = variance of $\bar{d}_{hi}$ given s_h

Hence

$$MSE(\hat{\bar{Y}})=\frac{1}{M^2}\sum_{h=1}^{H}[\frac{1}{2}\{\sum_{i\neq}^{M_h}\sum_{j=1}^{M_h}(\pi_{hi}\pi_{hj}-\pi_{hi,hj})(\frac{D_{hi}}{\pi_{hi}}-\frac{D_{hj}}{\pi_{hj}})^2\}+\{\sum_{i=1}^{M_{hi}}\frac{\delta_{hi}^2(sys)}{\pi_{hi}}\}]$$

(3)

$$=\frac{1}{M^2}\sum_{h=1}^{H}[\{V_{GK}(\hat{D}_h)\}+\{\sum_{i=1}^{M_{hi}}\frac{\delta_{hi}^2(sys)}{\pi_{hi}}\}]$$

where $V_{GK}(\hat{D}_h)\}=\frac{1}{2}\{\sum_{i\neq}^{M_h}\sum_{j=1}^{M_h}(\pi_{hi}\pi_{hj}-\pi_{hi,hj})(\frac{D_{hi}}{\pi_{hi}}-\frac{D_{hj}}{\pi_{hj}})^2\}$

From Chaudhuri and Arnab (1982), we note that the variance (mean square error) $MSE(\hat{\bar{Y}})$, cannot be estimated unbiasedly since no unbiased estimator of $\delta_{hi}^2(sys)$ exist. Noting $V_{GK}(\hat{D}_h)$ can be estimated uubiasedly by

$$\hat{V}_{GK}(\hat{D}_h)=\frac{1}{2}\sum_{i\neq}\sum_{j\in s_h}\frac{\pi_{hi}\pi_{hj}-\pi_{hi,hj})}{\pi_{hi,hj}}(\frac{D_{hi}}{\pi_{hi}}-\frac{D_{hj}}{\pi_{hj}})^2$$

(4)

Since D_{hi} 's are unknown replacing D_{hi} by its unbiased estimator $\hat{D}_{hi}$ in (4), we find an approximate unbiased estimator of $\hat{V}_{GK}(\hat{D}_h)$ as

$$\hat{V}^*_{GK}(\hat{D}_h)=\frac{1}{2}\sum_{i\neq}\sum_{j\in s_h}\frac{(\pi_{hi}\pi_{hj}-\pi_{hi,hj})}{\pi_{hi,hj}}(\frac{\hat{D}_{hi}}{\pi_{hi}}-\frac{\hat{D}_{hj}}{\pi_{hj}})^2$$

(5)

Since no unbiased estimator of $\delta_{hi}^2(sys)=M_i^2\sigma_{hi}^2(sys)$ is available, we have proposed a few alternative estimators for $\sigma_{hi}^2(sys)$ based on the following section 2.3.

Variance Estimation in Systematic sampling

Let a sample s of size n be selected by linear systematic sampling procedure from a population $U=(1,..,i,..,N)$ of size N. In case $N/n=k$ is an integer, an unbiased estimator of the population mean $\bar{Y}=\sum_{i\in U}y_i/N$ based on a systematic sample $s_r=\{r,r+k,...,r+(n-1)k\}$ is

$$\bar{y}_r = \sum_{i\in s_r} y_i / n$$

(6)

The variance of $\bar{y}_r$ is given by

$$V(\bar{y}_r) = \frac{1}{k}\sum_{r=1}^{k}(\bar{y}_r - \bar{Y})^2$$

(7)

$$= \qquad [1+(n-1)\rho]\sigma_y^2 / n$$

(8)

where $\sigma_y^2 = \sum_{i\in U}(y_i - \bar{Y})^2 / N$ and $\rho = \sum_{r=1}^{k}\sum_{j\neq}\sum_{j'\in s_r}(y_j - \bar{Y})(y_{j'} - \bar{Y}) / \{kn(n-1)\sigma_y^2\} =$ intraclass correlation between pairs of units of the same systematic sample.

The variance $V(\bar{y}_{sys})$ cannot be estimable unbiasedly since the inclusion probabilities for a pair of units belonging to different systematic samples is zero i.e. $\pi_{ij} = 0$ for $i \in s_r$ and $j \in s_{r'}, r \neq r'$. However, the following approximate estimators of $V(\bar{y}_r)$ are suggested by Wolter (1985).

Method 1: Treating systematic sample as a SRSWOR sample, $V(\bar{y}_r)$ is estimated by

$$\hat{V}_1(sys) = (1-f)s_y^2 / n \qquad (9)$$

where $s_y^2 = \sum_{i\in s}(y_i - \bar{y}_{sys})^2 / (n-1)$ and $f = n / N$.

The estimator $\hat{V}_1(sys)$ possesses upward or if ρ greater than $-1/(N-1)$ otherwise it possesses downward bias.

Method 2: Let the sample be $s_r = \{r, r+k, ..., r+(n-1)k\}$ with n be even $2m$ (say). Divide the sample s_r in m groups taking two consecutive units in the same group. Now treating the systematic sample as a stratified sample with 2 units selected from each of the $m = n/2$ strata of size 2k each by SRSWOR method e.g. the first 2k units of the population U correspond to first stratum, next 2k units as second stratum and the last 2k units for the m stratum. The proposed variance estimator is given by

$$\hat{V}_2(sys) = (1-f)\sum_{i=1}^{m}\Delta_{ri}^2 / n^2$$

(10)

where $\Delta_{ri} = y_{r+(2i-1)k} - y_{r+2(i-1)k}$

Method 3 &4: Let us consider difference table based on the systematic sample s_r as follows:

s_r	$y-values$	$\Delta(i)$	$\Delta^2(i)$
r	y_r		
$r+k$	y_{r+k}	$y_{r+k}-y_r=\Delta(1)$	
$r+2k$	y_{r+2k}	$y_{r+2k}-y_{r+k}=\Delta(2)$	$\Delta(2)-\Delta(1)=\Delta^2(1)$
.	.	.	
.	.	.	
$r+(n-1)k$	$y_{r+(n-1)k}$	$y_{r+(n-1)k}-y_{r+(n-2)k}=\Delta(n-1)$	$\Delta(n-1)-\Delta(n-2)=\Delta^2(n-2)$

The variance of $V(\bar{y}_r)$ are estimated using the above difference tables as follows:

$$\hat{V}_3(sys)=\frac{(1-f)}{n}\frac{1}{2(n-1)}\sum_{j=1}^{n-1}\{\Delta(j)\}^2$$

(11)
and

$$\hat{V}_4(sys)=\frac{(1-f)}{n}\frac{1}{6(n-2)}\sum_{j=1}^{n-2}\{\Delta^2(j)\}^2$$

(12)
Similarly variance estimators based on higher order differences are also available in the literature and some of them have been listed by Wolter (1985).

Method 5: Divide the sample s of size n into g systematic sub-samples each of size $n/g=q$ (assuming integer). Let $\bar{y}_\alpha$ mean of α th sub-sample. Then an estimator of $V(\bar{y}_r)$ is given by

$$\hat{V}_5(sys)=\frac{(1-f)}{n}\frac{1}{g(g-1)}\sum_{\alpha=1}^{g}(\bar{y}_\alpha-\bar{y}_r)^2$$

(13)
Method 6: Cochran (1946) proposed the following estimator by estimating correlation between consecutive units of a population as

$$\hat{V}_6(sys)=\begin{cases}c(\hat{\rho})(1-f)s_y^2/n \text{ for } \hat{\rho}>0\\ (1-f)s_y^2/n \text{ for } \hat{\rho}\le 0\end{cases} \qquad (14)$$

where

$$c(\hat{\rho})=1+2/In\hat{\rho}+2/(\hat{\rho}^{-1}-1) \text{ and } \quad \hat{\rho}=\sum_{j=0}^{n-1}(y_{r+(j+1)k}-\bar{y}_{sys})(y_{r+jk}-\bar{y}_{sys})/\{(n-1)s_y^2\}$$

Method 7 & 8: The variance of systematic sample can be estimated unbiasedly form more than one systematic samples. Suppose we select q independent linear systematic samples each of size

n assuming $k(=N/n)$ is an integer and $\bar{y}_\alpha$ be the sample mean based on the systematic sample $\alpha\ (=1,..q)$. The combined estimator $\bar{y}_{sys}=\sum_{\alpha=1}^{q}\bar{y}_\alpha/q$ is unbiased for $\bar{Y}$. The variance of $\bar{y}_{sys}$ can be unbiasedly estimated by

$$\hat{V}_7(\bar{y}_{sys})=\frac{1}{q(q-1)}\sum_{\alpha=1}^{q}(\bar{y}_\alpha-\bar{y}_{sys})^2$$

(15)

In case we select q systematic samples by choosing q random start from 1 to k by SRSWOR method the variance of $\bar{y}_{sys}=\sum_{\alpha=1}^{q}\bar{y}_\alpha/q$ can be estimated unbiasedly by

$$\hat{V}_8(\bar{y}_{sys})=\frac{(1-f)}{q(q-1)}\sum_{\alpha=1}^{q}(\bar{y}_\alpha-\bar{y}_{sys})^2$$

(16)

Proposed estimators for variance:

Let us denote an unbiased estimator of $\sigma^2_{hi}(sys)$ based on the method j (j=1,…,8) given in section 2.3 be $\hat{\sigma}^2_{hi}(sys\,|\,j)$. Then proposed approximate estimators of $MSE(\hat{\bar{Y}})$ is given by

$$\hat{M}(\hat{\bar{Y}}\,|\,j)=\frac{1}{\hat{M}^2}\sum_{h=1}^{H}[\{\hat{V}^*_{GK}(\hat{D}_h)\}+\{\sum_{i=1}^{M_{hi}}\frac{\hat{\delta}^2_{hi}(sys\,|\,j)}{\pi_{hi}}\}] \qquad \text{for} \qquad j=1,..,8$$

(17)

where

$$\hat{V}^*_{GK}(\hat{D}_h)=\frac{1}{2}\sum_{i\neq}\sum_{j\in s_h}\frac{(\pi_{hi}\pi_{hj}-\pi_{hi,hj})}{\pi_{hi,hj}}(\frac{\hat{D}_{hi}}{\pi_{hi}}-\frac{\hat{D}_{hj}}{\pi_{hj}})^2 \qquad \text{and}$$

$\hat{\delta}^2_{hi}(sys\,|\,j)=M_i^2\hat{\delta}^2_{hi}(sys\,|\,j)$

Remark: It is difficult to compare performances of the proposed mean square (variance) estimators theoretically. Hence an empirical comparison of the proposed variance estimators based on HIES and BIASII survey data are undertaken. The findings of the empirical investigations are expected to be published in future publication.

References

Chaudhuri, A. and Arnab, R. (1982). On unbiased variance estimation with various multi-stage sampling strategies. *Sankhya Ser. B*,44, (I) 92-101

Cochran, W.G. (1946). Relative accuracy of systematic and random samples for a certainclass of populations. Ann. Math. Statist.,71, 164-177.

CSO (2004). Household Income and expenditure survey 2002/03 (HIES)

CSO (2005). Botswana aids impact survey II (BIAS II)

Goodman,R. and Kish, L.(1950). Controlled selection-a technique in probability sampling. J. Amer. Statist. Assoc., 45, 350-372

Wolter,K.M.(1985). Introduction to variance estimation. Springer-Verlag.N.Y.