Transactions of the Nigerian Association of Mathematical Physics
Volume 11, (January - June, 2020 Issue), **pp61 – 68**
© Trans. of NAMP

JACKKNIFE APPROACHES FOR IMPROVING DATA QUALITY

Olayiwola O. M[1], Akintunde A.A.[1], Yusuff K. M.[1], Ajibade F.B.[2]and Adekpe D.O.[1]

[1]Department of Statistics, College of Physical Sciences, Federal University of Agriculture, Abeokuta, Nigeria
[2]Department of General Studies, Mathematics units, Petroleum Training Institute, Effurun, Delta State Nigeria.

Abstract

Knowledge of the magnitude of error at each sampling step is necessary to know how to improve data quality. This study derived three jackknife approaches for reduction of sampling error. Jackknife leaving one cluster value out (JK1), Jackknife leaving one stratum out (JK2), and Jackknife within each stratum (JK3). The Statistical properties of these Jackknife approaches were examined and compared. JK1 is the most efficient approach for sampling error reduction followed by JK3 and JK2.

Keywords**:** Jackknife, Data quality, sampling error, Statistical properties

1.0 Introduction

Sampling error is the degree to which the estimate differs from its true value. Knowledge of magnitude of error at each sampling step is necessary to identify how to reduce the error. In [1], Quenouille's technique was applied to study its use in ratio estimation, using $\vartheta = 2$ groups. Optimal choice of ϑ for bias reduction in ratio estimation was studied in [2], and $\vartheta = n$ was shown to be the optimal choice. With stratified sampling design using jackknife method under unequal probability sampling, an attempt was made in [3] to develop an efficient scheme for variance estimation.

A generalized jackknife variance estimator was also proposed and a general definition of a jackknife pseudo value that is applicable for unequal probability sampling and stratification was given. This generalized the jackknife variance estimate so that it applied to any sample design for which the variance of an estimated mean can be estimated (exist) [4]. Therefore, the re-sampling techniques, in the context of sampling survey, has been widely studied and developed to handle stratified multistage sampling and the properties of various forms of the jackknife estimator for this case have been studied theoretically and empirically [5].

In [6], the consistency of Campbell’s generalized jackknife variance estimator was established. The study also compared the performance of Campbell's jackknife in a single stage context with standard single stage jackknife and a modified Campbell's estimator by proposing a simple jackknife variance estimator was achieved in [7]. Berger's estimator was consistent under unistage stratified sampling without replacement. A brief overview of early uses of re-sampling methods in survey sampling, and an appraisal of more recent re-sampling methods for variance estimation and inference for small areas were provided in [2]. While in[8], the problem of approximating the sampling relative error of point estimates derived from large sample surveys on a finite population using stratified random sampling design without replacement was studied. Three jackknife methods and compared it with the plug-in and two bootstrap techniques. The first one (JK1) was considered by removing a sample value at each iteration, the second one (JK2) constructed by removing a stratum at each iteration, and the third estimate (JK3) constructed by considering the variance of $\hat{\theta}$ as a linear combination of variances of statistics constructed at stratum level, and these variances were previously estimated by jackknife in each stratum. The different procedures were examined and compared by extensive simulation study [8]. This work investigated reduction of bias and sampling error in stratified cluster sampling using jackknife approaches.

2.0 Methodology

2.1 Sampling Relative Error

The sampling error of $\hat{\theta}$ can be presented in relative terms, using the variation coefficient of the estimator given by:

$$E_{rel}(\hat{\theta}) = \sqrt{Var\,\frac{\hat{\theta}}{E(\hat{\theta})}} \qquad (1)$$

Supposing a population of *N* clusters of *M* units can be stratified into *L* strata such that $N_1, N_2, ..., N_L$ cluster units. Each stratified cluster units contains M_{N_h} *for* $h = 1, ..., L$ and a sample *n* is selected using stratified random sampling such that each stratified samples n_h

Correspondence Author: Olayiwola O.M., Email: akabosede@gmail.com, Tel: +2348134592275

contain M_{n_h} units. Let Y_{hij} be the value of the characteristics under study for the jth element, $j=1,...,M$, for the ith cluster, $i=1,...,N_h$ in the hth stratum.

2.2 Notations

$$\bar{Y}_{hi.}=\frac{1}{M}\sum_{j=1}^{M}Y_{hij} \qquad (2)$$

$$\bar{\bar{Y}}_{h..}=\frac{1}{N_h}\sum_{i=1}^{N_h}\bar{Y}_{hi.} \qquad (3)$$

$$\bar{\bar{Y}}_{h..}=\frac{1}{N_hM}\sum_{i=1}^{N_h}\sum_{j=1}^{M}Y_{hij} \qquad (4)$$

$$\bar{Y}_{stcl}=\sum_{h=1}^{L}W_h\bar{\bar{Y}}_{h..} \qquad (5)$$

$$\sigma_{ih}^2=\frac{1}{M-1}\sum_{j=1}^{M}\left(Y_{hij}-\bar{Y}_{hi.}\right)^2 \qquad (6)$$

$$\sigma_{wh}^2=\frac{1}{N_h}\sum_{i=1}^{N_h}S_i^2 \qquad (7)$$

$$\sigma_{bh}^2=\frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(\bar{Y}_{hi.}-\bar{\bar{Y}}\right)^2 \qquad (8)$$

The population total

$$t=N\bar{\bar{Y}}=N\sum_{h=1}^{L}\frac{N_h}{N}\bar{Y}_{h.}=\sum_{h=1}^{L}Y_{h..} \qquad (9)$$

for

$$\bar{Y}_{h.}=\frac{1}{N_h}\sum_{i=1}^{N_h}Y_{hi.} \qquad (10)$$

and

$$Y_{h.}=\sum_{i=1}^{N_h}Y_{hi.} \qquad (11)$$

Let $\{y_{hij},h=1,...,L,i=1,...,n_h,j=1,...,M\}$ be a stratified random sample without replacement of *Y*, of size $n=\sum_{h=1}^{L}n_h$, n_h being the sample size within the *h*-th stratum.

Denoting by $F_h=\frac{N_h}{n_h}$ the elevation factors of each stratum, the unbiased estimators of the mean $\bar{Y}$ and the total *t* are obtained as follows.

An estimator of $\bar{\bar{Y}}$ is given by the mean of cluster in based on n_h samples

$$\bar{y}_{stcl}=\sum_{h=1}^{L}W_h\bar{\bar{y}}_{h..}=\sum_{h=1}^{L}\frac{N_h}{N}\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{M}\sum_{j=1}^{M}y_{hij}=\frac{1}{N}\sum_{h=1}^{L}F_hn_h\bar{y}_{h.}=\frac{1}{N}\sum_{h=1}^{L}F_hy_{h.} \qquad (12)$$

$$\bar{\bar{y}}_{h..}=\frac{1}{n_hM}\sum_{i=1}^{N_h}\sum_{j=1}^{M}y_{hij} \qquad (13)$$

An estimate of *t* is given as
The population total

$$\hat{t}=N\bar{y}_{stcl}=N\sum_{h=1}^{L}N_h\bar{y}_{h.}=\sum_{h=1}^{L}F_hn_h\bar{y}_{h..}=\sum_{h=1}^{L}F_hy_{h..} \qquad (14)$$

Where

$$\bar{y}_{h.}=\frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi.} \qquad (15)$$

And

$$Y_{h.}=\sum_{i=1}^{n_h}Y_{hi.} \qquad (16)$$

Note that, by definition, $N_h = n_h$ for the strata self- represented in the sample, and hence the elevation factor of these strata is $F_h = 1$. The unbiasedness properties of $\bar{y}_{stcl}$ and $\hat{t}$ as estimators of $\bar{Y}$ and t, respectively, following from their construction as convex linear combinations of sample means are as follows

$$E(\bar{y}_{stcl}) = E\left(\sum_{h=1}^{L} W_h \frac{1}{n_h M} \sum_{i=1}^{n_h} \sum_{j=1}^{M} y_{hij}\right)$$

$$E(\bar{y}_{stcl}) = E\left(\sum_{h=1}^{L} W_h \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{1}{M} \sum_{j=1}^{M} y_{hij}\right)$$

$$E(\bar{y}_{stcl}) = \sum_{h=1}^{L} W_h \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{1}{M} \sum_{j=1}^{M} E(y_{hij})$$

$$E(\bar{y}_{stcl}) = \sum_{h=1}^{L} W_h \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{1}{M} \sum_{j=1}^{M} \bar{Y}_{h.}$$

$$E(\bar{y}_{stcl}) = \sum_{h=1}^{L} W_h \bar{Y}_{h.} = \bar{Y} \tag{17}$$

And

$$E(\hat{t}) = NE(\bar{y}_{stcl}) = N\bar{\bar{Y}} = t \tag{18}$$

The variances of these estimators are

$$V(\bar{y}_{stcl}) = \frac{1}{N^2} \sum_{h=1}^{L} F_h^2 n_h^2 \frac{N_h - n_h}{N_h n_h} \sigma_{bh}^2$$

$$= \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} \sigma_{bh}^2 \tag{19}$$

And

$$V(\hat{t}) = V(N\bar{y}_{stcl}) = \sum_{h=1}^{L} \frac{N_h - n_h}{N_h n_h} \sigma_{bh}^2 \tag{20}$$

σ_{bh}^2 being the variance of the finite population in the *h*-th stratum, given as

$$\sigma_{bh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(\bar{Y}_{hi.} - \bar{\bar{Y}}\right)^2 \tag{21}$$

And the corresponding sample variances corrected by their degrees of freedom are given as

$$S_{bh}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{N_h} (\bar{y}_{hi.} - \bar{y}_{stcl})^2 \tag{22}$$

Then the variance estimator becomes

$$\hat{V}(\bar{y}_{stcl}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} S_{bh}^2 \tag{23}$$

And

$$\hat{V}(\hat{t}) = \sum_{h=1}^{L} \frac{N_h - n_h}{N_h n_h} S_{bh}^2 \tag{24}$$

The absolute and relative sampling errors of estimator $\bar{y}_{stcl}$ take the form

$$E_{abs} = \sqrt{V(\bar{y}_{stcl})} = \frac{1}{N} \left(\sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} \sigma_{bh}^2\right)^{1/2} = \frac{1}{N} E_{abs}(\hat{t}) \tag{25}$$

$$E_{rel} = \frac{1}{t} \left(\sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} \sigma_{bh}^2\right)^{1/2} = E_{rel}(\hat{t}) \tag{26}$$

Hence the estimate of the variance estimator of these errors is

$$\hat{E}_{abs}(\bar{y}_{stcl}) = \frac{1}{N} \left(\sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} S_{bh}^2\right)^{1/2} = \frac{1}{N} \hat{E}_{abs}(\hat{t}) \tag{27}$$

$$\hat{E}_{abs}(\bar{y}_{stcl}) = \frac{1}{t} \left(\sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} S_{bh}^2\right)^{1/2} = \hat{E}_{rel}(\hat{t}) \tag{28}$$

The rest of the study will focus on the estimation of the relative error of parameter *t*, the population total.

2.3 The Jackknife Method

The Grouped Quenouille-Tukey (QT) Jackknife Method is given as follows:
The sample of size n independent and identically (iid) observations $x_1, x_2, ..., x_n$ is divided into g no-overlapping groups $G_1, G_2, ..., G_g$ each of the size d, assuming that $n = dg$ with the ithjacknife sample $S_{(1)} = (S_1, S_2, ..., S_{(i-d)d}, S_{id+1}, S_{id+2}, ..., S_{dg})$ and ith group $(S_{(i-1)d+1}, S_{(i-1)d+2}, ..., S_{id})$ are deleted in turn and the "delete-group" estimates $\hat{\theta}_i, i = 1,2,..., g$ are computed, where $\hat{\theta}_i$ denotes the estimator of θ based on the sample of size $n - d = d(g-1)$, which are named pseudo estimates.
Quenouille showed that the estimator

$$\hat{\theta}_j = \sum_{i=1}^{g} \frac{\hat{\theta}}{g} \tag{29}$$

Where $\hat{\theta}_k = g\hat{\theta} - (g-1)\hat{\theta}_k$ and these are named pseudo values.

$$\hat{\theta} = \frac{1}{g} \sum_{k=1}^{g} \hat{\theta}_k \tag{30}$$

$\hat{\theta}_j$ can be expressed in terms of the pseudo estimates as

$$\hat{\theta}_j = g\hat{\theta} - \left(\frac{g-1}{g}\right) \sum_{k=1}^{g} \hat{\theta}_k \tag{31}$$

It can also be expressed as,

$$\hat{\theta}_j = \hat{\theta} + (g-1)\left(\hat{\theta} - \hat{\theta}_{(.)}\right) \tag{32}$$

With

$$\hat{\theta}_{(.)} = \frac{1}{g} \sum_{k=1}^{g} \hat{\theta}_k$$

Tukey suggested regarding the $\tilde{\theta}_k$ as iid for general $\hat{\theta}$ and then using

$$V\left(\hat{\theta}\right) = \frac{(n-1)}{n} \sum_{k=1}^{n} \left(\hat{\theta}_k - \hat{\theta}_{(.)}\right) \tag{33}$$

$$\hat{\theta}_{(.)} = \sum_{k=1}^{n} \frac{\hat{\theta}_i}{n}$$

As the "jackknife" variance estimator of $\boldsymbol{\widehat{\theta}_j}$ or $\boldsymbol{\widehat{\theta}}$

In a stratified sampling, the jackknife pseudo-values can be constructed following one of the two possible criteria: either removing a sample value at each iteration or removing a stratum at each iteration. Application of these criteria leads to two different jackknife estimators for the variance of $\bar{y}_{stcl}$ and *t* can be expressed as a linear combination of independent statistics, each one being separately constructed from the subsample of each stratum. Consequently, the variance of $\bar{y}_{stcl}$ can be calculated as a linear combination of variances of $\bar{y}_{stcl}$ statistics constructed at stratum level. If these variances are previously estimated by jackknife in each stratum, then there will be a third way of using the jackknife to approximate the variance. Each of the three jackknife proposals are described more in detail below

2.4 The Cao [8] Jackknife Method

In [8], three jackknife estimators for the variance of $\hat{t}$was proposed. This work extended these estimators to stratified cluster sampling.

2.4.1 Case 1: Jackknife Leaving One Clustered Sample Value Out:

Each jackknife pseudo value is constructed by removing a single data value from the overall sample. The pseudo value obtained when eliminating the *s*-thcluster of the *r*-th stratum, Y_{rs}, takes form

$$\hat{t}_{(rs)}^{(1)} = \sum_{h=1, h\neq r}^{L} F_h y_h + F_h' y_r^{(s)} = \hat{t} + F_r' y_r^{(s)} - F_h y_r \tag{34}$$

Where

$$F_r' = \frac{N_h}{n_r - 1}$$

And

$$y_h^{(s)} = \sum_{j=1,\, j\neq s}^{n_h} y_{ij}$$

$$F_r' y_r^{(s)} - F_h y_r = \frac{N_r}{n_r - 1} \sum_{j=1,\, j\neq s}^{n_h} y_{rj} - \frac{N_r}{n_r}\sum_{j=1}^{n_h} y_{rj}$$

$$= \left(\frac{N_r}{n_r - 1} - \frac{N_r}{n_r}\right)\sum_{j=1,\, j\neq s}^{n_h} y_{rj} - \frac{N_r}{n_r}\sum_{j=1}^{n_h} y_{rs}$$

$$= \frac{N_r}{(n_r - 1)n_r} y_r - \frac{N_r}{n_r}\left(\frac{1}{n_r - 1} + 1\right) y_{rs}$$

$$= \frac{N_r}{n_r - 1}(\bar{y}_r - y_{rs})$$

and hence

$$\hat{t}_{(rs)}^{(1)} = \hat{t} + \frac{N_r}{n_r - 1}(\bar{y}_r - y_{rs}) \tag{35}$$

By averaging all the pseudo values, we obtain

$$\hat{t}_{(.)}^{(1)} = \frac{1}{n}\sum_{h=1}^{L}\sum_{s=1}^{n_r} \hat{t}_{(rs)}^{(1)} = \frac{1}{n}\sum_{h=1}^{L}\sum_{s=1}^{n_r}\left(\hat{t} + \frac{N_r}{n_r - 1}(\bar{y}_r - y_{rs})\right)$$

$$= \hat{t} + \frac{1}{n}\sum_{r=1}^{L}\frac{N_r}{n_r - 1}(n\bar{y}_r - y_{rs}) = \hat{t} \tag{36}$$

Hence, the jackknife estimator of $V(\hat{t})$ is given by

$$\hat{V}_{jack,1}(\hat{t}) = \frac{n-1}{n}\sum_{h=1}^{L}\sum_{s=1}^{n_r}\left(\frac{N_r}{n_r - 1}(\bar{y}_r - y_{rs})\right)^2$$

$$= \frac{n-1}{n}\sum_{h=1}^{L}\frac{n_r N_r^2}{(n_r - 1)^2}\left(\frac{1}{n_r}\sum_{s=1}^{n_r} y_{rs}^2 - \bar{y}_r^2\right)$$

$$= \frac{n-1}{n}\sum_{h=1}^{L}\frac{N_r^2}{n_r - 1} S_r^2 \tag{37}$$

and the first variant of the jackknife estimator for the relative error is

$$\hat{E}_{rel,\, jack,1}(\hat{t}) = \frac{1}{\hat{t}}\left(\frac{n-1}{n}\sum_{h=1}^{L}\frac{N_r^2}{n_r - 1} S_r^2\right)^{1/2} \tag{38}$$

2.4.2 Case 2: Jackknife Leaving One Stratum Out

To calculating each pseudo-value, removing all the cluster(s) of one stratum. Thus, the *r*-th jackknife pseudo-value is based on the original sample without the clusters of stratum *r*, i.e.

$$\hat{t}_{(r)}^{(2)} = \frac{N}{N - N_r}\sum_{i=1, i\neq r}^{L} F_h y_h = \frac{N}{N - N_r}\left(\hat{t} - F_r y_r\right) \tag{39}$$

Now, two variants of the jackknife estimator are introduced by considering different ways of averaging the pseudo-values $\hat{t}_{(r)}^{(2)}$ .First, using a weighted mean, where each pseudo-value is weighted by the population size of the stratum removed in the calculation. Thus, we have

$$\hat{t}_{(r)}^{(2A)} = \sum_{r=1}^{L}\frac{N_r}{N}\hat{t}_{(r)}^{(2)} = \sum_{r=1}^{L}\frac{N_r}{N - N_r}\left(\hat{t} - F_r y_r\right) = \hat{t}\sum_{r=1}^{L}\frac{N_r}{N - N_r} - \sum_{r=1}^{L}\frac{N_r}{N - N_r} F_r y_r \tag{40}$$

Then, the jackknife estimator of $V(\hat{t})$ takes the form

$$\hat{V}_{Jack,2A}(\hat{t}) = \sum_{r=1}^{L}\frac{N_r(N - N_r)}{N^2}\left(\hat{t}_{(r)}^{(2)} - \hat{t}_{(.)}^{(2A)}\right)$$

$$= \sum_{r=1}^{L}\frac{N_r(N - N_r)}{N^2}\left(\frac{N\left(\hat{t} - F_r y_r\right)}{N - N_r} - \sum_{h=1}^{L}\frac{N\left(\hat{t} + F_h y_h\right)}{N - N_r}\right)^2 \tag{41}$$

The jackknife estimator of the relative error is then calculated as

$$\hat{E}_{rel,Jack,2A}(\hat{t}) = \frac{1}{\hat{t}}\left(\hat{V}_{Jack,2A}(\hat{t})\right)^{\frac{1}{2}} \tag{42}$$

An alternative variant of the jackknife leaving a stratum out is obtained if all the strata contribute with the same weight in the estimation, i.e. the pseudo-values are directly averaged as follows

$$\hat{t}_{(.)}^{(2B)} = \frac{1}{L}\sum_{r=1}^{L}\hat{t}_{(r)}^{(2)} = \frac{N}{L}\sum_{r=1}^{L}\frac{N_r}{N-N_r}\left(\hat{t}-F_r y_r\right) = \frac{N}{L}\left(\hat{t}\sum_{r=1}^{L}\frac{1}{N-N_r}-\sum_{r=1}^{L}\frac{F_r y_r}{N-N_r}\right) \tag{43}$$

Then using $\hat{t}_{(.)}^{(2B)}$ the jackknife estimator of $V(\hat{t})$ becomes

$$\hat{V}_{Jack,2B}\left(\hat{t}\right) = \frac{L-1}{L}\sum_{r=1}^{L}\left(\hat{t}_{(r)}^{(2)}-\hat{t}_{(.)}^{(2B)}\right)^2 = \frac{L-1}{L}\sum_{r=1}^{L}\left(\frac{N\left(\hat{t}-F_r y_r\right)}{N-N_r}-\frac{N}{L}\sum_{h=1}^{L}\frac{\left(\hat{t}+F_h y_h\right)}{N-N_h}\right)^2 \tag{44}$$

The jackknife estimator of the relative error with this criterion is

$$\hat{E}_{rel,Jack,2B}\left(\hat{t}\right) = \frac{1}{\hat{t}}\left(\hat{V}_{Jack,2B}\left(\hat{t}\right)\right)^{\frac{1}{2}} \tag{45}$$

2.4.3 Jackknife Within Each Stratum

The population variance can be expressed as a linear combination of variances of the sample means within each stratum

$$V(\hat{t}) = \sum_{h=1}^{L} N_h^2\, V(\bar{y}_h) \tag{46}$$

Hence, a new jackknife approximation to the variance of *t* can be obtained by estimating each $V(\bar{y}_h)$ with the jackknife method and replacing these estimators in (46). For the jackknife estimator of $V(\bar{y}_h)$, the pseudo-values are defined as

$$u_h^{(s)} = \bar{y}_{h.}^{(s)} = \frac{1}{n_h-1}\sum_{j=1,\, j\neq s}^{n_h} y_{ij} = \frac{1}{n_h-1}\, y_{h.}^{(s)} \tag{47}$$

For *s* =1, 2,…,*n*h, and their mean is given by

$$u_h^{(s)} = \frac{1}{n_h}\sum_{s=1}^{n_h} u_h^{(s)} = \frac{1}{n_h\left(n_h-1\right)}\sum_{s=1}^{n_h}\sum_{j=1,\, j\neq s}^{n_h} y_{ij} = \frac{1}{n_h\left(n_h-1\right)}\sum_{j=1}^{n_h}\left(n_h-1\right)y_{ij} = \bar{y}_{h.} \tag{48}$$

Then, the jackknife estimator of the variance of the sample mean of the *h*-th stratum is

$$\hat{V}_{JACK}\left(\bar{y}_h\right) = \frac{n_h-1}{n_h}\sum_{s=1}^{n_h}\left(u_h^{(s)}-u_h^{(.)}\right)^2$$

$$= \frac{n_h-1}{n_h}\sum_{s=1}^{n_h}\left(\frac{y_{hj}}{\left(n_h-1\right)n_h}-\frac{y_{hs}}{n_h}\right)^2$$

$$= \frac{n_h-1}{n_h}\sum_{s=1}^{n_h}\left(\sum_{j=1}^{n_h}\frac{y_{hj}}{\left(n_h-1\right)n_h}-\frac{y_{hs}}{n_h-1}\right)^2$$

$$= \frac{n_h-1}{n_h}\sum_{s=1}^{n_h}\left(\sum_{j=1}^{n_h}\left(\frac{y_{hj}}{\left(n_h-1\right)n_h}\right)^2+\left(\frac{y_{hs}}{n_h-1}\right)^2\right)-\frac{2y_{hj}}{n_h-1}\sum_{j=1}^{n_h}\left(\frac{y_{hj}}{\left(n_h-1\right)n_h}\right)$$

$$= \frac{1}{n_h\left(n_h-1\right)}\sum_{s=1}^{n_h} y_{hs}^2-\frac{1}{\left(n_h-1\right)n_h^2}\left(\sum_{j=1}^{n_h} y_{hj}\right)^2$$

$$= \frac{1}{\left(n_h-1\right)}\left(\frac{1}{n_h}\sum_{j=1}^{n_h} y_{hj}^2-\bar{y}_{h.}^2\right)$$

$$\frac{1}{\left(n_h-1\right)n_h}\left(\sum_{j=1}^{n_h}\left(y_{hj}-\bar{y}_{h.}\right)^2\right) = \frac{S_h^2}{n_h} \tag{49}$$

and using these previous jackknife estimations we obtain

$$\hat{V}_{JACK,3}\left(\hat{t}\right) = \sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n_h} \tag{50}$$

The corresponding jackknife estimation of the relative error is given by

$$\hat{E}_{rel,JACK,3}\left(\hat{t}\right) = \frac{1}{\hat{t}}\left(\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n_h}\right)^{\frac{1}{2}} \tag{51}$$

3.0 Results

The data considered were on weights of junior secondary school students of Egba-Odeda High school, Odeda Local Government, Ogun State. Gender was considered as stratifying factor with class as cluster of students. The data was checked and arranged with the use of the statistical packages for social sciences (SPSS) and R.

The descriptive statistics shows the distribution of students according to gender for the classes.

Table 1: Distribution of students' average weight by gender

	N	Minimum	Maximum	Mean
JSS 2A Male Students	46	30	60	44.24
JSS 2B Male Students	51	30	70	43.14
JSS 3A Male Students	65	30	55	42.72
JSS 3B Male Students	59	30	55	41.96
JSS 2A Female Students	50	30	60	44.13
JSS 2B Female Students	57	30	55	44.41
JSS 3A Female Students	49	30	63	44.96
JSS 3B Female Students	51	30	60	43.53

3.1 Jackknife Estimator for Stratified Cluster Sampling Analysis

There are two strata (male and female), there are eight clusters in all, 4 clusters in each stratum. Three clusters were formed and a sample out of the three clusters was chosen at random from each stratum to have stratified clusters of size 6 clusters. The Jackknife mechanism for mean was computed for the clusters and the Jackknife estimate, variance and relative error for the total were computed.

N = 8, N_1 = 4, N_2 = 4, n_1 = 3, n_2 = 3, g = n

$$\hat{t} = N\bar{y}_{stcl} = N\sum_{h=1}^{L} N_h \bar{y}_{h.} = \sum_{h=1}^{L} F_h n_h \bar{y}_{h..} = \sum_{h=1}^{L} F_h y_{h..} \quad = 22{,}564$$

$$\hat{V}(\hat{t}) = \sum_{h=1}^{L} \frac{N_h - n_h}{N_h n_h} S_{bh}^2 = 649946.4$$

Relative error=0.033

Table 2: Jackknife variance Analysis

Clusters	$\bar{y}$ m	$\bar{y}$ f	$\bar{y}$ mJK	$\bar{y}$ Fjk
CL 1	2035	2030	2101.667	2184.333
CL 2	2200	2265	2046.667	2106.000
CL 3	1965	2068	2125.000	2171.667
CL 4	2140	2220	2066.667	2121.000

3.2 Jackknife leaving one cluster value out

$$\hat{t}_{(.)}^{(1)} = \frac{1}{n}\sum_{h=1}^{L}\sum_{s=1}^{n_r}\hat{t}_{(rs)}^{(1)} = \frac{1}{n}\sum_{h=1}^{L}\sum_{s=1}^{n_r}\left(\hat{t} + \frac{N_r}{n_r - 1}(\bar{y}_r - y_{rs})\right)$$

$$= \hat{t} + \frac{1}{n}\sum_{r=1}^{L}\frac{N_r}{n_r - 1}(n\bar{y}_r - y_{rs}) = \hat{t} \quad = 21{,}758$$

Hence, the jackknife estimator of $V(\hat{t})$ is given by

$$\hat{V}_{jack,1}(\hat{t}) = \frac{n-1}{n}\sum_{h=1}^{L}\sum_{s=1}^{n_r}\left(\frac{N_r}{n_r - 1}(\bar{y}_r - y_{rs})\right)^2$$

$$= \frac{n-1}{n}\sum_{h=1}^{L}\frac{N_r^2}{n_r - 1}S_r^2 = 40{,}181.35$$

and the first variant of the jackknife estimator for the relative error is

$$\hat{E}_{rel,jack,1}(\hat{t}) = \frac{1}{\hat{t}}\left(\frac{n-1}{n}\sum_{h=1}^{L}\frac{N_r^2}{n_r - 1}S_r^2\right)^{1/2} \quad = 0.008884$$

3.3 Jackknife Leaving One Stratum Out

$$\hat{t}_{(r)}^{(2A)} = \sum_{r=1}^{L}\frac{N_r}{N}\hat{t}_{(r)}^{(2)} = \sum_{r=1}^{L}\frac{N_r}{N - N_r}(\hat{t} - F_r y_r)$$

$$= \hat{t}\sum_{r=1}^{L}\frac{N_r}{N - N_r} - \sum_{r=1}^{L}\frac{N_r}{N - N_r}F_r y_r = 22{,}880$$

Then, the jackknife estimator of $V(\hat{t})$ takes the form

$$\hat{V}_{Jack,2A}(\hat{t}) = \sum_{r=1}^{L}\frac{N_r(N - N_r)}{N^2}\left(\hat{t}_{(r)}^{(2)} - \hat{t}_{(.)}^{(2A)}\right)$$

$$=\sum_{r=1}^{L}\frac{N_r\left(N-N_r\right)}{N^2}\left(\frac{N\left(\hat{t}-F_r y_r\right)}{N-N_r}-\sum_{h=1}^{L}\frac{N\left(\hat{t}+F_h y_h\right)}{N-N_r}\right)^2 = 501{,}670{,}464$$

The jackknife estimator of the relative error is then calculated as

$$\hat{E}_{rel,Jack,2A}\left(\hat{t}\right)=\frac{1}{\hat{t}}\left(\hat{V}_{Jack,2A}\left(\hat{t}\right)\right)^{\frac{1}{2}} = 1.985$$

An alternative variant of the jackknife leaving a stratum out is obtained if all the strata contribute with the same weight in the estimation, i.e. the pseudo-values are directly averaged as follows

$$\hat{t}_{(.)}^{(2B)}=\frac{1}{L}\sum_{r=1}^{L}\hat{t}_{(r)}^{(2)}=\frac{N}{L}\sum_{r=1}^{L}\frac{N_r}{N-N_r}\left(\hat{t}-F_r y_r\right)$$

$$=\frac{N}{L}\left(\hat{t}\sum_{r=1}^{L}\frac{1}{N-N_r}-\sum_{r=1}^{L}\frac{F_r y_r}{N-N_r}\right) = 11{,}443$$

Then using $\hat{t}_{(.)}^{(2B)}$ the jackknife estimator of $V\left(\hat{t}\right)$ becomes

$$\hat{V}_{Jack,2B}\left(\hat{t}\right)=\frac{L-1}{L}\sum_{r=1}^{L}\left(\hat{t}_{(r)}^{(2)}-\hat{t}_{(.)}^{(2B)}\right)^2$$

$$=\frac{L-1}{L}\sum_{r=1}^{L}\left(\frac{N\left(\hat{t}-F_r y_r\right)}{N-N_r}-\frac{N}{L}\sum_{h=1}^{L}\frac{\left(\hat{t}+F_h y_h\right)}{N-N_h}\right)^2=30{,}714{,}316{,}213$$

The jackknife estimator of the relative error with this criterion is

$$\hat{E}_{rel,Jack,2B}\left(\hat{t}\right)=\frac{1}{\hat{t}}\left(\hat{V}_{Jack,2B}\left(\hat{t}\right)\right)^{\frac{1}{2}} = 10.98$$

3.4 Jackknife Within Each Stratum

$$\hat{V}_{JACK,3}\left(\hat{t}\right)=\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n_h} = 57146.82$$

The corresponding jackknife estimation of the relative error is given by

$$\hat{E}_{rel,JACK,3}\left(\hat{t}\right)=\frac{1}{\hat{t}}\left(\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n_h}\right)^{\frac{1}{2}}=0.01059$$

4.0 Conclusion

This study was designed jackknife approaches for sampling error reduction in stratified cluster sampling. The data considered were on weights of students of Egba-Odeda High School, Odeda Local Government, Ogun State. Gender was considered as stratifying factor with class as cluster of students. Three clusters were formed and a sample out of the three clusters was chosen at random from each stratum to have stratified clusters of size 6 clusters. Three Jackkknife approaches [Jackknife leaving one cluster value out, (JK1), Jackknife leaving one stratum out, (JK2), and Jackknife within each stratum, (JK3)] for sampling error reduction were derived. There are two strata with 428 students in each. There are eight clusters, 4 in each stratum with sizes 46, 51, 65, 59 and 50, 57, 49, 51 respectively for both strata. The mean weights of the male and female students were 42.605 and 44.257 respectively. The mean weights of the male students in JSS 2A, JSS 2B, JSS 3A and JSS 3B were 44.239, 43.137, 42.717 and 41.961 respectively, and for female were 44.13, 44.412, 44.957 and 43.529 respectively. The total weight of students for JK1, JK2, JK3, and existing estimator (EE) were 21758, 22880, 11443, and 22564 respectively, where the variances were 40181.35, 501670, 57146.82 and 649946.4 respectively. The relative error for JK1, JK2, JK3 and EE were 0.0089, 0.0309, 0.01059 and 0.033 respectively. The variance and relative error for the three derived Jackknife approaches are less than that of the existing estimator with JK1 having minimum variance and sampling error. The study revealed that the three jackknife approaches were more efficient than EE with JK1 the most efficient.

References

[1] Durbin, J. (1959). A note on the application of quenouille’s method of bias reduction to the estimation of ratios, *Biometrika* 46,477-480

[2] Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. JISA 3 173-180

[3] Lee, K. (1973). Variance estimation in stratified sampling. *Journal of the American Statistical Association*66, 336- 342

[4] Campbell, C. (1980). A different view of finite population estimation. Proc. Surv. Res.Meth. Sect. Am.Stat. Assoc. 319-324

[5] Jones, H.L. (1974) Jackknife estimation of functions of stratum means. *Biometrika*61,343-348.

[6] Berger, Y.G. and Skinner, C.J. (2005).A jackknife variance estimator for unequal probability sampling. J. Roy. Statist. Soc. Series B 67, 79-89

[7] Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. Biometrika, 94, 953-964.

[8] Cao R., Vilar, J. A, Vilar, J. M. and López, A. K. (2013). Sampling Error Estimation in Stratified Surveys. *Open Journal of Statistics*, 3, 200-212http://dx.doi.org/10.4236/ojs.2013.33023(http://www.scirp.org/ journal/ojs).