Abstract

In survey sampling, the use of auxiliary information can greatly improve the precision of estimates of population total and/or means. Calibration estimation has developed into an important field of research in survey sampling where the auxiliary information plays an important role. In this paper, calibration estimator for cluster sampling with equal clusters have been introduced to improve variance estimator with the aid of auxiliary information and proposed the estimator of variance of calibration estimator. Six distances measures, presented the estimator of the variance of calibration approach estimators are introduced and a simulation study has been conducted to compare between the performance of calibration estimators against Horvitz-Thompson estimator using R statistical package.

Keywords: Auxiliary Information; Calibration; Cluster Sampling: Distance Measures; Estimation of Variance; Equal Clusters.

1. Introduction

Calibration has become a widely used procedure for estimation in sample survey. Deville and Särndal (1992) presented calibration estimators in survey firstly. They used auxiliary information to produce efficient estimates. Calibration requires that knowing population totals for one or more auxiliary variables. The efficiency of the calibration estimator depends on how well auxiliary variables x explain the variability of y, the variable of interest. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources. Calibration can deal effectively with surveys where auxiliary information exists at different levels.

The main aim of this paper is to deal with the case of cluster simple random sampling when only X is known for equal cluster sample. The calibration estimator for the parameters of interest is derived. The rest of this paper is organized as follows: A review of cluster sampling has been introduced in section (2) general Settings for cluster sample in case of equal clusters has been introduced in section (3) HT (Horvitz-Thompson) estimator for cluster sample has been presented in section (4) model of calibration estimation has been introduced in section (5) suggested calibration estimation for cluster sampling using six different distances in section (6) in section (7) suggested estimators of variance have been introduced while a simulation study has been conducted in section(8).

2. Some Review of cluster sample

Bensmail, et al (1997), has been introduced a new approach to cluster analysis based on parsimonious geometric modelling of the within-group covariance matrices in a mixture of multivariate normal distributions, using hierarchical agglomeration and iterative relocation. It works well and it is widely used via MCLUST software that available in S-PLUS and StatLib. However, it has several limitations: there is no assessment of the uncertainty about the classification, the partition can be suboptimal, parameter estimates are biased, the shape matrix has to be specified by the user, prior group probabilities are assumed to be equal, the method for choosing the number of groups is based on a crude approximation, and no formal way of choosing between the various possible models is included. So, they proposed a new approach which overcomes all these difficulties. It consists of exact Bayesian inference via Gibbs sampling, and the calculation of Bayes factors (for choosing the model and the number of groups) from the output using the Laplace-Metropolis estimator. It works well in several real and simulated examples.

Brown and Manly (1998), proposed modification restricted adaptive cluster sampling where a limit is placed on the sample size prior to sample and quadrats are selected sequentially for the initial sample size. As a result there is less variation in the final sample size and the total sampling effort can be predicted with some certainty, which is important for many ecological studies. Estimates of density are biased with the restricted design but under some circumstances the bias can be estimated well by bootstrapping.

Salehi (2003). Compared the properties of two estimators (modified Hansen-Hurwitz (HH) and Horvitz-Thompson (HT)). Some results are obtained in favor of the modified HT estimator so that practitioners are strongly recommended to use HT estimator despite easiness of computations for the HH estimator.

Kennel and Valliant (2010), developed point and variance estimators for totals of finite population characteristics from a clustered sample assisted by a logistic regression model. Using a national Public Use Micro data set they compared the design-based properties of the new estimator to the GREG and the Horvitz-Thompson estimator under two clustered sample designs.

Andrews and McNicholas (2012) began with an introduction to model-based clustering and a succinct account of the state-of-the-art. They put forth novel family of mixture models wherein each component is modeled using a multivariate t - distribution with an eigen-decomposed covariance structure. This family, which is largely a t-analogue of the well-known MCLUST family, is known as the tEIGEN family. The efficacy of this family for clustering, classification, and discriminant analysis is illustrated with both real and simulated data. The performance of this family is compared to its Gaussian counterpart on three real data sets.

Jeelani, et al (2018), had made to cover all the developments towards the cluster sampling for several authors. From the manuscript it can be concluded that the cluster sampling is one of the designs of sampling which has taken the keen interest not only of statisticians but also non-statisticians have contributed a lot in the field. Thus the cluster sampling has better usage rather than other types of sampling design in real life.

Murphy and Murphy (2020) addressed the equivalent aims of including covariates in Gaussian Parsimonious Cluster Models and incorporating parsimonious covariance structures into the Gaussian mixture of experts' framework. The MoEClust models demonstrate significant improvement from both perspectives in applications to univariate and multivariate data sets.

3. General Settings for cluster sample in case of equal clusters

Suppose that there are populations consist of into N clusters and each cluster of size M (M Number of elements in the cluster), a sample of n clusters is drawn by simple random sampling SRS (n number of clusters selected in SRS), n is taken without replacement (WOR). y_{ij} is the value of the characteristic under study for the j^{th} element, ($j = 1,2,3,.........M$) in the i^{th} cluster. Let, total population size $= NM$, total sample size $= nM$, and $\bar{y}_i = \sum_{j=1}^{M} y_{ij} / M$ (Mean of per element of the i^{th} cluster). $\bar{y}_n = \sum_{i=1}^{n} \bar{y}_i / n$ (Mean of cluster means in a sample of n clusters). $\bar{Y}_N = \sum_{i=1}^{N} \bar{y}_i / N$ is the mean of cluster means in the population while $\bar{Y} = \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} / NM$ is the mean per element in the population. For the M cluster S_{CL}^2 is the variance between clusters. Assume x_{ij} denote the value of the auxiliary variable for the j^{th} element, in the i^{th} cluster about which information known at the unit level or at the cluster X is total auxiliary variable. The samples are selected independently across the Clusters.

4. HT (Horvitz-Thompson) Estimator

A sample s of size n cluster is drawn without replacement according to a probabilistic sampling plan $\boldsymbol{P}$ with inclusion probabilities $\pi_{ij} = pr(i \in s)$ where π_i refer to probability to choice cluster $\boldsymbol{i}$ in clustering sample $\pi_{ij} = n / N$, where the number of elements is equal in all clusters. $d_{ij} = 1/\pi_{ij}$ denote the basic design weights.

Deville and Särndal (1992)

Such sampling design is called simple random cluster sampling.

HT (Horvitz-Thompson) estimator for auxiliary variable x_{ij} is:

$$\hat{X}_{CL.HT} = \sum_{j\in s} d_{ij} x_{ij} = \sum_{j\in s} \frac{1}{\pi_{ij}} X_{ij} = \frac{N}{n} \sum_{n\in s} X_{ij}$$

HT (Horvitz-Thompson) estimator for variable of interest y_{ij} variable is:

$$\hat{Y}_{CL.HT} = \sum_{j\in s} d_{ij} y_{ij} = \sum_{j\in s} \frac{1}{\pi_{ij}} Y_{ij} = \frac{N}{n} \sum_{n\in s} Y_{ij}$$

Variance Estimation

Consider the estimator of variance of HT in cluster sampling is given as:

$$\boldsymbol{Var}(\hat{\boldsymbol{y}}_{\boldsymbol{CL.HT1}}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{M}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{M}} \frac{\boldsymbol{\pi_{njk} - \pi_{nj}\pi_k}}{\boldsymbol{\pi_{njk}}} \left(\boldsymbol{d_{nj}}(\boldsymbol{y_{nj}} - \hat{\boldsymbol{B}}\boldsymbol{x_{nj}}) \right) \left(\boldsymbol{d_{nk}}(\boldsymbol{y_{nk}} - \hat{\boldsymbol{B}}\boldsymbol{x_{nk}}) \right) \tag{1}$$

With $\hat{B}$ satisfying the normal equation

$$\left(\sum_{i=1}^{n} \sum_{j=1}^{M} q_{ij} x_{ij} x_{ij}^{'} \right) \hat{B} = \sum_{i=1}^{n} \sum_{j=1}^{M} q_{ij} x_{ij} y_{ij}$$

The variance estimator proposed as:

$$var_{AH.HT2} = W \sum\sum_{j<k\in s} \left[\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k} \right]^2 \tag{2}$$

Where, $W = \frac{N-n}{N(N-1)}$ with n fixed sample size.

5. Model of Calibration Estimator

Deville and Särndal (1992) introduced a calibration estimator of $\hat{Y}_C$, which is constructed as

$$\hat{Y}_C = \sum_{i=1}^{n} w_i y_i, \tag{3}$$

Where, the calibration weights w_i's are chosen to minimize their average distance $\Phi_s = \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i q_i}$, where, q_i's are known positive weights unrelated to d_i. From the basic design weights, $d_i = 1/\pi_i$ that is used in HT estimator $\hat{X}_{HT} = \sum_{i=1}^{n} d_i x_i$, subject to the constraint (calibration equation)

$$\sum_{i=1}^{n} w_i x_i = X \tag{4}$$

Where, X are known as a population totals for auxiliary variable.

The resulting calibration estimator that wants to be proof is:

$$\hat{Y}_C = \sum_{i=1}^{n} w_i y_i = \hat{Y}_{HT} + (X - \hat{X}_{HT})' \hat{B} \tag{5}$$

Where, $\hat{X}_{HT} = \sum_{i=1}^{n} d_i x_i$ and $\hat{B} = [\sum_{i=1}^{n} d_i q_i x_i^T x_i]^{-1} \sum_{i=1}^{n} d_i q_i x_i y_i$.

$T = \sum_{i=1}^{n} d_i q_i x_i{}^t x_i$ Assuming that the inverse of T exist. The uniform weights $q_i = 1$ are used in most applications, but unequal weights can also be motivated as in example 1 of Deville and Särndal (1992).

Proof

Consider the Lagrange function,

$$\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i q_i} - 2.\lambda' (\sum_{i=1}^{n} w_i x_i - X)$$

With partial derivative $\partial / \partial w_i$

$$2(w_i - d_i)/d_i q_i - 2\lambda' x_i$$

Where, λ' is a vector of Lagrange factors. Equating the partial derivative to zero yields:

$$w_i = \lambda' x_i d_i q_i + d_i \tag{6}$$

By substituting from (6) in (4) to get,

$$\lambda' = [X - \sum_{i=1}^{n} d_i x_i] \quad [\sum_{i=1}^{n} d_i q_i x_i' x_i]^{-1}$$

Then.

$$w_i = d_i + [X - \sum_{i=1}^{n} d_i x_i] \quad [\sum_{i=1}^{n} d_i q_i x_i' x_i]^{-1} d_i q_i x_i$$

Therefore,

$$\hat{Y}_C = \sum_{i=1}^{n} d_i y_i + [X - \sum_{i=1}^{n} d_i x_i] \quad [\sum_{i=1}^{n} d_i q_i x_i' x_i]^{-1} \sum_{i=1}^{n} d_i q_i x_i' y_i \tag{7}$$

The estimator $\hat{Y}_C$ in equation (7) is equivalent to Generalized Regression Estimator (GREG) (Cassel, et al 1977; Särndal 1980)

6. The Calibration Estimators for Cluster sampling

1) Generalized Distance (GD)

The calibration estimator is defined as:

$$\hat{Y}_{CCL(GD)} = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}y_{ij} \qquad (8)$$

Where $\hat{Y}_{CCL(GD)}$is the calibration total for cluster sample and w_{ij} are the calibration weights for j^{th} element in the i^{th} cluster, the calibration are chosen to minimize their average distance Φ_s from the basic design weights, $d_{ij} = 1/\pi_{ij}$, that are used in HT estimator $\hat{X}_{HT} = \sum_{i=1}^{n}\sum_{J=1}^{M} d_{ij}x_{ij}$, subject to the constraint $\sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}x_{ij} = X$ (calibration equation)

Proof

Since, Φ_s has the form:

$$\Phi_s = \sum_{i=1}^{n}\sum_{j=1}^{M} \frac{(w_{ij}-d_{ij})^2}{d_{ij}q_{ij}} \qquad (9)$$

Where q_{ij}'s is known as positive weights unrelated to d_{ij}. The uniform weights $q_{ij} = 1$ are used in most applications. And calibration condition

$$\sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}x_{ij} = X \qquad (10)$$

Consider the Lagrange function.

$$L(\Phi_s; w_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{M} \frac{(w_{ij}-d_{ij})^2}{d_{ij}q_{ij}} - 2\lambda'(\sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}x_{ij} - X)$$

$$\frac{\partial L}{\partial w_{ij}} = 2(w_{ij} - d_{ij})/d_{ij}q_{ij} - 2\lambda' x_{ij}$$

So,

$$w_{ij} = d_{ij}(1 + q_{ij}\lambda x_{ij})$$

And as a result,

$$\lambda' = (X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}x_{ij})(\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x'_{ij}x_{ij})^{-1}$$

Therefore,

$$w_{ij} = [d_{ij} + (X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}x_{ij})(\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ii}q_{ii}x'_{ij}x_{ij})^{-1} d_{ij}q_{ij}x_{ij}] \qquad (11)$$

Then by substituting from (11) in (8), $\hat{Y}_{CCL(GD)} = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}y_{ij}$

2) Multiplicative Distance (MD)

The calibration estimator is defined as:

$$\hat{Y}_{CCL(MD)} = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}y_{ij} \qquad (12)$$

Where,

$$\Phi_s = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}\log(\frac{w_{ij}}{d_{ij}q_{ij}}) - w_{ij} + d_{ij}$$

Calibration for cluster sampling by using MD will be derived to get

Proof:

Consider the Lagrange function

$$L(\Phi_s; w_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}\, log(\frac{w_{ij}}{d_{ij}q_{ij}}) - w_{ij} + d_{ij} - \lambda^{'}(\sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}x_{ij} - X)$$

$$\frac{\partial L}{\partial w_{ij}} = w_{ij}\frac{d_{ij}q_{ij}}{w_{ij}}.\frac{1}{d_{ij}q_{ij}} + log\frac{w_{ij}}{d_{ij}q_{ij}} - 1 - \lambda^{'}x_{ij}$$

So,

$$w_{ij} = d_{ij}q_{ij}e^{\lambda^{'}x_{ij}} \tag{13}$$

By substituting from (13) in (10) to get

$$\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}e^{\lambda^{'}x_{ij}}x_{ij} = X \tag{14}$$

Since, $e^{\lambda^{'}x_{ij}} = 1 + \frac{\lambda^{'}x_{ij}}{1!} + \frac{(\lambda^{'}x_{ij})^2}{2!} + \frac{(\lambda^{'}x_{ij})^3}{3!} + \ldots..$

For approximation for linearity of auxiliary variable, the first two terms will be used

$$e^{\lambda^{'}x_{ij}} \approx 1 + \lambda^{'}x_{ij} \tag{15}$$

Then substituting from (15) in (14) to get

$$\lambda^{'} = [X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x_{ij}][\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x^{'}_{ij}x_{ij}]^{-1} \tag{16}$$

Based on (16) w_{ij} will be,

$$w_{ij} = d_{ij}q_{ij}e^{[X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x_{ij}][\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x'_{ij}x_{ij}]^{-1}x_{ij}}$$

By using (15) w_{ij} can be rewritten as:

$$w_{ij} = d_{ij}q_{ij}(1 + [X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x_{ij}][\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x'_{ij}x_{ij}]^{-1}x_{ij}$$

And as a results, (12) can be obtained.

3) Hellinger Distance HD

The calibration estimator is defined as:

$$\hat{Y}_{CCL(HD)} = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}y_{ij} \tag{17}$$

Where,

$$\Phi_s = \sum_{i=1}^{n}\sum_{j=1}^{M}\frac{(\sqrt{w_{ij}}-\sqrt{d_{ij}})^2}{q_{ij}}$$

Calibration for cluster sampling using HD will be derived to get

Proof:

Consider the Lagrange function

$$L(\Phi_s, w_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{M}\frac{(\sqrt{w_{ij}} - \sqrt{d_{ij}})^2}{q_{ij}} - 2\lambda^{'}(\sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}x_{ij} - X)$$

By the same way as above w_{ij} will be,

$$w_{ij} = d_{ij}(1 - \frac{q_{ij}x_{ij}(X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}x_{ij})(\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x^{'}_{ij}x_{ij})^{-1}}{2})^{-2} \tag{18}$$

By applying Binomial Theorem in (18) to get

$$w_{ij} = d_{ij}(1 + q_{ij}x_{ij}\big(X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}x_{ij}\big)\big(\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}x^{'}_{ij}x_{ij}\big)^{-1}\big]$$

And so, (17) can be obtained

4) New Distance 1 (ND1)

The calibration estimator is defined as:

$$\hat{Y}_{CCL(ND1)} = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}y_{ij} \qquad (19)$$

Following El-Sheikh and Mohamed (2013). Assume that Φ_s has the form:

$$\Phi_S = \sum_{i=1}^{n}\sum_{j=1}^{M} \frac{(w_{ij}-d_{ij})^2}{d_{ij}q_{ij}\rho_i s_i{}^2} \qquad (20)$$

Where q_{ij}'s are known positive weights unrelated to d_{ij}, ρ_i is the correlation between x_{ij}, y_{ij} in cluster $\boldsymbol{i}$, which takes the values (ρ_i (< = >) 1, $\rho_i = 1$ is used and s_i^2 is the variance between auxiliary variable for the $\boldsymbol{i}^{th}$ cluster. Calibration using ND1 will be derived as:

Proof:

Consider the Lagrange function

$$L(\Phi_S; w_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{M} \frac{(w_{ij}-d_{ij})^2}{d_{ij}q_{ij}\rho_i s_i{}^2} - 2\lambda'(\sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}x_{ij} - X)$$

By the same way as above,

$$w_{ij} = d_{ij}[1 + (X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}x_{ij})(\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}\rho_i s_i^2 x'_{ij}x_{ij})^{-1} q_{ij}\rho_i s_i^2 x_{ij}] \qquad (21)$$

Then, by substituting from (21 in 19) to get the result.

5) New Distance 2 (ND2)

The calibration estimator is defined as:

$$\hat{Y}_{CCL(ND2)} = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}y_{ij} \qquad (22)$$

Assume that Φ_s has the form:

$$\Phi_S = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}\, log(\frac{w_{ij}}{d_{ij}q_{ij}\rho_i s_i^2}) - w_{ij} + d_{ij} \qquad (23)$$

Calibration using ND2 will be derived as:

Proof:

Consider the Lagrange function

$$L(\Phi_S; w_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}\, log(\frac{w_{ij}}{d_{ij}q_{ij}\rho_i s_i^2}) - w_{ij} + d_{ij} - \lambda'(\sum_{i=1}^{n}\sum_{j=1}^{M} w_{ij}x_{ij} - X)$$

$$\frac{\partial L}{\partial w_{ij}} = w_{ij}\frac{d_{ij}q_{ij}\rho_i s_i^2}{w_{ij}} \cdot \frac{1}{d_{ij}q_{ij}\rho_i s_i^2} + log\frac{w_{ij}}{d_{ij}q_{ij}\rho_i s_i^2} - 1 - \lambda' x_{ij}$$

As before w_{ij} can be obtained as:

$$w_{ij} = d_{ij}q_{ij}(1 + [X - \sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}\rho_i s_i^2 x_{ij}][\sum_{i=1}^{n}\sum_{j=1}^{M} d_{ij}q_{ij}\rho_i s_i^2 x'_{ij}x_{ij}]^{-1} x_{ij} \qquad (24)$$

Then by substituting from (24) to get (22)

6) New Distance 3 (ND3)

The calibration estimator is defined as:

$$\hat{Y}_{CCL(ND3)} = \sum_{i=1}^{n} \sum_{j=1}^{M} w_{ij} y_{ij} \tag{25}$$

In this case, Φ_s has the form:

$$\Phi_s = \sum_{i=1}^{n} \sum_{j=1}^{M} \frac{(\sqrt{w_{ij}} - \sqrt{d_{ij}})^2}{q_{ij} \rho_i s_i^2} \tag{26}$$

Calibration using HD distance can easily be derived (proof is omitted)

7. Variance Estimation

The estimator of variance of combined regression estimator is given by

$$\boldsymbol{Var}(\hat{\boldsymbol{y}}_{CL.HT1}) = \sum_{i=1}^{n} \sum_{j=1}^{M} \sum_{k=1}^{M} \frac{\pi_{njk} - \pi_{nj}\pi_k}{\pi_{njk}} \left(\boldsymbol{d}_{nj} (\boldsymbol{y}_{nj} - \hat{\boldsymbol{B}} \boldsymbol{x}_{nj}) \right) \left(\boldsymbol{d}_{nk} (\boldsymbol{y}_{nk} - \hat{\boldsymbol{B}} \boldsymbol{x}_{nk}) \right) \tag{27}$$

With $\hat{B}$ satisfying the normal equation

$$\left(\sum_{i=1}^{n} \sum_{j=1}^{M} q_{ij} x_{ij} x_{ij}^{'} \right) \hat{B} = \sum_{i=1}^{n} \sum_{j=1}^{M} q_{ij} x_{ij} y_{ij}$$

It is acceptable to use the design weights d_{ij} in the variance estimation but suggested calibration weights w_{ij} be used in Equation (27) as this makes the variance estimator both design consistent and nearly model-unbiased.

$$var(\hat{y}_{CCL1}) = \sum_{i=1}^{n} \sum_{j=1}^{M} \sum_{k=1}^{M} \frac{\pi_{njk} - \pi_{nj}\pi_k}{\pi_{njk}} \left(w_{nj} (y_{nj} - \hat{B} x_{nj}) \right) \left(w_{nk} (y_{nk} - \hat{B} x_{nk}) \right)$$

Moreover, since the calibration estimator is asymptotically equivalent to the GREG estimator, it can be inferred that calibration estimators are more efficient compared to HT estimator if there is a strong correlation between y_{ij} andx_{ij}.

Since, $var\ \hat{y}_{AH.HT2} = W \sum \sum_{j<k\in s} \left[\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k} \right]^2$ (28)

Where, $W = \frac{N-n}{N*(N-1)}$ and with n as a fixed sample size.

So, the calibration weights w_{ij} can be used in Equation (28) to suggest the following variance estimator;

$$var\ \hat{y}_{AH.CL2} = W \sum \sum_{j<k\in s} \left[w_{nj} y_{nj} - w_{nk} y_{nk} \right]^2$$

And as a result, based on the above distances w_{ij} will be change from one to another and so the variance estimator will be changed respectively.

8. Simulation Study

In this section, the performance of the calibration estimator using distance functions against the HT estimator has been tested. So, a Monte Carlo simulation is carried out to investigate the finite sample performance of the estimators of $\hat{Y}_{CCL(GD)}$ proposed in (6), the population consists of $C = 10$ clusters with cluster sizes $M = 5$ elements and samples were drawn using cluster random sampling 2 and 4 clusters. The auxiliary variable, x as generated as a iid random Gamma distribution sample (Scale=1, Shape=0.5). The study variable, y represents a simple regression of the form: $y_{jj} = 2 + x_{ij} + e$, where, e distributed as Lognormal (0, 1).

The performance of the various estimators in cluster simple random sample was measured by the simulated relative efficiency of standard deviation 1 (2), Relative Bias (RB), Mean Square Error (MSE), and the Weight Ratio (WR). The number of iteration $B = 10000$. For the b^{th} run $b = (1,2,..., B)$.

From table (1 and 2), whatever the number of clusters, it can be concluded that: RESD1(2) is equal or close to 1 in most distance (D1,D3,ND1 and ND3); that mean SD are equal for calibration estimator and HT estimator in this distances,RB_C (RB_{HT}) of Bias for calibration (Horvitz-Thompson) estimator are equal 1 in most cases and it can be noted that D3 and ND3 have the least MSE_C (MSE_{HT}) with respect the others.

Table 1: Performance of Different Distance (n=2 clusters)

Distances	RESD1	RESD2	RB_C	RB_{HT}	MSE_C	MSE_{HT}	WR_i
D1	1	1	1	1	13.57	13.57	1
D2	0.0002	0.002	-0.99	3	13.49	121.6	0.001
D3	1	1	1	1	4.52	4.52	1
ND1	1	1	1	1	13.57	13.57	1
ND2	0.0003	0.0003	-0.999	0.6	12.94	116.6	0.003
ND3	1	1	1	1	4.52	4.52	1

Table 2: Performance of Different Distance (n=4 clusters)

Distances	RESD1	RESD2	RB_C	RB_{HT}	MSE_C	MSE_{HT}	WR_i
D1	1	1	1	1	13.57	13.57	1
D2	0.0003	0.0003	-0.999	3	12.92	116.6	0. 003
D3	1	1	3	3	4.52	4.52	1
ND1	1	1	3	3	158.12	158.12	1
ND2	0.0002	0.002	-0.999	3	13.49	121.6	0.001
ND3	1	1	3	3	4.52	4.52	1

Conclusion

In this paper, calibration estimators of finite population total for cluster sampling (with equal clusters) have been derived where six distances measures are introduced and so the weight for each one has been obtained and the estimators have been calculated. Based on that, estimators of the variance have been suggested. Comparing the performance of calibration estimators against the Horvitz-Thompson estimator has been established based on simulation study which show that the suggested calibration estimators performed well in many settings. Also, they have the same efficient as the HT. As was confirmed by the results of Deville and Särndal (1992), the limited version of the HT estimator showed essentially the same behavior as the CAL in terms of Standard Deviation, Bias, Weight Ratio and MSE for each sample size.

References

[1] Andrews, J.L., and Mcnicholas, P.D. (2012). "Model-Based Clustering, Classification and Discriminant Analysis via Mixtures of Multivariate t- Distributions: The tEIGEN Family", Statistics and Computing, 22(5), 1021– 1029.

[2] Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C.P. (1997). Inference in Model-Based Cluster Analysis. Statistics and Computing. 7, 1–10.

[3] Brown, J.A. and Manly, B.J.F. (1998). Restricted Adaptive Cluster Sampling, Environmental and Ecological Statistics 5, 49-63

[4] Cassel, C., Särndal, C. and Wretman, J. H. (1977). Foundations of inference in Survey Sampling. New York: John Wiley & Sons, Inc.

[5] Deville, J.-C., and Särndal, C.-E (1992). Calibration estimators in survey Sampling. Journal of the American Statistical Association, 87, 376-382.

[6] El-Sheikh, A.A. and Mohamed, H.A. (2013). Calibration estimation in stratified Sampling. The 48th Annual Conference on Statistics, Computer Sciences and Operation Research, Institute of Statistical Studies and Research, Cairo University, Egypt.

[7] Kennel, T.L., and Valliant, R. (2010). Logistic generalized regression (LGREG) Estimator in cluster samples, Proceedings of the Section on Survey Research Methods, 4756–4770.

[8] Jeelani M, I. Danish, F. and Gul, M. (2018). A Review on the Recent Development on the Cluster Sampling, Biostatistics and Biometrics. 5 (5), 146-150.

[9] Murphy, k. and Murphy, T.B. (2020). Gaussian parsimonious clustering models With covariates and a noise component. Advances in Data Analysis and Classification, 14, 293–325.

[10] Salehi, M.M. (2003). Comparison between Hansen–Hurwitz and Horvitz–Thompson estimators for adaptive cluster sampling. Environmental and Ecological Statistics, 10, 115–127.

[11] Särndal , C-E. (1980). On π-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling, Biometrika. 67(3), 639-650.