International Journal of Statistics and Applications 2018, 8(5): 259-266
DOI: 10.5923/j.statistics.20180805.04

Bias Reduction Technique for Estimating Finite Population Distribution Function under Simple Random Sampling without Replacement

Winnie Mokeira Onsongo[1,*], Romanus Odhiambo Otieno[2], George Otieno Orwa[2]

[1]Department of Mathematics, Pan African University Institute of Basic Sciences, Technology and Innovation, Nairobi, Kenya
[2]Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Abstract The problem of nonparametric estimation of finite population distribution function using multiplicative bias correction technique is considered in this paper. A robust estimator of the finite population distribution function based on multiplicative bias correction is derived with the aid of a super population model. The properties of the estimator are developed and comparative study with the existing model based and design based estimators is carried to assess the performance of the estimator developed using the simulated sets of data. It is observed that the estimator is asymptotically unbiased and statistically consistent when certain conditions are satisfied. It has been shown that when the model-based estimators are used in estimating the finite population total, there exists bias-variance trade-off along the boundary. The multiplicative bias corrected estimator has recorded better results in estimating the finite population distribution function by correcting the boundary problems associated with existing model based estimators. The simulation results led to the suggestion that the multiplicative bias corrected estimator can be highly recommended in survey sampling estimation of the finite population distribution function.

Keywords α-Quantile, Multiplicative Bias Correction, Pilot Smoother

1. Introduction

In most scenarios of sample survey, auxiliary information is available for all elements in the population under consideration. Auxiliary information aids in the prediction of finite population parameters and as such it forms a central part of sample surveys. The main idea of nonparametric statistics is to make inferences about unknown quantities without resorting to parametric reduction of the problem. It therefore follows that a model–based approach is used to increase the precision of the estimators by incorporating auxiliary variables. As an approach to such a problem, a super population model is used to describe the relationship between the auxiliary variable and the study variable. Various estimation procedures have been developed to estimate the distribution of a random variable in the past (Zhao et al., 2013).

(Chambers and Dunstan, 1986) studied a simple method for estimating the distribution function and the associated quantiles from sample survey data. The study showed that the model based estimator offers significant gains when there exists a strong linear relationship between the survey variable and the auxiliary variable. However, the estimator tends to be positively biased when the true variance is overstated and negatively biased when the true variance is understated. Kuk (1993) used auxiliary information to improve the estimation of population distribution function. Empirical results suggest that the proposed estimator has good robustness properties not enjoyed by the model-based estimator of (Chambers and Dunstan, 1986). In survey sampling, concern is with the proportion of values, say Y_i, in the finite population that are bounded by a given constant. Such a proportion is one particular value of the distribution function for the finite population. In particular, estimation of the distribution function is an important objective mainly because it helps to identify the proportion in the population whose values for particular variables lie substantially below or above the population average (Chambers and Dunstan, 1986).

Previously studied estimation procedures used kernel smoothers which tend to have boundary problems and require modifications at the boundary points. That is, towards the boundary points the estimators exhibit trade-off between the bias and variance of the estimators. However, alternative bias reduction techniques have been formulated. For a detailed review see Hardle (1986), (Muller and

* Corresponding author:
onsongowinnie@gmail.com (Winnie Mokeira Onsongo)
Published online at http://journal.sapub.org/statistics

Stadmuller, 1987) and Fan (1992). This study therefore aims at coming up with a nonparametric estimator for the distribution function of finite populations using a bias corrected technique to counter the shortcomings of the previously studied methods of estimation. (Linton and Nielsen, 1994) used the multiplicative bias correction technique in estimating a nonparametric regression function and the results obtained showed that the estimator of the regression function had desirable properties compared to existing estimators including solving the boundary problems. Onsongo (2018) also used the approach by (Linton and Nielsen, 1994) in estimating finite population total.

Outline of the paper

In section 2, we propose an estimator for finite population distribution function using a bias correction technique. Asymptotic properties of the estimator are derived in section 3. Empirical simulation of the results is given in section 4 and the conclusion of the findings is given in section 5.

2. Proposed Estimator

In this section, the exact procedure of estimating the population distribution function is now presented.

Suppose that $X_1, X_2, \dots, X_N$ are independent and identically distributed with corresponding survey measurements $Y_1, Y_2, \dots, Y_N$ from a common univariate distribution function.

The empirical distribution function for finite population is then defined by

$$F_N(t) = \frac{1}{N}\sum_{i=1}^{N} I(y_i \le t) \quad (1)$$

Where *I* denotes the indicator function of a given set and *t* is the α - quantile.

Let *s* be a sample of *n* units drawn from a finite population via simple random sampling without replacement and $j \in r = p - s$ be the non-sampled units of the finite population. Suppose that *Y* is the survey variable associated with the auxiliary variable *X*. Then the auxiliary information is known for all elements in the population while the survey variable is only observed for the sample elements.

Under the model-based framework, *X* and *Y* are assumed to follow a super population model. This study restricts attention to the linear regression model

$$Y = \mu(x_i) + \sigma(x_i)\varepsilon_i \quad (2)$$

For $i = 1,2, \dots, N$.

Where the ε_i's are independent and identically distributed and $E(Y) = \mu(x_i)$ and

$$Cov(Y_i, Y_j) = \begin{cases} \sigma^2(x_i) for i = 1,2, \dots, N \\ 0 \; otherwise \end{cases}$$

Where $\mu(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth functions of x_i.

Using model (1) as a guide, the predictive form of the proposed estimator of the distribution function under the model based approach is

$$F_N(t) = \frac{1}{N}\left\{\sum_{i\in s} I(y_i \le t) + \sum_{j\in r} I\left(y_j \le t\right)\right\} \quad (3)$$

In this paper, the estimator for equation (3) is proposed as

$$\hat{F}_{MBC}(t) = \frac{1}{N}\left\{\sum_{i\in s} I(y_i \le t) + \sum_{j\in r} \hat{H}\left(t - \hat{\mu}(x_j)\right)\right\} \quad (4)$$

Where $\hat{\mu}(x_j)$ is the model-based nonparametric estimator for $\mu(x_j)$ and $\hat{H}\left(t - \hat{\mu}(x_j)\right)$ is the estimated distribution function of the residuals defined by $e_j = y_j - \hat{\mu}(x_j)$.

The task is to estimate the second part of equation (4) and to do this, the multiplicative bias correction technique is employed.

Suppose that $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ are *N* independent pairs of random variables (X, Y) with real values.

Define a pilot smoother of the regression function as

$$\tilde{\mu}_n(x_i) = \sum_{j=1}^{n} w_i(x; l) y_j \quad (5)$$

Where $w_i(x; l)$ are the Nadaraya-Watson kernel weights defined by $w_i(x; l) = \frac{K_l(x - x_i)}{\sum_{i=1}^{n} K_l(x - x_i)}$ and *l* is the bandwidth.

Then the ratio $\beta_i = \frac{y_i}{\tilde{\mu}_n(X_i)}$ is a noisy estimate of the inverse relative estimation error of the smoother $\tilde{\mu}_n$ given by $\frac{\mu(X)}{\tilde{\mu}_n(X_i)}$.

(Burr et al., 2010) showed that this ratio significantly smoothens out the regression function since the residuals in the numerator will cancel out with the residuals in the denominator.

Smoothing β_i yields

$$\hat{\alpha}(x) = \sum_{i=1}^{n} w_i(x; l)\, \beta_i = \sum_{i=1}^{n} w_i(x; l) \frac{y_i}{\tilde{\mu}_n(X_i)} \quad (6)$$

Equation (6) can then be used as a multiplicative correction of the pilot smoother in equation (5) which can now be defined by

$$\hat{\mu}_n(x_i) = \hat{\alpha}(x_i)\tilde{\mu}_n(x) \quad (7)$$

Assumptions

The following assumptions are made in the estimation of $\mu_n(x_i)$

1. The regression function is twice continuously differentiable everywhere.
2. The bandwidth l is such that $l \to 0, nl \to \infty \; as \; n \to \infty$.

Using equation (6) in equation (7) easily yields

$$\hat{\mu}_n(x_i) = \sum_{i=1}^{n} w_i(x; l) \frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(X_i)} y_i \quad (8)$$

Now suppose that

$$E[\tilde{\mu}_n(x)/x_1, \dots, x_N] = \sum_{i=1}^{n} w_i(x; l)\, E[Y_i] = \sum_{i=1}^{n} w_i(x; l)\, \mu(X_i) = \bar{\mu}_n(x) \quad (9)$$

Then $\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_i)}$ in equation (8) can be expanded as follows

$$\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_i)} = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \times \frac{\tilde{\mu}_n(x)}{\bar{\mu}_n(x)} \times \left(\frac{\tilde{\mu}_n(x_i)}{\bar{\mu}(X_i)}\right)^{-1} = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \times \left(1 + b_n(x)\right) \times \left(1 + b_n(X_i)\right)^{-1} \quad (10)$$

Where $\frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)} = b_n(x)$ and $\frac{\tilde{\mu}_n(x_i)-\bar{\mu}(X_i)}{\bar{\mu}(X_i)} = b_n(X_i)$

Applying the binomial expansion to $\left(1 + b_n(x)\right) \times \left(1 + b_n(X_i)\right)^{-1}$ gives

$\left(1 + b_n(x)\right) \times \left(1 + b_n(X_i)\right)^{-1} = [1 + b_n(x)][1 - b_n(X_i) + b_n(X_i)^2]$ which further reduces to

$$\left(1 + b_n(x)\right) \times \left(1 + b_n(X_i)\right)^{-1} = 1 + b_n(x) - b_n(X_i) + r_i(x, X_i) \quad (11)$$

where $r_i(x, X_i)$ is the remainder term that involves the terms *x* and X_i.

Using equation (11) in equation (10) yields

$$\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_i)} = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \times [1 + b_n(x) - b_n(X_i) + r_i(x, X_i)] \quad (12)$$

Substituting equation (11) into equation (8) and using the model $Y_i = \mu(X_i) + \varepsilon_i$ one obtains

$$\hat{\mu}_n(x_i) = \sum_{i=1}^{n} w_i(x; l) \left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \times [1 + b_n(x) - b_n(X_i) + r_i(x, X_i)]\right\} (\mu(X_i) + \varepsilon_i) \quad (13)$$

$$\Rightarrow \hat{\mu}_n(x_i) = \sum_{i=1}^{n} w_i(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \mu(X_i) + \sum_{i=1}^{n} w_i(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \{\varepsilon_j + \mu(X_i)[b_n(x) - b_n(X_i)]\}$$
$$+ \sum_{i=1}^{n} w_i(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \varepsilon_i [b_n(x) - b_n(X_i)] + \sum_{i=1}^{n} w_i(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} r_i(x, X_i)[\mu(X_i) + \varepsilon_i] \quad (14)$$

Using the assumption $nl \to \infty$, the remainder terms converge to zero in probability. Therefore $r_i(x, X_i)[\mu(X_i) + \varepsilon_i] = O_P\left(\frac{1}{nl}\right)$ and equation (14) reduces to

$$\hat{\mu}_n(x_i) = \sum_{i=1}^{n} w_i(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \mu(X_i) + \sum_{i=1}^{n} w_i(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \{\varepsilon_j + \mu(X_i)[b_n(x) - b_n(X_i)]\}$$
$$+ \sum_{i=1}^{n} w_i(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_i)} \varepsilon_i [b_n(x) - b_n(X_i)] + O_P\left(\frac{1}{nl}\right) \quad (15)$$

Our estimator for the distribution function for finite population therefore becomes

$$\hat{F}_{MBC} = \frac{1}{N}\left\{\sum_{i\in s} I(y_i \le t) + \sum_{j\in r} \hat{H}\left(t - \begin{bmatrix} \sum_{j\in r} w_j(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \mu(X_j) + \\ \sum_{j\in r} w_j(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \{\varepsilon_j + \mu(X_j)[b_n(x) - b_n(X_j)]\} + \\ \sum_{j\in r} w_j(x; l) \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \varepsilon_j [b_n(x) - b_n(X_j)] + O_P\left(\frac{1}{nl}\right) \end{bmatrix}\right)\right\}$$

3. Properties of the Estimator under Simple Random Sampling without Replacement

3.1. Asymptotic Unbiasedness of the Proposed Estimator

The asymptotic bias of the nonparametric estimator is defined as

$$\frac{1}{N} E\left[\hat{F}_{MBC} - F_N(t)\right] \quad (16)$$

where $\hat{F}_{MBC} - F_N(t)$ is the estimated bias.

In order to estimate $\hat{H}(t)$ in equation (4), (Chambers et al., 1993) recommended local linear smoothing whereby $\hat{H}$, is estimated by averaging only over the sample residuals with *X*- values that are close enough to X_j.

Therefore $\hat{H}(t) = \frac{\sum_{i\in s} w_{ij} I(y_i - \hat{\mu}(x_j))}{\sum_{i\in s} w_{ij}}$ where *t* is the α -quantile and w_{ij} are the weights that only take non-zero values for sample units *i* with X_j close to X_i so that

$$\hat{H}(t) = \sum_{i\in s} w_{ij}^* I(y_i - \hat{\mu}(x_j)) \text{ where } w_{ij}^* = \frac{w_{ij}}{\sum_{i\in s} w_{ij}}$$

Therefore equation (4) becomes

$$\hat{F}_{MBC}(t) = \frac{1}{N}\left\{\sum_{i\in s} I(y_i \le t) + \sum_{j\in r} \sum_{i\in s} w_{ij}^* I\left(\hat{e}_j \le t\right)\right\}$$

As a result,

$$E\left[\hat{F}_{MBC}(t)\right] = E\left[\frac{1}{N}\left\{\sum_{i\in s} I(y_i \le t) + \sum_{j\in r} \sum_{i\in s} w_{ij}^* I\left(\hat{e}_j \le t\right)\right\}\right]$$

$$E\left[\hat{F}_{MBC}(t)\right] = \frac{1}{N}\left\{\sum_{i\in s} E[I(y_i \le t)] + \sum_{j\in r}\sum_{i\in s} w_{ij}^{*} E\left[I\left(\hat{e}_j \le t\right)\right]\right\}$$

$$\Rightarrow E\left[\hat{F}_{MBC}(t)\right] = \frac{n}{N}F_y(t) + \frac{1}{N}\sum_{j\in r}\sum_{i\in s} w_{ij}^{*} F_e(t) \quad (17)$$

Next,

$$E[F_N(t)] = E\left[\frac{1}{N}\sum_{i=1}^{N} I(y_i \le t)\right] = F_y(t) \quad (18)$$

Substituting the results in equation (17) and equation (18) back to equation (16) yields

$$\frac{1}{N}E\left[\hat{F}_{MBC} - F_N(t)\right] = \frac{1}{N}\left[\frac{n}{N}F_y(t) + \frac{1}{N}\sum_{j\in r}\sum_{i\in s} w_{ij}^{*} F_e(t) - F_y(t)\right]$$

$$\Rightarrow \frac{1}{N}E\left[\hat{F}_{MBC} - F_N(t)\right] = \frac{1}{N^2}F_e(t) - \left(\frac{N-n}{N^2}\right)F_y(t) \text{ since } \sum_{j\in r}\sum_{i\in s} w_{ij}^{*} = 1$$

3.2. Asymptotic Variance of the Proposed Estimator

Thus $\hat{F}_{MBC}(t)$ is asymptotically unbiased.

The estimated bias is given by

$$\hat{F}_{MBC} - F_N(t) = \frac{1}{N}\sum_{j\in r}\sum_{i\in s} w_{ij}^{*} I(y_i - \hat{\mu}_i \le t - \hat{\mu}_i) - \frac{1}{N}\sum_{j\in r} I\left(y_j \le t\right)$$

Therefore the variance of the estimated bias is

$$Var\left[\hat{F}_{MBC} - F_N(t)\right] = Var\left[\frac{1}{N}\sum_{j\in r}\sum_{i\in s} w_{ij}^{*} I(y_i - \hat{\mu}_i \le t - \hat{\mu}_i) - \frac{1}{N}\sum_{j\in r} I\left(y_j \le t\right)\right]$$

$$Var\left[\hat{F}_{MBC} - F_N(t)\right] = \frac{1}{N^2}\left\{Var\left(\sum_{j\in r}\sum_{i\in s} w_{ij}^{*} I(y_i - \hat{\mu}_i \le t - \hat{\mu}_i)\right) + Var\left(\sum_{j\in r} I\left(y_j \le t\right)\right)\right\} \quad (19)$$

Since the errors are assumed to be independent and identically distributed and therefore have zero covariance.

Consider $Var\left(\sum_{j\in r}\sum_{i\in s} w_{ij}^{*} I\left(y_i - \hat{\mu}_i \le t - \hat{\mu}_j\right)\right)$ and let $\hat{\Psi}_{\mathrm{j}}(t) = \sum_{i\in s} w_{ij}^{*} I\left(y_i - \hat{\mu}_i \le t - \hat{\mu}_j\right)$

Then

$$Var\left(\sum_{j\in r}\sum_{i\in s} w_{ij}^{*} I\left(y_i - \hat{\mu}_i \le t - \hat{\mu}_j\right)\right) = Var\left(\sum_{j\in r}\hat{\Psi}_{\mathrm{j}}(t)\right) = \sum_{j\in r}\sum_{\mathrm{k}\in \mathrm{r}}\mathrm{Cov}(\hat{\Psi}_{\mathrm{j}}(t), \hat{\Psi}_{\mathrm{k}}(t)) \quad (20)$$

With $\hat{\Psi}_{\mathrm{k}}(t) = \sum_{i\in s} w_{ik}^{*} I(y_i - \hat{\mu}_i \le t - \hat{\mu}_k)$

Define $H_i(u) = P(y_i - \mu_i \le u) = P(e_i \le u)$

$$Cov\left(\hat{\Psi}_{\mathrm{j}}(t), \hat{\Psi}_{\mathrm{k}}(t)\right) = \sum_{i\in s} w_{ij}^{*} w_{ik}^{*}\left[H_i\left(t - max\left(\hat{\mu}_j, \hat{\mu}_k\right)\right) - H_i\left(t - \hat{\mu}_j\right)H_i\left(t - \hat{\mu}_k\right)\right] \quad (21)$$

Suppose that $\hat{\mu}_j < \hat{\mu}_k$ whenever $j < k$ and suppose that the non-sampled units are labelled from 1 to $N - n$.

Then

$$Var\left(\sum_{j\in r}\sum_{i\in s} w_{ij}^{*} I\left(y_i - \hat{\mu}_i \le t - \hat{\mu}_j\right)\right) = \sum_{i\in s}\left\{\sum_{j=1}^{N-n}\sum_{k=1}^{N-n} w_{ij}^{*} w_{ik}^{*}\left[H_i\left(t - max\left(\hat{\mu}_j, \hat{\mu}_k\right)\right) - H_i\left(t - \hat{\mu}_j\right)H_i\left(t - \hat{\mu}_k\right)\right]\right\} \quad (22)$$

Next,

$$Var\left(\sum_{j\in r} I\left(y_j \le t\right)\right) = \sum_{j\in r} Var\left[I\left(y_j \le t\right)\right] = (N-n)P\left(y_j \le t\right)\left[1 - P\left(y_j \le t\right)\right] \quad (23)$$

Substituting equations (22) and (23) into equation (19) yields

$$Var\left[\hat{F}_{MBC} - F_N(t)\right] = \frac{1}{N^2}$$

$$\left[\sum_{i\in s}\left\{\sum_{j=1}^{N-n}\sum_{k=1}^{N-n} w_{ij}^{*} w_{ik}^{*}\left[H_i\left(t - max\left(\hat{\mu}_j, \hat{\mu}_k\right)\right) - H_i\left(t - \hat{\mu}_j\right)H_i\left(t - \hat{\mu}_k\right)\right]\right\} + (N-n)P\left(y_j \le t\right)\left[1 - P\left(y_j \le t\right)\right]\right] \quad (24)$$

4. Results

In this section, simulation experiments were done to study the performance of the multiplicative bias corrected estimator.

A population of 1, 000 auxiliary values x_i are generated as independent and identically distributed uniform random variables.

The corresponding survey values y i are generated using the super-population model

$y_i = \mu(x_i) + \varepsilon_i$ with the mean functions being linear, quadratic and cosine.

Nadaraya-Watson kernel weights are used in the smoothing of y_i to obtain the rough estimator,

$\tilde{\mu}_n(x_i) = \sum_{j=1}^{n} w_i\,(x; l)y_j$, of the mean function (x_i). A ratio $\beta_i = \frac{y_i}{\tilde{\mu}_n(X_i)}$ is evaluated and is smoothed further to obtain the correction factor $\hat{\alpha}(x_i)$ which is then used together with the rough estimator to obtain the multiplicative bias corrected estimator, $\hat{\mu}_n(x_i)$, of the mean function.

The existing estimators for distribution functions for finite populations that were used for comparison with our developed estimator $\hat{F}_{MBC}(t)$ are:

i. $\hat{F}_{NW}(t) = \frac{1}{n}\sum_{i\in s} I(y_i \le t)$which was suggested by Nadaraya-Watson (1968).

ii. $\hat{F}_{CD}(t) = \frac{1}{N}\left[\sum_{i\in s} I(y_i \le t) + \sum_{j\in r} \hat{G}\left(t - \hat{a} - \hat{b}x_j\right)\right]$ (Chambers & Dunstan 1986)

iii. $\hat{F}_{RKM}(t) = \frac{1}{n}\sum_{i\in s} I(y_i \le t) + \frac{1}{N}\sum_{j\in r} \hat{G}\left(t - \hat{a} - \hat{b}x_j\right) - \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i\in s} \hat{G}\left(t - \hat{a} - \hat{b}x_j\right)$ (Rao et al 1990).

iv. $\hat{F}_{DH}(t) = \frac{1}{N}\left[\sum_{i\in s} I(y_i \le t) + \sum_{j\in r} \hat{G}(t - \hat{\mu}(x_j))\right]$ (Dorfman & Hall (1993) where $\hat{\mu}$ is the linear estimator of the mean function.

Table 1 shows the unconditional Relative Mean Error (RME) and Relative Root Mean Error (RRME) for the estimators at various values of the quantile α (i.e. 0.25, 0.5 and 0.75). Linear and quadratic mean functions were used to obtain the tabulated results. Similar results and conclusions can be obtained using other mean functions such as sine, cosine, bump etc.

The unconditional Relative Mean Error and Relative Root Mean Error for the estimator $\tilde{F}_N(t)$ are calculated as:

$RME = \frac{1}{F_N(t)}\left\{\frac{1}{200}\sum_{r=1}^{200}\left[\tilde{F}_N{}^r(t) - F_N(t)\right]\right\}$ and $RRME = \frac{1}{F_N(t)}\sqrt{\left\{\frac{1}{200}\sum_{r=1}^{200}\left[\tilde{F}_N{}^r(t) - F_N(t)\right]\right\}}$ respectively where r represents the level of iteration.

Table 1. Unconditional Relative Mean Errors and Relative Root Mean Errors

	Unconditional Relative Mean Error and Relative Root Mean Error			
	Linear Function		***Quadratic Function***	
	$\boldsymbol{\alpha = 0.25}$			
Estimator	**RME**	**RRME**	**RME**	**RRME**
$\hat{F}_{MBC}(t)$	0.005485	0.008889	0.000452	0.008575
$\hat{F}_{NW}(t)$	-1.347148	0.570421	-0.679841	0.46868
$\hat{F}_{CD}(t)$	-3.387451	1.435164	-1.662266	1.14489
$\hat{F}_{RKM}(t)$	0.049355	0.135994	0.007001	0.082798
$\hat{F}_{DH}(t)$	-0.024113	0.036181	0.041498	0.417668
	$\boldsymbol{\alpha = 0.5}$			
	RME	**RRME**	**RME**	**RRME**
$\hat{F}_{MBC}(t)$	0.000572	0.001902	0.000693	0.002504
$\hat{F}_{NW}(t)$	-0.485538	0.481123	-0.340887	0.364297
$\hat{F}_{CD}(t)$	-0.671669	0.667073	-0.596499	0.638570
$\hat{F}_{RKM}(t)$	-0.009321	0.052715	0.002349	0.037726
$\hat{F}_{DH}(t)$	-0.118574	0.061423	-0.013618	0.017422
	$\boldsymbol{\alpha = 0.75}$			
	RME	**RRME**	**RME**	**RRME**
$\hat{F}_{MBC}(t)$	0.001265	0.004539	0.000385	0.001642
$\hat{F}_{NW}(t)$	-0.42853	0.565463	0.35509	0.476310
$\hat{F}_{CD}(t)$	-0.339878	0.449670	-0.337274	0.452659
$\hat{F}_{RKM}(t)$	0.005271	0.045202	0.001099	0.033943
$\hat{F}_{DH}(t)$	-0.016679	0.028441	-0.010832	0.025169

$\hat{F}_{MBC}(t)$ can be seen to be a very efficient estimator of the empirical distribution function at all levels of the α −quantile followed closely by $\hat{F}_{RKM}(t)$ and $\hat{F}_{DH}(t)$. $\hat{F}_{CD}(t)$ proved to be a very inefficient estimator at all levels of α.

Further, graphical comparison of estimators was done which further affirmed the results tabulated above. Figures 1 & 2 gives a plot of all the estimators listed above.

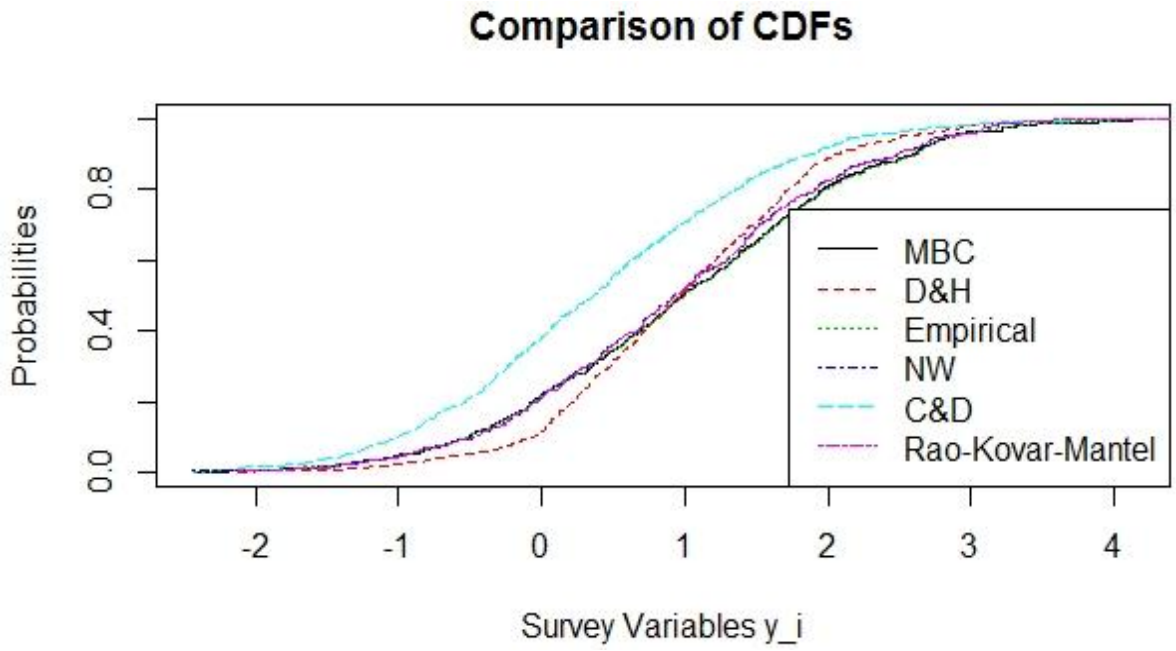

Figure 1. Plot of various Distribution functions using a linear function

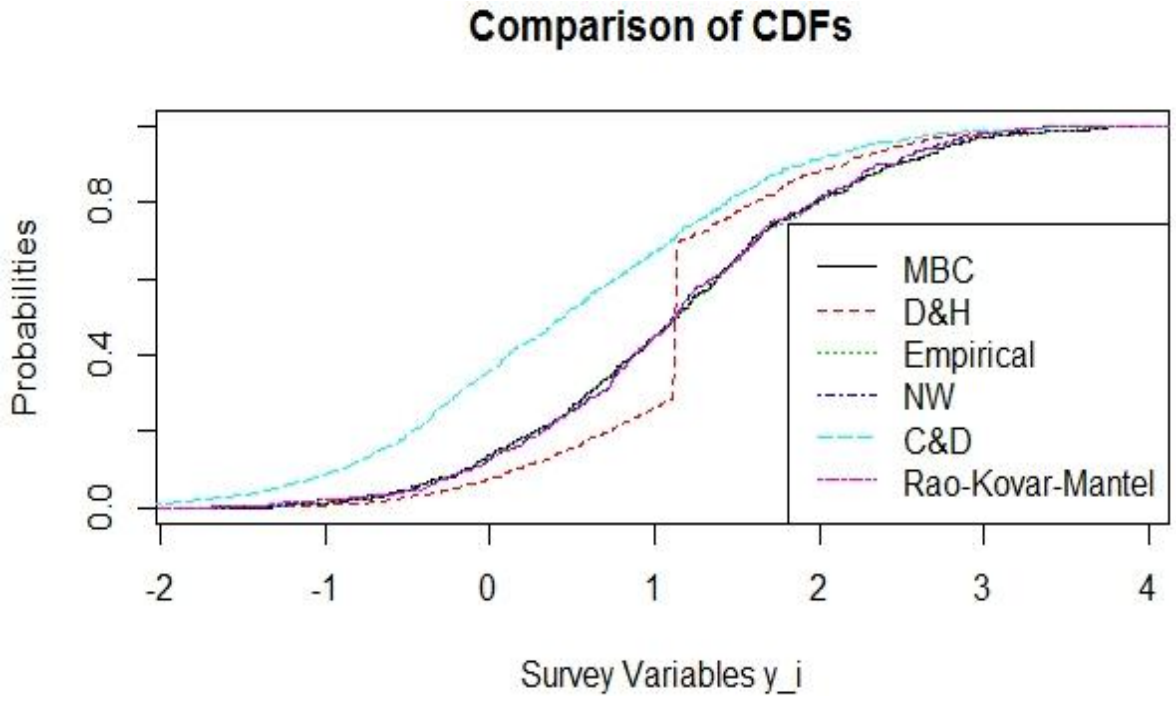

Figure 2. Plot of various Distribution functions using a quadratic function

$\hat{F}_{CD}$ overestimates the empirical distribution function at all points while $\hat{F}_{MBC}$ and $\hat{F}_{RKM}(t)$ give an almost perfect estimation of the empirical distribution function.On the other hand $\hat{F}_{DH}$ underestimates the true function at some points towards the lower tail while it overestimates the same function at other points along the upper tail.

The conditional performance of the estimator was done and was compared with the performance of the estimator.To do this, 200 random samples, all of size 400, were selected and the mean of the auxiliary values x_i was computed for each sample to obtain 200 values of $\bar{X}$.

These sample means were then sorted in ascending order and further grouped into clusters of size 20 such that a total of 10 goups was realized. Further, group means of the means of auxiliary variables was calculated to get $\bar{\bar{X}}$.

Empirical means and biases were then computed for all the estimators $\hat{F}_{MBC}(t)$ and $\hat{F}_{DH}$(t).

The conditional biases were plotted against $\bar{\bar{X}}$ to provide a good understanding of the pattern generated. Figures 3 & 4 show the behavior of the conditional biases realized by all the estimators of distribution functions.

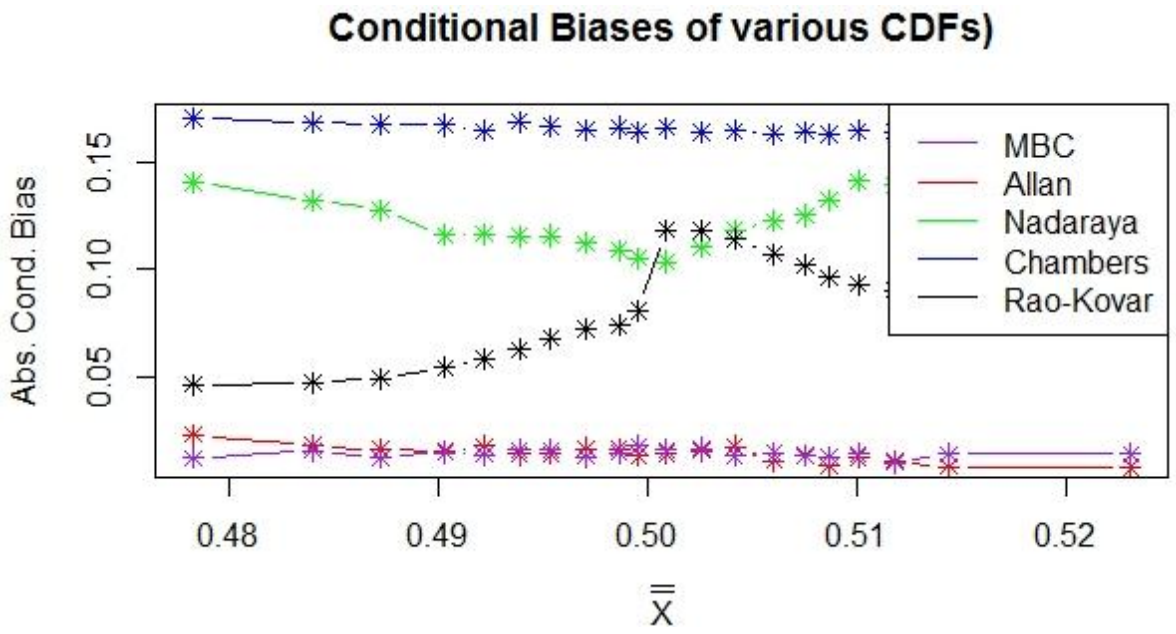

Figure 3. Absolute conditional biases for the estimators using a linear mean function

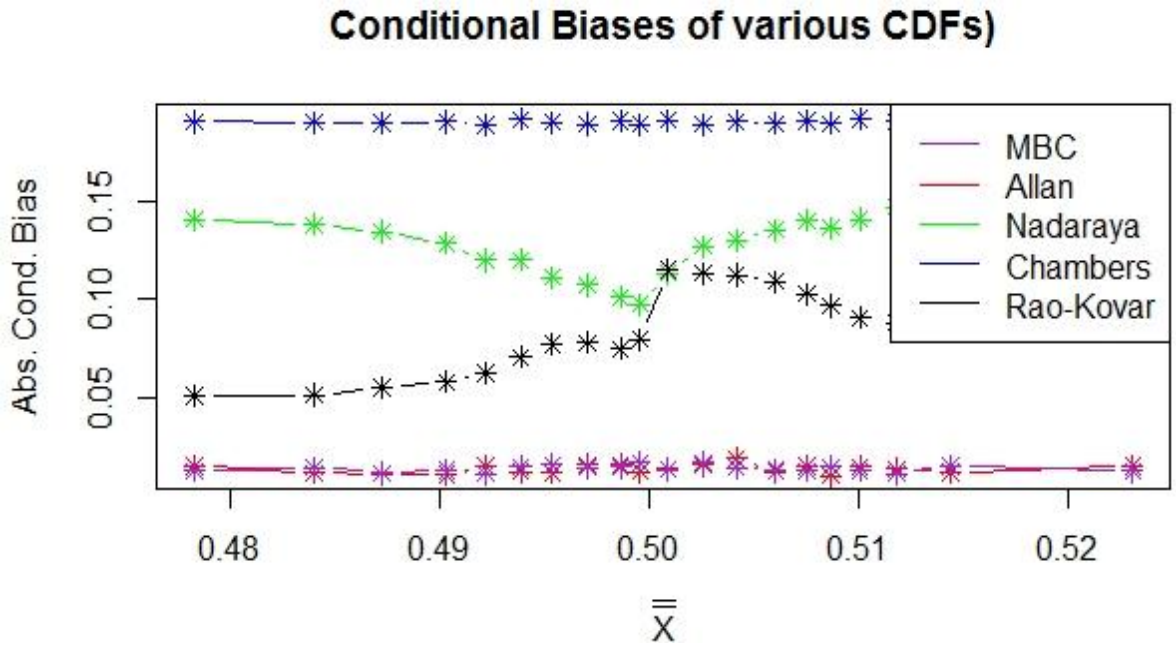

Figure 4. Absolute conditional biases for the estimators using a quadratic mean function

$\hat{F}_{MBC}(t)$ and $\hat{F}_{RKM}(t)$ performed equally better than all other estimators of the true distribution function and it can be seen that sample balancing does not affect the performance of the estimators.

5. Conclusions

In conclusion, using the results from Table 1 and the Figures 3 & 4 $\hat{F}_{MBC}(t)$ was found to be an efficient estimator of the distribution function for finite population. $\hat{F}_{CD}$ was found to be very inefficient of all the estimators with large conditional bias compared to the other estimators. $\hat{F}_{MBC}(t)$ can therefore be used in estimating distribution functions for various units in the population in various sectors of the economy.

REFERENCES

[1] Burr, T., Hengartner, N., Matzner-Lober, E., Myers, S., and Rouviere, L. (2010). Smoothing low resolution gamma spectra. IEEE Transactions on Nuclear Science, 57(5): 2831–2840.

[2] Chambers, R. and Clark, R. (2012). An introduction to model-based survey sampling with applications, volume 37.

OUP Oxford.

[3] Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. Journal of the American Statistical Association, 88(421): 268–277.

[4] Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. Biometrika, 73(3): 597–604.

[5] Dorfman, A. H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. The Annals of Statistics, pages 1452–1475.

[6] Fan, J. (1992). Design-adaptive nonparametric regression. Journal of the American statistical Association, 87(420): 998–1004.

[7] Hardle, W. (1986). A note on jackknifing kernel regression function estimators (corresp.). IEEE transactions on information theory, 32(2): 298–300.

[8] Kuk, A. Y. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. Biometrika, 80(2): 385–392.

[9] Linton, O. and Nielsen, J. P. (1994). A multiplicative bias reduction method for nonparametric regression. Statistics & Probability Letters, 19(3): 181–187.

[10] Muller, H. G. and Stadtmuller, U. (1987). Variable bandwidth kernel estimators of regression curves. The Annals of Statistics, pages 182–201.

[11] Onsongo, W. M. (2018). Nonparametric Estimation of Finite Population Total. PhD thesis, JKUAT-PAUSTI.

[12] Rao, J. N. K., Kovar, J. G., and Mantel H. J. (1990). On Estimating Distribution Functions and Quantiles from Survey data Using Auxiliary Information. Biometrika, pages 365-375.

[13] Zhao, P. Y., Tang, M. L., and Tang, N. S. (2013). Robust estimation of distribution functions and quantiles with non-ignorable missing data. Canadian Journal of Statistics, 41(4): 575–595.