Thailand Statistician
October 2022; 20(4): 860-872
http://statassoc.or.th
Contributed paper

Estimation of Finite Population Quantile by Analytical and Re-scaling Bootstrap Techniques

Sanghamitra Pal [a] and Purnima Shaw*[b]

[a] Faculty of Department of Statistics, West Bengal State University, West Bengal, India.
[b] Department of Statistics and Information Management, Reserve Bank of India, New Delhi, India.
*Corresponding author; e-mail: purnimashaw2011@gmail.com

Received: 18 July 2020
Revised: 25 January 2021
Accepted: 2 February 2021

Abstract

Using a large sample drawn from the population by employing a general sampling design, Chaudhuri and Shaw (2020) used a model-assisted design-based approach in deriving asymptotically design-unbiased estimators of finite population distribution function, the associated quantiles and their mean square errors. Anticipating improvement in the accuracy of estimates of population distribution function and quantiles, this paper uses two alternative techniques in revising estimator of the distribution function; a non-linear function of five population totals. This paper presents a linear approximation of this estimator by using Taylor series expansion neglecting the higher order terms and uses this in deriving its approximate mean square error and the mean square error estimate. Alternatively, Rao and Wu (1988) re-scaling bootstrap technique is also used to modify the estimator of the population distribution function. Estimation of population distribution function using the above two alternative techniques, the population quantiles, their standard errors and related confidence intervals are derived. Numerical findings based on real data show gain in efficiency of estimates of both distribution function and quantiles using the two alternative techniques.

Keywords: Bootstrap, distribution function, quantile, super-population, unequal probability sampling.

1. Introduction

Consider a finite population $U=(1,2,\ldots,i,\ldots,N)$ of a known number N of individuals and let y be a real variable with values y_i for individuals labelled i in U. The distribution function of y in the population for any real number ‘ a ’ is defined as $F(a)$, which denotes the proportion of y -values in the population not exceeding a real number ‘ a ’, i.e.,

$$F(a)=\frac{1}{N}\sum_{i=1}^{N}w_i, \tag{1}$$

where $w_i=\begin{cases}1, \text{ if } y_i\le a,\\ 0, \text{ otherwise}.\end{cases}$ For q in $[0,1]$ the q^{th} quantile of y is a number θ_q for which,

$$\theta_q = F^{-1}(q), \tag{2}$$

It may be noted that θ_q is non-linear function as it cannot be expressed as a linear function of the y-values. As design-based estimator for θ_q cannot be derived, hence, Chaudhuri and Shaw (2020) postulated a super-population model to propose a generalized regression (Greg) estimator for the distribution function. The procedure of estimating finite population distribution function is briefed below.

Consider a sample s drawn from U using an unequal probability sampling design with a pre-assigned probability $p(s)$ admitting positive 1st and 2nd order inclusion-probabilities $\pi_i = \sum_{s \ni i} p(s)$ and $\pi_{ij} = \sum_{s \ni i,j} p(s)$, for $j \neq i$ and $i = 1,2,\ldots,N$. Following the super-population model approach and taking $\underline{Y} = (y_1, y_2, \ldots, y_N)$ as a random vector of y_i's, $i \in U$, the following model is considered,

$$y_i = \beta x_i + \varepsilon_i,\ i \in U, \tag{3}$$

with β being an unknown real constant, x_i's being known real numbers which are well and positively correlated with y_i's and ε_i's being independently distributed random variables with zero means and positive standard deviations σ_i for i in U. Let E_m, V_m and C_m denote the operators for model-based expectation, variance and covariance, respectively and let E_P and V_p denote the design-based expectation and variance operators, respectively. A Greg estimator for the distribution function $F(a)$ of y is

$$\widehat{F(a)} = \frac{1}{N}\left\{\sum_{i \in s}\frac{w_i}{\pi_i} + \frac{\sum_{i \in s} w_i x_i Q_i}{\sum_{i \in s} x_i^2 Q_i}\left(X - \sum_{i \in s}\frac{x_i}{\pi_i}\right)\right\}, \tag{4}$$

where $X = \sum_{i=1}^{N} x_i$ with a suitable choice of Q_i as a positive constant free of the elements of $\underline{Y}$, e.g., $\frac{1}{x_i}, \frac{1}{x_i^2}, \frac{1}{\pi_i x_i}$ or $\frac{1-\pi_i}{\pi_i x_i}$. From Equation (4), it may be observed that $\widehat{F(a)}$ is a non-linear function of several linear estimators. So, the design-based properties of $\widehat{F(a)}$, using exact probability distribution are difficult to obtain. To study the design-based features relating to the expectation, mean square error and estimate of mean square error of the above Greg estimator, asymptotic properties relevant to finite population were applied by Chaudhuri and Shaw (2020) following Brewer (1979) 'model-assisted design-based' approach and it was found that $\widehat{F(a)}$ is asymptotically design unbiased for $F(a)$. Once the distribution function of y is estimated using the large sample at hand, asymptotic design-unbiased estimator of the q^{th} quantile of the distribution of y is obtained as

$$\widehat{\theta_q} = L + \frac{(q-b)}{f} l, \tag{5}$$

where (i) L is the lower class boundary of the q^{th} quantile class in the estimated distribution function, (ii) $b = \widehat{F(L)}$, the estimated distribution function at L, (iii) $f = \widehat{F(L')} - \widehat{F(L)}$, with L' being the upper-class boundary of the q^{th} quantile-class and $\widehat{F(L')}$ being the estimated distribution function at

L' and (iv) $l = L' - L.$ The approximate mean square error of $\widehat{F(a)}$ and asymptotic design unbiased estimator for the mean square error are then derived. Using Chebychev's inequality and neglecting design-based bias of $\widehat{F(a)}$ for large samples, related confidence interval for $F(a)$ are then estimated. This assists in estimating the asymptotic and approximate confidence interval for θ_q followed by estimation of standard error of the quantile estimate $\widehat{\theta_q}$. The two requirements of the above theory are that the sample size should be large enough and there should be high correlation between the explanatory variable and the variable of interest. If these conditions are satisfied, then the estimates computed thereafter are asymptotically unbiased.

The expression $\widehat{F(a)}$ is a non-linear function of different linear estimators. Following the plug-in principle of substituting each population total in a non-linear function, by its unbiased estimator, we anticipate that estimating N by its unbiased estimator $\sum_{i\in s}\frac{1}{\pi_i}$ may improve the accuracy of $\widehat{F(a)}.$ So, we propose an alternative estimator of $F(a)$ as

$$\widehat{F(a)}^* = \frac{1}{\sum_{i\in s}\frac{1}{\pi_i}}\left\{\sum_{i\in s}\frac{w_i}{\pi_i} + \frac{\sum_{i\in s} w_i x_i Q_i}{\sum_{i\in s} x_i^2 Q_i}\left(X - \sum_{i\in s}\frac{x_i}{\pi_i}\right)\right\}. \qquad (6)$$

As $\widehat{F(a)}^*$ cannot be converted conveniently into a Horvitz Thompson (1952) estimator, hence $\widehat{F(a)}^*$ is approximated to a linear function by using Taylor series expansion and neglecting the higher order terms. This helps in deriving the approximate mean square error of $\widehat{F(a)}^*$ and its mean square error estimate followed by estimation of the related confidence interval. From the revised estimated distribution function, the approximate estimators of quantiles, their standard errors and confidence intervals are obtained. These derivations are presented explicitly in Section 2.

The bootstrap technique given by Efron (1982) is used to obtain the empirical distribution when the original distribution of the random variable is unknown. In this technique, independent random samples are drawn from a random sample at hand to produce an empirical distribution. Although there is a huge literature in this area, we are interested in those in which sample selection from the population can be performed using a general sampling scheme. In this aspect, Chaudhuri and Saha (2004) used the bootstrap technique to evaluate the accuracy in estimation of total area under cultivation in Indian districts by using a two-stage unequal probability sampling scheme. Rao and Wu (1988) presented re-scaling bootstrap technique for complex survey design to find out confidence interval and mean square error of non-linear statistic. Later, Pal (2009) extended their method for non-fixed size design. Section 3 uses Rao and Wu (1988) re-scaling bootstrap technique in presenting a further modified estimator of $F(a),$ its mean square error, mean square error estimate and related confidence interval. The approximate quantile estimates, their standard errors and confidence intervals are further derived from the estimated distribution function thereof. To examine the performance of the above two methodologies, a numerical illustration is provided in Section 4, followed by the concluding remarks.

2. Linearization Technique

Here, we express $\widehat{F(a)}^*$ as a non-linear function of five linear estimators. Writing $\widehat{F(a)}^*$ as,

$$\widehat{F(a)}^{*} = \frac{t_1 + \frac{t_2}{t_3}\left(X - t_4\right)}{t_5} = f\left(t_1, t_2, \ldots, t_5\right) = f\left(\underline{\boldsymbol{t}}\right), \tag{7}$$

where

$$t_1 = \sum_{i \in s} \frac{w_i}{\pi_i},\ t_2 = \sum_{i \in s} \frac{w_i x_i Q_i \pi_i}{\pi_i},\ t_3 = \sum_{i \in s} \frac{x_i^{\,2} Q_i \pi_i}{\pi_i},\ t_4 = \sum_{i \in s} \frac{x_i}{\pi_i},\ t_5 = \sum_{i \in s} \frac{1}{\pi_i}$$

are unbiased estimators of $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$, respectively, with

$$\theta_1 = \sum_{i=1}^{N} w_i,\ \theta_2 = \sum_{i=1}^{N} w_i x_i Q_i \pi_i,\ \theta_3 = \sum_{i=1}^{N} x_i^{\,2} Q_i \pi_i,\ \theta_4 = \sum_{i=1}^{N} x_i,\ \theta_5 = N,$$

and noting that

$$\frac{\theta_1 + \frac{\theta_2}{\theta_3}\left(X - \theta_4\right)}{\theta_5} = f\left(\theta_1, \theta_2, \ldots, \theta_5\right) = f\left(\underline{\boldsymbol{\theta}}\right) = \frac{\theta_1}{\theta_5} = F\left(a\right). \tag{8}$$

We use Taylor series expansion on $f\left(\underline{\boldsymbol{t}}\right)$ and neglect the higher order terms,

$$f\left(\underline{\boldsymbol{t}}\right) = f\left(\underline{\theta}\right) + \left.\frac{\partial f\left(\underline{t}\right)}{\partial t_1}\right|_{\underline{t}=\underline{\theta}} \left(t_1 - \theta_1\right) + \left.\frac{\partial f\left(\underline{t}\right)}{\partial t_2}\right|_{\underline{t}=\underline{\theta}} \left(t_2 - \theta_2\right) + \left.\frac{\partial f\left(\underline{t}\right)}{\partial t_3}\right|_{\underline{t}=\underline{\theta}} \left(t_3 - \theta_3\right). \tag{9}$$

Neglecting the higher order terms it may be observed that $f\left(\underline{\boldsymbol{t}}\right)$ is an approximate estimator of $f\left(\underline{\boldsymbol{\theta}}\right)$, such that $E_P\left\{f\left(\underline{\boldsymbol{t}}\right)\right\} \simeq f\left(\underline{\boldsymbol{\theta}}\right)$. The approximate variance of $f\left(\underline{\boldsymbol{t}}\right)$ for any fixed-sample size design, is given by

$$V_P\left\{f\left(\underline{\boldsymbol{t}}\right)\right\} = V_P\left\{ \left.\frac{\partial \mathrm{f}\left(\underline{\boldsymbol{t}}\right)}{\partial \mathrm{t}_1}\right|_{\underline{\boldsymbol{t}}=\underline{\boldsymbol{\theta}}} t_1 + \left.\frac{\partial \mathrm{f}\left(\underline{\boldsymbol{t}}\right)}{\partial \mathrm{t}_2}\right|_{\underline{\boldsymbol{t}}=\underline{\boldsymbol{\theta}}} t_2 + \left.\frac{\partial \mathrm{f}\left(\underline{\boldsymbol{t}}\right)}{\partial \mathrm{t}_3}\right|_{\underline{\boldsymbol{t}}=\underline{\boldsymbol{\theta}}} t_3 + \left.\frac{\partial \mathrm{f}\left(\underline{\boldsymbol{t}}\right)}{\partial \mathrm{t}_4}\right|_{\underline{\boldsymbol{t}}=\underline{\boldsymbol{\theta}}} t_4 + \left.\frac{\partial \mathrm{f}\left(\underline{\boldsymbol{t}}\right)}{\partial \mathrm{t}_5}\right|_{\underline{\boldsymbol{t}}=\underline{\boldsymbol{\theta}}} t_5 \right\}$$

$$= V_P\left\{ \frac{1}{\theta_5} \sum_{i \in s} \frac{w_i}{\pi_i} + \frac{X - \theta_4}{\theta_3 \theta_5} \sum_{i \in s} \frac{w_i x_i Q_i \pi_i}{\pi_i} + \frac{\theta_2\left(\theta_4 - X\right)}{\theta_3^{\,2} \theta_5} \sum_{i \in s} \frac{x_i^{\,2} Q_i \pi_i}{\pi_i} - \frac{\theta_2}{\theta_3 \theta_5} \sum_{i \in s} \frac{x_i}{\pi_i} + \frac{\frac{\theta_2}{\theta_3}\left(\theta_4 - X\right) - \theta_1}{\theta_5^{\,2}} \sum_{i \in s} \frac{1}{\pi_i} \right\}$$

$$= V_P\left\{ \sum_{i \in s} \frac{\Psi_i}{\pi_i} \right\} \tag{10}$$

where

$$\Psi_i = \frac{w_i}{\theta_5} + \frac{\left(X - \theta_4\right) w_i x_i Q_i \pi_i}{\theta_3 \theta_5} + \frac{\theta_2\left(\theta_4 - X\right) x_i^{\,2} Q_i \pi_i}{\theta_3^{\,2} \theta_5} - \frac{\theta_2 x_i}{\theta_3 \theta_5} + \frac{\frac{\theta_2}{\theta_3}\left(\theta_4 - X\right) - \theta_1}{\theta_5^{\,2}},$$

$$V_P\left\{f\left(\underline{\boldsymbol{t}}\right)\right\} = V_P\left\{ \sum_{i \in s} \frac{\Psi_i}{\pi_i} \right\} = \frac{1}{N^2} \sum_{i<}^{N} \sum_{j}^{N} \left(\pi_i \pi_j - \pi_{ij}\right) \left(\frac{\Psi_i}{\pi_i} - \frac{\Psi_j}{\pi_j} \right)^2 . \tag{11}$$

An asymptotic design unbiased estimator for $V_P\left\{f\left(\underline{\boldsymbol{t}}\right)\right\}$ is

$$v = \frac{1}{N^2} \sum_{i<} \sum_{j \in s} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{\widehat{\Psi}_i}{\pi_i} - \frac{\widehat{\Psi}_j}{\pi_j} \right)^2 , \tag{12}$$

where

$$\widehat{\Psi_i} = \frac{w_i}{t_5} + \frac{\left(X - t_4\right) w_i x_i Q_i \pi_i}{t_3 t_5} + \frac{\left\{t_2\left(t_4 - X\right)\right\} x_i^{\,2} Q_i \pi_i}{t_3^{\,2} t_5} - \frac{t_2 x_i}{t_3 t_5} + \frac{\frac{t_2}{t_3}\left(t_4 - X\right) - t_1}{t_5^{\,2}}.$$

$$P\left[\left|\widehat{F(a)}^{*} - F(a)\right| \le \frac{1}{\sqrt{\alpha}} \sqrt{V_p\left(\widehat{F(a)}\right)}\right] \ge (1-\alpha). \tag{13}$$

Hence, the confidence interval for $F(a)$, denoted by $\left(\widehat{F_L(a)}, \widehat{F_U(a)}\right)$ with a confidence coefficient greater than or equal to $(1-\alpha),$ are

$$\widehat{F_L(a)} = \widehat{F(a)}^{*} - \lambda\sqrt{v},\ \widehat{F_U(a)} = \widehat{F(a)}^{*} + \lambda\sqrt{v}, \tag{14}$$

where $\lambda = 1/\sqrt{\alpha}$. The confidence interval $\left(\widehat{F_L(a)}, \widehat{F_U(a)}\right)$ are computed for each of the classes of the estimated distribution function. The asymptotic and approximate confidence interval for θ_q is $\left(\widehat{\theta_q^L}\ ,\ \widehat{\theta_q^U}\right),$ where

$$\widehat{\theta_q^L} = \widehat{F_U}^{-1}\left(q\right),\ \widehat{\theta_q^U} = \widehat{F_L}^{-1}\left(q\right). \tag{15}$$

This is followed by computation of the asymptotic and approximate estimator of standard error of $\widehat{\theta_q}$ as given below,

$$\widehat{SE}\left(\widehat{\theta_q}\right) = \frac{\sqrt{\alpha}\left(\widehat{\theta_q^U} - \widehat{\theta_q^L}\right)}{2}. \tag{16}$$

3. Re-scaling Bootstrap Technique

For a sample chosen by a complex sampling design, Rao and Wu (1988) proposed a re-scaling bootstrap technique to compute mean square error in estimating a non-linear function of multiple (say, k) population totals, each linear in y_i's. From the original samples, bootstrap samples are drawn in such a way that for the special case of $k = 1,$ the bootstrap based expectation and the bootstrap variance for the estimator matches in the linear case exactly to the sample based estimator and the standard unbiased sample based variance estimator for the population total, respectively. The related confidence intervals make allowance for the skewness in the distributions of the estimate avoiding Normality assumption. Hence, we use this method to estimate the mean square error of $\widehat{F(a)}^{*}$ and also estimate the related confidence interval for $F(a).$ Considering $\widehat{F(a)}^{*}$, the Yates Grundy estimate of variance of $t_1 = \sum_{i \in s} \frac{w_i}{\pi_i},$ is

$$v\left(t_1\right) = \sum_{i<j \in s} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{w_i}{\pi_i} - \frac{w_j}{\pi_j}\right)^2. \tag{17}$$

Following re-scaling method by Rao and Wu (1988), we consider all the $n\left(n-1\right)$ pairs of units $\left(i, j\right)$ in the sample, such that $i \ne j, i, j \in s$ and select m pairs of units with probability $\lambda_{ij} = \frac{1}{n\left(n-1\right)}$

with replacement. Let the selected pairs be represented by $\left(i^*,j^*\right), i^* \neq j^*$. Next, we modify t_1 to obtain the bootstrap estimator t_1^*,

$$t_1^* = t_1 + \sum_{\left(i^*,j^*\right)=1}^{m} \sqrt{\frac{\pi_{i^*}\pi_{j^*} - \pi_{i^*j^*}}{2\pi_{i^*j^*}}} \left(\frac{w_{i^*}}{\pi_{i^*}} - \frac{w_{j^*}}{\pi_{j^*}} \right). \tag{18}$$

Denoting E_* and V_* as the bootstrap expectation and variance operators, we get

$$E_*\left(t_1^*\right) = t_1, \tag{19}$$

$$V_*\left(t_1^*\right) = \sum_{i\neq j\in s} \left(\frac{\pi_i\pi_j - \pi_{ij}}{2\pi_{ij}} \right) \left(\frac{w_i}{\pi_i} - \frac{w_j}{\pi_j} \right)^2 = \sum_{i<j\in s} \left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{w_i}{\pi_i} - \frac{w_j}{\pi_j} \right)^2 = v\left(t_1\right). \tag{20}$$

Similarly, the revised estimators for t_2, t_3, t_4 and t_5 are,

$$t_2^* = t_2 + \sum_{\left(i^*,j^*\right)=1}^{m} \sqrt{\frac{\pi_{i^*}\pi_{j^*} - \pi_{i^*j^*}}{2\pi_{i^*j^*}}} \left(\frac{w_{i^*}x_{i^*}Q_{i^*}\pi_{i^*}}{\pi_{i^*}} - \frac{w_{j^*}x_{j^*}Q_{j^*}\pi_{j^*}}{\pi_{j^*}} \right), \tag{21}$$

$$t_3^* = t_3 + \sum_{\left(i^*,j^*\right)=1}^{m} \sqrt{\frac{\pi_{i^*}\pi_{j^*} - \pi_{i^*j^*}}{2\pi_{i^*j^*}}} \left(\frac{x_{i^*}{}^2Q_{i^*}\pi_{i^*}}{\pi_{i^*}} - \frac{x_{j^*}{}^2Q_{j^*}\pi_{j^*}}{\pi_{j^*}} \right), \tag{22}$$

$$t_4^* = t_4 + \sum_{\left(i^*,j^*\right)=1}^{m} \sqrt{\frac{\pi_{i^*}\pi_{j^*} - \pi_{i^*j^*}}{2\pi_{i^*j^*}}} \left(\frac{x_{i^*}}{\pi_{i^*}} - \frac{x_{j^*}}{\pi_{j^*}} \right), \tag{23}$$

$$t_5^* = t_5 + \sum_{\left(i^*,j^*\right)=1}^{m} \sqrt{\frac{\pi_{i^*}\pi_{j^*} - \pi_{i^*j^*}}{2\pi_{i^*j^*}}} \left(\frac{1}{\pi_{i^*}} - \frac{1}{\pi_{j^*}} \right). \tag{24}$$

Then, the bootstrap estimator for $F\left(a\right)$ is given by

$$\widetilde{\widetilde{F\left(a\right)}} = \frac{t_1^* + \frac{t_2^*}{t_3^*}\left(X - t_4^*\right)}{t_5^*}. \tag{25}$$

The above steps from (18) to (25) are replicated independently for a large number of times, (say $B = 1000$ or 10,000 or larger) to calculate the corresponding estimates $\widetilde{\widetilde{F\left(a\right)}}_1, \widetilde{\widetilde{F\left(a\right)}}_2, \ldots, \widetilde{\widetilde{F\left(a\right)}}_B$. Then, the approximate bootstrap estimator of $F\left(a\right)$, denoted by $\widetilde{\widetilde{F\left(a\right)}}^*$ and the bootstrap variance estimator, denoted by $\tilde{v}$, are

$$\widetilde{\widetilde{F\left(a\right)}}^* = \frac{1}{B}\sum_{b=1}^{B} \widetilde{\widetilde{F\left(a\right)}}_b, \tag{26}$$

$$\tilde{v} = \frac{1}{B-1}\sum_{b=1}^{B} \left(\widetilde{\widetilde{F\left(a\right)}}_b - \widetilde{\widetilde{F\left(a\right)}}^* \right)^2. \tag{27}$$

For obtaining the confidence interval for $F(a)$, the percentile method is followed. Two values $\widetilde{\widetilde{F(a)}}_L$ and $\widetilde{\widetilde{F(a)}}_U$ are found such that $100\alpha/2\%$ of the values $\widetilde{\widetilde{F(a)}}_1, \widetilde{\widetilde{F(a)}}_2, .., \widetilde{\widetilde{F(a)}}_B$ lie below $\widetilde{\widetilde{F(a)}}_L$ and $100\alpha/2\%$ of the values $\widetilde{\widetilde{F(a)}}_1, \widetilde{\widetilde{F(a)}}_2, .., \widetilde{\widetilde{F(a)}}_B$ lie above $\widetilde{\widetilde{F(a)}}_U$. Then, $\left(\widetilde{\widetilde{F(a)}}_L, \widetilde{\widetilde{F(a)}}_U\right)$ form a $100(1-\alpha)\%$ confidence interval for $F(a)$ with a confidence coefficient of $100(1-\alpha)\%$.

Using the large sample s, the distribution function and its related confidence interval are thus estimated and derived. Following Chaudhuri and Shaw (2020), the asymptotic design unbiased estimator of the q^{th} quantile is obtained from the estimated distribution function as shown in (5). Now, $\left(\widetilde{\widetilde{F(a)}}_L, \widetilde{\widetilde{F(a)}}_U\right)$ can be computed for several classes of the estimated distribution. The asymptotic and approximate confidence interval and standard error estimate of $\widehat{\theta}_q$ can be obtained following (15) and (16), respectively.

4. Numerical Illustration

In order to examine the performances of the above two techniques described in Sections 2 and 3 in estimating the distribution function and consequently, the quantiles, we use the same variables in the dataset used by Chaudhuri and Shaw (2020) i.e., “SaratogaHouses” (available in the package “mosaicData” in R). The data contains $N = 1728$ observations on 16 variables on the characteristics of houses in the country Saratoga in New York, USA in 2006. The variables used for our study are, y, x and z denoting house price in 1000’s of US dollars, living area in square feet and number of bathrooms (half bathrooms having no shower or tub), respectively. The variable y is well correlated with x as well as with z. The observation with $z_{1494} = 0$ i.e., no bathroom in the 1494^{th} house was removed from the population vide Chaudhuri (2018) which says that a necessary and sufficient condition for existence of an unbiased estimator for a finite population parameter using a sample selected according to a general sampling design $p(s)$ is $\pi_i > 0$ for each i in U. To estimate the 1st, 2nd and 3rd quartiles of y using known information on x, a sample of size $n = 190$ households was drawn by employing the Hartley and Rao (1962) sampling method in which a systematic sample by probability proportional to size (PPS) method is used after the random arrangement of the population units. The variable z is considered as the size measure for sampling the units from the population.

Using (2), we estimate the distribution function followed by estimation of its variance using the linearization technique discussed in Section 2. Utilizing this estimated distribution function, the q^{th} quartile, $\widehat{\theta}_q$ is then estimated using Chaudhuri and Shaw (2020), followed by $\widetilde{F\left(\widehat{\theta}_q\right)}^*$ using (2), estimate of its variance, denoted by $\widehat{V}_P\left(\widetilde{F\left(\widehat{\theta}_q\right)}^*\right)$, using (9). Then following Chaudhuri and Shaw (2020), we compute confidence intervals for $F\left(\theta_q\right)$ and θ_q, assuming $\lambda = 2$ and 3 and standard error of $\widehat{\theta_q}$, denoted by $SE\left(\widehat{\theta_q}\right)$ and estimate of coefficient of variation (CV) of

$$\widehat{F\left(\widehat{\theta_q}\right)}^* = 100\frac{\sqrt{\widehat{V_P}\left(\widehat{F\left(\widehat{\theta_q}\right)}^*\right)}}{\widehat{F\left(\widehat{\theta_q}\right)}^*}, \text{and estimated CV of } \widehat{\theta_q} = 100\frac{\sqrt{SE\left(\widehat{\theta_q}\right)}}{\widehat{\theta_q}}$$

The above measurements are obtained for 1st, 2nd and 3rd quartiles (i.e., $q = 1,2,3$) and by assuming different choices of Q_i as $\frac{1}{x_i}, \frac{1}{x_i^2}, \frac{1}{\pi_i x_i}$ and $\frac{1-\pi_i}{\pi_i x_i}$. To judge the efficacy of the proposed procedures, the above computations were done each time for 100 independent samples drawn from the population data and then to examine the quality of the estimates, we calculate AE (average estimate) of $F\left(\theta_q\right)$ denoted by $\overline{\widehat{F\left(\widehat{\theta_q}\right)}^*} = \frac{1}{100}\sum_{i=1}^{100}\widehat{F\left(\widehat{\theta_q}\right)}^*_i$, AE of θ_q denoted by $\overline{\widehat{\theta_q}} = \frac{1}{100}\sum_{i=1}^{100}\widehat{\theta}_{q\,i}$, where $\widehat{F\left(\widehat{\theta_q}\right)}^*_i$ and $\widehat{\theta}_{q\,i}$ are the estimates of $F\left(\theta_q\right)$ and θ_q, respectively, obtained from the i^{th} re-sample. Also, ACV (average coefficient of variation) is the average of the coefficient of variation over the 100 replicates for both $\widehat{F\left(\widehat{\theta_q}\right)}^*$ and $\widehat{\theta_q}$.

ARB (absolute relative bias) $= \left|\frac{\overline{\widehat{F\left(\widehat{\theta_q}\right)}^*} - F\left(\theta_q\right)}{F\left(\theta_q\right)}\right|$ and $\left|\frac{\overline{\widehat{\theta_q}} - \theta_q}{\theta_q}\right|$, ACP (actual coverage percentage) for $F\left(\theta_q\right)$ and θ_q is the percentage of replicates out of 100 for which the estimated CI's cover $F\left(\theta_q\right)$ and θ_q, respectively, AL (average length) is the average length of the CI's for $F\left(\theta_q\right)$ and θ_q over 100 replicates and AVE (average variance estimate) $= \frac{1}{100}\sum_{i=1}^{100}\widehat{V_p}\left(\widehat{F\left(\widehat{\theta_q}\right)}^*\right)_i$ and $\frac{1}{100}\sum_{i=1}^{100}\left\{SE\left(\widehat{\theta_q}\right)\right\}_i^2$ were computed.

In order to check the performance of Rao and Wu's (1988) re-scaling bootstrap technique discussed in Section 3, we take $B = 1000$ bootstrap samples and estimate the distribution function followed by estimation of its variance using (19) and (20), respectively. Utilizing this estimated distribution function, the q^{th} quartile, $\widehat{\theta_q}$ is then estimated using Chaudhuri and Shaw (2020), followed by $\widetilde{\widetilde{F\left(\widehat{\theta_q}\right)}}^*$, estimate of its variance, denoted by $\widehat{V_P}\left(\widetilde{\widetilde{F\left(\widehat{\theta_q}\right)}}^*\right)$, using (19) and (20), respectively. Then, we compute confidence intervals (CI) for $F\left(\theta_q\right)$ by using the percentile method as discussed in Section 3 and then estimate the confidence interval for θ_q, taking $\alpha = \frac{1}{\lambda^2}$ with values of λ as 2 and 3, as taken by Chaudhuri and Shaw (2020) and then estimate standard error of $\widehat{\theta_q}$, denoted by

$SE\left(\widehat{\theta_q}\right)$ and Estimate of CV of $\widetilde{\widehat{F\left(\widehat{\theta_q}\right)}}^{*}=100\frac{\sqrt{\widehat{V_P}\left(\widetilde{\widehat{F\left(\widehat{\theta_q}\right)}}^{*}\right)}}{\widetilde{\widehat{F\left(\widehat{\theta_q}\right)}}^{*}}$, and estimated CV of $\widehat{\theta_q}=100\frac{\sqrt{SE\left(\widehat{\theta_q}\right)}}{\widehat{\theta_q}}$.

The performances of estimates of distribution function and quantiles obtained from the linearization technique (denoted as Method-II) and re-scaling bootstrap technique (denoted by Method-III) *vis-à-vis* the performances of estimates by Chaudhuri and Shaw (2020) method, denoted as Method-I, are demonstrated in Tables 1 and 2 below. Results show that the proposed methodologies are more efficient than Method I in terms of improvement in ACV, ARB and AL for estimates of distribution function. Quantile estimates obtained using Methods II and III are better than the estimates using Method-I in terms of ACV and AL. As compared to Method I, Method II shows improvement in respect of ARB, ACV and AL. However, quantile estimates from Method III outperform the quantile estimates from both Methods I and II in respect of ACV and AL.

5. Conclusion

This paper, using a large sample drawn from the population using a general sampling scheme, intends to improve Chaudhuri and Shaw (2020) Greg estimator for the finite population distribution function of a variable and estimator of quantile of the variable. At first, it presents a modified Greg estimator for the distribution function and finds that the estimator is a non-linear function of several linear estimators. Then, this estimator is approximated to a linear function and using the linearization technique, the accuracy of the alternative estimator of the distribution function is derived. Rao and Wu (1988) re-scaling bootstrap technique is used for complex survey designs to find out confidence interval and mean square error of non-linear statistic. The technique has been used in this paper to assess the accuracy of the above mentioned modified estimator of the finite population distribution function and the associated quantiles. This method is proposed as an alternative to the linearization technique for assessing the accuracy of the estimators of distribution function and quantile. Based on real data, it is observed that the estimates of both distribution function and quantiles obtained from the linearization technique are relatively more efficient than those obtained from the method by Chaudhuri and Shaw (2020), in terms of average relative bias, average coefficient of variation and average length of the estimated confidence intervals. For quantile estimation, the re-scaling bootstrap technique provides the most efficient estimates in comparison to the other two methods.

Table 1 Performance of the estimates of $F\left(\theta_q\right)$ for different values of q and λ

q	θ_q	Q_i	Method	AE	AVE	ACV	ARB	ACP		AL	
								λ		λ	
								2	3	2	3
0.25	139199.3	$\frac{1}{x_i}$	I	0.217	0.00132	16.9	0.023	99.0	100.0	0.145	0.218
			II	0.224	0.00039	8.8	0.008	86.0	94.0	0.077	0.116
			III	0.227	0.00112	14.7	0.020	100.0	100.0	0.077	0.103
		$\frac{1}{x_i^2}$	I	0.219	0.00134	16.8	0.014	98.0	99.0	0.146	0.219
			II	0.224	0.00040	8.9	0.007	78.0	92.0	0.078	0.118
			III	0.223	0.00119	15.5	0.002	100.0	100.0	0.078	0.109
		$\frac{1}{\pi x_i}$	I	0.223	0.00134	16.5	0.002	100.0	100.0	0.146	0.219
			II	0.221	0.00039	8.9	0.005	80.0	95.0	0.077	0.116
			III	0.219	0.00119	15.9	0.014	75.0	100.0	0.079	0.109
		$\frac{1-\pi_i}{\pi_i x_i}$	I	0.221	0.00135	16.7	0.005	98.0	100.0	0.147	0.220
			II	0.220	0.00041	9.2	0.009	82.0	94.0	0.079	0.119
			III	0.228	0.00120	15.4	0.026	66.7	83.3	0.079	0.109
0.5	189519.8	$\frac{1}{x_i}$	I	0.503	0.00205	9.0	0.010	100.0	100.0	0.181	0.272
			II	0.504	0.00040	3.9	0.011	97.0	98.0	0.079	0.119
			III	0.487	0.00149	7.9	0.022	66.7	100.0	0.086	0.122
		$\frac{1}{x_i^2}$	I	0.500	0.00205	9.1	0.002	100.0	100.0	0.181	0.271
			II	0.503	0.00040	3.9	0.010	98.0	100.0	0.078	0.118
			III	0.508	0.00144	7.5	0.020	100.0	100.0	0.088	0.118
		$\frac{1}{\pi_i x_i}$	I	0.498	0.00206	9.1	0.002	100.0	100.0	0.181	0.272
			II	0.499	0.00041	4.0	0.001	95.0	99.0	0.080	0.120
			III	0.491	0.00150	7.9	0.016	100.0	100.0	0.089	0.124
		$\frac{1-\pi_i}{\pi_i x_i}$	I	0.501	0.00208	9.1	0.004	100.0	100.0	0.182	0.273
			II	0.499	0.00040	4.0	0.001	97.0	100.0	0.079	0.118
			III	0.512	0.00146	7.5	0.028	100.0	100.0	0.088	0.122
0.75	267121.0	$\frac{1}{x_i}$	I	0.777	0.00165	5.2	0.001	100.0	100.0	0.162	0.244
			II	0.772	0.00041	2.6	0.007	97.0	100.0	0.080	0.119
			III	0.767	0.00103	4.2	0.013	100.0	100.0	0.074	0.102
		$\frac{1}{x_i^2}$	I	0.773	0.00174	5.4	0.005	100.0	100.0	0.167	0.250
			II	0.770	0.00039	2.5	0.009	98.0	100.0	0.077	0.116
			III	0.770	0.00101	4.1	0.009	100.0	100.0	0.072	0.099
		$\frac{1}{\pi_i x_i}$	I	0.773	0.00168	5.3	0.006	100.0	100.0	0.164	0.246
			II	0.773	0.00042	2.6	0.006	100.0	100.0	0.082	0.122
			III	0.761	0.00100	4.1	0.020	75.0	100.0	0.072	0.100
		$\frac{1-\pi_i}{\pi_i x_i}$	I	0.775	0.00167	5.3	0.003	100.0	100.0	0.163	0.245
			II	0.774	0.00040	2.5	0.005	98.0	100.0	0.079	0.118
			III	0.777	0.00097	4.0	0.000	91.7	91.7	0.071	0.099

Table 2 Performance of the estimates of θ_q for different values of q and λ

q	θ_q	Q_i	Method	AE	ACV		ARB	ACP	
					λ			λ	
					2	3		2	3
0.25	139199.3	$\frac{1}{x_i}$	I	139133.4	4.5	4.5	0.0005	98.0	100.0
			II	139421.7	3.5	3.5	0.0016	99.0	100.0
			III	137999.8	2.4	2.2	0.0086	66.7	100.0
		$\frac{1}{x_i^2}$	I	138915.4	4.5	4.6	0.0020	99.0	100.0
			II	138610.3	3.4	3.4	0.0042	96.0	100.0
			III	142487.2	2.2	2.0	0.0236	66.7	66.7
		$\frac{1}{\pi_i x_i}$	I	139559.4	4.4	4.4	0.0026	99.0	100.0
			II	139384.7	3.5	3.5	0.0013	97.0	99.0
			III	140356.1	2.1	2.0	0.0083	100.0	100.0
		$\frac{1-\pi_i}{\pi_i x_i}$	I	140030.0	4.4	4.5	0.0060	97.0	99.0
			II	139522.1	3.6	3.6	0.0023	98.0	98.0
			III	142490.9	2.2	2.0	0.0236	75.0	83.3
0.5	189519.8	$\frac{1}{x_i}$	I	190847.0	5.0	5.1	0.0070	95.0	99.0
			II	189737.1	3.3	3.4	0.0011	86.0	100.0
			III	188930.7	2.1	1.9	0.0031	100.0	100.0
		$\frac{1}{x_i^2}$	I	192043.9	5.1	5.1	0.0133	96.0	100.0
			II	190443.5	3.3	3.4	0.0049	89.0	98.0
			III	194890.6	2.4	2.3	0.0283	100.0	100.0
		$\frac{1}{\pi_i x_i}$	I	188626.3	4.8	5.0	0.0047	96.0	99.0
			II	189879.2	3.3	3.4	0.0019	93.0	100.0
			III	188122.3	2.2	2.1	0.0074	75.0	83.3
		$\frac{1-\pi_i}{\pi_i x_i}$	I	189402.7	5.0	5.2	0.0006	99.0	100.0
			II	189779.2	3.3	3.4	0.0014	89.0	100.0
			III	188207.3	2.0	1.9	0.0069	83.3	83.3
0.75	267121.0	$\frac{1}{x_i}$	I	265820.2	5.4	5.6	0.0049	97.0	100.0
			II	267325.2	3.4	3.5	0.0008	95.0	100.0
			III	261041.7	2.5	2.2	0.0228	33.3	66.7
		$\frac{1}{x_i^2}$	I	264835.1	5.6	5.9	0.0086	96.0	100.0
			II	266522.3	3.7	3.8	0.0022	91.0	99.0
			III	266570.4	2.4	2.3	0.0021	100.0	100.0
		$\frac{1}{\pi_i x_i}$	I	269076.8	5.5	5.8	0.0073	96.0	100.0
			II	267284.8	3.5	3.6	0.0006	95.0	100.0
			III	268397.1	2.3	2.1	0.0048	83.3	83.3
		$\frac{1-\pi_i}{\pi_i x_i}$	I	267078.2	5.4	5.8	0.0002	97.0	100.0
			II	268263.9	3.5	3.7	0.0043	94.0	99.0
			III	262583.0	2.2	2.0	0.0170	66.7	83.3

Table 2 (Continued)

q	θ_q	Q_i	Method	AL		AVE	
				λ		λ	
				2	3	2	3
0.25	139199.3	$\frac{1}{x_i}$	I	24740.2	37586.5	38885999.2	39885095.1
			II	19332.2	29081.5	23862657.7	23998200.3
			III	13051.8	18071.7	11094502.8	9453381.2
		$\frac{1}{x_i^2}$	I	25206.9	38287.6	40255534.7	41273621.1
			II	18990.3	28555.9	23030807.0	23143000.8
			III	12558.3	17343.3	9915426.0	8412446.4
		$\frac{1}{\pi_i x_i}$	I	24406.8	37098.5	37939595.8	38953098.6
			II	19273.5	28993.5	23663302.3	23798503.3
			III	11903.8	16671.1	9022764.4	7867307.3
		$\frac{1-\pi_i}{\pi_i x_i}$	I	24849.9	37794.4	39276246.7	40384125.4
			II	19912.2	29950.7	25249814.7	25388215.2
			III	12303.7	17114.1	9644582.7	8277486.5
0.5	189519.8	$\frac{1}{x_i}$	I	38497.8	59017.4	95016816.4	98317087.2
			II	25257.5	38714.2	41324183.2	42895844.4
			III	16110.5	22138.7	16308973.6	13778144.7
		$\frac{1}{x_i^2}$	I	38982.0	59383.4	97399871.0	99454311.0
			II	25382.9	39335.7	41647270.1	44252803.7
			III	18697.0	26640.6	21895401.3	19801361.1
		$\frac{1}{\pi_i x_i}$	I	36367.6	56258.7	84781099.0	89492001.3
			II	25104.9	38603.4	40530120.7	42460330.0
			III	16695.7	23352.5	17862056.3	15476258.6
		$\frac{1-\pi_i}{\pi_i x_i}$	I	37942.7	58669.8	92060660.2	97040214.3
			II	24799.5	38406.4	39746087.7	42139418.0
			III	15168.9	21650.5	14762063.7	13393387.2
0.75	267121.0	$\frac{1}{x_i}$	I	57271.2	90001.0	209703313.6	229558665.9
			II	36802.2	56471.6	86473386.3	91000329.1
			III	25496.6	35124.9	41118752.6	34769371.0
		$\frac{1}{x_i^2}$	I	59307.5	93354.2	224194539.1	246577734.1
			II	38887.8	60749.0	97183071.2	106561992.7
			III	25743.9	35965.3	41672248.6	36221730.7
		$\frac{1}{\pi_i x_i}$	I	58728.3	93947.9	221507213.3	251407223.5
			II	37131.1	57150.3	88823045.4	94648089.2
			III	24130.9	33836.2	36726961.8	32086685.2
		$\frac{1-\pi_i}{\pi_i x_i}$	I	57669.8	93015.4	210946844.3	245076613.1
			II	37817.4	59251.0	93314579.0	102607466.9
			III	22886.1	31672.6	33183830.2	28221715.8

Acknowledgements

The authors are thankful to the reviewers and the editor for providing useful comments which have helped in improving the manuscript.

Disclosure Statement

The views expressed in this paper are personal and not of the Reserve Bank of India. The data used in this paper is publicly available.

References

Brewer KRW. A class of robust sampling designs for large-scale surveys. J Am Stat Assoc. 1979; 74(368): 911-915.

Chaudhuri A. Survey sampling. Boca Raton: CRC Press; 2018.

Chaudhuri A, Saha A. Extending Sitter's mirror-match bootstrap to cover Rao-Hartley-Cochran sampling in two-stages with simulated illustrations. Sankhya. 2004; 66(4): 791-802.

Chaudhuri A, Shaw P. A finite population quantile estimation by unequal probability sampling. Commun Stat Theory Methods. 2020; 49(22): 5419-5426.

Efron B. The jackknife, the bootstrap and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics; 1982.

Hartley HO, Rao JNK. Sampling with unequal probabilities and without replacement. Ann Math Stat. 1962; 33(2): 350-374.

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952; 47(260): 663-685.

Pal S. Rao and Wu's re-scaling bootstrap modified to achieve extended coverages. J Stat Plan Inference. 2009; 139(10): 3552-3558.

Rao JNK, Wu CFJ. Resampling inference with complex survey data. J Am Stat Assoc. 1988; 83(401): 231-241.