Mathematics

AIMS Mathematics, 10(2): 4457–4472.
DOI: 10.3934/math.2025205
Received: 06 December 2024
Revised: 15 February 2025
Accepted: 20 February 2025
Published: 28 February 2025

https://www.aimspress.com/journal/Math

Research article

Non-parametric calibration estimation of distribution function under stratified random sampling

Abdullah Mohammed Alomair[1]**, Weineng Zhu**[2]**, Usman Shahzad**[3,*] **and Fawaz Khaled Alarfaj**[4]

[1] Department of Quantitative Methods, School of Business, King Faisal University, Al-Ahsa 31982, Saudi Arabia
[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom
[3] Department of Management Science, College of Business Administration, Hunan University, Changsha 410082, China
[4] Department of Management Information Systems, School of Business, King Faisal University, Al-Ahsa 31982, Saudi Arabia

* **Correspondence:** Email: usman.stat@yahoo.com, usman@hnu.edu.cn.

Abstract: We introduced an innovative kernel-based nonparametric estimator for the cumulative distribution function (CDF) in finite populations, addressing the critical need to evaluate the proportion of values in a target variable that are less than or equal to specific thresholds. By leveraging auxiliary information under a stratified random sampling (StRS) framework, the proposed methodology employs multiple calibration constraints with a chi-square distance measure to derive calibrated weights, enhancing estimation efficiency. The estimators incorporate key descriptive measures of auxiliary variable, including the CDF and coefficient of variation, and tackle the challenge of bandwidth selection using advanced techniques such as plug-in selectors and cross-validation approaches. Simulation studies using datasets on apple production in Turkey and wheat production in Pakistan were conducted to assess the performance of the proposed estimators.

Keywords: cumulative distribution function (CDF); nonparametric; calibration approach; auxiliary information; stratified random sampling
Mathematics Subject Classification: 62A86, 62G07

1. Introduction

In the field of sampling surveys, the efficiency of estimators can be boosted for unknown population parameters by appropriately using auxiliary information. There are many estimators for estimating population parameters such as the mean, quantile, sum, distribution function, and median, which exist and require information regarding auxiliary variables in addition to the study variable parameters. Using auxiliary information in sampling theory is very beneficial as it enhances the efficiency of the estimator. Also, it is very prevalent and a regular practice in the field of sampling surveys as it plays a productive role in the development of sampling schemes.

There are many examples from daily life related to the linear relationship between the auxiliary variable and the study variable. For example, mass and weight are linearly related. Weight increases with mass. Similarly, there is a linear relationship between demand and the price of objects. The price also increases as demand increases. In such situations, auxiliary variables can be used to improve the estimation results of the study variables. Auxiliary information can be acquired in various forms from different sources, such as census data, outcomes of data (from previous experiments), and expert opinions. This information can be utilized in different methods. For instance, distributions of parameters of interest such as age, gender, and family income can be acquired using census data. Auxiliary information can be adequately used in the estimation phase, the design phase, or both stages. For further discussion on auxiliary information, interested readers may refer to Koyuncu [1], Abid et al. [2], Naz et al. [3], and Zaman and Kadilar [4].

When we use auxiliary information to estimate population parameters, it enhances the efficiency of the results. These estimates are based on traditional ratio, regression, and calibration methods. There are many studies available in the literature that encompass these methods for estimating population parameters using auxiliary information. For instance, Cochran [5] defined the ratio method of estimation, Watson [6] suggested the traditional regression estimation method, Deville and Sarndal [7] defined calibration-type estimators for parameter estimation, and Shahzad et al. [8,9] expanded on this work by introducing linear moments with calibration technique. These estimators are generally considered a reference to assess the efficiency of the proposed estimators by different authors. However, most of these researchers focus on estimating the mean and variance. In this paper, our focus is on estimating the CDF.

The issue arises regarding the CDF for the estimation of finite populations, especially in cases where the evaluation of the proportion of the data is based on the study variable, and the proportion can be more or less compared to a specific value. In such situations, it becomes necessary to estimate the CDF. To understand this concept, let us consider an example where some analysts suggest that the proportion of deaths due to COVID-19 is 2% or more of the total reported cases worldwide. Researchers have estimated the CDF using one or more auxiliary variables. For the estimation of CDF, Chambers and Dunstan [10] introduced an estimator that requires information from both the study and auxiliary variables for CDF estimation. Similarly, for CDF estimation using traditional sampling designs, Rao et al. [11] and [12] suggested ratio-type and regression type estimators. By utilizing auxiliary information for the estimation of CDF under the kernel method, Kuk [13] proposed an estimator. By using data with multiple auxiliary variables, Ahmed and Abu-Dayyeh [14] introduced the idea of estimating the CDF. In this paper, we estimate kernel-based CDF using multiple calibration constraints.

Calibration is a technique that is used to adjust the original weights. By following this technique, the original weights W_h are replaced by calibration weights V_h and adjusted. This advancement has greatly improved the efficiency of the CDF estimate. These adjusted weights are known as calibration weights. In this technique, the original weights W_h are upgraded using the chi-square and some other appropriate loss functions. However, chi-square is the most suitable loss function. This function depends on the appropriate calibration constraints that are associated with auxiliary variables. The idea of calibrations based on the estimation of parameters was initiated by Deville and Sarndal [7]. This concept was further modified by Tracy et al. [15] with the use of double stratified random sampling scheme. The extension of the idea was made by Koyuncu and Kadilar [16] by introducing some novel constraints by comparing the calibrated and original weights. Koyuncu [1] extended the idea by converting it into a rank set sampling technique for parameter estimation. Shahzad et al. [8,9] used the descriptive of linear moments (L-moments, TL moments) such as L-scale, L-location, and L-CV.

In this study, we were motivated to propose an estimator for the estimation of CDF using the multiple constraints-based calibration technique described above. It is worthy to note that we are using a kernel based nonparametric CDF function for the purposes of this article. From the literature mentioned above, we also know that the use of auxiliary information at the estimation stage provides better estimates. Therefore, keeping this fact in mind, calibration constraints on the basis of kernel-based nonparametric CDF, along with some traditional measures of descriptive statistics, namely the mean and coefficient of variation of the auxiliary variable, can provide better estimates. Based on this literature and following a calibration-based technique, an estimator for the population CDF of a study variable is proposed in this article using different bandwidths. The best estimator of the class is identified in light of different bandwidth selectors, and its efficiency is compared based on a number of simulation trials.

The remaining article is arranged in the following manner. Section 2 preliminarily consists of a kernel-based cumulative distribution function (CDF). Different bandwidth selection methods are also discussed in this section, such as Altman and Leger [17], Polansky and Baker [18] plug-in estimates, and cross validation bandwidth of Bowman et al. [19]. In section 3, empirical studies are conducted to determine the performance of the estimator and discuss the results. Our conclusion of this study is presented in section 4.

2. Kernel based CDF with different bandwidth selectors

2.1. Nonparametric CDF estimator

Suppose a data set of continuous random variable X, that is $(x_1, x_2, x_3, \ldots\ldots x_n)$ having density f and distribution function F. The empirical distribution function has a natural estimator that can be expressed at any point x in the following way,

$$F_n(y) = n^{-1} \sum_{j=1}^{n} I_{(-\infty,y]}(y_j) \tag{1}$$

The Rosenblatt-Parzen kernel estimator is also a renowned estimator of the density function and can be expressed as $\hat{f}_\lambda(y) = n^{-1} \sum_{j=1}^{n} k_\lambda(y - y_j)$, as indicated in Parzen [20]. However, $k_\lambda(u) = \lambda^{-1} k(u/\lambda)$. In this expression, k and λ signify the kernel and bandwidth of the function, respectively.

By using the connection between density and distribution function, the kernel estimator can be written as

$$\acute{F}_{\lambda}(y) = \int_{-\infty}^{y} \hat{f}_{\lambda}(t)dt \tag{2}$$

It can also be expressed in the form of kernel estimator of the density function as

$$\hat{F}_{\lambda}(y) = n^{-1} \sum_{j=1}^{n} H\left(\frac{y-y_j}{\lambda}\right) \tag{3}$$

Therefore, $H(y) = \int_{-\infty}^{y} k(t)dt$. Nadraya [21], Reiss [22], and Peter D. [23] have also given some theoretical properties for the estimator $\acute{F}_{\lambda}$.

In accordance with Eq (3), the kernel estimator depends on the kernel function k and smoothing parameter λ (bandwidth). It is not very problematic to select k, as functions can be used that give appealing results. However, the selection of bandwidth λ is more complicated. Choosing a small bandwidth may lead to an under-smoothed estimator with high variability. Therefore, selecting a large bandwidth results in low variability, making the estimator smoother and achieving the desired results.

The issue of bandwidth selection has been discussed in various methods and techniques, especially in regression and density estimation, with contributions made by Jones et al. [24] and del Rio [25]. Two well-known approaches in distribution estimation include the ‘plug-in bandwidth selection method’ by Altman and Leger [17] and Polanski and Baker [18], and ‘least squares cross validation method’ by Sarda [26]. Altman and Leger [17], however, noted that the latter demands a large sample size for valid results. Among all the cross-validation methods ever suggested, Bowman et al.’s [19] modified cross-validation is best suited for real-world datasets.

There are many applications of distribution function estimation in different fields e.g., hydrology, natural sciences, agricultural sciences, biological science, environmental sciences, and seismology. Thus, using nonparametric techniques, diverse methodologies have emerged that are directly associated with the risk term. Scientists have keen interest in knowing about the risk of a high magnitude earthquake, probability of the occurrence of a hurricane, and the risk of high-frequency flows.

2.2 Traditional bandwidth selectors

It is a well-known fact that bandwidth selection plays a vital role in nonparametric kernel-based methods. Some commonly used bandwidth selectors will be described in upcoming lines, which will be used for this article.

2.2.1. Plug-in bandwidth selection

The usual plug-in bandwidth selection procedure often works by minimizing a particular quadratic error between the estimator and the underlying true function like the mean integrated square error (MISE). After this, the selection of bandwidth minimizes the asymptotic estimation.

$$MISE\left(\acute{F_\lambda}\right) = \int_{-\infty}^{\infty} (\acute{F_\lambda}(y) - F(y))^2\, dy \tag{4}$$

According to Altman and Leger [17], under the assumptions of smoothness, it can be written as:

$$MISE\left(\acute{F_\lambda}\right) = \lambda^4 \int_{-\infty}^{\infty} C_U^2(y)\, dy + \frac{1}{n}\int_{-\infty}^{\infty} F(y)(1 - F(y)\, dy - \frac{\lambda}{n}\int_{-\infty}^{\infty} T_f^2(y) dy + o(MISE)(\lambda)) \tag{5}$$

Therefore,

$$C_F(y) = \frac{1}{2}(\acute{f}(y))^2 (\int_{-\infty}^{\infty} y^2 k(y) dy) \quad \text{and} \quad T_F^2(y) = 2f(y)(\int_{-\infty}^{\infty} yk(y)G(y)dy) \tag{6}$$

This can be in an asymptotically optimal bandwidth as:

$$\lambda_{AMISE}\left(\hat{F}_\lambda\right) = Bn^{-1/3} = (\frac{\frac{1}{2}\int_{-\infty}^{\infty} T_F^2(y)dy}{\int_{-\infty}^{\infty} C_F^2(y)dy})^{1/3}\, n^{-1/3} \tag{7}$$

According to Eq (7), the optimal bandwidth order is $n^{-1/3}$ than $n^{-1/5}$. As per Silverman's [27] recommendation, this is the optimal order for kernel nonparametric density estimation. However, in the case of nonparametric distribution estimation, as the sample size increases, the optimal bandwidth decreases compared to density estimation. A smaller bandwidth size for nonparametric density estimation leads to a closer estimation of the actual density. Therefore, the X-axis and the area under the estimated curve provide better estimation results for the actual area.

Therefore, a large bandwidth is recommended to achieve a smoother estimator. The value of C in Eq (7) is based on the kernel function.

$$\hat{\lambda} = \hat{C} n^{-1/3} \tag{8}$$

where $\hat{C}$ represents data sample.

2.2.2. AL_{bw} (Plug-in bandwidth of Altman and Leger)

Altman and Leger [17] introduced a plug-in technique that is commonly known as a nonparametric estimation technique used to estimate the unknown terms of Eq (7), which represents a function with asymptotically optimal bandwidths. By applying the technique of Altman and Leger [17], Eq (7) can be expressed as

$$\lambda_{AMISE}\left(\hat{F}_\lambda\right) = (\frac{\frac{1}{4}T_2}{C_3})^{1/3}\, n^{-1/3}$$

$$T_2 = \emptyset(k) \int_{-\infty}^{\infty} [f(y)]^2 dy \qquad \emptyset(k) = 2\int_{-\infty}^{\infty} yk(y)G(y)dy \tag{9}$$

$$C_3 = \frac{1}{4}\left(\omega_2(k^2)\right) \int_{-\infty}^{\infty} [f'(y)]^2 f(y) dy\, ; \omega_2(k) = \int_{-\infty}^{\infty} y^2 k(y) dy$$

The plug-in bandwidth can be written as:

$$\lambda_{AL}\left(\hat{F}_{\lambda}\right) = \left(\frac{\frac{1}{4}T_2}{C_3}\right)^{1/3} n^{-1/3} \tag{10}$$

Therefore,

$$\hat{T}_2 = \emptyset(k)\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq 1}^{n}\frac{1}{\gamma}k\left(\frac{y_i-y_j}{\gamma}\right) \tag{11}$$

$$\hat{C}_3 = \frac{1}{4}\hat{Z}_3(F)(\omega_2(k))^2 \tag{12}$$

Thus,

$$\hat{Z}_3(F) = \frac{1}{n^3\gamma_b^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}k_b'\left(\frac{y_i-y_j}{\gamma_b}\right)k_b'\left(\frac{y_i-y_k}{\gamma_b}\right). \tag{13}$$

In the above function, k_b' represents the derivative of the kernel function k_b. It is not required for $k_b$to be equal to k. The bandwidth parameter related to this is represented by γ_b. For implementation, it can be chosen as $\gamma_b = \gamma$ and $k_b = k$.

2.2.3. PBbw (Polansky and Baker Plug-in bandwidth)

For the estimation of $R(f')$, a nonparametric estimate is used. Mathematically, it can be written as

$$\varphi_m = \int_{-\infty}^{\infty} f^m(y)\, f(y)dy, \tag{14}$$

Therefore, m is considered an even integer, and $m \geq 2$. Integration by parts is applied by considering the adequate smoothness assumptions on f, and we get $R(f^{(a)} = (-1)^a\varphi_{2y}$. The concept of kernel estimates for φ_m was initiated by Hall and Marron [28], and Jones and Sheather [29] amended the idea. By utilizing the "diagonal-in" method to estimate φ_m, it can be written as:

$$\hat{\varphi}_m(v) = n^{-2}v^{-m-1}\sum_{i=1}^{n}\sum_{j=1}^{n}H^m\left\{\frac{Y_i-Y_j}{v}\right\} \tag{15}$$

where H represents the kernel function, and it is not required for L to be equal to k. However, v is a positive parameter that represents bandwidth and is known as a smoothing parameter. According to the conditions of $v \to 0$ and $nv^{2m+1} \to \infty$ as $n \to \infty$, the bandwidth factor v minimizes $E[\{\hat{\varphi}_m(v) - \varphi_m\}^2]$ and was introduced by Jones and Sheather [29]. It can be written as:

$$v_m = \left[\frac{2H^m(0)}{-n\omega_2(H)\varphi_{m+2}}\right]^{1/m+3} \tag{16}$$

The findings of Eq (16) are utilized to find the estimate of the following function

$$\lambda_d = \left[\frac{\emptyset(k)}{n\omega_2^2(k)R(f')}\right]^{1/3} \tag{17}$$

By solving Eqs (16) and (17), we get:

$$\hat{\lambda}_d = [\frac{\emptyset(k)}{-n\omega_2^2(k)\hat{\varphi}_2(v_2)}]^{1/3} \tag{18}$$

However

$$v_2 = [\frac{2H^{(2)}(0)}{-n\omega_2(H)\varphi_4}]^{1/5} \tag{19}$$

We found that v_2 relies on f succeeding φ_4, which is also necessary to estimate. This can be performed by estimating φ_4 with $\hat{\varphi}_4(v_4)$; therefore, bandwidth factor φ_4 relies on φ_6 continuously. According to Sheather and Jones [29], it is also important to estimate φ_m at some step by using some distribution; generally, a normal distribution is taken as a reference distribution. Consider the function f as normal i.e., the mean of f is μ and variance is σ^2. It can be written as:

$$\varphi_m = \frac{(-1)^{r/2}r!}{(2\emptyset)^{m+1}(m/2)!\pi^{1/2}} \tag{20}$$

Thus, the normal scale estimate of φ_m can be written as:

$$\hat{\varphi}_m^{NR} = \frac{(-1)^{r/2}r!}{(2\hat{\emptyset})^{m+1}(m/2)!\pi^{1/2}} \tag{21}$$

Therefore, $\hat{\emptyset}$ is either considered the standard deviation of sample or $\hat{\emptyset} = \min\{S, \frac{IQR}{1.349}\}$.

In this situation, it is recommended to use a c-stage estimator of λ_d that can be estimated by applying the algorithm outlined below. Here, c is an integer greater than zero $(c > 0)$. The estimation process is divided into following steps:

1) Evaluate $\hat{\varphi}_{2c+2}^{NR}$ using Eq (21)
2) Use $j = b$ for evaluation and continuously proceed the iterations until $j = 1$, then compute $\hat{\varphi}_{2j}\left(\hat{v}_{2j}\right)$, as follows:

$$\hat{v}_{2j} = [\frac{2H^{(2j)}(0)}{-n\omega_2(H)\hat{\varphi}_{2j+2}}]^{1/(2j+3)} \tag{22}$$

where

$$\hat{\varphi}_{2j+2} = \begin{cases} \hat{\varphi}_{2c+2}^{NR} & when\ j = b \\ \hat{\varphi}_{(2j+2)}\left(\hat{v}_{2j+2}\right) & when\ j < b \end{cases} \tag{23}$$

Evaluate

$$\hat{\lambda}_c = [\frac{\emptyset(k)}{-n\omega_2^2(k)\hat{\varphi}_2(\hat{v}_2)}]^{1/3} \tag{24}$$

This results in the c-stage estimator.

2.2.4. CV_{bw} (Cross-Validation bandwidth of Bowman et al.)

There are many approaches for selecting the bandwidth for kernel smoothing of distribution functions. Sarda [26] and Altman and Leger [17] suggested the “plug-in” and “leave-one-out” methods primarily used for density estimation. In contrast, Bowman et al. [19] proposed a cross-validation method that is more advantageous for smoothing distribution functions. This approach proved to be a more accurate analogue compared to other density estimation approaches. The bandwidth selection parameter depends on the unbiased estimation of MISE. According to Bowman et al. [19], this method minimizes the MISE value, indicating an optimal smoothing parameter. Thus, it demonstrates asymptotically optimal bandwidth selection, where kernel approaches enhance the general empirical distribution function.

In nonparametric statistics, the cross-validation technique is based on the estimated value of MISE of the function. After this, selection of bandwidth is to be done for minimization of this function. According to Sarda [26],

$$CV(\lambda) = \sum_{i=1}^{n}(U_n(x_i) - \hat{U}_{-i}(x_i))^2 \tag{25}$$

In the above equation, $CV(.)$ is used to minimize the differences among the observed distribution function $U_n(x) = n^{-1}\sum_{j=1}^{n} I_{(-\infty,x]}(x_j)$ and leave-one-out class of the kernel distribution estimator. The following estimator is expressed as an estimator that uses all the particulars (points) apart from x_i,

$$\hat{U}_{-i}(x) = \frac{1}{n-1}\sum_{j\neq 1} H\left(\frac{x-x_j}{\lambda}\right) \tag{26}$$

Regardless of the numerical value of asymptotic optimality given by Sarda [26], this technique does not yield practically effective results. Therefore, the cross-validation technique improved by Bowman et al. [19] provides better results in simulation studies. This technique is asymptotically optimal and involves minimizing the function.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(I(x - x_i) - \hat{U}_{-i}(x))^2 dx \tag{27}$$

In this function, $I(x - x_i) = 1$, therefore, $x - x_i \geq 0$ and elsewhere it is 0. A simulation study was carried out by Bowman [19] who compared the results with the plug-in-one method suggested by Altman and Leger [17]. The drawback of this method is that it affects the performance related to computational time. By using the method of cross-validation, the minimization of the function is carried out for the term n^2. Thus, it is necessary to search for a larger bandwidth grid. However, in the evaluation of an integral term, it provides good results.

It is worthy to note that we will use all three bandwidth selectors, AL_{bw}, PB_{bw}, and CV_{bw}, in our proposed estimator analysis in the upcoming section.

3. Calibration approach and estimation of distribution function

In this section, we present a new method for estimating the cumulative distribution function (CDF), incorporating distribution function insights and auxiliary variable descriptive measures under stratified random sampling. Comparative results reveal that the calibration-based estimator achieves higher accuracy than conventional techniques in finite population scenarios. The proposed estimator is

$$F_{st(j)} = \sum_{h=1}^{H''} V_h F'_\lambda(y_h) \quad \text{for j = 1,2} \tag{28}$$

where V_h represents calibration weights and $F'_\lambda(y_h)$ stands for the kernel-based CDF in this formulation. Now, considering chi-square loss function:

$$L(V_h, \pi_h) = \sum_{h=1}^{H''} \frac{(V_h - \pi_h)^2}{\pi_h \triangle_h} \tag{29}$$

and using these calibration constraints

$$\sum_{h=1}^{H''} V_h \hat{\mu}_{xh} = \sum_{h=1}^{H''} \pi_h \mu_{xh} \tag{30}$$

$$\sum_{h=1}^{H''} V_h F'_\lambda(x_h) = \sum_{h=1}^{H''} \pi_h F(x_h) \tag{31}$$

$$\sum_{h=1}^{H''} V_h c_{xh} = \sum_{h=1}^{H''} \pi_h C_{xh} \tag{32}$$

where c_{xh} is the CV of the sample auxiliary variable X. To elaborate on various forms of estimators, $\triangle_h$ is an appropriately selected weight. For more information about appropriate weight selection and the use of descriptive measures, refer to Koyuncu [1] and Shahzad et al. [8,9].

The Lagrange function can be written as:

$$\Omega = \sum_{h=1}^{H''} \frac{(V_h - \pi_h)^2}{\pi_h \triangle_h} - 2\theta'_1\left(\sum_{h=1}^{H''} V_h \hat{\mu}_{xh} - \sum_{h=1}^{H''} \pi_h \mu_{xh}\right) - 2\theta'_2\left(\sum_{h=1}^{H''} V_h F'_\lambda(x_h) - \sum_{h=1}^{H''} \pi_h F(x_h)\right) - 2\theta'_3\left(\sum_{h=1}^{H''} V_h c_{xh} - \sum_{h=1}^{H''} \pi_h C_{xh}\right) \tag{33}$$

The Chi-square loss function Eq (29) is minimized by considering the calibration constraints Eqs (30)–(32), which provide the calibration weights in the case of stratified sampling.

$$v = \pi_h + \pi_h \triangle_h \left(\theta'_1 \hat{\mu}_{xh} + \theta'_2 F'_\lambda(x_h) + \theta'_3 c_{xh}\right) \tag{34}$$

By substituting Eq (34) into Eqs (30)–(32), the given relations are provided.

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & A_{22} & A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} \begin{bmatrix} \theta'_1 \\ \theta'_2 \\ \theta'_3 \end{bmatrix} = \begin{bmatrix} A_{10} \\ A_{20} \\ A_{30} \end{bmatrix} \tag{35}$$

By solving the above system of Eq (35) for θ'_s , we get

$$\theta'_1 = \frac{(A_{13}A_{23} - A_{12}A_{33})(A_{12}A_{20} - A_{22}A_{10}) - (A_{13}A_{22} - A_{12}A_{23})(A_{12}A_{30} - A_{23}A_{10})}{(A^2{}_{12} - A_{11}A_{22})(A_{13}A_{23} - A_{12}A_{33}) - (A_{13}A_{22} - A_{12}A_{23})(A_{12}A_{13} - A_{11}A_{23})}$$

$$\theta'_2 = \frac{(A_{13}A_{23} - A_{12}A_{33})(A_{12}A_{10} - A_{11}A_{20}) - (A_{12}A_{13} - A_{23}A_{11})(A_{13}A_{20} - A_{12}A_{30})}{(A^2{}_{12} - A_{11}A_{22})(A_{13}A_{23} - A_{12}A_{33}) - (A_{13}A_{22} - A_{12}A_{23})(A_{12}A_{13} - A_{11}A_{23})}$$

$$\theta'_3 = \frac{(A^2{}_{12} - A_{11}A_{22})(A_{13}A_{20} - A_{12}A_{30}) - (A_{13}A_{22} - A_{12}A_{23})(A_{12}A_{10} - A_{11}A_{20})}{(A^2{}_{12} - A_{11}A_{22})(A_{13}A_{23} - A_{12}A_{33}) - (A_{13}A_{22} - A_{12}A_{23})(A_{12}A_{13} - A_{11}A_{23})}$$

whereas

$$A_{11} = \sum_{h=1}^{H''} \pi_h \triangle_h \hat{\mu}_{xh} \ , A_{22} = \sum_{h=1}^{H''} \pi_h \triangle_h F'^2_\lambda x_h, \quad A_{33} = \sum_{h=1}^{H''} \pi_h \triangle_h c^2 x_h,$$

$$A_{12} = \sum_{h=1}^{H''} \pi_h \triangle_h \hat{\mu}_{xh} F'_\lambda(x_h) \ , \quad A_{13} = \sum_{h=1}^{H''} \pi_h \triangle_h \hat{\mu}_{xh} c_{xh}, \quad A_{23} = \sum_{h=1}^{H''} \pi_h \triangle_h F'_\lambda(x_h) c_{xh}$$

$$A_{10} = \sum_{h=1}^{H''} \pi_h \mu_{xh} - \hat{\mu}_{xh} \ , \quad A_{20} = \sum_{h=1}^{H''} \pi_h F(x_h) - F'_\lambda(x_h), \quad A_{30} = \sum_{h=1}^{H''} \pi_h (C_{xh} - c_{xh})$$

By substituting the value of θ'_s into Eq (34), we obtain Eq (28). Therefore, considering $\triangle_h$ = 1, we get the proposed non-parametric kernel estimator $F_{st(j)}$ for population CDF as given below:

$$F_{st(j)} = \sum_{h=1}^{H''} \pi_h F'_\lambda(y_h) + E_{1h} A_{10} + E_{2h} A_{20} + E_{3h} A_{30} \tag{36}$$

whereas

$$E_{1h} = \frac{d_{12}[d_{14}(d_{22}d_{33} - d^2{}_{23}) + d_{24}(d_{13}d_{23} - d_{12}d_{23}) + d_{34}(d_{12}d_{23} - d_{13}d_{22})]}{(d^2{}_{12} - d_{11}d_{22})(d_{13}d_{23} - d^2{}_{12}) - (d_{13}d_{22} - d_{12}d_{23})(d_{12}d_{23} - d_{11}d_{23})}$$

$$E_{2h} = \frac{d_{12}[d_{14}(d_{13}d_{23} - d_{12}d_{33}) + d_{24}(d_{11}d_{33} - d^2{}_{13}) + d_{34}(d_{12}d_{13} - d_{11}d_{23})]}{(d^2{}_{12} - d_{11}d_{22})(d_{13}d_{23} - d^2{}_{12}) - (d_{13}d_{22} - d_{12}d_{23})(d_{12}d_{13} - d_{11}d_{23})}$$

$$E_{3h} = \frac{d_{12}[d_{14}(d_{13}d_{22} - d_{12}d_{23}) + d_{24}(d_{12}d_{13} - d_{11}d_{23}) + d_{34}(d_{11}d_{22} - d^2{}_{12})]}{(d^2{}_{12} - d_{11}d_{22})(d_{13}d_{23} - d^2{}_{12}) - (d_{13}d_{22} - d_{12}d_{23})(d_{12}d_{13} - d_{11}d_{23})}$$

$$d_{11} = \sum_{h=1}^{H''} \pi_h \hat{\mu}^2{}_{xh} \ , \qquad d_{22} = \sum_{h=1}^{H''} \pi_h F'^2_\lambda(x_h) \ , d_{33} = \sum_{h=1}^{H''} \pi_h c^2{}_{xh}$$

$$d_{12} = \sum_{h=1}^{H''} \pi_h \hat{\mu}_{xh} F'_\lambda(x_h) \ , \qquad d_{13} = \sum_{h=1}^{H''} \pi_h \hat{\mu}_{xh} c_{xh} \ , \ d_{14} = \sum_{h=1}^{H''} \pi_h \hat{\mu}_{xh} F'_\lambda(y_h),$$

$$d_{23} = \sum_{h=1}^{H''} \pi_h F'_\lambda(x_h) c_{xh} \ , \quad d_{24} = \sum_{h=1}^{H''} \pi_h F'_\lambda(x_h) F'_\lambda(y_h) \ , \quad d_{34} = \sum_{h=1}^{H''} \pi_h c_{xh} F'_\lambda(y_h)$$

Now, the performance of the proposed non-parametric kernel estimator $F_{st(j)}$ of the CDF will be assessed using a simulation study in the next sub-sections.

3.1. Numerical study

We evaluated the proposed estimator in this section through different populations. For the simulation, two datasets from different populations were considered to ensure that the efficiency of the suggested estimator surpasses that of traditional estimators. In the population-1 dataset, apple fruit production data from 1999, with respect to the number of apple trees, was used from 4 different regions. In the population-2 dataset, we considered data on wheat production in Pakistan from 1960 to 2020 with respect to the area used for wheat cultivation each year. The Percentage Relative Efficiency (PRE) of the proposed estimators was computed as:

$$MSE\left(F_{st(j)}\right) = \frac{1}{\binom{N}{n}} \sum_{i-1}^{\binom{N}{n}} \left(F_{st(j)} - F(y_h)\right)$$

where

$$F(y_h) = \frac{1}{\binom{N}{n}} \sum_{i-1}^{\binom{N}{n}} \left(F_{st(j)}\right)$$

Therefore,

$$PRE\left(F_{st(j)}\right) = \frac{MSE(d_0)}{MSE\left(F_{st(j)}\right)} \times 100$$

3.1.1. Population-1

In the first analysis, we used the dataset of apple fruit for a simulation [30]. Here, X represents the number of trees. We set the scale such that 100 trees are considered 1 unit. Therefore, Y represents the production quantity. According to the scale settings, 100 tonnes are equal to 1 unit. It is important to note that 477 villages are considered in 4 strata: Stratum 1 represents Marmarian, Stratum 2 indicates Aegean, Stratum 3 shows Mediterranean, and Stratum 4 denotes Central Anatolia. The PREs of the suggested estimators are evaluated using the mentioned dataset. The value of PRE for AL$_{bw}$ is 102.4369, PB$_{bw}$ is 106.949, and CV$_{bw}$ is 109.8822, as shown in Figure 1.

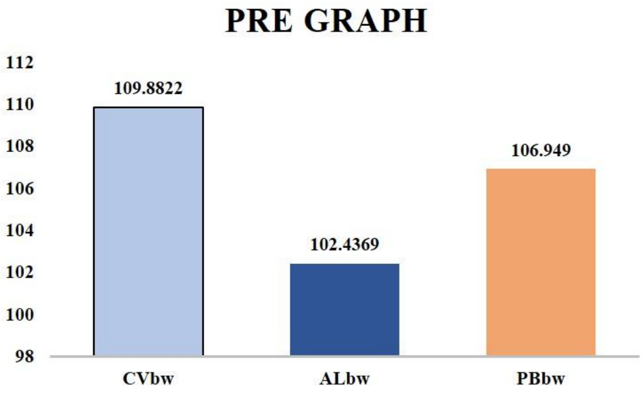

Figure 1. PRE for apple data.

3.1.2. Population-2

In the second empirical analysis of the estimators, we considered a dataset consisting of two variables, X and Y. Variable X denotes the production of wheat crop in Pakistan from the year 1960 to 2020, while Y represents the area under cultivation every year. Four strata are created, each representing one province of Pakistan. Stratum 1 represents the province of Punjab, Stratum 2 denotes the province of Sindh, Stratum 3 shows the province of KPK, and Stratum 4 indicates the province of Baluchistan. PREs for the proposed estimators are calculated using this data. The PRE for AL$_{bw}$ is 103.6978, PB$_{bw}$ is 102.9949, and CV$_{bw}$ is 104.0085, as indicated in Figure 2.

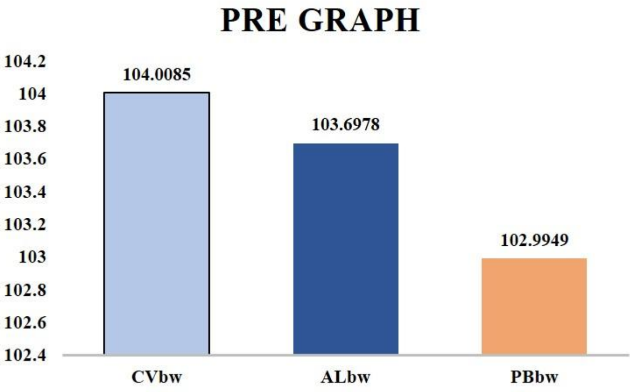

Figure 2. PRE for wheat data.

3.2. Simulation study

For the simulation study, a population size of 500 is considered with a total sample size of 100 from both strata. A sample of size 50 is selected from each stratum using equal allocation. The auxiliary variable X for the first and second stratum is generated using gamma distributions with parameters $G(2.5, 3.7)$ and $G(1.9, 2.9)$. The study variable Y is generated for each stratum as follows:

$$Y_h = T + WX_h + JX_h^e,$$

where J follows a standard normal distribution, and $T = 4, e = 1.6,\ and\ W = 2$. For the detailed steps of the simulation study, interested readers may refer to Shahzad et al. [8,9]. The PRE for ALbw using a simulation study is 102.3112, for PBbw it is 101.2999, and for CVbw it is 106.0021, as shown in Figure 3.

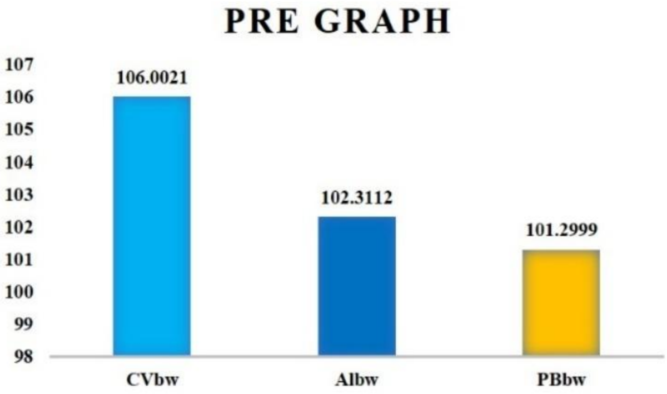

Figure 3. PRE for simulated data.

3.3. Results and discussion

Figures 1–3 show the results and indicate that the proposed kernel-based non-parametric estimator outperforms traditional methods very well in PRE values larger than 100 for all datasets. The results demonstrate the estimator's effectiveness in improving CDF estimation with stratified random sampling. Additionally, calibration constraints and auxiliary information are included more effectively for improved efficiency.

In this study, the CVbw method proved to be the most reliable bandwidth selection method of all the tested bandwidth selection methods, as it was the most consistent in terms of efficiency. CVbw performed better than both PBbw and ALbw, indicating that the adaptive bandwidth selection in CVbw-based algorithms is better for optimizing the estimator. Robust numerical results are derived from the proposed estimator, and it is a promising solution to improve the estimation accuracy in biological sciences, agriculture, and food sciences, where auxiliary variables could be used to enhance the estimation quality (tree count, land area, etc.). It is also likely to be useful in environmental studies, finance, and social sciences, given the necessity of accurate estimations of distribution.

3.4. Limitations of the study

The proposed estimator is shown to possess superior efficiency and robustness; however, some limitations should be noted. This methodology is based on the availability and quality of auxiliary information and can be affected if these data are incomplete or inaccurate.

4. Conclusions

We present a kernel-based nonparametric estimator of the CDF under the framework of finite population estimation for stratified random samples. The use of auxiliary information and the application of calibration constraints improve the accuracy and robustness of the proposed methodology with the help of a chi-square loss function. Comparisons of results from simulation analyses involving apple production data from Turkey and wheat production data from Pakistan also reflect the increased performance of the proposed estimator when PREs are considered, with values above 100%. These results support the reliability of the study and the real-life applicability of the estimator in various industries such as agricultural research. The graphical results also emphasize how efficient the estimator is in helping to solve real-life problems in the estimation of CDF. This leads to a discussion of how the presented work can be used for the following studies: The improvement and expansion of the methodology to suit more complex population structures and the examination of the domains relevant to further fields of environmentalism, social issues, and other branches of science. In future studies, the work can be extended for cluster or multi-stage sampling in light of [31–33]. Moreover, additional theoretical studies on loss function choice beyond the chi-square approach could be pursued, namely entropy based measures, to make the estimator robust.

Author contributions

Abdullah Mohammed Alomair: Conceptualization, Methodology, Investigation, Funding acquisition, Writing original draft; Usman Shahzad: Conceptualization, Methodology, Investigation, Writing original draft; Weineng Zhu: Conceptualization, Writing-review & editing, Investigation; Fawaz Khaled Alarfaj: Methodology, Writing-eview & editing. All authors have read and approved the final manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors thank and extend their appreciation to the funder of this work: the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU250644].

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU250644].

Conflict of interest

The authors declare no competing interests.

References

1. N. Koyuncu, Calibration estimator of population mean under stratified ranked set sampling design, *Commun. Stat.-Theor. M.*, **47** (2018), 5845–5853. https://doi.org/10.1080/03610926.2017.1402051
2. M. Abid, S. Ahmed, M. Tahir, H. Z. Nazir, M. Riaz, Improved ratio estimators of variance based on robust measures, *Sci. Iran.*, **26** (2019), 2484–2494. https://doi.org/10.24200/sci.2018.20604
3. F. Naz, T. Nawaz, T. Pang, M. Abid, Use of nonconventional dispersion measures to improve the efficiency of ratio-type estimators of variance in the presence of outliers, *Symmetry*, **12** (2020), 16. https://doi.org/10.3390/sym12010016
4. T. Zaman, C. Kadilar, Exponential ratio and product type estimators of the mean in stratified two-phase sampling, *AIMS Math.*, **6** (2021), 4265–4279. https://doi.org/10.3934/math.2021252
5. W. G. Cochran, The estimation of the yields of cereal experiments by sampling for the ratio gain to total produce, *J. Agr. Sci.*, **30** (1940), 262–275. https://doi.org/10.1017/S0021859600048012
6. D. J. Watson, The estimation of leaf area in field crops, *J. Agr. Sci.*, **27** (1937), 474–483. https://doi.org/10.1017/S002185960005173X
7. J.-C. Deville, C. E. Särndal, Calibration estimators in survey sampling, *J. Amer. Stat. Assoc.*, **87** (1992), 376–382. https://doi.org/10.1080/01621459.1992.10475217

8. U. Shahzad, I. Ahmad, I. Almanjahie, N. H. Al-Noor, M. Hanif, A new class of L-moments based calibration variance estimators, *CMC-Comput. Mater. Con.*, **66** (2021), 3013–3028. https://doi.org/10.32604/cmc.2021.014101
9. U. Shahzad, I. Ahmad, I. Almanjahie, N. H. Al-Noor, L-moments based calibrated variance estimators using double stratified sampling, *CMC-Comput. Mater. Con.*, **68** (2021), 3411–3430. https://doi.org/10.32604/cmc.2021.017046
10. R. L. Chambers, R. Dunstan, Estimating distribution functions from survey data, *Biometrika*, **73** (1986), 597–604. https://doi.org/10.1093/biomet/73.3.597
11. J. N. K. Rao, J. G. Kovar, H. J. Mantel, On estimating distribution functions and quantiles from survey data using auxiliary information, Biometrika, **77** (1990), 365–375. https://doi.org/10.1093/biomet/77.2.365
12. J. N. K. Rao, Estimating totals and distribution functions using auxiliary information at the estimation stage, *J. Off. Stat.*, **10** (1994), 153–165.
13. A. Y. C. Kuk, A kernel method for estimating finite population distribution functions using auxiliary information, Biometrika, **80** (1993), 385–392. https://doi.org/10.1093/biomet/80.2.385
14. M. S. Ahmed, W. Abu-Dayyeh, Estimation of finite-population distribution function using multivariate auxiliary information, *Statistics in Transition*, **5** (2001), 501–507.
15. D. S. Tracy, S. Singh, R. Arnab, Note on calibration in stratified and double sampling, *Surv. Methodol.*, **29** (2003), 99–104.
16. N. Koyuncu, C. Kadilar, Calibration weighting in stratified random sampling, *Commun. Stat.-Simul. C.*, **45** (2016), 2267–2275. https://doi.org/10.1080/03610918.2014.901354
17. N. Altman, C. Léger, Bandwidth selection for kernel distribution function estimation, *J. Stat. Plan. Infer.*, **46** (1995), 195–214. https://doi.org/10.1016/0378-3758(94)00102-2
18. A. M. Polansky, E. R. Baker, Multistage plug-in bandwidth selection for kernel distribution function estimates, *J. Stat. Comput. Sim.*, **65** (2000), 63–80. https://doi.org/10.1080/00949650008811990
19. A. W. Bowman, P. Hall, T. Prvan, Cross-validation for the smoothing of distribution functions, Biometrika, **85** (1998), 799–808. https://doi.org/10.1093/biomet/85.4.799
20. E. Parzen, On estimation of a probability density function and mode, *Ann. Math. Statist.*, **33** (1962), 1065–1076. https://doi.org/10.1214/aoms/1177704472
21. E. A. Nadaraya, On estimating regression, *Theor. Probab. Appl.*, **9** (1964), 141–142. https://doi.org/10.1137/1109020
22. R.-D. Reiss, Nonparametric estimation of smooth distribution functions, *Scand. J. Stat.*, **8** (1981), 116–119.
23. P. D. Hill, Kernel estimation of a distribution function, *Commun. Stat.-Theor. M.*, **14** (1985), 605–620. https://doi.org/10.1080/03610928508828937
24. M. C. Jones, J. S. Marron, S. J. Sheather, A brief survey of bandwidth selection for density estimation, *J. Amer. Stat. Assoc.*, **91** (1996), 401–407. https://doi.org/10.1080/01621459.1996.10476701
25. A. Q. del Rio, Comparison of bandwidth selectors in nonparametric regression under dependence, *Comput. Stat. Data Anal.*, **21** (1996), 563–580. https://doi.org/10.1016/0167-9473(95)00028-3
26. P. Sarda, Smoothing parameter selection for smooth distribution function, *J. Stat. Plan. Infer.*, **35** (1993), 65–75. https://doi.org/10.1016/0378-3758(93)90068-H

27. B. W. Silverman, *Density estimation for statistics and data analysis*, New York: Chapman and Hall, 1998. https://doi.org/10.1201/9781315140919
28. P. Hall, J. S. Marron, Estimation of integrated squared density derivatives, *Stat. Probabil. Lett.*, **6** (1987), 109–115. https://doi.org/10.1016/0167-7152(87)90083-6
29. M. C. Jones, S. J. Sheather, Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Stat. Probabil. Lett.*, **11** (1991), 511–514. https://doi.org/10.1016/0167-7152(91)90116-9
30. N. Koyuncu, C. Kadilar, Ratio and product estimators in stratified random sampling, *J. Stat. Plan. Infer.*, **139** (2009), 2552–2558. https://doi.org/10.1016/j.jspi.2008.11.009
31. T. H. Ali, Modification of the adaptive Nadaraya-Watson kernel method for nonparametric regression (simulation study), *Commun. Stat.-Simul. C.*, **51** (2022), 391–403. https://doi.org/10.1080/03610918.2019.1652319
32. T. H. Ali, H. A. A. M. Hayawi, D. S. I. Botani, Estimation of the bandwidth parameter in Nadaraya-Watson kernel non-parametric regression based on universal threshold level, *Commun. Stat.-Simul. C.*, **52** (2023), 1476–1489. https://doi.org/10.1080/03610918.2021.1884719
33. U. Shahzad, I. Ahmad, I. M. Almanjahie, N. H. Al-Noor, M. Hanif, Adaptive Nadaraya-Watson kernel regression estimators utilizing some non-traditional and robust measures: a numerical application of British food data, *Hacet. J. Math. Stat.*, **52** (2023), 1425–1437. https://doi.org/10.15672/hujms.1167617

AIMS Press