Ann Inst Stat Math
https://doi.org/10.1007/s10463-018-0668-7

Simultaneous confidence bands for the distribution function of a finite population in stratified sampling

Lijie Gu[1] · **Suojin Wang**[2] · **Lijian Yang**[3]

Received: 16 August 2017 / Revised: 24 April 2018

Abstract Stratified sampling is one of the most important survey sampling approaches and is widely used in practice. In this paper, we consider the estimation of the distribution function of a finite population in stratified sampling by the empirical distribution function (EDF) and kernel distribution estimator (KDE), respectively. Under general conditions, the rescaled estimation error processes are shown to converge to a weighted sum of transformed Brownian bridges. Moreover, simultaneous confidence bands (SCBs) are constructed for the population distribution function based on EDF and KDE. Simulation experiments and illustrative data example show that the coverage frequencies of the proposed SCBs under the optimal and proportional allocations are close to the nominal confidence levels.

Keywords Confidence band · Stratified population distribution · Allocation · Brownian bridge · Kernel · Superpopulation

This research was supported in part by Jiangsu Specially-Appointed Professor Program SR10700111, Jiangsu Province Key-Discipline Program ZY107992, National Natural Science Foundation of China Awards NSFC 11371272, 11771240, 11701403, Research Fund for the Doctoral Program of Higher Education of China Award 20133201110002, 2017 Jiangsu Overseas Visiting Scholar Program for University Prominent Young and Middle-aged Teachers and Presidents, and the Simons Foundation Mathematics and Physical Sciences Program Award #499650. Helpful comments from a reviewer are greatly appreciated.

✉ Lijian Yang
yanglijian@mail.tsinghua.edu.cn

1 School of Mathematical Sciences and Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou 215006, China

2 Department of Statistics, Texas A&M University, College Station, TX 77843, USA

3 Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

Published online: 21 May 2018

1 Introduction

Survey sampling is one of the most important branches of statistics. Its classic research contents are to design appropriate sampling methods to obtain some units from the finite population, and to make inference about the population using the sampled observations, such as the population total, population mean, and population quantiles; see Cochran (1977) and Lohr (2009). Therefore, one particular task is to estimate the cumulative distribution function (CDF) of the finite population and make corresponding inferences.

There are many approaches in the current literature for estimating the population distribution function under various conditions; see, for instance, Chambers and Dunstan (1986), Wang and Dorfman (1996), Chen and Wu (2002), Frey (2009), and O'Neill and Stern (2012) . Most of these existing works focus on studying the consistency and asymptotic properties at any fixed point of the proposed distribution estimator. However, this is often unsatisfactory as investigators may be interested in making statistical inferences on the unknown distribution function or testing whether the population distribution function is significantly different from a known distribution, for which a simultaneous confidence band (SCB) is desirable.

SCB is a powerful and appropriate inference tool for an entire unknown curve or function, which is a direct analogous concept of a confidence interval, regarded as a collection of confidence intervals over the whole range of function. For instance, the earlier work of Bickel and Rosenblatt (1973) is about SCB for a probability density function. The constructions of SCB for nonparametric regression function can be found in Härdle (1989), Xia (1998), Wang and Yang (2009). Recent theoretical development on SCB has appeared in various contexts, see Degras (2011), Zhu et al. (2012), Cao et al. (2012), Ma et al. (2012), Zheng et al. (2014), Gu et al. (2014), Song et al. (2014), and Cao et al. (2016) for functional data analysis; Song and Yang (2009) and Cai and Yang (2015) for a conditional variance function; Wang et al. (2013) about smooth SCB for cumulative distribution function in the independent and identically distributed (i.i.d.) settings based on continuous kernel distribution estimator (KDE); Gu and Yang (2015) for a single-index link function; Shao and Yang (2012) and Wang et al. (2014) for time series analysis; and Cardot and Josserand (2011) and Cardot et al. (2013) for SCB for functional data mean curve under survey sampling.

In the context of SCB for the finite population distribution function, Frey (2009) constructed the nonsmooth Kolmogorov–Smirnov type of SCB for the population distribution function under simple random sampling by a recursive algorithm to compute exact coverage frequency of SCB designed for the case when the sample size and population size are both small. More recently, Wang et al. (2016) considered large sample asymptotics based SCB for the CDF of the finite population also under simple random sampling.

However, in practice, stratified sampling is a standard technique in survey methodology commonly employed to increase efficiency over simple random sampling. Consider, for instance, a farm acres dataset in Example 3.2 of Lohr (2009), which consists of four census regions of the United States–Northeast, North Central, South, and West. Figure 8 depicts the distribution functions and histograms of the four regions, clearly showing significant variation among the four, as is also seen in Figure 3.1 in

Lohr (2009). Thus, stratified sampling is highly recommended in such a case. Yet, there do not exist any results on SCB for the CDF of the finite population under stratified sampling in previous works. If one naively treats a stratified sample as a simple random sample (SRS) and applies the method proposed by Wang et al. (2016) to construct SCBs for a stratified population distribution, the performance of the SCBs is not satisfying for a stratified sample as expected. In particular, the empirical coverage frequencies are much different from the nominal confidence level. Some detailed results are shown in Tables 3, 4, 5 and 6 and explained in Sects. 4 and 5. Hence, in this paper, we consider estimators of the finite population distribution function under stratified sampling based on the nonsmooth empirical distribution function (EDF) defined in (4) below and the smooth kernel distribution estimator (KDE) in (6), and develop their corresponding SCBs to provide a powerful tool to make statistical inferences for stratified population distribution function.

The rest of the paper is organized as follows. Section 2 provides the main theoretical results as well as technical conditions needed in our theoretical development. Section 3 describes the actual procedures to implement the SCBs. Simulation studies are given in Sect. 4, and illustrative data example in Sect. 5. Some technical proofs are given in the Appendix.

2 Main results

In stratified sampling, a finite population π of N units is first divided into several nonoverlapping subpopulations called strata. Once the strata have been determined, a simple random sample is drawn from each stratum with the drawings being made independently across the different strata. The total sample size is denoted by n.

In sample surveys, usually the population size N is large but finite. Thus, the classic framework of asymptotics in statistics may not directly apply for a finite population. On the other hand, in sampling theory the finite population π may be viewed as a sample of size N drawn from a superpopulation which has a continuous distribution $F(x)$. This is the setting we assume in this work. Specifically, in the spirit of Rosén (1964) for finite population asymptotics, we assume that there is a sequence of populations $\{\pi_k\}_{k=1}^{\infty}$ as i.i.d. random samples of sizes N_k ($N_k \to \infty$ as $k \to \infty$) generated from a superpopulation with a mixture continuous distribution function $F(x) = \sum_{s=1}^{S} W_s F_s(x)$, where each $F_s(x)$ is a continuous distribution function and W_s, $s = 1, 2, \ldots, S$, are weights satisfying $W_s \in (0, 1)$, $\sum_{s=1}^{S} W_s = 1$. Each π_k can be viewed as a “post-stratified” population in the sense that it can be divided into S strata $\pi_{1k}, \pi_{2k}, \ldots, \pi_{Sk}$ of $N_{1k}, N_{2k}, \ldots, N_{Sk}$ units respectively according to the superpopulation components $F_s(x)$, $s = 1, 2, \ldots, S$. It is clear that π_{sk} can be regarded as an i.i.d. random sample from $F_s(x)$ conditional on N_{sk}. Then stratified random sampling is applied to the population π_k. Notice that in this framework, the stratum sizes N_{sk}, $s = 1, 2, \ldots, S$, are treated as observed random variables due to the randomness of drawing π_k from the superpopulation and $N_k = N_{1k} + N_{2k} + \cdots + N_{Sk}$. Let $W_{sk} = N_{sk} N_k^{-1}$ be the stratum weights of population π_k, which are easily seen to satisfy that W_{sk} converges to W_s almost surely, and of course in probability, as $k \to \infty$.

Denote $y_{si,k}$, $1 \le s \le S$, $1 \le i \le N_{sk}$ as the characteristic of interest for the ith unit in the sth stratum of π_k. Then one has the population $\pi_k = \left\{y_{si,k}\right\}_{s=1,i=1}^{S,N_{sk}}$ and the subpopulations (strata) $\pi_{sk} = \left\{y_{si,k}\right\}_{i=1}^{N_{sk}}$, $s = 1, \ldots, S$. Mathematically, the discrete CDF of the population π_k can be defined as

$$F_{N_k}(x) = N_k^{-1} \sum_{s=1}^{S} \sum_{i=1}^{N_{sk}} I\left(y_{si,k} \le x\right) = \sum_{s=1}^{S} W_{sk} F_{N_{sk}}(x), \tag{1}$$

where $F_{N_{sk}}(x)$ is the CDF of π_{sk} given by

$$F_{N_{sk}}(x) = N_{sk}^{-1} \sum_{i=1}^{N_{sk}} I\left(y_{si,k} \le x\right), \quad s = 1, \ldots, S. \tag{2}$$

By the law of large numbers, we have $F_{N_k}(x) \to_p F(x)$ and $F_{N_{sk}}(x) \to_p F_s(x)$ as $k \to \infty, \forall x \in \mathbb{R}$, where $\to_p$ means convergence in probability. Our goal is to estimate the population distribution $F_{N_k}(x)$ as well as the superpopulation distribution $F(x)$ with their corresponding SCBs in stratified sampling.

For each $k \ge 1$, denote $\left\{Y_{s1,k}, \ldots, Y_{sn_{sk},k}\right\}$ as an SRS drawn from the sth stratum of population π_k with sample size n_{sk} $(n_{sk} \le N_{sk})$, $s = 1, \ldots, S$. Hence, the total sample size is $n_k = n_{1k} + n_{2k} + \cdots + n_{Sk}$. Define $F_{n_{sk}}(x)$ and $F_{n_k}(x)$ as the EDFs of sth stratum π_{sk} and population π_k, which are both random functions due to the randomness of sampling, given by

$$F_{n_{sk}}(x) = n_{sk}^{-1} \sum_{i=1}^{n_{sk}} I\left(Y_{si,k} \le x\right), \quad s = 1, \ldots, S, \tag{3}$$

$$F_{n_k}(x) = \sum_{s=1}^{S} W_{sk} F_{n_{sk}}(x), \quad x \in \mathbb{R}. \tag{4}$$

In addition, we adopt KDE, an integral form of a distribution estimator that appeared in Reiss (1981), Liu and Yang (2008), Wang et al. (2013) for i.i.d. and stationary sequences, to each stratum CDF $F_{N_{sk}}$ and the population CDF F_{N_k}. They are respectively defined as

$$\hat{F}_{sk}(x) = \int_{-\infty}^{x} n_{sk}^{-1} \sum_{i=1}^{n_{sk}} K_{h_s}\left(u - Y_{si,k}\right) \mathrm{d}u, \quad s = 1, \ldots, S, \; x \in \mathbb{R}, \tag{5}$$

$$\hat{F}_k(x) = \sum_{s=1}^{S} W_{sk} \hat{F}_{sk}(x), \quad x \in \mathbb{R} \tag{6}$$

in which K is a kernel function rescaled as $K_{h_s}(u) = K\left(u/h_s\right)/h_s$ with bandwidth $h_s = h_{n_{sk}} > 0$.

For convenience, denote the maximal deviation between any two distribution functions F_1 and F_2 as

$$M(F_1, F_2) = \|F_1 - F_2\|_\infty = \sup_{x\in\mathbb{R}} |F_1(x) - F_2(x)|, \tag{7}$$

and for nonnegative integer p and $\gamma \in (0, 1]$, the collection of functions whose pth derivatives satisfy Hölder conditions of order γ as

$$C^{(p,\gamma)}(\mathbb{R}) = \left\{ g : \mathbb{R} \to \mathbb{R} \,\middle|\, \|g\|_{p,\gamma} = \sup_{x_1\neq x_2, x_1, x_2\in\mathbb{R}} \frac{\left|g^{(p)}(x_1) - g^{(p)}(x_2)\right|}{|x_1 - x_2|^\gamma} < +\infty \right\}.$$

We assume some general technical conditions as follows:

(C1) $\forall s \in \{1, \ldots, S\}$, $\min(n_{sk}, N_{sk} - n_{sk}) \to_p \infty$, *as* $k \to \infty$.
(C2) *There exist constants* $w_s \in (0, 1)$ *and* $C \in [0, 1)$, *such that as* $k \to \infty$, $w_{sk} = n_{sk}/n_k \to_p w_s$ *and* $n_k/N_k \to C$.
(C3) *There exist an integer* $p \geq 0$ *and* $\gamma \in (1/2, 1]$ *such that* $F_s \in C^{(p,\gamma)}(\mathbb{R})$, *and* $F_s(x)$ *is uniformly continuous over* $x \in \mathbb{R}$.
(C4) *The bandwidth* $h_s = h_{n_{sk}} > 0$ *satisfies* $\lambda_{sk} h_{n_{sk}}^{p+\gamma} \to_p 0$, *as* $k \to \infty$, *in which* $\lambda_{sk} = \left(n_{sk}^{-1} - N_{sk}^{-1}\right)^{-1/2}$.
(C5) *The kernel* K *is a continuous and symmetric function, supported on* $[-1, 1]$, *and is a* q *th order kernel for some even integer* $q > p + \gamma$, *i.e., its moments* $\mu_r(K) = \int v^r K(v)\,dv$ *satisfy* $\mu_0(K) \equiv 1$, $\mu_q(K) \neq 0$, $\mu_r(K) \equiv 0$ *for any integer* $r \in (0, q)$.

Remark 1 In many convergence statements of the conditions, we use convergence in probability because we treat the population as an i.i.d. random sample from the superpopulation. In this framework, both n_{sk} and N_{sk} are treated as observed random variables. In the rest of the paper, as a convention we may drop "in probability" for brevity if there is no confusion when discussing convergence in probability.

Since in stratified sampling an SRS is independently drawn from each stratum, we first state the following Theorem 1 which is essentially Theorems 1 and 2 in Wang et al. (2016). It is a finite population version of Donsker's Theorem and the equivalence of EDF and KDE under simple random sampling. Its proof can be found in Wang et al. (2016).

Theorem 1 *Under Condition (C1), as* $k \to \infty$, *for the* s*th stratum of population* π_k, $F_{N_{sk}}(x)$, $F_{n_{sk}}(x)$ *and* $\hat{F}_{sk}(x)$ *respectively given in* (2), (3) *and* (5) *satisfy*

$$\lambda_{sk}\left\{F_{n_{sk}}(x) - F_{N_{sk}}(x)\right\} \xrightarrow{d} B_s(F_s(x)), \quad s = 1, \ldots, S, \tag{8}$$

where $\lambda_{sk} = \left(n_{sk}^{-1} - N_{sk}^{-1}\right)^{-1/2} = n_{sk}^{1/2}(1 - n_{sk}/N_{sk})^{-1/2}$ *is a finite population corrected scale factor and* $B_s(\cdot)$ *represents the Brownian bridge for the* s*th stratum. In addition, under Conditions (C1)–(C5),*

$$\lambda_{sk} M\left(\hat{F}_{sk}, F_{n_{sk}}\right) = \lambda_{sk} \sup_{x\in\mathbb{R}} \left|\hat{F}_{sk}(x) - F_{n_{sk}}(x)\right| = o_p(1), \quad s = 1, \ldots, S. \quad (9)$$

The above theorem can be extended to finite stratified sampling. To properly formulate the extension, we first provide a simple lemma.

Lemma 1 *Under Condition* (*C2*)*,* $\forall s \in \{1, \ldots, S\}$*,* $w_s^{-1} - CW_s^{-1} \geq 0$*, and there exists at least one* $s \in \{1, \ldots, S\}$ *such that* $w_s^{-1} - CW_s^{-1} > 0$.

The above lemma ensures that all $\left(w_s^{-1} - CW_s^{-1}\right)/(1 - C)$ in the next theorem are nonnegative with at least one of them being positive. The proofs of Lemma 1, Theorem 2 and others are given in the Appendix.

Theorem 2 *Under Conditions* (*C1*)*,* (*C2*)*, as* $k \to \infty$*,* $F_{N_k}(x)$ *and* $F_{n_k}(x)$ *given in* (1)*,* (4) *satisfy*

$$\lambda_k \left\{F_{n_k}(x) - F_{N_k}(x)\right\} \xrightarrow{d} B^*(x),$$

in which $\lambda_k = \left(n_k^{-1} - N_k^{-1}\right)^{-1/2}$ *and*

$$B^*(x) = \sum_{s=1}^{S} W_s \sqrt{\left(w_s^{-1} - CW_s^{-1}\right)/(1 - C)} B_s\left\{F_s(x)\right\} \quad (10)$$

where the transformed Brownian bridges $B_s\{F_s(x)\}$*,* $s = 1, \ldots, S$*, are independent of each other.*

Corollary 1 *Under Conditions* (*C1*)*–*(*C2*)*, as* $k \to \infty$*, one has*

$$\mathbb{P}\left[\lambda_k M\left(F_{n_k}, F_{N_k}\right) \leq t\right] \to D^*(t) = \mathbb{P}\left[\max_{x\in\mathbb{R}} \left|B^*(x)\right| \leq t\right], t \geq 0,$$

where $B^*(x)$ *is as defined in* (10)*, and* $D^*(t)$ *represents the extreme value distribution of* $B^*(x)$*. Hence, for* $\alpha \in (0, 1)$*, an asymptotic* $100(1-\alpha)\,\%$ *SCB based on* F_{n_k} *for the finite population CDF* $F_{N_k}(x)$ *is*

$$\left[\max\left\{F_{n_k}(x) - \lambda_k^{-1} D^*_{1-\alpha}, 0\right\}, \min\left\{F_{n_k}(x) + \lambda_k^{-1} D^*_{1-\alpha}, 1\right\}\right], \quad x \in \mathbb{R}, \quad (11)$$

where $D^*_{1-\alpha} = (D^*)^{-1}(1-\alpha)$ *is the* $100(1-\alpha)$*th percentile of* D^* *with* $(D^*)^{-1}$ *being the inverse function of* D^*.

Theorem 3 extends the asymptotic property of the finite population CDF $F_{N_k}(x)$ to the superpopulation CDF $F(x)$.

Theorem 3 *Under Conditions* (*C1*)*–*(*C2*) *and if* $n_k/N_k \to C \equiv 0$ *as* $k \to \infty$*, one has* $\lambda_k M(F_{N_k}, F) = o_p(1)$*. Hence,*

$$\mathbb{P}\left[\lambda_k M(F_{n_k}, F) \leq t\right] \to D^*(t), \quad t \in \mathbb{R}.$$

Furthermore, for $\alpha \in (0, 1)$*, an asymptotic* $100(1-\alpha)\,\%$ *SCB based on* F_{n_k} *for the superpopulation CDF* $F(x)$ *is constructed as* (11).

Theorem 4 below shows that nonsmooth F_{n_k} and smooth $\hat{F}_k$ are asymptotically uniformly equivalent. By Slutsky's Theorem, $\hat{F}_k$ automatically inherits the asymptotic properties of F_{n_k} . Therefore, one can successfully construct smooth SCBs for finite population CDF $F_{N_k}(x)$ and superpopulation CDF $F(x)$ based on $\hat{F}_k$.

Theorem 4 *Under Conditions (C1)–(C5), as $k \to \infty$, $F_{n_k}(x)$ and $\hat{F}_k(x)$ given in (4), (6) satisfy $\lambda_k M\left(\hat{F}_k, F_{n_k}\right) = o_p(1)$. Consequently,*

$$\mathbb{P}\left[\lambda_k M(\hat{F}_k, F_{N_k}) \leq t\right] \to D^*(t), \quad t \in \mathbb{R}.$$

Furthermore, under $n_k/N_k \to C \equiv 0$ in Condition (C2),

$$\mathbb{P}\left[\lambda_k M(\hat{F}_k, F) \leq t\right] \to D^*(t), \quad t \in \mathbb{R}.$$

Hence, an asymptotic $100(1-\alpha)\,\%$ smooth SCBs based on $\hat{F}_k$ for finite population CDF $F_{N_k}(x)$ and superpopulation CDF $F(x)$ are constructed as

$$\left[\max\left\{\hat{F}_k(x) - \lambda_k^{-1} D^*_{1-\alpha}, 0\right\}, \min\left\{\hat{F}_k(x) + \lambda_k^{-1} D^*_{1-\alpha}, 1\right\}\right], \quad x \in \mathbb{R}. \tag{12}$$

Remark 2 When $S = 1$, stratified sampling is reduced to simple random sampling and the extreme value of $B^*(x)$ in Theorem 2 follows the Kolmogorov distribution, i.e., $D^*(t) = D(t) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp\left(-2j^2t^2\right)$, $t > 0$; $D(t) = 0, t \leq 0$. The critical value $D^*_{1-\alpha} = (D^*)^{-1}(1-\alpha)$ can be computed by a standard numerical method.

When $S > 1$, however, the distribution of $B^*(x)$ is not distribution-free, and $D^*_{1-\alpha}$ can be obtained only by non-trivial Monte-Carlo methods. More details about the implementation of the SCBs will be presented in the following Sect. 3.

3 Implementation

In this section, we outline a practical procedure to implement the construction of the SCBs for finite population CDF $F_{N_k}(x)$ and superpopulation CDF $F(x)$ based on estimators EDF $F_{n_k}(x)$ and KDE $\hat{F}_k(x)$ given in (11) and (12), respectively.

The kernel function used in our proposed KDE $\hat{F}_k(x)$ is the quartic kernel, $K(u) = 15\left(1-u^2\right)^2 I\{|u| \leq 1\}/16$, which satisfies Condition (C6) with $q = 2$. The bandwidth for each stratum is taken to be $h_{s1} = \text{IQR}_s \times \lambda_{sk}^{-2}$, or $h_{s2} = \text{IQR}_s \times \lambda_{sk}^{-2/3}$, $s = 1, \ldots, S$, where IQR_s stands for the Inter-Quartile Range of an SRS drawn from the sth stratum, and h_{s1} automatically satisfies Condition (C5), while h_{s2} satisfies (C5) if $p + \gamma > 3/2$, especially taking $p = 1$ since $\gamma \in (1/2, 1]$.

The most important part of the procedure to construct the SCBs is to obtain the critical values $D^*_{1-\alpha} = (D^*)^{-1}(1-\alpha)$, i.e., the $100(1-\alpha)$th percentiles of D^*.

According to the stratified sample drawn from each stratum, it is easy to construct the empirical CDF of each stratum $F_{n_{sk}}$, $s = 1, \ldots, S$, and then to generate the transformed Brownian bridges $B_{sl}\left\{F_{n_{sk}}(x)\right\}$, $l = 1, \ldots, L$, where L is a preset large integer, the default of which is set to be 1000 due to lengthy simulation time. One takes the maximal absolute value over a equally divided values of x for each copy of the weighted sum of Brownian bridges $\sum_{s=1}^{S} \lambda_k W_{sk} \lambda_{sk}^{-1} B_{sl}\left\{F_{n_{sk}}(x)\right\}$ (SCBs for F_{N_k}) or $\sum_{s=1}^{S} W_{sk} w_{sk}^{-1/2}\ B_{sl}\left\{F_{n_{sk}}(x)\right\}$ (SCBs for F), and estimates $D_{1-\alpha}^{*}$ by the empirical quantile of these maximum values. The above limiting method is based on the limiting distribution stated in Theorems 2–4. In our implementation, a is taken to be 401, which appears to be sufficient in this application; see more discussion on this in the next section.

An alternative method for obtaining the critical values in finite sample cases is the bootstrap technique. Suppose for each stratum, $N_{sk} = t_{sk} n_{sk} + r_{sk}$, $1 \leq r_{sk} \leq n_{sk} - 1$, $s = 1, \ldots, S$. Adopting the idea of McCarthy and Snowden (1985), one replicates each sample elements t_{sk} times, together with selecting r_{sk} elements from the sample without replacement to construct an artificial population, from which L repeated stratified samples are drawn without replacement. For each stratified bootstrap sample, the bootstrap EDF and KDE are computed. Then one takes the maximal absolute value of the difference between each bootstrap EDF (KDE) and real sample EDF (KDE), and then estimates the critical values for the SCB based on EDF (KDE) using the proper quantiles of these maximum values.

For stratified random sampling, there are two common methods to allocate the number of units to be drawn from each stratum. One is called proportional allocation as n_{sk} the number of sampled units in each stratum is proportional to N_{sk} the size of the stratum, i.e., $n_{sk} = (n_k/N_k)\, N_{sk}$. The other is the Neyman allocation, a special case of the optimal allocation in which the costs in the strata are assumed to be equal. In this case, n_{sk} is proportional to $N_{sk} V_{sk}$, where V_{sk}^2 is population variance in stratum s. For the estimation of the population CDF,

$$V_{sk}^2 = \int \frac{N_{sk}}{N_{sk} - 1} F_{N_{sk}}(x) \left\{1 - F_{N_{sk}}(x)\right\} \mathrm{d}x, \quad s = 1, \ldots, S,$$

where the integration is approximated by the sum over 401 equally spaced grid points on the range of a pilot SRS drawn from stratum s and the unknown $F_{N_{sk}}(x)$ may be estimated by the corresponding pilot sample EDF.

4 Simulation study

In this section, we present some simulation results to show the finite-sample performance of the proposed estimators and the corresponding SCBs.

In our simulation settings, we consider a population with four strata to match the real data "agpop.dat" discussed in Sect. 5. These four strata are generated, respectively from $\Gamma\,(1.41, 0.66)$, $\Gamma\,(1.45, 2.25)$, $\Gamma\,(0.67, 2.98)$, $\Gamma\,(0.76, 9.56)$, which are regarded as four superpopulations following Gamma distribution $\Gamma\,(\upsilon, \beta)$ with the probability density function (PDF):

$$f(x) = \frac{1}{\beta^{\upsilon}\Gamma(\upsilon)} x^{\upsilon-1} e^{-x/\beta}, \quad x > 0.$$

The parameters of the above four Gamma distributions are taken to be the rescaled moment estimation based on agpop.dat.

The number of units for each stratum is taken to be $N_{1k} = 200$, $N_{2k} = 1000$, $N_{3k} = 1400$, $N_{4k} = 400$, respectively, so that the total number of units in the entire population is $N_k = N_{1k} + N_{2k} + N_{3k} + N_{4k} = 3000$, while the total sample size n_k is taken to 150, 300, 600, 900. Denote m, M as the minimum and maximum of a stratified random sample. We examine the global discrepancy between $F_{n_k}(x)$ and $\hat{F}_k(x)$ under the proportional and optimal allocations measured by the Integrated Squared Error (ISE):

$$\mathrm{ISE}(F_{n_k}, F_{N_k}) = \int \{F_{n_k}(x) - F_{N_k}(x)\}^2 \mathrm{d}x,$$

$$\mathrm{ISE}(\hat{F}_k, F_{N_k}) = \int \{\hat{F}_k(x) - F_{N_k}(x)\}^2 \mathrm{d}x,$$

$$\mathrm{ISE}(F_{n_k}, F) = \int \{F_{n_k}(x) - F(x)\}^2 \mathrm{d}x,$$

$$\mathrm{ISE}(\hat{F}_k, F) = \int \{\hat{F}_k(x) - F(x)\}^2 \mathrm{d}x,$$

in which the integration is approximated by the sum over 401 equally spaced grid points from $m - (M - m) n_k^{-2}$ to M. One then computes the Mean Integrated Squared Error (MISE) as the average of ISE over 1000 replications. Under the proportional and optimal allocations, Figs. 1, 2, 3 and 4 show respectively, the boxplots of the random values $\mathrm{ISE}(\hat{F}_k, F_{N_k})$, $\mathrm{ISE}(\hat{F}_k, F)$ and the random ratios $\mathrm{ISE}(\hat{F}_k, F_{N_k})/\mathrm{ISE}\left(F_{n_k}, F_{N_k}\right)$, $\mathrm{ISE}(\hat{F}_k, F)/\mathrm{ISE}(F_{n_k}, F)$. These plots indi-

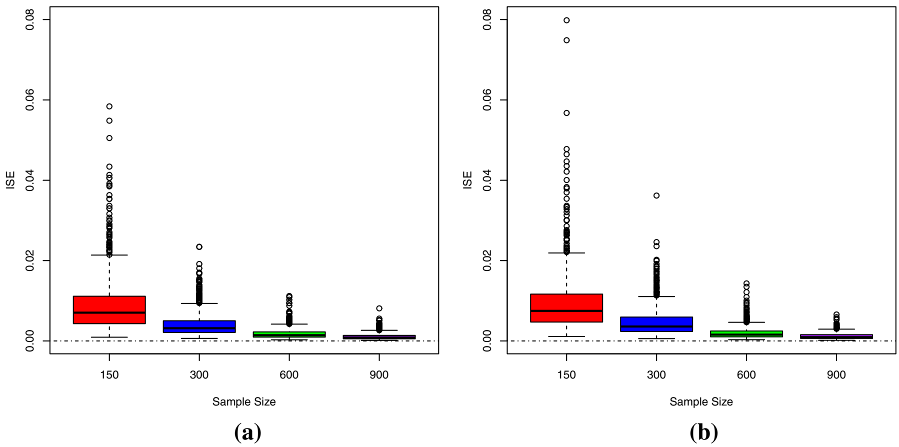


Fig. 1 Boxplots of $\mathrm{ISE}(\hat{F}_k, F_{N_k})$ with bandwidth h_{s1} under the optimal/proportional allocation, respectively. **a** Optimal allocation. **b** Proportional allocation

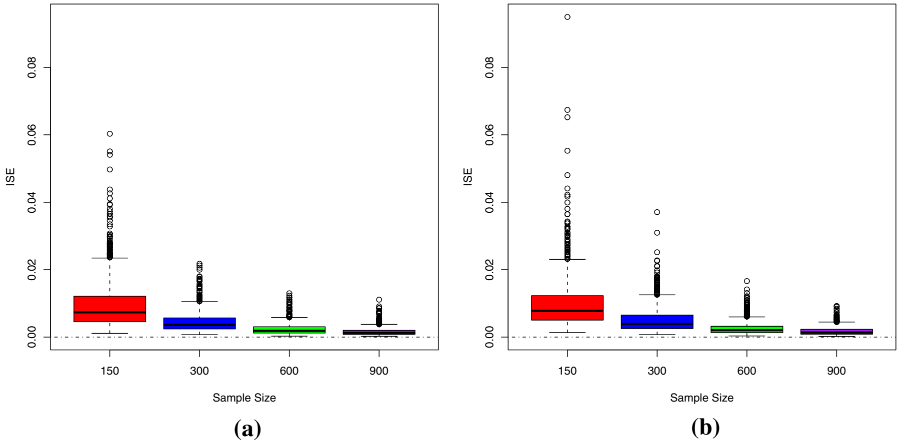


Fig. 2 Boxplots of ISE($\hat{F}_k, F$) with bandwidth h_{s1} under the optimal/proportional allocation, respectively. **a** Optimal allocation. **b** Proportional allocation

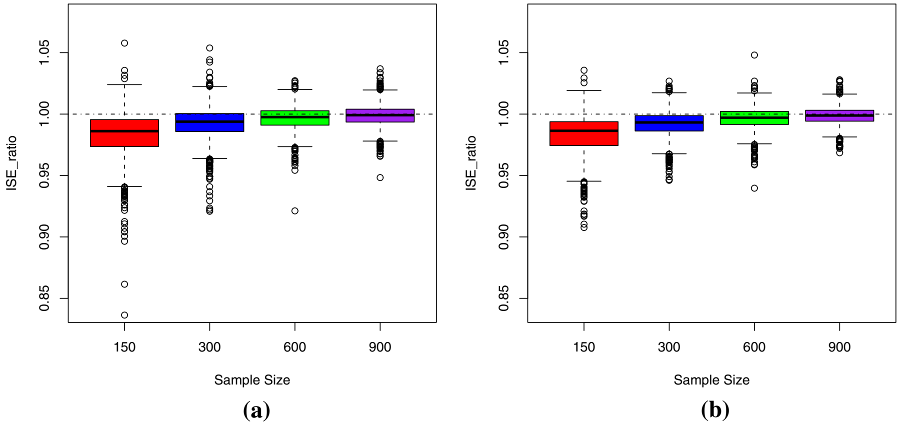
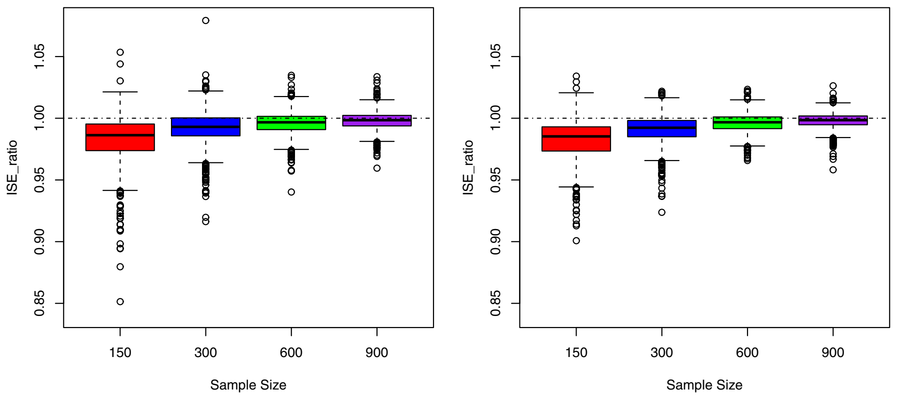


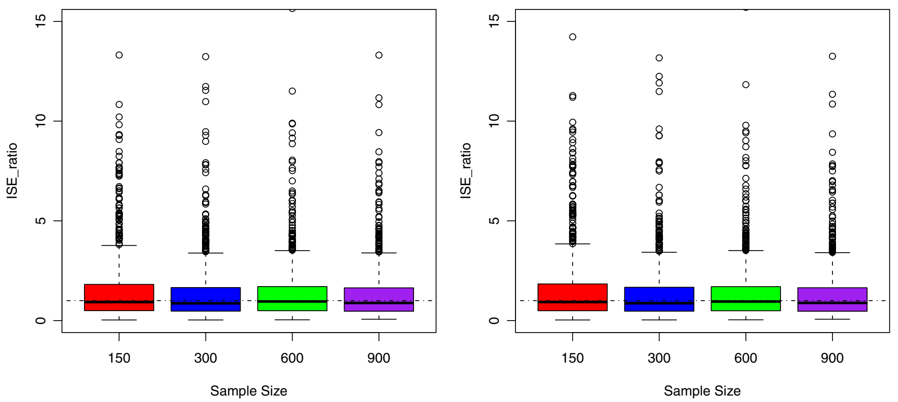
Fig. 3 Boxplots of the ratio ISE($\hat{F}_k, F_{N_k}$)/ISE(F_{n_k}, F_{N_k}) with bandwidth h_{s1} under the optimal/proportional allocation, respectively. **a** Optimal allocation. **b** Proportional allocation

cate that for both allocations ISE($\hat{F}_k, F_{N_k}$) $\rightarrow_p$ 0, ISE($\hat{F}_k, F$) $\rightarrow_p$ 0 and ISE($\hat{F}_k, F_{N_k}$)/ISE$\left(F_{n_k}, F_{N_k}\right)$ $\rightarrow_p$ 1, ISE($\hat{F}_k, F$)/ISE(F_{n_k}, F) $\rightarrow_p$ 1. Moreover, Figs. 5 and 6 show that the differences of ISE between the two allocations are not significant. Table 1 contains MISE($\hat{F}_k, F_{N_k}$), MISE($\hat{F}_k, F$) and the ratios MISE($\hat{F}_k, F_{N_k}$)/MISE(F_{n_k}, F_{N_k}), MISE($\hat{F}_k, F$)/MISE(F_{n_k}, F), which are compared under the two allocations. It shows that as n_k increases, under either the proportional or optimal allocation, MISE($\hat{F}_k, F_{N_k}$) goes to zero and MISE($\hat{F}_k, F_{N_k}$)/MISE(F_{n_k}, F_{N_k}) is smaller than (but close to) 1. MISE($\hat{F}_k, F$) and MISE($\hat{F}_k, F$) /MISE(F_{n_k}, F) behave similarly. All these findings are consistent with the asymptotic properties. In addition, although MISE($\hat{F}_k, F_{N_k}$) and MISE($\hat{F}_k, F$)



(a) **(b)**

Fig. 4 Boxplots of the ratio $\mathrm{ISE}(\hat{F}_k, F)/\mathrm{ISE}(F_{n_k}, F)$ with bandwidth h_{s1} under the optimal/proportional allocation respectively. **a** Optimal allocation. **b** Proportional allocation



(a) **(b)**

Fig. 5 Boxplots of the ratio **a** $\mathrm{ISE}_{\mathrm{opt}}(F_{n_k}, F_{N_k})/\mathrm{ISE}_{\mathrm{prop}}(F_{n_k}, F_{N_k})$ and **b** $\mathrm{ISE}_{\mathrm{opt}}(\hat{F}_k, F_{N_k})/\mathrm{ISE}_{\mathrm{prop}}(\hat{F}_k, F_{N_k})$ with bandwidth h_{s1}

under the optimal allocation are smaller than that under the proportional allocation, the differences are small.

Next, we compare the SCBs constructed by $\hat{F}_k$, F_{n_k} with the confidence levels $1-\alpha = 0.99, 0.95, 0.9, 0.8$. Under the proportional allocation and optimal allocation, respectively, Tables 3 and 4 report the coverage frequencies over 1000 replications that the true curve was covered by SCBs at the 401 equally spaced grid points from $m - (M - m)\, n_k^{-2}$ to M. For visualization of actual function estimates, Fig. 7 depicts curves of the true population CDF F_{N_k}, EDF F_{n_k}, KDE $\hat{F}_k$ taking bandwidth h_{s1}, together with corresponding asymptotic 95% nonsmooth and smooth SCBs at $n_k = 300$. Other settings yielded similar results, but are not included to save space. Our main findings are summarized as follows:

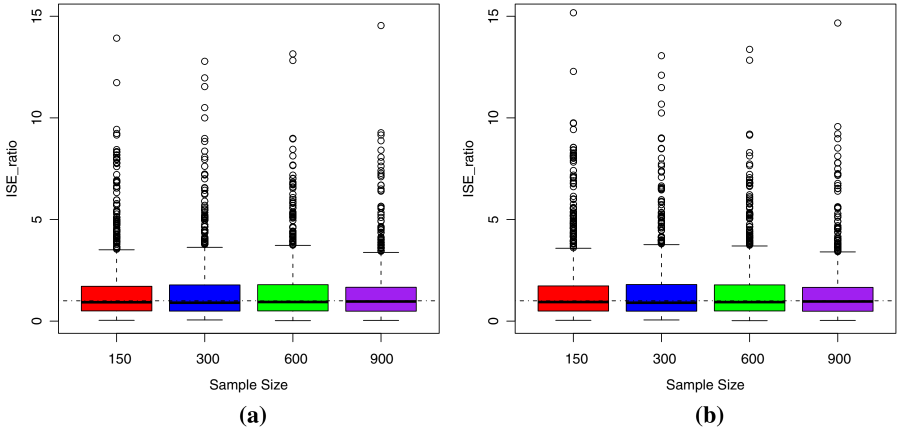


Fig. 6 Boxplots of the ratio **a** $\mathrm{ISE}_{\mathrm{opt}}(F_{n_k}, F)/\mathrm{ISE}_{\mathrm{prop}}(F_{n_k}, F)$ and **b** $\mathrm{ISE}_{\mathrm{opt}}(\hat{F}_k, F)/\mathrm{ISE}_{\mathrm{prop}}(\hat{F}_k, F)$ with bandwidth h_{s1}

Table 1 Comparing MISEs of $\hat{F}_k$ (with bandwidth h_{s1}) and F_{n_k} under the optimal/proportional allocation, respectively

	n_k			
	150	300	600	900
$\mathrm{MISE}_{\mathrm{opt}}(\hat{F}_k, F_{N_k})$	0.0089	0.0041	0.0018	0.0011
$\mathrm{MISE}_{\mathrm{prop}}(\hat{F}_k, F_{N_k})$	0.0094	0.0047	0.0020	0.0012
$\mathrm{MISE}_{\mathrm{opt}}(\hat{F}_k, F_{N_k})/\mathrm{MISE}_{\mathrm{prop}}(\hat{F}_k, F_{N_k})$	0.9489	0.8588	0.9011	0.8987
$\mathrm{MISE}_{\mathrm{opt}}(\hat{F}_k, F_{N_k})/\mathrm{MISE}_{\mathrm{opt}}(F_{n_k}, F_{N_k})$	0.9888	0.9948	0.9975	0.9992
$\mathrm{MISE}_{\mathrm{prop}}(\hat{F}_k, F_{N_k})/\mathrm{MISE}_{\mathrm{prop}}(F_{n_k}, F_{N_k})$	0.9882	0.9945	0.9977	0.9991
$\mathrm{MISE}_{\mathrm{opt}}(\hat{F}_k, F)$	0.0096	0.0046	0.0024	0.0016
$\mathrm{MISE}_{\mathrm{prop}}(\hat{F}_k, F)$	0.0100	0.0051	0.0026	0.0018
$\mathrm{MISE}_{\mathrm{opt}}(\hat{F}_k, F)/\mathrm{MISE}_{\mathrm{prop}}(\hat{F}_k, F)$	0.9627	0.8913	0.9298	0.9122
$\mathrm{MISE}_{\mathrm{opt}}(\hat{F}_k, F)/\mathrm{MISE}_{\mathrm{opt}}(F_{n_k}, F)$	0.9888	0.9944	0.9973	0.9988
$\mathrm{MISE}_{\mathrm{prop}}(\hat{F}_k, F)/\mathrm{MISE}_{\mathrm{prop}}(F_{n_k}, F)$	0.9878	0.9939	0.9974	0.9988

(1) From Tables 3, 4 and Fig. 7, one can see that there are no significant differences between the two allocations regarding the performance of SCBs. The coverage frequencies of SCBs for population CDF F_{N_k} based on EDF F_{n_k} and KDF $\hat{F}_k$ with bandwidth h_{s1} are close to the nominal levels. Meanwhile, three curves of F_{N_k}, F_{n_k}, $\hat{F}_k$ are very close in Fig. 7. All these results reveal that the smooth KDE $\hat{F}_k$ is asymptotically as efficient as the nonsmooth EDF F_{n_k}, and automatically inherits the asymptotic properties of F_{n_k}, which is consistent with Theorem 4.

(2) It is not surprising that the SCBs based on $\hat{F}_k$ with bandwidth h_{s2} do not work well. This is because in our simulation settings the above four Gamma distribution $\Gamma(1.41, 0.66)$, $\Gamma(1.45, 2.25)$, $\Gamma(0.67, 2.98)$, $\Gamma(0.76, 9.56)$ as F_s, s =

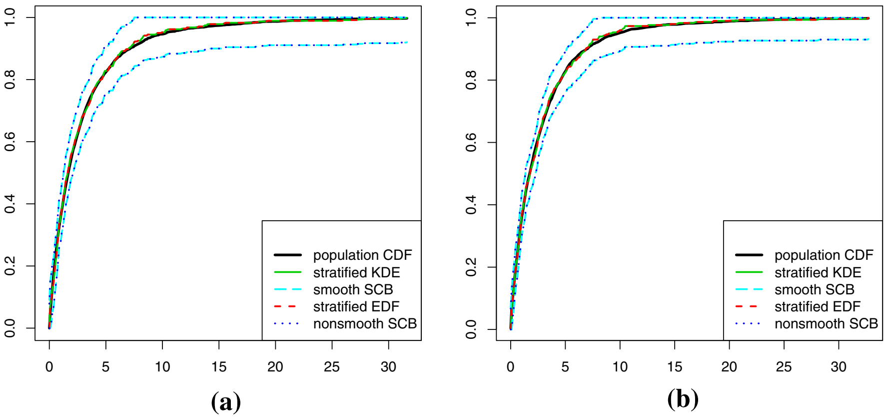


Fig. 7 Plots of 95% SCBs for the population CDF F_{N_k} with bandwidth h_{s1} at $n_k = 300$ by optimal/proportional allocation, respectively. **a** Optimal allocation. **b** Proportional allocation

Table 2 Comparisons of the computing time (minutes) of SCBs' coverage frequencies given the confidence levels $1-\alpha = 0.99, 0.95, 0.9, 0.8$ based on the bootstrap/limiting method with bandwidth h_{s1} over 1000 replications under the proportional allocation

Computing time	$n_k = 150$	$n_k = 300$	$n_k = 600$	$n_k = 900$
Limiting	6.89	7.58	8.35	8.83
Bootstrap	215.30	350.17	596.49	868.00

1, 2, 3, 4, do not satisfy $p+\gamma > 3/2$. To examine this, we ran further simulation studies by choosing $\Gamma(2, 0.5)$, $\Gamma(3, 1)$ as F_s, $s = 1, 2$, which satisfy $p+\gamma > 3/2$, and the results with $N_{1k} = 2000$, $N_{2k} = 1000$ under the proportional allocation are shown in Table 5. However, bandwidth h_{s2} still does not perform as well as h_{s1}. Since bandwidth h_{s2} satisfies Condition (C5) only when F_s has smoothness order $p+\gamma > 3/2$ and is also sensitive to the simulation results, we recommend using bandwidth h_{s1} in practice.

(3) Tables 3, 4 and 5 show that the coverage frequencies of SCBs for F_{N_k} based on F_{n_k} and $\hat{F}_k$ with bandwidth h_{s1} by the bootstrap method are generally as good as that of SCBs by the limiting method. But in Table 2, one can see that the computing time for the bootstrap is much longer than that for the limiting method.

(4) With N_k fixed at 3000, the performance of the SCBs for F deteriorates as n_k increases from 150 to 900, because the condition $\lim_{k\to\infty} n_k/N_k = 0$ for Theorem 3 is increasingly violated.

(5) For comparisons, we use the same stratified samples but incorrectly treat them as if they were simple random samples and apply the method developed in Wang et al. (2016) to construct SCBs for F_{N_k} and F. As shown in Tables 3, 4 and 5, the coverage frequencies of SCBs for simple random sampling are much lower than the nominal levels under the optimal allocation, while are generally higher than the nominal levels under the proportional allocation. In one word, the SCBs

Table 3 Coverage frequencies of SCBs for F_{N_k} and F under optimal allocation based on bootstrap/limiting method compared to simple random sampling: left of the parentheses-$\hat{F}_k$ with h_{s1}; right of the parentheses-$\hat{F}_k$ with h_{s2}; inside the parentheses-F_{n_k}

n_k	SCB	$1-\alpha=0.99$	$1-\alpha=0.95$	$1-\alpha=0.90$	$1-\alpha=0.80$
150	Bootstrap, F_{N_k}	0.988 (0.987) 0.984	0.945 (0.947) 0.937	0.892 (0.899) 0.857	0.774 (0.778) 0.648
	Limiting, F_{N_k}	0.988 (0.984) 0.999	0.938 (0.931) 0.979	0.892 (0.883) 0.950	0.794 (0.774) 0.880
	SRS, F_{N_k}	0.958 (0.956) 0.982	0.864 (0.858) 0.911	0.801 (0.804) 0.849	0.696 (0.692) 0.782
	Limiting, F	0.983 (0.979) 0.997	0.936 (0.927) 0.978	0.871 (0.856) 0.950	0.762 (0.744) 0.880
	SRS, F	0.955 (0.953) 0.981	0.861 (0.858) 0.919	0.790 (0.782) 0.860	0.664 (0.656) 0.777
300	Bootstrap, F_{N_k}	0.984 (0.984) 0.981	0.954 (0.956) 0.927	0.901 (0.901) 0.779	0.804 (0.792) 0.477
	Limiting, F_{N_k}	0.985 (0.985) 0.993	0.952 (0.950) 0.971	0.912 (0.908) 0.941	0.799 (0.787) 0.812
	SRS, F_{N_k}	0.916 (0.915) 0.936	0.747 (0.746) 0.809	0.616 (0.615) 0.698	0.448 (0.445) 0.543
	Limiting, F	0.974 (0.976) 0.989	0.929 (0.923) 0.968	0.879 (0.871) 0.932	0.758 (0.732) 0.777
	SRS, F	0.888 (0.888) 0.928	0.723 (0.722) 0.798	0.590 (0.588) 0.695	0.424 (0.417) 0.548
600	Bootstrap, F_{N_k}	0.983 (0.985) 0.980	0.955 (0.956) 0.765	0.906 (0.909) 0.448	0.814 (0.809) 0.126
	Limiting, F_{N_k}	0.984 (0.983) 0.992	0.945 (0.944) 0.942	0.903 (0.904) 0.756	0.804 (0.795) 0.395
	SRS, F_{N_k}	0.695 (0.694) 0.730	0.432 (0.430) 0.479	0.299 (0.296) 0.360	0.176 (0.172) 0.205
	Limiting, F	0.964 (0.964) 0.986	0.833 (0.833) 0.915	0.817 (0.810) 0.719	0.697 (0.689) 0.353
	SRS, F	0.639 (0.639) 0.711	0.392 (0.391) 0.477	0.278 (0.276) 0.356	0.162 (0.161) 0.201
900	Bootstrap, F_{N_k}	0.989 (0.991) 0.948	0.941 (0.940) 0.474	0.895 (0.894) 0.156	0.805 (0.806) 0.019
	Limiting, F_{N_k}	0.986 (0.986) 0.986	0.951 (0.949) 0.753	0.904 (0.904) 0.388	0.796 (0.793) 0.091
	SRS, F_{N_k}	0.419 (0.421) 0.448	0.200 (0.200) 0.224	0.083 (0.085) 0.098	0.040 (0.040) 0.028
	Limiting, F	0.955 (0.956) 0.967	0.850 (0.852) 0.694	0.746 (0.748) 0.341	0.606 (0.601) 0.076
	SRS, F	0.361 (0.360) 0.440	0.167 (0.167) 0.205	0.092 (0.093) 0.121	0.046 (0.047) 0.051

Table 4 Coverage frequencies of SCBs for F_{N_k} and F under the proportional allocation based on the bootstrap/limiting method compared to simple random sampling: left of the parentheses-$\hat{F}_k$ with h_{s1}; right of the parentheses-$\hat{F}_k$ with h_{s2}; inside the parentheses-F_{n_k}

n_k	SCB	$1-\alpha=0.99$	$1-\alpha=0.95$	$1-\alpha=0.90$	$1-\alpha=0.80$
150	Bootstrap, F_{N_k}	0.989 (0.987) 0.989	0.951 (0.951) 0.939	0.904 (0.896) 0.849	0.790 (0.776) 0.625
	Limiting, F_{N_k}	0.985 (0.985) 0.996	0.945 (0.934) 0.977	0.898 (0.885) 0.949	0.805 (0.771) 0.881
	SRS, F_{N_k}	0.997 (0.997) 1.000	0.972 (0.972) 0.990	0.943 (0.940) 0.972	0.876 (0.875) 0.941
	Limiting, F	0.979 (0.978) 0.995	0.934 (0.926) 0.973	0.880 (0.867) 0.944	0.775 (0.739) 0.885
	SRS, F	0.995 (0.995) 0.998	0.966 (0.965) 0.989	0.929 (0.929) 0.989	0.866 (0.858) 0.932
300	Bootstrap, F_{N_k}	0.988 (0.990) 0.984	0.942 (0.938) 0.892	0.874 (0.874) 0.718	0.765 (0.759) 0.375
	Limiting, F_{N_k}	0.988 (0.986) 0.993	0.944 (0.941) 0.971	0.894 (0.889) 0.936	0.794 (0.774) 0.763
	SRS, F_{N_k}	0.996 (0.996) 0.997	0.977 (0.976) 0.988	0.944 (0.943) 0.974	0.874 (0.871) 0.923
	Limiting, F	0.977 (0.974) 0.990	0.921 (0.917) 0.967	0.862 (0.854) 0.918	0.733 (0.718) 0.749
	SRS, F	0.990 (0.990) 0.996	0.957 (0.956) 0.981	0.916 (0.914) 0.957	0.839 (0.836) 0.907
600	Bootstrap, F_{N_k}	0.995 (0.994) 0.978	0.942 (0.941) 0.683	0.894 (0.893) 0.348	0.764 (0.770) 0.083
	Limiting, F_{N_k}	0.991 (0.990) 0.995	0.952 (0.952) 0.927	0.911 (0.909) 0.734	0.818 (0.820) 0.330
	SRS, F_{N_k}	0.997 (0.997) 0.998	0.979 (0.982) 0.992	0.955 (0.954) 0.971	0.898 (0.898) 0.890
	Limiting, F	0.972 (0.971) 0.988	0.914 (0.911) 0.900	0.843 (0.844) 0.723	0.713 (0.707) 0.342
	SRS, F	0.989 (0.989) 0.997	0.948 (0.949) 0.975	0.915 (0.914) 0.945	0.824 (0.824) 0.855
900	Bootstrap, F_{N_k}	0.990 (0.991) 0.916	0.960 (0.962) 0.380	0.931 (0.931) 0.104	0.818 (0.818) 0.009
	Limiting, F_{N_k}	0.988 (0.987) 0.975	0.946 (0.945) 0.665	0.892 (0.890) 0.290	0.792 (0.789) 0.053
	SRS, F_{N_k}	0.997 (0.997) 0.997	0.981 (0.981) 0.979	0.947 (0.947) 0.918	0.884 (0.882) 0.643
	Limiting, F	0.942 (0.939) 0.954	0.843 (0.838) 0.633	0.754 (0.744) 0.287	0.596 (0.594) 0.049
	SRS, F	0.976 (0.975) 0.987	0.910 (0.909) 0.944	0.845 (0.842) 0.863	0.719 (0.717) 0.597

Table 5 Coverage frequencies of SCBs for CDF of population with two strata from Gamma distribution under the proportional allocation based on the bootstrap/limiting method compared to simple random sampling: left of the parentheses-$\hat{F}_k$ with h_{s1}; right of the parentheses-$\hat{F}_k$ with h_{s2}; inside the parentheses-F_{n_k}

n_k	SCB	$1-\alpha=0.99$	$1-\alpha=0.95$	$1-\alpha=0.90$	$1-\alpha=0.80$
150	Bootstrap, F_{N_k}	0.990 (0.989) 0.987	0.948 (0.946) 0.942	0.899 (0.896) 0.888	0.809 (0.791) 0.781
	Limiting, F_{N_k}	0.982 (0.983) 0.993	0.946 (0.940) 0.979	0.889 (0.883) 0.957	0.788 (0.776) 0.900
	SRS, F_{N_k}	0.996 (0.996) 0.999	0.985 (0.984) 0.994	0.970 (0.968) 0.988	0.923 (0.908) 0.975
	Limiting, F	0.978 (0.976) 0.994	0.937 (0.928) 0.973	0.882 (0.873) 0.952	0.765 (0.738) 0.899
	SRS, F	0.996 (0.996) 0.999	0.980 (0.979) 0.995	0.964 (0.961) 0.987	0.904 (0.897) 0.970
300	Bootstrap, F_{N_k}	0.985 (0.983) 0.982	0.932 (0.934) 0.929	0.888 (0.885) 0.869	0.783 (0.779) 0.750
	Limiting, F_{N_k}	0.990 (0.991) 1.000	0.944 (0.942) 0.978	0.904 (0.901) 0.945	0.802 (0.796) 0.891
	SRS, F_{N_k}	1.000 (1.000) 1.000	0.989 (0.988) 1.000	0.971 (0.969) 0.990	0.922 (0.921) 0.965
	Limiting, F	0.983 (0.981) 0.996	0.926 (0.922) 0.971	0.864 (0.856) 0.937	0.719 (0.710) 0.870
	SRS, F	0.997 (0.997) 1.000	0.984 (0.984) 0.995	0.953 (0.951) 0.984	0.885 (0.883) 0.955
600	Bootstrap, F_{N_k}	0.988 (0.986) 0.984	0.942 (0.942) 0.931	0.892 (0.899) 0.881	0.801 (0.809) 0.767
	Limiting, F_{N_k}	0.995 (0.995) 0.998	0.959 (0.959) 0.979	0.920 (0.921) 0.950	0.819 (0.816) 0.894
	SRS, F_{N_k}	1.000 (1.000) 1.000	0.996 (0.995) 0.999	0.976 (0.976) 0.995	0.935 (0.934) 0.966
	Limiting, F	0.970 (0.969) 0.992	0.903 (0.902) 0.954	0.834 (0.831) 0.915	0.697 (0.690) 0.832
	SRS, F	0.997 (0.997) 1.000	0.971 (0.968) 0.991	0.937 (0.933) 0.974	0.855 (0.853) 0.938
900	Bootstrap, F_{N_k}	0.992 (0.992) 0.991	0.941 (0.944) 0.927	0.893 (0.895) 0.867	0.795 (0.796) 0.746
	Limiting, F_{N_k}	0.984 (0.983) 0.988	0.948 (0.948) 0.963	0.909 (0.906) 0.934	0.801 (0.799) 0.861
	SRS, F_{N_k}	0.996 (0.996) 0.997	0.983 (0.982) 0.990	0.964 (0.964) 0.980	0.923 (0.923) 0.946
	Limiting, F	0.951 (0.952) 0.981	0.860 (0.856) 0.925	0.761 (0.760) 0.863	0.598 (0.597) 0.756
	SRS, F	0.990 (0.990) 0.995	0.957 (0.955) 0.981	0.901 (0.899) 0.954	0.797 (0.796) 0.891

for simple random sampling do not perform well for a stratified sample. This is a strong motivation for us to propose the new method in this paper.

A reviewer pointed out that the number of grid points can be sensitive to the results. There are two issues associated with this point: (1) how the number, a, of grid points affects the simulated critical values and (2) how the number of grid points, b, affects the computation of the coverage frequencies. To investigate these two questions. we conducted some further simulations in the case of proportional allocation with $n_k = 150$. The results with the replication number of weighted sum of Brownian bridges $L = 20{,}000$ show that the differences of the critical values using $a = 401$ and $a = 801$ grid points are negligibly small at all significance levels under consideration, and thus the differences of coverage frequencies between using $a = 401$ and $a = 801$ are also negligibly small. Moreover, the coverage frequencies between using $b = 401$ and $b = 801$ points are in fact identical at all significance levels. We believe that one major contributing factor to this phenomenon is the smooth and monotone increasing features of the distribution function. In any event, using $a = b = 401$ appears to be sufficient in our simulation studies.

In conclusion, we recommend the smooth SCB based on KDE $\hat{F}_k$ with bandwidth h_{s1} for population CDF F_{N_k}. One can use the limiting method to obtain the critical values for SCBs to save computing time. The nonsmooth SCB based on EDF F_{n_k} is a good choice for validation purposes.

5 Illustrative data example

As an illustration, we apply the proposed method to the USA Census of Agriculture dataset introduced in the Introduction which is described and analyzed in Lohr (2009) (see Example 3.2). The population level data file is called agpop.dat. We focus on the 1992 information on acreage devoted to farms for each of the $N_k = 3078$ counties and county-equivalents in the United States as our population. According to the census regions, the population is divided into four strata: Northeast, North Central, South, and West. After omitting the 19 missing data, the numbers of each stratum are $N_{k1} = 213$, $N_{k2} = 1052$, $N_{k3} = 1376$, $N_{k4} = 418$. Figure 8 shows the four discrete population distribution functions $F_{N_{sk}}(x)$, $s = 1, \ldots, 4$, and their smoothed relative frequency plots. These relative frequency plots appear to be from different Gamma distributions. Hence it is useful to stratify the overall population.

Table 6 shows the coverage frequencies of the SCBs for the agpop distribution $F_{N_k}(x)$ under the proportional allocation over 1000 replications, including the smooth SCB based on KDE $\hat{F}_k$ with bandwidth h_{s1} and the nonsmooth SCB based on EDF F_{n_k} by the bootstrap and limiting methods, compared to SRS. In contrast, the coverage frequencies of the smooth SCB by the limiting method approach the nominal levels, while the SCBs for SRS are inaccurate as their coverage frequencies are much higher than the nominal levels. Figure 9 depicts the true population CDF F_{N_k}, EDF F_{n_k}, KDE $\hat{F}_k$ with bandwidth h_{s1} and the corresponding asymptotic 95% nonsmooth and smooth SCBs at sample size $n_k = 300$ in one of the 1000 runs. One sees that not only the curves of F_{n_k}, $\hat{F}_k$ fit the true population CDF F_{N_k} well, but also the 95% nonsmooth and smooth SCBs both contain F_{N_k}. All these numerical and graphical

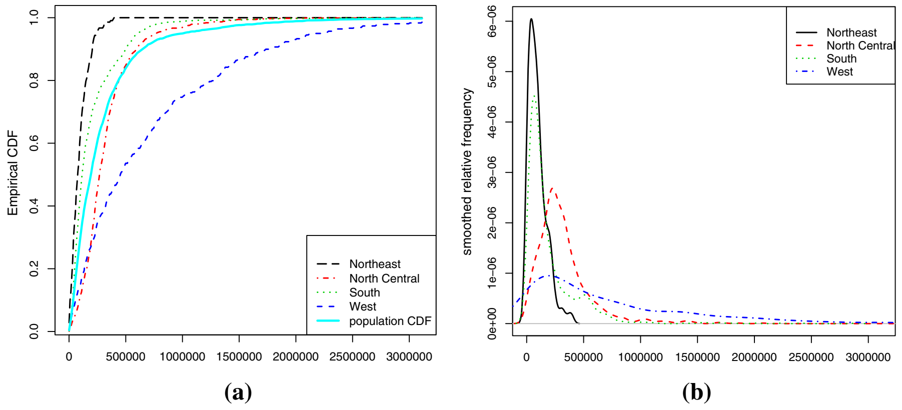


Fig. 8 Plots of **a** empirical CDFs of each stratum and population, **b** smoothed relative frequencies for each stratum in agpop.dat

Table 6 Coverage frequencies of the SCBs for agpop CDF $F_{N_k}(x)$ with bandwidth h_{s1} based on 1000 replications under proportional allocation

n_k	SCB	$1-\alpha$			
		0.99	0.95	0.90	0.80
150	Bootstrap, nonsmooth	0.982	0.935	0.843	0.733
	Bootstrap, smooth	0.980	0.934	0.859	0.744
	Limiting, nonsmooth	0.984	0.952	0.876	0.777
	Limiting, smooth	0.985	0.952	0.893	0.786
	SRS, nonsmooth	1.000	0.990	0.976	0.925
	SRS, smooth	1.000	0.990	0.977	0.931
300	Bootstrap, nonsmooth	0.980	0.935	0.870	0.752
	Bootstrap, smooth	0.985	0.941	0.876	0.760
	Limiting, nonsmooth	0.989	0.944	0.874	0.746
	Limiting, smooth	0.991	0.950	0.881	0.755
	SRS, nonsmooth	0.999	0.983	0.955	0.892
	SRS, smooth	0.999	0.984	0.956	0.895
600	Bootstrap, nonsmooth	0.995	0.966	0.915	0.818
	Bootstrap, smooth	0.996	0.964	0.909	0.819
	Limiting, nonsmooth	0.988	0.953	0.903	0.819
	Limiting, smooth	0.989	0.955	0.905	0.825
	SRS, nonsmooth	0.996	0.989	0.963	0.902
	SRS, smooth	0.996	0.989	0.962	0.902
900	Bootstrap, nonsmooth	0.994	0.968	0.915	0.785
	Bootstrap, smooth	0.994	0.967	0.921	0.790
	Limiting, nonsmooth	0.990	0.965	0.910	0.794
	Limiting, smooth	0.990	0.964	0.916	0.793
	SRS, nonsmooth	0.998	0.990	0.971	0.915
	SRS, smooth	0.998	0.990	0.971	0.916

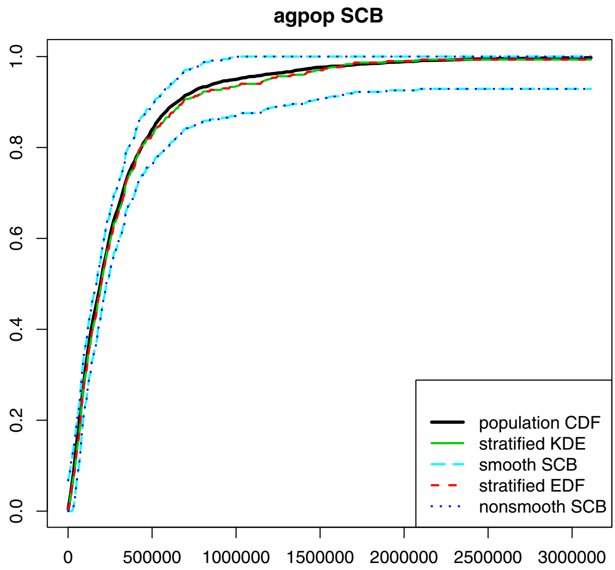


Fig. 9 Plot of 95% SCBs for the "agpop" CDF with bandwidth h_{s1} at $n_k = 300$ under proportional allocation

results suggest that our proposed SCBs are robust tools for statistical inference on the finite population distribution in stratified random sampling.

The R code used to produce the results in the simulations and example is available on the website: https://github.com/gulijie2018/StratifiedSCB.

Appendix

In this Appendix, we use $a_n = o(b_n)$ to denote that $\lim_{n\to\infty} a_n/b_n = 0$, and $a_n = O(b_n)$ to denote that $\lim\sup_{n\to\infty} a_n/b_n = c$, where c is a constant. In addition, we denote by o_p $\left(O_p\right)$ and $o_{a.s.}$ a sequence of random variables of order o (O) in probability and almost surely, respectively, while $u_{a.s.}$ means $o_{a.s.}$ uniformly in the domain.

In the following we will prove Lemma 1 and Theorems 2–4.

A.1 Proof of Lemma 1

Our framework given in Sect. 2 and Condition (C2) ensure that, for any $s \in \{1, \ldots, S\}$,

$$\lim_{k\to\infty} (N_{sk}/N_k) = W_s, \quad \lim_{k\to\infty} (n_{sk}/n_k) = w_s, \quad \lim_{k\to\infty} (n_k/N_k) = C.$$

Hence,

$$CW_s^{-1}w_s = \lim_{k\to\infty} (n_k/N_k)\,(N_{sk}/N_k)^{-1}\,(n_{sk}/n_k) = \lim_{k\to\infty} (n_{sk}/N_{sk}) \leq 1.$$

Making use of the simple inequality

$$n_k/N_k = \frac{\sum_{s=1}^{S} n_{sk}}{\sum_{s=1}^{S} N_{sk}} \geq \min_{1\leq s\leq S}(n_{sk}/N_{sk})$$

and letting $k \to \infty$, one obtains that

$$\begin{aligned} C &= \lim_{k\to\infty} n_k/N_k \geq \lim_{k\to\infty}\min_{1\leq s\leq S}(n_{sk}/N_{sk}) = \min_{1\leq s\leq S}\lim_{k\to\infty}(n_{sk}/N_{sk}) \\ &= \min_{1\leq s\leq S} CW_s^{-1}w_s, \end{aligned}$$

since $C < 1$ according to Condition (C2), one obtains that $\min_{1\leq s\leq S} CW_s^{-1}w_s < 1$. The Lemma 1 is proved. □

A.2 Proof of Theorem 2

For $s = 1, \ldots, S$, combining (8) in Theorem 1 with Skorohod's Representation Theorem shown in Theorem 6.7 of Billingsley (1999), there exits a version $\tilde{B}_{sk}(\cdot)$ of Brownian bridge $B_s(\cdot)$ that satisfies $\tilde{B}_{sk}(F_s(x)) \stackrel{d}{\to} B_s(F_s(x))$ as $k \to \infty$ such that

$$\sup_{x\in\mathbb{R}}\left|\lambda_{sk}\left\{F_{n_{sk}}(x) - F_{N_{sk}}(x)\right\} - \tilde{B}_{sk}(F_s(x))\right| \to 0,\ a.s.,$$

which implies that

$$F_{n_{sk}}(x) - F_{N_{sk}}(x) = \lambda_{sk}^{-1}\tilde{B}_{sk}(F_s(x)) + u_{a.s.}\left(\lambda_{sk}^{-1}\right).$$

Recalling the definitions of $F_{N_k}(x)$ and $F_{n_k}(x)$ given in (1) and (4), one has

$$\begin{aligned} \lambda_k\left\{F_{n_k}(x) - F_{N_k}(x)\right\} &= \lambda_k\left\{\sum_{s=1}^{S} W_{sk}F_{n_{sk}}(x) - \sum_{s=1}^{S} W_{sk}F_{N_{sk}}(x)\right\} \\ &= \lambda_k\sum_{s=1}^{S} W_{sk}\left\{F_{n_{sk}}(x) - F_{N_{sk}}(x)\right\} \\ &= \lambda_k\sum_{s=1}^{S} W_{sk}\left\{\lambda_{sk}^{-1}\tilde{B}_{sk}(F_s(x)) + u_{a.s.}\left(\lambda_{sk}^{-1}\right)\right\}. \end{aligned}$$

According to Condition (C2), as $k \to \infty$,

$$\frac{n_{sk}}{N_{sk}} = \frac{n_{sk}}{n_k}\cdot\frac{n_k}{N_k}\cdot\frac{N_k}{N_{sk}} \to_p w_sCW_s^{-1},$$

and

$$\lambda_k W_{sk} \lambda_{sk}^{-1} = \frac{N_{sk}}{N_k} \sqrt{\frac{n_k}{n_{sk}} \cdot \frac{1 - n_{sk}/N_{sk}}{1 - n_k/N_k}} \to_p W_s \sqrt{w_s^{-1} \frac{1 - w_s C W_s^{-1}}{1 - C}}. \tag{A.1}$$

Hence,

$$\lambda_k \left\{ F_{n_k}(x) - F_{N_k}(x) \right\} \xrightarrow{d} \sum_{s=1}^{S} W_s \sqrt{\left(w_s^{-1} - C W_s^{-1} \right) / (1 - C)} B_s \left\{ F_s(x) \right\}.$$

The proof of Theorem 2 is completed. □

A.3 Proof of Theorem 3

Note that $\lambda_k N_k^{-1/2} = \left(n_k^{-1} - N_k^{-1} \right)^{-1/2} N_k^{-1/2} = (n_k/N_k)^{1/2} (1 - n_k/N_k)^{-1/2} \to 0$ when $n_k/N_k \to C \equiv 0$ as $k \to \infty$. Because of a sequence of populations $\{\pi_k\}_{k=1}^{\infty}$ as i.i.d. random samples generated from $F(x)$, Donsker's Theorem entails that $N_k^{1/2} \left\{ F_{N_k}(x) - F(x) \right\} \xrightarrow{d} B\{F(x)\}$. Hence, as $k \to \infty$,

$$\lambda_k M(F_{N_k}, F) = \lambda_k O_p \left(N_k^{-1/2} \right) = o_p(1).$$

Then Theorem 3 follows by Theorem 2 and Slutsky's Theorem. □

A.4 Proof of Theorem 4

According to the definitions of $F_{n_k}(x)$ and $\hat{F}_k(x)$ given in (4) and (6), one has

$$\lambda_k \left\{ F_{n_k}(x) - \hat{F}_k(x) \right\} = \lambda_k \left\{ \sum_{s=1}^{S} W_{sk} F_{n_{sk}}(x) - \sum_{s=1}^{S} W_{sk} \hat{F}_{sk}(x) \right\}.$$

Then (9) and (A.1) imply that

$$\begin{aligned} \lambda_k M(\hat{F}_k, F_{N_k}) &= \lambda_k \sup_{x \in \mathbb{R}} \left| \hat{F}_k(x) - F_{n_k}(x) \right| \\ &\leq \lambda_k \sum_{s=1}^{S} W_{sk} \sup_{x \in \mathbb{R}} \left| \hat{F}_{sk}(x) - F_{n_{sk}}(x) \right| \\ &= \lambda_k \sum_{s=1}^{S} W_{sk} \lambda_{sk}^{-1} \times o_p(1) = o_p(1). \end{aligned}$$

Applying Theorems 2 and 3 and Slutsky's Theorem, Theorem 4 is proved. □

References

Bickel, P. J., Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics*, *1*, 1071–1095.

Billingsley, P. (1999). *Convergence of Probability Measures* (2nd ed.). New York: Wiley.

Cai, L., Yang, L. (2015). A smooth simultaneous confidence band for conditional variance function. *TEST*, *24*, 632–655.

Cao, G., Yang, L., Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, *24*, 359–377.

Cao, G., Wang, L., Li, Y., Yang, L. (2016). Oracle-efficient confidence envelopes for covariance functions in dense functional data. *Statistica Sinica*, *26*, 359–383.

Cardot, H., Josserand, E. (2011). Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, *98*, 107–118.

Cardot, H., Degras, D., Josserand, E. (2013). Confidence bands for Horvitz–Thompson estimators using sampled noisy functional data. *Bernoulli*, *19*, 2067–2097.

Chambers, R. L., Dunstan, R. (1986). Estimation distribution functions from survey data. *Biometrika*, *73*, 597–604.

Chen, J., Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, *12*, 1223–1239.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Degras, D. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, *21*, 1735–1765.

Frey, J. (2009). Confidence bands for the CDF when sampling from a finite population. *Computational Statistics and Data Analysis*, *53*, 4126–4132.

Gu, L., Yang, L. (2015). Oracally efficient estimation for single-index link function with simultaneous confidence band. *Electronic Journal of Statistics*, *9*, 1540–1561.

Gu, L., Wang, L., Härdle, W., Yang, L. (2014). A simultaneous confidence corridor for varying coefficient regression with sparse functional data. *TEST*, *23*, 806–843.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis*, *29*, 163–179.

Liu, R., Yang, L. (2008). Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics*, *20*, 661–677.

Lohr, S. (2009). *Sampling: Design and analysis* (2nd ed.). Boston: Brooks/Cole.

Ma, S., Yang, L., Carroll, R. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica*, *22*, 95–122.

McCarthy, P. J., Snowden, C. B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, *73*, 1–23.

O'Neill, T., Stern, S. (2012). Finite population corrections for the Kolmogorov–Smirnov tests. *Journal of Nonparametric Statistics*, *24*, 497–504.

Reiss, R. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, *8*, 116–119.

Rosén, B. (1964). Limit theorems for sampling from finite population. *Arkiv för Matematik*, *5*, 383–424.

Shao, Q., Yang, L. (2012). Polynomial spline confidence band for time series trend. *Journal of Statistical Planning and Inference*, *142*, 1678–1689.

Song, Q., Yang, L. (2009). Spline confidence bands for variance function. *Journal of Nonparametric Statistics*, *21*, 589–609.

Song, Q., Liu, R., Shao, Q., Yang, L. (2014). A simultaneous confidence band for dense longitudinal regression. *Communications in Statistics-Theory and Methods*, *43*, 5195–5210.

Wang, J., Yang, L. (2009). Polynomial spline confidence bands for regression curves. *Statistica Sinica*, *19*, 325–342.

Wang, J., Cheng, F., Yang, L. (2013). Smooth simultaneous confidence bands for cumulative distribution functions. *Journal of Nonparametric Statistics*, *25*, 395–407.

Wang, J., Liu, R., Cheng, F., Yang, L. (2014). Oracally efficient estimation of autoregressive error distribution with simultaneous confidence band. *Annals of Statistics*, *42*, 654–668.

Wang, J., Wang, S., Yang, L. (2016). Simultaneous confidence bands for the distribution function of a finite population and of its superpopulation. *TEST*, *25*, 692–709.

Wang, S., Dorfman, A. (1996). A new estimator for the finite population distribution function. *Biometrika*, *83*, 639–652.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society Series B*, *60*, 797–811.

Zheng, S., Yang, L., Härdle, W. (2014). A smooth simultaneous confidence corridor for the mean of sparse functional data. *Journal of the American Statistical Association*, *109*, 661–673.

Zhu, H., Li, R., Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Annals of Statistics*, *40*, 2634–2666.