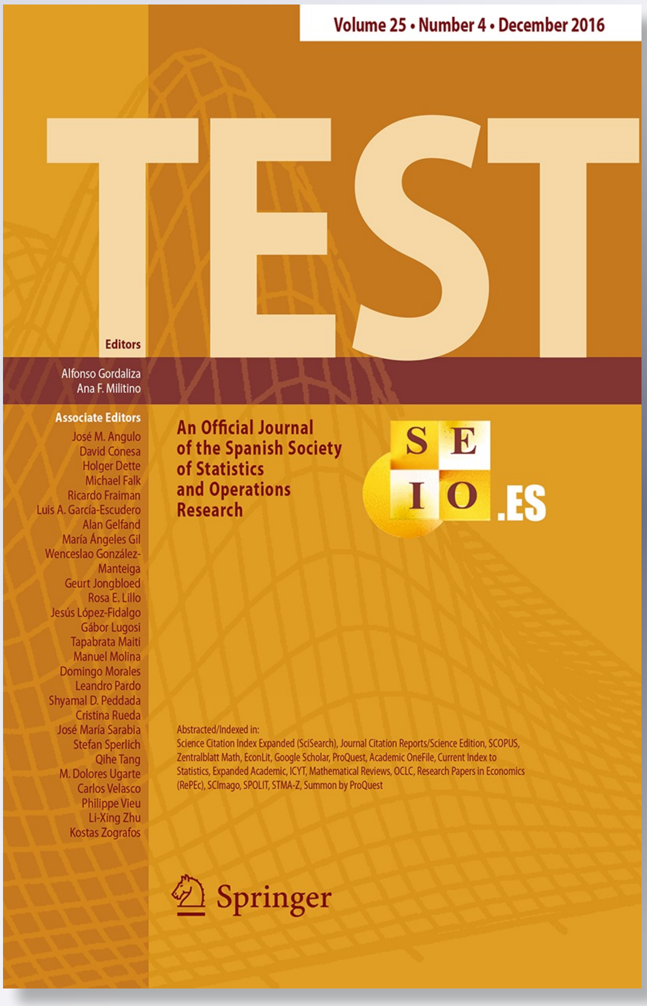

Volume 25 • Number 4 • December 2016

TEST

Editors
Alfonso Gordaliza
Ana F. Militino

Associate Editors
José M. Angulo
David Conesa
Holger Dette
Michael Falk
Ricardo Fraiman
Luis A. García-Escudero
Alan Gelfand
María Ángeles Gil
Wenceslao González-Manteiga
Geurt Jongbloed
Rosa E. Lillo
Jesús López-Fidalgo
Gábor Lugosi
Tapabrata Maiti
Manuel Molina
Domingo Morales
Leandro Pardo
Shyamal D. Peddada
Cristina Rueda
José María Sarabia
Stefan Sperlich
Qihe Tang
M. Dolores Ugarte
Carlos Velasco
Philippe Vieu
Li-Xing Zhu
Kostas Zografos

An Official Journal of the Spanish Society of Statistics and Operations Research

SEIO.ES

Abstracted/Indexed in:
Science Citation Index Expanded (SciSearch), Journal Citation Reports/Science Edition, SCOPUS, Zentralblatt Math, EconLit, Google Scholar, ProQuest, Academic OneFile, Current Index to Statistics, Expanded Academic, ICYT, Mathematical Reviews, OCLC, Research Papers in Economics (RePEc), SCImago, SPOLIT, STMA-Z, Summon by ProQuest

Springer

TEST (2016) 25:692–709
DOI 10.1007/s11749-016-0491-5

ORIGINAL PAPER

Simultaneous confidence bands for the distribution function of a finite population and of its superpopulation

Jiangyan Wang[1] · Suojin Wang[2] · Lijian Yang[3]

Received: 14 August 2015 / Accepted: 8 April 2016 / Published online: 5 May 2016

Abstract Simultaneous confidence bands (SCBs) are proposed for the distribution function of a finite population and of the latent superpopulation via the empirical distribution function (nonsmooth) and kernel distribution estimator (smooth) based on a simple random sample (SRS), either with or without finite population correction. It is shown that both nonsmooth and smooth SCBs achieve asymptotically the nominal confidence level under standard assumptions. In particular, the uncorrected nonsmooth SCB for superpopulation is exactly the same as the Kolmogorov–Smirnov SCB based on an independent and identically distributed sample as long as the SRS size is infinitesimal relative to the finite population size. Extensive simulation studies confirm the asymptotic properties. As an illustration, the proposed SCBs are constructed for the population distribution of the well-known baseball data (Lohr, Sampling: design and analysis, 2nd edn. Brooks/Cole, Boston, 2009).

Keywords Bandwidth · Brownian bridge · Kernel · Kolmogorov distribution · Sample survey

Mathematics Subject Classification 62D05 · 62G05 · 62G15 · 62G20

Electronic supplementary material The online version of this article (doi:10.1007/s11749-016-0491-5) contains supplementary material, which is available to authorized users.

✉ Lijian Yang
yanglijian@tsinghua.edu.cn

[1] 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

1 Introduction

Consider a finite population (population for brevity) of real numbers $\pi = \{b_1, b_2, \ldots, b_N\}$ where N denotes the population size and b_i's are continuous measurements, such as household income and body mass index. With equal probability mass N^{-1} assigned to each element b_i, $1 \le i \le N$, the (discrete) cumulative distribution function (cdf) of population π is defined as:

$$F_N^*(x) = N^{-1} \sum\nolimits_{i=1}^{N} I(b_i \le x). \tag{1}$$

While $F_N^*(x)$ is mathematically well defined, it is often the case that the entire population π is unavailable to a data analyst, who has only a random sample $X_1, X_2, \ldots, X_n$ of a much smaller size n, drawn from the population.

Following Rosén (1964), we denote by Ω the set of all the permutations $\omega = \{i_1, i_2, \ldots, i_N\}$ of the numbers $1, 2 \ldots, N$, which is made into a probability space by introducing the measure $P(\omega) = (N!)^{-1}, \forall \omega \in \Omega$. The following N-dimensional vector function

$$T_\pi(\omega) = T_\pi(i_1, i_2, \ldots, i_N) = (b_{i_1}, b_{i_2}, \ldots, b_{i_N}), \quad \forall \omega = \{i_1, i_2, \ldots, i_N\} \in \Omega, \tag{2}$$

is called a random permutation of the elements in π, denoted as:

$$T_\pi(\omega) = (X_1, X_2, \ldots, X_N) = \{X_1(\omega), X_2(\omega), \ldots, X_N(\omega)\}. \tag{3}$$

A simple random sample (SRS) of size n $(n \le N)$ is simply $(X_1, X_2, \ldots, X_n)$, which consists of n consecutive random drawings from π without replacement. In the sample survey literature, one important question to answer is the following: how to estimate $F_N^*(x)$ based on the data $(X_1, X_2, \ldots, X_n)$?

Estimation of the finite population distribution function $F_N^*(x)$ and the related population quantile functions has been studied by many researchers. Noteworthy among existing works include Francisco and Fuller (1991) which focused on the estimation problem in the setting of a stratified cluster sampling, Chen and Wu (2002) which proposed the pseudo-empirical likelihood methods, Chambers and Dunstan (1986) which described model-based estimation using auxiliary information, and Wang and Dorfman (1996) on further improving estimation of the finite population distribution by synthesizing model-based and design-based methods.

The natural estimator of $F_N^*(x)$ is the empirical distribution function $F_n^*(x) = F_n^*(x, \omega)$, a random function defined as:

$$F_n^*(x, \omega) = n^{-1} \sum\nolimits_{i=1}^{n} I(X_i(\omega) \le x), \quad \omega \in \Omega, \quad x \in \mathbb{R}. \tag{4}$$

Large sample properties of $F_n^*(x, \omega)$ in relation to the true distribution have been studied extensively for the independent and identically distributed (iid) sample case; see for instance, Cheng and Peng (2002), Falk (1985), Liu and Yang (2008), Reiss (1981), Wang et al. (2013), Xue and Wang (2010) and Yamato (1973).

Table 1 Asymptotic critical values for the Kolmogorov–Smirnov goodness-of-fit test

$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
1.63	1.36	1.22	1.07

While obviously both $F_N^*(x)$ and $F_n^*(x)$ are discrete, the population size N is usually quite large, and thus one may conceptualize that $F_N^*(x) \approx F(x)$ for a continuous distribution $F(x)$. To formalize this approximation, the framework of Rosén (1964) for finite population asymptotics assumes that there is a sequence of populations $\{\pi_k\}_{k=1}^{\infty}$, $\pi_k = \left\{b_{k1}, b_{k2}, \ldots, b_{kN_k}\right\}$ with population size $N_k \to \infty$ as $k \to \infty$, and that

$$F_{N_k}^*(x) = N_k^{-1} \sum_{i=1}^{N_k} I\left(b_{ki} \leq x\right) \to F(x), \tag{5}$$

for a continuous distribution $F(x)$. For each $k \geq 1$, we denote by $X_{k1}, \ldots, X_{kn_k}$ a simple random sample drawn from population π_k without replacement, of sample size n_k $(n_k \leq N_k)$ and $F_{n_k}^*$ the empirical distribution based on $X_{k1}, \ldots, X_{kn_k}$

$$F_{n_k}^*(x) = n_k^{-1} \sum_{i=1}^{n_k} I\left(X_{ki} \leq x\right). \tag{6}$$

Intuitively one may think about each π_k as an iid random sample drawn from a continuous superpopulation F, with the increasing size N_k. However, since π_k is typically unobserved or too large, the classic Kolmogorov–Smirnov simultaneous confidence band (SCB) for F based on π_k is unavailable. Although the object of greater interest $F_{N_k}^*$ is technically a discrete function, as we explained above for all practical purposes it can be viewed as a continuous function when the population size N_k is reasonably large and the b_{ki}'s are continuous measurements. It is, therefore, desirable to obtain large sample SCBs for both $F_{N_k}^*$ and F, none of which exist in the current literature.

To begin with, we denote the maximal deviation between two distribution functions G_1 and G_2 as:

$$D\left(G_1, G_2\right) = \|G_1 - G_2\|_\infty = \sup_x |G_1(x) - G_2(x)|. \tag{7}$$

Denoting by $B(t)$ the Brownian bridge, the distribution of $\max_{t \in [0,1]} |B(t)|$ is the Kolmogorov distribution

$$L(t) = \left\{1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2t^2)\right\} I\left(t > 0\right). \tag{8}$$

Table 1 contains $100(1-\alpha)$ %-th percentiles $L_{1-\alpha}$ of $L(t)$, $L_{1-\alpha} = L^{-1}(1-\alpha)$. If all the samples $X_{k1}, \ldots, X_{kn_k}$, $k \geq 1$ were iid with the same continuous distribution F, Donsker's Theorem would ensure the convergence of stochastic processes $n_k^{1/2}\{F_{n_k}^*(x) - F(x)\}$, $x \in \mathbb{R}$ to $B\{F(x)\}$ and hence

$$\mathbb{P}\left[n_k^{1/2} D\left(F_{n_k}^*, F\right) \leq t\right] \to L(t), \quad t \in \mathbb{R}, \quad \text{as } k \to \infty, \tag{9}$$

which leads to the well-known Kolmogorov–Smirnov SCB for $F(x)$ and the associated goodness-of-fit test.

A finite population version of Donsker's Theorem, Theorem 1 states that in uniform metric on the cadlag space $\mathcal{D}[0, 1]$

$$\lambda_k \left\{F_{n_k}^*(x) - F_{N_k}^*(x)\right\} \xrightarrow{d} B\left\{F(x)\right\}, \tag{10}$$

in which $\lambda_k = \left(n_k^{-1} - N_k^{-1}\right)^{-1/2} = n_k^{1/2}\left(\text{fpc}_k\right)^{-1/2}$ is a finite population corrected scale factor similar to $n_k^{1/2}$ for an iid sample of size n_k, with $\text{fpc}_k = 1 - n_k/N_k$ the *finite population correction* (fpc) factor. Equation (10) allows one to construct a "corrected" Kolmogorov–Smirnov SCB based on $F_{n_k}^*(x)$ for each $F_{N_k}^*(x)$, $k \geq 1$ with predetermined asymptotic coverage $1 - \alpha$:

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

Interestingly, under assumption (A4′), the above SCB also covers the superpopulation distribution $F(x)$ with probability $1 - \alpha$. Furthermore, the following "uncorrected" Kolmogorov–Smirnov SCB also covers both $F(x)$ and $F_{N_k}^*(x)$ with probability $1 - \alpha$ asymptotically:

$$\left[\max\left(F_{n_k}^*(x) - n_k^{-1/2} L_{1-\alpha}, 0\right), \min\left(F_{n_k}^*(x) + n_k^{-1/2} L_{1-\alpha}, 1\right)\right], \quad x \in \mathbb{R}, \tag{12}$$

since (A4′) implies that $\lim_{k\to\infty} n_k^{-1/2}/\lambda_k^{-1} = 1$. In other words, one obtains standard Kolmogorov–Smirnov SCB for the superpopulation distribution $F(x)$ as if the dependent sample $X_{k1}, \ldots, X_{kn_k}$ were actually an iid sample from $F(x)$.

Using the empirical cdf is one logical way to compare the random sample with $F(x)$, but in the following we further extend the recent work by Wang et al. (2013), which offers a useful smooth estimator, to the case of finite population to provide a smooth SCB for $F_{N_k}^*(x)$ and $F(x)$. The kernel distribution estimator (KDE) for $F_{N_k}^*$ is defined as:

$$\hat{F}_k^*(x) = \int_{-\infty}^{x} n_k^{-1} \sum\nolimits_{i=1}^{n_k} K_h\left(u - X_{ki}\right) \mathrm{d}u, \quad x \in \mathbb{R}, \tag{13}$$

where $h = h_{n_k} > 0$ is the bandwidth and K is a kernel function, and $K_h(u) = K(u/h)/h$. We show in Theorem 2 that $D\left(F_{n_k}^*, \hat{F}_k^*\right)$ is of order $o_p(\lambda_k^{-1})$, which leads us to propose the "corrected" smooth SCB for $F_{N_k}^*(x)$ and $F(x)$

$$\left[\max\left(\hat{F}_k^*(x) - \lambda_k^{-1} L_{1-\alpha}, 0\right), \min\left(\hat{F}_k^*(x) + \lambda_k^{-1} L_{1-\alpha}, 1\right)\right], \quad x \in \mathbb{R}, \tag{14}$$

by replacing $F^*_{n_k}(x)$ in (11) with the smooth estimator $\hat{F}^*_k(x)$. There is also the "uncorrected" smooth SCB which covers both $F^*_{N_k}(x)$ and $F(x)$ with probability $1-\alpha$ asymptotically:

$$\left[\max\left(\hat{F}^*_k(x) - n_k^{-1/2}L_{1-\alpha}, 0\right),\ \min\left(\hat{F}^*_k(x) + n_k^{-1/2}L_{1-\alpha}, 1\right)\right],\quad x \in \mathbb{R}; \tag{15}$$

see Corollaries 1 and 2 for conditions on these SCBs. In particular, the smooth SCB of (15) is exactly the same smooth SCB in Wang et al. (2013). This phenomenon is not an accident. In fact, the dependent sample $X_{k1}, \ldots, X_{kn_k}$ can be viewed as an iid sample from the superpopulation cdf $F(x)$. Corollary 2 thus provides a new way of looking at the classic Kolmogorov–Smirnov SCB for $F(x)$ using an SRS from the finite population.

SCB is a powerful tool for statistical inference when the object of interest is an entire curve or function, see for instance, the construction of SCB for nonparametric regression curve in various contexts in Wang and Yang (2009), Cai and Yang (2015) and Gu and Yang (2015). Moreover, SCBs have been made available to new areas of statistics research such as functional data analysis; see Degras (2011), Cao et al. (2012), Ma et al. (2012), Song et al. (2014), Zheng et al. (2014), Gu et al. (2014) and Cao et al. (2016), and in particular Cardot and Josserand (2011) and Cardot et al. (2013) for SCB of functional data mean curve under survey sampling. The aforementioned SCB in (14) has also been extended by Wang et al. (2014) to least squares residuals to provide simultaneous coverage of the distribution function for latent errors in autoregressive time series.

To the best of our knowledge, the only existing work on SCB for the finite population distribution $F^*_{N_k}$ is Frey (2009), which cleverly devised recursive combinatorial algorithm to compute exact coverage probability of Kolmogorov–Smirnov type SCB without having to use asymptotic arguments. It is useful when the sample size and population size are both small. The recursive algorithm becomes intractable for larger sample size n . In fact, the sample and population sizes considered in the paper were $2 \leq n \leq 20$ and $4 \leq N \leq 320$. In addition, in this method the user cannot set the coverage probability to a predetermined confidence level $1-\alpha \in (0, 1)$ as it is determined by a critical value parameter in an implicit fashion. As a good complement of that approach, the SCBs proposed in our paper rely on the limiting distribution as they are designed for sample size $n_k \to \infty$, so that the asymptotic coverage probability can be set to any $1-\alpha \in (0, 1)$. As an illustration, Table 2 and Fig. 2 show the empirical coverage frequencies and plots of SCBs with sample sizes $n_k = 60, 100, 200, 300, 400$ for the baseball data from Lohr (2009). Furthermore, our large sample SCBs, both nonsmooth and smooth, "corrected" and "uncorrected", are also good for the superpopulation distribution F, which is beyond the combinatorial approach of Frey (2009).

A closely related work is by O'Neill and Stern (2012) who provided an appropriate adjustment factor for the unsmoothed empirical cdf but without detailed theory for the limiting distribution in the case of a sample drawn without replacement from a finite population. Their focus is on the population corrections for the Kolmogorov–Smirnov tests. We will make use of the same adjustment factor in our methodology. However, our focus is on developing both smoothed and unsmoothed SCBs for both the finite

Table 2 Coverage frequencies of log-salary distribution $F^*_{N_k}(x)$ using the corrected smooth SCBs with bandwidth h_1 and corrected nonsmooth SCBs

$N_k = 797$	SCB	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
$n_k = 60$	Smooth, h_1	0.995	0.975	0.944	0.876
	Nonsmooth	0.995	0.971	0.940	0.851
$n_k = 100$	Smooth, h_1	0.993	0.963	0.935	0.842
	Nonsmooth	0.993	0.964	0.926	0.833
$n_k = 200$	Smooth, h_1	0.995	0.966	0.928	0.832
	Nonsmooth	0.995	0.970	0.934	0.837
$n_k = 300$	Smooth, h_1	0.992	0.952	0.896	0.792
	Nonsmooth	0.996	0.965	0.918	0.825
$n_k = 400$	Smooth, h_1	0.985	0.946	0.897	0.783
	Nonsmooth	0.992	0.958	0.925	0.846

population distribution function and the superpopulation. Moreover, we will develop a rigorous asymptotic theory for our approach.

The rest of the paper is organized as follows. Section 2 contains the main theoretical results: a finite population analog of Donsker's Theorem and uniform closeness between the two estimators $F^*_{n_k}(x)$ and $\hat{F}^*_k(x)$ up to order $o_p(\lambda_k^{-1})$ and the rationale behind the two SCBs. In Sect. 3, we describe the steps to implement the SCBs. In Sect. 4, we carry out simulation studies with detailed results in the Supplementary Material. We also apply the smooth SCB to analyze a real data. In Sect. 5, we discuss the contributions of the proposed SCBs in relation to the existing literature and possible extensions to more sophisticated settings. All technical proofs are given in the Appendix.

2 Main results

In this section, the weak limit of stochastic processes $\lambda_k \left\{ F^*_{n_k}(x) - F^*_{N_k}(x) \right\}$ is established, as well as the uniform closeness between $F^*_{n_k}(x)$ and $\hat{F}^*_k(x)$, and between $F^*_{N_k}(x)$ and $F(x)$, under some general regularity conditions. These probabilistic results lead to SCBs in Corollaries 1 and 2 for the finite cdf $F^*_{N_k}(x)$ and its superpopulation cdf $F(x)$, based on $F^*_{n_k}(x)$ and $\hat{F}^*_k(x)$.

For any $\mu \in (0, 1]$ and nonnegative integer ν, denote by $C^{(\nu,\mu)}(\mathbb{R})$ the space of functions whose ν-th derivatives satisfy Hölder conditions of order μ

$$C^{(\nu,\mu)}(\mathbb{R}) = \left\{ \varphi : \mathbb{R} \to \mathbb{R} \,\middle|\, \|\varphi\|_{\nu,\mu} = \sup_{x,y\in\mathbb{R}} \frac{\left|\varphi^{(\nu)}(x) - \varphi^{(\nu)}(y)\right|}{|x-y|^{\mu}} < +\infty \right\}.$$

We assume the following general technical conditions:

(A0) $\lim_{k\to\infty}\min(n_k, N_k - n_k) = \infty$, and there exists a continuous distribution function $F(x)$ such that $\lim_{k\to\infty} F^*_{N_k}(x) = F(x), \forall x \in \mathbb{R}$.

(A1) There exists an integer $\nu \geq 0$ and $\mu \in (1/2, 1]$ such that $F \in C^{(\nu,\mu)}(\mathbb{R})$, and $F(x)$ is uniformly continuous over $x \in \mathbb{R}$.

(A2) The bandwidth $h = h_{n_k} > 0$ and $\lim_{k\to\infty} \lambda_k h_{n_k}^{\nu+\mu} = 0$.

(A3) The kernel K is a continuous and symmetric function, is supported on $[-1, 1]$, and is an l-th order kernel for some even integer $l > \nu + \mu$, i.e., its moments $\mu_r(K) = \int K(w) w^r \mathrm{d}w$ satisfy $\mu_0(K) \equiv 1$, $\mu_l(K) \neq 0$, $\mu_r(K) \equiv 0$ for any integer r, $0 < r < l$.

(A4) For each $k \geq 1$, the population $\pi_k = \{b_{k1}, b_{k2}, \ldots, b_{kN_k}\}$ is an iid random sample from population F, and $\sup_k n_k/N_k < 1$, i.e., $\mathrm{fpc}_k \geq C$ for some $C > 0$.

(A0) requires that n_k and $N_k - n_k$ go to infinity simultaneously as in Rosén (1964), while (A1)–(A3) parallel those in Wang et al. (2013) with a few new features. (A1) includes uniform continuity of $F(x)$ which does not necessarily follow from $F \in C^{(\nu,\mu)}(\mathbb{R})$ except when $\nu = 0$. (A2) has the scale factor λ_k in place of the typical $n_k^{1/2}$. (A3) allows the kernel K to have order higher than 2, and thus $\nu + \mu$ can be greater than 2. In contrast, Wang et al. (2013) limit K to be second-order nonnegative kernel, thus a probability density, and $\nu = 0, 1$ so $\nu + \mu$ is always less than 2. (A4) ensures that the difference $F^*_{N_k}(x) - F(x)$ is asymptotically $N_k^{-1/2} B\{F(x)\}$ and the modulus of continuity $F^*_{N_k}(x)$ is of order $o_p\left(\lambda_k^{-1}\right)$; see (22) in the proof of Theorem 2.

The following additional condition is imposed if we wish to ensure the convergence of $F^*_{N_k}(x)$ to $F(x)$ at the rate of $o_p\left(\lambda_k^{-1}\right)$ because $N_k^{-1/2} = o\left(\lambda_k^{-1}\right)$ and $\lim_{k\to\infty} n_k^{-1/2}/\lambda_k^{-1} = 1$; see Theorem 3.

(A4′) For each $k \geq 1$, the population $\pi_k = \{b_{k1}, b_{k2}, \ldots, b_{kN_k}\}$ is an iid random sample from population F, and $\lim_{k\to\infty} n_k/N_k = 0$, i.e., $\lim_{k\to\infty} \mathrm{fpc}_k = 1$.

The following theorem is analogous to Theorem 14.3 of Billingsley (1999) for the case of iid samples:

Theorem 1 *Under* (A0)*, there exists a version* B^* *of Brownian bridge such that as* $k \to \infty$, $\sup_{x\in\mathbb{R}} \left|\lambda_k \left\{F^*_{n_k}(x) - F^*_{N_k}(x)\right\} - B^*\{F(x)\}\right| \overset{a.s.}{\to} 0$ *and consequently* $\lambda_k \left\{F^*_{n_k}(x) - F^*_{N_k}(x)\right\} \overset{d}{\to} B\{F(x)\}$.

The next theorem extends Wang et al. (2013) to finite population:

Theorem 2 *Under (A0)–(A4), as* $k \to \infty$*, the maximal deviation* $D\left(F^*_{n_k}, \hat{F}^*_k\right)$ *satisfies* $\lambda_k D\left(F^*_{n_k}, \hat{F}^*_k\right) = \lambda_k \sup_{x\in\mathbb{R}} \left|F^*_{n_k}(x) - \hat{F}^*_k(x)\right| = o_p(1)$*, and consequently* $\lambda_k \left\{\hat{F}^*_k(x) - F^*_{N_k}(x)\right\} \overset{d}{\to} B\{F(x)\}$.

Theorems 1 and 2 imply the next corollary, with the Kolmogorov distribution $L(t)$ defined in (8) and $1 - \alpha \in (0, 1)$ a predetermined confidence level.

Corollary 1 *Under* (A0), $\lim_{k\to\infty}\mathbb{P}\left[\lambda_k D\left(F^*_{n_k}, F^*_{N_k}\right) \le t\right] = L(t)$ *implying that the SCB in* (11) *is asymptotically* $100(1-\alpha)\%$ *for the finite cdf* $F^*_{N_k}(x)$. *Under (A0)–(A4),* $\lim_{k\to\infty}\mathbb{P}\left[\lambda_k D\left(\hat{F}^*_k, F^*_{N_k}\right) \le t\right] = L(t)$ *implying that the smooth SCB in* (14) *is asymptotically* $100\,(1-\alpha)\ \%$ *for the finite cdf* $F^*_{N_k}(x)$.

The next theorem extends the relationships between $F^*_{n_k}$ and $F^*_{N_k}$ in Theorem 1 and between $\hat{F}^*_k$ and $F^*_{N_k}$ in Theorem 2 to the relationships between $F^*_{n_k}$ and F and between $\hat{F}^*_k$ and F.

Theorem 3 *Under* (A0) *and* (A4′), *as* $k \to \infty$, $n_k^{-1/2}/\lambda_k^{-1} \to 1$, $n_k^{1/2} D\left(F^*_{N_k}, F\right) \xrightarrow{d} 0$ *and* $\lambda_k D\left(F^*_{N_k}, F\right) \xrightarrow{d} 0$. *Hence,* $\lim_{k\to\infty}\mathbb{P}\left[n_k^{1/2} D\left(F^*_{n_k}, F^*_{N_k}\right) \le t\right] = L(t)$,

$$\lim_{k\to\infty}\mathbb{P}\left[n_k^{1/2} D\left(F^*_{n_k}, F\right) \le t\right] = L(t) \text{ and } \lim_{k\to\infty}\mathbb{P}\left[\lambda_k D\left(F^*_{n_k}, F\right) \le t\right] = L(t), t \in \mathbb{R}.$$

Under (A0)–(A3) *and* (A4′), *as* $k \to \infty$, $\lim_{k\to\infty}\mathbb{P}\left[n_k^{1/2} D\left(\hat{F}^*_k, F^*_{N_k}\right) \le t\right] = L(t)$,

$$\lim_{k\to\infty}\mathbb{P}\left[n_k^{1/2} D\left(\hat{F}^*_k, F\right) \le t\right] = L(t) \text{ and } \lim_{k\to\infty}\mathbb{P}\left[\lambda_k D\left(\hat{F}^*_k, F\right) \le t\right]=L(t), t \in \mathbb{R}.$$

Corollary 2 *Under* (A0) *and* (A4′), *the "uncorrected" SCB in* (12) *is asymptotically* $100(1-\alpha)\%$ *for the finite cdf* $F^*_{N_k}(x)$ *and superpopulation cdf* $F(x)$, *the SCB in* (11) *is also asymptotically* $100(1-\alpha)\%$ *for the superpopulation cdf* $F(x)$. *Under* (A0)–(A3) *and* (A4′), *the "uncorrected" smooth SCB in* (15) *is asymptotically* $100(1-\alpha)\%$ *for the finite cdf* $F^*_{N_k}(x)$ *and superpopulation cdf* $F(x)$, *the smooth SCB in* (14) *is also asymptotically* $100(1-\alpha)\%$ *for the superpopulation cdf* $F(x)$.

Note that the weak convergence in Corollaries 1 and 2 is according to the uniform metric on the cadlag space, hence it facilitates nicely the construction of asymptotic SCBs. Another important point to make here is that these SCBs are for large samples ($\lim_{k\to\infty}\min(n_k, N_k - n_k) = \infty$), and they differ fundamentally from those in Frey (2009), which are designed for small population sizes. For instance, the average GPA data analyzed in Frey (2009) have $N = 65$, $n = 15$, while the analysis of the baseball salary data in Sect. 4 has $N = 797$, $60 \le n \le 400$.

3 Implementation

In this section, we describe procedures to construct the SCBs based on estimators $F^*_{n_k}(x)$ and $\hat{F}^*_k(x)$ defined in (6) and (13), respectively. According to Corollaries 1 and 2, for sample size $n_k > 50$, the corrected and uncorrected nonsmooth and smooth SCBs for the finite population distribution function and superpopulation cdf are computed and named as follows:

SCB in (11), “corrected, nonsmooth” ,
SCB in (14), “corrected, smooth”,
SCB in (12), “uncorrected, nonsmooth” ,
SCB in (15), “ uncorrected, smooth” .

Using the quartic kernel $K(u) = 15(1-u^2)^2 I\{|u| \leq 1\}/16$, the proposed function $\hat{F}_k^*(x)$ is computed as:

$$\hat{F}_k^*(x) = n_k^{-1} \sum\nolimits_{i=1}^{n_k} \int_{-\infty}^{x} h_j^{-1} K\left(\frac{u - X_{ki}}{h_j}\right) \mathrm{d}u, \quad j = 1, 2$$

in which $h_1 = \mathrm{IQR} \times \lambda_k^{-2}$ and $h_2 = \mathrm{IQR} \times \lambda_k^{-2/3}$, where IQR stands for the Inter-Quartile Range of $\left\{X_{k1}, \ldots, X_{kn_k}\right\}$. The bandwidth h_1 automatically satisfies (A2), while h_2 satisfies (A2) if $\nu + \mu > 3/2$, which is the case if $\nu = 1$ since $\mu > 1/2$, these bandwidths are similar to those used in Wang et al. (2013).

4 Simulated and real data examples

4.1 General simulation studies

In this section, we numerically investigate the coverage frequency performances of the proposal methods by simulation studies. The focus here is to display the performance of the various SCBs based on estimators $\hat{F}_k^*(x)$ and $F_{n_k}^*(x)$ for the finite distribution function $F_{N_k}^*(x)$ and superpopulation distribution function $F(x)$.

In the simulation studies, we generate the finite population π_k from different distributions to study the coverage sensitivity of the superpopulation. Specifically, we let the superpopulation be one of the standard normal, exponential, Cauchy, and beta distributions with the form given below:

$$\begin{aligned}
&F(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-u^2/2) \mathrm{d}u, \\
&F(x) = \left(1 - e^{-x}\right) I\,(x > 0)\,, \\
&F(x) = \pi^{-1} \arctan x + 1/2, \text{ or} \\
&F(x) = \{\Gamma(\beta_1)\,\Gamma(\beta_2)\}^{-1}\,\Gamma(\beta_1 + \beta_2) \int_0^x u^{\beta_1 - 1} (1-u)^{\beta_2 - 1} \mathrm{d}u
\end{aligned}$$

in which $\Gamma(\cdot)$ represents the Gamma function. An iid sample of size N_k is first drawn from distribution F to be used as the population π_k. We define the coverages of both the finite population distribution $F_{N_k}^*(x)$ and the superpopulation $F(x)$ as the relative frequencies of these functions at all the points in π_k covered by the band of the smooth/non-smooth type estimator. The coverage is precise for $F_{N_k}^*(x)$ but is a close approximation for $F(x)$ since all the points in π_k are used as check points for the entire domain of $F(x)$. This approximation may lead to some slight coverage differences for $F(x)$ apart from simulation errors among different population sizes N_k even if the sample sizes n_k stay the same. Samples $\left\{X_{kn_1}, \ldots, X_{kn_k}\right\}$ of size n_k are then drawn without replacement from the finite population π_k.

The population and sample sizes are selected as $(n_k, N_k) = (60, 200)$, $(100, 200)$, $(60, 500)$, $(100, 500)$, with confidence levels $1 - \alpha = 0.99, 0.95, 0.90, 0.80$ for constructing SCBs. Tables S.1–S.5 in the Supplementary Material display the coverage frequencies over 1000 replications of the various SCBs at all data points. Main findings are summarized as follows:

1. Coverage frequencies of uncorrected SCBs are always significantly higher than the corrected ones. Smooth SCBs with bandwidth h_1 have almost the same coverage frequencies as nonsmooth SCBs, while smooth SCBs with bandwidth h_2 almost always have higher coverage frequencies.
2. All SCBs are conservative except the corrected SCBs for F with sampling fraction $n_k/N_k > 1/5$, which comes short of the condition $\lim_{k\to\infty} \mathrm{fpc}_k = 1$ in (A4'). Thus, Corollary 2 fails and SCBs for F do not work well.
3. In general, smaller sampling fraction n_k/N_k (or higher finite population correction fpc_k) leads to more satisfactory performance of corrected SCBs for F and the uncorrected SCBs for $F^*_{N_k}$ (rows 2-3 of each cell), e.g., $(n_k, N_k) = (60, 500)$ being better than $(n_k, N_k) = (60, 200)$ due to Theorem 2 that imposes (A4) on n_k/N_k to have upper bound less than 1. The corrected SCBs for $F^*_{N_k}$ (row 1 in each cell) are not affected by n_k/N_k as Theorem 1 does not require n_k/N_k to be small.

To visualize the SCBs, Figure S.1 in the Supplementary Material depicts the superpopulation cdf F (thick) in the normal distribution case, the finite cdf $F^*_{N_k}$ (solid), the smooth KDE $\hat{F}^*_k$ together with its 95 % SCB (solid), the empirical cdf $F^*_{n_k}$ together with its 95 % SCB (dotted), for one specifically simulated finite population and one sample in each simulation setting. As one reviewer suggested, in all the subfigures we use the sample with the median confidence band width in the 1000 runs. To save space, only the combinations of $(n_k, N_k) = (60, 200)$ and $(100, 200)$ are included. The figures for other distributions are similar and thus omitted. One clearly sees that the SCBs constructed based on $F^*_{n_k}$ and $\hat{F}^*_k$, and the estimators themselves are nearly indistinguishable from each other. For the same population size of 200, the SCBs tend to be narrower as the sample size n_k increases and λ_k^{-1} decreases, which confirms Corollaries 1 and 2.

4.2 Further simulation studies

In response to reviewers' comments, we have conducted further simulations in the following cases: (1) The finite population size is fixed to be $N_k = 5000$, while the sample size is taken to be $n_k = 50$, 250, 500, 1000, and 4000, respectively; (2) The finite sample size is fixed to be $n_k = 60$, while the population sizes are $N_k = 120$, 200, 600, 2000, and 10,000, respectively. Again, the finite population π_k is generated from the standard normal, exponential, Cauchy, and beta distributions with 1000 simulation replications, but to save space we only report the results for the standard normal distribution; the results for other distributions are similar. Representative numerical results are given in Tables S.6 and S.7 in the Supplementary Material.

Patterns similar to those in the general simulation studies are observed. The performance of the corrected SCBs for $F^*_{N_k}$ and of the uncorrected SCBs for F is quite good regardless of the sample and population sizes. Meanwhile, the performance of the uncorrected SCBs for $F^*_{N_k}$ and of the corrected SCBs for F is not that bad when the sampling fraction n_k/N_k is small. Therefore, the sampling fraction n_k/N_k plays an important role in the performance of the estimation and SCB coverage frequencies. In addition, with n_k/N_k fixed increasing both n_k and N_k help to improve the performance as expected. These findings are again consistent with our theoretical results.

Furthermore, Figure S.2 in the Supplementary Material shows $D(\hat{F}^*_k, F^*_{N_k})/D(F^*_{n_k}, F^*_{N_k})$ and $D(\hat{F}^*_k, F)/D(F^*_{n_k}, F)$ with fixed finite population size $N_k = 5000$ for the case of the normal distribution. In this figure, we see that for the "corrected" version the medians of the ratios are all below 1, indicating that the "smooth" estimator $\hat{F}^*_k$ is a better one for the cdf. This was also observed in Wang et al. (2013, 2014) due to its smoothness for an essentially continuous distribution function.

In general, the corrected SCBs, the nonsmooth one and the smooth one with bandwidth h_1 both generally perform well for estimating the finite population distribution function $F^*_{N_k}$, and the smooth SCB is the natural choice for general use since we essentially estimate the cdf of a continuous random variable. The corrected smooth SCB with bandwidth h_2 can be used when one is reasonably certain that F has smoothness order $\nu + \mu > 3/2$. Although the smooth bands have higher coverage frequencies than the unsmoothed ones, they are especially useful for the superpopulation cdf when the random variable is continuous. Furthermore, the smooth estimators have asymptotically smaller Mean Integral Squared Error (MISE) than the empirical estimator of cdf $F(x)$, as observed and studied in Wang et al. (2013).

We recommend using the corrected smooth SCB with bandwidth h_1 in practice, together with the corrected nonsmooth SCB for validation purposes. The corresponding uncorrected SCBs can be used when the sampling fraction $n_k/N_k < 1/2$. In the next subsection, we apply the recommended procedures to a real data set for illustration.

4.3 Real data application

In this subsection, we apply the proposed method to an illustrative data set obtained for a collection of baseball activity, named as baseball data in Lohr (2009). The data file contains information on 797 baseball players from the rosters of all major league teams in November, 2004. Of particular interest is the salary of players, the smoothed relative frequency plot of which in Fig. 1b shows strong skewness. Following Lohr (2009), we work with natural logarithm of the salary (log-salary), whose smoothed relative frequency plot in Fig. 1a is much less skewed. Our goal is to examine if the proposed SCBs contain the distribution function $F^*_{N_k}(x)$ of log salary for the $N_k = 797$ players with reasonable coverage frequencies.

For each sample size $n_k = 60, 100, 200, 300, 400$, 1000 SRS samples are drawn, respectively. Table 2 shows the coverage frequencies of the finite population distri-

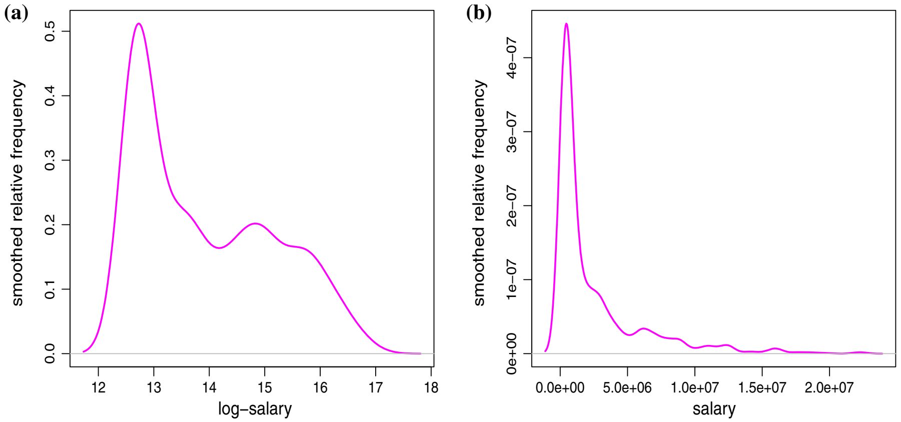

Fig. 1 Smoothed relative frequency plot of salary and log-salary of the baseball data: **a** log-salary; **b** salary

bution $F^*_{N_k}(x)$ by the corrected smooth SCB with bandwidth h_1 and the corrected nonsmooth SCB. It is seen that while both SCBs are conservative, the coverage frequencies of the smooth SCB approach the nominal levels with increasing sampling fraction n_k/N_k, especially when $n_k \geq 300$. The high coverage frequencies may be due to the existence of many repeated values. In general, the more discrete the population is, the more conservative the bands tend to be.

Figure 2 depicts the corrected SCBs described previously with sample size $n_k =$ 60, 100, 200, 300 in one of the 1000 runs, respectively. In each of subfigures (a)–(d), the center dotted, solid and thick lines are the empirical cdf, the proposed smooth kernel distribution and the finite distribution function (d. f.), while the upper (lower) solid (dotted) lines represent the corrected smooth (nonsmooth) SCBs for the finite distribution. These numerical and graphical results provide ample evidence that the proposed SCBs, especially the smooth one, are robust devices for statistical inference on the finite population distribution.

5 Discussion

In this paper, we have proposed both smooth and nonsmooth SCBs for the distribution function of a finite population and of its latent superpopulation when a simple random sample is drawn from the finite population. These SCBs have been shown to achieve the confidence levels theoretically and empirically. For the finite population, these SCBs have extended the classic Kolmogorov–Smirnov SCB based on an iid sample, while for the superpopulation, they have given new ways of viewing the classic SCBs. Extending these SCBs to more sophisticated (and often more effective) sampling designs such as stratified sampling (or more generally high entropy sampling designs) is highly desirable but at the same time quite non-trivial. Further investigations along this line are ongoing.

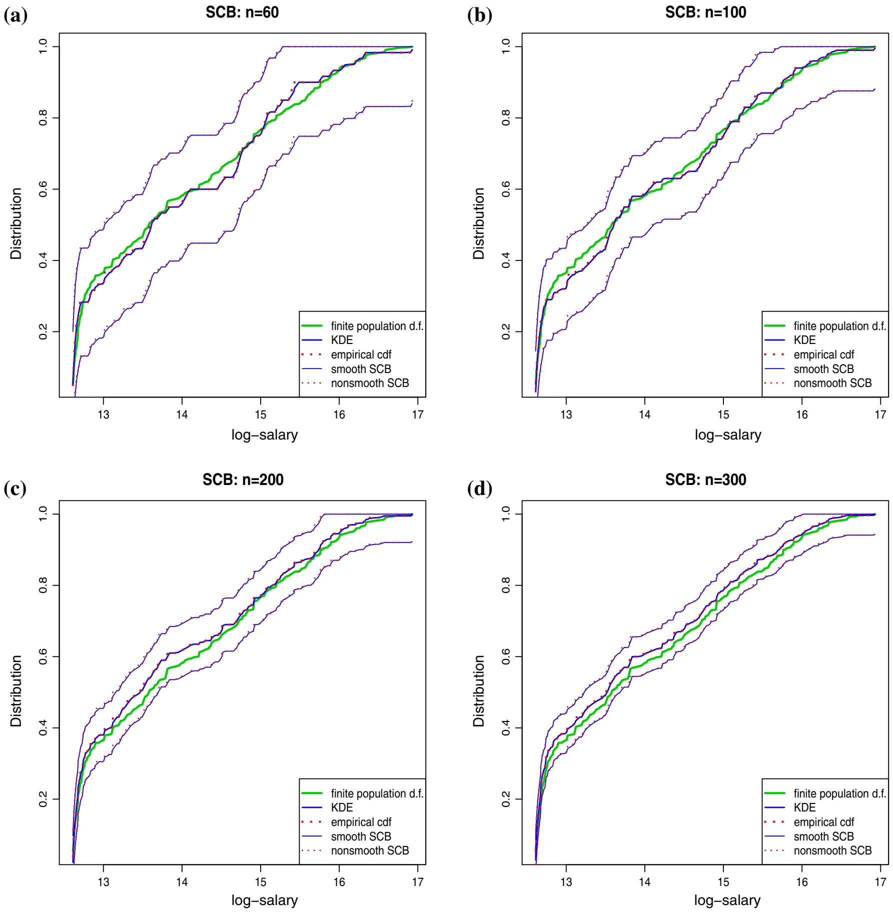

Fig. 2 Corrected smooth SCBs with bandwidth h_1 and corrected nonsmooth SCBs for log-salary at $\alpha = 0.1$ and **a**–**d** $n_k = 60, 100, 200, 300$ with $N_k = 797$, respectively

6 Supplementary material

The Supplementary Material published online along with this work contains extensive supporting numerical and displaying results of the simulation studies. It is available jointly with the published paper.

Acknowledgments We thank four reviewers, an Associate Editor and the Editor-in-Chief Ana F. Militino for their helpful comments and suggestions that have led to a much improved version of this paper. This research was supported in part by Jiangsu Key-Discipline Program (Statistics) ZY107002, ZY107992, National Natural Science Foundation of China award 11371272, Research Fund for the Doctoral Program of Higher Education of China award 20133201110002, Soochow University Excellent Doctoral Dissertation Project 233200113 and Jiangsu Graduate Students' Innovative Research Project KYZZ_0331.

Appendix

Throughout this section, c denotes any positive constant and O_p (or o_p) a sequence of random variables of certain order in probability. For instance, $o_p(\lambda_k^{-1})$ means a smaller order than λ_k^{-1} in probability. In addition, u_p denotes a sequence of random functions which are o_p uniformly defined in the domain.

Recall that $\lambda_k = \left(n_k^{-1} - N_k^{-1}\right)^{-1/2}$ and b_{ki}, $i = 1, 2, \ldots N_k$, are the elements of π_k. The following is a restatement of Theorem 14.1 of Rosén (1964).

Lemma 1 *If the sequence* $\{\pi_k\}_1^\infty$ *satisfies:* (a) $\lim_{k\to\infty} \min(n_k, N_k - n_k) = \infty$ *and* (b) *the elements of* π_k *are in* $[0, 1]$, *and* $\lim_{k\to\infty} F_{N_k}^*(t) = t$ *for* $0 \le t \le 1$, *then as* $k \to \infty$, $\lambda_k \left\{F_{n_k}^*(t) - F_{N_k}^*(t)\right\} \xrightarrow{d} B(t)$ *where* $B(\cdot)$ *represents the Brownian bridge.*

Next, we will use Lemma 1 to prove Theorem 1.

Proof of Theorem 1

Define $u_{ki} = F(b_{ki})$, $i = 1, 2, \ldots N_k$ as the elements in a population $\pi_{k,U} = \{u_{k1}, u_{k2}, \ldots, u_{kN_k}\}$ and that $U_{ki} = F(X_{ki})$, $1 \le i \le n_k$, as a simple random sample from population $\pi_{k,U}$. The finite cdf $F_{U,N_k}^*(t)$ of $\pi_{k,U}$ and the empirical cdf F_{U,n_k}^* of $\pi_{k,U}$ are

$$F_{U,N_k}^*(t) = N_k^{-1} \sum_{i=1}^{N_k} I\{u_{ki} \le t\}, \tag{16}$$

$$F_{U,n_k}^*(t) = n_k^{-1} \sum_{i=1}^{n_k} I\{U_{ki} \le t\}, \tag{17}$$

analogous to (5) and (6). For any $x \in \mathbb{R}$, let $t = F(x) \in [0, 1]$. Then, one can show that

$$F_{N_k}^*(x) = F_{U,N_k}^*(t),\ F_{n_k}^*(x) = F_{U,n_k}^*(t). \tag{18}$$

For instance, it is straightforward to verify that

$$F_{U,N_k}^*(t) = N_k^{-1} \sum_{i=1}^{N_k} I\{u_{ki} \le F(x)\} = N_k^{-1} \sum_{i=1}^{N_k} I\{b_{ki} \le x\} = F_{N_k}^*(x).$$

By (A0), $\lim_{k\to\infty} F_{N_k}^*(x) = F(x) = t$. Hence, $\lim_{k\to\infty} F_{U,N_k}^*(t) = t$ for $0 \le t \le 1$ according to (18). It is also clear that for $i = 1, 2, \ldots N_k$, $0 \le u_{k_i} \le 1$. Thus, all the conditions of Lemma 1 are fulfilled, indicating that the following holds according to the uniform metric on the space $\mathcal{D}[0, 1]$ of cadlag functions:

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

Applying Skorohod's Representation Theorem (Theorem 6.7, Billingsley 1999), there exists a version B^* of Brownian bridge such that

$$\sup_{t\in[0,1]} \left|\lambda_k \left\{F^*_{U,n_k}(t) - F^*_{U,N_k}(t)\right\} - B^*(t)\right| \to 0, \text{ a.s.}$$

Combining this and (18) completes the proof of Theorem 1.

Proof of Theorem 2

Define $G(x) = \int_{-\infty}^{x} K(u)\,\mathrm{d}u$. By the definition of $\hat{F}_k^*(x)$, one obtains

$$\hat{F}_k^*(x) = n^{-1}\sum_{i=1}^{n_k}\int_{-\infty}^{x} K_h(u - X_{ki})\,\mathrm{d}u = n_k^{-1}\sum_{i=1}^{n_k} G\left(\frac{x - X_{ki}}{h}\right).$$

Therefore, by the definition of $F^*_{n_k}(x) = n_k^{-1}\sum_{i=1}^{n_k} I(X_{ki} \le x)$ in (6)

$$\begin{aligned}\hat{F}_k^*(x) &= \int_{-\infty}^{+\infty} G\left(\frac{x-u}{h}\right)\mathrm{d}F^*_{n_k}(u) = \int_{-\infty}^{+\infty} h^{-1}K\left(\frac{x-u}{h}\right)F^*_{n_k}(u)\,\mathrm{d}u\\ &= \int_{-1}^{1} K(w)\,F^*_{n_k}(x - hw)\,\mathrm{d}w\end{aligned}$$

using integration by parts and a change of variable $w = (x-u)/h$. The following decomposition plays an important role:

$$\hat{F}_k^*(x) - F^*_{n_k}(x) = \int_{-1}^{+1}\left\{F^*_{n_k}(x - hw) - F^*_{n_k}(x)\right\}K(w)\,\mathrm{d}w. \tag{19}$$

It is easy to see the following inequalities:

$$\begin{aligned}&\sup_{w\in[-1,1],x\in\mathbb{R}}\left|\lambda_k\left\{F^*_{n_k}(x-hw) - F^*_{N_k}(x-hw)\right\} - \lambda_k\left\{F^*_{n_k}(x) - F^*_{N_k}(x)\right\}\right|\\ &\quad\le \sup_{x,x'\in\mathbb{R},|x-x'|\le h}\left|\lambda_k\left\{F^*_{n_k}(x') - F^*_{N_k}(x')\right\} - \lambda_k\left\{F^*_{n_k}(x) - F^*_{N_k}(x)\right\}\right|\\ &\quad\le 2\sup_{x}\left|\lambda_k\left\{F^*_{n_k}(x) - F^*_{N_k}(x)\right\} - B^*\{F(x)\}\right|\\ &\qquad + \sup_{x,x'\in\mathbb{R},|x-x'|\le h}\left|B^*\{F(x')\} - B^*\{F(x)\}\right|.\end{aligned} \tag{20}$$

Denote $\omega(F,h) = \sup_{x,x'\in\mathbb{R},|x-x'|\le h}\left|F(x') - F(x)\right|$ as the modulus of continuity for F, which is uniformly continuous by (A1) so that $\omega(F,h) \to 0$ as $k \to \infty$. According to Theorem 1, $\sup_{x\in\mathbb{R}}\left|\lambda_k\left\{F^*_{n_k}(x) - F^*_{N_k}(x)\right\} - B^*\{F(x)\}\right|$

$\overset{a.s.}{\to} 0$, as $k \to \infty$, and thus in probability. Thus, the expression in (20) is bounded by

$$2 \sup_{x \in \mathbb{R}} \left| \lambda_k \left\{ F^*_{n_k}(x) - F^*_{N_k}(x) \right\} - B^* \{F(x)\} \right| + \sup_{t,t' \in [0,1], |t-t'| \leq \omega(F,h)} \left| B^*\left(t'\right) - B^*(t) \right|$$
$$= o_p(1) + o_p(1) = o_p(1).$$

In other words,

$$\left\{ F^*_{n_k}(x - hw) - F^*_{n_k}(x) \right\} - \left\{ F^*_{N_k}(x - hw) - F^*_{N_k}(x) \right\} = u_p\left(\lambda_k^{-1}\right). \quad (21)$$

Similarly, since (A4) implies that $N_k^{1/2} \left\{ F^*_{N_k}(x) - F(x) \right\} \xrightarrow{d} B\{F(x)\}$, the following holds

$$\left\{ F^*_{N_k}(x - hw) - F^*_{N_k}(x) \right\} - \{F(x - hw) - F(x)\} = u_p\left(N_k^{-1/2}\right). \quad (22)$$

Since (A4) requires that $\sup_k n_k/N_k < 1$ and consequently $N_k^{-1/2} = O\left(\lambda_k^{-1}\right)$, (21) and (22) imply that as $k \to \infty$

$$\left| \left\{ F^*_{n_k}(x - hw) - F^*_{n_k}(x) \right\} - \{F(x - hw) - F(x)\} \right| = u_p\left(\lambda_k^{-1}\right). \quad (23)$$

Note that by (A3) $\int_{-1}^{1} K(w) w^r \mathrm{d}w \equiv 0, r = 1, \ldots, l-1$, and by (A1) $F(x) \in C^{(\nu,\mu)}(\mathbb{R})$. Hence

$$\int_{-1}^{1} \{F(x - hw) - F(x)\} K(w) \mathrm{d}w$$
$$= \int_{-1}^{1} \left\{ F(x - hw) - \sum_{r=0}^{\nu} \frac{F^{(r)}(x)}{r!} (-hw)^r \right\} K(w) \mathrm{d}w.$$

Furthermore, by (A1) $F(x) \in C^{(\nu,\mu)}(\mathbb{R})$ and

$$\left| \int_{-1}^{1} \{F(x - hw) - F(x)\} K(w) \mathrm{d}w \right|$$
$$\leq \int_{-1}^{1} \left| F(x - hw) - \sum_{r=0}^{\nu} \frac{F^{(r)}(x)}{r!} (-hw)^r \right| K(w) \mathrm{d}w$$
$$\leq \int_{-1}^{1} c h^{\nu+\mu} |w|^{\nu+\mu} K(w) \mathrm{d}w = O\left(h^{\nu+\mu}\right) = o\left(\lambda_k^{-1}\right), \quad (24)$$

which follows from (A2) that $\lim_{k \to \infty} \lambda_k h_{n_k}^{\nu+\mu} = 0$.

By applying (19), (23) and (24), the following holds

$$\sup_{x\in\mathbb{R}}\left|\hat{F}_k^*(x)-F_{n_k}^*(x)\right|=\sup_{x\in\mathbb{R}}\left|\int_{-1}^{1}\left\{F_{n_k}^*(x-hw)-F_{n_k}^*(x)\right\}K(w)\,\mathrm{d}w\right|=o_p\left(\lambda_k^{-1}\right).$$

Applying Theorem 1, one has $\lambda_k\left\{\hat{F}_k^*(x)-F_{N_k}^*(x)\right\}\overset{d}{\to}B\{F(x)\}$. Thus, the proof of Theorem 2 is complete.

Proof of Theorem 3

Notice that under (A4′), $n_k^{-1/2}/\lambda_k^{-1}\to 1$, $N_k^{-1/2}=o\left(\lambda_k^{-1}\right)$, $N_k^{-1/2}=o\left(n_k^{-1/2}\right)$ as $k\to\infty$, and that $N_k^{1/2}\left\{F_{N_k}^*(x)-F(x)\right\}\overset{d}{\to}B\{F(x)\}$. Hence, as $k\to\infty$

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

Likewise, $\lambda_k D\left(F_{N_k}^*,F\right)=o_p(1)$. The rest of Theorem 3 follows by applying Slutsky's Theorem.

References

Billingsley P (1999) Convergence of probability measures, 2nd edn. Wiley, New York

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

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

Cao G, Yang L, Todem D (2012) Simultaneous inference for the mean function based on dense functional data. J Nonpar Stat 24:359–377

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

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

Chambers RL, 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. Stat Sin 12:1223–1239

Cheng M, Peng L (2002) Regression modeling for nonparametric estimation of distribution and quantile functions. Stat Sin 12:1043–1060

Degras D (2011) Simultaneous confidence bands for nonparametric regression with functional data. Stat Sin 21:1735–1765

Falk M (1985) Asymptotic normality of the kernel quantile estimator. Ann Stat 13:428–433

Francisco C, Fuller W (1991) Quantile estimation with a complex survey design. Ann Stat 19:454–469

Frey J (2009) Confidence bands for the CDF when sampling from a finite population. Comput Stat Data Anal 53:4126–4132

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

Gu L, Yang L (2015) Oracally efficient estimation for single-index link function with simultaneous confidence band. Electron J Stat 9:1540–1561

Liu R, Yang L (2008) Kernel estimation of multivariate cumulative distribution function. J Nonpar Stat 20:661–677

Lohr s (2009) Sampling: design and analysis, 2nd edn. Brooks/Cole, Boston

Ma S, Yang L, Carroll R (2012) A simultaneous confidence band for sparse longitudinal regression. Stat Sin 22:95–122

O'Neill T, Stern S (2012) Finite population corrections for the Kolmogorov-Smirnov tests. J Nonpar Stat 24:497–504

Reiss R (1981) Nonparametric estimation of smooth distribution funtions. Scand J Stat 8:116–119

Rosén B (1964) Limit theorems for sampling from finite populations. Arkiv för Matematik 5:383–424

Song Q, Liu R, Shao Q, Yang L (2014) A simultaneous confidence band for dense longitudinal regression. Commun Stat Theory Methods 43:5195–5210

Wang J, Cheng F, Yang L (2013) Smooth simultaneous confidence bands for cumulative distribution functions. J Nonpar Stat 25:395–407

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

Wang J, Yang L (2009) Polynomial spline confidence bands for regression curves. Stat Sin 19:325–342

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

Xue L, Wang J (2010) Distribution function estimation by constrained polynomial spline regression. J Nonpar Stat 22:443–457

Yamato H (1973) Uniform convergence of an estimator of a distribution function. Bull Math Stat 15:69–78

Zheng S, Yang L, Härdle W (2014) A smooth simultaneous confidence corridor for the mean of sparse functional data. J Am Stat Assoc 109:661–673