CrossMark

Inference on a distribution function from ranked set samples

Lutz Dümbgen[1] · Ehsan Zamanzade[2]

Abstract

Consider independent observations (X_i, R_i) with random or fixed ranks R_i, while conditional on R_i, the random variable X_i has the same distribution as the R_i-th order statistic within a random sample of size k from an unknown distribution function F. Such observation schemes are well known from ranked set sampling and judgment post-stratification. Within a general, not necessarily balanced setting we derive and compare the asymptotic distributions of three different estimators of the distribution function F: a stratified estimator, a nonparametric maximum-likelihood estimator and a moment-based estimator. Our functional central limit theorems generalize and refine previous asymptotic analyses. In addition, we discuss briefly pointwise and simultaneous confidence intervals for the distribution function with guaranteed coverage probability for finite sample sizes. The methods are illustrated with a real data example, and the potential impact of imperfect rankings is investigated in a small simulation experiment.

1 Introduction

Ranked set sampling and judgment post-stratification are both sampling strategies in situations in which ranking several observations is possible and relatively easy without referring to exact values, whereas obtaining complete observations is much more

✉ Ehsan Zamanzade
e.zamanzade@sci.ui.ac.ir

1   Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland

2   Department of Statistics, University of Isfahan, Isfahan 81746-73441, Iran

involved. For instance, this occurs often in agriculture or forestry when the quantities of interest are yields on different plots or of different trees. Good overviews of theory and applications of ranked set sampling are given by Wolfe (2004, 2012) and Chen et al. (2004). Let us explain the two sampling schemes just mentioned in a simple hypothetical example: Suppose we want to estimate the distribution of body heights among all men of age 20–25 in a certain population. Whenever we have obtained a precise measurement X_i of such a man, we could compare him to $k - 1$ additional young men and note the rank $R_i \in \{1, 2, \ldots, k\}$ of X_i within this small group without measuring the heights of the additional men precisely. This sampling scheme is called judgment post-stratification (JPS), see MacEachern et al. (2004). Alternatively, for each observation we could recruit a group of k young men, rank them with respect to their heights and then obtain the precise body height X_i of the person with rank $R_i \in \{1, \ldots, k\}$ only. Here the ranks $R_1, R_2, \ldots, R_n$ have been specified in advance. This sampling scheme, called ranked set sampling (RSS), was introduced by McIntyre (1952). If the empirical distribution of the ranks R_i is (approximately) uniform on $\{1, \ldots, k\}$, one talks about balanced RSS, otherwise unbalanced RSS. For instance, if we are mainly interested in the upper tail of the distribution of body heights, we could favor larger ranks R_i.

In general, we consider independent random pairs $(X_1, R_1), (X_2, R_2), \ldots, (X_n, R_n)$ with fixed or random ranks $R_i \in \{1, 2, \ldots, k\}$. Conditional on $R_i = r$, the random variable X_i has the same distribution as the r-th order statistic of a random sample of size k from F. That means, X_i has distribution function

$$F_r(x) := \mathbb{P}(X_i \leq x \mid R_i = r) = B_r(F(x)),$$

where $B_r : [0, 1] \to [0, 1]$ denotes the distribution function of the beta distribution with parameters r and $k + 1 - r$. Thus for $p \in [0, 1]$,

$$B_r(p) = \sum_{i=r}^{k} \binom{k}{i} p^i (1 - p)^{k-i} = \int_0^p \beta_r(u) \, du$$

with

$$\beta_r(u) = C_r u^{r-1} (1 - u)^{k-r} \quad \text{and} \quad C_r = k \binom{k-1}{r-1} = k \binom{k-1}{k-r},$$

see David and Nagaraja (2003). The vector $N_n = (N_{nr})_{r=1}^{k}$ of stratum sizes

$$N_{nr} := \sum_{i=1}^{n} 1_{[R_i = r]}$$

plays a key role. In RSS, the ranks $R_1, R_2, \ldots, R_n$ and thus the whole vector N_n are fixed. In JPS, the R_i are independent and uniformly distributed on $\{1, \ldots, k\}$, whence N_n follows a multinomial distribution Mult $(n; 1/k, \ldots, 1/k)$.

Several estimators of the c.d.f. F have been proposed. Of course one could just ignore the rank information and compute the empirical c.d.f. $\widehat{F}_n$,

$$\widehat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \leq x]}.$$

In the JPS setting, this estimator is unbiased and $\sqrt{n}$-consistent. However, the stratified estimator

$$\widehat{F}_n^{\mathrm{S}} := \frac{1}{\#\{r : N_{nr} > 0\}} \sum_{r \,:\, N_{nr} > 0} \widehat{F}_{nr}$$

with the empirical c.d.f.

$$\widehat{F}_{nr}(x) := \frac{1}{N_{nr}} \sum_{i=1}^{n} 1_{[R_i = r, \, X_i \leq x]}$$

within stratum $\{i : R_i = r\}$ is usually more efficient. It has been introduced and analyzed in a balanced RSS setting by Stokes and Sager (1988). Refinements and modifications of this estimator $\widehat{F}_n^{\mathrm{S}}$ in the JPS setting have been proposed by Frey and Ozturk (2011) and Wang et al. (2012). In particular, these authors consider situations with small or moderate sample sizes so that some stratum sizes N_{nr} may be zero or the empirical c.d.f.s $\widehat{F}_{nr}$ may fail to satisfy order relations which are known for their theoretical counterparts F_r.

A second approach to estimating the c.d.f. F which can also handle empty strata was introduced by Kvam and Samaniego (1994). They propose to estimate $F(x)$ by maximizing the conditional log-likelihood function

$$
\begin{aligned}
L_n(x, p) &:= \sum_{i=1}^{n} \left[1_{[X_i \leq x]} \log B_{R_i}(p) + 1_{[X_i > x]} \log(1 - B_{R_i}(p)) \right] \\
&= \sum_{r=1}^{k} N_{nr} \left[\widehat{F}_{nr}(x) \log B_r(p) + (1 - \widehat{F}_{nr}(x)) \log(1 - B_r(p)) \right]
\end{aligned}
$$

of the indicator vector $(1_{[X_i \leq x]})_{i=1}^{n}$, given the rank vector $\boldsymbol{R}_n = (R_i)_{i=1}^{n}$. The resulting estimator $\widehat{F}_n^{\mathrm{L}}$ is given by

$$\widehat{F}_n^{\mathrm{L}}(x) := \underset{p \in [0,1]}{\arg\max} \, L_n(x, p).$$

Huang (1997) provides a detailed asymptotic analysis of this estimator $\widehat{F}_n^{\mathrm{L}}$ in the special setting when $n = k\ell$, $N_{nr} = \ell$ for $1 \leq r \leq k$, and $\ell \to \infty$.

A third approach, introduced by Chen (2001), is to estimate F by a moment equality for the naive empirical c.d.f. $\widehat{F}_n$. Note that

$$\mathbb{E}\big(n\widehat{F}_n(x) \,\big|\, \boldsymbol{R}_n\big) \;=\; \sum_{r=1}^{k} N_{nr} B_r(F(x)).$$

Hence, one can estimate $F(x)$ by the unique number $\widehat{F}_n^{\mathrm{M}}(x) \in [0, 1]$ such that

$$n\widehat{F}_n(x) \;=\; \sum_{r=1}^{k} N_{nr} B_r\big(\widehat{F}_n^{\mathrm{M}}(x)\big). \tag{1}$$

In the RSS setting with proportions N_{nr}/n converging to fixed numbers $\pi_r > 0$ as $n \to \infty$, Chen (2001) proves asymptotic normality of $\sqrt{n}\big(\widehat{F}_n^{\mathrm{M}}(x) - F(x)\big)$ for finitely many points x and shows that the supremum norm of $\widehat{F}_n^{\mathrm{M}} - F$ converges to zero in probability. (Note that Chen (2001) formulates the moment equality (1) with $n\pi_r$ in place of N_{nr}, but this would introduce an unnecessary estimation bias.)

In Sect. 2, we present some elementary properties of the estimators $\widehat{F}_n^{\mathrm{S}}$, $\widehat{F}_n^{\mathrm{L}}$ and $\widehat{F}_n^{\mathrm{M}}$ and comment briefly on the computation of the latter two. In addition, we describe two methods to obtain pointwise and simultaneous confidence intervals for F, respectively. The former procedure is just an adaptation of a method by Terpstra and Miller (2006) and closely related to the estimator $\widehat{F}_n^{\mathrm{M}}$. Inverting the underlying tests yields honest confidence intervals for any given quantile of F as proposed by Balakrishnan and Li (2006) for balanced RSS. The confidence bands are a generalization of the confidence bands described by Stokes and Sager (1988). Here it turns out that the estimator $\widehat{F}_n^{\mathrm{M}}$ is particularly convenient to work with.

Section 3 provides a detailed analysis of the asymptotic distribution of the estimators $\widehat{F}_n^{\mathrm{S}}$, $\widehat{F}_n^{\mathrm{L}}$ and $\widehat{F}_n^{\mathrm{M}}$ as $n \to \infty$ while k is fixed and $N_{nr}/n \to_p \pi_r > 0$ for $1 \le r \le k$. Our analyses provide linear stochastic expansions and functional central limit theorems for the processes $\sqrt{n}(\widehat{F}_n^{\mathrm{Z}} - F)$, $\mathrm{Z} = \mathrm{S}, \mathrm{L}, \mathrm{M}$. These results generalize the findings of Stokes and Sager (1988) about $\widehat{F}_n^{\mathrm{S}}$, of Huang (1997) about $\widehat{F}_n^{\mathrm{L}}$ in balanced RSS and of Chen (2001) and Ghosh and Tiwari (2008) about $\widehat{F}_n^{\mathrm{M}}$. We obtain explicit expressions for the asymptotic covariance functions of $\sqrt{n}(\widehat{F}_n^{\mathrm{Z}} - F)$ which enable efficiency considerations. The most important findings are that (i) the estimator $\widehat{F}_n^{\mathrm{L}}$ is always superior to the other two, (ii) the estimators $\widehat{F}_n^{\mathrm{S}}$ and $\widehat{F}_n^{\mathrm{M}}$ are asymptotically equivalent in case of $\pi_1 = \cdots = \pi_k = 1/k$ and (iii) in unbalanced settings the estimator $\widehat{F}_n^{\mathrm{S}}$ can be substantially worse than the other two estimators. Moreover, the efficiency gain of $\widehat{F}_n^{\mathrm{L}}$ over $\widehat{F}_n^{\mathrm{M}}$ is bounded and typically rather small. In addition, we analyze the estimators' asymptotic behavior in the tails of the distribution F where they turn out to be essentially equivalent.

A detailed analysis of a real data example is presented in Sect. 4. It involves population sizes of Swiss municipalities and illustrates that sampling from finite populations without replacement may render our confidence regions conservative, even if the rankings are not perfect. The impact of imperfect rankings itself is investigated in a small simulation study based on the model of Dell and Clutter (1972).

The main proofs are deferred to an appendix. Further technical details and additional material, including references to computer code in R (R Core Team 2013), are collected in a supplement.

2 Computation of the estimators and exact inference

Computations In what follows let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics of $X_1, X_2, \ldots, X_n$, augmented by $X_{(0)} := -\infty$ and $X_{(n+1)} := \infty$. One can easily verify that for $Z = S, M, L$, the estimator $\widehat{F}_n^Z$ is constant on each interval $[X_{(y)}, X_{(y+1)})$, $0 \le y \le n$, where $\widehat{F}_n^Z \equiv 0$ on $[X_{(0)}, X_{(1)})$ and $\widehat{F}_n^Z \equiv 1$ on $[X_{(n)}, X_{(n+1)})$.

While the computation of the stratified estimator $\widehat{F}_n^S$ is straightforward, the estimators $\widehat{F}_n^M$ and $\widehat{F}_n^L$ may be computed numerically by running a suitable bisection algorithm $n - 1$ times. Concerning $\widehat{F}_n^M$, note that $\sum_{r=1}^k N_{nr} B_r(p)$ is continuous and strictly increasing in $p \in [0, 1]$ with boundary values 0 and 1. Hence for $1 \le y < n$ and $X_{(y)} \le x < X_{(y+1)}$, the estimator $\widehat{F}_n^M(x)$ is the unique solution $p \in (0, 1)$ of $\sum_{r=1}^k N_{nr} B_r(p) = y$.

As to $\widehat{F}_n^L$, the next lemma provides some essential properties of the log-likelihood function $L_n(\cdot, \cdot)$. Its proof is given in the supplement.

Lemma 1 *For any $x \in \mathbb{R}$, the function $L_n(x, \cdot) : [0, 1] \to [-\infty, 0]$ is continuous and continuously differentiable on $(0, 1)$. Its derivative $L_n'(x, p) := \partial L_n(x, p)/\partial p$ is strictly decreasing in $p \in (0, 1)$ and equals*

$$L_n'(x, p) = \sum_{r=1}^k N_{nr} w_r(p) [\widehat{F}_{nr}(x) - B_r(p)]$$

with the auxiliary function

$$w_r(p) = \frac{\beta_r}{B_r(1 - B_r)}(p) = \frac{\beta_r(p)}{B_r(p) B_{k+1-r}(1 - p)}.$$

Moreover, in case of $X_{(1)} \le x < X_{(n)}$, the limits of $L_n'(x, \cdot)$ at the boundary of $(0, 1)$ are equal to $L_n'(x, 0) = \infty$ and $L_n'(x, 1) = -\infty$.

According to this lemma, for $y \in \{1, \ldots, n - 1\}$ and $X_{(y)} \le x < X_{(y+1)}$, the value of $\widehat{F}_n^L(x)$ is the unique number $p \in (0, 1)$ such that

$$\sum_{r=1}^k N_{nr} w_r(p) [\widehat{F}_{nr}(X_{(y)}) - B_r(p)] = 0.$$

The computation of $\widehat{F}_n^M$ and $\widehat{F}_n^L$ for one single data set is of similar complexity. There is, however, an important difference: The vector $(\widehat{F}_n^M(X_{(y)}))_{y=1}^{n-1}$ depends solely on the vector $N_n = (N_{nr})_{r=1}^k$ of stratum sizes. Hence, if we want to simulate the conditional distribution of $\widehat{F}_n^M$, given R_n, we have to compute the vector $(\widehat{F}_n^M(X_{(y)}))_{y=1}^{n-1}$ only

once. By way of contrast, the vector $\left(\widehat{F}_n^{\mathrm{L}}(X_{(y)})\right)_{y=1}^{n-1}$ depends on the whole matrix $(N_{nry})_{1 \le r \le k, 1 \le y \le n}$ of frequencies $N_{nry} = N_{nr}\widehat{F}_{nr}(X_{(y)}) = \sum_{i=1}^{n} 1_{[R_i = r, \, X_i \le X_{(y)}]}$. For given N_n, there are

$$\frac{n!}{N_{n1}! \, N_{n2}! \, \cdots \, N_{nk}!}$$

possibilities for that matrix, and this number grows exponentially with n, unless N_n is extremely unbalanced. As a consequence, for each new data set we have to compute $\widehat{F}_n^{\mathrm{L}}$ anew, even if N_n remains unchanged.

Basic distributional properties From now on, we condition on the rank vector $\boldsymbol{R}_n = (R_i)_{i=1}^{n}$. Hence the vector $N_n = (N_{nr})_{r=1}^{k}$ of stratum sizes is viewed as a fixed vector, and all probabilities, expectations and distributional statements refer to the conditional distribution of $X_n = (X_i)_{i=1}^{n}$, given $\boldsymbol{R}_n$.

All estimators $\widehat{F}_n$, $\widehat{F}_n^{\mathrm{S}}$, $\widetilde{\widehat{F}}_n^{\mathrm{M}}$ and $\widehat{F}_n^{\mathrm{L}}$ are distribution-free in the following sense: Let $\widehat{B}_n$, $\widehat{B}_n^{\mathrm{S}}$, $\widehat{B}_n^{\mathrm{M}}$ and $\widehat{B}_n^{\mathrm{L}}$ be defined analogously with raw observations from the uniform distribution on $[0, 1]$. That means, we replace the random variables $X_1, X_2, \ldots, X_n$ with random variables $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n \in [0, 1]$ which are independent, and $\widetilde{X}_i$ has (conditional) distribution function B_r if $R_i = r$. Then

$$\left(\widehat{F}_n^{\mathrm{Z}}(x)\right)_{x \in \mathbb{R}} \quad \text{has the same distribution as} \quad \left(\widehat{B}_n^{\mathrm{Z}}(F(x))\right)_{x \in \mathbb{R}},$$

where Z = S, M, L. Consequently, it suffices to analyze the distribution of the random processes $\left(\widehat{B}_n^{\mathrm{Z}}(t)\right)_{t \in [0, 1]}$.

Pointwise confidence intervals Recall that the estimator $\widehat{F}_n^{\mathrm{M}}(x)$ was defined by matching $n\widehat{F}_n(x)$ to its (conditional) mean. Comparing $n\widehat{F}_n(x)$ with its distribution function yields exact confidence bounds for $F(x)$. This approach has been used by Terpstra and Miller (2006) in the framework of balanced ranked set sampling. In the present general framework, this method works as follows: The (conditional) distribution of $n\widehat{F}_n(x)$ depends only on N_n and $F(x)$. Precisely, in case of $F(x) = p$, it has the same distribution as $\sum_{r=1}^{k} Y_{r,p}$ with independent random variables $Y_{1,p}, Y_{2,p}, \ldots, Y_{k,p}$, where

$$Y_{r,p} \sim \mathrm{Bin}(N_{nr}, B_r(p)).$$

Let $G_{N_n, p}$ be the corresponding distribution function, i.e.,

$$G_{N_n, p}(y) := \mathbb{P}\left(\sum_{r=1}^{k} Y_{r,p} \le y\right).$$

This is not a standard distribution but a convolution of binomial distributions which can be computed numerically quite easily. Elementary considerations reveal that for any $y \in \{0, 1, \ldots, n-1\}$, the distribution function $G_{N_n, p}(y)$ is continuous and strictly decreasing in $p \in [0, 1]$ with boundary values $G_{N_n, 0}(y) = 1$ and $G_{N_n, 1}(y) = 0$.

Further, $G_{N_n, p}(n) = 1$ and $G_{N_n, p}(-1) = 0$ for all $p \in [0, 1]$. Consequently, non-asymptotic p values for the null hypotheses "$F(x) \geq p$" and "$F(x) \leq p$" are given by $G_{N_n, p}(n\widehat{F}_n(x))$ and $1 - G_{N_n, p}(n\widehat{F}_n(x) - 1)$, respectively. These imply two different $(1 - \alpha)$-confidence regions for $F(x)$, namely

$$\left\{ p \in [0, 1] : G_{N_n, p}(n\widehat{F}_n(x)) \geq \alpha \right\} = \left[0, b_\alpha(N_n, n\widehat{F}_n(x)) \right],$$
$$\left\{ p \in [0, 1] : G_{N_n, p}(n\widehat{F}_n(x) - 1) \leq 1 - \alpha \right\} = \left[a_\alpha(N_n, n\widehat{F}_n(x)), 1 \right].$$

Here $b_\alpha(N_n, y)$ is the unique solution $p \in (0, 1)$ of the equation $G_{N_n, p}(y) = \alpha$ if $0 \leq y \leq n - 1$, and $b_\alpha(N_n, n) = 1$. Likewise, $a_\alpha(N_n, y)$ is the unique solution $p \in (0, 1)$ of the equation $G_{N_n, p}(y - 1) = 1 - \alpha$ if $1 \leq y \leq n$, and $a_\alpha(N_n, 0) = 0$. Obviously, one can combine lower and upper bounds and compute the Clopper and Pearson (1934) type $(1 - \alpha)$-confidence interval $\left[a_{\alpha/2}(N_n, n\widehat{F}_n(x)), b_{\alpha/2}(N_n, n\widehat{F}_n(x)) \right]$ for $F(x)$.

Note that the computation of all these confidence bounds for F boils down to determining only finitely many values $a_\lambda(N_n, y)$ and $b_\lambda(N_n, y)$ for $\lambda = \alpha, \alpha/2$ and $y \in \{0, 1, \ldots, n\}$.

If we would ignore the ranks R_i and just pretend that $X_1, X_2, \ldots, X_n$ are i.i.d. with distribution function F, then we would work with the distribution function $G_{n, p}$ of the binomial distribution $\text{Bin}(n, p)$ instead of $G_{N_n, p}$. This would lead to the traditional confidence bounds $a_\alpha^{\text{st}}(n, n\widehat{F}_n(x)), b_\alpha^{\text{st}}(n, n\widehat{F}_n(x))$ and the confidence interval of Clopper and Pearson (1934) with endpoints $a_{\alpha/2}^{\text{st}}(n, n\widehat{F}_n(x)), b_{\alpha/2}^{\text{st}}(n, n\widehat{F}_n(x))$ for $F(x)$.

Confidence bands We may compute Kolmogorov–Smirnov-type confidence bands for the unknown distribution function F as follows: Let $\kappa^Z(N_n, \alpha)$ be the $(1 - \alpha)$-quantile of the random variable $\|\widehat{B}_n^Z - B\|_\infty = \sup_{t \in [0, 1]} |\widehat{B}_n^Z(t) - t|$. Then, we may conclude with confidence $1 - \alpha$ that

$$F(x) \in \left[\widehat{F}_n^Z(x) \pm \kappa^Z(N_n, \alpha) \right] \quad \text{for all } x \in \mathbb{R}.$$

The quantiles $\kappa^Z(N_n, \alpha)$ may be estimated via Monte Carlo simulations. As explained before, this procedure is particularly convenient to implement for the moment-matching estimator $\widehat{F}_n^M$, whereas for the likelihood estimator $\widehat{F}_n^L$ it would be very computer-intensive.

Numerical example Figure 1 shows for $n = 210$ and $N_n = (70, 70, 70), (100, 70, 40)$ the estimator value $\widehat{F}_n^M(X_{(y)})$ and the two-sided 95%-confidence bounds $a_{2.5\%}(N_n, y)$, $b_{2.5\%}(N_n, y), a_{2.5\%}^{\text{st}}(n, y)$ and $b_{2.5\%}^{\text{st}}(n, y)$ as a function of $y \in \{0, 1, \ldots, n\}$. One sees that the additional rank information leads to more accurate confidence bounds in the balanced setting. In the unbalanced situation, ignoring the rank information and pretending the X_i to be i.i.d. would induce a severe bias, and the coverage probabilities would be substantially smaller than 95%.

For Kolmogorov–Smirnov-type confidence bands centered at $\widehat{F}_n^M$, we estimated the quantiles $\kappa^M(N_n, 5\%)$ in 10^5 Monte Carlo simulations and obtained

$$\widehat{\kappa}^M(N_n, 5\%) = \begin{cases} 0.0790 & \text{for } N_n = (70, 70, 70), \\ 0.0812 & \text{for } N_n = (100, 70, 40). \end{cases}$$
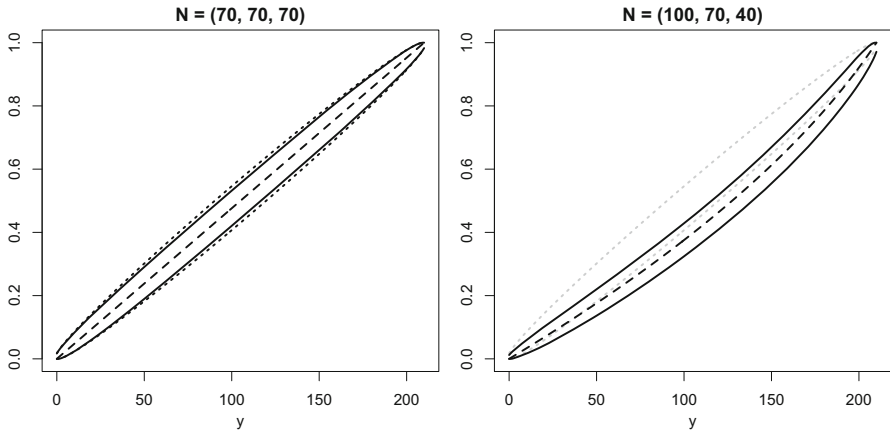
Fig. 1 Estimator $\widehat{F}_n^M$ and pointwise 95%-confidence intervals for F: For $y \in \{0, 1, \ldots, n\}$ one sees the value $\widehat{F}_n^M(X_{(y)})$ (dashed), the exact confidence bounds $a_{2.5\%}(N_n, y)$ and $b_{2.5\%}(N_n, y)$ (solid), and the classical bounds $a_{2.5\%}^{st}(n, y)$ and $b_{2.5\%}^{st}(n, y)$ (dotted)

For the usual Kolmogorov–Smirnov confidence band with $n = 210$ observations, the critical value would be $\kappa(n, 5\%) = 0.0927$.

Unequal group sizes The point estimators $\widehat{F}_n^L$, $\widehat{F}_n^M$ and the confidence regions just described may be extended easily to a more general setting with independent observations (X_i, R_i, k_i), $1 \le i \le n$, where $k_i \ge 1$ is a fixed integer, R_i is a fixed or random rank in $\{1, 2, \ldots, k_i\}$, and

$$\mathbb{P}(X_i \le x \mid R_i = r) = B_{r, k_i + 1 - r}(F(x)),$$

see, for instance, Bhoj (2001) or Chen (2001). Here $B_{r,s}$ denotes the distribution function of the beta distribution with parameters r and s.

3 Asymptotic considerations

We consider the asymptotic behavior of the estimators $\widehat{B}_n^S$, $\widehat{B}_n^M$ and $\widehat{B}_n^L$ for fixed k as $n \to \infty$ and

$$\pi_{nr} := \frac{N_{nr}}{n} \to \pi_r \quad \text{for } 1 \le r \le k.$$

Recall that we condition on the rank vector $\boldsymbol{R}_n$. The former condition is satisfied with $\pi_r = 1/k$ both in Huang's (1997) setting and in the JPS setting almost surely. In general, we assume that

$$\begin{cases} \pi_1, \ldots, \pi_k > 0 & \text{in connection with } \widehat{B}_n^S, \\ \pi_1, \pi_k > 0 & \text{in connection with } \widehat{B}_n^M, \widehat{B}_n^L. \end{cases}$$

Linear expansions and limit theorems In what follows, let

$$\mathbb{V}_{nr} := \sqrt{N_{nr}} \left(\widehat{B}_{nr} - B_r \right) \circ B_r^{-1}$$

for $1 \le r \le k$. Each stochastic process $\mathbb{V}_{nr}$ has the same distribution as a standardized empirical distribution function of N_{nr} independent random variables with uniform distribution on $[0, 1]$, see also "Appendix." Moreover, the processes $\mathbb{V}_{n1}, \ldots, \mathbb{V}_{nk}$ are stochastically independent. Our first result shows that the three estimators $\widehat{B}_n^{\mathrm{S}}$, $\widehat{B}_n^{\mathrm{M}}$ and $\widehat{B}_n^{\mathrm{L}}$ may be approximated by simpler processes involving $\mathbb{V}_{n1}, \ldots, \mathbb{V}_{nk}$.

Theorem 2 (Linear expansion). *For* $Z = \mathrm{S, M, L}$ *and any fixed* $\delta \in [0, 1/2)$,

$$\sup_{t \in (0,1)} \frac{\left| \sqrt{n} (\widehat{B}_n^{\mathrm{Z}}(t) - t) - \mathbb{V}_n^{\mathrm{Z}}(t) \right|}{t^{\delta} (1 - t)^{\delta}} \to_p 0,$$

where

$$\mathbb{V}_n^{\mathrm{Z}}(t) := \sum_{r=1}^{k} \gamma_{nr}^{\mathrm{Z}}(t) \, \mathbb{V}_{nr}(B_r(t))$$

with continuous functions $\gamma_{n1}^{\mathrm{Z}}, \ldots, \gamma_{nk}^{\mathrm{Z}} : [0, 1] \to [0, \infty)$. *Precisely, for* $t \in (0, 1)$,

$$\gamma_{nr}^{\mathrm{S}}(t) := \frac{1}{k \sqrt{\pi_{nr}}},$$

$$\gamma_{nr}^{\mathrm{M}}(t) := \sqrt{\pi_{nr}} \Big/ \sum_{s=1}^{k} \pi_{ns} \beta_s(t),$$

$$\gamma_{nr}^{\mathrm{L}}(t) := \sqrt{\pi_{nr}} \, w_r(t) \Big/ \sum_{s=1}^{k} \pi_{ns} w_s(t) \beta_s(t)$$

with $w_r = \beta_r / (B_r (1 - B_r))$. *Moreover,*

$$\sup_{t \in (0,c] \cup [1-c,1)} \frac{|\mathbb{V}_n^{\mathrm{Z}}(t)|}{t^{\delta} (1 - t)^{\delta}} \to_p 0 \quad \text{as } n \to \infty \text{ and } c \downarrow 0.$$

The next theorem shows that all estimators $\widehat{F}_n^{\mathrm{S}}$, $\widehat{F}_n^{\mathrm{M}}$, $\widehat{F}_n^{\mathrm{L}}$ are asymptotically equivalent in the tail regions. Moreover, the asymptotic behavior in the left and right tail is driven mainly by the processes $\mathbb{V}_{n1}$ and $\mathbb{V}_{nk}$, respectively.

Theorem 3 (Linear expansion in the tails). *For* $Z = \mathrm{S, M, L}$ *and any fixed* $\kappa \in [1/2, 1)$,

$$\sup_{t \in (0,c]} \frac{\left| \sqrt{n} (\widehat{B}_n^{\mathrm{Z}}(t) - t) - \mathbb{V}_n^{(\ell)}(t) \right|}{t^{\kappa}} \to_p 0$$

and

$$\sup_{t\in[1-c,1)} \frac{\left|\sqrt{n}(\widehat{B}_n^Z(t) - t) - \mathbb{V}_n^{(r)}(t)\right|}{(1-t)^\kappa} \to_p 0$$

as $n \to \infty$ *and* $c \downarrow 0$, *where*

$$\mathbb{V}_n^{(\ell)}(t) := \frac{\mathbb{V}_{n1}(B_1(t))}{k\sqrt{\pi_{n1}}} \quad and \quad \mathbb{V}_n^{(r)}(t) := \frac{\mathbb{V}_{nk}(B_k(t))}{k\sqrt{N_{nk}/n}}.$$

It follows from Donsker's theorem for the empirical process that $\mathbb{V}_{nr}$ behaves asymptotically like a standard Brownian bridge process $\mathbb{V} = (\mathbb{V}(u))_{u\in[0,1]}$. Together with Theorem 2, this leads to the following limit theorem:

Corollary 4 (Asymptotic distribution). *For* $Z = S, M, L$, *the stochastic process* $\mathbb{V}_n^Z$ *converges in distribution in the space* $\ell_\infty([0, 1])$ *to a centered Gaussian process* $\mathbb{V}^Z$ *with continuous paths on* [0, 1]. *Precisely, for* $t \in [0, 1]$,

$$\mathbb{V}^Z(t) = \sum_{r=1}^k \gamma_r^Z(t)\mathbb{V}_r(B_r(t))$$

with independent standard Brownian bridges $\mathbb{V}_1, \ldots, \mathbb{V}_k$ *and continuous functions* $\gamma_1^Z, \ldots, \gamma_k^Z : [0, 1] \to [0, \infty)$ *given by*

$$\gamma_r^S(t) := \frac{1}{k\sqrt{\pi_r}},$$

$$\gamma_r^M(t) := \sqrt{\pi_r} \Big/ \sum_{s=1}^k \pi_s \beta_s(t),$$

$$\gamma_r^L(t) := \begin{cases} \sqrt{\pi_r}\, w_r(t) \Big/ \sum_{s=1}^k \pi_s w_s(t)\beta_s(t) & \text{for } 0 < t < 1, \\ \sqrt{\pi_r}\, r/(\pi_1 k) & \text{for } t = 0, \\ \sqrt{\pi_r}(k + 1 - r)/(\pi_k k) & \text{for } t = 1. \end{cases}$$

Theorem 2 and Corollary 4 show that all three estimators $\widehat{F}_n^S$, $\widehat{F}_n^M$, $\widehat{F}_n^L$ are root-n-consistent. In the asymptotically balanced case with

$$\pi_1 = \pi_2 = \cdots = \pi_k = 1/k, \tag{2}$$

one can easily deduce from $\sum_{s=1}^k \beta_s \equiv k$ that

$$\gamma_r^M \equiv \gamma_r^S = 1/\sqrt{k} \quad \text{for } 1 \le r \le k.$$

Hence, in this particular case the estimators $\widehat{F}_n^S$ and $\widehat{F}_n^M$ are asymptotically equivalent. But otherwise $\widehat{F}_n^S$ may be substantially worse than $\widehat{F}_n^M$, as shown later.

Relative asymptotic efficiencies Let K be the covariance function of a standard Brownian bridge $\mathbb{V}$, i.e., $K(s,t) = \min\{s,t\} - st$ for $s, t \in [0, 1]$. Then the covariance function K^Z of the Gaussian process $\mathbb{V}^Z$ in Corollary 4 is given by

$$K^Z(s,t) = \sum_{r=1}^{k} \gamma_r^Z(s)\gamma_r^Z(t) K\big(B_r(s), B_r(t)\big).$$

In particular, for $0 < t < 1$ the asymptotic distribution of $\sqrt{n}\big(\widehat{B}_n^Z(t) - t\big)$ equals $\mathcal{N}\big(0, K^Z(t)\big)$ with $K^Z(t) := K^Z(t,t)$ given by

$$K^S(t) = \sum_{r=1}^{k} \frac{B_r(t)(1 - B_r(t))}{k^2 \pi_r},$$

$$K^M(t) = \sum_{r=1}^{k} \pi_r B_r(t)(1 - B_r(t)) \Big/ \left(\sum_{s=1}^{k} \pi_s \beta_s(t)\right)^2,$$

$$K^L(t) = \sum_{r=1}^{k} \pi_r w_r(t)^2 B_r(t)(1 - B_r(t)) \Big/ \left(\sum_{s=1}^{k} \pi_s \beta_s(t) w_s(t)\right)^2$$

$$= 1 \Big/ \sum_{s=1}^{k} \pi_s \beta_s(t) w_s(t).$$

The latter equation follows from $w_r = \beta_r/(B_r(1 - B_r))$. The next result provides a detailed comparison of these asymptotic variances.

Theorem 5 (Relative asymptotic efficiencies). *For arbitrary* $t \in (0, 1)$,

$$K^L(t) \leq K^S(t)$$

with equality for at most one $t \in (0, 1)$. *Furthermore,*

$$K^L(t) \leq K^M(t)$$

with equality if, and only if, $t = 1/2$ *and* $k = 2$. *On the other hand,*

$$\sup_{\pi} \frac{K^S(t)}{K^L(t)} = \infty,$$

$$\sup_{\pi} \frac{K^M(t)}{K^L(t)} = \frac{\rho(t) + \rho(t)^{-1} + 2}{4} \leq \frac{k + k^{-1} + 2}{4},$$
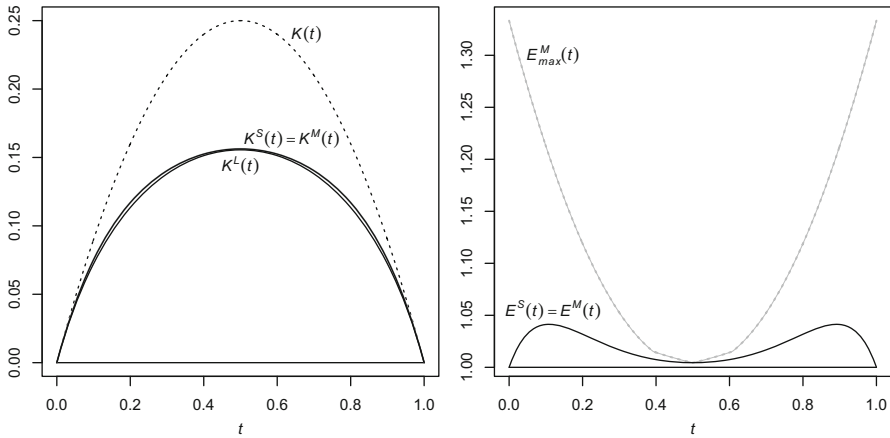
Fig. 2 Asymptotic variances of $\widehat{B}_n^L$, $\widehat{B}_n^S \equiv \widehat{B}_n^M$, $\widehat{B}_n$ (left panel) and relative efficiencies of $\widehat{B}_n^L$ versus $\widehat{B}_n^Z$ (right panel) in case of $\pi_1 = \pi_2 = \pi_3 = 1/3$

where the suprema are over all tuples $(\pi_r)_{r=1}^k$ with strictly positive components summing to one, and

$$\rho(t) := \max_{r=1,\ldots,k} w_r(t) \Big/ \min_{r=1,\ldots,k} w_r(t) \leq k.$$

Numerical examples In case of $k = 2$, the upper bound for $K^M(t)/K^L(t)$ equals $9/8 = 1.125$. More precisely,

$$\frac{\rho(t) + \rho(t)^{-1} + 2}{4} = 1 + \frac{u^2}{9 - u^2} \leq 1.125$$

with $u := 2t - 1 \in [-1, 1]$, see the supplement for more details.

In case of $k = 3$, the upper bound for $K^M(t)/K^L(t)$ equals $4/3 \approx 1.333$. Figures 2 and 3 show the asymptotic variance functions $K(\cdot)$ of $\widehat{B}_n$ and $K^Z(\cdot)$ of $\widehat{B}_n^Z$ for Z = S, M, L in the balanced and one unbalanced situation. Note that in the balanced setting, $\widehat{B}_n^S \equiv \widehat{B}_n^M$ and thus $K^S(\cdot) \equiv K^M(\cdot)$. In addition, one sees the asymptotic relative efficiencies

$$E^Z(t) := \frac{K^Z(t)}{K^L(t)}$$

of $\widehat{B}_n^L$ versus $\widehat{B}_n^Z$ together with the upper bound

$$E_{\max}^M(t) := \big(\rho(t) + \rho(t)^{-1} + 2\big)/4$$

for $E^M(t)$. One sees clearly that the inefficiency of $\widehat{B}_n^M$ versus $\widehat{B}_n^L$ is moderate, whereas the inefficiency of $\widehat{B}_n^S$ may become substantial in unbalanced settings. Note also that
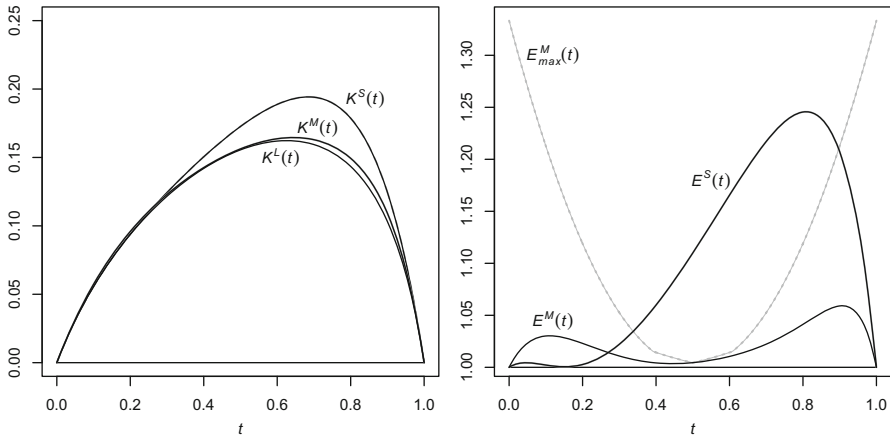
Fig. 3 Asymptotic variances of $\widehat{B}_n^S$, $\widehat{B}_n^M$, $\widehat{B}_n^L$ (left panel) and relative efficiencies of $\widehat{B}_n^L$ versus $\widehat{B}_n^Z$ (right panel) in case of $(\pi_1, \pi_2, \pi_3) = (10/21, 7/21, 4/21)$

in case of $\pi_1 > \pi_2 > \pi_3$ the accuracy in the left tail increases at the expense of larger errors in the right tail.

Implications for confidence intervals One can deduce from Corollary 4 that $n^{1/2}\kappa^Z(N_n, \alpha)$ converges to the $(1 - \alpha)$-quantile of the random supremum norm $\|\mathbb{V}^Z\|_\infty$. Moreover, for any $x \in \mathbb{R}$ with $0 < F(x) < 1$, the pointwise confidence bounds satisfy

$$a_\alpha(N_n, n\widehat{F}_n(x)) = \widehat{F}_n^M(x) - \frac{\sqrt{K^M(F(x))}}{\sqrt{n}}\, \Phi^{-1}(1 - \alpha) + o_p(n^{-1/2}) \qquad (3)$$

$$b_\alpha(N_n, n\widehat{F}_n(x)) = \widehat{F}_n^M(x) + \frac{\sqrt{K^M(F(x))}}{\sqrt{n}}\, \Phi^{-1}(1 - \alpha) + o_p(n^{-1/2}) \qquad (4)$$

with Φ^{-1} denoting the standard Gaussian quantile function, see the supplement.

4 A real data example and imperfect rankings

4.1 Population sizes of Swiss municipalities

Every 5 years, the Swiss Federal Office of Statistics releases data about all municipalities of Switzerland, including their population sizes. There are currently 2289 communities, and the two most recent data collections are from 2010 and 2015. Suppose we would have wanted to estimate the distribution function F of population sizes by the end of 2015 in early 2016. Back then only the data of 2010 would have been available, the data of 2015 having been released later in 2016 and corrected in 2017. In principle one could have approached each single municipality to obtain its population size by the end of 2015, but this would have been time-consuming of course. Hence, one could have applied RSS sampling as follows: One chooses randomly $n = 210$
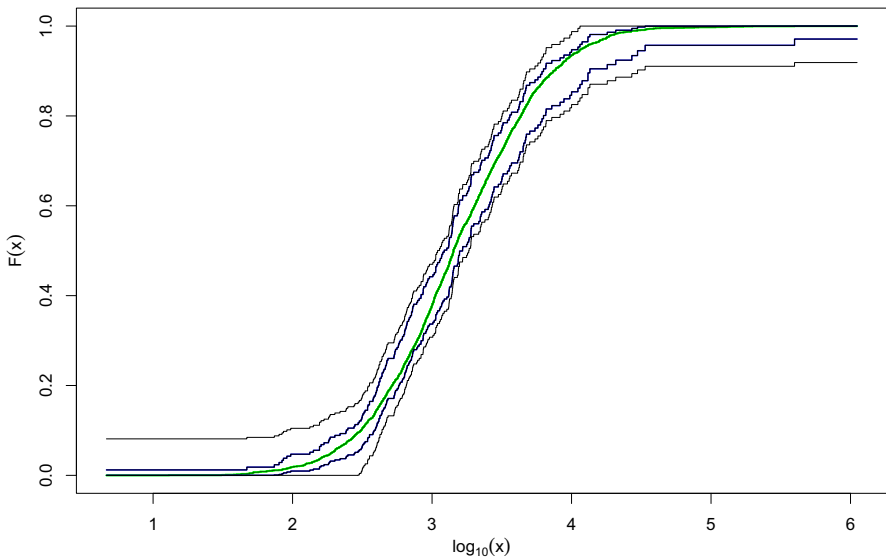
Fig. 4 Inference about the distribution of population sizes (Sect. 4.1) with $N_n = (100, 70, 40)$. The smoother function is the true c.d.f. F. The inner and outer two step functions are the pointwise and simultaneous 95%-confidence band for F

disjoint sets of $k = 3$ communities. Within the i-th set, one determines the unit with rank R_i according to population sizes in 2010 and obtains its precise population size X_i by the end of 2015. The ranks $R_1, \ldots, R_n \in \{1, 2, 3\}$ are prespecified. If one is particularly interested in smaller municipalities, one could choose $\boldsymbol{R}_n$ such that, say, $\boldsymbol{N}_n = (100, 70, 40)$.

Having the complete data of 2010 and 2015, one can easily simulate this sampling scheme. Figure 4 shows for one such sample the estimated distribution function $\widehat{F}_n^{\mathrm{M}}$ together with pointwise and simultaneous 95%-confidence intervals as described in Sect. 2. Since the distribution of population sizes is heavily right-skewed, the horizontal axis shows the decimal logarithms of population sizes. In the lower panel, the point estimator $\widehat{F}_n^{\mathrm{M}}$ is replaced with the true distribution function F, i.e., the empirical distribution function of all 2289 population sizes in 2015.

We simulated this sampling scheme 10^5 times and analyzed the performance of both $\widehat{F}_n^{\mathrm{M}}$ and the confidence intervals. The Monte Carlo estimator of

$$\mathrm{BIAS}(x) := \mathbb{E}\,\widehat{F}_n^{\mathrm{M}}(x) - F(x)$$

was everywhere between -10^{-4} and 10^{-3}, whereas the MC estimator of

$$\mathrm{RMSE}(x) := \left(\mathbb{E}\,(\widehat{F}_n^{\mathrm{M}}(x) - F(x))^2 \right)^{1/2}$$

was nowhere larger than 0.0263. The left panel of Fig. 5 depicts these two functions BIAS and RMSE. For each sample and any $x \in \mathbb{R}$, we obtained a pointwise and simultaneous 95%-confidence interval, denoted by $C_{\mathrm{pw}}(x)$ and $C_{\mathrm{sim}}(x)$, respectively. The
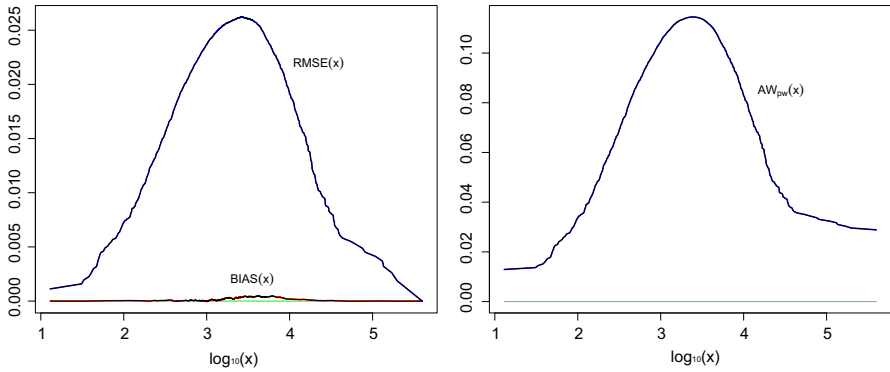
Fig. 5 Inference about the distribution of population sizes (Sect. 4.1) with $N_n = (100, 70, 40)$. Left panel: bias and root mean squared error of $\widehat{F}_n^M$. Right panel: Average width of pointwise 95%-confidence band for F

MC estimator of the error probability $\mathbb{P}\big(F(x) \notin C_{\mathrm{pw}}(x)\big)$ was nowhere larger than 4.22%, and the one of $\mathbb{P}\big(F(x) \notin C_{\mathrm{sim}}(x)$ for some $x \in \mathbb{R}\big)$ turned out to be smaller than 2.8%. The confidence intervals being conservative are probably a consequence of sampling without replacement, which results in more accurate estimators than sampling with replacement. The right panel of Fig. 5 shows MC estimates of the average widths

$$\mathrm{AW}_{\mathrm{pw}}(x) \; := \; \mathbb{E}\,\mathrm{width}(C_{\mathrm{pw}}(x)).$$

Here one sees clearly the effect of unbalanced sampling with $N_{n1} > N_{n2} > N_{n3}$, the benefit being shorter intervals in the left tail at the expense of longer intervals in the right tail.

Note that the ranking of municipalities within the n groups of size $k = 3$ was based on the population sizes in 2010 and thus imperfect. Indeed, a reasonable model for the pairs of log-transformed population sizes in 2010 and 2015 seems to be a bivariate Gaussian distribution with correlation 0.9986. As a consequence, in our MC simulations the average proportion of imperfect ranks R_i turned out to be 3.1%.

Analogous simulations for $k = 4, 5$ and different choices of N_n led to similar results. Enlarging k without changing n leads to larger coverage probabilities, presumably an effect of sampling without replacement, while the modulus of the bias of $\widehat{F}_n^M$ and the proportion of imperfect ranks get larger.

4.2 Imperfect rankings

In case of sampling *with* replacement, the previous data example would fit the model of Dell and Clutter (1972) for ranked set sampling with imperfect rankings quite well. They consider $2nk$ independent random variables $X_{ij} \sim F$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \tau^2)$ with $1 \le i \le n$ and $1 \le j \le k$. Instead of the true rank of X_{ij} among $X_{i1}, \ldots, X_{ik}$ one obtains the ranks
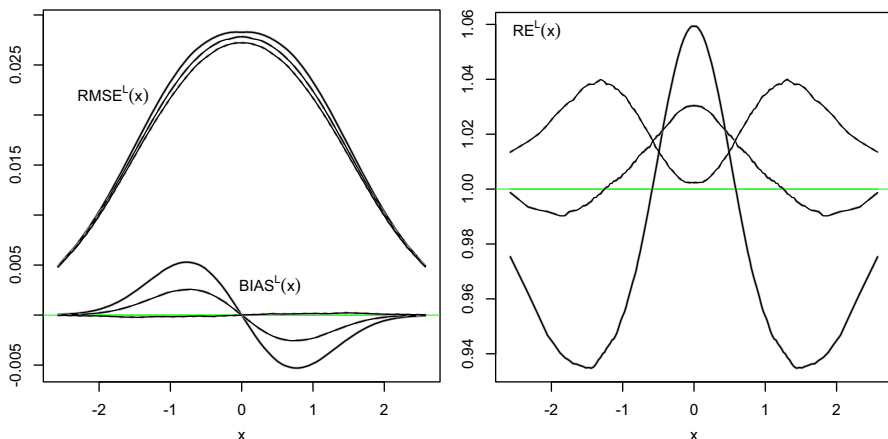
Fig. 6 Performance of $\widehat{F}_n^{\mathrm{L}}$ in balanced setting with $N_{n1} = N_{n2} = N_{n3} = 70$: Bias and root mean squared error (left panel) and relative efficiency versus $\widehat{F}_n^{\mathrm{S}}$ (right panel) for correlations $\rho = 1$ (dotted), $\rho = 0.95$ (dashed) and $\rho = 0.9$ (solid)

$$R_{ij} := \sum_{\ell=1}^{k} 1_{[Y_{i\ell} \leq Y_{ij}]}$$

of the concomitant variables $Y_{ij} := X_{ij} + \varepsilon_{ij}$. If $\sigma > 0$ denotes the standard deviation of the X_{ij}, the correlation between X_{ij} and Y_{ij} equals $\rho = (1 + \tau^2/\sigma^2)^{-1/2}$. Finally we obtain for $1 \leq i \leq n$ the observation $(X_i, R_i) = (X_{i1}, R_{i1})$ in JPS and $(X_{iJ(i)}, R_i)$ in RSS, where $J(i)$ is the unique index in $\{1, \ldots, k\}$ such that $R_{iJ(i)} = R_i$.

In this model, the stratified estimator $\widehat{F}_n^{\mathrm{S}}$ is still unbiased, see Presnell and Bohn (1999) for the RSS setting with $N_{n1}, \ldots, N_{nk} > 0$ and Dastbaravarde et al. (2016) for the JPS setting. For that reason, we considered $\widehat{F}_n^{\mathrm{S}}$ as a gold standard in our simulation study: We simulated 10^5 RSS data sets from this model with standard Gaussian distribution function $F = \Phi$, sample size $n = 210$ and different options for N_n and ρ. With these simulations, we estimated the bias and root mean squared error,

$$\mathrm{BIAS}^{\mathrm{Z}}(x) := \mathbb{E}\,\widehat{F}_n^{\mathrm{Z}}(x) - F(x) \quad \text{and} \quad \mathrm{RMSE}^{\mathrm{Z}}(x) := \left(\mathbb{E}\,(\widehat{F}_n^{\mathrm{Z}}(x) - F(x))^2\right)^{1/2},$$

for Z = S, M, L. In addition we estimated the relative efficiency

$$\mathrm{RE}^{\mathrm{Z}}(x) := \mathrm{RMSE}^{\mathrm{S}}(x)^2/\mathrm{RMSE}^{\mathrm{Z}}(x)^2$$

of $\widehat{F}_n^{\mathrm{Z}}$ versus the stratified estimator $\widehat{F}_n^{\mathrm{S}}$.

Firstly we considered $N_{n1} = N_{n2} = N_{n3} = 70$. Here $\widehat{F}_n^{\mathrm{S}} \equiv \widehat{F}_n^{\mathrm{M}} \equiv \widehat{F}_n$. In Fig. 6, one sees on the left-hand side the functions $\mathrm{BIAS}^{\mathrm{L}}$ and $\mathrm{RMSE}^{\mathrm{L}}$ for three different values of the correlation ρ. While $\widehat{F}_n^{\mathrm{S}} \equiv \widehat{F}_n^{\mathrm{M}}$ is unbiased, the bias of $\widehat{F}_n^{\mathrm{L}}$ gets worse as ρ decreases. For all three estimators $\widehat{F}_n^{\mathrm{Z}}$, the root mean squared error increases as ρ decreases. The right-hand side of Fig. 6 depicts the relative efficiency function
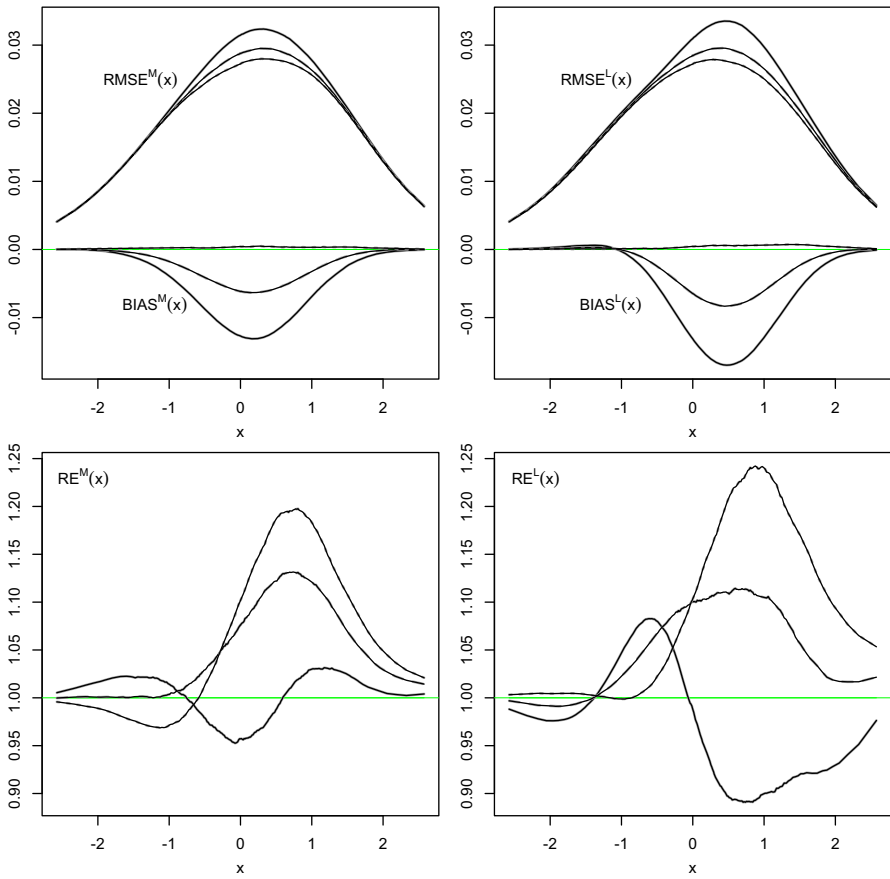
Fig. 7 Performance of $\widehat{F}_n^M$ and $\widehat{F}_n^L$ in unbalanced setting with $N_n = (100, 70, 40)$: Biases and root mean squared errors (upper panels) and relative efficiencies versus $\widehat{F}_n^S$ (lower panels) for correlations $\rho = 1$ (dotted), $\rho = 0.95$ (dashed) and $\rho = 0.9$ (solid)

RE^L. As predicted by asymptotic theory, $\mathrm{RE}^L > 1$ in case of $\rho = 1$, but for smaller correlations the relative efficiency drops below 1 in the tails.

Secondly we considered the unbalanced situation with $N_n = (100, 70, 40)$. Now the three estimators $\widehat{F}_n^Z$ are different, and only $\widehat{F}_n^S$ is unbiased. In Fig. 7, we show bias and root mean squared errors of $\widehat{F}_n^M$ and $\widehat{F}_n^L$. Clearly, the bias of $\widehat{F}_n^M$ and $\widehat{F}_n^L$ gets worse as ρ decreases, where $\widehat{F}_n^M$ is a bit more robust than $\widehat{F}_n^L$. Nevertheless, the plots of the relative efficiencies RE^M and RE^L show that for $\rho = 0.95$ the moment-matching estimator outperforms the stratified one everywhere, and also the likelihood estimator is better at most places. For $\rho = 0.9$, the likelihood estimator is less favorable than the other two.

Conclusions and future research

The present paper confirms and generalizes previous findings that the estimator $\widehat{F}_n^{\mathrm{L}}$ is the most efficient one in case of perfect ranking, both in balanced and unbalanced situations. In terms of computational efficiency, however, the estimator $\widehat{F}_n^{\mathrm{M}}$ has clear advantages and is particularly convenient as an ingredient for simultaneous confidence bands. Further, it is closely related to pointwise confidence bands for F. For now, we restricted ourselves to Kolmogorov–Smirnov-type bands, but other variants might be worthwhile to study.

The simulations in Sect. 4.2 indicate that even in case of imperfect rankings, both $\widehat{F}_n^{\mathrm{M}}$ and $\widehat{F}_n^{\mathrm{L}}$ perform well compared to $\widehat{F}_n^{\mathrm{S}}$, as long as the ranking precision is high. While $\widehat{F}_n^{\mathrm{L}}$ appears to be most sensitive to imperfect rankings, $\widehat{F}_n^{\mathrm{M}}$ seems to offer a good compromise in terms of efficiency (for perfect rankings) and robustness against ranking errors. Investigating and understanding these differences thoroughly would be an interesting topic for future research.

Appendix

We first recall two well-known facts about uniform empirical processes, see Shorack and Wellner (1986).

Proposition 6 *Let $U_1, U_2, U_3, \ldots$ be independent random variables with uniform distribution on $[0, 1]$. For $N \in \mathbb{N}$ and $u \in [0, 1]$ define*

$$\mathbb{V}^{(N)}(u) := N^{-1/2} \sum_{i=1}^{N} \big(1\{U_i \le u\} - u \big).$$

Then, as $N \to \infty$, $\mathbb{V}^{(N)}$ converges in distribution in $\ell_\infty([0, 1])$ to a standard Brownian bridge $\mathbb{V}$ on $[0, 1]$. Moreover, for any fixed $\delta \in [0, 1/2)$ and $\epsilon > 0$,

$$\sup_{N \ge 1} \mathbb{P} \left(\sup_{u \in (0,1)} \frac{|\mathbb{V}^{(N)}(u)|}{u^\delta (1 - u)^\delta} \ge C \right) \to 0 \quad as\ C \uparrow \infty,$$

$$\sup_{N \ge 1} \mathbb{P} \left(\sup_{u \in (0,c] \cup [1-c,1)} \frac{|\mathbb{V}^{(N)}(u)|}{u^\delta (1 - u)^\delta} \ge \epsilon \right) \to 0 \quad as\ c \downarrow 0.$$

For the estimators $\widehat{F}_n^{\mathrm{M}}, \widehat{F}_n^{\mathrm{L}}$ we need some basic facts and inequalities for the auxiliary functions w_k and B_k which are proved in the supplement:

Lemma 7 *(a) For $r = 1, 2, \ldots, k$, the function w_r on $(0, 1)$ may be written as $w_r(t) = \widetilde{w}_r(t)/(t(1-t))$ with $\widetilde{w}_r : [0, 1] \to (0, \infty)$ continuously differentiable. Moreover, for $r = 1, 2, \ldots, k$ and $t \in (0, 1)$,*

$$1 \leq \widetilde{w}_r(t) \leq \max(r, k+1-r).$$

(b) For any constant $c \in (0, 1)$ there exists a number $c' = c'(k, c) > 0$ with the following property: If $t, p \in (0, 1)$ such that

$$\frac{|p - t|}{t(1 - t)} \leq c,$$

then for $r = 1, 2, \ldots, k$,

$$\max \left\{ \left| \frac{w_r(p)}{w_r(t)} - 1 \right|, \left| \frac{B_r(p) - B_r(t)}{\beta_r(t)(p - t)} - 1 \right| \right\} \leq c' \frac{|p - t|}{t(1 - t)}.$$

Proof of Theorem 2 We start with the weight functions γ_{nr}^Z: Note that by Lemma 7,

$$\gamma_{nr}^{\mathrm{S}}(t) = \frac{1}{k\sqrt{\pi_{nr}}},$$

$$\gamma_{nr}^{\mathrm{M}}(t) = \sqrt{\pi_{nr}} \Big/ \sum_{s=1}^{k} \pi_{ns} \beta_s(t),$$

$$\gamma_{nr}^{\mathrm{L}}(t) = \sqrt{\pi_{nr}} \, \widetilde{w}_r(t) \Big/ \sum_{s=1}^{k} \pi_{ns} \widetilde{w}_s(t) \beta_s(t)$$

with the probability weights $\pi_{nr} := N_{nr}/n$ and continuous functions $\widetilde{w}_r : [0, 1] \to [1, k]$. Since the beta densities β_r are also continuous with $\beta_1(0) = \beta_k(1) = k$, this shows that γ_{nr}^Z is well-defined and continuous, provided that its denominator is strictly positive, i.e.,

$$\begin{cases} \pi_{n1}, \ldots, \pi_{nk} > 0 & \text{if } Z = S, \\ \pi_{n1}, \pi_{nk} > 0 & \text{if } Z = M, L. \end{cases}$$

For sufficiently large n this is the case, because $\lim_{n \to \infty} \pi_{nr} = \pi_r$ for all r. The functions γ_r^Z in Corollary 4 are continuous, too, and elementary considerations reveal that

$$\max_{t \in [0,1], \, 1 \leq r \leq k} \left| \gamma_{nr}^Z(t) - \gamma_r^Z(t) \right| \to 0 \qquad (5)$$

as $n \to \infty$. In particular, $\max_{t \in [0,1], 1 \leq r \leq k} \gamma_{nr}^Z(t) = O(1)$.

Note that for $n \geq 1$ and $1 \leq r \leq k$, the empirical process $\mathbb{V}_{nr}$ is distributed as $\mathbb{V}^{(N_{nr})}$ in Proposition 6. Note also that the distribution functions B_r satisfy $B_1 \geq B_2 \geq \cdots \geq B_k$, because for $1 \leq r < k$ the density ratio β_{r+1}/β_r is a positive multiple of $t/(1 - t)$ and thus strictly increasing. Consequently, for $1 \leq r \leq k$,

$$B_r(t) \leq B_1(t) \leq kt \quad \text{and} \quad 1 - B_r(t) \leq 1 - B_k(t) \leq k(1 - t),$$

so

$$\frac{B_r(t)(1 - B_r(t))}{t(1 - t)} \le k.$$

Consequently,

$$\sup_{t\in(0,1)} \frac{|\mathbb{V}_{nr}(B_r(t))|}{t^\delta(1 - t)^\delta} \le k^\delta \sup_{u\in(0,1)} \frac{|\mathbb{V}_{nr}(u)|}{u^\delta(1 - u)^\delta} = O_p(1) \quad \text{and}$$

$$\sup_{u\in(0,c]\cup[1-c,1)} \frac{|\mathbb{V}_{nr}(B_r(t))|}{t^\delta(1 - t)^\delta} \le k^\delta \sup_{u\in(0,kc]\cup[1-kc,1)} \frac{|\mathbb{V}_{nr}(u)|}{u^\delta(1 - u)^\delta} \to_p 0$$

as $n \to \infty$ and $c \downarrow 0$. All in all, we may conclude that

$$\sup_{t\in(0,1)} \frac{|\mathbb{V}_n^Z(t)|}{t^\delta(1 - t)^\delta} = O_p(1), \tag{6}$$

$$\sup_{t\in(0,c]\cup[1-c,1)} \frac{|\mathbb{V}_n^Z(t)|}{t^\delta(1 - t)^\delta} \to_p 0 \quad \text{as } n \to \infty \text{ and } c \downarrow 0. \tag{7}$$

It remains to be shown that the process $\sqrt{n}(\widehat{B}_n^Z - B)$ may be approximated by $\mathbb{V}_n^Z$. In case of $Z = S$ it follows from $\sum_{r=1}^k \beta_r \equiv k$ that $\sum_{r=1}^k B_r = kB$, and this implies that

$$\sqrt{n}(\widehat{B}_n^S - B) = \sum_{r=1}^k \frac{\sqrt{n}(\widehat{B}_{nr} - B_r)}{k} = \sum_{r=1}^k \gamma_{nr}^S \mathbb{V}_{nr} \circ B_r = \mathbb{V}_n^S.$$

For $Z = M, L$ it suffices to show that for any fixed number $b \ne 0$ and

$$p_n^Z(t) := t + \frac{\mathbb{V}_n^Z(t) + bt^\delta(1 - t)^\delta}{\sqrt{n}}$$

the following statements are true: If $b < 0$, then with asymptotic probability one,

$$\left.\begin{array}{c} \inf\limits_{t\in(0,1)} \left(n\widehat{B}_n(t) - \sum\limits_{r=1}^k N_{nr} B_r(p_n^M(t)) \right) \\ \inf\limits_{t\in(0,1)} L_n'(t, p_n^L(t)) \end{array}\right\} \ge 0. \tag{8}$$

If $b > 0$, then with asymptotic probability one,

$$\left.\begin{array}{c} \sup\limits_{t\in(0,1)} \left(n\widehat{B}_n(t) - \sum\limits_{r=1}^k N_{nr} B_r(p_n^M(t)) \right) \\ \sup\limits_{t\in(0,1)} L_n'(t, p_n^L(t)) \end{array}\right\} \le 0. \tag{9}$$

Here we use the conventions that $L_n'(t, \cdot) := \infty$ and $B_r := 0$ on $(-\infty, 0]$ while $L_n'(t, \cdot) := -\infty$ and $B_r := 1$ on $[1, \infty)$.

To verify these claims, we split the interval $(0, 1)$ into $(0, c_n]$, $[c_n, 1 - c_n]$ and $[1 - c_n, 1)$ with numbers $c_n \in (0, 1/2)$ to be specified later, where $c_n \downarrow 0$.

On $[c_n, 1 - c_n]$ we utilize Lemma 7: For $t \in [c_n, 1 - t_n]$ and $p \in (0, 1)$ such that $|p - t| \leq t(1 - t)/2$ we may write

$$
n\widehat{B}_n(t) - \sum_{r=1}^{k} N_{nr} B_r(p)
$$

$$
= \sum_{r=1}^{k} \sqrt{N_{nr}} \mathbb{V}_{nr}(B_r(t)) - \sum_{r=1}^{k} N_{nr}(B_r(p) - B_r(t))
$$

$$
= \sum_{r=1}^{k} \sqrt{N_{nr}} \mathbb{V}_{nr}(B_r(t)) - \sum_{r=1}^{m} N_{nr} \beta_r(t)(p - t) + \rho_n^{\mathrm{M}}(t, p)
$$

$$
= \sum_{r=1}^{k} N_{nr} \beta_r(t) \left(\frac{\mathbb{V}_n^{\mathrm{M}}(t)}{\sqrt{n}} - (p - t) \right) + \rho_n^{\mathrm{M}}(t, p)
$$

and

$$
L_n'(t, p) = \sum_{r=1}^{k} \sqrt{N_{nr}} w_r(p) \mathbb{V}_{nr}(B_r(t)) - \sum_{r=1}^{k} N_{nr} w_r(p)(B_r(p) - B_r(t))
$$

$$
= \sum_{r=1}^{k} \sqrt{N_{nr}} w_r(t) \mathbb{V}_{nr}(B_r(t)) - \sum_{r=1}^{k} N_{nr} w_r(t) \beta_r(t)(p - t) + \rho_n^{\mathrm{L}}(t, p)
$$

$$
= \sum_{r=1}^{k} N_{nr} w_r(t) \beta_r(t) \left(\frac{\mathbb{V}_n^{\mathrm{L}}(t)}{\sqrt{n}} - (p - t) \right) + \rho_n^{\mathrm{L}}(t, p),
$$

where

$$
|\rho_n^{\mathrm{M}}(t, p)| \leq \frac{O(n)|p - t|^2}{t(1 - t)},
$$

$$
|\rho_n^{\mathrm{L}}(t, p)| \leq \frac{O_p(\sqrt{n}) t^\delta (1 - t)^\delta |p - t|}{t(1 - t)} + \frac{O(n)|p - t|^2}{t^2(1 - t)^2}.
$$

Note that for $t \in [c_n, 1 - c_n]$,

$$
\frac{\left| p_n^{\mathrm{Z}}(t) - t \right|}{t(1 - t)} \leq \frac{O_p(1) t^\delta (1 - t)^\delta}{\sqrt{n} \, t(1 - t)} \leq \frac{O_p(1)}{\sqrt{n} \, c_n^{1-\delta}}.
$$

Hence we choose c_n such that $c_n \downarrow 0$ but $nc_n^{2(1-\delta)} \to \infty$. With this choice, we may conclude that uniformly in $t \in [c_n, 1 - c_n]$,

$$
\begin{aligned}
\left|\rho_n^{\mathrm{M}}(t, p_n^{\mathrm{M}}(t))\right| &\leq O_p(c_n^{\delta-1})t^\delta(1-t)^\delta, \\
\left|\rho_n^{\mathrm{L}}(t, p_n^{\mathrm{L}}(t))\right| &\leq O_p(c_n^{\delta-1})t^{\delta-1}(1-t)^{\delta-1}.
\end{aligned}
$$

On the other hand, since $\beta_1(t) + \beta_k(t) \geq \beta_1(1/2) + \beta_k(1/2) = k2^{2-k}$,

$$
\sum_{r=1}^{k} N_{nr}\beta_r(t) \geq k2^{2-k} \min\{N_{n1}, N_{nk}\},
$$

$$
\sum_{r=1}^{k} N_{nr}w_r(t)\beta_r(t) \geq \frac{k2^{2-k}c_w}{t(1-t)} \min\{N_{n1}, N_{nk}\}.
$$

Consequently,

$$
\begin{aligned}
n\widehat{B}_n(t) &- \sum_{r=1}^{k} N_{nr}B_r(p_n^{\mathrm{M}}(t)) \\
&= \sum_{r=1}^{k} N_{nr}\beta_r(t)\frac{-bt^\delta(1-t)^\delta}{\sqrt{n}} + \rho_n^{\mathrm{M}}(t, p_n^{\mathrm{M}}(t)) \\
&= \sum_{r=1}^{m} N_{nr}\beta_r(t)\frac{t^\delta(1-t)^\delta}{\sqrt{n}} \left(-b + O_p(c_n^{\delta-1}n^{-1/2})\kappa_n^{\mathrm{M}}(t)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
L_n'(t, p_n^{\mathrm{L}}(t)) &= \sum_{r=1}^{k} N_{nr}w_r(t)\beta_r(t)\frac{-bt^\delta(1-t)^\delta}{\sqrt{n}} + \rho_n^{\mathrm{L}}(t, p_n^{\mathrm{L}}(t)) \\
&= \sum_{r=1}^{k} N_{nr}w_r(t)\beta_r(t)\frac{t^\delta(1-t)^\delta}{\sqrt{n}} \left(-b + O_p\left(c_n^{\delta-1}n^{-1/2}\right)\kappa_n^{\mathrm{L}}(t)\right)
\end{aligned}
$$

for some random functions $\kappa_n^{\mathrm{M}}, \kappa_n^{\mathrm{L}} : [c_n, 1 - c_n] \to [-1, 1]$. These considerations show that (8) and (9) are satisfied with $[c_n, 1 - c_n]$ in place of $(0, 1)$.

It remains to verify (8) and (9) with $(0, c_n]$ in place of $(0, 1)$; the interval $[1 - c_n, 1)$ may be treated analogously. Note first that for $2 \leq r \leq k$,

$$
B_r(t) \leq B_2(t) \leq k(k-1)t^2/2 \quad \text{and} \quad \beta_r(t) \leq k2^{k-1}t,
$$

so

$$
\left|B_r(p) - B_r(t)\right| = \left|\int_t^p \beta_r(u)\,du\right| \leq O(\max(p, t))(p - t).
$$

Furthermore, since $B_1(t) = 1 - (1-t)^k$,

$$B_1(p) - B_1(t) = k(p-t) + O(\max(t,p))(p-t).$$

Hence for $t \in (0, c_n]$ and $p \in (0, 2c_n]$,

$$n\widehat{B}_n(t) - \sum_{r=1}^k N_{nr} B_r(p)$$

$$= \sum_{r=1}^k \sqrt{N_{nr}} \mathbb{V}_{nr}(B_r(t)) - \sum_{r=1}^k N_{nr}(B_r(p) - B_r(t))$$

$$= -N_{n1} k(p-t) + \rho_n^{\mathrm{M}}(t, p)$$

and

$$L'_n(t, p) = \sum_{r=1}^k \sqrt{N_{nr}} w_r(p) \mathbb{V}_{nr}(B_r(t)) - \sum_{r=1}^k N_{nr} w_r(p)(B_r(p) - B_r(t))$$

$$= -N_{n1} w_1(p) k(p-t) + \rho_n^{\mathrm{L}}(t, p),$$

where

$$|\rho_n^{\mathrm{M}}(t, p)| \le o_p(\sqrt{n}) t^\delta + O(n c_n)(p-t),$$

$$|\rho_n^{\mathrm{L}}(t, p)| \le o_p(\sqrt{n}) p^{-1} t^\delta + O(n c_n) p^{-1}(p-t).$$

Note also that

$$\sup_{t \in (0, c_n]} \left| \frac{\sqrt{n}(p_n^{\mathrm{Z}}(t) - t)}{t^\delta (1-t)^\delta} - b \right| \to_p 0.$$

In particular, $\sup_{t \in (0, c_n]} p_n^{\mathrm{Z}}(t) = c_n + o_p(n^{-1/2} c_n^\delta) = c_n(1 + o_p(1))$, and in case of $b > 0$, $\mathbb{P}\big(p_n^{\mathrm{Z}}(t) > 0 \text{ for } 0 < t \le c_n\big) \to 1$.

In case of $b > 0$, these considerations show that for $0 < t \le c_n$,

$$n\widehat{B}_n(t) - \sum_{r=1}^k N_{nr} B_r(p_n^{\mathrm{M}}(t))$$

$$= -N_{n1} k(p_n^{\mathrm{M}}(t) - t) + \rho_n^{\mathrm{M}}(t, p_n^{\mathrm{M}}(t))$$

$$\le \frac{N_{n1} k t^\delta (1-t)^\delta}{\sqrt{n}}(-b + o_p(1)) + o_p(\sqrt{n}) t^\delta + O(\sqrt{n} c_n) t^\delta$$

$$\le \frac{N_{n1} k t^\delta (1-t)^\delta}{\sqrt{n}}(-b + o_p(1))$$

and

$$
\begin{aligned}
L'_n(t, p_n^{\mathrm{L}}(t)) ={}& -N_{n1}w_1(p)k(p_n^{\mathrm{L}}(t) - t) + \rho_n^{\mathrm{L}}(t, p_n^{\mathrm{Z}}(t)) \\
\leq{}& \frac{N_{n1}w_1(p)kt^\delta(1-t)^\delta}{\sqrt{n}}(-b + o_p(1)) + o_p(\sqrt{n})p^{-1}t^\delta \\
& + O(\sqrt{n}c_n)p^{-1}t^\delta \\
\leq{}& \frac{N_{n1}w_1(p)kt^\delta(1-t)^\delta}{\sqrt{n}}(-b + o_p(1)).
\end{aligned}
$$

Analogously, in case of $b < 0$, for any $t \in (0, c_n]$ we obtain the inequalities

$$
n\widehat{B}_n(t) - \sum_{r=1}^{k} N_{nr} B_r(p_n^{\mathrm{M}}(t)) \geq
\begin{cases}
\dfrac{N_{n1}kt^\delta(1-t)^\delta}{\sqrt{n}}(-b + o_p(1)) & \text{if } p_n^{\mathrm{M}}(t) > 0, \\
0 & \text{if } p_n^{\mathrm{M}}(t) \leq 0,
\end{cases}
$$

$$
L'_n(t, p_n^{\mathrm{L}}(t)) \geq
\begin{cases}
\dfrac{N_{n1}w_1(p)kt^\delta(1-t)^\delta}{\sqrt{n}}(-b + o_p(1)) & \text{if } p_n^{\mathrm{L}}(t) > 0, \\
\infty & \text{if } p_n^{\mathrm{L}}(t) \leq 0.
\end{cases}
$$

Hence, (8) and (9) are satisfied with $(0, c_n]$ in place of $(0, 1)$. $\qquad\square$

Proof of Theorem 3 For symmetry reasons it suffices to prove the first part about the left tails. Let $(c_n)_n$ be a sequence of numbers in $(0, 1/2]$ converging to zero. Then for $t \in (0, c_n]$ and $\delta := \kappa/2 \in (0, 1/2)$,

$$
\left|\sqrt{n}\big(\widehat{B}_n^{\mathrm{S}}(t) - t\big) - \mathbb{V}_n^{(\ell)}(t)\right| = \left|\sum_{r=2}^{k} \frac{\mathbb{V}_{nr}(B_r(t))}{k\sqrt{N_{nr}/n}}\right| \leq t^{2\delta}o_p(1) = t^\kappa o_p(1).
$$

Concerning $\widehat{B}_n^{\mathrm{M}}$ and $\widehat{B}_n^{\mathrm{L}}$, for any $t \in (0, c_n]$ and $p \in (0, 1)$,

$$
\begin{aligned}
n\widehat{B}_n(t) &- \sum_{r=1}^{k} N_{nr} B_r(p) \\
&= \sum_{r=1}^{k} \sqrt{N_{nr}}\mathbb{V}_{nr}(B_r(t)) - \sum_{r=1}^{k} N_{nr}(B_r(p) - B_r(t)) \\
&= \sqrt{N_{n1}}\mathbb{V}_{n1}(B_1(t)) - N_{n1}k(p - t) + \rho_n^{\mathrm{M}}(t, p) \\
&= N_{n1}k\Big(\frac{\mathbb{V}_{n1}(B_1(t))}{k\sqrt{N_{n1}}} - (p - t)\Big) + \rho_n^{\mathrm{M}}(t, p)
\end{aligned}
$$

and

$$L'_n(t, p) = \sum_{r=1}^{k} \sqrt{N_{nr}} w_r(p) \mathbb{V}_{nr}(B_r(t)) - \sum_{r=1}^{k} N_{nr} w_r(p)(B_r(p) - B_r(t))$$

$$= \sqrt{N_{n1}} w_1(p) \mathbb{V}_{n1}(B_1(t)) - N_{n1} w_1(p) k(p - t) + \rho_n^{\mathrm{L}}(t, p)$$

$$= N_{n1} k w_1(p) \left(\frac{\mathbb{V}_{n1}(B_1(t))}{k\sqrt{N_{n1}}} - (p - t) \right) + \rho_n^{\mathrm{L}}(t, p),$$

where

$$|\rho_n^{\mathrm{M}}(t, p)| \le o_p(\sqrt{n}) t^{2\delta} + O(n) \max(t, p)(p - t),$$
$$|\rho_n^{\mathrm{L}}(t, p)| \le o_p(\sqrt{n}) p^{-1} t^{2\delta} + O(n) p^{-1} \max(t, p)(p - t).$$

Now we proceed similarly as in the proof of Theorem 2, defining

$$p_n(t) := t + \frac{\mathbb{V}_n^{(\ell)}(t) + bt^{\kappa}}{\sqrt{n}}$$

for some fixed $b \ne 0$. Note that for $t \in (0, c_n]$,

$$|p_n(t) - t| \le o_p(n^{-1/2}) t^{\delta} + O(n^{-1/2}) t^{\kappa} = o_p(n^{-1/2}) t^{\delta},$$

because $\kappa > \delta$. Note also that

$$t + \frac{\mathbb{V}_n^{(\ell)}(t)}{\sqrt{n}} = t + \frac{\mathbb{V}_{n1}(B_1(t))}{k\sqrt{N_{n1}}} = t - \frac{1 - (1 - t)^k}{k} + \frac{\widehat{B}_{n1}(t)}{k} > 0 \quad \text{on } (0, 1),$$

because $\widehat{B}_{n1} \ge 0$ and $t \mapsto t - (1 - (1 - t)^k)/k$ is strictly convex on $[0, 1]$ with derivative 0 at 0. Thus $p_n(t) > 0$ for all $t \in (0, c_n]$ in case of $b > 0$.

In case of $b > 0$, we may conclude that

$$n\widehat{B}_n(t) - \sum_{r=1}^{k} N_{nr} B_r(p_n(t))$$

$$= N_{n1} k \frac{-bt^{\kappa}}{\sqrt{n}} + \rho_n^{\mathrm{M}}(t, p_n(t))$$

$$\le \frac{N_{n1} k}{\sqrt{n}} \left(-bt^{\kappa} + o_p(1) t^{2\delta} + O(1)(t + o_p(n^{-1/2}) t^{\delta}) t^{\delta} \right)$$

$$\le \frac{N_{n1} k t^{\kappa}}{\sqrt{n}} (-b + o_p(1)),$$

and

$$L'_n(t, p_n(t)) \le \frac{N_{n1} k w_1(p) t^{\kappa}}{\sqrt{n}} (-b + o_p(1)).$$

Hence for any fixed $b > 0$,

$$\mathbb{P}\left(\sqrt{n}(\widehat{B}_n^Z(t) - t) \leq \mathbb{V}_n^{(\ell)}(t) + bt^\kappa \text{ for } t \in (0, c_n]\right) \rightarrow 0.$$

Similarly we can show that for any fixed $b < 0$, with asymptotic probability one, $\sqrt{n}(\widehat{B}_n^Z(t) - t) \leq \mathbb{V}_n^{(\ell)}(t) + bt^\kappa$ for all $t \in (0, c_n]$. □

Proof of Corollary 4 It follows from Proposition 6 that

$$\sup_{1 \leq r \leq k,\, u \in [0,1]} |\mathbb{V}_{nr}(u)| = O_p(1).$$

Together with (5) this entails that $\sup_{t \in [0,1]} |\mathbb{V}_n^Z(t) - \widetilde{\mathbb{V}}_n^Z(t)| \rightarrow_p 0$, where $\widetilde{\mathbb{V}}_n^Z :=$ $\sum_{r=1}^k \gamma_r^Z \mathbb{V}_{nr} \circ B_r$. But $\gamma_r^Z \equiv 0$ whenever $\pi_r = 0$. In case of $\pi_r > 0$ it follows from Proposition 6 that $\mathbb{V}_{nr}$ converges in distribution to $\mathbb{V}_r$. Consequently, $\widetilde{\mathbb{V}}_n^Z$ converges in distribution to the Gaussian process $\mathbb{V}^Z = \sum_{r=1}^k \gamma_r^Z \mathbb{V}_r \circ B_r$. □

Proof of Theorem 5 The asserted inequalities follow from Jensen's inequality. On the one hand, it follows from $w_r = \beta_r/(B_r(1 - B_r))$ and $\sum_{r=1}^k \beta_r \equiv k$ that

$$\begin{aligned}
K^S(t) &= \frac{1}{k} \sum_{r=1}^k \frac{\beta_r(t)}{k} \cdot (\pi_r w_r(t))^{-1} \\
&\geq \frac{1}{k} \left(\sum_{r=1}^k \frac{\beta_r(t)}{k} \cdot \pi_r w_r(t)\right)^{-1} \\
&= \left(\sum_{r=1}^k \pi_r \beta_r(t) w_r(t)\right)^{-1} = K^L(t).
\end{aligned}$$

Equality holds if, and only if,

$$\pi_1 w_1(t) = \pi_2 w_2(t) = \cdots = \pi_k w_k(t).$$

But

$$w_1(t) = \frac{k}{(1 - t)(1 - (1 - t)^k)} \quad \text{and} \quad w_k(t) = \frac{k}{t(1 - t^k)},$$

so

$$\frac{w_k(t)}{w_1(t)} = \frac{(1 - t)(1 - (1 - t)^k)}{t(1 - t^k)} = \frac{\sum_{j=0}^{k-1}(1 - t)^j}{\sum_{j=0}^{k-1} t^j}$$

is strictly decreasing in t. Hence there is at most one solution of the equation $\pi_1 w_1(t) = \pi_k w_k(t)$.

Similarly, with $a_r(t) := \pi_r \beta_r(t) / \sum_{s=1}^k \pi_s \beta_s(t)$,

$$K^{\mathrm{M}}(t) = \sum_{r=1}^k \pi_r \beta_r(t) \cdot w_r(t)^{-1} \Big/ \left(\sum_{s=1}^k \pi_s \beta_s(t) \right)^2$$

$$= \sum_{r=1}^k a_r(t) \cdot w_r(t)^{-1} \Big/ \sum_{s=1}^k \pi_s \beta_s(t)$$

$$\geq \left(\sum_{r=1}^k a_r(t) w_r(t) \right)^{-1} \Big/ \sum_{s=1}^k \pi_s \beta_s(t)$$

$$= \left(\sum_{r=1}^k \pi_r \beta_r(t) w_r(t) \right)^{-1} = K^{\mathrm{L}}(t).$$

Here the inequality is strict unless

$$w_1(t) = w_2(t) = \cdots = w_k(t).$$

But $w_1(t) = w_k(t)$ implies that $t = 1/2$. Moreover, $w_1(1/2) = 2k/(1 - 2^{-k})$ and

$$w_{k-1}(1/2) = \frac{2k(k-1)}{(k+1)(1 - (k+1)2^{-k})}$$

are identical if, and only if, $k^2 + k + 2 = 2^{k+1}$. But $2^{k+1} = 2 \sum_{j=0}^k \binom{k}{j}$ is strictly larger than $2(1 + k + k(k-1)/2) = k^2 + k + 2$ if $k \geq 3$.

As to the ratios $E^{\mathrm{Z}}(t) := K^{\mathrm{Z}}(t)/K^{\mathrm{L}}(t)$, note first that

$$E^{\mathrm{S}}(t) = \sum_{r=1}^k \frac{B_r(t)(1 - B_r(t))}{k^2 \pi_r} \sum_{s=1}^k \pi_s \beta_s(t) w_s(t)$$

$$\geq \min_{r,s=1,\ldots,k} \frac{B_r(t)(1 - B_r(t)) \beta_s(t) w_s(t)}{k^2} \Big/ \min_{r=1,\ldots,k} \pi_r$$

$$\to \infty \quad \text{as} \quad \min_{r=1,\ldots,k} \pi_r \downarrow 0.$$

On the other hand, with $a_r(t)$ as above,

$$E^{\mathrm{M}}(t) = \sum_{r=1}^k a_r(t) w_r(t)^{-1} \sum_{s=1}^k a_s(t) w_s(t) = \mathbb{E}(W) \mathbb{E}(W^{-1})$$

with a random variable W with distribution $\sum_{r=1}^k a_r(t) \delta_{w_r(t)}$. But with $\ell(t) := \min_r w_r(t)$ and $u(t) := \max_r w_r(t)$, convexity of $w \mapsto w^{-1}$ on $[\ell(t), u(t)]$ implies that

$$W^{-1} \leq \frac{W - \ell(t)}{u(t) - \ell(t)} u(t)^{-1} + \frac{u(t) - W}{u(t) - \ell(t)} \ell(t)^{-1},$$

so

$$
\begin{aligned}
\mathbb{E}(W)\,\mathbb{E}(W^{-1}) &\leq \mathbb{E}(W) \left(\frac{\mathbb{E}(W) - \ell(t)}{u(t) - \ell(t)} u(t)^{-1} + \frac{u(t) - \mathbb{E}(W)}{u(t) - \ell(t)} \ell(t)^{-1} \right) \\
&= \frac{\mathbb{E}(W)(\ell(t) + u(t) - \mathbb{E}(W))}{\ell(t) u(t)} \\
&\leq \frac{(\ell(t) + u(t))^2}{4\ell(t) u(t)} = \frac{\rho(t) + \rho(t)^{-1} + 2}{4}.
\end{aligned}
$$

This upper bound for $E^{\mathrm{M}}(t)$ is attained approximately, if the distribution of W approaches the uniform distribution on $\{\ell(t), u(t)\}$. Hence we should choose $(\pi_r)_{r=1}^k$ as follows: Let $r(1), r(2)$ be two different numbers in $\{1, \ldots, k\}$ such that $w_{r(1)}(t) = \ell(t)$ and $w_{r(2)}(t) = u(t)$. Then let

$$\pi_r \approx \begin{cases} \beta_r(t)^{-1}/(\beta_{r(1)}^{-1} + \beta_{r(2)}^{-1}) & \text{for } r \in \{r(1), r(2)\}, \\ 0 & \text{for } r \notin \{r(1), r(2)\}. \end{cases}$$

The inequality $\rho(t) \leq k$ follows from Lemma 7 and the fact that $\rho(t)$ remains unchanged if we replace $w_r(t)$ with $\widetilde{w}_r(t) = t(1 - t)w_t(t) \in [1, k]$. □

References

Balakrishnan, N., Li, T. (2006). Confidence intervals for quantiles and tolerance intervals based on ordered ranked set samples. *Annals of the Institute of Statistical Mathematics*, 58, 757–777.

Bhoj, D. S. (2001). Ranked set sampling with unequal samples. *Biometrics*, 57(3), 957–962.

Chen, Z. (2001). Non-parametric inferences based on general unbalanced ranked-set samples. *Journal of Nonparametric Statistics*, 13(2), 291–310.

Chen, Z., Bai, Z., Sinha, B. K. (2004). *Ranked set sampling. Theory and applications*. New York: Springer.

Clopper, C. J., Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4), 404–413.

Dastbaravarde, A., Arghami, N. R., Sarmad, M. (2016). Some theoretical results concerning non parametric estimation by using a judgment poststratification sample. *Communications in Statistics, Theory and Methods*, 45(8), 2181–2203.

David, H. A., Nagaraja, H. N. (2003). *Order statistics* (3rd ed.). Hoboken, NJ: Wiley-Interscience.

Dell, T. R., Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28(2), 545–555.

Frey, J., Ozturk, O. (2011). Constrained estimation using judgement post-stratification. *Annals of the Institute of Statistical Mathematics*, 63, 769–789.

Ghosh, K., Tiwari, R. C. (2008). Estimating the distribution function using k-tuple ranked set samples. *Journal of Statistical Planning and Inference*, 138(4), 929–949.

Huang, J. (1997). Properties of the Npmle of a distribution function based on ranked set samples. *Annals of Statistics*, 25(3), 1036–1049.

Kvam, P. H., Samaniego, F. J. (1994). Nonparametric maximum likelihood estimation based on ranked set samples. *Journal of the American Statistical Association*, 89(426), 526–537.

MacEachern, S. N., Stasny, E. A., Wolfe, D. A. (2004). Judgement post-stratification with imprecise rankings. *Biometrics*, 60, 207–215.

McIntyre, G. A. (1952). A method of unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, *3*, 385–390.

Presnell, B., Bohn, L. L. (1999). U-Statistics and imperfect ranking in ranked set sampling. *Journal of Nonparamatric Statistics*, *10*(2), 111–126.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed July 2018.

Shorack, G. R., Wellner, J. A. (1986). *Empirical processes with applications to statistics*. New York: Wiley.

Stokes, S. L., Sager, T. W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, *83*(402), 374–381.

Terpstra, J. T., Miller, Z. A. (2006). Exact inference for a population proportion based on a ranked set sample. *Communications in Statistics, Simulation and Computation*, *35*(1), 19–26.

Wang, X., Wang, K., Lim, J. (2012). Isotonized CDF estimation from judgement poststratification data with empty strata. *Biometrics*, *68*(1), 194–202.

Wolfe, D. A. (2004). Ranked set sampling: An approach to more efficient data collection. *Statistical Science*, *19*(4), 636–643.

Wolfe, D. A. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *ISRN Probability and Statistics*, *2012*, 568385. https://doi.org/10.5402/2012/568385.