Annals of the Institute of Statistical Mathematics (2023) 75:303–333
https://doi.org/10.1007/s10463-022-00845-3

On the choice of the optimal single order statistic in quantile estimation

Mariusz Bieniek[1] · **Luiza Pańczyk**[1]

Received: 3 January 2022 / Revised: 14 June 2022 / Accepted: 21 June 2022 /
Published online: 2 August 2022

Abstract
We study the classical statistical problem of the estimation of quantiles by order statistics of the random sample. For fixed sample size, we determine the single order statistic which is the optimal estimator of a quantile of given order. We propose a totally new approach to the problem, since our optimality criterion is based on the use of nonparametric sharp upper and lower bounds on the bias of the estimation. First, we determine the explicit analytic expressions for the bounds, and then, we choose the order statistic for which the upper and lower bound are simultaneously as close to 0 as possible. The paper contains rigorously proved theoretical results which can be easily implemented in practise. This is also illustrated with numerical examples.

Keywords Bias · Nonparametric statistics · Order statistics · Quantile estimation · Sharp bounds · Small sample

1 Introduction

One of the most important problems of mathematical and applied statistics is to determine the unknown distribution function F of some random quantity based on a sample of observations. In theory, this is always possible by the Glivenko-Cantelli theorem which says that the empirical distribution functions converge uniformly to F as the sample size goes to infinity. However, the application of this result demands large sample sizes, and even if F is continuous, every empirical distribution function

✉ Mariusz Bieniek
mariusz.bieniek@umcs.lublin.pl

Luiza Pańczyk
luizapanczyk@gmail.com

[1] Institute of Mathematics, Maria Curie Skłodowska University, pl. M. Curie Skłodowskiej 1, 20-031 Lublin, Poland

is a stepwise function. Therefore, this approach is of little use if the sample size is small.

On the other hand, it often suffices to determine not the whole distribution, but only some of its quantiles. A very good introduction to the problem of quantile estimation and a rich source of references can be found in Keating and Tripathi (2006). A very important class of estimators of population quantiles are appropriately chosen L-statistics, i.e. linear combinations of order statistics $\sum_{j=1}^{n} c_{j,n} X_{j:n}$. The general problem we address here is to choose the optimal L-statistics which best estimates an unknown quantile of the given order $p \in (0,1)$ assuming that the sample size n is fixed. Among L–statistics, two interesting special cases are single order statistics and linear combinations of two adjacent order statistics. See Hyndman and Fan (1996) and Parrish (1990) for extensive lists of L–statistics used in quantile estimation.

In this paper, we focus on the problem of estimation of population quantiles by appropriately chosen single order statistics, which are the most popular estimators of quantiles. Despite its simplicity, sample quantiles may be very poor estimators of quantiles, see e.g. Zieliński (2009). Therefore, the restriction to the choice of single order statistics may seem unimportant. However, this study is a part of a broader research topic devoted to the choice of optimal L–statistics (linear combinations of order statistics) in quantile estimation. It appears that the consideration of single order statistics is a very essential first step in general reasoning, so it has to be carried out on its own merits.

First, we recall some standard definitions and notations. For any random variable X with distribution function F, the quantile of F of order $p \in (0,1)$ is defined as any number $x_p = x_p(F)$ such that

$$\mathbb{P}(X \le x_p) \ge p, \quad \mathbb{P}(X \ge x_p) \ge 1 - p,$$

or equivalently $F(x_p^-) \le p \le F(x_p)$. Here and in the rest of the paper, we use standard notations $f(a^-)$ and $f(a^+)$ for the left- and right-hand limits of a function f at the point a. Since x_p is possibly non-unique, then we define the upper and lower quantile functions of F as

$$F^{\rightarrow}(p) = \sup\{x \in \mathbb{R} : F(x) \le p\},$$

and

$$F^{\leftarrow}(p) = \inf\{x \in \mathbb{R} : F(x) \ge p\}, \tag{1}$$

respectively. Then, for any $p \in (0,1)$, the values $F^{\rightarrow}(p)$ and $F^{\leftarrow}(p)$ are uniquely determined, and for any quantile $x_p(F)$, we have $F^{\leftarrow}(p) \le x_p(F) \le F^{\rightarrow}(p)$. Moreover, the functions $F^{\rightarrow}$ and $F^{\leftarrow}$ are right- and left-continuous, respectively.

Next, let $(X_1, X_2, \ldots, X_n)$ be the random sample consisting of independent and identically distributed copies of X. Then by $X_{1:n} \le X_{2:n} \le \ldots \le X_{n:n}$, we denote the order statistics of the sample. The choice of the single order statistic $X_{j:n}$ as an estimator of x_p is usually justified mainly by its asymptotic properties. Excellent references on this subject are monographs by Serfling (1980) and Reiss (1989).

Firstly, it is well known that $X_{j:n}$ is a strongly consistent estimator of x_p under the assumptions that $\frac{j}{n} \to p$ as $n \to \infty$ and x_p is uniquely determined. This says that

if the sample size is fixed, then j should be chosen as close to np as possible, so the usual choice is either $j = \lfloor np \rfloor$ or $j = \lceil np \rceil$. Here and in the rest of the paper, $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote usual floor and ceiling functions. The most often choice is $j = \lceil np \rceil$ since if I_{A} denotes the indicator of an event A and

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \le x\}} = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_{i:n} \le x\}}, \quad x \in \mathbb{R},$$

denotes the empirical distribution function of the sample, then using the notation (1) we obtain $F_n^{\leftarrow}(p) = X_{\lceil np \rceil : n}$ for all $p \in (0, 1)$. Then, $X_{\lceil np \rceil : n}$ is called sample quantile of order p. Finally, it is also well known that the sample quantiles are asymptotically normal. More precisely, assume that F is differentiable at $F^{\leftarrow}(p)$ and $F'(F^{\leftarrow}(p)) > 0$ which implies that x_p is uniquely defined. In this case, if $\sqrt{n}\left(\frac{j}{n} - p\right) \to 0$ as $n \to \infty$, then $X_{j:n}$ is asymptotically normal with mean x_p and variance $\frac{p(1-p)}{n(F'(\xi_p))^2}$. Unfortunately, the above results do not apply to small sample sizes. Namely, for large sample sizes, the difference between $X_{\lceil np \rceil : n}$ and $X_{\lfloor np \rfloor : n}$ is irrelevant, but for small sizes, it is not so obvious which choice gives "better" results.

To distinguish optimal estimators of quantiles, we introduce a totally new criterion of optimality based on sharp upper and lower bounds of the estimation of $x_p(F)$ by a single order statistic derived in Sect. 3. Using this criterion, we derive our main result which reveals rather simple conditions indicating which of the order statistics $X_{\lfloor np \rfloor : n}$ or $X_{\lceil np \rceil : n}$ or $X_{\lceil np \rceil + 1 : n}$ is the best estimator of the quantile of order $p \in (0, 1)$. Strictly speaking, we prove that for a fixed sample size $n \ge 2$, there exists the uniquely determined set of numbers $a_{j,n}$, $1 \le j < n$, such that if $p \in (a_{j-1,n}, a_{j,n})$, then the optimal estimator of x_p is $X_{j:n}$ and if $p = a_{j,n}$ then both $X_{j:n}$ and $X_{j+1:n}$ are optimal (see Theorem 5). Moreover, $a_{j,n} < \frac{j}{n} < a_{j+1,n}$ for $1 \le j < \frac{n}{2}$ and $a_{n-j,n} = 1 - a_{j,n}$. In particular, as a special case, we obtain rather surprising result concerning the case when $p = \frac{k}{n}$. Contrary to traditional choice, we show that the optimal estimator of $x_{k/n}$ is $X_{k:n}$ if $1 \le k < \frac{n}{2}$ and $X_{k+1:n}$ if $\frac{n}{2} < k \le n$. If $k = \frac{n}{2}$, then both $X_{n/2:n}$ and $X_{n/2+1:n}$ are optimal.

Obviously, the choice depends on the adopted criterion of optimality. In this paper and in our forthcoming papers, we propose a totally new approach based on the analysis of sharp upper and lower nonparametric bounds on the bias of the estimation of x_p by $X_{j:n}$. Since the sample size is fixed and we do not assume the uniqueness of x_p, our criterion of optimality is as follows. Consider the class $\mathcal{F}$ of all distribution functions F with mean μ_F and finite variance σ_F^2 which may be expressed in terms of the quantile functions as

$$\mu_F = \int_0^1 F^{\leftarrow}(u)du, \quad \sigma_F^2 = \int_0^1 \left(F^{\leftarrow}(u) - \mu_F\right)^2 du, \tag{2}$$

where $F^{\leftarrow}$ may be replaced by $F^{\rightarrow}$. For fixed $n \ge 2$, $1 \le j \le n$ and $p \in (0, 1)$ we denote by

$$\overline{B}(j,n,p) = \sup_{F\in\mathcal{F}} \frac{\mathbb{E}X_{j:n} - F^{\leftarrow}(p)}{\sigma_{\mathrm{F}}} \tag{3}$$

and

$$\underline{B}(j,n,p) = \inf_{F\in\mathcal{F}} \frac{\mathbb{E}X_{j:n} - F^{\rightarrow}(p)}{\sigma_{\mathrm{F}}} \tag{4}$$

the upper and lower sharp bounds on the bias of $X_{j:n}$ as an estimator of $F^{\leftarrow}(p)$ and $F^{\rightarrow}(p)$, respectively.

Definition 1 Fix $n \geq 2$ and $p \in (0,1)$. The order statistic $X_{k:n}$ is said to be the optimal estimator of x_p if k minimizes the function $s_{n,p}$ defined by

$$s_{n,p}(j) = \sqrt{p(1-p)}\left|\overline{B}(j,n,p) + \underline{B}(j,n,p)\right|$$

with respect to $j \in \{1,\ldots,n\}$.

In general, k does not have to be unique, but obviously it depends on n and p. To justify our criterion, we first note that, by (6) and (7) below, it follows that $\underline{B}(j,n,p) < 0 < \overline{B}(j,n,p)$ for $1 \leq j \leq n$. Therefore, the most reasonable criterion seems to be to minimize the difference $\overline{B}(j,n,p) - \underline{B}(j,n,p)$ with respect to j. However, from the results of Sect. 3, it follows that the above difference is at least $1/\sqrt{p(1-p)}$ independently of j. The numerical computations show that in most of the cases, the difference is equal to $1/\sqrt{p(1-p)}$ for at least few values of j. Therefore, such criterion would be useless. On the other hand, if the difference is replaced with the sum (as in the definition of $s_{n,p}$), then we measure the degree of the symmetry of the bounds about 0. If the values of $s_{n,p}$ are close to 0, then the values of the corresponding upper and lower bounds are more or less symmetric about 0. In other words, an appropriately chosen order statistic on the average does not overestimate or underestimate any quantile of order p regardless of the underlying distribution F. The normalizing factor $\sqrt{p(1-p)}$ is introduced to avoid large values of $s_{n,p}$ for p close to 0 or 1. More precisely, for all $1 \leq j \leq n$ taking into account the explicit values of the bounds given in Theorem 2 we obtain $\lim_{p\to 0^+} s_{n,p}(j) = \lim_{p\to 1^-} s_{n,p}(j) = 1$. Moreover, for any j fixed, $s_{n,p}(j)$ is a continuous function of $p \in [0,1]$, so it is bounded on this interval.

Another motivation for our criterion is the following. If x_p is any fixed value of pth quantile of F, then by definitions (3) and (4) for fixed $j = 1,\ldots,n$ we have

$$\left|\frac{\mathbb{E}X_{j:n} - x_p}{\sigma_{\mathrm{F}}}\right| \leq \max\left(\overline{B}(j,n,p), -\underline{B}(j,n,p)\right) =: c_{n,p}(j) \tag{5}$$

for all distributions $F \in \mathcal{F}$, but this bound cannot be sharp. The value $c_{n,p}(j)$ may be considered as the maximum bias of $X_{j:n}$ as the estimator of x_p, measured in standard deviation units of F, with respect to all $F \in \mathcal{F}$. The notion of maximum bias comes from robust statistics, but here we use it in the nonparametric context. Namely, the

idea is to choose j such that the maximum bias is as small as possible. This way we minimize maximal possible error of the estimation of pth quantile by single order statistic. However, we show in Theorem 6 that, for most values of p, the functions $s_{n,p}$ and $c_{n,p}$ are minimized by exactly the same value of j. Therefore, both the criteria are equivalent. This is not surprising since both criteria aim at balancing the upper and lower bounds on the bias, but still the equivalence is not obvious at all.

As it can be seen from the above discussion, our approach is nonparametric in the sense that we always aim at taking into account the worst case scenario. Also, we concentrate our attention on the optimization of the bias of the estimation, but in the last section of the paper, we will comment on its mean square error. We underline again that we do not restrict our considerations to absolutely continuous distribution functions, as it is usually done in the literature devoted to quantile estimation. We also stress the fact that we present complete theory composed with theoretical results which are proven rigorously. However, the results may be easily implemented in statistical practise.

The paper is organized as follows. In Sect. 2, we investigate some preliminary properties of the bounds $\overline{B}(j,n,p)$, $\underline{B}(j,n,p)$ and of the function $s_{n,p}$ without the knowledge of their explicit form. Next, in Sect. 3, we present the explicit values of the bounds and we study their further properties. In Sect. 4, we give the formulation and proof of our main results. In Sect. 5, we provide numerical examples illustrating obtained results, and Sect. 6 contains short summary and discussion of the results of the paper and remarks on mean square error of obtained estimators. Appendices contain the proofs of some technical results of Sects. 3 and 4.

2 Preliminary results

In this section, we introduce further notation and we prove preliminary properties of the optimal bounds. Let us stress that the properties are proved directly from the definitions without the assumptions of the knowledge of their exact form. Namely the proofs rely only on the definitions of $\underline{B}(j,n,p)$ and $\overline{B}(j,n,p)$ given by (3) and (4). First we note that since $F^{\leftarrow}(u) \leq F^{\rightarrow}(u)$ for $u \in (0,1)$, then obviously

$$\underline{B}(j,n,p) \leq \overline{B}(j,n,p). \tag{6}$$

Next, we show that it is sufficient to determine the values of the upper bounds $\overline{B}(j,n,p)$ or, strictly speaking,

$$\underline{B}(j,n,p) = -\overline{B}(n-j+1,n,1-p). \tag{7}$$

Indeed, consider the random variables X and $Y = -X$. Then, the distribution function of Y is $G(x) = 1 - F(-x)$ and obviously $\sigma_F = \sigma_G$. The detailed and careful analysis shows that for any distribution F we have $\sup\{y : F(y) \leq p\} = \sup\{y : F(y^-) \leq p\}$. This implies that $G^{\rightarrow}(u) = -F^{\leftarrow}(1-u)$ for $u \in (0,1)$. Moreover, if $Y_i = -X_i$ for $i = 1, \ldots, n$, then $Y_{i:n} = -X_{n-i+1:n}$ which implies

$$\inf_{F\in\mathcal{F}} \frac{\mathbb{E}X_{j:n} - F^{\rightarrow}(p)}{\sigma_{\mathrm{F}}} = -\sup_{G\in\mathcal{F}} \frac{\mathbb{E}Y_{n-j+1:n} - G^{\leftarrow}(1-p)}{\sigma_{\mathrm{G}}},$$

so (7) is proved.

To find j which minimizes $s_{n,p}$ first, we consider the function $r_{n,p}$ defined by

$$r_{\mathrm{n,p}}(j) = \overline{B}(j,n,p) + \underline{B}(j,n,p), \quad 1 \le j \le n. \tag{8}$$

By the definition we have $X_{j:n} \le X_{j+1:n}$ and therefore obviously

$$\overline{B}(j,n,p) \le \overline{B}(j+1,n,p) \quad \text{and} \quad \underline{B}(j,n,p) \le \underline{B}(j+1,n,p). \tag{9}$$

Later on we show that both above inequalities are strict (see Corollary 2). In other words, the functions $\overline{B}(j,n,p)$ and $\underline{B}(j,n,p)$ are strictly increasing with respect to j with fixed n and p. In consequence, $r_{n,p}$ is also strictly increasing function of $j \in \{1,\ldots,n\}$.

Theorem 1 *The optimal choice of k which minimizes $s_{n,p}$ is the following*:

(a) *if $|r_{n,p}(1)| < r_{n,p}(2)$, then $k = 1$, and if $r_{n,p}(n-1) > |r_{n,p}(n)|$, then $k = n$;*
(b) *otherwise $k = j$ or $k = j+1$, where $j \in \{2,\ldots,n-2\}$ is the unique index such that*

$$r_{n,p}(j) \le 0 < r_{n,p}(j+1); \tag{10}$$

(c) *in particular, if*

$$r_{n,p}(j) = -\,r_{n,p}(j+1), \tag{11}$$

then $s_{n,p}(j) = s_{n,p}(j+1)$ and both j and $j+1$ are optimal.

Proof If $r_{\mathrm{n,p}}(1) \ge 0$ then by the increasing monotonicity of $r_{n,p}$, the optimal choice is $k = 1$. If $r_{n,p}(1) < 0 < r_{n,p}(2)$ but $-r_{n,p}(1) < r_{n,p}(2)$ then again the optimal choice is $k = 1$. This proves (a). If j is determined by (10), then $s_{n,p}$ is nonnegative decreasing-increasing which proves (b) and (c). □

Let us define the set-valued function $k(n,p)$ which values, for fixed n and p, are subsets of $\{1,2,\ldots,n\}$ consisting of all indices minimizing $s_{n,p}$. The above theorem says that $k(n,p)$ is a singleton unless the condition (11) holds in which case $k(n,p) = \{j, j+1\}$. In the former case, for simplicity, we write $k(n,p) = j$ instead of $k(n,p) = \{j\}$.

We will compare the values of k for fixed n and various values of p, and in particular we write $k(n,p) \le k(n,q)$ if this inequality is satisfied in the usual sense for any values of $k(n,p)$ and $k(n,q)$. We will also write $k(n,p) \le j$ if both possible values of $k(n,p)$ are not greater than j. The following lemma informally says that $k(n,p)$ is nondecreasing with respect to $p \in (0,1)$.

Lemma 1 *For $0 < p < q < 1$, we have $k(n,p) \le k(n,q)$.*

Proof Assume that $0 < p < q < 1$. Since $F^{\leftarrow}$ and $F^{\rightarrow}$ are nondecreasing, then by the definitions (3) and (4) we get

$$\overline{B}(j,n,q) \le \overline{B}(j,n,p) \quad \text{and} \quad \underline{B}(j,n,q) \le \underline{B}(j,n,p). \tag{12}$$

In Corollary 2, we will show that both the above inequalities are sharp. Therefore, $r_{n,q}(j) < r_{n,p}(j)$ for $1 \le j \le n$.

Now let us assume that for some k and ℓ, we have $r_{n,p}(k) \le 0 < r_{n,p}(k+1)$ and $r_{n,q}(\ell) \le 0 < r_{n,q}(\ell+1)$. Assuming that $\ell < k$ we obtain $\ell + 1 \le k$, so $r_{n,q}(\ell+1) \le r_{n,q}(k) < r_{n,p}(k) \le 0$, which is a contradiction. Also, if $k = 1$ is optimal for q, then $\ell = 1$ is optimal for p. Thus, we proved that if k minimizes $s_{n,p}$ and ℓ minimizes $s_{n,q}$, then $k \le \ell$. □

By the above lemma, there exists the set of $n-1$ numbers $a_{1,n}, \ldots, a_{n-1,n}$ such that

(a) $0 = a_{0,n} < a_{1,n} < \ldots < a_{n-1,n} < a_{n,n} = 1$;
(b) $k(n, a_{j,n}) = \{j, j+1\}$ for $1 \le j \le n-1$;
(c) $k(n,p)$ has constant value $j+1$ for $p \in (a_{j,n}, a_{j+1,n})$ and $0 \le j \le n-1$.

Therefore, our problem may be reformulated as: (a) to find the numbers $a_{j,n}$ and (b) to determine their locations with respect to $\frac{j}{n}$, $1 \le j \le n$.

We first show that $a_{j,n} = 1 - a_{n-j,n}$ for $0 \le j \le n$. This is an easy consequence of the next lemma.

Lemma 2 *For $p \in (0,1)$ we have*:

(a) *$k(n,p) = j$ if and only if $k(n,1-p) = n-j+1$;*
(b) *$k(n,p) = \{j, j+1\}$ if and only if $k(n,1-p) = \{n-j, n-j+1\}$.*

Proof By (7) and the definitions of $r_{n,p}$ and $s_{n,p}$, we have

$$r_{n,1-p}(j) = -r_{n,p}(n-j+1), \quad s_{n,1-p}(j) = s_{n,p}(n-j+1). \tag{13}$$

for $p \in (0,1)$ and $1 \le j \le n$. Now it is straightforward to see that j minimizes $s_{n,p}$ if and only if $s_{n,1-p}$ is minimized by $n-j+1$. □

Now we show that if n is even then $a_{n/2,n} = \frac{1}{2}$, and if n is odd then $a_{(n-1)/2,n} < \frac{1}{2} < a_{(n+1)/2,n}$. This is easily inferred from the values of $k(n, 1/2)$ which are given below.

Lemma 3 *For $n \ge 2$, we have*

$$k(n,1/2)=\begin{cases}\frac{n+1}{2}, & \text{if } n \text{ is odd},\\ \left\{\frac{n}{2},\frac{n}{2}+1\right\}, & \text{if } n \text{ is even}.\end{cases}$$

Proof For $p=1/2$, the equation (13) yields $r_{n,1/2}(n-j+1)=-r_{n,1/2}(j)$. If n is even, then for $j=n/2$ we get

$$r_{n,1/2}\left(\frac{n}{2}+1\right)=-r_{n,1/2}\left(\frac{n}{2}\right).$$

Since $r_{n,p}$ is strictly increasing, we get $r_{n,1/2}(n/2)<0$ and $s_{n,1/2}(n/2)=s_{n,1/2}(n/2+1)$. If n is odd, then putting $j=(n+1)/2$ we get

$$r_{n,1/2}\left(\frac{n+1}{2}\right)=-r_{n,1/2}\left(\frac{n+1}{2}\right),$$

so $r_{n,1/2}((n+1)/2)=0$. Now the claim follows from Theorem 1(b) and (c). □

Remark 1 Note that if n is odd, then our criterion of optimality yields the classical estimator of the median $X_{\frac{n+1}{2}:n}$. If n is even, then both $X_{\frac{n}{2}:n}$ and $X_{\frac{n}{2}+1:n}$ are equally good. In this case using our approach we do not obtain classical estimator of the median, which is the mean of two middle order statistics. The obvious reason is that we consider estimators of population quantiles based on just single order statistic. The problem of the optimal choice of linear combination of neighbouring order statistics will be the subject of one of the forthcoming papers by the authors.

Since the problem of the optimal choice of the estimator of $x_{1/2}$ is solved, from now on we will assume that $p\neq\frac{1}{2}$. By Lemma 2 it suffices to determine $k(n,p)$ only for $p\in\left(0,\frac{1}{2}\right)$. The next lemma says that in this case, we should search for the optimal estimator only in the "lower half" of the order statistics.

Lemma 4 *For $p\in\left(0,\frac{1}{2}\right)$, it holds $k(n,p)\leq\left\lfloor\frac{n+1}{2}\right\rfloor$.*

Proof By Lemma 3, we obtain $k(n,1/2)\subset\left\{\left\lfloor\frac{n+1}{2}\right\rfloor,\left\lfloor\frac{n+1}{2}\right\rfloor+1\right\}$. By Lemma 1, this proves our claim. □

The conclusion of the last three lemmas is that to solve the problem of the optimal choice of order statistic in estimation of quantiles, it is sufficient to find the values of $a_{k,n}$ for $1\leq k\leq\frac{n}{2}$.

3 The values of optimal bounds

In this section, we determine the exact values of upper and lower bounds $\overline{B}(j,n,p)$ and $\underline{B}(j,n,p)$ for $1\leq j\leq n$ with fixed $n\geq 2$ and $p\in(0,1)$. It appears that the lower bounds are easily obtained once we find the upper ones and apply

the symmetry (7). In particular, we prove that for fixed $2 \le j \le n-1$, there exists some interval $(\theta_j(n), \xi_j(n))$ containing j/n such that for all $F \in \mathcal{F}$ and all $p \in (\theta_j(n), \xi_j(n))$

$$-\frac{F_{j:n}(p)}{\sqrt{p(1-p)}} \le \frac{\mathbb{E}X_{j:n} - F^{\rightarrow}(p)}{\sigma_F} \le \frac{\mathbb{E}X_{j:n} - F^{\leftarrow}(p)}{\sigma_F} \le \frac{1-F_{j:n}(p)}{\sqrt{p(1-p)}} \tag{14}$$

and the bounds are attained for appropriately chosen two-point distributions. For p outside this interval, either $\overline{B}(j,n,p)$ or $\underline{B}(j,n,p)$ has a much more complicated form. As a consequence, putting $p = \frac{j}{n}$ we obtain the correct version of Proposition 1 of Okolewski and Rychlik (2001). They claimed incorrectly that for the majority of j and n and $p = \frac{j}{n}$ the inequalities

$$-\frac{F_{j:n}(p)}{\sqrt{p(1-p)}} \le \frac{\mathbb{E}X_{j:n} - F^{\rightarrow}(p)}{\sigma_F} \le \frac{1-F_{j:n}(p)}{\sqrt{p(1-p)}} \tag{15}$$

are sharp which is not the case for the upper inequality. Moreover, they claimed that equalities are attained for the same two-point distribution which is obviously impossible. We also explain the flaw in their proof which led to this incorrect statement, see Remark 2.

In what follows we use the following notation. For $0 \le j \le n$ let

$$B_{j,n}(p) = \binom{n}{j} p^j (1-p)^{n-j}, \quad 0 \le p \le 1,$$

denote the Bernstein polynomials of order n. These polynomials have numerous interesting properties, but in fact we will need only two of them: variation diminishing property (VDP) and so-called Simmons' inequality. For the ease of readers, we recall them in Appendix 2 as Lemma 11 and 12. Then, for $1 \le j \le n$ and $p \in [0,1]$, we denote by

$$F_{j:n}(p) = \sum_{i=j}^{n} B_{i,n}(p)$$

and $f_{j:n}(p) = nB_{j-1,n-1}(p)$ the distribution function and density function, respectively, of j–th uniform order statistic of the sample of size n from the uniform distribution on the interval [0, 1]. From theoretical point of view, it is very interesting to note that in the proofs of the results of this section, we use the straightforward and very well-known connection between $F_{k:n}$ and the binomial distribution $\mathcal{B}(n,p)$ with parameters $n \in \mathbb{N}$ and $p \in (0,1)$. Namely, if $Y \sim \mathcal{B}(n,p)$ then

$$F_{j:n}(p) = \mathbb{P}(Y \ge j), \quad 1 \le j \le n,$$

and $1 - F_{j+1:n}(p) = \mathbb{P}(Y \le j)$ for $0 \le j \le n-1$. In particular, the median and the mode of $\mathcal{B}\left(n, \frac{j}{n}\right)$ distribution are both equal to j, see e.g. Kaas and Buhrman (1980). More precisely, in this case, j is the strong median of Y which is equivalent to

$$F_{j:n}\left(\frac{j}{n}\right) > \frac{1}{2} \quad \text{and} \quad F_{j+1:n}\left(\frac{j}{n}\right) < \frac{1}{2}. \tag{16}$$

Next, we define auxiliary numbers θ_j and ξ_j for $1 \le j \le n$. These numbers determine the form of the bounds (3) and (4), and the functions $r_{n,p}$ and $s_{n,p}$. For instance, it appears that if $p \in (\theta_j, \xi_j)$, then $r_{n,p}(j)$ has a simple form contrary to p outside this interval (see Corollary 3 for details).

For $2 \le j \le n-1$ we define $\theta_j = \theta_j(n)$ as the unique solution to the equation

$$1 - F_{j:n}(\theta_j) = (1-\theta_j) f_{j:n}(\theta_j) \tag{17}$$

in the interval $\left(0, \frac{j-1}{n-1}\right)$. Similarly, we define $\xi_j = \xi_j(n)$ as the unique solution to the equation $F_{j:n}(\xi_j) = \xi_j f_{j:n}(\xi_j)$ in the interval $\left(\frac{j-1}{n-1}, 1\right)$. Then, obviously $\theta_j < \xi_j$ for $2 \le j \le n-1$. Moreover, we adopt the convention $\theta_1 = \xi_1 = 0$ and $\theta_n = \xi_n = 1$. Taking into account obvious equalities

$$1 - F_{j:n}(u) = F_{n-j+1:n}(1-u) \tag{18}$$

and $f_{j:n}(u) = f_{n-j+1:n}(1-u)$ we infer that

$$\xi_j = 1 - \theta_{n-j+1}, \quad 1 \le j \le n. \tag{19}$$

Finally, although this is not relevant in this section, we prove here that both the sequences are strictly increasing.

Lemma 5 *For fixed $n \ge 3$, the sequences $\theta_j(n)$ and $\xi_j(n)$ are strictly increasing with respect to $j = 1, 2, \ldots, n$.*

Proof Since $\theta_1 = 0 < \theta_2$ and $\theta_{n-1} < 1 = \theta_n$, by the symmetry $\xi_j = 1 - \theta_{n-j+1}$ it is sufficient to consider only the sequence θ_j, $2 \le j \le n-1$. For $0 \le x \le 1$ and $2 \le j \le n-2$, we consider the function

$$g_j(x) = h_{j+1}(x) - h_j(x) = \frac{n}{n-j} B_{j,n-1}(x).$$

Then, $g_j(0) = g_j(1) = 0$ and g_j is increasing on $\left(0, \frac{j}{n-1}\right)$ and decreasing on $\left[\frac{j}{n-1}, 1\right)$. Since $\theta_j \in \left(0, \frac{j-1}{n-1}\right) \subset \left(0, \frac{j}{n-1}\right)$, then the sum $h_j + g_j$ is strictly increasing on $(0, \theta_j]$ and strictly decreasing on $\left[\frac{j}{n-1}, 1\right)$. Since the sum is continuously differentiable, there exist a and b such that $\theta_j < a < b < \frac{j}{n-1}$ and $h_j + g_j$ is increasing on $(0, a)$ and decreasing on $(b, 1)$. On the other hand $h_j + g_j = h_{j+1}$, so it has the unique maximum at θ_{j+1} and $\theta_{j+1} \in [a, b] \subset \left(\theta_j, \frac{j}{n-1}\right)$. In particular, $\theta_j < \theta_{j+1}$ for $2 \le j \le n-2$. □

The idea of the proof of the main result comes from Okolewski and Rychlik (2001), who used so-called Moriguti inequality obtained by Moriguti (1953).

Lemma 6 (Moriguti inequality) *Let $\Phi : [0,1] \to \mathbb{R}$ be a function of bounded variation continuous at both ends. Then, the inequality*

$$\int_0^1 x(t)d\Phi(t) \le \int_0^1 x(t)\overline{\varphi}(t)dt$$

holds for any nondecreasing function $x : [0,1] \to \mathbb{R}$ for which the integrals exist and are finite, where $\overline{\varphi}$ is the right-hand derivative of the greatest convex minorant $\overline{\Phi}$ of Φ. The equality holds if and only if the function x is constant in every interval where $\overline{\Phi}(t) < \min\{\Phi(t^-), \Phi(t^+)\}$, and if Φ is not continuous at some point t_0, then

(a) *if $\Phi(t_0^-) < \Phi(t_0^+)$, then x is right continuous;*
(b) *if $\Phi(t_0^-) > \Phi(t_0^+)$, then x is left continuous.*

Now we formulate and prove the main result of this section.

Theorem 2 *Let $n \ge 2$ and $p \in (0,1)$.*

(a) *Assume that $2 \le j \le n-1$. If $p \ge \theta_j$, then*

$$\overline{B}(j,n,p) = \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}}, \tag{20}$$

and the equality holds for the two-point distribution

$$\mathbb{P}\left(X = \mu - \sigma\sqrt{\tfrac{1-p}{p}}\right) = p = 1 - \mathbb{P}\left(X = \mu + \sigma\sqrt{\tfrac{p}{1-p}}\right). \tag{21}$$

Otherwise, if $p < \theta_j$, then

$$\overline{B}(j,n,p) = \overline{B}_j = \left(\frac{(1-F_{j:n}(p))^2}{p} + \int_p^{\theta_j} f_{j:n}^2(x)dx + \frac{(1-F_{j:n}(\theta_j))^2}{1-\theta_j}\right)^{1/2}. \tag{22}$$

The bound (22) is attained for the distribution function F given by

$$F(x) = \begin{cases} 0, & \text{for } \frac{x-\mu}{\sigma} < -\frac{1-F_{j:n}(p)}{p\overline{B}_j}, \\ p, & \text{for } -\frac{1-F_{j:n}(p)}{p\overline{B}_j} \le \frac{x-\mu}{\sigma} < \frac{f_{j:n}(p)}{\overline{B}_j}, \\ f_{j:n}^{-1}\left(\overline{B}_j\frac{x-\mu}{\sigma}\right), & \text{for } \frac{f_{j:n}(p)}{\overline{B}_j} \le \frac{x-\mu}{\sigma} < \frac{1-F_{j:n}(\theta_j)}{(1-\theta_j)\overline{B}_j}, \\ 1, & \text{for } \frac{x-\mu}{\sigma} \ge \frac{1-F_{j:n}(\theta_j)}{(1-\theta_j)\overline{B}_j}. \end{cases} \tag{23}$$

(b) *For $j = 1$ and all $p \in (0,1)$ we have*

$$\overline{B}(1,n,p)=\frac{1-F_{1:n}(p)}{\sqrt{p(1-p)}}=(1-p)^{n-1}\sqrt{\frac{1-p}{p}}$$

and the equality holds for the two-point distribution (21).

(c) *For $j=n$ and all $p\in(0,1)$ we have*

$$\overline{B}(n,n,p)=\overline{B}_n=\left(\frac{(1-p^n)^2}{p}+\frac{n^2}{2n-1}(1-p^{2n-1})\right)^{1/2}. \tag{24}$$

The bound (24) is attained by the distribution

$$F(x)=\begin{cases}0, & \text{for } \frac{x-\mu}{\sigma}<-\frac{1-p^n}{p\overline{B}_n},\\ p, & \text{for } -\frac{1-p^n}{p\overline{B}_n}\le\frac{x-\mu}{\sigma}<\frac{np^{n-1}}{\overline{B}_n},\\ \left(\overline{B}_n\frac{x-\mu}{\sigma n}\right)^{\frac{1}{n-1}}, & \text{for } \frac{np^{n-1}}{\overline{B}_n}\le\frac{x-\mu}{\sigma}<\frac{n}{\overline{B}_n},\\ 1, & \text{for } \frac{x-\mu}{\sigma}\ge\frac{n}{\overline{B}_n}.\end{cases}$$

Proof We start with the representations

$$\mathbb{E}X_{j:n}=\int_0^1 F^{\leftarrow}(u)dF_{j:n}(u)\quad\text{and}\quad F^{\leftarrow}(p)=\int_0^1 F^{\leftarrow}(u)d\mathbf{1}_{[p,1)}(u). \tag{25}$$

where $\mathbf{1}_A$ denotes the indicator function of a set $A\subset\mathbb{R}$. Using the expression for μ_F given by (2), we obtain

$$\mathbb{E}X_{j:n}-F^{\leftarrow}(p)=\int_0^1[F^{\leftarrow}(u)-\mu_F]\,dH_{j:n}(u),$$

where

$$H_{j:n}(u)=\begin{cases}F_{j:n}(u), & \text{for } u<p,\\ F_{j:n}(u)-1, & \text{for } u\ge p.\end{cases}$$

Obviously, $H_{j:n}$ depends also on p, but for simplicity, we will omit it in the notation.

(a) Assume that $2\le j\le n-1$. Then, $F_{j:n}$ is strictly increasing on $[0,1]$, convex on $\left(0,\frac{j-1}{n-1}\right)$ and concave on $\left(\frac{j-1}{n-1},1\right]$. Since $\theta_j<\frac{j-1}{n-1}$, then $F_{j:n}$ is convex on $(0,\theta_j)$ and on $(\theta_j,1)$ its graph lies entirely above the straight line through the points $(\theta_j,F_{j:n}(\theta_j))$ and $(1,1)$. So the function $H_{j:n}$ increases from 0 at 0 to the left-hand limit $H_{j:n}(p^-)=F_{j:n}(p)>0$, then it jumps down to the value $H_{j:n}(p)=F_{j:n}(p)-1<0$ at p and again it increases to 0 at 1. Moreover, since $p\ge\theta_j$, $H_{j:n}$ is either concave or convex-concave on $(p,1]$, but its graph lies entirely above the line through $(p,F_{j:n}(p)-1)$ and $(1,0)$. Therefore, its greatest

convex minorant is a two-piece broken line with knots at $(0, 0)$, $(p, F_{j:n}(p) - 1)$ and $(1, 0)$ given by

$$\overline{H}_{j:n}(u) = \begin{cases} \frac{F_{j:n}(p)-1}{p} u, & \text{for } u < p, \\ \frac{1-F_{j:n}(p)}{1-p}(u-1), & \text{for } u \geq p, \end{cases} \tag{26}$$

with the right-hand derivative

$$\overline{H}'_{j:n}(u) = \begin{cases} \frac{F_{j:n}(p)-1}{p}, & \text{for } u < p, \\ \frac{1-F_{j:n}(p)}{1-p}, & \text{for } u \geq p. \end{cases}$$

Obviously

$$\|\overline{H}'_{j:n}\| = \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}},$$

where $\|\cdot\|$ denotes the usual $\mathcal{L}^2$ norm with respect to Lebesgue measure on $[0, 1]$. By Lemma 6 and Schwarz inequality for each $F \in \mathcal{F}$, we get

$$\mathbb{E}X_{j:n} - F^{\leftarrow}(p) \leq \int_0^1 [F^{\leftarrow}(u) - \mu]\overline{H}'_{j:n}(u)du \leq \sigma_F \|\overline{H}'_{j:n}\|, \tag{27}$$

which proves that $\overline{B}(j, n, p) \leq \frac{1-F_{j:n}(p)}{\sqrt{p(1-p)}}$. The equality in latter inequality in (27) holds if and only if

$$F^{\leftarrow}(u) - \mu = \sigma \frac{\overline{H}'_{j:n}(u)}{\|\overline{H}'_{j:n}\|}, \quad u \in (0, 1), \quad u \neq p.$$

Combining this with Lemma 6 we see that the equality holds in both inequalities in (27) if and only if

$$\frac{F^{\leftarrow}(u) - \mu}{\sigma} = \begin{cases} -\sqrt{\frac{1-p}{p}}, & \text{for } u \leq p, \\ \sqrt{\frac{p}{1-p}}, & \text{for } u > p. \end{cases}$$

This implies equations (20) and (21). If $p < \theta_j$, then $H_{j:n}$ is convex on (p, θ_j) and on $(\theta_j, 1)$ its graph lies entirely above the line through $(\theta_j, F_{j:n}(\theta_j) - 1)$ and $(1, 0)$. In this case, the greatest convex minorant is given by

$$\overline{H}_{j:n}(u) = \begin{cases} \frac{F_{j:n}(p)-1}{p}\, u, & \text{for } 0 \le u \le p, \\ H_{j:n}(u), & \text{for } p < u \le \theta_j, \\ \frac{1-F_{j:n}(\theta_j)}{1-\theta_j}\,(u-1), & \text{for } \theta_j < u \le 1. \end{cases} \tag{28}$$

Now, analogous reasoning as in the case $p \ge \theta_j$ proves (22) and (23).

(b) For $j = 1$, $F_{1:n}(u) = 1 - (1-u)^n$ is concave on $[0, 1]$. Therefore, the greatest convex minorant $\overline{H}_{1:n}$ is given by (26) with $j = 1$. The rest of the proof is the same as in case (a) with $p \ge \theta_j$.

(c) For $j = n$, we have $\theta_n = 1$ and the greatest convex minorant $\overline{H}_{n:n}$ is given by (28) without the third case. So the proof of claim (c) follows the proof of (22) and (23).

□

Remark 2 Using the comparison of θ_{j+1}, ξ_j and j/n (see Corollary 4 in the next section) for $p = \frac{j}{n}$, we obtain the correct version of the result of Okolewski and Rychlik (2001). The reason of their error is the flaw in the proof of their Proposition 1. Namely, instead of the second equality in (25), they used the equality

$$F^{\rightarrow}(p) = \int_0^1 F^{\rightarrow}(u) d\mathbf{1}_{[p,1)}(u)$$

which is not true. The problem is that the above Riemann-Stieltjes integral does not exist, as both functions $F^{\rightarrow}$ and $\mathbf{1}_{[p,1)}$ are right-continuous at p. On the contrary, the integral used in the expression for $F^{\leftarrow}(p)$ in (25) does exist since $F^{\leftarrow}$ is left-continuous and $\mathbf{1}_{[p,1)}$ is right-continuous (see Rudin (1976), Exercise 3, p. 138). Moreover, Okolewski and Rychlik (2001) claimed that both bounds in (15) are attained for the same distribution (21) which is obviously impossible.

Remark 3 Assume that F is a continuous and strictly increasing distribution function on some interval $(a, b) \subset \mathbb{R}$, such that $F(a) = 0$ and $F(b) = 1$. Then, we have $F^{\leftarrow} = F^{\rightarrow} = F^{-1}$ which is just the usual inverse function of F. Then, obviously the middle inequality in (14) becomes equality. Therefore, it is of interest to study whether the upper and lower bounds in (14) may be improved in this case. However, analysing carefully the proof of Theorem 2, we may easily construct the sequence of continuous and strictly increasing quantile functions for which the bounds in (14) are attained in the limit.

From Theorem 2 using the identities (7) and (19), we obtain the values of lower bounds $\underline{B}(j, n, p)$ and of the function $r_{n,p}$ defined by (8).

Corollary 1 *For $j = 1$ and $p \in (0, 1)$, we have*

$$r_{n,p}(1) = \frac{1 - F_{1:n}(p)}{\sqrt{p(1-p)}} - \left(\frac{(F_{1:n}(p))^2}{1-p} + \int_0^p f_{1:n}^2(x) dx \right)^{\frac{1}{2}}$$

and analogously $r_{n,p}(n) = -r_{n,1-p}(1)$. *For* $2 \le j \le n-1$ *we have*

(a) *if* $p < \theta_j$, *then*

$$r_{n,p}(j) = \overline{B}_j - \frac{F_{j:n}(p)}{\sqrt{p(1-p)}}$$

with $\overline{B}_j$ defined in (22);

(b) *if* $\theta_j \le p \le \xi_j$, *then*

$$r_{n,p}(j) = \frac{1 - 2F_{j:n}(p)}{\sqrt{p(1-p)}};$$

(c) *if* $p > \xi_j$, *then*

$$r_{n,p}(j) = \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}} - \left(\frac{(F_{j:n}(\xi_j))^2}{\xi_j} + \int_{\xi_j}^{p} f_{j:n}^2(x)dx + \frac{(F_{j:n}(p))^2}{1-p} \right)^{\frac{1}{2}}.$$

We close this section with two properties of the bounds which have important implications for the auxiliary function $r_{n,p}$. The proofs of these results are given in Appendix 1. The next corollary says that the inequalities (9) are proper. Due to the symmetry (7), it suffices to consider the upper bounds only.

Corollary 2 $\overline{B}(j,n,p)$ *is a strictly increasing function of* j, *namely*,

$$\overline{B}(j,n,p) < \overline{B}(j+1,n,p), \quad 1 \le j \le n-1, \tag{29}$$

and strictly decreasing function of $p \in (0,1)$, *i.e.*

$$\overline{B}(j,n,q) < \overline{B}(j,n,p), \quad 0 < p < q < 1, \tag{30}$$

The last useful property of the bounds is given in the following lemma.

Lemma 7 *If* $p \in (\xi_j, 1)$ *with* $1 \le j \le n-1$

$$\underline{B}(j,n,p) < -\frac{F_{j:n}(p)}{\sqrt{p(1-p)}},$$

or equivalently if $p \in (0, \theta_j)$ *with* $2 \le j \le n$

$$\overline{B}(j,n,p) > \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}}.$$

In consequence, for all $1 \le j \le n$ *and* $p \in (0,1)$ *we have*

$$\underline{B}(j,n,p) \le -\frac{F_{j:n}(p)}{\sqrt{p(1-p)}} \quad \text{and} \quad \overline{B}(j,n,p) \ge \frac{1-F_{j:n}(p)}{\sqrt{p(1-p)}}. \tag{31}$$

4 The choice of the optimal order statistic

In this section, we prove our main result which says that for $n \ge 2$ the optimal order statistic to estimate any quantile of order p is either $X_{\lfloor np \rfloor:n}$ or $X_{\lceil np \rceil:n}$ or $X_{\lceil np \rceil+1:n}$ depending on p. First, we define and study auxiliary functions and numbers, which are next used in the solution to the problem of the optimal choice of order statistic in quantile estimation.

Note that in the trivial case $n = 2$, we have $a_{1,2} = \frac{1}{2}$, so $k(n,p) = 1$ for $p \in (0, \frac{1}{2})$ and $k(n,p) = 2$ for $p \in (\frac{1}{2}, 1)$. In other words, in this case, the optimal estimator of x_p, with $p \ne \frac{1}{2}$ is $X_{\lceil np \rceil:n}$, i.e. $X_{1:2}$ if $p \in (0, \frac{1}{2})$ and $X_{2:2}$ for $p \in (\frac{1}{2}, 1)$. So for the rest of this section we assume that $n \ge 3$.

4.1 Auxiliary numbers and functions

For fixed $n \ge 3$ and $p \in (0,1)$, we define the function $t_{n,p}$ as

$$t_{n,p}(j) = \frac{1-2F_{j:n}(p)}{\sqrt{p(1-p)}}, \quad 1 \le j \le n.$$

Obviously, $t_{n,p}$ is a continuous function of $p \in (0,1)$ and, since $F_{j:n}(p) > F_{j+1:n}(p)$, then $t_{n,p}$ is a strictly increasing function of j. A simple consequence of Corollary 1 and Lemma 7 is the following comparison of the values of two functions $r_{n,p}$ and $t_{n,p}$.

Corollary 3

(a) *For all $p \in (0,1)$, we have $r_{n,p}(1) < t_{n,p}(1)$ and $r_{n,p}(n) > t_{n,p}(n)$.*
(b) *For $2 \le j \le n-1$, we have $r_{n,p}(j) \le t_{n,p}(j)$ if $p \in (\theta_j, 1)$ and $r_{n,p}(j) \ge t_{n,p}(j)$ if $p \in (0, \xi_j)$.*

Next, for fixed $n \ge 2$ and $1 \le j \le n$, we define $p_j = p_j(n)$ as the unique solution to the equation

$$F_{j:n}(p) = \frac{1}{2}.$$

The uniqueness of p_j follows easily from the fact that $F_{j:n}$ is continuous and strictly increasing with $F_{j:n}(0) = 0$ and $F_{j:n}(1) = 1$. Since $F_{j:n} \ge F_{j+1:n}$ on $[0,1]$, then is obvious that $p_1 < \ldots < p_n$. By (18), it is also obvious that

$$p_{n-j+1} = 1 - p_j. \tag{32}$$

The importance of these numbers lies in the fact that for fixed $1 \le j \le n$, we have $t_{n,p_j}(j) = 0$, as well as $t_{n,p}(j) > 0$ for $p \in (0, p_j)$ and $t_{n,p}(j) < 0$ for $p \in (p_j, 1)$. Note that p_j is just the median of the jth uniform order statistic $U_{j:n}$.

Recall that our problem is to find the numbers $a_{j,n}$, $1 \le j \le \left\lfloor \frac{n-1}{2} \right\rfloor$, defined after the statement of Lemma 1. We will show that for $3 \le j \le n-2$, the value of $a_{j,n}$ is equal to the number $q_j(n)$ defined as follows. For $1 \le j \le n-1$, let $q_j = q_j(n)$ be the unique solution to the equation

$$|1 - 2F_{j:n}(q)| = |1 - 2F_{j+1:n}(q)| \tag{33}$$

in the interval (p_j, p_{j+1}). To prove the uniqueness of q_j, we note that if

$$Q_{j,n}(p) = |1 - 2F_{j:n}(p)| - |1 - 2F_{j+1:n}(p)|$$

then the function $Q_{j,n}$ is equal to $-2B_{j+1:n}$ on the interval $[0, p_j)$, to $2B_{j+1:n}$ on $(p_{j+1}, 1]$ and for $p \in [p_j, p_{j+1}]$ it holds

$$Q_{j,n}(p) = 2(F_{j:n}(p) + F_{j+1:n}(p) - 1).$$

By the definition of p_j, $1 \le j \le n$, we obtain $Q_{j,n}(p_j) < 0 < Q_{j,n}(p_{j+1})$. Since $Q_{j,n}$ is continuous and strictly increasing on $[p_j, p_{j+1}]$, it follows that q_j is defined uniquely as the solution to equation $F_{j:n}(q_j) + F_{j+1:n}(q_j) = 1$. Also, from (18), we see that

$$q_{n-j} = 1 - q_j. \tag{34}$$

Now we need to study various properties of the defined numbers and their mutual relations. Our aim here is to show that in most of the cases we have $\theta_{j+1} < q_j < \xi_j$. The reason is that this inequality implies that the values of $s_{n,q_j}(j)$ and $s_{n,q_j}(j+1)$ have simple analytic forms and they may be compared easily. The proofs of next 2 lemmas are very technical, so they are given in Appendix 2.

First, we investigate mutual relations between p_j, q_j and $\frac{j}{n}$. Therefore, the lemma is helpful to establish the location of $a_{j,n}$ with respect to $\frac{j}{n}$.

Lemma 8 *For* $1 \le j \le n-1$ *we have*

(a) *if* $1 \le j < \frac{n}{2}$, *then*

$$p_j < \tfrac{j}{n} < q_j < p_{j+1}; \tag{35}$$

(b) *if* $\frac{n}{2} < j \le n-1$, *then*

$$p_j < q_j < \tfrac{j}{n} < p_{j+1}; \tag{36}$$

(c) *if* $j = \frac{n}{2}$, *then*

$$p_{n/2} < q_{n/2} = \tfrac{1}{2} < p_{n/2+1}.$$

The next lemma gives mutual relations between $\theta_j(n)$, $\xi_j(n)$ and $q_j(n)$ for various values of j and n.

Lemma 9

(a) *For* $1 \le j < \frac{n}{2}$ *we have* $\theta_{j+1} < \frac{j}{n}$.
(b) *For* $3 \le j < \left\lfloor \frac{n-1}{2} \right\rfloor$ *and* $n \ge 7$ *or* $j = 2$ *and* $n = 5$ *we have* $q_j(n) < \xi_j(n)$.
(c) *For* $j = 2$ *and* $n = 6$ *we have* $\frac{2}{6} < \xi_2(6) < q_2(6) < \xi_3(6)$.
(d) *For* $j = 2$ *and* $n \ge 7$ *we have* $\xi_2(n) < \frac{2}{n} < q_2(n)$.
(e) *For* $j = 1$ *and* $n \ge 3$ *we have* $\xi_2(n) > q_1(n)$.

For the ease of the reference, all the conclusions of Lemmas 8 and 9 are summarized below.

Corollary 4

(a) $\left[\frac{1}{n}, q_1\right] \subset (\theta_2, \xi_2) \cap (p_1, p_2)$ *for* $n \ge 3$;
(b) $\frac{2}{6} \in \left(\theta_3, \xi_2\right)$ *and* $q_2 \in \left(\xi_2, \xi_3\right) \subset \left(\theta_3, \xi_3\right)$ *for* $n = 6$;
(c) $\left[\frac{2}{n}, q_2\right] \subset (\xi_2, \frac{1}{2}) \cap (\theta_3, \xi_3) \cap (p_2, p_3)$ *for* $n \ge 7$;
(d) $q_j \in \left(\theta_{j+1}, \xi_j\right) \cap (p_j, p_{j+1})$ *for* $n = 5$ *and* $j = 2$ *or* $n \ge 7$ *and* $3 \le j \le \left\lfloor \frac{n-1}{2} \right\rfloor$.

4.2 The solution to the problem of optimal choice

Now we are ready to solve the problem of determination of $a_{j,n}$. The main idea of the solution is as follows. By Corollary 3, we have

$$s_{n,p}(j) \ge \sqrt{p(1-p)}\left|t_{n,p}(j)\right|$$

for all $p \in (0, 1)$ and the equality holds if and only if $p \in (\theta_j, \xi_j)$. So the problem of the minimization of $s_{n,p}$ amounts in fact to the analysis of $t_{n,p}$. First, we consider the case $j \ge 3$, which appears to be the simplest one.

Theorem 3 *For* $n \ge 7$ *and* $3 \le j \le \left\lfloor \frac{n-1}{2} \right\rfloor$, *we have* $k(n, q_j) = \{j, j+1\}$, *so* $a_{j,n} = q_j(n)$. *Moreover*, $k(5, q_2) = \{2, 3\}$, *so* $a_{2,5} = q_2(5)$.

Proof Since $t_{n,p}$ is a strictly increasing function of j, it suffices to prove that $s_{n,q_j}(j) = s_{n,q_j}(j+1)$ taking into account the conclusion of Corollary 4(d). Since

$q_j \in (p_j, p_{j+1})$, we first infer that $t_{n,q_j}(j) < t_{n,p_j}(j) = 0$ and $t_{n,q_j}(j+1) > t_{n,p_{j+1}}(j+1) = 0$. Next, since $q_j \in (\theta_j, \xi_j) \cap (\theta_{j+1}, \xi_{j+1})$, then

$$s_{n,q_j}(j) = -\sqrt{p(1-p)}t_{n,q_j}(j) = 2F_{k:n}(q_j) - 1,$$

and

$$s_{n,q_j}(j+1) = \sqrt{p(1-p)}t_{n,q_j}(j+1) = 1 - 2F_{j+1:n}(q_j).$$

But by the definition of q_j (see (33)) the right-hand sides of both equations are equal. □

Now we turn to the more complicated cases $j = 1$ and $j = 2$. The proof of the next theorem follows analogous ideas, but it is more technical, so it is presented in Appendix 3.

Theorem 4

(a) *For $n \geq 3$, we have $k\left(n, \frac{1}{n}\right) = 1$ and $k(n, q_1) = 2$, so $a_{1,n} \in \left(\frac{1}{n}, q_1\right)$ is the unique solution to equation $s_{n,p}(1) = s_{n,p}(2)$ with respect to p.*

(b) *For $n \geq 6$, we have $k\left(n, \frac{2}{n}\right) = 2$ and $k(n, q_2) = 3$, so $a_{2,n} \in \left(\frac{2}{n}, q_2\right)$ is the unique solution to equation $s_{n,p}(2) = s_{n,p}(3)$ with respect to p.*

In the statement and the proof of our main result, we use the following notation. For $n \geq 3$ we denote

$$M = M(n) = \left\lfloor \tfrac{n-1}{2} \right\rfloor, \quad N = N(n) = \left\lceil \tfrac{n+1}{2} \right\rceil.$$

Then, $M < \frac{n}{2} < N$ and $N = M + 1$ if n is odd and $M = \frac{n}{2} - 1$, $N = \frac{n}{2} + 1$ if n is even. We also define the subsets of the interval (0, 1)

$$\mathcal{P}_n = \bigcup_{j=0}^{M-1} \left(a_{j,n}, \tfrac{j+1}{n}\right) \cup \left(a_{M,n}, \tfrac{1}{2}\right), \quad \mathcal{Q}_n = \bigcup_{j=1}^{M} \left(\tfrac{j}{n}, a_{j,n}\right).$$

Moreover, let

$$\mathcal{P}'_n = \left(\tfrac{1}{2}, a_{N,n}\right) \cup \bigcup_{j=N}^{n-1} \left(\tfrac{j}{n}, a_{j+1,n}\right), \quad \mathcal{Q}'_n = \bigcup_{j=N}^{n-1} \left(a_{j,n}, \tfrac{j}{n}\right).$$

Taking into account the symmetry $a_{n-j,n} = 1 - a_{j,n}$ we easily obtain $p \in \mathcal{P}'_n$ if and only if $1 - p \in \mathcal{P}_n$ and analogous statement holds for $\mathcal{Q}_n$ and $\mathcal{Q}'_n$.

Theorem 5 *Let $n \geq 3$ and $p \neq \frac{1}{2}$. The optimal estimators of x_p are as follows*:

(a) *both* $X_{\lfloor np\rfloor:n}$ *and* $X_{\lceil np\rceil:n}$ *if* $p \in \{a_{1,n}, \ldots, a_{M,n}\}$ *or both* $X_{\lceil np\rceil:n}$ *and* $X_{\lceil np\rceil+1:n}$ *if* $p \in \{a_{N,n}, \ldots, a_{n-1,n}\}$;
(b) $X_{\lfloor np\rfloor:n} = X_{\lceil np\rceil:n}$ *if* $p \in \left\{\frac{1}{n}, \ldots, \frac{M}{n}\right\}$ *and* $X_{\lfloor np\rfloor+1:n} = X_{\lceil np\rceil+1:n}$ *if* $p \in \left\{\frac{N}{n}, \ldots, \frac{n-1}{n}\right\}$;
(c) $X_{\lceil np\rceil:n}$ *if* $p \in \mathcal{P}_n \cup \mathcal{P}'_n$, $X_{\lfloor np\rfloor:n}$ *if* $p \in \mathcal{Q}_n$ *and* $X_{\lceil np\rceil+1:n}$ *if* $p \in \mathcal{Q}'_n$.

Proof For $p \in (0, \frac{1}{2})$ by Theorems 3 and 4, we have

$$k(n,p) = \begin{cases} \{\lfloor np\rfloor, \lceil np\rceil\}, & \text{if } p \in \{a_{j,n} : 1 \le j \le M\}, \\ \lceil np\rceil = \lfloor np\rfloor, & \text{if } p \in \{\frac{j}{n} : 1 \le j \le M\}, \\ \lceil np\rceil, & \text{if } p \in \mathcal{P}_n, \\ \lfloor np\rfloor, & \text{if } p \in \mathcal{Q}_n. \end{cases}$$

Furthermore, by Lemma 2 and the symmetry $\lfloor n(1-p)\rfloor = n - \lceil np\rceil$, we easily obtain the values of $k(n,p)$ for $p \in \left(\frac{1}{2}, 1\right)$. By the definition of the function $k(n,p)$, we get the thesis of the theorem. □

Recall that the case $p = \frac{1}{2}$ has been considered in Remark 1. Moreover, in particular case $p = k/n$, we obtain the following corollary.

Corollary 5 *The optimal estimator of* $x_{k/n}$ *is either* $X_{k:n}$ *if* $1 \le k < \frac{n}{2}$ *or* $X_{k+1:n}$ *if* $\frac{n}{2} < k < n$. *If* $n \ge 2$ *is even and* $k = \frac{n}{2}$, *then both* $X_{n/2:n}$ *and* $X_{n/2+1:n}$ *are optimal.*

This is different than the usual choice of $X_{k:n}$ for all $1 \le k \le n$ as an estimator of $x_{k/n}$. The advantage of our approach is that it yields a symmetric quantile estimator. The symmetry property is very desirable in quantile estimation (see Hyndman and Fan 1996). However, it agrees with the traditional method to estimate the population median by its sample counterpart for samples of odd size n.

4.3 An equivalent approach: the minimization of maximum bias

In this section, our aim is to choose j which minimizes the maximum bias of the estimation, so we want to minimize the function $c_{n,p}$ (see (5)) with respect to j . Let $\ell(n,p)$ denote the set of all indices j which minimize the function $c_{n,p}$. The next theorem says that for $p \in (q_2, q_{n-2})$, it is minimized for exactly the same j as the function $s_{n,p}$ is.

Theorem 6 *For* $n \ge 7$ *and* $p \in (q_2, q_{n-2})$, *we have* $\ell(n,p) = k(n,p)$.

Numerical computations suggest that the theorem is true for all $n \ge 3$ and $p \in (0,1)$, but we could not find the formal proof for such statement. In the proof of the theorem, we use the next elementary lemma which follows easily from $\max(x,y) \le z$ iff $x \le z$ and $y \le z$.

Lemma 10 *Fix any real numbers a, b, c and d such that $a < c$ and $b > d$. Then, the inequality* $\max(a,b) \le \max(c,d)$ *holds if and only if* $b \le c$. *Moreover,* $\max(a,b) \ge \max(c,d)$ *iff* $b \ge c$ *and* $\max(a,b) = \max(c,d)$ *iff* $b = c$.

Proof of Theorem 6 Since n and p are fixed, then in the proof for brevity we write $\overline{B}_j$ and $\underline{B}_j$ instead of $\overline{B}(j,n,p)$ and $\underline{B}(j,n,p)$, respectively. For $1 \le j \le n-1$, we denote

$$d_{n,p}(j) = \overline{B}_{j+1} + \underline{B}_j.$$

By Corollary 2, the function $d_{n,p}$ is strictly increasing with respect to j. By Lemma 10 with $a = \overline{B}_j$, $b = -\underline{B}_j$, $c = \overline{B}_{j+1}$ and $d = -\underline{B}_{j+1}$, we have obviously

$$c_{n,p}(j) < c_{n,p}(j+1) \quad \Leftrightarrow \quad d_{n,p}(j) > 0,$$

for $1 \le j \le n-1$. If the former inequality is reversed, then so is the latter. Therefore $c_{n,p}(j) = c_{n,p}(j+1)$ if and only if $d_{n,p}(j) = 0$.

Using the symmetry (7), we conclude that $c_{n,1-p}(j) = c_{n,p}(n-j+1)$. Therefore, it suffices to consider the case $p \le \frac{1}{2}$. We consider two cases:

- $p \in (q_j, q_{j+1})$ for some $2 \le j \le \frac{n}{2}$;
- $p = q_j$ for some $3 \le j \le \frac{n}{2}$;

In the first case, we have $k(n,p) = j+1$ and $p \in (q_j, q_{j+1}) \subset [\theta_{j+1}, \xi_{j+1}]$. Recall that $Q_{j,n}(p) = 2(F_{j:n}(p) + F_{j+1:n}(p) - 1)$ for $p \in [p_j, p_{j+1}]$ and $Q_{j,n}(p)$ is negative for $p < q_j$ and positive for $p > q_j$. Then, combining the values of the upper and lower bounds given in Theorem 2 with (31), we obtain

$$d_{n,p}(j) \le -\frac{Q_{j,n}(p)}{2\sqrt{p(1-p)}} \quad \text{and} \quad d_{n,p}(j+1) \ge -\frac{Q_{j+1,n}(p)}{2\sqrt{p(1-p)}}.$$

This implies that $d_{n,p}(j) < 0$ as $p > q_j$, and $d_{n,p}(j+1) > 0$ as $p < q_{j+1}$. Therefore, $d_{\mathrm{n,p}}(i)$ is negative for $i \le j$ and positive for $i \ge j+1$. Thus, $c_{n,p}(i)$ is strictly decreasing for $1 \le i \le j$ and strictly increasing for $j+1 \le i \le n$, and $\ell(n,p) = j+1 = k(n,p)$.

In the second case, we have $k(n,q_j) = \{j, j+1\}$ and $p = q_j \in [\theta_{j+1}, \xi_j]$. Note that for all p in this interval, we have $2\sqrt{p(1-p)}d_{n,p}(j) = -Q_{j,n}(p)$, so $d_{n,q_j}(j) = 0$ and $d_{n,q_j}(i)$ is negative for $i < j$ and positive for $i > j$. Thus, $c_{n,q_j}(j) = c_{n,q_j}(j+1)$, $c_{n,q_j}(i)$ is strictly decreasing for $i \le j$, and strictly increasing for $i \ge j+1$. So c_{n,q_j} is minimized simultaneously at j and $j+1$ and $\ell(n,q_j) = \{j, j+1\} = k(n,q_j)$. □

5 Numerical examples

In this section, we present examples of numerical computations which illustrate the results of the paper. Firstly, we note that all the quantities θ_k, ξ_k, p_k and q_k are hard to find explicitly, but their approximate values can be easily computed numerically.

Indeed, they are the roots of polynomial equations involving Bernstein polynomials, see (38), (19) and (43).

For instance, in Table 1, we present the values of p_k, k/n, q_k, θ_k and ξ_k for $n = 15$ and $1 \le k \le 8$. The corresponding values for $9 \le k \le 15$ can be obtained by respective symmetries of p_k, q_k and (19). These results confirm most of the conclusions of Lemmas 5, 8 and 9 (a) and (b).

Figure 1 shows the values of $s_{15,p}(j)$, $1 \le j \le 15$ for $p = 0.05, 0.1, \ldots, 0.5$. The graphs confirm the conclusions of Theorem 5, namely $k(15, 0.05) = 1$, $k(15, 0.1) = 2$, $k(15, 0.2) = 3$, $k(15, 0.3) = 5$, $k(15, 0.4) = 6$ and $k(15, 0.5) = 8$. More explicitly, for instance, the optimal estimator of $x_{0.4}$ based on the sample of size 15 is $X_{6:15}$.

Finally, Fig. 2 shows the comparison of values of $s_{n,p}(j), 1 \le j \le 15$ for $p = 3/15$, $p = q_3(15) \approx 0.2062$ and $p = 0.21$. The figure illustrates the fact that $k(15, 0.2) = 3$, $k(15, q_3) = \{3, 4\}$ and $k(15, 0.21) = 4$.

6 Summary and discussion

In the paper, we have presented completely new approach to the problem of the choice of the optimal single order statistic which estimates unknown quantile x_p of a fixed order $p \in (0, 1)$ based on the random sample of given size n. The novelty is the proposal of the new criterion of optimality based on the determination of sharp nonparametric bounds on the bias of the estimation of x_{p} by $X_{j:n}$. The main result obtained using this approach (Theorem 5) says that, contrary to traditional usage of $X_{\lceil np \rceil:n}$ for any $p \in (0, 1)$, for $p \in \mathcal{Q}_n$ and $p \in \mathcal{Q}'_n$ we should use $X_{\lfloor np \rfloor:n}$ and $X_{\lceil np \rceil+1:n}$, respectively, as the estimate of x_p. Since the length of each interval component of $\mathcal{Q}_n$ converges to 0 as n becomes larger, it may seem that the effort put in our approach is worthless. However, the total length of the set $\mathcal{Q}_n \cup \mathcal{Q}'_n$ is

$$S_n = 2 \sum_{k=1}^{N} \left(a_{k,n} - \frac{k}{n} \right).$$

It is hard to analyse the limit behaviour of S_n analytically, but the numerical computations show that the values of S_n increase as n increases, see Table 2. Some approximate values presented in the table show that the proportion of quantile orders for

Table 1 The values of p_k, k/n, q_k, θ_k and ξ_k for $n = 15$ and $1 \le k \le 8$

k	1	2	3	4	5	6	7	8
p_k	0.0451	0.1093	0.1743	0.2393	0.3045	0.3696	0.4348	0.5
k/n	0.0667	0.1333	0.2	0.2667	0.3333	0.4	0.4667	0.5333
q_k	0.0749	0.1407	0.2062	0.2715	0.3368	0.4021	0.4674	0.5326
θ_k	0	0.0051	0.0336	0.0779	0.1321	0.1934	0.2605	0.3326
ξ_k	0	0.1256	0.2332	0.3314	0.4229	0.5091	0.5905	0.6674

Table 2 Some approximate values of S_n

n	10	20	50	100
S_n	0.0587	0.0694	0.0769	0.0798

Table 3 Approximate values of $\sigma^2_{15}(j)$ for $1 \le j \le 8$

j	1	2	3	4	5	6	7	8
$\sigma^2_{15}(j)$	15	3.1235	1.8951	1.4334	1.2049	1.0818	1.0193	1

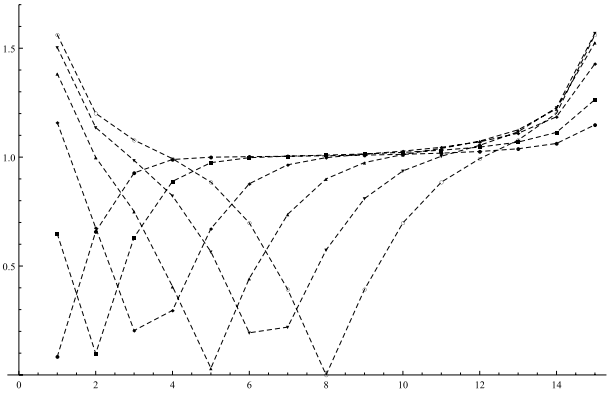

Fig. 1 Graphs of $s_{15,p}(j), 1 \le j \le 15$ for $p = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$

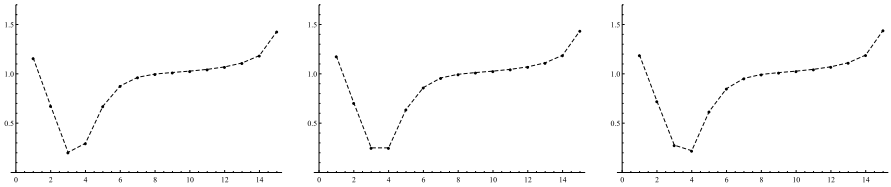

Fig. 2 Graphs of $s_{15,p}(j), 1 \le j \le 15$ for $p = 3/15$ (left), $p = q_3(15)$ (middle) and $p = 0.21$ (right)

which our approach is better is in the range of 6–7% for n from 10 to 100, so it should not be neglected. In our forthcoming papers, we use the same approach to find the optimal linear combinations of two neighbouring order statistics as quantile estimators. This class of L-statistics is much more often used in statistical packages, see e.g. Hyndman and Fan (1996).

The last problem we shortly discuss is the efficiency of proposed estimators. We have optimized the bias of order statistic as an estimator of a quantile only, not looking at the mean square error of the estimate. Let x_p denote any fixed value of quantile of order p. Obviously, the mean square error of $X_{j:n}$ as an estimator of x_p is

$$MSE_F(X_{j:n}, x_p) = Var(X_{j:n}) + \left(\mathbb{E}X_{j:n} - x_p\right)^2.$$

Furthermore, let

$$\sigma_n^2(j) = \sup_{F\in\mathcal{F}} \frac{Var(X_{j:n})}{\sigma_F^2}, \quad 1 \le j \le n,$$

denote the optimal upper bound on the variance of $X_{j:n}$ measured in σ_F units. The values of $\sigma_n^2(j)$ were established by Papadatos (1995). The values of the bounds $\sigma_n^2(j)$ are symmetric in the sense that $\sigma_n^2(n-j+1) = \sigma_n^2(j)$ and their approximate values for $n = 15$ and $1 \le j \le 8$ are given in Table 3.

Using the above notation, we obtain

$$\frac{MSE_F(X_{j:n})}{\sigma_F^2} \le \sigma_n^2(j) + \left(c_{n,p}(j)\right)^2 := w_{n,p}(j), \quad 1 \le j \le n,$$

for all distributions $F \in \mathcal{F}$. Obviously, this bound need not be sharp. In order to minimize possible mean square error, we should minimize the right-hand side of the above inequality with respect to j. Assuming without loss of generality that $p \in \left(0, \frac{1}{2}\right)$ it suffices to consider $1 \le j \le \left\lfloor \frac{n+1}{2} \right\rfloor$. Our Theorem 5 says that for the most of the quantile orders $p \in \left(0, \frac{1}{2}\right)$, it is best to estimate x_p by $X_{\lfloor np \rfloor:n}$, but for $p \in \mathcal{Q}_n$, we should switch to $X_{\lfloor np \rfloor:n}$. However, this makes the second term in $w_{n,p}(j)$ smaller. Unfortunately, this increases the first component according to the results presented in Table 3. This is an example of very well-known phenomenon called bias–variance trade-off, which is the conflict which we encounter while trying to minimize simultaneously the variance and the bias of an estimator. However, this implies that if $p = a_{j,n}$ for some $1 \le j < \frac{n}{2}$, then it is better to choose $j = \lceil na_{j,n} \rceil$ than $j = \lfloor na_{j,n} \rfloor$. This is an improvement of Theorem 5(a), which says that for $p = a_{j,n}$ both $X_{j:n}$ and $X_{j+1:n}$ are equally good.

Appendix 1: The proofs of corollary 2 and lemma 7

Proof of corollary 2 For $p \ge \theta_{j+1}$, inequality (29) is obvious, since $F_{j:n} > F_{j+1:n}$ on $(0, 1)$. For $p \in (0, \theta_{j+1})$, we use the first inequality in (9) and the fact that the bounds $\overline{B}(j, n, p)$ and $\overline{B}(j+1, n, p)$ cannot be equal since they are attained for different distributions. This proves (29).

To prove (30), first, we fix $1 \le j \le n-1$ and assume that $\theta_j < p < q < 1$. Then, (12) holds but $\overline{B}(j, n, q)$ and $\overline{B}(j, n, p)$ are attained for different distributions so they cannot be equal. For $2 \le j \le n$ and $p \in (0, \theta_j)$, it is obvious that the first two terms in (22) decrease when p increases. Moreover,

$$\overline{B}(j, n, \theta_j^-) = \overline{B}(j, n, \theta_j^+) = \frac{1 - F_{j:n}(\theta_j)}{\sqrt{\theta_j(1-\theta_j)}}$$

for $2 \le j \le n-1$, which completes the proof. □

Proof of lemma 7 Recall that $\xi_1 = 0$. Consider the function

$$k_j(t) = \int_{\xi_j}^{t} f_{j:n}^2(x)dx + \xi_j f_{j:n}^2(\xi_j) - t\bar{h}_j^2(t), \quad 0 \le t \le 1,$$

where the function $\bar{h}_j$ is defined by (37). Since $\bar{h}_j(\theta_j) = f_{j:n}(\xi_j)$, we get $k_j(\xi_j) = 0$. Moreover

$$k_j'(t) = f_{j:n}^2(t) - \bar{h}_j^2(t) - 2t\bar{h}_j(t)\bar{h}_j'(t).$$

Substituting (39) for $\bar{h}_j'$, we obtain after elementary computations

$$k_j'(t) = (\bar{h}_j(t) - f_{j:n}(t))^2 \ge 0$$

for all $0 \le t \le 1$, and the equality holds only for $t = \xi_j$. So k_j is strictly increasing on [0, 1]. In particular for $p > \xi_j$, we have $k_j(p) > k_j(\xi_j) = 0$ or

$$\int_{\xi_j}^{p} f_{j:n}^2(x)dx + \xi_j f_{j:n}^2(\xi_j) > p\bar{h}_j^2(p) = \frac{(F_{j:n}(p))^2}{p}.$$

Combining this with (22), we get

$$\underline{B}^2(j,n,p) > \left(\frac{F_{j:n}(p)}{\sqrt{p(1-p)}}\right)^2.$$

Since $\underline{B}(j,n,p) < 0$, this completes the proof. □

Appendix 2: The proofs of the results of subsection 4.1

In the proofs, we use the following two properties of Bernstein polynomials.

Lemma 11 (VDP) *The number of zeros of any linear combination* $\sum_{i=0}^{n} \alpha_i B_{i,n}$ *of Bernstein polynomials in* (0, 1) *does not exceed the number of sign changes in the sequence* $\alpha_0, \alpha_1, \ldots, \alpha_n$ *of its coefficients after deletion of zeros. Moreover, the first and the last signs of the combination are the same as the signs of the first and the last, respectively, nonzero element of the sequence.*

VDP of Bernstein polynomials was proved by Schoenberg (1959) and Gajek and Rychlik (1998). In fact, this is a simple consequence of well-known Descartes rule of signs (see e.g. Komornik 2006). See Bieniek (2007) for far reaching generalization of VDP to some special cases of Meijer's G-functions.

Lemma 12 (Simmons' inequality) *For* $1 \le k < \frac{n}{2}$, *we have*

$$\sum_{i=0}^{k-1} B_{i,n}\left(\tfrac{k}{n}\right) > \sum_{i=k+1}^{n} B_{i,n}\left(\tfrac{k}{n}\right)$$

and the equality holds if and only if $k = \frac{n}{2}$.

The reader is referred to Perrin and Redside (2007) for the latest proof of Simmons' inequality and to references therein for its older proofs.

Proof of lemma 8 The inequality $p_j < q_j < p_{j+1}$ for $1 \le j \le n-1$ follows from the proof of the uniqueness of q_j. To prove that $p_j < \frac{j}{n} < p_{j+1}$, for $1 \le j \le n-1$ we use the inequalities (16). Therefore, by the definition of p_j and p_{j+1}, we obtain

$$F_{j:n}\left(\tfrac{j}{n}\right) > \tfrac{1}{2} = F_{j:n}(p_j) \quad \text{and} \quad F_{j+1:n}\left(\tfrac{j}{n}\right) < \tfrac{1}{2} = F_{j+1:n}(p_{j+1}).$$

Since $F_{j:n}$ and $F_{j+1:n}$ are strictly increasing, we get desired inequality.

Now assume that $1 \le j < \frac{n}{2}$. Since both $\frac{j}{n}$ and q_j are in the interval (p_j, p_{j+1}), then $\frac{j}{n} < q_j$ if and only if $Q_{j,n}\left(\frac{j}{n}\right) < 0$ or equivalently

$$F_{j:n}\left(\tfrac{j}{n}\right) < 1 - F_{j+1:n}\left(\tfrac{j}{n}\right).$$

But this is equivalent to Simmons' inequality, which proves (35). To prove (36), it is enough to combine (32) and (34) with (35). If $j = \frac{n}{2}$, then Simmons' inequality becomes the equality equivalent to

$$1 - F_{n/2:n}\left(\tfrac{1}{2}\right) = F_{n/2+1:n}\left(\tfrac{1}{2}\right).$$

Therefore, $Q_{n/2,n}\left(\frac{1}{2}\right) = 0$, so $q_{n/2} = \frac{1}{2}$. □

For the next two proofs, we need another auxiliary functions. For $2 \le j \le n-1$, we define the h_j and $\bar{h}_j$ as

$$h_j(x) = \frac{1 - F_{j:n}(x)}{1 - x}, \quad 0 \le x < 1,$$

with $h_j(1) = h_j(1^-) = 0$, and

$$\bar{h}_j(x) = \frac{F_{j:n}(x)}{x}, \quad 0 < x \le 1 \tag{37}$$

with $\bar{h}(0) = \bar{h}(0^+) = 0$. Then $\bar{h}_j(x) = h_{n-j+1}(1-x)$ and

$$h_j'(x) = \frac{1}{(1-x)^2}\left[1 - F_{j:n}(x) - (1-x) f_{j:n}(x)\right].$$

Expanding $F_{j:n}$ as the sum of Bernstein polynomials we have

$$h_j'(x) = \frac{1}{(1-x)^2}\left[\sum_{i=0}^{j-2} B_{i,n}(x) - (n-j)B_{j-1,n}(x)\right]. \tag{38}$$

Applying VDP to the expression inside the brackets, we see that it is first positive, then negative ($+-$, for short), so equation (17) defining θ_j has exactly one solution. By VDP, we see that h_j' is also $+-$. Therefore, h_j is strictly increasing on $(0, \theta_j)$ from 1 at $x = 0$ to $h(\theta_j) > 1$ and then strictly decreasing on $(\theta_j, 1)$ to 0 at $x = 1$. Analogously,

$$\bar{h}_j'(x) = \frac{1}{x}\big(f_{j:n}(t) - \bar{h}_j(t)\big) = \frac{1}{x^2}\left[(j-1)B_{j,n}(x) - \sum_{i=j+1}^{n} B_{i,n}(x)\right]. \tag{39}$$

and we infer that $\bar{h}_j$ is strictly increasing on $(0, \xi_j)$ from 0 at $x = 0$, and strictly decreasing on $(\xi_j, 1)$ from $\bar{h}_j(\xi_j) > 1$ to 1 at $x = 1$. By these monotonicity properties of h_j and $\bar{h}_j$, we infer that for $2 \le j \le n-1$

$$\theta_j(n) > p \quad \text{if and only if} \quad h_j'(p) < 0, \tag{40}$$

and

$$\xi_j(n) > p \quad \text{if and only if} \quad \bar{h}_j'(p) > 0. \tag{41}$$

Proof of lemma 9(a) By (40), we have to check that $h_{j+1}'\left(\frac{j}{n}\right) < 0$. By (38), this is equivalent to

$$\sum_{i=0}^{j} B_{i,n}\left(\tfrac{j}{n}\right) < (n-j)B_{j,n}\left(\tfrac{j}{n}\right). \tag{42}$$

However, as explained in Sect. 3, the mode of $Y \sim \mathcal{B}\left(n, \frac{j}{n}\right)$ is equal to j, so $B_{i,n}\left(\frac{j}{n}\right) < B_{j,n}\left(\frac{j}{n}\right)$ for $0 \le i < j$. Now if $1 \le j < \frac{n}{2}$ then $j+1 \le n-j$. Combining this with the last inequality, we easily obtain (42), which completes the proof. □

Proof of lemma 9(b) By (41), we need to show that $\bar{h}_j'(q_j) > 0$. By the definition of q_j, we have

$$\sum_{i=j+1}^{n} B_{i,n}(q_j) = \sum_{i=0}^{j-1} B_{i,n}(q_j). \tag{43}$$

Using (39), we obtain for $2 \le j \le n-1$

$$\bar{h}_j'(q_j) = \frac{1}{q_j^2}\left[(j-1)B_{j,n}(q_j) - \sum_{i=0}^{j-1} B_{i,n}(q_j)\right].$$

For $2 \le j \le n-1$, we define the function

$$g_{j,n}(p) = (j-1)B_{j,n}(p) - \sum_{i=0}^{j-1} B_{i,n}(p).$$

Then, $\bar{h}_j'(q_j) = \frac{1}{q_j^2} g_{j,n}(q_j)$. By VDP, $g_{j,n}$ is first negative and then positive on (0, 1). Therefore, since by Lemma 8 we have $\frac{j}{n} < q_j$ for for $1 \le j < \frac{n}{2}$, it suffices to prove that

$$g_{j,n}\left(\frac{j}{n}\right) > 0 \tag{44}$$

for j satisfying the assumptions in part (b).

If $n = 5$ and $j = 2$, then $g_{2,5}(\frac{2}{5}) = \frac{3^3}{5^5} > 0$. Since $q_2(5) > \frac{2}{5}$ we also have $g_{2,5}(q_2) > 0$ and $\bar{h}_2'(q_2) > 0$. For $3 \le j < \frac{n}{2}$, the inequality (44) follows from two properties:

(i) $g_{3,n}\left(\frac{3}{n}\right) > 0$ for $n \ge 4$;

(ii) $g_{j,n}\left(\frac{j}{n}\right) < g_{j+1,n}\left(\frac{j+1}{n}\right)$ for $2 \le j \le n-2$.

To prove property (i), it suffices to study the expansion

$$g_{3,n}\left(\frac{3}{n}\right) = \frac{1}{n^3}\left(1 - \frac{3}{n}\right)^{n-3} c(n)$$

where $c(n) = 23n^3 - 60n^2 - 9n + 54$. It is elementary to show that $c(6) > 0$ and $c(n)$ is increasing for $n \ge 6$, which implies (i).

For the proof of (ii), first, we need to study monotonicity properties of $g_{j,n}$. By VDP, $g_{j,n}$ is first negative and then positive on (0, 1). Moreover, we have $g_{j,n}(0) = -1$, $g_{j,n}(1) = 0$ and using an elementary relation

$$B_{k,n}'(p) = n(B_{k-1,n-1}(p) - B_{k,n-1}(p)),$$

(with the convention that $B_{-1,n} \equiv 0$) we obtain

$$g_{j,n}'(p) = n(jB_{j-1,n-1}(p) - (j-1)B_{j,n-1}(p)).$$

Therefore, by VDP, the derivative $g_{j,n}'$ is $+-$. Furthermore, for $p \in (0,1)$, simple computations show that

$$g_{j,n}'(p) = 0 \Leftrightarrow p = \frac{j^2}{(n+1)j - n} := b_{j,n}.$$

So $g_{j,n}$ is increasing on the interval $\left(0, b_{j,n}\right)$ from -1 to $g_{j,n}(b_{j,n}) > 0$, and then decreasing to 0 at 1. Another series of elementary computations shows that for $p \in (0,1)$ and $2 \le j \le n-2$

$$g_{j,n}(p) = g_{j+1,n}(p) \Leftrightarrow p = \tfrac{j+1}{n+1}.$$

Then, obviously we have $\frac{j}{n} < \frac{j+1}{n+1} < b_{j,n}$. Therefore, $g_{j,n}$ is increasing on $\left(\frac{j}{n}, \frac{j+1}{n+1}\right)$ and $g_{j+1,n}$ is increasing on $\left(\frac{j+1}{n+1}, \frac{j+1}{n}\right)$. In consequence,

$$g_{j,n}\left(\tfrac{j}{n}\right) < g_{j,n}\left(\tfrac{j+1}{n+1}\right) = g_{j+1,n}\left(\tfrac{j+1}{n+1}\right) < g_{j+1,n}\left(\tfrac{j+1}{n}\right).$$

This completes the proof of (b). □

Proof of lemma 9(c) For $n = 6$, we have $Q_{2,6}\left(\frac{41}{120}\right) > 0$, so $q_2(6) < \frac{41}{120}$. Moreover, $g_{2,6}\left(\frac{41}{120}\right) < 0$, so as in the proof of (a) we get $\xi_2(6) < q_2(6)$. Moreover, $\bar{h}_2'\left(\frac{2}{6}\right) > 0$, so $\xi_2(6) > \frac{2}{6}$. Finally, $\bar{h}_3'(q_2) = \frac{1}{q_2^2} g(p)$ where

$$g(p) = 3B_{3,6}(p) - B_{0,6}(p) - B_{1,6}(p).$$

By VDP, the function g is $-+$. Moreover, $q_2 > \frac{2}{6}$ and $g\left(\frac{2}{6}\right) > 0$, and thus $g(q_2) > 0$. Therefore, $\bar{h}_3'(q_2) > 0$ and $\xi_3 > q_2$. □

Proof of lemma 9(d) To compare $\frac{2}{n}$ with $\xi_2(n)$, we write $\bar{h}_2'$ in the form

$$\bar{h}_2'(p) = \tfrac{1}{p^2}\left[(1-p)^n + np(1-p)^{n-1} + n(n-1)p^2(1-p)^{n-2} - 1\right].$$

Then,

$$\bar{h}_2'\left(\tfrac{2}{n}\right) = \tfrac{n^2}{4}\left[\tfrac{1}{n^2}\left(1 - \tfrac{2}{n}\right)^{n-2}\left(7n^2 - 12n + 4\right) - 1\right].$$

Moreover, $\bar{h}_2'\left(\frac{2}{7}\right) < 0$ and the expression in brackets is strictly decreasing with respect to $n \geq 7$. By (41), we obtain $\xi_2(n) < \frac{2}{n}$ for $n \geq 7$, which completes the proof of (9). □

Proof of lemma 9(e) By (41), we need to show that $\bar{h}_2'(q_1) > 0$. Using (43), we compute that

$$\bar{h}_2'(q_1) = \frac{1}{q_1^2}\left(2B_{2,n}(q_1) - B_{0,n}(q_1)\right).$$

It is elementary to show that $2B_{2,n}(p) > B_{0,n}(p)$ for $p \in \left(\frac{1}{n}, 1\right)$. Since $q_1 > \frac{1}{n}$, the claim follows. □

Appendix 3: The proof of theorem 4

Proof of theorem 4(a) Using Lemma 8, for $1 \le j \le \frac{n}{2}$ and $p \in \left[\frac{j}{n}, q_j\right]$ we have

$$t_{n,p}(j) < 0 < t_{n,p}(j+1). \tag{45}$$

Next, using Corollary 3(a), the inequality (45) and Corollary 4(a) for $n \ge 3$ and $p \in \left[\frac{1}{n}, q_1\right]$ we have

$$r_{n,p}(1) < t_{n,p}(1) < 0 < t_{n,p}(2) = r_{n,p}(2). \tag{46}$$

Thus, $1 \le k\left(n, \frac{1}{n}\right) \le k(n, q_1) \le 2$ by Theorem 1.

First, we prove that $k(n, q_1) = 2$ or equivalently $s_{n,q_1}(1) > s_{n,q_1}(2)$. For all $p \in \left[\frac{1}{n}, q_1\right]$ by (46), we have $s_{n,p}(2) = 1 - 2F_{2:n}(p)$. Moreover, using (46) again and the definition of q_1, we obtain $s_{n,q_1}(1) > 2 - F_{1:n}(q_1) - 1 = s_{n,q_1}(2)$.

To prove that $k\left(n, \frac{1}{n}\right) = 1$, we need to show that for $n \ge 3$

$$s_{n,\frac{1}{n}}(1) =: a_n < b_n := s_{n,\frac{1}{n}}(2). \tag{47}$$

Elementary but tedious computations, using (46) with $p = \frac{1}{n}$, show that

$$b_n = 2\left[\left(1 - \tfrac{1}{n}\right)^n + \left(1 - \tfrac{1}{n}\right)^{n-1}\right] - 1$$

is strictly decreasing to its limit $\frac{4}{\mathrm{e}} - 1 > \frac{4}{9}$ and $a_n < \frac{4}{9}$ for $n \ge 3$. Therefore, any element of the sequence $\{b_n\}$ is greater than any element of $\{a_n\}$. In particular, the inequality (47) holds for all $n \ge 3$, which completes the proof of (a). □

Proof of theorem 4(b) Assume that $n \ge 6$. By Corollary 4(b) and (c) we have $\left[\frac{2}{n}, q_2\right] \subset (\theta_3, \xi_3)$, so for p in the former interval by (45) we have

$$r_{n,p}(3) = t_{n,p}(3) > 0. \tag{48}$$

On the other hand, since $r_{n,p}$ is strictly decreasing with respect to p, then

$$r_{n,q_2}(2) < r_{n,\frac{2}{n}}(2) \le t_{n,\frac{2}{n}}(2) < 0. \tag{49}$$

Here, the second inequality follows from $\frac{2}{n} > \theta_2$ (in fact it becomes equality for $n = 6$), and the third one follows from (45). Summarizing, for $p = \frac{2}{n}$ and $p = q_2$ we have $r_{n,p}(2) < 0 < r_{n,p}(3)$ and therefore

$$2 \le k\left(n, \tfrac{2}{n}\right) \le k(n, q_2) \le 3.$$

First, we prove that $k(n, q_2) = 3$ or in other words $s_{n,q_2}(2) > s_{n,q_2}(3)$. Indeed, by (48) we have $s_{n,q_2}(3) = 1 - 2F_{3:n}(q_2)$. Moreover, by Corollary 4(b) and (c) we have

$q_2 > \xi_2 > \theta_2$, so $r_{n,q_2}(2) < t_{n,q_2}(2) < 0$. Therefore $s_{n,q_2}(2) > 2F_{2:n}(q_2) - 1 = s_{n,q_2}(3)$ by the definition of q_2.

Now we prove that $k\left(n, \frac{2}{n}\right) = 2$ or equivalently $s_{n,\frac{2}{n}}(2) < s_{n,\frac{2}{n}}(3)$. Since $\frac{2}{n} \in \left(\theta_3, \xi_3\right)$ then by (48)

$$s_{n,\frac{2}{n}}(3) = 1 - 2F_{3:n}\left(\frac{2}{n}\right). \tag{50}$$

For $n = 6$, we have also $\frac{2}{6} \in \left(\theta_3, \xi_2\right)$, so by (49) we obtain $s_{6,\frac{2}{6}}(2) = 2F_{2:n}\left(\frac{2}{6}\right) - 1 < s_{6,\frac{2}{6}}(3)$ by (50) and Simmons' inequality. For $n \geq 7$, we define the sequences

$$c_n := s_{n,\frac{2}{n}}(2), \quad d_n := s_{n,\frac{2}{n}}(3).$$

Tedious computations show that d_n is strictly decreasing to its limit $\frac{10}{e^2} - 1$ and $c_n < \frac{10}{e^2} - 1$ for $n \geq 7$. So $c_n < d_n$ for all $n \geq 7$, which completes the proof. □

References

Bieniek, M. (2007). Variation diminishing property of densities of uniform generalized order statistics. *Metrika, 65,* 297–309.

Gajek, L., Rychlik, T. (1998). Projection method for moment bounds on order statistics from restricted families. II. Independent case. *Journal of Multivariate Analysis, 64,* 156–182.

Hyndman, R. J., Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician, 50,* 361–365.

Kaas, R., Buhrman, J. (1980). Mean, median and mode in binomial distributions. *Statistica Neerlandica, 34,* 13–18.

Keating, J. P., Tripathi, R. (2006). Percentiles, estimation of. *Encyclopedia of statistical sciences* (pp. 6054–6060). New York: John Wiley.

Komornik, V. (2006). Another short proof of Descartes's rule of signs. *American Mathematical Monthly, 113,* 829–830.

Moriguti, S. (1953). A modification of Schwarz's inequality with applications to distributions. *Annals of Mathematical Statistics, 24,* 107–113.

Okolewski, A., Rychlik, T. (2001). Sharp distribution-free bounds on the bias in estimating quantiles via order statistics. *Statistics and Probability Letters, 52,* 207–213.

Papadatos, N. (1995). Maximum variance of order statistics. *Annals of the Institute of Statistical Mathematics, 47,* 185–193.

Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics, 46,* 247–257.

Perrin, O., Redside, E. (2007). Generalization of simmons theorem. *Statistics and Probability Letters, 77,* 604–606.

Reiss, R.-D. (1989). *Approximate distributions of order statistics*. New York: Springer-Verlag.

Rudin, W. (1976). *Principles of mathematical analysis* (3rd ed.). New York: McGraw-Hill Book Co.

Schoenberg, I. J. (1959). On variation diminishing approximation methods. In R. E. Langer (Eds.) *On numerical approximation. Proceedings of a symposium, Madison, April 21-23, 1958* (pp. 249–274). Madison: The University of Wisconsin Press.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: John Wiley.

Zieliński, R. (2009). Optimal nonparametric quantile estimators. Towards a general theory. A survey. *Communications in Statistics Theory and Methods, 38*(7), 980–992.