The Differences Among Methods for Computation of Quartiles Do Matter

Darja Rupnik Poklukar and **Janez Žerovnik**[1]
University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 6, SI-1000 Ljubljana, Slovenia
darja.rupnik@fs.uni-lj.si, janez.zerovnik@fs.uni-lj.si

Abstract

The choice of the method for computation of quartiles may have a direct impact on practical decisions. As the differences among the methods tend to vanish with growing sample size, a common belief is that the methods are practically equivalent. To the contrary, we show on several experiments with various sample sizes that in some examples, the differences may be very likely. For a discrete distribution, exact estimates of differences between expected values of sample quartiles, given by two different methods, are derived. This implies that it is crucial that in any application, the choice of the method for computation of quartiles (or, percentiles) is explicitly given.

Keywords: quantiles, percentiles, sample quartiles

1 Introduction

In the increasingly competitive global arena of business in the twenty first century, decision making is necessarily backed by statistics. No longer is the production of statistics confined to quantitative analysis and market research divisions only, but are increasingly useful when they are applied to improve decision making (Borozan, 2017; Gurgul and Machno, 2017; Hunjet et al., 2015; Marek and Vrabec, 2016). Statistics is based on probability theory, and thus mathematical rigor is expected in theory and application. In this paper, we discuss importance of clarification of some basic definitions of a notion that in practice may result in some unclarities in interpretation of results.

Quantiles play a fundamental role in statistics: they are the critical values used in hypothesis testing and interval estimation. Often they are the characteristics of distributions we usually wish to estimate. The use of quantiles as primary measure of performance has gained prominence, particularly in microeconomic, financial and environmental analyses and others.

In statistics a set of observed values (maybe with repetitions) is usually a sample drawn from some distribution, and a natural question is how to estimate the quartiles of the original distribution from the given sample. Sometimes there is additional information available about the type of distribution that may allow some assumptions giving arguments for various methods for estimation of quartiles yielding a variety of methods for computing the quartiles in the literature and in software (Hyndman and Fan, 1996; Langford, 2006). While there may be good reasons to use different algorithms (and different definitions!) for quartiles in various contexts, it is very important to know that various methods will either converge (the differences will vanish) or to understand that they may provide different results. It is well known among statisticians that there are a number of different definitions in the literature of the first and third quartile values of a finite data set. Furthermore, different methods based on these definitions are used by some statistical computing packages (Hyndman and Fan, 1996).

[1] Part time researcher at Institute of Mathematics, Physics and Mechanics, Jadranska 19, Ljubljana, Slovenia.

Langford (2006) answered on question *"Why worry? The differences are small so who cares? "* with words of Freund and Perles (1987):

> *"Before we go into any details, let us point out that the numerical differences between answers produced by the different methods are not necessarily large; indeed, they may be very small. Yet if quartiles are used, say to establish criteria for making decisions, the method of their calculation becomes of critical concern. For instance, if sales quotas are established from historical data, and salespersons in the highest quarter of the quota are to receive bonuses, while those in the lowest quarter are to be fired, establishing these boundaries is of interest to both employer and employee. In addition, computer-software users are sometimes unaware of the fact that different methods can provide different answers to their problems, and they may not know which method of calculating quartiles is actually provided by their software."*

We have discussed the problem from teachers' perspective elsewhere (Žerovnik and Rupnik Poklukar, 2017), and argued that Langford's method is the best choice for considering quartiles at elementary level. Langford (2006) writes *"the situation is, I believe, far worse than most realize."* He examined various methods which are actually used in elementary statistics textbooks and the methods employed by various commonly-used calculator and computer packages, and using a precise definition of percentile, identified which of the methods satisfy this definition. When discussing the issue with (anonymous) reviewers and editors it seems that in general, the statisticians do not find the issue to be very serious. Namely, it is believed that in statistics, due to consideration of large populations, large samples, and many repetitions the difference in the definitions of quartiles will vanish: *"It is of the utmost importance to emphasize that data quartiles and medians are providing estimates of theoretical or population values, and that differences we see in small samples in ways of calculating data values, do not tend to matter in large samples."* At first, we agreed with the statement, even illustrated the vanishing effect with an example (Žerovnik and Rupnik Poklukar, 2017). After a second thought, we however think that the inconsistencies may not be so easy to overcome, or at least a deeper argument is needed. In particular, the questions related may be much more tricky when considering discrete distributions.

The rest of this paper is organized as follows. We will recall some basic definitions in Section 2, describe two among the most popular methods for finding quartiles in Section 3 and we will look in detail at differences that occur in Section 4. Some exact estimates of those differences will be derived and the results will be discussed in conclusions.

2. Definitions

For a random variable X let F denote the (unknown) cumulative distribution function $F(x) = P(X \leq x)$. The $p-$th quantile q_p is given by $q_p = F^{-1}(p)$, where

$$F^{-1}(p) = \inf\{t \mid F(t) \geq p\},\ p \in (0,1) \tag{1}$$

denotes the generalized inverse of distribution function. Note that F^{-1} is nondecreasing as $F(q_p-) \leq p \leq F(q_p)$ and $F(q_p) = p$ in case of continuous distribution. The quartiles are special cases of percentiles: first quartile $Q_1 = q_{0.25}$, second quartile (or median) $Q_2 = M = q_{0.5}$, third quartile $Q_3 = q_{0.75}$.

The sample cumulative distribution function of a sample $X_1, \ldots, X_n$ of size n is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x),\ -\infty < x < \infty,$$

where $I(\cdot)$ denotes the indicator function. Then the sample $p-$th quantile based on the sample cumulative distribution function can be represented as

$$\hat{q}_p = \inf\{x \mid F_n(x) \geq p\},\ p \in (0,1). \tag{2}$$

Quite a lot of work has been done dealing with the limit properties of $\hat{q}_p$. Miao, Chen and Xu (2011) studied some asymptotic properties of the deviation between $p-$th quantile and the estimator including moderate and large deviations and also Bahadur representation. Ma, Genton and Parzen (2011) gave the definition of sample $p-$th quantile based on mid-distribution function to provide a unified framework for asymptotic properties of sample $p-$th quantile from discrete distributions.

For discrete quantitative random variables the most common idea is to arrange the data in ascending order and divide them into fixed number of roughly equal parts. (For finding quartiles, divide the ordered sample into four quarters having the same number of observations in each quarter, Joarder and Firozzaman (2001).) However, there are some differences in the details and also different definitions are used by various statistical software packages (Hyndman and Fan, 1996).

Langford (2006) uses the following definition of data percentiles.

Definition of data percentile. A i-th percentile value is a number which puts at least i percent of the data values at that number or below and at least (100 - i) percent of the data values at that number or above. If more than one such number exists, there will be an entire interval of such and we choose the i-th percentile value to be the midpoint of that interval. (We propose to call this value the **canonical value** of the quartile.)

3. Two methods for discrete data quartiles

Two methods among the most popular methods listed in Langford (2006) are described below. Both methods first compute the data median, and then compute a median of the two *halves*. In the first method (M1) the median is included in both halves if the number of data points in the entire set is odd and excluded if the number of data points is even (Tukey, 1977).

In the second method (M2), Langford (2006) suggests to divide the data set into two halves, a bottom half and a top half. If n is odd, include or exclude the median in the halves so that each half has an odd number of elements. The first and the third quartiles are then the medians of the bottom and top halves respectively. If n is even, the median is taken to be the average of the middle two values.

Both methods can be summarized by the following four rules (write $n = 4q + r$, where q is an integer, and assume the dataset is ordered $v_1 \leq v_2 \leq v_3 \cdots \leq v_n$):

Tukey's method (M1)

if $r = 3$, then $Q_1 = \frac{v_{q+1}+v_{q+2}}{2}$, $Q_2 = v_{2q+2}$ and $Q_3 = \frac{v_{3q+2}+v_{3q+3}}{2}$.
if $r = 2$, then $Q_1 = v_{q+1}$, $Q_2 = \frac{v_{2q+1}+v_{2q+2}}{2}$ and $Q_3 = v_{3q+2}$.
if $r = 1$, then $Q_1 = v_{q+1}$, $Q_2 = v_{2q+1}$ and $Q_3 = v_{3q+1}$.
if $r = 0$, then $Q_1 = \frac{v_q+v_{q+1}}{2}$, $Q_2 = \frac{v_{2q}+v_{2q+1}}{2}$and $Q_3 = \frac{v_{3q}+v_{3q+1}}{2}$.

Langford's method (M2)

if $r = 3$, then $Q_1 = v_{q+1}$, $Q_2 = v_{2q+2}$ and $Q_3 = v_{3q+3}$.
if $r = 2$, then $Q_1 = v_{q+1}$, $Q_2 = \frac{v_{2q+1}+v_{2q+2}}{2}$ and $Q_3 = v_{3q+2}$.
if $r = 1$, then $Q_1 = v_{q+1}$, $Q_2 = v_{2q+1}$ and $Q_3 = v_{3q+1}$.
if $r = 0$, then $Q_1 = \frac{v_q+v_{q+1}}{2}$, $Q_2 = \frac{v_{2q}+v_{2q+1}}{2}$and $Q_3 = \frac{v_{3q}+v_{3q+1}}{2}$.

We can see that there are only differences in case $n = 4q + 3$ so we will consider only this case in the following section.

4. Sample quartiles of discrete distributions

In case of absolutely continuous distribution it is well known that the asymptotic distribution of sample quantiles in the classical definition is normal (see Ma, Genton and Parzen (2011), Theorem 1 and further references therein). If the underlying distribution is discrete, the situation is much more delicate (Jentsch and Leucht, 2014). Sample quantiles may not even be consistent in general with the population quantiles in this case. In the following subsection we will consider an example of a discrete distribution.

4.1 Simulation throwing a (fair) dodecahedron

Suppose a dodecahedron is thrown independently n-times and we observe a sequence $X_1, \ldots, X_n$ of scores. Assume that the faces of dodecahedron are labelled with numbers $1, 2, \ldots, 12$.

Example 1[2]. In simple case with $n = 1000$ we found some interesting results (in only four attempts). Frequencies are written in Table 1, cumulative frequencies and sample quartiles, derived by definition (2), in Table 2.

Table 1: Frequencies of $n = 1000$ dodecahedron throws in four attempts.

	1	2	3	4	5	6	7	8	9	10	11	12
1	69	84	86	69	92	88	71	77	86	104	87	87
2	72	76	89	81	83	84	75	88	76	91	89	96
3	76	76	85	92	102	77	98	87	75	83	76	73
4	74	82	95	82	81	90	95	83	84	74	72	88

Table 2: Some cumulative frequencies and sample quartiles from Table 1.

	≤2	≤3	≤4	≤5	≤6	≤7	≤8	≤9	≤10	Q_1	Q_2	Q_3
1	153	239	308	400	488	559	636	722	826	4	7	10
2	148	237	318	401	485	560	648	724	815	4	7	10
3	152	237	329	431	508	606	693	768	851	4	6	9
4	156	251	333	414	504	599	682	766	840	3	6	9

As seen, sample quantiles are not consistent in general with the population quantiles. This issue occurs due to the fact that the cumulative distribution function (cdf) in this case is a step function. This leads to inconsistency if the level of the quartile lies in the image of the cdf and, consequently, the Central limit theorem does not hold anymore. In other words, the first quartiles of samples are most likely either 3, 4 or 3.5, regardless of the sample size. We may however anticipate that the expected value of the

[2] All numerical examples in this paper were done with R, version 3.1.0 (2014-04-10) The R Foundation for Statistical Computing.

sample quartiles will converge, hopefully to the quartile values of the population. We will look at this (phenomena) more closely below.

Suppose a dodecahedron is thrown independently $n = 4q + 3$ times. Clearly, we can assume that $P(X_i = k) = \frac{1}{12}$ for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, 12$. The random variables X_i are thus distributed uniformly over the set of twelve values. The canonical values of quartiles are thus $Q_1 = 3.5$, $Q_2 = M = 6.5$ and $Q_3 = 9.5$.

Let $v_1 \leq v_2 \leq \ldots \leq v_n$ be the ordered dataset of scores, each $v_i \in \{1, 2, \ldots, 12\}$. We will compare two different methods computing the first quartile. Recall that the first quartile of ordered sample is

(A) $Q_1^A = v_{q+1}$ by Langford's method (M2), and

(B) $Q_1^B = \frac{v_{q+1}+v_{q+2}}{2}$ by Tukey's method (M1) .

Let $E(X)$ denote the expected value of random variable X.

It is obvious that $E(Q_1^A) \leq E(Q_1^B)$ since $Q_1^A \leq Q_1^B$. The relationship between the two expected values is more precisely enlightened by the following two lemmas. (The proofs are given in the appendix A.)

Lemma 1 *At a fixed sample size* $n = 4q + 3$*,*

(a) the expected values of first quartiles due to different methods are different $E(Q_1^A) < E(Q_1^B)$*,*

(b) there is a constant $\Delta > 0$ *such that*

$$E(Q_1^B) - E(Q_1^A) > \Delta = \sum_{i=1}^{10} \Delta_i = \frac{1}{2}\binom{4q+3}{q+1} \sum_{i=1}^{10} \left(1 - \frac{i}{12}\right)^{3q+2} \cdot \left(\frac{i}{12}\right)^{q+1}.$$

Lemma 2 $\lim\limits_{n\to\infty} E(Q_1^B) - E(Q_1^A) = 0.$

Numerically, exact differences Δ from Lemma 1(b) for some different sample sizes $n = 4q + 3$ are given in Table 3.

Table 3: Some examples of differences lower bound Δ from Lemma 1.

n	23	43	203	403	803	1203
q	5	10	50	100	200	300
Δ	0.2499	0.1364	0.0338	0.0229	0.0162	0.0133

Our main concern is to observe how fast the difference $E(Q_1^B) - E(Q_1^A)$ vanishes. Namely, as quartiles are used in several ways in practical applications, it is important to know whether the choice of computational method has any impact on practical decisions. Bearing this in mind, we run several experiments with various population and sample sizes.

In the next example, 1000 fair dodecahedron throws of size $n = 4q + 3$ were randomly generated and the differences between the first quartiles given with two different methods were calculated.

Example 2. We randomly generated 1000-times a sample of size n of fair dodecahedron throws and then compared the difference D between the first quartiles calculated by Langford's method and by Tukey's method:

$$D = Q_1^B - Q_1^A.$$

Table 4: Percentage of differences D for samples of size n.

n	31	35	39	43	103	123	203	4003
$D = 0$	67.4	67.2	70.9	74.3	87.6	89.7	93.1	98.4
$D = 0.5$	30.6	31.7	28.4	25.2	12.4	10.3	6.9	1.6
$D = 1$	2.0	1.1	0.7	0.5	0	0	0	0

We measure the percentages of samples in which the differences are 0, 0.5 or 1 and write the results in Table 4.

Observe that for sample sizes up to $n = 43$ there is about 25.7% possibility to get a different result from a different methods. Samples of this size are often used in practical applications. Note that the differences appear often also for larger samples, even for sample of size $n = 4003$, the difference occurred in 1.6% cases in our simulation.

So, if quartiles are used to establish criteria for making decisions, the method of their calculation becomes of critical concern, as we mentioned earlier in the introduction, citing Langford (2006).

A more comprehensive batch of simulations was done for 1000 generated samples of each size $n = 4q + 3$ for $q = 7, 8, \ldots, 500$. The summary of the results is plotted in Figure 1.

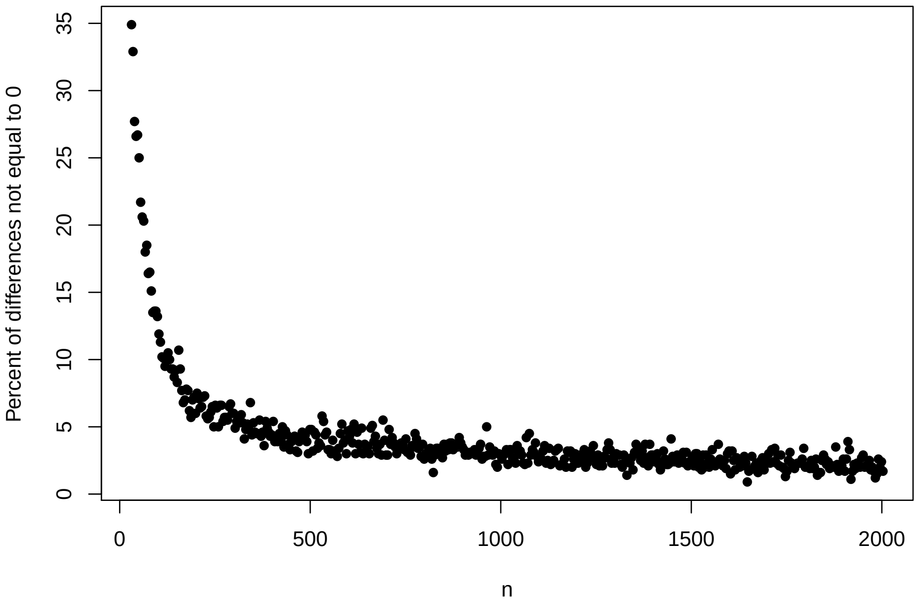

Figure 1: Differences between two methods.

Recall our main question: does the method of calculating quantiles matter? Theoretically, the different methods compute different estimates and have different properties. We have seen that there are examples with practical population and sample sizes, where the choice of the method can have important impact on the conclusions that are to be taken. Which method should be used? It depends on user's needs to choose the methods that suits him. However, it is crucial that the choice of the method is not hidden as a technical detail!

5. Conclusions

In this paper, our goal was to show that the choice of method for computation of quartiles may have a direct impact on practical decisions.

We have shown on an example of throwing fair dodecahedron that in contrary to rather common belief that because the differences between the methods tend to vanish, in some examples, they may be very likely. Recall the Example 2 with samples sizes 43 where the differences appear in almost 26% of cases!

If, for example, scores are results of the entrance exam for the University and those in the lowest quarter are not to be accepted, it may be very inconvenient to have two methods providing different results. In other words, there may be people that may enter the University by one method, and are not admitted to enter by another method. The discussion can easily be transformed into a legal issue!

Our experiment was chosen in a way to show the differences in quartile computations. However, for any given percentile, it is straightforward to design an experiment, where similar phenomena can be shown. For example, estimating the differences caused by different methods in calculating the $30-$th percentile $q_{0.30}$ we can use the simulation throwing of icosahedron (regular polyhedron with 20 faces).

It is therefore crucial that in any application, the choice of the method for computation of quartiles (or, percentiles) is explicitly given. Or in other words, any method that uses quartile (or percentile) values that come from a hidden computational method may be misleading.

A Detailed proofs

Before proving the Lemma 1 and Lemma 2, we write suitable expressions for $E(Q_1^A)$ and $E(Q_1^B)$.

Q_1^A is a discrete random variable with possible values $i \in \{1, 2, 3, \ldots, 12\}$, randomly selected with probabilities

$$p_i^{(q+1)} = P[v_{q+1} = i].$$

Thus,

$$E(Q_1^A) = 1 \cdot p_1^{(q+1)} + 2 \cdot p_2^{(q+1)} + \cdots + 12 \cdot p_{12}^{(q+1)} = \sum_{i=1}^{12} i \cdot P[v_{q+1} = i]. \tag{3}$$

For $j \geq i$ denote

$$p_{j,i}^{(q+2)} = P[v_{q+2} = j \,|\, v_{q+1} = i].$$

Then the expected value $E(Q_1^B)$ equals

$$\begin{aligned} E(Q_1^B) &= \sum_{i=1}^{12} \sum_{j=i}^{12} \frac{i+j}{2} \cdot P[v_{q+1} = i \wedge v_{q+2} = j] \\ &= \sum_{i=1}^{12} \sum_{j=i}^{12} \frac{i+j}{2} \cdot P[v_{q+1} = i] \cdot P[v_{q+2} = j \,|\, v_{q+1} = i] \\ &= \sum_{i=1}^{12} \sum_{j=i}^{12} \frac{i+j}{2} \cdot p_i^{(q+1)} \cdot p_{j,i}^{(q+2)}. \end{aligned} \tag{4}$$

Proof: (of Lemma 1) We will estimate each term in eq. (3). First observe that

$$p_1^{(q+1)} \cdot 1 \quad < \quad p_1^{(q+1)} \left[p_{1,1}^{(q+2)} \cdot 1 + p_{2,1}^{(q+2)} \cdot \frac{1+2}{2} + p_{3,1}^{(q+2)} \cdot \frac{1+3}{2} + \cdots + \right.$$
$$\left. + p_{12,1}^{(q+2)} \cdot \frac{1+12}{2} \right].$$

The expression in brackets is obviously greater than 1 since $\sum_{j=1}^{12} p_{j,1}^{(q+2)} = 1$, $p_{j,1}^{(q+2)} \geq 0$ and the multiplication factors are $\frac{1+j}{2} > 1, \quad j = 2, 3, \ldots, 12$.

Similarly,

$$\begin{aligned}
p_2^{(q+1)} \cdot 2 &< p_2^{(q+1)} \left[p_{2,2}^{(q+2)} \cdot \frac{2+2}{2} + p_{3,2}^{(q+2)} \cdot \frac{2+3}{2} + p_{4,2}^{(q+2)} \cdot \frac{2+4}{2} + \cdots + \right. \\
&\qquad \left. + p_{12,2}^{(q+2)} \cdot \frac{2+12}{2} \right], \\
p_3^{(q+1)} \cdot 3 &< p_3^{(q+1)} \left[p_{3,3}^{(q+2)} \cdot \frac{3+3}{2} + p_{4,3}^{(q+2)} \cdot \frac{3+4}{2} + p_{5,3}^{(q+2)} \cdot \frac{3+5}{2} + \cdots + \right. \\
&\qquad \left. + p_{12,3}^{(q+2)} \cdot \frac{3+12}{2} \right], \\
&\ldots\ldots \\
p_{11}^{(q+1)} \cdot 11 &< p_{11}^{(q+1)} \left[p_{11,11}^{(q+2)} \cdot \frac{11+11}{2} + p_{12,11}^{(q+2)} \cdot \frac{11+12}{2} \right], \\
p_{12}^{(q+1)} \cdot 12 &= p_{12}^{(q+1)} \cdot p_{12,12}^{(q+2)} \cdot \frac{12+12}{2},
\end{aligned}$$

since $p_{12,12}^{(q+2)} = 1$.

Summing together all those inequalities on the left side gives $E(Q_1^A)$ and on the right side exactly $E(Q_1^B)$. This proves (a) of Lemma 1. Now we consider the difference $E(Q_1^B) - E(Q_1^A)$.

We have

$$\begin{aligned}
&p_1^{(q+1)} \left[p_{1,1}^{(q+2)} \cdot 1 + p_{2,1}^{(q+2)} \cdot \frac{1+2}{2} + p_{3,1}^{(q+2)} \cdot \frac{1+3}{2} + \cdots + p_{12,1}^{(q+2)} \cdot \frac{1+12}{2} \right] \\
&\qquad > p_1^{(q+1)} \cdot \left[p_{1,1}^{(q+2)} \cdot 1 + \left(1 - p_{1,1}^{(q+2)}\right) \cdot \frac{3}{2} \right] \\
&\qquad = p_1^{(q+1)} \cdot \left[p_{1,1}^{(q+2)} \cdot 1 + \left(1 - p_{1,1}^{(q+2)}\right) \cdot \left(1 + \frac{1}{2}\right) \right] \\
&\qquad = p_1^{(q+1)} \cdot 1 + \triangle_1,
\end{aligned}$$

where

$$\triangle_1 = \frac{1}{2} p_1^{(q+1)} \cdot \left(1 - p_{1,1}^{(q+2)}\right).$$

Similarly, for $i = 2, 3, \ldots, 10$

$$\begin{aligned}
&p_i^{(q+1)} \left[p_{i,i}^{(q+2)} \cdot \frac{2i}{2} + p_{i+1,i}^{(q+2)} \cdot \frac{2i+1}{2} + p_{i+2,i}^{(q+2)} \cdot \frac{2i+2}{2} + \cdots + p_{12,i}^{(q+2)} \cdot \frac{i+12}{2} \right] \\
&\qquad > p_i^{(q+1)} \cdot \left[p_{i,i}^{(q+2)} \cdot i + \left(1 - p_{i,i}^{(q+2)}\right) \cdot \frac{2i+1}{2} \right] \\
&\qquad = p_k^{(q+1)} \cdot \left[p_{i,i}^{(q+2)} \cdot i + \left(1 - p_{i,i}^{(q+2)}\right) \cdot \left(i + \frac{1}{2}\right) \right] \\
&\qquad = p_i^{(q+1)} \cdot i + \triangle_i,
\end{aligned}$$

where

$$\triangle_i = \frac{1}{2} p_i^{(q+1)} \cdot \left(1 - p_{i,i}^{(q+2)}\right).$$

Summing together we get

$$E(Q_1^B) > E(Q_1^A) + \triangle_1 + \triangle_2 + \triangle_3 + \cdots + \triangle_{10}$$

or

$$E(Q_1^B) - E(Q_1^A) > \sum_{i=1}^{10} \triangle_i.$$

Hence we only need to find an estimate of

$$\sum_{i=1}^{10} \triangle_i = \sum_{i=1}^{10} \frac{1}{2} p_i^{(q+1)} \cdot \left(1 - p_{i,i}^{(q+2)}\right).$$

Since

$$\begin{aligned}
\triangle_i &= \frac{1}{2} p_i^{(q+1)} \cdot \left(1 - p_{i,i}^{(q+2)}\right) \\
&= \frac{1}{2} \cdot P[v_{q+1} = i] \cdot (1 - P[v_{q+2} = i \,|\, v_{q+1} = i]) \\
&= \frac{1}{2} \cdot P[v_{q+1} = i] \cdot P[v_{q+2} > i \,|\, v_{q+1} = i] \\
&= \frac{1}{2} \cdot P[v_{q+2} > i \wedge v_{q+1} = i] \\
&= \frac{1}{2} \binom{n}{q+1} \left(1 - \frac{i}{12}\right)^{n-(q+1)} \cdot \left(\frac{i}{12}\right)^{q+1},
\end{aligned}$$

it follows

$$\sum_{i=1}^{10} \triangle_i = \frac{1}{2} \binom{n}{q+1} \sum_{i=1}^{10} \left(1 - \frac{i}{12}\right)^{n-(q+1)} \cdot \left(\frac{i}{12}\right)^{q+1} > 0,$$

as claimed in Lemma 1(b). □

Proof: (of Lemma 2) From (3) and (4) we have

$$E(Q_1^B) - E(Q_1^A) = \sum_{i=1}^{12} \sum_{j=i}^{12} \frac{i+j}{2} \cdot p_i^{(q+1)} \cdot p_{j,i}^{(q+2)} - \sum_{i=1}^{12} i \cdot p_i^{(q+1)}.$$

Taking into account that $\sum_{j=i}^{12} p_{j,i}^{(q+2)} = 1$ and, in particular, $p_{12,12}^{(q+2)} = 1$, we have

$$\sum_{j=12}^{12} \frac{12+j}{2} \cdot p_{12}^{(q+1)} \cdot p_{j,12}^{(q+2)} - 12 \cdot p_{12}^{(q+1)} = 0$$

and, for $i < 12$,

$$\begin{aligned}
\sum_{j=i}^{12} \frac{i+j}{2} \cdot p_i^{(q+1)} \cdot p_{j,i}^{(q+2)} - i \cdot p_i^{(q+1)} &= \\
\sum_{j=i}^{12} \frac{i+j}{2} \cdot p_i^{(q+1)} \cdot p_{j,i}^{(q+2)} - i \cdot p_i^{(q+1)} \sum_{j=i}^{12} p_{j,i}^{(q+2)} &= p_i^{(q+1)} \sum_{j=i+1}^{12} \frac{j-i}{2} \cdot p_{j,i}^{(q+2)}.
\end{aligned}$$

Recall the meaning of $p_{j,i}^{(q+2)} \leq P[v_{q+2} > i \wedge v_{q+1} = i]$ and observe that the later event occurs exactly when the number of occurrences of values $\leq i$ is exactly $q+1$, hence

$$p_{j,i}^{(q+2)} \leq P[v_{q+2} > i \wedge v_{q+1} = i] = \binom{n}{q+1}\left(1 - \frac{i}{12}\right)^{n-(q+1)} \cdot \left(\frac{i}{12}\right)^{q+1}.$$

As $n = 4q + 3$ we have

$$\lim_{n\to\infty} \binom{n}{q+1}\left(1 - \frac{i}{12}\right)^{n-(q+1)} \cdot \left(\frac{i}{12}\right)^{q+1} = 0$$

and consequently

$$\lim_{n\to\infty} (E(Q_1^B) - E(Q_1^A)) = 0\,.$$

□

References

Borozan D, Borozan L (2017), Analyzing total-factor energy efficiency in Croatian counties: evidence from a non-parametric approach. *Cent. Eur. J. Oper. Res.*, 1 – 22. https://doi.org/10.1007/s10100-017-0493-8

Freund JE, Perles BM (1987), A new look at quartiles of ungrouped data. *Amer. Statist.* 41(3), 200 – 203.

Gurgul H, Machno A (2017), The impact of asynchronous trading on Epps effect on Warsaw Stock Exchange. *Cent. Eur. J. Oper. Res.* 25(2), 287 – 301. https://doi.org/10.1007/s10100-016-0442-y

Hunjet D, Neralić L, Wendell RE (2015), Evaluation of the dynamic efficiency of Croatian towns using Data Envelopment Analysis *Cent. Eur. J. Oper. Res.* 23(3), 675 – 686. https://doi.org/10.1007/s10100-014-0363-6

Hyndman RJ, Fan Y (1996), Sample quantiles in statistical packages. *Amer. Statist.* 50(4), 361 – 365.

Jentsch C, Leucht A (2014), Bootstrapping sample quantiles of discrete data. https://ub-madoc.bib.uni-mannheim.de/36588/1/Jentsch und Leucht 14-15.pdf (accessed 23.6.2017)

Joarder AH, Firozzaman M (2001), Quartiles for Discrete Data. *Teach. Stat.* 23(3), 86 – 89.

Langford E (2006), Quartiles in Elementary Statistics. *J. Stat. Educ.* 14(3), 16 pp. https://ww2.amstat.org/publications/jse/v14n3/langford.html

Ma Y, Genton MG, Parzen E (2011), Asymptotic properties of sample quantiles of discrete distributions. *Ann. Inst. Statist. Math.* 63, 227 – 243. DOI 10.1007/s10463-008-0215-z

Marek L, Vrabec M (2016), Using mixture density functions for modelling of wage distributions *Cent. Eur. J. Oper. Res.* 24(2), 389 – 405. https://doi.org/10.1007/s10100-015-0409-4

Miao Y, Chen YX, Xu SF (2011), Asymptotic properties of the deviation between order statistics and $p-$Quantile. *Comm. Statist. Theory Methods*, 40(1), 8 – 14.

Tukey JW (1977), *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Žerovnik J, Rupnik Poklukar D (2017), Elementary methods for computation of quartiles. *Teach. Stat.* 39(3), 88 – 91.