

Estimation of quantiles with the exact bootstrap method

Joanna Kisielińska¹

Abstract

A problem with the estimation of quantiles occurs when the sample comes from an unknown distribution. The estimation uses the bootstrap method in the version that the literature refers to as exact. Three bootstrap estimators were used: two of them based on one order statistic, and the third on a linear combination of two order statistics (for an integer). The distribution of the exact bootstrap estimator based on a single order statistic is known. It has been shown that there is no general form of the distribution of the exact bootstrap estimator based on two order statistics. However, it is possible to calculate such a distribution – the article presents the algorithm that performs such a task. The bootstrap confidence intervals were constructed using the exact percentile method. It has been shown that if the estimator is based on a single order statistic, it is known in advance which elements of the primary sample are the limits of the confidence intervals, so there is no need to resample. The intervals determined by the exact percentile method were compared with those constructed using other methods. It has been shown that the information on the direction of the asymmetry of the distribution that the sample comes from is worth considering when selecting the rank of the order statistic used as an estimator. Attention is paid to the influence of the quality of the pseudorandom number generators on the results of the Monte Carlo simulation.

Key words: quantile estimation, confidence intervals for quantile, exact bootstrap method, exact percentile method, Monte Carlo method.

1. Introduction


Let X be a continuous random variable with cumulative distribution (CDF) $F(x)$ and density function (PDF) $f(x)$. Let $p \in (0,1)$ be given and let ξ_p be p -quantile of F , such that $p = F(\xi_p)$, $\xi_p = F^{-1}(p)$, and $f(\xi_p)$ (e.g. Bahadur (1966, p. 577), Nagaraja and Nagaraja (2020, p. 75)). Bahadur (1966, p. 577), gives the conditions to be satisfied by F so that ξ_p be unique.

The p -quantile is most often defined as the left quantile (e.g. Serfling, 1980, p. 3):

$$\xi_p = F^{-1}(p) = \inf\{x: F(x) \geq p\}. \quad (1)$$

¹ Warsaw University of Life Sciences, Poland. E-mail: joanna_kisielinska@sggw.edu.pl.

ORCID: <https://orcid.org/0000-0003-3289-1525>.

© Joanna Kisielińska. Article available under the CC BY-SA 4.0 licence 

Sample quantiles are used to estimate quantiles. For a sample (X_1, X_2, \dots, X_n) from distribution F Serfling (1980, p. 74) defines the sample p -quantile ξ_{pn} as p -quantile of empirical distribution F_n :

$$\xi_{pn} = F_n^{-1}(p) = \inf\{x: F_n(x) \geq p\}. \quad (2)$$

Sample p -quantiles, used as quantile estimators (i.e. when $\hat{\xi}_{pn} = \xi_{pn}$), are presented using order statistics. Let X_{nz} denote the z th order statistic, i.e. the smallest z th element of the sample of size n , and then (Serfling, 1980, p. 88):

$$\hat{\xi}_{pn} = \begin{cases} X_{n,np}, & \text{if } np \text{ is integer} \\ X_{n,[np]+1}, & \text{if } np \text{ is not integer} \end{cases} \quad (3)$$

where $[\cdot]$ denotes the floor function.

Nagaraja and Nagaraja (2020, p. 75) identify the sample p -quantile with the following order statistic:

$$\hat{\xi}_{pn} = X_{n,[np]+1}. \quad (4)$$

Hyndman and Fan (1996, p. 361) give many other definitions of sample quantiles based on order statistics. Their general form is:

$$\xi_{pn} = (1 - \gamma)X_{nj} + \gamma X_{n,j+1}, \quad (5)$$

where $\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$ for some $m \in \mathbb{R}$ and $0 \leq \gamma \leq 1$. The γ parameter is a function of j and g , where $j = [pn + m]$ and $g = pn + m - j$. Formula (5) includes the definitions (3) and (4).

With some assumptions (Serfling, 1980 p. 74), the sample p -quantile $\hat{\xi}_{pn}$ defined by (2) is strongly consistent for estimation of ξ_p .

It is known that the sample p -quantile is asymptotically normal if f is continuous and positive at ξ_p (e.g. Serfling, 1980, p. 77), (Nagaraja and Nagaraja, 2020, p. 77). The limit distribution has a mean ξ_p and a variance $\frac{p(1-p)}{f^2(\xi_p)n}$. The sample quantile vector $(\hat{\xi}_{p_1}, \dots, \hat{\xi}_{p_k})$ is also asymptotically normal for $0 < p_1 < \dots < p_k < 1$ if f is continuous and positive at $\xi_{p_1}, \dots, \xi_{p_k}$. The parameters of this distribution are a mean vector $(\xi_{p_1}, \dots, \xi_{p_k})$ and a covariance matrix with elements: $\left(\frac{p_i(1-p_j)}{f(\xi_{p_i})f(\xi_{p_j})n} \right)$ (Serfling, 1980 p. 80). The consequence of the asymptotic normality of the sample quantile vector is the asymptotic normality of any linear combination of these quantiles.

Serfling (1980, p. 94) based on the Bahadur (1966) article indicates that the order statistic X_{nk_n} (where $\{k_n\}$ is a sequence of positive integers ($1 \leq k_n \leq n$) such that k_n/n tends to p sufficiently fast) and the sample p -quantile $\hat{\xi}_{pn}$ are roughly equivalent as estimates of ξ_p . Despite this, in the general case difference $X_{nk_n} - \xi_p$ has a limit normal distribution not centered at 0 (Serfling, 1980, p. 94).

If the sample of size n comes from a continuous distribution with CDF $F(x)$ and PDF $f(x)$, then the PDF of the z th order statistics X_{nz} is (e.g. David and Nagaraja, 2003, p. 10); (Evans, Leemis and Drew, 2006, p. 20):

$$f_{X_{nz}}(x) = \frac{n!}{(z-1)!(n-z)!} f(x)[F(x)]^{z-1}[1 - F(x)]^{n-z} \quad (6)$$

This formula is the same as that given by Serfling (1980 p. 85), which specifies PDF of the sample p -quantile.

The practical application of the expression (6) is cumbersome (Serfling, 1980, p. 87). First, it requires knowledge of the distribution the sample comes from, and secondly, the distribution of order statistics is not usually in the class of known and commonly used distributions. Pekasiewicz (2015, p. 23) gives the density functions of the order statistics X_{nz} for selected distributions the sample comes from. Using limit distribution is also troublesome due to the necessity of knowing $f(\xi_p)$. The bootstrap method proposed by Efron in 1979 does not have these disadvantages. It does not require knowledge of the distribution a sample comes from. Falk and Kaufmann (1991), Falk and Reiss (1989), Bickel and Freedman (1981) and Singh (1981) (among others) studied the convergence of the bootstrap estimators of the parameters (also quantiles). They showed that bootstrap error converges to 0 with probability one. This indicates the correctness of this approach, although one can discuss the order of this convergence.

In the bootstrap method, empirical distribution F_n is an estimator of the distribution F . And therefore, the bootstrap estimator distribution (dependent on the F_n) is an estimator of the estimator distribution (dependent on the F). Efron (1979, p. 4) proposes three methods of computing the bootstrap estimator distribution. The first is a theoretical calculation, the second is the Monte Carlo (MC) approximation, and the third is the Taylor series expansion. The MC approximation involves selecting many resamples of size n with replacement from the n -element primary sample. Fisher and Hall (1991) pointed out that instead of drawing resamples² (especially for small samples), one can generate all resamples. One can then determine all the realizations of the bootstrap estimator. This method was called the exact bootstrap method in order to distinguish it from the commonly used MC approximation with resampling. It should be noted that the distributions of the bootstrap estimators gained with the exact bootstrap method are equivalent to those obtained with the first method proposed by Efron. The difference is only in the method of their determination. The exact method relies on numerical calculations, Efron method on theoretical calculations. In the following considerations, the bootstrap method based on all resamples will be called the exact method, no matter how the calculations were made.

² There are n^n resamples in total, but the different resamples are $\binom{2n-1}{n}$ (Fisher and Hall 1991 p. 160). To calculate the number of resamples with the same elements, one should compute the number of its permutations. One should permute only the elements on positions with non-repeating elements.

Taking into consideration all resamples allows to eliminate errors caused by resampling (Hutson and Ernst, 2000, p. 94). Resampling may be interpreted as drawing bootstrap samples from their entire population (n^n). In the MC approximation, some resamples may be omitted, while others used multiple times. Kisieleńska (2013, p. 1068) presented the comparison of the exact bootstrap method and the bootstrap method with resampling for any parameter.

The bootstrap p -quantile estimators ξ_{pn}^* are also based on order statistics. The order statistics of the bootstrap resample X_{nz}^* is its z th smallest element. Evans, Leemis, Drew (2006, p. 23) give the distribution of such statistic – a case of a finite population, sampling with replacement. The distribution thus determined is of course the exact bootstrap distribution (formula (16) in section 2).

The bootstrap method is not the only method for estimating quantiles, which does not require to know the distribution the sample comes from. For an ample review of distribution-free methods to construct confidence intervals, see Nagaraja and Nagaraja (2020).

Confidence intervals for quantiles can be determined using an asymptotic approach based on the sample p -quantile. In simulation experiments, the samples come from a known distribution, and therefore the values of ξ_p and $f(\xi_p)$ are known. One can determine $1-\alpha$ confidence interval of the sample p -quantile from the limit distribution:

$$I_{pn}^{Aa} = [F_A^{-1}(\alpha/2), F_A^{-1}(1 - \alpha/2)] \quad (7)$$

where F_A is the normal distribution with mean ξ_p and variance $\frac{p(1-p)}{f^2(\xi_p)n}$.

Serfling (1980, p. 130) proposes to use an asymptotic approach based on order statistics to determine confidence intervals for quantiles. The confidence interval is as follows:

$$I_{pn}^{Ab} = [X_{nk_{1n}}, X_{nk_{2n}}] \quad (8)$$

where $k_{1n} = n \cdot \left(p - \frac{u_{\alpha} \sqrt{p(1-p)}}{\sqrt{n}} \right)$, $k_{2n} = n \cdot \left(p + \frac{u_{\alpha} \sqrt{p(1-p)}}{\sqrt{n}} \right)$, and u_{α} is the $100 \cdot (1-\alpha/2)$ th percentile point of standard normal distribution. If $n \rightarrow \infty$ confidence coefficient of the interval $I_{pn}^{Ab} \rightarrow 1-\alpha$ (Serfling, 1980 p. 104).

The percentile method enables to construct confidence intervals when using the bootstrap approach. The main objections to this method relate to applications in the cases of small samples. Many authors note that the percentile method produces confidence intervals of first-order accuracy only (e.g Falk and Kaufmann (1991), Efron and Tibshirani (1993), Nagaraja and Nagaraja, 2020). For this reason, many proposals for better solutions have been created.

Efron (1987) proposed the BCa method (i.e. bias-corrected and accelerated), which is second-order accurate (Efron and Tibshirani, 1993) and has higher coverage probability.

Nagaraja and Nagaraja (2020) proposed a method of adjacent spacings to construct quantiles confidence intervals. The method is easy to apply, yet it requires experimental selection of two parameters s and t , and knowledge of the critical values of the statistics $W_{(s+t)}$. Critical values of the $W_{(s+t)}$ are given by Nagaraja and Nagaraja (2020, p. 88). Parameters s and t determine the rank of order statistics, used to construct the confidence interval. Nagaraja and Nagaraja (2020) conducted simulation studies to compare quantiles confidence intervals obtained using various distribution free-methods. They assessed the effectiveness of the methods based on the width of confidence intervals and coverage probability.

The problem presented in the article is in the estimation of quantiles when a distribution the sample comes from is not known. The novelty of the approach presented in the paper consists in estimating the quantiles with an exact bootstrap quantile estimator based on a linear combination of two order statistics. The algorithm presented in Section 2 allows for determining its distribution exactly, not only in an approximate manner. It is also shown that in the case of quantile estimation, confidence intervals are much easier to determine with the exact percentile method than with the percentile method with resampling, if the estimator is based on a single order statistics. Moreover, it is shown that the information about the direction of asymmetry of the distribution the sample comes from can be used to determine the rank of the single order statistics used as an estimator.

All calculations were made in Excel using the VBA language for Application.

2. Distributions of quantiles bootstrap estimators

Let the n -element resample, drawn with replacement from the original sample (x_1, x_2, \dots, x_n) , be marked as $(X_1^*, X_2^*, \dots, X_n^*)$. Each variable X_i^* has a discrete empirical distribution F_n . Efron, 1979 assumed equal probabilities $p_i = 1/n$ for each element of the primary sample x_i . Due to a finite measurement accuracy of any values, elements of the observed sample can be repeated. The empirical distribution is determined by probabilities p_i for each x_i where $\sum_{i=1}^k p_i = 1$ and k is the number of distinct elements in a primary sample.

The elements of the resample are the discrete random variables X_i^* with PDF f_n , CDF F_n , and survival function (SF) S_n :

$$f_n(x) = P(X_i^* = x) = \begin{cases} p_j & x = x_j, j = 1, \dots, k \\ 0 & \text{for others } x \in R \end{cases}, \tag{9}$$

$$F_n(x) = P(X_i^* \leq x) = \sum_{j=1; x_j \leq x}^k p_j, \tag{10}$$

$$S_n(x) = P(X_i^* \geq x) = 1 - F_n(x) + f_n(x). \tag{11}$$

The bootstrap quantile estimator according to (5) can be written in the general form as:

$$\hat{\xi}_{pn}^* = (1 - \gamma)X_{nj}^* + \gamma X_{n,j+1}^* \quad (12)$$

where: $0 \leq \gamma \leq 1$, $j=[np]$ (wherein $[np]=np$ for integer np), and X_{nj}^* is the j th order statistics of the resample.

In the research three bootstrap quantile estimators were used:

$$\hat{\xi}_{pn}^{1*} = \begin{cases} X_{n,np}^* & \text{if } np \text{ is integer} \\ X_{n,[np]+1}^* & \text{if } np \text{ is not integer} \end{cases} \quad (13)$$

$$\hat{\xi}_{pn}^{2*} = X_{n,[np]+1}^* \quad (14)$$

$$\hat{\xi}_{pn}^{3*} = \begin{cases} (1 - \varepsilon)X_{n,np}^* + \varepsilon X_{n,np+1}^* & \text{if } np \text{ is integer} \\ X_{n,[np]+1}^* & \text{if } np \text{ is not integer} \end{cases} \quad (15)$$

where: $\varepsilon = (n+1)p - [(n+1)p]$ as Hutson 2002 p. 332 suggests.

The estimator $\hat{\xi}_{pn}^{1*}$ was obtained assuming $\gamma = 0$ for np integer and $\gamma = 1$ for np not integer, the estimator $\hat{\xi}_{pn}^{2*}$ assuming $\gamma = 1$, and the estimator $\hat{\xi}_{pn}^{3*}$ assuming $\gamma = \varepsilon$ for np integer and $\gamma = 1$ for np not integer.

In the formulae given below, it was assumed that the primary sample is ordered, viz. $x_1 \leq x_2, \dots, x_{n-1} \leq x_n$.

Distributions of bootstrap quantile estimators based on one order statistics result directly from the formula given by Evans, Leemis, Drew (2006, p. 23) and are as follows:

$$P(X_{nz}^* = x_l) = \begin{cases} \text{for } l = 1 \\ \sum_{w=0}^{n-z} \binom{n}{w} [f_n(x_1)]^{n-w} [S_n(x_2)]^w \\ \text{for } l = 2, \dots, k-1 \\ \sum_{u=0}^{z-1} \sum_{w=0}^{n-z} \binom{n}{u, n-u-w, w} [F_n(x_{l-1})]^u [f_n(x_l)]^{n-u-w} [S_n(x_{l+1})]^w \\ \text{for } l = k \\ \sum_{u=0}^{z-1} \binom{n}{u} [F_n(x_{k-1})]^u [f_n(x_k)]^{n-u} \end{cases} \quad (16)$$

where z is the rank of the order statistic used as the estimator.

When np is not integer the estimators $\hat{\xi}_{pn}^{1*}$, $\hat{\xi}_{pn}^{2*}$, and $\hat{\xi}_{pn}^{3*}$ are the same. The rank of the order statistic used as the bootstrap estimator of p -quantile is $z=[np] + 1$. When np is not integer, the rank of the order statistic used as the bootstrap estimator of p -quantile is $z=np$ for estimator $\hat{\xi}_{pn}^{1*}$ and $z=np$ for estimator $\hat{\xi}_{pn}^{2*}$.

Only elements of a primary sample can be realizations of the estimators based on one order statistics. Realizations of the estimator in the form of a linear combination of two order statistics may also be weighted means of all two-element combinations

chosen therefrom. This means that the estimator $\hat{\xi}_{pn}^{3*}$ for integer np has a considerably higher number of realizations. It is impossible to give a general expressions determining these estimator. Nevertheless, one can determine the probabilities that on positions np and $np+1$ in resamples either any l th primary sample element will occur or any two of its elements: l_1 and l_2 , with $l_1 < l_2$.

The probability that in an ordered resample element x_l occurs at least on two positions $z = np$ and $z + 1$ is:

$$P\left((X_{nz}^* = x_l) \wedge (X_{n,z+1}^* = x_l)\right) = \begin{cases} \text{for } l = 1 \\ \sum_{w=0}^{n-z-1} \binom{n}{w} [f_n(x_1)]^{n-w} [S_n(x_2)]^w \\ \text{for } l = 2, \dots, k - 1 \\ \sum_{u=0}^{z-1} \sum_{w=0}^{n-z-1} \binom{n}{u, n-u-w, w} [F_n(x_{l-1})]^u [f_n(x_l)]^{n-u-w} [S_n(x_{l+1})]^w \\ \text{for } l = k \\ \sum_{u=0}^{z-1} \binom{n}{u} [F_n(x_{k-1})]^u [f_n(x_k)]^{n-u} \end{cases} \quad (17)$$

The probability that in an ordered resample, element x_l occurs exactly z times, and the element x_l , for $l = 2, \dots, k - 1$ occurs at least once on position $z + 1$ is equal to:

$$P\left((X_{nz}^* = x_l) \wedge (X_{n,z+1}^* = x_l)\right) = \sum_{w=0}^{n-z-1} \binom{n}{z, n-z-w, w} [F_n(x_1)]^z [f_n(x_l)]^{n-z-w} [S_n(x_{l+1})]^w. \quad (18)$$

The probability that in an ordered resample, element x_l , for $l = 2, \dots, k - 1$ occurs at least once on position z , and element x_k occurs exactly $n-z$ times, is:

$$P\left((X_{nz}^* = x_l) \wedge (X_{n,z+1}^* = x_k)\right) = \sum_{u=0}^{z-1} \binom{n}{u, z-u, n-z} [F_n(x_{l-1})]^u [f_n(x_l)]^{z-u} [S_n(x_k)]^{n-z}. \quad (19)$$

The probability that element x_l occurs in an ordered resample exactly z times, and the elements x_k exactly $n-z$ times, is:

$$P\left((X_{nz}^* = x_l) \wedge (X_{n,z+1}^* = x_k)\right) = \binom{n}{z, n-z} [f(x_l)]^z [f_n(x_k)]^{n-z}. \quad (20)$$

The probability that in an ordered resample element x_{l_1} occurs at least once on position z , and element x_{l_2} at least once on position $z+1$ $\wedge_{l_1 < l_2} (l_1 < l_2) \in \{2, 3, \dots, k - 2\} \times \{3, 4, \dots, k - 1\}$, is:

$$P\left((X_{nz}^* = x_{l_1}) \wedge (X_{n,z+1}^* = x_{l_2})\right) = \sum_{u=0}^{z-1} \sum_{w=0}^{n-z-1} \binom{n}{u, z-u, n-z-w, w} W \quad (21)$$

$$W = [F_n(x_{l_1-1})]^u [f_n(x_{l_1})]^{z-u} [f_n(x_{l_2})]^{n-z-w} [S_n(x_{l_2+1})]^w$$

The exact distribution of bootstrap p -quantile estimator based on two order statistics (estimator $\hat{\xi}_{pn}^{3*}$ when np is an integer) is:

$$\begin{aligned}
 P(\hat{\xi}_{pn}^{3*} = x_l) &= P\left((X_{nz}^* = x_l) \wedge (X_{n,z+1}^* = x_l)\right), \text{ for } l = 1, \dots, k \\
 P(\hat{\xi}_{pn}^{3*} = (1 - \varepsilon)x_1 + \varepsilon x_l) &= P\left((X_{nz}^* = x_1) \wedge (X_{n,z+1}^* = x_l)\right), \\
 &\text{ for } l = 2, \dots, k - 1 \\
 P(\hat{\xi}_{pn}^{3*} = (1 - \varepsilon)x_l + \varepsilon x_k) &= P\left((X_{nz}^* = x_l) \wedge (X_{n,z+1}^* = x_k)\right), \\
 &\text{ for } l = 2, \dots, k - 1 \\
 P(\hat{\xi}_{pn}^{3*} = (1 - \varepsilon)x_1 + \varepsilon x_k) &= P\left((X_{nz}^* = x_1) \wedge (X_{n,z+1}^* = x_k)\right) \\
 P(\hat{\xi}_{pn}^{3*} = (1 - \varepsilon)x_{l_1} + \varepsilon x_{l_2}) &= P\left((X_{nz}^* = x_{l_1}) \wedge (X_{n,z+1}^* = x_{l_2})\right), \\
 \wedge_{l_1 < l_2} (l_1 < l_2) &\in \{2, 3, \dots, k - 2\} \times \{3, 4, \dots, k - 1\}. \tag{22}
 \end{aligned}$$

Some realizations of the $\hat{\xi}_{pn}^{3*}$ estimator may repeat themselves, therefore the probabilities corresponding to these realizations should be added. As the number and order of the ordered realizations of the estimator based on two order statistics depend on the primary sample, one cannot give the general form of its distribution.

The algorithm for determining the distribution of the estimator based on two order statistics is as follows:

1. For each pair $(l_1, l_2) \in \{1, 2, \dots, k\} \times \{1, 2, \dots, k\}$ such that $l_1 \leq l_2$, the corresponding realization of the estimator should be calculated: $y_j = (1 - \varepsilon)x_{l_1} + \varepsilon x_{l_2}$ and probability $P\left((X_{nz}^* = x_{l_1}) \wedge (X_{n,z+1}^* = x_{l_2})\right)$.
2. Calculate the sum of probabilities determined in point 1 for each unique y_j .
3. If necessary, the estimator realizations should be sorted (e.g. to use the percentile method).

The presented algorithm allows for an exact calculation of the distribution of a linear combination of two consecutive order bootstrap statistics. Nagaraja and Nagaraja (2020, p. 81) based on the previous work of other authors (Nyblom, 1992), (Hettmansperger and Sheather, 1986) give the formulas that allow calculating this distribution approximately.

A useful attribute of quantile estimators distributions based on single order statistics is that the probabilities for all estimator realizations are the same for all primary samples of a given size, provided that $k = n$ (probabilities given by the expression (16) depend only on n and p). Distributions of estimators in the form of a linear combination of two order statistics do not have such property. It is worth noting, however, that the probabilities given by the expressions (17)–(21) also depend only on n and p – if there were no repetition in the sample. The occurrence of repetitions only causes that the probabilities for an element occurring multiple times in the sample are added together.

Knowing the exact bootstrap distribution of the quantile estimator can be useful for constructing confidence intervals for quantile. Determining the expected value and variance does not require knowing it. These values can be calculated using exact analytical expressions for any L -estimator given by Hutson and Ernst (2000). All variants of estimators (13)–(15) are L -estimators. If an estimator is based on a single order statistics, the expressions for multipliers for a mean (Hutson and Ernst (2000, p. 91)) are equivalent to the probabilities given in (16). If the estimator is based on two order statistics, the multipliers can be easily obtained from formulas (17)–(21). It is worth recalling that the expressions for the mean and variance of the median bootstrap estimators were given by Maritz and Jarrett as early as 1978. It was before Efron presented the concept of the bootstrap method.

3. Confidence intervals for quantiles by the exact bootstrap percentile method

One may construct the quantiles bootstrap confidence intervals by the percentile method described in the paper Wilcox (2001, p. 88), among others. It should be noted that the resamples do not need to be drawn from their entire population (size n^n). The distributions of the bootstrap quantile estimators can be calculated. On this basis, one may easily find the limits of the confidence interval. We know in advance numbers of the primary sample elements, constituting the limits of confidence intervals when the estimator is based on a single order statistic³. The determination of the limits of the quantiles confidence intervals requires much larger calculations (when n is big⁴) if the estimator is based on two order statistics, due to the sorting of possible realizations – in that case, resampling may be justified but is not necessary. The percentile method using all resamples can be called the exact percentile method (by analogy with the exact bootstrap method).

Let y_j be a realization of a bootstrap p -quantile estimator $\hat{\xi}_{pn}^*$, for $j = 1, \dots, o$. If an estimator is based on a single order statistic, o is equal to k (or n if there were no repetitions in the primary sample). Let us mark the bootstrap confidence interval as $I_{pn}^* = [y_{z_1}^*, y_{z_2}^*]$. For a given confidence level of $1 - \alpha$, the lower limit is:

$$y_{z_1}^* = \sup \left\{ y_j : F_{\hat{\xi}_{pn}^*}(y_j) \leq \frac{\alpha}{2} \right\}, \tag{23}$$

where $F_{\hat{\xi}_{pn}^*}$ is the bootstrap quantile estimator distribution. The upper limit is:

$$y_{z_2}^* = \inf \left\{ y_j : F_{\hat{\xi}_{pn}^*}(y_j) \geq 1 - \frac{\alpha}{2} \right\}, \tag{24}$$

³ It results from the properties of the exact distribution of the bootstrap percentile estimator (the probabilities of individual realizations of the estimator are the same for all samples of a given size). Table 4 lists these numbers for $p = 0.5$ and $1 - \alpha = 0.95$.

⁴ Currently, due to the high computing efficiency of computers, computations even for big n are not long-lasting.

If the bootstrap p -quantile estimator is based on a single order statistic, the elements of the primary sample are the limits of the confidence intervals. If the estimator is based on a linear combination of order statistics, the linear combinations of these elements may also be the limits.

Due to the discrete nature of distributions of the bootstrap estimator, besides the assumed confidence level $1 - \alpha$, we have a confidence level that can be called actual.

The number of realizations of bootstrap quantile estimators based on single order statistics is much smaller than that based on their linear combination. This has two important consequences. First, estimators based on two order statistics will allow building narrower confidence intervals than those based on one. Secondly, we can suspect that the discrepancy between the assumed and the actual confidence level is smaller for the estimator based on two order statistics than for that based on a single one (Nagaraja and Nagaraja (2020, p. 81)⁵ pay attention to this discrepancy).

4. Monte Carlo method for quantile estimation

The bias and variance of the estimators, the widths of the confidence intervals, and the coverage probability can be estimated by the Monte Carlo (MC) simulation method. These measures can be used, for example, to compare different estimators. Calculating them requires drawing R random samples, the so-called replication. If the bootstrap estimators are used, a single replication is a single primary sample (which may be resampled).

For sampling, pseudorandom number generators are used, which generate real numbers from a uniform distribution on the interval $[0; 1]$. Let the drawn number (ω_i) be the value of a CDF of the distribution the sample comes from. Elements of the sample can be designated as $x_i = F^{-1}(\omega_i)$, for $i = 1, \dots, n$.

Let β denote some target quantity of interest, $\hat{\beta}_R$ its MC estimate from simulation experiment with R replications, and $\hat{\beta}_r$ the estimate based on the r th replication, $r = 1, \dots, R$ (Koehler et al. (2009)). The MC estimate of β is then:

$$\hat{\beta}_R = \frac{1}{R} \sum_{r=1}^R \hat{\beta}_r. \quad (25)$$

In statistical experiments, the distributions that the samples come from are known. So, it is possible to calculate the MC approximation of the estimator bias relative to the true value of the p -quantile. The MC estimate of the bias and variance of some p -quantile estimator is:

$$\widehat{bias}_{MC} = \frac{1}{R} \sum_{r=1}^R \sum_{j=1}^o \left(y_j^r \cdot P(\hat{\xi}_{pn}^r = y_j^r) \right) - \xi_p, \quad (26)$$

⁵ Nagaraja and Nagaraja point out the discrepancy between the assumed confidence level and the coverage probability (coverage probability will be discussed in Section 4). Since the coverage probability can be regarded as an estimate of the actual level of confidence, both statements are roughly equivalent.

$$\hat{V}_{MC} = \frac{1}{R} \sum_{r=1}^R \left(\sum_{j=1}^o \left((y_j^r)^2 \cdot P(\hat{\xi}_{pn}^r = y_j^r) \right) - \left(\sum_{j=1}^o \left(y_j^r \cdot P(\hat{\xi}_{pn}^r = y_j^r) \right) \right)^2 \right), \tag{27}$$

where $\hat{\xi}_{pn}^r$ is the p -quantile estimator in the r th replication.

Let the confidence interval determined in the r th replication be marked as $I_{pn}^r = [y_{z_1}^r, y_{z_2}^r]$. The MC estimate of its width and the coverage probability is:

$$\hat{d}_R = \frac{1}{R} \sum_{r=1}^R (y_{z_2}^r - y_{z_1}^r) \tag{28}$$

$$\varphi_R = \frac{\#\{[y_{z_1}^r, y_{z_2}^r] : \varphi_p \in [y_{z_1}^r, y_{z_2}^r]\}}{R} \tag{29}$$

When the coverage probability is close to the assumed confidence level but not lower than it, the method of determining the confidence intervals properly fulfills its task.

5. The median estimation – comparison of estimators

The simulation research using the Monte Carlo method was carried out. Samples come from six distributions: two with right asymmetry (LogNormal(1.0.75), Gamma(2.2)), two with left asymmetry (-LogNormal(1.0.6) + 5, Gamma(1.25, 2.5) + 5) and two symmetrical (N(3.0.5) and N(3.2)). The sample sizes were selected to include both small and large samples.

For different sample sizes n , $R = 2,000$ times n pseudorandom numbers from the interval $[0,1]$ were drawn, which were treated as a CDF value. The same CDF values were used for all distributions and methods, which allows for a better comparability of results. Such selection makes the results for individual cases independent of the quality of the pseudorandom number generator. Based on the n values of the CDF, random samples were determined for six distributions.

The first stage of the simulation studies was to estimate the bias and variance of the three bootstrap median estimators $\hat{\xi}_{0.5n}^{1*}$, $\hat{\xi}_{0.5n}^{2*}$, and $\hat{\xi}_{0.5n}^{3*}$, defined by formulas (13), (14) and (15). The bias and the variance were estimated according to the formulas (26) and (27) by the MC method.

In the second stage, confidence intervals for quantiles were determined using the following methods:

M1 – exact percentile method and estimator $\hat{\xi}_{0.5n}^{1*}$,

M2 – exact percentile method and estimator $\hat{\xi}_{0.5n}^{1*}$,

M3 – exact percentile method and estimator $\hat{\xi}_{0.5n}^{3*}$,

M4 – BC_a method (Efron and Tibshirani, 1993 p. 185) and estimator $\hat{\xi}_{0.5n}^{3*}$,

- M5 – the adjacent spacings method (Nagaraja and Nagaraja, 2020 p. 89) (calculations were made for several combinations of parameters s and t ⁶, choosing those for which confidence intervals were narrowest),
- M6 – using the limit distribution of the order statistics corresponding to the sample p -quantile (formula (8)),
- M7 – using the limit distribution of p -quantile (formula (7)).

The widths and coverage probabilities were used to compare confidence intervals constructed with different methods, calculated according to (28) and (29). The use of the M7 method requires a comment. The confidence intervals constructed using the M1 to M6 methods were estimated by the MC method based on R replications. The intervals constructed using the M7 method (from the limit distribution of p -quantile) are calculated from formula (7). One may suspect (especially for large samples) that the confidence intervals determined in this way are close to the real ones and may constitute a reference point for the intervals obtained with other methods. Note, however, that the use of formula (7) requires knowledge of ξ_p and $f(\xi_p)$. In fact, we know them very rarely.

In Figure 1, the bias of the bootstrap median estimators is given, depending on the sample size (for $n = 10, 15, \dots, 205$). The bias was calculated by the MC method based on $R = 2,000$ samples from six distributions. To improve the readability of the graph, the data series are presented as continuous lines.

If np is not integer (which in the case of the median corresponds to the odd n), the tree bootstrap median estimators are based on the same order statistics, so they are the same. Estimators differ when n is even. The case n odd was extracted as a separate data series to avoid oscillations when the sample size changes from even to odd. This was made because those oscillations would completely obscure the image (as in Parrish (1990 p. 253)). It is obvious that as the sample size increases, the bias on all estimators usually decreases (if bias jumps are omitted when the sample size changes from even to odd and vice versa). This does not mean, however, that the increase in n in the case of simulation by the MC method is always accompanied by a decrease in bias. The possible increase in bias results from the random selection of R samples.

The estimator $\hat{\xi}_{0.5n}^{3*}$ shows the smallest jumps in the bias with the change in the sample size from odd to even (and vice versa). The data series marked as E123 and E3 for all distributions almost coincide.

When samples came from right asymmetry distributions (for even n), the absolute value of the bias of the estimator $\hat{\xi}_{0.5n}^{1*}$ was usually the smallest, while that of the

⁶ Five parameter combinations were used: $s = 1$ and $t = 2$, $s = 2$ and $t = 1$, $s = 2$ and $t = 2$, $s = 2$ and $t = 3$, $s = 3$ and $t = 2$. The narrowest confidence intervals were obtained for the last two variants in all simulation experiments. The combination of $s = 3$ and $t = 2$ was best in the case of samples from right asymmetry distributions, while the combination of $s = 2$ and $t = 3$ in the case of samples from left asymmetry and symmetrical distributions.

estimator $\hat{\xi}_{0.5n}^{2*}$ the largest. When samples came from left asymmetry distributions, the opposite was true usually – the absolute value of the bias of the estimator $\hat{\xi}_{0.5n}^{2*}$ was the smallest, while that of the estimator $\hat{\xi}_{0.5n}^{1*}$ the largest. When samples came from symmetrical distributions, the absolute value of the bias of the estimator $\hat{\xi}_{0.5n}^{3*}$ was the smallest for almost all n , while that of the estimator $\hat{\xi}_{0.5n}^{1*}$ or $\hat{\xi}_{0.5n}^{2*}$ was the largest.

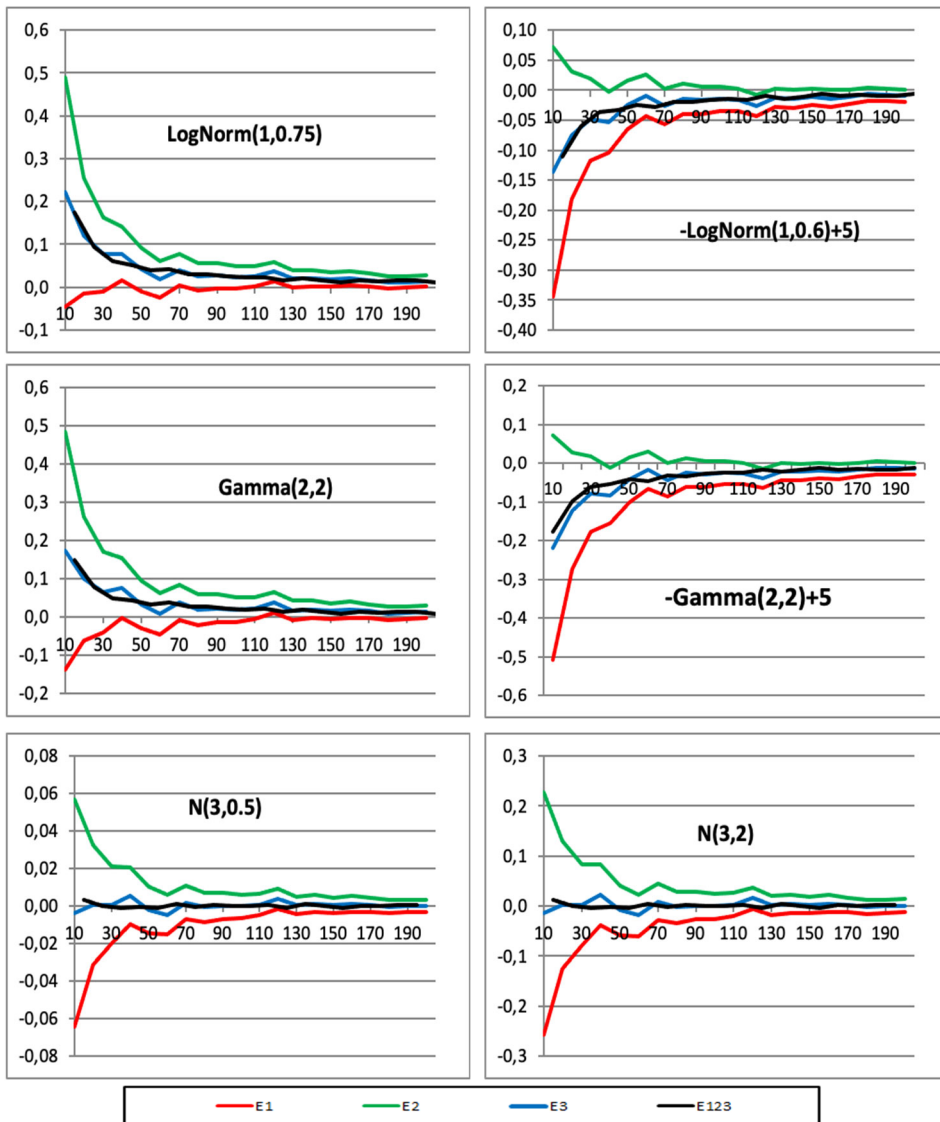


Figure 1: The bias of the bootstrap median estimators depending on the sample size ($n = 10, 15, \dots, 205$), calculated by the MC method. Note: In the charts, the data series marked E1, E2 and E3 correspond to the estimators $\hat{\xi}_{0.5n}^{1*}$, $\hat{\xi}_{0.5n}^{2*}$, and $\hat{\xi}_{0.5n}^{3*}$ for even n , while E123 corresponds to all estimators for odd n .

When samples came from right asymmetry distributions, the expected value of almost all estimators is greater than the median (except for even n and the estimator $\hat{\xi}_{0.5n}^{1*}$). When samples came from left asymmetry distributions, the expected value of almost all estimators is less than the median (except for even n and the estimator $\hat{\xi}_{0.5n}^{2*}$). When samples came from symmetrical distributions, the bias oscillates around zero – except for even n and the estimator $\hat{\xi}_{0.5n}^{1*}$ (always negative bias) or the estimator $\hat{\xi}_{0.5n}^{2*}$ (always positive bias).

Figure 2 shows the variance of the bootstrap median estimators depending on the sample size. The graphs were prepared only for small samples ($n = 10, 11, \dots, 35$). The differences in the case of large samples were very small. When samples came from symmetrical distributions, the variance of the $\hat{\xi}_{0.5n}^{3*}$ estimator was the smallest. When samples came from right asymmetry distributions, the variance of the $\hat{\xi}_{0.5n}^{2*}$ estimator was the biggest. When samples came from left asymmetry distributions, the variance of the $\hat{\xi}_{0.5n}^{1*}$ estimator was the biggest.

Figure 3 presents the width of 0.95 median confidence intervals depending on the sample size. The intervals were calculated by the MC method (the M1-M6 methods) and using the limit distribution (the M7 method). The graphs were made up only for small samples ($n = 10, 15, \dots, 35$). For large samples, the widths of confidence intervals constructed with different methods are very similar (except those obtained using the M5 method).

Confidence intervals constructed with the M7 method were usually narrowest, especially for samples from left asymmetry distributions and large samples (n above 115) from symmetrical distributions. Interval widths for M7, M4, and M3 (but only for even n) are very similar. If n was even, narrower confidence intervals were usually obtained using the M3 method rather than using the M4 for large samples (n above 180) and samples from left asymmetry distribution.

There are jumps in the widths of confidence intervals constructed with the exact bootstrap estimators (that is for M1, M2, and M3 methods) when n changes from even to odd. This is due to the changing the rank of the order statistic used as an estimator (we do not observe it for other methods). The narrowest confidence intervals were obtained by the M3 method, regardless of the distribution asymmetry type the samples came from. This is because the estimator based on two order statistics has much more realizations than when based on one order statistic only. If the samples came from right asymmetry distribution, narrower confidence intervals were obtained with the M1 method than with the M2. If the samples came from left asymmetry distribution, the effect was opposite. If the samples came from symmetrical distributions, the M1 method gave the narrower intervals for about half of the cases and the M2 for the other half. These conclusions are similar to those obtained for the variance and apply of course only to cases when n is even.

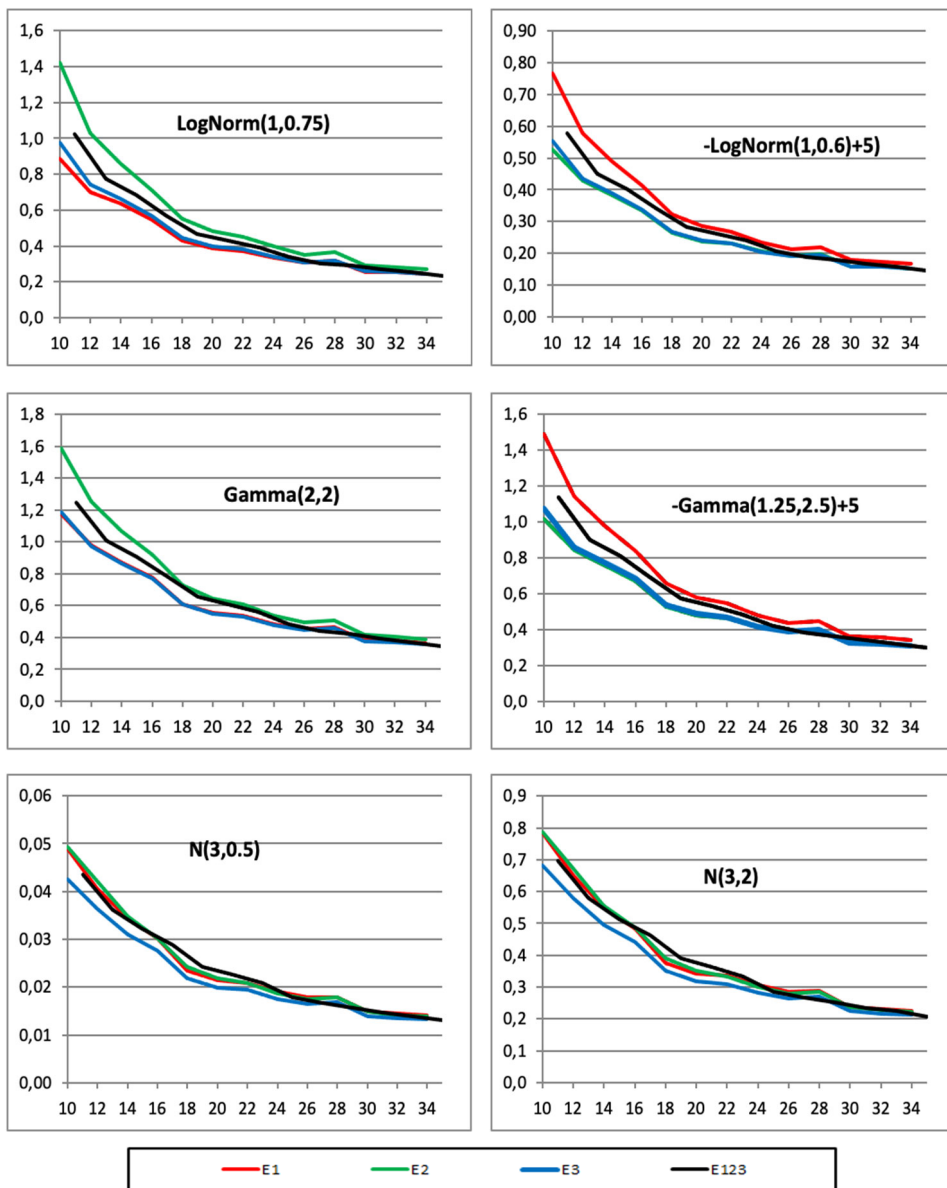


Figure 2: The variance of the bootstrap median estimators depending on the sample size ($n = 10, 11, \dots, 35$), calculated by the MC method. Note: as for Figure 1.

For almost all sample sizes, the widest confidence intervals were obtained by the M5 method (despite using a combination of parameters giving the best results). The authors of the method (Nagaraja, Nagaraja (2019 p. 75)) point out that although this method gives wider intervals than other methods, it can be used in the case of extreme quantiles even if the sample has only a few observations. The M6 method usually gave

wider confidence intervals than the M7, M4, and M3 methods (for n even), although the differences were small for large samples.

Figure 4 presents $1-\varphi$ (1-coverage probability) calculated for confidence intervals estimated by M1-M6 methods, depending on the sample size for a 0.95 confidence level. This probability was illustrated in three variants. The first variant covers all sample sizes, the second only odd sizes, and the third only even sizes. The charts are presented only for LogNorm(1.0.75) distribution. The results for the remaining distributions were very similar. The chart shows strong fluctuations in the coverage probability when the sample size changes, both for all sample sizes and separately for even and odd sizes. De Angelis et al. 1993 p. 526 believed the discrete nature of the bootstrap quantile estimators distributions to be the reason for the fluctuation. It should be noted that fluctuations occur for all methods. Therefore they can only result from the quality of random samples generated (for all methods the same pseudorandom numbers were used). Although relatively many samples were drawn ($R = 2,000$), the effect of the work of the pseudorandom number generator indicates its imperfection⁷. 4,000, 8,000, and 16,000 samples were drawn several times to check whether increasing the number of drawn samples would reduce the observed fluctuations. It turned out that the differences in the calculated $1-\varphi$ values in individual experiments were large. This means that a comparison across methods by simulation experiments requires the same conditions. It is advisable to use the same generated pseudorandom numbers in all experiments.

6. Conclusions

1. Information about the asymmetry direction of the distribution the sample came from may be a valuable indication when choosing a bootstrap quantile estimator (when np is an integer). The sample skewness coefficient can be used for this purpose. If the estimator is based on a single order statistic and a sample comes from a right asymmetry distribution, the order statistic of the rank np used as an estimator gives a smaller bias and narrower confidence intervals. If a sample comes from a left asymmetry distribution, an order statistic of the rank $np + 1$ is a better estimator. Most asymmetric distributions used in statistical experiments have a right asymmetry distribution. This is why the most common quantiles and sample quantiles are defined as left quantiles. For the distributions with left asymmetry, a right quantile would be more appropriate.

⁷ When using simulation methods using random numbers, one should take into account the limited possibilities of pseudorandom number generators. It is worth conducting experiments using various pseudorandom number generators. Examples of such studies were presented by Sulewski 2019.

2. Quantile estimators in the form of single order statistics are very simple to apply (they have the same PDF values for the same sample sizes). To use a linear combination of two order statistics as the estimator requires more effort of calculations. As the research shows, this effort pays off - the estimated confidence intervals are narrower, and the coverage probability is closer to the assumed

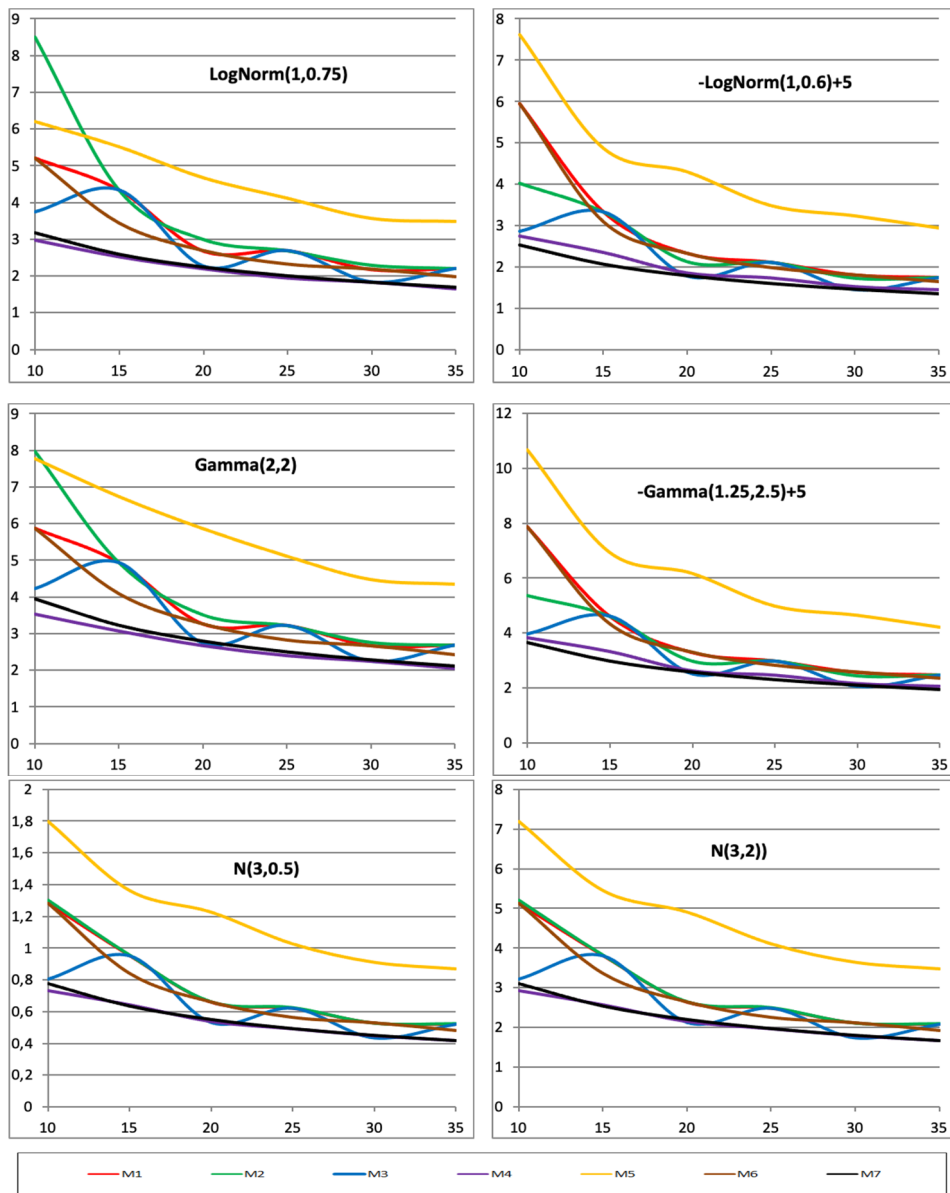


Figure 3: The width of median confidence intervals ($1-\alpha = 0.95$) depending on the sample size ($n = 10, 15, \dots, 35$).

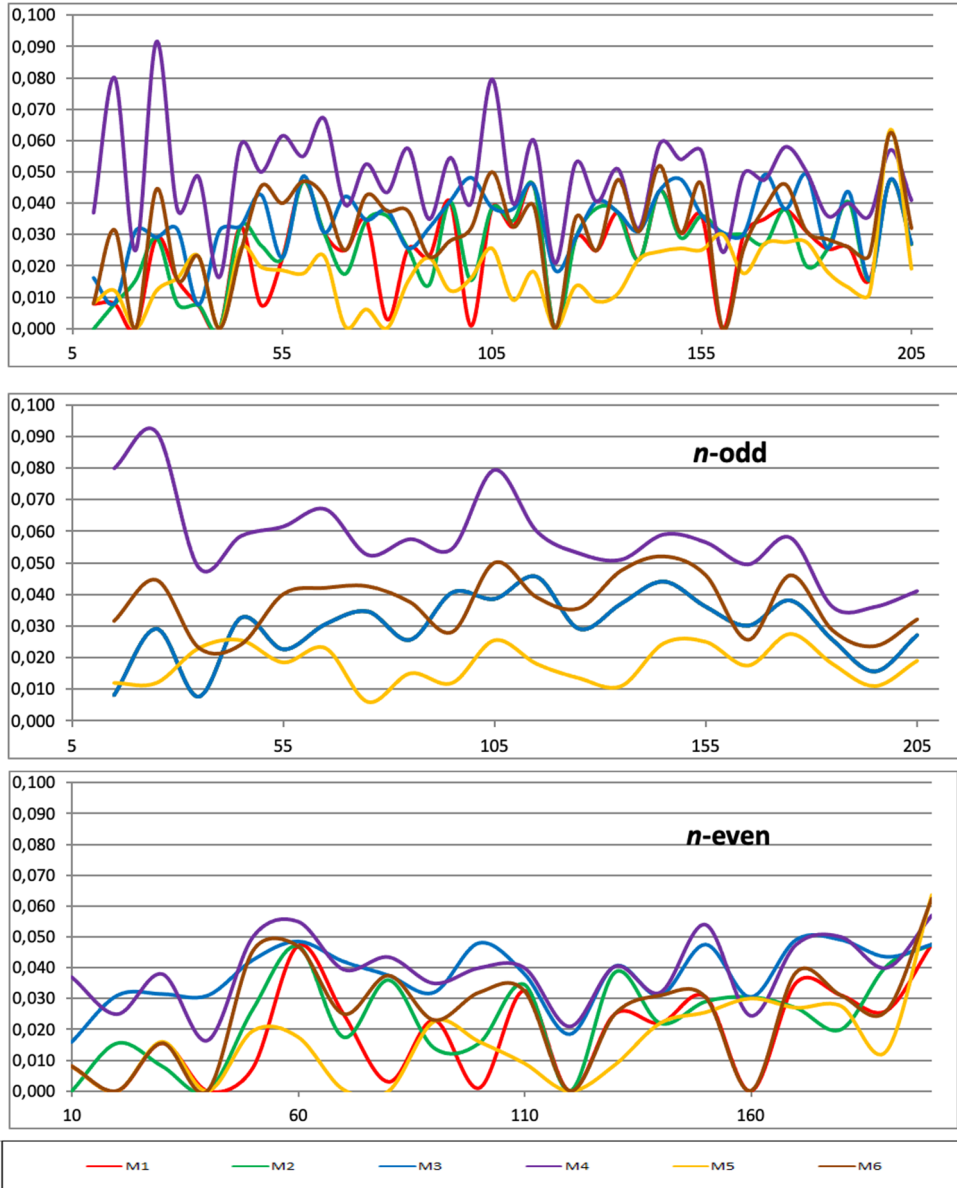


Figure 4: $1-\phi$ depending on the sample size ($n = 10, 15, \dots, 205$) for a 0.95 confidence level. Samples came from the LogNorm(1.0.75) distribution. Note: The top chart is for all numbers, the middle chart for odd n , and the bottom chart for even n .

confidence level. The possibility to construct narrower confidence intervals results from a bigger number of realizations of the estimator based on two order statistics. When np is not an integer, you may consider using the estimator as a linear combination of the three order statistics of the ranks $[np]$, $[np] + 1$, and $[np] + 2$.

The algorithm for calculating the distribution of such an estimator would be similar to the algorithm given in Section 2. Also, in this case, the probability that the given elements of the primary sample will occur on three positions: $[np]$, $[np] + 1$, and $[np] + 2$ in the ordered resample, is the same for all samples without repetition of the same size. This makes it possible to construct statistical tables one can use to compute the exact distribution of the estimator for a given primary sample (as it is possible for a combination of two order statistics).

3. There is no need to use the percentile method with resampling for interval estimation of quantiles. The application of the exact percentile method is much simpler. When one uses an estimator based on a single order statistic, it is known in advance which elements of the ordered primary sample constitute the limits of the confidence interval. When one uses an estimator based on two order statistics, the computational effort resulting from sorting all its possible realizations is probably comparable with the time needed to sort its realizations determined from the drawn resamples.⁸
4. The coverage probability fluctuations (on changes in the sample size) result from the limited capabilities of the pseudorandom number generator. One can conclude so because fluctuations occur for all methods. The conducted experiments indicate that increasing the number of repetitions in Monte Carlo simulations does not reduce the fluctuations. This conclusion was made based on experiments with $R = 2,000, 4,000, \text{ and } 8,000$. This means that it is better to use the same samples when you compare different estimation methods.
5. The bias and the variance of the bootstrap median estimators, as well as the width of the median confidence intervals were estimated using the MC method. Fluctuations of these parameters resulted mainly from the change in rank of order statistics used as estimators when the sample size changed from even to odd. This fact was particularly clear for the estimators in the form of a single order statistic. If even and odd samples are considered separately, there are no fluctuations. There are also no fluctuations in the width of the confidence intervals estimated with the other methods. This is because the discussed measures are calculated as average from all replications. The coverage probability is determined for all R repetitions.

The research conducted and presented in the article is based on a limited set of distributions. However, one can assume that conclusions can be generalized for their wider collection. Only one pseudorandom number generator was used – an Excel generator. The results showed that it is worth researching various pseudorandom numbers generators and examining their impact on the quality of the Monte Carlo simulations.

⁸ For a sample with 50 elements, the number of realizations of the estimator based on two order statistics is maximally equal to $50 + 49 \cdot 25 = 1275$.

References

- Altman, E. I., (1968). Financial Ratios, Discriminant Analysis and the Prediction of the Corporate Bankruptcy. *The Journal of Finance*, vol. 23, pp. 589–609.
- Chen, J. H., Williams, M., (1999). The determinants of business failures in the US low-technology and high-technology industries. *Applied Economics*, vol. 31, pp. 1551–1563.
- European Commission, (2003). BEST project on restructuring, bankruptcy and a fresh start: Final report of the expert, Enterprise Directorate-General.
- Bahadur, R. R., (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, vol. 37(3), pp. 577–580.
- Bickel, P. J., Freedman, D. A., (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, vol. 9(9), pp. 1196–1217.
- David, H. A., Nagaraja, H. N., (2003). Order Statistics, Wiley & Sons, Inc.
- De Angelis, D., Hall, P., Young, G. A., (1993). A note on coverage error of bootstrap confidence intervals for quantiles. *Math. Proc. Camb. Phil. Soc.*, vol. 114(3), pp. 517–531.
- Efron, B., (1979). Bootstrap Methods: Another look at the jackknife. *The Annals of Statistics*, vol. 7, no. 1, 1–26.
- Efron, B., (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.*, vol. 82(39), pp. 171–185.
- Efron, B., Tibshirani, R. J., (1993). An introduction to the Bootstrap, New York: Chapman & Hall.
- Evans, D. L., Leemis, L. M., Drew, J. H., (2006). The distribution of order statistics for discrete random variables with applications of bootstrapping. *Journal on Computing*, vol. 18(1), pp. 19–30.
- Falk, M., Kaufmann, E., (1991). Coverage probabilities of bootstrap-confidence intervals for quantiles. *Ann. Statist.*, vol. 19(1), pp. 485–495.
- Falk, M., Reiss, R. D., (1989). Weak convergence of smoothed and nonsmoothed bootstrap quantile estimates. *Ann. Probab.*, vol. 17(1), pp. 362–371.
- Fisher, N. I., Hall, P., (1991). Bootstrap algorithms for small samples. *Journal of Statistical Planning and Inference*, vol. 27, pp. 157–169.

- Hettmansperger, T. P., Sheather, S. J., (1986). Confidence intervals based on interpolated order statistics. *Statist. Probab. Lett.*, vol. 4(2), pp. 75–79.
- Hutson, A. D., (2002). A semi-parametric quantile function estimator for use in bootstrap estimation procedures. *Stat. Comput.*, vol. 2(4), pp. 331–338.
- Hutson, A. D., Ernst, M. D., (2000). The exact bootstrap mean and variance of an L-estimator. *Journal of the Royal Statistical Society: Series B*, vol. 62(1), pp. 89–94.
- Hyndman, R. J., Fan, Y., (1996). Sample quantiles in statistical packages. *The American Statistician*, vol. 50(4), pp. 361–365.
- Koehler, E., Brown, E., Haneuse, J.-P.A., (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *Amer. Statist.*, vol. 63(2), pp. 155–162.
- Kisielińska, J., (2013). The exact bootstrap method shown on the example of the mean and variance estimation. *Computational Statistics*, vol. 28(3), pp. 1061–1077.
- Maritz, J. S., Jarrett, R. G., (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, vol. 73(361), pp. 194–196.
- Nagaraja, C. H, Nagaraja, H. N., (2020). Distribution-free approximate methods for constructing confidence intervals for quantiles. *International Statistical Review*, vol. 88(1), pp. 75–100.
- Nyblom, J., (1992). Note on interpolated order statistics. *Statist. Probab. Lett.*, vol. 14(2), pp. 129–131.
- Parrish, R. S., (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, vol. 46, pp. 247–257.
- Pekasiewicz, D., (2015). Order statistics in estimation procedures and their applications in economic research, University of Lodz, Lodz.
- Serfling, R. J., (1980). Approximation theorems of mathematical statistics, John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore.
- Sulewski, P., (2019). Comparison of normal random number generators, *Wiadomości Statystyczne. The Polish Statistician*, vol. 64, pp. 5–31.
- Singh, K., (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.*, vol. 9(6), pp. 1187–1195.
- Wilcox, R. R., (2001). Fundamentals of modern statistical methods, Springer, New York.