STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 737–748

INTERVAL ESTIMATION OF HIGHER ORDER QUANTILES. ANALYSIS OF ACCURACY OF SELECTED PROCEDURES

Dorota Pekasiewicz[1]

ABSTRACT

In the paper selected nonparametric and semiparametric estimation methods of higher orders quantiles are considered. The construction of nonparametric confidence intervals is based on order statistics of appropriate ranks from random samples or from generated bootstrap samples. Semiparametric bootstrap methods are characterized by double bootstrap simulations. The values of bootstrap sample below the prearranged threshold are generated by the empirical distribution and the values above this threshold are generated by the distribution based on the asymptotic properties of the tail of the random variable distribution. The results of the study allow one to draw conclusions about the effectiveness of the considered procedures and to compare these methods.

Key words: accuracy of estimation, order statistic, percentile bootstrap method, quantile, semiparametric bootstrap method, Value at Risk.

1. Introduction

Quantiles of a random variable distribution are used in different kinds of economic and financial research. They are applied in defining, for example, measures of poverty and wealth in the analysis of population income and Value at Risk measure in the studies of market risk. Value at Risk is defined as p-quantile of random variable being the value of losses from investments.

Nonparametric and semiparametric quantile estimation methods are the subject of interest when the quantile order is greater than 0.9. In the group of nonparametric procedures, bootstrap and non-bootstrap methods are considered. One of them is the percentile bootstrap method (Efron, Tibshirani 1993), and the other is the best exact nonparametric method (Zieliński, Zieliński 2005). Bootstrap semiparametric methods are based on information about the tail distribution of the random variable (Pandey et al. 2003). The accuracy of the

[1] Department of Statistical Methods, University of Łódź, Poland. E-mail:pekasiewicz@uni.lodz.pl.

estimation, defined as the length of the confidence interval, is analysed. The accuracy of quantile estimation for selected distributions, for the nonparametric non-bootstrap procedure of quantile estimation, known as the best exact method, is determined analytically. But for bootstrap methods the simulation study is used. In the paper Pareto and Student t-distribution are considered. The selection of distributions is associated with the possibility of choosing these parameters for which the distribution is characterized by a thin or fat tail. Simulation methods allow one to estimate the probability that the confidence interval includes the real value of the quantile, and additionally they allow to investigate whether this probability is approximately equal to the prearranged confidence coefficient. The application of the semiparametric methods require estimating the generalized Pareto distribution parameters. This distribution is used in approximation of the tail of the random variable distribution. In the paper, two methods of estimating the generalized Pareto distribution parameters are considered. One of them is the probability weighted moments method based on the classical empirical distribution and the other is the probability weighted moments method based on the level crossing empirical distribution (Huang, Brill 1999).

2. The best exact nonparametric estimation of quantile

Let us assume that we investigate a population with regard to random variable *X* with unknown continuous distribution *F*. Let $X_1, X_2, ..., X_n$ be a simple random sample drawn from this population and $1-\alpha$ be the fixed confidence coefficient.

Nonparametric interval estimation of quantile Q_p of order $p \in (0, 1)$ is associated with a random variable *K*, which denotes the number of observations in the sample smaller than the quantile Q_p. Random variable *K* has binomial distribution with the probability function:

$$P(K=k)=\binom{n}{k}(p)^k(1-p)^{n-k} \qquad \text{for} \quad k = 0,\ 1,2,\ \ldots,\ n. \tag{1}$$

Let $X_{(r)}^{(n)}$ and $X_{(s)}^{(n)}$ denote order statistics of rank *r* and *s*, $(1 \le r \le s \le n,$ $r, s \in N)$ respectively. The probability that the value of the quantile Q_p is in the interval $\left(X_{(r)}^{(n)}, X_{(s)}^{(n)}\right)$ is calculated by the formula:

$$P\left(X_{(r)}^{(n)} \le Q_p \le X_{(s)}^{(n)}\right)=\sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i}. \tag{2}$$

The values of order statistics are determined so that the right side of this formula is equal to the fixed confidence coefficient, i.e. $\sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i}=1-\alpha.$ Sometimes it results in an unequivocal confidence interval, especially for small sample sizes, when choosing ranks of order statistics to obtain the confidence interval for the *p*th quantile at the confidence level $1-\alpha$ is impossible. In these cases it is necessary to randomize (Zieliński 2008):

$$\lambda P\left(X_{(r)}^{(n)}\le Q_p\le X_{(s)}^{(n)}\right)+(1-\lambda)P\left(X_{(r')}^{(n)}\le Q_p\le X_{(s')}^{(n)}\right)=1-\alpha. \tag{3}$$

We calculate λ and we take $\left(X_{(r)}^{(n)},X_{(s)}^{(n)}\right)$ as the confidence interval for Q_p with probability λ or $\left(X_{(r')}^{(n)},X_{(s')}^{(n)}\right)$ with probability $1-\lambda$. The obtained interval is not always symmetric under the value of the quantile estimator.

The accuracy of the interval estimation is given by the formula:

$$d=\lambda\left(E\left(X_{(s)}^{(n)}\right)-EX_{(r)}^{(n)}\right)+(1-\lambda)\left(E\left(X_{(s')}^{(n)}\right)-EX_{(r')}^{(n)}\right), \tag{4}$$

where

$$E\left(X_{(k)}^{(n)}\right)=\int_{-\infty}^{\infty}xg_{k;n}(x)dx=\frac{n!}{(k-1)!(n-k)!}\int_{-\infty}^{\infty}x[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)dx\ ,$$

$$k=r,s,r',s'. \tag{5}$$

Then, for the random variable with known distribution *F*, it is possible to calculate the precision of quantile estimation.

In Table 1 the minimum sample sizes for selected quantiles and different confidence coefficients are presented. In Table 2 are shown the ranks of order statistics allow one to obtain confidence interval for quantile on the confidence level approximately equals 0.95 for selected *p*-quantiles and selected sample. sizes.

Table 1. The minimum sample sizes for nonparametric estimation of *p*-quantiles

p	$1-\alpha$				
	0.9	0.925	0.95	0.975	0.99
0.80	11	12	14	15	21
0.90	22	25	29	23	44
0.95	45	51	59	72	90
0.99	230	258	299	368	459

Source: Own calculations.

Table 2. The ranks of order statistics in estimation of Q_p

p	Sample sizes					
	300		600		1000	
0.80	227, 254 (0.9491)	228, 256 (0.9515)	459, 498 (0.9495)	460, 499 (0.9527)	773, 823 (0.9479)	774, 824 (0.9506)
0.90	261, 281 (0.9451)	261, 282 (0.9524)	528, 560 (0.9499)	528, 561 (0.9509)	884, 923 (0.9494)	884, 924 (0.9514)
0.95	277, 292 (0.9491)	278, 293 (0.9548)	561, 582 (0.9467)	561, 583 (0.9512)	938, 965 (0.9474)	937, 964 (0.9504)
0.99	291, 300 (0.9499)	290, 300 (0.9507)	590, 599 (0.9412)	590, 600 (0.9558)	985, 998 (0.9495)	985, 999 (0.9517)

Source: Own calculations.

3. The percentile bootstrap method of quantile estimation

The next analysed procedure of quantile estimation is the percentile bootstrap method (Domański, Pruska 2000).

Based on the simple random sample $X_1, X_2, ..., X_n$ we generate *N* bootstrap samples $X_1^*, X_2^*, ..., X_n^*$, from the bootstrap distribution:

$$P\left(X^* = x_i\right) = \frac{1}{n}, \quad \text{for } i = 1, 2, \ldots, n, \tag{6}$$

where $x_1, x_2, ..., x_n$ are elements of the sample $X_1, X_2, ..., X_n$.

Next, for each bootstrap sample we compute the quantile $X_{p,k}^*$, where k =1, 2, …, *N*. Therefore, after *N* replications we get the sequence of ordered quantiles (sorted from least to greatest) $X_{p,(1)}^*, ..., X_{p,(N)}^*$, which allow one to approximate the distribution of quantile Q_p. Using this sequence we determine the percentiles of ranks $N\frac{\alpha}{2}$ and $N - N\frac{\alpha}{2}$.

The confidence bootstrap interval for Q_p has the following form:

$$P\left(X_{\frac{\alpha}{2}}^* < Q_p < X_{1-\frac{\alpha}{2}}^*\right) \approx 1 - \alpha, \tag{7}$$

where statistics $X^*_{\frac{\alpha}{2}}$ and $X^*_{1-\frac{\alpha}{2}}$ are the percentiles of ranks $N\frac{\alpha}{2}$ and $N-N\frac{\alpha}{2}$, respectively.

The number of repetitions N is selected so as $\frac{N\alpha}{2}$ and $N-\frac{N\alpha}{2}$ are integers.

4. Semiparametric bootstrap methods of quantile estimation

Semiparametric bootstrap estimation methods are characterized by double bootstrap simulations, i.e. *n–k* values of bootstrap sample below the fixed threshold *u* are generated using empirical distribution F_n, but *k* values above this threshold are generated using the distribution which takes into account asymptotic properties of tail distribution (Pandey et al. 2003).

In this case the bootstrap distribution has the form:

$$F^*(x|u)=\begin{cases}(1-F_n(u))F_0(x)+F_n(u), & \text{for} \quad x>u,\\ F_n(x), & \text{for} \quad x\le u,\end{cases} \tag{8}$$

where F_n is the empirical distribution and F_0 is the generalized Pareto distribution.

The generalized Pareto distribution $GPD(\xi,\beta)$ is expressed by the formula:

$$F_0(x)=\begin{cases}1-\left(1+\frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}} & \text{for} \quad \xi\ne 0,\\ \\ 1-\exp\left(-\frac{x}{\beta}\right) & \text{for} \quad \xi=0,\end{cases} \tag{9}$$

so the estimated distribution (8) has the following forms:
for $\xi \ne 0$:

$$\hat{F}(x|u)=1-\frac{k}{n}\left(1+\hat{\xi}\frac{x-u}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}} \quad \text{for } x>u, \tag{10}$$

and for $\xi = 0$:

$$\hat{F}(x|u)=1-\frac{k}{n}\exp\left(-\frac{x-u}{\hat{\beta}}\right) \tag{11}$$

where *k* is the number of elements of random sample greater than the fixed threshold and $\hat{\xi},\hat{\beta}$ are estimators of parameters ξ,β.

Thus, *p*-quantile has the form:

$$X_{p;n} = \begin{cases} u + \dfrac{\hat{\beta}}{\hat{\xi}}\left(\left(\dfrac{n}{k}(1-p)\right)^{-\hat{\xi}} - 1\right) & \text{for } \hat{\xi} \neq 0, \\ u - \hat{\beta}\ln\left(\dfrac{n}{k}(1-p)\right) & \text{for } \hat{\xi} = 0. \end{cases} \qquad (12)$$

The values of parameters ξ and β can be estimated by moments method, probability weighted moments method, or maximum likelihood method. The initial estimation of ξ can be obtained by generalized Hill estimator, moment estimator, Peng estimator or *W*–estimator (see Pekasiewicz 2015).

Two methods of estimation of parameters ξ, β are considered. In the first method, called semiparametric method I, parameters are estimated by probability weighted moments method (Landwehr, et al. 1979) and in the second one, called semiparametric method II – by modified probability weighted moments method, which is proposed in Pekasiewicz (2015).

In the method I, the estimators of parameters ξ, β have the following forms:

$$\hat{\xi}^{m(I)} = 2 - \frac{k\bar{Y}}{k\bar{Y} - 2\sum_{i=1}^{k}\frac{k-i}{k-1}Y_{(i)}^{(k)}}, \qquad (13)$$

$$\hat{\beta}^{m(I)} = \frac{2\sum_{i=1}^{k}\frac{k-i}{k-1}Y_{(i)}^{(k)}\bar{Y}}{k\bar{Y} - 2\sum_{i=1}^{k}\frac{k-i}{k-1}Y_{(i)}^{(k)}}, \qquad (14)$$

where $Y = X - u.$

In the semiparametric procedure with modified probability weighted moments method of estimation of ξ and β, the level crossing empirical distribution is used (Huang, Brill 1999):

$$F_k(y) = \begin{cases} 0 & \text{for } \; y < y_{(1)}^{(k)}, \\ \dfrac{1}{2}\left[1 - \dfrac{k-2}{\sqrt{k(k-1)}}\right] & \text{for } \; y_{(1)}^{(k)} \le y < y_{(2)}^{(k)}, \\ \dfrac{1}{2}\left[1 - \dfrac{k-2i}{\sqrt{k(k-1)}}\right] & \text{for } \; y_{(i)}^{(k)} \le y < y_{(i+1)}^{(k)}, \; i = 2,3,\ldots,k-1, \\ 1 & \text{for } \; y \ge y_{(k)}^{(k)}. \end{cases} \qquad (15)$$

The estimators of parameters ξ, β are the following:

$$\hat{\xi}^{m(II)} = 2 - \frac{\bar{Y}}{\bar{Y} - 2\nu}, \tag{16}$$

$$\hat{\beta}^{m(II)} = \frac{2\nu\bar{Y}}{\bar{Y} - 2\nu}, \tag{17}$$

where

$$\nu = \frac{1}{k}\left(\frac{1}{2} + \frac{k-2}{2\sqrt{k(k-1)}}\right)Y_{(1)}^{(k)} + \frac{1}{k}\sum_{i=2}^{k-1}\left(\frac{1}{2} + \frac{k-2i}{2\sqrt{k(k-1)}}\right)Y_{(i)}^{(k)}. \tag{18}$$

5. Analyses of interval quantile estimation accuracy

The aim of the study is to compare the length of the confidence interval obtained by considered nonparametric and semiparametric methods. The accuracy of the best exact confidence interval is calculated by formula (4) and bootstrap procedures are analysed by simulation methods. The presented procedures are applied in the estimation of quantiles of orders higher than 0.9 for selected distributions.

The following distributions with fat tails are considered:

− Pareto $Pa(\theta, a)$, where $\theta, a > 0$,

− Student *t*-distribution S(*k*), where *k* is degrees of freedom.

Depending on parameters the distributions are characterized by the expected or non-expected value.

Quantiles of higher orders are estimated by the best exact nonparametric method, the percentile bootstrap method and two semiparametric bootstrap procedures (method I, method II). In the case of semiparametric bootstrap methods it is necessary to use information about the values from the tail distribution and the tail estimation using generalized Pareto distribution.

The construction of bootstrap nonparametric and semiparametric confidence intervals means that these methods can be studied and compared only using simulation analysis. The mean length of the confidence interval, and probability γ that the real value of the quantile is contained in the constructed interval are computed by repeating the estimation procedure 1000 times. This probability should be approximately equal to the predetermined confidence coefficient $1-\alpha$.

In the case of estimating higher order quantile, the changes in distribution parameters causes a significant change of the value of the quantile. The relationship between the values of selected quantiles and parameter *a* of Pareto distribution is presented in Figure 1. In Figure 2 the relationship between selected quantile values and the degree of freedom of Student *t*-distribution is shown.

The results of the analysis of 0.99-quantile estimation for Pareto and Student *t*-distribution are presented in Table 3 and Table 4. The random samples must be rather big (in the tables - 1000 elements), which allows one to estimate generalized Pareto distribution parameters with small mean squared errors (the number of elements above the threshold is equals 100).

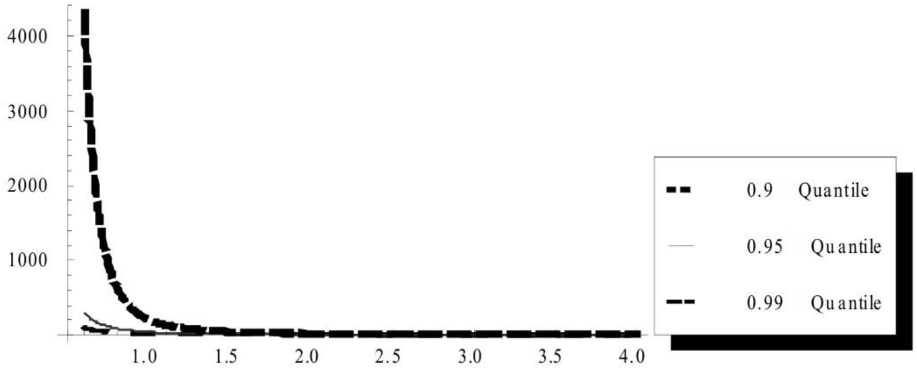

Figure 1. Relationship between higher order quantiles of Pareto distribution $Pa(2,a)$ and parameter *a*

Source: Own elaboration.

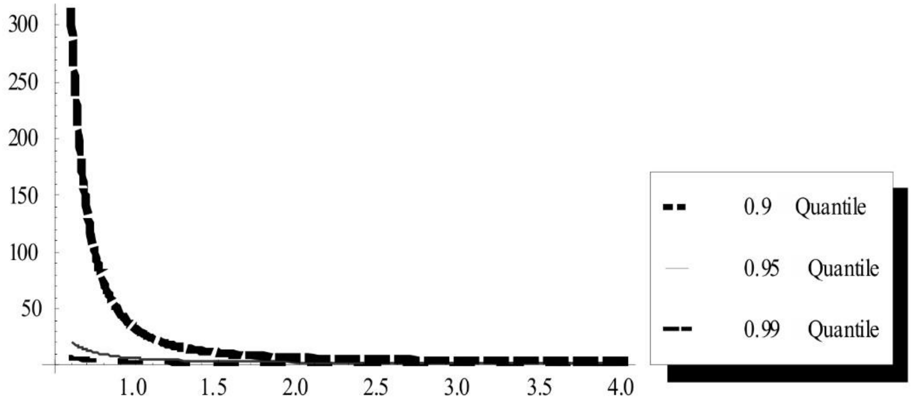

Figure 2. Relationship between higher order quantiles of Student *t*-distribution $S(k)$ and degree of freedom k

Source: Own elaboration.

Table 3. Accuracy d and probability γ of 0.99-quantile Pareto distribution estimation

Distribution	Best exact nonparametric method	Percentile method		Semiparametric method (I)		Semiparametric method (II)	
	d	d	γ	d	γ	d	γ
Pa(2, 1.25)	265.6520	86.7278	0.925	62.7672	0.921	61.8472	0.919
Pa(2, 1.5)	101.3190	37.0402	0.924	29.6164	0.926	30.0481	0.928
Pa(2, 1.75)	50.2623	20.0566	0.925	17.0014	0.946	16.9164	0.947
Pa(2, 2)	29.3498	12.6772	0.919	10.8831	0.926	12.6491	0.944
Pa(2, 2.25)	19.1083	8.7414	0.936	7.6362	0.943	7.6150	0.941
Pa(2, 2.5)	13.4312	6.3572	0.928	5.6160	0.942	5.7064	0.954
Pa(2, 2.75)	9.9807	4.8764	0.918	4.4066	0.956	4.4007	0.948
Pa(2, 3)	7.7489	3.8876	0.937	3.4652	0.949	3.5462	0.957
Pa(2, 3.25)	6.2154	3.1110	0.928	2.8326	0.954	2.9202	0.951
Pa(2, 3.5)	5.1188	2.6086	0,923	2.3731	0.949	2.4622	0.950
Pa(2, 3.75)	4.3070	2.2758	0.932	2,0576	0.945	2.0999	0.958
Pa(2, 4)	3.6884	1.9676	0.925	1.8047	0.953	1.8310	0.964

Source: Own calculation based on Mathematica 8.

The relative precision (the ratio of confidence length and the real value of quantile) in percentages is shown in Figure 3 and Figure 4.

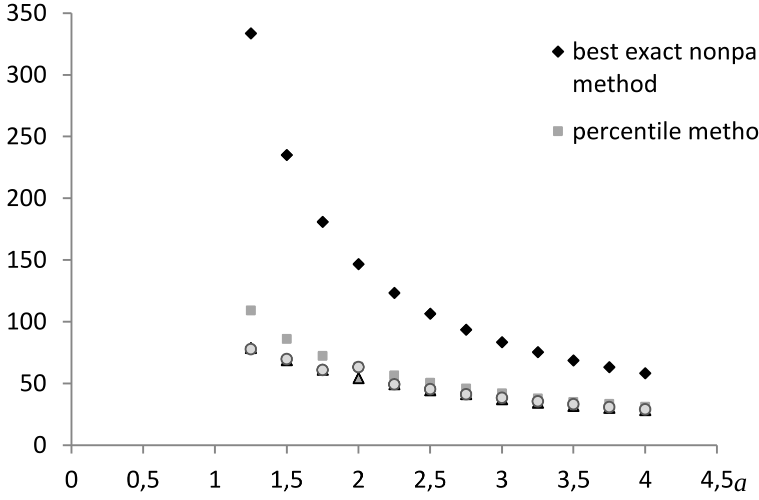

Figure 3. Relative precision of 0.99-quantile estimation for $Pa(2,a)$ distribution for considered methods

Source: Own calculation.

Table 4. Accuracy d and probability γ of 0.99-quantile Student t-distribution estimation

Distribution	Best exact nonparametric method	Percentile method		Semiparametric method (I)		Semiparametric method (II)	
	d	d	γ	d	γ	d	γ
$S(1.25)$	56.1500	18.8406	0.940	13.2522	0.916	13.0903	0.914
$S(1.5)$	26.4871	9.8648	0.921	7.9817	0.935	8.0148	0.940
$S(1.75)$	15.5615	6.3563	0.924	5.4952	0.940	5.5240	0.947
$S(2)$	10.4578	4.5764	0.910	4.0748	0.960	4.0828	0.957
$S(2.25)$	7.6799	3.4644	0.911	3.1792	0.945	3.2461	0.960
$S(2.5)$	5.9998	2.9216	0.926	2.6388	0.951	2.6384	0.953
$S(2.75)$	4.9027	2.4135	0.915	2.2014	0.950	2.2738	0.963
$S(3)$	4.1436	2.1148	0.940	1.9098	0.955	1.9402	0.948
$S(3.25)$	3.5942	1.8519	0.924	1.7452	0.959	1.7670	0.960
$S(3.5)$	3.1820	1.6880	0.932	1.5552	0.958	1.5813	0.954
$S(3.75)$	2.8635	1.5579	0.932	1.4434	0.963	1.4624	0.949
$S(4)$	2.6113	1.4072	0.921	1.3266	0.956	1.3604	0.961

Source: own calculation based on Mathematica 8.

The results of analysis imply that the application of bootstrap methods in estimation of quantile of Pareto distribution and Student t-distribution is more effective.

The choice of the quantile estimation method is important (see Figure 3 and 4) particularly for estimating the quantile heavy tailed distributions, i.e. $Pa(2, 1.25)$ or $S(1.25)$. It is associated with high values of distribution quantiles (see Figure 1 and 2).

In these cases the interval length obtained by the best exact nonparametric method (non-bootstrap procedure) is even three times longer than the bootstrap methods.

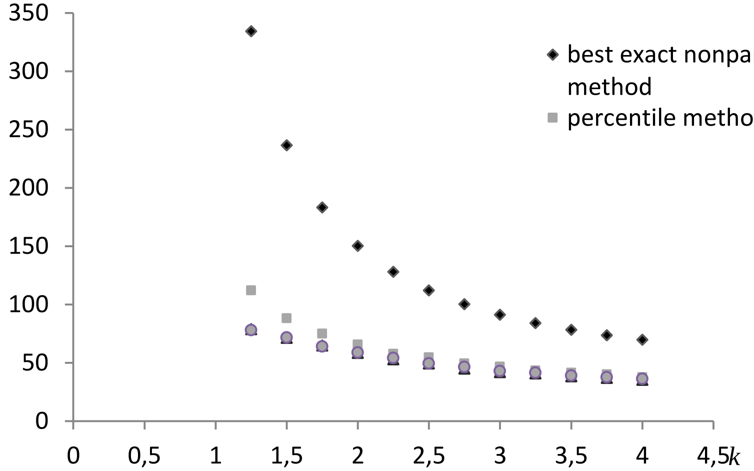

Figure 4. Relative precision of 0.99-quantile estimation for Student *t*-distribution for considered methods

Source: Own calculation.

5. Conclusions

In the paper different estimation procedures of higher order quantiles, including nonparametric and semiparametric methods, are considered. The application of bootstrap methods leads to confidence intervals which have smaller lengths than intervals derived from nonparametric, non-bootstrap methods. The best exact nonparametric confidence interval lengths are greater than the lengths of confidence interval obtained from the percentile bootstrap method. Semiparametric estimation methods allow one to get even shorter confidence intervals. Moreover, the probability that the confidence interval contains the real value of the distribution quantile is usually closer to the predetermined confidence level for semiparametric methods.

The generalized Pareto distribution parameters estimation method, which is used to approximate the tail distribution of the random variable, turns out to be less important in comparison with choosing the quantile estimation procedure.

The results obtained indicate that the choice of the estimation method is of greater importance when heavy tailed distribution quantiles are estimated.

The analysed procedures may be used to estimate measures based on higher order quantiles and may be applied in different economic and financial research.

REFERENCES

DOMAŃSKI, C., PRUSKA, K., (2000). Nieklasyczne metody statystyczne, [Non-classical Statistical Methods], Polskie Wydawnictwo Ekonomiczne, Warszawa.

EFRON, B., TIBSHIRANI, R. J., (1993), An Introduction to the Bootstrap, Chapman & Hall, New York.

HUANG, M. L., BRILL, P. H., (1999). A Level Crossing Quantile Estimation Method, Statistics & Probability Letters, 45, pp. 111–119.

LANDWEHR, J. M., MATALAS, N. C., WALLIS, J. R., (1979). Probability Weighted Moments Compared with Some Traditional Techniques in Estimating Gumbel Parameters and Quantiles, Water Resources Research 15(5), pp. 1055–1064.

PANDEY, M. D., VAN GELDER, P. H. A. J. M., VRIJLING, J. K., (2003). Bootstrap Simulations for Evaluating the Uncertainty Associated with Peaks-over-Threshold Estimates of Extreme Wind Velocity, Environmetrics, 14, pp. 27–43.

PEKASIEWICZ, D. (2015). Statystyki pozycyjne w procedurach estymacji i ich zastosowania w badaniach społeczno-ekonomicznych, [Order Statistics in Estimation Procedures and their Applications in Economic Research], Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

ZIELIŃSKI, R, ZIELIŃSKI, W., (2005). Best Exact Nonparametric Confidence Intervals for Quantiles, Statistics, 34, pp. 353–355.

ZIELIŃSKI, W., (2008). Przykład zastosowania dokładnego nieparametrycznego przedziału ufności dla VaR, [Example of Application of Exact Nonparametric Interval Confidence for VaR] Metody Ilościowe w Badaniach Ekonomicznych, 9, pp. 239–244.