

Agnieszka LACH¹
Łukasz SMAGA²

Comparison of the goodness-of-fit tests for truncated distributions

1. INTRODUCTION

The role of statistical testing has been the subject of discussion for years. An overview of this topic was given decades ago for instance in Cox et al. (1977). Nevertheless, the matter is still important, which confirms for example the paper on statistical testing in finance written by Kim, Ji (2015). As was mentioned in that paper, many statistical tests are used in practice with little consideration of their key characteristics as size and power. These characteristics should be intensively studied at least in simulations as, for example, in Pavia (2015), Górecki, Smaga (2015) or Orzeszko (2014). In this paper, we investigate the finite sample behavior of some goodness-of-fit testing procedures for truncated distributions known in the literature. Such behavior was not considered in the original paper introducing these tests. The paper is an extension of the results obtained in the bachelor thesis by Lach (2017).

The shape of a distribution in the tails is very important in many areas of science. Chernobai et al. (2015) adapted the standard goodness-of-fit tests for left-truncated distributions. The modifications of standard procedures help to take the decision, whether the tail belongs to a specified distribution or not. The tests were implemented in the R package `truncgof` (R Core Team, 2017; Wolter, 2012). The detailed description of their seven tests is given in section 2. Five of them are the commonly used standard tests with the modified null hypothesis cumulative distribution function. Following the original notation, these tests will be referred to as the AD^* (supremum Anderson-Darling), AD^{2*} (quadratic Anderson-Darling), KS^* (Kolmogorov-Smirnov), V^* (Kuiper) and W^{2*} (Cramér-von Mises) tests, respectively, in the remainder of the article. The other two tests are specially designed for the upper tails. They use the modified null hypothesis cumulative distribution function and the new weighing function. These are modified Anderson-Darling tests, which will be referred to as AD_{up}^* and AD_{up}^{2*} tests, respectively.

¹ Poznan University of Economics and Business, Faculty of Informatics and Electronic Economy, Operations Research Department, 10 Al. Niepodległości St., 61–875 Poznań, Poland, corresponding author – e-mail: agnlach1@gmail.com.

² Adam Mickiewicz University, Faculty of Mathematics and Computer Science, Department of Probability and Mathematical Statistics, 87 Umultowska St., 61–614 Poznań, Poland.

The tests by Chernobai et al. (2015) are often used in the literature, especially in the field of the operational risk calculation. Here, the choice of appropriate severity distribution is of crucial importance. In the process of calculating the aforementioned risk, Fischer, Jakob (2016) used a compound severity distribution, which involves dividing it into the body and the tail by a threshold. The Authors conclude that positive tempered α -stable distribution better fits empirical data in the tail than lognormal, Weibull, gamma and generalized gamma distributions. To assess goodness-of-fit of the distributions in the tail they used among others the AD_{up}^{2*} test for truncated distributions. Chernobai et al. (2006) considered the following severity distributions: exponential, lognormal, Weibull, Burr, generalized Pareto (GPD) and log α -stable. The null hypothesis that the cumulative distribution function belongs to truncated versions of the families of these distributions was verified by using the procedure described in Chernobai et al. (2015). The tests for truncated distributions were also used by Chernobai et al. (2010), who analyzed the effects of model misspecifications on Value-at-Risk and Conditional Value-at-Risk figures.

Examples of applications of the tests by Chernobai et al. (2015) can also be found in hydrology and social sciences. To estimate flood peaks, Brunner et al. (2017) used among others modified Anderson-Darling test for the upper tail to verify fitting of GPD and generalized extreme value distribution (GEV) to observed flood hydrographs. As was stated in the study, the test confirmed that the GPD fits well to the peak discharges and the GEV distribution fits well to the flood volumes. In the field of social sciences, Fagiolo et al. (2010) studied distributional properties of Italian household consumption expenditures. To study the tails of the distributions, they truncated distributions in several points and then they used standard truncated goodness-of-fit normality tests. Clementi et al. (2012) proposed a new model for income distribution: the κ -generalized distribution. As the fit in the right tail was of greater importance here, they decided to compare it with Singh-Maddala or Dagum type I distributions using upper tail goodness-of-fit tests.

In the majority of the studies listed above, the truncated tests were not the only ones, upon which the decisions were taken. However, it is clear that they had impact on the researchers' final decisions and that the range of possible applications of them is wide. Until now no studies concerning the size and the power of these tests for the left truncated distributions were published. The aim of this paper is to fill this gap.

The research of this paper is similar to that conducted by Pavia (2015). The main difference is that Pavia concentrated on complete distributions, while this paper refers to truncated ones. Pavia conducted the research for different sample sizes (10, 20, 50, 100, 200, 500). In this paper, the research is conducted for the sample of size 1000. Pavia verified the empirical sizes and the empirical powers of several goodness-of-fit tests available in the R packages, including five tests from the *truncgof* package (AD^* , AD^{2*} , KS^* , V^* and W^{2*}). As the Author was interested only in complete distributions, he omitted the tests from the *truncgof* package de-

signed for the upper tail. When the truncation point is not set, the tests from this package can be used also in the case of complete samples. According to the analysis of the sizes of the tests, the main conclusion was that most of the tests from the *truncgof* package are giving unacceptable results, especially for bounded distributions. Only the AD^* test from this package achieved acceptable rejecting rates for all examined distributions except for the uniform one. In case of the exponential distribution, the V^* test also gave reasonable results. When analyzing the power of the tests, in most of the examples the tests implemented in the *truncgof* package showed superior power over the rest of the tests taken into comparison. The results for the bounded distribution were again unacceptable.

The study in this paper is based on the artificial data generated from the distributions that are used to describe the tails of asset returns. The shape of the tails has great importance in the assessment of the risk. The origins of the studies on the distribution of asset returns dates back to the year 1900. At that time Louis Bachelier noticed, that according to the Central Limit Theorem the distribution of the asset returns in long term should be Gaussian (Haas, Pigorsch, 2009). That implies that the tails of the distributions should be thin and tend to zero faster than exponentially (Feller, 1950). This conception was prevailing until 1963, when Mandelbrot (1963) noticed fat tails of distributions of the cotton prices logarithms. One of the first distributions proposed to replace the normal distribution was the *t*-distribution with power decaying tails (Haas, Pigorsch, 2009). However, recent studies show that most of the asset returns have semi-heavy tails (Echaust, 2014; Piasecki, Tomasik, 2013). The power-exponential distribution could be proposed here as alternative. Depending on parameters, its tails can change from thinner tails than those of normal distribution to fat ones. Another example might be Weibull distribution, whose tails vary from thin to fat. The distributions mentioned in this paragraph were chosen for the research due to their historical meaning or the possibilities they offer. For the details of the Central Limit Theorem and these distributions the reader is referred to Krzyśko (2000) and Magiera (2005).

The remainder of this paper is organized as follows. In section 2, the tests for truncated distributions introduced in Chernobai et al. (2015) are presented. Section 3 contains the results of the simulation studies. Finally, section 4 draws some conclusions.

2. TESTS FOR TRUNCATED DISTRIBUTION

This section contains description of seven goodness-of-fit tests for truncated distributions, which were introduced in Chernobai et al. (2015). Five of these tests are modifications of standard goodness-of-fit tests. The remaining testing procedures are specifically constructed for upper tails of distributions. Before the description of the tests, short information about upper tails in finance is given.

Upper tails in finance are defined as $\overline{F}(x) = P(X > x)$, where x is sufficiently high, which means that $x \rightarrow \infty$ (Haas, Pigorsch, 2009). In case of the asset returns, even 5% is enough high to be set as the truncation point (Haas, Pigorsch, 2009). However, banks and other financial institutions may wish to define the tails in terms of the quantiles of the distribution. When calculating risk measures, like VaR or CVaR, usually quantiles of level 0.95, 0.975 or 0.99 are taken into account, although higher quantiles also appear (Haas, Pigorsch, 2009). On the other hand, when choosing an investment strategy, investors might be interested in much lower quantiles of distributions.

Chernobai et al. (2015) adapted the Kolmogorov-Smirnov, Kuiper, Cramér-von Mises and Anderson-Darling tests, which are standard goodness-of-fit tests, for truncated distributions. Anderson, Darling (1952) enabled giving different weights to specific parts of a distribution function, multiplying classical Kolmogorov and Cramér-von Mises statistics by the weight function $\Psi(x)$ (where $\Psi(x) \geq 0$ for $x \in [0,1]$). Anderson and Darling considered two weight functions: $\Psi(x) = 1$ and $\Psi(x) = 1/[x(1-x)]$. While for the first function test statistics reduce to the standard Kolmogorov and Cramér-von Mises statistics, the second function gives greater importance to the tails of the distribution function.

Let us assume that we have a sample $\mathbf{X} = (X_1, \dots, X_{n_c})'$ of i.i.d. variables with an unknown distribution function F . To formulate a goodness-of-fit problem for truncated distributions, Chernobai et al. (2015) used the appropriate distribution function for the truncated sample. Let F_0 denote distribution function for the complete sample and let H be the truncation point. The modified distribution function for the truncated sample is then defined by the following formula:

$$F_0^*(x) = \begin{cases} \frac{F_0(x) - F_0(H)}{1 - F_0(H)} & , \text{for } x \geq H, \\ 0 & , \text{for } x < H. \end{cases} \quad (1)$$

The complete sample of observations consists of n_c items. The ordered sample of observations $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n_c)}$ has the empirical distribution function (Krzyśko, 2004):

$$F_{n_c}(x; \mathbf{X}) = \frac{\#\{1 \leq j \leq n_c; X_j \leq x\}}{n_c}, \quad x \in R, \mathbf{X} \in R^{n_c}. \quad (2)$$

The difference between the values of the empirical distribution function for two neighboring points is equal to $1/n_c$. In case of left truncated distribution, the complete sample is the sum $n_c = m + n$, where m denotes the number of unknown observations below the truncation point and n is the number of observations equal to or greater than the truncation point. The empirical distribution function for the truncated sample is the same as for the complete sample, but the difference between the values of the empirical distribution function for two

neighboring points is equal to $1/n$. The empirical distribution function for the observed part of the whole population is then (Chernobai et al., 2015):

$$F_n(x; \mathbf{X}) = \begin{cases} F_0(H) & x < x_{(1)}, \\ \frac{j}{n}(1 - F_0(H)) + F_0(H) & x_{(j)} \leq x < x_{(j+1)}, j = 1, \dots, n-1, \\ 1 & x \geq x_{(n)}. \end{cases} \quad (3)$$

Thus, the null and alternative hypothesis can be formulated as follows:

$$\begin{aligned} H_0: F &= F_0^*, \\ H_1: F &\neq F_0^*. \end{aligned} \quad (4)$$

To test the null hypothesis against the alternative one, the null distributions of the test statistics (described below) are approximated by the Monte Carlo method. The detailed procedure for computing the corresponding p -values is as follows:

1. Compute the test statistic T_{obs} for the original data.
2. Generate a sample of n observations from the theoretical distribution function F_0^* . Each observation has to be greater than or equal to H .
3. Compute the test statistic T for the data generated in step 2.
4. Repeat steps 2 and 3 N times. Let T_1, \dots, T_N denote the obtained values of the test statistic.
5. Compute the p -value according to the formula $(1/N) \sum_{i=1}^N I(T_i \geq T_{obs})$, where $I(S)$ denotes the indicator function of a set S .

The null hypothesis is rejected, when the p -value is less than or equal to the nominal significance level α . Otherwise, we do not have any evidence to reject the null hypothesis. The asymptotic distributions of the test statistics considered in this paper are not known, which is one of the reasons of using the above procedure.

Following Chernobai et al. (2015), the test statistics applied to verify the null hypothesis are divided into three groups: (1) the supremum class, (2) the quadratic class, (3) the test statistics specifically designed to test goodness-of-fit in the upper tail.

The first group is made up of three modified statistics: Kolmogorov-Smirnov (KS^*), Kuiper (V^*) and Anderson-Darling (AD^*) in supremum version. The Kolmogorov-Smirnov test is based on the statistic called Kolmogorov distance, which measures the distance between empirical distribution function and given distribution function. The modified version of the Kolmogorov-Smirnov statistic is as follows:

$$KS^* = \sqrt{n} \sup_x |F_n(x; \mathbf{X}) - F_0^*(x)|. \quad (5)$$

The Kuiper test statistic derives from the Kolmogorov distance. It is the sum of the greatest positive and negative difference between the empirical distribution function and the given distribution function. The modified version of the Kuiper test statistic is given by the following formula:

$$V^* = \sqrt{n}(\sup_x\{F_n(x; \mathbf{X}) - F_0^*(x)\} + \sup_x\{F_0^*(x) - F_n(x; \mathbf{X})\}). \tag{6}$$

The Anderson-Darling test statistic in the supremum version is also based on the distance between two distribution functions, but it put more emphasis on the tails:

$$AD^* = \sqrt{n}\sup_x \frac{|F_n(x; \mathbf{X}) - F_0^*(x)|}{\sqrt{F_0^*(x)(1 - F_0^*(x))}} \tag{7}$$

The second group of statistics consists of two statistics: Cramér-von Mises (W^{2*}) and Anderson-Darling (AD^{2*}) in quadratic version. Both statistics measure the area between empirical distribution function and given distribution function, but they assign different weights to observations. Cramér-von Mises statistic has the weight function equal to one, and its customized version is of the form:

$$W^{2*} = n \int_H^\infty (F_n(x; \mathbf{X}) - F_0^*(x))^2 dF_0^*(x). \tag{8}$$

The Anderson-Darling statistic in quadratic version again puts more weight in the tails:

$$AD^{2*} = n \int_H^\infty \frac{(F_n(x; \mathbf{X}) - F_0^*(x))^2}{F_0^*(x)(1 - F_0^*(x))} dF_0^*(x). \tag{9}$$

New statistics proposed in Chernobai et al. (2015) are based on the Anderson-Darling statistics and give more importance to the upper tail of the distribution. The Authors introduced a new weight function, namely $\Psi(x) = 1/(1 - x)$. After substituting this function, the Anderson-Darling statistics for the truncated samples in supremum (AD_{up}^*) and quadratic (AD_{up}^{2*}) version are respectively as follows:

$$AD_{up}^* = \sqrt{n}\sup_x \frac{|F_n(x; \mathbf{X}) - F_0^*(x)|}{1 - F_0^*(x)}, \tag{10}$$

$$AD_{up}^{2*} = n \int_H^\infty \frac{(F_n(x; \mathbf{X}) - F_0^*(x))^2}{(1 - F_0^*(x))^2} dF_0^*(x). \tag{11}$$

The computational formulas for the test statistics for truncated distributions (for quadratic versions of the statistics and for the new statistics) can be found in Chernobai et al. (2015).

3. SIMULATION STUDIES

This section contains the results of the simulation studies, conducted for seven modified goodness-of-fit statistics presented in section 2 and for the selected distributions described in section 1. The aim of the studies was to evaluate the size and the power of the goodness-of-fit tests for truncated distributions on the basis of artificial data. The simulation studies were conducted for different tail thickness and truncation points. This section is organized as follows: first part describes the methodology of the studies, next the results of the evaluation of the size and the power are presented, finally some details of implementation in R program are given.

3.1. Description of simulation experiments

To compute the empirical sizes of the analyzed tests, the following procedure was applied:

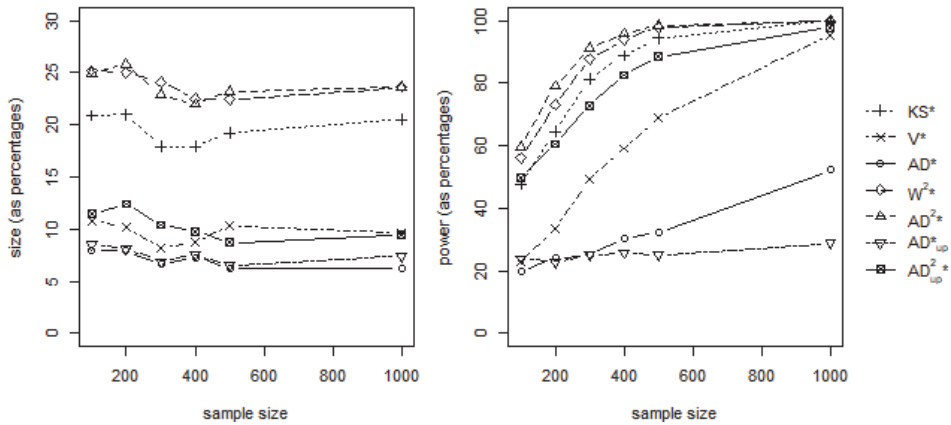
1. Generate n observations from the theoretical distribution that appears in the null hypothesis.
2. Apply all the analyzed tests to the data generated in point 1. Note the p -values of the tests.
3. Repeat the steps described in points 1 and 2 M times, where M is sufficiently large number.
4. Compute the empirical size of each test as the mean of a number of rejections of the null hypothesis.

To compute the empirical power of the tests, in point 1 of the above procedure, the data were generated from a different distribution than it was stated in the null hypothesis. The steps from 2 to 4 remained the same.

To determine the p -values, the testing procedures described in section 2 were carried out. The p -values were calculated on the basis of $N = 100$ Monte Carlo samples, which is the default value of N in the `truncgof` package. Within each simulation a sample of $n = 1000$ observations was generated. The number of simulation replicates was $M = 1000$. The studies were conducted for 12 distributions described in the next paragraph and for 5 truncation points $H = 2, 4, 6, 8, 10$. Altogether 60 experiments were conducted to evaluate both the empirical size and power. The results were verified on the significance level $\alpha = 5\%$.

The sample size $n = 1000$, was determined on the basis of the simulation studies conducted for the selected cases. Namely, figure 1 presents the empirical size and power of the analyzed tests for some cases under t -distribution. The power of all analyzed tests improved with the increase of the sample size to $n = 1000$. In many cases also improvement in the size is visible.

Figure 1. The size ($t(5)$ distribution, $H=6$) and the power ($t(5)$ vs $t(6)$ distributions, $H=6$) of the analyzed tests with respect to the sample size



Source: own calculation.

Values of the cumulative distribution functions for the chosen truncation points range from 0.4592 to 0.9999. In insurance data studies even lower levels are considered. Respective values in Chernobai et al. (2006) amounted to 0.0387 and 0.8212 for the conditional distributions. On the other hand, in finance, some of the risk measures like VaR and CVaR, are based on as high quantiles of distributions as 0.95, 0.975 or 0.99, as was mentioned at the beginning of section 2.

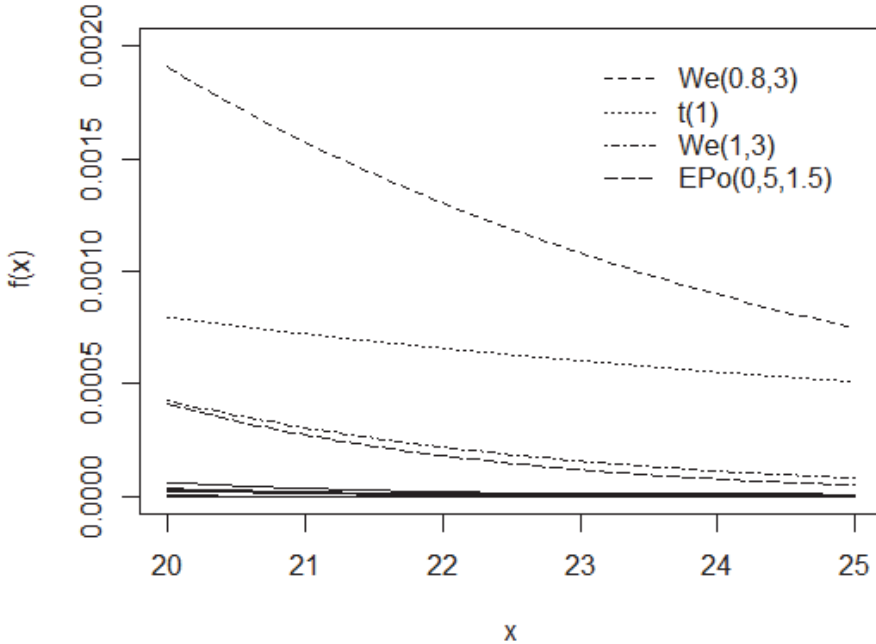
The actual distributions in experiments are: normal, t-distribution, power-exponential and Weibull. The notation used for these distributions in the paper is as follows: $N(\mu, \sigma^2)$ for the normal distribution, where $\mu \in \mathbb{R}$ is the location parameter and $\sigma > 0$ is the scale parameter; $t(n)$ for the t-distribution, where $n \in \mathbb{N}$ denotes the degrees of freedom; $EPo(\mu, \sigma_p, p)$ for the power-exponential distribution, where $\mu \in \mathbb{R}$, $\sigma_p > 0$ and $p > 1$ are the location, scale and shape parameters, respectively; $We(\alpha, \beta)$ for the Weibull distribution with the shape parameter $\alpha > 0$ and the scale parameter $\beta > 0$. Normal distribution, t-distribution and power-exponential distribution were used among other distributions by Piasecki, Tomasik (2013) to verify the shapes of the log asset returns on the polish market. The values of the estimators of these distributions' parameters, e.g. for the WIG index for the chosen period labeled as "h3", were as follows: $N(0.1956, 1.8623)$, $t(2.1561)$, $EPo(0.1982, 1.5881, 1.3618)$. The t-distribution used by Piasecki, Tomasik (2013) is the generalized t-distribution with $n \in \mathbb{R}^+$, while EPo is referred to as generalized error distribution (GED). Burnecki et al. (2015) studied the tails of the asset returns and considered among others: normal distribution $N(0, 2)$ and t-distribution $t(4)$. Weibull distribution was used for instance to assess the operational risk by Guegan, Hassani (2018). The estimated parameters of the distributions for the whole analyzed period were as follows $We(0.5896, 182.9008)$. However, the standard version of the Weibull distribution is rarely used. If X has Weibull distribu-

tion, then $-X$ has extreme value distribution of type III and is used in extreme value theory (Magiera, 2005). For the detailed overview of domestic and external researches in the field of asset returns distributions please refer to Piasecki, Tomasiak (2013).

For each distribution, three sets of parameters were considered. As average return rates on the stock exchange are not significantly different from zero, all the location parameters of the considered distributions (if they exist) were set to zero. To obtain different thickness of tails, the remaining parameters were changed. In case of the normal distribution, the thickness of the exponential tail was controlled by the standard deviation, that was set to $\sigma \in \{3,4,5\}$. In case of the t-distribution, the thickness of the power tail was controlled by the number of degrees of freedom. They were set to $n \in \{1,3,5\}$. For the power-exponential distribution, the parameter p was set to 1.5, so the distribution has the tail thicker than exponential. Here the thickness of the semi-heavy tail was controlled by $\sigma_p \in \{3,4,5\}$. In case of the Weibull distribution, $\alpha \in \{0.8,1,1.2\}$, which results in the power, exponential and faster than exponential decaying tail.

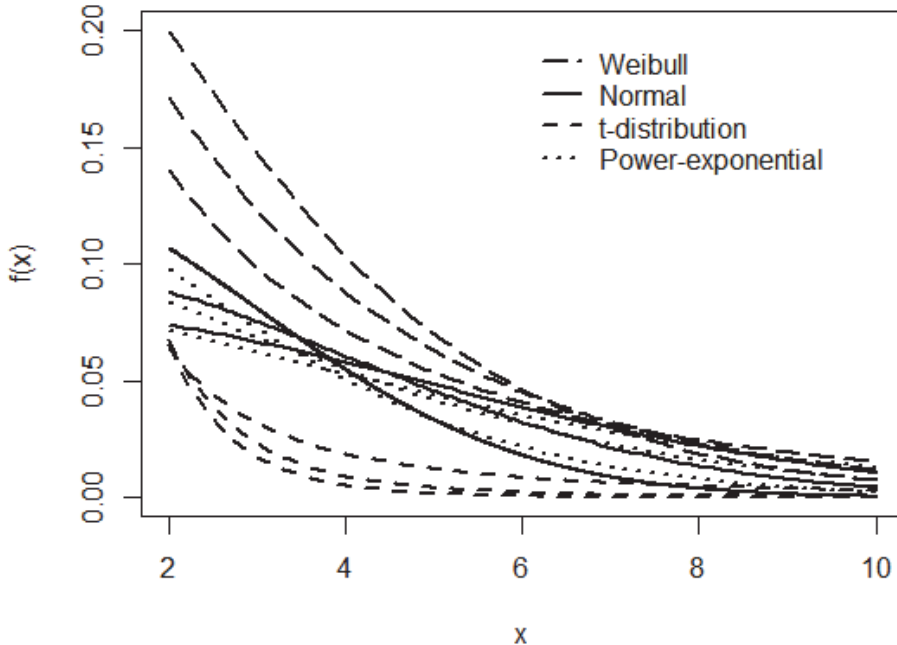
The tails of the distributions considered in the simulation studies are visualized in figure 2. The greatest probability mass in the tails appear in the $We(0.8,3)$, $t(1)$, $We(1,3)$ and $EPo(0,5,1.5)$ distributions, respectively. The tails of the remaining distributions practically disappeared.

Figure 2. Density for the tails of the distributions



Source: own calculation.

Figure 3. Probability distributions used in power study



Source: own calculation.

3.2. Discussion of simulation results

Empirical sizes and powers of a test are identified with the number of rejected null hypothesis. The empirical size of a test should be close to a determined significance level. The empirical power should be as large as possible.

The empirical sizes of the tests obtained in the simulation studies are presented in table 1. The results suggest dividing the tests into three groups. First group contains the KS^* , W^{2*} and AD^{2*} tests. In this group, the empirical sizes were on average 8-times higher than the determined significance level. The tests achieved visibly better results for the t-distribution, but the average rate of rejection was here still 4-times higher than the determined significance level. The second group includes the following tests: V^* and AD_{up}^{2*} . For these tests, the average rate of rejection was 3-times higher than the significance level. These tests also noted better results for the t-distribution. Here probability of rejection of the true null hypothesis was twice higher than the significance level. The third group consists of the remaining two tests: AD^* and AD_{up}^* . In this group, the rate of rejection of the null hypothesis was on average 1.5-times higher than the determined significance level. No clear differences among the distributions were detected.

On the basis of figure 1 and the above results it can be stated, that the considered tests require large number of observations to control the type I error. Unfortunately, the tests in the first and second group may not keep the preassigned

type I error level even for samples of 1000 observations. In case of the fat tailed distribution like $t(1)$, the Authors made additional calculations to verify the size of the tests, when threshold levels are very high, $H = 50, 100, 150, 200, 250$. The size of the tests remained similar to the ones given in table 1 or figure 1. The tests from the first group are the most liberal, and the results of the powers of these tests will not be further analyzed. The results for the tests from the second group will be presented only for the illustrative purposes.

Table 1. EMPIRICAL SIZES OF THE TESTS (AS PERCENTAGES, $\alpha = 5\%$)

Actual distribution	H	KS^*	V^*	AD^*	W^{2*}	AD^{2*}	AD_{up}^*	AD_{up}^{2*}
$N(0,3)$	2	48.1	16.3	7.0	51.8	50.6	7.5	14.9
	4	44.9	15.0	7.0	49.6	48.3	7.9	16.7
	6	42.3	13.8	6.9	46.9	45.0	8.2	17.5
	8	40.4	12.5	7.2	44.3	42.1	8.3	16.6
	10	37.0	12.2	7.4	41.8	39.6	7.9	15.8
$N(0,4)$	2	49.3	17.0	6.7	52.9	51.5	7.5	14.6
	4	46.6	15.5	7.0	50.8	48.9	7.6	15.6
	6	44.2	14.7	6.9	49.2	47.5	8.0	17.2
	8	43.0	13.9	7.0	46.8	44.8	8.3	17.9
	10	41.2	13.5	7.2	45.4	43.7	8.3	17.6
$N(0,5)$	2	49.5	16.8	6.7	53.0	52.2	7.4	14.2
	4	47.4	16.2	7.0	51.6	50.2	7.6	15.2
	6	45.5	15.2	7.0	50.2	48.4	7.8	16.5
	8	44.0	14.3	6.9	48.4	46.7	8.0	16.5
	10	42.4	13.3	7.1	46.7	44.4	8.4	17.0
$t(1)$	2	18.8	9.7	6.4	22.9	22.8	7.5	9.4
	4	20.9	9.9	6.3	24.3	24.2	7.5	9.3
	6	21.1	9.9	6.3	24.6	24.3	7.5	9.3
	8	21.2	9.9	6.3	24.6	24.4	7.5	9.3
	10	21.2	9.9	6.3	24.6	24.4	7.5	9.3
$t(3)$	2	16.0	9.3	6.4	20.4	20.2	7.4	9.8
	4	18.8	9.8	6.4	23.1	23.2	7.4	9.6
	6	20.7	9.8	6.3	24.0	24.0	7.5	9.4
	8	21.1	9.9	6.3	24.3	24.2	7.5	9.3
	10	21.1	9.9	6.3	24.5	24.3	7.5	9.3
$t(5)$	2	15.5	9.2	6.5	18.8	19.6	7.7	10.0
	4	18.3	9.7	6.4	22.8	22.8	7.4	9.6
	6	20.5	9.7	6.3	23.6	23.6	7.4	9.5
	8	20.9	9.9	6.3	24.3	24.2	7.5	9.4
	10	21.1	9.9	6.3	24.4	24.2	7.5	9.3
$EPo(0,3,1.5)$	2	52.5	24.5	7.6	61.0	61.6	9.7	24.7
	4	49.0	19.9	7.9	55.6	55.8	10.7	26.8
	6	45.9	17.2	8.0	51.9	51.1	10.7	25.4
	8	43.3	16.0	7.8	48.6	47.8	9.9	23.7
	10	41.7	14.7	7.8	46.1	45.9	9.5	22.2
$EPo(0,4,1.5)$	2	53.9	25.4	7.6	62.5	62.7	9.2	24.2
	4	51.3	21.7	7.8	58.2	58.7	10.4	26.2
	6	49.2	19.1	7.9	54.8	55.4	10.9	27.1
	8	45.6	17.2	8.0	51.3	50.4	10.5	24.9
	10	43.4	16.1	8.0	49.1	47.8	9.9	22.9

Table 1. EMPIRICAL SIZES OF THE TESTS (AS PERCENTAGES, $\alpha = 5\%$) (cont.)

Actual distribution	H	KS^*	V^*	AD^*	W^{2*}	AD^{2*}	AD_{up}^*	AD_{up}^{2*}
$EPo(0,5,1.5)$	2	54.6	25.6	7.6	62.6	64.2	9.3	23.1
	4	51.2	23.3	7.8	60.6	59.8	10.2	25.7
	6	50.0	20.2	7.9	56.5	56.0	10.7	27.0
	8	47.7	18.4	7.8	53.1	53.1	10.4	25.6
	10	45.1	16.7	7.6	50.7	50.0	10.4	23.9
$We(0.8,3)$	2	47.4	15.6	7.0	51.2	49.8	7.6	15.8
	4	44.7	14.8	6.9	49.5	48.5	7.8	17.0
	6	44.0	14.3	7.0	48.4	46.2	7.9	17.9
	8	43.1	14.0	7.1	47.3	45.5	8.2	18.3
	10	42.7	13.8	7.2	46.3	45.0	8.2	18.5
$We(1,3)$	2	48.0	15.6	7.0	51.5	50.1	7.6	15.7
	4	44.3	14.5	6.9	49.2	47.9	7.8	17.2
	6	43.3	14.3	7.0	47.7	45.7	8.2	18.0
	8	42.5	13.7	7.2	46.4	45.1	8.2	18.5
	10	41.7	13.2	7.3	45.3	44.7	8.2	19.0
$We(1.2,3)$	2	48.3	16.0	7.0	51.8	50.6	7.6	15.4
	4	44.3	14.5	6.9	49.1	47.5	7.9	17.3
	6	43.0	14.0	7.1	47.2	45.6	8.2	18.3
	8	41.8	13.5	7.3	45.6	44.7	8.3	18.7
	10	40.5	13.2	7.4	44.4	44.1	8.3	19.2

Source: own calculation.

Empirical powers of the tests are presented in table 2. The visualization of the probability distributions used in the power study is presented in figure 3. The tests from the second group, that is the V^* and AD_{up}^{2*} tests, have high empirical powers, that on average amount to 97% for the first test and 98% for the second test. However, it has to be reminded, that these tests are too liberal. The AD^* and AD_{up}^* tests from the third group are much more realistic. While the average rate of rejection of the false null hypothesis for the first test is 64%, it is only 30% for the second test. The results for the AD_{up}^* are very irregular. For the distributions with the fast decaying tails, that is for the normal and power-exponential ones, the average powers are lower than 3%.

The powers of the V^* , AD_{up}^{2*} and AD^* tests show common behavior with respect to the decaying rates of the tails. With regard to the distributions with the fast decaying tails, that is the normal and power-exponential distributions, the empirical powers of the tests decrease with the growing thickness of the tail. In case of the distributions with thicker tails, that is the t-distribution and Weibull distribution, the relation is opposite, the powers of the tests increase with the growing thickness of the tail. Summarizing, the powers of the tests are higher for extreme tails, that are decaying exponentially or powerly. In case of the distributions with semi-heavy tails, the considered tests had more problems with recognizing the actual distribution. It is also worth noting that the powers of the V^* , AD_{up}^{2*} and AD^* tests were increasing with the growth of the truncation point.

Table 2. EMPIRICAL POWERS OF THE TESTS (AS PERCENTAGES, $\alpha = 5\%$)

Null hypothesis	Actual distribution	H	V^*	AD^*	AD_{up}^*	AD_{up}^{2*}
$N(0,3.5)$	$N(0,3)$	2	100.0	71.3	0.4	99.9
		4	100.0	81.0	0.5	100.0
		6	100.0	87.0	0.5	100.0
		8	100.0	89.5	0.4	100.0
		10	100.0	90.5	0.5	100.0
$N(0,4.5)$	$N(0,4)$	2	99.2	35.1	1.2	98.6
		4	99.7	45.5	1.2	98.8
		6	99.7	53.2	1.2	99.5
		8	99.8	59.8	1.2	99.6
		10	99.9	64.7	1.2	99.7
$N(0,5.5)$	$N(0,5)$	2	93.4	15.6	1.8	91.4
		4	95.8	19.8	1.7	94.8
		6	97.1	23.6	1.8	96.7
		8	98.3	27.9	1.8	97.4
		10	98.5	32.7	1.8	97.6
$t(2)$	$t(1)$	2	100.0	100.0	100.0	100.0
		4	100.0	100.0	100.0	100.0
		6	100.0	100.0	100.0	100.0
		8	100.0	100.0	100.0	100.0
		10	100.0	100.0	100.0	100.0
$t(4)$	$t(3)$	2	99.3	83.6	48.8	99.6
		4	100.0	93.4	53.1	100.0
		6	100.0	95.3	54.2	100.0
		8	100.0	96.0	55.7	100.0
		10	100.0	96.2	55.9	100.0
$t(6)$	$t(5)$	2	56.0	28.4	24.3	83.9
		4	90.4	44.9	27.5	95.1
		6	95.4	52.2	28.9	97.8
		8	97.4	56.1	29.5	98.4
		10	98.0	57.7	29.9	98.5
$EPo(0,3.5;1.5)$	$EPo(0,3,1.5)$	2	99.9	63.4	1.4	100.0
		4	100.0	69.7	1.4	100.0
		6	100.0	73.8	1.5	100.0
		8	100.0	74.0	1.3	100.0
		10	100.0	74.4	1.3	99.9
$EPo(0,4.5,1.5)$	$EPo(0,4,1.5)$	2	97.7	26.4	2.1	98.3
		4	98.1	32.7	2.2	98.9
		6	98.2	36.4	2.1	99.3
		8	98.3	38.4	2.0	99.2
		10	98.2	38.9	2.0	99.3
$EPo(0,5.5,1.5)$	$EPo(0,5,1.5)$	2	91.3	14.4	2.8	91.0
		4	91.9	16.8	2.6	94.1
		6	92.0	19.0	2.8	95.6
		8	91.5	19.9	2.7	96.1
		10	91.5	20.0	2.4	95.9

Table 2. EMPIRICAL POWERS OF THE TESTS (AS PERCENTAGES, $\alpha = 5\%$) (cont.)

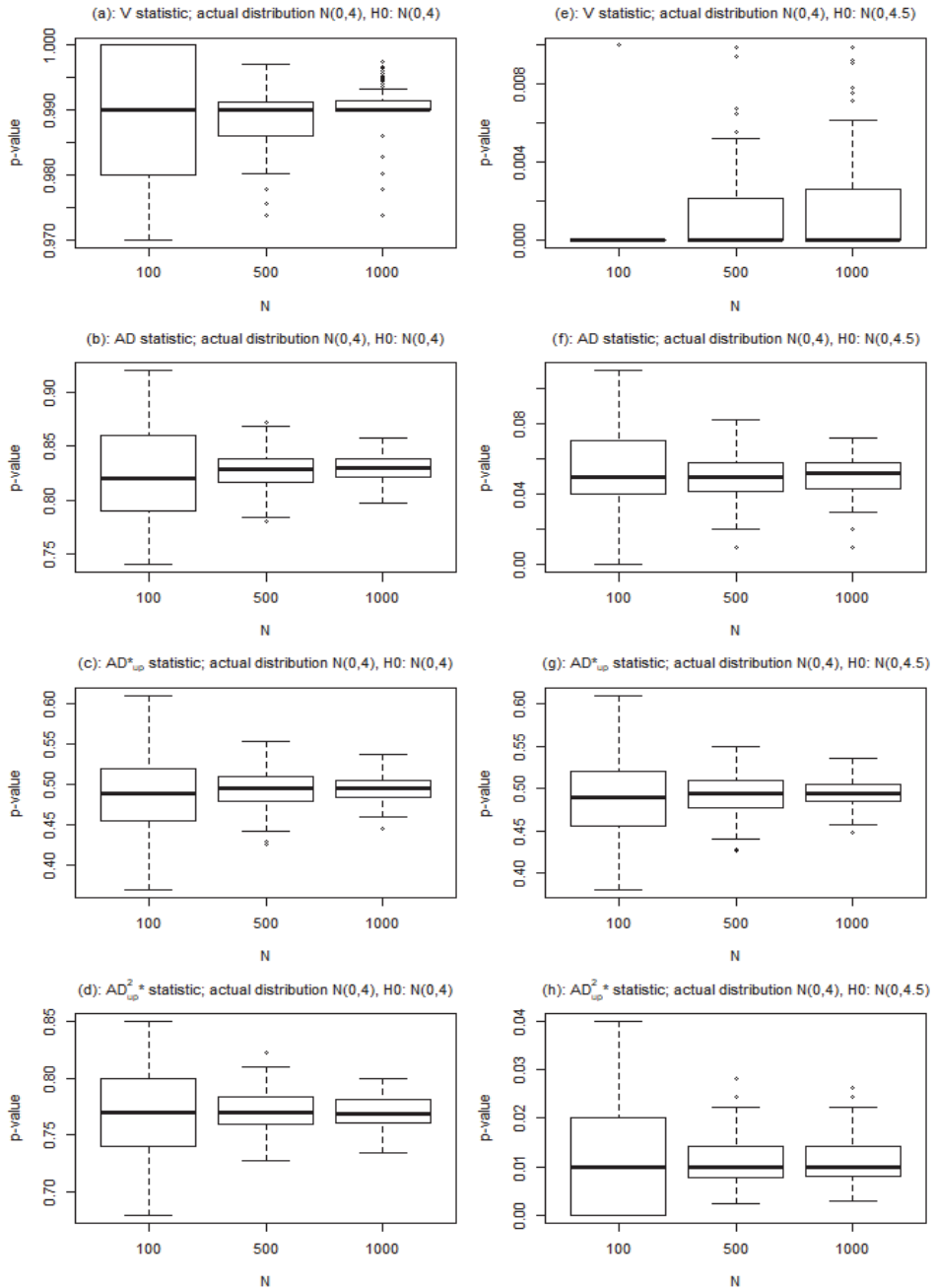
Null hypothesis	Actual distribution	H	V^*	AD^*	AD_{up}^*	AD_{up}^{2*}
$We(0.9,3)$	$We(0.8,3)$	2	95.8	80.3	58.5	99.7
		4	99.9	90.4	64.8	100.0
		6	99.9	94.9	69.3	100.0
		8	100.0	97.2	72.5	100.0
		10	100.0	98.3	75.9	100.0
$We(1.1,3)$	$We(1,3)$	2	82.9	58.4	43.0	98.2
		4	96.9	74.2	49.3	99.8
		6	99.8	84.7	55.3	100.0
		8	99.8	92.0	59.3	100.0
		10	99.9	95.2	62.3	100.0
$We(1.3,3)$	$We(1.2,3)$	2	67.4	43.6	35.2	93.1
		4	90.7	60.1	39.8	99.2
		6	97.8	73.7	43.9	99.9
		8	99.8	82.9	49.0	100.0
		10	99.8	89.5	53.4	100.0

Source: own calculation.

Due to the time-consuming procedures, the p -values were calculated on the basis of $N = 100$ Monte Carlo samples, the default value of N in the truncgof package. To justify the obtained results, the randomness of the p -values was studied for the selected cases (similar analysis was considered in Smaga, 2017). The tests chosen to the power study were applied 100 times to a single data set, with different values of N . The study was performed for the actual distribution $N(0,4)$, under the true and false null hypothesis. Figure 4 presents the results. The median of each analysed case does not vary considerably between different numbers of N . The variance of p -values decreases with the increase of N , therefore, in the unconvincing cases, it is recommended to repeat the tests with a higher value of N . This may at least slightly improve the results, for example, the empirical sizes of the tests AD^* , AD^{2*} , KS^* , V^* , W^{2*} , AD_{up}^* and AD_{up}^{2*} were equal to 7.5, 42.8, 39.4, 13.6, 44.8, 7.6 and 13 respectively for $N = 1000$, actual distribution $N(0,4)$ and $H = 6$, while for $N = 100$, they were equal to 6.9, 47.5, 44.2, 14.7, 49.2, 8 and 17.2 respectively.

All the calculations were done in R statistical environment (R Core Team, 2017). Except the R package truncgof (Wolter, 2012), the R packages normalp (Mineo, 2014) and doParallel (Calaway et al., 2017) were also used, since they deal with power-exponential distributions and parallel computing, respectively. When writing the code in R program, many tips and hints were drawn from the handbook written by Górecki (2011).

Figure 4. Boxplots for the randomness of the p-values



For each boxplot the testing procedure was repeated 100 times. For a-d the null, and for e-h the alternative hypothesis was true. The p-values determined on the basis of $N = 100000$ for a-h were as follows: 0.99010, 0.80513, 0.49084, 0.76880, 0.00053, 0.04464, 0.48980, 0.00680.

Source: own calculation.

4. CONCLUSIONS

The aim of this paper was to present the results of the simulation studies, that were evaluating the finite sample behavior of the tests for truncated distributions introduced in Chernobai et al. (2015) and implemented in the R package *truncgof* (Wolter, 2012). The tests were performed with default values of parameters used in the *truncgof* package. The research was based on artificial data generated from the distributions that are often describing the tails of asset returns. The study was conducted for different tail thickness and for changing truncation point. In the cases considered in the article, the KS^* , W^{2*} , AD^{2*} , V^* and AD_{up}^{2*} tests did not maintain the preassigned type I error level. The remaining two tests, AD^* and AD_{up}^* , obtained reasonable rejection rates for the true null hypothesis. The power of the AD^* test was much higher than the power of the AD_{up}^* test. While the average rate of rejection of the false null hypothesis for the first test is 64%, it is only 30% for the second test. It was also noticed, that the power of AD^* and AD_{up}^* tests is higher for extreme tails and it grows with the truncation point. On the basis of the obtained results, it is recommended to assess the behavior of the tests analyzed in this article, in terms of the sample size, theoretical distribution and truncation point, before every application. In the unconvincing cases (e.g., when the p -value is close to the significance level), it is suggested to use greater number N of Monte Carlo samples to estimate the p -values of the tests than $N = 100$, which is the default value of N in the *truncgof* package.

REFERENCES

- Anderson T. W., Darling D. A., (1952), Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes, *The Annals of Mathematical Statistics*, 23 (2), 193–212.
- Brunner M. I., Viviroli D., Sikorska A. E., Vannier O., Favre A.-C., Seibert J., (2017), Flood Type Specific Construction of Synthetic Design Hydrographs, *Water Resources Research*, 53 (2), 1390–1406.
- Burnecki K., Chechkin A., Wylomanska A., (2015), Discriminating Between Light- and Heavy-Tailed Distributions with Limit Theorem, *PLoS ONE*, 10 (12), 1–23.
- Chernobai A., Burnecki K., Rachev S., Trück S., Weron R., (2006), Modelling Catastrophe Claims with Left-Truncated Severity Distributions, *Computational Statistics*, 21 (3), 537–555.
- Chernobai A., Menn C., Rachev S. T., Trück S., (2010), Estimation of Operational Value-at-Risk in the Presence of Minimum Collection Threshold: An Empirical Study, Working Paper Series in Economics, 4, Karlsruhe Institute of Technology (KIT), Department of Economics and Business Engineering.
- Chernobai A., Rachev S. T., Fabozzi F. J., (2015), Composite Goodness-of-Fit Tests for Left-Truncated Loss Samples, in: Lee C. F., Lee J., (eds.), *Handbook of Financial Econometrics and Statistics*, Springer, New York, NY.
- Clementi F., Gallegati M., Kaniadakis G., (2012), A New Model of Income Distribution: The κ – Generalized Distribution, *Journal of Economics*, 105 (1), 63–91.
- Cox D. R., Spjøtvoll E., Johansen S., Van Zwet W. R., Bithell J. F., Barndorff-Nielsen O., (1977), The Role of Significance Tests, *Scandinavian Journal of Statistics*, 4 (2), 49–70.

- Echaust K., (2014), *Ryzyko zdarzeń ekstremalnych na rynku kontraktów futures w Polsce*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań.
- Fagiolo G., Alessi L., Barigozzi M., Capasso M., (2010), On the Distributional Properties of Household Consumption Expenditures: The Case of Italy, *Empirical Economics*, 38 (3), 717–741.
- Feller W., (1950), *An Introduction to Probability Theory and Its Applications*, Wiley, New York.
- Fischer M., Jakob K., (2016), pTAS Distributions with Application to Risk Management, *Journal of Statistical Distributions and Applications*, 3 (1), 1–18.
- Górecki T., (2011), *Podstawy statystyki z przykładami w R*, Wydawnictwo BTC, Legionowo.
- Górecki T., Smaga Ł., (2015), A Comparison of Tests for the One-Way ANOVA Problem for Functional Data, *Computational Statistics*, 30, 987–1010.
- Guegan D., Hassani B.K., (2018), More Accurate Measurement for Enhanced Controls: VaR vs ES?, *Journal of International Financial Markets, Institutions & Money*, 54, 152-165.
- Haas M., Pigorsch C., (2009), Financial Economics, Fat-Tailed Distributions, *Encyclopedia of Complexity and Systems Science*, 4, 3404–3435.
- Kim J. H., Ji P. I., (2015), Significance Testing in Empirical Finance: A Critical Review and Assessment, *Journal of Empirical Finance*, 34, 1–14.
- Krzyśko M., (2000), *Wykłady z teorii prawdopodobieństwa*, Wydawnictwa Naukowo-Techniczne, Warszawa.
- Krzyśko M., (2004), *Statystyka matematyczna*, Wydawnictwo Naukowe UAM, Poznań.
- Lach A., (2017), Testy zgodności z rozkładami uciętymi, *Praca licencjacka*, Uniwersytet im. Adama Mickiewicza, Poznań.
- Magiera R., (2005), *Modele i metody statystyki matematycznej. Cz. 1: Rozkłady i symulacja stochastyczna*, Oficyna Wydawnicza GIS, Wrocław.
- Mandelbrot B., (1963), The Variation of Certain Speculative Prices, *The Journal of Business*, 36, 394-419.
- Mineo A.M., (2014), normalp: Routines for Exponential Power Distribution, R package version 0.7.0, <https://CRAN.R-project.org/package=normalp>.
- Orzeszko W., (2014), Symulacyjna ocena rozmiaru testu BDS, *Przegląd Statystyczny*, 4 (61), 335–361.
- Pavia J. M., (2015), Testing Goodness-of-Fit with the Kernel Density Estimator: GoFKernel, *Journal of Statistical Software*, 66, 1–27.
- Piasecki K., Tomasiak E., (2013), *Rozkłady stóp zwrotu z instrumentów polskiego rynku kapitałowego*, Wydawnictwo edu-Libri, Kraków, Warszawa.
- R Core Team, (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>.
- Revolution Analytics, Weston S., (2015), doParallel: Foreach Parallel Adaptor for the 'parallel' Package, R package version 1.0.10, <http://CRAN.R-project.org/package=doParallel>.
- Smaga Ł., (2017), A Two-Sample Test Based on Cluster Subspaces for Equality of Mean Vectors in High Dimension, *Discusiones Mathematicae Probability and Statistics*, 37, 147–156.
- Wolter T., (2012). Truncgof: GoF Tests Allowing for Left Truncated Data, R package version 0.6-0, <https://CRAN.R-project.org/package=truncgof>.

PORÓWNANIE TESTÓW ZGODNOŚCI DLA ROZKŁADÓW UCIĘTYCH

Streszczenie

Celem artykułu jest empiryczne zbadanie mocy i rozmiaru siedmiu testów zgodności, zaprezentowanych w pracy Chernobai i inni (2015), przeznaczonych dla rozkładów lewostronnie uciętych. Badania symulacyjne oparto na danych, wygenerowanych z rozkładów, które były w przeszłości lub są obecnie sto-

sowane do opisu ogonów rozkładów stóp zwrotu. Badania przeprowadzono dla różnych grubości ogonów rozkładów oraz zmieniających się poziomów ucięcia. Wyniki symulacji wskazują na istnienie znacznych różnic pomiędzy poszczególnymi procedurami testowymi. Ponadto otrzymanie zadowalających wyników w przypadku niektórych procedur wymaga dość dużej liczby obserwacji.

Słowa kluczowe: moc testu, program R, rozkłady ucięte, rozmiar testu, testy zgodności

COMPARISON OF THE GOODNESS-OF-FIT TESTS FOR TRUNCATED DISTRIBUTIONS

Abstract

The aim of this paper is to investigate the finite sample behavior of seven goodness-of-fit tests for left truncated distributions of Chernobai et al. (2015) in terms of size and power. Simulation experiments are based on artificial data generated from the distributions that were used in the past or are used nowadays to describe the tails of asset returns. The study was conducted for different tail thickness and for changing truncation point. Simulation results indicate that the testing procedures do not work equally well under finite samples, and some of them require quite large number of observations to perform satisfactorily.

Keywords: goodness of fit tests, power of test, R program, size of test, truncated distributions