Теория вероятностей и математическая статистика

Probability theory and mathematical statistics

УДК 519.254

ОБ ОБНАРУЖЕНИИ ВЫБРОСОВ С ПОМОЩЬЮ НЕРАВЕНСТВА ЧЕБЫШЕВА

М. А. ЧЕПУЛИС[1)], ***Г. Л. ШЕВЛЯКОВ***[1)]

[1)]*Санкт-Петербургский политехнический университет Петра Великого, ул. Политехническая, 29, 195251, г. Санкт-Петербург, Россия*

Рассматриваются алгоритмы, основанные на использовании неравенства Чебышева. Эти алгоритмы сравниваются с такими классическими методами, как боксплот Тьюки, правило *N*-сигм и его робастные модификации с *MAD*- и *FQ*-оценками масштаба. Для настройки алгоритмов используется процедура выбора параметров на основе полного знания модели распределения данных. Строятся области субоптимальных параметров при неполном знании модели засорения. Показывается, что непосредственное применение неравенства Чебышева приводит к классическому правилу *N*-сигм. При использовании неклассического неравенства Чебышева получается робастное правило отбраковки, которое зачастую не уступает, а иногда и превосходит прочие рассматриваемые алгоритмы.

Ключевые слова: аномалия; обнаружение выбросов; неравенство Чебышева; робастность.

Благодарность. Это исследование частично поддержано грантом РФФИ № 18-29-03250.

Образец цитирования:
Чепулис МА, Шевляков ГЛ. Об обнаружении выбросов с помощью неравенства Чебышева. *Журнал Белорусского государственного университета. Математика. Информатика.* 2020;3:28–35 (на англ.).
https://doi.org/10.33581/2520-6508-2020-3-28-35

For citation:
Chepulis MA, Shevlyakov GL. On outlier detection with the Chebyshev type inequalities. *Journal of the Belarusian State University. Mathematics and Informatics.* 2020;3:28–35.
https://doi.org/10.33581/2520-6508-2020-3-28-35

Авторы:
Михаил Артемович Чепулис – магистрант высшей школы прикладной математики и вычислительной физики Института прикладной математики и механики. Научный руководитель – Г. Л. Шевляков.
Георгий Леонидович Шевляков – доктор физико-математических наук, профессор; профессор высшей школы прикладной математики и вычислительной физики Института прикладной математики и механики.

Authors:
Michael A. Chepulis, master's degree student at the department of applied mathematics and mechanics, high school of applied mathematics and computational physics.
michael.chepulis@yandex.ru
https://orcid.org/0000-0001-7340-9323
Georgy L. Shevlyakov, doctor of science (physics and mathematics), full professor; professor at the department of applied mathematics and mechanics, high school of applied mathematics and computational physics.
georgy.shevlyakov@phmf.spbstu.ru
https://orcid.org/0000-0001-7559-5633

ON OUTLIER DETECTION WITH THE CHEBYSHEV TYPE INEQUALITIES

M. A. CHEPULIS[a], ***G. L. SHEVLYAKOV***[a]

[a]*Peter the Great St. Petersburg Polytechnic University,*
29 Polytechnicheskaya Street, Saint Petersburg 195251, Russia
Corresponding author: M. A. Chepulis (michael.chepulis@yandex.ru)

This work considers algorithms of outlier detection based on the Chebyshev inequality. It compares these algorithms with such classical methods as Tukey's boxplot, the *N*-sigma rule and its robust modifications based on *MAD* and *FQ* scale estimates. To adjust the parameters of the algorithms, a selection procedure is proposed based on the complete knowledge of the data distribution model. Areas of suboptimal parameters are also determined in case of incomplete knowledge of the distribution model. It is concluded that the direct use of the Chebyshev inequality implies the classical *N*-sigma rule. With the non-classical Chebyshev inequality, a robust outlier detection method is obtained, which slightly outperforms other considered algorithms.

Keywords: anomaly; outlier detection; Chebyshev inequality; robustness.

Acknowledgements. This research is partially supported by the Russian Foundation for Basic Research (number of grant 18-29-03250).

The problem of outlier detection is one of the oldest in statistics. However, despite the large number of publications on this topic, there is no general method for solving this problem in mathematical statistics. In this work, we suggest one more novel method of outlier detection, in this case based on the classical Chebyshev inequality [1].

This method performs quite well in the processing of data from laser ranging of the Moon [2]. Therefore, it seems possible to effectively apply it to the classical problem of one-dimensional outlier detection. In [2], this method is proposed for data of a specific structure when the data points are two-dimensional (the mean and its standard error), and it is impossible to apply it to a one-dimensional problem in its original form. Thus, it is required to adapt it for such data.

The aim of the work is to develop an algorithm based on the Chebyshev inequality for one-dimensional data and compare its performance with such classical methods as the *N*-sigma rule, its robust modifications based on the highly robust median absolute deviation (*MAD*) and the fast *Qn*-scale (*FQ*) estimates of scale [3], Tukey's boxplot at the standard normal distribution with its «shift» and «scale» contaminated versions and at the Cauchy distribution.

As a rule, according to the Neyman – Pearson approach to outlier detection, the performance evaluation of outlier detection is associated with the power of the detection rule and the probability of its false alarm. However, it is difficult to keep these both parameters stable simultaneously in Monte Carlo studies, especially the small value of the false alarm probability. Therefore, the *H*-measure proposed in [4] is chosen as a comparison criterion, which naturally combines the power of detection and the probability of a false alarm.

The Chebyshev inequality

The classical Chebyshev inequality. The classical Chebyshev inequality [1] estimates the probability of deviation of a random variable from its mean by a certain value through the moment characteristics of a distribution.

For a random variable X from an arbitrary distribution with the known mean μ and standard deviation σ, the Chebyshev inequality has the form

$$P\{|X-\mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2},$$

where ε is an arbitrary positive number.

Based on this estimate with $\varepsilon = \lambda\sigma$, a confidence interval for a random variable X is constructed:

$$\Delta(\lambda) = (\mu - \lambda\sigma, \mu + \lambda\sigma).$$

The observations lying within this confidence interval are declared regular, and outside it as outliers. However, in cases where the distribution parameters are unknown, we estimate the mean and standard deviation by the sample mean and square mean deviation, and thus we arrive at the classical *N*-sigma rule.

Robust version of the Chebyshev inequality

Now we use another probability metrics in the Chebyshev inequality derivation scheme:

$$D_p = \int_{-\infty}^{\infty} |x-\mu|^p f(x)dx \geq \int_{|x-\mu|\geq\varepsilon} |x-\mu|^p f(x)dx \geq \varepsilon^p \int_{|x-\mu|\geq\varepsilon} f(x)dx = \varepsilon^p P\{|x-\mu|\geq\varepsilon\}.$$

The sought probability is of the form

$$P\{|x-\mu|\geq\varepsilon\} \leq \frac{D_p}{\varepsilon^p}.$$

For $p = 2$, we get the classical Chebyshev inequality. For $p = 1$, the distance D_1 in the inequality is the mean absolute deviation. Here, the median is a natural estimate of location as it is the solution minimising the mean absolute deviation. The obtained outlier detection rule is of the same structure as in the classical case. It is worth noting that the median and the mean absolute value are robust estimates of location and scale, respectively.

H-measure of performance evaluation

In the case of research on model data in the form of a mixture of two distributions, we will call regular observations from F_0 or the «main» distribution, and anomalous from F_1

$$F = (1-\alpha)F_0(x) + \alpha F_1(x),$$

where α is the fraction of contamination. Here, we identify the concepts of anomalous value and outlier.

The problem of outlier detection can be reduced to the problem of binary classification, when it is necessary to determine which of the two distributions the observation belongs to: regular data (hypothesis H_0) or anomalous (hypothesis H_1).

In order to compare algorithms with each other, you need to select the comparison criteria. The power of the criterion (P_D) and the probability of the false alarm (P_F) are classical for statistics

$$P_D = \frac{\text{TN}}{\text{total number of otliers}},$$

$$P_F = \frac{\text{FN}}{\text{total number of regular points}},$$

where TN (true negative) is the number of outliers that have been classified as outliers; FN (false negative) is the number of regular points that have been classified as outliers.

To assess the quality of the classification results, we use the *H*-measure (the harmonic mean between P_D and $1 - P_F$) introduced in [4] and calculated by the following formula:

$$H(t) = \frac{2P_D(t)(1-P_F(t))}{P_D(t)+1-P_F(t)},$$

where t is a parameter of an outlier detection algorithm.

The higher is the value of the *H*-measure, the better the quality of the classification. One can also calculate the parameter at which the maximum of the *H*-measure is reached:

$$t^* = \arg\max_t H(t).$$

Comparison study

The following methods are compared:

- *N*-sigma rule [4];
- robust *N*-sigma rule (*MAD* estimate of scale) [4];
- robust *N*-sigma rule (*FQ* estimate of scale) [4];
- Tukey's boxplot [5].

The following distributions are considered:

- mixture of normal («shift»): $(1-\alpha)N(x, 0, 1) + \alpha N(x, k, 1)$;
- mixture of normal («scale»): $(1-\alpha)N(x, 0, 1) + \alpha N(x, 0, k)$;
- mixture of normal and Cauchy: $(1-\alpha)N(x, 0, 1) + \alpha C(x, 0, 1)$.

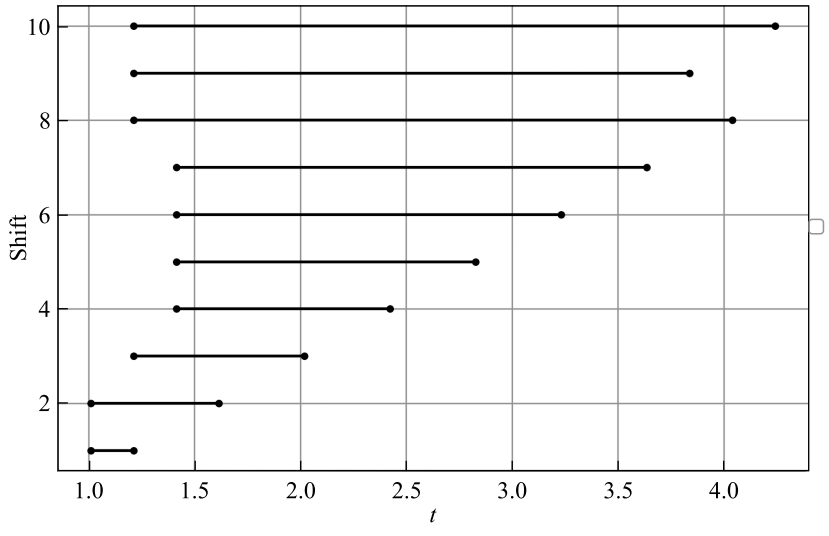

Fig. 1. Suboptimal areas with «shift» contamination

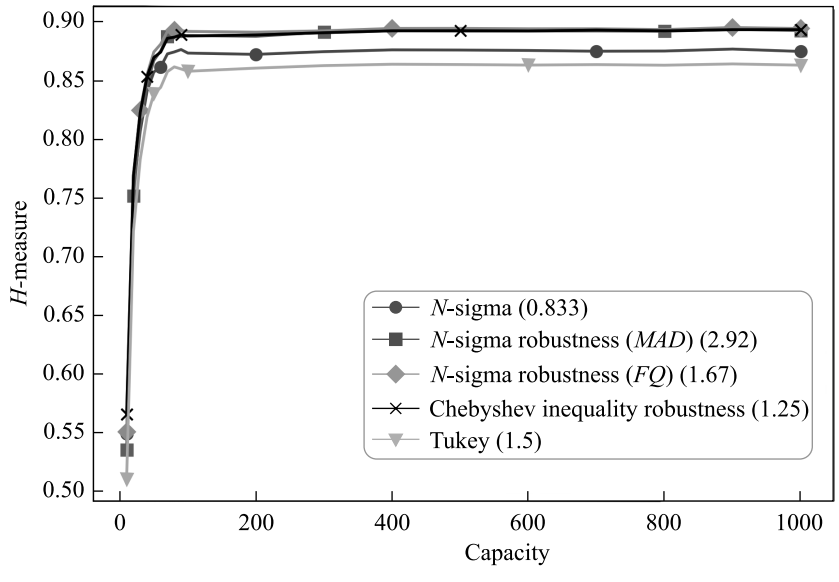

Fig. 2. Dependence *H*-measure on the sample size with «scale» ($k = 10$) contamination

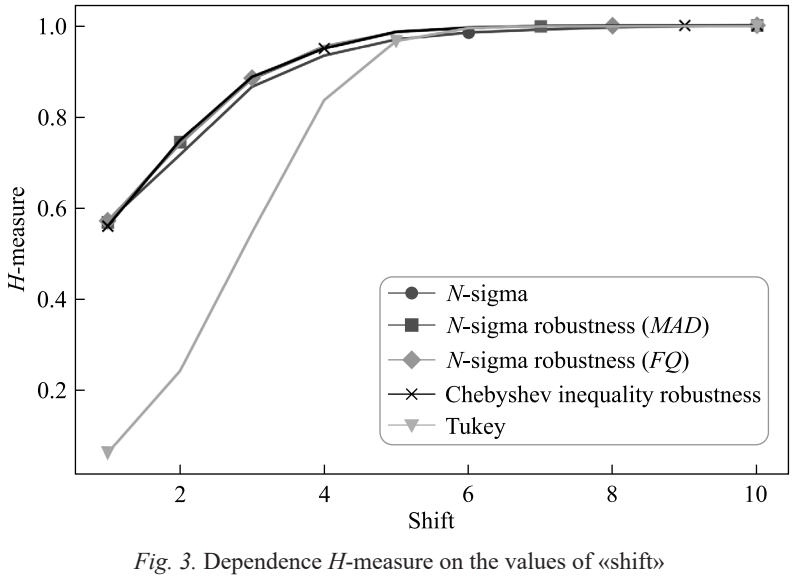

Fig. 3. Dependence *H*-measure on the values of «shift»

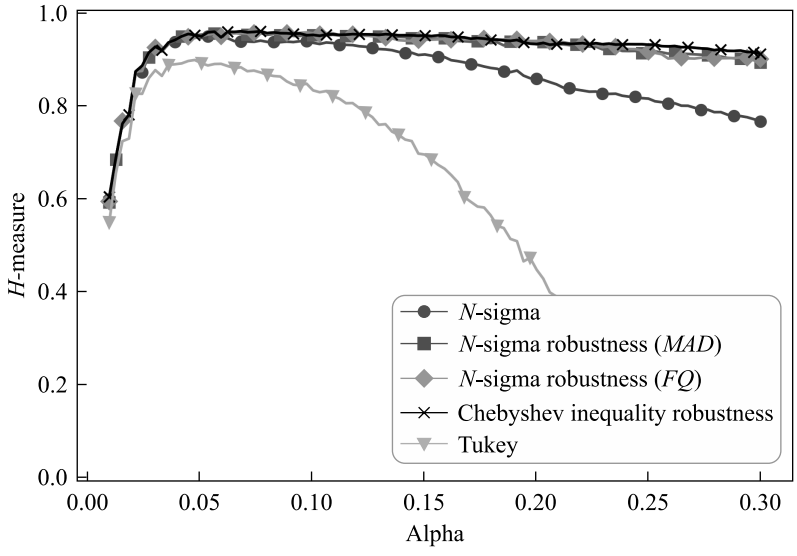

Fig. 4. Dependence *H*-measure on the contamination fraction with «shift» ($k = 4$) contamination

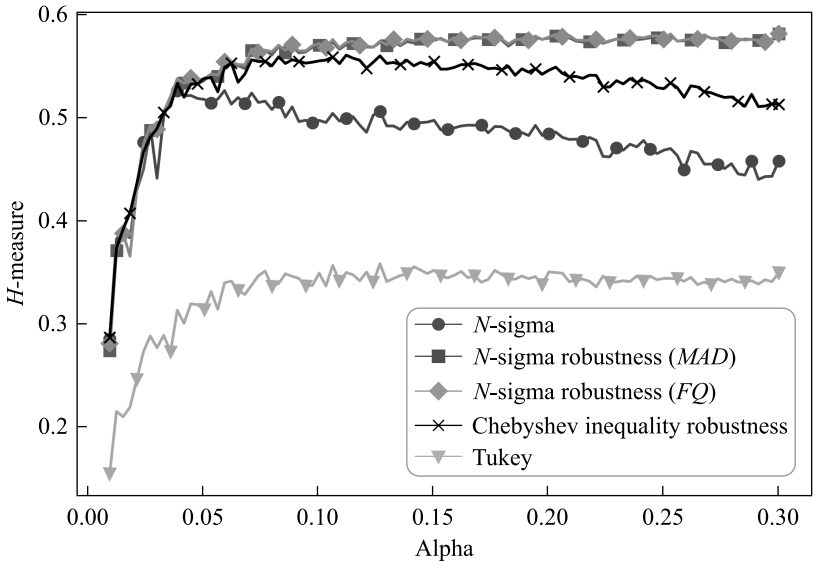

Fig. 5. Dependence *H*-measure on the contamination fraction with Cauchy contamination

The main interest is the dependence of the quality of detection on such parameters of contamination as the «shift», «scale» and percentage of contamination.

The quality of outlier detection strongly depends on the parameters of the algorithms, for example, *N* in the *N*-sigma rule. We adjust the parameters by iterating over the grid, choosing the parameter with the maximum *H*-measure.

In order to reduce the influence of randomness on the processes under consideration, the study will be carried out by the Monte Carlo method, averaging the values over all repetitions.

Algorithm

1. The distribution and parameters of the studied algorithms are selected.
2. For each cycle of repetition of the Monte Carlo method:
   a) a sample is generated with automatically marked «regular» and «anomalous» points;
   b) outlier detection methods are applied to the sample;
   c) the information is collected about the results of this application.
3. Further, the data for all repetition cycles are averaged and analysed.

Choice of algorithm parameters

Based on the full knowledge of the data model, we calculate the value of the *H*-measure for various parameters of algorithms on a certain distribution: the optimal one we take is the argument for which the *H*-measure is maximal.

Next, instead of choosing the parameter according to the maximum of the *H*-measure, we consider all parameters that differ from the optimal value by no more than 5 % (the larger the percentage, the wider the area, and the lower the detection quality). In this case, we get a certain area of suboptimal parameters. Such an area can be calculated for each distribution of interest, and the parameter of the algorithm can be chosen so that it falls into the maximum number of areas.

For example, fig. 1 exhibits the regions of suboptimal parameters for the robust modification of the method based on the Chebyshev inequality for contamination of the «shift» type: $F(x) = 0,9N(x, 0, 1) + 0,1N(x, k, 1)$.

Appendix shows the areas of suboptimal parameters of the considered methods for some contamination models.

Performance evaluation: results

Dependence on the sample size. The statistics computed in each method can be highly dependent on the sample size. As a rule, with its growth, the value of statistics stabilises. Consequently, the detection quality is also stabilised. This is observed in fig. 2.

With an increase in the sample size, the performance of the method based on the Chebyshev inequality does not deteriorate and stabilises at a certain level. When dealing with the contamination types of «shift» and «scale», it shows results similar to the robust modifications of *N*-sigma. On contamination with the Cauchy distribution, it works much better than the classical *N*-sigma method and Tukey's boxplot, but is slightly inferior to the robust modifications of the *N*-sigma rule.

Dependence on the values of shift and scale

All of the above methods behave in almost the same way. The quality of detection monotonically increases and with a shift of about 5 it practically reaches 1. The classical *N*-sigma method is slightly inferior in quality to the robust method based on the Chebyshev inequality in outlier detection (see fig. 3).

When dealing with the contamination of the «scale» type, the situation is similar to the contamination of the «shift» type: the algorithm based on the Chebyshev inequality gives similar results as the classical *N*-sigma rule and its robust modifications, which are much better than the results of Tukey's boxplot.

Dependence on the contamination fraction

It can be seen in fig. 4 that in case of contamination of the «shift» type, the Tukey and *N*-sigma methods are less resistant to an increase in the contamination fraction than the robust modification of the Chebyshev inequality, the rejection quality of which fluctuates at the same level as the robust *N*-sigma modifications. A similar situation occurs with contamination of the «scale» type.

Slightly different behaviour can be seen in contamination of the Cauchy distribution type (see fig. 5). The method based on the Chebyshev inequality lags behind the robust modifications of the *N*-sigma rule in the quality of detection, but it still shows results significantly better than the boxplot method or the classical *N*-sigma rule.

Conclusion

1. Based on the classical Chebyshev inequality, an outlier detection algorithm is obtained, which, according to its classification rule, coincided with the *N*-sigma rule. Using the non-classical robust version of the Chebyshev inequality with $p = 1$, an outlier detection method is proposed, which proved to be very effective.

2. Based on the results of the comparative analysis, one can draw conclusions about the effectiveness of the method based on the Chebyshev inequality. It manifests itself as a fairly robust algorithm. On contamination such as «shift» and «scale», it is not inferior, and sometimes even outperforms the robust modifications of the *N*-sigma method. For all considered samples, it works better than Tukey's boxplot and the classical *N*-sigma rule. However, when the contamination is with the Cauchy distribution, it can give slightly worse results than robust modifications of the *N*-sigma rule.

3. It should be noted that the computational complexity of the statistics of the mean absolute deviation used in the method based on the Chebyshev inequality is slightly lower than the computational complexity of *MAD* or *FQ* scale estimates.

4. Practical recommendations are given on the choice of algorithm parameters on certain data models.

Appendix

Table 1

Boundary of suboptimal parameters area with «shift» contamination

k	N-sigma	N-sigma robustness (MAD)	N-sigma robustness (FQ)	Chebyshev inequality robustness
1	[0.606, 1.010]	[1.010, 1.414]	[0.606, 1.010]	[0.808, 1.212]
2	[0.808, 1.212]	[1.212, 2.020]	[0.808, 1.212]	[1.010, 1.616]
3	[1.010, 1.414]	[1.616, 2.828]	[1.010, 1.818]	[1.212, 2.020]
4	[1.010, 1.616]	[2.020, 3.434]	[1.212, 2.020]	[1.414, 2.424]
5	[1.010, 1.818]	[2.222, 4.444]	[1.414, 2.626]	[1.414, 2.828]
6	[1.010, 1.818]	[2.222, 5.455]	[1.414, 3.232]	[1.414, 3.232]
7	[1.010, 2.020]	[2.222, 6.667]	[1.414, 4.040]	[1.212, 3.636]
8	[1.010, 2.020]	[2.222, 8.080]	[1.414, 4.848]	[1.212, 4.040]
9	[0.808, 2.020]	[2.222, 8.889]	[1.414, 5.450]	[1.212, 4.040]
10	[0.808, 2.222]	[2.424, 10.303]	[1.414, 6.262]	[1.212, 4.444]

Note. Data distribution model: $F(x) = 0{,}9N(x, 0, 1) + 0{,}1N(x, k, 1)$.

Table 2

Boundary of suboptimal parameters area with «scale» contamination

k	N-sigma	N-sigma robustness (MAD)	N-sigma robustness (FQ)	Chebyshev inequality robustness
1	[0.606, 0.808]	[0.808, 1.212]	[0.606, 0.808]	[0.606, 1.010]
2	[0.808, 1.010]	[1.212, 1.818]	[0.808, 1.212]	[1.010, 1.414]
3	[0.808, 1.212]	[1.212, 2.424]	[0.808, 1.414]	[1.010, 1.818]
4	[0.808, 1.212]	[1.414, 2.626]	[1.010, 1.616]	[1.010, 1.818]
5	[0.606, 1.212]	[1.616, 3.030]	[1.010, 1.818]	[1.010, 2.020]
6	[0.606, 1.212]	[1.616, 3.232]	[1.010, 2.020]	[1.010, 2.020]
7	[0.606, 1.212]	[1.818, 3.636]	[1.212, 2.222]	[1.010, 2.020]
8	[0.606, 1.212]	[1.616, 3.636]	[1.010, 2.222]	[1.010, 2.020]
9	[0.606, 1.010]	[1.818, 3.838]	[1.212, 2.222]	[1.010, 2.020]
10	[0.606, 1.010]	[1.818, 3.838]	[1.212, 2.424]	[1.010, 2.222]

Note. Data distribution model: $F(x) = 0{,}9N(x, 0, 1) + 0{,}1N(x, 0, k)$.

Table 3

Boundary of suboptimal parameters area with Cauchy contamination

α	N-sigma	N-sigma robustness (MAD)	N-sigma robustness (FQ)	Chebyshev inequality robustness
0.01	[0.606, 1.000]	[1.010, 1.616]	[0.808, 1.010]	[0.808, 1.414]
0.031	[0.606, 0.808]	[1.010, 1.616]	[0.606, 1.010]	[0.808, 1.212]
0.052	[0.606, 0.606]	[0.808, 1.616]	[0.606, 1.212]	[0.606, 0.808]
0.073	[0.404, 0.808]	[1.010, 1.616]	[0.606, 1.010]	[0.808, 1.212]
0.094	[0.404, 0.808]	[1.010, 1.818]	[0.808, 1.212]	[0.808, 1.212]
0.115	[0.404, 0.606]	[1.010, 1.616]	[0.606, 1.010]	[0.808, 1.010]
0.136	[0.404, 0.606]	[1.010, 1.818]	[0.808, 1.212]	[0.606, 1.212]
0.157	[0.404, 0.606]	[1.010, 1.818]	[0.606, 1.212]	[0.606, 1.010]
0.178	[0.404, 0.404]	[1.010, 1.818]	[0.606, 1.010]	[0.606, 1.010]
0.2	[0.404, 0.404]	[1.010, 1.616]	[0.606, 1.010]	[0.606, 1.010]

Note. Data distribution model: $F(x) = (1 - \alpha)N(x, 0, 1) + \alpha C(x, 0, 1)$.

Библиографические ссылки

1. Tchebichef P. Des valeurs moyennes. *Journal de Mathematiques Pures et Appliquees.* 1867;12:177–184.

2. Shevlyakov G, Kan M. Stream data preprocessing: outlier detection based on the Chebyshev inequality with applications. In: *Proceeding of 26th Conference of Open Innovations Association (FRUCT); 2020 April 20–24; Yaroslavl, Russia.* [S. l.]: IEEE; 2020. p. 402–407. DOI: 10.23919/FRUCT48808.2020.9087459.

3. Shevlyakov GL, Oja H. *Robust correlation: theory and applications.* [S. l.]: Wiley; 2016. 352 p. (Wiley series in probability and statistics). DOI: 10.1002/9781119264507.

4. Андрэа К. *Методы и алгоритмы разведочного анализа данных, основанные на робастных модификациях боксплотов* [диссертация]. Санкт-Петербург: Санкт-Петербургский политехнический университет Петра Великого; 2013. 164 с.

5. Tukey JW. *Exploratory data analysis.* Reading, MA: Addison Wesley; 1977. 711 p.

References

1. Tchebichef P. Des valeurs moyennes. *Journal de Mathematiques Pures et Appliquees.* 1867;12:177–184.

2. Shevlyakov G, Kan M. Stream data preprocessing: outlier detection based on the Chebyshev inequality with applications. In: *Proceeding of 26th Conference of Open Innovations Association (FRUCT); 2020 April 20–24; Yaroslavl, Russia.* [S. l.]: IEEE; 2020. p. 402–407. DOI: 10.23919/FRUCT48808.2020.9087459.

3. Shevlyakov GL, Oja H. *Robust correlation: theory and applications.* [S. l.]: Wiley; 2016. 352 p. (Wiley series in probability and statistics). DOI: 10.1002/9781119264507.

4. Andrea K. *Metody i algoritmy razvedochnogo analiza dannykh, osnovannye na robastnykh modifikatsiyah boksplotov* [Methods and algorithms for exploratory data analysis based on robust boxplot modification] [dissertation]. Saint Petersburg: Peter the Great St. Petersburg Polytechnic University; 2013. 164 p. Russian.

5. Tukey JW. *Exploratory data analysis.* Reading, MA: Addison Wesley; 1977. 711 p.

Received by editorial board 28.09.2020.