UDC 004.02:519.234.7

OUTLIER DETECTION TECHNIQUE FOR HETEROGENEOUS DATA USING TRIMMED-MEAN ROBUST ESTIMATORS

Shved A. V. – Dr. Sc., Associate professor, Associate professor of Department of Software Engineering, Petro Mohyla Black Sea National University, Mykolayiv, Ukraine.

Davydenko Ye. O. – PhD, Associate professor, Head of Department of Software Engineering, Petro Mohyla Black Sea National University, Mykolayiv, Ukraine.

ABSTRACT

Context. Fortunately, the most commonly used in parametric statistics assumptions such as such as normality, linearity, independence, are not always fulfilled in real practice. The main reason for this is the appearance of observations in data samples that differ from the bulk of the data, as a result of which the sample becomes heterogeneous. The application in such conditions of generally accepted estimation procedures, for example, the sample mean, entails the bias increasing and the effectiveness decreasing of the estimates obtained. This, in turn, raises the problem of finding possible solutions to the problem of processing data sets that include outliers, especially in small samples. The object of the study is the process of detecting and excluding anomalous objects from the heterogeneous data sets.

Objective. The goal of the work is to develop a procedure for anomaly detection in heterogeneous data sets, and the rationale for using a number of trimmed-mean robust estimators as a statistical measure of the location parameter of distorted parametric distribution models.

Method. The problems of analysis (processing) of heterogeneous data containing outliers, sharply distinguished, suspicious observations are considered. The possibilities of using robust estimation methods for processing heterogeneous data have been analyzed. A procedure for identification and extraction of outliers caused by measurement errors, hidden equipment defects, experimental conditions, etc. has been proposed. The proposed approach is based on the procedure of symmetric and asymmetric truncation of the ranked set obtained from the initial sample of measurement data, based on the methods of robust statistics. For a reasonable choice of the value of the truncation coefficient, it is proposed to use adaptive robust procedures. Observations that fell into the zone of smallest and lowest ordinal statistics are considered outliers.

Results. The proposed approach allows, in contrast to the traditional criteria for identifying outlying observations, such as the Smirnov (Grubbs) criterion, the Dixon criterion, etc., to split the analyzed set of data into a homogeneous component and identify the set of outlying observations, assuming that their share in the total set of analyzed data is unknown.

Conclusions. The article proposes the use of robust statistics methods for the formation of supposed zones containing homogeneous and outlying observations in the ranked set, built on the basis of the initial sample of the analyzed data. It is proposed to use a complex of adaptive robust procedures to establish the expected truncation levels that form the zones of outlying observations in the region of the lowest and smallest order statistics of the ranked dataset. The final level of truncation of the ranked dataset is refined on the basis of existing criteria that allow checking the boundary observations (minimum and maximum) for outliers.

KEYWORDS: outliers, robust estimates, trimmed mean, symmetric and asymmetric truncation.

NOMENCLATURE

X is an initial data set;

X^* is an ordered sample, constructed on the basis of X;

n is a data sample size;

$x_{(i)}$ is an order statistics from the data set X^*;

α is a trimming proportion;

α_L is a trimming proportion of smallest order statistics of X^* (an asymmetric case);

α_U is a trimming proportion of largest order statistics of X^* (an asymmetric case);

ε is a proportion of outliers;

$L(\alpha)$ is an average of $[\alpha n]$ smallest order statistics of X^*;

$U(\alpha)$ is an average of $[\alpha n]$ largest order statistics of X^*;

$M(\alpha)$ is an average of $[\alpha n]$ middle order statistics of X^*;

Es_{sym} is a group of robust estimators based on symmetric trimming;

Es_{asym} is a group of robust estimators based on asymmetric trimming;

$T_j(\alpha)$ is a symmetric trimmed mean obtained in accordance with the selected adaptive robust estimator;

$T_j(\alpha_L, \alpha_U)$ is an asymmetric trimmed mean obtained in accordance with the selected adaptive robust estimator;

$e_{T_j(\alpha)}$ is a standard error of the symmetric trimmed mean;

$e_{T_j(\alpha_L,\alpha_U)}$ is a standard error of the asymmetric trimmed mean;

Al is a set of possible truncation levels;

$u_{(i)}$ is a statistics of Smirnov criterion;

u_α is a critical value of the Smirnov criterion;

$G(x;\ \bar{x};\ \sigma^2)$ is a normal distribution with mean $\bar{x}$ and variance σ^2;

$[y]$ is an integer part of y.

INTRODUCTION

In the act of processing of real data that are collected on the basis of registration (observation) methods, measurements (experiments, tests) or the participation of third parties (interviews, focus groups, expert assessment methods), analysts are often faced with a situation when data sets (samples) have observations that, to one degree or another, differ (stand out) from the rest in terms of the analyzed attribute.

DOI 10.15588/1607-3274-2022-3-5

In statistical analysis, such observations were called “abnormal”, “outliers”, “sharply distinguished”, “suspicious”, etc. The gross errors, measurement errors, failures of measuring equipment, human factors (operator errors), as well as short-term abrupt changes in measurements (experiments), for example, vibration, can cause of them. The share of outliers is usually about 5% to 10% of observations in the total dataset [1–4], which disturbs their homogeneity. The appearance of abnormal data in the samples is the reason for so-called “heavy tails”, multimodality, pronounced asymmetry, kurtosis, etc. This, in turn, does not allow such data samples to be represented by rigorous parametric models that are described by the well-known probability distributions (e. g. Normal, Poisson, Student distributions, etc.) and characterized by two main parameters: the location parameter, which is the population or sample mean, median, mode; the scale parameter, which can be represented by variance, standard deviation, peak-to-peak, etc.

The use of generally accepted assessment procedures in such conditions, which are based on an explicit or implicit assumption of normality, entails the bias increasing and reducing the effectiveness of the obtained estimates.

In the context of the analysis of expert information, the use of survey data processing methods, which are based on the procedure of their averaging, will be justified only if there is a sufficiently high consistency (proximity) of expert assessments.

The object of study is the process of detecting and excluding anomalous objects from the heterogeneous data sets.

The subject of study is procedures, methods and technique for finding an unusual observations (outliers) in analyzed data sets.

The purpose of the work is to develop a procedure for anomaly detection in heterogeneous data sets, and the rationale for using a number of trimmed-mean robust estimators as a statistical measure of the location parameter of distorted parametric distribution models.

1 PROBLEM STATEMENT

Let, $X = (x_1, x_2, \dots, x_n)$ be a set of values measured by some parameter, where n is a sample X size.

The task is to identify the $X' \subseteq X$ region, that contains homogeneous data and the $X'' \subset X$, $X' \cup X'' = X$ region, that is anomalous in relation to the X' region values (outliers).

By a homogeneous component $X' \subseteq X$ we mean an area such that all $x_j \in X'$, $j = \overline{1, |X'|}$ are independent of each other and have the same probability distribution density (same values of characteristics (parameters) under the same distribution).

If the proportion of anomalous data ($\varepsilon = |X''| / |X|$) is more then 0.5, the further analysis is impractical; if the level of “clogging” is 0.5, only a sample median can be recommended as a robust estimate for finding the average of such a sample.

2 REVIEW OF THE LITERATURE

To process data sets containing heterogeneous observations, the following approaches are used [1, 2, 4, 5, 6]: the identification and exclusion of outliers; the application of robust methods for statistical analysis of data with outliers. To solve the first problem, there is currently a whole class of parametric and nonparametric methods for anomalous observations identification [5–9]. Parametric methods are based on complete a priori information about observations, their application presupposes a priori knowledge of the theoretical distribution of the investigated values or its determination from empirical data. Nonparametric methods do not use detailed a priori information about observations and can be used when the distribution of the indicator under study is unknown and there is no need for its analytical description. An important condition for using nonparametric methods is that the distribution functions of analyzed measurements must be continuous. The effectiveness of the application of these procedures largely depends on the size of the sample under study and the power of the selected criteria.

One aim of robust statistics is to develop evaluation procedures that are resistant to the appearance of outliers in data sets, and to obtain unbiased (or slightly biased) and effective estimates. Currently, there are next classes of robust estimates [1–4, 10]: robust estimators based on Maximum-likelihood argument (*M*-estimators); robust estimators based on rank statistics (*R*-estimators); robust estimators based on a linear combination of order statistics (*L*-estimators). Robust *L*-estimators are most widespread due to the simplicity of their computational implementation [10–13]. These estimates include truncated, censored, Winsorized means, sample median, etc. The main problem of the considered estimators is the choice of the trimming proportion α, which can be sufficiently solved by using adaptive robust procedures for statistical data estimation [13–16].

3 MATERIALS AND METHODS

Most of the existing criteria for testing outlying observations are based on the assumption that the distribution of measurements corresponds to the normal distribution [5–9]. To search for and filter out sharply distinguished observations in small samples, the most widespread and theoretical justification was obtained by Grubbs-type statistics such us Smirnov (Grubbs) test, Tietjen-Moore test, Dixon test [6, 8, 9]. These criteria provide checking either one outlier (smallest or largest), or two (two smallest or two largest in the sample). At the same time, there is a problem of searching for outliers if their share in the total set of measurement data is unknown.

In this regard, the problem of finding a homogeneous component of the set of measurement data is urgent. It is assumed that the measurement results have approximately normal distribution. Modern research has shown that the procedure for checking the analyzed data samples for compliance with the Gaussian distribution is a rather difficult task, especially for analyzing samples with limited data volume ($n \le 50$). Currently, there is a fairly extensive

DOI 10.15588/1607-3274-2022-3-5

class of goodness-of-fit tests [17–19], applicable for small data samples, for example, the nonparametric Shapiro-Wilk test [18], the Sarkadi test [19]. At the same time, it was proved that in small samples case it is not always possible to distinguish the normal distribution from other types of distributions.

Under these conditions, the paper proposes to use a complex of adaptive robust estimates to identify a homogeneous component of the analyzed data set. However, in this case, the problem arises of choosing robust estimates that can recommend different levels of truncation of the ordered sample built on the basis of the analyzed sample of measurement data. The trimming proportion has a direct impact on the size of the area containing homogeneous data, i.e. an area that does not contain outliers. To clarify (expand, or narrow) the area of homogeneous data, the Grubbs-type test was used.

Let us consider the main stages of the proposed procedure for searching and eliminating outliers in the studied data set.

Stage 1. Formation of truncation levels of the ordered sample X^*.

1.1. Construction on the basis of a set of measurement data X the ordered sample values X^*: $x_{(1)} \le x_{(2)} \le \ldots \le x_{(i)} \le \ldots \le x_{(n)}$. Denote by $x_{(i)}$ the order statistics from the data set X^*.

1.2. Calculation the robust measure of tail length [13–15]:

$$Q = (U_{0.05} - L_{0.05}) / (U_{0.5} - L_{0.5}) , \tag{1}$$

$$Q_1 = (U_{0.2} - L_{0.2}) / (U_{0.5} - L_{0.5}) . \tag{2}$$

1.3. Estamation the tail length of distribution, for example [14, 15]:

$$\begin{array}{ll} Q < 2.0 & \text{light tailed;} \\ 2.0 \le Q \le 2.6 & \text{medium tailed;} \\ 2.6 \le Q \le 3.2 & \text{heavy tailed;} \\ Q > 3.2 & \text{very heavy tailed.} \end{array} \tag{3}$$

$$\begin{array}{ll} Q_1 < 1.81 & \text{light tailed;} \\ 1.81 \le Q_1 \le 1.87 & \text{medium tailed;} \\ Q_1 > 1.87 & \text{heavy tailed.} \end{array} \tag{4}$$

1.4. Test X^* for the symmetry of the distribution, based on measure of the tail length, *e.g.* according to robust measure HeQ_1 [14, 20]:

$$\begin{array}{ll} HeQ_1 < 0.7 & \text{left - skewed model;} \\ 0.7 \le HeQ_1 \le 1.4 & \text{symmetric model;} \\ HeQ_1 > 1.4 & \text{right - skewed model.} \end{array} \tag{5}$$

where HeQ_1 can be defined by next formula:

$$HeQ_1 = U_{(0.05)} - M(0.5) \big/ M(0.5) - L_{(0.05)} \tag{6}$$

1.5. Selection of a group of robust estimators based on order statistics Est (Es_{sym}, Es_{asym}):

$$Est = \begin{cases} Es_{sym}, & (HeQ_1 < 0.7) \vee (HeQ_1 > 1.4); \\ Es_{asym}, & 0.7 \le HeQ \le 1.4. \end{cases} \tag{7}$$

1.6. Calculation the value of robust estimators within the selected group Est.

1.7. Selection of estimator $T_j(\alpha) \in Est$, which are satisfied the condition $\min\limits_j (e_{T_j(\alpha)})$ or $\min\limits_j (e_{T_j(\alpha_L,\alpha_U)})$, $j = \overline{1,k}$.

Setting the truncation levels α (a symmetric case), or α_L and α_U (an asymmetric case; $\alpha = \alpha_L + \alpha_U$).

1.8. Formation of possible truncation levels, Fig.1:

– a symmetric case

$$Es_{sym}\text{:}\ \ Al = \{\alpha_t \mid t = \overline{1,z}\},\ \alpha_t > \alpha,\ \forall T(\alpha) \in Est;$$

– an asymmetric case

$$Es_{asym}\text{:}\ \ Al_L = \{\alpha_L^t \mid t = \overline{1,z_1}\},\ \alpha_L^t > \alpha_L,$$
$$\forall T(\alpha_L,\alpha_U) \in Est;\ \ Al_U = \{\alpha_U^s \mid s = \overline{1,z_2}\},\ \alpha_U^s > \alpha_U.$$

The obtained values are sorted in ascending order.

Stage 2. Checking the boundary elements $x_{(g)}$, $x_{(n-g+1)}$, $g = [\alpha n]$, (symmetric case) or $x_{(g1)}$, $x_{(n-g2+1)}$, $g_1 = [\alpha_L n]$, $g_2 = [\alpha_U n]$ (asymmetric case) of the X^* on anomaly.

2.1 Formulation of the null and alternative hypothesis:

2.1.1 the E_{sym} group has been selected.

H_0: the boundary observation ($x_{(g)}$ or $x_{(n-g+1)}$) belongs to the same general population as the rest ($q = n - 2g - 1$) of the central values of the ordered sample.

H_1: the boundary observation ($x_{(g)}$ or $x_{(n-g+1)}$) is outlier.

2.1.2 the E_{asym} group has been selected.

H_0: the boundary observation ($x_{(g1)}$ or $x_{(n-g2+1)}$) belongs to the same general population as the rest ($q = n - g_1 - g_2 - 1$) of the central values of the ordered sample.

H_1: the boundary observation ($x_{(g1)}$ or $x_{(n-g2+1)}$) is outlier.

IF ($u_{(g1)} \le u_\alpha$), or ($u_{(n-g2+1)} \le u_\alpha$), THEN H_0 is accepted.

The elements $x_{(g1)}$ and $x_{(n-g2+1)}$ are checked alternately.

Stage 3. Outliers exclusion, Fig. 2–3.

3.1 IF H_0 is accepted:

3.1.1 for $x_{(n-g+1)}$ (or $x_{(n-g2+1)}$ respectively), THEN the next senior member of the ordered sample is tested until the element $x_{(s)}$ is found, for which H_0 will be rejected.

Then the group of senior members of the ordered sample $x_{(s)} \le \ldots \le x_{(n)}$ is considered as outliers.

3.1.2 for $x_{(g)}$ (or $x_{(g1)}$ respectively), THEN the previous junior member of the ordered sample is tested until the element $x_{(s)}$ is found, for which H_0 will be rejected.

DOI 10.15588/1607-3274-2022-3-5

Then the group of junior members of the ordered sample $x_{(1)} \le \ldots \le x_{(s)}$ is considered as outliers

3.2. IF H_1 is accepted:

3.2.1 for $x_{(n-g+1)}$ (or $x_{(n-g2+1)}$ respectively), THEN the truncation levels $\alpha_t \in Al$ (or $\alpha_U^s \in Al_U$ respectively) are consistently selected.

Repeat the procedure of items 2.1–3.2.

3.2.2 for $x_{(g)}$ (or $x_{(g1)}$ respectively), THEN the truncation levels $\alpha_t \in Al$ (or $\alpha_L^t \in Al_L$ respectively) are consistently selected.

Repeat the procedure of items 2.1–3.2.

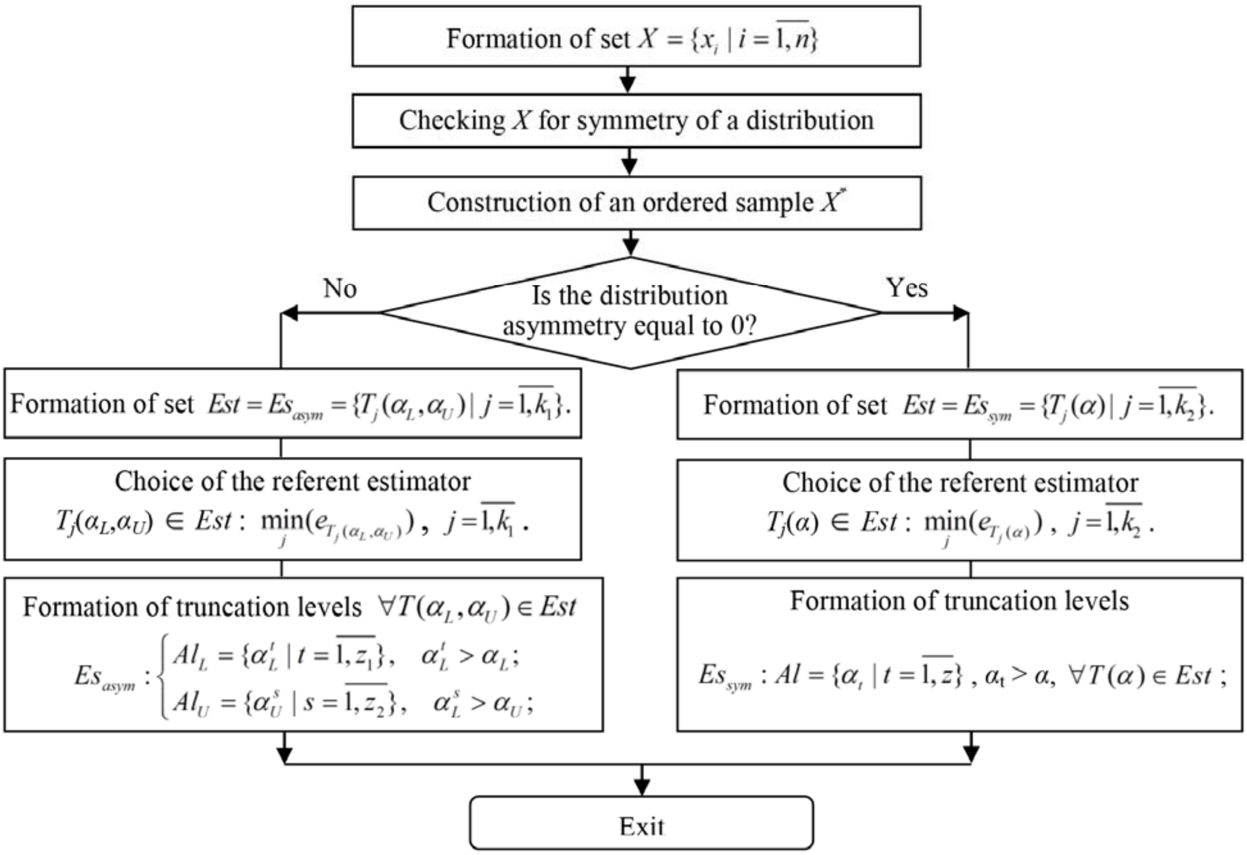
Formation of set $X = \{x_i \mid i = \overline{1,n}\}$
Checking X for symmetry of a distribution
Construction of an ordered sample X^*
Is the distribution asymmetry equal to 0?
No
Yes
Formation of set $Est = Es_{asym} = \{T_j(\alpha_L,\alpha_U) \mid j = \overline{1,k_1}\}$.
Formation of set $Est = Es_{sym} = \{T_j(\alpha) \mid j = \overline{1,k_2}\}$.
Choice of the referent estimator $T_j(\alpha_L,\alpha_U) \in Est$: $\min_j(e_{T_j(\alpha_L,\alpha_U)})$, $j = \overline{1,k_1}$.
Choice of the referent estimator $T_j(\alpha) \in Est$: $\min_j(e_{T_j(\alpha)})$, $j = \overline{1,k_2}$.
Formation of truncation levels $\forall T(\alpha_L,\alpha_U) \in Est$
$Es_{asym} : \begin{cases} Al_L = \{\alpha_L^t \mid t = \overline{1,z_1}\}, & \alpha_L^t > \alpha_L; \\ Al_U = \{\alpha_U^s \mid s = \overline{1,z_2}\}, & \alpha_L^s > \alpha_U; \end{cases}$
Formation of truncation levels
$Es_{sym} : Al = \{\alpha_t \mid t = \overline{1,z}\}$, $\alpha_t > \alpha$, $\forall T(\alpha) \in Est$;
Exit

Figure 1 – Algorithm for truncation levels formation

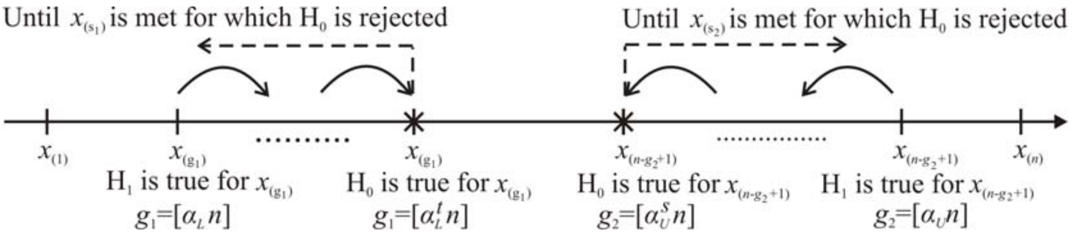
Until $x_{(s_1)}$ is met for which H_0 is rejected
Until $x_{(s_2)}$ is met for which H_0 is rejected
$x_{(1)}$
$x_{(g_1)}$
$x_{(g_1)}$
$x_{(n-g_2+1)}$
$x_{(n-g_2+1)}$
$x_{(n)}$
H_1 is true for $x_{(g_1)}$
$g_1=[\alpha_L n]$
H_0 is true for $x_{(g_1)}$
$g_1=[\alpha_L^t n]$
H_0 is true for $x_{(n-g_2+1)}$
$g_2=[\alpha_U^s n]$
H_1 is true for $x_{(n-g_2+1)}$
$g_2=[\alpha_U n]$

Figure 2 – The procedure for selection the initial truncation levels in asymmetric case

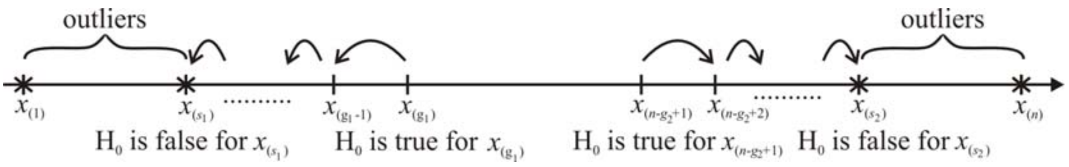
outliers
outliers
$x_{(1)}$
$x_{(s_1)}$
$x_{(g_1-1)}$
$x_{(g_1)}$
$x_{(n-g_2+1)}$
$x_{(n-g_2+2)}$
$x_{(s_2)}$
$x_{(n)}$
H_0 is false for $x_{(s_1)}$
H_0 is true for $x_{(g_1)}$
H_0 is true for $x_{(n-g_2+1)}$
H_0 is false for $x_{(s_2)}$

Figure 3 – The procedure for outliers’ determination in asymmetric case

DOI 10.15588/1607-3274-2022-3-5

4 EXPERIMENTS

Let us consider an example of the implementation of the proposed method. To model the set of measurement data, a mathematical model of independent observations with a Gaussian distribution $X = \{x_i \mid i = \overline{1,n}\}$ (volume n = 500) has been constructed.

The constructed model is a two-component symmetric mixture of normal distributions of the form:

$$F(x) = (1-\varepsilon)G(x;\bar{x};\sigma_0^2) + \varepsilon G(x;\bar{x};\sigma_1^2), \qquad (8)$$

where $\sigma_0^2 < \sigma_1^2$; $0 \le \varepsilon \le 0.5$.

Using the Tukey contamination model (8), a mixture was generated with the following parameters:

– the main Gaussian distribution $G(x, \bar{x}, \sigma_0^2)$ with parameters $\bar{x} = 0$, $\sigma_0 = 0.7$;

– the contaminated Gaussian distribution $G(x, \bar{x}, \sigma_1^2)$ with parameters $\bar{x} = 0$, $\sigma_1 = 1.2$;

– the proportion of outliers ε = 25%.

The test for symmetry of distribution was performed on the basis of the HeQ_1 robust estimator (6). Based on the results of such verification, a group of symmetric truncation estimators $Est = Es_{sym}$ was selected [10, 13–16].

5 RESULTS

The results of the analysis of the *Est*-group estimators are presented in the Table 1.

Table 1 – The analysis of *Est*-group estimators

№, *j*	Estimator	$T_j(\alpha)$	α, %	$e_{T_j}(\alpha)$	Region without outliers
1	NH_1	–0.0109	12.5	0.0353	$[X^*_{63}; X^*_{438}]$
2	NH_2	–0.0126	18.75	0.03537	$[X^*_{94}; X^*_{407}]$
3	HGP	–0.0078	25	0.0380	$[X^*_{126}; X^*_{375}]$
4	HG_1	0.0018	0	0.0369	$[X^*_{1}; X^*_{500}]$
5	HG_2	–0.0078	25	0.0380	$[X^*_{126}; X^*_{375}]$
6	PAR	–0.0078	25	0.0380	$[X^*_{126}; X^*_{375}]$
7	PR_1	–0.0116	20	0.0355	$[X^*_{101}; X^*_{400}]$
8	PR_2	0.0018	0	0.0369	$[X^*_{1}; X^*_{500}]$
9	A_0	–0.0081	15	0.03538	$[X^*_{76}; X^*_{425}]$
10	A_1	–0.0081	15	0.03538	$[X^*_{76}; X^*_{425}]$
11	A_2	–0.0122	15	0.03538	$[X^*_{76}; X^*_{425}]$
12	A_3	–0.0116	20	0.0355	$[X^*_{101}; X^*_{400}]$

The results of Table 1 show that $\min_j(e_{T_j(\alpha)})$ corresponds to the Hogg estimator NH_1, (α = 12.5%), respectively the next subset of truncation levels was formed Al={0.15; 0.1875; 0.2; 0.25}, $\alpha_t > \alpha$, $\forall T(\alpha) \in Est$.

Based on the results of the verification of the boundary values $x_{(g)}$, $x_{(n-g+1)}$, at α = 12.5% (NH_1 estimator) according to Smirnov test at 0.05 significance level H_0 was accepted, the limit observations belong to the same general population, as well as other central values of the ordered sample.

Thus, two regions $[X^*_1; X^*_{62}]$ and $[X^*_{439}; X^*_{500}]$ were formed for testing the limit values for the anomaly by Smirnov criterion in accordance with the scheme shown in Fig. 3.

According to the obtained results, two regions of outliers were obtained: $[X^*_1; X^*_{15}]$ and $[X^*_{484}; X^*_{500}]$.

6 DISCUSSION

The procedure for outliers searching in analyzed data set paper has been proposed. To highlight a homogeneous component of the set of measurement data, it was proposed to use robust estimates based on a linear combination of order statistics, which are used the procedure of both symmetric and asymmetric truncation. In the symmetric case, [α*n*] smallest and [α*n*] largest observations are considered as outliers, which violate homogeneity of analyzed data set. In the asymmetric case, the truncation ratio α is additionally divided into the proportions α_L and α_U, which corresponds to the truncation levels of the [α*n*] smallest and [α*n*] largest observations. The main problem of such type estimators is the choice of the value of the truncation coefficient α, which can be sufficiently solved by using adaptive robust procedures. With the adaptive approach, a specific type of auxiliary measures of tail-length, skewness and kurtosis. To choice the initial (reference) level of truncation of ordered sample, an adaptive estimate was chosen that minimizes the value of the standard error of the truncated mean.

CONCLUSIONS

The problem of outliers detection in heterogeneous data sets has been studied.

The scientific novelty of obtained results is that the methods of detection and exclusion of observations distorted by measurement errors (anomalous observations) due to hidden defects of the equipment, operating conditions of the equipment, and other conditions, are received the further development. The mathematical apparatus of nonparametric statistics was used to process the results of observations and search for outliers. The proposed approach is based on the procedure of truncation of ordered samples obtained on the basis of the initial sample of measurement data. Observations that fall into the region of [$\alpha_L n$] smallest and [$\alpha_U n$] largest order statistics are considered outliers. To form possible levels of truncation of the ordered sample, data processing algorithms using adaptive robust statistical estimation procedures were used, which allowed to formalize the procedure for selecting the level of truncation.

The practical significance of the obtained results lies in the fact that the proposed approach allows, along with the traditional tests for outliers detection, to single out a set of abnormal measurement results, if their share in the total set of measurement data is unknown. This, in turn, expands the capabilities of existing algorithms for searching the outliers, which is ultimately aimed at increasing the reliability of statistical processing the results of observations (primary measurements).

Prospects for further research are aimed at developing a procedure for a smoother selection of the truncation coefficients in asymmetric case.

DOI 10.15588/1607-3274-2022-3-5

ACKNOWLEDGEMENTS

The work is supported by the state research project of Petro Mohyla Black Sea National University "Development of modern information and communication technologies for the management of intellectual resources to decision-making support of operational management" (research project no. 0121U107831, financed by the Government of Ukraine).

REFERENCES

1. Hampel F. R., Ronchetti E. M., Rousseeuw P. J., Stahel W. A. Robust statistics: the approach based on influence functions. New York, John Wiley & Sons, 1986, 526 p. DOI: 10.1002/9781118186435.
2. Hoaglin D. C., Mosteller C. F., Tukey J. W. Understanding robust and exploratory data analysis. New York, Wiley-Interscience, 2000, 447 p.
3. Staudte R. G., Sheather S. J. Robust estimation and testing. New York, Wiley, 1990, 351 p.
4. Huber P., Ronchetti E. M. Robust statistics. New York, Wiley, 2009, 363 p. DOI: 10.1002/9780470434697.
5. Cousineau D., Chartier S. Outliers detection and treatment: a review, *International Journal of Psychological Research,* 2010, Vol. 3(1), pp. 58–67. DOI: 10.21500/20112084.844.
6. Hawkins D. M. Identification of outliers. London, Chapman and Hall, 1980, 188 p. DOI: 10.1002/bimj.4710290215
7. Aggarwal C. Outlier analysis. New York, Springer, 2016, 488 p. DOI 10.1007/978-3-319-47578-3
8. Manoj K., Senthamarai K. Comparison of methods for detecting outliers, *International Journal of Scientific & Engineering Research,* 2013, Vol. 4, pp. 709–714.
9. Saleem S., Aslam M., Shaukat M. R. A review and empirical comparison of univariate outlier detection methods, *Pakistan Journal of Statistics,* 2021, Vol. 37(4), pp. 447–462.
10. Andrews D. F., Bickel P. J., Hampel F. R., P. J. Huber, W. H. Rogers, J. W. Tukey Robust estimates of location: survey and advances. Princeton, NJ, Princeton University Press, 1972, 372 p.
11. Balakrishnan N. Parameters, order statistics, outliers and robustness, *Revista Matematica Complutence,* 2007, Vol. 20(1), pp. 7–107. DOI: 10.5209/rev_rema.2007.v20.n1.16528
12. Ramsey P. H., Ramsey P. P. Optimal trimming and outlier elimination, *Journal of Modern Applied Statistical Methods,* 2007, Vol. 6, pp. 355–360. DOI: 10.22237/jmasm/1193889660
13. Reed III J. F., Stark D. B. Robustness estimators of locations: a classification of linear and adaptive estimators, *Journal of Applied Statistics,* 1994, Vol. 21(3), pp. 95–124. DOI: 10.1080/757583650
14. Reed III J. F., Stark D. B. Contributions to adaptive estimators, *Journal of Applied Statistics*, 1998, Vol. 25(5), pp. 651–669. DOI: 10/1080/02664769822882
15. Hogg R. V., Lenth R. V. A review of some adaptive statistical techniques, *Communications in Statistics – Theory and Methods*, 1984, Vol. 13, pp. 1551–1579. DOI: 10.1080/03610928408828779
16. Prescott P. Selection of trimming proportions for robust adaptive trimmed mean, *Journal of the American Statistical Association,* 1978, No. 73(361), pp. 133–140. DOI: 10.2307/2286534
17. Zylstra R. R. Normality tests for small sample sizes, *Quality Engineering,* 1994, Vol. 7(1), pp. 45–58. DOI: 10.1080/08982119408918766
18. Shapiro S. S., Wilk M. B. An analysis of variance test for normality (complete samples), *Biometrika,* 1965, Vol. 52(3/4), pp. 591–611. DOI: 10.2307/2333709
19. Sarkadi K. Testing for normality, *Banach Center Publications,* 1980, Vol. 6, pp. 281–287. DOI: 10.4064/-6-1-281-287
20. Hersgaard D. Distributions of asymmetric trimmed means, *Communications in Statistics: Simulation and Computation*, 1979, Vol. 8(4), pp. 359–367. DOI: 10.1080/03610917908812125

Received 13.06.2022.
Accepted 27.08.2022.

УДК 004.02:519.234.7

ПРОЦЕДУРА ВИЯВЛЕННЯ АНОМАЛЬНИХ СПОСТЕРЕЖЕНЬ У НАБОРАХ НЕОДНОРІДНИХ ДАНИХ З ВИКОРИСТАННЯМ РОБАСТНИХ УСІЧЕНИХ ОЦІНОК

Швед А. В. – д-р техн. наук, доцент, доцент кафедри інженерії програмного забезпечення Чорноморського національного університету імені Петра Могили, Миколаїв, Україна.

Давиденко Є. О. – канд. техн. наук, доцент, завідувач кафедри інженерії програмного забезпечення Чорноморського національного університету імені Петра Могили, Миколаїв, Україна.

АНОТАЦІЯ

Актуальність. Загальноприйняті припущення в параметричній статистиці, такі як нормальність, лінійність, незалежність, далеко не завжди виконуються у реальній практиці. Основною причиною тому є поява спостережень у досліджуваних вибірках даних, що відрізняються від основної маси даних, внаслідок чого вибірка стає неоднорідною. Застосування в таких умовах загальноприйнятих процедур оцінювання, наприклад, вибіркового середнього, тягне за собою збільшення зсуву та зниження ефективності одержуваних оцінок. Це в свою чергу висуває задачу пошуку можливих шляхів вирішення проблеми обробки масивів даних, що містять аномальні спостереження, особливо в умовах обробки вибірок малого обсягу. Об'єкт дослідження – процес виявлення та виключення аномальних спостережень у вибірках неоднорідних даних. Мета роботи – розробка процедури пошуку аномальних спостережень у вибірках неод-

DOI 10.15588/1607-3274-2022-3-5

норідних даних, та обґрунтування використання низки усічених оцінок типу «середнє» для оцінювання параметру положення спотворених параметричних моделей розподілів.

Метод. Розглянуті питання аналізу (обробки) неоднорідних даних, що містять аномальні, підозрілі спостереження. Проаналізовано можливості використання робастних процедур оцінювання, стійких до наявності викидів у вибірках неоднорідних даних. Запропоновано процедуру виявлення та виключення аномальних спостережень, причиною яких можуть бути помилки вимірювань, приховані дефекти апаратури, вироблення ресурсів, умови проведення експерименту тощо. В основу запропонованого підходу покладено процедуру симетричного та несиметричного усічення варіаційного ряду, отриманого на основі вихідної вибірки даних, на основі методів робастної статистики. Для обґрунтованого вибору величини коефіцієнта усічення α, запропоновано використовувати адаптивні робастні процедури статистичного оцінювання. Спостереження, що потрапили до зони молодших та зони старших порядкових статистик, визнані аномальними.

Результати. Запропонований підхід дозволяє на відміну від традиційних критеріїв пошуку аномальних значень, таких як критерій Смірнова(Граббса), критерій Діксона та ін., розбивати аналізовану сукупність даних на однорідну складову та виявляти сукупність аномальних спостережень, при припущенні, що їх частка у загальній сукупності аналізованих даних невідома.

Висновки. У статті запропоновано використання методів робастної статистики для формування передбачуваних зон, що містять однорідні та аномальні спостереження у варіаційному ряді, побудованому за вихідною вибіркою аналізованих даних. Запропоновано використовувати комплекс адаптивних робастних процедур для встановлення рівнів усічення, що утворюють зони аномальних спостережень в області старших та молодших порядкових статистик. Остаточний рівень усічення варіаційного ряду уточнюється на основі існуючих критеріїв, що дозволяють перевіряти граничні спостереження (мінімальне та максимальне) на аномальність.

КЛЮЧОВІ СЛОВА: викиди, робастні оцінки, усічене середнє, симетричне і несиметричне усічення.

УДК 004.802:519.234.7

ПРОЦЕДУРА ВЫЯВЛЕНИЯ АНОМАЛЬНЫХ НАБЛЮДЕНИЙ В НАБОРАХ НЕОДНОРОДНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ РОБАСТНЫХ УСЕЧЕННЫХ ОЦЕНОК

Швед А. В. – д-р техн. наук, доцент, доцент кафедры инженерии программного обеспечения Черноморского национального университета имени Петра Могилы, Николаев, Украина.

Давыденко Е. А. – канд. техн. наук, доцент, заведующий кафедрой инженерии программного обеспечения Черноморского национального университета имени Петра Могилы, Николаев, Украина.

АННОТАЦИЯ

Актуальность. Общепринятые предположения в параметрической статистике, такие как нормальность, линейность, независимость, далеко не всегда выполняются в реальной практике. Основной причиной этого является появление наблюдений в исследуемых выборках данных, отличающихся от основной массы данных, вследствие чего выборка становится неоднородной. Применение в этих условиях общепринятых процедур оценивания, например, выборочного среднего, влечет за собой увеличение смещенности и снижение эффективности получаемых оценок. Это в свою очередь выдвигает задачу поиска возможных путей решений проблемы обработки массивов данных (выборок), содержащих аномальные наблюдения, особенно в условиях обработки выборок малого объема. Объект исследования – процесс обнаружения и исключения аномальных объектов выборки неоднородных данных. Цель работы – разработка процедуры поиска аномальных наблюдений в выборках неоднородных данных, и обоснование использования ряда усеченных оценок типа «среднее» для оценивания параметра положения искаженных параметрических моделей распределений.

Метод. Рассмотрены вопросы анализа (обработки) неоднородных данных, содержащие аномальные, резко выделяющиеся, подозрительные наблюдения. Проанализированы возможности использования робастных процедур оценивания, устойчивых к наличию в выборках данных «засоряющих» значений, для обработки неоднородных данных. Предложена процедура выявления и исключения аномальных наблюдений, причиной которых могут быть ошибки измерений, скрытые дефекты аппаратуры, выработка ресурсов, условия проведения эксперимента и т.д. В основу предложенного подхода положена процедура симметричного и несимметричного усечения вариационного ряда, полученного на основе исходной выборки данных, на основе методов робастной статистики. Для обоснованного выбора величины коэффициента усечения α, предложено использовать адаптивные робастные процедуры статистического оценивания. Наблюдения, которые находятся в зоне младших и зоне старших порядковых статистик, признаны аномальными.

Результаты. Предложенный подход позволяет в отличие от традиционных критериев поиска аномальных значений, таких как критерий Смирнова(Граббса), критерий Диксона и др., разбивать анализируемую совокупность данных на однородную составляющую и выявлять совокупность аномальных наблюдений, при предположении, что их доля в общей совокупности анализируемых данных неизвестна.

Выводы. В статье предложено использование методов робастной статистики для формирования предполагаемых зон, содержащих однородные и аномальные наблюдения в вариационном ряде, построенном по исходной выборке анализируемых данных. Предложено использовать комплекс адаптивных робастных процедур, для установления

DOI 10.15588/1607-3274-2022-3-5

предполагаемых уровней усечения, образующих зоны аномальных наблюдений в области старших и младших порядковых статистик вариационного ряда. Окончательный уровень усечения вариационного ряда уточняется на основе существующих критериев, позволяющих проверять граничные наблюдения (минимальное и максимальное) на аномальность.

КЛЮЧЕВЫЕ СЛОВА: выбросы, робастные оценки, усеченное среднее, симметричное и не симметричное усечение.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Hampel F. R. Robust statistics: the approach based on influence functions / F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel. – New York: John Wiley & Sons, 1986. – 526 p. DOI: 10.1002/9781118186435.
2. Hoaglin D. C. Understanding robust and exploratory data analysis / D. C. Hoaglin, C. F. Mosteller, J. W. Tukey. – New York: Wiley-Interscience, 2000. – 447 p.
3. Staudte R. G. Robust estimation and testing / R. G. Staudte, S. J. Sheather. – New York : Wiley, 1990. – 351 p.
4. Huber P. Robust statistics / P. Huber, E. M. Ronchetti. – New York: Wiley, 2009. – 363 p. DOI: 10.1002/9780470434697.
5. Cousineau D. Outliers detection and treatment: a review / D. Cousineau, S. Chartier // International Journal of Psychological Research. – 2010. – Vol. 3(1). – P. 58–67. DOI: 10.21500/20112084.844.
6. Hawkins D. M. Identification of outliers / D. M. Hawkins. – London: Chapman and Hall, 1980. – 188 p. DOI: 10.1002/bimj.4710290215
7. Aggarwal C. Outlier analysis / C. Aggarwal. – New York: Springer, 2016. – 488 p. DOI 10.1007/978-3-319-47578-3
8. Manoj K. Comparison of methods for detecting outliers / K. Manoj, K. Senthamarai // International Journal of Scientific & Engineering Research. – 2013. – Vol. 4. – P. 709–714.
9. Saleem S. A review and empirical comparison of univariate outlier detection methods / S. Saleem, M. Aslam, M. R. Shaukat // Pakistan Journal of Statistics. – 2021. – Vol. 37(4). – P. 447–462.
10. Andrews D. F. Robust estimates of location: survey and advances / D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, J. W. Tukey. Princeton, NJ: Princeton University Press, 1972. – 372 p.
11. Balakrishnan N. Parameters, order statistics, outliers and robustness / N. Balakrishnan // Revista Matematica Complutence. – 2007. – Vol. 20(1). – P. 7–107. DOI: 10.5209/rev_rema.2007.v20.n1.16528
12. Ramsey P. H. Optimal trimming and outlier elimination / P. H. Ramsey, P. P. Ramsey // Journal of Modern Applied Statistical Methods. – 2007. – Vol. 6. – P. 355–360. DOI: 10.22237/jmasm/1193889660
13. Reed III J. F. Robustness estimators of locations: a classification of linear and adaptive estimators / J. F. Reed III, D. B. Stark // Journal of Applied Statistics. – 1994. – Vol. 21(3). – P. 95–124. DOI: 10.1080/757583650
14. Reed III J. F. Contributions to adaptive estimators / J. F. Reed III, D. B. Stark // Journal of Applied Statistics. – 1998. – Vol. 25(5). – P. 651–669. DOI: 10/1080/02664769822882
15. Hogg R. V. A review of some adaptive statistical techniques / R. V. Hogg, R. V. Lenth // Communications in Statistics – Theory and Methods. – 1984. – Vol. 13. – P. 1551–1579. DOI: 10.1080/03610928408828779
16. Prescott P. Selection of trimming proportions for robust adaptive trimmed mean / P. Prescott // Journal of the American Statistical Association. – 1978. – No. 73(361). – P. 133–140. DOI: 10.2307/2286534
17. Zylstra R. R. Normality tests for small sample sizes / R. R. Zylstra // Quality Engineering. – 1994. – Vol. 7(1). – P. 45–58. DOI: 10.1080/08982119408918766
18. Shapiro S. S. An analysis of variance test for normality (complete samples) / S. S. Shapiro, M. B. Wilk // Biometrika. – 1965. – Vol. 52(3/4). – P. 591–611. DOI: 10.2307/2333709
19. Sarkadi K.Testing for normality / K. Sarkadi // Banach Center Publications. – 1980. – Vol. 6. – P. 281–287. DOI: 10.4064/-6-1-281-287
20. Hersgaard D. Distributions of asymmetric trimmed means / D. Hersgaard // Communications in Statistics: Simulation and Computation. – 1979. – Vol. 8(4). – P. 359–367. DOI: 10.1080/03610917908812125

DOI 10.15588/1607-3274-2022-3-5