УДК 519.25 DOI: 10.15222/TKEA2016.4-5.42

В. С. ПОПУКАЙЛО

Республика Молдова, г. Тирасполь, Приднестровский государственный университет им. Т. Г. Шевченко
E-mail: vsp.science@gmail.com

ОБНАРУЖЕНИЕ АНОМАЛЬНЫХ ИЗМЕРЕНИЙ ПРИ ОБРАБОТКЕ ДАННЫХ МАЛОГО ОБЪЕМА

Рассмотрена мощность критериев обнаружения аномальных измерений в зависимости от объема малой выборки. Исследованы и наглядно проиллюстрированы возможности критериев Граббса, Диксона, Титьена–Мура, Ирвина, Шовене, Львовского и Романовского при объеме исследуемых данных от 5 до 20 измерений. Сделаны выводы о возможности применения каждого из критериев для обнаружения аномальных измерений при обработке данных малого объема.

Ключевые слова: малая выборка, грубые ошибки, аномальные измерения, критерии обнаружения выбросов.

Одним из этапов предварительной обработки сигналов является устранение аномальных измерений, которые даже при небольшой частоте их появления могут внести большие погрешности в результаты восстановления сообщений или в оценки их статистических характеристик. В этом контексте аномальными называют значения, резко выделяющиеся по величине и статистическим свойствам из основной группы [1]. Причины возникновения таких аномальных значений, называемых также грубыми ошибками, могут заключаться как в сбое или отказе определенного оборудования, так и в кратковременном повышении допустимого уровня шумов из-за какого-либо внешнего воздействия.

Существует большое количество критериев, предназначенных для решения задачи отсева аномальных измерений из полученной экспериментальным путем информации [2, 3]. Эффективность этих методов во многом зависит от объема исследуемой выборки. На практике часто встречаются ситуации, когда исследуемый входной процесс представляется ограниченным объемом данных, и поэтому нахождение оптимального метода обнаружения аномальных измерений в условиях малого количества исходной информации является актуальной задачей.

В [4] были исследованы возможности 18 критериев обнаружения грубых ошибок и выбраны критерии, наиболее достоверно их обнаруживающие при объеме исходных данных в 10 элементов. Однако, поскольку малыми считаются выборки, содержащие менее 20 элементов [5], целью настоящей работы была проверка полученных результатов на выборках различного объема n — от 5 до 20 элементов.

Для решения поставленной задачи статистический эксперимент проводился по следующему алгоритму.

1. Для каждого из исследуемых значений n с помощью генератора случайных чисел были получены массивы данных, содержащие несколько тысяч выборок с нормальным законом распределения и заданными характеристиками: средними величинами и дисперсиями. При этом среднее значение было принято равным 10 для всех выборок, а величина дисперсии варьировалась от 0,25 до 25.

2. Каждая выборка упорядочивалась по возрастанию, после чего на место максимального элемента добавлялось значение из диапазона $[1\sigma; 5\sigma]$, где σ — известное генеральное среднеквадратическое отклонение (**СКО**), т. е. показатель рассеивания случайной величины.

3. Выборки проверялись на наличие аномальных измерений каждым из 8 исследуемых критериев.

4. Полученные результаты вносились в общую таблицу, далее подсчитывалась доля обнаруженных аномальных измерений для каждого из критериев при выбранном значении σ.

5. Производились анализ и интерпретация полученных результатов.

Из теории математической статистики известно, что значения нормально распределенной случайной величины с вероятностью более 99% лежат в интервале $(\bar{x} - 3\sigma ; \bar{x}+3\sigma)$, где $\bar{x}$ — истинная величина среднего арифметического. При этом в интервале $(\bar{x}-2\sigma; \bar{x}+2\sigma)$ находится более 95% значений, а в интервале $(\bar{x}-1\sigma; \bar{x}+1\sigma)$ — более 68%. Очевидно, значения выше $\bar{x}+3\sigma$ с большой вероятностью окажутся аномальными измерениями, в то время как значения менее $\bar{x}+2\sigma$ таковыми являться не будут.

Критерии обнаружения грубых ошибок, как правило, делятся на две группы:

— методы, используемые при известном генеральном СКО;

— методы, используемые при неизвестном генеральном СКО.

Первая группа методов применяется в тех случаях, когда исследователь обладает сведениями как о функции полезной составляющей сигналов, так и о параметрах распределения шумовой составляющей, что на практике встречается довольно редко. Выборочная же оценка таких параметров по малому числу измерений приводит к высокой погрешности [6]. Так, в качестве базового критерия в ГОСТ Р ИСО 5725-2—2002 «Точность (правильность и прецизионность) методов и результатов измерений» рекомендуется использовать критерий Граббса, значение которого при известном генеральном СКО и будем считать оптимальным.

Рассмотрим критерии, которые подлежали анализу [2, с. 544—553].

1. Критерий Граббса при неизвестном генеральном СКО. Данный критерий основан на оценивании выборочных отклонений — среднего арифметического $\overline{x}$ и среднеквадратического s — и рассчитывается по формуле

$$\tau = \frac{x_n - \overline{x}}{s}.$$

2. Критерий Диксона используется для быстрого выявления аномальных измерений в выборках небольшого объема по отношению размаха и подразмахов. При этом статистиками являются следующие:

— для проверки одного сомнительного наблюдения

$$r_{10} = \frac{x_n - x_{n-1}}{x_n - x_1};$$

— для проверки одного сомнительного наблюдения независимо от противоположного крайнего наблюдения

$$r_{11} = \frac{x_n - x_{n-1}}{x_n - x_2};$$

— для проверки одного сомнительного наблюдения независимо от следующего по величине

$$r_{20} = \frac{x_n - x_{n-2}}{x_n - x_1};$$

— для проверки одного сомнительного наблюдения независимо от следующего по величине и крайнего противоположного

$$r_{21} = \frac{x_n - x_{n-2}}{x_n - x_2}.$$

3. Критерий Титьена—Мура является обобщением критерия Граббса в случае выявления нескольких выбросов в выборке. Для выделения k наибольших выбросов используется статистика

$$L_k = \frac{\sum_{i=1}^{n-k}(x_i - \overline{x}_k)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

где $\overline{x}_k = \frac{1}{n-k}\sum_{i=1}^{n-k} x_i$.

4. Критерий Ирвина. Использование данного критерия также требует выборочной оценки СКО. Значение критерия вычисляется по формуле

$$\tau = \frac{x_n - x_{n-1}}{s}.$$

5. Критерий Шовене. Элемент выборки x_i объема n является аномальным измерением, если вероятность его отклонения от среднего значения не более $1/(12n)$. Значение критерия вычисляется по формуле

$$K = \frac{x_n - \overline{x}}{s}.$$

6. Критерий Львовского используется для нахождения аномальных измерений при малом их числе, при этом в метод максимального относительного отклонения вводится уточняющий коэффициент, зависящий от объема выборки. Значение критерия вычисляется по формуле [7, с. 24]

$$\tau = \frac{x_n - \overline{x}}{s \cdot \sqrt{\frac{n-1}{n}}}.$$

7. Критерий Романовского. Гипотеза о наличии аномальных измерений в подозрительных результатах подтверждается, если выполняется неравенство

$$|x_i - \overline{x}| \geq t_p s,$$

где t_p — квантиль распределения Стьюдента при заданной доверительной вероятности с числом степеней свободы $k = n - k_n$ (k_n — число подозрительных результатов наблюдений).

Точечные оценки распределения среднего значения и СКО вычисляются без учета подозрительных результатов наблюдений.

В [4] было показано, что в определении грубых ошибок в выборках малого объема лучше всего зарекомендовали себя критерии Диксона, Ирвина, Львовского и Титьена—Мура. Проиллюстрируем это утверждение данными, полученными в ходе статистического эксперимента.

На графике **рис. 1** по оси абсцисс располагаются значения величины i из выражения $\overline{x}+i\sigma$, а по оси ординат — доля обнаруженных аномальных измерений при $n = 10$. Здесь видно, что критерии Диксона, Ирвина и Львовского дают практически одинаковые результаты (с точностью до 1%), и все они могут быть рекомендованы к использованию при данном объеме выборки. Критерий Титьена—Мура обладает пониженной мощностью обнаружения аномальных измерений, однако это компенсируется малым количеством ошибок I рода (так называемая ложная тревога). Остальные крите-

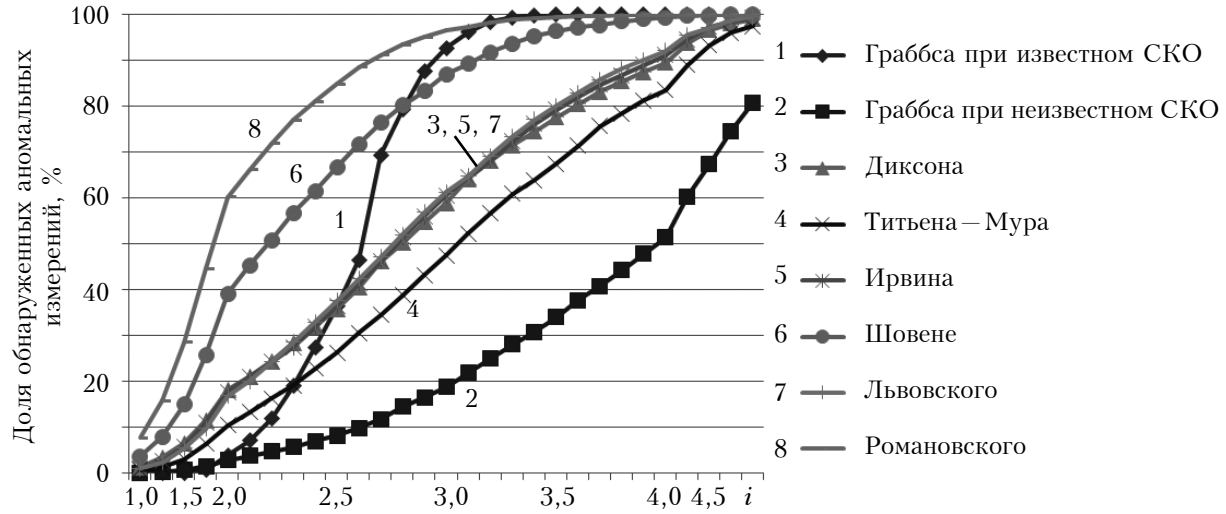

Рис. 1. Мощность критериев обнаружения аномальных измерений при n = 10

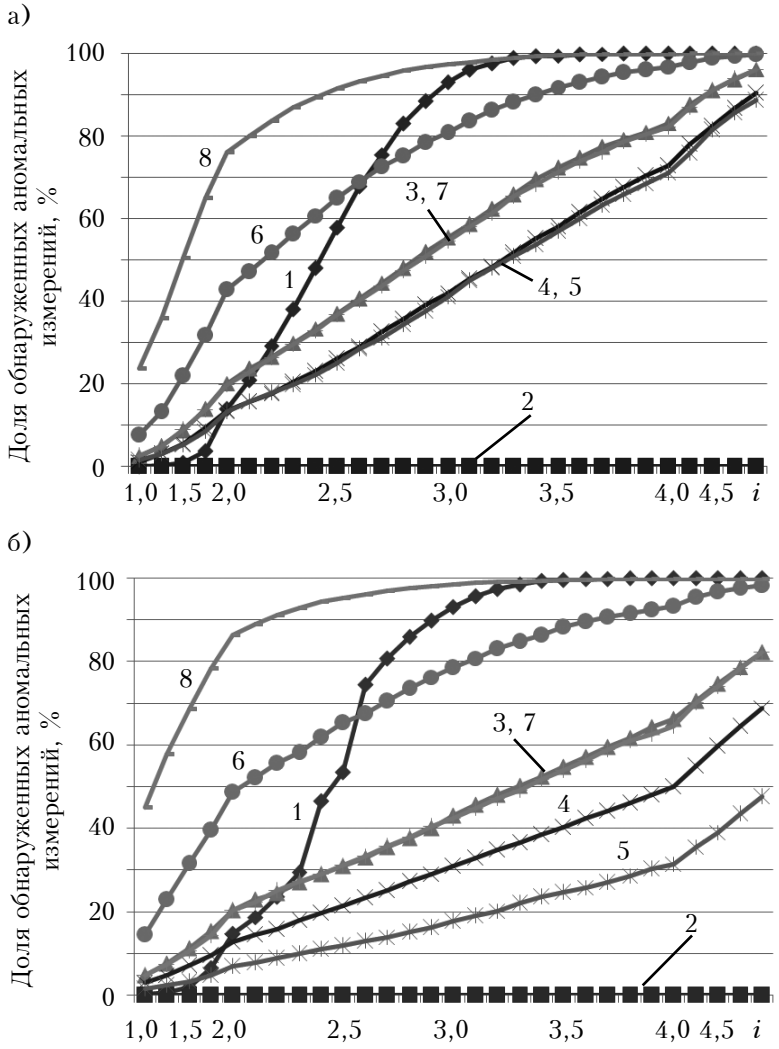

Рис. 2. Мощность критериев обнаружения аномальных измерений при n = 7 (*а*) и n = 5 (*б*) (обозначения те же, что и на рис. 1)

рии использовать не рекомендуется из-за низкой мощности (критерий Граббса при неизвестном генеральном СКО) либо большим количеством ошибок I рода (так, критерии Шовене и Романовского считают грубой ошибкой значение $\bar{x}+2\sigma$ в 39% и 60% случаев соответственно, что не может быть признано удовлетворительным). Таким образом, результаты предыдущего исследования [4] подтверждаются.

Рассмотрим, как изменяется мощность критериев обнаружения аномальных измерений при уменьшении исследуемого объема выборки. На **рис. 2**, где изображены графики, полученные для n = 7 и n = 5, видно, что при таких объемах выборки обнаружить аномальные измерения с помощью критерия Граббса при неизвестном генеральном СКО невозможно. При использовании критериев Шовене и Романовского количество ошибок I рода увеличивается с уменьшением объема выборки (при n = 7 критерий Шовене считает значение $\bar{x}+2\sigma$ аномальным результатом в 42,84% случаев, а критерий Романовского — в 76,08%; при n = 5 — соответственно, в 48,76% и в 86,52% случаев), что указывает на зависимость достоверности получаемых результатов от объема выборки. Также из графиков видна прямая зависимость от объема исследуемой выборки мощности критерия Ирвина. Оптимальными для обнаружения аномальных измерений в выборках объемом менее 10 значений можно

ISSN 2225-5818

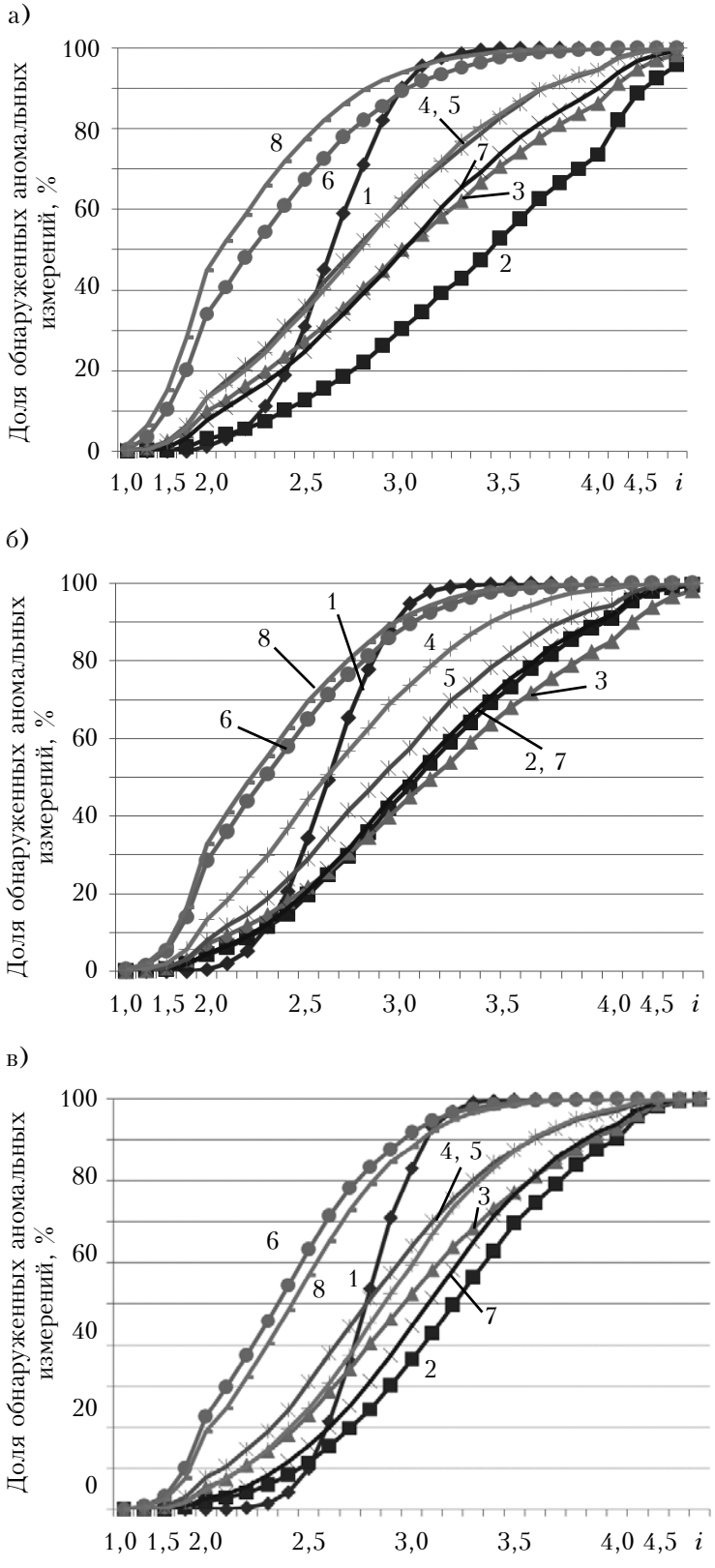

Рис. 3. Мощность критериев обнаружения аномальных измерений при n = 13 (*а*), n = 16 (*б*) и n = 20 (*в*)

1 — Граббса при известном СКО; *2* — Граббса при неизвестном СКО; *3* — Диксона; *4* — Титьена—Мура; *5* — Ирвина; *6* — Шовене; *7* — Львовского; *8* — Романовского

считать критерии Диксона и Львовского, которые дают статистически неразличимые результаты.

Теперь рассмотрим, как изменяется мощность критериев обнаружения аномальных результатов при увеличении исследуемого объема выборки. Как видно из графиков на **рис. 3**, полученных для n = 13, 16, 20, наилучшие результаты по обнаружению аномальных измерений в выборках при $n > 10$ показывают критерии Ирвина и Львовского. Следует отметить, что при увеличении объемов исследуемых данных возрастает точность критерия Граббса при неизвестном СКО и значительно уменьшается количество ошибок I рода у критериев Шовене и Романовского.

Выводы

Таким образом, проведенное исследование позволяет сделать следующие выводы.

1. Оптимальным для обнаружения аномальных измерений является критерий Львовского, который дает приемлемые результаты даже в случае малого количества исходной информации, т. е. вне зависимости от объема данных (это связано с использованием поправочного коэффициента, зависящего от величины выборки), и при этом его точность возрастает с увеличением количества исследуемой информации.

2. Критерий Диксона можно рекомендовать при обработке выборок с объемом не более 10, поскольку он прост в расчетах, а его результаты статистически неотличимы от критерия Львовского. При этом следует обратить внимание, что для объемов выборок больше 10 исследовалась мощность статистик Диксона для проверки одного сомнительного наблюдения независимо от следующего по величине, как это рекомендуется в [2, с. 550]. Направлением дальнейших исследований может быть проверка данных рекомендаций и изучение мощности различных критериев Диксона в зависимости от объема выборки.

3. Мощность критерия Ирвина прямо зависит от объема выборки. Это связано с использованием величины выборочной дисперсии, которая при малом объеме выборки не отражает достоверно величину дисперсии генеральной совокупности. Однако критерий Ирвина можно применять для выборок объемом не менее 10.

4. Критерий Титьена—Мура может быть рекомендован для обнаружения аномальных значений в малых выборках (больше 5), поскольку он довольно хорошо распознает ошибки в значе-

ниях более $\bar{x}+4\sigma$ и обладает наименьшим количеством ошибок I рода.

5. Критерий Граббса при неизвестном генеральном среднеквадратическом отклонении можно применять при объемах выборки не менее 15.

6. Критерии Шовене и Романовского не рекомендуется применять при объеме исследуемых данных менее 20.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Марчук В. И., Токарева С.В. Способы обнаружения аномальных значений при анализе нестационарных случайных процессов. — Шахты: ЮРГУЭС, 2009.

2. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. — М.: ФИЗМАТЛИТ, 2012.

3. Charu C. Aggarwal. Outlier Analysis. — NY: Springer, 2013.

4. Попукайло В. С. Исследование критериев грубых ошибок применительно к выборкам малого объема // Радіоелектронні і комп'ютерні системи.— 2015. — № 3(73).— С. 39—44.

5. Столяренко Ю. А. Контроль кристаллов интегральных схем на основе статистического моделирования методом точечных распределений / Автореф. дис. …канд. техн. наук.— М.: ГУП НПЦ «СПУРТ», 2006.

6. Громыко. Г. Л. Теория статистики.— М.: ИНФРА-М, 2005.

7. Львовский Е. Н. Статистические методы построения эмпирических формул: учеб. пособие для вузов.— М.: Высшая школа, 1988.

Дата поступления рукописи в редакцию 27.04 2016 г.

В.С. ПОПУКАЙЛО

Республіка Молдова, м. Тирасполь, Придністровський державний університет ім. Т. Г. Шевченка
E-mail: vsp.science@gmail.com

ВИЯВЛЕННЯ АНОМАЛЬНИХ ВИМІРЮВАНЬ ПРИ ОБРОБЦІ ДАНИХ МАЛОГО ОБСЯГУ

Розглянуто потужність критеріїв виявлення аномальних вимірювань в залежності від обсягу малої вибірки. Досліджено та наочно проілюстровано можливості критеріїв Граббса, Діксона, Тітьєна—Мура, Ірвіна, Шовене, Львівського та Романовського при обсягах досліджуваних даних від 5 до 20 вимірювань. Зроблено висновки про можливість застосування кожного з критеріїв для виявлення аномальних вимірювань при обробці даних малого обсягу.

Ключові слова: мала вибірка, грубі помилки, аномальні вимірювання, критерії виявлення викидів.

DOI: 10.15222/TKEA2016.4-5.42
UDC 519.25

V. S. POPUKAYLO

Republic of Moldova, Tiraspol, Taras Shevchenko Transnistria State University
E-mail: vsp.science@gmail.com

DETECTION OF OUTLIERS IN PROCESSING OF SMALL SIZE DATA

This article describes the criteria for detection of outliers power depending on a small size sample. Removing outliers is one of the stages of signals pre-processing. Statistical experiment, in which using a random number generator were received arrays of data, containing several thousand samples with normal distribution, with the given mean averages and standard deviation for each n-value, was conducted to solve this problem. Thus, we researched and vividly illustrated the possibility of Grubbs, Dixon, Tietjen—Moore, Irving, Chauvenet, Lvovsky and Romanovsky criteria at studied data sizes from 5 to 20 meterages. Conclusions about the applicability of each criterion for the outliersdetection in processing of small size data were made. Lvovsky criterion was recognized the optimal criterion. Dixon's criterion was recommended for $n \le 10$. Irwin's criterion was recommended when $n \ge 10$. Tietjen—Moore'scriterion can be recommended for the detection of outliers in small samples for $n > 5$, since it recognizes errors well in the values of a $\bar{x}+4\sigma$ and has the least amount of I type mistakes. Grubb's with an unknown standard deviation may be used in samples for $n \ge 15$. Chauvenet and Romanovsky criteria cannot be recommended for the detection of outliers in small size data.

Keywords: small size data, outlier detection criteria, anomalous meterages, outlier analysis.

REFERENCES

1. Marchuk V. I., Tokareva S. V. *Sposoby obnaruzheniya anomal`nykh znachenii pri analize nestatsionarnykh sluchainykh protsessov* [Methods for detection of outliers in the analysis of non-stationary random processes]. Shakhty, SRSUES, 2009. (Rus)

2. Kobzar` A. I. P*rikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotneykov* [Applied mathematical statistics.For engineers and scientists]. Moscow, FIZMATLIT, 2012. (Rus)

3. Charu C. Aggarwal. *Outlier Analysis*. NY, Springer, 2013, 446 p.

4. PopukailoV.S. [The outlier criteria research in relation to small volume samples]. *Radioelektronni i komp'yuterni sistemi*, 2015, 3(73), pp. 39-44. (Rus)

5. Stolyarenko Yu.A. [*The crystals control of integrated schemes on the basis of statistical modeling by pointed distributions method*]. Extended abstract of dissertation… Ph.D. in Engineering Science. Moscow, SUE NPTs "SPURT", 2006. (Rus)

6. Gromyko G. L. *Teoriya statistiki* [Theory of Statistics]. Moscow, INFRA-M, 2011, 476 p. (Rus)

7. L`vovskii E. N. *Statisticheskie metody postroeniya empiricheskikh formul: ucheb. posobie dlya vuzov* [Statistical methods for constructing empirical formulas: a textbook for high schools]. Moscow, Vysshaya shkola, 1988.