УДК 519.233.22

К. Андреа, Г.Л. Шевляков

ОБНАРУЖЕНИЕ ВЫБРОСОВ С ПОМОЩЬЮ БОКСПЛОТОВ, ОСНОВАННЫХ НА НОВЫХ ВЫСОКОЭФФЕКТИВНЫХ РОБАСТНЫХ ОЦЕНКАХ МАСШТАБА

K. Andrea, G.L. Shevlyakov

OUTLIER DETECTION WITH BOXPLOTS BASED ON NEW HIGHLY EFFICIENT ROBUST ESTIMATES OF SCALE

Разработаны улучшенные робастные версии классического боксплота Тьюки на основе быстрых высокоэффективных робастных оценок масштаба. Методом Монте–Карло проведен сравнительный анализ классического и предложенных боксплотов при оценке качества обнаружения выбросов для двух больших групп моделей распределения: симметричных и асимметричных. Полученные результаты подтверждают преимущество новых предложенных методов при засоренных данных.

БОКСПЛОТ. ВИЗУАЛИЗАЦИЯ ДАННЫХ. РОБАСТНОСТЬ. ЭФФЕКТИВНОСТЬ.

Robust versions of Tukey's boxplot based on the low-complexity highly efficient robust estimates of scale are presented. Boxplots' performance is tested experimentally using Monte–Carlo method on two big groups of data distributions: symmetric and skewed models of distribution. The proposed versions of boxplots prove to be a better choice when dealing with contaminated data.

BOXPLOT. DATA VISUALIZATION. ROBUSTNESS. EFFICIENCY.

Основной вклад в развитие робастной статистики внесли пионерские работы Тьюки [1], Хьюбера [2] и Хампеля [3]. Робастность в статистике означает в широком смысле устойчивость статистических оценок в условиях отклонений от предполагаемой модели распределения данных. Методы робастной статистики основаны на асимптотическом подходе, однако их использование на малых по объему выборках данных также дает хорошие результаты, что обеспечивает возможность их широкого применения при решении практических задач.

В данной статье представлены новые результаты в области робастных методов обработки данных. Мы предлагаем робастные модификации классического боксплота. Классический одномерный боксплот Тьюки наглядно обобщает параметры распределения данных, позволяя быстро сравнивать выборки. Боксплот содержит информацию о параметрах положения и масштаба, а также о весе «хвостов» распределения, асимметрии распределения данных и наличии выбросов. В нашей статье основное внимание уделено повышению эффективности выявления выбросов и наглядности представления «хвостов» распределения.

Одномерный боксплот [4] определяется пятью параметрами: минимумом и максимумом выборки, нижним LQ и верхним UQ квартилями, интерквартильной широтой IQR (оценка масштаба) и выборочной медианой. Экстремумы («усы» боксплота) вычисляются следующим образом:

$$x_L = \max\{x_{(1)}, \mathrm{LQ} - \tfrac{3}{2}\,\mathrm{IQR}\},$$
$$x_U = \min\{x_{(n)}, \mathrm{UQ} + \mathrm{IQR}\}. \qquad (1)$$

Все элементы выборки, расположенные за пределами экстремумов, принято рассматривать как выбросы и отображать графически вместе с соответствующим боксплотом.

Графически внутренняя часть боксплота представлена как коробчатая конструкция с границами, равными нижнему и верхнему

квартилям, содержащая 50 % центральных порядковых статистик выборки. Выборочная медиана обозначается линией внутри коробки и делит интерквартильную область на две части. Прямые, исходящие из противоположных сторон коробки, обозначают «хвосты» распределения выборки, их длина определяется экстремумами по формуле (1).

Для улучшения стандартной модели были предложены различные модифицированные варианты боксплота. Среди таких модификаций можно выделить боксплот, графическое представление которого содержит информацию о доверительном интервале оценки параметра положения – медиане [5], что обеспечивает возможность сравнения качества оценки медианы нескольких выборок. В [6] предложена иная модификация боксплота, позволяющая включить в графическое представление информацию о плотности распределения данных. На базе боксплота, в графической структуре которого отображена информация о плотности распределения, разработаны другие виды боксплотов [7–9].

Различные модификации классического боксплота, как правило, расширяют набор доступных для визуального восприятия параметров, однако это требует дополнительных вычислительных ресурсов. В данной статье мы предлагаем простую по вычислительной сложности модификацию параметров классического боксплота через высокоэффективные робастные оценки масштаба.

Высокоэффективные робастные оценки масштаба в предложенных модификациях классического боксплота

На практике классический боксплот применяется для выявления выбросов, присутствующих в выборке. Мы можем повысить качество обнаружения аномальных данных посредством повышения качества определения экстремумов. В классическом варианте боксплота экстремумы выборки определяются через интерквартильную широту IQR, но эта оценка не обладает высокой эффективностью и робастностью.

Методы робастной статистики предлагают более устойчивые статистические оценки для случаев, когда в выборке данных присутствуют выбросы, в частности, робастная, высокоэффективная, но вычислительно сложная Q_n-оценка масштаба [11]. В [12] предложена «быстрая» робастная высокоэффективная FQ_n-оценка масштаба, основанная на аппроксимации функции влияния Q_n-оценки. Показано, что максимальная эффективность предложенной FQ_n-оценки достигает 96 %, а минимальное возможное ее значение не опускается ниже уровня 81 % на нормальном распределении, при этом их пороговая точка (breakdown point [10]) достигает максимального значения 50 %. Используемая далее FQ_n-оценка масштаба имеет максимально высокую робастность, так же как медианное абсолютное отклонение $\mathrm{MAD} = \mathrm{med}|x\text{-med } x|$, но значительно более высокую эффективность (эффективность MAD на нормальном распределении равна 37 %).

Обобщим уравнения вычисления экстремумов боксплота следующим образом:

$$x_L = \{x_{(1)}, \mathrm{LQ} - kS_n\}, x_U = \{x_{(n)}, \mathrm{UQ} + kS_n\}, \quad (2)$$

где k – пороговое значение; S_n – робастная высокоэффективная оценка масштаба.

Предложенные нами версии боксплотов отличаются от классического подстановкой робастных оценок масштаба MAD_n и FQ_n в формулу (2). Результатом применения такого приема является изменение оценки экстремумов, что влияет на обнаружение выбросов. Проведем сравнительный анализ качества обнаружения выбросов классического и модифицированных боксплотов.

Сравнительный анализ качества обнаружения выбросов в модели Тьюки–Хьюбера

В статистике выбросом является наблюдаемое значение, заметно отличающееся от остальных элементов выборки [13]. Выбросы могут возникать случайным образом в любой модели распределения. Во многих случаях выбросы либо являются ошибками измерения, либо указывают на наличие тяжелых «хвостов» в модели распределения. В первом случае наличие таких аномальных результатов наблюдений – итог неправильной работы системы. Причиной появления

Таблица 1

Значения гармонического среднего H для модели засорения типа «масштаб» (3), μ = 0, s = 3

ε = 0,1	20	50	100	1000	10 000
БП Тьюки	0,64	**0,72**	0,72	0,72	0,72
MAD-БП	**0,67**	**0,72**	**0,73**	**0,73**	**0,73**
FQ-БП	0,66	**0,72**	0,72	0,72	**0,73**
Граббс	0,17	0,29	0,30	0,30	0,30

выбросов нередко является смешение двух моделей распределений. Последняя ситуация формализуется моделью засорения Тьюки–Хьюбера [1]:

$$f(x) = (1 - \varepsilon)N(x; 0, 1) + \varepsilon N(x; \mu, s), \quad (3)$$

где $0 \le \varepsilon < 1$ – вероятность появления выбросов; μ – параметр положения; s – параметр масштаба для распределения «плохих» данных.

В классическом методе определения выбросов, известном как тест Граббса [13], выбросом объявляется наблюдение x, для которого справедливо неравенство $|x - \bar{x}|/S > k_\alpha$, где $\bar{x}$ – выборочное среднее; S – среднеквадратическое отклонение; k_α – пороговое значение, определяемое вероятностью ложной тревоги (вероятность ошибки первого рода) для нормального распределения.

В данной статье мы обозначаем наблюдение x как выброс в случае $x < x_L$ и $x > x_U$, где x_L и x_U – соответственно нижняя и верхняя границы боксплота. Пороговое значение k_α выбирается экспериментально при условии, что вероятность ложной тревоги α = 0,1.

Для сравнительного анализа эффективности тестов воспользуемся мерами чувствительности (SE) и специфичности (SP). Чувствительность представляет собой мощность критерия, а специфичность – вероятность отсутствия ложной тревоги. Оба эти показателя комбинируем в один, вычисляя их гармоническое среднее H = 2 SE SP / /(SE + SP). Выбор гармонического среднего в качестве оценки эффективности обнаружения выбросов обусловлен тем свойством, что с его использованием можно успешно оценивать разные критерии выявления выбросов, отличающиеся по вероятности ложной тревоги. В нашей статье вероятности ложной тревоги классического боксплота и модифицированных боксплотов равны α = 0,06 и α = 0,1 соответственно.

Результаты экспериментов представлены в табл. 1, 2, где лучшие показатели выделены жирным шрифтом.

По результатам, указанным в табл. 1, 2, в обеих моделях засорения («сдвиг» и «масштаб») боксплоты не отличаются значительно по эффективности обнаружения выбросов, за исключением теста Граббса, демонстрирующего очень низкие показатели эффективности. Такое поведение теста Граббса можно объяснить неробастностью используемых в нем оценок сдвига и масштаба.

Таблица 2

Значения гармонического среднего H для модели засорения типа «сдвиг» (3), μ = 3, s = 1

ε = 0,1	20	50	100	1000	10 000
БП Тьюки	**0,75**	0,79	0,80	0,80	0,80
MAD-БП	0,73	**0,80**	0,80	0,80	0,80
FQ-БП	0,73	0,79	**0,81**	**0,81**	**0,81**
Граббс	0,32	0,39	0,40	0,39	0,39

Таблица 3

Значения гармонического среднего H для модели засорения типа «сдвиг» с различными значениями ε

ε	0,05	0,10	0,20	0,30	0,4	0,5
БП Тьюки	0,63	0,62	0,59	0,55	0,51	0,43
MAD-БП	0,65	0,65	0,60	**0,56**	**0,52**	**0,44**
FQ-БП	**0,67**	**0,67**	**0,61**	**0,56**	0,50	0,40
Граббс	0,65	0,56	0,41	0,31	0,25	0,21

Робастные версии боксплотов немного превышают по эффективности классический боксплот.

В табл. 3 представлены результаты исследования зависимости эффективности обнаружения выбросов от параметра засорения ε. Среди малых и умеренных уровней засорения FQ-боксплот имеет самые лучшие показатели эффективности. В сильно засоренных выборках MAD-боксплот отличается более высоким уровнем эффективности по сравнению с остальными. Такое поведение MAD-боксплота обусловлено принадлежностью оценки масштаба MAD к классу минимаксных робастных оценок параметра масштаба [10].

Сравнительный анализ качества обнаружения выбросов в асимметричных моделях распределения данных

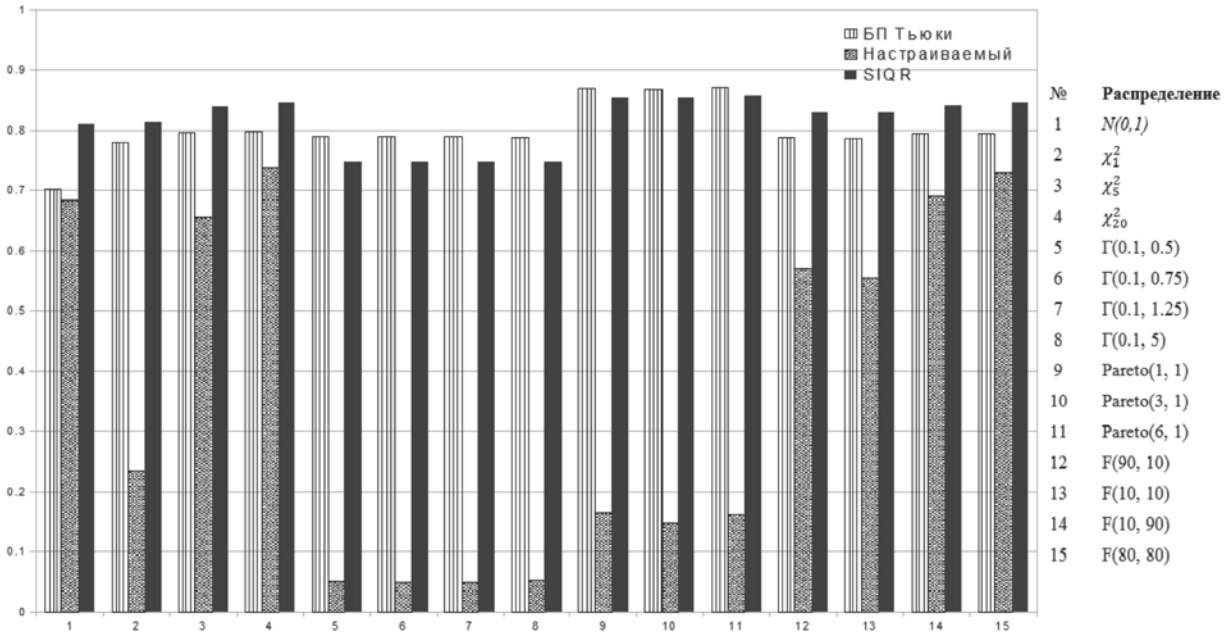
Несмотря на способность классического боксплота выявлять асимметрию распределения данных, подход к определению экстремумов боксплота, наоборот, предполагает наличие симметрии. Значения экстремумов зависят только от значения интерквартильной широты. Минимальное и максимальное значения боксплота симметричны относительно параметра сдвига в формуле (1). В условиях нормального закона распределения такая симметрия естественна, причем выбор порогового значения $k = 1{,}5$ обусловлен правилом трех сигм. Последнее означает, что вероятность определения выбросом регулярного наблюдения составляет 0,003. Такое правило справедливо для нормально распределенных данных, но в общем случае оно хорошо работает и для остальных симметричных моделей распределения.

Разнообразие асимметричных моделей не позволяет обобщить вычисления экстремумов (параметров «усов») боксплота. Для правильного вычисления экстремумов важно точно определить квантили изучаемой модели, чтобы с заданной вероятностью не пропускать регулярные наблюдения. Применение классического боксплота для асимметрично распределенных данных приводит к сильным искажениям в визуализации (неверная наглядная информация) и повышает ошибку обнаружения выбросов (когда регулярные наблюдения попадают в категорию выбросов).

Предложенные модификации классического боксплота для работы с асимметричными моделями основаны на оценке асимметрии данных выборки. Дополнительное требование, предъявляемое к асимметричным боксплотам – классическое поведение при наличии симметрии изучаемой выборки. Проведем сравнительный анализ следующих методов по эффективности обнаружения выбросов:

- классический боксплот Тьюки;
- SIQR-боксплот [14], вычисления экстремумов которого определяются по верхней и нижней половинам интерквартильного размаха $\mathrm{SIQR}_U = \mathrm{Q}_3 - \mathrm{Q}_2$, $\mathrm{SIQR}_L = \mathrm{Q}_2 - \mathrm{Q}_1$ соответственно. Экстремумы вычисляются следующим образом: $x_L = \max\{x_{(1)}, Q_1 - 3\mathrm{SIQR}_L\}$, $x_U = \min\{x_{(1)}, Q_3 + 3\mathrm{SIQR}_U\}$;
- настраиваемый боксплот (adjusted boxplot) [15], алгоритм которого реализован в статистической среде **R**.

Монте–Карло эксперимент проведен тысячу раз на выборке объемом 1000 с засорением типа «сдвиг» при ε = 0,05. Моделированы данные пяти групп распределений: стандартное нормальное распределение, χ^2, **Г**, Pareto и F-распределение.



Результаты эффективности выявления выбросов
для различных асимметричных моделей распределения и ε = 0,05.
Порядковый номер на оси абсцисс соответствует распределению в таблице справа

На рисунке преобладают высокие оценки H для SIQR-бокспота и бокспота Тьюки. Настраиваемый бокспот выдает самые худшие показатели по эффективности выявления выбросов, что можно объяснить его настройкой на низкую вероятность ложной тревоги. Средние значения H по всем моделям распределения равны 0,81 для SIQR-бокспота, 0,8 для классического бокспота и 0,37 для настраиваемого бокспота. Таким образом, классический бокспот Тьюки можно рекомендовать использовать и при асимметричном засорении.

Исходя из результатов проведенного анализа, можно рекомендовать применение MAD-бокспота для сильно засоренных выборок, в то время как для умеренного и малого засорения выборки лучше применять FQ-бокспот. Пороговые значения, определяющие экстремумы, необходимо настроить в зависимости от желаемого уровня вероятности ложной тревоги. При α = 0,1 мы рекомендуем воспользоваться значениями k_{MAD} = 1,44 и k_{FQ} = 0,97 в случае нормального распределения.

СПИСОК ЛИТЕРАТУРЫ

1. **Tukey, J.W.** A survey of sampling from contaminated distributions [Text] / J.W. Tukey // Contributions to Probability and Statistics. – 1960. – P. 448–485.

2. **Huber, P.J.** Robust estimation of a location parameter [Text] / P.J. Huber // Ann. Math. Statist. – 1964. – № 1. – Vol. 35. – P. 73–101.

3. **Hampel, F.R.** Contributions to the Theory of Robust Estimation [Text] / F.R. Hampel // Ph.D. thesis. – University of California, Berkeley, 1968.

4. **Tukey, J.W.** Exploratory Data Analysis [Text] / J.W. Tukey. –Addison-Wesley–Reading, MA, 1977.

5. **McGill, R.** Variations of box plots [Text] / R. McGill, J.W. Tukey, W.A. Larsen // The American Statistician. – 1978. – № 1. – Vol. 32. – P. 12–16.

6. **Potter, K.** Methods for Presenting Statistical Information: The Box Plot [Text] / K. Potter // Visualization of Large and Unstructured Data Sets, University of Utah. – 2006. – Vol. S-4. – P. 97–106.

7. **Benjamini, Y.** Opening the box of a boxplot [Text] / Y. Benjamini // The American Statistician. – 1988. – № 4. – Vol. 42. – P. 257–262.

8. **Esty, W.W.** The box-percentile plot [Text] /

W.W. Esty, J. Banfield // J. of Statistical Software. – 2003. – Vol. 8. – № 17. – P. 181–184.

9. **Hintze, J.L.** Violin plots: A box plot-density trace synergism [Text] / J.L. Hintze, R.D. Nelson // The American Statistician. – 1998. – Vol. 52. – № 2. – P. 181–184.

10. **Hampel, F.R.** Robust statistics: the approach based on influence functions [Text] /F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel // Wiley series in probability and mathematical statistics. Probability and mathematical statistics. – Wiley&Sons, 2005.

11. **Rousseeuw, P.J.** Alternatives to the median absolute deviation [Text] / P.J. Rousseeuw, C. Croux // J. of the American Statistical Association. – 1993. – Vol. 88. – № 424. – P. 1273–1283.

12. **Shevlyakov, G.L.** On approximation of the Q_n-estimate of scale by fast M-estimates [Text] / G.L. Shevlyakov, P.O. Smirnov // Internat. Conf. on Robust Statistics. – Parma, Italy, 2010.

13. **Grubbs, F.E.** Procedures for detecting outlying observations in samples [Text] / F.E. Grubbs // Technometrics. – 1969. – Vol. 11. – № 1. – P. 1–21.

14. **Kimber, A.C.** Exploratory data analysis for possibly censored data from skewed distributions [Text] / A.C. Kimber // Applied Statistics. – 1990. – Vol. 39. – № 1. – P. 21–30.

15. **Hubert, M.** An adjusted boxplot for skewed distributions [Text] / M. Hubert, E. Vandervieren // Computational Statistics & Data Analysis. – 2008. – Vol. 52. – № 12. – P. 5186–5201.

REFERENCES

1. **Tukey J.W.** A survey of sampling from contaminated distributions / Contributions to Probability and Statistics. – 1960. – P. 448–485.

2. **Huber P.J.** Robust estimation of a location parameter / Ann. Math. Statist. – 1964. – № 1. – Vol. 35. – P. 73–101.

3. **Hampel F.R.** Contributions to the Theory of Robust Estimation: Ph.D. thesis. – University of California, Berkeley, 1968.

4. **Tukey J.W.** Exploratory Data Analysis. – Addison-Wesley – Reading, MA, 1977.

5. **McGill R., Tukey J.W., Larsen W.A.** Variations of box plots / The American Statistician. – 1978. – № 1. – Vol. 32. – P. 12–16.

6. **Potter K.** Methods for Presenting Statistical Information: The Box Plot / Visualization of Large and Unstructured Data Sets, University of Utah. – 2006. – Vol. S-4. – P. 97–106.

7. **Benjamini Y.** Opening the box of a boxplot / The American Statistician. – 1988. – № 4. – Vol. 42. – P. 257–262.

8. **Esty W.W., Banfield J.** The box-percentile plot / Journal of Statistical Software. – 2003. – Vol. 8. – № 17. – P. 181–184.

9. **Hintze J.L., Nelson R.D.** Violin plots: A box plot-density trace synergism / The American Statistician. – 1998. – Vol. 52. – № 2. – P. 181–184.

10. **Hampel F.R, Ronchetti E.M., Rousseeuw P.J., Stahel W.A.** Robust statistics: the approach based on influence functions / Wiley series in probability and mathematical statistics. Probability and mathematical statistics. – Wiley&Sons, 2005.

11. **Rousseeuw P.J., Croux C.** Alternatives to the median absolute deviation. / Journal of the American Statistical Association. – 1993. – Vol. 88. – № 424. – P. 1273–1283.

12. **Shevlyakov G.L., Smirnov P.O.** On approximation of the Q_n-estimate of scale by fast *M*-estimates / Internat. Conf. on Robust Statistics. – Parma, Italy, 2010.

13. **Grubbs F.E.** Procedures for detecting outlying observations in samples / Technometrics. – 1969. – Vol. 11. – № 1. – P. 1–21.

14. **Kimber A.C.** Exploratory data analysis for possibly censored data from skewed distributions / Applied Statistics. – 1990. – Vol. 39. – № 1. – P. 21–30.

15. **Hubert M., Vandervieren E.** An adjusted boxplot for skewed distributions / Computational Statistics & Data Analysis. – 2008. – Vol. 52. – № 12. – P. 5186–5201.

АНДРЕА Клитон – *аспирант кафедры прикладной математики Санкт-Петербургского государственного политехнического университета.*

195251, Россия, Санкт-Петербург, ул. Политехническая, д. 29.

E-mail: kliton.andrea@gmail.com

ANDREA, Kliton *St. Petersburg State Polytechnical University* .

195251, Politechnicheskaya Str. 29, St.-Petersburg, Russia.

E-mail: kliton.andrea@gmail.com

ШЕВЛЯКОВ Георгий Леонидович – *профессор кафедры прикладной математики Санкт-Петербургского государственного политехнического университета, доктор физико-математических наук.*

195251, Россия, Санкт-Петербург, ул. Политехническая, д. 29.
E-mail: Georgy.Shevlyakov@phmf.spbstu.ru

SHEVLYAKOV, Georgy L. *St. Petersburg State Polytechnical University.*
195251, Politechnicheskaya Str. 29, St.-Petersburg, Russia.
E-mail: Georgy.Shevlyakov@phmf.spbstu.ru