 doi:10.20944/preprints202007.0226.v1

Outsiders: The Definition and Critical Analysis

Lev B. Klebanov[a] and Georgy L. Shevlyakov[b]

Abstract

A critical analysis of the classical definition of outsiders is given. Some examples show that this notion is not universal and has at least two drawbacks. Particularly, the set of outsiders may have the probability arbitrary close to one half. On the other hand, the deleting of the set of outsiders may dramatically change the value of the median.

Keywords: outsiders; outliers; instability of the median; gamma-distribution

1 Main Results

As far as we know, the discussed below notion of outsiders had been given by Tukey [1] in the content of a robust estimation of the location parameter. However, many recent publications use outsider notions in a more general situation and, sometimes, instead of the outlier notion. Our opinion is that such use may lead to some essential mistakes. We support this point providing some examples of distributions having rather unexpected properties.

Let us remind the definition of outsiders. Suppose that X is a random variable having probability distribution function $F(x)$. Denote by $Q = (Q_1, Q_2, Q_3)$ the quartiles $Q_1 \leq Q_3$ and the median Q_2 of $F(x)$. And let

[a]Department of Probability and Mathematical Statistics, Charles University, Prague, Czech Republic; e-mail: lev.klebanov@mff.cuni.cz

[b]Department of Applied Mathematics, Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russia

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2020 doi:10.20944/preprints202007.0226.v1

$IQR = Q_3 - Q_1$ be an interquartile range. Lower and upper fences are defined by

$$Lf = Q_1 - 1.5IQR$$

and

$$Uf = Q_3 + 1.5IQR$$

correspondingly. All possible values of X lying outside of the interval $[Lf, Uf]$ are called **outsiders**. Tukey gave no precise definition of **outliers** but said outliers are among outsiders. Sometimes, scientists make no difference between outliers and outsiders (see, for example, [2]). On the level of the intuition, both outliers and outsiders are such values of the random variable X which are far from the "main part" of the values. Of course, this definition has a precise sense. Some critics and variants of possible definitions may be found in [3] and [4].

Example 1.1. *Consider a random variable X taking four values: $-5, -1, 1, 5$ with probabilities*

$$\mathbb{P}\{X = -5\} = \mathbb{P}\{X = 5\} = 1/4 - \varepsilon/2,$$

$$\mathbb{P}\{X = -1\} = \mathbb{P}\{X = 1\} = 1/4 + \varepsilon/2,$$

where $\varepsilon \in (0, 1/4)$ is a fixed number. It is easy to calculate

$$Q = (-1, 0, 1); \quad Lf = -4; \quad Uf = 4.$$

Therefore, the set of outsiders consists of two points -5 and 5 and the probability for X to be an outsider is $1/2 - \varepsilon$.

Example 1.1 shows the probability of outsiders may be arbitrary close to one half. It seems (on the level of intuition) not natural to ignore the event having the probability arbitrary close to $1/2$. Of course, we have a mathematically strong definition of outsiders (not that of outliers) and Example 1.1 does not contradict it. However, it would be interesting to understand why does one needs such definition. In our opinion, the notion of outsiders has a sense of some special problems only.

Example 1.1 provides us a discrete distribution having a high probability set of outsiders. Is it possible to construct an example of a continuous distribution with a similar property? The next example gives an affirmative answer.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2020 doi:10.20944/preprints202007.0226.v1

Example 1.2. *Let $G(x, a)$ be a function of gamma distribution with unite scale and shape parameter $a > 0$. Define $F(x, a) = \big(G(x, a)+1-G(-x, a)\big)/2$ and suppose that X has $F(x, a)$ as its distribution function. It is not difficult to calculate the probabilities of outsiders set. On the Picture 1 is the plot of this probability as a function of a parameter.*

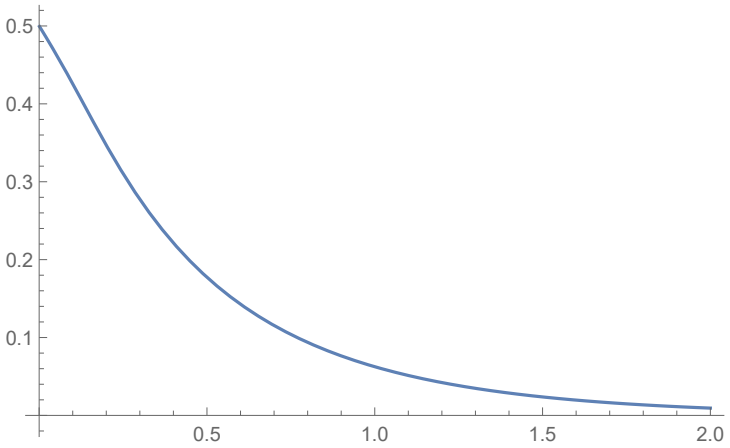

Figure 1: Probability of outsiders as a function of shape parameter a

The next Figure shows the dependence of fences on the shape parameter.

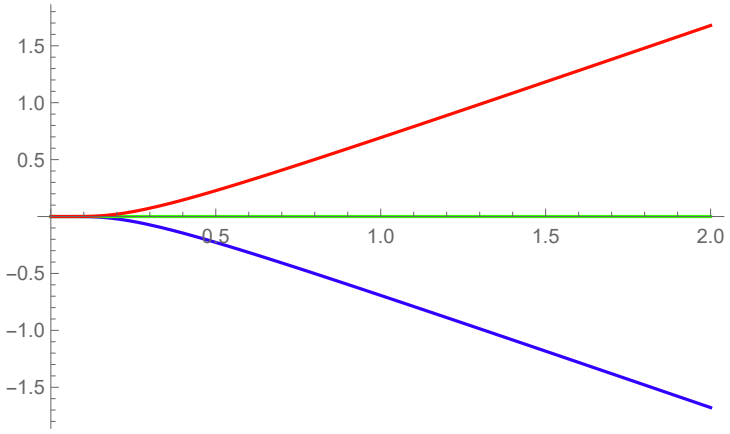

Figure 2: Lower (blue) and upper (red) fences as the functions of the shape parameter a

For the case $a = 0.001$ the probability of outsiders is 0.499306. *We see*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2020 doi:10.20944/preprints202007.0226.v1

that the situation with outsiders is very similar to that of the Example 1.1. However, the distribution here is continuous. Its density is given on the Picture 3.

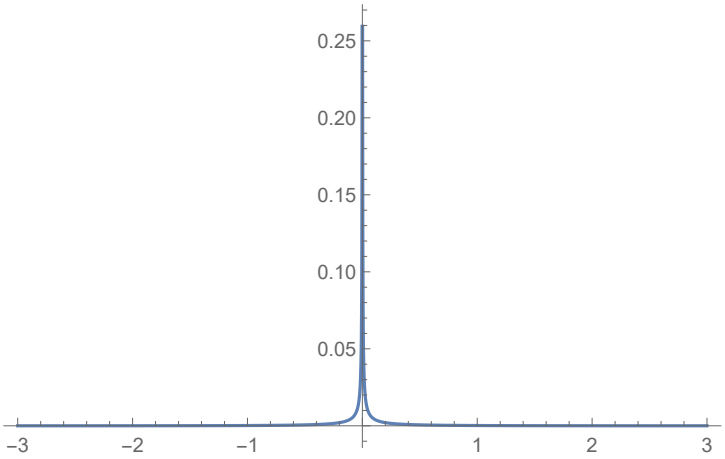

Figure 3: Density of the random variable X for the case $a = 0.001$. The peak at zero is infinite

The use of lower and upper fences is usually motivated by the wish to have "robust" boundaries instead of these based on empirical variance. Such a robustification has a sense for the case when such characteristics as a median are "robust". However, the robustness here is not just a small sensitivity concerning to heaviness of the tail of the distribution. One may pass from the "real" distribution to its truncation on the interval (Lf, Uf) and the robustness has to be understood as a small change of the median after such truncation. The next examples show such robustness may be absent.

Example 1.3. *Let $G(x, a, b)$ be a distribution function of a gamma distribution with shape parameter a and scale b. Consider*

$$F(x) = \big(G(x, 400, 1) + 1 - G(-x, 4, 100)\big)/2.$$

It is possible to calculate the quartiles and fences of $F(x)$. We have

$$Q = (-399.667, 0, 367.206), \quad Lf = -1549.98, \quad Uf = 1517.52.$$

The probability of outsiders is 0.000091629. *The truncation of $F(x)$ on the interval $[Lf, Uf]$ has quartiles*

$$Q_T = (-399.669, -329.472, 367.141).$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2020 doi:10.20944/preprints202007.0226.v1

Now we see that the median changes dramatically from the initial zero value to the value of -329.472.

It is not difficult to see the Example 1.3 may be modified to obtain arbitrary large changes of median while the corresponding truncation.

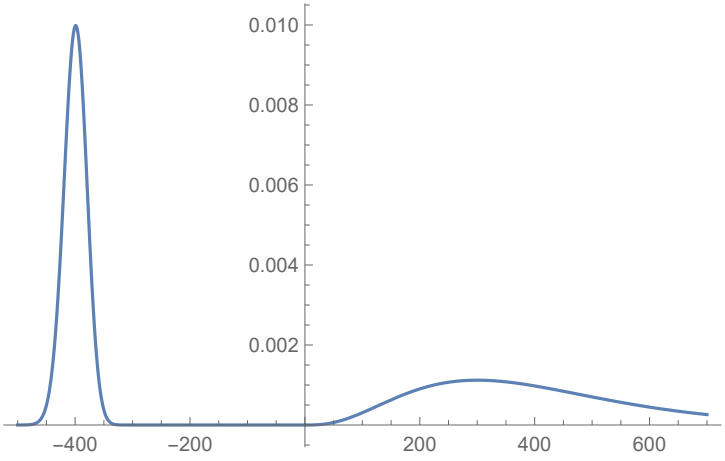

Figure 4: Density of the random variable X for the case under consideration in Example 1.3

Acknowledgments

The study was partially supported by grant GAČR 19-04412S (Lev Klebanov).

References

[1] Tukey, J. W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.

[2] Devore, J.L. (2015) Probability and Statistics for Engineering and the Sciences. Cengage Learning, Australia, Brazil, Mexico, Singapur, United Kindom, United States.

[3] Jordanova, P. (2019) Tails and probabilities for p-outliers. arXiv:1902.03810v1, 1 –45.

 doi:10.20944/preprints202007.0226.v1

[4] Lev B. Klebanov, Jaromir Antoch, Andrea Karlova and Ashot V. Kakosyan (2017) Outliers and related problems, arXiv:1701.06642v1, 1–17.