Method for Automatically Processing Outliers of a Quantitative Variable

NIANGORAN Aristophane Kerandel[1]
DIAKO Doffou Jérome[4]
Ecole Doctoral Polytechnique – UMRI 78
Institut National Polytechnique Houphouet Boigny
Yamoussoukro
Côte d'Ivoire

MENSAH Edoété Patrice[2]
INP-HB, Institut National Polytechnique Houphouet Boigny
Yamoussoukro, Côte d'Ivoire

ACHIEPO Odilon Yapo M[3]
Université Virtuelle de côte d'Ioire
UVCI, Abidjan, Côte d'Ivoire

***Abstract*—In data analysis processes, the treatment of outliers in quantitative variables is very critical as it affects the quality of the conclusions. However, despite the existence of very good tools for detecting outliers, dealing with them is not always straightforward. Indeed, statisticians recommend modeling the process underlying outliers to identify the best way to deal with them. In the context of Data Science and Machine Learning, the identification of processes that generate outliers remains problematic because this work requires a visual human interpretation of certain statistical tools. The techniques proposed so far, are systematic imputations by a central tendency characteristic, usually the arithmetic mean or median. Although adapted to the framework of Data Science and Machine Learning, these different approaches cause a fundamental problem, that of modifying the distribution of the initial data. The purpose of our paper is to propose an algorithm that allows the automatic processing of outliers by a software while preserving the distributional structure of the treated variable, whatever the law of probability is. The method is based on the moustache box theory developed by John Tukey. The procedure is tested with existing real data. All treatments are performed with the R programming language.**

***Keywords*—Outliers; boxplot; exploratory data analysis; Programming R; data science**

I. Introduction

Today, with the evolution of information collection techniques and their processing by means of computer tools, the problem of outliers has taken on a significant proportion in data analysis processes [1]. According to the definition of Grubbs (1969), "An outside observation, or 'outlier', is an observation that appears to deviate markedly from other members of the sample in which it occurs" [2]. Their treatment is a crucial problem when analysing the data. Their presence can lead to biased estimates of population parameters and erroneous results, especially in the implementation of statistical tests [3]. Ensuring high quality results when analyzing quantitative data involves detecting outliers and then processing them [4].

Several tools can be used to indicate the presence of these atypical values. Some are based on graphical techniques and others on statistical tests [5]. In those analysis, the quality of the data is one of the determining factors that contribute to a conclusion of a value. However, the methods and techniques used to process the values still do not comply with the methodology indicated by the rules of statistics. Ruilin REN in [6] notes this fact when using imputation methods, developed for missing values problems, to deal with outliers; this would be tantamount to treating outliers as missing data. Whereas statistically, the missing value problem and the outlier problem are different in nature.

It should be noted that the issue of detecting outliers through efficient procedures is solved and is even integrated in most statistical software. However, the treatment of outliers is still problematic because the methods are only diverse, but their effects on the structure of the variables are not taken into account. Among these methods, the best known are the arithmetic mean imputation method and the k-nearest neighbor algorithm, which do not have clear rules for optimal use.

Our work is aimed to automatically deal with outliers in a quantitative variable by minimizing perturbations in the probability law. We propose the determination and processing of outliers by a software in an automatic way without any human intervention. The proposed method consists of exploiting the boxplot, the main tool for outlier detection, proposed by John Tukey. The basic idea is to subdivide the data distribution into several intervals from which surrogate values to be used to replace outliers will be randomly drawn. It is thus a non-monotonic version of imputation techniques in which outliers are not replaced by a single value, but different values, all of them randomly drawn within a specific interval. The determination of the imputation interval is carried out in such a way that it retains the original distributional structure of the data.

In the paper, we first present methods that allow the detection and treatment of outliers. Then, we expose our method as well as the results obtained from the simulations performed on real data.

II. Outlier Detection

An outlier is a data item that deviates significantly from the rest of the data, as if it had been generated by a different mechanism [7]. Barnett and Lewis (1994), define an outlier in

a data set as an observation (or set of observations) that appears to be inconsistent with the rest of the data [8]. In other words, an outlier occurs when one of the observations in a data set is inconsistent with the other observations.

Outlier detection is one of the key pillars of data mining technology [9]. Many graphical methods are available to detect the presence of these outliers. These include the moustache box, bar graph, quantum diagrams, histogram, run sequence plot, etc. [1],[10],[11]. In the field of Artificial Learning, the presence of outliers in training databases is a problem for the development of good predictive models for many algorithms. Indeed, they not only make the learning time longer but they also make the obtained predictive model less optimal.

From a practical point of view, outliers may not be errors in some cases. They may be indicative of extraordinary or exceptional situations such as fraudulent behavior, rare events, etc. [12].

III. Treatment of Outliers

Dealing with outliers is a complex task in data analysis. This activity is very often neglected or neglected by analysts precisely because of its complexity or lack of knowledge of its effects on analytical results. An outlier can lead to completely wrong analytical results if it is not properly handled [5]. The detection and subsequent treatment of atypical individuals is a crucial preliminary step in data analysis [3].

These values can be, after their identification, either deleted or corrected [4]. If an atypical value is deleted, an explanation must justify the decision [13], [14]. However, this objective must take into account their origin (random or determined outliers). In the case where the emphasis is placed on the inferential characteristics of a model, during the analysis, the objective is the treatment of the atypical values in order to minimize their negative impact on the parameter estimates and the results of the analysis. In this situation, it is necessary to use appropriate methods for their treatment. This is our case in this article.

There are two main ways of dealing with outliers. The first is to correct them if the sources for producing these data are available. The second is to make a correction for them using an imputation method (by the mean or median method in general).

IV. Proposed Method of Treatment

A. Exploitation of the Boxplot

Most of the algorithms implemented in software for outlier processing use the method of imputation by the positional parameters arithmetic mean and median. This technique has the advantage of simplicity, but in most cases, it greatly alters the distributional structure of the data. Indeed, it remains acceptable when the distribution of the variable is close in practice. The principle of our method is to replace each outlier in the distribution by another value very close to it. However, these imputation values are subject to constraints that eliminate some potential candidates:

a) If the outlier is among the smallest values of the variable, it may not be replaced by a value higher than the mean value of the 1st quartile and the 3rd quartile. In its treatment, preference will be given to values closest to the lower bound of the boxplot (left side of the box).

b) If the outlier is among the largest values of the variable, it may not be replaced by a value lower than the mean value of the 1st quartile and the 3rd quartile. In its treatment, preference will be given to values closest to the upper bound of the moustache box (right side of the box).

c) The imputation values should be drawn within a close range of these values without necessarily being the same for all outliers. This last constraint prevents the cumulation of modality for a large number of outliers, which is one of the sources of distortion of the distribution of the series under study.

To do this, we will exploit John Tukey's moustache box [15]. The latter will be divided into several intervals taking into account the indicators that are the first quartile Q_1, the third quartile Q_3 and the median M_e (see Fig. 1).

Q_1 , Q_3 : are respectively the first and third quartile.

M_e : the median of the distribution.

$IQ = Q_3 - Q_1$: This is the interquartile range.

I_i represents the *i*th interval of the moustache box (i ranging from 1 to 10).

B. Description of the Method

The objective of our method is to automatically process outliers detected with the boxplot. This involves substituting them with another value so as to obtain reliable knowledge from the data, i.e. one that reflects reality. To do this, the boxplot will be decomposed into several intervals (10 in total) of 0.5*IQ length.

The different terminals are a function of the distribution parameters which are the quartiles (Q_1, M_e, Q_3). The use of quartiles is justified by the fact that they are insensitive to outliers. They are robust to the presence of outliers, unlike the mean. This, in fact, will make a real difference.

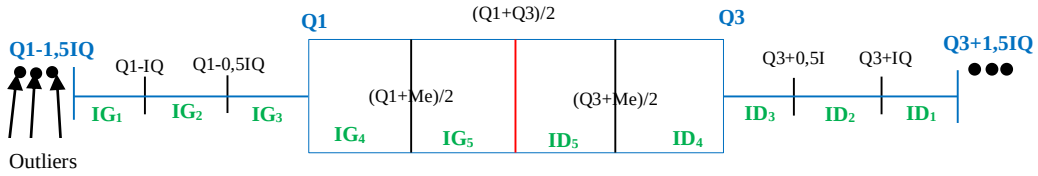

Fig. 1. Boxplot with Interval Decomposition.

To understand the principle of our method, let us consider a group of atypical individuals with respect to a given quantitative variable. Let us assume that the outliers corresponding to these individuals are located on the left side of the box (left moustache); in this case, the algorithm will find replacement values for them by traversing the calculated intervals from left to right. These new values will be randomly drawn from the current interval under consideration. This choice will first be made in interval I1. If, after replacement, it is found that there are new outliers, then this interval is abandoned in favor of the next interval (I_2). This process will be repeated until an interval provides replacement values that eliminates all outliers. It should be noted that in the course of this process, terminal value $(Q_1 + Q_3) / 2$, which represents the middle of the box, will not be crossed when randomly selecting new values for all outliers on the left side of the box.

This same process will be done with the atypical individuals located on the right side of the box. At the end of the execution of our algorithm, all outliers will be processed.

V. Proposed Algorithm for Automatic Outlier Processing

1) *Input: quantitative variable X= (x1, x2,...,xn)*
2) *Calculate the median of X*
3) *Calculate the first quartile Q1 of X*
4) *Calculate the third quantile Q3 of X*
5) *Calculate the interquartile range IQ = (Q3 - Q1)*
6) *Determine the limits of the left moustache intervals.*
    a) *binf = Q1 - 1.5 IQ*
    b) *minf = Q1 - 1IQ*
    c) *finf = Q1 - 0.5IQ*
7) *Determine the box terminals*
    a) *MQ1 = (Q1+Me) / 2*
    b) *MQQ = (Q1+Q3) / 2*
    c) *QM3 = (Q3+Me) / 2*
8) *Determine the limits of the intervals of the right moustache.*
    a) *fsup = (Q3 + 0.5*IQ)*
    b) *msup = (Q3 + 1.0*IQ)*
    c) *bsup = (Q3 + 1.5*IQ)*
9) *Determine the intervals at which the left-hand imputation values are drawn*
    a) *IG1 = [binf, minf]*
    b) *IG2 = [minf, finf]*
    c) *IG3 = [finf, Q1]*Tapez une équation ici.
    d) *IG4 = [Q1, MQ1]*
    e) *IG5 = [MQ1, MQQQ]*
10) *Determine the draw intervals for the right-hand imputation values*
    a) *ID1 = [msup, bsup]*
    b) *ID2 = [fsup, msup]*
    c) *ID3 = [Q3, fsup]*
    d) *ID4 = [MQ3, Q3]*
    e) *ID5 = [MQQQ, MQ3]*
11) *For each left outlier (less than binf) :*
    a) *For i ranging from 1 to 5 :*
        i) *Drawing a random value in IGi*
        ii) *Replace the outlier with a randomly drawn outlier*
    b) *If there are still left outliers :*
        i) *Take i = i+1*
        ii) *Go to a)*

    Else go to 12)
12) *For each right outlier (greater than bsup)*
    a) *For i ranging from 1 to 5 :*
        i) *Drawing a random value in Idi*
        ii) *Replace the outlier with a randomly drawn outlier*
    b) *If there are still left outliers :*
        i) *Take i = i+1*
        ii) *Go to a)*

    Else go to 13)
13) *End of treatment*

VI. Simulations and Results with Real Data

The outlier treatment method we proposed has been implemented under R. It is a programming language for statistics and data science [16].

A. Test with Data Iris from R

1) Description of the database: For the first simulation, we performed the test with the Iris dataset contained in the R environment. This database contains data on 150 iris flowers. For each iris, the length and width of the petals as well as the length and width of the sepals were measured. For this first test of our method, we limit ourselves to the *Sepal.Width* variable (variable measuring the width of the sepals).

2) Simulation and results: First, we plot the moustache box of the variable Sepal.Width with the plot function of R (Fig. 2), i.e. without intervention of our method. We can notice on Fig. 2 the presence of atypical individuals on both sides of the box mustaches (outliers).

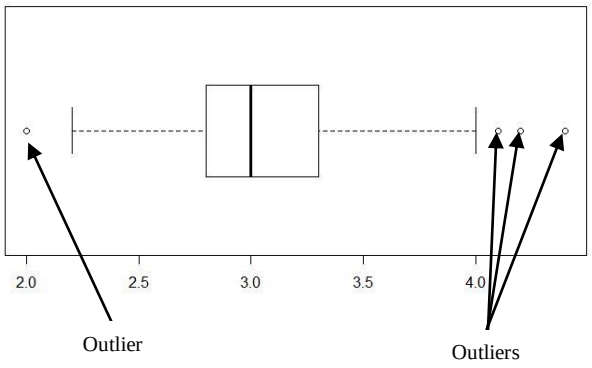

Fig. 2. Drawing of the Mustache Box of the Variable Sepal.Width of Iris with the Original Data.

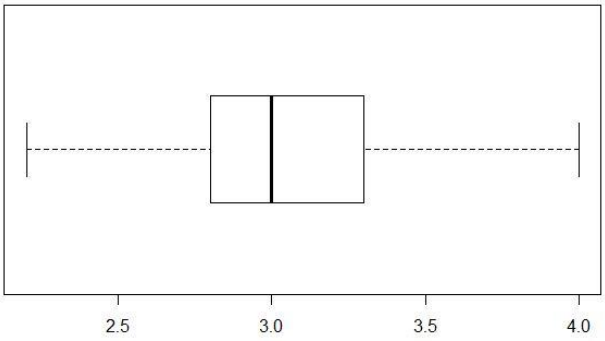

Fig. 3. Drawing of the Sepal.width Variable Moustache Box after Application of the Algorithm.

In the second phase, we apply our outlier processing algorithm to the same data. The algorithm with its execution principle explained above (in 3) will replace these outliers by new ones. We have on Fig. 3 the result obtained with our method.

We can notice that after applying our algorithm on the data, the outliers have been treated so they no longer appear on the moustache box plot.

B. Test with Data from Open Data

1) Description of the database: For the second test, we used data transcribing the monthly performance of the hotel sector in the Brussels region as a resource [17]. This file contains information on the variables *Occupancy.rate* (Room occupancy rate), *Average.Price* (Average price per room) and RevPAR (Income per available room).

2) Simulation and results: First, we plot the moustache box of the variable *Occupancy.rate* with the plot function of R (Fig. 4), i.e. without intervention of our method. We can notice on Fig. 4, the presence of atypical individuals on the left side of the boxplot.

In the second phase, we apply our outlier processing algorithm to the same data. On Fig. 5, we have the result obtained with our method.

One can notice the presence of outliers on the side of the left moustache of the box.

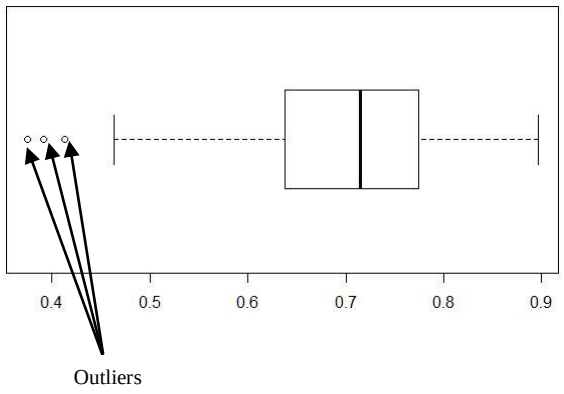

Fig. 4. Plotting the Mustache Box of the Variable Occupancy.rate with the Original Data.

Fig. 5 shows the absence of outliers after application of our algorithm on the data. All the outliers have been processed.

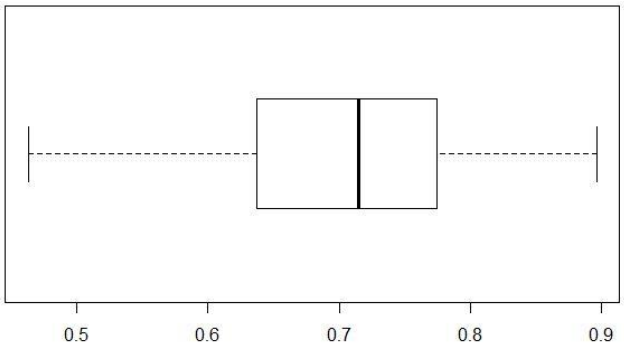

Fig. 5. Drawing of the Boxplot on the Variable Occupancy.rate after Application of the Algorithm.

VII. Conclusion

The algorithm developed by us in this article is mainly based on the principle of John Tukey's boxplot. It exploits its properties to automatically handle outliers in quantitative variables. While the statistical approaches used to identify outliers are effective, current methods of dealing with these outliers are based on imputation techniques that change the probability distribution of the variable of interest. Our method has the particularity of preserving the structure of the distribution of the treated variable. Simulations have shown that the method always manages to treat correctly atypical individuals. This algorithm is part of the more general search for solutions to automate the process of quantitative variables analysis. It will make it possible to automate correct data analysis methodologies with a view of making as reliable as possible the results of data analysis by a machine without human intervention. Another interest of our approach is to be able to write statistical software for novice statisticians that can produce highly reliable results by minimizing the possibilities of common analytical errors due to a lack of knowledge of the rules, limits and conditions of validity of statistical data analysis methods. Theoretically, it seems that our method may lead to a situation where no interval is suitable for removing outliers. However, in practice, the method is always able to deal adequately with outliers. One avenue for reflection would therefore be to examine the theoretical convergence of the

algorithm regardless of the nature of the distribution of the quantitative variable used. This work is part of a series of research projects aimed at automating the analysis of a quantitative variable. It follows on from a method we have developed and published, which allows the automatic identification of the symmetrical or non-symmetrical nature of the distribution of a quantitative variable without the use of graphs or statistical tests. Future research may address the development of a method for automating the process underlying the Stem-and-Leaf tool, which will make it possible to automatically analyze a quantitative variable according to the Exploratory Data Analysis approach as advocated by John Turkey. Such a perspective is necessary if we want to make it possible to develop effective analytical solutions in the context of BI 4.0 and pave the way for real-time statistics for the analytical needs of the Internet of Things.

REFERENCES

[1] V. Planchon, « Traitement des valeurs aberrantes : concepts actuels et tendances générales », Biotechnol. Agron. Soc. Environ., p. 16, 2005.

[2] F. E. Grubbs, « Procedures for Detecting Outlying Observations in Samples », Technometrics, vol. 11, no 1, p. 1-21, févr. 1969, doi: 10.1080/00401706.1969.10490657.

[3] S. Seo, « A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets », p. 59.

[4] N. S. Henia, « Synth`ese des M´ethodes sur les Donn´ees Aberrantes », p. 51.

[5] E. Ftoutou, M. Chouchane, et N. Besbès, « Etude des effets de l'élimination et d'imputation des valeurs aberrantes sur la classification d'un défaut d'injection dans un moteur Diesel », p. 2, 2014.

[6] «Méthodes D'imputation De Valeurs Aberrantes Pour Des Données D'enquêtes », p. 14, 2002.

[7] J. Han, M. Kamber, et J. Pei, « Outlier Detection », in Data Mining, Elsevier, 2012, p. 543-584.

[8] V. Barnett et T. Lewis, Outliers in statistical data, 3rd ed. Chichester ; New York: Wiley, 1994.

[9] F. Angiulli, « Data Mining: Outlier Detection », in Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2019, p. 456-462.

[10] M. N. A. Zerbet, « Détection des observations aberrantes par des méthodes statistiques », p. 28.

[11] B. Mogoş, « Exploratory data analysis for outlier detection in bioequivalence studies », Biocybernetics and Biomedical Engineering, vol. 33, no 3, p. 164-170, janv. 2013, doi: 10.1016/j.bbe.2013.07.005.

[12] V. J. Hodge et J. Austin, « A Survey of Outlier Detection Methodologies », p. 42.

[13] J. I. E. Hoffman, « Outliers and Extreme Values », in Basic Biostatistics for Medical and Biomedical Practitioners, Elsevier, 2019, p. 149-155.

[14] H. Mark et J. Workman, « Outliers—Part 3: Dealing With Outliers ☆ », in Chemometrics in Spectroscopy, Elsevier, 2018, p. 931-936.

[15] R. L. Nuzzo, « The Box Plots Alternative for Visualizing Quantitative Data », PM&R, vol. 8, no 3, p. 268-272, mars 2016, doi: 10.1016/j.pmrj.2016.02.001.

[16] «R: The R Project for Statistical Computing ». https://www.r-project.org/ (consulté le mai 16, 2020).

[17] «Performance Hôtelière - Jeux de données - Open Data: portail des données ouvertes de la Région de Bruxelles-Capitale ». http://opendatastore.brussels/fr/dataset/hotels-performance (consulté le avr. 03, 2020).