ISBN: 978-1-948012-08-9

Contents List available at VOLKSON PRESS

International Symposium on Computer Science and Artificial Intelligence(ISCSAI)

The outlier test based on normally distributed data

Chen Peifan[a], Chen Fen[a]

[a]*Department of Economics and Management, Nanjing University of Science and Technology, Xiao Ling Wei Street, Nanjing, China*
**Corresponding Author: chenpf273@163.com*

ARTICLE DETAILS

Article History:

Received 02 october 2017
Accepted 06 october 2017
Available online 11 october 2017

Keywords:

The outlier test, Quartile model, PauTa model, Chauvenet model, Grubbs model

ABSTRACT

In this paper, the simulation experiments are used to research the efficiency of different outlier test. R software is used to generate the normally distributed data, which contains the outliers defined differently. Then each model are applied to test the outliers in the dataset. The experimental result shows that the efficiencies of the same model differed greatly when the definition of the abnormal value varies. So we can draw the conclusion that the efficiency of the test model depends on the definition of the outlier. And the efficiency of Quartile model is higher than other models. So Quartile model is more effective and applicable than other models in actual data test.

1. Introduction

In recent years, as people pay more attention to statistical data, the quality of statistical data is also getting higher and higher, in order to better ensure the quality of statistical data, in particular, pay attention to statistical data in the abnormal value.

An outlier is an individual value that is significantly deviated from most observed values in the sample [1]. At present, there are many ways to detect outliers, including statistical methods based on clustering, density-based methods [2]. Based on the perspective of statistics, sample size, data set distribution from different detection model and its corresponding to the advantages and disadvantages of point of view, we propose a variety of detection models, Moreover, they used the residual t-test of regression analysis to test the binary variable outliers [3]. The test method Dixon Grubbs, has conducted test for outliers using R language [4]. In the outliers were tested from the Quartile model using the box graph function in SPSS [5,6].

The corresponding test model is applied in a certain field, without distinction in different outlier definitions, test methods of test results. Therefore, according to the definition of abnormal value different from that on a comparative study of different model test results, our current outlier test fill blank in the research, has a strong practical significance.

2. THE MODEL OF OUTLIER TEST

According to the characteristics of commonly used outlier detection methods in document, the most commonly used Quartile model, PauTa model, Chauvenet model and Grubbs model are given below [7].

2.1 Quartile model

Quartile model, also known as four-point method, also called box diagram method. Boxplot method is the use of five statistical data points: minimum, four quantile (Q1), median (Q2), four quantile (Q3), and the maximum value to describe data.

The box diagram consists of three parts: reference frame (coordinate axis), sign (box, top and bottom four quantiles, median, outlier's truncation point), detection data (extension line and abnormal value at both ends of the box), as shown in figure 1. The two ends of the box correspond to the lower four quantile Q1 and the upper four quantile Q3, respectively, and Q1 and Q3 are called the four Inter Quartile (IQR Range). The point on the upper four points is 1.5 times the IQR on the right and the 1.5 IQR on the left of the next four points is the outlier truncation point, and the truncation point between the outliers is the internal limit. The 3-point IQR on the right and the 3 IQR position on the left of the four sub loci correspond to the cutoff point of the extreme value, and the extreme value between the truncation points is the outer limit of the four point. The abnormal value of truncation points outside data called outliers, including limit and abnormal value limit between moderate outlier value (Outlier, mild, outliers) in the outer limit except for extreme outliers or extreme values (Extreme, extreme, outliers) [7].

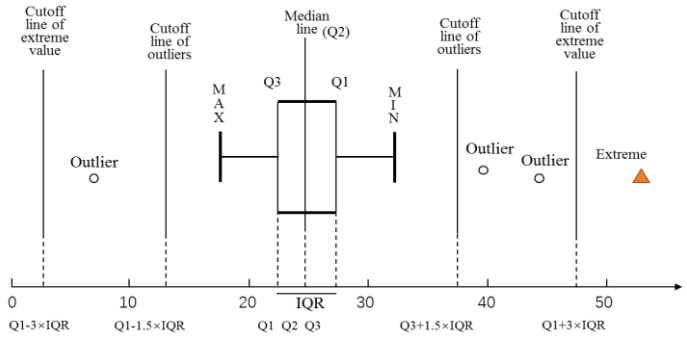

Figure 1: The structure of the boxplot

2.2 Grubbs model

2.2.1 PauTa model

The PauTa model is also called the 3σ criterion. The model tests the data based on the normal distribution. According to the normal distribution theory, the $|v_d| \leqq 3\sigma$ probability is 99.7%, and the σ is the standard deviation of the sample. Therefore, in the limited detection, the probability of a test error is greater than 0.3%, the test value is considered as the abnormal value of gross error value, which needs to be eliminated [8].

2.2.2 Chauvenet model

The Chauvenet model, also known as the Chauvenet criterion, is also based

on the normal distribution of data. Several hypotheses testing repeatedly measure the value, residual error of a detected value is$|\Delta X_i|>z_c\sigma$, this data is rejected.

The N≦10 sample containing them are as the discriminant coefficient N, $Z_C$value can be obtained. The Chauvenet criterion has make up PauTa criterion to a certain extent is not sufficient, can be used for N≦10 of outlier's judgment. Compared with the PauTa criterion, Chauvenet criterion is more stringent [8].

2.2.3 Grubbs model

The Grubbs model is also called the Grubbs criterion, which is similar to the Chauvenet criterion, but the determinant coefficients are different. Several detection in the Grubbs code assuming repeated measure the value, when the residual error of a detection value greater than T_o (n,α), this data is rejected. T_o is the sample contents of them are as the discriminant coefficient n, by T_o (n,α)value table access. With the Chauvenet criterion in different Z_c values of the coefficient of determination is that, the T_o value and the repeated measurement times of n and α were related to the confidence probability, The probabilistic meaning of the Grubbs criterion is more explicit. At the same time, Grubbs criterion also requires normal distribution of test data [8].

3 COMPARISON OF FOUR OUTLIER DETECTION MODELS

3.1 The first group randomized experiment

3.1.1 Based on the different definition of abnormal value

The experiment should meet the following assumptions:

(1) Outliers are defined based on the PauTa criterion.

(2) Assume that the number of randomly generated standard normal data sets contains 1 or 2 of the specified outliers.

(3) Based on the assumption (1) and (2), the following steps are used to generate the corresponding data set:

Step 1: Randomly generate standard normal data sets of the sample size of n is G:N(μ,σ).

Step2: Do the PauTa test for G, and check whether the number of outliers is 1 or 2.

Step3: After Step2 test, enter Step4; otherwise, go back to Step1.

Step4: the abnormal values which are checked out in Step2 are used as the specified outliers so that the required data set G is obtained.

3.1.2 Determination of Random Experimental Conditions

The generation of the above data set relies heavily on the selection of the data set sample size. Choosing a good initial sample size has a great impact on the rapid generation of specific data sets. According to the standard normal distribution of the 3σ criterion, from the standard normal random sample in a sample, which fall within the interval [3, 3] between the probability of 99.73%. Therefore, the data randomly generated by the subject falls standard normal distribution [3, 3] is the probability of other than 0.27% <0.01.When the sample size increases, the number of randomly generated data other than [3, 3] will increase. Therefore, for the outliers based on the PauTa criterion, it is easier to select the appropriate data set for the random data set sample size, and to improve the efficiency of the data set. And the data set of the sample size is too large or too small, are difficult to pass Step2, and can't carry out the subsequent model effect experiment.

According to the relevant criteria in 1.2, based on Chauvenet, Grubbs criteria for the abnormal value of the judge and the sample size n, the significance level. Through the literature can be [6,9]:
[1] Based on the Chauvenet criterion, the sample size n = 1,2,3, ... 100, 200, 500, the value of Zn.
[2] Based on the Grubbs criterion, the value of $T_{(n,\ \alpha)}$decreases with the significance levelαand the sample size n simultaneously.

Considering the efficiency of data set generation, Zn and $T_{(n,\ \alpha)}$, and finally determine the sample size based on the PauTa criterion, the Chauvenet criterion, the Quartile criterion is n = 100; n = 200; n = 500, the randomly generated sample size and significance level based on the Grubbs criterion = 100, α = 10%; n = 100, α= 5%; n = 100, α= 2.5%.

3.1.3 Experimental results and analysis

Using the above data set to generate the algorithm and the sample size, generate the corresponding data set. Here we designed four groups of experiments, random simulation of the number of 20, the use of R software "outliers" package Grubbs. test () and Boxplot () function can be generated on the data set to do the value of the test, and then through the experimental results Analyze the applicability of various models [4,10-12]. Table 2, 3, 4, 5, respectively, contains the definition of different criteria based on different data, the use of various models to test the correct rate, the results are as follows:

Table 1: The experimental results of abnormal value defined based on PauTa model

Sample size n	n=100	n=200	n=500	average value
Quartile model	90%	100%	100%	97%
PauTa model	100%	100%	100%	100%
Chauvenet model	100%	100%	40%	80%
Grubbs model	45%	35%	15%	32%

The experimental results of abnormal value defined based on PauTa model
Table 2: The experimental results of abnormal value defined based on Chauvenet model

Sample size n	n=100	n=200	n=500	average value
Quartile model	75%	100%	95%	90%
PauTa model	40%	100%	100%	80%
Chauvenet model	100%	100%	100%	100%
Grubbs model	30%	20%	35%	28%

Table 3: The experimental results of abnormal value defined based on Grubbs model

Experimental conditions	(100,10%)	(100,5%)	(100,2.5%)	average value
Quartile model	85%	100%	100%	95%
PauTa model	100%	100%	100%	100%
Chauvenet model	100%	100%	100%	100%
Grubbs model	100%	100%	100%	100%

Table 4: The experimental results of abnormal value defined based on Quartile model

Sample size n	n=100	n=200	n=500	average value
Quartile model	100%	100%	100%	100%
PauTa model	25%	25%	20%	23%
Chauvenet model	50%	20%	10%	27%
Grubbs model	20%	15%	10%	15%

Analysis of tables 1, 2, 3, 4 can be the following conclusions:

• When the outliers are different in the data set, the detection effect of the same abnormal value detection model is very different. Based on the different values defined by different criteria, the correct rate is 100%. However, when the cross test is carried out, the correct rate Have declined.

• The Quartile model's test accuracy is greater than 90%, regardless of which exception value definition criteria Quartile model for the value of less than 2 out of the data set, the overall test efficiency is higher.

• The Quartile model applies to a variety of data, while the other three models apply only to normal data. Therefore, when the number of abnormal values is small, or the data is not normal, Quartile model compared to other models, the applicability of a wider, more robust.

• The Grubbs model can only be used for the detection of the number of outliers ≤ 2 and is only applicable to datasets with smaller sample sizes. For the exception values defined in other ways, the test efficiency is low, although the anomaly of industrial data In general recommended Grabus law, but its applicability and poor health. In the case of uncertain number of outliers, or when the amount of data is large, the Grabus method is not appropriate.

3.2 Randomized trials

3.2.1 Experiment and results

The first set of random experiments requires that the number of outliers is less than 2, and there is no comparison between the number of abnormal values. Therefore, to further compare the suitability and robustness of each model, we designed a second randomized experiment. In the second group, we did not limit the number of outliers. We produced stochastic data with different sample sizes and subject to normal distribution. Twenty random experiments were carried out to directly compare the test results of different models. In this group of experiments, because the Grubbs model does not apply to the larger sample size and the number of abnormal values ≥ 2 data set, so no Grubbs model for outliers and comparison, only compared to the Quartile model, PauTa model the accuracy of the Chauvenet model. Tables 5, 6, and 7 give the correct values for the anomalies defined by different criteria, and the correctness of the tests using different models. The results are as follows:

Table 5: The experimental results of abnormal value defined based on PauTa model

Sample size n	n=100	n=200	n=500	average value
Quartile model	92.5%	95%	100%	95.8%
PauTa model	100%	100%	100%	100%
Chauvenet model	100%	100%	52.9%	84.3%

Table 6: The experimental results of abnormal value defined based on Chauvenet model

Sample size n	n=100	n=200	n=500	average value
Quartile model	82.5%	100%	100%	94.1%
PauTa model	40%	100%	100%	80%
Chauvenet model	100%	100%	100%	100%

Table 7: The experimental results of abnormal value defined based on Quartile model

Sample size n	n=100	n=200	n=500	average value
Quartile model	100%	100%	100%	100%
PauTa model	33.3%	30.4%	39.4%	34.3%
Chauvenet model	54.6%	29.1%	16%	33.2%

It can be found from Table 6, 7 and 8 that the accuracy of each model is 100% under the corresponding definition, and the efficiency is reduced when the cross test is carried out. Quartile model for the different criteria defined by the outliers, the accuracy of its cross test is greater than 90%, are higher than the accuracy of another models cross-test. However, in any case, the Quartile model has high efficiency and the Quartile model has no normality requirement, which further shows that the Quartile model has good applicability and robustness.

4 CONCLUSION

In this paper, we introduce a series of experiments by introducing the commonly used anomaly detection model and then defining the different outliers. At the same time for the actual life of the abnormal value of the test, the need to note the following:

(1) Abnormal value is a relative concept, the exception is looking at the 0-1 variable, that is, either unusual, or not unusual. In the actual outlier verification process, according to the background of the data set, select the appropriate definition of abnormal value.

(2) The experimental results show that the detection effect of the same anomaly detection model is very different when the definition of outliers in the data set is different. However, regardless of which outlier definition method, Quartile model test accuracy is high. And Quartile model for a variety of data, the data is not a normal requirement. Quartile model has a good applicability, compared with other abnormal value test method of its best robustness.

In short, the outlier detection process is a complex multi-dimensional analysis process. Need to consider the data set of the sample size, probability distribution, the actual background, the definition of abnormal values and other factors, and then to the data set to do the value of the test.

REFERENCES

[1] Deran, Z. 2003. Test Method for Outliers in Statistics [J]. Statistical Analysis, 5 (1), 53-55.

[2] Hongding, W., Yunhai, T., Shaohua, T. 2006. Development of outlier detection [J]. Journal of Intelligent Systems, 1 (1), 61-73.

[3] Rongwu, X., Xiao, Z., Dejiang, T. 2013. Discrimination of common outliers in drug quality control [J]. Chinese Journal of Pharmaceutical Analysis, 33 (11), 1845-1848.

[4] Huailiang, W. 2012. Identification of Outliers of Statistical Data and Implementation of R Language [J]. Transactions of China Electronics Technology, 5 (2), 6-7.

[5] Zhongwen, C., Xuelian, Y., Jianjun, Z. 2014. Application of SPSS software in data review and processing of water census [J]. Water Resources Science and Technology Economy, 20 (1), 153-293.

[6] Xianghong, S., Yongjun, L., Wenwen, C. 2010. The application of box plot method in the anomaly testing of animal health data [J]. Statistical Analysis, 27 (7), 66-68.

[7] Lifeng, G., Yutao, C. 2001. Mechanical Engineering Materials Test Manual [M]. Shenyang: Liaoning Science and Technology Press, 82-84.

[8] Yanyan, X., Xianqiu, W. 2010. Comparison and Application of Four Criteria of Gross Error [J]. University Physics Experiment, 23 (1), 66-68.

[9] Bingchang, Z., Liang, W. 2011. R Language Beginner's Guide [M]. Xi'an: Xi'an Jiaotong University Press, 81-93.

[10] Junping, J., Xiaoqun, H., Yongjin, J. 2012. Statistics [M] .5 Beijing: Renmin University of China Press, 157-160.

[11] Tao, G., Nan, X., Gang, C. 2013. R Language Practice [M]. Beijing: People's Posts and Telecommunications Press, 40-81.

[12] Yu, Z., Wenyi, C., Ying, Z. 2011. R language data manipulation [M]. Xi'an: Xi'an Jiaotong University Press, 12-40.