

Comparison of methods for detecting outliers

Manoj K, Senthamarai Kannan K

Abstract - An outlier is an observations which deviates or far away from the rest of data. There are two kinds of outlier methods, tests discordance and labeling methods. In this paper, we have considered the medical diagnosis data set finding outlier with discordancy test and comparing the performance of outlier detection. Most of the outlier detection methods considered as extreme value is an outlier. In some cases of outlier detection methods no need to use statistical table. The suggested outlier detection methods using the context of detection sensitivity and difficulties of analyzing performance for outlier detections are compared.

Index Terms — Discordance test, Dixon, Generalized ESD, Grubbs, Hampel, Outlier Detection

1 INTRODUCTION

Outlier is an interesting field of data mining. The identification of outliers can lead to the unexpected knowledge discovery in the areas such as credit card fraud detection, criminal behaviors detection, computer intrusion detection, calling card fraud detection etc. Applications such as outlier detection customized marketing, network intrusion detection, weather prediction, pharmaceutical research and exploration in science databases require the detection of outliers.

Barnett and Lewis (1978) defined as in a sample of moderate size taken from a certain population it appears that one or two values are surprisingly far away from the main group. D.M. Hawkins (1980) gives definition to outlier as: An outlier is an observation, which so much deviates from other observations as to arouse suspicions that it was generated by a different mechanism. Example, dataset from Laurie Davies (1993)

9.1, 79.5, 26.8, 81.5, 19.1, 15.2, 22.6, 28.8, 24.1, 23.6,
18.6, 17.3, 25.8, 78.8, 23.1, 11.9, 20.1, 20.3, 14.1, 26.5

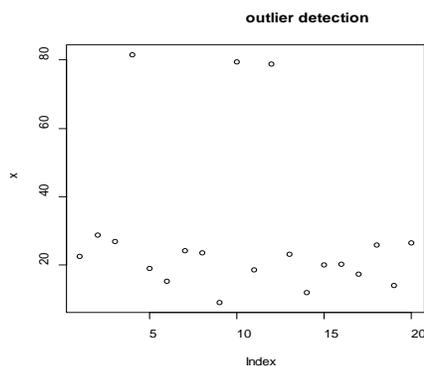


Fig - 1. Scatter plot for outlier detection

The fig-1 represents the three points 81.5, 79.5, and 78.8 are far away from the data set. In the three values are considered as outlier.

Anscombe (1960), have attempted to categorize the different ways in which outliers may arise. It is relevant to consider them in rather more detail. In taking observations, different sources of variability can be encountered. We can distinguish three of these.

Inherent variability:

This is the expression of the way in which observations intrinsically vary over the population; such variation is a natural feature of the population and uncontrollable. Thus, for example, measurements of heights of men will reflect the amount of variability indigenous to that population.

Measurement error:

Often we must take measurements on members of a population under study. Inadequacies in the measuring instrument superimpose a further degree of variability on the inherent factor. The rounding of obtaining values, or mistakes in recording, compound the measurement error: they are part of it. Some control of this type of variability is possible.

Execution error:

A further source of variability arises in the imperfect collection of our data. We may inadvertently choose a biased sample or include individuals who are not truly representative of the population we aimed to sample. Again, sensible precautions may reduce such variability.

Treatment of Outliers

The various outlier methods are using to test and compared in this paper. Recently, most of people affected by the blood pressure. They have to resort to the hospital to check their health conditions. The treatments cannot cure in single day. They need every time after consumption of drugs, blood pressure is checking by physician. Sometimes measuring the blood pressure referred to false measurements. It may be negligence of the physician or the measuring error instrumented. It is not a valid measure of treatment. In this situation using outlier detection method

- K. Manoj, Research Scholar, Department of Statistics Manonmaniam Sundaranar University, Tamil Nadu, India, E-mail: manojstatms@gmail.com
- K. Senthamarai Kannan, Professor, Department of Statistics Manonmaniam Sundaranar University, Tamil Nadu, India, E-mail: senkannan2002@gmail.com

is very useful to find the right treatment.

2 RELATED WORK

The previous studies using outliers methods to find the different methodologies and results. Armin Bohrer (2008) proposed method for using Dixon's outlier test has been calculated using Monte Carlo simulation one sided two-sided case critical values are determined. Barbato G. et.al (2011) discussed about a several statistical methods that are currently in use for outlier identification and their performance are compared theoretically for typical statistical distributions of experimental data and considering values derived from the distribution of extreme order statistics as reference terms.

Grubbs (1969) describes the procedures are given for determining statistically whether the highest observation, the highest and lowest observations, the two highest observations, the lowest observations, or more of the observation in the sample are statistical outliers. Khrominski (2010) using various methods of outlier detection in medical diagnoses. They discussed investigated the usefulness of selected outlier detection methods in the context of detection speed and performance analysis and the difficulty of automating the performance analysis by using the test methods for outlier detection.

Thomas et. al., (1988) describes the outlier test procedure was found to influence the interlaboratory standard deviations (SDs), but not the averages. It was shown that even small number of differences in the numbers of outliers detected can change the SD severely. Comparing the outliers test procedures for Hampel, Grubbs and Graf-Henning, it was found that Hampel test detected the most outliers. Tietjen (1973) proposed a procedure of studentizing or standardizing the residuals by dividing them by their estimated standard deviations is proposed for testing for outliers in simple linear regression.

Paul and Fung (1991) are concerned with describes the procedures for detecting multiple y outliers in the linear regression. The generalized extreme studentized residual (GESR) procedure, controls which type I error rate, is developed and approximate formula to calculate the percentile is given for large samples and more accurate percentiles for $n \leq 25$ are tabulated. The procedure performance is compared with others by Monte Carlo techniques and found to be superior. However, the procedure fails in detecting y outliers that are on high-leverage cases. They suggest a two-phase procedure. The phase- 1 a set of suspect observations is identified by GESR and one of the diagnostics applied sequentially and phase- 2 a backward testing is conducted using the GESR procedure to see which of the suspect cases are outliers. They analyzed several examples in this paper.

Quesenberry (1961) discussed on the rejection and

location of outlying observations that there might be several ways of approaching the problem, which depended to a large extent on the object in view. One might, for instance, be primarily interested in pruning the observations in order to secure a more accurate analysis of what was left, example to obtain the most reliable estimate of a mean. Or one might be particularly interested in identifying the genuinely exceptional observations, in order to a new insight into the phenomena under study. In the first case the criterion of what was best might be the effect on the standard error of estimation, in the second case the risk of wrongly deciding whether an observation was exceptional or not. The procedures discussed in the following paper start from the basis of risks of misclassification rather than of estimation errors.

McMillan (1971) describes performances of three procedures for treatments of outliers in normal samples are evaluated. The first procedure is the continuous application of the usual maximum residual test. The largest value is an outlier if the largest studentized residual exceeds a already determined value. If one outlier is detected, the test is repeated on the remaining observations, in the process to continue until no further outliers are detected. The second procedure is two largest observations are declared to outliers if the sum of two largest studentized residuals exceeds a predetermined value. In the third procedure of the two largest values are considered outliers if the ratio of the corrected sum of squares omitting these values to the total corrected sum of squares is less than a critical ratio. The procedure performances are evaluated for samples in which two of the values have means different from the common mean of the remainder of the sample.

Tietjen and Moore (1972) are described problems of repeated application and "masking". They suggested as appropriate to over-come these problems are two new statistics: L_k which is based on the k largest (observed) values and E_k which is based on the k largest (in absolute value) residuals. Jacqueline and Hawkins (1981) proposed method for accurate bounds a represented for the fractiles of the maximum normed residual (which is often used to test for a single outlier) for two way and three way layouts and its shown that the second Bonferroni bound of the critical value is an excellent approximation of the critical value being much more accurate the first B on ferroni upper bound. The third Bonferroni (upper) bound is expensive to compute and agrees with the second bound to at least four decimal places for all factor combinations considered.

Laurie Davies and Ursula Gather (1993) approach to identifying outliers is to assume that the outliers have a different distribution from the remaining observations. They define outliers in terms of their position relative to the model for the good observations. The identification outlier problem is then the problem of identifying those observations that lie in a so-called outlier region. A more detailed analysis shows that methods based on robust

statistics perform better with respect to worst-case behavior. They given a concrete outlier identifier based on a suggestion of Hampel.

Rosner (1975) proposed with "many outlier" procedures that can detect more than one outlier in a sample. various many outlier procedures are proposed via power, comparisons in Section 3 are found to be much superior to one-outlier procedures in detecting many outliers. They compare several different many outlier procedures find that the procedure based on the extreme studentized deviate (ESD) is slightly the best. Finally, 5%, 1% and .5% points are given for the ESD procedure for various sample sizes.

3 METHODOLOGY

In this section discussed about some formal tests using outlier detection. The described method consists of the information about way of counting the outlier values for the tests. The method testing with a formula necessary to find the outliers in the data set. In these methods final description discussed some conditions under which a decision whether checking data is an outlier or not is made.

There are two kinds of outlier methods, Formal Method and Informal Method. It is usually called, 'Tests of Discordance' and 'Labeling Methods' respectively. A detection test procedure must need to a statistical test, termed here a test of discordance. They are usually based on assuming some well-behaving distribution, and test if the target of extreme value point is an outlier in the distribution.

3.1 Grubbs Test

Grubbs (1969) used to detect a single outlier in a univariate data set. The data set that follows an approximately normal distribution. Grubbs' test is defined as the following two hypotheses:

H0: There is no outlier in the data set

H1: There is at least single outlier in the data set

The general formula for Grubbs' test statistic is defined as:

$$G = \frac{\max |Y_i - \bar{Y}|}{s}$$

Where y_i is the element of the data set, \bar{Y} and s denoting the sample mean and standard deviation and the test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation. The calculated value of parameter G is compared with the critical value for Grubb's test. When the calculated value higher or lower than the critical value of choosing statistical significance, then the calculated value can be accepted as and outlier. The statistical significance (α) describes the maximum mistake level which a person searching for outlier can accept.

3.2 Quartile Method

Quartile method is no need to use in statistical tables. To find the outlier using the quartile method it is necessary to carry out the following steps:

Step: 1

Calculate the upper quartile: Q3 - 75% of the data in the data set are lower than this.

Step: 2

Calculate the lower quartile: Q1 - 25% of the data in the data set are higher than this.

Step: 3

Calculate the gap between the quartiles: $H=Q3 - Q1$

A value lower than $Q1 - 1.5.H$ and higher than $Q3+1.5.H$ is considered to be a mild outlier. A value lower than $Q1-3.H$ and higher than $Q3+3.H$ is considered to be an extreme outlier.

3.3 Dixon's Test

The test developed by Dixon (1950) and used to the test is appropriate for small sample size. The test has some limitations to $n \leq 30$, were later on extended to $n \leq 40$ (UNI 9225: 1988). The test first step for organizing the data in an ascending order, and then the next step is to count parameter R.

The test has various test statistics. Suppose for testing large set of element to be an outlier, the sample arranged in ascending order $X_1 \leq X_2 \leq \dots \leq X_n$ Implying that the large sample element is given by X_n . Dixon proposed the following test statistics defined as

$$R_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}, \quad \text{for } 3 \leq n \leq 7$$

$$R_{11} = \frac{x_n - x_{n-1}}{x_n - x_2}, \quad \text{for } 8 \leq n \leq 10$$

$$R_{21} = \frac{x_n - x_{n-2}}{x_n - x_2}, \quad \text{for } 11 \leq n \leq 13$$

$$R_{22} = \frac{x_n - x_{n-2}}{x_n - x_3}, \quad \text{for } 14 \leq n \leq 30$$

For testing the smallest sample element to be an outlier, the sample is ordered in descending order implying that the smallest sample element is labeled X_n . All the selection of the test statistics depends on the Dixon's criteria.

The variable X_n is marked as an outlier, when the corresponding statistic $R^{(n)}$ exceeds a critical value, which depends on the selected significance level α .

The calculated value of the parameter R is compared with the Dixon's test critical value for choosing statistical significance. When the calculated value of parameter R is bigger than the critical value then it is possible to accept data from the data set as an outlier.

3.4 Hampel Method

To calculate Hampel's test statistical tables are not necessary. Theoretically, this method is resistant, which means that it is not sensitive to outliers, it also has no restrictions as to the abundance of the data set.

Hampel's test performs the steps for data sets are as follows:

- i. Compute the median (Me) for the total data set. The median is described as the numeric value and separating the higher half of a data set from the lower half.
- ii. Compute the value of the deviation r_i from the median value; this calculation should be done for all elements from the data set:

$$r_i = (x_i - Me)$$

where, x – simple data from the data set,

i – belongs to the set for 1 to n .

n – number of all element of the set

Me – median

- iii. Calculate the median for deviation $Me_{|r_i|}$
- iv. Check the conditions: $|r_i| \geq 4.5Me_{|r_i|}$

If the condition is executed, then the value from the data set can be accepted as an outlier.

3.5 Generalized ESD Test for Outliers

Rosner (1983) used in the generalized (extreme Studentized deviate) ESD test to detect one or more outliers in a univariate data set that follows an approximately normal distribution.

The generalized ESD test (Rosner 1983) only requires that an upper bound for the suspected number of outliers be specified.

Given the upper bound, r , the generalized ESD test essentially performs r separate tests: a test for single outlier, a test for two outliers, and so on up to r outliers. The generalized ESD test is defined for the hypothesis:

H_0 : There is no outlier found in the data set

H_a : There are up to r outliers in the data set

Test Statistic: Compute

$$R_i = \frac{\max_i |x_i - \bar{x}|}{s}$$

Remove the observation that maximizes $|x_i - \bar{x}|$ and then compute the above statistic with $n - 1$ observations. Repeat and continues the process until r observations have been removed. Then the results in r test statistics R_1, R_2, \dots, R_r .

Significance Level: α

Critical Region: Corresponding test statistics r to calculate the following r critical values

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t_{p,n-i-1}^2)(n-i+1)}}$$

where $i = 1, 2, \dots, r$, $t_{p,v}$ is the 100_p percentage point from the t distribution with v degrees of freedom and

$$p = 1 - \frac{\alpha}{2(n-i+1)}$$

Number of outliers is determined by finding the largest I such that $I > \lambda_i$. Simulation studies by Rosner (1983) indicate that this critical value approximation is very accurate for $n \geq 25$. It is used to test with higher number of outliers than expected when testing for outliers among data coming from a normal distribution.

4 Results and Discussion

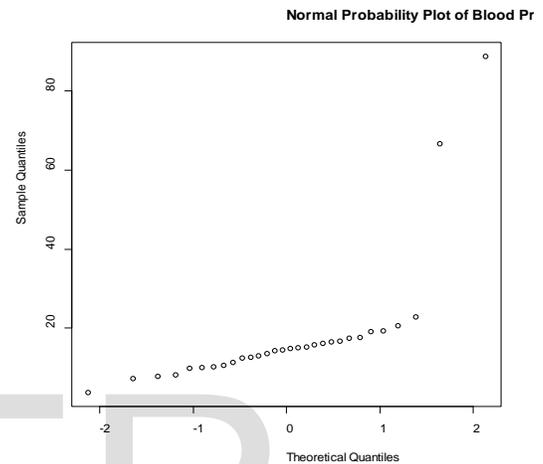


Fig - 2. Normal probability plot for outlier detection

In this experiment, we use blood pressure reduction in after taking the drug reading data. The data were collected from Tirunelveli Government health center. For the test purpose we take only 30 samples from the data set.

The normal probability plot fig. 1 representing the data with outlier value deviates from the original data. The plot indicates the outliers point far away from samples. The fig. 2 shows that the outlier values removed by using outlier detection methods and it follow as a normally distributed.

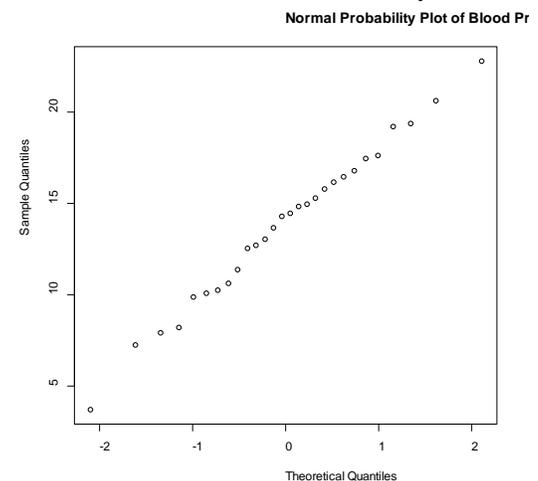


Fig - 3. Outlier removing after Normal probability plot
The various discordancy methods are used in the

experiment to detect the outliers. Table-1 represents the total number of outliers detected by the experiments. In these experiments Grubb's and Dixon tests have given the same results in repeated experiments. Three other methods such as Hampel, Quartile and Generalized ESD test are same results in the experiments. The first two methods are detected 3 outliers for each. But in case the two numbers of outliers only strongly detected. Remaining one outlier is the small number of the observation. The two tests only needed for repeated experiments after detecting outliers.

Table - 1 Total number of outlier detection in the blood pressure after taking drug-reading data

Outlier Tests (Two-tailed test)	Sig- α	Sample Size N=30			
		Number of Outlier Detected			
		Test With outliers	Outliers removing after test		Total
			1st	2nd	
Grubbs Test	Critical value 0.5%	1	1	1	3
Dixon Test		1	1	1	3
Hampel		2	0	0	2
Quartile Method		2	0	0	2
Generalized ESD		2	0	0	2

The other three outlier methods strongly detect outliers in a single experiment. The major outlier is finding easy and quick in the experiments. In these experiments no need critical value for Hampel and Quartile methods and other tests must needed for critical value to detect the outliers.

The R software tested the experimental purpose of the tested methods used for R scripts. Lukasz Komsta (2006) is used for example for the R codes for Dixon, Generalized ESD Test and Grubb's tests.

5 Conclusions

The table-1 describes that outlier values detected by the five-outlier detection methods. Grubbs and Dixon test had low sensitivity for outlier detection in the experiment (every test detected single outlier and find only minimum or maximum value). The other three methods can find single experiment to identify the maximum outliers. The methods Hampel, Quartile and Generalized ESD test can find easy and average detection levels are equal to find the maximum outliers. The result reveals that the three methods (Hampel, Quartile and Generalized ESD) are much better than Grubbs and Dixon test.

Acknowledgement

The first author acknowledges the UGC for awarding the Scheme of Rajiv Gandhi National Fellowship (RGNF) for providing financial support to carry out this

work.

References

1. Anscombe F. J and Irwin Guttman (1960). Rejection of outliers. *Technometrics*, Vol. 2, No. 2, pp. 123-147.
2. Armin Bohrer (2008). One-sided and Two sided Critical Values for Dixon's Outlier Test for Sample Sizes up to n=30, *Economic Quality Control*, Vol. 23, No. 1, 5-13.
3. Barbato, G., Barini, E. M., Genta, G., & Levi, R. (2011). Features and performance of some outlier detection methods, *Journal of Applied Statistics*, 38:10, 2133-2149.
4. Barnett V. and Lewis T. (1978). *Outliers in statistical data*. John Wiley & Sons.
5. Chrominski Kornel, Magdalena TKACZ (2010). Comparison of outlier detection methods in biomedical data, *Journal of Medical Informatics & Technologies* Vol. 16, ISSN 1642-6037.
6. Dixon, W.J. (1950). Analysis of extreme values. *Ann. Math. Stat.* 21, 4, 488-506.
7. Grubbs F. E. (1969), Procedures for detecting outlying Observations in Samples. *American Statistical Association and American Society for Quality. Technometrics*, Vol. 11. No. pp. 1-21.
8. Hawkins D. M. (1980), *Identification of Outliers*, Chapman & Hall, London.
9. Jacqueline S. Galpin and Douglas M. Hawkins (1981). Rejection of a Single Outlier in Two- or Three-Way Layouts, *Technometrics*, Vol. 23, No. 1, pp. 65-70.
10. Laurie Davies and Ursula Gather (1993). The Identification of Multiple Outliers. *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 782- 792.
11. Lukasz Komsta (2006). Processing data for outlier: R News, Vol 6/2.
12. McMillan R. G. (1971). Tests for One or Two Outliers in Normal Samples with Unknown Variance, *Technometrics*, Vol. 13, No. 1, pp. 87-100.
13. Paul S. R, and Karen Y. Fung(1991). A Generalized Extreme Studentized Residual Multiple-Outlier-Detection Procedure in Linear Regression, *Technometrics*, Vol. 33, No. 3, pp. 339-348.
14. Quesenberry C. P. and David H. A. (1961). Some Tests for Outliers, *Biometrika*, Vol. 48, No. 3/4, pp. 379-390.
15. Rorabacher, D.B. (1991). Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon Q Parameter and Related Subrange Ratios at the 95 percent Confidence Level. *Anal. Chem.* 83, 2, 139-146.
16. Rosner Bernard(1975), On the Detection of many outliers, *Technometrics*, Vol. 17, No. 2 (May, 1975), pp. 221-227.
17. Rosner, Bernard (1983), Percentage Points for a Generalized ESD Many-Outlier Procedure,

Technometrics, 25(2), pp. 165-172.

18. Tietjen G. L., Moore R. H., Beckman R. J. (1973). Testing for a Single Outlier in Simple Linear Regression. *Technometrics*, Vol. 15, No. 4, pp. 717-721.
19. Tietjen Gary L. and Moore Roger H. (1972). Some Grubbs-Type Statistics for the Detection of Several Outliers, *Technometrics*, Vol. 14, No. 3, pp. 583-597.
20. UNI 9225(1988), Precision of Test Methods: Determination of Repeatability and Reproducibility by Inter-laboratory Tests, Ente Nazionale Italiano di Unificazione, Milano.

IJSER