Study on Sample Outliers Statistical Test Methods

ZhiYong CHEN
Qinghai Normal University
Xining 810008, China

***Abstract*—The exist ence of outliers reduces the quality of data to a great extent, and makes the corresponding results of data analysis significantly change more, which will lead to people's inaccurate judgments. In this article, we discuss the characteristics and origin of outliers, and introduce some test methods of normal distribution sample: sample quantity method, Dixon test method, Grubbs test method and Nair test method. Finally through example comparison, we have found that the test effect is associated with significant level.**

Keywords- Outliers; Statistical Test; Sample Data; Test Statistics

I. INTRODUCTION

In the information age, the problems humans encounter in the research fields of science and technology are more and more profound and complicated. It is important to effectively collect, analyze and process the experimental data that contains a large amount of information. However in actual research work, for various reasons, there are often some outliers in sample data, which refers to certain individual data that significantly deviates from the rest observation values. These outliers will increase the system error, so some of the classic analytical methods become worthless, and even lead to mistakes in macroeconomic policy-making. There are two scenarios of outliers generation: a) they are extreme manifestation of the random change within the data; b) they are generated by some unrelated factors in experimental process, such as observation deviation and record error, so these outliers are not belonging to the same overall as other data, with relatively greater deviation.

How to detect these outliers? It is a research problem for statisticians and data analyst in practical application. Now in some ways it has been formed some effective methods and new theory, but still lack of effective methods and complete theoretical system. Many statisticians are exploring deep into this field, to find ways to solve these problems.

For most outliers, there is a common statistical test method: in a certain distribution assumption, obtain the test statistics that well reflect the difference between the outliers and most of the rest of the normal data, and do hypothesis test at a given significance level, to examine whether there is a significant difference between the outliers and the main data. For outliers from different causes, we should take a different approach. For example, some outliers are generated from unusual circumstances in data collection process, so we should eliminate these data; however, for some outliers that generated from inherent variability of the data, not statistical error, we should not simple eliminate them, otherwise there may be missing some important hidden information. Complementary, some outliers may also bring us valuable information.

II. OUTLIERS TEST OF NORMAL DISTRIBUTION SAMPLE

Suppose Xobeys normal distribution, then its Distribution function is as Formula (1):

(1)

In Formula (1), $X_1, X_2, \cdots, X_n$are simple random samples with capacity of n; $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are order statistics of the samples; $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$are observed values of sample order statistics. The outliers test methods of normal distribution sample are as follow:

A. Sample Quantile Method

If the variance$\sigma^2$of normal distribution is unknown, we can take sample quantile method. If doubt $X_{(n)}$ ($X_{(1)}$) is outliers, then calculate the upper end (lower end) outliers test statistic S_n (S_1) (as shown in Formula (2)); if cannot ascertain which side outliers are, then calculate the bilateral test statistic MRS (as shown in Formula (3)):

(2)

(3)

In the Formula,

(4)

If the variance is known, similarly can take 1/4 sample quantile method, and use σ as scale parameter to replace ($X_{(n2)}$-$X_{(n1)}$) to construct test statistic. Optimized combination of sample quantile has good Robustness and high estimation efficiency, and it can be used continuously to test multiple outliers.

B. Dixon Test

Dixon proposed Dixon type statistic to test whether the twoextreme value$X_{(1)}$, $X_{(n)}$ are outliers. Statistics to test $X_{(n)}$are as Formula (5):

(5)

We can see from the statistics, when Distinguish$X_{(1)}$ *and* $X_{(n)}$, the method is replace σ with range$X_{(n)}$- $X_{(1)}$ or pseudo range$X_{(n)}$- $X_{(2)}$, $X_{(n)}$- $X_{(3)}$ as graduation, to estimate differences between adjacent data. If the adjacent difference is too large, then it is outliers. This test method can produce good result in the case of one outlier sample. However, if the outliers are more than one and in the same, Dixon test results are not good enough, and vulnerable to the shielding effect.

C. Grubbs Test

If the variance$\sigma^2$of normal distribution is unknown, Grubbs test utilize the modified sample standard deviation S^*_n to replace overall standard deviationσ, and obtain Grubbs statistics as Formula (6):

(6)

In the Formula, . Grubbs test method utilize to estimate overall central location parameters. Its ability to resist outliers contamination is quite poor. When the sample contains multiple outliers in the same side, it is ineffective. Of course, it can be considered to replace in statistics with sample median, but the interference of outliers in still cannot be eliminated.

D. Nair Test

Nair test is an important method to test sample of normal distribution whether there are outliers. This method was introduced into China as national standard. Statistics of Nair test method are as Formula (7):

(7)

We can see that Nair test method belongs to step test, where the known variance of normal population is essential. Statistic $R_{(1)}$ and $R_{(n)}$Respectively describe the Differences between samples extreme value $X_{(1)}$ and $X_{(n)}$ with sample center . However, to estimate overall center position parameters with will bring weak robustness, and weak ability to resist outliers contamination. If there are multiple outliers in the same side, due to the mutual influence of outliers, its statistics will no longer reflect the difference between sample extreme value and center position, so Nair test method cannot ensure good results in this case.

In order to avoid such pitfalls of Nair test statistics and enhance their contamination resistance abilities. Researchers have proposed some improvement programs. One is to utilize sample median$X_{[n/2]+1}$ to estimate overall center position replace parameters, in place of Nair statistic . The modified test statistics proved to be quite robust, shown as Formula (8):

(7)

The modified Nair test method is also step test. These two test methods both need known normal population variance.

III. Outliers Test Example

In order to compare the test methods above, now we take an example. Here is a group of data, which strictly obey normal distribution, as shown in Table 1:

Table 1. Normal Distribution Data Residuals

No.	$X_{i(m)}$	$X_{(i)}$	$X^2_{(i)}$	$X'_{(i)}$	$X'^2_{(i)}$	$X''_{(i)}$	$X''^2_{(i)}$
1	0.270	-0.102	0.010404	-0.040	0.001600	-0.013	0.000169
2	-0.002	-0.040	0.001600	-0.013	0.000169	-0.005	0.000025
3	0.018	-0.013	0.000169	-0.005	0.000025	-0.002	0.000004
4	0.008	-0.005	0.000025	-0.002	0.000004	-0.001	0.000001
5	0.011	-0.002	0.000004	-0.001	0.000001	0.003	0.000009
6	0.028	-0.001	0.000001	0.003	0.000009	0.004	0.000016
7	0.012	0.003	0.000009	0.004	0.000016	0.008	0.000064
8	-0.001	0.004	0.000016	0.008	0.000064	0.010	0.000100
9	-0.102	0.008	0.000064	0.010	0.000100	0.011	0.000121
10	0.003	0.010	0.000100	0.011	0.000121	0.012	0.000144
11	0.018	0.011	0.000121	0.012	0.000144	0.018	0.000324
12	0.004	0.012	0.000144	0.018	0.000324	0.018	0.000324
13	0.010	0.018	0.000324	0.018	0.000324	0.027	0.000729
14	-0.005	0.018	0.000324	0.027	0.000729	0.028	0.000784
15	-0013	0.027	0.000729	0.028	0.000784	-	-
16	-0.040	0.028	0.000784	-	-	-	-

(1) Sample Quantile Method

Utilize 1/4 sample quantile test method to calculate 16 residuals $X_{(i)}$. According to Formula (4), we can obtain: $n_3=4$, $n_4=14$, and then according to Formula (2) obtain S_1 and S_n:

$S_1=4.717>S_1(16, 0.05)=2.051$, $S_{16}=0.652<S_1(16, 0.05)=2.051$

So we can determine $X_{(1)}=0.102$ is outliers. Eliminate $X_{(1)}$ and repeat the same procedure, then we can find that $X'_{(1)}=-0.040$ is also outliers. Eliminate $X'_{(1)}$ and repeat again:

$S^*_{(1)}=1.132< S^*_{(1)}(14, 0.05)=2.455$, $S^*_{(14)}=1.026< S^*_{(14)}(14, 0.05)=2.455$

So $X'_{(1)}$ and $X'_{(14)}$ are not outliers.

(2) Dixon Test

First, we distinguish the maximum residual $X_{(16)}$:

So we think that $X_{(16)}$ is not outliers. Then distinguish the minimum residual $X_{(1)}=-0.102$:

So there is the evident to think $X_{(1)}$ is outliers. Eliminate the first residual $X_{(1)}$, and re-examine the two ends. The result is $X'_{(15)}$ is not outliers. It shows that the test is associated with significant level.

(3) Grubbs Test

By Grubbs test method, we can obtain:

So, $X_{(16)}$ is not outliers. Similarly, we can conclude that $X_{(1)}$ is not outliers neither.

IV. CONCLUSIONS

In this article, we discuss the characteristics and origin of outliers. It is obvious that, no matter in the field of basic research, or quality control, network security, insurance claim, etc., for the data item thatdeviate from the normal patterns of data, to do essential causal analysis is an indispensable job, because it determines whether the application activity can run and produce expected effect or not. Then we introduce some test methods of normal distribution sample: sample quantile method, Dixon test method, Grubbs test method and Nair test method, together with their statistics calculation. Finally, in order to compare the test methods above, we take an example to test. The result shows that the test is associated with significant level. In the practical application, we should better use multiple methods, and conclude by comprehensive analysis.

REFERENCES

[1] Barnett V, Lewis T. Outliers in Statistical Data [M]. New York: John Wiley & Sons, 1994.

[2] Hawkins D M. Identification of outliers[M]. London: Chapman and Hall, 1980.

[3] Hampel F R, Ronchetti E M, Rousseeuw P J, et al. Robust statistics: the approach based on influence functions[M]. John Wiley & Sons, 2011.

[4] Hotta L K, Tsay R S. Outliers in GARCH processes[J]. Economic time series, 2012: 337.

[5] Verma S P, Quiroz-Ruiz A. Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering[J]. Revista Mexicana de Ciencias Geológicas, 2006, 23(2): 133-161.

[6] Tietjen G L, Moore R H. Some Grubbs-type statistics for the detection of several outliers[J]. Technometrics, 1972, 14(3): 583-597.

[7] Liu X, Chen F, Lu C T. On detecting spatial categorical outliers[J]. GeoInformatica, 2014, 18(3): 501-536.