Asian Journal of Probability and Statistics

Volume 26, Issue 5, Page 42-49, 2024; Article no.AJPAS.117004
ISSN: 2582-0230

Methods of Assigning Labels to Detect Outliers

Shashank Kirti [a++*] and Rajeev Pandey [a#]

[a] University of Lucknow, Lucknow, India.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/ajpas/2024/v26i5617

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/117004

Original Research Article

Received: 11/03/2024
Accepted: 14/05/2024
Published: 15/05/2024

Abstract

Outlier identification is a crucial field within data mining that focuses on identifying data points that significantly depart from other patterns in the data. Outlier identification may be categorized into formal and informal procedures. This article discusses informal approaches, sometimes known as labelling methods. The study focused on the analysis of real-time medical data to identify outliers using outlier labelling techniques. Various labelling approaches are used to calculate realistic situations in the dataset. Ultimately, using the anticipated outcomes of the outliers is a more suitable approach for addressing the needs of the larger populations.

Keywords: Outlier detection; informal methods; labeling methods; median absolute deviation.

[++] *Research Scholar;*
[#] *Professor;*
**Corresponding author: Email: Shashank.stats@gmail.com;*

Cite as: *Kirti, S., & Pandey, R. (2024). Methods of Assigning Labels to Detect Outliers. Asian Journal of Probability and Statistics, 26(5), 42–49. https://doi.org/10.9734/ajpas/2024/v26i5617*

1 Introduction

Data mining is the process of extracting concealed predictive insights from extensive datasets. Outlier detection is a potent data mining approach. Various writers have provided various definitions for outliers. Hawkins [1] defines an outlier as "an observation that deviates significantly from other observations, leading to suspicions that it was generated by a different mechanism." Outliers, in the context of data mining and statistics, are also known as discordance, deviants, abnormalities, or anomalies (Aggarwal, 2005). Tietjen and Moore [2] demonstrated the issues associated with the repeated use of a method or technique, as well as the phenomenon of "masking". In their study, Thomas et al. [3] outlined a test protocol that was seen to have an impact on the interlaboratory standard deviations (SDs), rather than the averages. It has been shown that even a modest number of variations in the number of outliers may significantly alter the standard deviation. Rousseeuw and Croux [4] introduced an alternate approach to the Mean Absolute departure technique (MAD), which calculates the average departure from the median. In a recent work, [5] conducted a comparison of several formal approaches for identifying outliers. According to Leys et al. [6] the standard deviation approach was deemed unsuitable for outlier identification and was hence regarded a subpar method. The MAD approach was used as a robust estimator for calculating the median absolute deviation around the median. Obikee et.al. (2014) conducted a comparison of several outlier detection strategies, including the modified Z-Scores method. This method was used in simulation research using a normal distribution produced from a disease group dataset.

Iglewicz and Hoaglin [7] classified the three subsequent matters concerning outliers.

1. Outlier labelling involves identifying data points that are likely to be errors and do not fit well with the distributional model. These points need to be further examined.
2. Outlier accommodation refers to the use of robust statistical methods that are not considerably affected by outliers. If we cannot confirm that suspected outliers are erroneous data, should we modify our statistical methodology to better include these observations?
3. Outlier identification is a procedure used to formally determine if data may be categorized as outliers. This study focuses on the outlier labelling method and the difficulties related to detecting outliers.

2 Methods and Materials

In this paper, several outlier labeling methods are used namely Z-Score, Modified Z-Scores, Median Absolute Deviation (MADe) and Tukey Method (Boxplot).

2.1 Z – Score

The Z-score approach, which use the mean and standard deviation, may be employed to find outliers in the sample.

$$Z_{score}\, i \;=\; \frac{x_i - \bar{x}}{s} \text{ , where } X_i \sim N\,(\mu, \sigma^2) \text{ , and } s \;=\; \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} x_i \;-\; \bar{x}^2} \text{ .}$$

The Z-scores are derived from the assumption that X is normally distributed.
$N(\mu, \sigma^2)$ then Z follows a standard normal distribution, $z = \frac{x-\mu}{\sigma} \sim N(0,1)$ and Z-scores that exceed 3 in absolute value are generally considered as outliers.

2.2 Modified Z-Scores

The previous issue of Z-Scores used two estimators, namely the sample mean (x) and the sample standard deviation (s), which might be influenced by a small number of extreme values or even a single extreme value. In order to address this issue, the updated Z-Scores use the median and the median absolute deviation (MAD) instead of the sample's mean and standard deviation, respectively [7].

$MAD = median\;|x_i - \tilde{x}|$, where $\tilde{x}$ is the sample median.
$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$, where E(MAD) = 0.675σ for large normal data.

Lglewicz and Hoaglin, [7] suggested that observations are labeled outliers when $|M_i| > 3.5$ through the simulation based on pseudo – normal observations for sample size of 10, 20, and 40. The M_i score is effective for normal data in the same way as the Z-score [8,9].

2.3 Median Absolute Deviation (MADe)

The Median Absolute Deviation (MADe) technique is a fundamental robust approach that is very resistant to the influence of extreme values in a dataset. This strategy has resemblance to the SD method. Instead of using the mean and standard deviation, this technique uses the median and MADe. The term is defined in the following manner:

$$2MAD_eMethod : Median \pm 2MAD_e$$
$$3MAD_eMethod : Median \pm 3MAD_e$$

Where $MAD_e = 1.483 \times MAD$ for large normal data and is an estimator of the spread in a data similar to the standard deviation.

$$MAD = median\ |x_i - median(x)|, \qquad \text{i = 1,2,……,n}$$

The MAD is scaled by a factor of 1.483 it also similar to the standard deviation in normal distribution or Absolute Deviation around the Median as stated in the title is a robust measure of central tendency [10,11].

2.4 Tukey's method (box plot)

Tukey's Method, which involves generating a boxplot, is a well-recognized and straightforward graphical method used to visually represent information about continuous univariate data. This includes important statistics like as the median, lower quartile, higher quartile, lower extreme, and upper extreme of a given dataset. This approach utilizes the interquartile range to identify and exclude very high or low values, hence detecting outliers. The formulas are:

$$Low\ Outliers = Q_1 - 1.5\ (Q_3 - Q_1) = Q_1 - 1.5\ IQR$$
$$High\ Outliers = Q_1 + 1.5\ (Q_3 - Q_1) = Q_1 + 1.5\ IQR$$

Where $Q_1 = First\ Quartile$, $Q_3 = Third\ Quartile$, IQR = Interquartile range
These equations provide two distinct values, sometimes referred to as "fences". A barrier that segregates the extreme values from the majority of the data.

3 Computation Results and Discussion

In this study the datasets are taken from Career Institute of Medical Sciences and Hospital, Utter Pradesh. Here for illustration 100 observations are considered for the study. The variable Blood Pressor in mm is taken. It has computed with Z-score and Modified Z-score labeling methods and Median Absolute Deviation (MADe). The given methods are computed by SPSS V29 Software. From several labeling methods we employed Z-score, Modified Z-score and Median Absolute Deviation (MADe) for identifying outliers in the data set [12].

3.1 Z-Score

In Table 1, case 1 with all data has included, it appears that the value 99 and 100 are outliers, yet no observations exceed the absolute value 3. For case 2, the most extreme value 99 and 100 have excluded in the data, 96, 97 and 98 has considered as outliers [13].

3.2 Modified Z-Score

For this method, the computation results are tabulated below and it is compared with Z-score. Table 2 shows that the computed data value of the modified Z-Scores |Mi| > 3.5 in absolute value, out of these, these 5 observations (260, 260, 260, 390, 400) may well be outliers.

Table 1. Computation and masking problem of the Z-Score

Case - 1 $(\bar{x} = 145.86, sd = 45.355)$						Case - 2 $(\bar{x} = 140.78, sd = 28.160)$					
ID	**xi**	**Z - Score**	**ID**	**xi**	**Z - Score**	**ID**	**xi**	**Z - Score**	**ID**	**xi**	**Z - Score**
1	101	-0.9891	51	140	-0.1292	1	101	-1.41247	51	140	-0.02754
2	104	-0.92295	52	140	-0.1292	2	104	-1.30594	52	140	-0.02754
3	108	-0.83476	53	140	-0.1292	3	108	-1.16389	53	140	-0.02754
4	108	-0.83476	54	140	-0.1292	4	108	-1.16389	54	140	-0.02754
5	110	-0.79066	55	140	-0.1292	5	110	-1.09287	55	140	-0.02754
6	110	-0.79066	56	140	-0.1292	6	110	-1.09287	56	140	-0.02754
7	110	-0.79066	57	140	-0.1292	7	110	-1.09287	57	140	-0.02754
8	110	-0.79066	58	140	-0.1292	8	110	-1.09287	58	140	-0.02754
9	110	-0.79066	59	140	-0.1292	9	110	-1.09287	59	140	-0.02754
10	110	-0.79066	60	140	-0.1292	10	110	-1.09287	60	140	-0.02754
11	114	-0.70247	61	142	-0.08511	11	114	-0.95083	61	142	0.04348
12	115	-0.68042	62	142	-0.08511	12	115	-0.91531	62	142	0.04348
13	117	-0.63632	63	142	-0.08511	13	117	-0.84429	63	142	0.04348
14	118	-0.61427	64	144	-0.04101	14	118	-0.80878	64	144	0.1145
15	118	-0.61427	65	145	-0.01896	15	118	-0.80878	65	145	0.15002
16	120	-0.57017	66	145	-0.01896	16	120	-0.73776	66	145	0.15002
17	120	-0.57017	67	148	0.04718	17	120	-0.73776	67	148	0.25655
18	120	-0.57017	68	150	0.09128	18	120	-0.73776	68	150	0.32757
19	120	-0.57017	69	150	0.09128	19	120	-0.73776	69	150	0.32757
20	120	-0.57017	70	150	0.09128	20	120	-0.73776	70	150	0.32757
21	120	-0.57017	71	150	0.09128	21	120	-0.73776	71	150	0.32757
22	120	-0.57017	72	150	0.09128	22	120	-0.73776	72	150	0.32757
23	120	-0.57017	73	150	0.09128	23	120	-0.73776	73	150	0.32757
24	120	-0.57017	74	150	0.09128	24	120	-0.73776	74	150	0.32757
25	120	-0.57017	75	150	0.09128	25	120	-0.73776	75	150	0.32757
26	124	-0.48198	76	150	0.09128	26	124	-0.59572	76	150	0.32757
27	125	-0.45993	77	152	0.13538	27	125	-0.5602	77	152	0.39859
28	125	-0.45993	78	152	0.13538	28	125	-0.5602	78	152	0.39859
29	125	-0.45993	79	160	0.31177	29	125	-0.5602	79	160	0.68268
30	125	-0.45993	80	160	0.31177	30	125	-0.5602	80	160	0.68268
31	125	-0.45993	81	160	0.31177	31	125	-0.5602	81	160	0.68268
32	128	-0.39379	82	160	0.31177	32	128	-0.45367	82	160	0.68268
33	128	-0.39379	83	160	0.31177	33	128	-0.45367	83	160	0.68268
34	130	-0.34969	84	160	0.31177	34	130	-0.38265	84	160	0.68268
35	130	-0.34969	85	160	0.31177	35	130	-0.38265	85	160	0.68268
36	130	-0.34969	86	164	0.39996	36	130	-0.38265	86	164	0.82473
37	130	-0.34969	87	165	0.42201	37	130	-0.38265	87	165	0.86024
38	130	-0.34969	88	170	0.53225	38	130	-0.38265	88	170	1.03779
39	130	-0.34969	89	170	0.53225	39	130	-0.38265	89	170	1.03779
40	130	-0.34969	90	170	0.53225	40	130	-0.38265	90	170	1.03779
41	132	-0.30559	91	170	0.53225	41	132	-0.31163	91	170	1.03779
42	132	-0.30559	92	172	0.57635	42	132	-0.31163	92	172	1.10881
43	132	-0.30559	93	174	0.62045	43	132	-0.31163	93	174	1.17984
44	134	-0.2615	94	178	0.70864	44	134	-0.24061	94	178	1.32188
45	134	-0.2615	95	178	0.70864	45	134	-0.24061	95	178	1.32188
46	134	-0.2615	96	260	2.51662	46	134	-0.24061	96	260	4.23378
47	135	-0.23945	97	260	2.51662	47	135	-0.20509	97	260	4.23378
48	136	-0.2174	98	260	2.51662	48	136	-0.16958	98	260	4.23378
49	138	-0.1733	99	390	5.38293	49	138	-0.09856	99	-	-
50	138	-0.1733	100	400	5.60341	50	138	-0.09856	100	-	-

Table 2. Computation of Z-Scores compared with the modified Z-Scores

ID	xi	Z - Score		Modified	ID	xi	Z - Score		Modified
		Case - 1	Case - 2	Z - Score			Case - 1	Case - 2	Z - Score
1	101	-0.9891	-1.41247	-1.83	51	140	-0.1292	-0.02754	0.05
2	104	-0.92295	-1.30594	-1.69	52	140	-0.1292	-0.02754	0.05
3	108	-0.83476	-1.16389	-1.49	53	140	-0.1292	-0.02754	0.05
4	108	-0.83476	-1.16389	-1.49	54	140	-0.1292	-0.02754	0.05
5	110	-0.79066	-1.09287	-1.4	55	140	-0.1292	-0.02754	0.05
6	110	-0.79066	-1.09287	-1.4	56	140	-0.1292	-0.02754	0.05
7	110	-0.79066	-1.09287	-1.4	57	140	-0.1292	-0.02754	0.05
8	110	-0.79066	-1.09287	-1.4	58	140	-0.1292	-0.02754	0.05
9	110	-0.79066	-1.09287	-1.4	59	140	-0.1292	-0.02754	0.05
10	110	-0.79066	-1.09287	-1.4	60	140	-0.1292	-0.02754	0.05
11	114	-0.70247	-0.95083	-1.2	61	142	-0.08511	0.04348	0.14
12	115	-0.68042	-0.91531	-1.16	62	142	-0.08511	0.04348	0.14
13	117	-0.63632	-0.84429	-1.06	63	142	-0.08511	0.04348	0.14
14	118	-0.61427	-0.80878	-1.01	64	144	-0.04101	0.1145	0.24
15	118	-0.61427	-0.80878	-1.01	65	145	-0.01896	0.15002	0.29
16	120	-0.57017	-0.73776	-0.92	66	145	-0.01896	0.15002	0.29
17	120	-0.57017	-0.73776	-0.92	67	148	0.04718	0.25655	0.43
18	120	-0.57017	-0.73776	-0.92	68	150	0.09128	0.32757	0.53
19	120	-0.57017	-0.73776	-0.92	69	150	0.09128	0.32757	0.53
20	120	-0.57017	-0.73776	-0.92	70	150	0.09128	0.32757	0.53
21	120	-0.57017	-0.73776	-0.92	71	150	0.09128	0.32757	0.53
22	120	-0.57017	-0.73776	-0.92	72	150	0.09128	0.32757	0.53
23	120	-0.57017	-0.73776	-0.92	73	150	0.09128	0.32757	0.53
24	120	-0.57017	-0.73776	-0.92	74	150	0.09128	0.32757	0.53
25	120	-0.57017	-0.73776	-0.92	75	150	0.09128	0.32757	0.53
26	124	-0.48198	-0.59572	-0.72	76	150	0.09128	0.32757	0.53
27	125	-0.45993	-0.5602	-0.67	77	152	0.13538	0.39859	0.63
28	125	-0.45993	-0.5602	-0.67	78	152	0.13538	0.39859	0.63
29	125	-0.45993	-0.5602	-0.67	79	160	0.31177	0.68268	1.01
30	125	-0.45993	-0.5602	-0.67	80	160	0.31177	0.68268	1.01
31	125	-0.45993	-0.5602	-0.67	81	160	0.31177	0.68268	1.01
32	128	-0.39379	-0.45367	-0.53	82	160	0.31177	0.68268	1.01
33	128	-0.39379	-0.45367	-0.53	83	160	0.31177	0.68268	1.01
34	130	-0.34969	-0.38265	-0.43	84	160	0.31177	0.68268	1.01
35	130	-0.34969	-0.38265	-0.43	85	160	0.31177	0.68268	1.01
36	130	-0.34969	-0.38265	-0.43	86	164	0.39996	0.82473	1.2
37	130	-0.34969	-0.38265	-0.43	87	165	0.42201	0.86024	1.25
38	130	-0.34969	-0.38265	-0.43	88	170	0.53225	1.03779	1.49
39	130	-0.34969	-0.38265	-0.43	89	170	0.53225	1.03779	1.49
40	130	-0.34969	-0.38265	-0.43	90	170	0.53225	1.03779	1.49
41	132	-0.30559	-0.31163	-0.34	91	170	0.53225	1.03779	1.49
42	132	-0.30559	-0.31163	-0.34	92	172	0.57635	1.10881	1.59
43	132	-0.30559	-0.31163	-0.34	93	174	0.62045	1.17984	1.69
44	134	-0.2615	-0.24061	-0.24	94	178	0.70864	1.32188	1.88
45	134	-0.2615	-0.24061	-0.24	95	178	0.70864	1.32188	1.88
46	134	-0.2615	-0.24061	-0.24	96	260	2.51662	4.23378	5.83
47	135	-0.23945	-0.20509	-0.19	97	260	2.51662	4.23378	5.83
48	136	-0.2174	-0.16958	-0.14	98	260	2.51662	4.23378	5.83
49	138	-0.1733	-0.09856	-0.05	99	390	5.38293	-	12.09
50	138	-0.1733	-0.09856	-0.05	100	400	5.60341	-	12.57

3.3 Median absolute deviation

This method was computed from the data set results as follows, from the equations Median = 139, MADe = 14. Here the 2 MADe method has identifying 13 outliers which are: 170, 170, 170, 170, 172, 174, 178,178, 260,260,260, 390 and 400. Also, the 3MADe method has identifying 5 outliers which are: 260, 260,260, 390 and 400. In Fig. 1 the extreme values above 250 clearly shown as outliers.

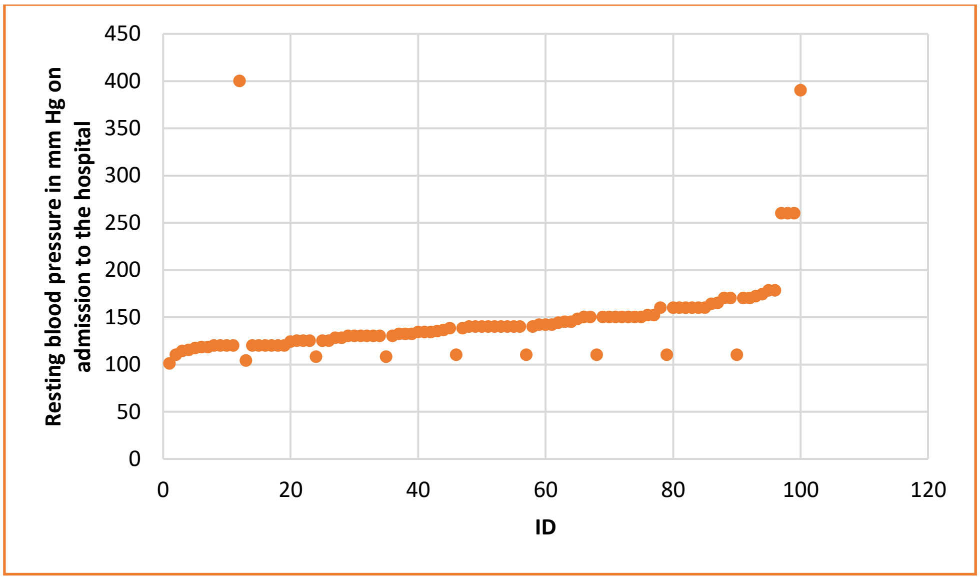

Fig. 1. Dotplot for visualize the data with outliers

Table 3. Tukey method outlier detection using IQR and box whisker plot

Resting blood pressure in mm Hg on admission to the hospital	
Sample Size	100
Mean	145.86
Median	139.00
Std. Deviation	45.355
Skewness	3.829
Std. Error of Skewness	0.241
Kurtosis	17.882
Minimum	101
Maximum	400
Suspected Outliers (Tukey, 1977)	
Outside values	260, 260,260
Far-out values	390, 400

Table 4. Number of outliers detected by different outlier labeling methods

Methods	**Cases**	**Cutoff Value**	**Outliers**
Z-Score	Case - 1	Zi > 3	390, 400
	Case - 2		260, 260, 260
Modified Z-Scores	MAD	|Mi| > 3.5	260, 260, 260, 390, 400
MADe	2 MADe	MAD > 2	170, 170, 170, 170, 172, 174, 178,178, 260,260,260, 390, 400
	3MADe	MAD > 3	260, 260,260, 390, 400
Tukey Method	Outside values	[77.5, 193.5]	260, 260, 260
	Far - out values	[34, 237]	390, 400

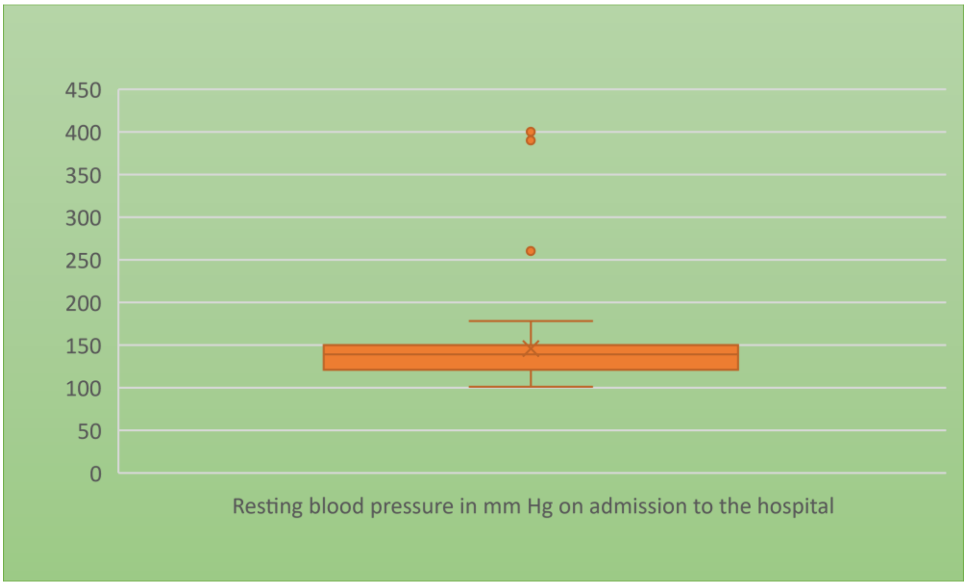

Fig. 2. Box plot

3.4 Tukey method

The Interquartile Range (IQR) is defined as the difference between the first quartile $Q_1 = 121$ and the third quartile $Q_3 = 150$, resulting in an IQR of 29. The coordinates for the inner fences are [77.5, 193.5], whereas the coordinates for the outer fences are [34, 237]. The five extreme values, 260, 260, 260, 390 and 400 have been recognized as possible outliers in this approach and boxplot for the dataset.

4 Conclusion

A statistical analysis was conducted to examine the effectiveness of several outlier labelling approaches - Z-Score, Modified Z-Scores, MADe, and Tukey method in finding and managing outliers. Blood Pressor dataset is used for this study. Intervals are often used to detect potential outliers in outlier labelling techniques that are specifically designed for the normal distribution. The masking issue significantly reduces the detection sensitivity of Z-Scores and Tukey techniques. MADe is a widely used method for identifying outliers in one-dimensional data. It involves labelling every point that is more than two standard deviations apart as a probable outlier. The MADe and Modified Z-scores are used in the MAD technique. The outliers have been discovered as the numbers 260, 260, 260, 390 and 400 nearly five in total. However, all of the approaches are able to identify that the highest distant value is 390 and 400. In the MADe approach, MAD>2 identifies thirteen outliers (170, 170, 170, 170, 172, 174, 178, 178, 260, 260, 260, 260, 390, 400) whereas MAD>3 identifies five outliers (260, 260, 260, 390 and 400). When dealing with a single variable, the Median Absolute Deviation is a very reliable measure of dispersion, particularly when there are outliers present therefore, we suggest using the MADe approach for detecting outliers.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Hawkins DM. Identification of Outliers, Chapman & Hall, London; 1980.

[2] Tietjen and Moore Some Grubbs-Type Statistics for the Detection of Outliers, Technometrics. 1972;14(3):583-597.

[3] Thomas Peter Josef Linsinger, Wolfgang Kandler, Rudolf Krska, Manfred Grasserbauer The influence of different evaluation techniques on the results of interlaboratory comparisons. Springer-Verlag. 1998;3: 322–327.

[4] Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. Journal of the American Statistical Association, Link to the paper. This paper introduces alternatives to MAD and discusses its scale dependence and the need for scaling to achieve consistency with the standard deviation. 1993; 88(424):1273-1283.

[5] Manoj K, Senthamarai Kannan K. Comparison of methods for detecting outliers. International Journal of Scientific & Engineering Research. 2013;4(9):709–714.

[6] Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, Link to the paper. This source discusses the advantages of MAD over standard deviation and highlights its robustness against outliers. 2013;49(4):764-766.

[7] Iglewicz B, Hoaglin DC. How to detect and handle outliers. The ASQC Basic References in Quality Control: Statistical Techniques, This reference manual addresses the efficiency of MAD and its application in outlier detection, including the issue of threshold selection. 1993;16:1-85.

[8] Barbato G, Barini EM, Genta G, Levi R. Features and performance of some outlier detection methods, Journal of Applied Statistics. 2011;38:10:2133-2149.

[9] David M, Rocke David L. Woodruff, Identification of Outliers in Multivariate Data, Journal of the American Statistical Association. 1996;91(435):1047-1061.
Available:http://dx.doi.org/10.2307/2291724

[10] Jacqueline S, Galpin and Douglas M. Hawkins Rejection of a Single Outlier in Two - or Three Way Layouts, Technometrics. 1981;23(1):65-70.

[11] Peter J, Rousseeuw and Christophe Croux. Alternatives to the Median Absolute Deviation, Journal of the American Statistical Association. 1993;88;424:1273 - 1283.

[12] Senthamarai Kannan K, Manoj K, Arumugam S. Labeling Methods for Identifying Outliers. International Journal of Statistics and Systems. 2015;10(2):231-238.

[13] Senthamarai Kannan K, Manoj K. Outlier detection in multivariate data, Applied Mathematical Sciences. 2015;9(45-48):2317-2324

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle5.com/review-history/117004