Aceh Int. J. Sci. Technol., 11(1): 38-45
April, 2022
doi: 10.13170/aijst.11.1.23809

Aceh International Journal of Science and Technology
ISSN: p-2088-9860; e-2503-2398
Journal homepage: http://jurnal.unsyiah.ac.id/aijst

Comparing Outlier Detection Methods using Boxplot Generalized Extreme Studentized Deviate and Sequential Fences

Anwar Fitrianto[1*], Wan Zuki Azman Wan Muhamad[2], Suliana Kriswan[3], Budi Susetyo[1]
[1]Department of Statistics, IPB University, Bogor, Indonesia;
[2]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Malaysia;
[3]Department of Mathematics, Universiti Putra Malaysia, Serdang, Malaysia;
*Corresponding author email: anwarstat@gmail.com

Received : December 12, 2021
Accepted : March 3, 2022
Online : April 30, 2022

Abstract– Outliers identification is essential in data analysis since it can make wrong inferential statistics. This study aimed to compare the performance of Boxplot, Generalized Extreme Studentized Deviate (Generalized ESD), and Sequential Fences method in identifying outliers. A published dataset was used in the study. Based on preliminary outlier identification, the data did not contain outliers. Each outlier detection method's performance was evaluated by contaminating the original data with few outliers. The contaminations were conducted by replacing the two smallest and largest observations with outliers. The analysis was conducted using SAS version 9.2 for both original and contaminated data. We found that Sequential Fences have outstanding performance in identifying outliers compared to Boxplot and Generalized ESD.

Keywords: Outlier, ESD, Boxplot, Performance, Errors, Generalized

Introduction

An outlier is an observation that lies far from the other observation. It is also a strange point in a data set that differs greatly from the others. An outlier may affect numerical measures in a distribution. Outliers may give inaccurate results of analysis, especially for a small sample. Meanwhile, the effect of an outlier on means might seem smaller for the large sample, but increased variance can change the statistical significance of regression estimates.

Many different methods have been used to detect outliers. In general, an outlier is a point in data that lies far away outside the norm for a variable or population (Kuna *et al.*, 2014). It is also known as observations inconsistent with the remainder of the data (Sun et al., 2017; Swersky et al., 2016). Meanwhile, Swersky *et al.* (2016) described an observation that diverges so much from other observations to arouse suspicions as an outlier. Outliers may arise from several different mechanisms or causes. They can be those that appear from the inherent variability of the data and those that appear from the data's errors.

Outlier is unavoidable (Bashiri & Moslemi, 2013; Mahapatra *et al.*, 2020); Bailey, 2018). This is because when using either parametric or nonparametric tests, the existence of outliers can induce the inflation of error rates and substantial distortions of parameter and statistic estimates (Liao *et al.*, 2016; (Parrinello *et al.*, 2016)). Since the outliers may substantially impact most parametric tests, awareness of the outliers has been concentrated on identifying outliers (Schwertman *et al.*, 2004; Yang *et al.*, 2011).

Outliers should be investigated carefully, as their presence and effect will cause misinterpretation in the statistical analysis (Benhadi-Marín, 2018). When a potential outlier is detected, a careful investigation must be performed well as the mistake in calculations or a data coding will lead to an inaccurate conclusion. The suspected outlier may be a good observation that may provide beneficial information (Erdogan, 2014). The aim of the

study is to compare the performance of the Boxplot, Generalized Extreme Studentized Deviate (Generalized ESD), and Sequential Fences method in identifying outliers. The study is expected to provide the best performance in identifying outliers.

Materials and Methods

Data

The data was obtained from a laboratory analysis of calories and sodium content of major hot dog brands available in Moore & McCabe (1989). The researchers for Consumer Reports analyzed three types of hot dogs: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). There are about 54 observations with the variable of hotdog (beef, meat, or poultry) and calories (calories per hot dog). In order to make the comparisons, the data was modified by contaminating with a few outliers to observe the performance of several outlier detection methods in identifying outliers.

Boxplot for Identifying Outlier

Information such as the data's location, spread, skewness, and tails can be obtained easily in a boxplot. This graphical plot is widespread to identify outliers (Dawson, 2011; Walker *et al.*, 2018). This method is very simple and does not use the extreme potential outliers, a point beyond an outer fence, which may disturb the computing of the spread of the data (Schwertman *et al.*, 2004). The fence procedure uses the estimated interquartile range (IQR), often referred to as the *H*-spread. *H*-spread is known as the averages of the first and third quartiles. The inner fences, $f_1$and $f_3 A$nd outer fences, F_1 and F_3, of the Tukey's boxplot are defined as follows:

$$f_1 = q_1 - 1.5(H - spread), \qquad f_3 = q_3 + 1.5(H - spread),$$
$$F_1 = q_1 - 3(H - spread), \qquad F_3 = q_3 + 3(H - spread), \tag{1}$$

where q_1 and q_3 are first and third sample quartile, respectively. Meanwhile, the $H - spread$ equals to $q_3 - q_2$. Observations that fall in the middle of the inner and outer fences are referred to as "outside" outliers, while those that fall under the outer fence, F_1, or beyond the outer fence, F_3, are referred as "far out" outliers. For constructing a boxplot, a five-number summary need to be computed. The five-number summary of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation.

Extreme Studentized Deviate (ESD) for Identifying Outlier

Extreme studentized deviate (ESD) is an outlier identification method in univariate data set (Ryu *et al.*, 2021). It is also called as Grubb test. However, Grubbs method is only able to detect a single outlier. Thus, another procedure was proposed to detect more than one outlier in a sample, such as generalized extreme studentized deviate by Rosner, (1983). Brant, 1990) compared Rosner's (1983) generalized ESD and an extension of Tukey's boxplot, which was known as the boxplot rule. The generalized ESD method is computed as follows:

$$G_i = \frac{\max|x_i - \bar{x}|}{s}, \qquad i = 1, \ldots, n \tag{2}$$

where $\bar{x}$ sample mean and s is sample standard deviation. Observation with highest $|x_i - \bar{x}|$ value needs to be removed. Then, the G_i needs to be recalculated for the remaining $n - 1$ observations. These processes are repeated until mth contaminated observations have been removed. The mean and standard deviation are recalculated sequentially after the observation is deleted with the largest absolute standard deviation.

The values obtained from Equation (2) will then be compared to the critical value, λ, at $\alpha = 0.05$, as shown in Iglewicz & Hoaglin (1993). An observation is considered an outlier when G_i exceeds the critical value, λ. But if G_i does not exceed the critical value, λ, then it is unnecessary to remove the observation and continue the process on the remaining $n - 1$ observations.

Sequential Fences for Identifying Outlier

Many modifications have been built on Tukey's original boxplot (Walker *et al.*, 2018; Babura *et al.*, 2017). Schwertman *et al.* (2004) had come out with new modifications by proposing simple yet more general fences method than Tukey's boxplot model as the outer fences, F_1 and F_3 maybe too conservative for many practitioners, which may tend to neglect many outliers. Sequential Fences enable flexibility in setting the "outside rate," which is the probability that an observation from a non-contaminated normal population is outside a specified limit or boundary. However, the effect of outside rates for small samples showed incorrect identification of observations as an outlier (Hoaglin *et al.*, 1986). It will reduce the probability of misclassifying an observation as an outlier in large data sets and correctly identifying multiple outliers. According to Schwertman *et al.* (2004), calculating the fences in the method of sequential fences is as follows:

$$F_n = q_2 \pm \frac{\text{IQR}}{k_n} t_{\text{df},\alpha}, \tag{3}$$

where

n = the sample size ($20 \leq n \leq 100$),

k_n = the appropriate adjustment relating the expected values of the IQR to the standard deviation for various sample sizes.

df = degrees of freedom which is calculated using the following formula: $7.6809524 + 0.05294156n - 0.00237n^2$.

Schwertman & de Silva (2007) clearly describes plotting sequential fences. The method can formulate a sequence of fences to detect whether more outliers are expected to occur. For contaminated observations, the existence of an outlier is identified by using the "inward" testing. It is done by formulating a series of fences sequentially and checking until the number of observations beyond a particular fence is not more than m. This procedure will also be applied to the other sides of the fences. Therefore, if more than m observations are detected beyond a particular fence, these observations are said to be outliers.

Methodology

The boxplot, generalized ESD, and sequential fences methods will be applied to identify possible outliers in the data. In case of no outliers in the original data, the original data will be contaminated with few outliers. The two highest and the two lowest observations in the original data set were replaced with outlier observations. For the two highest values, observations 22 and 34 were arbitrarily replaced with $\bar{x} + 3\sigma$ and $\bar{x} + 5\sigma$, respectively. Meanwhile, for the two lowest values, observations 44 and 50 were replaced by $\bar{x} - 3\sigma$ and $\bar{x} - 5\sigma$ respectively. The $\bar{x}$ and σ are the mean and median, respectively.

The PROC BOXPLOT in SAS program version 9.2 was used to build the boxplot of the data. After conducting the boxplot analysis, it proceeded to identify outliers using generalized ESD. Then PROC MEANS will be used to compute Equation (2). Observation with highest $|x_i - \bar{x}|$ the value will be removed. The process will be repeated with $n - 1$ observations until removed all contaminated observations. Comparing the value that maximizes $|x_i - \bar{x}|$ with the critical value will let us know whether the observation is an outlier.

Results

Outlier Detection for the Original Data

A boxplot of original data is displayed to identify whether the data has an outlier (Figure 1). There is no observation in both the lower and upper fences of the boxplot. The second method that was employed to identify outliers is the generalized ESD method. Not like in the boxplot method, identifying outliers using the generalized ESD requires us to specify the significance level. In this study, a significance level of 0.05 was specified.

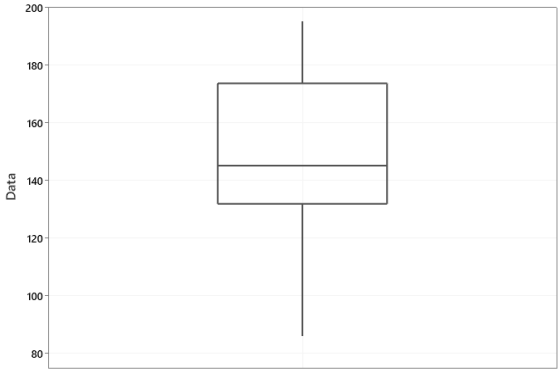

Figure 1. Boxplot of the original data

Table 1 compares the values for each observation that maximize absolute standardized deviation, G_i with the critical values, λ at $\alpha = 0.05$ to test for outliers in the data. The table displays only 4 smallest observations. If the values exceed the critical value, then the observation is an outlier. However, if we look through each value of G_i, none of them exceeds the value of λ. Thus, based on the generalized ESD method, we found no outliers in the original data.

Table 1. Outlier identification using generalized ESD for the original data.

Observation Number, i	Mean for ith observation	Standard deviation	G_i	Critical Value, λ
50	145.444	29.383	2.033	3.159
44	146.566	28.474	2.092	3.151
42	147.712	27.491	1.954	3.144
45	148.765	26.684	1.865	3.136

While for the sequential fences method, after conducting Step 1 of the algorithm, it was obtained $q_1 = 132, q_2 = 145, q_3 = 173$ and $\text{IQR} = 41$ from the original data. At the sample size of $n = 54$, the conversion table from Schwertman & de Silva (2007) gives us $k_{54} = 1.34285$ in Step 2. Since $\gamma = 0.05$ was chosen, the C_m values were obtained from row 4 of the probability table of Schwertman & de Silva (2007) corresponds to the $1 - \gamma$ value is 0.95. The values of C_m are divided by the sample size to compute the α_{nm} values and obtain Table 2.

Table 2. Values for α_{nm} for $n = 54$ and $\gamma = 0.05$.

m	1	2	3	4	5	6
α_{nm}	0.0009	0.0066	0.0151	0.0253	0.03648	0.0484

The procedures are continued by computing t with a degree of freedom, $\text{df} = 29$. The result of this step is displayed in Table 3.

Table 3. t Statistics for $n = 54$ and $\gamma = 0.05$ in Sequential Fences

m	1	2	3	4	5	6
t	-3.4118	-2.6394	-2.2765	-2.0384	-1.8597	-1.7156

Then lower and upper fences are calculated based on Equation (3), and the results are displayed in Table 4.

Table 4. Values for upper fences and lower fences

m	Lower Fences	Upper Fences
1	40.829	249.171
2	64.415	225.585
3	75.494	214.506
4	82.763	207.237
5	88.218	201.782
6	92.620	197.380

Once the lower and upper fences table is obtained, a scatter plot with the y axis against the *x*-axis is developed. The scatter plot values of fences are drawn sequentially from $m = 1$ on both sides. But when there are no observations more than an upper fence or lower than a lower fence on $m = 1$, the sequential procedure is stopped. Based on the plot (Figure 2), it can be observed that there are no observations beyond the first upper fence and first lower fence, $m = 1$. There is no outlier in the original data based on the sequential fences method.

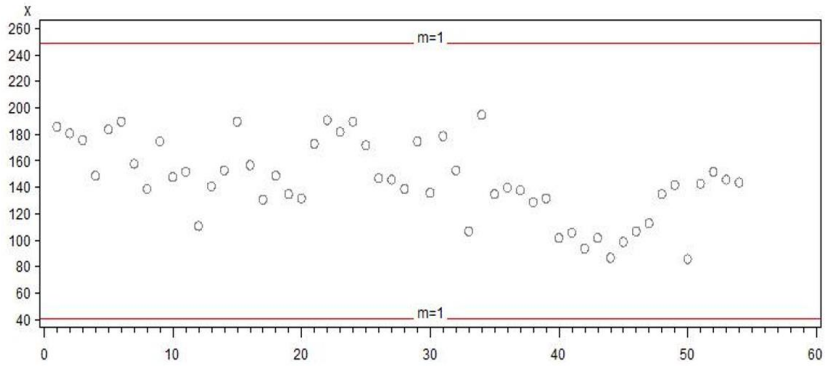

Figure 2. Scatter plot with sequential fences for original data

Outlier Detection for the Contaminated Data

The boxplot of the contaminated data is displayed in Figure 3, in which only observations 34, 44, and 50 are identified as outliers; even four observations were contaminated.

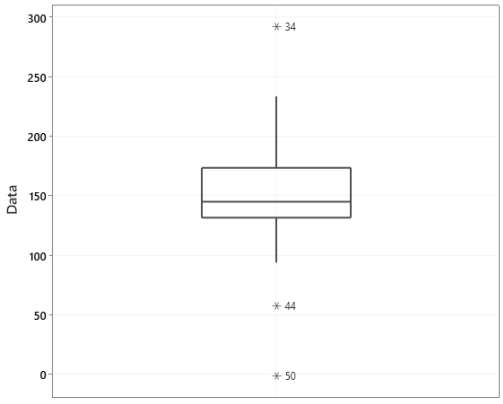

Figure 3. Boxplot of the contaminated data

For the generalized ESD method, the values for each largest absolute standardized deviation, G_i of the contaminated data were recalculated for the two largest and smallest observations. The result of the calculation

is presented in Table 5. The first largest absolute standardized deviation comes from observation 50. Since the value of $G_1 = 3.518 \geq \lambda = 3.159$; hence, based on the generalized ESD, observation 50 is an outlier. The second largest of G_2 corresponds to observation 34 with $G_2 = 3.893 \geq \lambda = 3.151$. Thus observation 34 is also an outlier.

Table 5. Outlier identification using generalized ESD for the contaminated data for the two largest and smallest observations

Observation number	Mean for ith observation	Standard deviation	G_i	Critical Value, λ
50	145.866	41.885	3.518	3.159
34	148.646	36.915	3.893	3.151
44	145.882	31.252	2.835	3.144
22	147.620	28.916	2.973	3.136

Meanwhile, the third largest of $G_3 C$orresponds to observation 44. But, the value of $G_3 = 2.835 \leq \lambda = 3.144$.Then, observation 44 is not an outlier. With the same calculation, observation 22 is also not an outlier. Hence, the generalized ESD method can detect two of the four potential outliers. For the sequential fences method, the same procedure eas run for the contaminated data based on Algorithm 1. Based on Table 6, there are four observations suspected as outliers: the contaminated observations. To confirm whether all suspected outliers are true outliers, comparing the observations with the fences based on the "inward" testing is needed. Based on Figure 4, few observations lie far apart from the other observations that are suspected as outliers.

Table 6. Outlier identification outliers using the sequential fences method for the contaminated data

Suspected Outliers	Contaminated Values	Original Values	Lower Fence	Upper Fence
34	292.361	195	—	$m = 1$ (249.171)
22	233.595	191	—	$m = 2$ (225.585)
44	57.294	87	$m = 2$ (64.415)	—
50	-1.473	86	$m = 1$ (40.829)	—

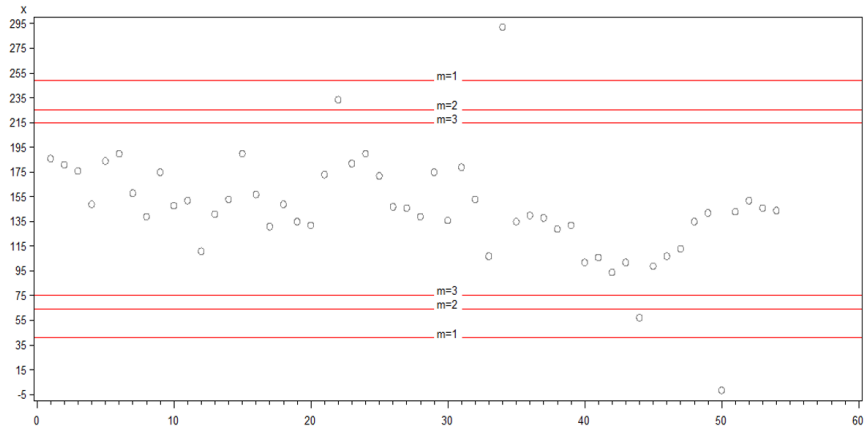

Figure 4. Scatter plot with sequential fences for the contaminated data

In Figure 4, observation 34 is located beyond the first upper fence, $m = 1$. By using suspected "inward" testing, if the number of observations $\geq m$, these observations are outliers. Thus, observation 34 is an outlier.

The next potential outlier is observation 22, which is located between the second upper fence, $m = 2$, and the first upper fence $m = 1$. Hence, observation 22 is also an outlier. Since at the third upper fence, $m = 3$ with the value of the fence is 214.506, there are no more suspected outliers, the procedure of constructing the next series of upper fences is stopped.

An observation is located below the fence for the first lower fence $m = 1$, which is observation 50. Thus it is an outlier. For the next lowest observation, 44 is also an outlier since the observation is located between the second lower fence, $m = 2$ and the first lower fence $m = 1$. Meanwhile, there are no more potential outliers for $m = 3$ with the value of the lower fence being 75.494, so the process of constructing the next series of lower fences is ended.

Discussion

One of the important stages before conducting data analysis is to explore the data. Through data exploration, researchers will know the behavior and characteristics of the data to be analyzed so that later they will be able to determine a more appropriate analysis method. At the data exploration stage, one thing that is important to do is whether the data to be analyzed contains outliers. As we commonly know, the presence of outliers in the data will have many consequences and result in inaccurate or valid analysis results.

Many methods have been developed to identify outliers in univariate data. The methods that have been developed take into account the distribution of the data. What has been done in this research is to compare several outlier detection methods, including Boxplot, Generalized Extreme Studentized Deviate, and Sequential Fences. Boxplot has a good enough performance to detect one or more outliers. Unfortunately, the performance of this method is only good for symmetrical data. Generalized Extreme Studentized Deviate has a fairly good performance on symmetrically distributed data, but this method is only good for detecting a single outlier. Meanwhile, Sequential Fence is a method that can be used to identify one or more outliers for both symmetrical and skewed distributed data. This is because the formula used has taken into account the validity of the data.

Conclusion

The overall results show that boxplot, generalized ESD, and sequential fences methods could not identify any outliers in the original data. After contaminating the original data with four outliers, not all methods can identify all of them as outliers. Both boxplot and generalized ESD methods could only identify three potential outliers, few of them were different observations. The boxplot could not identify observation 22 as an outlier, and the generalized ESD method could not identify observation 44 as an outlier. Meanwhile, the sequential fences method can identify all four contaminated observations as outliers. The sequential fences method for contaminated data shows that this method is very sensitive to the presence of outliers as it can detect all the contaminated observations correctly. Therefore based on these results, the Sequential Fences method was much more effective than the Boxplot and Generalized ESD method at detecting multiple outliers.

References

Babura, B.I., Adam, M.B. Fitrianto, A. and Rahim, A.S.A. 2017. Modified boxplot for extreme data. AIP Conference Proceedings, 1842(1): 030034.

Bailey, D. 2018. Why OUTLIERS are good for science. Significance, 15(1): 14–19.

Bashiri, M. and Moslemi, A. 2013. Simultaneous robust estimation of multi-response surfaces in the presence of outliers. Journal of Industrial Engineering International, 9(1): 1–12.

Benhadi-Marín, J. 2018. A conceptual framework to deal with outliers in ecology. Biodiversity and Conservation, 27(12): 3295–3300.

Brant, R. 1990. Comparing classical and resistant outlier rules. Journal of the American Statistical Association, 85(412): 1083–1090.

Dawson, R. 2011. How significant is a boxplot outlier? Journal of Statistics Education, 19(2):1-13.

Erdogan, B. 2014. An outlier detection method in geodetic networks based on the original observations. Boletim de Ciências Geodésicas, 20: 578–589.

Hoaglin, D.C., Iglewicz, B. and Tukey, J.W. 1986. Performance of some resistant rules for outlier labeling. Journal of the American Statistical Association, 81(396): 991–999.

Iglewicz, B., and Hoaglin, D.C. 1993. How to detect and handle outliers. Asq Press.

Kuna, H.D., García-Martinez, R. and Villatoro, F.R. 2014. Outlier detection in audit logs for application systems. Information Systems, 44: 22–33.

Liao, H., Li, Y. and Brooks, G. 2016. Outlier impact and accommodation methods: Multiple comparisons of Type I error rates. Journal of Modern Applied Statistical Methods, 15(1): 23.

Mahapatra, A.P.K., Nanda, A. Mohapatra, B.B., Padhy, A.K. and Padhy, I. 2020. Concept of outlier study: The management of outlier handling with significance in Inclusive education setting. Asian Research Journal of Mathematics: 7–25.

Moore, D.S., McCabe, G.P. 1989. Introduction to the practice of statistics WH Freeman and Company. New York.

Parrinello, C.M., Grams, M.E. Sang, Y. Couper, D. Wruck, L.M. Li, D. Eckfeldt, J.H. Selvin, E. and Coresh, J. 2016. Iterative outlier removal: a method for identifying outliers in laboratory recalibration studies. Clinical Chemistry, 62(7): 966–972.

Rosner, B. 1983. Percentage points for a generalized ESD many-outlier procedure. Technometrics, 25(2): 165–172.

Ryu, M., Lee, G. and Lee, K. 2021. Online sequential extreme studentized deviate tests for anomaly detection in streaming data with varying patterns. Cluster Computing: 1–13.

Schwertman, N.C. and de Silva, R. 2007. Identifying outliers with sequential fences. Computational Statistics & Data Analysis, 51(8): 3800–3810.

Schwertman, N.C., Owens, M.A. and Adnan, R. 2004. A simple more general boxplot method for identifying outliers. Computational Statistics & Data Analysis, 47(1): 165–174.

Sun, D., Zhao, H. Yue, H. Zhao, M. Cheng, S. and Han, W. 2017. ST TD outlier detection. IET Intelligent Transport Systems, 11(4): 203–211.

Swersky, L., Marques, H.O. Sander, J. Campello, R.J.G.B. and Zimek, A. 2016. On the evaluation of outlier detection and one-class classification methods. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA): 1–10.

Walker, M.L., Dovoedo, Y.H. Chakraborti, S. and Hilton, C.W. 2018. An improved boxplot for univariate data. The American Statistician, 72(4): 348–353.

Yang, J., Xie, M. and Goh, T.N. 2011. Outlier identification and robust parameter estimation in a zero-inflated Poisson model. Journal of Applied Statistics, 38(2): 421–430.