ASM Sc. J., 12, Special Issue 5, 2019 for ICoAIMS2019, 107-115

Adjusted Sequential Fences for Detecting Univariate Outliers in Skewed Distributions

H.S. Wong[1,2*] and Anwar Fitrianto[3]

[1]*Department of Business Administration, School of Business and Management, University College of Technology Sarawak, Sibu, Sarawak, Malaysia.*

[2]*Laboratory of Computational Statistics and Operation Research, Institute for Mathematical Research, Universiti Putra Malaysia, Serdang, Selangor, Malaysia*

[3]*Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Indonesia*

Boxplot is a simple graphical method for identifying of outliers. It indicates the location, spread and skewness of the data. The common fences procedures are too liberal and conservative. When the data are skewed, many observations cross the whisker and are erroneously declared as outliers. Sequential fences is a multiple outliers' detection method based on specified probability. However, this method works nicely for detection of outliers for symmetric and fairly symmetric distributions. It is unable to detect outliers in skewed distributions but misclassify some observations as outliers. This paper presents a solution to address this problem and proposes an adjusted sequential approach to detect outliers in skewed distributions. Simulation technique has been applied by constructing fences for different sample sizes from chi square, gamma, weibull, normal and lognormal distributions to check the performance of the method. Several real problems have also been used to show the benefits of this new adjusted approach. The results showed that the new approach performed better in reducing the swamping rate and increasing the accuracy than the standard boxplot and ordinary sequential fences.

Keywords: univariate data; boxplot; sequential fences; bowley coefficient of skewness; generalized extreme value distribution; outliers

I. INTRODUCTION

Outliers are observations that lies an abnormal distance from the majority of the data and distinct from overall pattern of a distribution. The existence of outliers in data set can bring some consequences to statistical data analysis and might further affect decision making. Thus, it is vital for data analysts to inspect for the outliers in the data before conduct the data analysis. One of the common graphical methods is boxplot which was introduced by Tukey (Tukey, 1977). The method involves the interquartile range or known as *H*-spread which is the difference between third quartile, Q_3, and first quartile, Q_1, or $IQR = Q_3 - Q_1$. In addition, the boxplot gives insight of the shape of univariate data distribution, namely, minimum, maximum, first quartiles, third quartiles and median (Tukey, 1977). Extreme outliers are the observations that lie above three times IQR or more above the Q_3 or three times IQR or more below the Q_1, $[Q_1 - 3IQR,\ Q3 + 3IQR]$. Observations that fall 1.5 times interquartile range apart from the first and third quartile, $[Q_1 - 1.5IQR,\ Q3 + 1.5IQR]$ are known as suspected outliers.

The constant of fences which is fixed as 1.5 is considered too liberal for detecting outliers in random normally distribution data (Hoaglin *et al.*, 1986). In addition, the utilization of first quartile and third quartile in the formulation of Tukey's boxplot are considered too conservative and may lead to some outlying data are neglected (Schwertman *et al.*, 2004). Furthermore, the classical boxplot method use single criterion to identify outliers regardless of the different sample sizes of data (Schwertman *et al.*, 2004). This may resulted in some outlying data in small sample will be unsuccessfully to be

*Corresponding author's e-mail: huishein.wong@gmail.com

detected while the observation in large sample will be incorrectly classified as an outlier.

In order to overcome this issue, Schewertman and de Silva (2007) proposed an approach to adjust the fences for various sample sizes of data. The sequential fences were constructed based on different sample sizes using the Poisson distribution to reduce the tail probabilities. Outstanding performance can be seen when this method is applied on normal or approximately normal distributions but problems arise with the construction of sequential fences in skewed distributions. In symmetric distribution, many uncontaminated data will fall beyond the fence at heavy tail whereas regular data from light tailed distributions will scarcely exceed the outlier cut off value. Identical issue happens in skewed distributions when the thickness of tails of data distributions are in different directions. For instance, chi-square distribution is positively skewed, so the observations have low probability to exceed the lower fence whereas the regular observation exceed the upper fence with high probability. Consequently, this outlier labelling approach cannot work effectively when data are skewed.

In the past literatures, numerous outlier identification methods have been proposed. Most of these methods work efficiently under assumption that data come from a symmetric distribution but fail in skewed distributions (Dovoedo & Chakraborti, 2012). For instance, Tukey's boxplot is the most common outlier identification technique but this method tends to give false outcomes when the distributions are being considered as skewed (Hubert & Vandervieren, 2008). Myriad of methods have been suggested to increase the precision in detecting of outliers in skewed distribution data. Meanwhile, the shape of the underlying distribution is vital in labelling the outliers (Dovoedo & Chakraborti, 2012). Kimber (1990) introduced the fences procedure of a constant multiple of the lower semi-interquartile range, $SIQR_L = Q_2 - Q_1$, and the upper semi-interquartile range, $SIQR_U = Q_3 - Q_2$. Carling (2000) made modification to the boxplot method as the distance from median instead of first and third quartiles and constant 1.5 was substituted by a size-dependent formula to regulate the fences of boxplot. Another simple outlier identification procedure which was proposed by Banerjee and Iglewicz (2007) for random sample in case of univariate data according to the popular boxplot outlier labeling rule. Huber and Vandervieren (2008) proposed an adjustment to the Tukey's boxplot by introducing a robust measure of skewness, namely medcouple (MC) which was introduced by Aucremanne et al. (2004). Constant 3.5 and 4 were used on the both sides of fences and thus this lead to the changes to the constant location with respect to the sign of the MC. Dovoedo and Chakraborti (2015) remodeled the traditional boxplot by formulating the fences with multiples of the lower and upper semi-interquartile range measured from the sample median. The outside rate per sample is the probability that minimum one data is misclassified as an outlier which is derived for the location-scale distributions family which is utilized to determine to the constants of fences. Adil and Irshad (2015) and Babangida et al. (2017) modified the approach which was proposed by Hubert and Vandervieren (2008) by suggesting the use of classical skewness measurement instead of constants. This modified method generates broader fences especially for those data which come from skewed distributions and thus this overcome the trouble in using these constants as power of exponential times MC.

Babangida et al. (2017) improved the boxplot for extreme data by adjusting fences constant using a robust skewness measure, namely Bowley coefficient. This modified boxplot able to identify unusual data and solve the major restriction to outlier detection in different distributions for generalization aim. Besides, this approach is capable to show the location parameter region of Gumbel or Generalized Extreme Value Distribution (GEV) fitted extreme data. Extreme data are history of events that are more extraordinary than any that have been noticed. The extreme value can be low extreme or high extreme which are known as minima or maxima respectively. The recent growth in global warming which mark a substantial interest in financial crisis such as volatility in financial area and environment issue that resulted in a universal interest in modeling and forecasting of an extreme events. Exploratory data analysis towards extreme data is not highly given attention in spite of its importance toward confirmatory data analysis.

Hence, these provide us an idea of incorporating robust measure of skewness in regulating the sequential fences. Robust measure of skewness can display the pattern of the distribution curve whether is symmetric or being distorted with negative or positive skewness. The shape and asymmetry of a distribution can be assessed by interpreting the skewness of distribution. For any symmetric distribution, the skewness is approaching zero and the tails on either side of the curve reflect as mirror images of each other. For unimodal distribution, the value of skewness indicate the direction of the tail in the distribution. Positive skewness and negative skewness indicate the tail is on right

and left respectively. Similarly, positively skewed distribution display larger right tail compared to the left tail.

In this research, an adjustment to the sequential fences based on the skewness of the distribution to improve the outlier identification in skewed distributions is presented. A technique of incorporating the Bowley coefficient, a robust skewness in the formulation of sequential fences is proposed in order to detect the multiple outliers in symmetric and skewed distributions. This research focus on detecting extreme data in right skewed underlying distribution univariate data. The objective of this study is to improve sequential fences which is robust to outlier and applicable in skewed distributions. Thus, a modified sequential fences method is proposed by considering the skewness of the underlying distributions in order to enhance the performance of the coverage of sequential fences in outlier detection.

A. Tukey Boxplot

One of the standard and common graphical method in detecting outliers is boxplot which is proposed by Tukey (1977). The lower and upper fences of boxplot is defined as

$$[Q_1 - 1.5\,IQR\,;\; Q_3 + 1.5\,IQR] \qquad (1)$$

where Q_1 and Q_3 are first and third quartile respectively while the interquartile range is denoted by $IQR = Q_3 - Q_1$. Any observation that fall beyond the fences is considered as potential outlier.

B. Sequential Fences

Since our purpose is based on checking for outliers by adjusting sequential fences which was proposed by Schewertman and de Silva (2007), we first describe how the outliers are detected in this method. Schewertman and de Silva (2007) proposed a graphical approach to construct the outlier identification fences with sequential procedure for detecting multiple outliers. The fences are defined as

$$F_{n,m} = q_2 \pm \frac{t_{df,\alpha_{nm}}}{k_n} IQR \qquad (2)$$

where IQR is the interquartile range which is the difference between third quartile and first quartile. The k_n is the value that are the appropriate adjustment regarding to the expected value of the IQR to the standard deviation for different sample sizes. The conversion coefficients for IQR to the standard deviation can be referred to Schwertman and Silva (2007). Due to the sensitivity of the fourth moment of a distribution toward the tail thickness, thus probabilities associated with the tails. Hence, the degree of freedom for the approximating t distribution is determined based on the fourth moment of the t distribution to the fourth empirical moment acquired from 10 000 simulations of the IQR using sample sizes between 20 to 100 from a standard normal distribution. Based on the sample size between 20 and 100, the least squares quadratic equation for obtaining the degree of freedom which is approaching t distribution is

$$df = 7.6809524 + 0.5294156n - 0.00237n^2. \qquad (3)$$

The sample sizes were adjusted for the construction of sequential fences using the Poisson distribution in order to reduce the tail probabilities which is similar to the adjustment done in Davies and Gather (1993) and Gather and Becker (1997). This method increased the accuracy in identifying the outliers, reduced the swamping effect and lower the chance of mislabelling an uncontaminated observations as outliers in large sample size data sets. By using Poisson model, m contaminated observations can be checked. Let X be the number of observations outside the computed fences. The Poison model is written as

$$P(X < m) = e^{-n\alpha_{nm}}\left(1 + n\alpha_{nm} + \frac{(n\alpha_{nm})^2}{2!} + \cdots + \frac{(n\alpha_{nm})^{m-1}}{(m-1)!}\right) = 1 - \gamma, \qquad (4)$$

where m is the number of contaminated observations, n is sample size, and γ is probability that m or more observations beyond the fences are uncontaminated. The solution of Equation (4) for $n\alpha_{nm}$ value can be obtained from Schwertman and Silva (2007). The constant is divided by the sample size in order to obtain the value of α_{nm}. The probability of at most $(m-1)$ uncontaminated observations beyond the constructed fences is $1-\gamma$. The identification of outlier has to be checked until the $(m+1-th)$ fence which has only m observations beyond the fence. For example, if there is no observation detected beyond the first fence $(m=1)$, this meaning that there is no outlier present, then the investigation procedure stop. Otherwise, repeat the outlier checking procedures by constructing the fences continuously until there is no additional observations fall outside the next fence.

II. ADJUSTMENT OF SEQUENTIAL FENCES FOR ASYMMETRIC DISTRIBUTIONS

The adjusted sequential fences with the combination of a robust skewness is proposed.

A. Generalized Extreme Value Distribution (GEV)

Sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ exist, as $n \to \infty$, such that $Pr\{(M_n - b_n)/a_n \leq x\} \to G(x)$ where M_n is a block maximum of n observations and G is a non-degenerate distribution function, then G is a member of the GEV family. The GEV includes three types of extremal distribution which are determined by the shape parameter ξ in the distribution function

$$G(x) = exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad (5)$$

which is defined on $\left\{x: 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$, where $\sigma > 0$ and $\mu, \xi \in \Re$.

Hence, $G(x)$ is said to be Weibull if $\xi < 0$, Gumbel when $\xi = 0$, and Freichet if $\xi > 0$ with re-expression of the limiting distribution as (Coles *et al.*, 2001; Thas *et al.*, 1997)

$$G(x) = exp\left\{-exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}. \quad (6)$$

B. Robust Measure of Skewness

Bowley (1920) proposed a robust coefficient of skewness and is denoted as

$$\zeta = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}, \quad (7)$$

where Q_i's are the i^{th} quartiles of x_i (Kim & White, 2004).

Any reasonable coefficient of skewness $\zeta(x_i)$ should satisfy the following four characteristics which are categorized by Groneveld and Meeden (1984) whereby (i) for any $a \in (0, \infty)$ and $b \in (-\infty, \infty), \zeta(x_i) = \zeta(ax_i + b)$; (ii) for symmetric distribution, then $\zeta(x_i) = 0$; (iii) if $y_i = -x_i$, then $-\zeta(x_i) = \zeta(-x_i)$; (iv) if F and G are cumulative distribution functions of x_i and y_i, and $F < G$ (component wise), then $\zeta(x_i) \leq \zeta(y_i)$.

C. Incorporating Bowley Coefficient of Skewness into Sequential Fences

In order to construct sequential fences for the skewed data, we incorporate Bowley coefficient of skewness ζ into the sequential fences. The coefficient of IQR is substituted by some functions related to ζ, such that $f_l(\zeta)$ and $f_l(\zeta)$. The proposed fences is defined by

$$[Q_2 - f_l(\zeta)IQR, Q_2 + f_u(\zeta)IQR]. \quad (8)$$

Let $f_l(0) = f_u(0)$ to similar to the standard sequential fences at symmetric distribution. For asymmetric distributions, $f_l(\zeta)$ and $f_u(\zeta)$ can be utilised to adjust the fences according to the skewness of the distribution. Comparison of different functions has been made.

The following two functions were considered:

(1) Linear function is described as

$$f_l(\zeta) = \frac{t_{df,\alpha_{nm}}}{k_n} + a\zeta, \quad (9)$$
$$f_u(\zeta) = \frac{t_{df,\alpha_{nm}}}{k_n} + b\zeta.$$

(2) Exponential function is written as

$$f_l(\zeta) = \frac{t_{df,\alpha_{nm}}}{k_n} e^{c\zeta}, \quad (10)$$
$$f_u(\zeta) = \frac{t_{df,\alpha_{nm}}}{k_n} e^{d\zeta},$$

where $a, b, c, d \in \Re$.

D. Determination of the Functions Parameters

In order to determine the functions parameters a, b, c, d, the expected percentage of outliers was set to 0.7% according to the rule of Tukey boxplot of the Gaussian distribution. According to the rule, the fences cut off should satisfy

$$Q_2 - f_l(\zeta)IQR \approx Q_\alpha, \quad (11)$$
$$Q_2 + f_u(\zeta)IQR \approx Q_\beta,$$

where Q_α and Q_β denotes the α^{th} and β^{th} quantile of the distribution respectively, and $\alpha = \frac{0.7\%}{2} = 0.0035$ and $\beta = 1 - \alpha = 0.9965$.

Thus, the functions (9), (10) and (11) can be combined and written as

$$\begin{cases} \frac{Q_2 - Q_\alpha}{IQR} - \frac{t_{df,\alpha_{nm}}}{k_n} \approx a\zeta, \\ \frac{Q_\beta - Q_2}{IQR} - \frac{t_{df,\alpha_{nm}}}{k_n} \approx b\zeta, \end{cases} \quad (12)$$

$$\begin{cases} ln\left(\frac{k_n}{t_{df,\alpha_{nm}}} \cdot \frac{Q_2 - Q_\alpha}{IQR}\right) \approx c\zeta, \\ ln\left(\frac{k_n}{t_{df,\alpha_{nm}}} \cdot \frac{Q_\beta - Q_2}{IQR}\right) \approx d\zeta, \end{cases} \quad (13)$$

The parameter values of the two functions could be derived using median resistant line fit. In order to simulate the quantities in both sides of Equations (12) and (13), GEV distribution is generated with a sample size of 100, location and scale parameters are set to 0 and 1 respectively with various shape parameter ξ within the interval of -0.8 and 10 with an increment of 0.01. The setting of the interval for ξ is based on the Extreme Value Theory practical recommendation in the past literature. For non-typical cases for the estimation of parameters of GEV distribution has been explained by Coles et al. (2001) such that the GEV distribution will possess short upper tail when $\xi \leq -0.5$. Hence, simulation can be generated without given weight to the non-typical cases. This evolution shows to be reasonable to consider ξ =-0.8 as a lower limit for ξ in the simulation study. In spite of that, the setting of 10 as the upper limit for ξ due to the value Bowley coefficient of skewness is very close to 1.0 when $\xi = 10$ (Babura *et al.*, 2017). In the simulation, thus we fixed the sample size as 100 and computed the first fence, $m = 1$ with $\gamma = 0.0035$ such that there is 0.9965 probability of no uncontaminated observations exceed the fence and only 0.0035 probability that an observation exceed the fence is uncontaminated. The simulated samples with the constant location, scale and shape was repeated for 5000 replications. The average ζ and corresponding average quantities in Equations (12) and (13) from each sample were obtained as finite sample estimate of the quantities. The returned simulated quantities is sorted whereby those with nonnegative skewness ζ were chosen since the assumption is that the estimated parameters can be easily exchanged with the coefficient of function with negative skewness. The returned quantities were further sorted into two categories whereby $\zeta \leq 0.9$ is for fitting the lower fence function whereas $0 \leq \zeta \leq 0.6$ is for the upper fence function. In order to compare the behaviour of the two functions, these quantities were applied to fit lower and upper fences. The results show that the exponential function fit the samples better than linear function. Thus, the exponential function was chosen to conduct the adjusted sequential fences.

E. Proposed Adjusted Sequential Fences

In order to simplify the practical implementation of proposed approach, the estimated parameters of the exponential function in Equation (13) obtained are rounded off to the nearest smaller integer. The estimated values $c = -4.35$ and $d = 6.68$ are rounded off to $c = -4$ and $d = 6$. Thus, the proposed adjusted sequential fences can be written as

when $\zeta \geq 0$,

$$\begin{cases} Q_2 - \frac{t_{df,\alpha_{nm}}}{k_n} e^{-4\zeta} IQR, \\ Q_2 + \frac{t_{df,\alpha_{nm}}}{k_n} e^{6\zeta} IQR, \end{cases} \tag{14}$$

when $\zeta < 0$,

$$\begin{cases} Q_2 - \frac{t_{df,\alpha_{nm}}}{k_n} e^{6\zeta} IQR, \\ Q_2 + \frac{t_{df,\alpha_{nm}}}{k_n} e^{-4\zeta} IQR. \end{cases} \tag{15}$$

III. SIMULATION STUDY

The proficiency comparisons between adjusted sequential fences against standard Tukey boxplot and standard sequential fences are presented in this section due to the interest to know the effectiveness of outlier identification of these three methods.

A simulation study has been carried out for the verification of the claim in varying levels of asymmetric distributions. From the stimulation study, it is desired to know the possible outlier percentage of the methods detected in uncontaminated data and also which method is more robust according to the two determinants, skewness of the underlying distributions and sample size. Besides standard normal distribution, $N(0,1)$, with a mean of 0 and a standard deviation of 1, and the lognormal which represent symmetric and fairly symmetric distributions, χ^2, gamma and weibull distributions are chosen as the representative of asymmetric distribution for the purpose of testing the efficiency of outlier detection in skewed data. Sample size has been taken equivalent to 20, 50, and 100 in all the distributions. The study has been done on the lognormal with a mean of 5 and a standard deviation of 0.6 while for the χ^2 with 2 and 4 degree of freedom. For the gamma distribution, the shape parameter is 0.5 and the scale parameter is 1, and for the weibull distribution are with a shape parameter of 1 and scale parameter of 2. For each sample, 10 000 replications have been conducted in the SAS software.

For the implementation of sequential fences, in order to identify the first potential outlier, we compute the first fence $m = 1$ with $\gamma = 0.05$ whereby there is 0.95

probability of no good data identified as outliers and only 0.05 probability that a data beyond the fence is uncontaminated. By calculating the mean percentage of observations flagged as potential outliers (MP), the comparison results between the three methods are displayed in Table 1. The MP is defined as

$$\text{MP} = \frac{\text{number of good data identified as outliers}}{\text{total number of good data}} \times 100. \quad (16)$$

Table 1. Mean percentage of observations flagged as Potential Outliers in uncontaminated data.

Distribution	n	Tukey	Standard	Adjusted
Normal (0,1)	20	0.0568	0.0607	0.0391
	50	0.0391	0.0835	0.0104
	100	0.0272	0.0416	0.0183
Log Normal (5, 0.6)	20	0.5034	0.4288	0.3844
	50	0.5319	0.4374	0.2938
	100	0.7168	0.7082	0.2701
Chi-Squared (2)	20	0.6140	0.5157	0.1681
	50	0.7463	0.7304	0.3189
	100	0.8351	0.8280	0.4623
Chi-Squared (4)	20	0.4596	0.3590	0.0873
	50	0.6833	0.5733	0.1611
	100	0.8330	0.8487	0.4081
Gamma (0.5,1)	20	0.7285	0.6599	0.3148
	50	0.9278	0.8811	0.4165
	100	0.9947	0.9893	0.5125
Weibull (1,2)	20	0.6196	0.5214	0.1768
	50	0.9503	0.9191	0.2877
	100	0.9599	0.9273	0.5050

For the symmetric or fairly symmetric distributions, the performance of three methods in misdetection minimum one observation as outlier in normal and lognormal distributions are very similarly or slightly better. However, in asymmetric case, adjusted sequential fences is outperforming compared to the other two methods. For χ^2 distribution with degree of freedom 2 and 4, the proportion of mistakenly labelling at least one observation as an outlier for proposed adjustment method is recorded much lower than Tukey boxplot and standard sequential fences. As the sample size grow, the proportion of falsely labelling at least one observation as an outlier increases. In high skewed data such as gamma and weibull distribution, for Tukey boxplot and standard sequential fences, the proportion of incorrectly classifying at least one observation as an outlier is approaching 1 as sample size is increased to 100. However, better performance can be seen for adjusted sequential fences with lower proportion of misidentifying regular observations as outlying data when simulating no outlier.

IV. NUMERICAL EXAMPLES USING REAL DATA

Two real data examples are presented to illustrate the usefulness of the proposed adjusted sequential fences in identifying multiple outliers and comparison has been made with both Tukey boxplot and standard sequential fences. For both examples, a one sided 80% confidence interval was necessary to identify these extreme data in order to have a clearer picture on the comparison among the methods. Then $\boldsymbol{\gamma = 0.20}$, which indicates that there is 0.20 or less probability of misidentifying a good data as an outlier for being too extreme and similarly at most 0.20 probability of misidentifying a good data as an outlier for being too mild.

A. Oil Yield for the Belle Ayr Liquefactioin Data

First, we consider a sample which is "Belle Ayr liquefaction" data in Montgomery et al. (2001). The data is about thermal liquefaction of coal. The variable "Oil Yield" is the results of conversion of coal to (production of) oil for 27 runs of the experiment. The robust Bowley measure of skewness value is -0.0228. This shows that the underlying distribution of this data is slightly asymmetric with two higher value observations which can be observed from the Normal Q-Q plot in Figure 1. These two observations are distant from majority of the observations. In Figures 3 and 4, for both sequential fences methods, it can be noted that there are two observations fall beyond upper first fence $m = 1$. Then, the checking procedures keep continuing till $m = 3$, there are still two observations be over the upper fence. Thus, both sequential fences detect two outliers. As illustrated in Figures 2-4, all the three methods are able to identify the two outliers. The fences of proposed adjusted sequential fences are slightly shifted upward due to the adjustment using robust skewness.

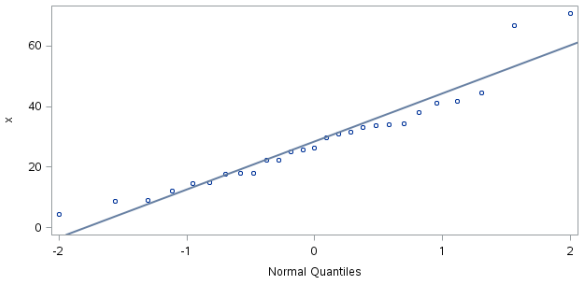

Figure 1. Normal Q-Q plot of Belle Ayr liquefaction data

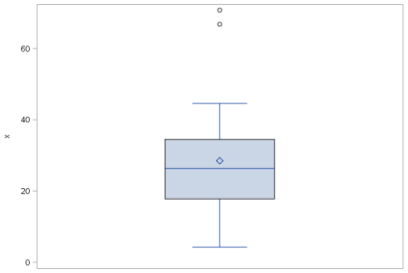

Figure 2. Tukey boxplot of Belle Ayr liquefaction data

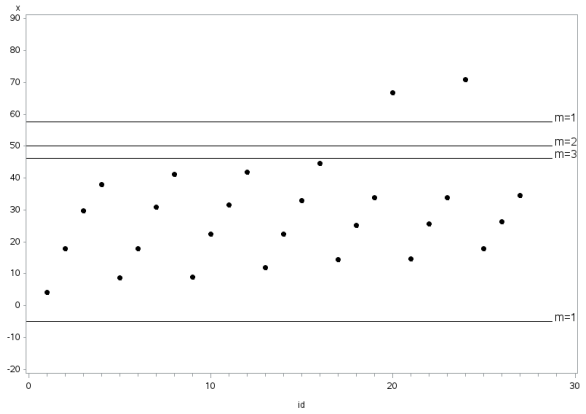

Figure 3. Standard sequential fences of Belle Ayr liquefaction data

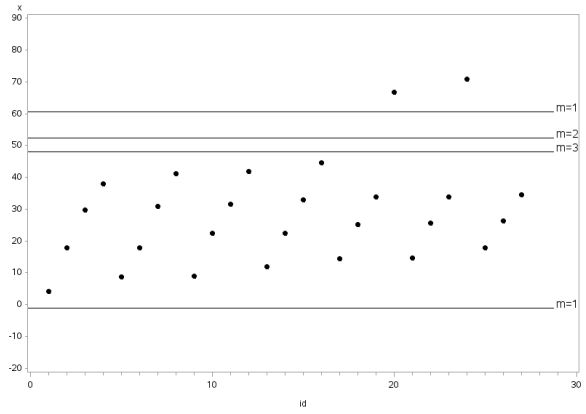

Figure 4. Adjusted sequential fences of Belle Ayr liquefaction data

B. Belgium CPI Data

As a second illustration, Belgium consumer price index (CPI) for monthly relative price differences of data which consists of 60 items is selected. The data is asymmetrically distributed with a positively skewness with two outliers on the right tail (Dutter *et al.*, 2003). The value of Bowley coefficient is 0.3614, which indicates that the distribution is highly skewed. From Normal Q-Q plot in Figure 5, there is obviously two observations deviate from the majority of the data. The upper whisker of the adjusted sequential fences extends further away from the median than that of the Tukey boxplot as shown in Figure 8. From both sequential fences in Figures 6-8, the Tukey boxplot and standard sequential fences are affected by the asymmetrical property distribution whereas the adjusted sequential fences is able to adjust the fences to the skewed data. As a consequence of this substantial asymmetry, the upper fence of Tukey boxplot and standard sequential fences inaccurately capture uncontaminated observations as outliers. However, the adjusted boxplot performs better in capturing all the extreme data as atypical but indicates that several observations below the lower fence as potential outliers that require special attention.

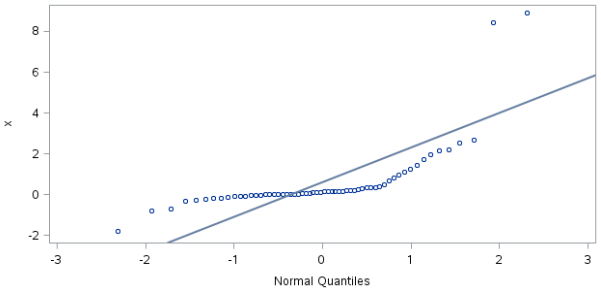

Figure 5. Normal Q-Q plot of Begium CPI data

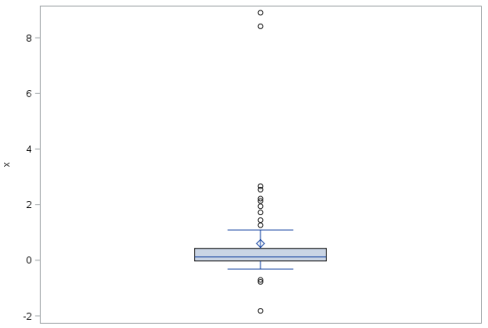

Figure 6. Tukey boxplot of Belgium CPI data

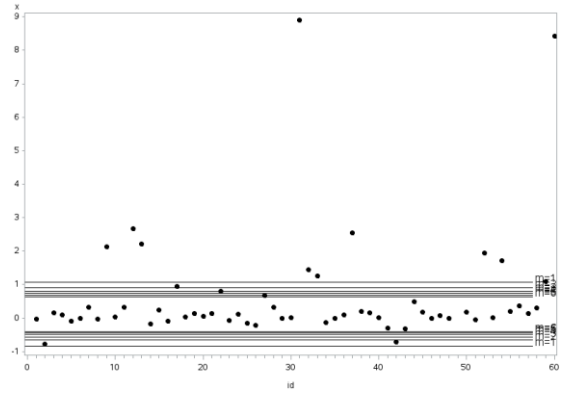

Figure 7. Standard sequential fences of Belgium CPI data

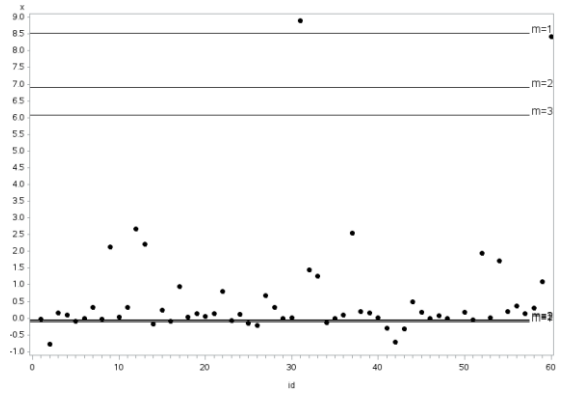

Figure 8. Adjusted sequential fences of Belgium CPI data

V. CONCLUSION

In this paper, we presented an adjusted sequential fence to create fences for detecting outliers with a continuous distribution that might be skewed and/or heavy-tailed. This approach can be utilised to identify atypical observations as it is exclusively constructed based on the skewness of distributions. Moreover, it is robust with respect to the outliers. The adjusted sequential fences method takes into account the skewness of the underlying distribution of data by incorporating the robust Bowley coefficient of skewness into the sequential fences to adjust lower and upper cut off values. It has clear advantages over the Tukey boxplot as well as the standard sequential fences.

Comparative study between proposed adjusted sequential fences and Tukey's boxplot and standard sequential fences method has been done. In symmetric and fairly symmetric case, these three methods behave very similarly. However, the results on simulation study and real data indicate that the proposed method perform better in identifying the outliers and is stable with lower error of misclassified regular observations as outlying data as the data is skewed and sample size increases. This makes the adjusted sequential fences an appropriate approach in detecting outliers for variety of distributions data.

In this paper, we focused on the identification of univariate outliers in symmetric and moderately skewed distributions. In future, the extension of the idea of adjusted sequential fences method can be applied to more highly skewed distributions to detect multivariate outliers.

VI. ACKNOWLEDGEMENT

The authors wish to thank University College of Technology Sarawak for funding assistance: UCTS/RESEARCH/1/2019/06

VII. REFEREENCES

Adil, IH & Irshad, AUR 2015, 'A Modified Approach for Detection of Outliers', *Pakistan Journal of Statistics and Operation Research*, vol. 11, no. 1, pp. 91.

Aucremanne, L, Brys, G, Hubert, M, Rousseeuw, PJ & Babura, BI, Adam, MB, Fitrianto, A & Abdul Samad, AR 2017, 'Modified boxplot for extreme data', The 3rd ISM International Statistical Conference 2016 AIP Conference Proceedings, pp. 1842.

Banerjee, S & Iglewicz, B 2007, 'A simple univariate outlier identification procedure designed for large samples', *Communications in Statistics - Simulation and Computation*, vol. 36, no. 2, pp. 249–263.

Bowley, AL 1920, Elements of statistics (London: P.S. King & Son, ltd), vol. 2.

Carling, K 2000, 'Resistant outlier rules and the non-Gaussian case', *Computational Statistics & Data*

Analysis, vol. 33, no. 3, pp. 249–258.

Coles, S, Bawa, J, Trenner, L & Dorazio, P 2001, 'An introduction to statistical modeling of extreme values' (London: Springer), vol. 208, pp. 47-54.

Davies, L & Gather, U 1993, 'The Identification of Multiple Outliers', *Journal of the American Statistical Association*, vol. 88**,** pp. 782–792.

Dovoedo, YH & Chakraborti, S 2012, 'Boxplot-based phase I control charts for time between events', *Quality and Reliability Engineering International*, vol. 28, no. 1, pp. 123–130.

Dovoedo, YH & Chakraborti, S 2015, 'Boxplot-Based Outlier Detection for the Location-Scale Family', *Communications in Statistics - Simulation and Computation*, vol. 44, no. 6, pp. 1492–1513.

Dutter, R, Filzmoser, P, Gather, U & Rousseeuw, PJ 2003, 'A comparison of some new measures of skewness', Developments in robust statistics: International Conference on Robust Statistics 2001 (Heidelberg: Springer-Verlag), pp. 98–113.

Gather, U & Becker, C 1997, 'Outlier identification and robust methods', *Handbook of Statistics Robust Inference*, vol. 15, pp. 123–143.

Groeneveld, RA & Meeden, G 1984, 'Measuring Skewness and kurtosis', *Journal of Royal Statistical Society Series D* (*The Statistician)*, vol. 33, no. 4, pp. 391-399.

Hoaglin, DC, Iglewicz B & Tukey JW 1986, 'Performance of some resistant rules for outlier labeling', *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 991–999.

Hubert, M & Vandervieren, E 2008, 'An adjusted boxplot for skewed distributions' *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5186–5201.

Kim, T & White, H 2004, 'On more robust estimation of skewness and kurtosis: simulation and application to the S & P500 index', *Finance Research Letters*, vol. 1, no. 1, pp. 56-73.

Kimber, AC 1990, 'Exploratory data analysis for possibly censored data from skewed distributions', *Applied Statistics*, vol. 39, no. 1, pp. 21–30.

Montgomery, DC, Peck, EA & Vining, GG 2001, Introduction to linear regression analysis, 3rd edn, John Wiley, New York.

Schwertman, NC & Silva, RD 2007, 'Identifying outliers with sequential fences', *Computational Statistics & Data Analysis*, vol. 51, no. 8, pp. 3800–3810.

Schwertman, NC, Owens, MA & Adnan, R 2004, 'A simple more general boxplot method for identifying outliers', *Computational Statistics & Data Analysis*, vol. 47, pp. 165–174.

Struyf, A 2004, 'A Study of Belgian inflation, relative prices and nominal rigidities using new robust measures of skewness and tail weight', *Theory and Applications of Recent Robust Methods*, pp. 13–15.

Thas, O, Vanrolleghem, P, Kops, B, Van Vooren, L & Ottoy, J P 1997, 'Extreme value statistics: potential benefits in water quality management', *Water Science and Technology*, vol. 36, no. 5, pp. 133–140.

Tukey, JW 1977, Exploratory Data Analysis (Reading, MA: Addison-Wesley), pp. 39-49.