ASM Sc. J., 17, 2022

https://doi.org/10.32802/asmscj.2022.500

Split Sample Sequential Fences based on Bootstrap Cut Off Points for Identifying Outliers and Parameter Estimations

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*

Sequential fences is a simple graphical method that is used for detecting outliers. This method is advantageous to the analysts in constructing fences which adjusted for various sample sizes. It is a helpful way to to detect the single and multiple outliers especially in normal or approximately normal data. However, when the distributions of the data are skewed, sequential fences method tends to result in misleading outcome. This paper proposes solution to deal with this problem. The proposed approach with modified algorithm namely Split sample sequential fences with bootstrap (SSFB) is an alternative way to improve sequential fences which can lead to higher accuracy in the detection of outliers and can be applied to a wide range of distributions data. The validity of the new technique has been checked by constructing fences around the true 95% values of different distributions. It was found that the sequential fences constructed by the modified technique not only can detect the outliers in positively skewed distribution but also has smaller bias and smaller root of mean squares error (RMSE) which proves it superiority on the existing techniques.

Keywords: sequential fences; bootstrap; split sample; outliers; parameter estimations

I. INTRODUCTION

Outliers are the observations which lie far away from the majority of cases in the dataset. Boxplot is one of the widely used graphical method which is introduced by Tukey (1977) to identify outliers. Inner fences are situated at a distance of 1.5 IQR below q_1 and above q_3 which are denoted as

$$[q_1 - 1.5\,IQR,\ \ q_3 + 1.5\,IQR]. \quad (1)$$

Outer fences can be constructed at a distance of 3 IQR below q_1 and above q_3 which are computed as

$$[q_1 - 3\,IQR,\ \ q_3 + 3\,IQR]. \quad (2)$$

The Tukey's box plot displays distribution of data based on statistics summary which are minimum, maximum, median, first and third quartiles. The whiskers of boxplot are plotted 1.5 multiply the interquartile range from the median. In the past years, the boxplot has been modified and changed by different researchers, as well as variations in the definition of quantile.

Schewertman and de Silva (2007) proposed sequential fences as another useful technique to detect the outliers in the data. In this study, the sequential fences proposed by Schewertman and de Silva is henceforth referred as SDSF. This technique identifies outliers sequentially based on the specific sample size and the pre-specified outside rate. The rate is the probability that an uncontaminated observation beyond the fences. The fences are defined as

$$F_{n,m} = q_2 \pm \frac{t_{df,\alpha_{nm}}}{k_n} IQR \quad (3)$$

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

where q_2 is median, IQR is interquartile range, k_n and α_{nm} are values that are obtained from Table 1 and Table 2 in Schewertman and de Silva (2007) to construct m^{th}fences, $t_{df,\alpha_{nm}}$ is the value obtained from t distribution based on specified outside rate, α_{nm} and degree of freedom, df which is calculated using Equation (4). For sample size between 20 and 100, the least squares quadratic equation for obtaining the degree of freedom, df approaching t distribution based on the sample size is

$$df = 7.6809524 + 0.5294156n - 0.00237n^2. \quad (4)$$

For the construction of sequential fences, the sample sizes are adjusted using Poisson model in order to decrease the tail probabilities. The adjustment is similar to the adjustment done in Davies and Gather (1993) and Gather and Becker (1997). SDSF increases the accuracy to identify the outliers, reduces the swamping effect and less likely to misclassify uncontaminated observations as outliers in large sample size. By using Poisson model, the m contaminated observations can be checked. Let X be the number of observations outside the computed fences. Based on the Poison model,

$$P(X < m) = e^{-n\alpha_{nm}}\left(1 + n\alpha_{nm} + \frac{(n\alpha_{nm})^2}{2!} + \cdots + \frac{(n\alpha_{nm})^{m-1}}{(m-1)!}\right) = 1 - \gamma. \quad (5)$$

The solution of Equation (5) for $n\alpha_{nm}$ for γ 0.05, 0.025 and 0.005 can be referred to Table 2 which is adapted from Schewertman and de Silva (2007). The probability of at most $(m-1)$ uncontaminated observations beyond the constructed fences is $1-\gamma$. In short, there is γ probability of at least m uncontaminated observations outside the fence. For instance, in order to check for the first outlier, compute the first fence, $m = 1$, with $\gamma = 0.10$. This means that there is 0.10 probability that an observation which lies beyond the first fence is uncontaminated. SDSF method can identify outliers using different outside rates in lower and upper tail separately. Thus, the characteristic of SDSF which allows to use different outside rates on both tails is suitable to some occasions when the number of outliers in either tail of the distribution are unequal.

Table 1. Conversion coefficients for IQR to σ (IQR = $k_n\sigma$)

n	k_n	n	k_n	n	k_n
5	1.65798	22	1.33333	39	1.38071
6	1.28351	23	1.4023	40	1.34165
7	1.51475	24	1.33753	41	1.38021
8	1.32505	25	1.40096	42	1.34104
9	1.50427	26	1.33587	43	1.37779
10	1.31212	27	1.39455	44	1.34226
11	1.45768	28	1.33894	45	1.37737
12	1.32968	29	1.39355	46	1.34175
13	1.45268	30	1.3377	47	1.37536
14	1.32353	31	1.38876	48	1.34278
15	1.42975	32	1.34004	49	1.37501
16	1.33318	33	1.38799	50	1.34235
17	1.42684	34	1.33909	60	1.34394
18	1.32959	35	1.38428	70	1.34429
19	1.41322	36	1.34092	80	1.34514
20	1.33568	37	1.38367	90	1.34535
21	1.41132	38	1.34017	∞	1.34898

Table 2. Constants $C_m = n{\alpha_{nm}}^a$

1-γ	m=1	2	3	4
0.75	0.287682	0.961279	1.72730	2.53532
0.80	0.223144	0.824388	1.53504	2.29679
0.90	0.1053605	0.531812	1.10207	1.74477
0.95	0.0512932	0.355362	0.817691	1.36632
0.975	0.025318	0.242209	0.618672	1.08987
0.99	0.0100503	0.148555	0.436045	0.823249
0.995	0.005013	0.103495	0.337873	0.672207

Hyndman and Fan (1996) made a conclusion by recommending the use of median-unbiased estimator, since median contains most of the desirable properties of a quantile estimator and can be defined independently from the underlying distribution. Besides, some boxplots use multipliers other than 1.5 for the whiskers of boxplots or substitute the extreme with constant quantile such as minimum and maximum or 2% and 98% (Frigge *et al.,* 1989). There are also other graphical elements used to display the distributional characteristics such as kurtosis (Aslam & Khurshid, 1991), skewness and multimodality (Choonpradub & McNeil, 2005), and mean and standard error (Marmolejo-Ramos & Tian, 2010).

For skewed distribution data, some researchers (Kimber, 1990; Aucremanne *et al.,* 2004) made adjustments for the boxplot by introducing lower and upper semi-interquartile range, $SIQRL = Q_2 - Q_1$ and $SIQRU = Q_3 - Q_2$, in order to replace the interquartile range. In that case, the fences are defined as $[Q_1 - 3SIQRL;\ Q_3 + 3SIQRU]$. Huber and Vandervieren (2008) pointed out that the boxplot using $SIQR$ does not sufficiently adjust itself for skewness of distribution. There is enlargement in the whisker part and consequently

lead to a number of uncontaminated observations is flagged as outliers.

Babura *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 generalisation aim. Besides, this approach is capable to show the location parameter region of Gumbel or Generalised Extreme Value Distribution (GEV) fitted extreme data. Wong and Fitrianto (2019) proposed an outlier detection method, namely Adjusted Sequential Fences (ASF) by making some adjustments to the sequential fences (2007) with the combination of a robust skewness. The adjusted sequential fences method takes into account the skewness of the underlying distribution of data by incorporating the robust Bowley coefficient of skewness, $\zeta = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$, into the sequential fences to adjust lower and upper cut off values. The fences for detecting outliers with a continuous distribution are able to identify outliers in skewed and/or heavy-tailed data. This technique is able to identify atypical observations as it is exclusively constructed based on the skewness of distributions. Moreover, it is robust with respect to the outliers.

Thus, the adjusted sequential fences (ASF) are as follows:

when $\zeta \geq 0$,

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

when $\zeta < 0$,

$$\begin{cases} Q_2 - \frac{t_{df,\alpha_{nm}}}{k_n} e^{6\zeta} IQR, \\ Q_2 + \frac{t_{df,\alpha_{nm}}}{k_n} e^{-4\zeta} IQR. \end{cases} \quad (7)$$

Median and interquartile range (IQR) are commonly used in the computation of boxplot and fences which are robust way to provide better summaries of data. Median and IQR are non-parametric univariate statistics which is used to obtain center and spread for quantitative variables. The benefit of this non-parametric technique is that it requires less assumptions, thus the non-parametric technique can be applied for a wider range of applications. In addition, the non-parametric technique is simpler compared to parametric technique. When the data are not normally distributed, not measured on an interval scale or the sample size is small, median and IQR are good measures to summarise the distribution (Iftikhar, 2011). Therefore, median and IQR shares some similarity as the mean and standard deviation to measure the centre and spread of data. However, mean and standard deviation are easily distorted by the presence of outliers.

Both Tukey's boxplot and SDSF use median and interquartile range in the formulation of fences for the outlier detection. These methods concern only the central half of the data which is the interval from 25th percentile to 75th percentile of the data. Meanwhile, skewed distributions have narrower tail in either side and the skewness of interval between 12.5th and 37.5th percentile and 62.5th and 87.5th percentile are dissimilar. This motivates us to make adjustment to the sequential fences by considering the different skewness of the distribution on both sides. The data lying in the tails parts should be considered in computing the fences.

Data screening is necessary before beginning a data analysis. (Tabachnick & Fidell, 2001). Monte Carlo simulation is a type of simulation that relies on repeated random sampling and statistical analysis to compute the results (Samik, 2008). Data analysis is done based on the data that is generated by computer. However, researchers do not aware about the accuracy of the generated data and just proceed the data analysis process by interpreting and making conclusions from the results (Anwar & Habshah, 2011). This can lead to erroneous conclusions. Thus, the data generating process does not give warranty the data is free from outliers. It is necessary to screen for the data before conducting data analysis.

Process of data screening has become a part in data analysis. This process has become a procedure that researchers have to face with. Statistical tests for outliers are included in the data validation process. After the data are screened and evaluated in numerous ways, the data are then kept in data bank and to be used for population parameters estimations and decisions making. The data screening for air quality data and validation procedures have been discussed by Nelson *et al.* (1980). Bootstrapping is a statistical technique for making an estimation of the sampling

distribution of an estimator by sampling with replacement from the actual sample. The technique allocates measures of precision to sample estimates with bias and root of mean square error.

In order to address the problems, in this study, an adjustment to the sequential fences SDSF involving bootstrap resampling is proposed. The objectives of this study are i) to increase the coverage of the data that lie on either side of the tails so that it can be applied to different types of distributions and various sizes of data, ii) to modify the algorithm of the sequential fences technique by incorporating procedure of data screening, and iii) to show accuracy of parameter estimations after outliers is detected by the proposed method and the existing methods, SDSF and Tukey's boxplot.

The remainder of the paper is organized as follows. Section 2 reviews data screening and robust estimation. Section 3 presents the procedures of the proposed approach in detailed. In section 4, we illustrated the performance of the techniques in outliers detection in various distributions with different sample sizes data. Furthermore, bootstrap resampling was performed to estimate the robust parameters in uncontaminated and contaminated data in order to show the efficiency of the techniques in different conditions.

II. LITERATURE REVIEWS

A. Data Screening

In statistical analysis, it is important to screen the data before analysing the data to prevent the occurrence of misleading results. Computer is utilised to simulate and analyse the data. However, the generated data from a simulation might consist of potential outliers or unusual observations.

Tabachnick and Fidell (2001) recommended a procedure for screening a data with appropriate sequences. The order of the data screening can affect the final decision making. It is related to data distribution. When the data is not normal and contamination of the data might create problems which lead to decision whether discard or transform the data. The screening process can make sure the data set is suitable to be used. It will help in identification of unusual observation and enable us to do necessary adjustment to the data for further analysis.

According to the past literatures such as Beckman and Cook (1983), Ahmad *et al. (*2011), Anwar and Habshah (2011), and Tabachnick and Fidell (2011), outlier detection is a segment of data screening procedures which should be conducted prior any statistical analysis. Outliers can present in univariate and multivariate data. In symmetrically distributed data, the observations which are located at both end of the tails are suspected to be outliers. Meanwhile, for skewed data, the observations which fall on the longer tail might likely to be outlying observations. Therefore, screening data is an important first step before beginning a statistical analysis.

B. Bootstrapping in Outlier Detection and Parameter Estimation

Simulation is a procedure to generate random sampling from probability distributions such that the simulated data is approximately representing real world outcomes. Based on simulated outcomes, researchers are able to have a perception of the real world.

Efron (1979) introduced a bootstrap technique. This technique implies that the sample has the same connection to the population as an empirical distribution produced by resampling N samples of the same size as the original sample with replacement from the original distribution (Denise, 2021). Determination of the parameter value of a population is typically impossible to measure directly from the population and might incur a high cost.

Hence, bootstrap resampling provides solution to overcome these problems. The simplicity of bootstrap helps in building an estimate of the sampling distribution by drawing a large amount of random sample of size *n* from a population. From the estimated standard errors, hypothesis testing and confidence intervals, we can make interpretation about the corresponding of population parameter. Furthermore, bootstrap method is also easily implemented in statistical softwares such as *R* and *SAS* programming to make a summary of a sample in order to obtain a general conclusion regarding the population.

Another advantage of the bootstrap can be applied in circumstances where standard statistical tools are unavailable or in situations where the usual statistical methods are inappropriate due to the violation of the

underlying assumptions (Hansen *et al.,* 1999). Furthermore, using bootstrap resampling and other theoretical computations, standard error can be calculated for any complex estimator. For example, the basic summary statistic such as sample mean fluctuates from sample to sample. Analysts would like to identify the magnitude of the fluctuations about the corresponding population parameter in an overall sense which is called margin of errors. All the possible values of the sample statistics of the entire display is presented in probability distribution or sampling distribution. Thus, it provides a good approximation of the sample distribution.

Bootstrap method has been applied to outlier detection procedure. Singh and Xie (2003) proposed a bootstrap based outlier detection plot or known as bootlier plot which is a non-parametric graphical tool to identify the outliers. The bootstrapping sample statistics called 'mean – trimmed mean' (MTM) was introduced. From the bootlier plot, outliers can be detected by checking the multimodality in the density plot of that bootstrap sample statistic. The distribution of bootstrap sample statistic MTM is expressed as a combination of normal distributions with multiple modes when the sample has outliers.

C. Estimation of Robust Estimators

In statistics, based on information acquired from a sample, one can make inferences about a population through estimation. Researchers utilise sample statistics to estimate population parameters. For instance, sample means and sample standard deviation are commonly used to estimate the population parameters means and standard deviation, respectively. However, sample means and standard deviations are vulnerable and sensitive to the outlying values.

When outliers are present in the data, the distribution of the data is usually heavy tailed and outlying observations are found in higher proportion and far away from the mean. In order to cope with the problem, robust estimators of the population mean such as trimmed and winsorised means were introduced. These robust estimators are relatively less sensitive to the outliers. Therefore, trimming and winsorisation are techniques for decreasing the impacts of extreme values in the sample.

The trimmed mean or truncated mean is a measure of the population mean which its standard error is less distorted by the departures from normality where the extreme observations are removed (Lix & Keselman, 1998). The calculation is usually done after discarding low and high part of the end of a probability distribution or sample and usually discarding an equal amount on both tails. The number of values to be discarded is typically based on the percentage of the total number of values, α but may also be a fixed number of values to be removed (Boos & Stefanski, 2013). The trimmed mean can be defined as the mean of the central $1 - 2\alpha$ part of the distribution. Before computing the trimmed mean, the percentage of trimming has to be specified. This can be implemented by eradicating αn of the data from each end of the distribution. For instance, trimming 0% equal to mean while trimming 50% gives second quartile or median.

The winsorised mean is another robust measure of mean. It is a winsorised statistical measure of central tendency which is less sensitive to extreme values (Gross, 1976). It involves the calculation of the mean after substituting given parts of a probability distribution or sample at the low and high end with the values that are closest to them (Wilcox, 1995). Commonly, a similar amount of both extreme values are replaced.

Meanwhile, the trimmed standard deviation is a robust estimator of scale (Capéràa & Rivest, 1995). The computation of trimmed standard deviation is the average trimmed sum of squared deviations around the trimmed mean after discarding a certain percentage of observations from the tails. For instance, the 50% trimmed standard deviation is the standard deviation of the observations between the upper and lower quartiles.

III. IMPLEMENTATION OF PROPOSED TECHNIQUE

In this study, a new approach of sequential fences involves split sample method and determination of cut off points based on bootstrap resampling was proposed. We call this approach as Split Sample Sequential Fences based on Bootstrap Resampling (SSFB). The technique was then compared with Tukey's boxplot and SDSF method which was proposed by Schewertman and de Silva (2007). Positively skewed distribution data is considered and a one sided 95%

confidence level was required for the SDSF and SSFB to detect the outliers. The efficiency of the techniques in detecting the number of outliers in different skewed distributions were examined. Robust estimators such as trimmed mean and trimmed standard deviation are also computed based on the outliers detected by each method to validate the productivity of the methods in correctly identify the real outliers.

A. Screening of Data to Generate Clean Data

In this section, simulation study was performed to generate data based on various sample sizes, *n*, and distributions. The sample sizes considered were 20, 50 and 100. Different distributions such as normal, log normal and chi-square distribution with different parameters were considered.

In order to ensure the generated data is clean, first, the data was screened using the outlier detection method, namely Adjusted Sequential Fences method (ASF) (Wong & Anwar, 2019). The steps to screen a data was proposed and discussed by Fitrianto and Habshah (2011). The procedure that is incorporating the ASF approach into the algorithm to generate a set of clean data is named as GCD.

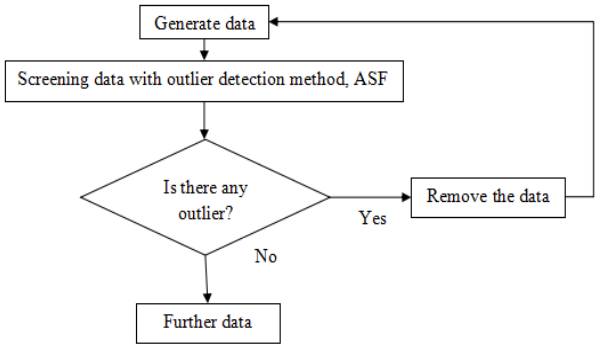

Figure 1. Flow chart of the GCD algorithm

The algorithm is shown in Figure 1 and the steps are as follows:

Step 1. The regular observations are generated from a specified distribution with size n.

Step 2. The data is screened based on a condition of the specified outlier detection technique. In this study, we use ASF to check for the existence of outliers in the data.

Step 3. Carry out outlier identification procedures in the generated data using the ASF.

Step 4. Let Out be a variable to record the presence of outlier. If an observation is flagged as outlying observation, then assign '1' and '0' if otherwise. In other words, the variable Out consists of a set of binary variable which containing 0 and 1 values only.

Step 5. The frequency of number '0' value in variable Out is counted. If the number of '0' equals to n, then the screening procedures is completed, otherwise repeat steps 1 until 4.

B. Detecting Outliers using Split Sample Sequential Fences with determination of Cut Off Points based on Bootstrap Resampling

Next, the clean distributions data was contaminated with no outlier, one outlier, or multiple outliers. All outliers were situated in the upper tail. It is a common practice that any observation is located beyond the extremes (maximum and minimum) as outliers (Georgy *et al.,* 2013). In order to generate outlier in the simulated data, the single outlier for the simulations was 10 standard deviations above the largest observation. When simulations with two outliers for the distribution data were done, the outliers were an addition of fixed shift of 10 standard deviation distance from the two observations. Similarly, the procedure for three outliers contamination was done by contaminating three observations with addition of 10 standard deviations.

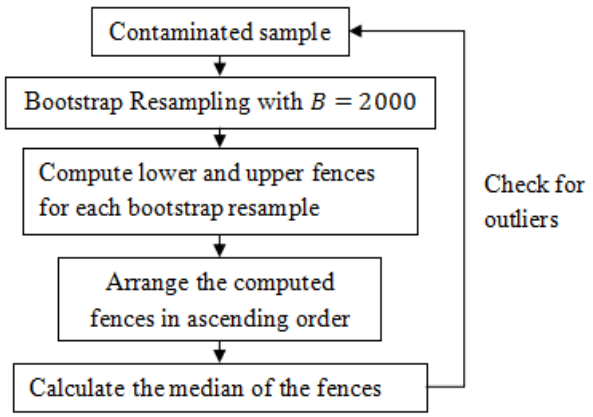

Figure 2. Flow chart of the SSFB algorithm

For the proposed approach, bootstrap resampling with replacement is used to draw the samples randomly of size n from the original sample as shown in Figure 2.

Bootstrap technique is used to estimate the cut off points of sequential fences. The following are the details of the proposed SSFB procedures:

Step 1. Contaminate the data by adding 10 standard deviation as a fixed shift to the highest observation.

Step 2. Generate a bootstrap sample $B^* = (B_1, B_2, B_e, \dots, B_n)$. The sample of size n is generated by common standard bootstrap resampling with replacement from the contaminated sample. The number of bootstrap replications is set as 2000.

Step 3. For each bootstrap sample, compute the fences using the proposed outlier detection method. The procedurs of the construction of the newly proposed fences is discussed further in the end of this subsection. In this study, we also compute the fences based on Tukey's boxplot and SDSF for each bootstrap sample, so that a clear picture of the efficiency of each method can be obtained. One sided significance level, 0.05 was used with the SDSF and SSFB methods.

Step 4. Arrange the computed fences which were produced in each resample in ascending order, then obtain the median of the fences. The median of the fences are the cut off points which will be used to identify the outliers.

Step 5. Apply the cut off points to the contaminated sample. If an observation falls beyond the cut off points, then it is classified as an outlier.

The fences of SDSF are computed using median and interquartile range. It is done by separating the data into two parts from the median of the data. Then, first and third quartile are obtained from each part of the data to calculate interquartile range. Split sample method is used to obtain a better coverage of the fences to the data. The procedures to split the sample are based on Iftikhar (2011). First, the data is separated into two parts from the median. Then, the data is divided again into four parts for each of the separated part in order to obtain 12.5th, 37.5th, 62.5th and 87.5th percentile.

The previous SDSF technique concerns only on the central half of the data sets while the proposed approach considers for wider coverage of the data. In the skewed distribution case, the new approach is expected to be able to build coverage that close to the middle 95% values of data which is higher coverage compared to the SDSF. The SDSF leaves 2.5% data on each side of the distribution. Since the simulated data is skewly distributed, the skewness of interval between 12.5th and 37.5th percentile is different with the skewness of interval between 62.5th and 87.5th percentile. The following are the steps to implement the proposed method.

Let $x = \{x_1, x_2, x_3, \dots, x_n\}$ be sample of size n from an unspecified probability distribution. Then, a bootstrap sample of sample size, $x^* = \{x_1^*, x_2^*, x_3^*, \dots, x_n^*\}$, is generated. After ranking the resample values in each bootstrap sample from lowest to highest, let us denote these bootstrap values as $\{x_{(1)}^*, x_{(2)}^*, x_{(3)}^*, \dots, x_{(n)}^*\}$. From the ordered bootstrap values, determine the 12.5th, 37.5th, 62.5th and 87.5th percentiles in each resample. Two thousand observations for each percentile $P_{12.5} = 12.5$ percentile, $P_{37.5} = 37.5$ percentile, $P_{62.5} =$ 62.5 percentile and $P_{87.5} = 87.5$ percentile are generated by continuously repeating the procedures.

In each sample, the percentiles are then used to calculate the lower and upper interquartile range, IQR_L and IQR_R. The computation of the IQR_L and IQR_R are written as

$$IQR_L = P_{37.5} - P_{12.5}, \tag{8}$$

$$IQR_R = P_{87.5} - P_{62.5}. \tag{9}$$

The IQR_L and IQR_R are used in the construction of the proposed method. Thus, the split sample sequential fences, $SSFB$, are defined as

$$\text{Lower } SSFB_{n,m} = P_{12.5} - \frac{t_{df,\alpha nm}}{k_n} IQR_L, \tag{10}$$

$$\text{Upper } SSFB_{n,m} = P_{87.5} + \frac{t_{df,\alpha nm}}{k_n} IQR_R \tag{11}$$

which are lower and upper fences, respectively.

Compared to the SDSF method, instead of median, 12.5 percentile and 87.5 percentile are used in order to construct the fences. The SSFB technique combines the benefits of boxplots and sequential fences (easy of understanding, capacity to compare several data sets at the same time) with other percentile plots (display all the data, no arbitrary choices in construction). The idea is to highlight the middle of the data by using width (as in the SDSF) and to provide less attention to the more extreme data by continuing to utilise width.

As a result, the SSFB fences are wide in the middle and very narrow at the extreme. The width, unlike the boxplot and SDSF, provide precise information about the data distribution. They contain all of the information found in SSFB and allow for quick and precise symmetry evaluation.

Furthermore, IQR_L and IQR_R are chosen to substitute the interquartile range in SDSF method so that the proposed sequential fences can be determined according to the skewness of the underlying distribution. If IQR_L is smaller than IQR_R, then the distribution is right skewed. If IQR_L is greater than IQR_R, then the distribution is left skewed. The procedures stated above are summarised as follows:

Step 1. From the data set $\{x_1, x_2, x_3, \dots, x_n\}$, generate a bootstrap sample $x^* = \{x_1^*, x_2^*, x_3^*, \dots, x_n^*\}$ with replacement.

Step 2. Rank the observations in each bootstrap sample in ascending order.

Step 3. Determine the 12.5^{th}, 37.5^{th}, 62.5^{th} and 87.5^{th} percentiles in each sample set.

Step 4. Calculate the lower and upper interquartile range, IQR_L and IQR_R by using the Equations (8) and (9).

Step 5. Determine the degree of freedom, df.

Step 6. Obtain the k_n values from Table 1 of Schwertman and de Silva (2007) according to the sample size.

Step 7. For construction of each m fence, obtain constant α_{nm} under specific outside rate by dividing the C_m values with n. The values of C_m can be referred to Table 2 of Schewertman and de Silva (2007).

Step 8. Construct the proposed fence, m for lower and upper side using Equations (10) and (11). Initiate it with first fence $(m = 1)$.

Step 9. There are as many as B number of lower and upper fences generated. Then, sort the fences in ascending order and obtain the median among the fences and use the obtained median as the cut off points.

Step 10. Apply the cut off points to the original sample data. Then, check any outlier that fall outside lower and upper cut off points.

Step 11. If the number of outliers detected is $(m - 1)$, then the procedure is stopped. If there is more than or equal to m outliers beyond the fence, then continue the outliers identification by repeating steps 8, 9 and 10 until the construction of the next fence $(m + 1 - th)$ fence does not capture any additional extreme observations beyond the fences.

If there is no observation lies beyond the first fence $(m = 1)$, this means that there is no outlier detected and the identification of outlier is accomplished. Otherwise, if there is at least one observation beyond the first fence, then the procedures continues by constructing the next fence to check whether there are other outlying observations. It is important to check the outliers by constructing the fences continuously until there are only m observations beyond the $(m + 1 - th)$ fence.

C. *Computation of Robust Estimators based on the proposed Bootstrap Resampling*

In order to study the efficacy of the proposed approach, robust estimators such as trimmed mean and trimmed standard deviation are studied. Trimmed mean and trimmed standard deviation for one sided and two sided are calculated. The determination of number of observation to be trimmed is based on the number outliers detected by a particular outlier detection method.

Singh and Xie (2003) introduced bootlier plot which is bootstrap based statistical framework involving trimming procedures to identify outliers. The direction of the trimming is determined by which side(s) of outlying observations located. Let k be the number of outliers detected by a particular method. The trimmed mean for one sided and two sided can be obtained as follows:

Lower-sided trimmed mean:

$$\bar{x}_{TL} = \frac{1}{n-k}\sum_{i=1}^{n-k} X_{(i)}^* \tag{12}$$

Upper-sided trimmed mean:

$$\bar{x}_{TU} = \frac{1}{n-k}\sum_{i=k+1}^{n} X_{(i)}^* \tag{13}$$

Two-sided trimmed mean:

$$\bar{x}_{T2} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} X_{(i)}^* \tag{14}$$

For one-sided trimming, the trimmed mean estimator, $\bar{x}_L$ and $\bar{x}_U$, are sensitive to the outliers in both lower side and upper side. However, for two-sided trimmed mean, similar number of data points are trimmed on both sides. In this study, since the focus is only on outliers problem that occur in positively skewed distribution, one-sided trimmed mean is

used to trim the outlying observations that are found on either side of tail. Two-sided trimmed mean is also calculated whereby equal number of observation is trimmed on both tails when outliers are detected on either tail of the data. If the number of outliers detected on both tails are different, then the larger number of outliers will be selected to be trimmed equally on both sides. Similar trimming procedures are carried out for estimating trimmed standard deviation which are simulated as follows:

Lower-sided trimmed standard deviation,

$$s_{TL} = \sqrt{\frac{\sum_{i=1}^{n-k}(X_{(i)}^{*}-\bar{x}_{TL})^2}{n-k}} \quad (13)$$

Upper-sided trimmed standard deviation,

$$s_{TU} = \sqrt{\frac{\sum_{i=k+1}^{n}(X_{(i)}^{*}-\bar{x}_{TU})^2}{n-k}} \quad (14)$$

Two-sided trimmed standard deviation,

$$s_{T2} = \sqrt{\frac{\sum_{i=k+1}^{n-k}(X_{(i)}^{*}-\bar{x}_{T2})^2}{n-2k}} \quad (15)$$

This procedure is denoted as trimmed estimators based on bootstrap resampling (TEB). The algorithm of this procedure is presented in Figure 3.

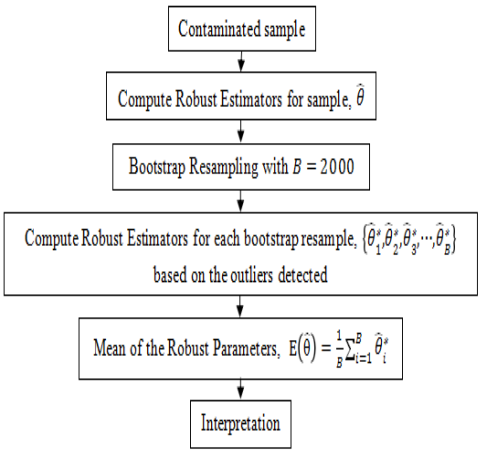

Figure 3. Flow chart of the TEB algorithm

IV. RESULT AND DISCUSSION

In this section, detail comparisons between the proposed technique with Tukey's boxplot and SDSF in symmetric and asymmetric distributions were done with various sample sizes. Normal distribution and several positively skewed distributions such as log normal distribution with parameters (5,0.4), (5,0.6) and (5,0.8) and chi square distribution with degree freedom 2, 4 and 8. Small (n=20), medium (n=50) and large (n=100) sample sizes were investigated. Table 3 shows a total of twenty one different distributions with various sample sizes that are used in simulation study.

The efficacy of the proposed approach (SSFB) was examined in terms of the number of outliers detected when the clean data was contaminated with certain amount of outliers. In order to see the sensitivity of the techniques to the outliers, number of the outliers detected by proposed SSFB, SDSF and Tukey's boxplot (TB) are compared.

The construction of fences using SSFB is mainly based on the cut off points which were determined using bootstrapping. In order to see the performance of bootstrapping using the existing methods in identifying outliers, determination of cut off points of SDSF and Tukey's boxplot (TB) using the proposed bootstrapping techniques, namely SDSFB and TBB were utilised.

Table 3. Twenty one different distributions with various sample sizes in simulation study

No.	Distribution	Sample size
1	N(0,1)	20
2	Log Normal(5,0.4)	20
3	Log Normal(5,0.6)	20
4	Log Normal(5,0.8)	20
5	$\chi^2(8)$	20
6	$\chi^2(4)$	20
7	$\chi^2(2)$	20
8	N(0,1)	50
9	Log Normal(5,0.4)	50
10	Log Normal(5,0.6)	50
11	Log	50

	Normal(5,0.8)	
12	$\chi^2(8)$	50
13	$\chi^2(4)$	50
14	$\chi^2(2)$	50
15	N(0,1)	100
16	Log Normal(5,0.4)	100
17	Log Normal(5,0.6)	100
18	Log Normal(5,0.8)	100
19	$\chi^2(8)$	100
20	$\chi^2(4)$	100
21	$\chi^2(2)$	100

The results of Figure 4 reveal that in the absence of outlier with 0% contamination, most of the methods had similar result where no observation was misidentified as outlier. The proposed SSFB method also did not flag any observation as outlier. However, TB marked two and three non-contaminated observations in average as outliers in χ_4^2 and χ_2^2 distributions, respectively. Besides that, TBB misclassified an observation as outlier in both χ_4^2 and χ_2^2 distributions.

When there exists single outlier, SSFB performed consistently with the detection of only one outlier in the data. However, the presence of outlier affected the performance of SDSF, SDSFB, TB and TBB. When there was single outlier in χ_4^2 and χ_2^2 data with size $n = 20$, SDSF, SDSFB and TB and TBB misidentified more than one outlier.

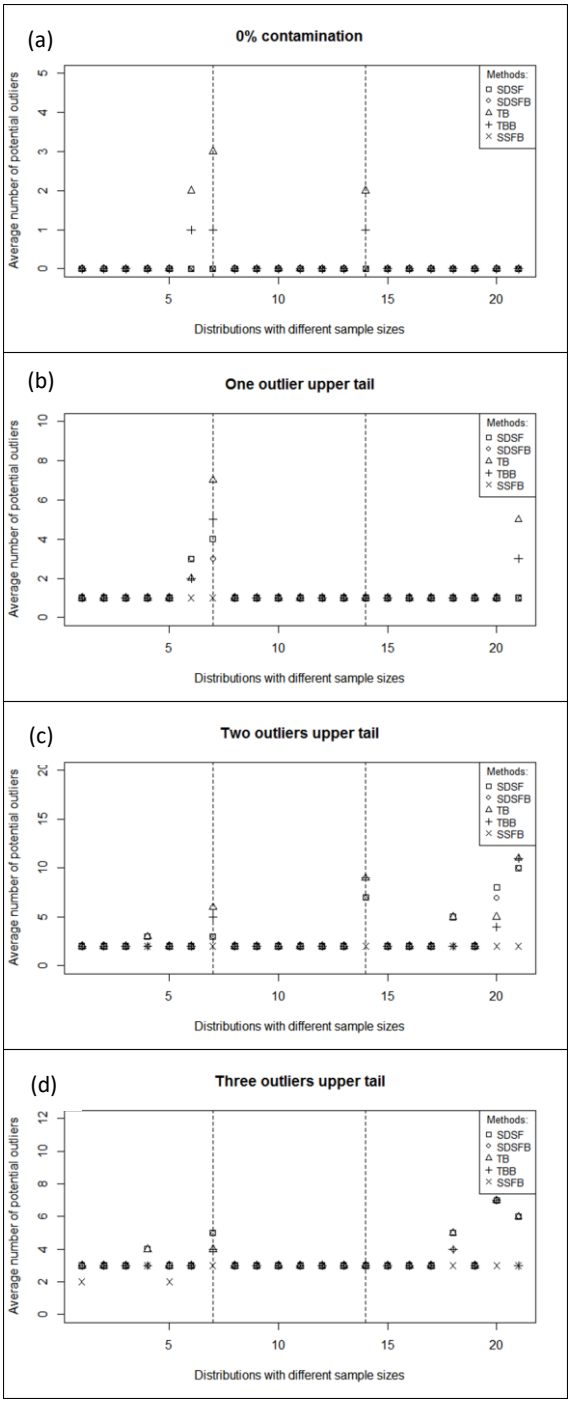

Figure 4. Average number of data beyond the upper and lower whiskers when (a) 0% contamination, (b) one outlier, (c) two outliers and (d) three outliers at upper tail

The benefits of the SSFB over other methods are even more apparent in data sets with medium and large size with multiple outliers. The SSFB method was substantially better in not misclassifying uncontaminated data as outlier while SDSF, SDSFB, TB and TBB methods were good in correctly identifying all the outliers but more likely to mislabel the uncontaminated observation as outlying observation. When

there were multiple outliers, SSFB still can identify the outliers correctly in symmetric and positively skewed distributions. There were only exceptions when $n = 20$ with three outliers where other approaches were slightly better than SSFB in correctly identifying all the real outliers. More specifically, for sample size 20 with three outliers, SSFB was only able to detect 2 outliers in both normal and χ^2_8 distribution which are symmetric and less skewed distribution. However, the other methods misidentified too many uncontaminated observations as outliers in skewed distributions such as log normal (5, 0.8), χ^2_4 and χ^2_2 in small and large size data.

The results of the simulation revealed that determination of fences using bootstrap can decrease the occurrence of misclassification uncontaminated data. For instance, when the sample size is 100 with two outliers in χ^2_4 distribution data, the SDSF detected 8 outliers while SDSFB classified only 7 outliers. Using the boxplot method, the TB identified 5 outliers while TBB detected 4 outliers. The rate of misclassification of uncontaminated observation as outliers of SDSFB and TBB are reduced compared to SDSF and TB.

The obtained fences were then applied to each bootstrap sample. Based on the number of outliers detected in each bootstrap sample, the robust mean and robust standard deviation were calculated using the proposed method TEB in two different ways. The first way is trimming one sided while the second way is trimming both sides equally when outliers are detected on either side of the distribution. This procedure was done in both clean data and screened contaminated data. Performances of the techniques on different scenarios can be observed in Figures 5-12.

Figure 5 and Figure 6 display bias and root mean square error (RMSE) of the one-sided trimmed mean of the proposed bootstrap method, SSFB, SDSFB and TBB at various number of outliers and sample size. In 0% contamination, it can be observed that after B = 2000 bootstrap resampling, most of the bias and RMSE produced by SSFB method are smaller than the SDSFB and TBB.

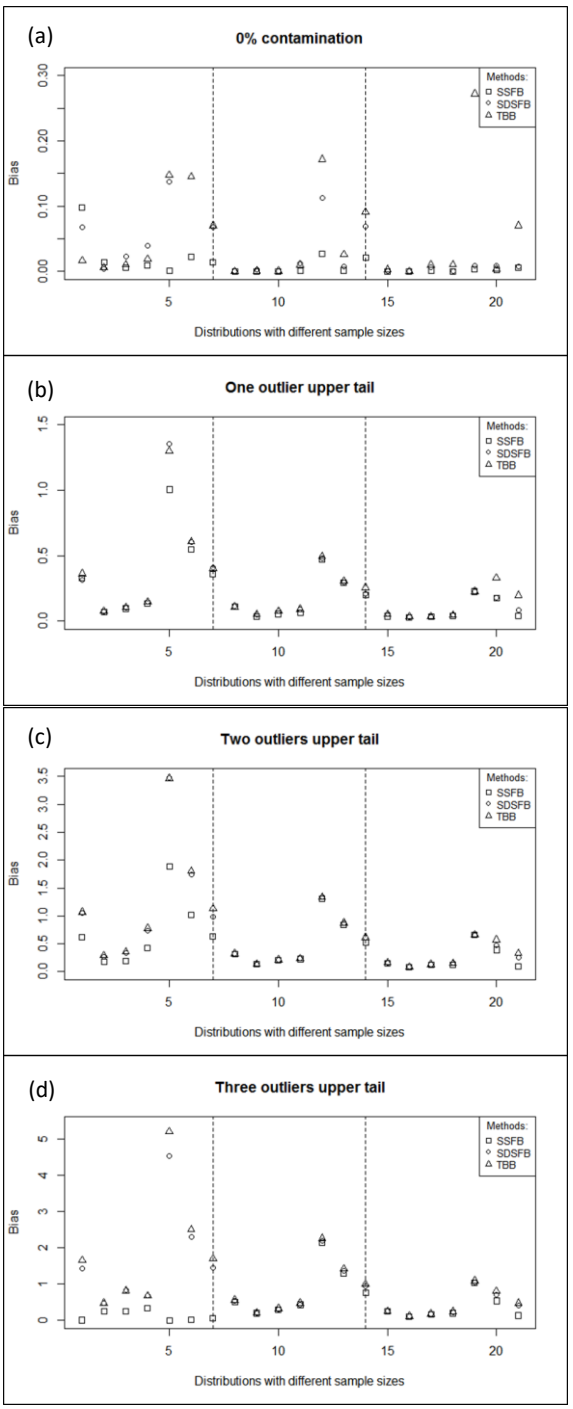

Figure 5. Average Bias for the One-sided Trimmed Mean for (a) 0% contamination, (b) one outlier, (c) two outliers, and (d) three outliers upper tail

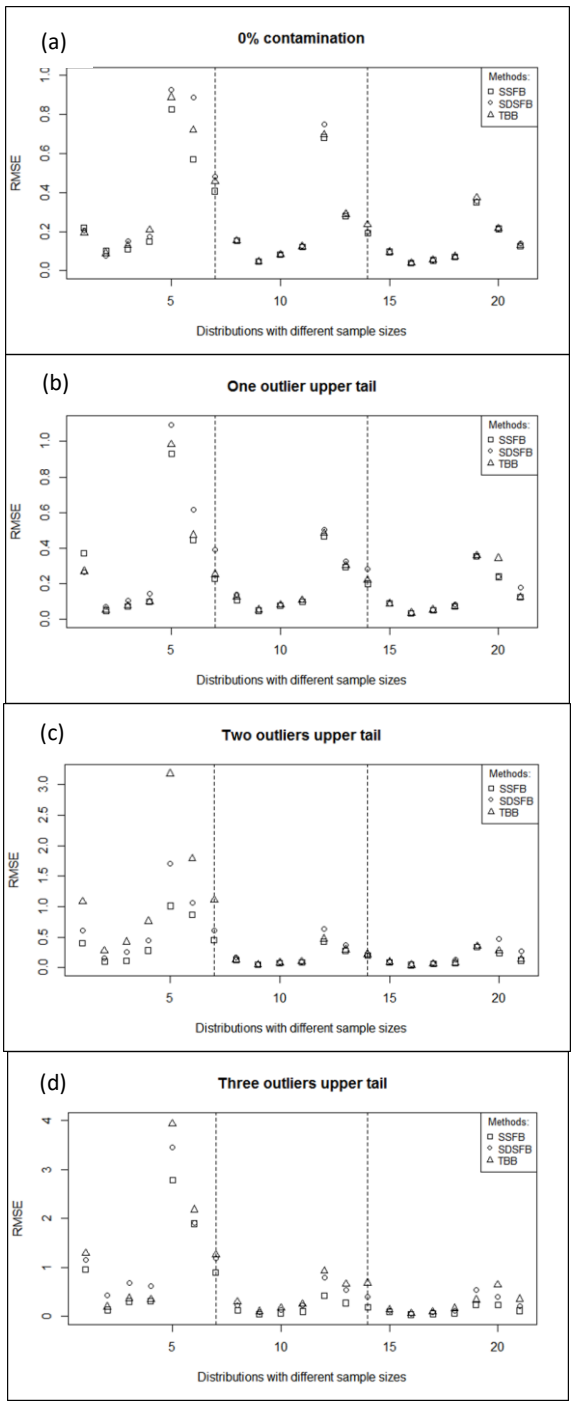

Figure 6. Average RMSE for the One-sided Trimmed Mean for (a) 0% contamination, (b) one outlier, (c) two outliers, and (d) three outliers upper tail

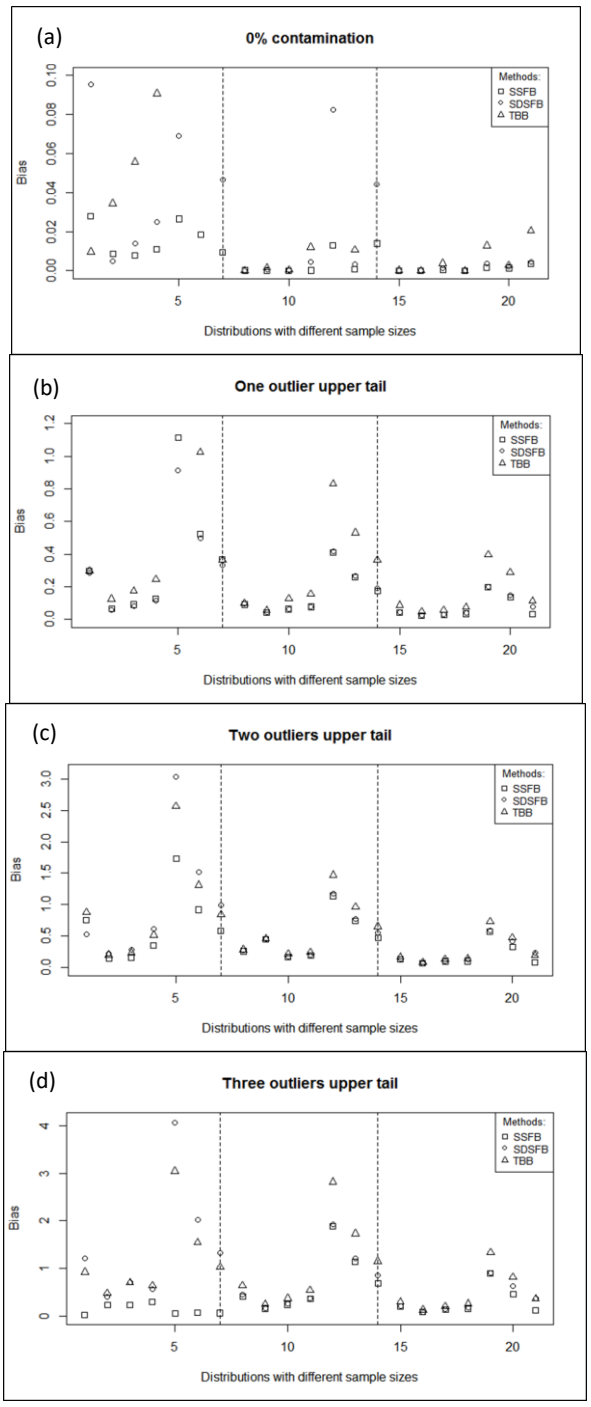

Figure 7. Average Bias for the Two-sided Trimmed Mean for a) 0% contamination, (b) one outlier, (c) two outliers, and (d) three outliers upper tail

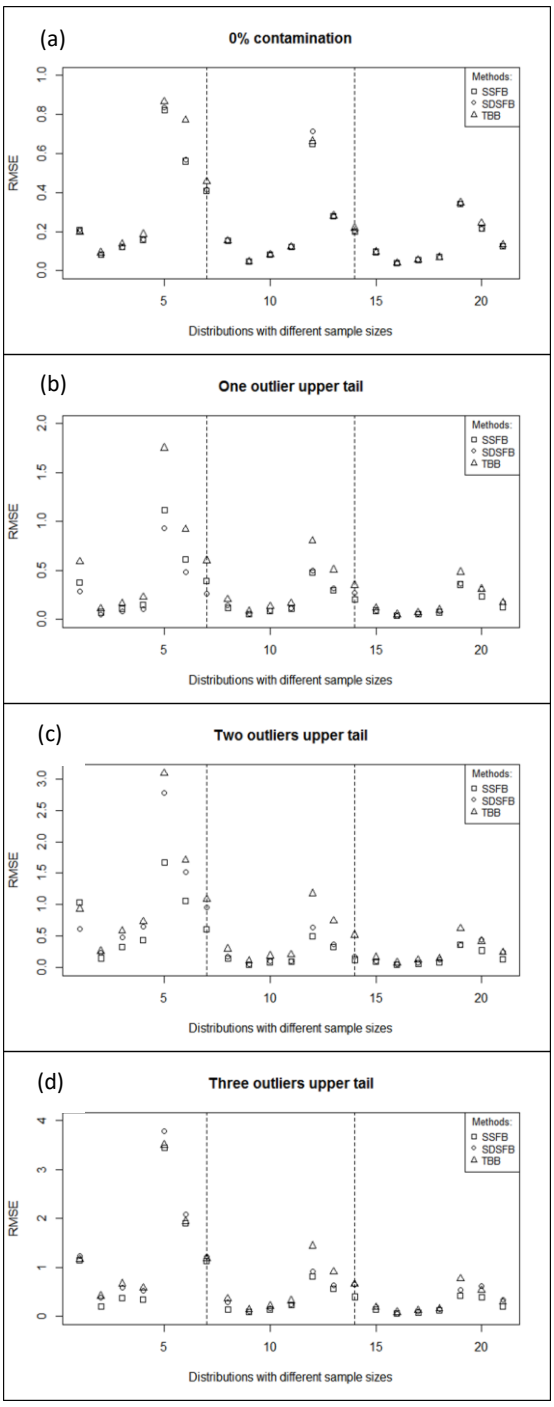

Figure 8. Average RMSE for the Two-sided Trimmed Mean for (a) 0% contamination, (b) one outlier, (c) two outliers, and (d) three outliers upper tail

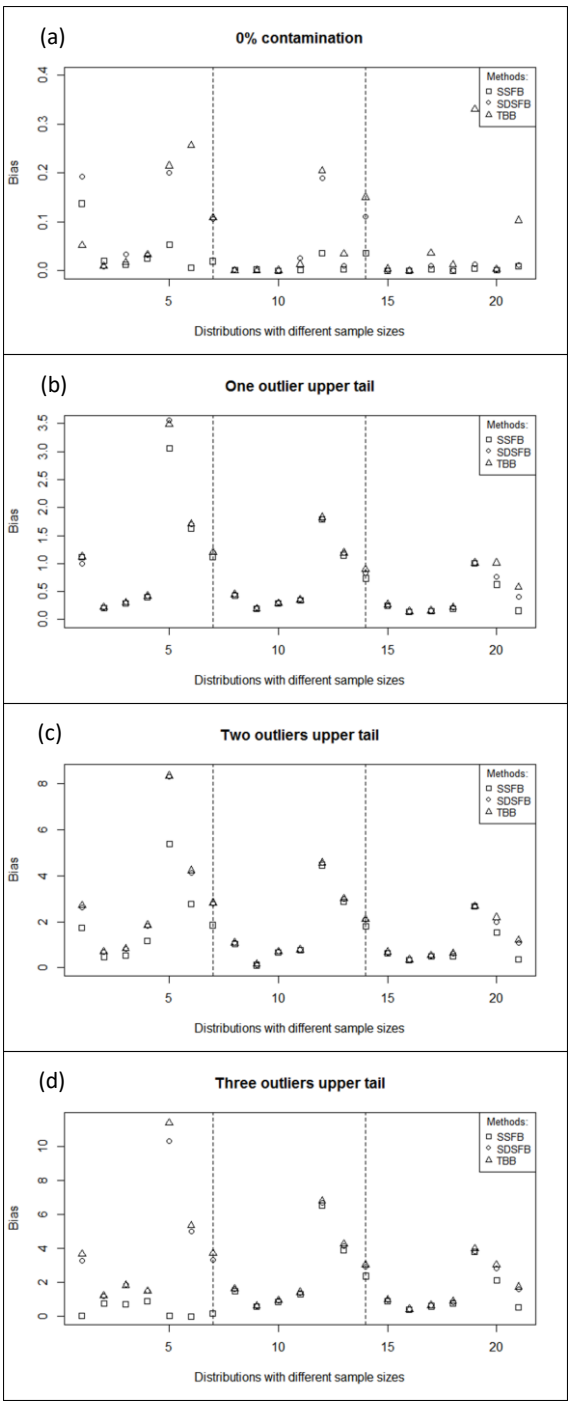

Figure 9. Average Bias for the One-sided Trimmed Standard deviation for (a) 0% contamination, (b) one outlier, (c) two outliers, and (d) three outliers upper tail

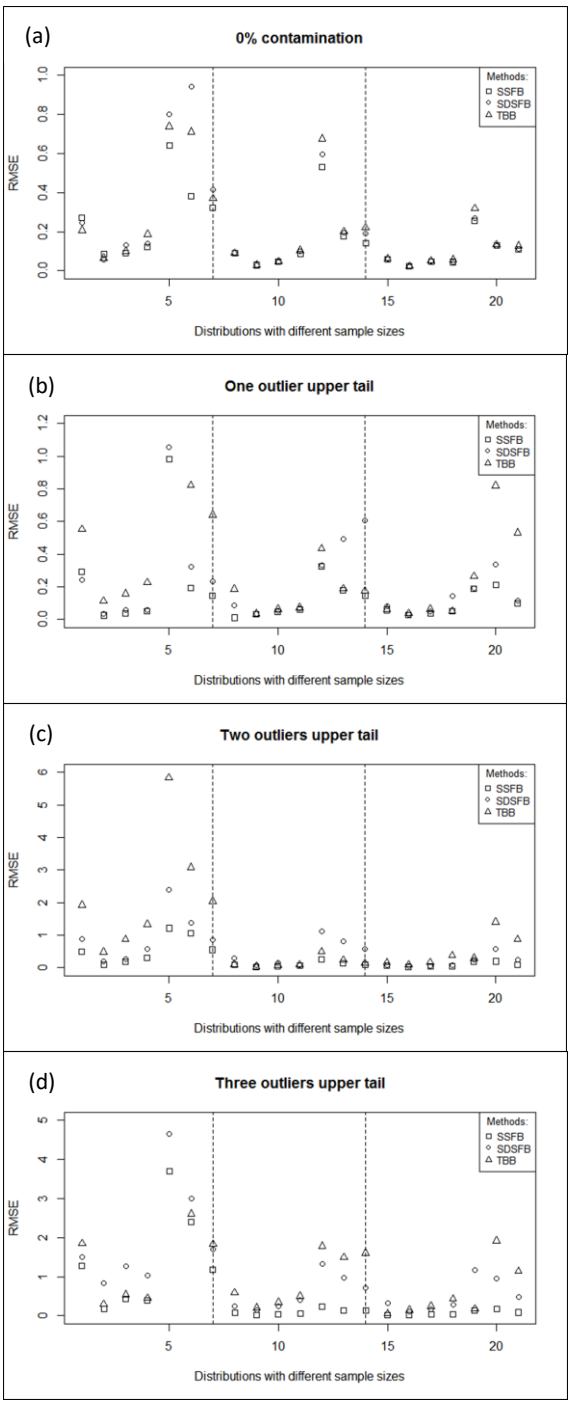

Figure 10. Average RMSE for the One-sided Trimmed Standard deviation for (a) 0% contamination, (b) one outlier, (c) two outliers, and (d) three outliers upper tail

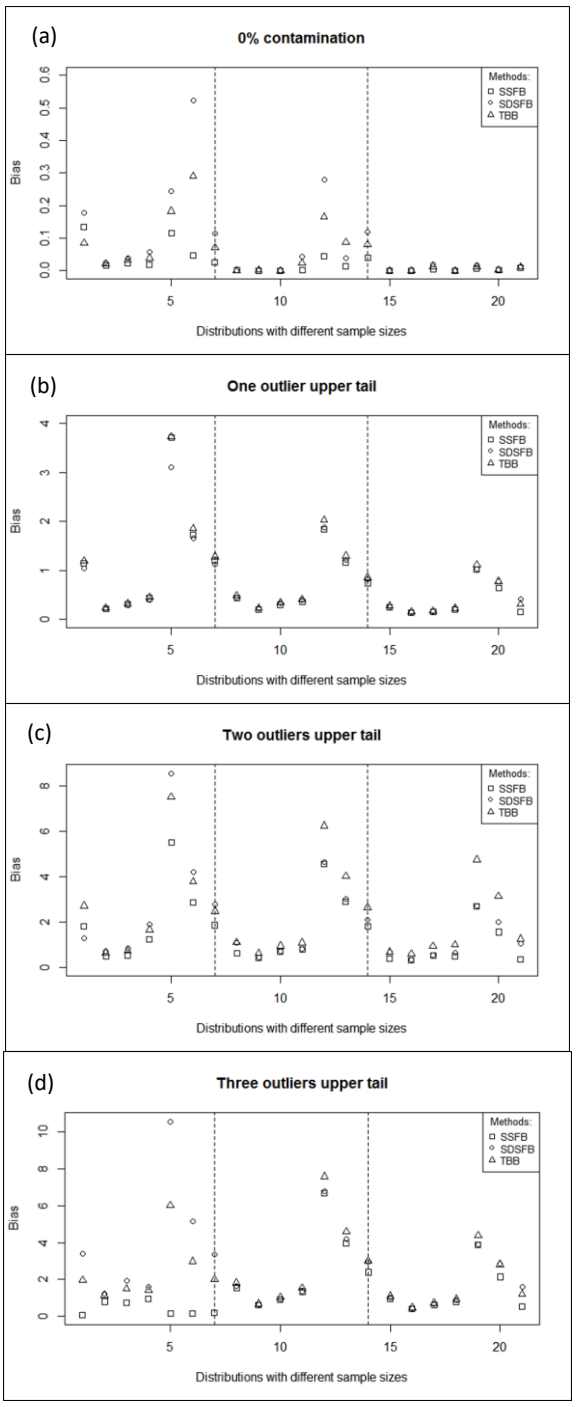

Figure 11. Average Bias for the Two-sided Trimmed Standard deviation for (a) 0% contamination, (b) one outlier, (c) two outliers, and (d) three outliers upper tail

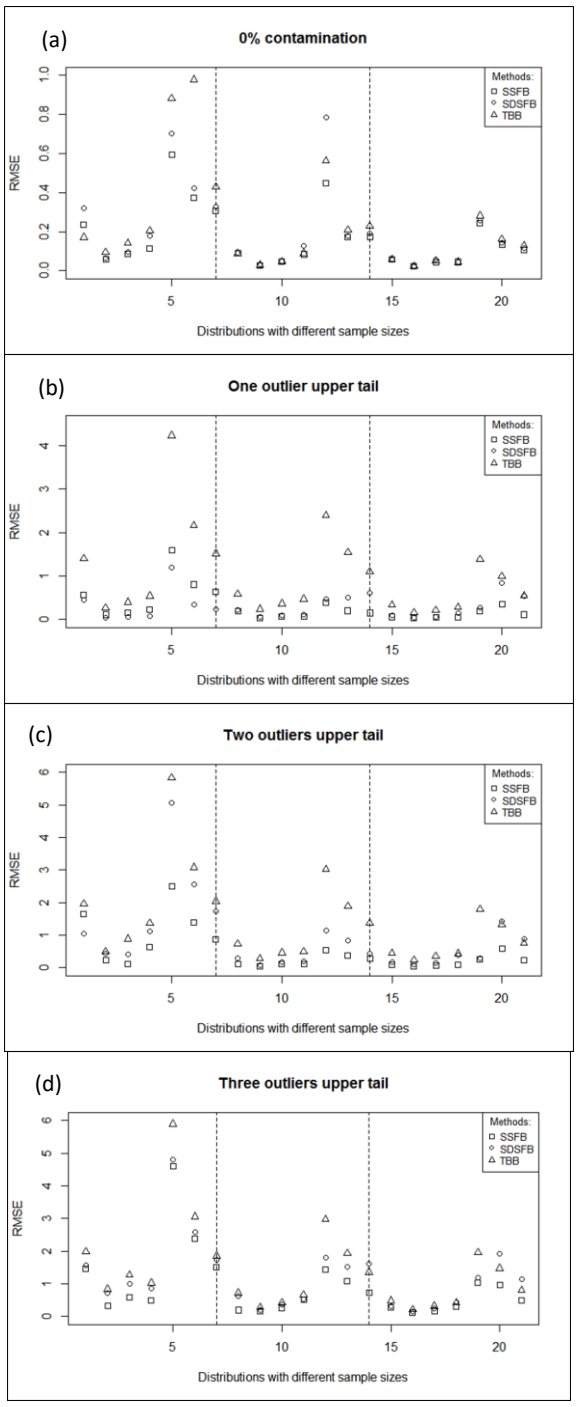

Figure 12. Average RMSE for the Two-sided Trimmed Standard deviation for (a) 0% contamination, (b) one outlier, (c) two outliers, and (d) three outliers upper tail

However, based on Figure 5 and Figure 6, at sample sizes of $n = 20$ and $n = 50$ with absence of outlier, TBB showed substantially lower bias and RMSE when the data is normally distributed. Moreover, SDSFB also showed closer mean estimator for log normally distributed data by producing lower bias and RMSE. The performance of SSFB was obviously improved at the larger number of outliers compared to SDSFB and TBB. There was only one exception when sample size of 20 with single outlier in normally distributed data where the SDSFB showed lower bias and RMSE.

Figure 7 and Figure 8 present bias and root mean square error (RMSE) of the two-sided trimmed mean. SSFB showed lower bias and RMSE in estimating two-sided trimmed mean. Thus, the estimated two-sided trimmed means based on SSFB method are closer to the two sided trimmed mean of the sample data compared to SDSFB and TBB methods. When there were three outliers in various sample sizes, SSFB performed substantially better by showing lower bias and RMSE in different skewed distribution data.

Similarly, a general pattern can be seen for estimating trimmed standard deviation for trimming one-sided and two-sided cases. Lower bias and RMSE of proposed SSFB method can be observed from results as shown in Figure 9 -12. TBB was better in estimating trimmed standard deviation when there are no outliers in normally distributed data with small and medium sample sizes. In the absence of outlier, SDSFB presented better performance with lower bias and RMSE in log normally distributed data with sample size 100 and also in log normally distributed data with sample size 20.

In summary, the results of bootstrap resampling signify that the SSFB technique performs well in identification of outliers and better estimation of robust estimators in most of the scenarios.

In overall, SSFB performed substantially better compared to SDSFB and TBB by showing closer estimation values to the sample estimators. The outstanding performances of SSFB over other techniques are more obvious in data sets with larger number of outlying observations and skewed distribution data.

V. CONCLUSION AND RECOMMENDATIONS

In this study, a new algorithm of sequential fences involving bootstrap resampling with some adjustments was proposed. Adjustment was made to SDSF using split sample method to increase the coverage of the data that lie on either side of the tails. Thus, this can increase the flexibility of the sequential fence in identifying outliers in different distributions and various sizes of data. Next is the determination of the cut-off

points in sequential fences are based on the bootstrap resampling technique.

The result of outlier detection shows that SSFB was better in detecting the outlier without misdetection of outliers whereas SDSFB and TBB are better in correctly identify the outliers but more likely to misidentify uncontaminated observations as outlier. On the other hand, with the use of bootstrap resampling technique, SDSFB and TBB emerges to be more efficient in identifying the outliers as the rate of misclassification of the outliers was reduced.

Based on the number of outliers detected in each bootstrap resample, the suspected outliers were trimmed by one-sided or two-sided from the tails in order to estimate the trimmed mean and trimmed standard deviation. TBB method presented better bias and RMSE when no outlier and more likely applicable to small and normally distributed data. SDSFB method performed better when there are single outliers in small size data. Apparently, in overall, it can be seen that SSFB consistently showed the lowest bias and RMSE among the methods in skewed distributions with higher number of outliers. In summary, the overall results indicate that SSFB approach performs well in most of the circumstances. In future, research regarding SSFB can be extended to identify outliers in bivariate or multivariate cases.

VI. REFERENCES

Ahmad, S, Midi, H & Norazan, MR 2011, 'Diagnostics for residual outliers using deviance component in binary logistic regression', World Applied Sciences Journal, vol. 14, no. 8, pp. 1125-1130.

Aucremanne, L, Brys, G, Hubert, M, Rousseeuw, PJ & 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. 13-25.

Aslam M & Khurshid A 1991, 'Shape-finder box plots', ASQC Statistics Division Newsletter, pp. 9–11.

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.

Beckman, RJ & RD Cook 1983, 'Outlier….s', Technometrics, vol. 25, pp. 119-149.

Boos, DD & Stefanski, LA 2013, 'Essential statistical inference: Theory and methods'.

Capéràa, P & Rivest, LP 1995, 'On the variance of the trimmed mean', Statistics Probability Letters, vol. 22, no. 1, pp. 79-85.

Choonpradub, C & McNeil, D 2005, 'Can the box plot be improved?', Songklanakarin Journal of Science and Technology, vol. 27, no. 3, pp. 649–657.

Davies, L & Gather, U 1993, 'The identification of multiple outliers', Journal of the American Statistical Association, vol. 88, pp. 782-792.

Efron, B 1979, 'Bootstrap methods: another look at the Jackknife', The Annals of Statistics, vol. 7, no. 1, pp. 1-26.

Fitrianto, A & Midi, H 2011, 'Procedures of generating a true clean data in simple mediation analysis', World Applied Sciences Journal, vol. 15, no. 7, pp. 1046-1053.

Frigge, M, Hoaglin, DC & Iglewicz, B 1989, 'Some implementations of the boxplot', The American Statistician, vol. 43, no. 1, pp. 50-54.

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

Shevlyakov, G, Andrea, K, Choudur, L, Smirnov, P, Ulanov, A & Vassilieva, N 2013, 'Robust versions of the Tukey boxplot with their application to detection of outliers', 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.

Gross, A 1976, 'Confidence interval robustness with long-tailed symmetric distribution', Journal of the American Statistical Association, vol. 71, no. 354, pp. 409-416.

Hansen, CM, Evans, MA & Shultz, TD 1999, 'Application of the bootstrap procedure provides an alternative to standard statistical procedures in the estimation of the vitamin B-6 requirement', The Journal of Nutrition, vol. 129, no. 10, pp. 1915–1919.

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

Hyndman, RJ & Fan, Y 1996, 'Sample quantiles in statistical packages', The American Statistician, vol. 50, no. 4, pp. 361.

Iftikhar, HA 2011, 'Robust outlier detection techniques for skewed distributions and applications to real data', Doctoral thesis, International Islamic University, Islamabad, Pakistan.

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

Lix, LM & Keselman, HJ 1998, 'To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality', Educational and Psychological Measurement, vol. 58, no. 3, pp. 409-429.

Marmolejo, RF & Tian, TS 2010, 'The shifting boxplot: A boxplot based on essential summary statistics around the mean', International Journal of Psychological Research, vol. 3, no. 1, pp. 37–45.

Nelson, AC, Armentrout, DW & Johnson, TR 1980, 'Validation of air monitoring data', EPA-600/4-80-030. U. S. Environmental Protection Agency, Research Triangle Park, N. C.

Denise, L 2021, 'The history of bootstrapping: tracing the development of resampling with replacement,' The Mathematics Enthusiast, vol. 18, no. 1.

Samik, R 2008, 'Introduction to Monte Carlo simulation', Winter Simulation Conference, 2008, pp. 91-100. doi: 10.1109/WSC.2008.4736059.

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

Singh, K & Xie, M 2003, 'Bootlier-plot: bootstrap based outlier detection plot', Sankhyā: The Indian Journal of Statistics (2003-2007), vol. 65, no. 3, pp. 532-559.

Tabachnick, BG & Fidell, LS 2001, Using Multivariate Statistics, 4th edn, New York: Boston and Bacon.

Tukey, JW 1977, Exploratory Data Analysis. New York: Addison-Wesley.

Wilcox, RR 1995, ANOVA: A paradigm for low power and misleading measures of effect size? Review of Educational Research, vol. 65, no. 1, pp. 51-77.

Wong, HS & Fitrianto, A 2019, 'Adjusted sequential fences for detecting univariate outliers in skewed distributions', ASM Science Journal, vol. 12, no. 5, pp. 107–115.