Thailand Statistician
July 2021; 19(3): 450-471
http://statassoc.or.th
Contributed paper

Robust Outliers Detection Method for Skewed Distribution

Prem Junsawang, Mintra Promwongsa and Wuttichai Srisodaphol*
Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding author; e-mail: wuttsr@kku.ac.th

Received: 18 June 2020
Revised: 4 October 2020
Accepted: 7 December 2020

Abstract

The aim of this study is to propose the robust outliers detection method called MH boxplot for skewed distribution. The proposed method is modified from Hubert's boxplot by embedding the Bowley coefficient, the ratio of lower split interquartile range and upper split interquartile range into the fences of the boxplot. The performance of the boxplot is evaluated by the percentage of outlier ratio mean in three cases of simulated data (truncated, uncontaminated and contaminated data) and real data. Furthermore, the existing boxplots for outliers detection are used to make a comparison with the MH boxplot as well. The results from simulated and real data show that the MH boxplot efficiently detects outliers and is robust to skewness of data over the other boxplots for any sample size. Moreover, the MH boxplot efficiently detects outliers as the shape of real data.

Keywords: Robust outlier detection, skewed data, split interquartile range.

1. Introduction

An outlier is an observation which differs or deviates so much from the other observations. The outliers could be very large or very small when it is compared to the other in the data set. The outliers might occur from incorrect measurements, including data entry errors, or different population. Outliers might have a negative influence on the real data characteristics, e.g. outliers go against the normality of data, increases in variance value and reduces the power of statistical hypothesis tests. Therefore, the outliers detection methods play an important role in data preprocessing step to filter them out before further data analysis. In recent years, outliers detection methods have been developed to detect and remove them from the original data. One type of outliers detection methods is the outliers labelling techniques or informal test in which potential outliers could be considered as extreme values (Iglewicz and Hoaglin 1993). The main idea of outliers labelling methods is to construct an interval for detecting observations which are outside the interval and then are labelled as outliers. The type of these methods is also considered as statistical outliers detection approach. A suitable interval is constructed by various location and scale parameters without hypothesis testing.

For univariate data, one of the traditional and popular methods for outliers detection is a boxplot, which was introduced by Tukey (1977). The outliers are labelled by the observations outside a defined interval called fences such as $[\mathrm{Q}_1 - 1.5\mathrm{IQR}\,, \mathrm{Q}_3 + 1.5\mathrm{IQR}]$ where Q_1, Q_3 and IQR stand for the first and the third quartiles, and the interquartile range, respectively. Several researchers have reported that the Tukey's boxplot is fitted to symmetrical data (Walker and Chakraborti 2013; Adil and Irshad 2015; Zhao and Yang 2019). For skewed data, especially, there are too many observations as being potential outliers. Generally, some of the marked observations were presumed to occur

naturally in skewed data rather than the real outliers Hubert and Vandervieren (2008). For instance, for symmetrical data, the lower and upper fences, obtained from Tukey's boxplot for standard normal distribution, contain 99.3% of all observations, approximately. Hence, the left observations outside the fences are labelled as outliers. While, for skewed data, the fences, obtained from Tukey's boxplot for χ^2 distribution with one degree of freedom, approximately contain 92.44% of all observations and so, 7.56% of the data are outliers which is rather more than usual.

To enhance the efficiency of boxplot-based outliers detection for skewed data, a variety of boxplot techniques has been proposed in the literature. Kimber (1990) proposed the fences of boxplot for skewed data, called split interquartile range (SIQR), in which a position of the split was at the median of the data. Carling (2000) replaced Q_1 and Q_3 in Tukey's fences by the median and mentioned that the constant 1.5 fold of IQR should be varied and depended on sample size. To reduce the effect of the sample size on the number of detected outliers, he proposed a reasonable constant of 2.3 instead of 1.5. Barnett and Cohen (2000) proposed the modified boxplot based on lognormal distribution to solve problems of right censoring with high skewness in lifetime data. Hubert and Vandervieren (2008) proposed the adjusted boxplot by using a robust measure of skewness, namely a medcouple (MC) which was introduced by Brys et al. (2004). In their work, they also used the families of skewed distributions for choosing the appropriate constant to insert into exponential terms of fences for efficient applying with skewed data. Walker and Chakraborti (2013) extended Tukey's fences based on SIQR to insert the ratios of SIQR for skewed data. Adil and Irshad (2015) proposed the modified boxplot for solving extreme fences problem by incorporating a moment coefficient of skewness to construct lower and upper fences. Babura et al. (2017) extended the adjusted Hubert's boxplot by using the Bowley coefficient which is a robust measure of skewness and they estimated the constant on lower and upper fences by conducting the simulation on extreme data from Generalized Extreme Value (GEV) distribution. Recently, Promwongsa et al. (2018) proposed a variation of Kimber, called MK. In their work, the lower and upper fences were modified by using the ratio of lower and upper SIQR.

In this study, we propose a modified boxplot by using the Bowley coefficient with the ratio of SIQR for constructing the proper fences which are robust to data skewness. The modified boxplot improves the performance of detecting outliers with any data regardless of the distribution. For evaluation, simulated data of both symmetric and skewed data distributions are generated. Moreover, real data sets in various situations are also tested. The standard and popular boxplot-based methods are used to make a comparison with the proposed boxplot, as well.

2. The Proposed Boxplot

Let $X_n = \{x_1, x_2, \ldots, x_n\}$ be a set of n univariate samples. The skewness of the univariate data could be measured by computing medcouple (MC) value (Brys et al. 2004). If $MC > 0$, then the distribution of X_n is right-skewed and if $MC < 0$, then the distribution of X_n is left-skewed. MC is mathematically expressed as follows:

$$\mathrm{MC} = \underset{x_i \le m_n \le x_j}{\mathrm{med}} h(x_i,\, x_j),\ \textit{for}\ x_i \neq x_j,$$

where m_n is the median of X_n, and $h(x_i,\, x_j)$ is a kernel function given by

$$h(x_i,\, x_j) = \frac{(x_j - m_n) - (m_n - x_i)}{x_j - x_i}. \tag{1}$$

For the special case $x_i = m_n = x_j$, the kernel function is defined as follows. Let $m_1 < m_2 < \ldots < m_k$ be the indices of the observations which are tied to the median m_n, i.e. $x_{m_l} = m_n$ for all $l = 1, 2, ..., k$. Then,

$$h(x_{m_i},\, x_{m_j}) = \begin{cases} -1 & \text{if } i + j - 1 < k, \\ 0 & \text{if } i + j - 1 = k, \\ +1 & \text{if } i + j - 1 > k. \end{cases}$$

In Hubert and Vandervieren (2008), the lower and upper fences for the right-skewed data are defined by

$$[\mathrm{Q}_1 \ - 1.5\mathrm{e}^{-4\mathrm{MC}}\mathrm{IQR} \ , \ \mathrm{Q}_3 + 1.5\mathrm{e}^{3\mathrm{MC}}\mathrm{IQR}], \tag{2}$$

and the fences for the left-skewed data are defined by

$$[\mathrm{Q}_1 \ - 1.5\mathrm{e}^{-3\mathrm{MC}}\mathrm{IQR} \ , \ \mathrm{Q}_3 + 1.5\mathrm{e}^{4\mathrm{MC}}\mathrm{IQR}], \tag{3}$$

where Q_1, Q_3 and IQR stand for the first and the third quartiles, and the interquartile range, respectively. Kimber (1990) introduced a lower and upper split interquatile range, named SIQR_L and SIQR_U, respectively, by splitting IQR at the median location for expressing the spread of the data.

In this work, the proposed method is modified from the Hubert boxplot by substituting the exponent of exponential terms with the ratio between SIQR_L and SIQR_U, and Bowley coefficient (δ) in (2.2) and (2.3). The proposed method is called Modified Hubert or MH boxplot and given by

$$\left[\mathrm{Q}_1 - 1.5\mathrm{e}^{\left(\frac{\mathrm{SIQR}_\mathrm{L}}{\mathrm{SIQR}_\mathrm{U}}\delta\right)}\mathrm{IQR} \ , \quad \mathrm{Q}_3 + 1.5\mathrm{e}^{\left(\frac{\mathrm{SIQR}_\mathrm{U}}{\mathrm{SIQR}_\mathrm{L}}\delta\right)}\mathrm{IQR}\right], \tag{4}$$

where $\mathrm{SIQR}_\mathrm{L} = \mathrm{Q}_2 - \mathrm{Q}_1$, $\mathrm{SIQR}_\mathrm{U} = \mathrm{Q}_3 - \mathrm{Q}_2$, and $\delta = \frac{\mathrm{Q}_3+\mathrm{Q}_1-2\mathrm{Q}_2}{\mathrm{Q}_3-\mathrm{Q}_1}$.

The proposed MH boxplot is capable of adapting automatically fences to the shape of data, since the data are right-skewed, then the upper fence is longer, and the lower fence is shorter. Otherwise, the data are left-skewed, then the upper fence is shorter, and the lower fence is longer.

3. Experiments

For performance evaluation, the proposed MH boxplot method was evaluated on both simulated and real univariate data sets with the various distributions. Six standard and popular boxplot-based methods for outliers detection including Tukey (1977), Kimber (1990), Hubert and Vandervieren (2008), Walker and Chakraborti (2013), Adil and Irshad (2015), and Promwongsa et al. (2018) called MK were used to make a comparison with the proposed MH method.

Case I: Truncated data: n simulated samples of each of eight distributions, namely $N(0,1)$, χ^2_1, χ^2_5, χ^2_{20}, $F_{(10,10)}$, $F_{(10,90)}$, $F_{(90,10)}$, and $F_{(90,90)}$, were generated. 40% out of data including 20% on the leftmost and rightmost were trimmed. The experiments were conducted in the following steps:

Step 1: **For** each distribution $\tilde{D} \in \{N(0,1), \chi^2_1, \chi^2_5, \chi^2_{20}, F_{(10,10)}, F_{(10,90)}, F_{(90,10)}, F_{(90,90)}\}$ **do** Steps 2-15.
Step 2: **For** each method $i \in \{1, 2, \ldots, 7\}$
Step 3: **For** $m = 1, 2, \ldots, M$ **do** Steps 4-12.
Step 4: Simulate a set of n samples $X = \{x_1, x_2, \ldots, x_n\}$, where $X \sim \tilde{D}$.
Step 5: Sort the samples in X in ascending order.
Step 6: Create a data set $X_{trimmed} = \{x'_1, x'_2, \ldots, x'_{n_{trimmed}}\}$ by trimmimg 40% out of data in X.
Step 7: Compute the lower fence (l_{im}) and the upper fence (u_{im}) based on $X_{trimmed}$.
Step 8: **For** each $x \in X_{trimmed}$ **do**
Step 9: **If** $x < l_{im}$ or $x > u_{im}$ **then** label x as an outlier.
Step 10: **End** For in Step 8.
Step 11: Count the number of detected outliers (do_{im}).
Step 12: Compute outlier ratio (or_m) by

$$or_m = \frac{do_{im}}{n_{trimmed}}$$

Step 13: **End** For in Step 3.
Step 14: Compute the percentage of the i^{th} outlier ratio mean $(\bar{or}_i\%)$ by

$$\bar{or}_i\% = \frac{\sum_{m=1}^{M} or_m}{M} \times 100\%$$

Step 15: **End** For in Step 2.
Step 16: **End** For in Step 1.

Case II: Uncontaminated data: n simulated samples of each of eight distributions, namely $N(0,1)$, χ^2_1, χ^2_5, χ^2_{20}, $F_{(90,10)}$, $F_{(10,90)}$, $F_{(90,90)}$, and $F_{(10,10)}$, were generated. The experiments were conducted in the following steps:

Step 1: **For** each distribution $\tilde{D} \in \{N(0,1), \chi^2_1, \chi^2_5, \chi^2_{20}, F_{(90,10)}, F_{(10,90)}, F_{(90,90)}, F_{(10,10)}\}$, **do** Steps 2-13.
Step 2: **For** each method $i \in \{1, 2, \ldots, 7\}$
Step 3: **For** $m = 1, 2, \ldots, M$ **do** Steps 4-10.
Step 4: Simulate a set of n samples $X = \{x_1, x_2, \ldots, x_n\}$, where $X \sim \tilde{D}$.
Step 5: Compute the lower fence (l_{im}) and the upper fence (u_{im}) based on X.
Step 6: **For** each $x \in X$ **do**
Step 7: **If** $x < l_{im}$ or $x > u_{im}$ **then** label x as an outlier.
Step 8: **End** For in Step 6.
Step 9: Count the number of detected outliers (o_{im}).
Step 10: Compute an outlier ratio (or_m) by

$$or_m = \frac{o_{im}}{n}$$

Step 11: **End** For in Step 3.
Step 12: Compute the percentage of the i^{th} outlier ratio mean $(\bar{or}_i\%)$ by

$$\bar{or}_i\% = \frac{\sum_{m=1}^{M} or_m}{M} \times 100\%$$

Step 13: **End** For in Step 2.
Step 14: **End** For in Step 1.

Case III: Contaminated data: n simulated samples of each of eight distributions, namely $N(0,1)$, χ^2_1, χ^2_5, χ^2_{20}, $F_{(90,10)}$, $F_{(10,90)}$, $F_{(90,90)}$, and $F_{(10,10)}$, were generated. For symmetrical data, $r\%$ out of samples on both the leftmost and rightmost half by half were contaminated. For skewed data, $r\%$ out of samples on either the leftmost or rightmost are contaminated. The experiments were conducted in the following steps:

Step 1: **For** each distribution $\tilde{D} \in \{N(0,1), \chi^2_1, \chi^2_5, \chi^2_{20}, F_{(90,10)}, F_{(10,90)}, F_{(90,90)}, F_{(10,10)}\}$, **do** Steps
Step 2: **For** each method $i \in \{1, 2, \ldots, 7\}$
Step 3: **For** $m = 1, 2, \ldots, M$ **do** Steps 4-10.
Step 4: Simulate a set of n samples $X = \{x_1, x_2, \ldots, x_n\}$, where $X \sim \tilde{D}$.
Step 5: Sort the samples in X in ascending order.
Step 6: **If** $\tilde{D}$ is symmetric **then** go to Step 7, **else** go to Step 8.
Step 7: $\frac{r}{2}\%$ of the lower and upper tails of the data in X are multiplied by a constant c and $c > 0$, and go to Step 9.

Step 8: r% of the lower and upper tails of the data in X are multiplied by a constant c and $c > 0$.
Step 9: Compute the lower fence (l_{im}) and the upper fence (u_{im}).
Step 10: **For** each $x \in X$ **do**
Step 11: **If** $x < l_{im}$ or $x > u_{im}$ **then** label x as an outlier.
Step 12: **End** For in Step 10.
Step 13: Count the number of detected outliers o_{im}.
Step 14: Compute outlier ratio or_m by

$$or_m = \frac{o_{im}}{n}$$

Step 15: **End** For in Step 3.
Step 16: Compute the percentage of the i^{th} outlier ratio mean $(\bar{or}_i\%)$ by

$$\bar{or}_i\% = \frac{\sum_{m=1}^{M} ro_m}{M} \times 100\%$$

Step 17: **End** For in Step 2.
Step 18: **End** For in Step 1.

3.1. Simulated data sets

In this section, the experiments on three different scenarios of simulated data, including truncated, uncontaminated and contaminated data, were conducted. The evaluation was not only performed on symmetrical data but also skewed data. For symmetrical data, standard normal distribution $N(0, 1)$ was selected as a case study. For skewed data, χ^2 and F distributions with having mildly and moderately skewed levels were selected. By considering the coefficient of skewness, the mildly skewed distributions consisted of χ^2_5, χ^2_{20}, $F_{(10,90)}$ and $F_{(90,90)}$. The moderately skewed distributions consisted of χ^2_1 , $F_{(10,10)}$ and $F_{(90,10)}$. Let $\tilde{D}$ and M be a given distribution and the number of simulation times, respectively.

3.2. Real data sets

In this section, four real data sets were used to evaluate the performance of the proposed MH method. The description of each data set is briefly explained as follows:

1) Coal mine data set (Jarrett 1979) contains 190-time intervals in days between explosions in coal mines from 15^{th} March 1851 to 22^{nd} March 1962 inclusive.
2) Mississippi River Maximum Daily Discharge data set (Gumbel 1941) contains the maximum daily discharge of the Mississipi river for 50 years from 1890-1939.
3) Indian Liver Patient data set (Ramana et al. 2012) contains Alamine Aminotransferase of 416 liver patients and 167 non-liver patients that were collected from the northeast of Andhra Pradesh, India.
4) Facebook metrics data set (Moro et al. 2016) contains the number of people who clicked anywhere in all posts published in the Facebook's page of 500 worldwide renowned cosmetic brands between January 1^{st}, 2014 to December 31^{st}, 2014.

From the above four data sets, we created a histogram and a plot of the sorted data points for each data set to identify that which observations were far from the most of observations visually and they were flagged as potential outliers. Afterwards, we computed the lower and upper fences of each boxplot for detecting outliers. For evaluation, the numbers of the detected outliers obtained from all boxplots were compared to the number of potential outliers. The descriptive statistics for each data set such as minimum (min), maximum (max), mean $(\bar{x})$, median (med), the first quartile (Q_1), the third quartile (Q_3), medcouple (MC), and Bowley coefficient (δ) were computed and given in Table 1.

Table 1 Descriptive statistics for each data set

Data set	min	max	$\bar{x}$	med	Q_1	Q_3	MC	δ
Coal mine	0	2,366	213.42	113.50	37.75	270	0.40	0.35
Daily discharge	760	2,334	1,355.78	1,355	1,063.25	1,507	-0.13	-0.32
Alamine	10	4,929	109.91	25	42	87	0.54	0.46
Number of clicks	9	11,328	798.78	551.50	332.50	955.50	0.36	0.30

3.3. Experimental results on the simulated data

In this section, the simulated data with sample size *n* for each distribution were generated, repeatedly and independently. The number of repetitions (*M*) was 100. The samples size (*n*) was varied from 15, 20, 25, 30, 50, 100, 250, 500, 750 and 1000. The sample size n was considered as small size if $n \leq 30$ (n = 15, 20, 25 and 30). Otherwise, it was considered as a large size (n = 50, 100, 250, 500, 750 and 1000). For evaluation, the percentage of outlier ratio mean $(\bar{o}r\%)$ was computed. The average and range statistics of $\bar{o}r\%$ for small and large sample sizes on each distribution were measured. The range statistic was computed from the maximum value minus minimum value of $\bar{o}r\%$ in each group of small or large size. All experiments were implemented by *R* programming. The experimental results of each case are given as follows:

Case I: Truncated data

For each of simulation, 40% out of samples including 20% on the leftmost and rightmost are trimmed. Since the trimmed data were generated by the given distribution, the average of outlier ratio mean ($\bar{o}r\%$) should be zero in this case. The best average of $\bar{o}r\%$ among seven methods is the value closest to zero. In Table 2, the first, second and third averages of $\bar{o}r\%$ are boldface and noticed by the superscript with the number 1, 2 and 3 in parentheses, respectively.

From Table 2, it is shown that for symmetric distribution, Turkey's, Adil's and Kimber's methods are of the first three best averages of $\bar{o}r\%$ in both small and large sample sizes. The average of $\bar{o}r\%$ values of proposed MH are slightly less than these of Kimber and Adil in large sample size. For moderately skewed distribution, the averages of $\bar{o}r\%$ values of Adil, the prpposed MH, and Tukey are of the first three bests in both small and large sample sizes, respectively. For mildly skewed distribution, the averages of $\bar{o}r\%$ of Turkey and Adil are of the first two bests in both small and large sizes. The averages of $\bar{o}r\%$ of the proposed MH and Kimber are quite small, especially in large sample size. The individual average of $\bar{o}r\%$ of each method for each distribution in small and large sample sizes are shown in Figures 1 and 2 in which the width of a sign shows the range of the $\bar{o}r\%$ and the red dot line is located at the optimal value of the average $\bar{o}r\%$.

Case II: Uncontaminated data

As suggested in Adil and Irshad (2015), they assumed that the extream values of the given distribution are considered as outliers. So, the central 95% of the simulated data are normal data and the rest data are outliers. The efficient boxplot should detect the number of outliers less than 5% of sample size n. In Table 3, the averages of $\bar{o}r\%$ that is less than 5% are boldface.

From Table 3, it is shown that for symmetric distribution, the averages of $\bar{o}r\%$ of the five methods namely Turkey, Kimber, Walker, Adil and the proposed MH are less than 5% in both small and large sizes. Hubert and MK only provide good results in a large size. For moderately skewed distribution, only Adil and the proposed MH, their averages of $\bar{o}r\%$ are less than 5% in both small and large sizes. The averages of $\bar{o}r\%$ of Kimber, Hubert, Walker and MK are less than 5% only in a large size. The average of $\bar{o}r\%$ of Turkey is more than 5% for all distributions in the moderately skewed. For mildly skewed distribution, Tukey's, Kimber's, Walker's, Adil's and the proposed MH methods perform well in both small and large sizes but Hubert and MK work well only in large size. Additionally, the individual average of $\bar{o}r\%$ of each method for each distribution in small and large sample sizes are shown in Figures 3 and 4 in which the width of a sign implies the range of the $\bar{o}r\%$ and the red dot line is placed at the optimal value of the $\bar{o}r\%$.

Table 2 Average and range of $\bar{or}_i\%$ for small sample size $(n = 15, 20, 25, 30)$ and for large sample size $(n = 50, 100, 250, 500, 750, 1000)$ on each distribution in truncated data simulation

Skewed level	$\tilde{D}$	n	Stats.	Tukey	Kimber	Hubert	Walker	Adil	MK	MH
Symmetric	$N(0,1)$	$n \leq 30$	range	3.83	3.53	5.39	4.13	4.33	5.61	4.11
			average	$\mathbf{1.55}^{(1)}$	$\mathbf{2.24}^{(3)}$	4.82	3.71	$\mathbf{1.80}^{(2)}$	6.44	3.09
		$n > 30$	range	0.40	0.50	2.17	1.27	0.47	2.57	1.30
			average	$\mathbf{0.07}^{(1)}$	$\mathbf{0.08}^{(2)}$	0.54	0.26	$\mathbf{0.08}^{(2)}$	0.64	$\mathbf{0.24}^{(3)}$
Moderately Skewed	χ^2_1	$n \leq 30$	range	2.22	3.67	2.50	4.39	1.56	4.72	2.33
			average	$\mathbf{2.89}^{(3)}$	2.93	4.98	3.33	$\mathbf{1.40}^{(1)}$	5.93	$\mathbf{2.19}^{(2)}$
		$n > 30$	range	1.29	0.53	1.57	0.77	0.27	2.40	0.37
			average	0.29	$\mathbf{0.11}^{(3)}$	0.55	0.16	$\mathbf{0.05}^{(1)}$	0.60	$\mathbf{0.08}^{(2)}$
	$F_{(10,10)}$	$n \leq 30$	range	2.61	4.11	3.50	4.44	2.91	4.06	4.58
			average	$\mathbf{1.79}^{(2)}$	2.88	5.18	3.90	$\mathbf{1.50}^{(1)}$	6.85	$\mathbf{2.54}^{(3)}$
		$n > 30$	range	0.80	0.77	2.60	1.20	0.27	2.63	0.83
			average	$\mathbf{0.19}^{(2)}$	0.18	0.65	0.30	$\mathbf{0.06}^{(1)}$	0.66	$\mathbf{0.17}^{(3)}$
	$F_{(90,10)}$	$n \leq 30$	range	3.17	5.51	5.28	6.67	3.11	5.78	3.20
			average	$\mathbf{2.26}^{(2)}$	3.47	5.55	4.52	$\mathbf{1.75}^{(1)}$	7.08	$\mathbf{2.88}^{(3)}$
		$n > 30$	range	0.47	0.43	2.07	0.70	0.20	2.33	0.40
			average	$\mathbf{0.08}^{(3)}$	$\mathbf{0.08}^{(3)}$	0.46	0.13	$\mathbf{0.03}^{(1)}$	0.50	$\mathbf{0.07}^{(2)}$
Mildly Skewed	χ^2_5	$n \leq 30$	range	1.76	3.72	3.78	4.44	1.17	4.56	2.39
			average	$\mathbf{1.43}^{(2)}$	$\mathbf{2.48}^{(3)}$	5.04	3.90	$\mathbf{1.10}^{(1)}$	6.97	2.57
		$n > 30$	range	0.07	0.57	1.90	0.97	0.10	2.93	0.50
			average	$\mathbf{0.02}^{(1)}$	0.11	0.44	0.19	$\mathbf{0.03}^{(2)}$	0.63	$\mathbf{0.09}^{(3)}$
	χ^2_{20}	$n \leq 30$	range	1.78	3.06	3.11	4.00	2.00	5.39	3.28
			average	$\mathbf{1.16}^{(1)}$	$\mathbf{2.51}^{(3)}$	4.57	3.71	$\mathbf{1.14}^{(2)}$	5.78	2.77
		$n > 30$	range	0.07	0.40	2.07	1.23	0.07	3.00	0.73
			average	$\mathbf{0.01}^{(1)}$	$\mathbf{0.07}^{(2)}$	0.39	0.23	$\mathbf{0.01}^{(1)}$	0.59	$\mathbf{0.12}^{(3)}$
	$F_{(10,90)}$	$n \leq 30$	range	1.80	2.93	2.61	3.58	1.56	3.94	3.60
			average	$\mathbf{1.15}^{(2)}$	$\mathbf{2.25}^{(3)}$	4.92	3.35	$\mathbf{1.02}^{(1)}$	6.03	2.34
		$n > 30$	range	0.17	0.37	1.67	1.27	0.13	2.73	0.20
			average	$\mathbf{0.03}^{(1)}$	0.08	0.32	0.25	$\mathbf{0.03}^{(1)}$	0.55	$\mathbf{0.05}^{(3)}$
	$F_{(90,90)}$	$n \leq 30$	range	2.00	3.72	2.67	4.50	2.06	5.22	4.33
			average	$\mathbf{1.08}^{(1)}$	$\mathbf{2.37}^{(3)}$	4.17	3.69	$\mathbf{1.14}^{(2)}$	6.25	3.02
		$n > 30$	range	0.23	0.47	1.87	0.87	0.17	2.23	0.63
			average	$\mathbf{0.04}^{(2)}$	$\mathbf{0.08}^{(3)}$	0.40	0.22	$\mathbf{0.03}^{(1)}$	0.64	0.12

Case III: Contaminated data

For generating contaminated data, 5% of data are selected and multiplied by a constant c. The constant c is used to move the selected data away from the true position. The more value of c is the more move get. The selected data are randomly and equally obtained from the upper and lower tails for the normal distribution and only obtained from the upper tail for the χ^2 and F distributions. In this case, two types called Type I and Type II of contaminated data were generated based on the constant c. The selected data were multiplied by 2 and 10 for Type I and Type II, respectively. Since the percentage of contaminated data is 5%, the optimal value of the average of $\bar{or}\%$ would be 5 as well. In Tables 4 and 5, the first, second and third averages of $\bar{or}\%$ are also boldface and noticed by the superscripts with the number 1, 2 and 3 in parentheses, respectively.

From the results of Type I of contamination shown in Table 4, it is shown that for symmetric distribution, Turkey, Walker and Adil are of the first three best averages of $\bar{or}\%$ in both small and large sample sizes. Kimber and the proposed MH provide good results only for the large size. For moderately and mildly skewed distributions, the average of $\bar{or}\%$ of the proposed MH outperforms these of the other methods. From the results of Type II of contamination shown in Table 5, it is shown that for symmetric distribution, Turkey, Kimber and Adil are of the first three bests of $\bar{or}\%$ in both small and large sample sizes. The proposed MH method provides good results only for the large size. For moderately and mildly skewed distributions, the average of $\bar{or}\%$ of the proposed MH still outperforms these of the other methods. Additionally, the individual average $\bar{or}\%$ of each method

Table 3 Average and range of $\bar{or}_i\%$ for small sample size $(n = 15, 20, 25, 30)$ and for large sample size $(n = 50, 100, 250, 500, 750, 1000)$ on each distribution in untruncated data simulation

Skewed level	$\tilde{D}$	n	Stats.	Tukey	Kimber	Hubert	Walker	Adil	MK	MH
Symmetric	$N(0,1)$	$n \leq 30$	range	2.47	3.10	3.00	2.21	2.17	2.12	3.21
			average	**2.20**	**3.20**	5.82	**4.13**	**2.24**	6.45	**3.84**
		$n > 30$	range	0.89	1.22	2.79	1.89	0.83	2.85	0.98
			average	**0.94**	**1.10**	**1.85**	**1.36**	**0.93**	**1.82**	**1.04**
Moderately Skewed	χ^2_1	$n \leq 30$	range	1.60	1.89	4.53	1.49	2.07	4.77	0.90
			average	8.13	6.24	**4.87**	**4.05**	**3.02**	5.55	**3.45**
		$n > 30$	range	0.74	0.59	1.71	0.46	0.95	1.47	0.83
			average	7.54	**4.82**	**0.57**	**1.86**	**0.71**	**1.03**	**1.57**
	$F_{(10,10)}$	$n \leq 30$	range	0.84	1.28	1.57	1.87	0.82	2.45	1.13
			average	5.94	5.56	7.27	5.50	**4.11**	7.30	**4.08**
		$n > 30$	range	0.56	0.58	3.60	1.19	1.48	2.93	0.49
			average	5.38	**4.16**	**3.30**	**2.95**	**2.61**	**2.71**	**3.27**
	$F_{(90,10)}$	$n \leq 30$	range	0.93	1.23	2.63	1.80	1.37	2.80	1.23
			average	5.76	5.54	7.40	5.58	**4.17**	7.33	**3.96**
		$n > 30$	range	0.11	0.42	3.00	1.26	1.51	2.89	0.34
			average	5.13	**3.96**	**3.42**	**2.84**	**2.54**	**2.64**	**3.11**
Mildly Skewed	χ^2_5	$n \leq 30$	range	0.90	2.17	2.20	2.73	1.20	3.00	1.97
			average	**4.03**	**4.18**	6.07	**4.71**	**2.80**	6.80	**3.35**
		$n > 30$	range	0.52	0.95	3.99	1.53	0.50	3.79	0.44
			average	**2.84**	**2.17**	**1.80**	**1.63**	**1.57**	**1.78**	**1.72**
	χ^2_{20}	$n \leq 30$	range	1.40	2.77	3.20	2.93	1.27	3.50	2.64
			average	**2.54**	**3.76**	6.02	**4.80**	**2.37**	6.77	**3.41**
		$n > 30$	range	0.68	0.99	2.55	1.83	0.60	3.23	0.67
			average	**1.51**	**1.32**	**1.75**	**1.32**	**1.22**	**1.72**	**1.12**
	$F_{(10,90)}$	$n \leq 30$	range	0.53	1.67	1.59	2.20	0.90	2.73	1.10
			average	**3.58**	**4.08**	6.34	**4.93**	**2.97**	7.05	**3.21**
		$n > 30$	range	0.26	0.63	2.78	1.40	0.41	3.05	0.36
			average	**2.49**	**2.00**	**1.88**	**1.67**	**1.67**	**1.83**	**1.74**
	$F_{(90,90)}$	$n \leq 30$	range	1.02	1.13	2.70	1.60	0.94	2.73	1.15
			average	**3.08**	**3.88**	6.24	**4.69**	**2.79**	6.87	**3.48**
		$n > 30$	range	0.70	0.92	3.14	1.61	0.73	2.95	0.61
			average	**1.69**	**1.52**	**2.01**	**1.57**	**1.44**	**1.90**	**1.34**

for each distribution in small and large sample sizes is shown in Figures 5-8 in which the width of a sign implies the range of the $\bar{or}\%$ and the red dot line is placed at the optimal value of the $\bar{or}\%$.

3.4. Experimental results on the real data

In this section, the proposed MH and six existing boxplots were evaluated on four real data sets namely time interval in days of the coal mine, maximum daily discharge in cubic, alamine(U/L), and the number of clicks in Facebook's page. The descriptive statistics for each data set are shown in Table 1. The histogram and a plot of the sorted data points for each data set are also given to identifying the number of potential outliers visually.

For coal mine data set, referring to the MC and Bowley values given in Table 1, we obtain that the distribution of this data is right-skewed. The potential outliers are the observations which are so far from the majority of the data. From the histogram plot in Figure 9(a), the majority of the data is between 0 and 1000. So, there are six points of potential outliers or 3.16% out of total data. The potential outliers are marked by the red plus sign as shown in Figure 9(b). From Table 6, The numbers of outliers detected by Tukey's and Kimber's boxplots are 11 and 13, respectively which are more than the others. Since the upper fences of Hubert's and Adil's boxplots are greatly extended, the numbers of detected outliers are only 3 and 1, respectively, which are much less than the observed number of potential outliers. We see that Walker's, MK and the proposed MH boxplots could detect 6, 5 and 6 outliers, respectively. The numbers of outliers, detected by these three boxplots, are satisfied with the number of potential outliers identified from the relevant histogram and plot of sorted data

Table 4 Average and range of $\bar{or}_i\%$ for small sample size $(n = 15, 20, 25, 30)$ and for large sample size $(n = 50, 100, 250, 500, 750, 1000)$ on each distribution for Type I of contamination

Skewed level	$\tilde{D}$	n	Stats.	Tukey	Kimber	Hubert	Walker	Adil	MK	MH
Symmetric	$N(0,1)$	$n \leq 30$	range	5.30	5.00	5.03	4.47	5.07	4.27	5.50
			average	$\mathbf{8.96}^{(1)}$	9.16	10.00	$\mathbf{9.07}^{(3)}$	$\mathbf{9.04}^{(2)}$	9.99	9.13
		$n > 30$	range	1.05	0.99	1.67	0.91	1.05	1.41	1.11
			average	$\mathbf{5.20}^{(2)}$	$\mathbf{5.20}^{(2)}$	5.42	$\mathbf{5.19}^{(1)}$	$\mathbf{5.20}^{(2)}$	5.34	$\mathbf{5.23}^{(3)}$
Moderately Skewed	χ^2_1	$n \leq 30$	range	1.55	2.25	2.47	1.48	1.40	2.98	1.25
			average	9.57	8.30	6.27	$\mathbf{5.95}^{(3)}$	$\mathbf{4.29}^{(2)}$	6.60	$\mathbf{4.78}^{(1)}$
		$n > 30$	range	0.96	2.09	1.42	0.60	2.45	0.55	0.43
			average	7.90	5.87	3.12	$\mathbf{5.12}^{(1)}$	1.46	$\mathbf{4.28}^{(3)}$	$\mathbf{4.54}^{(2)}$
	$F_{(10,10)}$	$n \leq 30$	range	2.13	2.21	1.58	2.23	2.12	1.90	1.86
			average	8.18	8.30	9.88	$\mathbf{8.12}^{(3)}$	$\mathbf{7.86}^{(2)}$	9.40	$\mathbf{6.57}^{(1)}$
		$n > 30$	range	1.70	1.89	2.59	1.56	0.53	2.65	1.08
			average	5.98	$\mathbf{5.48}^{(3)}$	6.92	$\mathbf{5.35}^{(2)}$	6.82	5.68	$\mathbf{5.27}^{(1)}$
	$F_{(90,10)}$	$n \leq 30$	range	1.89	2.43	2.43	2.08	1.07	3.53	1.86
			average	8.18	8.30	9.88	$\mathbf{8.12}^{(3)}$	$\mathbf{7.86}^{(2)}$	9.40	$\mathbf{6.57}^{(1)}$
		$n > 30$	range	1.49	1.47	2.89	1.36	0.52	2.14	0.82
			average	5.78	$\mathbf{5.42}^{(3)}$	6.98	$\mathbf{5.35}^{(2)}$	6.86	5.70	$\mathbf{5.25}^{(1)}$
Mildly Skewed	χ^2_5	$n \leq 30$	range	2.71	2.38	3.07	2.17	2.45	3.40	1.63
			average	$\mathbf{7.27}^{(2)}$	7.90	9.21	7.67	$\mathbf{7.30}^{(3)}$	9.21	$\mathbf{6.27}^{(1)}$
		$n > 30$	range	1.38	1.48	2.63	1.62	1.69	2.48	0.68
			average	$\mathbf{5.30}^{(2)}$	$\mathbf{5.31}^{(3)}$	5.90	5.34	5.37	5.54	$\mathbf{5.17}^{(1)}$
	χ^2_{20}	$n \leq 30$	range	2.46	2.81	2.32	2.33	1.89	2.77	1.90
			average	$\mathbf{7.19}^{(1)}$	$\mathbf{8.02}^{(3)}$	10.34	8.51	8.38	10.47	$\mathbf{7.45}^{(2)}$
		$n > 30$	range	1.06	1.17	2.21	1.64	1.75	2.51	1.00
			average	$\mathbf{5.23}^{(1)}$	$\mathbf{5.26}^{(2)}$	6.12	$\mathbf{5.42}^{(3)}$	5.65	5.77	$\mathbf{5.23}^{(1)}$
	$F_{(10,90)}$	$n \leq 30$	range	2.83	2.23	4.53	2.13	3.07	2.64	2.03
			average	$\mathbf{7.14}^{(1)}$	$\mathbf{7.57}^{(3)}$	9.60	8.13	7.86	9.70	$\mathbf{7.27}^{(1)}$
		$n > 30$	range	1.12	1.34	2.82	1.66	1.70	2.68	0.91
			average	$\mathbf{5.23}^{(2)}$	$\mathbf{5.27}^{(3)}$	6.29	5.37	5.59	5.73	$\mathbf{5.15}^{(1)}$
	$F_{(90,90)}$	$n \leq 30$	range	2.65	2.76	2.89	2.76	2.19	2.60	2.65
			average	$\mathbf{7.18}^{(1)}$	$\mathbf{8.12}^{(3)}$	10.81	9.07	9.31	10.86	$\mathbf{7.79}^{(2)}$
		$n > 30$	range	1.20	1.34	3.25	1.95	2.30	2.78	1.16
			average	$\mathbf{5.26}^{(2)}$	$\mathbf{5.34}^{(3)}$	6.45	5.54	6.01	5.97	$\mathbf{5.25}^{(1)}$

points visually.

For discharge data, referring to the MC and Bowley values given in Table 1, we obtain that the distribution of this data is left-skewed. From the histogram in Figure 10(a), the majority of the data is between 0 and 2000. So, there are three points of potential outliers or 6.00% out of total data. The potential outliers are marked by the red plus sign as shown in Figure 10(b). From Table 7, Kimber and Hubert provide the numbers of three detected outliers which are satisfied with the number of potential outliers identified from the relevant histogram and plot of sorted data points visually. Turkey, Walker, Adil and the proposed MH equally provide two percentage away from the number of observed potential outliers.

For Alamine data set, referring to the MC and Bowley values given in Table 1, we obtain that the distribution of this data is right-skewed. From the histogram plot in Figure 11(a), the majority of the data is between 0 and 700. So, there are 16 points of potential outliers or 2.74% out of total data. The potential outliers are marked by the red plus sign as shown in Figure 11(b). Among these outliers, there are at least two or four potential outliers which are so far away from the rest of the data and could affect the real mean or standard deviation of the data. From Table 8, the numbers of detected outliers, obtained from the proposed MH, MK, and Walker, are 26, 32 and 38 which are rather close to the number of potential outliers, whereas the numbers of detected outliers obtained from Tukey and Adil are 68 and 136 which are quite the larger numbers of outliers than usual. It is simply noticed that these far away outliers could affect the range of the fences obtained from these two methods.

Table 5 Average and range of $\bar{or}_i\%$ for small sample size $(n = 15, 20, 25, 30)$ and for large sample size $(n = 50, 100, 250, 500, 750, 1000)$ on each distribution for Type II contamination

Skewed level	$\tilde{D}$	n	Stats.	Tukey	Kimber	Hubert	Walker	Adil	MK	MH
Symmetric	$N(0,1)$	$n \leq 30$	range	6.70	7.40	7.60	7.13	7.30	8.20	7.40
			average	$\mathbf{10.12}^{(1)}$	$\mathbf{10.62}^{(3)}$	12.13	11.21	$\mathbf{10.52}^{(2)}$	12.84	10.92
		$n > 30$	range	1.21	1.49	3.07	1.91	1.49	3.46	1.41
			average	$\mathbf{5.24}^{(1)}$	$\mathbf{5.30}^{(3)}$	5.73	5.42	$\mathbf{5.29}^{(2)}$	5.81	$\mathbf{5.30}^{(3}$
Moderately Skewed	χ^2_1	$n \leq 30$	range	1.52	1.82	3.67	2.64	1.62	3.15	1.91
			average	9.21	8.36	9.34	$\mathbf{7.73}^{(2)}$	$\mathbf{8.11}^{(3)}$	9.17	$\mathbf{6.52}^{(1}$
		$n > 30$	range	0.94	1.72	1.62	1.26	1.63	1.68	0.98
			average	7.78	5.74	$\mathbf{5.33}^{(2)}$	5.29	6.26	$\mathbf{5.34}^{(3)}$	$\mathbf{5.26}^{(1}$
	$F_{(10,10)}$	$n \leq 30$	range	1.56	2.23	2.42	2.63	1.73	2.40	2.14
			average	$\mathbf{8.04}^{(2)}$	$\mathbf{8.18}^{(3)}$	12.14	8.95	11.23	11.72	$\mathbf{7.38}^{(1}$
		$n > 30$	range	1.44	1.55	2.68	1.52	1.40	2.77	1.48
			average	5.95	$\mathbf{5.46}^{(3)}$	6.81	$\mathbf{5.36}^{(1)}$	9.20	5.73	$\mathbf{5.38}^{(2}$
	$F_{(90,10)}$	$n \leq 30$	range	1.83	2.11	2.47	2.84	2.20	2.76	1.58
			average	$\mathbf{8.04}^{(2)}$	$\mathbf{8.18}^{(3)}$	12.14	8.95	11.23	11.72	$\mathbf{7.38}^{(1}$
		$n > 30$	range	1.48	1.63	2.96	1.72	0.49	3.32	1.52
			average	5.86	$\mathbf{5.46}^{(3)}$	7.25	$\mathbf{5.45}^{(2)}$	9.26	6.01	$\mathbf{5.40}^{(1}$
Mildly Skewed	χ^2_5	$n \leq 30$	range	2.57	2.55	4.58	3.42	3.65	4.18	2.52
			average	$\mathbf{7.23}^{(2)}$	$\mathbf{7.62}^{(3)}$	10.74	8.57	9.95	10.43	$\mathbf{7.19}^{(1}$
		$n > 30$	range	1.10	1.22	2.97	1.48	2.35	2.40	1.06
			average	$\mathbf{5.27}^{(2)}$	$\mathbf{5.27}^{(2)}$	6.09	$\mathbf{5.32}^{(3)}$	6.26	5.60	$\mathbf{5.24}^{(1}$
	χ^2_{20}	$n \leq 30$	range	3.18	3.09	4.53	2.68	2.33	2.91	3.37
			average	$\mathbf{7.15}^{(1)}$	$\mathbf{8.39}^{(3)}$	11.73	9.53	10.83	11.50	$\mathbf{7.96}^{(2}$
		$n > 30$	range	1.20	1.47	3.39	2.16	3.05	3.25	1.36
			average	$\mathbf{5.25}^{(1)}$	$\mathbf{5.32}^{(3)}$	6.29	5.52	6.31	5.91	$\mathbf{5.29}^{(2}$
	$F_{(10,90)}$	$n \leq 30$	range	2.36	2.25	3.38	2.09	1.92	2.95	2.20
			average	$\mathbf{7.32}^{(1)}$	$\mathbf{7.89}^{(3)}$	11.40	8.51	10.50	10.77	$\mathbf{7.49}^{(2}$
		$n > 30$	range	1.22	1.44	3.32	1.94	2.91	3.36	1.30
			average	$\mathbf{5.25}^{(1)}$	$\mathbf{5.30}^{(3)}$	6.37	5.41	6.52	5.81	$\mathbf{5.27}^{(2}$
	$F_{(90,90)}$	$n \leq 30$	range	2.50	2.47	3.75	2.82	3.37	3.60	3.05
			average	$\mathbf{7.28}^{(1)}$	$\mathbf{8.20}^{(3)}$	11.90	9.55	11.10	11.73	$\mathbf{8.13}^{(2}$
		$n > 30$	range	1.19	1.59	3.05	1.95	2.81	2.93	1.50
			average	$\mathbf{5.26}^{(1)}$	$\mathbf{5.39}^{(3)}$	6.44	5.57	6.47	6.02	$\mathbf{5.34}^{(2}$

Table 6 The lower and upper fences, and the number of detected outliers of coal mine data set computed from the seven boxplots

Method	[lower fence, upper fence]	Range	#Outliers(%)
Tukey	[-310.63, 618.38]	929.00	13 (6.84%)
Kimber	[-189.50, 739.50]	929.00	11 (5.79%)
Hubert	[-33.06, 1,420.78]	1,453.84	3 (1.57%)
Walker	[-130.87, 989.74]	1,120.62	6 (3.16%)
Adil	[-47.34, 1,696.26]	1,743.61	1 (0.53%)
MK	[-72.24, 1,239.99]	1,312.23	5 (2.63%)
MH	[-374.47, 984.51]	1,358.99	6 (3.16%)

For the number of clicks data set, referring to the MC and Bowley values given in Table 1, we obtain that the distribution of this data is right-skewed. From the histogram plot in Figure 12(a), the majority of the data is between 0 and 3000. So, there are 11 points of potential outliers or 2.20% out of the total data. The potential outliers are also noticed by the red plus sign as shown in Figure 12(b). Like Alamine data set, among these outliers, there are at least one outliers which considerably deviate from the rest of the data. From Table 9, the numbers of detected outliers, obtained from MK,

Table 7 The lower fences, upper fences and the number of outliers of discharge data set computed from the seven boxplots

Method	[lower fence, upper fence]	Range	#Outliers(%)
Tukey	[397.63, 2,172.63]	1,775.00	1 (2.00%)
Kimber	[188.00, 1,963.00]	1,775.00	3 (6.00%)
Hubert	[118.22, 1,924.13]	1,805.91	3 (6.00%)
Walker	[-214.36, 1,853.79]	2,068.15	5 (10.00%)
Adil	[347.37, 2,125.90]	1,778.53	1 (2.00%)
MK	[-616.71, 1,744.57]	2,361.28	7 (14.00%)
MH	[699.58, 2,071.90]	1,372.32	1 (2.00%)

Table 8 The lower fences, upper fences and the number of outliers of Alamine data set computed from the seven boxplots

Method	[lower fence, upper fence]	Range	#Outliers(%)
Tukey	[-68.00, 180.00]	248.00	68 (11.66%)
Kimber	[-26.00, 222.00]	248.00	54 (9.26%)
Hubert	[14.32, 558.33]	544.01	41 (7.03%)
Walker	[-10.13, 333.18]	343.31	38 (6.51%)
Adil	[24.68, 27,222.11]	27,197.43	136 (23.33%)
MK	[5.73, 444.35]	438.62	26 (4.46%)
MH	[-85.30, 394.37]	479.67	32 (5.49%)

Walker, and the proposed MH, are 11, 13, and 16, respectively which are very close to the number of potential outliers, while the other methods give the number of detected outliers more than this of potential outliers.

Table 9 The lower fences, upper fences and the number of outliers of number of clicks data set computed from the seven boxplots

Method	[lower fence, upper fence]	Range	#Outliers(%)
Tukey	[-602.00, 1890.00]	2492.00	38 (7.60%)
Kimber	[-324.50, 2167.50]	2492.00	32 (6.40%)
Hubert	[108.46, 3683.00]	3574.54	25 (5.00%)
Walker	[-174.07, 2678.42]	2852.49	13 (2.60%)
Adil	[175.89, 6531.74]	6355.85	43 (8.60%)
MK	[-23.65, 3191.34]	3214.99	11 (2.20%)
MH	[-765.21, 2571.66]	3336.87	16 (3.20%)

4. Conclusions and Discussion

The boxplot is a practical and simple tool for detecting outliers in a univariate data set. Unfortunately, when drawing the boxplot of a skewed distribution, many more observations are typically labelled as potential outliers. In this work, we proposed the boxplot-based method called MH boxplot for handling the outliers detection problem in a skewed distribution. The proposed MH boxplot was modified from Hubert's boxplot by embedding the Bowley's coefficient, the ratio of lower and upper split interquartile ranges into the fences of the boxplot. For performance evaluation, the family of boxplots-based methods, including Tukey (1977), Kimber (1990), Hubert and Vandervieren (2008),

Walker and Chakraborti (2013), Adil and Irshad (2015), and Promwongsa et al. (2018), was used to make a comparison in terms of the percentage of outlier ratio mean ($\bar{or}\%$). The experiments were conducted on three cases of simulated data i.e. truncated, uncontaminated and contaminated data, and four real data sets. For truncated data, the $\bar{or}\%$ of the proposed MH is very close to these of Tukey, Kimber, Walker and Adil in large sample size for all of symmetric, moderately and mildly skewed distributions. Moreover, Turkey and Adil perform well the same and better than other methods. For uncontaminated, the $\bar{or}\%$ values from all methods are the same level of efficiency but Tukey is not good in moderately skewed distribution. Contaminated data, the proposed MH method provides good results for moderately and mildly skewed distributions, especially in moderately skewed distribution, but Turkey performs well in symmetric and mildly skewed distributions, Kimber gives good results in performance for symmetric and mildly skewed distribution and Adil gives good results in performance for symmetric skewed distribution. For real dataset, the proposed MH performs quite well in all real data sets but the performances of the others drop in some data sets such as Turkey, Kimber and Adil by which they flag many observations as outliers on Alamine data set. The results from simulated and real data show that the proposed MH boxplot efficiently detects outliers and is robust to skewness of data for any sample size. Moreover, the proposed MH boxplot efficiently detects outliers as the shape of real data, especially right-skewed distribution considering with real data sets.

References

Adil IH, Irshad AR. A modified approach for detection of outliers. Pak J Stat Oper Res. 2015; 11(1): 91-102.

Babura BI, Adam MB, Fitrianto A, Abdul Rahim AS. Modified boxplot for extreme data. AIP Conference Proceedings 1842(1); 2017. 1-9.

Barnett O, Cohen A. The histogram and boxplot for the display of lifetime data. J Comput Graph Stat. 2000; 9(4): 759-778.

Brys G, Hubert M, Struyf A. A robust measure of skewness. J Comput Graph Stat. 2004; 13(4): 996-1017.

Carling K. Resistant outlier rules and the non-Gaussian case. Comput Stat Data Anal. 2000; 33: 249-258.

Gumbel EJ. The return period of flood flows. Ann Math Stat. 1941; 12: 163-190.

Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Comput Stat Data Anal. 2008; 52: 5186-5201.

Jarrett RG. A note on the intervals between coal mining disasters. Biometrika. 1979; 66(1): 191-193.

Kimber AC. Exploratory data analysis for possibly censored data from skewed distributions. J R Stat Soc C. 1990; 39(1): 21-30.

Iglewicz B, Hoaglin DC. How to detect and handle outliers. Wisconsin: ASQC Quality Press; 1993.

Promwongsa M, Srisodaphol W, Junsawang P. The Modified boxplot for outlier detection. ICAS2018: Proceeding of International Conference on Applied Statistics; 2018 Oct 24-26; Thailand. pp. 121-125.

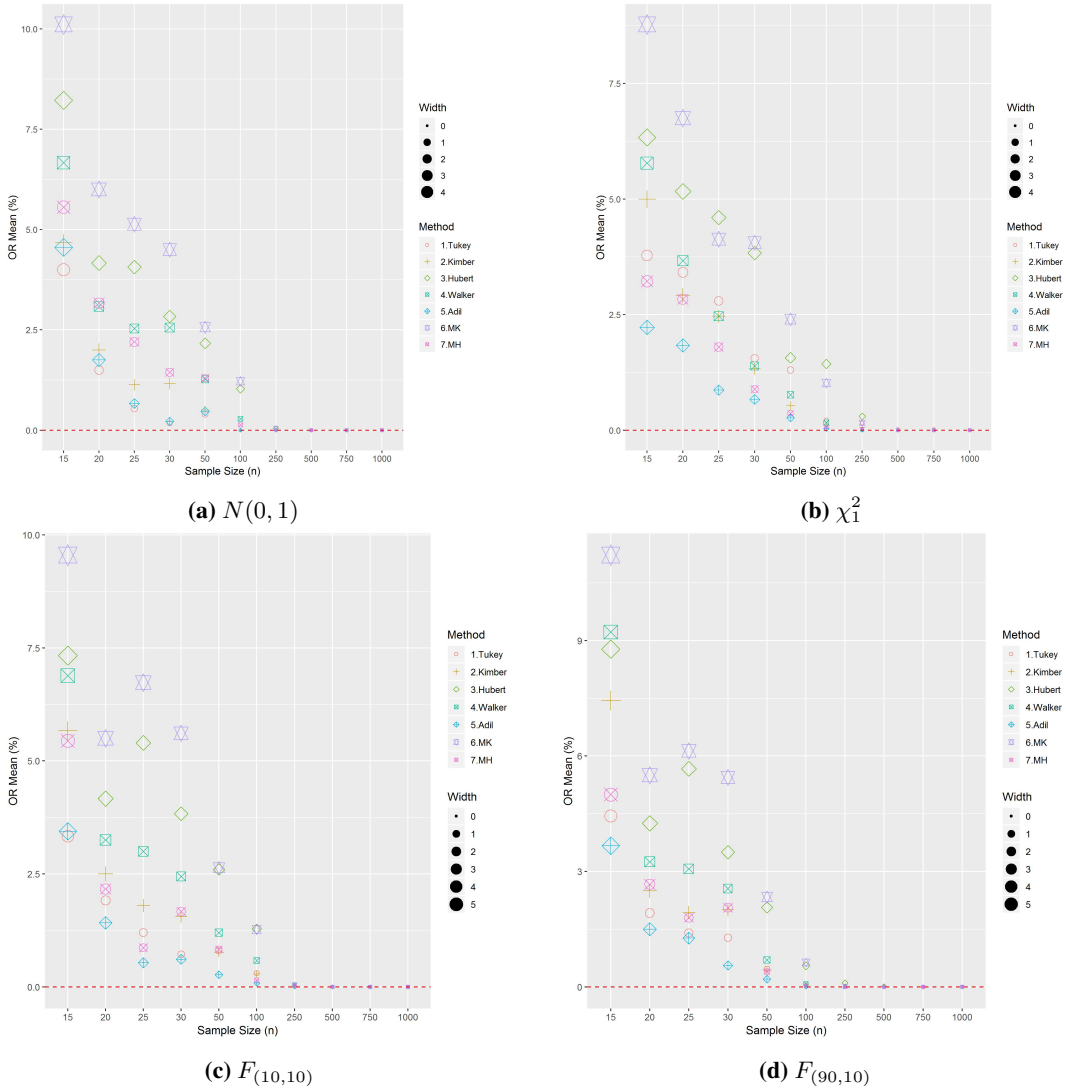
Moro S, Rita P, Vala B. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. J Bus Res. 2016; 69: 3341-3351.

Ramana BV, Babu MSP, Venkateswarlu N. A critical comparative study of liver patients from USA and INDIA: An exploratory analysis. Int J Comput Sci Issues. 2012; 9(2): 506-516.

Tukey JW. Exploratory data analysis. Massachusetts: Addison-Wesley; 1977.

Walker M, Chakraborti S. An asymmetrically modified boxplot for exploratory data analysis. 2013. Available from: https://docplayer.net/46093068-An-asymmetrically-modified-boxplot-for-exploratory-data-analysis.html

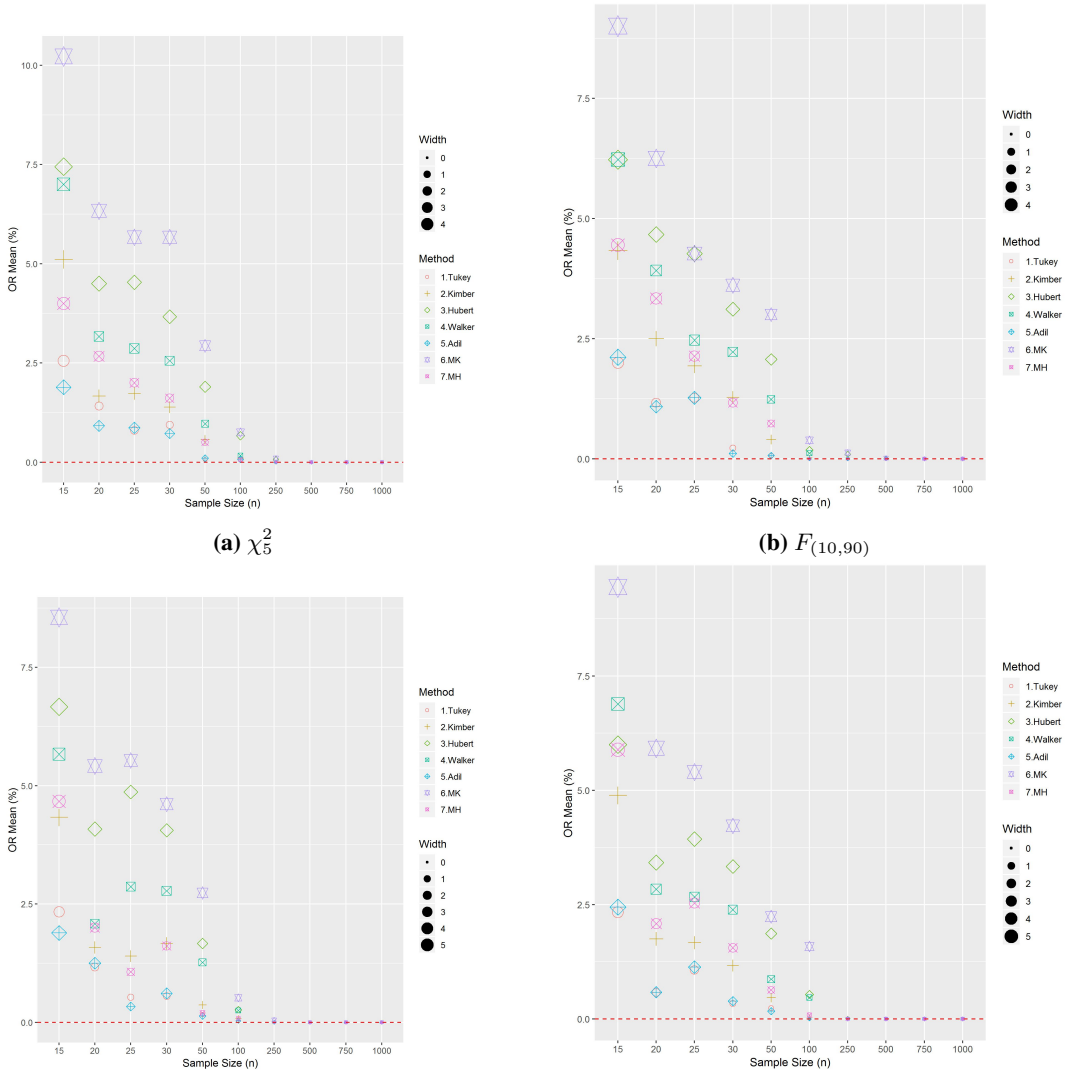
Zhao C, Yang J. A robust skewed boxplot for detecting outliers in rainfall observations in real-time flood forecasting. Adv Meteorol. 2019; Volume 2019, Article ID 1795673, 7 pages. https://doi.org/10.1155/2019/1795673



(a) $N(0,1)$ **(b)** χ^2_1

(c) $F_{(10,10)}$ **(d)** $F_{(90,10)}$

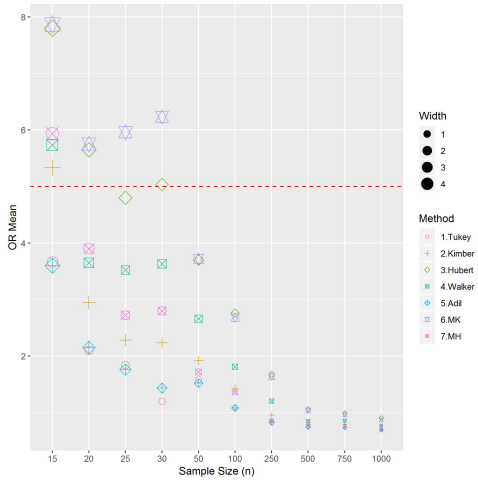
Figure 1 Symmetric and moderately skewed distribution for case I



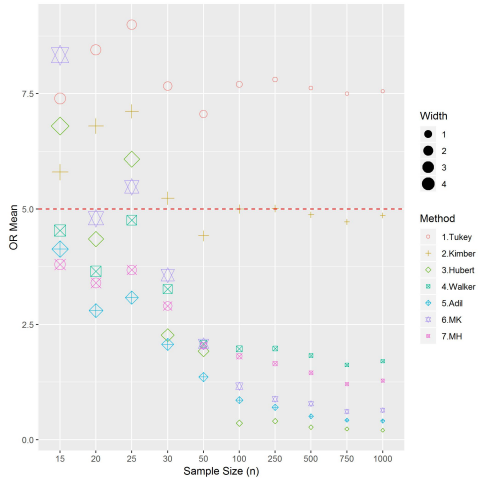
(a) χ^2_5 **(b)** $F_{(10,90)}$

(c) $F_{(10,10)}$ **(d)** $F_{(90,90)}$

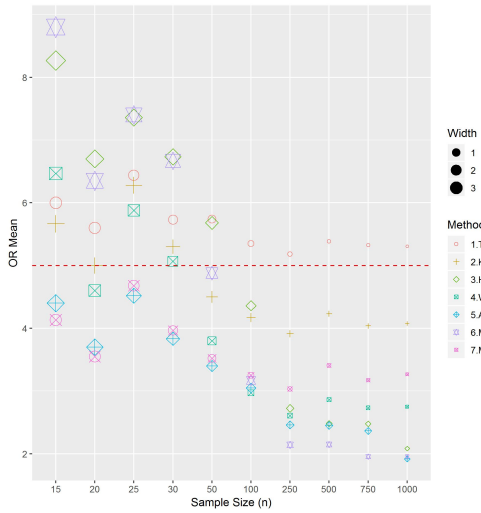
Figure 2 Mildly skewed distribution for case I



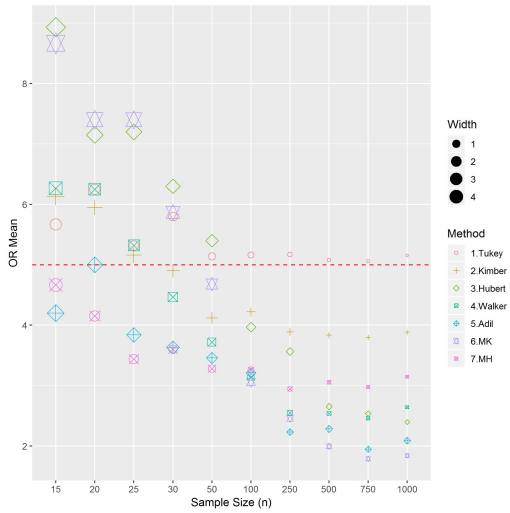
(a) $N(0,1)$



(b) χ^2_1

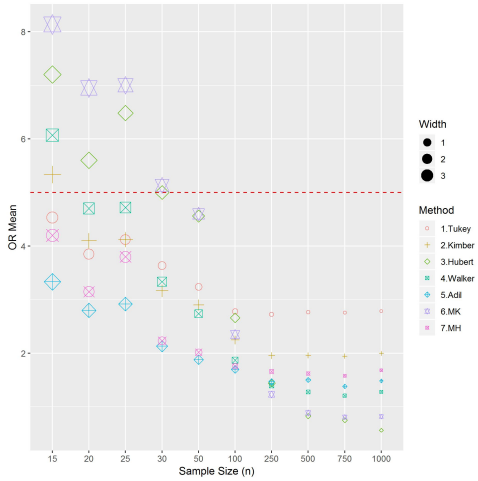


(c) $F_{(10,10)}$

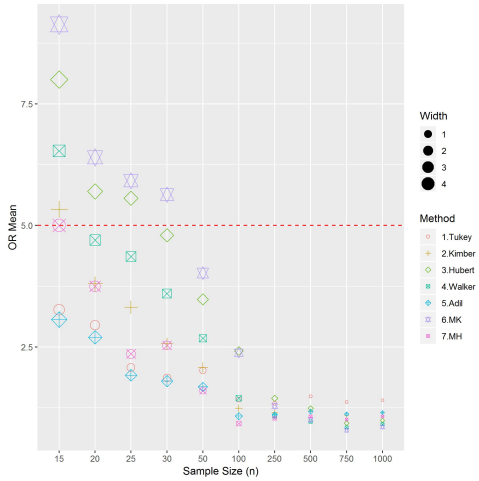


(d) $F_{(90,10)}$

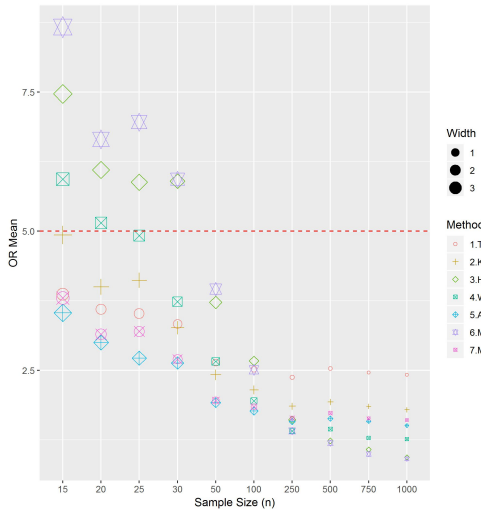
Figure 3 Symmetric and moderately skewed distribution for case II



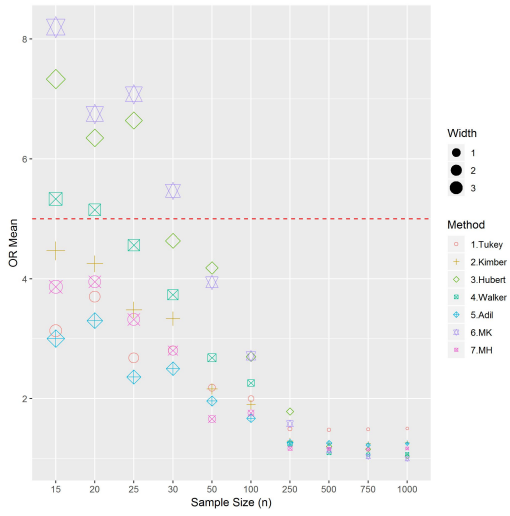
(a) χ^2_5



(b) χ^2_{20}

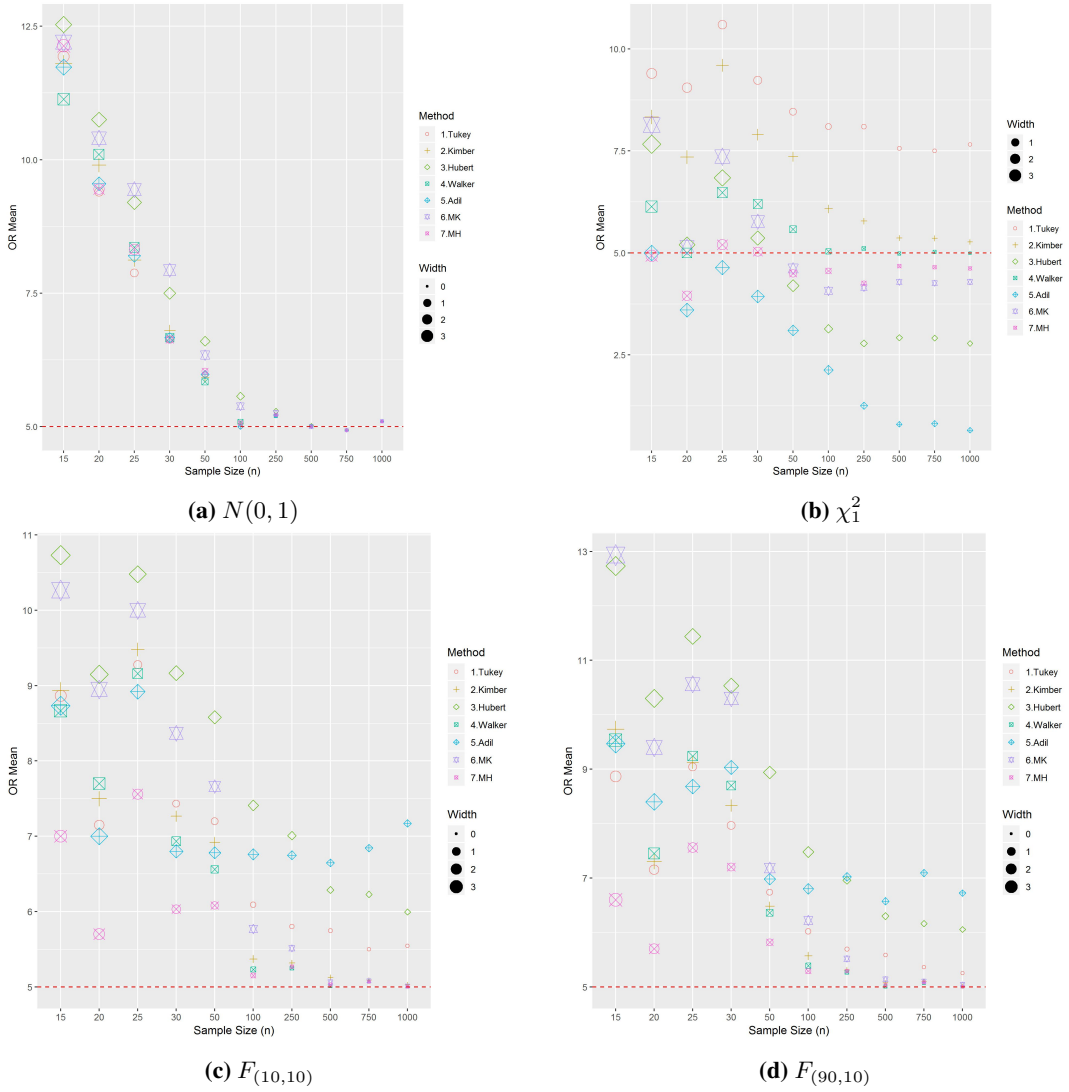


(c) $F_{(10,90)}$



(d) $F_{(90,90)}$

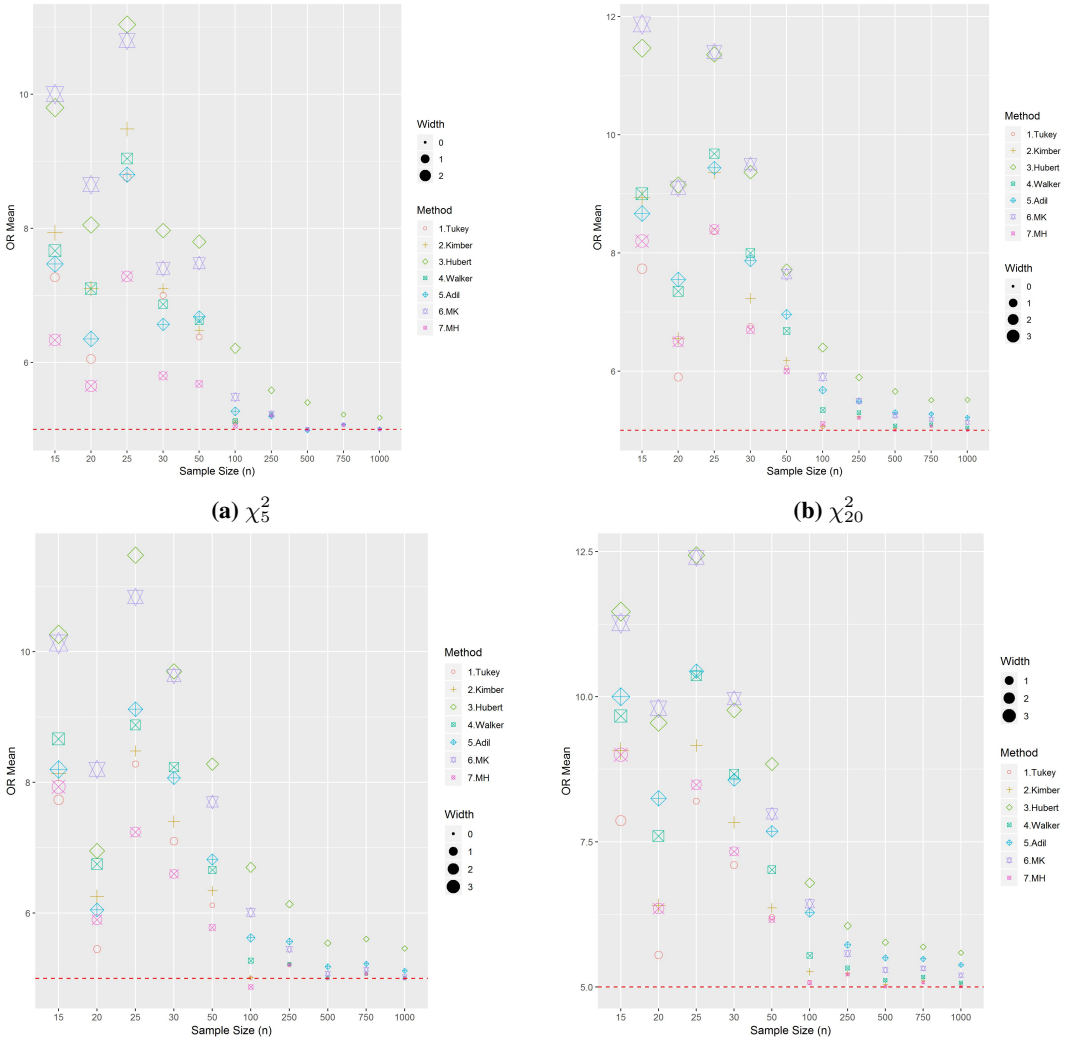
Figure 4 Mildly skewed distribution for case II



(a) $N(0,1)$ **(b)** χ^2_1

(c) $F_{(10,10)}$ **(d)** $F_{(90,10)}$

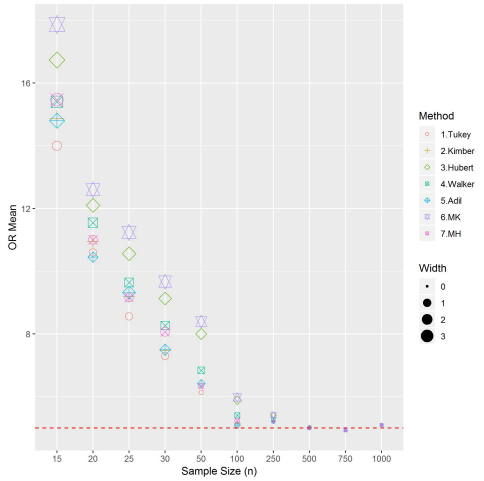
Figure 5 Symmetric and moderately skewed distribution for type I contamination



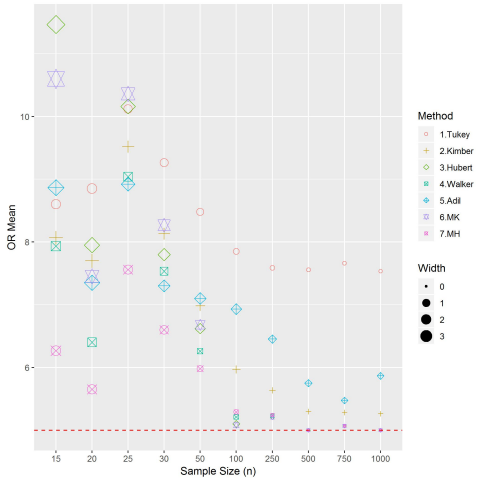
(a) χ^2_5 **(b)** χ^2_{20}

(c) $F_{(10,90)}$ **(d)** $F_{(90,90)}$

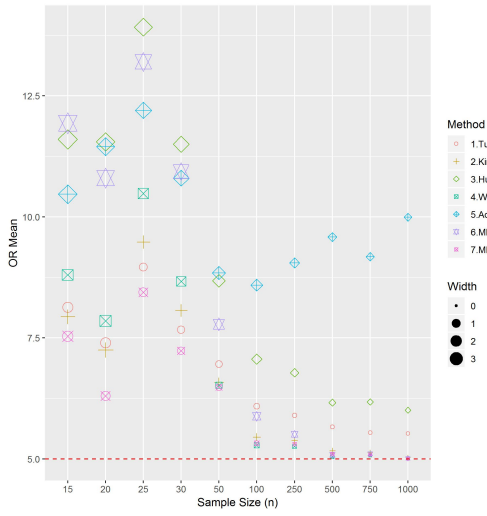
Figure 6 Mildly skewed distribution for type I contamination



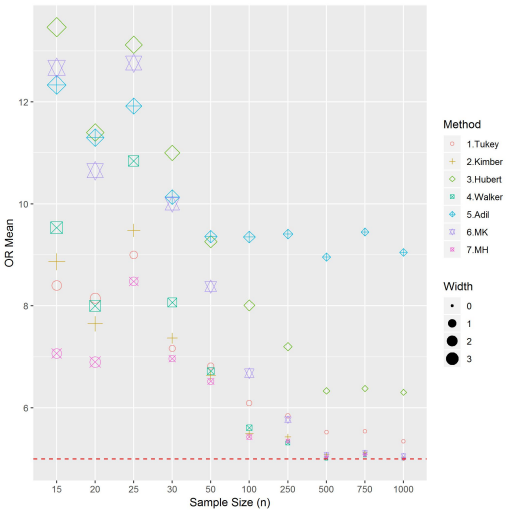
(a) $N(0,1)$



(b) χ^2_1

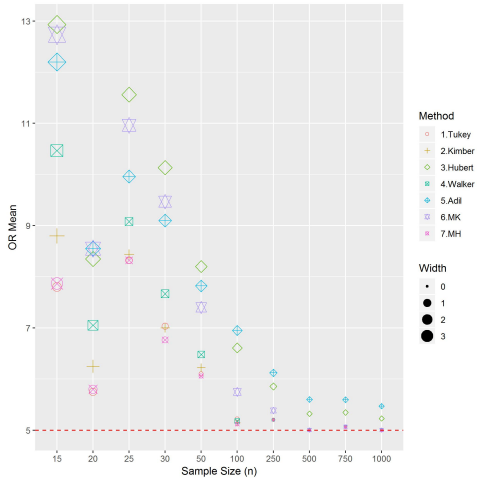


(c) $F_{(10,10)}$

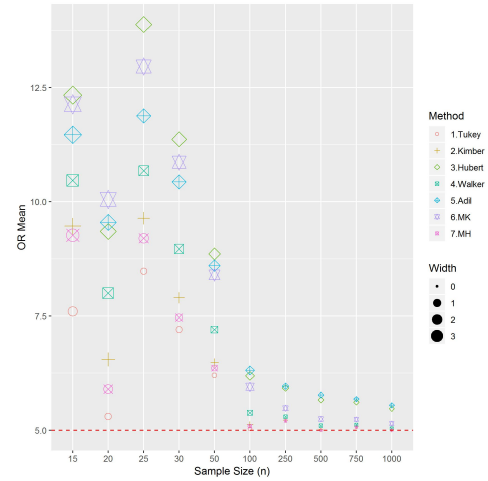


(d) $F_{(90,10)}$

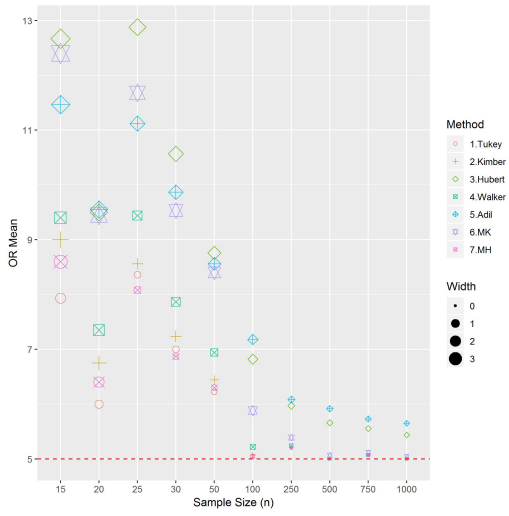
Figure 7 Symmetric and moderately skewed distribution for type II contamination



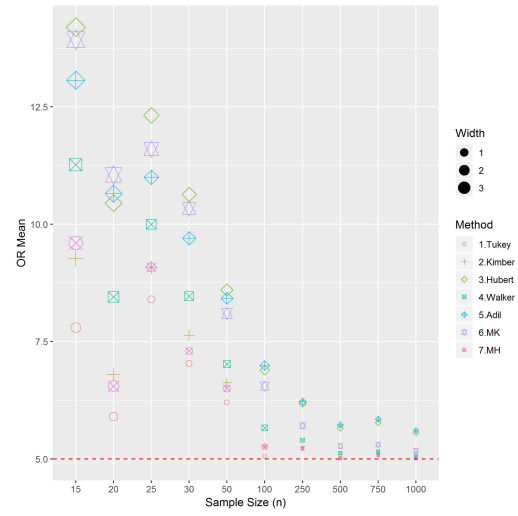
(a) χ^2_5



(b) χ^2_{20}

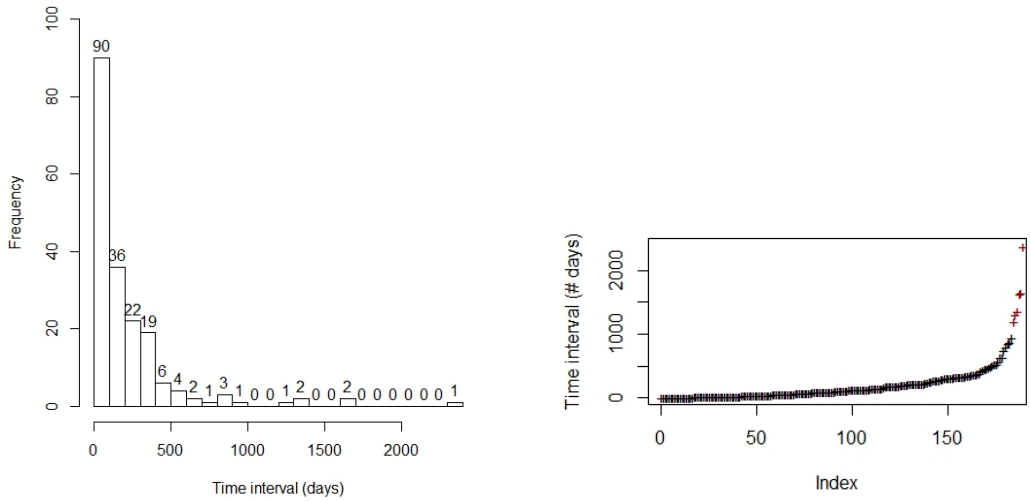


(c) $F_{(10,90)}$



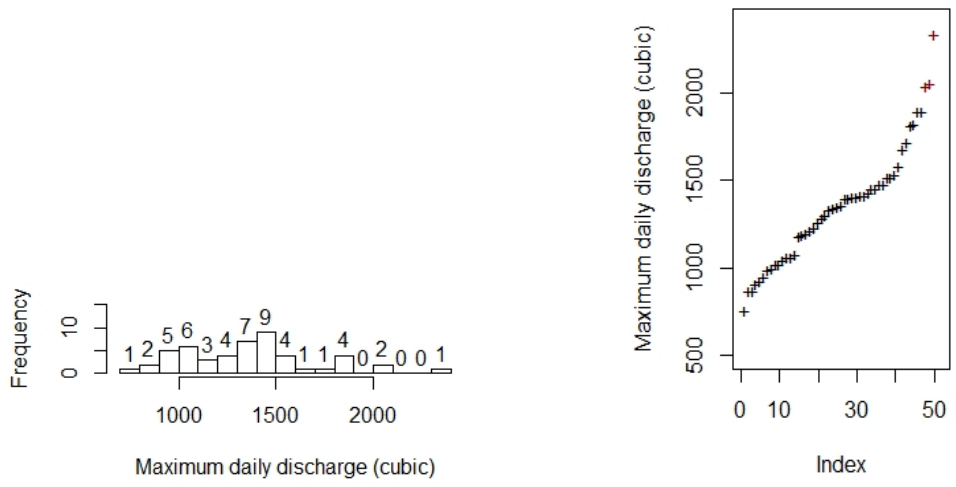
(d) $F_{(90,90)}$

Figure 8 Mildly skewed distribution type II contamination



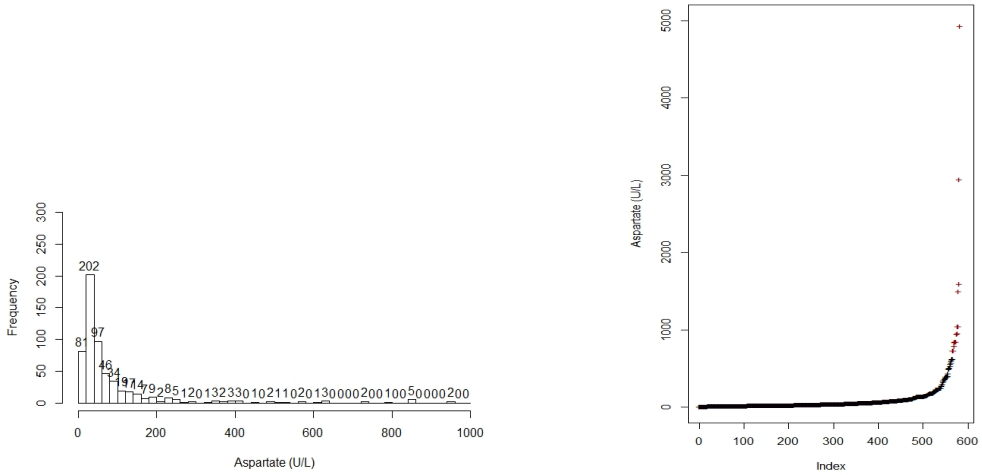
(a) Histogram plot **(b)** Plot of sorted data points

Figure 9 Histogram and the plot of sorted data points of coal mine data set



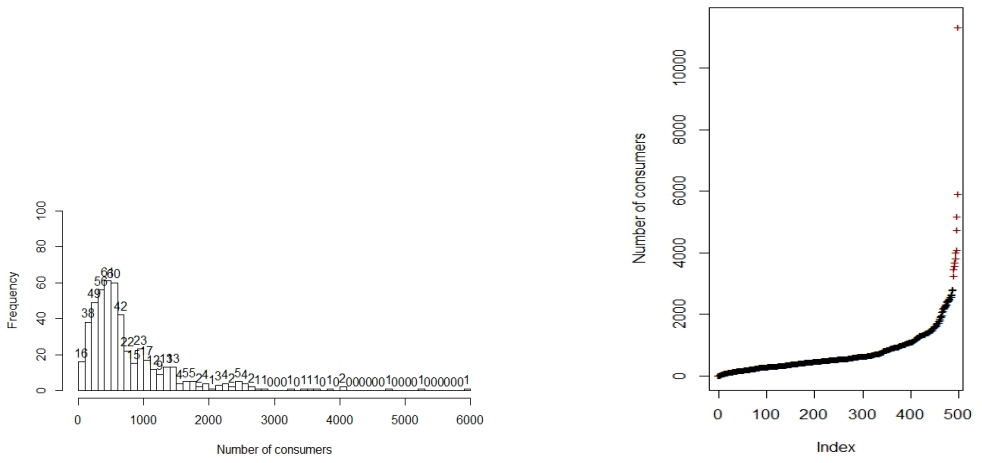
(a) Histogram plot **(b)** Plot of sorted data points

Figure 10 Histogram and the plot of sorted data points of discharge data set



(a) Histogram plot **(b)** Plot of sorted data points

Figure 11 Histogram and the plot of sorted data points of Alamine data set



(a) Histogram plot **(b)** Plot of sorted data points

Figure 12 Histogram and the plot of sorted data points of clicks data set