Gulf Journal of Mathematics
Vol 16, Issue 2 (2024) 298-312

ISSN: 2309-4966
https://doi.org/10.56947/gjom.v16i2.1813

RELATIVE RANGE FOR SKEWED DISTRIBUTIONS: A TOOL FOR OUTLIER DETECTION

DANIA DALLAH[1*], HANA SULIEMAN[2], AND AYMAN ALZAATREH[3]

Abstract. The ongoing exploration of outlier detection involves the development and refinement of diverse statistical analysis procedures. In this paper, we examine the performance of the standardized range and relative range in detecting outliers in a skewed data. The proposed approach considers Weibull distribution as a placeholder for skewed data. The empirical construction of the PDFs of the standardized range and relative range from the Weibull distribution along with the simulation work for outlier detection reveal that while the relative range possesses comparable probability behavior to the standardized range, it outperforms the standardized range in detecting outliers from skewed data. Furthermore, the simulated outlier detection framework has been applied to real datasets. The results showed that relative range continue to outperform standardised range in identifying outliers.

1. Introduction

In the field of data analysis, the identification and examination of outliers have emerged as a critical component that helps unveil significant patterns that change the data. Outliers are data points that deviate markedly from majority of the data. The presence of such outliers contains crucial information to data analysts since these points tend to disrupt many statistical characteristics of the data. Identifying these anomalies is essential because they can indicate errors in data collection, reveal unique insights, or even point to potential fraud or malfunction in systems. Outlier detection plays a pivotal role in ensuring the robustness and reliability of statistical analyses and machine learning models. The reasons for their occurrence are diverse, encompassing human error, such as data entry mistakes, sampling error involving items or individuals outside the intended population, and natural fluctuations in the data. To address this, statisticians have developed numerous methods to identify these unusual observations. Vigilantly monitoring outliers and employing suitable outlier detection techniques can pinpoint errors in the dataset, effectively cleansing it from irrelevant outcomes and enhancing its suitability for more accurate results. It is crucial to acknowledge that outliers can sometimes carry valuable information for the study; they may

Date: Received: Feb 21, 2024; Accepted: Mar 17, 2024.

* Corresponding author.

2020 *Mathematics Subject Classification.* Primary 46L55; Secondary 44B20.

Key words and phrases. Outlier detection, Range statistic, Interquartile range, Skewness, Relative Range.

be an integral part of the data structure. Consequently, understanding these data values and recognizing their role as a distinct variable becomes essential.

This paper is organized as follows. Section 2 discuss the numerous existing outlier detection techniques found in the literature. Section 3 introduces the methodology and proposes a more robust-to-outliers standardization of the range statistics. The resulting statistic is called "relative range" and denoted by K. The probability distribution of K is also explored empirically based on univariate skewed data according to the Weibull distribution. Further extensive simulation work is carried out to investigate the performance of K in detecting outliers. Section 4 presents the numerical results to assess the performance of K in detecting outliers and compares it to the performance of W by using real datasets. Finally, conclusions and some insight for future work are presented in section 5.

2. Background

In the field of statistics there exists numerous outlier detection approaches, each comes with its own characteristics and practical benefits. Below is a brief description of the most widely used methods.

A boxplot is a method to present outliers graphically using numerical data based on their quartiles. It was proposed by Tukey (1977) who defined outliers as the observations that are below the boxplot lower whisker defined by $Q_3 - 1.5(IQR)$ or above the boxplot upper whisker defined by $Q_3 + 1.5(IQR)$. However, its effectiveness is compromised when dealing with skewed data distributions. In addition, the outer fences may be too conservative for many practitioners causing many real outliers to be overlooked. Therefore, the fences have undergone several adjustments in the literature as can be seen in Hubert and Vandervieren (2008) and Dovoedo and Chakraborti (2015). These works achieved modifications to the whiskers of the boxplot to allow the detection of outliers in skewed distributions.

Rosner (1983) proposed the Extreme Studentized Deviate (ESD), a modification of the Grubbs' test for detecting outliers in a univariate dataset. The Grubbs' test is based on the assumption of normality and is used to identify a single outlier in a univariate dataset. The ESD test, on the other hand, is designed to detect multiple outliers by iteratively removing the most extreme value from the dataset and recalculating the test statistic until no outliers remain. It involves removing the most extreme value from the dataset and recalculating the test statistic iteratively until no more outliers are detected. The upper bound 'r' is the maximum number of outliers the test is designed to detect. The effectiveness of the proposed method is demonstrated through simulations and an application to real-world data. The ESD is known for its effectiveness in identifying outliers in datasets, even when there is a high proportion of potential outliers. The generalized ESD test is used in the following hypotheses:
H_0 = There are no outliers in the data set.

H_a = There are up to r outliers in the data set. The test statistic is:

$$R_i = \frac{max_i|x_i - \bar{x}|}{s}$$

with $\bar{x}$ and s denoting the sample mean and sample standard deviation, respectively.

Furthermore, in statistics, the range indicates roughly how diverse the data values are. It is a useful measure of the spread out of the dataset and can also be effective tool to gesture the existence of outliers. The existence of one extreme value in either end of the dataset will result in an entirely different and inflated range. Harter (1961) was the first to introduce the standardized range statistic and use it in an attempt to detect outliers. He defined this test statistic W as:

$$W = \frac{R}{\sigma} \tag{2.1}$$

from a sample of normal distribution, where R is the range of a sample and is the population standard deviation. In a random sample of n observations drawn from a normal population with standard deviation, Harter (1961) defined W in a way that ensures R is non-negative, and the distribution function of R is independent of the standard deviation, σ . While W can be a useful tool for detecting outliers, it has a primary shortcoming. When an independent estimate of is used, the statistic W becomes highly influenced by the existence of unusual values in the data because the standard deviation itself is inflated in value by the unusual or outlier values (Dallah and Sulieman, 2024). Consequently, using W as a tool to identify outliers can be misleading.

3. Methodology

Using W for outlier detection comes with a notable limitation. When an external estimation of σ is utilized (e.g., from the sample itself), W can be strongly affected by outliers. This occurs because outliers can inflate the standard deviation, potentially causing W to erroneously detect the presence of outliers, even when their significance may be minimal. It was previously mentioned the proposed statistic that standardizes the range R by a more robust statistic defined by the IQR. We call this statistic K, the Relative Range and define it as:

$$K = \frac{R}{IQR} \tag{3.1}$$

For a given pdf $f_x(x)$ of a random variable X, it can be shown that PDF of **R** is

$$f_R(r) = \int_{-\infty}^{\infty} n(n-1)[F(x_1 + r) - F(x_1)]^{n-2} f(x_1) f(x_1 + r)\, dx_1 \tag{3.2}$$

where $0 < r < \infty$ and $F_x(x)$ is the cumulative probability density function of X. Further, the PDF of **IQR** is derived to be:

$$\begin{aligned} f_Q(q) = \frac{n!}{\left(\frac{n}{4}-\frac{3}{4}\right)!\left(\frac{n}{2}-\frac{1}{2}\right)!\left(\frac{n}{4}-\frac{1}{4}\right)!} \int_{-\infty}^{\infty} [F_x(x_i)]^{\frac{n}{4}-\frac{3}{4}} [F_x(q + x_i) - F_x(x_i)]^{\frac{n}{2}-\frac{1}{2}} \\ \cdot [1 - F_x(q + x_i)]^{\frac{n}{4}-\frac{1}{4}} \cdot f_x(x_i) f_x(q + x_i)\, dx_i \end{aligned} \tag{3.3}$$

where Q = IQR and $0 < q < \infty$

Considering the complex form seen in equations (3.2) and (3.3), it is evident that constructing a closed analytical form of the pdf of K is nearly impossible. Dallah and her colleagues (2024) used the kernel density estimation based on simulated observation to construct the pdf of K for a normal distribution. In this paper, we conduct comparative analysis of the performance of W and K in detecting outliers from skewed distribution based on the widely used Weibull distribution.

3.1. **Weibull Distribution.** The Weibull distribution is an extensively applied distribution in reliability and life data analysis across many fields including engineering, biomedical sciences, and ecology. It proves particularly effective in describing wear-out or fatigue failures. The two-parameter Weibull distribution also finds practical utility in a range of applications. These include wind energy assessment, rainfall amount estimation, and analysis of material lifetimes. Depending on the parameter values, the Weibull distribution can effectively model various life behaviors. The shape and scale parameters play a crucial role, influencing distribution characteristics like the shape of the probability density function, reliability, and failure rate. The probability density function of the three-parameter Weibull distribution is:

$$f(x;\alpha,\beta,\mu) = \frac{\alpha}{\beta}\left(\frac{x-\mu}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x-\mu}{\beta}\right)^{\alpha}}, x > \mu, \alpha > 0, \beta > 0 \tag{3.4}$$

where α is the shape parameter, β is the scale parameter, and μ is the location parameter. While the cumulative density function is denoted as:

$$1 - e^{-\left(\frac{x-\mu}{\beta}\right)^{\alpha}}, x > \mu, \alpha > 0, \beta > 0 \tag{3.5}$$

Without loss of generality, we assume $\mu = 0$ in equation (3.2) and (3.3) The expected value and variance of X:

$$\begin{aligned} E[X] &= \beta\Gamma\left(1+\tfrac{1}{\alpha}\right) \\ \mathrm{Var}[X] &= \beta^2\left[\Gamma\left(1+\tfrac{2}{\alpha}\right) - \left(r\left(1+\tfrac{1}{\alpha}\right)\right)^2\right] \end{aligned} \tag{3.6}$$

such that $\Gamma(z) = \int_0^{\infty} x^{z-1}e^{-x}\,\mathrm{d}x$

The maximum likelihood estimates of α and β are:

$$\widehat{\beta} = \left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i^{\widehat{\alpha}}\right]^{\frac{1}{\widehat{\alpha}}} \quad \text{and} \quad \widehat{\alpha} = \frac{n}{\left(\frac{1}{\widehat{\beta}}\right)\sum_{i=1}^{n} x_i^{\hat{r}}\log x_i - \sum_{i=1}^{n}\log x_i} \tag{3.7}$$

$\widehat{\alpha}$ and $\widehat{\beta}$ are unbiased estimators of the parameters α and β Moreover, the Weibull distribution is related to other distributions. When $\alpha = 1$, the $Weibull(1,\beta)$ simplifies to the exponential distribution and when $\alpha = 2$ and $\beta = \sigma\sqrt{2}$ we have the Raleigh Distribution. Consequently, the Weibull distribution serves as a versatile placeholder for skewed distributions, showcasing its adaptability and role

in representing various distributional forms. The following subsection presents the results of the empirical work of W and K for skewed distribution.

3.2. **Empirical construction of the pdf of W and K.** RStudio is used as a programming tool for the empirical construction of probability distributions of W and K. Without loss of generality, we will set $\mu = 0$, for this work. The way the simulation works is by randomly sampling from Weibull distribution with shape $= \alpha = 2$ and scale $= \beta = 1$. While dealing with skewed distributions, only a one-sided fence (say on the upper side) is of interest. 100,00 random samples of specified sample size, n, are generated and the statistics W and K are calculated for each sample. The selected samples sizes are $n = 30, 50, 100, 250, 500,$ and 1000. Using the Kernel Density Estimation procedure, the following graphs show the pattern of the probability distribution of W and K. The distribution indicated

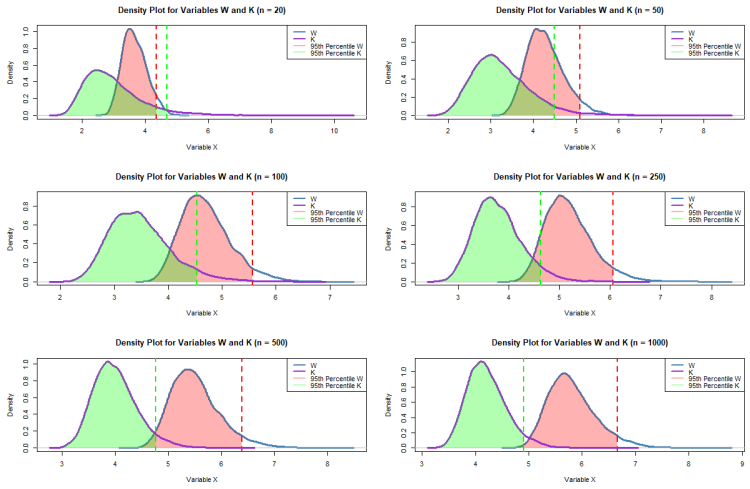

FIGURE 1. Empirical exploration of the distributions of W and K for n = 30, 50, 100, 250, 500, and 1000

by the blue curve with red highlighted area represents the empirical distribution for W while the purple curve with green highlighted area represents the empirical probability distribution for K. Both have comparable shapes. Moreover, we can see that the overlapping decreases as the sample size increases.

3.3. 95^{th} **percentile of the empirical distribution.** With the construction of the empirical distributions of W and K in the preceding subsection, it becomes clear that large values of these non-negative two statistics can flag the existence of outliers. Hence, an outlier region can be defined in the upper tail of their distributions. Table 1 gives the simulated 95^{th} percentiles (upper 5 percent of the distribution tail) for various sample sizes $n = 30, 50, 100, 250, 500$ and $1,000$.

By looking at the percentiles, we can expect the relative range to have a higher likelihood at detecting outliers since its percentile is smaller than the percentile of W for all sample sizes considered.

TABLE 1. 95^{th} percentiles of W and K

Sample Size	95^{th} percentile for W	95^{th} percentile for K
30	4.690923	4.567135
50	5.097515	4.492128
100	5.563626	4.534597
250	6.053882	4.629895
500	6.385800	4.763639
1000	6.666635	4.900996

4. Outlier Detection using Range Statistics

In this section, we will demonstrate how K, the relative range, is used to detect outliers in comparison to W. In subsection 4.1, the performance of K detecting outliers will be demonstrated through simulated work. After that, in subsection 4.2, the framework in the simulation work will be applied for the real datasets. In the interest of the simulation work, we will consider transforming the random variable X that follows Weibull distribution whose density is given in equation (3.4) for $\mu = 0$. Therefore, for $X \sim Weibull(\alpha, 1)$, one can obtain a transformed random variable Y such that:

$$Y = \frac{X}{\beta} \sim Weibull(\alpha, 1) \tag{4.1}$$

We can obtain the PDF of Y, $f(y; \alpha)$ which is clearly $X \sim Weibull(\alpha, 1)$:

$$f(y; \alpha) = \alpha y^{\alpha-1} e^{-(y)^{\alpha}}, y > 0, \alpha > 0 \tag{4.2}$$

4.1. **Simulation Work.** Using the 95th percentile we previously generated as the threshold, we are able to obtain the percentage of detecting at least one outlier. For each sample size, n, we contaminate each generated sample with 1, 3, and 5 points from a Weibull distribution ($\alpha = 2, \beta = 1$) with different locations (location = 2, 3, 4, 5, and 10). The figure below shows how the distribution of the contaminated values differs for different location values. Contaminating with location = 2 somewhat overlaps with our original Weibull distribution while contaminating with location = 5 significantly differs from the original distribution. The following equation describes how we obtained the belows graphs:

$$ContaminatedX = OriginalX + location\_contamination\_value \tag{4.3}$$

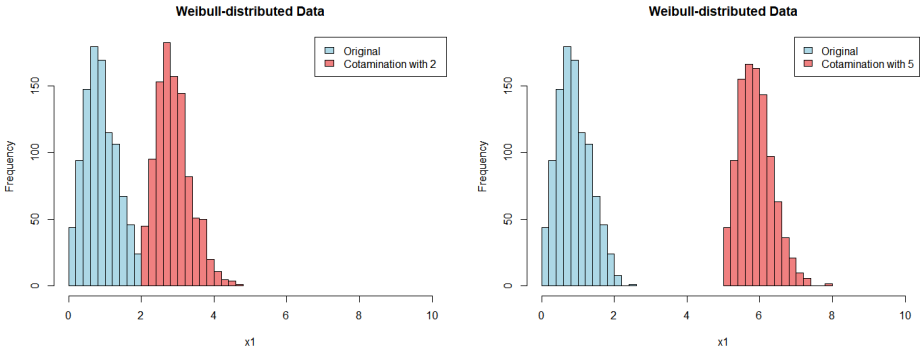

FIGURE 2. The effect of contaminating with different location values.

The relative frequency of detecting at least one outlier is calculated for each sample size and each location contamination. The results are presented in Table 2.

TABLE 2. Summary of the percent of outliers detected from simulated data for different sample size and mean

		cont = 1			cont = 3			cont = 5		
Sample Size	Location cont	Method W	Method K	Method ESD	Method W	Method K	Method ESD	Method W	Method K	Method ESD
$n = 30$	location = 2	47.60%	38.46%	75.41%	3.47%	39.56%	44.82%	0.09%	22.67%	9.70%
	location = 3	93.50%	83.74%	99.88%	0.37%	81.71%	67.43%	0.00%	63.44%	3.91%
	location = 4	9993%	98.73%	100.00%	0.02%	98.07%	81.24%	0.00%	91.03%	1.02%
	location = 5	100.00%	99.95%	100.00%	0.00%	99.88%	89.95%	0.00%	98.80%	0.25%
	location = 10	100.00%	100.00%	100.00%	0.00%	100.00%	99.97%	0.00%	100.00%	0.00%
$n = 50$	location = 2	50.4796	41.02%	76.79%	19.07%	53.16%	75.44%	2.96%	47.14%	44.54%
	location = 3	97.45%	90.18%	99.99%	20.32%	94.41%	99.09%	0.30%	91.65%	58.17%
	location = 4	100.00%	99.81%	100.00%	13.74%	99.85%	100.00%	0.01%	99.75%	63.40%
	location = 5	100.00%	100.00%	100.00%	6.61%	100.00%	100.00%	0.00%	100.00%	63.72%
	location = 10	100.00%	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%	100.00%	41.70%
$n = 75$	location = 2	49.65%	41.67%	76.88%	35.96%	59.74%	85.36%	12.62%	62.43%	74.46%
	location = 3	98.69%	93.90%	100.00%	68.21%	97.90%	99.97%	7.74%	98.03%	98.14%
	location = 4	100.00%	99.97%	100.00%	86.68%	9998%	100.00%	2.32%	99.99%	100.00%
	location = 5	100, 00%	100.00%	100.00%	93.89%	100.00%	100.00%	0.58%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	91.02%	100.00%	100.00%	0.00%	100.00%	100.00%
$n = 100$	location = 2	48.50%	41.87%	76.07%	46.77%	64.41%	89.56%	25.11%	69.01%	86.84%
	location = 3	99.13%	95.18%	100.00%	90.68%	98.99%	100.00%	31.66%	99.33%	99.96%
	location = 4	100.00%	999936	100.00%	99.74%	100.00%	100.00%	26.30%	100.00%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	16.57%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0.11%	100.00%	100.00%
$n = 250$	location = 2	42.35%	38.17%	73.06%	61.47%	66.97%	93.52%	62.22%	78.52%	96.72%
	location = 3.	99.16%	97.18%	100.00%	99.85%	99.82%	100.00%	99.27%	99.96%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 5	100, 00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
$n = 1,000$	location = 2	28.99%	28.27%	71.04%	56.32%	58.15%	93.02%	68.51%	73.73%	97.97%
	location = 3	96.95%	95.62%	100.00%	99.94%	99.95%	100.00%	99.99%	100.00%	100.0036
	location = 4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 5	100, 00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

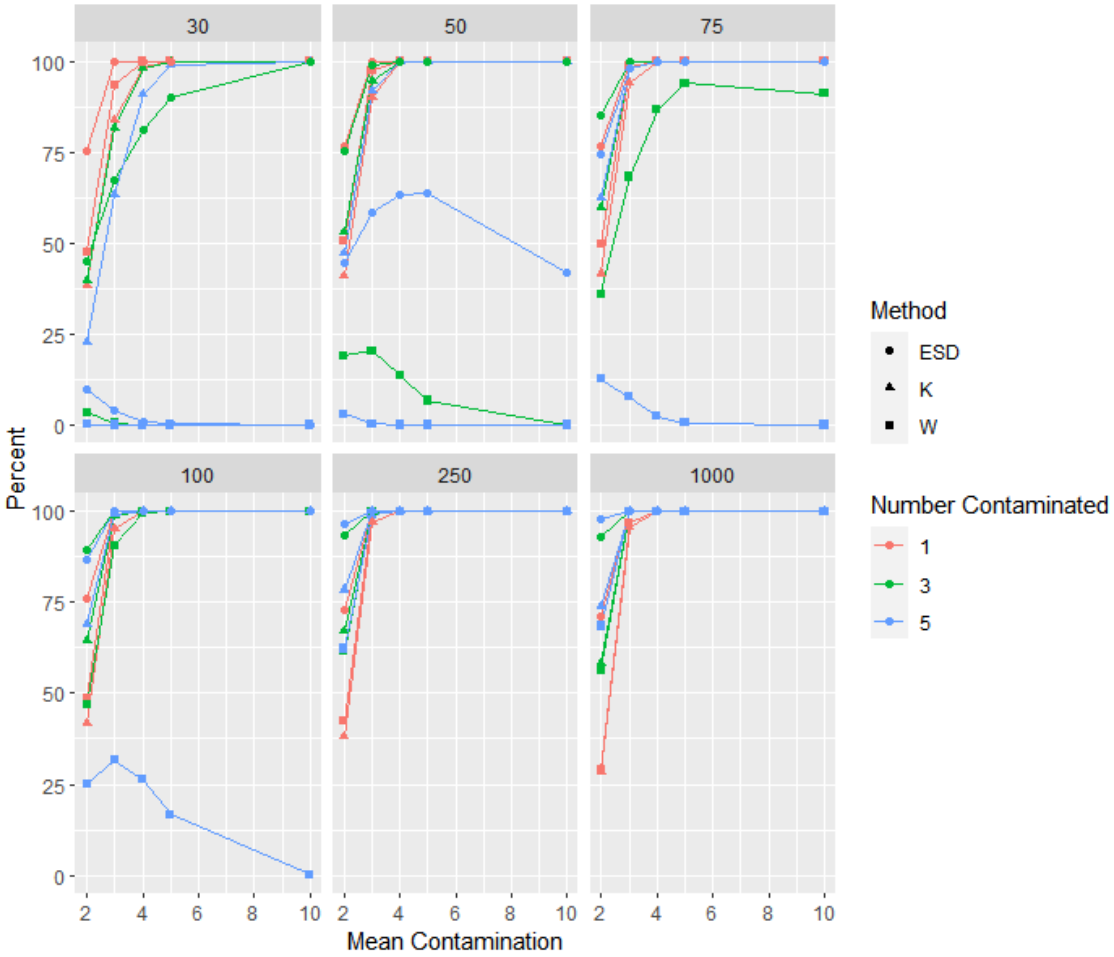

FIGURE 3. Line graphs plotted by method and the contamination number where: y-axis represents the percentage of detecting at least one outlier and the x-axis represents the value of the mean contaminated.

Analyzing Figure 3 reveals that the performance of the relative range surpasses that of variable W. This difference is illustrated by considering a scenario where, with a sample size of 30, the introduction of five data points as contaminants represents a significant proportion, constituting 20 percent of the dataset. However, after standardizing these points, they become "normal" within the dataset and no longer act as outliers, despite originating from a different location. Consequently, the detection of at least one outlier shows a low percentage. This pattern is consistent for sample sizes up to 100, as the presence of five contaminated data points constitutes 5 percent of the dataset, marking the threshold for outlier detection. The W statistic is heavily influenced by unusual values in the data, as the standard deviation is inflated by these outliers. Beyond this sample size, both statistics demonstrate comparable behavior. It becomes evident that variables W and relative range perform similarly, aligning with expectations, as

the number of contaminated data points remains insignificantly small compared to the overall dataset. Similarly, the ESD method performed comparably well with the relative range with a few instances showing the relative range having a higher percentage.

4.2. **Real Dataset.** In this section, we will be using three different real datasets to detect outliers using the standardised range and relative range. Below is the summary of the datasets:

TABLE 3. Summary of the real datasets used

Dataset	Sources	Application	# of Observation	Variable Used	Type
Spectrometer	UCI Repository	Kamalov et al. (2020)	531	Red_base_2	Continuous
Bladder Cancer	R Documentation	Bhatti et al. (2019)	128	Bladder Cancer	Continuous
Survival Times	R Documentation	Bjerkedal (1960)	72	Survival Times	Continuous

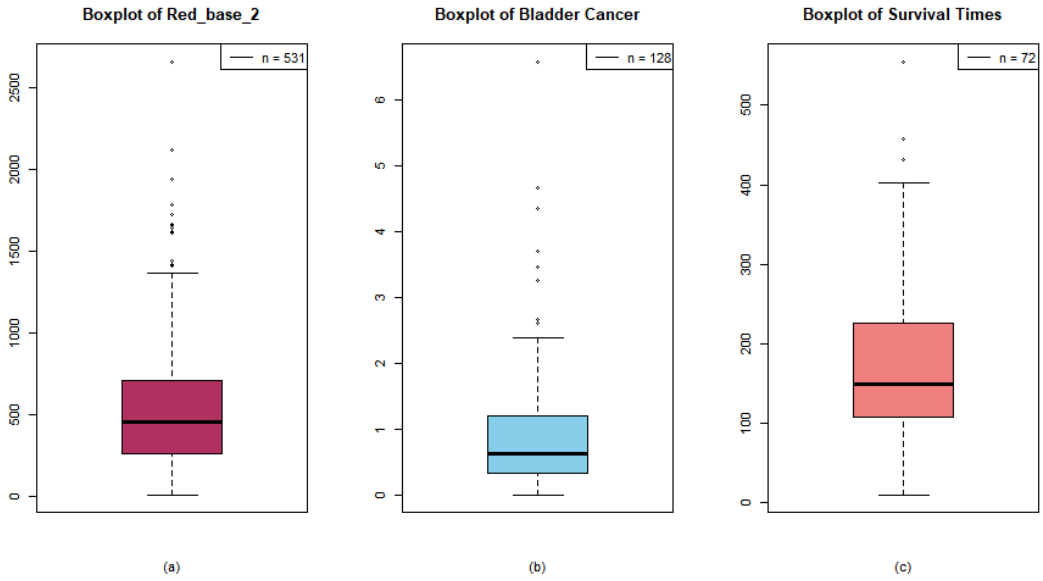

FIGURE 4. Boxplots of variable "Red_base_2" from spectrometer data, Bladder Cancer datapoints, and Survival Times datapoints

- **Red_base_2 - Spectrometer dataset:**
  As seen in the boxplot above, there seem to be possible outlier observations. By estimating the parameters α and β from equation (3.7), we obtain $\hat{\alpha} = 1.45$ and $\hat{\beta} = 539.94$. According to Kolmogorov-Smirnov test ($p\text{-}value = 0.5$), the data follows a Weibull distribution with the estimated parameters. Further, the data was transform according to equation (4.1). The resulting transformed data follows $X \sim Weibull(\hat{\alpha}, 1)$. Following the empirical construction of the 95^{th} percentile of Weibull distribution explained in subsection 3.3, the obtained thresholds for $\alpha = 1.45$ are W

= 6.99 and K = 5.46. Let WC and KC be the calculated W and K for the variable Red_base_2. The results are WC = 7.40 and KC = 5.97. By comparing these statistics with the above percentiles, one can conclude that the maximum is an outlier because $WC > W$ and $KC > K$.
It would be of interest to test if the second maximum observation shown in the boxplot in Figure 4(a) can be identified as an outlier by W, K, or both. To do that, we obtain WC and KC for the second maximum as 6.09 and 4.77, respectively. According to the above percentiles, it is clear that the second maximum cannot be considered an outlier. This outcome applies to both W and K.

- **Bladder Cancer dataset:**
The same procedure was applied to the continuous variable Bladder Cancer with the following results.
- $\hat{\alpha} = 1.05$ and $\hat{\beta} = 9.56$
- The 95^{th} percentiles of W = 7.01 and K = 7.41
- WC = 6.80 and KC = 9.14
Given that $WC < W$ and $KC > K$, we conclude that the relative range statistic has succeeded in detecting the maximum observation as an outlier unlike the standardised range. Subsequently, we can test out if the second maximum is an outlier. The new WC and KC are:
WC = 6.58 and KC = 7.60
From the percentiles provided above, it's evident that $WC < W$ and $KC > K$. Consequently, we deduce that the relative range statistic effectively identifies the second maximum observation as an outlier. This underscores K's robustness in outlier detection compared to W.
- **Survival Times dataset:**
The same procedure was applied to the continuous variable Survival Times with the following results.
- $\hat{\alpha} = 1.83$ and $\hat{\beta} = 199.87$
- The 95^{th} percentiles of W = 5.41 and K = 4.62
- WC = 5.27 and KC = 4.70
Given that $WC < W$ and $KC > K$, the relative range statistic has succeeded in detecting the maximum observation as an outlier. This aligns with our conclusion from the simulation work, indicating that K demonstrates superior performance for smaller sample sizes.

5. Conclusion

The area of statistics offers numerous techniques for detecting outliers, providing vital insights into dataset characteristics. Outlier detection is a widely employed research area with applications spanning various domains. In this paper, we explore the use of the relative range, a more robust-to-outliers standardization of the range statistics, to identify outliers in univariate skewed data. Our exploration extends to the evaluation of Harter's (1961) proposed technique alongside our proposed relative range, aiming to assess their robustness in outlier detection considering Weibull distribution as a placeholder for skewed data. The empirical

construction of the PDFs of the standardized range and relative range from the Weibull distribution along with the simulation work for outlier detection reveal that while the relative range possesses comparable probability behavior to the standardized range, it surpasses the standardized range in effectively detecting outliers from skewed data. Simulated outlier detection framework has been applied to real datasets. The relative range (K) proved more robust in simulation and with real data due to its use of the interquartile range (IQR), which is less sensitive to outliers, in contrast to W, which utilizes the standard deviation, a measure that can be inflated by outliers.

References

1. Bhatti, F., Hamedani, G., Korkmaz, M., Cordeiro, G., Yousof, H. & Ahmad, M. On Burr III Marshal Olkin family: development, properties, characterizations and applications. *Journal Of Statistical Distributions And Applications*. **6** pp. 1-21 (2019) https://doi.org/10.1186/s40488-019-0101-7
2. Bjerkedal, T. & Others Acquisition of Resistance in Guinea Pies infected with Different Doses of Virulent Tubercle Bacilli.. *American Journal Of Hygiene*. **72**, 130-48 (1960)
3. Dallah, D., Sulieman, H., & Alzaatreh, A. Outlier Detection Based on Range Statistic: Empirical Evaluation and Comparisons. *Statistical Outliers and Related Topics*. Taylor and Francis (2024)
4. Dallah, D. & Sulieman, H. Outlier Detection Using the Range Distribution. *Advances In Mathematical Modeling And Scientific Computing*. pp. 687-697 (2024) https://doi.org/10.1007/978-3-031-41420-6_57
5. Dovoedo, Y. & Chakraborti, S. Boxplot-based outlier detection for the location-scale family. *Communications In Statistics-simulation And Computation*. **44**, 1492-1513 (2015) https://doi.org/10.1080/03610918.2013.813037
6. Harter, H. Use of tables of percentage points of range and studentized range. *Technometrics*. **3**, 407-411 (1961)
7. Hubert, M. & Vandervieren, E. An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*. **52**, 5186-5201 (2008) https://doi.org/10.1016/j.csda.2007.11.008
8. Kamalov, F. & Leung, H. Outlier detection in high dimensional data. *Journal Of Information & Knowledge Management*. **19**, 2040013 (2020) https://doi.org/10.1142/S0219649220400134
9. Rosner, B. Percentage points for a generalized ESD many-outlier procedure. *Technometrics*. **25**, 165-172 (1983)
10. Tukey, J. & Others Exploratory data analysis. (Reading, MA,1977)

6. Appendix

Table 4. Summary of the percent of outliers detected from simulated data for different sample sizes and means for $\alpha = 1$

$\alpha = 1$							
		cont = 1		cont = 3		cont = 5	
Sample Size	Location Contamination	Method W	Method K	Method W	Method K	Method W	Method K
$n = 30$	location = 2	3.40%	4.26%	1.81%	2.17%	0.98%	0.67%
	location = 3	4.30%	6.28%	0.66%	4.20%	0.15%	1.33%
	location = 4	9.20%	11.27%	0.20%	8.55%	0.02%	3.07%
	location = 5	18.92%	21.02%	0.04%	15.56%	0.00%	6.21%
	location = 10	86.13%	89.75%	0.00%	77.60%	0.00%	48.43%
$n = 50$	location = 2	4.20%	4.43%	2.70%	2.86%	1.98%	1.58%
	location = 3	3.75%	5.37%	1.48%	4.40%	0.73%	2.88%
	location = 4	6.19%	8.47%	0.95%	8.50%	0.32%	5.51%
	location = 5	14.62%	16.34%	0.69%	17.38%	0.15%	11.93%
	location = 10	93.46%	93.69%	0.07%	90.74%	0.00%	81.25%
$n = 75$	location = 2	4.26%	4.96%	320%	3.62%	2.56%	2.28%
	location = 3	4.23%	5.55%	2.25%	5.04%	1.52%	3.70%
	location = 4	5.14%	7.77%	1.80%	8.57%	0.87%	7.32%
	location = 5	10.54%	13.71%	1.76%	16.99%	0.51%	15.56%
	location = 10	95.87%	96.06%	0.92%	95.49%	0.01%	92.94%
$n = 100$	location = 2	4.48%	4.87%	3.60%	3.73%	2.98%	3.00%
	location = 3	4.14%	5.49%	2.83%	4.94%	1.86%	4.26%
	location = 4	4.99%	7.25%	2.52%	8.01%	1.32%	7.72%
	location = 5	8.83%	11.66%	2.80%	15.88%	1.07%	15.49%
	location = 10	96.97%	97.00%	3.85%	97.37%	0.12%	96.45%
$n = 250$	location = 2	4.69%	4.62%	4.29%	4.18%	3.88%	3.73%
	location = 3	4.50%	4.79%	3.76%	4.79%	3.12%	4.66%
	location = 4	4.63%	5.39%	3.87%	6.49%	2.99%	6.75%
	location = 5	5.69%	7.25%	4.95%	10.44%	3.42%	12.14%
	location = 10	98.05%	97.70%	48.84%	99.01%	7.69%	99.33%
$n = 1,000$	location = 2	4.70%	4.64%	4.58%	4.57%	4.47%	4.41%
	location = 3	4.61%	4.68%	4.51%	4.73%	4.19%	4.66%
	location = 4	4.65%	4.86%	4.45%	5.22%	4.29%	5.48%
	location = 5	4.85%	5.25%	4.99%	6.46%	4.72%	7.34%
	location = 10	87.61%	86.60%	80.05%	97.12%	63.58%	98.85%

[1] Department of Mathematics, American University of Sharjah, UAE.
Email address: g00078476@aus.edu

[2] Department of Mathematics, American University of Sharjah, UAE
Email address: hsulieman@aus.edu

[3] Department of Mathematics, American University of Sharjah, UAE
Email address: aalzaatreh@aus.edu

TABLE 5. Summary of the percent of outliers detected from simulated data for different sample sizes and means for $\alpha = 3$

$\alpha = 3$							
		contamination = 1		contamination = 3		contamination = 5	
Sample Size	Location Contamination	Method W	Method K	Method W	Method K	Method W	Method K
$n = 30$	location = 2	96.94%	86.39%	2.48%	88.04%	0.00%	75.79%
	location = 3	100.00%	99.82%	0.02%	99.80%	0.00%	98.93%
	location = 4	100.00%	100.00%	0.00%	100.00%	0.00%	99.99%
	location = 5	100.00%	100.00%	0.00%	100.00%	000%	100.00%
	location = 10	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
$n = 50$	location = 2	98.85%	92.59%	48.56%	97.26%	1.30%	96.40%
	location = 3	100.00%	99.99%	34.25%	100.00%	0.00%	100.00%
	location = 4	100.00%	100.00%	14.26%	100.00%	0.00%	100.00%
	location = 5	100.00%	100.00%	3.37%	100.00%	0.00%	100.00%
	location = 10	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
$n = 75$	location = 2	99.45%	95.93%	94.95%	99.39%	27.00%	99.40%
	location = 3	100.00%	100.00%	99.88%	100.00%	7.01%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	0.58%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	0.03%	100.00%
	location = 10	100.00%	100.00%	99.90%	100.00%	0.00%	100, 00%
$n = 100$	location = 2	99.60%	96.96%	99.46%	99.74%	74.39%	99.84%
	location = 3	100.00%	100.00%	100.00%	100.00%	69.52%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	46.79%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	20.98%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%
$n = 250$	location = 2	99.77%	98.71%	100.00%	100.00%	100.00%	99.99%
	location = 3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	100.00%	100, 00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
$n = 1,000$	location = 2	99.41%	99.05%	100, 00%	100.00%	100.00%	100.00%
	location = 3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

TABLE 6. Summary of the percent of outliers detected from simulated data for different sample sizes and means for $\alpha = 4$

$\alpha = 4$							
		contamination = 1		contamination = 3		contamination = 5	
Sample Size	Location Contamination	Method W	Method K	Method W	Method K	Method W	Method K
$n = 30$	location = 2	99.96%	98.37%	0.55%	98.90%	0.00%	96.23%
	location = 3	100.00%	100.00%	0.01%	100.00%	0.00%	99.99%
	location = 4	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 5	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 10	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
$n = 50$	location = 2	100.00%	99.62%	49.54%	99.95%	0.16%	99.94%
	location = 3	100.00%	100.00%	20.28%	100.00%	0.00%	100.00%
	location = 4	100.00%	100.00%	3.37%	100.00%	0.00%	100.00%
	location = 5	100.00%	100.00%	0.22%	100.00%	0.00%	100.00%
	location = 10	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
$n = 75$	location = 2	100.00%	99.90%	99.69%	99.98%	18.84%	100.00%
	location = 3	100.00%	100.00%	100.00%	100.00%	1.04%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	0.01%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%
	location = 10	100.00%	100.00%	90.69%	100.00%	0.00%	100.00%
$n = 100$	location = 2	100.00%	99.96%	100.00%	100, 00%	81.58%	100.00%
	location = 3	100.00%	100.00%	100.00%	100.00%	56.67%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	20.48%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	3.42%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%
$n = 250$	location = 2	100.00%	100.00%	100.00%	100.00%	100.00%	99.99%
	location = 3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
$n = 1,000$	location = 2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

TABLE 7. Summary of the percent of outliers detected from simulated data for different sample sizes and means for $\alpha = 10$

$\alpha = 10$							
		contamination = 1		contamination = 3		contamination = 5	
Sample Size	Location Contamination	Method W	Method K	Method W	Method K	Method W	Method K
$n = 30$	location = 2	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 3	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 4	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 5	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 10	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
$n = 50$	location = 2	100.00%	100.00%	0.51%	100.00%	0.00%	100.00%
	location = 3	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 4	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 5	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
	location = 10	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
$n = 75$	location = 2	100.00%	100.00%	75.18%	100.00%	0.00%	100.00%
	location = 3	100.00%	100.00%	26.47%	100.00%	0.00%	100.00%
	location = 4	100.00%	100.00%	3.83%	100.00%	0.00%	100.00%
	location = 5	100.00%	100.00%	0.26%	100.00%	0.00%	100.00%
	location = 10	100.00%	100.00%	0.00%	100.00%	0.00%	100.00%
$n = 100$	location = 2	100.00%	100.00%	100.00%	100.00%	98.90%	100.00%
	location = 3	100.00%	100.00%	100.00%	100.00%	0.80%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	0.00%	100, 00%
	location = 5	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%
$n = 250$	location = 2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100, 00%	100.00%
$n = 1,000$	location = 2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	location = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%