SIMULATION STUDY ON THE PERFORMANCE OF ROBUST OUTLIER LABELLING METHODS

BY

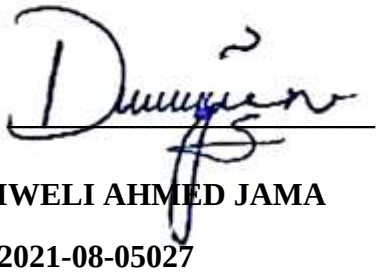
ABDIWELI AHMED JAMA
REG: 2021-08-05027

A RESEARCH THESIS SUBMITTED TO THE SCHOOL OF MATHEMATICS IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE MASTER OF SCIENCE IN STATISTICS OF KAMPALA INTERNATIONAL UNIVERSITY

OCTOBER, 2023.

DECLARATION

I am Abdiweli Ahmed Jama, student at Kampala international University Uganda, I declare that t his work is as a result of an independent investigation and in circumstances where it's under obli gation to the work of other people, due acknowledgment has been made.

Sign:  **Date: 22-10-2023**

ABDIWELI AHMED JAMA

Reg. 2021-08-05027

APPROVAL

I affirm that the study conducted "***SIMULATION STUDY ON THE PERFORMANCE OF ROBUST OUTLIER LABELLING METHODS***" has been under my supervision and is now ready for submission to Kampala International University with my supervision.

Sign: _____ **Date: 22-10-2023**

NAME OF THE SUPERVISOR

DR. BABANGIDA IBRAHIM BABURA

DEDICATION

Many thanks to the Almighty ALLAH for granting me with life, strength, wisdom and health as I dedicate this research to my beloved parents and especially my mother and father for all the love, encouragement, material and moral support but also a special appreciation to all my brothers, sisters, colleagues and other relatives who's support me one day for financial, advice and any kind of encouragement, and to my supervisor for his guidance and help; without them my studies would not have been a success.

ACKNOWLEDGEMENTS

I thank the Almighty Allah for enabling me maneuver through all the tough, hard times and trying moments I have had in life. My dream of this award would not have become true without His guidance, protection and assurance that all things are possible if you believe in him.

I acknowledge the staff members of Kampala International university especially my supervisor, Furthermore, I acknowledge all my lecturers for having sacrificed their time and efforts to ensure my success during the course of the study plus all my panelists for their guidance.

My special thanks also go to my parents for the financial, emotional and moral support during my study. May Allah Almighty bless you. Special regards to my family, who have always supported, protected and wished me all the best for life. I don't have enough words to thank you but all I can say is that I will always be grateful for everything you have done for me and pray to the Allah to grant each one of you all your wishes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC Computation	…….....Approximation Bayesian
ANOVA	…….....Analyze of variance
AO	……..... Additive Outliers
AR	…….....Autoregressive Models
ARMA Average	…….....Autoregressive Moving
ATM	…….....Automated Teller Machine
CDF function	…….....Cumulative Distribution
CDR Ratio	…….....Coefficient of Determination
CFTP	…….....Coupling from The Past
CODB	…….....Class Outlier Distance Based
DGP	…….....Data generating process
ESD	……..... Extreme Studentized deviate
GEV	…….....Generalized Extreme Value
IO	…….....Innovational outliers
MADe	…….....Median Absolute Deviation
MCMC	…….....Markov chain Monte Carlo
MDRs	……..... Member Dissatisfaction Rates
PCPs	……..... Primary Care Physicians
PDF	…….....Probability Density Function
Q1	……..... First quartile and
Q3	…….....Third quartile
SIQR	……..... Split Inter Quarterlife Range
SMC	……..... Sequential Monte Carlo
SMC	…….....Sequential Monte Carlo

ABSTRACT

The identification and labeling of outliers play a crucial role in data analysis and modeling tasks. Robust outlier labeling methods aim to accurately identify observations that deviate significantly from the majority of the data points while being resilient to noise, measurement errors, and data corruption. In this simulation study, we evaluate the performance of various robust outlier labeling methods using synthetic datasets.

To conduct the study, we defined the simulation setup by specifying the characteristics of the datasets, including the number of variables, sample size, distributional assumptions, and proportion of outliers. Synthetic datasets were generated based on these specifications, incorporating both normal observations and outliers with known characteristics.

A set of robust outlier labeling methods was selected for evaluation. These methods were designed to effectively handle outliers and provide reliable labels. Implementation of the selected methods was carried out using a programming language, ensuring proper application to the generated datasets.

Performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) were defined to assess the effectiveness of the outlier labeling methods. Each method was applied to the synthetic datasets, and the results were recorded. The performance metrics were calculated based on the known labels of the synthetic outliers.

The collected results were analyzed and compared to identify the strengths and limitations of each robust outlier labeling method. The performance metrics were used to assess accuracy, robustness, and computational efficiency. To ensure the reliability of the findings, the simulation study was repeated with different simulation setups and datasets, validating the consistency of the results across multiple iterations.

Based on the findings, conclusions were drawn regarding the performance of the evaluated robust outlier labeling methods. The most effective methods for the specific characteristics of the datasets used in the study were identified. These findings provide valuable insights for researchers, practitioners, and data analysts in choosing appropriate outlier labeling methods for their data analysis and modeling tasks.

In summary, this simulation study contributes to the understanding of the performance of robust outlier labeling methods and provides a systematic evaluation framework for comparing and selecting suitable methods in the presence of outliers.

CHAPTER ONE:

GENERAL INTRODUCTION

1.0. Introduction

This chapter consists different sections: the Background statement of problem Outlier Detection Method, Importance of Detecting Outliers and Causes of Outliers. In the basic ideas of an outlier are discussed such as definitions, features, and reasons to detect outliers. In the Outlier Detection Method section, characteristics of the two kinds of outlier detection methods are described briefly: formal and informal tests. Importance of Detecting Outliers in generally and finally what causes of outlier.

1.1. Background of the study

Outliers in data collection with unusually large or small values are frequently seen in observed variables. Some data sets may come from groups that are comparable to one another, while others may come from mixed groups with diverse features related to a particular variable, such as height data that is not gender- stratified. Inaccurate measurements, such as data entry errors, or coming from a different than demographic the rest of the data might be the source of outliers. If the measurement is accurate, it shows an uncommon occurrence. Considerable outliers have three characteristics.

1) Outliers generally serve to increase error variance and reduce the power of statistical tests.
2) If non-randomly distributed, they can decrease normality (and in multivariate analyses, violate assumptions of sphericity and multivariate normality), altering the odds of making both Type I and Type II errors.
3) They can seriously bias or influence estimates that may be of substantive interest.

The following example simply shows how one outlier can highly distort the mean, variance, and 95% confidence interval for the mean. Let's suppose there is a simple data set composed of data points 3, 4, 5, 6, 7, 8, 9 and its basic statistics are as shown in Table 1. let's substitute data point 9 with 99. As shown in Table 2, the mean and variance of the data are much higher than that of the original data set due to one unusual data value, 99. The 95% confidence interval for the mean

is also much broader because of the large variance. It may cause potential problems when data analysis that is sensitive to a mean or variance is conducted.

Table 1.1. Basic statistics example.

Mean	**median**	**Variance**	**95% confidence for the mean**
6	6	4.67	4.00 to 8.00

Table 1.2: After changing 99 our example one

Mean	median	Variance	95% confidence for the mean
18	6	1251.81	-13.86 to 51.58

The second aspect of outliers is that they can provide useful information about data when we look into an unusual response to a given study. They could be the extreme values sitting apart from the majority of the data regardless of distribution assumptions. The following two cases are good examples of outlier analysis in terms of the second aspect of an outlier: 1) to identify medical practitioners who under- or over-utilize specific procedures or medical equipment, such as an x-ray instrument; 2) to identify Primary Care Physicians (PCPs) with inordinately high Member Dissatisfaction Rates (MDRs) (MDRs = the number of members complaints / PCP practice size) compared to other PCPs

In summary, there are two reasons for detecting outliers. The first reason is to find outliers which influence assumptions of a statistical test, for example, outliers violating the normal distribution assumption in an ANOVA test, and deal with them properly in order to improve statistical analysis. This could be considered as a preliminary step for data analysis. The second reason is to use the outliers themselves for the purpose of obtaining certain critical information about the data as was shown in the above examples.

1.2. Statement of Problem

The robust outlier labelling method especially the boxplot methods are quite effective when working with large data sets that are fairly normally distributed, many distributions of real-world data do not follow a normal distribution. Such us highly skewed, usually to the right, and in such cases the distributions are frequently closer to a skewed distributions like lognormal or Weibull

distributions than a normal one (Felix Famoye1, 2018) Recent studies on outlier labelling such as (Hubert and Babura, 2017). incorporate skewness in the labelling mark-up so that the labelling rule can overcome the limitation of the initial boxplot methods by (Tukey, 1977) and others. Since the limitation of all the procedures cannot handle in the general complexity of datasets, so it's important to investigate the performance of the procedures and make much informed decisions and suggestions using them.

1.3 Main objective

To compare algorithm method and identifying the best method in outlier labelling methods according to sample size.

1.4. Objective of the study

1. Review the existing boxplot outlier labeling methods.
2. Formulate a simulation algorism to compare the performance of revised methods.
3. Explore the application of the best method in outlier labelling for some real-life datasets

1.5 Research Questions

1. Is the outlier labelling method wrongly marked regular observations as outlier?
2. Is the outlier labelling method wrong marked outliers as regular observations?
3. Which among the outlier labelling methods has the best labelling mark up?

1.6. Significance of the study

This study will bring to the literature on decision making on chosen a best outlier labelling method in a different data structure.

1.7. Outlier Detection Method

There are two kinds of outlier detection methods: formal tests and informal tests. Formal and informal tests are usually called tests of discordancy and outlier labeling methods, respectively. Most formal tests need test statistics for hypothesis testing. They are usually based on assuming some well-behaved distribution, and test if the target extreme value is an outlier of the distribution, i.e., whether or not it deviates from the assumed distribution. Some tests are for a single outlier and others for multiple outliers. Selection of these tests mainly depends on

numbers and type of target outliers, and type of data distribution. Many various tests according to the choice of distributions are discussed in (Barnett and Lewis , 1994) and (Iglewicz and Hoaglin , 1993). reviewed and compared five selected formal tests which are applicable to the normal distribution, such as the Generalized extreme Studentized deviate Kurtosis statistics, Shapiro-Wilk, the Boxplot rule, and the Dixon test, through simulations. Even though formal tests are quite powerful under well-behaved statistical assumptions such as a distribution assumption, most distributions of real-world data may be unknown or may not follow specific distributions such as the normal, gamma, or exponential. Another limitation is that they are susceptible to masking or swamping problems.

1.8. Masking effect

It is said that one outlier masks a second outlier if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier.

1.9. Swamping effects

It is said that one outlier swamps a second observation if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation.

Many studies regarding these problems have been conducted by (Barnett and Lewis , 1994) (Hoaglin, 1993) (Davies and Gather, 1993) and (Kale, 1987) On the other hand, most outlier labeling methods, informal tests, generate an interval or criterion for outlier detection instead of hypothesis testing, and any observations beyond the interval or criterion is considered as an outlier. Various location and scale parameters are mostly employed in each labeling method to define a reasonable interval or criterion for outlier detection. There are two reasons for using an outlier labeling method. One is to find possible outliers as a screening device before conducting a formal test. The other is to find the extreme values away from the majority of the data regardless of the distribution. While the formal tests usually require test statistics based on the distribution assumptions and a hypothesis to determine if the target extreme value is a true outlier of the distribution, most outlier labeling methods present the interval using the location and scale parameters of the data. Although the labeling method is usually simple to use, some observations

outside the interval may turn out to be falsely identified outliers after a formal test when the outliers are defined as only observations that deviate from the assuming distribution.

However, if the purpose of the outlier detection is not a preliminary step to find the extreme values violating the distribution assumptions of the main statistical analyses such as the t-test, ANOVA and regression, but mainly to find the extreme values away from the majority of the data regardless of the distribution, the outlier labeling methods may be applicable.

In addition, for a large data set that is statistically problematic, e.g., when it is difficult to identify the distribution of the data or transform it into a proper distribution such as the normal distribution, labeling methods can be used to detect outliers

(Wiley, 1998)

1.10. Importance of Detecting Outliers

Outlier detection plays an important role in modeling, inference and even data processing because outlier can lead to model misspecification, biased parameter estimation and poor forecasting (Tsay, Pena and Pankratz, 2000) and (Fuller, 1987).Outlier detection as a branch of data mining has many important applications, and deserves more attention from data mining community.

The identification of outliers may lead to the discovery of unexpected knowledge in areas such as credit card and calling card fraud, criminal behaviors, and cybercrime, etc. (Mansur and Sap, 2005).Detection of outliers in the data has significant importance for continuous as well as discrete data sets (Chen, Miao and Zhang, 2010). (Justel and Pena , 1996). proved that the presence of a set of outliers that mask each other will result in failure of the Gibbs sampling (In Bayesian parametric model Gibbs sampling is an algorithm which provides an accurate estimation of the marginal posterior densities, or summaries of these distributions, by sampling from the conditional parameter distributions) with the result that posterior distributions will be inadequately estimated.

(Iglewicz and Hoaglin , 1994). recommend that data should be routinely inspected for outliers because outliers can provide useful information about the data. As long as the researchers are interested in data mining, they will have to face the problem of outliers that might come from the real data generating process (DGP) or data collection process. Outliers are likely to be present

even in high quality data sets and a very few economic data sets meet the criterion of high quality (Zaman, Robusseeuw and Orhan, 2001).

1.11. Causes of outliers

Outlies are data points that significantly deviate from the normal or expected pattern of dataset. There are several causes of outliers, including

1.11.1 Human errors for example data entry errors.

Data entry errors, such as transposition mistakes, typographical errors, and omissions, are common human errors that can result in inaccurate or incomplete data records, impacting the reliability and quality of the entered information. Implementing robust quality control measures and providing comprehensive training can help minimize these errors and ensure accurate data entry.

1.11.2 Human errors for example measurement errors

Measurement errors can occur due to human errors in the process of taking or recording measurements. These errors may involve incorrect readings, inaccurate calibration of instruments, or inconsistent measurement techniques, leading to deviations from the true values. Proper training, standardized procedures, and regular equipment calibration can help reduce measurement errors caused by human factors and improve the accuracy of recorded measurements.

1.11.3 Data processing errors for example data manipulation

Data processing errors, such as data manipulation mistakes, refer to errors that occur during the manipulation and transformation of data. These errors can involve miscalculations, improper data formatting, incorrect data joins or merges, or flawed data filtering or sorting operations, leading to inaccurate or distorted results. Implementing rigorous quality assurance processes, conducting thorough data validation checks, and utilizing automated data processing tools can help minimize these errors and ensure the integrity and reliability of processed data.

1.11.4 Sampling errors for example extracting data from wrong source

Sampling errors can occur when data is extracted from the wrong source during the sampling process. This could involve mistakenly selecting data from an incorrect population or sampling frame, resulting in biased or unrepresentative samples that do not accurately reflect the target population of interest. Careful attention to the sampling methodology, ensuring proper identification and selection of the correct data sources, can help mitigate these errors and improve the validity of the sample.

1.11.5 Not an error, the value is extreme, just a Novelty in the data

When a data point exhibits an extreme value or novelty, it signifies an outlier in the dataset. Outliers can indicate potential anomalies, errors, or unique observations that differ significantly from the rest of the data, and they should be carefully examined to determine their validity and impact on the analysis or modeling process.

CHAPTER TWO

REVIEW OF LITERATURE OF STUDY

2.1. Outlier defined

An outlier is a data point that differs significantly from other data points in the dataset, deviates from expected normal behavior, or closely resembles an abnormal behavior that has been characterized by (Hodge and Austin, 2004). and (Chandola, Banerjee , 2009) This definition uses the terms "substantially different," "does not conform to the anticipated normal conduct," and "conforms well to a defined deviant behavior" are quite arbitrary and demand careful examination; as a result, the definition of an outlier is ambiguous. Outliers are frequently referred to as anomalous data points; a data point is considered anomalous if it deviates from predicted typical behavior. We use the term "inlier" to describe a data point that closely resembles the expected typical behavior. Data points that are not outliers are frequently referred to as "inliers." Different domains' outliers have different characteristics from one another.

An outlier in a credit card transaction differs greatly from one in a set of weather data. As a result, different applications define outliers differently. A dataset may contain outliers for a variety of reasons, including nefarious activity, instrument mistakes, setup errors, changes in the environment, human error, and disaster. Regardless of the cause, outliers may be fascinating to the user since they provide them with information that is different from that of typical data. Although some people view outliers as issues and others as attractive objects, in any case they cannot be avoided (V.Chandola, A. Banerjee, and V. Kumar , 2009), (Barnett & Lewis , 1994), in brief: "Outliers are fascinating and come in a variety of shapes and sizes in many kinds of applications.

2.2. Boxplot outlier labelling

To enhance the efficiency of boxplot-based outliers' detection for skewed data, a variety of box-plot techniques have 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 Q1 and Q3 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.

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 [Q1 – 1.5 IQR*;* Q3 + 1.5 IQR] where Q1, Q3 and IQR stand for the first and the third quartiles, and the interquartile range, respectively. Several researchers have reports that the Tukey's boxplot is fitted to symmetrical data (Walker and Chakrabori, 2013).

2.3. Types of outliers

Types of outliers is defined is into three types according to the Isolated individual data points in a dataset, A data point is isolated with respect to the context. A particular group of data points appear as outliers with respect to the entire dataset.

2.3.1. Type I Outliers

Isolated individual data points in a dataset are termed as Type I outliers. By definition they are the simplest type and very easy to identify. Intuitively they are far from other data points in the dataset in terms of attribute values.

2.3.2. Type II Outliers

A data point is isolated with respect to the context. Typically, data in this type of dataset has other contextual attributes (e.g., time and location). An outlier is far from other data points in the same context in terms of value. This is a little bit different from a Type I outlier; a Type I outlier is a data point isolated from all the other data points in the dataset. A Type II outlier was first investigated in time series data in the late seventies. (Barnett & Lewis , 1994) defined Type II outliers as the Additive Outliers (AO) for time series data. The good thing about additive outliers is that they do not influence the other data points in the context, hence they are easy to identify

2.3.3. Type III Outliers

A particular group of data points appear as outliers with respect to the entire dataset. No data point in a small subset is an outlier with respect to the other points in the subset, but as a group, they are the outliers. For contextual data like time series, the entire dataset forms a sequence; hence a particular subsequence is an outlier with respect to the entire sequence. (Barnett & Lewis, 1994) called them Innovations Outliers (IO) for time series data. The bad thing about innovations outliers is that they influence other data points of the same context and try to hide themselves; therefore, it is difficult to identify innovations outliers.

A data stream has one temporal context with each data point; so, it might have a type II or type III outlier but never a type I outlier. This is because data streams are considered as infinite series

and the processing has to be online. Therefore, at any particular moment, only a subset of the entire dataset is present, and so a data point cannot be an outlier with respect to an entire dataset. Regardless of the type of outliers, outlier detection is a popular branch of application. We discuss the problem of outlier detection and its significance in the next section

2.4. History of outliers

Outliers have been found in data sets since the 18th century through analysis. About 200 years ago, (Bernoulli, 1777) brought attention to the technique of removing the outliers. Although deleting outliers is not the best way to deal with them, this was a popular practice in the past. The first statistical method was created in 1850 to deal with the issue of outliers in the data (Beckman and Cook, 1983).

Extreme observations, according to some researchers, should be maintained with the data since they provide important details about the data. As an illustration, (Bessel and Baeuer, 1838) asserted that simply because severe observations stand out from the rest of the data. (Barnett, 1978)

(Legendre, 1805) advises against erasing the extreme observations that have been "judged too enormous to be tolerable." Some researchers choose to remove extreme observations from the data since they skew estimates. Boscovitch, a 19th-century astronomer, disregarded Legendre's advice and caused them to be deleted (ad hoc correction), maybe favoring (Pierce, 1852), (Chauvenet, 1863) or (Wright, 1884).

According to (Chartier, 2010) outliers should be removed because they are almost often the result of erroneous activity. The debate over whether to remove or maintain outliers from data is as contentious today as it was 200 years ago.

Studies on how to deal with outliers have been done by (Bendre and Kale , 1987), Davies and (Gather, 1993), (Iglewicz and Hoaglin, 1994) and (Barnett and Lewis, 1994).Finding unique examples in a dataset known to contain distance-based outliers is frequently done by defining outliers according to their distance to nearby examples.

Approach for detection. Class Outlier Distance Based (CODB) outlier's detection approach was introduced by (Saad and Hewahi , 2009) who demonstrated its superiority to the distance-based

outlier's detection method. Because it yields a more accurate estimate of the mean and other statistical parameters in an international geological reference material, (Verma, 1997)

places emphasis on the detection of outliers in univariate data as opposed to accommodating the outliers (RM).

2.5. Simulation and bootstrap

The bootstrap's effectiveness can be attributed in part to its simplicity, which allows for the replacement of theoretical knowledge by repetitive calculations using random samples. What a wise choice of a name! The most common concept is known as the plug-in rule or, less formally, the substitution principle, which explicitly acknowledges that frequentist inference entails the replacement of an unknown probability distribution F with an estimate F. A random sample Y = (Y1,...,Yn) is supplied in the simplest scenario, and the empirical distribution function F serves as the nonparametric estimate. A parametric model is then used. F(y;ψ) with a parameter ψ Depending on the situation, one may choose between parametric and nonparametric estimates. Semiparametric estimates are also frequently used, notably in regression situations. The imposition of constraints, as in situations involving hypothesis testing, or for technical reasons, such as to increase a rate of convergence, can change the estimate of F. The second notion was to replace analytical calculation of an estimator's $\hat{\theta}$ attributes with computational methods, foreseeing that the coming era of affordable computing will democratize data analysis.

The second idea was to substitute analytical calculation of attributes of an estimator $\hat{\theta}$ f an unknown parameter $\hat{\theta}$ =$\hat{\theta}$ (F) in anticipation of the era of low-cost computing that would democratize data analysis. (Efron, B, 1979)

This provides the well-known method of independent sampling from the fitted model F and the use of the accompanying estimate to generate R duplicate bootstrap samples $(Y_1^*, \ldots, Y_n^*)$. $\hat{\theta}_1^*, ,.. \hat{\theta}_R^*$ to determine the repeated sampling characteristics of $\hat{\theta}$ We refer to this type of resampling, specifically in nonparametric circumstances, as the conventional bootstrap. The fact that, under mild conditions, $\hat{\theta}$ (.) can be the result of an algorithm of nearly arbitrary complexity is significant because it dispels the misconception that a parameter is a Greek letter that appears in a probability distribution and demonstrates the potential for uncertainty analysis for the sophisticated procedures that are used today but were unthinkable 25 years ago.

Combining these two concepts results in a very adaptable inference tool that is appealing from a variety of angles. Because of its clear connections to current practices, it can be used by non-expert practitioners in a wide range of applications and is still open to theoretical investigation. (Bickel and Freedman , 1981). among others, looked at the circumstances under which bootstrap inference is reliable and developed useful mathematical tools in the process. Researchers were motivated to expand the applicability of the first sample technique as a result of several "smoking guns" such as (Bretagnolle, 1983). pointing at instances of boot-strap failure; we go over two similar methods to this. (Beran, 1977) and (Putter and van Zwet, 1996) provide additional recent theoretical analyses. The discovery that the bootstrap might provide higher-order accuracy for confidence intervals, similar to Edgeworth correction of traditional normal intervals, but less unpleasant and error-prone, and thus more reliable in reality, was a significant step forward (Singh, 1981). Peter Hall and his coauthors are largely responsible for the latter entwining of classical asymptotic and the boot-strap, as outlined by (Hall, 1992) (Hall, 1986) was especially influential.

The creation of generally accurate non-parametric confidence intervals is a major topic in theoretical bootstrap literature. The two major methods for doing this are based on the direct usage of quantiles from the bootstrap replicates and the building of Studentized pivots $\hat{\theta}_1^*,,..\hat{\theta}_R^*$

The first method, which takes its cue from the student's t statistic, calls for an estimated variance V^* for $\hat{\theta}^*$ based on the same bootstrap sample. The quantiles of $Z^* = (\hat{\theta}^* - \hat{\theta})/V^{*1/2}$ then offer consistent estimators of the quantiles of $Z = (\hat{\theta} - \hat{\theta})/V^{1/2}$ for a wide class of estimators that fall into the "smooth function model" (Hall, 1992). This is demonstrated via an Edgeworth expansion argument. Due to this, second-order accurate so-called Studentized bootstrap or bootstrap t confidence intervals for $\hat{\theta}$ are produced, i.e., the likelihood that a one-sided interval with nominal level $1-\alpha$ includes θ, then it is $1-\alpha+O(n^{-1})$. With a difference of $O(n^{-1/2})$ from the nominal probability, the coverage error for the associated normal confidence interval is reduced. According to one interpretation of the second strategy, sampling from the posterior distribution of O given 0 is approximated by using re-sampling of $\widehat{\theta^*}$ conditional on θ. By doing this, two-sided confidence intervals of the pattern $(\widehat{\theta^*}_{(R\alpha i)}, \widehat{\theta^*}_{(R(1-\alpha_2)}$ are produced, where $\widehat{\theta^*}$) is the rth ordered boot-strap replicate. The percentile intervals are produced using the simplest and crudest option, $\alpha_1 = \alpha_2 = \alpha$, but the associated one-sided intervals are only first-order

accurate, therefore improvements have been sought that empirically define α_1 and α_2 ((Efron, 1987); (DiCicco, 2004 and 1992); and (Efron, 1996). Similar to Studentized bootstrap intervals, the resulting bias-corrected and accelerated (BC_a) intervals and their variations are second-order accurate. BC_a intervals, on the other hand, are transformation-invariant, unlike the Studentized intervals. The former intervals perform slightly better, in part because sporadic instability in the variance estimate V can sometimes result in overly long intervals, according to numerical studies, which has shown that both Studentized bootstrap and BC_a intervals typically show a modest under coverage. Prepivoting ((Beran, 1987, 1988). which involves bootstrap correction of bootstrap techniques and typically entails a double or nested bootstrap computation, can sometimes significantly improve poor coverage of confidence intervals.

A null hypothesis H_0, which imposes restrictions on the distribution of the data, such as fixing a mean value, is one of a hypothesis test's key components and a test statistic T, whose high values provide evidence against H_0. The significant probability or P-value Pobs = Pro(T > tobs), where tobs is the actual value of T observed, and the probability is calculated under a null hypothesis distribution, measures the degree of disagreement between the data and Ho. Calculation under the null hypothesis distribution is required for bootstrap estimation of Pobs. This is typically done by simulating an estimate $\overline{F_0}$ that fulfills H_0.

In the nonparametric scenario, the null hypothesis might include altering F's support, the resampling probabilities associated with Y1,..., Y, or the empirical distribution function in some other way; for a discussion of this, opening paragraphs. The resulting bootstrap tests are often nearly identical to permutation tests in comparison test scenarios; the key distinction is the use of sampling with and without replacement.

If the test is based on a pivot, the sample plan doesn't need to be altered. Assume, for instance, that Ho means that 0 equals some fixed number 0o and that the test's foundation is (0 - 0)/V1/2. If the null hypothesis is true, the observed value of a random variable with a distribution that is very similar to (0*-0) is tobs = (Oobs- 0) / Vobs12! Due to its pivotality, V*l/2 was produced by simulation from either Fo or F. Therefore, in this relatively unique situation, simulation using a properly designed null distribution is not required (Hall and Wilson, 1991). A less complicated method that is occasionally accessible is to compare including a null hypothesis value of 0 in a

confidence range for 0 with Pobs > a ((Beran, 1986). This coincides with the previous use of a pivot,

if the confidence set was created using that pivot. However, it should be noted that the shape of the confidence set needs to be carefully considered for null hypotheses other than points; for instance, one-sided tests equate to one-sided intervals with scalar parameters. Also,

Finding exchangeable components to which resample-ling can be done is essential when the data are independent but not symmetrically distributed. When a regression model sets $Y_1 = h_1(\psi, \epsilon_1), \ldots . Y_n = h_n(\psi, \epsilon_n)$ where the ϵ_j form a random sample with distribution function G, many times these components are residuals of some kind.

When $Y_j = h_j(\widehat{\psi}, \widehat{\varepsilon_j})$ where if is an estimate of 4, is used, then G is normally estimated using the empirical distribution function of residuals 'j.

Then, independent sampling of $\varepsilon_1^*, \ldots, \varepsilon_n^*$ for $\widehat{\varepsilon_1}, \ldots, \widehat{\varepsilon_n}$ and setting $Y_J^* = h_j(\widehat{\psi}, \widehat{\varepsilon_j})$ can be used to create a bootstrap data set. When a specific parametric model is crucial to the study, such as when determining if time series data follow a particular introgressive moving average model, such model-based bootstrapping may be extended to dependent data scenarios. In these situations, pure nonparametric bootstrapping is challenging and parametric or semi parametric models are typically required. In certain circumstances, stratified resampling is required.

where the strata are picked based on empirical data to exhibit variation that is roughly uniform. The "wild" boot strap ((Wu, Hardle, 986,1989, 1990); (Mammen, 1993), in which each residue establishes its own stratum, may be the most extreme example of this.

The subjects outlined above received in- depth coverage from (Davison and Hinkley , 1997) with references to the primary literature.

2.5.1 Simulation study

Simulation studies are computer experiments that involve creating data by pseudo-random sampling. A key strength of simulation studies is the ability to understand the behavior of statistical methods because some "truth" (usually some parameter/s of interest) is known from the process of generating the data. (Davison, A. C., & Hinkley, D. V. , 1997)

2.6. Significance of Outlier Detection

Outlier detection refers to the problem of identifying the outliers in a dataset. Since the definition of outliers is vague and application-dependent, a formal method for outlier detection is not yet developed. By definition, an outlier detection technique takes a dataset as input and outputs the outliers. Despite the vagueness of outliers, several approaches are popular for outlier detection based on the state of the input data. The first approach is called the supervised approach where the outlier detection technique assumes the availability of labeled data A supervised technique collects

knowledge from labeled data and applies the collected knowledge to unlabeled data for outlier detection. The second approach is called the semi-supervised approach which requires only the inliers or the outliers to be labeled. Both of these approaches are less popular due to the lack of labeled datasets. (Hodge, V. J., & Austin, J. , 2004)

The third and final approach is called the unsupervised approach which does not require any type of labeling, hence, is very popular for outlier detection. However, unsupervised techniques often suffer from higher false alarms. Outlier detection in a dataset is a crucial step in the data assimilation process since outliers are less intuitive than regular data points and prompt a user's curiosity to examine their reasons. Outlier detection is carried performed by many applications for various goals. Intrusion detection is one of the most frequently used functions. Outliers are typically caused by incursion; hence their existence is a strong indicator of an intrusion. Other significant goals include data cleaning, damage detection, defect identification, novelty detection, and data damage detection (for example, for medical and public health data). There are many uses for and endless potential in outlier detection for data streams. To spot irregularities instantly, practically every monitoring system needs live outlier detection. In this part, we go over a few illustrations of outlier detection in single- and multiple-stream applications.

Outlier detection technique for single streams is appropriate where data points from one stream is independent from those from other streams. In that case each stream can be processed independently and therefore we call it single stream application. Typical applications of outlier detection for single stream include fraud detection, fault detection, error detection, etc. Fraud detection refers to the problem of detecting unauthorized transactions in bank accounts, credit cards, insurance agencies, cell phone companies, etc. (Aggarwal, C. C., & Yu, P. S. , 2001)

The transactions from one user are referred to as one stream here. Because transactions from one user are independent of transactions from other users, this form of application carries independent streams. Fraudulent transactions can be recognized through outlier detection because they are, by definition, markedly different from legitimate ones.

Nowadays, many of these programs are available online, allowing for real-time transaction monitoring. This sector, where transactions are tracked and fraudulent activity is immediately identified, has considerable potential for outlier detection for single data streams and (Basu, S., & Meckesheimer, M. , 2007) proposed the use of data stream outlier detection to identify instrument failures. Practically, this method can be applied in any industry where it is possible to continuously monitor machine state. Each machine generates a single stream of machine status information, and the statuses of different machines are unrelated to one another. By using outlier detection, a malfunction of any instrument can be identified, helping to prevent major harm caused by a catastrophic fault.

Since errors are found offline using the offline storage and process approach, they may go unnoticed for a while and result in serious repercussions. Each weather station, which is installed in various geographic regions, measures distinct meteorological characteristics. Due to their excessive distance from one another, weather stations frequently exhibit very low correlation and are analyzed as separate streams. These stations are frequently situated in inaccessible areas that are challenging to directly monitor. They may generate incorrect values for a variety of reasons, including instrument malfunction, erratic behavior, and incorrect configuration. Error-prone values can be found by using the outlier detection method. (Zhang, J., & Wang, C. , 2003)

Discordant observations are ones that, when joined with other data, exhibit a distinct pattern in terms of their law of frequency (Edgeworth, 1887) cited by (Beckman and Cook, 1983). Another definition of a discordant observation is one that the researcher perceives as unexpected or discrepant (Iglewicz and Hoaglin, 1993).An outlier is an observation that stands out significantly from the other observations in the sample in which it is found. These claims serve as examples of how an outlier is a post-data, subjective notion. In the past, dealing with outliers required using "objective" procedures only after the outliers had been discovered through a visual examination of the data (Grubbs, 1969) cited by (Beckman and Cook, , 1983). An observation that comes

from a distribution that differs from the distribution of the other data is referred to as a contamination. The investigator might or might not record any contaminants (Barnett, 1984)

Outliers are known to both be contaminants and conflicting findings. Iglewicz therefore defined outliers as findings that are discordant with the remaining data (Iglewicz and Hoaglin, 1994) According to Hawkins, an outlier is an observation that differs so significantly from other observations that it raises questions about whether it was produced by a separate mechanism (Hawkins, 1980)An outlier is an observation "that appears to deviate markedly from other members of the sample in which it occurs" (Grubbs, 1969)

2.7. Outliers in Survey Data Sets

2.7.1. Problem in Questionnaire

Outliers are anticipated to arise in the data because the questionnaire design may include certain unclear questions that neither the enumerator nor the responder can understand. If only revenue is given in the questionnaire, for instance, without mentioning the time frame (monthly, annual, etc.), the responder may interpret it as monthly income in most cases, leading to an outlier in the data on annual income. Similar to the information on monthly income, an economic Incorrectly creating an outlier on the positive side of the monthly income distribution, a graduate respondent would interpret it as annual income rather than salary. (Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R., 2009)

2.7.2. Problem Arising out of Enumerators‟ Mistakes

Outliers may potentially have been created by the enumerators themselves. In the same scenario as before, if one out of twenty enumerators misunderstand the difference between yearly and monthly income, roughly 5% of the data will be flagged as outliers. (Fowler Jr, 2014)

2.7.3. Problem in Explaining Question by the Enumerator to Respondent

Similar to the previous situation, an outlier may show up if the enumerator fails to explain the question to the respondent while the questionnaire is being filled out. For instance, if an enumerator asks a respondent about their family's income but does not define it for some of the respondents, there may be outliers. (Dillman, D. A., Smyth, J. D., & Christian, L. M. , 2014)

2.7.4. Outliers Arising out of Misunderstanding on the Part of Respondent

Most respondents in the poor world lack familiarity with the questionnaire design, most likely due to illiteracy. As a result, they give a lackadaisical response or simply guess at an answer until they comprehend the subject. The appearance of outliers may also be caused by lack of interest in the response or by responses based on a hunch or assumption. (Ghosh, 2006)

2.7.5. Poor Handwriting of the Enumerator

One of the potential reasons for outliers in survey data could be the consequence of the enumerators' illegible handwriting, which the data entry operator might not comprehend and fill out incorrectly.

2.7.6. Problem in Data Entry by the Data Entry Operator

Outliers could result from a data entry operator error. An advertent increase of one zero may result in a massive income gain of 70,000 to 700,000, creating outliers. Such situations may occur when the data entry operator's task is to copy data from a questionnaire to a database but he is not sufficiently conversant with the project at hand.

Any such outliers that result from an error made during the data collection or documentation process can be removed or corrected to reflect the true population. However, it is necessary to keep the outliers resulting from natural variation since they are anticipated to reveal an intriguing background to the data collection procedure and that particular observation. (Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. , 2013)

2.8. Effects of Outliers

Outliers may affect the statistics in a positive or negative way. If these observations are accurate, they reveal several intriguing facets of the data. Due to an anomaly, the infamous (Hadlum v. Hadlum) case from (1949) (Barnett, 1978) is of statistical relevance. After Mr. Hadlum had departed for his duties in the military, Mrs. Hadlum gave birth to a child 349 days later. Given that the average gestation time lasts about 280 days, a gestation term this abnormally long will be viewed as an oddity. Due to the court's unique and statistically erroneous limit of 360 days for gestation, Mr. Hadlum's petition was rejected.

This anomaly appears to deviate from the gestation period distribution, however it actually occurred and was a natural anomaly. However, if the outlier develops as a result of a mistake, it will hinder data analysis. For instance, if ten dice are thrown ten times, and the numbers of sixes are recorded as 2, 0, 3, 12, 0, 1, 1, then 12 will undoubtedly be an outlier in the data in addition to displaying a missing value. Without paying attention to the outlier, such data analysis will produce inaccurate or deceptive conclusions.

2.8.1. Damaging Effects of Outliers

Outliers in the data can raise the errors variance and reduce test power, which can have a significant impact on parameter estimation (Zimmerman, 1994, 1995, 1998).If there are any outliers in the errors, these outliers.

In the case of univariate data, they reduce their normalcy, and in the case of multivariate data, they sphericity and multivariate normality, which changes the likelihood of making Type I and Type II mistakes. In this approach, Type I and Type II errors are the fault of outliers. Finally, the outlier's skew regression results that could be of real interest (Overbay, 2004)

CHAPTER THREE

METHODOLOGY

3.1. Introduction

In this chapter, we introduce the necessary statistical tools that are required for the implementation of the methodological process in achieving the aims and objectives of this research work. We begin with sample quantiles which are essential for boxplot construction. The boxplot construction methods and their assessment tools using outside rate. The robust skewness measures especially the Bowley coefficient and medcouple are introduced in detail along with their important properties. The bootstrap method along with the Monte Carlo simulation process were also introduced.

3.2 Quantile Measures

3.2.1. Sample Quantiles

Let Q(p) be the sample quantile at quantile position p [0,1], and let $X = \{x_{(1)} \ldots., x_{(n)}\}$ denote the order statistics of a sample X of n independently identically distributed random variables (i.i.d. Next, a distribution F's quantile is given by $Q(p) = F^{-1}(p) = \inf\{x: F(x) \geq p\}$.

There are several similar approaches to define quantile estimates, however in this article we focus on two of the order statistics-based formulations offered by (Hyndman, 1996)

These definitions have a general form that is given by $\hat{Q}_1(P) = (1-\gamma)x_j + \gamma x_{j+1}$)and is a representation based on weighted averages of sequential order statistics.

Where $\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$ for some $m \epsilon \mathbb{R}$ and $0 \leq \gamma \leq 1$. The value of γ is a function of $j = [pn + m] and\ g = pn + m - j\ with$ $\lfloor \cdot \rfloor$ denoting the greatest integer operation

3.2.2. Boxplot Quantiles

The three sample quartiles Q_1 , Q_2 and Q_2 that are used to compute the boxplot five summary statistics in the majority of statistical software are known as the boxplot quantiles [Hyndman and Fan (1996)]. Using the criterion that the two "hinges" are variations of the first and third quartiles, or near to quantiles of $at\ p = \{0.25,0,75\}.$, the quartiles are retrieved. For odd n, the hinges match the quartiles, while for even n, they diverge. The hinges do so additionally for $n \cong 2$ mod 4, the quartiles do so just for $n \cong 1$ mod 4, and they are in between two consecutive observations in all other cases. In an ordered sample X of size n, the first and third quartiles are given by $Q_1 = x_{(k)}$and $Q_3 = x_{(n-k+1)}$ $where\ k = \frac{1}{2}[(\frac{n+1)}{n}$.

3.2.3. Median-unbiased and Distribution-Free Quantiles

According to Hyndman and Fan (1996), the quantile of a sample X is defined as the median position given by $MF(X_{(k)} \approx k - \frac{1/3}{n+1/3}$ such that the sample quantile is determined by setting $P_k = k - \frac{\frac{1}{3}}{n+\frac{1}{3}}$. Thus, $\hat{Q}_i(Pk) = x_k$.represents the quantile estimate. Linear interpolation can also be used to derive the quantile estimates at point $\hat{Q}_i(P_k)$. As a result, Reiss (1989) found that the quantile that results from p is median unbiased of order $o(\frac{1}{\sqrt{n}})$ According to Reiss (1989), the quantile $\hat{Q}_i(P_k)$ is distribution-free and superior to all other median unbiased quantile estimators. Among other characteristics, p_k possesses translation equivariant.

Due to the distribution-free nature of $\hat{Q}_i(P_k)$ if we assume that X follows the GEV distribution, we can still derive the estimate $\hat{Q}_i(P_k)$ on X even if the fitting parameters of X with the GEV distribution are unknown. Unless otherwise noted, we implemented all estimation processes that called for quantile estimates in this research effort using the benefits of the quantile concept in this section.

3.3. The Boxplot Construction

The three quartiles $Q_1, Q_2\ and\ Q_3$must typically be estimated before a boxplot can be constructed. These quartiles are used to summarize a set of data by highlighting its key characteristics. The rectangular box that represents the center half of the dataset and is bound at the first and third quartiles, respectively, gives the chapter 1 discussion its name. The position of a typical central value known as the second quartile or median $Q_2$is indicated by a line drawn across the box. The spread and placement of the batch are represented by these two aspects, which normally draw the majority of the viewer's attention.

The outermost observations that are not extreme enough to be labeled as potential outliers by an exploratory rule of thumb or Tukey's resistance rules are captured by the two lines that are drawn outward from the two ends of the box to the two neighboring values marked as fence values. Any anomalous observations will be represented on the display as individual points, drawing attention to them and allowing for further investigation as potential anomalies.

The most crucial metrics still employed in boxplot development are quartiles. Simply put, they serve as the foundation for the fence description that generates the rules for identifying probable outliers. Specifically, if Q_1 and Q_3 represent the lower and higher quartiles, respectively, the fences are calculated at $Q_1 - k(Q_3 - Q_1)$ and $Q_3 + k(Q_3 - Q_1)$), where k = 1.5 for the traditional boxplot design. This suggests that, as highlighted by (Frigge et al, 1989)the number of observations that may be possible outliers can be affected by the definition of the quartiles and that of k.

We list all eight of the (Frigge et al, 1989).examined lower quartile Q1 definitions, which are all expressed in terms of the ordered sample points $x_1, x_2, \dots \dots, x_n$from a univariate distribution. But lower quartile definitions need that $(\frac{n}{4}) = j + b$ with $j = 1,2, \dots n$, such that $0 \leq b \leq 1$. The eight lower quartile Q1 definitions are as follows

Definition 1: The lower quartile is the Weighted Average at $x_{(n/4)}$ and is given by

1. $Q_1 = (1 - b)x_j + bx_{j+1}$

Definition 2: The lower quartile is the Observation Numbered Closest to $\frac{n}{4}$, given $Q_1 = x_i$ with I being the integer part of $\frac{n}{4} + 0.5$

Definition 3: The lower quartile is the Empirical Distribution Function given as $Q_1 = x_{(j)}b = 0, x_{(j+1)}b > 0$

Definition 4: The lower quartile is the Weighted Average Aimed at $x((n + 1/4$, given by $Q_1 = (1 - b)x_j + b_{j+1}$ where $\frac{n+1}{4} = j + b.$

Definition 5: The lower quartile is the Empirical Distribution Function with Averaging given by

$$Q_1 = \frac{x_{(j)} + x_{(j+1)}}{2} b = 0, x_{(j+1)}b > 0$$

Definition 6: The lower quartile is the Standard Fourths for Hinges given by

$Q_1(1 - b)x_j + bx_{(j+1)}$. Where $\frac{1}{2}\left(\frac{n+3}{2}\right) = j + b \; with \; b = 0 \; or \; b = 0.5$

Definition 7: The lower quartile is the Ideal Fourths and is given by

$$Q_1 = (1 - b)x_{(j)} + bx_{(j+1)}$$

Where $\frac{n}{4} + \frac{5}{12} = j + b, with \; 0 \leq b \leq 1$

Definition 8: The lower quartile is the Weighted Average Aimed at $x = (\frac{n}{2} + \frac{1}{2})$ and is given by $Q_1 = (1 - b)x_j + bx_{(j+1)}$. Where $\frac{n}{4} + \frac{1}{2} = j + b, with \; 0 \leq b \leq 1$

By inserting p =¼4 into the appropriate definitions of the 100pth percentile, definitions 1 through definition 5 and definition 8 can be obtained. In a similar manner, substituting p =¾4 will yield the upper quartiles. However, (Frigge, 1989) recommended only using Definitions 4 to 8 as the acceptable quartile estimates for boxplot development. Details regarding the effects of these selections for the quartile definition for boxplot building were also provided by (Frigge et al, 1989). Keep in mind that the boxplot quantiles definition in Subsection 1.2 and Definition 6 yield the same outcome. For the inner fence, (Tukey, 1977)advised using k = 1.5, and for the outer fence, k = 3.

Table 3.1: Simulation studies of some outside rate per sample (1 B(k, n)) of a Gaussian samples, on selected values of k and n (Frigge et al., 1989).

N				
	1.0	1.5	2.0	3.0
10	0.424	0.198	0.094	0.026
20	0.577	0.232	0.082	0.011
30	0.705	0.284	0.094	0.008
50	0.837	0.365	0.094	0.004
100	0.967	0.523	0.115	0.003

for the outer fence, which is known as the general guideline for identifying observations that deviate from the norm. Other early suggestions include those made by (Macneil, 1977) for k= 1.0 and K=1.5 and by (Ingelfinger et.al , 1983)for k=2.

By offering some guidance on the outside rate per sample using a statistical measure, (Hoaglin et al., 1986) explored inconsistency over a standard choice of the value of k. The measure can be calculated by calculating the likelihood that a random sample of n will contain one or more external observations. For a Gaussian sample, (Hoaglin et al, 1986) denote the rate as 1-B(k,n), where B(k, n) is the probability that the sample of size n does not contain any outliers at any constant k. In Definition 6, the quartile produces values of 1-B(k, n) whose relationship to n for a specific k depends in part on the remainder of n mod 4. As n gets bigger, 1-B(k, n) for every fixed k must get closer to 1. n Table 1.1, the column for k = 1.0 clearly illustrates that the rate 1-B(k, n) approaches 1 as n increases, and the column for k = 1.5 obviously but less dramatically demonstrates the same tendency of the rate 1-B(k, n).

According to (Frigge et al, 1989) k = 1.5 is the recommended value for the majority of exploratory rules for Gaussian samples. However, they also note circumstances in which k = 2.0 or k = 3.0 should be used and point out that k = 1.0 is too small based on the strong numerical evidence.

3.4. Robust Measures of Skewness

Based on a distribution's skewness, one can gauge a distribution's degree of asymmetry and shape. Asymmetric distributions with long tails to the right typically have positive skewness, while those with longer tails to the left typically have negative skewness. Symmetric distributions typically have zero skewness. According to traditional statistics, the skewness coefficient of a univariate dataset, $Y_n = \{Y_1, Y_2, \ldots\ldots\ldots\ldots, Y_n\}$, sampled from a continuous distribution, is given by the formula: $\rho(Y_n) = \frac{m_3(Y_n)}{m_2(Y_2)}$where m3 and m2 refer to the dataset's third and second distributional moments, respectively.

The measurement's sensitivity to data contamination (outliers) is a drawback. Any contamination in the right tail of such a sample can increase the measure, which will lead to incorrect interpretation in either scenario. A single outlier in the left tail of a symmetric or right-tailed sample can greatly affect the value of to become negative. Another drawback of is that the measure cannot be obtained if any of the two moments disappears.

These drawbacks are overcome by the widely used robust measures of skewness, which also present a moment-independent measure that is resistant to outliers. The two robust skewness measures, the Bowley coefficient and the medcouple Skewness measure introduced by (Bowley, 1926)and (Brys et al, 2004). respectively, are of interest to the technique in this framework.
These drawbacks are overcome by the widely used robust measures of skewness, which also present a moment-independent measure that is resistant to outliers. The two robust skewness measures, the Bowley coefficient and the medcouple skewness measure introduced by (Bowley, 1926) and (Brys et al, 2004). respectively, are of interest to the technique in this framework.

3.4.1. Bowley Coefficient of Skewness

The literature has long recognized the validity of the robust metrics of location and dispersion. For instance, the interquartile range (IQR) can be used for dispersion and the median Q2 can be used for location. Using quantile estimates of a random sample $\boldsymbol{Y_n}$., the median and interquartile

range can be calculated. According to this custom, Bowley (1926) suggests a skewness coefficient that is based on quantile estimates and is denoted by $\delta(\boldsymbol{Y_n}) = \frac{Q_3+Q_1+2Q_2}{Q_3-Q_1}$ where the first three quartiles of $Q_i, i = 1,2,3$. On the range (-1, 1), the coefficient will have limited values. The numbers 1 and -1 denote severe right and left skewness, respectively, whereas 0 denotes absolute symmetry. However, just like other traditional skewness measures, the coefficient has preserved the skewness ordering of two distributions, making it relatively easy to understand as a skewness measure.

In their general representation of Equation for the random variable Y with the distribution function (F, Groeneveld and Meeden , 1984) provided.

2. $\gamma_{(\alpha)} = \frac{F^{-1}(1-\alpha)+F^{-1}(\alpha)-2v_y}{F^{-1}(1-\alpha)-F^{-1}(\alpha)}$

With $0 < \alpha < 0.5$ and v_y is the median determine at F^{-1} (0.5). The quantile of Y is represented by the measure F^{-1} (α) It is clear to see that at = 0.25, Equation becomes Equation. To get two further versions of the robust skewness measure $Y_{(\alpha)}$ (Groeneveld and Meeden, 1984) first integrated the numerator and denominator of Equation with respect to on the interval (0, 0.5).

3. $\gamma_1 = \int_0^{0.5} \frac{[F^{-1}(1-\alpha)+F^{-1}(\alpha)-2v_y]d\alpha}{\int_0^{0.5}[F^{-1}(1-\alpha)-F^{-1}(\alpha)]d\alpha} = \frac{\mu_y-v_y}{E|Y-\mu_y|}$

Allowing 0 in Equation (1.2) to obtain the limiting value as well as the second iteration of the skewness coefficient across a finite period (a,b).

$\gamma_2 = \lim_{a\to 0} \gamma_{(\alpha)} = \frac{b+a-2v_y}{b-a}$.

According to (Groeneveld and Meeden's , 1984) classification of tolerable skewness, all four of the measures, $\delta, \gamma_\alpha, \gamma_1 \text{ and } \gamma_2$ met this benchmark:

(i) The measure is location scale invariant and has the formula Y = aX+b for the r.v.s X and Y, where an is equal to 0 and b is equal to 1, and X = Y.

(ii) When the distribution F is symmetric, λ(F) = 0.

(iii) The measure λ is sign equivariant, that is if Y=-X then λ(Y) = −λ(X).

(iv) If F and G are c.d.f's for X and Y as above and F <c G then λ(X) ≤ λ(Y).

According to (oja, 1981) the ordering F c G, which means "F c-precedes G," is a method for comparing the degree of skewness in two univariate distributions, F and G.

3.4.2. Medcouple Skewness Measure

Let $Y_n = \{y_{(1)}, y_{(2)}, \dots y_{(n)}\}$. be a continuous univariate distribution and a continuous sample of n independent observations. F. Think of Q2 as the Yn median. According to Brys et al. (2004), the definition of a medcouple is given by the formula $MC = \underset{y_1 \le Q_2 \le y_{(j)}}{\text{med}} h(y_{(i)}, y_{(j)})$

where for $y_{(1)} \neq y_{(j)}$ A kernel function called h is defined as $h(y_{(i)}, y_{(j)} = \frac{(y_{(j)} - Q_2) - (Q_2 - y_{(i)})}{y_{(j)} - y_{(i)}}$

The kernel function is defined in the following manner for the situation where $y_{(i)} = y_{(J)} = Q_2$ Let the observation indices $t_1 < t_2 < \cdots t_k$ be those that are related to the median Q2, which is $y_{(ti)} = Q_2$for all 1=1, 2,...k. The

$$h(y_{(tj)}, y_{(tj)} = \begin{Bmatrix} -1 \; if \; i + j - 1 < k, \\ 0 \;\; if \; i + j - 1 = k, \\ +1 \; if \; i + j - 1 > k. \end{Bmatrix}$$

It is evident from looking at the denominator in the kernel function's initial statement that both h $h(y_{(i)}, y_{(j)}$ and MC always lie between -1 and 1. The kernel calculates the standard deviation between the $y_{(j)}$ and $y_{(i)}$distances from the median Q_2 If $y_{(J)}$ deviates from the median more than $y_{(i)}$h becomes positive; if $y_{(i)}$ deviates, h becomes negative.

When a value is obtained at $y_{(j)} - Q_2 = Q_2 - y_{(i)}$ a symmetric case is seen. When a single point in $\mathbf{Y_n}$coincides with the sample median $Q_2, h(Q_2, Y_{(j)} = +1$for every $y_{(j)} > Q_2$, indicating that Y_J is infinitely farther from the population median than Q2 is. In other words, for any $h(y_{(i)}, Q_2 = -1$ for all $y_{(i)} < Q_2$. Here, we can see that there are roughly equal numbers of data points greater than the median and smaller than the median, indicating that we have roughly equal numbers of +1 and -1. As a result, these extreme values have no effect on the medcouple MC. When numerous data points fall within the median, it is possible that there are more strictly larger observations than strictly smaller observations, therefore we will include more strictly positive values (+1) than strictly negative values (–1).

We can also see that the second expression of h's number of zeros added equals the number of data values that are tied with the median. The medcouple is drawn toward zero as a result, supporting the idea that having numerous points that are equal to the median reduces a distribution's skewness. Although relatively redundant, the first and third equations in Equation are included to prevent the use of undefined kernels and to make it easier to perform the quick

method that determines the final value of the medcouple. The medcouple belongs to the family of incomplete generalized L-statistics since the first statement of the kernel function h does not apply to all couples $(y_{(i)}, y_{(j)})$from $\boldsymbol{Y_n}$ (ossjer, 1996) However, the medcouple MC can be identified from the medcouple functional form (MCF), defined at any continuous distribution F, if the distribution of $\boldsymbol{Y_n}$ is known. Equation corresponding definition of the MCF is

$$MC_F = \underset{y_{(i)} \le m_F \le y_{(j)}}{\text{med}} h(y_{(i)}, y_{(j)}),$$

where sample points from F, and $m_F = F^{-1}(0.5)$ with $y_{(i)}$ and $y_{(j)}$If we swap out the finite-sample median Q_2 for mF, the kernel in Equation (1.6) is equal to the kernel in Equation. I'll act as the indicator function; after that, define

$$H_F(v) = 4 \int_{m_F}^{+\infty} \int_{-\infty}^{m_F} 1\left[h(y_{(i)}, y_{(j)} < v\right] dF\left(y_{(i)}\right)$$

and obtain a shorter form of Equation (1.6) as

$$MC_F = H_F^{-1}(0.5)$$

We can observe that the domain of $H_F$is [1, 1] and that the conditions $h(y_{(i)}, y_{(j)} \le v, y_{(i)} \le m_F \; and \; y_{(j)} \ge m_F$ are equivalent to $y_{(i)} \le y_{(j)}(v-1) + 2m_F(v+1) \; and \; y_{(j)} \ge m_F$ and hence equation (1.7) can be re-written in simpler form as

$$H_F(v) = 4 \int_{m_F}^{+\infty} F\left[\frac{y_{(i)}(v-1) + 2m_F}{v+1}\right] df(y_{(j)}.$$

It's critical to remember that MC can be thought of as an estimate of MCF.

In addition to its resilience, the medcouple has an intriguing trait that the traditional skewness measure lacks: because it is only based on ranks in its structure, it can be computed at distributions where the moments vanish.

According to von (Zwet, 2012)and (Oja, 1981)definitions of skewness measures, (al. B. e., 2004) demonstrate that the functional medcouple meets these conditions naturally, just as the Bowley coefficient. In other words, the robust measure of skewness medcouple MC has the following characteristics given a random variable X with a continuous distribution F.

(i) The measure is a location and scale invariant that is, for $a \in (0,\infty)$ and $b \in (-\infty,\infty)$, $MC(FaX+b) = MC(FX)$.

(ii) In a symmetric distribution F, $MC(F) = 0$.

(iii) The measure is sign equivariant, that is if $Y=-X$ then $MC(Y) = -MC(X)$.

(iv) If F and G are continuous distributions functions for X and Y, and $F <_c G$ then $MC(X) \leq MC(Y)$.

3.5. Robust Outlier Methods

The term "robust" in the context of boxplots refers to a method or strategy that is resistant to the impact of outliers. Even with extreme values present, a robust boxplot is made to accurately depict the distribution of data. This is accomplished by modifying the way that whiskers and outliers are shown, enabling more accurate visualization and analysis.

The whiskers of a typical boxplot are defined by the interquartile range (IQR), and they reach up to a particular multiple of the IQR from the top and lower quartiles. Outliers are prospective data points that fall outside of these whiskers, and they are plotted separately. However, these outliers can have a significant impact on the whisker length and ultimately the interpretation of the plot in datasets with extreme values or skewed distributions.

Strong boxplot methods take a different tack to deal with outliers, such Tukey's fences. Robust approaches determine modified whisker lengths based on more reliable estimators, such as the median absolute deviation (MAD) or percentile-based procedures, as opposed to utilizing a predetermined multiple of the IQR. These estimators offer a more accurate depiction of the data's spread and are less affected by extreme results.

Depending on the features of the dataset, the whiskers in a robust boxplot may be shorter or longer than those in a regular boxplot. The criteria used to determine whether or not an outlier is present are still used to identify and plot them, but the robust estimates are taken into consideration. This makes sure that the median and center box appropriately represent the data's central tendency, while the whiskers and outliers accurately display the data's real distribution while accounting for the occurrence of extreme values.

Comparative studies and exploratory data analysis both benefit greatly from the use of robust boxplots. They make it possible to visualize data distributions in a more solid and trustworthy

manner, assisting in the detection of potential outliers and the evaluation of the dataset's overall shape and variability. Boxplots that incorporate robustness present the data in a clearer, more accurate manner, allowing researchers and analysts to draw reasonable conclusions and take appropriate action based on a solid understanding of the data's features (Carling, 2002)

Algorism that focuses on:

I. Tukey's method
II. Kimber method
III. Hubert method
IV. Babura method

3.5.1 Tukey's Method (Boxplot)

Tukey's Method, constructing a boxplot, is well known simple graphical tool to display information about continuous univariate data, such as the median, lower quartile, upper quartile, lower extreme and upper extreme of a data set. This method for finding outliers uses the interquartile range to filter out very large or very small numbers. The formulas are:

The terms "lower fence" and "upper fence" are used in statistics and data analysis in the context of outlier detection and reliable data display, particularly in boxplots. The idea of Tukey's fences, which are thresholds used to identify probable outliers in a dataset, is where they got their inspiration. (Hubert, 2009)

The lower fence is the lower threshold below which data points are considered potential outliers. It is calculated as:

$$Lower\ fence = Q_1 - (K \times IQR)$$

Q_1 is the dataset's first quartile (the 25th percentile), IQR is the distance between the first and third quartiles, and k is a multiplier that determines the fence's height. K typically has a value of 1.5, however it can be changed to suit the analysis's particular requirements.

The upper fence serves as a similar maximum limit above which data points are thought to be potential outliers. It is determined by:

$$Upper\ fence = Q_3 + (K \times IQR)$$

where Q_3 is the third quartile (75th percentile) of the dataset.

Step 1: Find the Interquartile Range and Median.

Step 2: Find Q_1 and Q_3. Q_1 can be thought of as a median in the lower half of the data. Q3 can be thought of as a median for the upper half of data. Subtract Q_1 from Q_3.

Step 3: Calculate 1.5 * IQR and subtract from Q_1 to get lower fence

Step 4: Add to Q_3 to get upper fences

Step 5: Add fences to the data to identify outliers.

3.5.2. Kimber Method (Boxplot)

(Kimber, 1990) introduced a lower and upper split interquartile 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, to cope with this, several adjusted boxplot methods have been proposed in case of skewed data. Kimber. suggested the use of the semi-interquartile range (SIQR) rather than IQR, i.e., the fence of the SIQR boxplot is defined as $q_1 - 3 * (q_2 - q_1)$, $q_3 + 3 * (q_3 - q_2)$. The SIQR boxplot has also been applied to the real hourly rain observations from the Wuyigong rain gauge

Lower fence $[Q_1 - 3SIQR_L]$

Upper fence $[Q_3 - 3SIQR_u]$

where $SIQR_L = Q_2 - Q1$, $SIQR_U = Q3 - Q2$,

3.5.3 Hubert method (Boxplot)

A fence rule was introduced by (Hubert and Vandervieren , 2008)to take the degree of data skewness into account. According to the Tukey's rule of thumb, the lower fence estimate overestimates the idle fence position and underestimates the upper fence position when the data is skewed to the right. Therefore, (Hubert and Vandervieren, 2008) redefined the fence cut-off location F as follows using medcouple (MC), a reliable metric of skewness:

$$F = \begin{matrix} [Q_1 - 1.5e^{-4MC}IQR, Q_3 + 1.5e^{3MC}IQR] \text{ if} MC \geq 0, \\ [Q_1 - 1.5e^{-MC}IQR, Q_3 + 1.5e^{4MC}IQR] \text{ if } MC < 0, \end{matrix}$$

where MC is a reliable skewness measurement termed the medcouple that was put forth by (Brys et al, 2004). where we discuss robust skewness metriin detail, we also provide more information about medcouple measure incorporated in the Huberts Method.

3.5.4. Babura Method

(Babura, 2017) incorporate δ in the new definition of boxplot fences by replacing the constant multiple of *IQR* with a relation say $f_l(\delta)$ and $f_u(\delta)$ which are functions of δ into the outliers cut off value. So, the proposed fences will be given by:

$Q_1 - 1.5e^{-4\delta}$ IQR for lower fence with $\delta \geq 0$

$Q_3 + 1.5e^{6\delta}$ IQR for upper fence with $\delta \geq 0$

$Q_1 - 1.5e^{6\delta}$ IQR for lower fence with $\delta < 0$

$Q_3 + 1.5e^{-4\delta}$ IQR for upper fence with $\delta < 0$

3.6. Simulation Methods

3.6.1 The Bootstrap Method

A trusted and reliable statistical methodology is the Bootstrap method. Although bootstrap has been used extensively in signal processing and the social science, there are no reports of its use in electronics design. The original sample is randomly replicated as part of the Bootstrap methodology. Based on a sample x and a desired statistical estimate, S, Where.

$$\mathrm{x} = \mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_\mathrm{i}, \ldots \mathrm{x}_\mathrm{n}$$

To create a bootstrap sample is the first step in the Bootstrap technique, $x_B = x_1, x_2, \ldots x_b, \ldots x_m$

where $\mathrm{x_b} = \mathrm{x}_{1\mathrm{b}}, \mathrm{x}_{2\mathrm{b}}, \ldots \mathrm{x}_{\mathrm{nb}}$

The components of x b are randomly selected from the initial sample, and x b has the same dimension as x. One of the bootstrap samples might be 1.0, 3.1, 3.1, 6.7, for instance, if the set is 1.0, 3.1, 9.2, and 6.7. S_B=$S_1$1, S_2,… S_m. can be used to produce a succession of fresh statistical estimates from the input x_n. The result provides a reliable estimate of S. S*=E_{sb}

This works especially well when the original sample is expensive or challenging to obtain.

3.6.1. Desirable properties of the bootstrap method

For statistical estimating, the Bootstrap has at least two favorable asymptotic features The estimate's asymptotic property comes first. The confidence interval will typically converge as the number of bootstrap samples increases. The second is the distribution's asymptotic property. Bootstrap samples have an asymptotic normal distribution even with a limited number of Monte Carlo simulation runs.

3.6.2. Monte Carlo simulation Method

Monte Carlo methods, or Monte Carlo experiments, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. The underlying concept is to use randomness to solve problems that might be deterministic in principle. They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other approaches. Monte Carlo methods are mainly used in three problem classes: optimization, numerical integration, and generating draws from a probability distribution (Kroese, D. P.; et. Al. , 2014)

Monte Carlo simulation has been extensively used in the design and study of electronics, including testing, defect finding, sensitivity analysis, production discrimination, etc. In most electronics design and analysis, the high computing cost of Monte Carlo simulation acts as a bottleneck. So that we can lower this computing cost, we recommend the Bootstrap approach. By contrasting the mean and variance produced from 100 Monte Carlo fault simulation runs of an operational amplifier with a smaller bootstrap sample acquired from only 10 runs, the accuracy of the approach is shown.

3.7. Hypothetical Distributions

Hypothetical Distribution are distribution deployed for scenarios captured in our simulation processes. We will be considering two fundamental scenarios for symmetric distributions (normal and Uniform) and Asymmetric Distribution (lognormal and chi squared).

3.7.1. Normal Distribution

A continuous probability distribution that is symmetric and bell-shaped is called a normal distribution, commonly referred to as a Gaussian distribution. It is one of the distributions that is used most frequently in probability and statistics. For the purpose of studying and working with data that follows a normal distribution, it is crucial to comprehend the density function and distribution function of this distribution.

Probability Density Function, also known as the density function:

The relative likelihood of detecting various values from the distribution is expressed by the density function of a normal distribution. The following formula defines it.

$$f(x) = \frac{1\, e^{\frac{-(x-x)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

In this equation, μ stands for the distribution's mean (or average), for σ standard deviation (a measure of dispersion or spread), e for the natural logarithm's base (about 2.718), and for the mathematical constant pi (roughly 3.14159). The density function approaches 0 as the deviation from the mean grows and is symmetrical around the mean.

3.7.2. Uniform Distribution

A probability distribution known as the uniform distribution has a constant probability density function (PDF) over a given range. It is frequently applied to simulate scenarios in which all values within that range are equally likely.

Probability Density Function (PDF)

The PDF of a uniform distribution on the interval [a, b] is given by $F(x) = \frac{1}{(b-a)}$, for $a \leq x \leq b$

Where,

x stands for a specific value within the interval [a,b)

a is the lower bound of the interval

b is the upper bound of the interval

cumulative Distribution function CDF

According to the interval [a, b], the cumulative distribution function (CDF) of a uniform distribution is given by:

$F(x) = \frac{(x-a)}{(b-a)}$ for $a \leq$ x $\leq$ b 0, for x $<$ a 1. for x $\geq$ b

For x is the value at which the CDF is evaluated

a is the lower bound of the interval

b is the upper bound of the interval

3.7.3. Log normal Distribution

The logarithm of a randomly distributed quantity yields the lognormal distribution, a type of probability distribution. It is frequently used to represent positively skewed data with non-negative values.

The probability Density Function (PDF)

The probability density function (PDF) of a lognormal distribution with parameters $\mu \, and \, \sigma$ is given by

$$F(x) = \left(\frac{1}{(x * \sigma * \sqrt{2\pi})} exp \frac{(-\ln(x) - \mu)^2}{(2\sigma^2)} \right)$$

For x represents a specific value of the random variable (x>0)

μ is the mean of the logarithm of the distribution

σ Is the standard deviation of the logarithm of the distribution

π Is a mathematical constant (approximately (3.14159)

Exp () denotes the exponential function

ln () is the natural logarithm

3.7.4. Chi-square Distribution

The chi-square distribution is a continuous probability distribution that arises from the sum of squares of independent standard normal random variables. It is widely used in statistical inference and hypothesis testing.

The probability Density Function PDF: of a chi-square distribution with K degree of freedom is given by

$$F(x) = \{\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \quad \text{for } x>0, \text{ for } x\leq 0,$$

For x represents a specific value of the random variable (x > 0)

k is the number of degrees of freedom of the distribution

e denotes the exponential function

Γ() denotes the gamma function

3.8. Limitations of the study

Complexity of Outliers Real-world outliers can vary in terms of their characteristics, such as the number of outliers, their severity, and their relationship to the main data distribution. A simulation study may not capture this complexity accurately.

Sample Size The sample size in a simulation study may not be representative of the variety of sample sizes encountered in real data. The performance of outlier labeling methods can vary with sample size.

CHAPTER FOUR

IMPLEMENTATION AND DATA ANALYSIS

We develop and implement the following algorithms using bootstrap and Monte Carlo's simulation principle to obtain result as presented

4.1. Performance of the robust method using uncontaminated data

4.11 Simulation algorithm 1

Step 1 Generate regular observations X_1+X_2+X_3+... ... . X_n from hypothesized distribution F with press-specified parameters

Step 2 Estimate that lower and upper fences for the generated sample in Step 1. According to Tukey, Kimber, Hubert and Babura methods

Step 3 Count the number of outlying observations according to each fence estimates in Step 2 and record the percentage of outlying observations

Step 4 Repeat Step 1 to Step 3 until a specified number of replications is achieved

4.2. Performance of the robust method using contaminated data

4.2.1. **Simulation algorithm 2**

Step 1 Generate regular observations X_1+X_2+X_3+... ... . X_n from hypothesized distribution F with press-specified parameters

Step 2 Estimate that lower and upper fences for the generated sample in Step 1. According to Tukey, Kimber, Hubert and Babura methods

Step 3 Generate a single value Z from arbitrary chosen distribution H (usually normal) with parameters that makes the value of Z bigger or smaller than maximum or minimum values from F respective. Make large deviation of parameters of H from values of F.

Step 4 Repeat Step 3 to generate $n_o = \alpha\ x\ n$ number of outlying of observation where α is the contamination level. Check if Z satisfy outlier criteria in Step 2

Step 5 A set of n_o observations is selected randomly from observations in Step 1 and replaced with n_o outliers generated in Step 4 to contaminated the data

Step 6 Count the number of outlying observations according to each fence estimates in Step 2 and record the percentage of outlying observations

Step 7 **Repeat 1 to 6 until specified numbers of replication is achieved.**

4.3 Simulation Study

In a simulation analysis, we compare the outlier labeling approaches to the (Tukey, 1977) (Kimber, 1990) Hubert, and Babura (2017) methods for both uncontaminated and contaminated data. Below is a discussion of the findings.

4.4 Simulation Scheme

The simulation scheme describes how the simulated datasets were generated.

We use four Algorithms methods that we want to use to create a comparative analysis of how they handled outliers in the data. These four algorithms are Tukey method, Kimber Method, Hubert Method and Babura Method. These methods we make comparison that allows detect which method is best performance revised methods according to the sample sizes and sample sizes categorized into three categories small sample size, medium sample size and large sample sizes. Also, use four distributions that are separated into two groups symmetric and asymmetric. For symmetric distribution we used Normal distribution and Uniform distribution with different sample size and for asymmetric distribution we focus on Lognormal and chi-square distribution with different degree of freedom. Simulating data, data are generate using codes that are run in R programming language and then making replications.

To generate random numbers from a normal distribution in R, we can use the rnorm() function. The rnorm() function generates random numbers from a normal distribution with specified mean and standard deviation. As shown bellow

rnorm(n, mean = 0, sd = 1)

N: The number of random values to generate.

Mean: The mean of the normal distribution. The default is 0.

Sd: The standard deviation of the normal distribution. The default is 1.

To generate random numbers from a uniform distribution in R, we can use the runif() function. The runif() function generates random numbers from a uniform distribution between specified minimum and maximum values. Here's the basic syntax.

random_numbers <- runif(n, min = 1, max = 10),

N: The number of random values to generate.

Min: The minimum value for the uniform distribution. The default is 0.

Xax: The maximum value for the uniform distribution. The default is 1.

To generate random numbers from a log-normal distribution in R, we can use the rlnorm() function. The rlnorm() function generates random numbers from a log-normal distribution with specified mean and standard deviation of the logarithm of the distribution. As we see here.

rlnorm(n, meanlog = 0, sdlog = 1)

N The number of random values to generate.

Meanlog, The mean of the logarithm of the distribution. The default is 0.

Sdlog, The standard deviation of the logarithm of the distribution. The default is 1.

To generate random numbers from a chi-squared distribution in R, we can use the rchisq() function. The rchisq() function generates random numbers from a chi-squared distribution with a specified degrees of freedom. Here's the basic syntax.

rchisq(n, df)

N: The number of random values to generate.

df: The degrees of freedom parameter, which determines the shape of the chi-squared distribution.

For each algorithms have formulas that we can calculate the lower and upper fence of the data, then we combine the formulas for algorithms and codes that we going to generated the distributions. Data will be classified into uncontaminated dataset and contaminated dataset.

For Uncontaminated dataset we will display tables while contaminated dataset will display by graphs, these graphs constructed by R programming using codes.

4.5 Performance of the Boxplot Methods for Uncontaminated dataset

The analysis of uncontaminated data (data without any outliers) uses four families of distributions. The Normal, Uniform, Chi-square with various degrees of freedom (2,5,10), and the Log Normal Distribution are some of these. Each distribution yields datasets with n = 10, 20, 30, 50, 100, 150, 200, 500, and 1000 are generated from each distribution and 5000 simulations

are run. The mean proportions of observations falsely (incorrectly) flagged as outliers are calculated. A subset of the results is displayed in table

In the table we classified data in three different categories with small size, medium size and large sample size as shown below in these three tables.

Table 4.1: Simulation Result: Comparison of Outside rates for uncontaminated Normal, Uniform, Chi-squared and lognormal distributions with small sample size.

Sample size (n)	**Method**	**Distributions**					
		Symmetric		Asymmetric			
		Normal	Uniform	Chiq(n,2)	Chiq(n,5)	Chiq(n,10)	Log normal
10	Tukey	**0.02782**	**0.0099**	0.0494	0.0464	0.0494	**0.0280**
	Kimber	0.0326	0.0162	**0.0183**	**0.0172**	**0.0194**	0.0339
	Hubert	0.0555	0.0319	0.0540	0.0431	0.0380	0.0574
	Babura	0.11354	0.0921	0.0619	0.0387	0.0288	0.1151
20	Tukey	**0.01645**	**0.00126**	0.0358	0.0306	0.03077	**0.0169**
	Kimber	0.0223	0.0045	**0.0209**	**0.0148**	**0.0141**	0.0227
	Hubert	0.0476	0.0188	0.0553	0.0462	0.0421	0.0476
	Babura	0.0899	0.0626	0.0818	0.0552	0.0432	0.0896
30	Tukey	**0.01523**	**0.00031**	0.0319	0.0239	0.0233	**0.0146**
	Kimber	0.0172	0.0018	**0.0242**	**0.0162**	**0.0137**	0.01714
	Hubert	0.0421	0.0130	0.0546	0.0479	0.0410	0.0422
	Babura	0.0769	0.0465	0.0897	0.0657	0.0496	0.0766

Table 4.1 shows simulation for uncontaminated Normal, Uniform, Chi-square with (2,5,10) degree of freedom and lognormal distributions with small sample size outside rate: As the results of the simulation for uncontaminated samples indicate that the small value in outside rates shows the best performance method, it is classified into two major types: symmetric and asymmetric. For symmetric distributions, we focus on normal and uniform distributions, while for asymmetric

distributions, we focus on chi-square and lognormal distributions with different samples in small sample sizes, medium sample sizes, and large sample sizes.

For the case of small sizes (n = 10, 20, and 30), symmetric distribution in normal and uniform distribution Tukey is the best performance method in this case because it has the smallest values, while asymmetric distribution in lognormal Tukey is the best performance method. In both symmetric and asymmetric cases in small sizes, Tukey has the best performance in normal, uniform, and lognormal distributions, while for other asymmetric cases in chi-square with 2, 5, and 10 degrees of freedom, Kimber is the best performance method since it has the lowest value of outside rates.

Table 4.2: Simulation Result: Comparison of Outside rates for uncontaminated Normal, Uniform, Chi-squared and lognormal distributions with medium sample size.

Sample size (n)	**Method**	**Distributions**					
		Symmetric		Asymmetric			
		Normal	Uniform	Chiq(n,2)	Chiq(n,5)	Chiq(n,10)	Log normal
50	Tukey	**0.0115**	**0.0000**	0.0488	0.0482	0.0487	**0.0116**
	Kimber	0.0125	0.0002	0.0213	0.0250	0.0272	0.0123
	Hubert	0.0328	0.0048	0.0272	0.0153	0.0103	0.0324
	Babura	0.0590	0.0271	**0.0183**	**0.0086**	**0.0046**	0.0580
100	Tukey	0.0091	**0.0000**	0.0299	0.0285	0.0284	0.0094
	Kimber	0.0085	**0.0000**	**0.0137**	**0.0144**	0.0153	**0.0091**
	Hubert	**0.0226**	0.0008	0.0325	0.0216	0.0165	0.0230
	Babura	0.0398	0.0103	0.0284	0.0155	**0.0105**	0.0404
150	Tukey	0.00867	**0.0000**	0.0218	0.0202	0.0203	0.0086
	Kimber	**0.0076**	0.0000	**0.0109**	**0.0100**	**0.0111**	**0.0076**
	Hubert	0.0186	0.0001	0.0330	0.0226	0.0181	0.0186
	Babura	0.0306	0.0039	0.0349	0.0192	0.0139	0.0304

Table 4.2 shows simulation for uncontaminated Normal, Uniform, Chi-square with (2,5,10) degree of freedom and lognormal distributions with medium sample size outside rate. This second table shows simulation for uncontaminated outside rates in medium sample sizes, and comparing the four algorithms in both symmetric and asymmetric distributions that focus on the smallest outside rates indicates the best performance method for samples in this case of n = 50, 100, and 150. For the case n = 50 in normal and uniform, symmetric distributions, the Tukey method gets the smallest value, and it is the best performance method according to the sample size of n = 50.

Also, in case n = 100, normal distribution Hubert is the best performance method, while n = 150, Kimber is the best performance method. For case samples n = 100 and 150 in uniform distribution, both Tukey and Kimber get the same values, which indicates that two methods are the best methods according to the samples. In cases of symmetric distribution, Tukey, Kimber, and Hubert are the best methods, while the Babura method is not good in these cases, according to the sample sizes.

In asymmetric distributions for lognormal distributions, when sample size n = 50, Tukey is the best method, while when n = 100 and 150, Kimber is the best method. In chi-square with 2, 5, and 10 degrees of freedom, when n = 50, Babura has the lowest values, showing it is the best performance method according to the sample size.

While Kimber is the best method at n = 100 and 150 with 2, 5, and 10 degrees of freedom, when n = 100 with 10 degrees of freedom, Babura is the best method.

The Kimber method is good to use in asymmetric distribution. According to the samples and some degree of freedom

Table 4.3: Simulation Result: Comparison of Outside rates for uncontaminated Normal, Uniform, Chi-squared and lognormal distributions with large sample size.

Sample size (n)	Method	Distributions					
		Symmetric		Asymmetric			
		Normal	Uniform	Chiq(n,2)	Chiq(n,5)	Chiq(n,10)	Log normal
200	Tukey	0.0081	**0.0000**	0.0488	0.0482	0.0480	**0.0081**
	Kimber	**0.0071**	**0.0000**	0.0282	0.0299	0.0304	0.0070
	Hubert	0.01603	**0.0000**	0.0081	0.0039	0.0031	0.0161
	Babura	0.0257	0.0020	**0.0029**	**0.0005**	**0.0002**	0.02585
500	Tukey	0.0075	**0.0000**	0.0285	0.0283	0.0281	0.0076
	Kimber	**0.0065**	**0.0000**	0.0164	0.0184	0.0189	**0.0066**
	Hubert	0.0111	**0.0000**	0.0138	0.0082	0.0059	0.0113
	Babura	0.0150	**0.0000**	**0.0085**	**0.0033**	**0.0015**	0.0152
1000	Tukey	0.0073	**0.0000**	0.0197	0.0196	0.0195	0.0072
	Kimber	**0.0066**	**0.0000**	0.0112	0.0129	0.0136	**0.0066**
	Hubert	0.0093	**0.0000**	0.0154	0.0103	0.0084	0.0092
	Babura	0.0108	**0.0000**	**0.0108**	**0.0058**	**0.0038**	0.0109

Table 4.3 simulation for uncontaminated Normal, Uniform, Chi-square with (2,5,10) degree of freedom and lognormal distributions with large sample size outside rate: the above table shows that for uncontaminated data in large sample sizes with n = 200, 500, and 1000 and both symmetric and asymmetric distributions in normal uniform and lognormal in all samples, Kimber is the best performance method according to the samples, while the other asymmetric distribution, the chi-square Babura method, is the best method in all samples and all degrees of freedom. Also, with this large sample size, both Tukey and Hubert are not good, according to the sample size.

4.5 Case of contaminated data

For the contaminated data, we use the sample size n= 50,100,150,200,500 and 1000. We use graphical display for the percentage outlier value on y axis plotted against sample sizes on x axis. The plots that are contaminated with symmetric and asymmetric distributions with normal, uniform, lognormal, and chi-square with different degrees of freedom (2, 5, and 10), with replications of 5000 times each, and classified with a 6% contamination level and a 10% contamination level. And the best performance of the revised method is the one closest to the contaminated line on the y-axis.

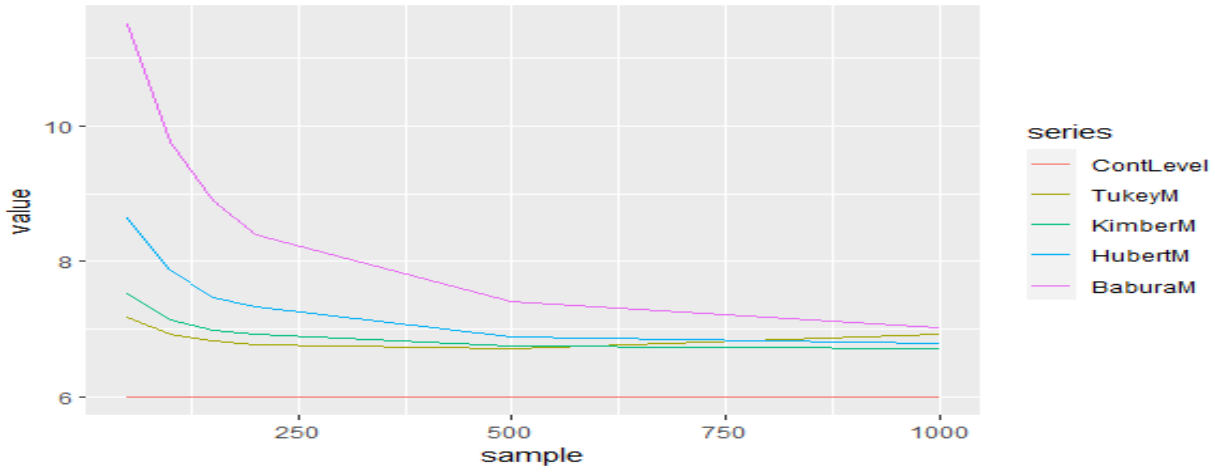


Figure 4.1: Simulation for contaminated normal distribution with Tukey, Kimber, Hubert and Babura methods at 6% contamination level.

Figure 4.1 shows simulations for contaminated normal distributions with Tukey, Kimber, Hubert, and Babura methods at a 6% contamination level. The sample size on the x-axis was plotted against the percentage value of outliers on the y axis. As the results show, all the methods are above the 6% contamination level; Tukey and Kimber are closer to the contamination level, when comparing to the ither methods, and its method with the best performance that is revised, but when the sample size is above 500, Hubert is closer to the 6% contamination level and meets the Tukey and Kimber methods, but as the sample size increases, the Babura method falls below the contamination level. In this case, with a symmetrical normal distribution simulation of a 6% contamination level, two methods have the best performance in all sample sizes, and other methods perform equally well as the sample size increases.

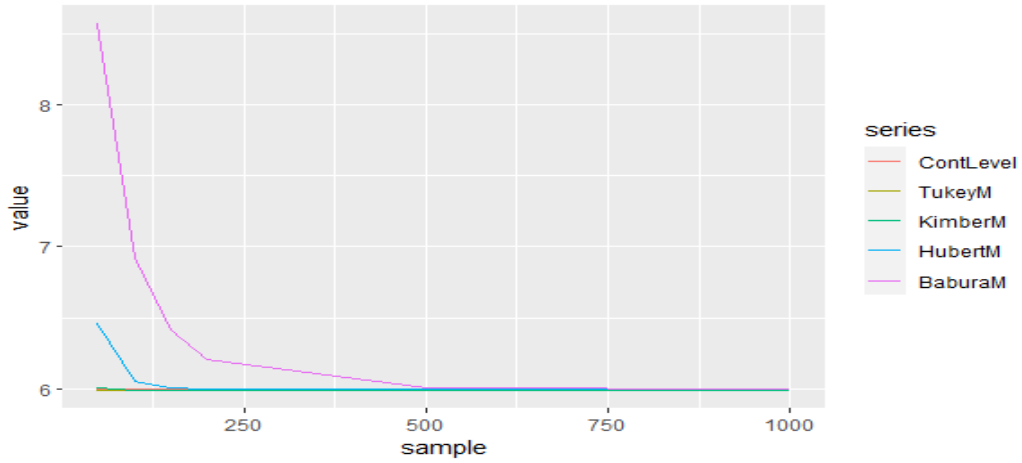


Figure 4.2: Simulation for contaminated data uniform distribution with Tukey, Kimber, Hubert and Babura methods at a 6% contamination level.

Figure 4.2 shows a simulation for contaminated data uniform distribution with Tukey, Kimber, Hubert, and Babura methods at a 6% contamination level. Also, this is also a symmetric distribution; the three methods Tukey, Kimber, and Hubert are met in the 6% contaminated line, while Babura is also met when sample size is 500 and they are the same for all other samples. In this case, Tukey, Kimber, and Hubert are the best performing revised methods according to the all-sample size, and Babura is good when sample size increases.

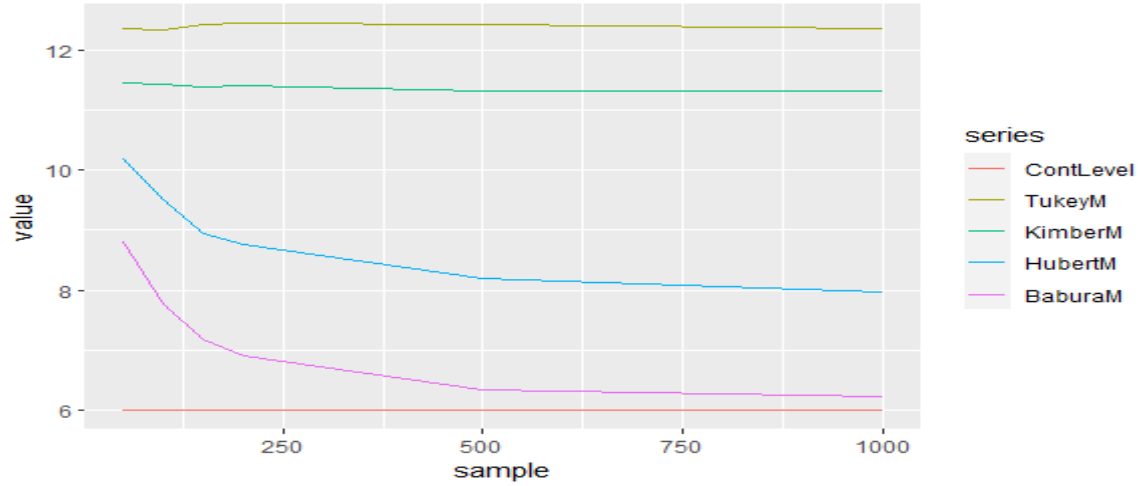


Figure 4.3: Simulation for contaminated data log normal distribution with Tukey, Kimber, Hubert and Babura methods at a 6% contamination level.

Figure 4.3 shows a simulation for contaminated data log normal distribution with Tukey, Kimber, Hubert, and Babura methods at a 6% contamination level. In an asymmetric distribution, all methods are above and far above the line, but the Babura method is close to the line when compared to the other methods. Babura is the best performing revised method based on sample size.

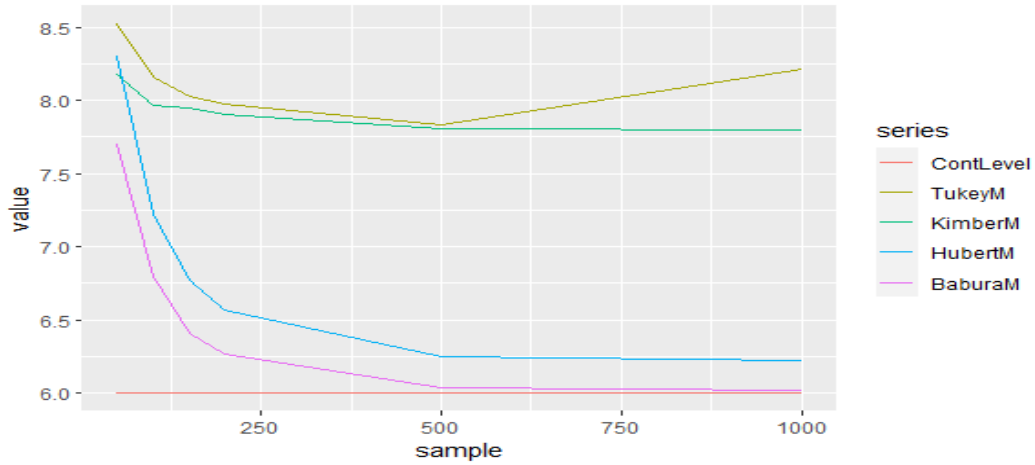


Figure 4.4: Simulation for contaminated data chi-square with 2 degree of freedom using Tukey, Kimber, Hubert and Babura methods at a 6% contamination level.

Figure 4.4 shows a simulation for contaminated data with a chi-square with a 2-degree of freedom distribution. Tukey, Kimber, Hubert, and Babura methods at a 6% contamination level Ch-square is also an asymmetric distribution; this figure shows that the Babura method is the best performing revised method for the beginning The Babura line is too far for the contamination level, and Hubert is the next method that is closer to the contamination level. Other methods are too far off the mark.

In cases of asymmetric distribution and a 6% contamination level with 2 degrees of freedom, the Babura method provides the best performance.

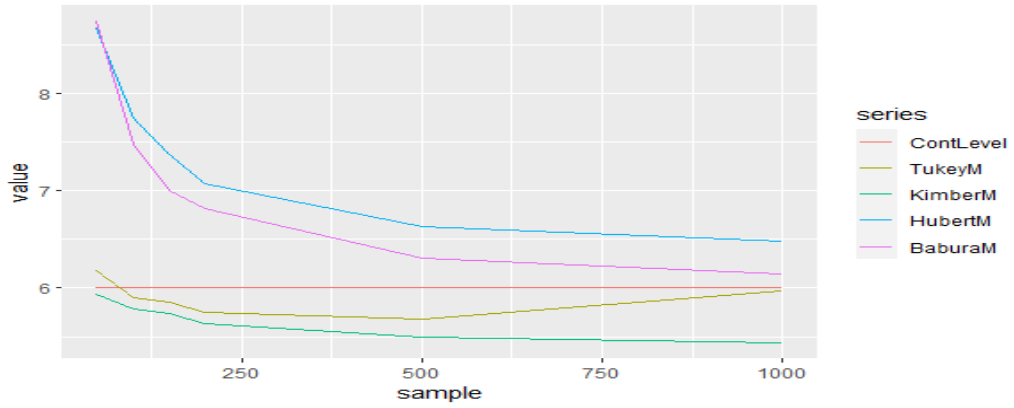


Figure 4.5: Simulation for contaminated data chi-square with 5 degree of freedom using Tukey, Kimber, Hubert and Babura methods at a 6% contamination level.

Figure 4.5 shows us a simulation for contaminated data chi-square distribution with 5 degrees of freedom. Tukey, Kimber, Hubert, and Babura methods at a 6% contamination level This asymmetric distribution shows that the four methods, Hubert and Babura, are above contamination level, while Tukey and Kimber are below. In this case, all methods are shown to have the best performance of the revised methods, but Babura and Tukey are close to the line. These two methods are the best permeance to use chi-square with degrees of freedom of 5 and 6% contamination level.

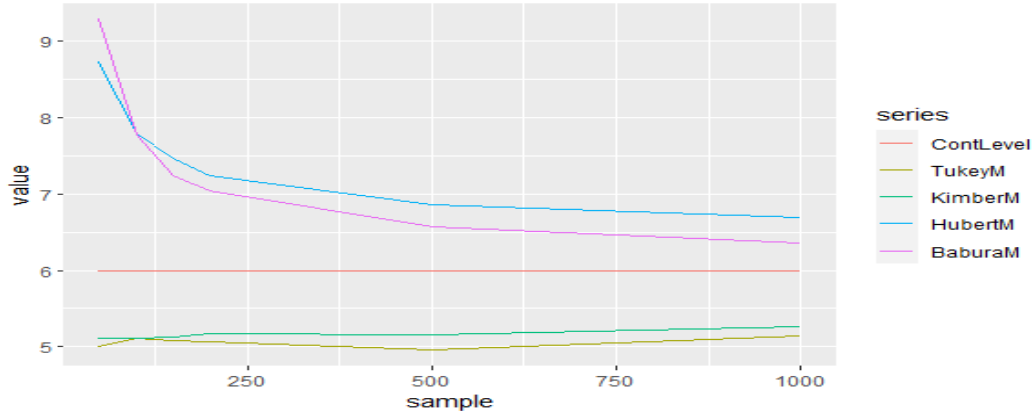


Figure 4.6: Simulation for contaminated data chi-square with 10 degree of freedom using Tukey, Kimber, Hubert and Babura methods at a 6% contamination level.

Figure 4.6 shows a simulation for contaminated data with a chi-squared with 10 degrees of freedom. Tukey, Kimber, Hubert, and Babura methods at a 6% contamination level in this

asymmetric distribution, it shows that four methods, Hubert and Babura, are above contamination level, while Tukey and Kimber are below. In this case, all methods are shown to have the best performance of the revised methods, but Babura and Tukey are close to the line. These two methods are the best to use with a chi-square with a degree of freedom of 10 and a 6% contamination level. Same as the chi-square with a degree of freedom of 5.

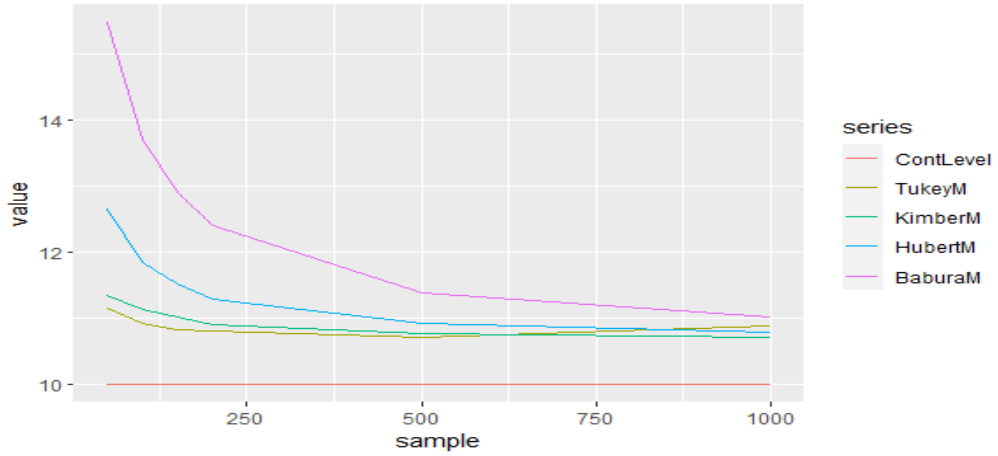


Figure 4.7: Simulation for contaminated normal distribution with Tukey, Kimber, Hubert and Babura methods at 10% contamination level.

Figure 4.7 shows simulations for contaminated normal distributions using Tukey, Kimber, Hubert, and Babura methods at a 10% contamination level. Normal is one of the symmetric distributions, as the results show for the contaminated normal distribution, with 6% contaminated. Tukey and Kimber are the best-performing revised methods, while other methods change when sample sizes increase.

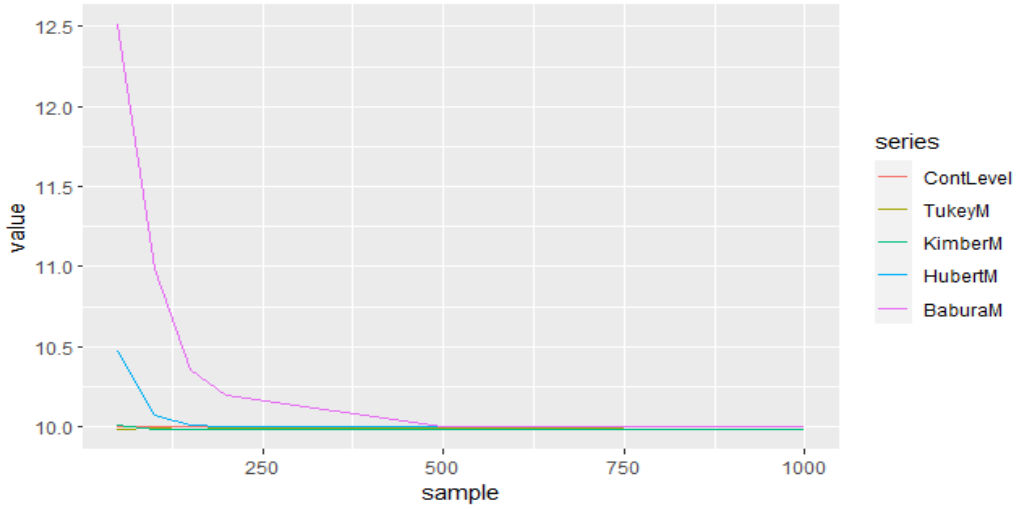


Figure 4.8: Simulation for contaminated uniform distribution with Tukey, Kimber, Hubert and Babura methods at 10% contamination level.

Figure 4.8 shows us simulations for contaminated uniform distribution using Tukey, Kimber, Hubert, and Babura methods at a 10% contamination level. In the case that the symmetric distribution shows that all the methods except the Babura method meet the contamination level, this means that Tukey, Kimber, and Hubert are the best performing revised methods.

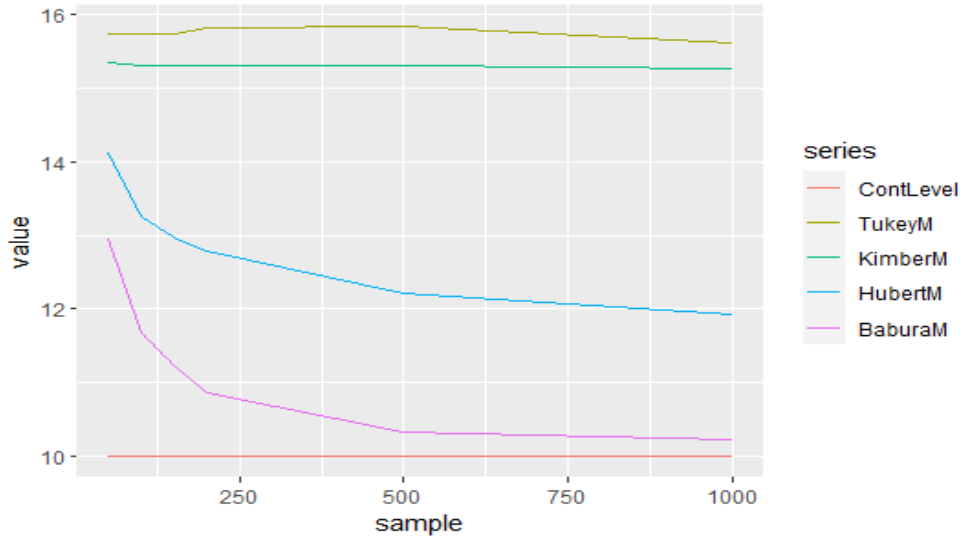


Figure 4.9: Simulation for contaminated log normal distribution with Tukey, Kimber, Hubert and Babura methods at 10% contamination level.

Figure 4.9 shows simulation for contaminated lognormal distribution. Tukey, Kimber, Hubert, and Babura methods at a 10% contamination level for asymmetric distributions, Babura has the best performance, as shown in the plots. This is not much different from a lognormal distribution with a 6% contamination level.

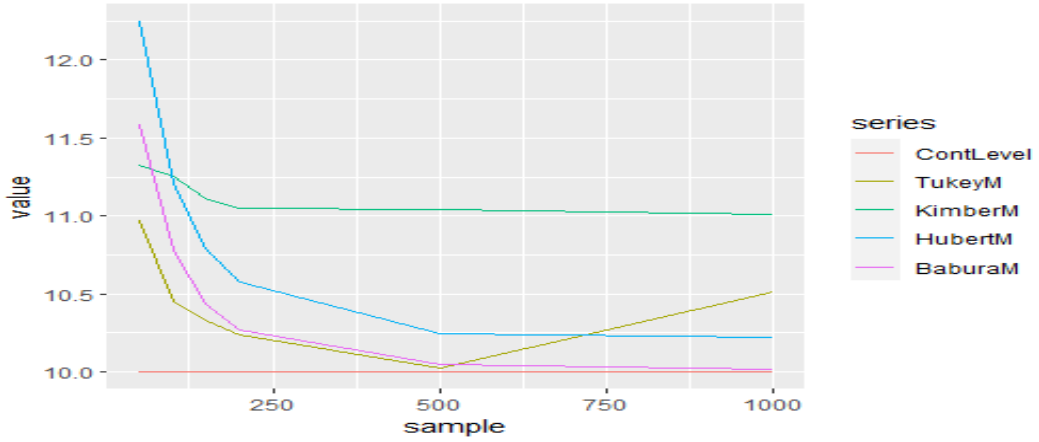


Figure 4.10: Simulation for contaminated data chi-square with 2 degree of freedom using Tukey, Kimber, Hubert and Babura methods at a 10% contamination level.

Figure 4.10 shows a simulation for a contaminated data chi-square distribution with 2 degrees of freedom for the Tukey, Kimber, Hubert, and Babura methods at a 10% contamination level. The resultant asymmetric distribution shows that the Tukey and Babura methods are the best performance methods since they are close to the line, while other methods are too far.

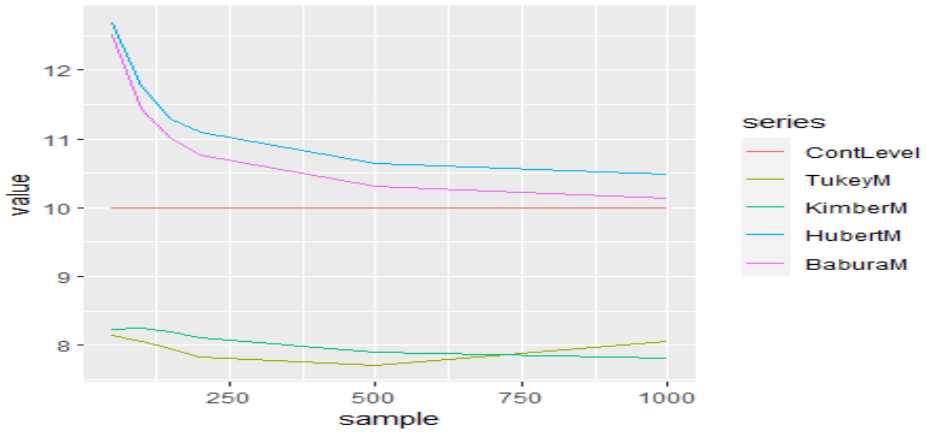


Figure 4.11: Simulation for contaminated data chi-square with 5 degree of freedom using Tukey, Kimber, Hubert and Babura methods at a 10% contamination level.

Figure 4.11 is simulation result for contaminated data from chi-square distribution with 5 degrees of freedom using the Tukey, Kimber, Hubert, and Babura methods at a 10% contamination level the results show that two methods are above the contaminated line while the other two methods are below the contaminated line, and it shows that the four methods are the best performing revised methods.

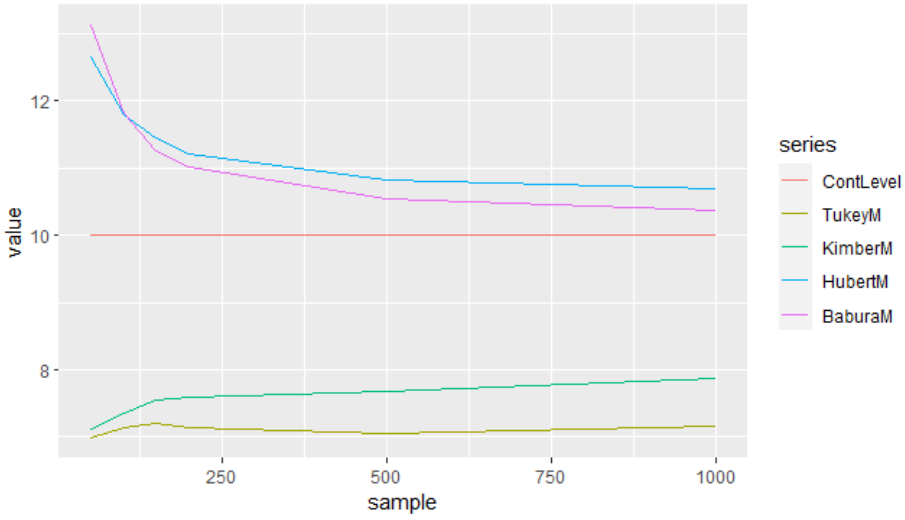


Figure 4.12 Simulation for contaminated data chi-square with 10 degree of freedom using Tukey, Kimber, Hubert and Babura methods at a 10% contamination level.

Figure 4.12 is a simulation of a contaminated data chi-square distribution with 5 degrees of freedom using the Tukey, Kimber, Hubert, and Babura methods at a 10% contamination level. The results show that two methods are above the contaminated line while the other two methods are below the contaminated line, and it shows that the four methods deviate away from the contamination level. The best-performing Babura method is followed by Hubert method considering the magnitude of deviation from the contamination level especially for sample sizes above 100. This demonstrates the effect of in cooperating skewness in fence definition by both Babura and Hubert methods.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1 Discussion.

Since Tukey's (1977) Exploratory Data Analysis, which was published 46 years ago, the boxplot has been a commonplace tool for data visualization across many academic disciplines. Even if there are more advanced graphical techniques available now, the boxplot is still useful because of its straightforwardness, readability, and relative efficacy. The conventional boxplot, however, is not without its drawbacks; in particular, it may not be able to accurately flag probable outliers in data from skewed distributions. Since this problem has been studied for so long, the conventional boxplot has undergone a number of changes. While many of these intricate suggestions have clear applications when working with particular kinds and amounts of data because Tukey's initial aims of simplicity and easy of interpretation in boxplots and EDA in general appear to have increasingly been lost in attempts to modify the boxplot for public use. The interesting to ascertain claims of improvement of the Tukey method through a simulation study as employed in this thesis.

The bootstrap Simulation method employed in this research has present an interesting result with power of computational approach. The simulation study has assessed performance of some outlier labelling methods namely: Turkey (1977), Kimber (1985), Hubert (2002) and Babura (2017). Based on the fact that the performance of any outlier method is specifically identified based on the concepts of masking and swamping effects. Consequently, we develop the two simulation algorithms for uncontaminated and contaminated datasets for assessment of the masking and swamping effects respectively. The algorithms were deployed over symmetric dataset (normal & uniform distribution) and asymmetric data (chi-square and log normal distributions).

5.2 Summary of Findings

The performance analysis as presented in chapter 4 is summarized as follows; For symmetric distribution at a 6% and 10% contamination level with small sample sizes categorize as medium and large sample sizes generated from normal distribution, the Tukey and Kimber methods are

the best-performing methods. As a result, they are so closely related to the contamination level that two methods are best when compared with other methods. Also, for uniform distribution all the methods arrived at the same performance level except Babura methods which deviate when the sample size reaches n = 500. For asymmetric distribution case at 6% and 10% contamination level for both medium sample, and large sample sizes. with lognormal data, Babura method outperform other methods. While in the chi-square distribution case, Babura, and Tukey outperform other methods by demonstrating similar performance level. Conclusively Tukey and Kimber methods are the best-performing revised methods for symmetric distribution. Babura is the best-performing revised method for asymmetric distribution then followed by Hubert methods.

5.3 Recommendations

The boxplot outlier labelling methods are meant for exploratory data analysis usually deployed to ascertain diagnostic insight into a dataset undergoing statistical test or analysis. This study isolates the performance of individual outlier labelling methods under review to recommend the following insight,

- The Tukey method is specifically design for symmetric data with Gaussian assumptions.
- As such it is best performing in real life datasets such as records of specie height, weight etc.
- The Kimber method has similarly assumptions with Tukey but reflect some level of asymmetry using semi-interquartile range. The method interestingly performs better in the uniform distribution sense. The real-life scenario when the Kimber method is highly recommended are records of persons birthdays, points of a dartboard, numbered raffle tickets etc.
- Hubert and Babura Methods clams to capture symmetric scenario but our simulation study indicate their limitations, and are not recommended for such case. The two methods demonstrate an edge advantage in asymmetric case with Babura method being the best.

  The recommended real-life data to deploy both Hubert and Babura methods are record of household income data, maximum rainfall records, etc.
- This study can be extended to consider more family of symmetric and asymmetric distributions.

REFERENCES

Abraham, B., & Box, G. E. (1979). Bayesian analysis of some outlier problems in time series. Biometrika, 66(2), 229-236.

Acuna, E., & Rodriguez, C. (2004). A meta-analysis study of outlier detection methods in classification. Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, 1, 25.

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. In Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data (pp. 37-46).

Ampanthong, P., & Suwattee, P. (2009, March). A comparative study of outlier detection procedures in multiple linear regression. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1).

Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002, June). Models and issues in data stream systems. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 1-16).

BABURA, B. I. (2017). Modified Boxplot and Stairboxplot for Generalized Extreme Value Distribution (Doctoral dissertation, Phd dissertation, Institute for Mathematical Research, Universiti Putra Malaysia).

BABURA, B. I. (2017). *Modified Boxplot and Stairboxplot for Generalized Extreme Value Distribution* (Doctoral dissertation, Phd dissertation, Institute for Mathematical Research, Universiti Putra Malaysia)

Bain, L. J., & Engelhardt, M. (1992). Introduction to probability and mathematical statistics (Vol. Belmont, CA: Duxbury Press.

Barnett, G., Kohn, R., & Sheather, S. (1996). Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. Journal of Econometrics, 74(2), 237-254.

Barnett, V., & Lewis, T. (1994). Outliers in statistical data (Vol. 3, No. 1). New York: Wiley.

Barnett, V., & Lewis, T. (1994). Outliers in statistical data (Vol. 3, No. 1). New York: Wiley.

Basu, S., & Meckesheimer, M. (2007). Automatic outlier detection for time series: an application to sensor data. Knowledge and Information Systems, 11, 137-154.

Bendre, S. M., & Kale, B. K. (1987). Masking effect on tests for outliers in normal samples. Biometrika, 74(4), 891-896.

Ben-Gal, I. (2005). Outlier detection. In Data mining and knowledge discovery handbook (pp. 131-146). Springer, Boston, MA.

Bickel, Peter J., Friedrich Götze, and Willem R. van Zwet. *Resampling fewer than n observations: gains, losses, and remedies for losses*. Springer New York, 2012.

Bonar, Scott A., Wayne A. Hubert, and David W. Willis. "Standard methods for sampling North American freshwater fishes." (2009).

Box, G. E., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical association, 70(349), 70-79.

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

Brown, J. (1957). A. C. The Lognormal Distribution.

Brys, G., Hubert, M., & Rousseeuw, P. J. (2005). A robustification of independent component analysis. Journal of Chemometrics: A Journal of the Chemometrics Society, 19(5- 7), 364-375

Brys, G., Hubert, M., & Struyf, A. (2003). A comparison of some new measures of skewness. In Developments in robust statistics (pp. 98-113). Physica, Heidelberg.

Brys, G., Hubert, M., & Struyf, A. (2004). A robust measure of skewness. Journal of Computational and Graphical Statistics, 13(4), 996-1017.

Chen, C., & Liu, L. M. (1993). Forecasting time series with outliers. Journal of forecasting, 12(1), 13-35.

Chen, C., & Liu, L. M. (1993). Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association, 88(421), 284-297.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91(434), 883-904.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91(434), 883-904.

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap Methods and Their Application. Cambridge University Press.

de Aquino, A. L., Figueiredo, C. M. S., Nakamura, E. F., Buriol, L. S., Loureiro, A. A. F., Fernandes, A. O., & Coelho Jr, C. J. N. (2007, June). A sampling data stream algorithm for wireless sensor networks. In 2007 IEEE International Conference on Communications (pp. 3207-3212). IEEE.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method (4th ed.). John Wiley & Sons.

Dubes, R., & Jain, A. K. (1980). Clustering methodologies in exploratory data analysis. Advances in computers, 19, 113-228.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1-26.

Felix Famoye1, Eno Akarawak2, Matthew Ekum (2018). Weibull-Normal Distribution and its Applications, Journal of Statistical Theory and Applications DOI: 10.2991/jsta.2018.17.4.12; ISSN 1538-7887 https://www.atlantis-press.com/journals/jsta

Fill, J. A. (1997, May). An interruptible algorithm for perfect sampling via Markov chains. In Proceedings of the twenty-ninth annual ACM symposium on Theory of computing (pp. 688-695).

Fill, J. A., Machida, M., Murdoch, D. J., & Rosenthal, J. S. (2000). Extension of Fill's perfect rejection sampling algorithm to general chains. Random Structures & Algorithms, 17(3- 4), 290-316.

Fowler Jr, F. J. (2014). Survey Research Methods (5th ed.). SAGE Publications.

Fox, A. J. (1972). Outliers in time series. Journal of the Royal Statistical Society: Series B (Methodological), 34(3), 350-363.

Frigge, Michael, David C. Hoaglin, and Boris Iglewicz. "Some implementations of the boxplot." *The American Statistician* 43.1 (1989): 50-54.

Ghosh, A. (2006). Mixed Methods Research for TESOL. TESOL Quarterly, 40(3), 601-614.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics), 41(2), 337-348.

Golab, L., & Özsu, M. T. (2003). Issues in data stream management. ACM Sigmod Record, 32(2), 5-14.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey methodology (Vol. 561). John Wiley & Sons.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11(1), 1-21.

Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). Multivariate Data Analysis (7th ed.). Pearson.

Hekimoglu, S., Erenoglu, R. C., & Kalina, J. (2009). Outlier detection by means of robust regression estimators for use in engineering science. Journal of Zhejiang university-science A, 10(6), 909-921.

Hodge, V. J., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22(2), 85-126.

Hubert, C., & Voordouw, G. (2007). Oil field souring control by nitrate-reducing Sulfurospirillum spp. that outcompete sulfate-reducing bacteria for organic electron donors. Applied and environmental microbiology, 73(8), 2644-2652.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. Computational statistics & data analysis, 52(12), 5186-5201.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, *52*(12), 5186-5201.

Hyndman, Rob J., and Yanan Fan. "Sample quantiles in statistical packages." *The American Statistician* 50.4 (1996): 361-365.

Iglewicz, B., & Hoaglin, D. C. (1993). How to detect and handle outliers (Vol. 16). Asq Press.

Ismail, S. O. (2008). Accommodation of outliers in time series data: An alternative method. Asian Journal of Mathematics and Statistics, 1(1), 24-33.

K. Carling, "Resistant outlier rules and the non-Gaussian case," *Computational Statistics & Data Analysis*, vol. 33, no. 3, pp. 249–258, 2000.

Kaya, A. (2010). Statistical modelling for outlier factors. Ozean Journal of Applied Sciences, 3(1), 185-194.

Kimber, A. C. (1990). Exploratory data analysis for possibly censored data from skewed distributions. Journal of the Royal Statistical Society: Series C (Applied Statistics), 39(1), 21-30.

Liu, B. (2006). Classifying Data Streams Using a Concept Drifting Indicator (Doctoral dissertation, University of Oklahoma).

Macneil, Ian R. "Contracts: Adjustment of long-term economic relations under classical, neoclassical, and relational contract law." *Nw. UL Rev.* 72 (1977): 854.

McCulloch, R. E., & Tsay, R. S. (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. Journal of Time Series Analysis, 15(2), 235-250.

Nkechi, E. M., Chekwube, B. D., Paul, O. C., & Chizoba, K. L. (2022). A Monte Carlo Simulation Comparison of Methods of Detecting Outliers in Time Series Data.

Nkechi, E. M., Chekwube, B. D., Paul, O. C., & Chizoba, K. L. (2022). A Monte Carlo Simulation Comparison of Methods of Detecting Outliers in Time Series Data.

Nkechi, E. M., Chekwube, B. D., Paul, O. C., & Chizoba, K. L. (2022). A Monte Carlo Simulation Comparison of Methods of Detecting Outliers in Time Series Data.

Oja, Hannu. "On location, scale, skewness and kurtosis of univariate distributions." *Scandinavian Journal of statistics* (1981): 154-168.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research, and Evaluation, 9(1), 6.

Penny, K. I., & Jolliffe, I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. Journal of the Royal Statistical Society: Series D (The Statistician), 50(3), 295-307.

Propp, J. G., & Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures & Algorithms, 9(1- 2), 223-252.

Rousseeuw, P. J., & Ruts, I. (1998). Constructing the bivariate Tukey median. Statistica Sinica, 827-839.

Sadik, M., & Gruenwald, L. (2010, August). DBOD-DS: Distance based outlier detection for data streams. In International Conference on Database and Expert Systems Applications (pp. 122-136). Springer, Berlin, Heidelberg.

Satman, M. H. (2013). A new algorithm for detecting outliers in linear regression. International Journal of statistics and Probability, 2(3), 101.

Siegel, A. Statistics and data analysis: An Introduction, Wiley, New York, 1988

Smith, A. F. M., & West, M. (1983). Monitoring renal transplants: an application of the multiprocess Kalman filter. Biometrics, 867-878.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. Journal of the American Statistical Association, 81(393), 132-141.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. Journal of forecasting, 7(1), 1-20.

Tukey, J. W. (1977). Exploratory data analysis (Vol. 2, pp. 131-160).

Zakaria, A., Howard, N. K., & Nkansah, B. K. (2014). On the detection of influential outliers in linear regression analysis.

Zhang, J., & Wang, C. (2003). Outlier detection techniques for credit card fraud. Information Assurance Workshop, 2003. Proceedings from the Fourth Annual IEEE SMC (pp. 322-329).