Chapter 1

Introduction

The term "outlier" can generally be defined as an observation that is significantly different from the other values in a data set.
The outliers may be instances of error or indicate events. The task of outlier detection aims at identifying such outliers in order to improve the analysis of data and further discover interesting and useful knowledge about unusual events within numerous applications domain.

Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains. Outlier detection has been a widely researched problem and finds immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, intrusion detection for cyber security, fault detection in safety critical systems, military surveillance for enemy activities and many other areas.

The importance of outlier detection is due to the fact that outliers in data translate to significant (and often critical) information in a wide variety of application domains.

Outliers can also translate to critical entities such as in
military surveillance, where the presence of an unusual region in a satellite image of enemy area could indicate enemy troop movement.
Or anomalous readings from a space craft would signify a fault in some component of the craft. Outlier detection has been found to be directly applicable in a large number of domains. This has resulted in a huge and highly diverse literature of outlier detection Outlier detection methods can be classified into univariate methods and multivariate methods.
Or, one can classify them based on parametric methods (Hawkins, 1980;
Rousseeuw and Leroy, 1987; Barnett and Lewis, 1994) and non-parametric methods (Williams, Baxter, He, Hawkins and Gu, 2002).

1.1 Statement of the problem

Outlier in the data presents one of the main problems that usually face researchers. This is so since they can not usually be ignored and must be treated. This mean that in any set of data the researcher must test for the presence of outliers if he suspected that there is a possibility of their presence. Different methods are suggested in the literature for the detection of outliers. And although the efficiency of most of these methods are studied by many authors, yet there is still a need to investigate their performance under various conditions particularly under different sample size and different distributions. This is the problem that motivated this study.

1.2 Research objective:

The aim of this thesis is to:

(1) Investigate the sensitivity of various outliers' detection methods under different conditions.

(2) Propose a new method for outlier detection.

1-3 Research Approach

The analytical approach is adopted in this thesis. Univariate

techniques employed in outlier detection will be used. A simulation

experiment is performed to enable the comparison of the different

methods by using MATLAB Software.

1.4 Structure of the Thesis:

The thesis is organized as follows:

Chapter (1) contains the introduction , statement of the problem, research objective and research Approach.

Chapters(2) presents different definitions of outliers and their sources and Classifications. It also discusses their importance and applications.

A chapter (3) reviews previous work in the literature on outlier detection.

Chapter (4) a new method is proposed for detection of outliers.

Chapter (5) compares through a simulation study different methods

of outlier detections.

Finally, in chapter (6) a conclusion is presented that contains the main research Findings.

CHAPTER 2

2-1 Introduction:

Observed variables often contain outliers that have unusually large or small values when compared with others in a data set.

Some data sets may come from homogeneous groups; others from heterogeneous groups that have different characteristics regarding a specific variable, such as height data not stratified by gender.

Outliers can be caused by incorrect measurements, including data entry errors, or by coming from a different population than the rest of the data. In this chapter the various Definition , Masking and Swamping effect are reviewed . Source and type of outlier and the importance of detection outlier and some of the Applications of Outlier Detecting Techniques are also presented,Finally we discuss the Skewenss and Brokndownpoint.

2-2 Defintions of outlier:

Hadi et al.(2009) stated, "There are numerous definitions of outliers in the statistical and machine learning literatures." One commonly used definition is that outliers are a minority of observations in a dataset that have different patterns from that of the majority of observations in the dataset. The assumption here is that there is a core of at least 50% of observations in a dataset that are homogeneous (that is, represented by a common pattern) and that the remaining observations (hopefully few) have patterns that are inconsistent with this common pattern. However, given that there is currently no universally accepted definition for an outlier, the seven most-commonly used definitions of outliers are provided.

1. **Grubbs (1969)**:

"defines outlier as one that appears to deviate markedly from other members of the sample in which it occurs".

2. **Hawkins (1980):**

"defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism".

3. **Johnson (1992):**

"defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data".

4. **Mendenhall et al. (1993)**

apply the term "outliers" to values "that lie very far from the middle of the distribution in either direction".

5. **Barnet and Lewis (1994):**

"Indicate that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs".

6. **Pyle (1999):**

"states that an outlier is a single, or very low frequency, occurrence of the value of a variable that is far away from the bulk of the values of the variable".

7. **Hair, Anderson, Tatham and Black, (2005, pg 64) :**

"explains that outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations".

These definitions all refer to an observation that is surprisingly different from the rest of the data. However, the words "appears to deviate", "arouse suspicion", "inconsistent" and "distinctly different" imply some kind of subjectivity or preconceived ideas about what the data should look like.

The detection and study of outliers present a significant challenge to the data analyst in many areas of application.

Sometimes, the outliers themselves may be of interest as they might lead to new knowledge and discovery. In other cases, the presence of outliers can be a problem as they can significantly distort classical analysis of data and the inferences drawn from that analysis. Thus outlier detection has received and continues to receive considerable attention both inside and outside of the statistics literature

(see for example, Barnett (1978), Barnett and Lewis (1994), Cao et al. (2010), Cerioli (2010), Dang and Serfling (2010), Hawkins (2006), Louni (2008), Schwertman et al. (2004), Schwertman and de Silva (2007), Tukey (1977)).

2.3 Masking and Swamping effect:

There are many definitions of masking and swamping effect (see, Hawkins, 1980; Iglewics and Martinez, 1982; Davies and Gather, 1993; Barnett and Lewis, 1994,seo 2006) ((*Nazrina Aziz ((2010))*

2.3.1 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. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.

2.3.2 Swamping effect :

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. Swamping occurs when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers. A single step procedure with low masking and swamping is given in (Iglewics and Martinez, 1982).

2.4 type of outlier:

There are some of type of outlier as following :

2.4. 1 Type I Outliers:

In a particular dataset an individual outlying instance is termed as a Type I outlier. That single point of data is an outlier because of its attribute value which is inconsistent with values taken by normal instances.

Many of the existing outlier detection schemes focus on this single outlier. These techniques analyze the relationship of this single point of data with regard to the rest of the points in the dataset. For example, in credit card data or medical data each data represents a single transaction or a single patient.

2.4.2-Type II Outliers:

These outliers are caused because of the occurrence of an individual data instance in a specific context in the given data. These outliers are also individual data instances. But these type II outliers are in the context of a particular dataset especially in relation to its structure and problem formulation..

2.4.3-Type III Outliers:

These are not individual observations but rather are an entire subset of the entire dataset which are outliers. Their occurrence

together constitute an anomalous formulation. They are usually meaningful when the data has a sequential nature. These outliers are either subgraphs or subsets.

2.5 Source of outliers:

Outliers may arise coincidently without any anticipation by a researcher. Sometimes it cannot be explained. However, there are a few possible reasons for the existence of outliers in the data set.

Barnett and Lewis (1994) classified outlier source into three types.

The initial source is named as inherent variability, which implies a situation beyond one's control since it might arise from the natural characteristics of the individual variable.

For example, if the data collection involves time duration, it may cause an occurrence of outliers since some of the observations might be influenced by any event that might occur unexpectedly throughout the period of the study.

The next cause is measurement error such as reading, computing and typing errors during the data entry process. This possibly makes the observation peculiar compared to the other observations in the data set.

The last reason is the execution error, related to the research design where one might choose a biased sample or include individuals that are not true representatives of the population that is to be sampled.

No matter what the causes of outliers are, the most important aspect of outlier issue is the technique to identify whether there are outliers in the data set or not.

By identifying the existence of outliers, one may identify the source of the outliers.

Anscombe (1960) (cited by Beckman and Cook, 1983) divided outliers into two major categories. First, there might be errors in the data due to some mistake/error and second, outliers may be present due to natural variability. There might be the third category of outliers when they come from outside the sample. Ludbrook (2008) discussed a number of reasons of outlier‟s existence and methods of handling them.

Outliers can arise from several different mechanisms or causes.

It is therefore important to consider the range of causes that may be responsible for outliers in a given set of data.

1- Outliers from data errors:-

outliers are often caused by human error, such as errors in data collection, recording or entry. Data from an interview can be recorded incorrectly, or mistaken upon data entry. Errors of this nature can often be corrected by returning to the original documents or even the subjects if necessary and possible and entering the correct value.

2-Outliers from intentional or motivated mis-reporting: -

There are times when participants purposefully report incorrect data to experimenters or surveyors. A participant may make conscious effort to sabotage the research, Huck (2000) or may be acting from other motives. Social desirability and self presentation motives can be powerful. This can also happen for obvious reasons when data are sensitive (e.g. teenagers under-reporting drug or alcohol use, misreporting of sexual behaviour). If all but few teens under-report a behavior the few honest responses might appear to be outliers when in fact they are legitimate and valid scores. Motivated over-reporting can occur when the variable in question is socially desirable (e.g. income, educational attainment, grades, study times, church attendance, and sexual experience). Environmental conditions can motivate over-reporting or misreporting, such as if an attractive female researcher is interviewing male undergraduates about attitude on gender equality in marriage. Depending on the details of the research, one of two things can happen: inflation of all estimates, or production of outliers. If all subjects respond the same way, the distribution will shift upward, not generally causing outliers. However, if only a small subsample of the group responds this way to the experimenter, or if multiple researchers conduct interviews, then outliers can be created.

3-Outliers from sampling error:-

another cause of outliers is sampling. It is possible that a few members of a sample were inadvertently drawn from a different

population than the rest of the sample. These cases should be removed as they do not reflect the target population.

4-Outliers from standardization failure:-

outliers can be caused by research methodology, particularly if something anomalous happened during a particular subject experience one might argue that a study of stress levels in school children around the country might have found some significant outliers. Unusual phenomena such as a construction noise outside a research laboratory or an experimenter feeling particularly grouchy, or even events outside the context of the research laboratory ,such as a student protest, a rape or murder on campus, observations in the classroom the day before a big holiday recess and so on can produce outliers.

Faulty or non-calibrated equipments is another common cause of outliers. These data can be legitimately discarded if the researchers are not interested in studying the particular phenomenon in question .

5-Outliers from faulty distributed assumptions:-

incorrect assumptions about the distribution of the data can also lead to the presence of suspected outliers, Iglewieze and Hoaglin,(1993). may give rise to bimodal, skewed, asymptotic or fled is attributions, depending upon the sampling design. The data may have a different structure than the researcher originally assumed, and long or short-term trends may affect the data in unanticipated ways.

Depending on the goal of the research, these extreme values may or may not represent an aspect of the inherent variability of the data, and may have a legitimate place in the data set.

6-Outliers as legitimate cases sampled from the correct population:-

It is possible that an outlier can come from the population being sampled legitimately through random chance, it is important to note that sample size plays a role in the probability of outlying values. Within a normally distributed population, it is more probable that a given data point will be drawn from the most densely concentrated area of the distribution, rather than one of the tails Evans,(1999); Sachs, (1982). As a researcher casts a

wider net and the data set becomes larger, the more the sample resembles the population from which it was drawn and thus the likelihood of outlying values become greater. In other words, there is only about one percentage chance you will get an outlying data point from a normally distributed population, this means that, on the average, about one percentage of your subjects should be three standard deviations from the mean. In the case that outliers occur as a function of the inherent variability of the data, opinions differ widely on what to do. Due to the dexterous effect on power, accuracy and error rates that outliers can have, here it might be desirable to use a transformation or recoding/truncation strategy to both keep the individual in the data set and at the same time minimize the harm to statistical inference: Osborne, (2002).

7-Outliers as potential focus of inquiry:-

we all know that interesting research is often as much a matter of serendipity as planning and inspiration. Outliers can represent a nuisance error, or legitimate data. They can also be inspiration for inquiry. When researchers in Africa discovered that some women were living with HIV just fine for years and years, untreated, those rare cases were outliers compared to most untreated women, who die fairly rapidly. They could have been discarded as noise or error, but instead they serve as inspiration for inquiry. This extreme score might shed light on an important principal or issue. Before discarding outliers, researchers need to consider whether those data contain valuable information that may not necessarily relate to the intended study, but has importance in a more global sense. The presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistics estimates when using either parametric or nonparametric tests(Zimmerman, 1994, 1995, 1998). Casual observation of the literature suggests that researchers rarely report checking for outliers of any sort. This inference is supported empirically by Osborne, Christiansen and Gunter (2001), who found that authors reported testing assumptions of the statistical procedures used in their studies – including checking for the presence of outliers – only eight per cent of the time. Given what we know of the importance of

assumptions to accuracy of estimates and error rates, this in itself is alarming.

Wainer (1976) also introduced the concept of the "froigelier" referring to "unusual events which occur more often than Seldom" (p.286). These points lie near three standard deviations from the mean and hence may have a disproportionately strong influence on parameter estimates, yet are not so obvious or easily identified as ordinary outliers due to their relative proximity to the distribution center.

As fringeless are a special case of outliers, for much of the rest of this study we will use the generic term "outlier" to refer to any single data point of dubious origin or disproportionate influence.

occurring in the data.

2.6 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 cyber crime, 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, Rousseeuw and Orhan, 2001). Some techniques designed for skewed distributions such as the boxplot introduced by Mia Hubert and Ellen Vandervieren (2008) and some other techniques introduced by Banner and Iglewicz (2007) are designed for large sample sizes but there are also some techniques which are designed for smaller sample size like Dixon test (Constantinos E. Efstathiou, 2006). Some techniques like 2SD (standard deviation) perform well in the symmetric distributions but fail in the skewed distribution due to the fact that they construct large intervals of critical values around the means of asymmetrically centered distributions on the compressed side while short it on the skewed side of the distribution according to the level of skewness.

2.7 Applications of Outlier Detecting Techniques:

Outlier‟s detection can be applied on lot of data sets for various purposes. Some of which are discussed below:

2.7.1 Intrusion Detection Systems:

In many host-based or networked computer systems, different kinds of data are collected about the operating system calls, network traffic, or other activity in the system.

This data may show unusual behavior because of malicious activity.

The detection of such activity is referred to as intrusion detection.

2.7.2 Credit Card Fraud:

Credit card fraud is quite prevalent, because of the ease with which sensitive information such as a credit card number may be compromised. This typically leads to unauthorized use of the credit card. In many cases, unauthorized use may show different patterns, such as a buying spree from geographically obscure locations. Such patterns can be used to detect outliers in credit card transaction data.

2.7.3 Interesting Sensor Events:

Sensors are often used to track various environmental and location parameters in many real applications.

The sudden changes in the underlying patterns may represent events of interest. Event detection is one of the primary motivating applications in the field of sensor networks.

2.7.4 Medical Diagnosis:

In many medical applications the data is collected from a variety of devices such as MRI scans, PET scans or ECG time-series. Unusual patterns in such data typically reflect disease conditions.

2.7.5 Law Enforcement:

Outlier detection finds numerous applications to law enforcement, especially in cases, where unusual patterns can only be discovered over time through multiple actions of an entity.

Determining fraud in financial transactions, trading activity, or insurance claims typically requires the determination of unusual patterns in the data generated by the actions of the criminal entity.

2.7.6 Earth Science:

A significant amount of spatiotemporal data about weather patterns, climate changes, or land cover patterns is collected through a variety of mechanisms such as satellites or remote sensing. Anomalies in such data provide significant insights about hidden human or environmental trends, which may have caused such anomalies.

2.8 Skewness:

Asymmetry in the probability distribution of the random variable is known to be the skewness of that random variable. Using the conventional third moment measure, the value of skewness might be positive or negative or may be undefined. If the distribution is negatively skewed, it implies that tail on the left side of the probability density function is longer than the right hand side of the distribution. It also shows that larger amount of the values including median lie to the right of the mean. Alternatively, positively skewed distribution indicates that the tail on the right side is longer than the left side and the bulk of the values lie to the left of the mean. If the value of the skewness is

exactly zero, this suggests symmetry of the distribution. The third moment is a crude measure of symmetry, and in fact highly asymmetric distributions may have zero third moment. In addition, the third moment is extremely sensitive to outliers, which makes it unreliable in many practical situations. It is therefore useful to develop alternative measures of skewness which are insensitive to outliers and more direct measures of symmetry (Iftikhar 2011).

2.9 **Breakdownpoiont:**

The notion of breakdown point was introduced by Hodges (1967) and Hampel (1968, 1971). It's a robustness measure of an estimator such as the mean and median or a related procedure using the estimators. The breakdown point of an estimator generally can be defined as the largest percentage of the data that can be changed into arbitrary values without distorting the estimator.

For example, if even one observation of a univariate data set is moved to infinity, the estimators of the data set such as the mean and variance go to infinity. Thus, the breakdown point of these estimators is zero. In contrast, the breakdown point of the median is approximately 50% and it varies slightly according whether the sample size n is odd or even. The exact breakdown point of the median is $50(1-1/n)$ % and $50(1-2/n)$ % for odd sample size n and even sample size n, respectively. Therefore, if the breakdown point of an estimator is high, the estimator is robust.

CHAPTER 3

3.1 **Introduction:**

An outlier is a data point which is significantly different from the remaining data Outliers are also referred to as *abnormalities*, *discordants*, *deviants*,or *anomalies* in the data mining and statistics literature. In most applications,the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers, so in this chapter we review different methods for identify in outliers.

3.2 A historical review of Detection methods of Univariatc Outliers:

Detection of outliers in the analysis of the data sets dates back to 18th century. Bernoulli (1777) pointed out the practice of deleting the outliers about 200 years ago. Deletion of outliers is not a proper solution to handle the outliers but this remained a common practice in past. To address the problem of outliers in the data, the first statistical technique was developed in 1850 (Beckman and Cook, 1983).

Some of the researchers argued that extreme observations should be kept as a part of data as these observations provide very useful information about the data. For example, Bessel and Baeuer (1838) claimed that one should not delete extreme observations just due to their gap from the remaining data (cited in Barnett, 1978). The recommendation of Legendre (1805) is not to rub out the extreme observations "adjudged too large to be admissible". Some of the researchers favored to clean the data from extreme observations as they distort the estimates. An astronomer of 19th century, Boscovitch, put aside the recommendations of the Legendre and led them to delete (ad hoc adjustment) perhaps favoring the Pierce (1852), Chauvenet (1863) or Wright (1884). Cousineau and Chartier (2010) said that outliers are always the result of some spurious activity and should be deleted. Deleting or keeping the outliers in the data is as hotly discussed issue today as it was 200 years ago.

Bendre and Kale (1987), Davies and Gather (1993), Iglewicz and Hoaglin (1994) and Barnett and Lewis (1994) have conducted a number of studies to handle issues of outliers. Defining outliers by their distance to neighboring examples is a popular approach to finding unusual examples in a dataset known to be distance based outlier detection technique.

Saad and Hewahi (2009) introduced Class Outlier Distance Based (CODB) outlier‟s detection procedure and proved that it is better than distance based outlier‟s detection method.

Surendra P. Verma (1997) emphasize for detection of outliers in univariate data instead of accommodating the outliers because it provides better estimate of mean and other statistical parameters in an international geochemical reference material (RM).

Hadi and Simonoff (1993) provided distributional results for testing, multiple outliers in regression analysis. The test is based on the deletion residual. Beckman and Cook (1983) encountered a serious problem of "masking" if there are several outliers.

Least square estimation of the parameter of the model may lead to small residuals for the outlying observations. Single detection methods (for example Cook and Weisberg,1982; Alkinson, 1985) may fail and the outliers will go undetected.

Hawkins (1983) argues for exclusion of all possible outlying observations, which are then tested sequentially for reinclusion. The drawback to this procedure is that it is unclear how many observations should be deleted,and because of masking, which ones, before reinclusion and testing begin.

Carling (1998) introduces the median rule for identification of outliers through studying the relationship between target outlier percentage and Generalized Lambda Distributions (GLDs).

GLDs with different parameters are used for various moderately skewed distributions.

The use of the forward search in regression is described in Atkinson and Riani (2000) whereas in Atkinson (1994) the emphasis oninformative plots and their interpretations.

Although the forward search is a powerful general method for the detection of multiple outliers and unidentified clusters and of their influential effects.

The interest here is in Atkinson (1994) on information plots and the information it provides about the adequacy of our simple approximation to the distribution of the test statistic.

Possible sources of outliers are recording and measurement errors incorrect distributional assumptions, unknown data structure, or novel phenomenon (Iglewiez, 1993).

A data set indicative of a novel phenomenon can be often labeled as an outlier. For instance, the measurements indicating existence of the hole in the ozone layer were initially thought to be outliers and they were automatically discarded.

This join delayed the discovery of the phenomenon by several years (Berthouex,1994).

The first step in data analysis is to label suspected outliers for further study.

Three different methods are available to the investigation for normally distributed data: z score method, (Iglewiez, 1993; Barnett,1984). All of the experimental observations are standardized and the standardized values outside a predetermined bound are labeled as outliers (Rousseeuw, 1987).

Outliers can arise from several different mechanisms as causes. Anscombe (1960)sorts outliers into categories from intentional or motivated misreporting; a participant may make a conscious effort to sabotage the research (Huck, 2000) or may be acting from other motives. In outliers from faulty distributional assumptions, incorrect assumption about the distribution of the data can also lead to the presence of suspected outliers (Iglewiez and Hoaglin, 1993).

Due to the deleterious effect on power accuracy, and error rates that outliers can have, it might be desirable to use a transformation or recording strategy to both keep the individual in the data set and at the same time minimize the harm to statistical inference (Osborne, 2002).

Rosner's Test identifies outliers that are both high and low; it is therefore always two tailed (Gibbon, 1994).

The R. Statistics is compared with a critical value (Gilbert, 1987). Rosner's (1983) "many outlier" sequential procedures

is an improved version of Rosner's (1983)"extreme studentized deviate" outlier test.

Simonoff (1982) found this earlier well compared to other outlier test, although Rosner (1983) points out that it tends to detect more outliers than are actually present.

Rosner's (1983) method assumes that the main body of data is from a normal distribution.

Rosner's tests are two tailed since the procedure identifies either suspiciously large or suspiciously small data. When a one tailed test is needed, that is when there is interest in detecting only large values or only small values, then the skewness test for outliers discussed by Barnett and Lewis (1994) is suitable.

Hamilton, L.C. (1982) gove a graphical procedure for identifying outliers from bivariate normal or bivariate log normal distributions.

Rather than identifying outliers and discarding them before doing least square regression, one could do robust regression, as discussed and illustrated by Rousseeuw and Leroy (1987)who cautioned that robust regression should be applied only after the investigator is satisfied that less weight should be applied to the divergent data.

Non-parametric regression discussed by Holander and Wolfe (1973), and Reckhow and Chapra (1983) is an alternative to either standard least squares regression or robust regression.

Methods for detecting outliers have received a great deal of attention recently Cook and Wainer, 1976 and Steven, 1984). Leverages are related to an alternate regression diagnostic, Mahalanobis distance (Stevens,1984).

Mixture regression occurs when there is an omitted categorical predictor like gender,species or location and different regression occur in each category. It has long been recognised that a lurking variable, a variable that has an important effect but is not present among the predictors under consideration (Box, 1966; Joiner, 1981; Moore, 1997) can complicate regression analyses.

Atkinson, (1994) have applied Akaike Criterion (AIC) in detection of outliers by using (quasi) Bayesian approach with

Predictive likelihood in place of the usual likelihood function otherwise, detection of outliers has a long history. The main theme, however, has been around univariate and single outliers. Recently, some promising results have been obtained in detecting multiple outliers also in multivariate analysi(Hadi, 1992).

An approach to the identification of aberrant points is the construction of outliers'diagnostics.

These are quantities computed from the data with the purpose of pinpointing influential points, after which these outliers

are to be removed or corrected, followed by a least square analysis on the remaining cases.

When there is only a single outlier, some of these methods work quite well by looking at the effect of deleting one point at a time (Atkinson, 1985 ;) Cook and Weisberg, 1982 and Hawkins, 1980). Unfortunately, it is much more difficult to diagnose outliers when there are several of them, due to the so-called masking effect which says that one may mask another. The naira extensions of classical diagnostics to such multiple outliers often give rise to extensive computations.

Recent work by Atkinson (1986), Hawkins, Bradu

and Kass (1984), and Rousseeuw and Van Zomeran (1999) indicates that one needs to use robust methods in one way or another to safely identify multiple outliers.

Some researchers prefer visual inspection of the data. Others (Lornez, 1987) argue that outlier detection is merely a special case of the examination of data for influential data points.

In analysis of variance, the biggest issue after screening for univariate outliers is the issue within cell outliers or the distance of an individual from the subgroup. Standardised residuals represent the distance from the subgroup and thus are effective in assisting analysis in examining data for multivariate outliers. Tabachnick and Fidell (2000) discuss data cleaning in the context of other analyses.

Where outliers are illegitimately included in the data, it is only common sense that those data points should be removed (Barnett and Lewis, 1994).

Few should disagree with that statement. When the outlier is either a legitimate part of the data or the cause is unclear, the issue becomes unclear. Murkier Judd and McClelland (1989) make several strong points for removal even in these cases in order to get the most honest estimate of population parameters. (Barnett and Lewis, 1994).

One mean of accommodating outliers is the use of transformations (Osborne, 2002).

By using transformation extreme scores can be kept in the data set, and the relative ranking of scores remains yet the skew and error variance present in the variable can be recorded (Hamilton, 1992).

However, transformations may not be appropriate for the model being tested or may affect its interpretation in undesirable ways.

Taking the logarithms of a variable makes a distribution less skewed, but it also alters the relationship between the original variables in the model (Newton and Rudestam, 1999; Osborne, 2001).

Hodge and Austin (2004) have pointed towards the significance of outliers in various contexts such as making decision about the loan application of problematic customers, intrusion detection, activity monitoring, network performance, fault diagnosis, structural defect detection, satellite image analysis, detecting novelties in images, motion segmentation, time-series monitoring, medical condition monitoring, pharmaceutical research, motion segmentation, detecting image features moving independently, detecting novelty in text, detecting unexpected entries in database and detecting mislabeled data in a training data set besides many other situations.

Instead of transformation, researchers sometimes use various robust procedures to protect their data from being distorted by the presence of outliers. These techniques "accommodate the outliers at no serious inconvenience or are robust against the presence of outliers (Barnett and Lewis, 1994; p.35). Certain parameter estimates, especially the mean and least square estimates, are particularly vulnerable to outliers, or have "low breakdown"

values. For this reason, researchers turn to robust or "high breakdown" methods to provide alternative estimates for these important aspects of data.

A common robust estimation method of the univariate distributions involves the use of trimmed mean, which is calculated by temporarily eliminating extreme observationsof both ends of the sample (Anscombe, 1960).

Alternatively, researchers may choose to compute a winsorized mean, for which the highest and lowest observations are temporarily censored, and replaced with adjacent values from the remaining data(Barnett and Lewis, 1994).

Assuming that the distribution of prediction errors is close to normal, several common robust regression techniques can help reduce the influence of outlying data points.

The least trimmed squares (LTS) and the least median of squares (LMS) estimators are conceptually similar to the trimmed mean, helping to minimize the scatter of the prediction errors by eliminating a specific percentage of the largest positive and negative outliers (Rousseeuw and Leroy, 1987).

While Winsorized regression smoothes Y-data by replacing extreme residuals with the next closest value in the dataset (Lane, 2002).

In correlations, we are expected to see the effect of outliers on two different types of correlations. These are correlations close to zero (to demonstrate the effect of outliers on Type II error rates) correlations will be calculated in each sample both before removal of outliers and after. If a sample correlation leads to a decision that deviated from the "correct" state of affairs it was considered an error of inference. In most cases the incidence of errors of inference was lower with cleaned than unclean data.

For the T-test and Analysis of Variance (ANOVA) this deals with analysis that look at group mean differences, such as the t-test and analysis of variance. For the purpose of simplicity these analyses are simple t-tests but these results would be generalized to any analysis of variance. For these analyses two different conditions are examined when there were no significant differences between the groups in the population and when there

were significant group differences in the population. For both variables the effects of having outliers in only one cell as compared to both cells were examined.

Removal of outliers will produce a significant change in the mean differences between two groups. It will also produce significant change in the t-statistics. Evidence of outliers may produce type I or type II errors. Removal of outliers may tend to have a significant beneficial effect on error rates.

Most analysts argue that removal of extreme scores produces undesirable outcomes; they are in the minority especially when the outliers are illegitimate.

When the data points are suspected of being legitimate, some authors Orr, Sacketts, P.R. and Du Bois(1991), argue that data are more likely to be

representative of the population as a whole if outliers are not removed. Conceptually, there are strong arguments for removal or alteration of outliers. In some analyses the benefits of outliers' removal are reported.

Both correlations and t-tests may show significant changes in statistics as a function of removal of outliers. In most cases errors of inference were significantly reduced, a prime argument for screening and removal of outliers.

It is straightforward to argue that the benefits of data cleaning extend to simple and multiple regressions to different types of ANOVA procedures. There are other procedures outside these but the majority of social science research utilizes one of these procedures. Other researches (e.g. Zimmerman, 1995) have dealt with the effects of extreme scores in less commonly used procedures, such as nonparametric analyses.

Thus, checking for the presence of outliers and understanding how they impact data analysis are extremely part of a complete analysis, especially when any statistical technique is involved.

3.3 Influence Measures:

The hat diagonal and residual measures are useful diagnostic measures to quantify an observation's remoteness in X-space and the distance of the regression surface. Influence diagnostic

measures have been developed to help in making the decision of what to do with an unusual observation.

Cook (1977) proposed Cook's squared distance ($CD^2_{(i)}$). This distance measure can be expressed in a general form as

$$CD^2_{(i)} = (\hat{\underline{\beta}}_{(i)} - \hat{\underline{\beta}})'(X'X)(\hat{\underline{\beta}}_{(i)} - \hat{\underline{\beta}}) / p\hat{\sigma}^2, \quad \ldots\ldots\ldots \quad (3.1)$$

for $i = 1,2, ,n$. The point with $CD^2_{(i)} > 1.0$ indicates that the i^{th} observation has influence on the determination of $\hat{\underline{\beta}}$.

Cook's squared distance of the i^{th} unit is a measure based on the square of the maximum distance between the least-squares estimate based on all n points of $\hat{\underline{\beta}}$, and $\beta_{(i)}$ is the least-squares estimate of $\hat{\beta}$, with the i_{th} observation deleted.

Cook and Weisber (1989) suggested that the cases where ($CD^2_{(i)} > 1.0$) $CD_i > 1.0$ should always be carefully noted.

Andrews and Predibon (1978) proposed the determinant ratio

$$AP(1) = \frac{\det\{z'(1)z(1)\}}{\det\{z'z\}}, \quad \ldots\ldots\ldots \quad (3.2)$$

where Z is the X matrix with the response variables y_i,

($z = \{(x_{11}, \ldots x_{1p}, y_1), (x_{21} \ldots, x_{2p}, y_2), \ldots, (x_{n1}, \ldots x_{np}, y_n)$,

and $z_{(I)}$ is the part of Z obtained by deleting the rows corresponding to the indicated I.

Rousseeuw and Leroy (1987) mentioned that Belsley et al., (1980) used

$$DFFITS_{(1)} = (h_{ii}^{1/2} r_1) / (\sigma_{(1)}(1 - h_{ii})), \quad \ldots\ldots\ldots \quad (3.3)$$

for $i = 1,2, ,n$, where h_{ii} is the i^{th} diagonal element of $X_1 (X^1 X)^{-1} X$.

$DFFITS_{(i)}$ is the number of standard deviations from the i^{th} component of $\hat{y} - \hat{y}_{(i)}$.

So $DFFITS_{(i)}$ measures the influence on the prediction when an observation is deleted. The criterion suggests that when

$$DFFITS_{(1)} > 2\sqrt{\frac{p}{n}}$$

attention to outliers is warranted.

In another way, Belsey et al., (1980) defined a diagnostic measure based on the change in the j^{th} regression coefficient of $\beta_{j(\iota)}$. The statistic is

$$DFBETAS = \left(\frac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c^2_{jk}}}\right)\left(\frac{r_i}{\hat{\sigma}^2_{(1)}(\widehat{1-h_{11}})}, \right) \quad \ldots\ldots\ldots\ldots\ldots \quad (3.4)$$

where the n elements in the j_{th} row of C produce the leverage element $(XX)^{-1}X'$ and h_{ii} is the i^{th} diagonal element of $x^1(X^1X)^{-1}X$.

Specifically, if $DFBETAS\,j^{(i)} \quad 2 / > \sqrt{n}$
then the i^{th} observation warrants an examination for outliers.
Atkinson (1985) modified Cook's squared distance and denoted Atkinson's modified statistic as

$$A_i = DFFITS_{(i)}\,[(n-p)\,]^{\frac{1}{2}}. \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.5)$$

Atkinson's modified statistic is closely connected to $DFFITS_{(i)}$. The measure obtains another set related to Cook's squared distance, suggesting that, when $A_i > \sqrt[2]{n-\frac{n}{P}}$,
there should always be outliers. The measures
$CD^2_{(i)}$, $DFFITS_{(i)}$ and Atkinson's modified statistic are very similar, therefore usually only one of them is used.
The COVRATIO's statistic measures the change in the determinant of a covariance matrix of the estimates by deleting the *th i* observation. Therefore, the design point *i x* would be affected if the i^{th} case were deleted. any observation where *i*

$$COVRATO_I > 1+3p/\,n$$

or $COVRATIO_{\,i} < 1-3p/\,n$

warrants attention for outliers, where *p* is the number of parameters in the model and *n* is the number of observations used to fit the model.
Kempthorne and Mendel (1990) discuss the inadequacies of these single row influence diagnostics when applied to multiple observations.
Cook (1998) gives guidance on numerous other modern graphical procedures that can provide insight into outliers and influence in regression.the problem is that they can fail if there are multiple outliers.

3.4 Various methods for identification of outliers:

In the remainder of this chapter the various methods of outlier's identification are discussed in some detail There are many methods available for the identification of outliers.
All of these methods can basically be grouped into two categories, namely the univariate method and the multivariate method (see Hawkins, 1980; Barnett and Lewis, 1994).

The univariate method is performed independently on each variable, whereas the multivariate method investigates the

relationship of several variables (Franklin, Thomas and Brodeur, 2000). One can also classify the methods in both categories into parametric and nonparametric approaches.

Other classifications of outlier detection methods can be found in Papadimitriou, Kitawaga, Gibbons and Faloutsos (2002), Hu and Sung (2003) and Acuna and Rodriguez (2004).

This chapter will briefly explain outlier identification methods for high-dimensional data. Detailed explanations about those methods can be found in Hawkins (1980), Barnett and Lewis (1994), Papadimitriou et al. (2002), Hu and Sung (2003) and Acuna and Rodriguez (2004).

This chapter does not attempt to summarize literature covering the univariate method but some major concepts are reviewed before moving to the multivariate method.

Most univariate methods assume a known distribution of the data (i.e. often independent and identically distributed) and often assume that the distribution parameters and the type of expected outliers are also known (Barnett and Lewis, 1994).

Ben-Gal (2005) notes that these assumptions are often violated in the real world of data-mining applications. Seo (2006) wrote a thesis comparing different methods for detecting outliers in univariate data sets. Many methods have been proposed for univariate outlier detection. The test of discordance, i.e. a formal test, and outlier labeling methods, i.e. informal test are the most popular approaches.

3.4.1 Test of discordance:

The test of discordance needs test statistics for hypothesis testing and it is usually based on the assumption of well-behaved distribution.

Normally the distribution is assumed to be identically and independently distributed. Additionally, the type of expected outlier and the distribution parameters are assumed to be known.

From Barnett and Lewis (1994), there are hundreds of discordance tests that have been developed for different conditions depending on

(i). the data distribution, i.e. whether the distribution parameters are known or not;

(ii) the number of expected outliers;

(iii) the types of expected outliers.

The test of discordance is quite powerful since it is based on distribution assumption. However, it is noted that most real world data may not followa specific distribution or the distribution is unknown.

The discordance test is thoroughly discussed in Barnett and Lewis (1994) and Iglewicz and Hoaglin (1993). Examples of discordance test are generalized extreme studentized deviate (ESD), kurtosis statistics and the Dixon test and Gurps test.

3.4.1.1Grubs test:

Grubbs (1969) introduced a test for detection of outliers for the univariate normal distribution with the sample size greater than 3.

Grubbs statistics is given as:

$$G = Max \quad \frac{(x-\bar{x})}{SD} \qquad (3.6)$$

Where $\bar{x}$ and SD are the sample mean and standard deviation respectively. Null hypothesis of Grubbs test is that data have no outliers while the alternative is that at least one outlier in the data is present.

As given in the above statistics largest absolute value of G is suspected as the outlier and the decision whether the observation is outlier or not is made by looking it in the table of critical values (Grubbs, 1969).

3.4.1.2 Extreme Studentized Deviate (ESD):

The ESD test is suitable to use if we want to identify a single outlier in a normally distributed data. It is also known as the Grubb test.

The maximum deviation from the mean is given as

$$\tau = \frac{|x_i - \bar{x}|}{SD} \qquad (3.7)$$

where x_i is the observation, $\bar{x}$and SD are the mean and standard deviation of the data set, respectively. Equation 2.9 is calculated for each observation and the value is compared to the critical value, $\bar{\tau}$ at the selected α. If τ is greater than the $\bar{\tau}$ (see Iglewicz and Hoaglin(1993) for ESD test critical values), then the observation under consideration is an outlier.

3.4.1.3 **Dixon test:**

The Dixon test is based on the ratio of the ranges and it is g enerally used for detecting a small number of outliers. There are six test statistics from Dixon for normal univariate samples. It is a very simple test. The algorithm is as follows:

• Step 1: Observations in the data set are sorted in ascending order,

$x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ where $x_{(1)}$ is the lowest a

• Step 2: Compute the suitable test statistics and depending on the number of suspected outliers, different test statistics are used to identify potential outliers. The corresponding test statistics are given in Table 3.1

Tests r_{10}, r_{11}, r'_{11}, r_{12} and r'_{12} are the test statistics for an extreme outlier, $x_{(n)}$ or $x_{(1)}$ in a normal sample with population variance unknown, whereas tests r_{20}, r'_{20}, r_{21}, r'_{21}, r_{22} and r'_{22} are for two extreme observations either the upper-pair $x_{(n)}$, $x_{(n-1)}$ or the lower-pair $x_{(1)}$, $x_{(2)}$ in a similar normal sample;

• Step 3: Next the value of test statistics is compared to the critical value, $r*$ for a given number of observations n .

Table 3.1 Dixon tests for univariate normal samples

Applicability of test $n_{min} - n_{max}$	Value$_{(s)}$ tested	Test Statistics
3-30	Upper $x_{(n)}$	$r_{10} = \frac{(x_{(n)} - x_{(n-1)})}{(x_{(n)} - x)_{(1)})}$
4-30	Upper $x_{(n)}$	$r_{11} = \frac{(x(n) - x(n-1))}{(x_{(n)} - x_{(2)}}$
4-30	Lower $x_{(1)}$	$r'_{11} = \frac{(x_{(2)} - x_{(1)}))}{(x(n-1) - x(1))}$
5-30	Upper $x_{(n)}$	$r_{12} = \frac{(x_{(n)} - x_{(n-1)})}{(x_{(n)} - xx_{(3)})}$
5-30	Lower $x_{(1)}$	$r'_{12} = \frac{(x_{(2)} - x_{(1)})}{(x_{(n-2)} - x_{(1)})}$
4-30	Upper pair $x_{(n),}X_{(n-1)}$	$r_{20} = \frac{(x_{(n)} - x_{(n-2)})}{(x_{(n)} - x_{(1)})}$
4-30	Lower pair $x_{(1)}$,$x_{(2)}$	$r'_{20} = \frac{(x_{(3)} - x_{(1)})}{(x_{(n)} - x_{(1)}}$
5-30	Upper pair $x_{(n)}$, $x_{(n-1)}$)	$r_{21} = \frac{(x_{(n)} - x_{(n-2)})}{(x_{(n)} - x_{(2)}}$
5-30	Lower pair $x_{(1)}$, $x_{(2)}$	$r'_{21} = \frac{(x_{(3)} - x_{(1)})}{(x_{(n-1)} - x_{(1)})}$
6-30	Upper pair $x_{(n)}$, $x_{(n-1)}$	$r_{22} = \frac{(x_{(n)} - x_{(n-2)})}{(x_{(n)} - x_{(3)})}$
6-30	Lower pair $x_{(1)}$, $x_{(2)}$	$r'22 = \frac{(x_{(3)} - x_{(1)})}{(x_{(n-2)} - x_{(1)})}$

given significance α. (The r^* critical value can be found in Kanji (1993));

• Step 4: If the test statistic is less than the critical value $r*$, there are no outliers present.

However, if the test statistic is greaterthan the critical value, the null hypothesis is rejected and the conclusion is that the most extreme value is an outlier. The test is applied consecutively for other extreme values until the null hypothesis is true.

3.5 Outlier labeling methods:

Outlier labeling methods use the interval for identification of outliers. The interval will separate outliers into 'good region' and 'bad region'. Bad region refers to the area outside the interval. Any observations that fall in the bad region are considered as

outliers. Normally, outlier labeling methods are appropriate to use if one is only interested in finding an observation that is extremely different from the majority data. This method is not suitable to be applied if one wants to identify the observation that violates the distribution assumption of statistical analyses, such as regression.

Another reason for using the outlier labeling method is when we have a large data set.

Note that it is difficult to identify the distribution of a large data set. Therefore, in this condition, the labeling method is appropriate for outlier detection rather than discordance tests Here will divide outlier labeling methods into two grubs as follows:-

3.5.1 univariate outlier detection:

3.5.1.1 STANDARD DEVIATION (SD) METHOD:

The simple classical approach to screen outliers is to use the SD (Standard Deviation) method. It is defined as

2 SD Method: $x \pm 2$ SD (3.8)

3 SD Method: $x \pm 3$ SD, (3.9)

where the mean is the sample mean and SD is the sample standard deviation.

The observations outside these intervals may be considered as outliers. According to the Chebyshev inequality, if a random variable X with mean μ and variance σ^2 exists, then for any k > 0,

$$P[\,|X-\mu|] \geq k\sigma \leq \frac{1}{k^2}$$

$$P[\,|X-\mu|] \geq k\sigma \leq 1-\frac{1}{k^2}, \quad k > 0$$

the inequality $[1-\frac{1}{k^2}]$ enables us to determine what proportion of our data will be within *k* standard deviations of the mean3. For example, at least 75%, 89%, and 94% of the data are within 2, 3, and 4 standard deviations of the mean, respectively. These results may help us determine the likelihood of having extreme values in the data.

Although Chebychev's therom is true for any data from any distribution, it is limited in that it only gives the smallest proportion of observations within *k* standard deviations of the mean. In the case of when the distribution of a random variable is known, a more exact proportion of observations centering around the mean can be computed. For instance, if certain data follow a

normal distribution, approximately 68%,95%, and 99.7% of the data are within 1, 2, and 3 standard deviations of the mean, respectively; thus, the observations beyond two or three SD above and below the mean of the observations may be considered as outliers in the data.

3.5.1.2 Z-SCORE:

Another method that can be used to screen data for outliers is the Z-Score, using the mean and standard deviation.

$$Z_i = \frac{(\ x_i - \bar{x}\ )}{SD} \qquad (3.10)$$

Where $X_i \sim N(\mu, \sigma^2)$, and *SD* is the standard deviation of data.

The basic idea of this rule is that if X follows a normal distribution, N (μ, σ^2), then *Z* follows a standard normal distribution, N (0, 1), and Z-scores that exceed 3 in absolute value are generally considered as outliers.

This method is simple and it is the same formula as the 3SD method when the criterion of an outlier is an absolute value of a Z-score of at least 3.

It presentsa reasonable criterion for identification of the outlier when data follow the normal distribution.

According to Shiffler (1988), a possible maximum Z-score is dependent on sample size, and it is computed as $\frac{(n-1)}{n}$. Since no z-score exceeds 3 in a sample size less than or equal to 10, the z-score method is not very good for outlier labeling, particularly in small data sets. Another limitation of this rule is that the standard deviation can be inflated by a few or even a single observation having an extreme value. Thus it can cause a masking problem, i.e., the less extreme outliers go undetected because of the most extreme outliers and vice versa. When masking occurs, the outliers may be neighbors.

3.5.1.3THE MODIFIED Z-SCORE:

Two estimators used in the Z-Score, the sample mean and sample standard deviation, can be affected by a few extreme values or by even a single extreme value. To avoid this problem, the median and the median of the absolute deviation of the median (MAD) are employed in the modified Z-Score instead of the mean

and standard deviation of the sample, respectively(Iglewicz and Hoaglin, 1993).

$$MAD = median\{|x_i - \bar{x}|\} \quad (3.11)$$

where $\bar{x}$is the sample median.

The modified Z-Score (M_i) is computed as

$$M_i = \frac{0.6745(x_i - \bar{x})}{MAD} \quad (3.12)$$

where E(*MAD*)=0.6745 σ for large normal data.

Iglewicz and Hoaglin (1993) suggested that observations are labeled outliers when$|M_i| > 3.5$

through the simulation based on pseudo-normal observations for sample sizes of 10, 20, and 40 The M_i score is effective for normal data in the same way as the Z-score.

3.5.2 Univariate Robust methods:

3.5.2.1 MADE METHOD:

The MADe method, using the median and the Median Absolute Deviation (MAD), is one of the basic robust methods which are largely unaffected by the presence of extreme values of the data set. This approach is similar to the SD method. However, the median and MADe are employed in this method instead of the mean and standard deviation. The MADe method is defined as follows;

$2\ MAD_e$ Method: Median $\pm$ $2\ MAD_e$ (3.13)

$3\ MAD_e$ Method: Median $\pm$ $3\ MAD_e$, (3.14)

where MAD_e=1.483×MAD for large normal data.

MAD is an estimator of the spread in a data, similar to the standard deviation, but has an approximately 50% breakdown point like the median.

MAD= median ($|x_i$ – median(x)|) i=1,2,…,n)

When the MAD value is scaled by a factor of 1.483, it is similar to the standard deviation in a normal distribution. This scaled MAD value is the MAD_e.

Since this approach uses two robust estimators having a high breakdown point, i.e., it is not unduly affected by extreme values even though a few observations make the distribution of the data skewed, the interval is seldom inflated, unlike the SD method.

3.5.2.2 MEDIAN RULE:

The median is a robust estimator of location having an approximately 50% breakdown point. It is the value that falls exactly in the center of the data when the data are arranged in order. That is, if $x_1, x_2, \ldots, x_n$ is a random sample sorted by order of magnitude, then the median is defined as:

Median, $\tilde{x} = x_m$ when n is odd

$\tilde{x} = (x_m + x_m + 1)/2$ when n is even

where m=round up (n/2)

For a skewed distribution like income data, the median is often used in describing the average of the data.

The median and mean have the same value in a symmetrical distribution.

Carling (1998) introduces the median rule for identification of outliers through studying the relationship between target outlier percentage and Generalized Lambda Distributions (GLDs).

GLDs with different parameters are used for various moderately skewed distributions.

The median substitutes for the quartiles of Tukey's method, and a different scale of the (IQR) is employed in this method. It is more resistant and its target outlier percentage is less affected by sample size than Tukey's method in the non-Gaussian case.

The scale of IQR can be adjusted depending on which target outlier percentage and GLD are selected. In my thesis, 2.3 is chosen as the scale of IQR; when the scale is applied to normal distribution, the outlier percentage turns out to be between Tukey's method of 1.5 IQR and that of 3 IQR, i.e., 0.2 %.

It is defined as:

$$[C1, C2] = Q_2 \pm 2.3\ IQR \qquad (3.15)$$

where Q_2 is the sample median.

3.5.2.3 Boxplot:

One of the well known and widely used labeling methods is the Boxplot. The Boxplot was introduced by Tukey in 1977.

Tukey introduced the Boxplot as a graphical display on which outliers can be indicated.

The observation that falls between the inner fence and outer fence, or beyond the outer fence is labeled as an outlier. The inner fence is calculated as

$$[Q_1 - 1.5\ IQR,\ Q_3 + 1.5\ IQR], \qquad (3.16)$$

where $IQR = Q_3 - Q_1$ is the inter quartile range of the data set with$Q_3$and Q_1 are the upper quartile of the data set and the lower quartile of the data set, respectively. One can compute the outer fence as

$$[Q_1 - 3\ IQR,\ Q_3 + 3\ IQR]. \qquad (3.17)$$

Notice that the upper and lower quartiles, Q_3 and Q_1 are used to obtain the robust measures for mean,$\frac{(Q_1+Q_3)}{2}$ and the standarddeviation, $Q_3 - Q_1$, which can replace $\bar{X}$ and s in equation.

The Boxplot is applicable to skewed data since it makes no distributional assumptions and it does not depend on a mean or standard deviation. However it is not suitable for a small sample size and it is noted that the more skewed the data are, the more observations may be detected as outliers.

3.5.2.4 ADJUSTED BOXPLOT:

Although the boxplot proposed by Tukey (1977) may be applicable for both symmetric and skewed data, the more skewed the data, the more observations may be detected as outliers,

This results from the fact that this method is based on robust measures such as lower and upper quartiles and the IQR without considering the skewness of the data.

Vanderviere and Huber (2004) introduced an adjusted boxplot taking into account the medcouple (*MC*), a robust measure of skewness for a skewed distribution.

When X_n={ x , x ,..., x_n } is a data set independently sampled from a continuous univariate distribution and it is sorted such as $n\ x \leq x \leq ... \leq x$ the *MC* of the data is defined as:

$$MC = \underset{\substack{x_i \leq \bar{x} \leq x_i \\ x_i \neq x_i}}{median}\ h\left(x_{i\,,}\ x_{i\,,}\right)(\ ,...,\ )\ 1\ ,$$

where *k med* is the median of X*n*,

$$h\left(x_j x_j\right) = \frac{(x_{j-}\ \bar{x}) - (\bar{x} - x_j)}{x_j - x_j} \qquad (3.18)$$

and j have to satisfy $i\ x \leq k\ med \leq j\ x$, and $i\ x \neq j\ x$. The interval of the adjusted boxplot is as

follows (G. Bray et al. (2005)):

$$[L, U] = [Q_1\text{-}1.5 * e^{(-3.5MC)} * IQR,\ Q_3\text{+}1.5 * e^{(4MC)} * IQR] \text{ if } MC \geq 0 ,$$

$$[Q_1\text{-}1.5 * e^{(-4MC)} * IQR,\ Q_3\text{+}1.5 * e^{(3.5MC)} * IQR] \text{ if } MC \leq 0,$$

where L is the lower fence, and U is the upper fence of the interval. The observations which fall outside the interval are considered outliers.

The value of the *MC* ranges between -1 and 1. If *MC*=0, the data is symmetric and the Adjusted boxplot becomes Tukey's box plot.

If *MC*>0, the data has a right skewed distribution,Where as if *MC*<0, the data has a left skewed distribution.

of the intervals of two boxplot methods, Tukey's method and the adjusted boxplot, for the example data set.

The vertical dotted lines are the lower and upper bound of the interval of each method. Although the example data set is artificial and is not large enough to explain their difference, we can see a general trend that the interval of the adjusted boxplot, especially the upper fence, moves to the side of the skewed tail, compared to Tukey's method.

Inner fences of Tukey Method (Q_1-1.5*IQR, Q_3+1.5*IQR)

Outer fences of Tukey Method (Q_1-3*IQR, Q_3+3* IQR)

Single fence of adjusted box plot

$$(Q_1\text{-}1.5 * e^{(-3.5MC)} * IQR,\ Q_3\text{+}1.5 * e^{(4MC)} * IQR) \quad (3.19)$$

Van derviere and Huber (2004) computed the average percentage of outliers beyond thelower and upper fence of two types of boxplots, the adjusted Boxplot and Tukey's Boxplot, for several distributions and different sample sizes. In the simulation, less observations, especially in the right tail, are classified as outliers compared to Tukey's method when the data are skewed to the right.

In the case of a mildly right-skewed distribution, the lower fence of the interval may move to the right and more observations in the left side will be classified as outliers compared to Tukey's method. This difference mainly comes from a decrease in the lower fence and an increase in the upper fence from Q_1 and Q_3, repectively.

3.5.2.5 Adjusted Boxplot:

As a solution to the Boxplot, Vanderviere and Hubert (2008) presented an adjusted Boxplot. The difference between the former and latter Boxplot is the inner and outer fence. In the adjusted Boxplot,

the medcouple (MC) is introduced. The MC value is between -1 and 1.

If MC = 0, the data is symmetric and the adjusted Boxplot becomesTukey's Boxplot. In addition, if MC > 0, the data is right skewed; if MC < 0, the data is left skewed.

Let $X = x_1, x_2, \ldots, x_n$ be the independent sample of a continuous univariate distribution. Sort each observation inX, fromthe smallest value to the largest value, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Therefore, one can define the MC as

$$MC(x_1, x_2, \ldots, x_n) = \frac{\text{med}(x_j - \text{med}') - (\text{med}' - x_i)}{x_i - x_i,}$$

where med'= the median of x_i and j have to satisfy $x_i \leq \text{med}' \leq x_i$ and $x_i \neq x_i$. If MC ≥ 0

one can develop the fence as below

$$L\ U = [Q_1 - 1.5 * IQR * e^{-3.5MC} Q_3 + 1.5 * IQR * e^{4MC})] \text{ If MC} \geq 0$$

$$L\ U = [Q_1 - 1.5 * IQR * e^{-4MC} Q_3 + 1.5 * IQR * e^{3.5MC})] \text{ If MC} \leq 0$$

Observations situated outside the fence are labeled as outliers.

3.5.2.6 (SSSBB) Boxplot:

Split Sample Skewness Based Boxplot test and its modifications are designed on the basis of first quartile for the lower side Q_{1L}, third quartile for the lower side Q_{3L} and inter-quartile range for the lower side IQR_L. \

Similarly, first quartile for upper side Q_{1R}, third quartile for the upper side Q_{3R} and inter-quartile range for the upper side IQR_R In order to construct boundaries for labeling an observation as an outlier, 1.5 times IQR_L is subtracted from Q_{1L} for lower threshold and 1.5 times IQR_R in added to the Q_{3L} for upper threshold

the boundaries for the complete data set are as under:

Q_{1L} = 12.5th percentile, Q_{3R} = 87.5th percentile,

$IQR_L = Q_{3L} - Q_{1L}$=37.5th percentile - 12.5th percentile,

$IQR_R = Q_{3R} - Q_{1R}$ = 87.5th percentile - 62.5th percentile

Lower and upper boundaries are defined as

$$L\ U = [Q_{1L} - 1.5 * IQR_L\ Q_{3R} + 1.5 * IQR_R] \quad (3.21)$$

Where L is the lower critical value and U is the upper critical value of the data. An observation outside these boundaries $L\ U$ would be labeled as an outlier.

3.6 Multivariate methods:

Outliers become more difficult to detect in high dimensional data. One cannot claim multivariable observations as outliers if each variable is considered independently.

Another scenario that could happen in multivariate cases is the masking and swamping problem.

Recall that the masking problem occurs when the appearance of one outlier covers the appearance of another outlier, whereas the swamping problem arises when the observation is identified as an outlier even if it is not. In other words, swamping is the opposite of masking.

Instead of declaring too few outliers, the method declares more outliers than there actually are (Hawkins, Bradu and Kass, 1984).

Some of the multivariate outliers have been modified from the univariate method, so that it can take into account a multivariable. Examples are the generalized distance with studentized residual (Siotani, 1959), the ratio of generalized distance with all observations (Wilk, 1963) and the W statistics for normality (Shapiro and Wilk, 1965).

There are also examples of multivariate outlier detection method that are based on residuals.

Cook (1977) recommended using plot of residuals or examining the standardized residuals or studentized residuals.

Other suggestions of multivariate outlier detection method that are based on residuals can be found in David (1978) and Cook (1986).

3.6.1 Statistical methods:

Observations that are situated far from the centre of the data distribution is labeled as outliers in the statistical method. One of the most widely used approaches for the detection of multivariate outlier in the statistical method is called the Mahalanobis distance.

According to Stevens (1984), the Mahalanobis distance is a measure of the distance in factor space.

Let

X_i = $P \times 1$ vector of observation for the ith unit

X = matrix of the original data set with column centred by the mean

$\bar{x}$ = $P \times 1$ dimensional vector with the means of each variable

S = $\frac{1}{n-1}$ $(x^1\ x)$, covariances matrix of the p variables

Now, one can develop the Mahalanobis distance, D

$$MD(x, \bar{x}) = \{(x - \bar{x})` S^{-1} (x - \bar{x})\}^{\frac{1}{2}} \quad (3.22)$$

where D is the distance of x to the mean of the data set. For multivariate normally distributed data, the values of the Mahalanobis distance are approximately chi-square distributed with p degrees of freedom (χ^2 p).

An observation with large Mahalanobis distance can be considered as an outlier.

The Mahalanobis distance works well when identifying scattered outliers (Rocke andWoodruff, 1996). However, it fails to perform when a data set contains clustered outliers.

This is supported by Filzmoser (2004), who mentions that a single extreme observation or a group of observations far away from the main data structures can have a significant influence on the Mahalanob is distance.

They are subject both to the masking and swamping effect because both estimators, i.e. mean and covariance, are usually estimated in a non-robust manner, Robust estimators mean they are less affected by outliers.

Penny and Jolliffe (2001) explain the scenario of the masking and swamping effects if the Mahalanobis distance is used for identification of outliers. In the situation of masking effects, a value of Mahalanobis distance for outliers will decrease as the outliers will pull x and S towards themselves.In contrast, in the swamping effect, Mahalanobis distance values for non outliers might increase since outliers attract $\bar{x}$ and blow S away from themajority of observations.

3.6.2 Multivariate robust measures:

As a consequence of the Mahalanobis distance and Wilk's statistics problem in the statistical methods, many robust means and covariances have been introduced in previous studies. Examples are minimum volume ellipsoid (MVE) estimators (Rousseeuw and von Zomeren, 1990) and minimum covariance determinant (MCD) estimators by Rousseeuw and Driessen(1999).

These estimators have the desirable properties of high breakdown point and affine equivariance.

Originally, the breakdown point definition was given by Hodges (1967), where the definition is limited to a one-dimensional estimation of location.

Nevertheless, Hampel (1971) proposed a much more general formulation.

The breakdown point is a percentage of outliers which will make the estimator take on the large values. Therefore, estimators with a large breakdown point are more robust. It is noted that the highest breakdown point

value can possibly reach 50%. If the value goes beyond 50%, one cannot

decide which data are outliers and which are from the main distribution.

Another desirable property of an estimator is affine equivariance.

3.6.2.1 M-estimator

M-estimator is an early version of robust estimators, which are developed by a simple adjustment of the classical estimators. Maronna (1976) studied affinely equivariant M-estimators for covariance matrices and Campbell (1980) proposed using the Mahalanobis distance computed using the M-estimators for the mean and covariance matrix.

To compute these estimators, each observation is given a weight.

The given weight depends on the di(x_i, $\bar{x}$) values of each observation.

Observation with a high value of di(x_i, $\bar{x}$)will be down weighted.

Full weight is given to the observations with normal di(x_i, $\bar{x}$)value.

Note that the observations with the large value of di(x_i, $\bar{x}$)could be considered as outliers. Therefore by giving a reduced weight to the outlying observation in the data set, it hardly influences the estimator.

However, the M-estimator has a low breakdown point, which p+1.

It means the performance of these estimators is not consistent.

Considering the M-estimator has a low breakdown point, a different approach has been proposed to overcome the difficulty.

3.6.2.2 Minimum Volume Ellipsoide (MVE) estimator

Minimum volume ellipsoid estimators are the mean and covariance matrix of subsample size h, where h≤ n. It minimizes the volume of the covariance matrix associated with the subsample. The basic idea of the MVE is to search among all such ellipsoids for the one having the smallest value. Therefore, the main problem of MVE is to find h that produces the smallest ellipse because the number of all subsamples containing half of the data is so large that determining the subsample with the minimum volume is impractical. It is noted that h is taken to be

$h = (n + p + 1)^2$ which is the integer function.

The h value can be assumed as the minimum number of instances that must not be outlying. Otherwise, one can state this approach has a breakdown point of approximately 50%.

3.6.2.3 Minimum Covariance Determinant (MCD) estimator

The minimum covariance determinant (MCD) estimator also has a breakdown point of approximately 50%. The MCD estimator is the mean and covariance of a subsample of size h (h≤ n) that minimizes the determinant of the covariance matrix that corresponds to the subsample. As with MVE, it is impractical to consider all subsets of half of the data since it is computationally expensive.

3.6.3 Application of multivariate robust measures:

Rousseeuw and von Zomeren (1990) used theMVE estimators to develop a method for outlier detection. The method was based on the basic resampling algorithm and they named it the Robust Distance method. However, Hadi (1992) pointed out three

weaknesses of this method, particularly a problem related to the situation when the covariance matrix has a zero determinant. Therefore, he solved this weakness by presenting an idea that makes outliers appear in one subset, with the other subset highly unlikely to contain outliers.

The new approach still applies the MVE estimator, but it is easier to compute and the method is not dependent on the basic resampling algorithm.

Later, Hadi (1994) modified his idea by giving an alternative step to the existing algorithm. The findings of this approach were almost similar to the findings of the previous solution in 1992.

The minimum covariance determinant (MCD) estimator had been used by Hawkins in 1994 to develop a feasible solution algorithm (FSA) to discover outliers. This approach still uses the subset to divide a data set from outliers. The disadvantage of this method is that large number of subsets need to be constructed from a data set, especially when one has a data set with a large sample size and variables. Therefore, in order to solve this problem, Rousseeuw and Driessen (1999) suggested the fast algorithm using the MCD estimator called FAST-MCD.

They introduced two techniques, which are, the selective iteration and the nested extension. They also presented *C-Step* where the *'C'* means concentration.

The word concentration could be interpreted as their focus on *h* observations with least distances. It also could be described as the most recent chosen subset that provides a minimum determinant. The *C-Step* has four steps which are repeated until the last process fulfils the latter definition of *'C'*.Hawkins and Olive (1999) also tried to improve the *FSA* by adding acondition called *C-Condition*. However, their approach still retainedthe similar computational complexity as *FSA* since it is only reducesthe computation time for studies that use the fixed sample size, i.e. asubset with the same sample size .

Chapter 4

A proposed method for outlier detection

4.1 Introduction:

As already stated an outlier is an observation that does not conform to normal behavior, but defining a normal behavior is very challenging. Some of the difficulties that are encountered in the process are :

1- A normal region which will encompass all possible normal behavior is difficult.

2-A normal region which is defined at present may not be normal in the Future due to evolution of data.

3- in many cases of malicious behavior the hacker often disguises the Hacking as normal behavior causing difficulty in identification.

4-for those data which lie at the bordering area between normal and Outlier region represent difficulty in classification.

5- Noise in data is often confused with outliers.

Chapter(3)contains a review of these methods and numerous methods for the detection of outliers have been explored in disciplines like data mining ,machine, learning and statistics, however an important graphical technique that is often used as an aid in outlier detection is the boxplot ,it is easy to see that the boxplot produces a misleading data summary for bimodal data, since both the measures of central tendency and spread can be very far off descriptively. For example a mixture of N (0,1) data in equal proportion will be characterized as having a median of 5 and spread (IQR) of 10 .these statistics do not describe either subpopulation distribution . it is suggested that different data summaries may be useful for different data sets. Unimodality is an important assumption for boxplote .Thus a test for Unimodality , either formal or informal ,should routinely accompany these boxplot. As shown in chapter (3) several method are suggested for the detection of outliers. In this chapter we propose a new method for outlier detection. The suggested method is a modification of

tukey's method (tukey1977). in chapter (5) the new method is compared with tukey's method and three other methods based on it. To see the proposed method and these four method brief review presented below for theus.

4.2 Tukey's method:

Tukey's method for outlier detection is based on the median and the interquarticl range (IQR),defined as the difference($Q_3 - Q_1$)between

The third quartile (Q_3) and the first quartile(Q_1), the method consist of two fence :

An inner fence and an outer fence, the inner fence is defined as

$$[Q_1 - 1.5\ \text{IQR}, Q_3 + 1.5\ \text{IQR}] \quad (4.1)$$

While the outer fence is defined as

$$[Q1 - 3\ \text{IQR}, Q_3 + 3\ \text{IQR}] \quad (4.2)$$

Formally ,the lower (L)and the upper(U) boundaries for the two fence can writer as:

$$L\ U = [\ Q_1 - g * \text{IQR}\ , Q_3 + g * \text{IQR})]$$

Where

$$g \begin{cases} 1.5 \text{ for inner fence} \\ 3 \text{ for outer fence} \end{cases}$$

adjusting Tukey‟s fence at true 95% fence is possible by using the formula below:

$$L\ U = [Q1 - 0.95 * \text{IQR}\ , Q3 + 0.95 * \text{IQR}] \quad (4.3)$$

since the introduction of Tukey's method in (1977) several modification are suggested . Among these are the methods of kimber's , Hubert and Vanderviere, and Iftikhar.

4.2.1 Kimber's method

Kimber's (1990) modified Tukey's method by replacing the(IQR) by the difference (Q_2-Q_1) for the lower boundary and (Q_3-Q_2) for the upper boundary ,is the two fence as an attempt to solve the problem of showed Formally the modified form proposed by Kimber's is

$$L\ U = [Q_1 - 1.5*(Q_2 - Q_1),\ Q_3 + 1.5*(Q_3 - Q_2)] \quad (4.4)$$

where Q_2 is the sample median.

4.2.2 Hubert and Vandervieren method:

This method is based on the medcouple (MC) which range between -1 and +1 with symmetry a achieved for MC is zore

Hubert and Vandervieren (2008) proposed a technique for detection of outliers, called HV boxplot defined as

$$L\ U = [Q_1 - 1.5*IQR*e^{-3.5MC}\ Q_3 + 1.5*IQR*e^{4MC})] \text{ If MC}\geq 0 \quad (4.5)$$

$$L\ U = [Q_1 - 1.5*IQR*e^{-4MC}\ Q_3 + 1.5*IQR*e^{3.5MC})] \text{ If MC}\leq 0 \quad (4.6)$$

when the value of MC is zero the HV method coincide with Tukey's method .

4.2.3 SSSBB method

Iftikhar (2011) proposed a methods for detection of outliers, called SSSBB boxplot defined as:

$$L\ U = [Q_{1L} - 1.5*IQR_L\ ,\ Q_{3R} + 1.5*IQR_R] \quad (4.7)$$

Where L is the lower critical value and U is the upper critical value of the data. An observation outside these boundaries would be labeled as an outlier.

adjusting SSSBB fence at true 95% fence is possible by using the formula below:

$$L\ U = [\ Q_{1L} - 0.97*\ IQR_L\ ,\ Q_{3R} + 0.97*\ IQR_R] \quad (4.8)$$

4.3 the proposed method

The suggested method is based on an upper fence and a lower fence and is intended to be more conservative than the previous methods. The steps for construction of boundaries of the fence are as follows:

Step 1

The absolute values of the data are obtained and the median M of absolute values calculated.

Step 2

With $|X_{LL}|$ and $|X_{SL}|$ the largest absolute value and smallest absolute value respectively to the left of the median, Calculate the deference:

$$b = |X_{LL}| - |X_{SL}| \qquad (4.9)$$

Step 3

form the fence by adding b to $(Q_1 - 1.5(IQR))$ to obtain the lower boundary, and subtract it form$(Q_3 - 3\ IQR)$ to obtain the upper boundary. i.e the fence is given by:

$$LU(Q_1 - 1.5(IQR) + b, (Q_3 - 3IQR) - b) \qquad (4.10)$$

Where L is the lower boundary and U the upper boundary.
Any value outside this fence will be labeled as an outlier .
The above fence is found most suited for data distribution as chi-square and log normal .
For the beta distribution data the alternation form :

$$LU[(Q_3 - 1.5(IQR) - b, (Q_1 - 1.5\ IQR) + b)] \qquad (4.11)$$

is formal more suited.

4.4 methodology:

Outlier detection methods suggest a fence such that observations outside the fence would be labeled as outliers. Five percent probability of Type I error is allowed. We make the fence such that there is 5% chance of the random draw to be labeled as outlier when in fact it is not.

All points outside the central 95% fence are treated as outliers. In a distribution with no outliers, this leads to a 5% type I error probability. The main theme of this thesis is that the central 95% points are not symmetric around the median in skewed distributions. Tukey's technique is symmetric around the median and will therefore construct a fence which is too short on the right hand side and too long on the left hand side for a distribution which is skewed to the right. For any given distribution F, let

$$LCV = F^{-1}\ 2.5\%\ and$$

$$UCV = F^{-1}\ 97.5\%$$

then [LCV, UCV]

are the true upper and lower fence values of the distribution F.

Different methods will be assessed according to their ability to approach these true values. As this thesis is dealing with skewness and outliers in skewed data sets, the performance will be different on the two sides. Only distributions skewed to the right will be considered. It is important to note that this thesis is adopting the 95% fence to compare methods instead of comparing the percentage of outliers as in previous studies. This methodology has the advantage observing a 95% boundary as 95% fence is a robust measure than the extreme values True boundaries are considered at 95% central values of the distribution leaving 2.5% on each side of the distribution and fences of all methods are calculated by substituting theoretical values of the distribution in their respective formulae.

chapter 5

5.1 introduction:

The purpose of this chapter is to provide a comparison of outlier detection methods through a simulation experiment. The comparison is based on the matching the fence with the true distribution of the data. the distributions considered are chi-square ,beta and the lognormal distribution. If the distribution of the data is skewed, the classical outlier detection methods tend to treat symmetrically both sides of the data distribution. Therefore it leaves a lot of data on the long-tail side of the distribution and covers extra area on the shorter tail of the distribution. The theoretical fence is calculated by allowing 5% probability of type I error. That is 5% of the data is allowed to remain outside the fence with 2.5% on each side. The method investigated in the chapter is nonparametric methods. Non-parametric techniques make fewer assumptions; the range of applications of the non parametric techniques is therefore wider than that of parametric techniques. Another benefit of techniques is that these are often simpler than parametric techniques. The plan in remaining sections of this chapter is to compare the method for each of the three distributions. , this is done for various sample size and different parameters of the distribution. The chi-square, beta and log normal are chosen because are skewed distribution .the simulation experiment is used to enable investigating the performance of the method and its sensitivity to change in parameters and sample size can than be ascertained . A method is more efficient the closer its fence to the true 95% fence of the distribution. The simulation is executed using MATLAB Software.

5.2 The chi square distribution:

For chi square six parameters are tried namely 2, 5, 10, 15, 20, and 25. Samples of size 25,100 and 500 are taken. There are thus 18 problems. Each problem is repeated 10.000 times. For each problem the lower and upper fence of $\chi 2$ Distribution are determined. The lower and upper fences for each of the five methods are calculated by using the formula in chapter (4).

5.2.1 Small sample size :

Table (5.1) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting SSSBB, HV, Kimber's and our proposed) for 6 parameter of the $\chi 2$ Distribution when a small sample of size is used.

Table 5.1 Fences of five methods and True boundary in $\chi 2$ Distribution at Small sample size(n=25)

	skewenss	1.54	1.06	1.50	-0.45	0.76	1.04
Sample size	**parameter**	**2**	**5**	**10**	**15**	**20**	**25**
True lower fence (2.5%)		**0.05**	**0.83**	**3.25**	**6.26**	**9.59**	**13.12**
Small sample size	**Adjusting** Tukey's	**-0.55**	**-1.36**	**1.56**	**3.65**	**6.95**	**12.25**
	Adjusting SSSBB	**-0.92**	**0.64**	**0.54**	**3.64**	**5.83**	**12.45**
	HV	**0.48**	**-2.16**	**0.78**	**4.46**	**9.34**	**-1.77**
	Kimber's	**0.29**	**2.62**	**2.48**	**6.16**	**10.47**	**12.97**
	Our proposed	**-0.48**	**4.66**	**4.31**	**5.40**	**11.24**	**16.23**
True upper fence (97.5%)		**7.38**	**12.83**	**20.48**	**27.49**	**34.17**	**40.65**
Small sample size	**Adjusting** Tukey's	**3.31**	**10.39**	**15.80**	**23.35**	**30.71**	**36.56**
	Adjusting SSSBB	**10.37**	**9.88**	**24.54**	**34.57**	**35.80**	**41.25**
	HV	**19.39**	**10.77**	**21.10**	**35.31**	**50.95**	**35.79**
	Kimber's	**3.62**	**10.39**	**14.76**	**23.18**	**30.96**	**33.92**
	Our proposed	**5.23**	**11.74**	**20.40**	**31.79**	**38.71**	**45.15**

Inspection of Table (5.1) reveals the following facts:-

At parameter (2) all methods perform badly with respect to the lower fence with Kimber's method being the best.

As to the upper fence our proposed method's fence is closest to the true 95% fence, so it is the best compared to the other methods while HV method's is the worst.

At parameter(5)in the lower fence the Adjusting SSSBB method is the closest to the true 95% fence, with the other's performing poorly.

However in the upper fence our proposed method is the closest to the true 95% fence, providing the best performance while the Adjusting SSSBB method is the worst.

At parameter (10) in the lower fence Kimber's method has best performance followed by our proposed method, while the Adjusting SSSBB method is the worst.

In the upper fence our proposed method is the closest to the true 95% fence, followed by HV method, while the Adjusting SSSBB method is the worst.

At parameter (15) in the lower fence Kimber's method it is the best, followed by our proposed method. While the Adjusting Turkey's and Adjusting SSSBB methods both have the worst performance.

In the upper fence the Adjusting Turkey's is the nearest to the true boundary but can not be considered good, while the HV method is the worst.

At parameter (20) in the lower fence HV method has the best performance followed by Kimber's method, while the Adjusting SSSBB method is the worst.

In upper fence the Adjusting SSSBB is the best, while HV method is the worst performs.

At parameter (25) in the lower fence Kimber's method is the best followed by the Adjusting SSSBB method, while HV method is the worst.

But in upper fence the Adjusting SSSBB is the best, while Kimber's method i the worst.

5.2.2 Medium sample size:

Table (5.2) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting SSSBB, HV, Kimber's and our proposed) for 6 parameter of the $\chi 2$ Distribution when a medium sample of size is used.

Table 5.2 Fences of the five methods and True boundary in $\chi 2$ Distribution at medium sample size (n=100)

	Skewenss	**1.37**	**0.71**	**0.60**	**0.13**	**0.34**	**0.28**
Sample size	**parameter**	**2**	**5**	**10**	**15**	**20**	**25**
True lower fence (2.5%)		**0.05**	**0.83**	**3.25**	**6.26**	**9.59**	**13.12**
medium sample size	**Adjusting** Tukey's	-1.81	-0.50	1.12	3.60	10.18	12.35
	Adjusting SSSBB	-1.01	-0.69	0.63	0.45	6.20	6.30
	HV	-0.52	0.21	1.30	0.51	5.18	11.66
	Kimber's	-0.62	0.78	2.39	5.11	11.83	14.44
	Our proposed	-1.99	1.05	4.56	10.25	16.75	22.27
Trueupperfence(97.5%)		**7.38**	**12.83**	**20.48**	**27.49**	**34.17**	**40.65**
medium sample size	**Adjusting** Tukey's	5.03	10.05	17.70	24.49	29.74	38.85
	Adjusting SSSBB	9.77	14.31	22.51	31.17	40.23	49.77
	HV	15.95	17.46	26.27	29.38	32.30	50.21
	Kimber's	5.28	9.88	16.68	23.11	28.69	37.29
	Our proposed	8.76	13.95	22.83	28.65	33.29	42.64

From the Table (5.2) we see the following**:-**

At parameter (2) in lower fence all methods did not work as properly.

But in the upper fence our proposed method has the best performance with HV method showing the worst performance.

At parameter (5) in the lower fence Kimber's method showed the best performance followed by our proposed method, while the Adjusting SSSBB method is the worst.

But in the upper fence it is our proposed which is the best, while HV method is the worst.

At parameter (10) in the lower fence Kimber's method is the best followed by our proposed method, while the Adjusting SSSBB method is the worst.

However in the upper fence it is the Adjusting SSSBB which is best, While HV method is the worst.

At parameter (15) in the lower fence Kimber's method is the best, while the Adjusting SSSBB method is the worst.

In the upper fence it is our proposed method which is best other than methods; while HV method has the worst performance.

At parameter (20) in the lower fence the Adjusting turkey's method is the best, while our proposed method is the worst.

But in the upper fence our proposed method is the best, while the Adjusting SSSBB method is the worst.

At parameter (25) in the lower fence the Adjusting Tukey's method is the best, while our proposed method is the worst.

But in the upper fence the Adjusting turkey's method is the best, while HV method is the worst.

5.2.3 Large Sample Size:

Table (5.3) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting SSSBB, HV,

Kimber's and our proposed) for 6 parameter of the $\chi 2$ Distribution when a large sample of size is used

Table 5.3 Fences of the five methods and True boundary in $\chi 2$ Distribution at large sample size (n=500)

	Skewenss	**1.96**	**1.27**	**0.79**	**0.72**	**0.91**	**0.55**
Sample size	**Parameter**	**2**	**5**	**10**	**15**	**20**	**25**
True lower fence (2.5%)		**0.05**	**0.83**	**3.25**	**6.26**	**9.59**	**13.12**
large sample size	**Adjusting** Tukey's	**-1.43**	**-0.92**	1.28	4.29	8.09	11.67
	Adjusting SSSBB	**-1.39**	**-2.22**	-1.43	-1.56	2.29	4.15
	HV	**-0.67**	**-0.89**	1.24	3.78	7.17	8.09
	Kimber's	**-0.58**	**0.15**	2.91	6.44	9.78	13.77
	Our proposed	**-1.25**	**1.19**	5.78	12.37	15.61	22.39
True upper fence (97.5%)		**7.38**	**12.83**	**20.48**	**27.49**	**34.17**	**40.65**
large sample size	**Adjusting** Tukey's	**4.70**	**10.12**	18.05	25.13	29.75	37.90
	Adjusting SSSBB	**13.19**	**23.03**	31.30	37.86	53.09	54.11
	HV	**10.70**	**15.59**	26.14	34.12	38.32	44.42
	Kimber's	**4.70**	**9.67**	17.36	24.40	28.45	36.38
	Our proposed	**7.70**	**13.27**	22.24	27.84	33.43	40.75

From the Table (5.3) show the follow**:-**

At parameter (2) in the lower fence all methods did not work as perform properly.

But in the upper fence our proposed method is the best, While the Adjusting SSSBB method is the worst.

At parameter (5) in the lower fence our proposed method is the best, While the Adjusting SSSBB method is the worst.

However in the upper fence it is our proposed method which is best, while the Adjusting SSSBB method is the worst.

At parameter (10) in the lower fence Kimber's method is the best, while the Adjusting SSSBB method is the worst.

But in the upper fence it is our proposed method which is best, while the Adjusting SSSBB method is the worst.

At parameter (15) in the lower fence Kimber's method is the best, while the Adjusting SSSBB method is the worst.

But in the upper fence is our proposed method which is best, while the Adjusting SSSBB method is the worst.

At parameter (20) in the lower fence Kimber's method is the best, while the Adjusting SSSBB method is the worst.

But in the upper fence it is our proposed method which is best, while the Adjusting SSSBB method is the worst.

At parameter (25) in the lower fence Kimber's method is the best, while our proposed method is the worst.

But in the upper fence it is our proposed method which is best, while the Adjusting SSSBB method is the worst.

Form the Tables (5.1), (5.2) and (5.3) we conclude the following concerning the proposed method:-

{1} Small sample size (Table 5.1)

A) At the lower fence the proposed method failed to be the best in nay parameter although it performed well in parameter (5) and (10) and is the worst in parameter (20) and (25).

B) At the upper fence it performed best in parameters (2), (5) and (10) and reasonably well in parameter(15).

{2} Medium sample size (Table 5.2)

A) At the lower fence the proposed method failed to be superior in any parameter, but it performed relatively well in parameter (10) and is the worst in parameter (20) and (25).

B) At upper fence it showed the best performance in parameters (2), (5), (15) and (20).

{3} Large sample size (Table5.3)

A) At the lower fence it is the best in parameter (5), and the worst in parameter (25).

B) At the upper fence it is the best in parameters (2),(5),(10),(15), (20)and(25)be in all parameters .

5.3 The lognormal Distribution:

The log normal distribution used in the simulation are, *lnN*(0, 0.2), *lnN* (0, 0.4), *lnN* (0, 0.6), *lnN* (0, 0.8),*lnN* (0, 1) with sample sizes of 25, 100 and 500 ,this yielded 15 problem in all. 10.000 replicates are performed for each problem; the lower and upper fences for each of the five methods are calculated by using the formula in chapter (4).

5.3.1 Small sample size:

Table (5.4) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting SSSBB, HV, Kimber's and our proposed) for 5 parameter of the lognormal Distribution when a small sample of size is used.

Table 5.4 Fences of the five methods and True boundary in lognormal Distribution at small sample size (n=25)

	Skewenss	**-0.35**	**0.07**	**1.03**	**2.43**	**1.81**
Sample size	**parameter**	**0,0.2**	**0,0.4**	**0,0.6**	**0,0.8**	**0,1.0**
True lower fence (2.5%)		**0.68**	**0.46**	**0.31**	**0.21**	**0.14**
Small sample size	**Adjusting** Tukey's	**0.61**	**0.08**	**0.11**	**0.69**	**0.54**
	Adjusting SSSBB	**0.73**	**0.30**	0.30	-0.04	-0.34
	HV	**0.67**	**-0.26**	0.51	0.58	-0.09
	Kimber's	**0.67**	**0.21**	0.42	-0.31	-0.08
	Our proposed	**0.75**	**0.36**	0.38	-0.61	-0.34
True upper fence (97.5%)		**1.48**	**2.19**	**3.24**	**4.80**	**7.10**
Small sample size	**Adjusting** Tukey's	**1.27**	**1.86**	1.96	2.78	2.69
	Adjusting SSSBB	**1.39**	**1.48**	3.06	2.77	2.53
	HV	**1.79**	**2.20**	5.90	4.78	6.22
	Kimber's	**1.29**	**1.74**	2.02	2.68	2.70
	Our proposed	**1.48**	**2.51**	2.64	4.49	4.16

Inspection of the table (5.4) leads to the following remarks:-

At parameter (0, 0.2) in the lower fence all methods performed equally well HV method and Kimber's method showing the best performance

But in the upper fence it is the proposed method which is best with the fence very close to the true are ,while HV method is the worst.

At parameter (0, 0.4) in the lower fence the proposed method is the best, while HV method is the worst.

But in the upper fence it is HV method that is best, while Adjusting SSSBB method is the worst.

At parameter (0, 0.6) in the lower fence Adjusting SSSBB method is the best, while Adjusting Tukey's method is the worst.

But in the upper fence it is the Adjusting SSSBB method which is best, while HV method is the worst.

At parameter (0, 0.8) in the lower fence HV method is the best, while the proposed method is the worst.

But in the upper fence it is the Adjusting SSSBB method which is best, while HV method is the worst.

At parameter (0, 1.0) in the lower fence the Adjusting Tukey's method is the best, while the proposed method and the Adjusting SSSBB method both the worst.

But in the upper fence HV method is the best, while the Adjusting SSSBB method is the worst.

5.3.2 Medium sample size:

Table (5.5) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting SSSBB, HV, Kimber's and our proposed) for 5 parameter of the lognormal Distribution when a medium sample of size is used.

Table 5.5 Fences of the five methods and True boundary in lognormal Distribution at medium sample size (n=100)

	Skewenss	**1.10**	**1.68**	**1.44**	**2.66**	**1.96**
Sample size	**parameter**	**0,0.2**	**0,0.4**	**0,0.6**	**0,0.8**	**0,1.0**
True lower fence (2.5%)		**0.68**	**0.46**	**0.31**	**0.21**	**0.14**
medium sample size	**Adjusting** Tukey's	**0.61**	**0.26**	**-0.12**	**-0.49**	**-0.86**
	Adjusting SSSBB	**0.65**	**0.36**	0.18	-0.04	-0.19
	HV	**0.49**	**0.37**	-0.06	0.04	0.19
	Kimber's	**0.67**	**0.44**	0.13	0.002	0.03
	Our proposed	**0.91**	**0.55**	0.23	-0.31	-0.82
True upper fence (97.5%)		**1.48**	**2.19**	**3.24**	**4.80**	**7.10**
medium sample size	**Adjusting** Tukey's	**1.39**	**1.86**	2.24	2.90	3.68
	Adjusting SSSBB	**1.34**	**2.02**	2.30	3.70	7.11
	HV	**1.57**	**2.98**	3.56	7.17	14.91
	Kimber's	**1.34**	**1.82**	2.16	2.92	3.94
	Our proposed	**1.50**	**2.40**	3.10	4.47	6.00

Table (5.5) reveals the following:-

At parameter (0, 0.2) in the lower fence Kimber's method is the best, while the proposed method is the worst.

But in the upper fence it is the proposed method which is best and it fence is very close to the true 95% fence, while HV method is the worst.

At parameter (0, 0.4) in the lower fence Kimber's method is the best, while the Adjusting Tukey's method is the worst.

But in the upper fence it is the Adjusting SSSBB method which is best, while the HV is the worst.

At parameter (0, 0.6) in the lower fence the proposed method is the best, while the Adjusting Tukey's method is the worst.

But in the upper fence the proposed method is the best, while the Adjusting SSSBB method is the worst.

At parameter (0, 0.8) in the lower fence HV method is the best, while the Adjusting Tukey's method is the worst performs.

But in the upper fence the proposed method is the best, while HV method is the worst.

At parameter (0, 1.0) in the lower fence HV method is the best, while the Adjusting Tukey's ey method and the proposed method both are the worst.

But in the upper fence the Adjusting SSSBB method is the best of the other methods, while HV method is the worst.

5.3.3 Large sample size:

Table (5.6) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting SSSBB, HV, Kimber and our proposed) for 5 parameter of the lognormal Distribution when a large sample of size is used.

Table 5.6 Fences of the five methods and True boundary in lognormal Distribution at large sample size (n=500)

	Skewenss	**0.82**	**1.72**	**2.06**	**5.16**	**3.36**
Sample size	**parameter**	**0,0.2**	**0,0.4**	**0,0.6**	**0,0.8**	**0,1.0**
True lower fence (2.5%)		**0.68**	**0.46**	**0.31**	**0.21**	**0.14**
Large sample size	**Adjusting** Tukey's	**0.62**	**0.27**	**-0.09**	**-0.49**	**-0.95**
	Adjusting SSSBB	**0.65**	**0.45**	0.17	0.10	-0.09
	HV	**0.58**	**0.23**	0.11	0.15**-**	0.33-
	Kimber's	**0.68**	**0.42**	0.16	-0.07	-0.33
	Our proposed	**0.93**	**0.66**	0.30	-0.17	-0.78
True upper fence (97.5%)		**1.48**	**2.19**	**3.24**	**4.80**	**7.10**
large sample size	**Adjusting** tukey's	**1.38**	**1.87**	2.19	2.94	3.56
	Adjusting SSSBB	**1.39**	**2.16**	2.51	4.18	4.73
	HV	**1.66**	**2.55**	**4.00**	5.82	8.45
	Kimber's	**1.34**	**1.80**	2.13	2.98	3.56
	Our proposed	**1.46**	**2.31**	2.98	4.40	5.73

From Table (5.5) we see the following:-

At parameter (0, 0.2) in the lower fence Kimber's method is the best with fence very close to the true 95% fence, while the proposed method is the worst.

But in the upper fence it is our proposed method which is best with fence very close to the true 95% fence, while HV method is the worst.

At parameter(0,0.4) in the lower fence the Adjusting SSSBB method is the best with fence very close to the true 95% fence, while HV method is the worst.

But in the upper fence the Adjusting SSSBB method is the best, while Kimber's method is the worst.

At parameter (0,0.6) in the lower fence the proposed method is the best with fence very close to true 95% fence, while the Adjusting tukey's method is the worst .

But in the upper fence the proposed method is the best, while kimber method is the worst.

At parameter (0, 0.8) in the lower fence the Adjusting SSSBB method is the best, while the Adjusting Tukey's method is the worst.

But in the upper fence our proposed method is the best, while the Adjusting Tukey's method is the worst.

At parameter (0,1.0) in the lower fence the Adjusting SSSBB method is the best, while the Adjusting Tukey's key method is the worst .

But in the upper fence the proposed method is best, while HV method is the worst.

Form tables (5.4), (5.5) and (5.6) we arrive at the following conclusion the proposed method:-

{1} Small sample size form Table (5.4

A- At the lower fence it is not the best or the worst in any parameter.

B-At the upper fence is the best in the parameter (0, 0.2), in the other parameters its performance is not good but not the worst.

{2} Medium sample size form Table (5.5)

A-At the lower fence it is the best in the parameter (0, 0.6), but the worst in the parameter (0, 0.2) and (0, 1.0).

B- At the upper fence it is the best in the parameters (0, 0.2),(0,0.6) and(0,0.8), while in the other parameter it performed well.

{3} Large sample size form Table (5.6)

A- At the lower fence it is the worst in (0, 0.2) and is not the best in any parameter.

B-At the upper fence it is the best in (0,0.2),(0,0.6) (0,0.8) and(0,1.0),and performed very good in the parameter(0,0.4).

5.4 Beta Distribution:

Beta distribution with the following values for the parameters α and β are used (35, 1), (35, 2), (35, 3), (35, 4), (35, 5), with sample sizes of 25, 100 and 500, the 15 problem are repeated 10,000 time, The lower and upper fences for each of the five methods are calculated by using the formula in chapter (4).

5.4.1 small sample size :

Table (5.7) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting SSSBB, HV, Kimber's and our proposed) for 5 parameter of the beta Distribution when a small sample of size is used.

Table 5.7 Fences of the five methods and True boundary in β Distribution at Small sample size (n=25)

	skewness	**-1.30**	**-1.38**	**-0.91**	**-1.87**	**-0.77**
Sample size	**parameter**	**35,1**	**35,2**	**35,3**	**35,4**	**35,5**
True lower fence (2.5%)		**0.90**	**0.85**	**0.82**	**0.79**	**0.76**
Small sample size	**Adjusting Tukey's**	**0.93**	**0.87**	**0.85**	**0.87**	**0.76**
	Adjusting SSSBB	**0.88**	**0.78**	0.77	0.72	0.74
	HV	**0.88**	**0.88**	0.84	0.87	0.60
	Kimber's	**0.94**	**0.88**	0.86	0.88	0.76
	Our proposed	**0.96**	**0.89**	0.91	0.83	0.92
True upper fence (97.5%)		**1.00**	**0.99**	**0.98**	**0.97**	**0.96**
Small sample size	**Adjusting Tukey's**	**1.02**	**1.01**	1.01	0.98	1.00
	Adjusting SSSBB	**1.01**	**1.02**	1.01	0.99	0.97
	HV	**1.00**	**1.09**	**1.07**	1.06	0.98
	Kimber's	**1.02**	**1.00**	1.00	0.98	0.97
	Our proposed	**1.00**	**0.98**	0.95	1.01	0.83

From Table (5.7) we not the following:-

At parameter (35,1) in the lower fence HV and the Adjusting SSSBB method both are the best, while the proposed method is the worst.

But in the upper fence it is our proposed method which is best , while Kimber's and the Adjusting tukey's method both are the worst.

At parameter (35, 2) in the lower fence the Adjusting tukey's method is the best, while the Adjusting SSSBB method is the worst.

But in the upper fence our proposed method is the best, while HV method is the worst.

At parameter (35, 3) in the lower fence HV method is the best, while our proposed method is the worst.

But in the upper fence our proposed method is the best, while HV method is the worst.

At parameter (35,4) in the lower fence our proposed method is the best, while Kimber's method is the worst .

But in the upper fence kimber and the Adjusting tukey's method both are the best, while HV method is the worst.

At parameter(35,5) in the lower fence the Adjusting Tukey's and Kimber's method both are the best, while HV and our proposed method both are the worst .

But in the upper fence Kimber's and the Adjusting SSSBB both are the best, while our proposed method is the worst.

5.4.2 Medium sample size:

Table (5.8) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting SSSBB, HV, Kimber's and our proposed) for 5 parameter of the beta Distribution when a medium sample of size is used.

Table 5.8 Fences of the five methods and True boundary in β Distribution at medium sample size (n=100)

	skewness	-1.69	-0.86	-0.83	-0.75	-0.57
Sample size	**parameter**	**35,1**	**35,2**	**35,3**	**35,4**	**35,5**
True lower fence(2.5%)		**0.90**	**0.85**	**0.82**	**0.79**	**0.76**
Medium sample size	**Adjusting Tukey's**	**0.94**	**0.88**	**0.85**	**0.80**	**0.80**
	Adjusting SSSBB	**0.86**	**0.80**	0.76	0.73	0.70
	HV	**0.82**	**0.80**	0.81	0.70	0.79
	Kimber's	**0.94**	**0.88**	0.86	0.81	0.82
	Our proposed	**0.93**	**0.91**	0.88	0.83	0.85
True upper fence (97.5%)		**1.00**	**0.99**	**0.98**	**0.97**	**0.96**
Medium sample size	**Adjusting Tukey's**	**1.02**	**1.01**	0.99	0.98	0.96
	Adjusting SSSBB	**1.01**	**1.01**	1.01	1.00	1.00
	HV	**1.01**	**1.01**	**1.01**	0.98	1.02
	Kimber's	**1.01**	**1.00**	0.98	0.96	0.96
	Our proposed	**1.03**	**0.98**	0.97	0.92	0.91

From table (5.8) we deduce the following:-

At parameter (35, 1) in the lower fence our proposed method is the best, while HV method is the worst.

But in the upper fence it is the Adjusting SSSBB, HV and Kimber's methods that are the best, while our proposed method is the worst.

At parameter (35, 2) in the lower fence the Adjusting Tukey's and Kimber's method both are the best s, while our proposed method is the worst.

But in the upper fence our proposed method is the best, while HV, the Adjusting SSSBB and the Adjusting Tukey's method are the worst.

At parameter (35,3) in the lower fence HV method is best, while our proposed and the Adjusting SSSBB method are both the worst .

But in the upper fence Kimber's method is the best, while HV and the Adjusting SSSBB method are both the worst.

At parameter (35,4) in the lower fence the Adjusting Tukey's method is the best while HV method is the worst performs.

But in upper fence Kimber's method is the best, while our proposed method is the worst.

At parameter (35, 5) in the lower fence HV method is the best, while our proposed method t is the worst.

But in the upper fence Kimber's and the Adjusting Tukey's are both the best, while HV method is the worst.

5.4.3 Large sample size :

Table (5.9) shows the true lower and upper fences and the calculated fences by five methods (Adjusting Tukey's, Adjusting

SSSBB, HV, Kimber's and our proposed) for 5 parameter of the beta Distribution when a small large of size is used.

Table 5.9 Fences of the five methods and True boundary in β Distribution at large sample size(n=500)

	skewness	**-1.77**	**-1.36**	**-0.95**	**-1.07**	**-0.80**
Sample size	**parameter**	**35,1**	**35,2**	**35,3**	**35,4**	**35,5**
True lower fence (2.5%)		**0.90**	**0.85**	**0.82**	**0.79**	**0.76**
Large sample size	**Adjusting** Tukey's	**0.94**	**0.89**	**0.85**	**0.80**	**0.79**
	Adjusting SSSBB	**0.82**	**0.75**	0.71	0.60	0.64
	HV	**0.84**	**0.81**	0.80	0.70	0.68
	Kimber's	**0.94**	**0.89**	0.86	0.81	0.80
	Our proposed	**0.90**	**0.86**	0.84	0.76	0.80
True upper fence (97.5%)		**1.00**	**0.99**	**0.98**	**0.97**	**0.96**
Large sample size	**Adjusting** Tukey's	**1.02**	**1.01**	**1.00**	0.99	0.98
	Adjusting SSSBB	**1.02**	**1.02**	1.02	1.03	1.00
	HV	**1.01**	**1.01**	**1.01**	0.98	0.97
	Kimber's	**1.01**	**1.00**	0.98	0.96	0.96
	Our proposed	**1.06**	**1.03**	1.01	1.03	0.96

From table (5.9) we absence the following:-

At parameter (35,1) in the lower fence our proposed method is the best, while the Adjusting SSSBB method is the worst.

But in the upper fence the Adjusting SSSBB and Kimber's methods are both better than the other methods, while our proposed method is the worst.

At parameter (35, 2) in the lower fence our proposed method is the best, while the Adjusting SSSBB method is the worst.

But in the upper fence Kimber's method is the best, while our proposed method is the worst.

At parameter (35, 3) in the lower fence HV method is the best, while the Adjusting SSSBB method is the worst .

But in the upper fence Kimber's method is the best, while the Adjusting SSSBB method is the worst.

At parameter (35,4) in the lower fence the Adjusting Tukey's method is the best, while the Adjusting SSSBB method is the worst .

But in the upper fence Kimber's method is the best, while our proposed and HV method are the worst.

At parameter (35, 5) in the lower fence the Adjusting tukey's method is the best, while the Adjusting SSSBB method is the worst.

But in the upper fence Kimber's and our proposed are both the best, while the Adjusting SSSBB method is the worst.

Form tables (5.7), (5.8) and (5.9) we concluded the following:-

{1} Small sample size (Table 5.7)

A- At lower fence the proposed method is better in the parameter (35, 4) and is the worst in the parameter (35, 5) and work-well in the other parameters.

B-At upper fence the proposed method is better in the parameter (35, 1) and (35, 2) and is worst in the parameter (35, 5) and work-well in the other parameters.

{2} Medium sample size (Table 5.8)

A- At the lower fence the proposed method is better in the parameter (35, 1) and is the worst in the

parameters (35,2),(35,3)and(35,5) and well-work in the

parameter (35,4).

B- At upper fence the proposed method is better in the parameter (35, 2) and is the worst in the parameter (35, 1) and (35,4)and well-work in the other parameters .

{3}- Large sample size (Table 5.9)

A- At lower fence the proposed method is better in the parameter (35, 1) and (35, 2), and well-work in the other parameters.

B- At upper fence the proposed method is better in the parameter(35,5) and is the worst in the parameter(35,1) ,(35,2)and(35,4),and well-work in the parameter(35,3).

Chapter 6

Conclusion

6.1 Main results:

This thesis considered the problem of outlier detection in univariate data with skewed distribution . Areviwe is provided for parametric and nonparametric methods of outlier detection. Interest in the thesis is focused on nonparametric methods and in particular the methods Adjuseting tukey's , Adjuseting SSSBB , Hubert and Vandervieren and Kimber's.

A new more conservative method is suggested. The performance of the new method relative to the four above mentioned methods is then studied under different distributional assumption and various sample sizes.

The distributions considered are the chi -square, log normal and beta distributions. Investigation of efficiency and sensitivinty of the methods is achieved through a suimulation experiment executed in MATLAB Software .

When the data is distributed as chi- square, the new method gave the best Performance (for all parameters considered) for large sample size with respect to the upper fence . For medium sample size the performance is the best for four parameters ,while for small sample size it is the best for three parameters . The

performance with respect to the lower fence ranges between good and worst.

For log -normally distributed data, the performance of the new method is in general the same as that for the chi -square distribution with slight improvement in the case of chi- square.

As for the beta distribution , the new method showed a relatively less efficiency with respect to the upper fence compared to chi -square and Lognormal distribution.

The general conclusion arrived at form the above discussion, is that the proposed method can be recommended for data with shewed distribution (right skewness) particulary when the sample size is large.

6.2 Recommendation for future work:

More work is needed to generally the suggested method to the multivariate case. Also more research is needed to consider application of the methods to regression and time series datas.

Bibliography

Acuna, E. and Rodriguez, C. (2004). Ameta analysis study of outlier detection methods in classification, *Technical report*, Department of Mathematics,University of Puerto Rico, In proceedings IPSI 2004, Venice.

Atkinson, A. C. (1981). Robustness, transformations and two graphical displays for outlying and influential observations in regression, *Biometrika* **68**: 13–20.

Banerjee, S., & Iglewicz, B. (2007). A Simple Univariate Outlier Identification Procedure Designed for Large Samples. Communications in Statistics- Simulation and Computation , 36, 249-263.

Barnett, V., & Lewis, T. (1994). Outliers in statistical data (3rd ed.) Wiley.

Beckman, R. J. and Cook, R. D. (1983). Outlier...s. *Technometrics 25,* 2, 119-149.

Ben-Gal I., (2005), Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

Brys. *G*, M. Hubert, and A. Struyf.(2003). A comparison of some new measuresof skewness. In R. Dutter, P. Filzmoser, U. Gather, and P.J. Rousseeuw,editors, Developments in Robust Statistics: International Conferenceon Robust Statistics 2001, volume 114, pages 98–113. Physika Verlag, Heidelberg.

Brys. G., M. H. (2004). A Robust Measure of Skewness. Journal of Computational and Graphical Statistics , 13 (4), 996-1017.

Brys. *G*, M. Hubert, and P.J. Rousseeuw(2005). A robustification of Independent ComponentAnalysis.Journal of Chemometrics, 19:364–375.

Brys, G., Hubert, M., and Struyf, A. (2006), "Robust measures of tail weight," *ComputationalStatistics and Data Analysis*, 50, 733–759.

Chandola, V., &A. Banerjee ,A., and Kumar, V. (2009). Anomaly Detection: ASurvey. *ACM Computing Surveys*, 41(3).

Carling, K. (2000). Resistant outlier rules and the non-Gaussian case Computational Statistics and Data Analysis , 33, 249-258.

Carter, N. J., Schwertman, N. C., & Kiser, T. L. (2009). A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry. Statistical Methodology , 6, 604-621.

Chatterjee, S., & Hadi, A. S. (2006). Regression Analysis by Example (Fourth Edition ed.). Hoboken, New Jersey: John Wiley and Sons, Inc.

Chen, Y., Miao, D., & Zhang, H. (2010). Neighbourhood Outlier Detection. Expert Systems with Applications , 37 (12).

Cousineau, D., & Chartier, S. (2010). Outlier Detection and Treatment; a review. International Journal of Pscyclogical Research , 3 (1), 59-68.

DAN,E, &AGATHA,O. *(2013)*.STATISTICAL ANALYSIS/METHODS OF DETECTING OUTLIERS IN AUNIVARIATE DATA IN A REGRESSION ANALYSIS MODELD**.** *International Journal of Education and Research,Vol. 1 No***5**

Davies, L. G. (1993). The identification of multiple outliers. Journal of the American Statistical Association , 88 (423), 782-792.

Dovoedo, Y. H., and Chakraborti, S. (2010), "A Modified Adjusted Boxplot for Skewed Distributions," American Statistical Association Proceedings of the Statistical graphics section

Efstathiou, C. E. (2006). Estimation of type I error probability from experimental Dixon's "Q" parameter on testing for outliers within small size data sets. Talanta , 69, 1068-1071.

Franklin, S., Thomas, S. and Brodeur, M. (2000). Robust multivariate outlier detetection using mahalanobis distance and modified staheldonoho estimators, *Proceeding International Conference on Establish-ment Surveys*, New York, pp. 697–706.

Frigge, M., Hoaglin, D. C., and Iglewicz, B. (1989), "Some Implementations of the Boxplot,"*Journal of the American Statistical Association*, 43, 50–54.

Fuller, W. A. (1987). Measurement Error Models. United States of America: Braun-Brumfield, lnc.

Groeneveld, R. A., & Meeden, G. (1984). Measuring Skewness and Kurtosis. Journal of the Royal Statistical Society , 33 (4), 391-399.

Grubbs F . E (1950) . (sample criteria for testing outlying observations , Ann moth , statist , 27-28

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. Technometrics , 11 (1), 1-21.

Hadi, A. S. (1992) "Identifying multiple outliers in multivariate data," Journal of the Royal Statistical Society. Series B, 54, 761-771.

Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the Identification of Multiple Outliers in Linear Models. Journal of the American Statistical Association , 88 (424), 1264-1272.

Hadi, A. S., Rahmatullah Imon, A. H. M., and Werner, M. (2009), "Detection of Outliers," WiresComputational Statistics, 1, 57-70.

Hawkins, D. M. (1980). Identification of Outliers. London: Chapman and Hall.

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics 26,* 3 (August), 197-208.

Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. University of York, Department of Computer Science. Kluwer Academic Publishers.

Hoaglin, D. C., and Iglewicz, B. (1987), "Fine Tuning Some Resistant Rules for OutlierLabeling," *Journal of the American Statistical Association,* 82, 1147–1149

Hubert, M., & Veeken, S. V. (2007). Outlier detection for skewed data. Katholieke Universiteit Leuven, DEPARTMENT OF MATHEMATICS. Technical Report

Hubert, M., and Van der veeken, S., (2008), "Outlier Detection for Skewed Data," Chemometrics, 22, 235-246.

Hubert, M., & Vandervieren, E. (2008). An Adjusted Boxplot for Skewed Distributions. Computational Statistics and Data Analysis , 52 , 5186–5201.

Hubert .M, P.J. Rousseeuw, and T. Verdonck. (2009).Robust PCA for skeweddata.Computational Statistics and Data Analysis, 53:2264–2274.

Hubert, M., Van der Veeken, S. (2010). Robust classification for skewed data, Advances in Data Analysis and Classification and Classification. in press.DOI 10.1007/s1 1634-010-0066-3.

Hubert, M., Van der Veeken, S. (2010). Fast and robust classifiersa djusted for skewness, Proceedings in Computational Statistic sedited by Y. Lechevallier and G. Saporta, Spring Springer-Verlag, Heidelberg,pp. 1135-1142, ISBN 978-3-7908-2603-6.

Hubert .M, G. Dierckx, and D. Vanpaemel. (2010., Detecting influential datapoints in pareto-type distributions.submitted

Hu, T. and Sung, S. Y. (2003). Detecting pattern-based outliers, *Pattern Recognition Letters* **24**: 3059–3068.

Iftikhar Hussain , Adil (2012) *Robust Outlier Detection Techniques For Skewed Distributions And Applications To Real Data.* PhD thesis, International Islamic University, Islamabad

Iglewics, B. and Martinez, J. (1982). Outlier detection using robust measures of scale, *Journal of Sattistical Computation and Simulation* **15**: 285–293.

Iglewicz, B., & Hoaglin, D. C. (1993). How to Detect and Handle Outliers. 16, Wisconsin: ASQC Quality Press.

Kafadar, K. (2003). John Tukey and Robustness. Statistical Science , 18 (3), 319-331.

Kimber, A. C. (1990). Exploratory Data Analysis for Possibly Censored Data From Skewed Distributions. Applied Statistics , 39 (1), 21-30.

Liu H., Shah S., Jiang W. (2004)"On-line outlier detection and data cleaning," Com-puters and Chemical Engineering, 28, 1635–1647.

Manoj ,K, Senthamarai ,K.K(2013) . Comparison of methods for detecting outliers. International Journal of Scientific & Engineering Research, Vol 4, Issue 9, September-2013 ISSN 2229-5518 http://www.ijser.org

Mahalanobis, P. C. (1930). On tests and measures of group divergence, *Journal of the asiatic society of bengal* **26**: 541–588.

Mansur, M. O., & Sap, M. N. (2005). Outlier Detection Technique in Data Mining. Postgraduate Annual Research Seminar.

Nazrina Aziz.(2010) ***((Analysis and Diagnostics for Censored*** **Regression and Multivariate data)) PHD** thesis. Victoria University of Wellington.

Olewuezi, N.P. (2011).Note on the Comparison of Some Outlier Labeling. Techniques Journal of Mathematics and Statistics 7 (4): 353-355

Osborne, J.W. (2002). Notes on the use of Data Transformation. Practical Assessment, Research and Evaluation,8, Available online at http://ericae.net|pare|getvn.asp?v=8&n=6.

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

Papadimitriou, S., Kitawaga, H., Gibbons, P. G. and Faloutsos, C. (2003).Loci: Fast outlier detection using the local correlation integral, *Proceedings of the 19th International Conference on Data Engineering* pp. 315–328.

Penny K. I., Jolliffe I. T. (2001) "A comparison of multivariate outlier detection methods for clinical laboratory safety data," The Statistician 50(3), 295-308.

Pragyan, P.D., &Maya N(2013).**Outlier Detection Methods--- An Analysis**. International Journal of Engineering Research & Technology (IJERT),Vol. 2 Issue 9.

Rand ,R. W.& H. J. Keselman(2003b). Modern Robust Data Analysis Methods: Measures ofCentral Tendency. American Psychological Association, Vol. 8, No. 3, 254–27.

Rocke D. M., and Woodruff D. L. (1996), "Identification of Outliers in Multivariate Data,"*Journal of the American Statistical Association*, 91, 1047–1061.

Rousseeuw, P.J. and Leroy, A.M. (1987), Robust Regression and Outlier Detection, Wiley-Interscience, New York (Series in Applied Probability and Statistics), 329 pages.

Rousseeuw, P. J. and von Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Associ- ation* **85**(411): 633–639

Rousseeuw P.J., I. Ruts, and J.W. Tukey. The bagplot: A bivariate boxplot. The American Statistician , 53:382–387,

Saad, M. K., & Hewahi, N. M. (2009). A Comparative Study of Outlier Mining and Class Outlier Mining. Computer Science Letters , 1 (1).

Schwertman, N. C., & Silva, R. d. (2007). Identifying outliers with sequential fences. Computational Statistics and Data Analysis , 51, 3800-3810.

Seo, S. (2006). A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. University of Pittsburgh, Graduate School of Public Health.

Sim, C. H., Gan, F. F., and Chang, T. C. (2005), "Outlier Labeling With Boxplot Procedures," Journal of the American Statistical Association, 100, 642-652.

Stuart, A., & Ord, J. K. (1994). Kendall's advanced theory of statistics. Distribution Theory (6th ed., Vol. 1). London.

Tabor, J. (2010). Investigating the Investigative Task: Testing for Skewness;An Investigation of Different Test Statistics and their Power to Detect Skewness. Journal of Statistics Education , 18 (2).

Tajuddin, I. H. (1999). A comparison between two simple measures of skewness. Journal of Applied Statistics , 26 (6), 767-774.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesely.

zhang N.F. (1998)"A Statistical Control Chart for Stationary Process Data," Technometrics, 40 (1), 24–38.

Zhang, Y., Meratnia, N., Havinga, P. J. M., (2007)"A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets", Technical Report, University of Twente

Zaman, A., Rousseeuw, P. J., & Orhan, M. (2001). Econometric applications of high-breakdown robust regression techniques. Economics Letters , 71, 1–8.

Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. Journal of Experimental Education , 64 (1), 71-78.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. Journal of General Psychology , 121 (4), 391-401.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. Journal of Experimental Education , 67 (1), 55-68.

Williams G. J., Baxter R. A., He H. X., Hawkins S., Gu L., (2002)"A Compar-ative Study of RNN for Outlier Detection in Data Mining," IEEE Inter-national Conference on Data-mining (ICDM'02), Maebashi City, Japan,CSIRO Technical Report CMIS-02/102.

APPENDIX

MTALB PROGRAMS USED IN SIMULATIONS

PROGRAMS 1

```
function  [out1 out2  out3 out4]=outlir(e)
m1=median(e);
t=0;
for i=1:length(e)
if e(i)<m1
t=t+1;
d1(t)=e(i);
end
end
tt=1;
for i=length(d1)+1:length(e)
if e(i)>m1
tt+1;
d2(tt)=e(i);
end
end

absd1=abs(max(d1)-min(d1));
absd2=abs(max(d2)-min(d2));
k1=absd2-absd1;
% k2=absd2+absd1;
iqr=quantile(e,0.75)-quantile(e,0.25);
if1=quantile(e,0.25)-1.5*iqr;
www1=if1+absd1;
of1=quantile(e,0.25)-3*iqr;
www2=of1+absd1;
if2=quantile(e,0.75)+1.5*iqr;
ww1=if2+absd1;
ww2=if2+absd1;
out1=ww1;
out2=ww2;
out3=www1;
out4=www2;
```

PROGRAMS 2

```
function  [out1 out2 out3 out4 out5 out6]=outlir(e)
                                        m1=median(e);
                                              t=0;
                                    for i=1:length(e)
                                        if e(i)<m1
                                            t=t+1;
                                        d1(t)=e(i);
                                            end
                                            end
                                            tt=1;
                                for i=length(d1)+1:length(e)
                                        if e(i)>m1
                                            tt+1;
                                        d2(tt)=e(i);
                                            end
                                            end

                                absd1=abs(max(d1)-min(d1));
                                absd2=abs(max(d2)-min(d2));
                                        k1=absd2-absd1;
                                    % k2=absd2+absd1;
                            iqr=quantile(e,0.75)-quantile(e,0.25);
                                    if1=quantile(e,0.25)-1.5*iqr;
                                        wwww1=if1-absd1;
                                    of1=quantile(e,0.25)-3*iqr;
                                        wwww2=of1-absd1;
                                    if2=quantile(e,0.75)+1.5*iqr;
                                        ww1=if2-absd1;
                                        ww2=if2+absd1;
                                    of2=quantile(e,0.75)+3*iqr;
                                        www1=of2-absd1;
                                        www2=of2+absd1;
                                mad=mean(abs(e-mean(e)));
                                    md=median(e)+3*mad;
                                    lr=k1+1.5*median(e);
                                        out1=ww1;
                                        out2=ww2;
                                    out3=www1;
                                    out4=www2;
                                    out5=wwww1;
                                    out6=wwww2;
```

PROGRAMS 3

```
function  [out1 out2 out3 out4]=hebrv(e)
                                    mm=mc(e);
                  iqr=quantile(e,0.75)-quantile(e,0.25);

          w1=quantile(e,0.75)+1.5*exp(3.5*mm)*iqr;

          w2=quantile(e,0.75)+1.5*exp(4*mm)*iqr;

ww1=quantile(e,0.25)-1.5*exp(-3.5*mm)*iqr;

          ww2=quantile(e,0.25)-1.5*exp(-4*mm)*iqr;
                                    out1=w1;
                                    out2=w2;
                                   out3=ww1;
                                   out4=ww2;
```

PROGRAMS 4

```
                                   Q1L=prctile(e,12.5)
                                   Q3R=prctile(e,87.5)

                                   Q3L=prctile(e,37.5)
                                  Q1R=prctile(e,62.5);
                                    TQRL=Q3L-Q1L;
                                    TQRR=Q3R-Q1R;
                                  L=Q1L-0.97*TQRL;
                                 U=Q3R+0.97*TQRR;
```

PROGRAMS 5

```
function [out1 out2]=kimber(e)
mm=median(e);
w1=quantile(e,0.25)-1.5*(mm-quantile(e,0.25));
w2=quantile(e,0.75)+1.5*(quantile(e,0.75)-mm );
out1=w1;
out2=w2;
```

PROGRAMS 6

```
function [L U]=ninetypercent(e)
Q1=quantile(e,0.25);
Q3=quantile(e,0.75);
TQR=quantile(e,0.75)-quantile(e,0.25);
L=Q1-(0.95*(Q3-Q1));
U=Q3+(0.95*(Q3-Q1));
```

PROGRAMS 7

```
function m = mc(e)

m1=median(e);
c1=1;
c2=1;
for i=1:length(e)
if e(i)<m1
x1(c1)=e(i);
c1=c1+1;
else
if e(i)>m1
x2(c2)=e(i);
c2=c2+1;
end
end
end
c=1;
for i=1:length(x1)
for j=1:length(x2)
```

```
mm(j,i)=((x2(i)-m1)-(m1-x1(j)))/(x2(i)-x1(j));
ff(c)= mm(j,i);
c=c+1;
end
end
m=median(ff);
```

PROGRAMS 8

```
function sk= skews(x)
xbar=mean(x);
s=std(x);
n=length(x);
for i=1:n
d=x(i)-xbar;
dd(i)=d^3;
end
sk=sum(dd)/(n-1)*s^3
end
```