ROBUST OUTLIER DETECTION TECHNIQUES FOR SKEWED DISTRIBUTIONS AND APPLICATIONS TO REAL DATA

Researcher:

IFTIKHAR HUSSAIN ADIL

Reg. No.04-SE/PhD (Et.) F04

Supervisor:

Prof. Dr. ASAD ZAMAN

INTERNATIONAL INSTITUTE OF ISLAMIC ECONOMICS

INTERNATIONAL ISLAMIC UNIVERSITY
ISLAMABAD

ROBUST OUTLIER DETECTION TECHNIQUES FOR SKEWED DISTRIBUTIONS AND APPLICATIONS TO REAL DATA

IFTIKHAR HUSSAIN ADIL

Reg. No.04-SE/PhD (Et.) F04

Submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy
in Econometrics at International Institute of Islamic Economics,
International Islamic University, Islamabad

Supervisor:

Professor Dr Asad Zaman December, 2011

DECLARATION

I hereby declare that this thesis, neither as a whole nor as a part thereof, has been copied out from any source. It is further declared that I have carried out this research by myself and have completed this thesis on the basis of my personal efforts under the guidance and help of my supervisor. If any part of this thesis is proven to be copied out or earlier submitted, I shall stand by the consequences. No portion of work presented in this thesis has been submitted in support of any application for any other degree or qualification in International Islamic University or any other university or institute of learning.

Iftikhar Hussain Adil

بسم الله الرحمن الرحيم

Dedicated to

Hadia Adil, Usman Adil, and Umaima Adil

APPROVAL SHEET

ROBUST OUTLIER DETECTION TECHNIQUES FOR SKEWED DISTRIBUTIONS AND APPLICATION TO REAL DATA

by

Iftikhar Hussain Adil

Reg. No.04-SE/PhD (Et) F04

Accepted by the Faculty of International Institute of Islamic Economics, International Islamic University, Islamabad for the **Doctor of Philosophy Degree** in **Econometrics**

Supervisor	Dr. Asad Zaman
Internal Examiner	Dr. Abdul Jabbar
External Examiner	Dr. Eatzaz Ahmed
External Examiner	Dr. Ijaz Ghanni

Dr. Pervez Zamurrad Janjua

Chairman
School of Economics,
International Institute of Islamic Economics,
International Islamic University,
Islamabad

Dr. Asad Zaman

Director General,
International Institute of Islamic Economics,
International Islamic University,
Islamabad

ABSTRACT

Most of the data sets belonging to the real world contain observations at the extremes that might not be in conformity with the remaining data set. These extreme observations known to be outliers might have positive or negative effect on the data analysis like regression estimates, forecasting and ANOVA etc. Outliers are powerful tools to identify the most interesting events of the world in cross sectional data and historically important events can be picked by detecting outliers in time series data sets. Numerous outlier detection techniques have been proposed in the literature. This study provides a survey of these techniques and their properties. Most of these techniques work well under the assumption that data come from a symmetric distribution and these techniques fail to work in skewed distributions. Because of this limitation, Hubert and Vandervieren (2008) proposed a technique for outlier's detection in skewed data sets. Our thesis presents a new technique to measure robust skewness (SSS) and a new outlier detection technique (SSSBB) for skewed data distributions. The study shows that the proposed technique measures skewness more accurately than existing techniques and the proposed technique for outlier's detections works better than Hubert's technique on a class of theoretically skewed and symmetric distributions. The study also compares the technique with other established outlier detection techniques in the literature. This study uses simulation technique for computer generated distributions and some real data sets for comparison purposes. The study also analyzes real life data sets and compares the baby birth weight data and stock returns, both of which are known to be skewed. These results will help us in making a choice of appropriate outlier detection technique for skewed data sets for different sample sizes which might be helpful in identifying underweight babies.

ACKNOWLEDGEMENTS

Praise to God, who taught man what he knew not (The Holy Quran: 96-5)

First and foremost I acknowledge with immense pleasure my esteemed supervisor, Professor Dr. Asad Zaman who has supported my studies, and made my stay possible at Islamabad.

I am thankful to HEC for financial support particularly Dr Atta-ur-Rahman (Ex chairman and founder of HEC) for launching the PhD programme in Pakistan for those people who otherwise could not have afforded to invest so much time in PhD programmes.

I do not know how to thank my parents for loving me, believing in me, and empowering my thoughts throughout my life. This thesis might not have been completed without cooperation of my wife Atia Adil, I am thankful to her for her love and encouragement to me and patience to stay alone for extended periods of time and supporting our children alone. I would also like to thank my children for being quiet and understanding my position during the writing of the thesis. I would also like to be grateful of my younger brother Waqar Hussain and his wife Nosheen to support me morally and to help my family a lot during my studies. My sister Kalsoom and her husband Noubahar Husain's prayers always remained with me for my success throughout my educational career.

I am pleased with the inspirational number of inspiration and passionate brainstorming sessions with Ali Muhammad Khan Dehlavi (WWF) who has contributed a great deal to this thesis via research exposure. I am also thankful to Muhammad Saeed Sial who motivated me for the higher studies; otherwise I would not have joined the PhD programme.

I wish to thank all my class fellows who encouraged me during my studies and moments of disappointments and spurred me to continue my studies with patience specially Attique-ur-Rahman for helping me a lot in programming in Matlab, Dr Ihtsham ul Haq Padda and Rafi Amir-ud-Din for their useful comments about the English language, Saud Ahmad Khan, Muhammad Jahanzeb Malik, Mumtaz Ahmad, Tanveer-ul-Islam and Javed Iqbal for inspiring me during my studies/research.

I am cordially thankful to Dr Fozia Qureshi (SUIT) who helped me a lot searching the data on baby birth weight and referred me to Dr Sarah Saleem (AKU) for collection of data.

It is not so easy for me to enumerate all of the supporting persons who helped me throughout my studies. I might not even know some of the people involved, but I am sure that without many of them, the thesis would not have become a reality. The helping people whom I remember are Dr Abdul Jabbar, Hamid Hassan, Staff of SOE and IT branch, Muhammad Arshad Dahar, Dr Nasir Abbas, Khadija Zaheer, Asif Iqbal, Muhammad Akram Zakki, Muhammad Kafeel Bhatti, Faisal Imran Bhatti, Muhammad Ilyas, Sajjad Haider, and Malik Muhammad Shafique. I appreciate the effort of all those people whom I could not mention here.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS II

LIST OF TABLES VII

LIST OF FIGURES IX

ACRONYMS XI

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 REVIEW OF LITERATURE 3

2.1 WHAT IS AN OUTLIER? 3
2.2 HISTORY OF OUTLIERS 4
2.3 IMPORTANCE OF DETECTING OUTLIERS 5
2.4 CAUSES OF OUTLIERS 6
2.4.1 OUTLIERS IN SURVEY DATA SETS 7
2.4.2 NATURAL VARIATION 9
2.4.3 CONTAMINATION 9
2.5 EFFECTS OF OUTLIERS 10
2.5.1 DAMAGING EFFECTS OF OUTLIERS 10
2.5.2 BENEFITS OF OUTLIERS IN THE DATA SET 11
2.6 MASKING AND SWAMPING EFFECTS OF THE OUTLIERS 11
2.6.1 MASKING EFFECT 12
2.6.2 SWAMPING EFFECT 12
2.7 APPLICATIONS OF OUTLIER DETECTING TECHNIQUES 13
2.7.1 FRAUD DETECTION 13
2.7.2 MEDICAL DATA 13
2.7.3 COMMUNITY BASED DISEASES 14
2.7.4 SPORTS DATA ANALYSIS 14
2.7.5 DETECTING MEASUREMENT ERRORS 14
2.8 PREVIOUS TECHNIQUES 15
2.8.1 TUKEY'S METHOD (BOXPLOT) 17
2.8.2 METHOD BASED ON MEDCOUPLE 18

CHAPTER 3 SPLIT SAMPLE SKEWNESS 20

3.1 INTRODUCTION 20

3.2 SKEWNESS 21
3.3 VARIOUS MEASURES OF SKEWNESS 22
3.3.1 MOMENT BASED MEASURE OF SKEWNESS 22
3.3.2 PEARSON SKEWNESS 23
3.3.3 QUARTILE SKEWNESS 23
3.3.4 OCTILE SKEWNESS 24
3.3.5 MEDCOUPLE 24
3.4 SPLIT SAMPLE SKEWNESS (SSS) 25
3.5 METHODOLOGY: BOOTSTRAP TESTS FOR SKEWNESS 28
3.6 POWER AND SIZE OF THE TEST 31
3.7 MERITS OF THE NEW TECHNIQUE 40

CHAPTER 4 SPLIT SAMPLE SKEWNESS BASED BOX PLOTS 41

4.1 INTRODUCTION 41
4.2 INTRODUCTION 41
4.3 PROBLEM STATEMENT 43
4.4 PROPOSED TECHNIQUE 44
4.4.1 CONSTRUCTION 45
4.4.2 BENEFITS/ADVANTAGES OF SPLIT SAMPLE SKEWNESS ADJUSTED TECHNIQUE 46
4.5 HYPOTHETICAL DATA EXAMPLE 48
4.6 HYPOTHESIS AND METHODOLOGY 48
4.7 THEORETICAL APPROACH 50
4.8 CONVENTIONAL APPROACH: BEST AND WORST CASE IN CONTEXT OF PERCENTAGE OUTLIERS 56
4.9 COMPARISON OF SSSBB TECHNIQUE WITH KIMBER'S APPROACH 58
4.9.1 COMPARISON IN SYMMETRIC DISTRIBUTION 59
4.9.2 COMPARISON IN SKEWED DISTRIBUTIONS 60
4.10 CONCLUSION 64

CHAPTER 5 MODIFIED HUBERT VANDERVIEREN BOXPLOT 65

5.1 INTRODUCTION 65
5.2 PROBLEM STATEMENT 66
5.3 MODIFIED HUBERT VANDERVIEREN BOXPLOT (MHVBP) 69
5.4 CONSTRUCTION OF TECHNIQUE BY PROPOSED MODIFICATION 69
5.5 HYPOTHESIS AND METHODOLOGY 70
5.6 THEORETICAL APPROACH AND SIMULATION STUDY 71
5.7 SIZE OF TESTS 72
5.8 POWER OF THE TEST 73
4.11 CONVENTIONAL APPROACH: BEST AND WORST CASE IN CONTEXT OF PERCENTAGE OUTLIERS 79

5.10 ARTIFICIAL OUTLIER EXAMPLE 80
5.11 CONCLUSION 82

CHAPTER 6 MEDCOUPLE BASED SPLIT SAMPLE SKEWNESS ADJUSTED TECHNIQUE 83

6.1 INTRODUCTION 83
6.2 PROPOSED MODIFICATION 84
6.3 MONTE CARLO SIMULATION STUDY 85
6.4 COMPARISON OF FENCES PRODUCED BY MCSSSBB AND HVBP TECHNIQUES WITH TRUE 95 PERCENT CENTRAL BOUNDARY 85
6.5 CONVENTIONAL APPROACH: BEST AND WORST CASE IN CONTEXT OF PERCENTAGE OUTLIERS 93
6.6 CONCLUSION 95

CHAPTER 7 APPLICATIONS 96

SUMMARY 96
7.1 STOCK RETURN DATA SET 96
7.2 BABY BIRTH WEIGHT DATA 99
7.3 COMPARISON OF TUKEY'S TECHNIQUE AND SSSBB IN BABY BIRTH WEIGHT DATA 101
7.4 COST BENEFIT ANALYSIS 103

CHAPTER 8 CONCLUSIONS AND RECOMMENDATIONS 105

8.1 CONCLUSIONS 105
8.2.1 ADVANTAGES OF SPLIT SAMPLE SKEWNESS BASED BOXPLOT 106
8.3 ADVANTAGES OF THE MODIFIED HUBERT VANDERVIEREN BOXPLOT 106
8.4 RECOMMENDATIONS 106
8.5.1 FUTURE WORK 107

BIBLIOGRAPHY 108

APPENDIX 112

PREVIOUS TECHNIQUES 112

LIST OF TABLES

Table: 3.1 Size of Different Tests for Standard Normal Distribution.................... 33

Table: 3.2 Power Computation of Skewness Tests in χ^2 Distribution.................. 34

Table: 3.3 Power Computations of Skewness Tests in Lognormal Distribution........ 36

Table: 3.4 Power Computation of Skewness Tests in β Distribution........................ 38

Table: 4.1 Fences of Tukey and SSSBB Techniques and True Boundary in $\chi 2$ Distribution.. 52

Table; 4.2 Fences of Tukey and SSSBB techniques and True Boundary in β Distribution.. 54

Table: 4.3 Fences of Tukey and SSSBB techniques and True 95% Boundary in Lognormal Distribution.. 55

Table: 4.4 Percentage Outliers Detected by Tukey and SSSBB Technique in Chi Square Distribution... 57

Table: 4.5 Percentage Outliers Detected by Tukey and SSSBB Technique Lognormal Distribution 58

Table: 5.1 Fences of HVBP and MHVBP Techniques and 95% True Boundary in χ^2 Distribution.. 73

Table: 5.2 Fences of HVBP and MHVBP Techniques and 95% True Boundary in β Distribution.. 75

Table: 5.3 Fences of HVBP and MHVBP Techniques and 95% True Boundary in Lognormal Distribution.. 77

Table: 5.4 Percentage Outliers Detected by HVBP and MHVBP Techniques in Chi Square Distribution... 79

Table: 5.5 Percentage Outliers Detected by HVBP and MHVBP Techniques in β Distribution.. 80

Table 5.6 One Sided Artificial Outliers Detected by HVBP and MHVBP......... 81

Table 5.7 Two Sided Artificial Outliers Detected by HVBP and MHVBP........ 81

Table: 6.1 Fences of HVBP and MCSSSBB Techniques and 95% True Boundary in Chi Square Distribution.. 87

Table: 6.2 Fences of HVBP and MCSSSBB Techniques and 95% True Boundary in β Distribution... 89

Table: 6.3 Fences of HVBP and MHVBP Techniques and 95% True Boundary in Lognormal Distribution.. 91

Table: 6.4 Percentage Outliers Detected by HVBP and MCSSSBB Techniques in β Distribution.. 93

Table: 6.5 Percentage Outliers Detected by HVBP and MCSSSBB Techniques in Lognormal Distribution.. 94

Table: 7.1 Synchronized Left and Right Outliers.. 97

Table: 7.2 Outliers and IW for all Techniques in Stock Return Data...................... 99

Table: 7.3 Summary of Baby Birth Weight Data.. 101

Table: 7.4 Outliers and IW for all Techniques in BBW Data............................. 102

Table: 7.5 Performance Comparison in BBW Data 102

List of Figures

Figure: 2.1 Naseer Somroo in Local Market.. 9

Figure: 3.1 Symmetric and Skewed Distributions...................................... 21

Figure: 3.2 Comparison of Size in Standard Normal Distribution.................... 33

Figure: 3.3 Power Comparisons of Skewness Tests in Chi Square Distribution..... 35

Figure: 3.4 Power Comparisons of Skewness Tests in Lognormal Distribution...... 37

Figure: 3.5 Power Comparisons of Skewness Tests in β Distribution................ 39

Figure: 4.1 Data Coverage by Tukey's Technique in Lognormal (0, 1) Distribution.. 43

Figure: 4.2 Data Coverage by Tukey's Technique Vs SSSBB........................ 47

Figure: 4.3 Tukey's and SSSBB Technique Fences vs. 95% Boundaries in Standard Normal Distribution... 51

Figure: 4.4 Fences of Tukey and SSSBB Techniques Matching with True95% Boundary in χ^2Distribution.. 53

Figure: 4.5 Fences of Tukey and SSSBB Techniques Matching with True 95% Fence in β Distribution.. 55

Figure: 4.6 Fences of Tukey and SSSBB Technique Matching with True 95% Fence in Lognormal Distribution... 56

Figure: 4.7 Fences of Kimber and SSSBB Technique Matching with True Fence in Standard Normal Distribution... 60

Figure: 4.8 Fences of Kimber and SSSBB Technique Matching with True 95%

Fence in Chi Square Distribution.. 61

Figure: 4.9 Fences of Kimber and SSSBB Technique Matching with True 95% Fence in β Distribution.. 62

Figure: 4.10 Fences of Kimber and SSSBB Technique Matching with True 95% Fence in Lognormal Distribution.. 63

Figure: 5.1 Fence Construction of HVBP Technique around True 95% Boundary in β Distribution.. 67

Figure: 5.2 HVBP and MHVBP Technique Fences Matching with True 95% Boundary in Chi Square Distribution.. 74

Figure: 5.3 HVBP and MHVBP Technique Fences Matching with True 95% Boundary in β Distribution.. 76

Figure: 5.4 HVBP and MHVBP Technique Fences Matching with True 95% Boundary in Lognormal Distribution.. 78

Figure: 6.1 HVBP and MCSSSBB Technique Fences vs. 95% Boundaries in Standard Normal Distribution.. 86

Figure: 6.2 HVBP and MCSSSBB Technique Fences Matching with True 95% Boundary in χ^2 Distribution.. 88

Figure: 6.3 HVBP and MCSSSBB Technique Fences Matching with True 95% Boundary in β Distribution.. 90

Figure: 6.4 HVBP and MCSSSBB Technique Fences Matching with True 95% Boundary in Lognormal Distribution.. 92

Figure: 7.1 Histogram for the Stock Returns UTPL for the year 2008............... 97

Figure: 7.2 Histogram for Baby Birth Weight... 100

ACRONYMS

ACRONYMS	DESCRIPTION
HVBP	Hubert Vandervieren Boxplot
MMS	Moment Measure of Skewness
MHVBP	Modified Hubert's Vandervieren Boxplot
MC	Medcouple
OS	Octile Skewness
QS	Quartile Skewness
SSS	Split Sample skewness
SSSBB	Split Sample Skewness Adjusted
MCSSSBB	Medcouple Based Split Sample Skewness Adjusted

CHAPTER 1

INTRODUCTION

Although there is a lot of literature on outlier detection, most of the existing techniques are suitable for symmetric distributions as discussed in detail in chapter 2. Some of the authors proposed outliers techniques for skewed data, but the performance of these techniques needs improvement. The major problem of the existing outlier detection techniques is that these work in symmetric distribution and fail to work in asymmetric distribution. Some techniques assume normality assumptions while most of the real data do not follow normal distribution. Literature needs techniques which work both in symmetric and asymmetric distributions equally. This thesis proposes a new technique for measuring skewness and new technique for detection of outliers in skewed data. This technique works well both in symmetric and skewed distribution. Its performance has been proved better than existing techniques by comparing their constructed fences with the true lower and upper boundaries defined around the central 95 percent of the distributions. These calculations are analytical and easy to understand. The study has been planned in the following way. In Chapter 2 this study provides literature review of various aspects of skewness, its measurements, existence of outliers in the real data sets due to natural effects and some time due to errors and contaminations. Benefits and deleterious effects of outliers in data have been discussed along with the application of outlier detection in real life. Existing outlier detection techniques have also been discussed.

Since this study is related to the skewed distributions, it is important to have robust tests for measuring skewness of the given data set. Chapter 3 provides a review of techniques of measuring skewness in the data. This study also introduces a new technique for measuring skewness (the Split Sample Skewness henceforth abbreviated as SSS) that splits the sample from the median as its name suggests. This study also compares SSS with previous non parametric techniques like quartile skewness, octile skewness and medcouple. A new methodology based on bootstrapping has been developed to compare these techniques. Since all the techniques except moment measure of skewness are designed to be robust measure of skewness, the performance of all robust techniques has been compared by matching the size in symmetric distribution and then comparing the power in skewed distributions adopting bootstrap simulation technique. Superiority of the technique has been proven by simulation results.

In Chapter 4, a new technique has been developed based on split sample methodology to detect outliers in the skewed distributions. This technique has been applied on different distributions (χ^2, β, and Lognormal) with different parameters, and the results are compared with a very popular method named box plot developed by Tukey (1977). Applications of the proposed technique show its dominance on Tukey's and Kimber's techniques in constructing the fence around the true central 95% boundaries of the different distribution and also in real data sets.

In Chapter 5, a modification is proposed in the HV box plot technique introduced by Mia Hubert and Ellen Vandervieren (2008) which is specially designed for detection of outliers in the skewed distribution. The main problem of HV boxplot is that it generates a larger fence around the 95% boundary of the distribution and increases the chance of type II error. Simulation study has been done on the skewed distributions, like χ^2 with different degrees of freedom, β, and lognormal with different parameters and different sample sizes and supremacy of proposed modification over HVBP has been proven by the results.

In Chapter 6, a robust measure of skewness known as medcouple, introduced by G. Brys, M. Hubert and A. Struyf (2004), has been incorporated in the technique developed in Chapter 4. Again simulation study has been done on the early tested distributions in the similar fashion.

Chapter 7 includes applications of the Tukey's technique, SSSBB technique introduced in Chapter 4, HVBP (2008) and MHVBP proposed in Chapter 5 and MCSSSBB technique proposed in Chapter 6 on the real data sets of stock return of United Trust of Pakistan (UTP-2008) and baby birth weight data followed up till 28th day. Chapter 8 comprises the conclusions and recommendations based on the theoretical and empirical evidence and directions for the future work.

CHAPTER 2

REVIEW OF LITERATURE

2.1 What is an outlier?

Discordant observations may be defined as those which look different from other observations with which they are combined with respect to their law of frequency (Edgeworth, 1887; cited by Beckman and Cook, 1983). Another definition of discordant observation is that observation which appears surprising or discrepant to the investigator (Iglewicz and Hoaglin, 1993). An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs. These statements illustrate that an outlier is a subjective, post-data concept. Historically, "objective" methods for dealing with outliers were employed only after the outliers were identified through a visual inspection of the data (Grubbs, 1969; cited by Beckman and Cook, 1983). A contaminant is defined as an observation coming from a distribution which is different from the distribution of the rest of the data. Contaminants may or may not be noted by the investigator (Barnett, 1984). Contaminants and discordant observations are jointly known to be outliers. So in the words of Iglewicz, inconsistent observations with respect to the remaining data may be defined as outliers (Iglewicz and Hoaglin, 1994). For Hawkins, an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism (Hawkins, 1980).

2.2 History of 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).

2.3 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 (3-12) 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.

2.4 Causes of 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 in the first category might arise from a variety of sources some of which are discussed in this section. Here are some of the possible sources of outliers which the researcher observed during carrying out a survey at Keenjhar Lake district Thatta (Sind) for the SANDEE study "Valuing Recreational Use of Pakistan Wetlands".

2.4.1 Outliers in Survey Data Sets

i. Problem in Questionnaire

The design of questionnaire might have some ambiguous question that neither enumerator can understand nor the respondent can follow so that outliers are expected to appear in the data. For example, if only income is mentioned in the questionnaire without specifying the period (monthly or annual etc.), the respondent might generally understand it as monthly income and thereby give rise to an outlier in the annual income data. Similarly for the monthly income data, an economic graduate respondent may understand it as annual income rather than monthly salary and will create an outlier mistakenly on the positive side of the monthly income distribution.

ii. Problem Arising out of Enumerators' Mistakes

The enumerators themselves may also be responsible for giving rise to outliers. Taking the same example as above, if one out of twenty enumerators confuses the annual income with the monthly income, nearly five percent of the data will be detected as outliers because of the mistake of one enumerator.

iii. Problem in Explaining Question by the Enumerator to Respondent

Similarly outlier might appear when enumerator fails to explain the question to the respondent during the time questionnaire being filled. For example, an enumerator asks the respondent for the family income but does not define family income to some of respondents then outliers might exist.

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

In the developing world, most of the respondents are not familiar with the design of questionnaire most probably because of illiteracy. As a result, they respond lose heartedly or just answer by guess up till they understand the question. Lack of interest in the response or responses based on hunch or guess may also result of appearance of outliers.

v. Poor Handwriting of the Enumerator

One of the possible causes of the outliers in the survey data might be the result of illegible handwriting of the enumerators, which the data entry operator may not understand and fills the data wrongly.

vi. Problem in Data Entry by the Data Entry Operator

Outliers might be due to the mistake of the data entry operator. An advertent increase of a single zero may register a huge increase of income of 70,000 to 700,000 thereby giving rise to outliers. Such cases may arise when the data entry operator is not adequately familiar with the project in hand and his job is to copy data from questionnaire to data base.

All such types of the outliers that arise from any mistake at any step of the collection or documentation of data may be deleted or may be adjusted according to the actual population. However the outliers arising from natural variation must be kept because they are expected to tell interesting story behind data generating process and that specific observation.

2.4.2 Natural Variation

Natural variation may also be responsible for outliers. Naseer Soomro, a 7' 8'' (233.6cm) tall man from Shikar Pur of the Sind province is one of the tallest person in the Pakistan. Naturally he is markedly different from the rest of the population in that area. Birth of a person with such a height seems to be unusual in that popoulation but all of us know this reality and these type of outliers must not be deleted or ignored without sound theoritical justification.

Figure 2.1 Naseer Soomro in Local Market

2.4.3 Contamination

Outliers of third category originate from mixing of two populations in an unbalanced way. For example, mixing 97% and 3% of two populations from two different samples respectively may show 3% of the population from one sample as outlier in the resulting pooled sample.

2.5 Effects of Outliers

Outliers may have good or bad effects on the data. If these are the real observations, they point to some interesting dimensions of the data. The famous case of Hadlum vs. Hadlum, held in 1949 (Barnett, 1978) is of statistical interest because of an outlier. Mrs. Hadlum gave birth to a child after 349 days after Mr. Hadlum had left home to take up his duty in the armed force. Such an unusually long gestation period will be considered as an outlier against common gestation period which usually lasts around 280 days. The claim of Mr. Hadlum was failed as the court drew the limit of gestation period of 360 days which is unusual and statistically unreasonable. This outlier seems to be away from the distribution of gestation period but in reality it happened and was a natural outlier. However, if the outlier appears due to some mistake, it will have negative effects in analyzing the data. e.g. If ten dice are thrown ten times and a guy records the numbers of sixes in the form 2,0,3,12,2,0,1,1,3 then surely 12 will be an outlier in the data besides showing a missing value. Analysis of such type of data without giving attention to the outlier will lead to incorrect or misleading results.

2.5.1 Damaging Effects of Outliers

Estimation of parameter is greatly influenced when outliers are present in the data (Zimmerman, 1994, 1995, 1998), because they may result in an increase in the errors variance and decrease the power of test. If the errors contain outliers, these outliers decrease their normality in univariate case and sphericity and multivariate normality in case of multivariate altering the odds of making both Type I and Type II errors. In this way outliers become responsible for committing Type I and Type II error. Finally regression estimates that might be of substantive interest are distorted by the outliers (Osborne and Overbay, 2004).

2.5.2 Benefits of Outliers in the Data Set

Main benefit of the outliers in cross sectional data is that they reveal interesting facts. Outlier has importance as they appear different from the remaining data and having some genuine causes. The researcher may be interested in the causes that generate outliers. In the time series data, they tell interesting stories about the past. Six sigma event, which is the probability that an extreme value which is six SDs away from the means of a normal distribution, was presented as a sop by the econometricians of the early years of 20th century to justify the remote probability of occurrence of economic change of a magnitude of Great Depression. The 'outlier' in this case is the Great Depression itself which has great historical significance in the world economy.

2.6 Masking and swamping effects of the outliers

Sometimes one outlier has a capability to hide the other outliers and sometimes one outlier has the capability to expose an observation as outlier while it is inlier in real terms. Iglewicz and Martinez (1982, cited by Maimon, Rockach and Bin-Gal, 2005) have defined these two properties of the outliers as follows. For the regression analysis, due to masking and swamping effects, false decisions are made but former is “false negative” decision and latter is “false positive” (Chatterjee and Hadi, 2006).

2.6.1 Masking Effect

If one observation is detected as inlier in the presence of the extreme observation and by deleting this extreme observation, the observations nearer to it are also found to be outliers, this phenomenon is considered as the masking effect. Masking occurs when mean and covariance estimates are skewed towards a group of outliers, and the resulting gap of the outlier from the mean is small. For example, let x be a univariate vector as

$$x = [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 8 \quad 10 \quad 20 \quad 35]$$

By use of Tukey method of outlier detection, it will just detect one outlier that is 35. But after deleting this outlier and again applying Tukey’s method, 20 will be detected as outlier. So it can be said that 35 masked 20. As the mask (35) is removed 20 appears to be outlier. Some well-known real life data sets having the masking effect are Pearson and Sekar (1936), Belgian Telephone data, Hertzsprung-Russell Stars data, HBK data (Hawkins, Bradu and Kass, 1984) and HS data (Hadi and Simonoff, 1993).

2.6.2 Swamping Effect

When an observation (inlier) appears to be outlier in presence of another outlier and by deleting the specific outlier that observation is detected as inlier, it is called the swamping effect. Swamping occurs when a cluster of outliers skews the mean and the covariance estimates toward it and away from other inliers on the other side of the distribution, and the resulting distance from these observations to the mean is large, making them look like outliers. For example

$$x = [-16, 2, 6, 10, 15, 18, 20, 20, 30, 110]$$

Tukey's technique detects -16 and 110 as outliers but when after deleting the observation 110,-16 appears inlier which suggests that 110 is swamping -16 (Maimon, Rockach and Bin-Gal, 2005). Execution

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 Fraud Detection

Credit card fraud may be discovered when purchasing discontinuously jumps upward. Generally purchasing pattern goes suddenly high when the credit card is stolen and the person doing high shopping can be detected as Fraudulent and abnormal use of the credit card can point out the holder as fake person.

2.7.2 Medical Data

Unusual indications or extraordinary test results may be found to be associated with health troubles of a patient and to test whether a specific medical test result is abnormal. It may depend on other characteristics of the patients (e.g. gender, age, race etc). While analyzing the data of birth weight of babies, extraordinary less weighted babies are at high risk and are treated as outliers. Similarly outliers in the data of blood pressure, patients with extraordinary high blood pressure can be treated as outliers.

2.7.3 Community Based Diseases

When a public disease such as tetanus, cholera or plague etc. is disproportionately congested in some parts of the area under study, it may be an indicator of ineffectiveness of the treatment caused by some systematic human error. It points towards troubles with the corresponding vaccination program in that city. Whether an occurrence is unusual or usual it depends on different characteristic like frequency, spatial correlation, etc.

2.7.4 Sports Data Analysis

Presence of outliers in any variable related to the performance of a player may give important clues about the intentions of the players. Match/spot fixing may be suspected by the appearance of outliers. Presence of outliers in the data on "no balls" and "wide balls" of a specific player or a group of players may raise the suspicion of match or spot fixing.

2.7.5 Detecting Measurement Errors

When data are collected through a scientific experiment, an outlier may readily point towards measurement error. A very large or a very small observation relative to the whole sample may be removed if it is measurement error and in case this outlier is a real observation it will open new doors for research.

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.

2.8 Previous Techniques

Outliers labeling techniques are of two types

I. Formal Techniques

II. Informal Techniques

Formal tests are designed to test any statistical hypothesis. Generally null hypothesis is assumed for a particular distribution and then this hypothesis is checked if the extreme values belong to the distribution or not at given level of significance. Some tests are for a single outlier and others for multiple outliers. The choice of technique to detect outliers depends on the objective of analysis. Selection might depend on type of target outliers, numbers and type of data distribution (Seo, 2006).

Chauvenet (1852), Stone and Pierce (1863) first proposed a method of deletion of outliers in the data sets and this practice prevailed till twentieth century. Irwin (1925) proposed that gap between the first and second and the second and third order statistics should be used to decide whether the extreme observations are from the same population or from a different population. He computed critical values for the test statistics based on the magnitude of variance. Walsh (1950) favored a non-parametric test to decide whether the extreme values belong to the same population. Dixon's (1950) had a similar view on the gap test. For outlier rejection, Ferguson (1961) considered a number of invariant tests and found that the tests based on sample skewness are locally best invariant for detection of outliers with a minor mean shift towards positive side while the invariant tests based on sample kurtosis are locally best invariant for outliers detection with minor mean shift on either side (cited by Beckman and Cook, 1983).

Grubbs (1950) introduced a technique for outlier detection for univariate normal data sets having sample size greater than 3. This technique is based on mean and standard deviation and the largest absolute value is treated as outlier. Commonly ±2SD and ±3SD are used for normal /symmetric distributions. . These tests perform well in symmetric distributions but fail in skewed distributions. Dixon was pioneer of the test for outlier detection based on the statistical distribution "sub range ratio" for the data transformed in any order (ascending or descending). This test is designed for small samples and used to test small number of outliers. In this test, critical values are checked by Sachs (1982) table (Gibbons, Bhaumic and Aryal, 1994). If one observation is suspected as outlier then by Dixon test statistic is checked in table of critical values if the specific observation is outlier or inlier. A major drawback of this test is that it cannot be applied on the remaining data set when one observation is deleted after being observed to be an outlier.

Iglewicz and Hoaglin (1993) suggested using the median and median of the absolute deviation and on the basis of these two parameters, they proposed the test statistic for outlier's detection in univariate distribution. Hair et.al (1998) introduced the method for outliers detection based on the leverage statistic and standard deviation. In MAD$_E$ method, median and median of the absolute deviation is used. Since this statistics is based on median, it has a very high break point value equal to 50%. Carlings (1998) introduced a technique based on the median and inter quartile range as against Tukey's which used first and third quartiles and inter quartile range.

2.8.1 Tukey's Method (Boxplot)

Tukey test and its modifications are designed on the basis of first and third quartiles and inter-quartile range in which Q$_1$ (first quartile) exist at 25th percentile, Q$_3$ (3rd quartile) at 75th percentile and Inter quartile range (IQR) is the difference between the 3rd and 1st quartile. In order to construct boundaries for labeling an observation as an outlier, 1.5 times IQR is subtracted from Q$_1$ for lower threshold and 1.5 times IQR in added to the Q$_3$ for upper threshold to get the "inner fence". To find the critical values of outer fence 3 is used instead of 1.5 as value of g, mathematically

$$[L \quad U] = [Q_1 - g * (Q_3 - Q_1) \qquad Q_3 + g * (Q_3 - Q_1)]$$

where g=1.5 for inner fence and 3 for outer fence. Kimber (1990) modified the Tukey's method by changing Q$_3$ and Q$_1$ by M (median) in the lower and upper range values respectively and tried to resolve the problem of skewness. The modified form of the Tukey's approach proposed by Kimber is

$$[L \quad U] = [Q_1 - g * (M - Q_1) \qquad Q_3 + g * (Q_3 - M)]$$

where M is the sample median. Kimber also used (like Tukey) g=1.5.Carling (1998) introduced median rule on the basis of quadrants as

$$[L \quad U] = [Q_2 - 2.3 * (Q_3 - Q_1)Q_2 + 2.3 * (Q_3 - Q_1)]$$

Where Q$_2$ represent sample median and 2.3 is not fixed but it depends on target outlier percentage.

2.8.2 Method Based on Medcouple

G. Brys, M. Hubert, and A. Struyf (2004) introduced a robust measure of skewness named medcouple and found that it combines the robustness of quartile skewness and sensitivity of octile skewness. If $X_n = \{x_1, x_2, x_3 \ldots\ldots\ldots\ldots x_n\}$ is the set of continuous univariate distribution and it is sorted such as $x_1 \leq x_2 \leq x_3 \ldots\ldots\ldots\ldots \leq x_{n-1} \leq x_n$ then medcouple of the data is defined as

$$MC(x_1, x_2, x_3, \ldots\ldots x_n) = med\frac{(x_j - med_k) - (med_k - x_i)}{x_j - x_i}$$

where *med*$_k$ is the median of X_n and i and j have to satisfy $x_i \leq med_k \leq x_j$and$x_i \neq x_j$

The idea of the medcouple is quite simple. It takes a pair of observations, one from below the median and another from above the median and compares the difference from the median. If the difference is zero, then the pair is symmetric about the median. A positive difference shows that the positive observation is farther away from the median than the negative. Instead of taking the absolute value of this difference, the MC takes a ratio which converts this to the proportional difference. All pairs of such differences are tabulated and the median of these is taken as the measure of skewness. Some complications are introduced in case of ties which are ignored in this study, since these do not matter for continuous distributions. For details, the reader may see the original article.

Hubert and Vandervieren (2008) used medcouple to modify Tukey's box plot and called it the Adjusted Box Plot for skewed distribution and defined the interval of critical values as

$$[L \quad U] = [Q_1 - 1.5 * IQR * e^{a*MC} \qquad Q_3 + 1.5 * IQR * e^{b*MC}]$$

where MC is the Medcouple introduced by Brys, Hubert and Struyf (2004), defined above. They selected a=-3.5 and b=4 for a simulation based study and were uncertain about the appropriateness of these values. They also proposed the different values of a and b as $a = -3.79, b = 3.87, -a = b = 4 \; and \; -a = b = 3$

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

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

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

The value of MC ranges between -1 and +1. Data are symmetric if MC is zero and when value of MC is zero then HV box plot takes the shape of original Tukey's method as discussed above.

CHAPTER 3

SPLIT SAMPLE SKEWNESS

3.1 Introduction

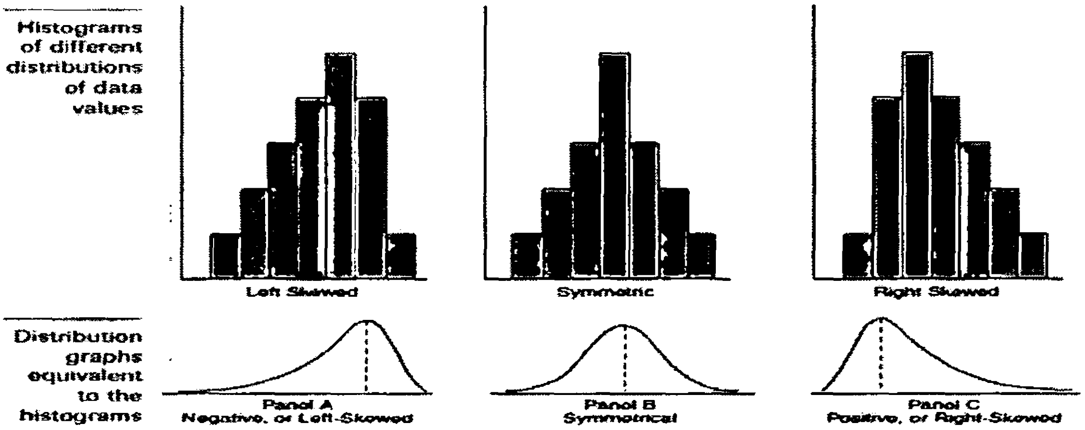
In this chapter, literature related to measuring skewness has been reviewed. This study also introduces a new measure of skewness based on the split sample.

For analyzing the observed data by nonparametric estimates it is important to evaluate different features of the distribution. In particular, the unimodality, bimodality and multimodality of the data distribution are essential for the validity of conventional descriptive statistics. If the distribution is unimodal, most of the test statistics for detection of outliers which will be reviewed in forthcoming pages will be valid and applicable. If the distribution is "two club", "twin peak" or multimodal, these tests are useless and will lead to the biased results. This study proceeds under the assumption that the data are unimodal.

3.2 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.

Figure 3.1 Symmetric and Skewed Distributions



3.3 Various Measures of Skewness

To find whether the data under consideration is symmetric or skewed, statisticians have developed different measures of skewness some of which are discussed below:

3.3.1 Moment Based Measure of Skewness

Generally, skewness of a random variable X is calculated by the third standardized moment. If X is a random variable and μ is the mean and σ standard deviation of the random variable then skewness (γ_1) can be defined as

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{\mu_3}{\sigma^3} = \frac{k_3}{k_2^{3/2}}$$

Where E is the expectation operator, k_2 and k_3 are second and third commulants respectively. This formula can also be transformed into non central moments just by expanding the above formula as

$$\gamma_1 = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{E[X^3] - 3\mu E[X^2] + 2\mu^3}{\sigma^3} = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

If $\{x_1, x_2, x_3, \ldots, x_{n-1}, x_n\}$ is a random sample, skewness of the sample is given as

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

Where n is the number of observations, $\bar{x}$ is the average of the sample. From the given sample of the population the above equation is treated as the biased estimator of the population skewness and unbiased skewness is given as

$$Moment\ Measure\ of\ skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(N-1)s^3}$$

Where s denoted the standard deviation of the sample while N shows the sample size. If the output is greater than zero, the distribution is considered to be positively skewed but if the output is less than zero ,distribution will be negatively skewed. However, if the classical skewness is statistically zero then typically distribution is treated as symmetric. Tabor (2010) discussed a number of techniques derived from the Tukey's boxplot, five point summary and from the ratio of mean to median to assess whether data are symmetric or not and evaluated which statistic performs best when sampling from various skewed populations.

3.3.2 Pearson Skewness

Karl Pearson introduced the coefficient of skewness which is estimated as

$$Sk = \frac{Mean - Mode}{Standard\ Deviation}$$

Sometimes mode can't be defined perfectly and is difficult to locate by simple methods. Therefore it is replaced by an alternative form as (Stuart and Ord, 1994)

$$Sk = \frac{3(Mean - Median)}{Standatd\ Deviation}$$

The coefficient of skewness usually varies between -3 and +3 and sign of the statistic indicate the direction of skewness.

3.3.3 Quartile Skewness

Arthur Lyon Bowley (1920, cited by Groeneveld and Meeden, 1984) proposed the quartile skewness based on the first, second and third quartiles. The co-efficient of skewness lies between -1and +1and is estimated as

$$QS = \frac{Q_1 + Q_3 - 2Median}{Q_3 - Q_1}$$

3.3.4 Octile Skewness

Hinkley (1976) introduced the octile skewness as

$$OS = \frac{Q_{0.875} + Q_{0.125} - 2 * Q_{0.50}}{Q_{0.875} - Q_{0.125}}$$

Its value also varies between -1 and +1.

3.3.5 Medcouple

Since the classical skewness is limited to the measurement of the third central moment, it may be affected by a few outliers. Keeping in view its limitations, Brys et al. introduced an alternative measure of skewness named medcouple (*MC*) which is a robust alternative to classical skewness (Brys, Hubert and Struyf, 2003). For any continuous distribution F, let $m_F = Q_2 = F^{-1}(0.5)$ be the median of F, medcouple for the distribution denoted as MC$_F$ or MC (f), is then defined as:

$$MC(F) = \underset{x_1 \leq m_F \leq x_2}{med} h(x_1, x_2)$$

Where $x_1 and\ x_2$are sampled from F and h denote the kernel. The kernel for the indicator function I is defined as

$$H_F(\mu) = 4 \int_{m_F}^{+\infty} * \int_{-\infty}^{m_F} I(h(x_1, x_2) \leq I(h(x_1, x_2) \leq \mu) dF(x_1) dF(x_2)$$

Median of this kernel is known to be the Medcouple. The domain of H_F is [-1, 1] with the conditions $h(x_1, x_2) \leq \mu, x_1 \leq m_F, x_2 \geq m_F$ are equivalent to $x_1 \leq \frac{x_2(\mu-1)+2m_F}{\mu+1}$ and $x_2 \geq m_F$. The simplified form of above equation is

$$H_F(\mu) = 4 \int_{m_F}^{+\infty} F(\frac{x_2(\mu-1)+2m_F}{\mu+1})dF(x_2)$$

If $X_n = \{x_1, x_2, x_3, \ldots, x_n\}$ is a random sample from the univariate distribution under consideration then MC is estimated as

$$MC = {}_{x_i \leq med_k \leq x_j}^{med} h(x_i, x_j)$$

Where $med_k$is the median ofX_n, and i and j have to satisfy$x_i \leq med_k \leq x_j$, and$x_i \neq x_j$.The kernel function $h(x_i, x_j)$ is given as $h(x_i, x_j) = \frac{(x_j - med_k) - (med_k - x_i)}{(x_j - x_i)}$.

The case of ties in data requires a somewhat more complex treatment, for which the reader may look at the original paper of HVBP.

The value of the MC ranges between -1 and 1. If MC=0, the data are symmetric. When MC > 0, the data have a positively skewed distribution, whereas if MC < 0, the data have a negatively skewed distribution.

3.4 Split Sample Skewness (SSS)

Classical measure of skewness is good when outliers are not present in the data. Being a third central moment, this classical measure of skewness is disproportionately affected even by a single outlier. For example, if exactly symmetric data are like{ -5,-4,-3,-2,-1, 0, 1, 2, 3, 4, 5} then its classical skewness is exactly zero but by replacing just last observation by 50, the classical skewness approaches to 2.66 whereas the other nonparametric measures perform much better in presence of this outlier. Although the latest measure of skewness (medcouple) is robust for outliers and has registered an improvement on the previously introduced measures of the skewness like quartile and octile skewness (Brys, Hubert and Struyf, 2003). It is difficult to compute the statistic even for only twenty observations without a computer. For high frequency data sets such as hourly stock exchange rates 5000 observations for example, a researcher has to construct a complex matrix of the order 2500 x 2500 and this is not possible without putting a heavy drain even on an efficient computing machine. This study introduces a new technique for measuring skewness based on natural log of the ratio of IQR_R to IQR_L where IQR_L is the inter quartile range of the lower side from the median (difference of 37.5th percentile to 12.5th percentile) while IQR_R is the inter quartile range of the upper side from the median (difference of 87.5th percentile to 62.5th percentile). Mathematically IQR_L =37.5th -12.5th percentiles and IQR_R =87.5th -62.5th percentiles. Then the split sample skewness is defined as

$$SSS = Ln\,(IQR_R/\,IQR_L)$$

Since the proposed statistic is log ratio of IQR_R to IQR_L, so if the distribution is fairly symmetric then IQR_R and IQR_L must be equal concluding their ratio equal to one and statistic value equal to 0 for the distribution to be symmetric. If the distribution is rightly skewed, IQR_R (numerator) will be greater than IQR_L (denominator) and their ratio will be greater than 1 that results the statistic value positive. If the IQR_L (denominator) is greater than IQR_R (numerator) then their ratio will be less than 1 which results the statistic value negative. So if the value of statistic is not statistically different from 0, distribution can be treated as symmetric otherwise it will be significantly skewed. If the ratio is statistically less than zero, the distribution will be negatively skewed and if it is greater than zero then it will be positively skewed.

A number of measures of skewness are suggested in the literature. It is well known fact that moment measure of skewness is not trustworthy measure of skewness (see, for example Groeneveld and Meeden, 1984; Li and Morris, 1991).

Van Zwet (1964) introduced the ordering of distributions with respect to the skewness values. According to Van Zwet, if X and Y random variables having cumulative distribution functions F(x) and G(x) and probability distribution functions f(x) and g(x) with interval support, then G(x) is more skewed to the right than F(x) if $R(x) = G^{-1}(F(x))$ is convex. One writes F $<_C$ G and says F c-precedes G. This ordering, sometimes called the convex ordering, is discussed in detail by Oja (1981). A sufficient condition for F<c G is that the standardized distribution functions$F\,s(x) = F(x\sigma_x + \mu_x)$ and $G\,s(x) = F(x\sigma_y + \mu_y)$ cross twice with the last change of sign of Fs(x) - Gs(x) being positive. Intuitively, the standardized F distribution has more probability mass in the left tail and less in the right tail than does the standardized G distribution. Gibbons and Nichols (1979) have shown that Pearson coefficient of skewness does not satisfy the ordering of Van Zwet (1964).

Oja (1981) and others (see, for example Arnold and Groeneveld, 1995) have found that any general skewness measure γ for any continuous random variable X should satisfy the following conditions (Tajuddin, 2010)

a. $\gamma(aX + b) = \gamma(X), \forall\, a > 0, -\infty < b < \infty$

b. $\gamma(-X) = -\gamma(X)$

c. If F is symmetric then $\gamma(F)=0$

d. If F <c G (F c-precedes G) then γ (F) < γ (G)

For SSS it can be observed that

a. Split sample skewness is unaffected by change of location and scale as IQR_L and IQR_R will remain the same even after changing the location or scale for the distribution with the result that SSS will remain same.

b. By changing the sign of complete data set, shape of distribution will be changed in the opposite direction. In case of SSS, IQR_L and IQR_R will be mutually changed thus changing the sign of SSS.

c. If the distribution under consideration is symmetric, IQR_L and IQR_R will be equal and the ratio of IQR_L and IQR_R will be close to 1 and natural log of 1 will be zero satisfying the third property given above.

d. Van Zwet (1964) introduced the concept of ordering two distributions with regard to skewness. According to Van Zwet, if F(x) and G(x) are cumulative distribution functions of two random variables and f(x) and g(x) are their probability distribution functions with interval support, then G(x) will be treated more skewed to the right than F(x) if $R(x) = G^{-1}(F(x))$ is convex. This property fails for split sample skewness

3.5 Methodology: Bootstrap Tests for Skewness

In the existing literature, most of the tests for skewness are based on asymptotic distributions of the test statistics. This study has introduced a new technique to measure skewness in both symmetric and asymmetric distributions using bootstrap method.

Alternatively, normalizing transformations are used, and tests based on normal distribution. Here a new method of testing for skewness based on the bootstrap is suggested. It is expected that this method produces better finite sample results.

This study seeks to solve the following problem. Let $\{x_1, x_2, x_3, \ldots\ldots\ldots x_n\}$ be an ordered sample from an i.i.d. distribution F. Is F symmetric around its median? In other words, is it true that $F(x + m) = 1 - F(m - x)$, where m is the median of F?

To solve this problem, let $T(x_1, x_2, x_3, \ldots\ldots\ldots x_n)$ be any statistic which measures skewness. We propose to reject the null hypothesis of symmetry if this statistic is significantly different from zero. The problem is: how do we determine the critical value to assess significance?

A natural solution to this problem can be based on the method of bootstrapping. Let m be the median of the sample $y_1 = x_1 - m, \ldots\ldots, y_n = x_n - m$ and $z_1 = m - x_1, \ldots\ldots, z_n = m - x_n$ *and* $G = [y_1, y_2, \ldots\ldots, y_n, z_1, z_2, \ldots\ldots\ldots z_n]$ be the sample symmetrized around the median. In a natural sense, this is the closest symmetric sample to the original one. The empirical distribution of the symmetrized sample G is the symmetric distribution which comes closest to observed empirical distribution of the actual data. So it is considered by the null hypothesis that the observed sample is i.i.d. from this symmetric distribution G.

To test this null hypothesis, generate a bootstrap sample B* = (B_1, B_2, B_3... B_n) i.i.d. from G – such a sample can be generated by standard bootstrap re-sampling with replacement from the symmetrized sample $G = [y_1, y_2, \ldots\ldots, y_n, z_1, z_2, \ldots\ldots\ldots z_n]$. Calculate the test statistic T (B*), which measures the skewness of the sample B*. By using repeated samples, 1000 i.i.d values were generated of this test statistic under the null hypothesis of symmetry: B_i is an i.i.d. sample from the symmetric distribution G. Arranging these values in order from T1 <T2 < ... < T1000, let T (25) and T (975) be the upper and lower 2.5% critical values for a test of skewness based on T.

Because there are many symmetric distributions and each distribution will have a separate set of critical values and researcher does not know what the appropriate critical values of the specific distribution are. The natural solution to the choice of critical values is therefore the distribution is symmetrized by the proposed technique. Resample from these symmetrized distributions would be used to calculate the critical values. This methodology will overcome the problem of choice of critical values and will be compatible with sample in hand. Calculations of the size and power of a test based on this bootstrap procedure and several skewness measures are reported below.

The procedure just mentioned above provides decision about the series whether it is symmetric or not. The above mentioned procedure is summarized step by step as follows:

1. Given any data series X_1, X_2, ...X_n, calculate the test statistics for skewness T(X)
2. Formulate the symmetrized series[X_1-m, X_2-m, ...X_n-m, -(X_1-m), -(X_2-m), ...-(X_n-m), m-X_1, m-X_2,m-X_3,.....m-X_{n-1},m-X_n]
3. Generate 1000 re-samples of length n from the symmetrized series and calculate the test statistics[T_1, T_2, ...T_{1000}] for each resample
4. Sort T_1, T_2, ...T_{1000} to calculate 2.5% upper critical value UCV and 2.5% lower critical value LCV

5. Compare the T(X) with the two critical values; if LCV<T(X) <UCV than the skewness will not be rejected.

3.6 Power and Size of the Test

Here we make a comparison of the size and power of the newly introduced technique SSS with the existing measures of the skewness to show the robustness of the split sample skewness measuring technique. Various symmetric and skewed distributions are taken to analyze the power and size of different tests of skewness. The bootstrap technique discussed in previous section will give a logical decision about the symmetry of the series under consideration. However it is interesting to know whether this bootstrap based skewness testing scheme can differentiate between samples from symmetric and asymmetric distributions.

For this purpose, following algorithm has been used to calculate the size/power of the bootstrap based skewness testing procedure.

1. Given any distribution F generate a sample of size n. i.e. $\{x_1, x_2, x_3, \dots\dots . x_n\}$
2. Apply the bootstrap skewness test algorithm discussed in section 3.5 to get the logical decision about symmetry.
3. Count the percentage of rejections of null hypothesis of symmetry.
4. If the underlying distribution F was symmetric (as selected, $N(0,1)$), the rate of rejection of symmetry would correspond to the size of skewness testing scheme.
5. If the underlying distribution was asymmetric (as selected $\chi^2, lnN \ and \ \beta$), rejection of symmetry corresponds to the power of the testing scheme.

Since Classical skewness is highly affected even by single outlier and cannot be compared to the robust measures of the skewness and hence it is omitted in this study for the purposes of comparison. Power has been compared to different levels of moment measure of skewness in χ^2, β and lognormal distributions. Moment measure of skewness of distributions under consideration is given below:

$Skewness\ of\ \chi^2 = \sqrt{8/_{\mathrm{k}}}$ Where k is degree of freedom of χ^2 distribution

$Skewness\ of\ lognormal\ distribution = \left(e^{\sigma^2} + 2\right)\sqrt{e^{\sigma^2} - 1}$ where σ is standard deviation of the lognormal distribution

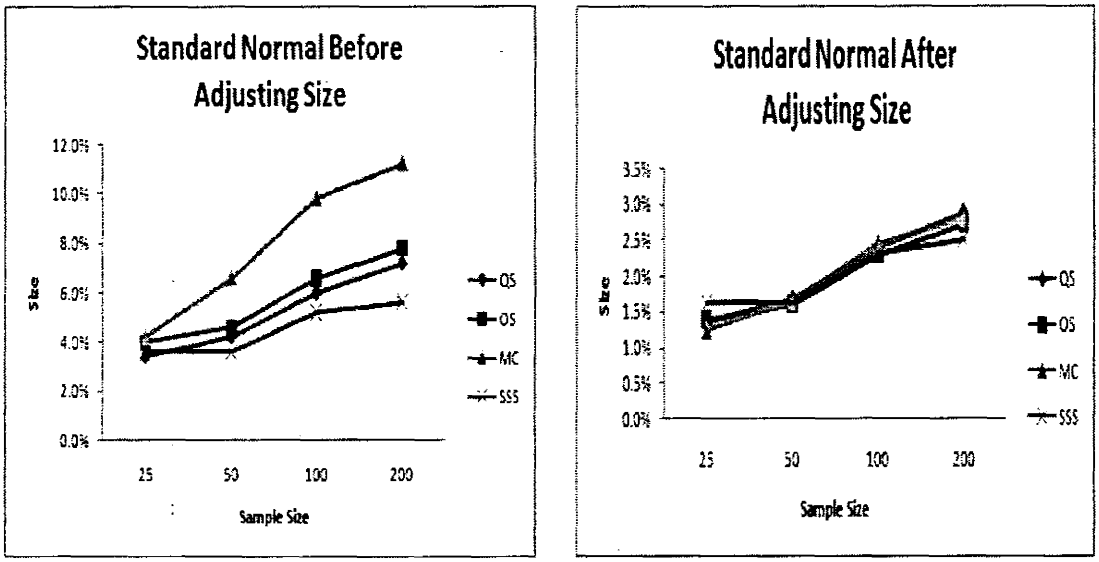
$Skewness\ of\ \beta\ distribution = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$ where α and β are the parameters of β distribution.

Left diagram for the standard normal distribution (that is symmetric theoretically), the size of the test, it can be observed that different measures of skewness have different sizes in figure 3.2 (left). It is also obvious that size of medcouple is greater than all competing techniques. When someone wants to compare the power of these statistics, size of statistic should be kept same. To equalize the size of tests, necessary adjustments in the formulae of these statistics are made. By multiplying quartile skewness with 0.8, octile skewness with 0.7, medcouple with 0.5 and SSS with 0.9 the size of the tests become similar. Sizes of the techniques under comparison are matched for different sample sizes 25, 50, 100 and 200. Sample size is taken on X-axis and on Y-axis size of the statistics is represented. Now having equaled the size of all the statistics under comparison, a researcher is able to compare the power of these statistics in skewed distributions by multiplying the power with same constants.

Table 3.1 Size of Different Tests for Standard Normal Distribution

Sample Size	Quartile Skewness	Octile Skewness	Medcouple	Split Sample Skewness
25	3.4%	4.0%	4.2%	3.6%
50	4.2%	4.6%	6.6%	3.6%
100	6.0%	6.6%	9.8%	5.2%
200	7.2%	7.8%	11.2%	5.6%

Figure 3.2 Comparison of Size in Standard Normal Distribution



Power of different statistics to is used to identify if the sample under consideration is generated from symmetric or from skewed distribution. Since chi square and lognormal distributions are skewed to the right, the statistic which will detect it skewed to the right maximum number of times in simulations will have the highest power.

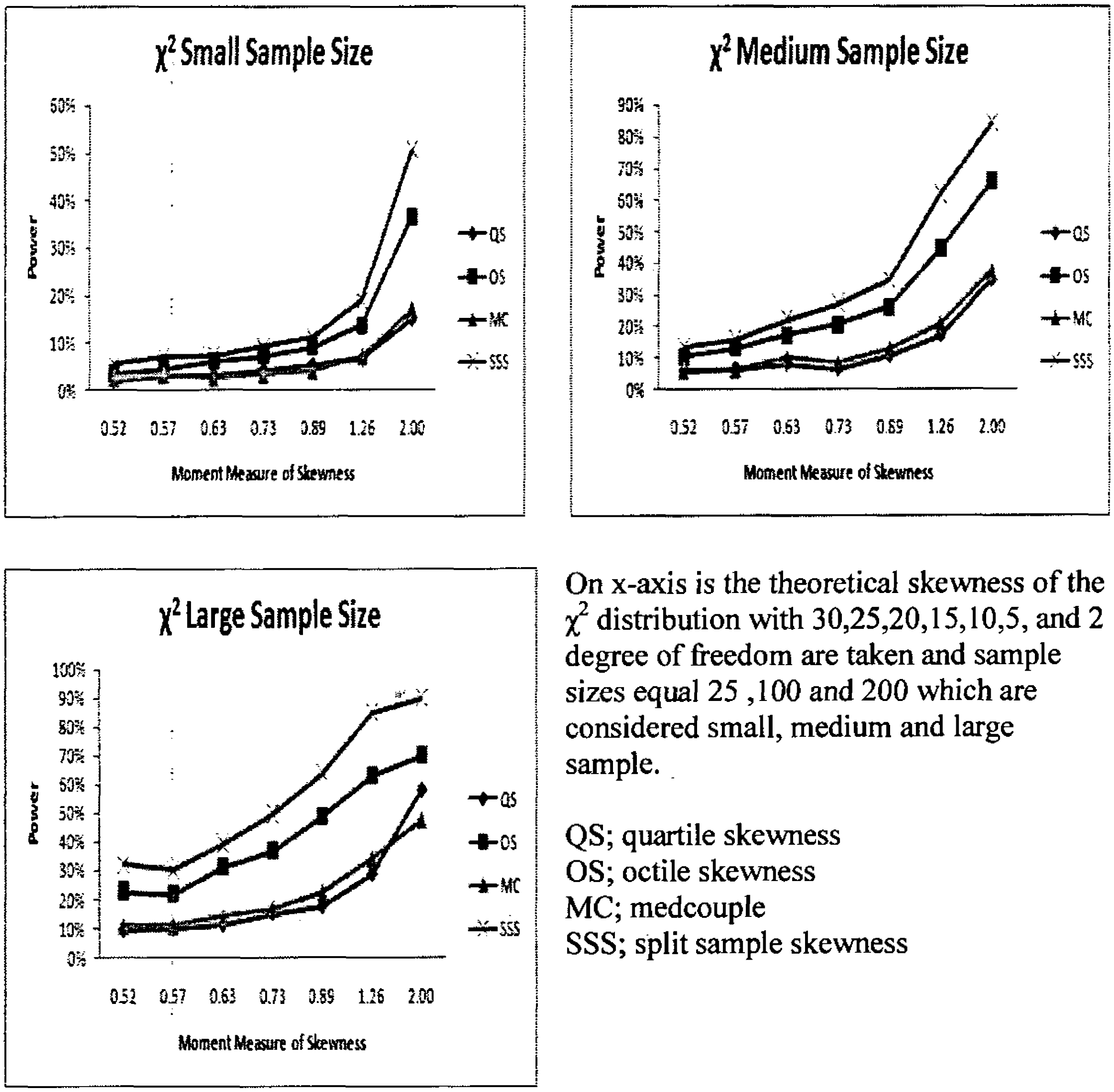
Figure 3.3 below shows the power of the different techniques using chi square distribution with different degrees of freedoms. For small sample size in chi square distribution, it can be observed that powers of medcouple are almost same for all levels of moment measure of skewness.

Table 3.2 Power Computation of Skewness Tests in χ^2 Distribution

Size	Degree of Freedom	Moment Measure Skewness	Quartile Skewness	Octile Skewness	Med-couple	Split Sample Skewness
Small Sample Size	30	0.52	1.92%	3.50%	2.60%	5.58%
	25	0.57	2.88%	4.62%	3.30%	7.02%
	20	0.63	2.88%	6.16%	2.70%	7.20%
	15	0.73	4.16%	7.28%	3.30%	9.36%
	10	0.89	5.12%	9.10%	4.10%	11.16%
	5	1.26	6.56%	13.72%	7.20%	18.90%
	2	2.00	14.88%	36.54%	16.90%	50.58%
Medium Sample Size	30	0.52	5.76%	10.64%	6.10%	13.68%
	25	0.57	6.24%	13.16%	6.30%	15.66%
	20	0.63	8.16%	17.22%	9.90%	22.32%
	15	0.73	6.56%	20.58%	8.30%	27.36%
	10	0.89	10.56%	26.18%	12.80%	34.74%
	5	1.26	17.12%	44.66%	20.60%	61.56%
	2	2.00	34.72%	65.80%	37.40%	84.24%
Large Sample Size	30	0.52	9.44%	22.82%	10.90%	32.04%
	25	0.57	10.08%	21.98%	10.90%	30.42%
	20	0.63	11.04%	31.22%	13.80%	39.24%
	15	0.73	15.36%	36.96%	16.70%	49.86%
	10	0.89	17.60%	48.86%	22.40%	63.72%
	5	1.26	28.80%	63.00%	34.10%	84.78%
	2	2.00	58.08%	70.00%	47.60%	89.82%

Power of SSS is slightly better than octile skewness at small level of moment measure of skewness while its power improves as level of skewness goes up. For medium sample size, medcouple performs better than quartile skewness while in overall comparison SSS performs better than all the robust measures of skewness under comparison followed by octile skewness in all the levels of moment measure of skewness. Again in large sample size, medcouple has higher power than the quartile skewness and quartile skewness has minimum power but when skewness level increases, quartile skewness power becomes better than medcouple.

Figure 3.3 Power Comparisons of Skewness Tests in Chi Square Distribution



On x-axis is the theoretical skewness of the χ^2 distribution with 30,25,20,15,10,5, and 2 degree of freedom are taken and sample sizes equal 25 ,100 and 200 which are considered small, medium and large sample.

QS; quartile skewness
OS; octile skewness
MC; medcouple
SSS; split sample skewness

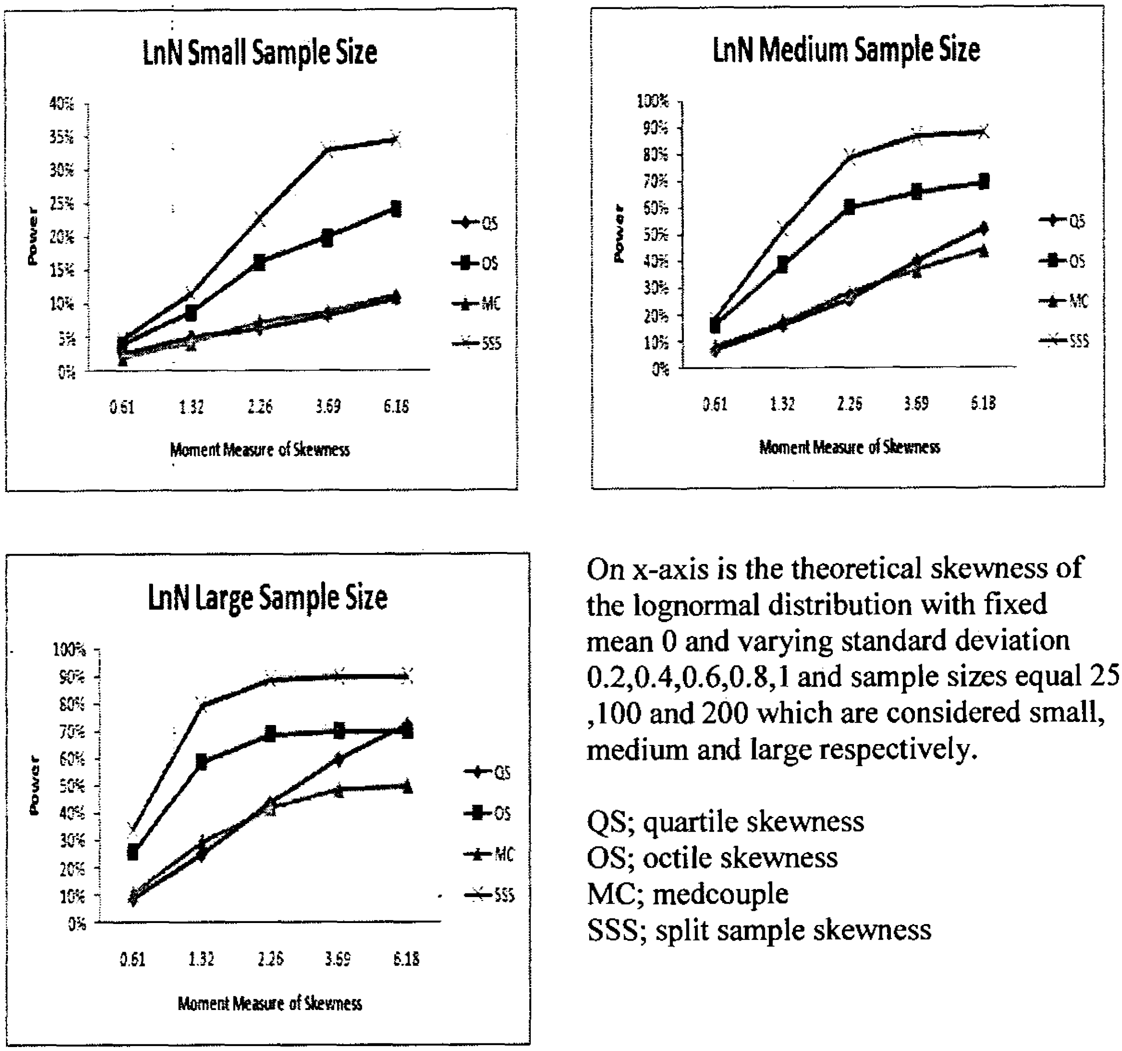
Split sample skewness has the highest power of all the measures of skewness under comparison for any level of moment measure of skewness. Hence for all sample sizes, SSS has the highest power followed by octile skewness to pick the asymmetry from the sample data sets coming from chi square distribution.

Table 3.3 Power Computation of Skewness Tests in Lognormal Distribution

Size	Parameters	Moment Measure of Skewness	Quartile Skewness	Octile Skewness	Medcouple	Split Sample Skewness
Small Sample Size	(0.0.2)	0.61	2.40%	3.92%	1.90%	4.68%
	(0,0.4)	1.32	4.96%	8.68%	4.10%	11.34%
	(0,0.6)	2.26	6.24%	16.10%	7.20%	22.68%
	(0,0.8)	3.69	8.32%	19.74%	8.70%	32.76%
	(0,1)	6.18	10.72%	24.08%	11.10%	34.38%
Medium Sample Size	(0.0.2)	0.61	6.72%	15.96%	8.30%	18.54%
	(0,0.4)	1.32	16.32%	38.50%	17.10%	52.02%
	(0,0.6)	2.26	25.60%	59.92%	28.00%	78.84%
	(0,0.8)	3.69	40.00%	65.52%	36.70%	86.76%
	(0,1)	6.18	52.16%	69.16%	44.10%	88.20%
Large Sample Size	(0.0.2)	0.61	8.64%	25.76%	10.30%	34.20%
	(0,0.4)	1.32	24.96%	58.94%	29.50%	79.56%
	(0,0.6)	2.26	43.52%	68.88%	41.80%	88.92%
	(0,0.8)	3.69	59.68%	70.00%	48.50%	90.00%
	(0,1)	6.18	72.00%	70.00%	50.00%	90.00%

Figure 3.4 reveals that in lognormal distribution, for any level of moment measure of skewness, quartile skewness and medcouple have almost the same power in small sample sizes while SSS has the highest power of all the robust measures of skewness followed by octile skewness. For medium sample size, again quartile skewness and medcouple have equal power initially for low and high moment measure of skewness but at very high moment measure of skewness (3.69 and 6.18), power of quartile skewness becomes better than medcouple. Also from start to end, SSS performed better than all other measures of skewness followed by octile skewness. For large sample sizes, it can be observed that quartile skewness and medcouple have equal power for small levels of moment measure of skewness but at high levels of skewness, quartile improves significantly from the medcouple. As usual, SSS has greater power than any other method while octile skewness has the highest power after split sample skewness.

Figure 3.4 Power Comparisons of Skewness Tests in Lognormal Distribution



On x-axis is the theoretical skewness of the lognormal distribution with fixed mean 0 and varying standard deviation 0.2,0.4,0.6,0.8,1 and sample sizes equal 25 ,100 and 200 which are considered small, medium and large respectively.

QS; quartile skewness
OS; octile skewness
MC; medcouple
SSS; split sample skewness

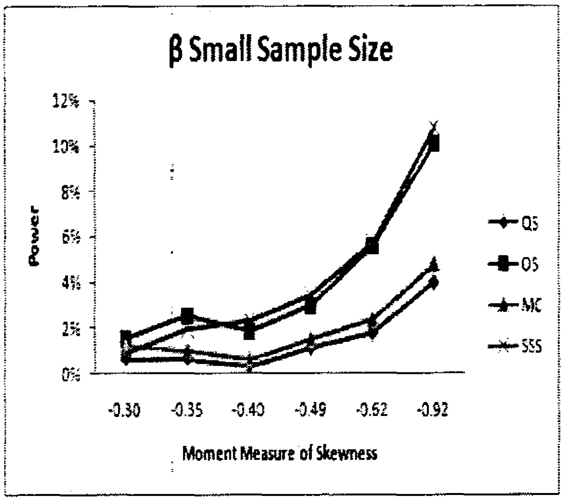
Figure 3.5 below reveals the power of different statistics to pick whether sample under consideration is generated from symmetric or from skewed distribution. As β distribution is skewed to the left, the statistic which will detect its skewness maximum number of times in simulations will be considered to be of highest power. For small sample sizes, octile skewness and SSS are performing almost equally and have a higher power as compared to quartile skewness and medcouple. Medcouple has a slightly higher power than quartile skewness. For the medium sample size, quartile skewness and MC perform equally at low levels of moment measure of skewness but MC improves for high levels of skewness. Split sample skewness has maximum power followed by octile skewness. For large sample sizes, SSS has maximum power while octile skewness seems to chase it. Remaining two measures i.e. quartile and medcouple perform almost equal.

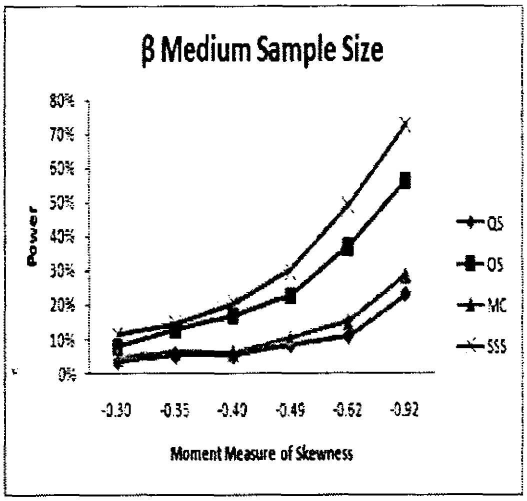
Table 3.4 Power Computation of Skewness Tests in β Distribution

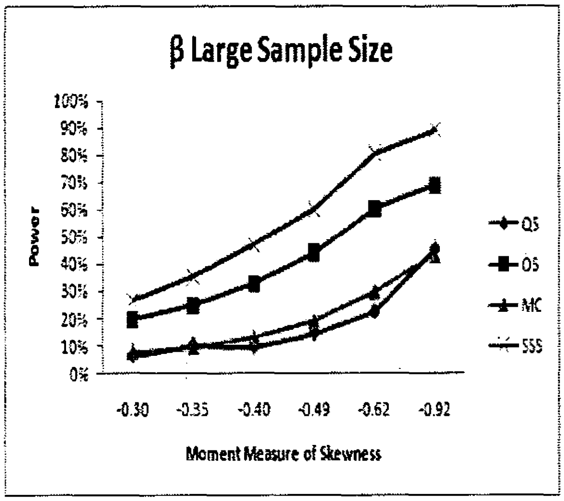
Size	Parameters	Moment Measure of Skewness	Quartile Skewness	Octile Skewness	Medcouple	Split Sample Skewness
Small Sample Size	(35,6)	-0.30	0.64%	1.54%	1.20%	0.90%
	(35,5)	-0.35	0.64%	2.52%	1.00%	1.98%
	(35,4)	-0.40	0.32%	1.82%	0.60%	2.34%
	(35,3)	-0.49	1.12%	2.94%	1.50%	3.42%
	(35,2)	-0.62	1.76%	5.60%	2.40%	5.76%
	(35,1)	-0.92	4.00%	10.08%	4.80%	10.80%
Medium Sample Size	(35,6)	-0.30	3.36%	7.70%	4.30%	11.34%
	(35,5)	-0.35	5.12%	12.60%	6.10%	14.58%
	(35,4)	-0.40	5.12%	16.80%	6.00%	20.52%
	(35,3)	-0.49	8.32%	22.68%	10.50%	30.06%
	(35,2)	-0.62	10.72%	36.96%	15.10%	49.32%
	(35,1)	-0.92	22.88%	56.14%	28.80%	72.72%
Large Sample Size	(35,6)	-0.30	6.40%	19.60%	7.70%	26.46%
	(35,5)	-0.35	10.08%	24.36%	9.50%	35.28%
	(35,4)	-0.40	9.60%	32.48%	13.20%	46.98%
	(35,3)	-0.49	14.72%	44.24%	19.20%	60.12%
	(35,2)	-0.62	22.40%	60.34%	30.00%	80.64%
	(35,1)	-0.92	45.76%	69.02%	43.40%	89.28%

For small sample sizes, no technique had any power to detect the skewness β distribution except classical skewness but for medium sample size (top right figure 3.4), split sample skewness show nearly 30% power while the rest of robust measures of skewness show zero power. For large sample sizes (bottom left figure 3.2), among the robust measures of skewness, just split sample skewness has power which approaches 100% while other robust measures have zero power in β distribution. Overall it is clear that split sample skewness has greater power than all the robust measures of skewness in all the distributions under consideration and in all sample sizes.

Figure 3.5 Power Comparisons of Skewness Tests in β Distribution







On x-axis is the theoretical skewness of the β distribution with (35, 2), (35, 3), (35, 4), (35, 5), (35, 6) and sample are taken as 25, 100 and 200 considered as small, medium and large respectively.

QS; quartile skewness
OS; octile skewness
MC; medcouple
SSS; split sample skewness

3.7 Merits of the New Technique

The important thing about the proposed technique is that it is robust in presence of outliers. Like quartile and octile skewness measurement techniques, it is easy to compute and is free from the complexity that characterizes the medcouple methodology. Its power to detect the discrepancy that distribution is symmetric is greater than all the robust measures of skewness. More weights are given to IQR_L and IQR_R because at the central octiles (37.5-50, 50-62.5 percentiles) data strength nearly remains same in general even in skewed distributions but spread declares its nature at IQR_L and IQR_R while for the external octiles (1-12.5 & 87.5-100 percentiles) outliers might exist in these locations and leaving these positions to calculate the robust skewness as it is well known that classical skewness is third central moment and it is highly affected even by 1 or 2 outliers. Major spread or compression of the data can be seen at IQR_L and IQR_R so that by giving more weight to this portion will identify the skewness that will not be affected even by 12.5% outliers on either side. It is therefore, a good addition to the class of robust methods for measuring skewness.

CHAPTER 4

SPLIT SAMPLE SKEWNESS BASED BOX PLOTS

4.1 Introduction

Tukey's boxplot continues to be extensively used to obtain data summaries. Particularly in cases where data are not normally distributed, it offers substantial advantages over other standard data summaries. It is also helpful in guiding one for detection of outliers in the data. Tukey's boxplot however tends to supply misleading results when distributions being considered are skewed (Hubert and Vandervieren, 2008). In this chapter this study suggests an alternative method to Tukey's boxplot which offers improved data summaries and greater accuracy in the identification of outliers when the data being considered belongs to skewed distributions.

4.2 Introduction

One of the important tools of exploratory data analysis (EDA) which has become widely used is the boxplot. The boxplot uses median and inter-quartile range (IQR), which are substantially more robust than the mean and the SD, and hence provide better data summaries of real data sets in most cases. In addition, the boxplot provides a useful guide to identification of outliers, which is an important activity for many reasons.

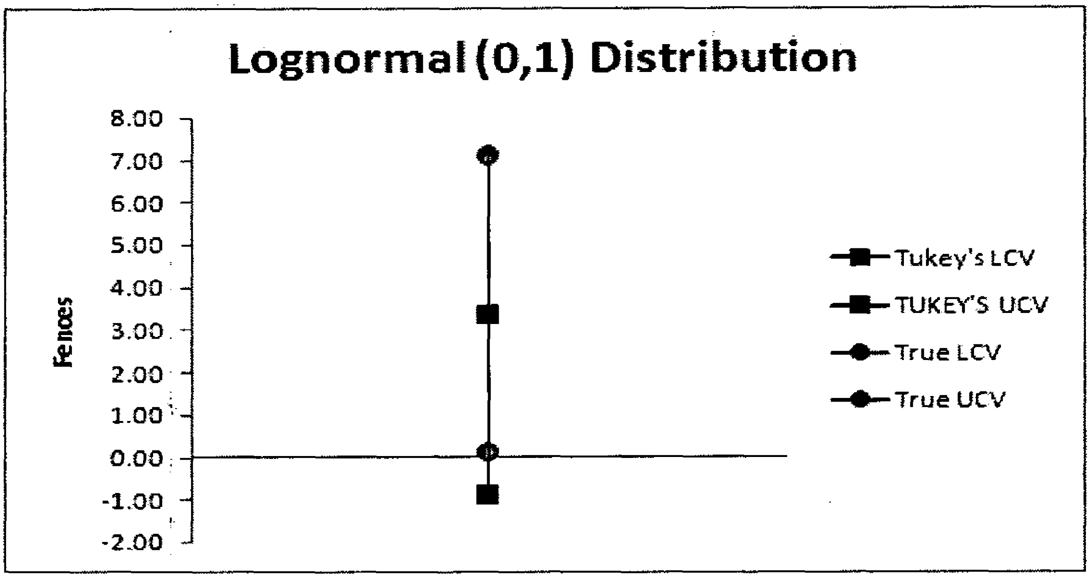
The mean and SD are perfectly adequate (in fact, theoretically optimal) data summaries for normal distributions, but fail for more general types of distributions. It is also obvious that the boxplot works well for unimodal and symmetric distribution but not so well for the distributions outside this category. It is suggested that the boxplot should routinely be accompanied by a test statistics for both unimodality and skewness. This will provide users a measure by which to assess the suitability of the boxplot in its applications to data summary.

Statistical techniques and analyses were developed in the early twentieth century under the assumption that data were normal. The assumption of normality offered theoretical convenience, simplicity and elegance of analysis, and relative ease of computational requirements. Developments in analysis of non-normal data sets required more sophisticated techniques and computational power both of which have only recently become available. Widespread use of the mean and standard deviation as data summaries is built upon the assumption of normality. These test statistics however fail badly in non-normal data sets. It has become clear that normality often fails in real data sets. Following the normal distribution assumption blindly as observed in many econometric models and in research on applied economics may affect the accuracy of inference and estimation procedures, in both cross-sectional and time series data sets. 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 nonparametric techniques is that these are often simpler than parametric techniques. In this chapter a non-parametric technique for outlier's detection has been introduced.

4.3 Problem Statement

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) and N (10, 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, nor is it a sensible description of the mixed distribution. It is suggested that different data summaries may be useful for different data sets. Unimodality is an important assumption both for Tukey's boxplot and the variant which is proposed here. Thus a test for Unimodality, either formal or informal, should routinely accompany these box plots.

Figure 4.1 Data Coverage by Tukey's Technique in Lognormal (0, 1) Distribution



True LCV and UCV corresponds 2.5% and 97.5% of the distribution respectively

It is also easy to see the failure of the boxplot for the case of skewed distributions as shown in the Fig 4.1. Since this is the topic of our chapter, this study has provided a detailed review of this issue. There are three main failings of the boxplot:

i. It does not provide a good coverage of the centre.

ii. It is ineffective at catching outliers on the narrow side of the distribution as is evident in the above figure 4.1. Lognormal distribution is rightly skewed and its narrow side is on the left. In figure 4.1 above, it can be seen that on lower side Tukey's fence has covered an area away from the true 95% distributional boundary.

iii. It detects an excessive number of outliers on the long tailed side of the skewed distribution. As shown in the above figure 4.1, Tukey's fence has dropped a lot of data on the extended side of the distribution between the Tukey's fence and the true 95% distributional boundary.

4.4 Proposed Technique

Tukey (1982) concluded his arguments on "The role of statistical graduate training" with the following lines : "We plan to influence what actually goes on, today and tomorrow.. . . We plan to help others in laying foundations for future" (Tukey, 1982, page 889, Cited by Kafadar, 2003).This study tries to bring Tukey's dream a step closer to reality through research on the foundations he laid while devising a method that is robust enough to detect outliers in skewed distributions.

As discussed earlier, Tukey's method depends on the estimation of lower and upper critical values resulting from the calculations of first and third quartile and the inter-quartile range of the complete data set. The technique being proposed in the present chapter separates the data into two parts from the median and subsequently applies Tukey's technique separately to both parts. In such a case, the first quartile, the third quartile and inter-quartile ranges lying to the left of the median generate the lower critical value, while those to the right hand side of the median generate the upper critical value.

In this study the present technique "Split Sample Skewness Based Boxplot" (henceforth abbreviated as SSSBB), computes sets of information lying on either side of the median, ranging from the 12.5 percentile to the 87.5 percentile of the complete data set. By contrast, Tukey's technique is concerned with the central half of the data, i.e., a range extending only from the 25thpercentile to the 75th percentile of the complete data set. One important advantage of the SSSBB technique over Tukey's technique in the presence of data with skewed distribution so that it is able to reliably produce coverage that approaches the middle 95% values of data more closely than Tukey's leaving 2.5% data on either side of the distribution. Here the intuition suggests that skewness of interval 12.5th percentile-37.5th percentile is different from the skewness of the interval 62.5th percentile-87.5th percentile. Skewness of the central 25% of the data (12.5% on either side of the median) is nearly equal and the extreme 25 % of the data (12.5 % on both extremes of the data) is assumed to contain outliers. Tukey's technique divides the data into four parts for detection of outliers while the SSSBB technique divides it into eight parts for detection of outliers in skewed distributions.

4.4.1 Construction

Here the procedure of the construction is discussed. Divide the data into two parts from the median, so that exactly 50% data lies on both lower and upper sides of the median. Treat these lower and upper sides as complete data sets and find the 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 is also computed. Lower and upper critical values for detecting outliers in the skewed distributions are computed by subtracting 1.5 times the inter quartile range of the lower side from the first quartile of the lower side of the median and adding 1.5 times the inter quartile range of the upper side with the third quartile of the right side of the median. Mathematically, 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

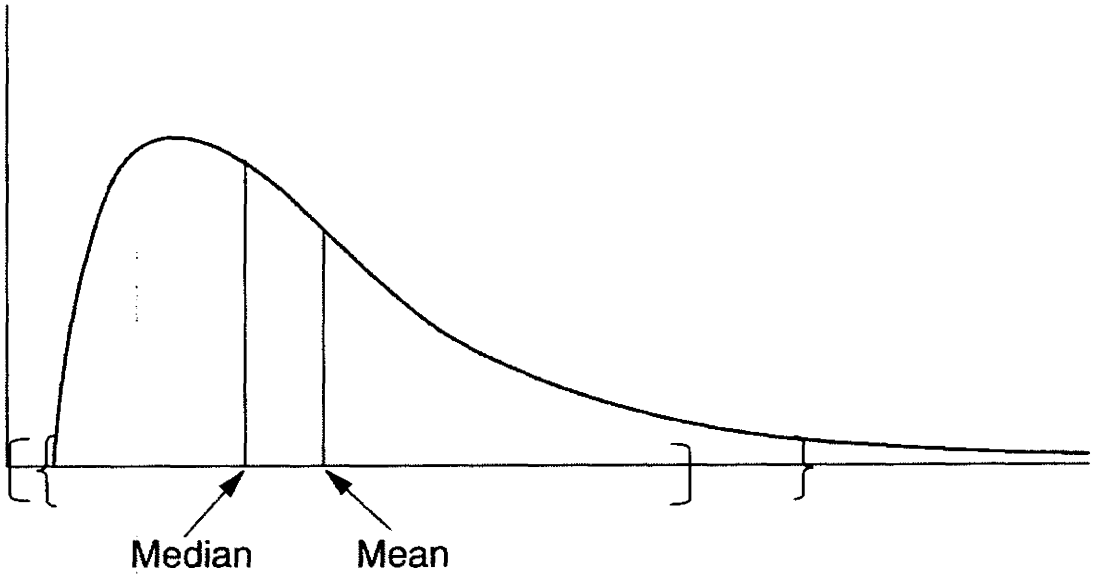
$$[L \quad U] = [Q_{1L} - 1.5 * IQR_L \quad Q_{3R} + 1.5 * IQR_R]$$

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

4.4.2 Benefits/Advantages of Split Sample Skewness Adjusted Technique

The split sample skewness adjusted (SSSBB) technique is superior to Tukey's technique when data are highly skewed. When data are moderately skewed or symmetric, performance of Tukey's technique is more or less equivalent to SSSBB technique with respect to outliers. However 95% true boundary is to remain close to the fence in the SSSBB technique. By applying the SSSBB technique, the interval of critical values moves towards the skewed side of the data. A common problem inherent in Tukey's and other techniques for detection of outliers is that these techniques extend the fence of critical values on the compressed side where data are not available and ignore the data on the side in which distribution is skewed. The SSSBB technique drags the interval of critical values towards the actual position of the data. In other words, it can be said that the interval of critical values moves towards where data are found to be more abundant. Fig.4.2 below compares the expected data coverage pattern of Tukey's with SSSBB technique. Here the square brackets represent the expected interval constructed by Tukey's technique while flower brackets represent the same for the SSSBB technique.

Figure 4.2 Data Coverage by Tukey's Technique Vs SSSBB



The SSSBB technique provides a placement of fences that improve upon the Tukey's technique. In particular it substitutes interval of critical values of Tukey's from the first and 3rd quartile to 12.5 percentile and 87.5 percentile respectively along with the selection of IQR_L and IQR_R that are helpful in determining the fence whether the distribution is skewed right or left. If IQR_L is less than IQR_R the distribution is right skewed and vice versa. Further, the SSSBB technique is an improvement over the Tukey's technique in that it is more effective in detection of outliers in theoretical as well as empirical data sets.

4.5 Hypothetical Data Example

Let us have the hypothetical data like X= [-200, 3, 7, 31, 63, 127, 255, 540]. Here it is clear that distribution is skewed to the right while -200 on the left side of the distribution is much away from the nearest observation and it should be treated as outlier.

Critical values of Tukey and SSSBB are given as

LCV (Tukey) = -223.5 UCV (Tukey) = 388.5

LCV (SSSBB) = -88.93 UCV (SSSBB) = 596.06

It is noticeable that Tukey's technique cannot detect -200 as an outlier which is a real outlier in the data and similarly it has detected the value of 540 as an outlier, when it is real observation on the right side of the data.

4.6 Hypothesis and Methodology

Outliers in a data set are small proportion coming from a different distribution from the rest of the data set comes from. The outlier detection techniques suggest a fence such that the observations outside the fence would be labeled as outliers. Five percent probability of Type I error is allowed as sizes of both techniques match at 95% true distributional boundary i.e. 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. However there are infinite types of distributions each giving different fence; if different fences are designed for different distributions, application to real data would demand prior knowledge of distributions of the data which a researcher usually does not have, so the fence is formulated for the data generated by normal distribution. In normal distribution, the size of both techniques under consideration is matched at 95% true boundary (shown in figure 4.3). So keeping in view the size of test statistics, five percent probability of Type I error is allowed, 2.5% on either side of the distribution.

For the purpose of comparisons, treat all points outside the central 95% 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 techniques will be assessed according to their ability to approach these true values. As this study is dealing with skewness and outliers in skewed data sets, the performance will be different on the two sides. Distributions only skewed to the right can be considered only. This can be done without loss of generality since if X is skewed left, -X is skewed to right.

It is important to note that this study is adopting the 95% fence to compare methodologies instead of comparing the percentage of outliers as in previous studies. This methodology has advantage to be stay at 95% boundary as 95% fence is robust measure than the extreme values i.e. Maximum and minimum so at the end percentage of outliers detected by each technique are compared.

4.7 Theoretical Approach

Every outlier detection technique makes a fence to discriminate between the usual observations and the outliers. The comparison of outlier detection techniques is based on the match between the fence and the true distribution of the data. If the distribution of the data is skewed, the classical outlier detection techniques tend to treat symmetrically both sides of the data. Therefore it leaves a lot of data on the skewed side of the distribution and covers extra area on the shorter tail of the distribution. As a result an unusual observation on the shorter tail of the distribution cannot be detected. In order to ensure the match between the distribution and the fence, the theoretical fence is calculated by allowing 5% probability of type I error. That is 5% data are allowed to remain outside the fence, 2.5% on each side. Therefore the theoretical fence for any distribution can be found in the following way

True Upper Fence= {U: P(x>U) = 2.5%}

True Lower Fence= {L: P(x<L) = 2.5%} *where x is a draw from the underlying distribution*

Fortunately, this demarcation of fence matches the fence of Tukey and SSSBB techniques when applied to the normal distribution. All outlier detection techniques are compared with respect to the match between the true fence and the fence designed by outlier detection techniques. In order to compare different techniques, adjustments in techniques are made to ensure that there is an exact match between the fences drawn by them at the standard N (0, 1) distribution. This ensures that they have equal sizes, so that a fair comparison is possible.

In order to undertake a theoretical approach, the third central moment of the distribution for normal and t distributions in case of symmetric distribution is computed to match the size of the test. Afterwards, fences of both techniques have been compared for the chi square distribution using different degrees of freedom and with different parameters of the lognormal and β distributions. Third central moments are found for the distributions under consideration using different parameters of β and lognormal distribution and different degree of freedom of chi square. True boundaries are considered at 95% central values of the distribution leaving 2.5% on each side of the distribution and fences of both techniques are calculated by substituting theoretical values of the distribution in their respective formulae.

Figure 4.3 Tukey's and SSSBB Technique Fences vs. 95% Boundaries in Standard Normal Distribution

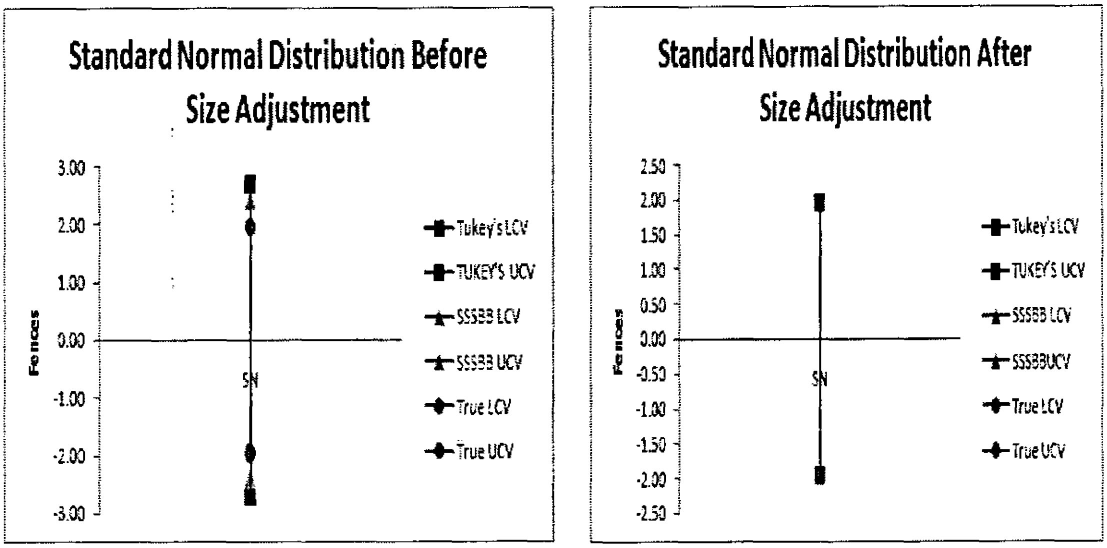


Figure 4.3(left) shows that fence of the SSSBB (marked by green triangles), fence of Tukey's technique marked by red square and true 95% fence marked as grey ball. As the size of both techniques is different from the true 95% fence which should be equal for the comparison of power of both techniques. So it is necessary to equalize the size of both

techniques at 95% true fence. Taking 95% true fence as base fence and then adjusting Tukey's fence at true 95% fence is possible by using the formula below:

$$[L \quad U] = [Q_1 - 0.95 * (Q_3 - Q_1) \qquad Q_3 + 0.95 * (Q_3 - Q_1)]$$

While SSSBB formula is adjusted as

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

After adjusting the size with respect to 95% fence, it can be observed that size of both techniques is matched at 95% fence shown in figure 4.3 (right). So the size of both techniques is matching at 5% probability of type I error. At 5%, true values of standard normal distributions are -1.96 and +1.96 (at 2.5% and 97.5%). The critical values computed by Tukey technique and SSSBB are same after adjustment in formulae of both techniques i.e. -1.96 and +1.96

Table 4.1 Fences of Tukey and SSSBB Techniques and True Boundary in χ^2 Distribution

Degree of Freedom	Moment Measure of Skewness	True LCV	Tukey's LCV	SSSBB LCV	True UCV	TUKEY'S UCV	SSSBB UCV
25	0.57	13.12	11.01	12.33	40.65	38.27	39.67
20	0.63	9.59	7.49	8.81	34.17	31.78	33.19
15	0.73	6.26	4.19	5.50	27.49	25.09	26.50
10	0.89	3.25	1.22	2.51	20.48	18.07	19.48
5	1.26	0.83	-1.08	0.18	12.83	10.38	11.80
2	2	0.05	-1.51	-0.39	7.38	4.86	6.29

Figure 4.4 Fences of Tukey and SSSBB Techniques Matching with True 95% Boundary in χ^2 Distribution

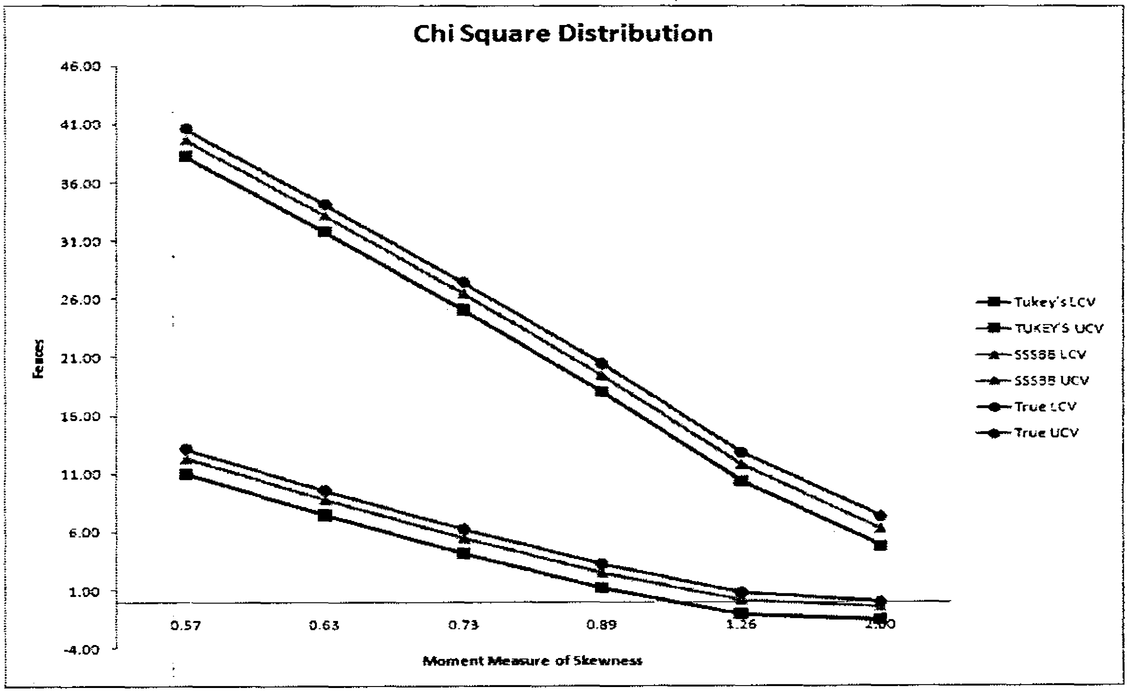


Figure 4.4 above shows that fence of SSSBB (represented by triangles) is close to true 95% fence (represented by balls) as compared to Tukey's fence (represented by squares) on both sides of the distribution. In other words SSSBB has more power to approach the reality as compared to Tukey's fence. It is obvious that for any level of moment measure of skewness (starting from 0.57 in chi square with 30 degree of freedom to 2 for chi square with 2 degree of freedom) fence of SSSBB is close to the true 95% fence. So it can be concluded that SSSBB performs better than Tukey's technique in constructing fence on both narrow and extended side of the distribution.

Table 4.2 **Fences of Tukey and SSSBB Techniques and True Boundary in β Distribution**

Parameters	Moment Measure of Skewness	True LCV	Tukey's LCV	SSSBB LCV	True UCV	TUKEY'S UCV	SSSBB UCV
(35,1)	-0.92	0.90	0.93	0.91	1.00	1.02	1.01
(35,2)	-0.62	0.85	0.88	0.87	0.99	1.02	1.00
(35,3)	-0.49	0.82	0.84	0.83	0.98	1.01	0.99
(35,4)	-0.4	0.79	0.81	0.79	0.97	0.99	0.98
(35,5)	-0.35	0.76	0.78	0.77	0.96	0.98	0.97

Figure 4.5 below provides the comparison of fences with the true distributional fence at 95% central values of the β distribution which shows that on the lower side of the distribution, SSSBB technique manages to construct a lower fence closer to the 2.5 percentile of the distribution. Again it is clear that for any level of skewness (in absolute terms) SSSBB technique is performing better as compared to the Tukey's technique. The upper fence reveals that SSSBB fence is more close to the true 95% boundary than Tukey's technique fenced as Tukey's fence is farther. So in β distribution, performance of SSSBB is better for any moment measure of skewness on both narrow and extended sides of the β distribution.

Figure 4.5 Fences of Tukey and SSSBB Techniques Matching with True 95% Fence in β Distribution

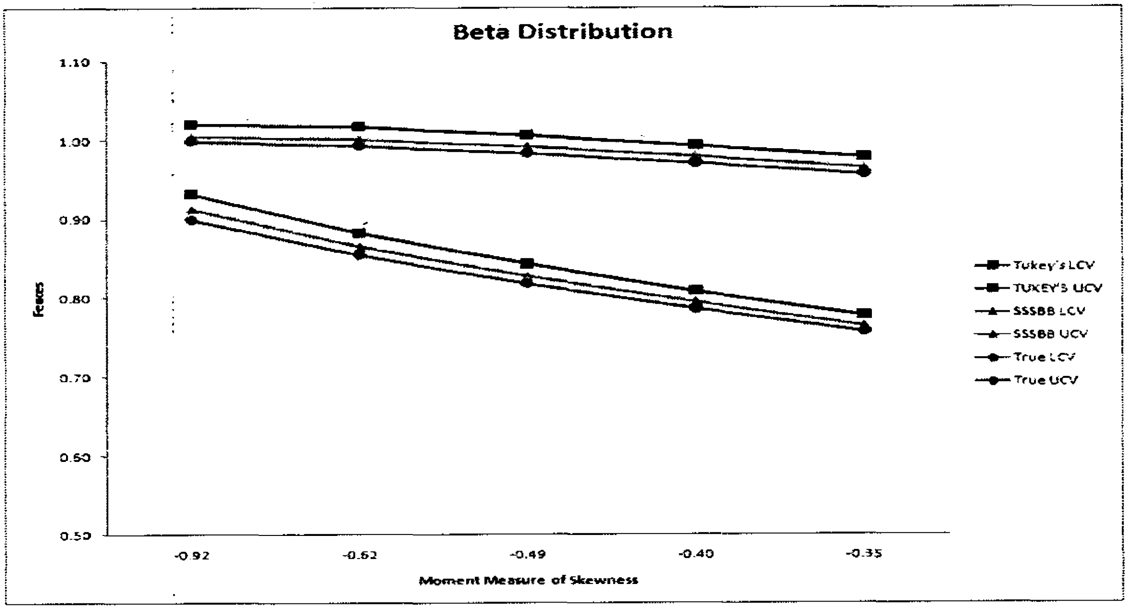


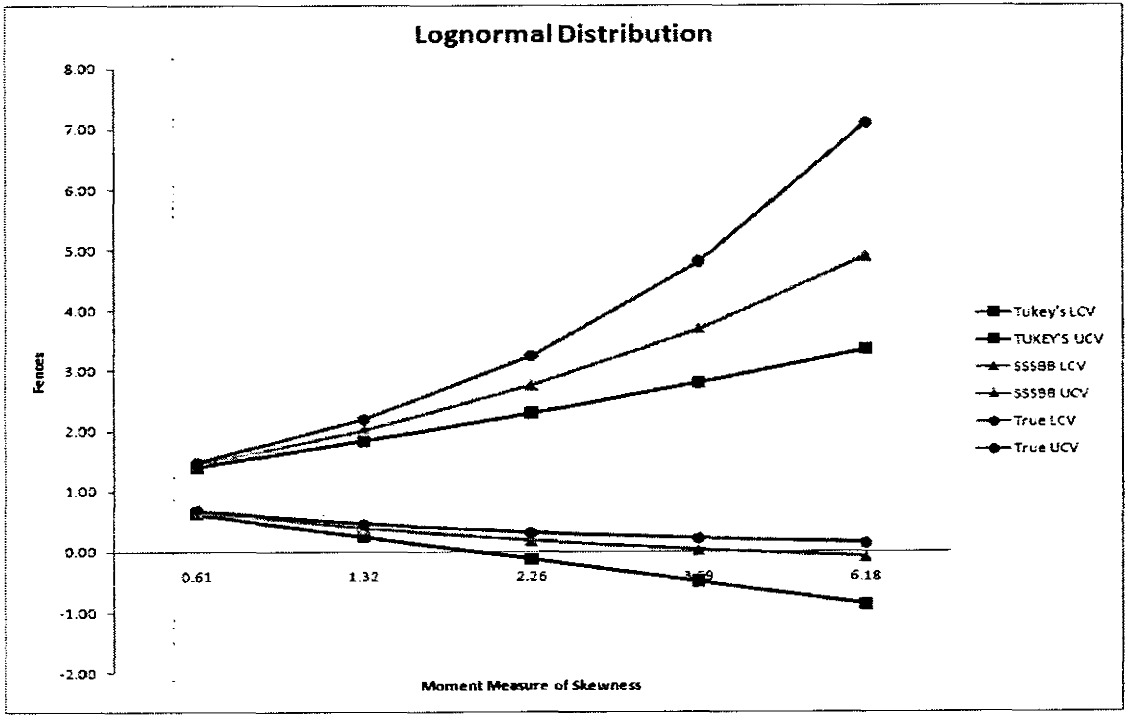
Figure 4.6 shows a comparison between Tukey's and the SSSBB fences approaching the 95% true values of the lognormal distribution. It is observed that at any level of moment, measure of skewness on the lower side of the distribution SSSBB fence is very close to the true fence while fence of Tukey's technique moving away from the true fence as the level of skewness increases.

Table 4.3 Fences of Tukey and SSSBB techniques and True 95% Boundary in Lognormal Distribution

Parameters	Moment Measure of Skewness	True LCV	Tukey's LCV	SSSBB LCV	True UCV	TUKEY'S UCV	SSSBB UCV
(0,0.2)	0.61	0.68	0.62	0.66	1.48	1.40	1.45
(0,0.4)	1.32	0.46	0.24	0.39	2.19	1.83	2.02
(0,0.6)	2.26	0.31	-0.12	0.19	3.24	2.29	2.75
(0,0.8)	3.69	0.21	-0.49	0.03	4.80	2.79	3.69
(0,1)	6.18	0.14	-0.87	-0.08	7.10	3.34	4.89

On the uppers side of the lognormal distribution, it is evident that true fence is moving away from the fences of both techniques as the skewness is going to increase but the fence of SSSBB is close to true fence as compared to Tukey's fence. In other words it can be said that SSSBB is constructing fence close to true fence around the central 95% of the data for any level of skewness on both sides of the distribution.

Figure 4.6 Fences of Tukey and SSSBB Technique Matching with True 95% Fence in Lognormal Distribution



4.8 Conventional Approach: Best and Worst Case in Context of Percentage Outliers

In conventional methodology it is frequently discussed how much percentage of outliers a technique has the power to detect in a specific distribution and performance of the technique is measured on the basis of this detected percentage of outliers. Both methodologies are more or less the same but the methodology introduced in previous section has an advantage in that it reflects the theoretical background of constructing fence around 95% by matching size of techniques under comparison and leaving 5% for type I error. In earlier studies it is common to compare the percentage of outliers but a confounding factor is that size is not matched. If a technique is generating larger fence it will detect small number of outliers. For this purpose a researcher has to match the size before power with respect to percentage of outliers which are to be compared.

In the section 4.6, it has been already discussed how to construct the fence around true 95% of the distribution. In this section, the best and worst cases in the context of percentage of outliers will be discussed. The best case discusses the best performance of SSSBB in random sample of the skewed distribution while the worst case means the worst performance of SSSBB technique. From the figures4.5, 4.6 and 4.7 above, it is clear that in all cases the fence of SSSBB is close to the 95% fence but here we wanted to assess this in terms of percentage of outliers as conventionally just percentage outliers has been discussed in earlier studies like in "A review and comparison of outliers detecting techniques" (Songwon Seo,2006).

Table 4.4 Percentage Outliers Detected by Tukey and SSSBB Techniques in Chi Square Distribution

Degree of Freedom	Skewness	Left Outliers Tukey	Left Outliers SSSBB	Right Outliers Tukey	Right Outliers SSSBB
30	0.52	1.09	1.89	4.30	3.15
25	0.57	0.98	1.78	4.47	3.14
20	0.63	0.76	1.68	4.66	3.24
15	0.73	0.52	1.50	4.94	3.35
10	0.89	0.22	1.17	5.40	3.41
2	2	0.00	0.00	8.71	4.23

A sample from chi square distribution of size equal to 100 is used for simulated study with different degree of freedom. Some 5000 simulations were run and compiled to compute the percentage of outliers detected by both techniques under comparison and it can be observed that on both sides of chi square distribution, SSSSBB has detected outliers close to 2.5 percent as compared to Tukey's technique.

Table 4.5 Percentage Outliers Detected by Tukey and SSSBB Techniques in Lognormal Distribution

Parameters	Skewness	Left Outliers Tukey	Left Outliers SSSBB	Right Outliers Tukey	Right Outliers SSSBB
(0.0.2)	0.61	1.01	1.88	4.61	3.20
(0.0.4)	1.32	0.15	1.18	6.51	3.92
(0.0.6)	2.26	0.00	0.60	8.33	4.52
(0.0.8)	3.69	0.00	0.22	9.84	5.01
(0,1)	6.18	0.00	0.06	11.22	5.48

Similarly for the same sample size in lognormal distribution, percentage of outliers detected by both techniques are presented in table 4.5 and it can be easily observed that SSSBB is performing well as compared to Tukey's technique as percentage of outliers detected by SSSBB are close to 2.5 percent than percentage of outliers detected by Tukey's technique.

4.9 Comparison of SSSBB Technique with Kimber's Approach

Kimber (1990) proposed modification in Tukey's technique for the skewed distribution

$$[L \quad U] = [Q_1 - g * (M - Q_1) \qquad Q_3 + g * (Q_3 - M)]$$

To address the problem for skewed distributions, he replaced Q_3 with median for the left critical value and Q_1 with median for the upper critical value. In the following section, the performance of the newly introduced technique is compared with Kimber technique.

4.9.1 Comparison in Symmetric Distribution

In symmetric distributions we do adjustment in SSSBB and Kimber technique to equalize the interval in symmetric distributions at 95% true fence. From the formula for critical values by Kimber method

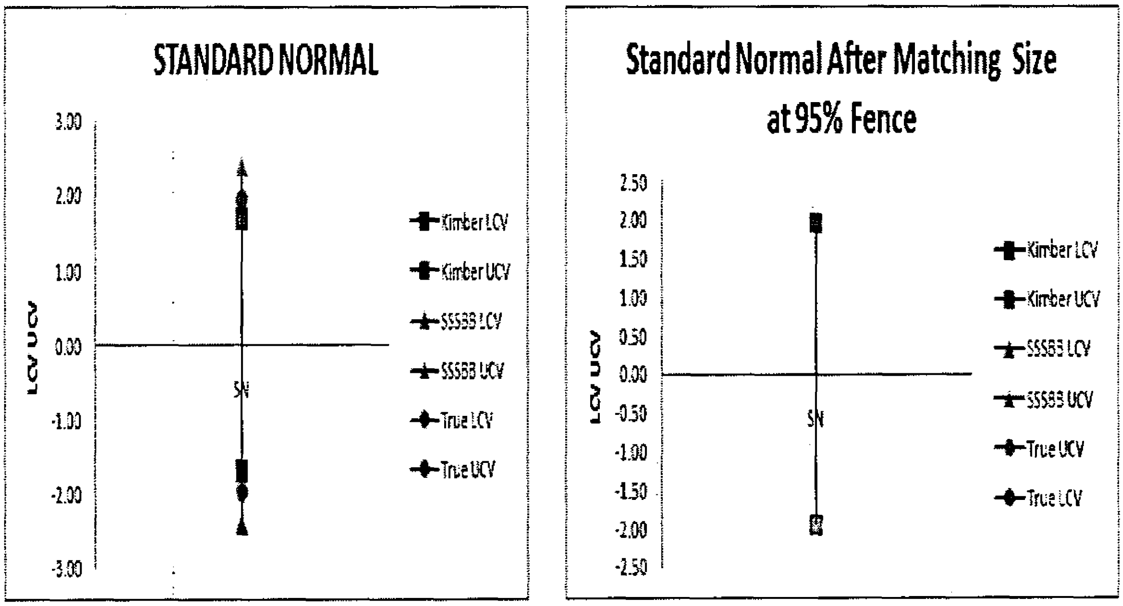
$$[L \quad U] = [Q_1 - g * (M - Q_1) \qquad Q_3 + g * (Q_3 - M)]$$

And the technique introduced in this chapter is defined as

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

But these tests do not construct equal interval in symmetric as given in the below figure 4.7 on left side. To equalize size at 95% fence, adjustment in SSSBB has been made and sizes match at 0.97 instead of original 1.5 at 95% fence (-1.96 and +1.96) and in Kimber technique it is found the value of g equal to 1.9 instead of 1.5 matches at true 95% fence. Here reader can observe the exact matching size at 95% fence in the figure 4.7 (right) below.

Figure 4.7 Fences of Kimber and SSSBB Technique Matching with True Fence in Standard Normal Distribution



In the figure golden round balls are at true 95% boundary of the distribution and red squares represent the Kimber's interval of critical values while the green triangles are for the SSSBB technique. Intervals of critical values of both techniques under comparison are overlapping with true 95% boundary.

4.9.2 Comparison in Skewed Distributions

In this section, the study compares the power of both tests in skewed distributions. The better technique constructing interval of critical values closer to the true 95% boundary is expected to perform better. The skewed distributions that were analyzed in previous section are also taken here for power comparison.

Figure 4.8 Fences of Kimber and SSSBB Technique Matching with True 95% Fence in Chi Square Distribution

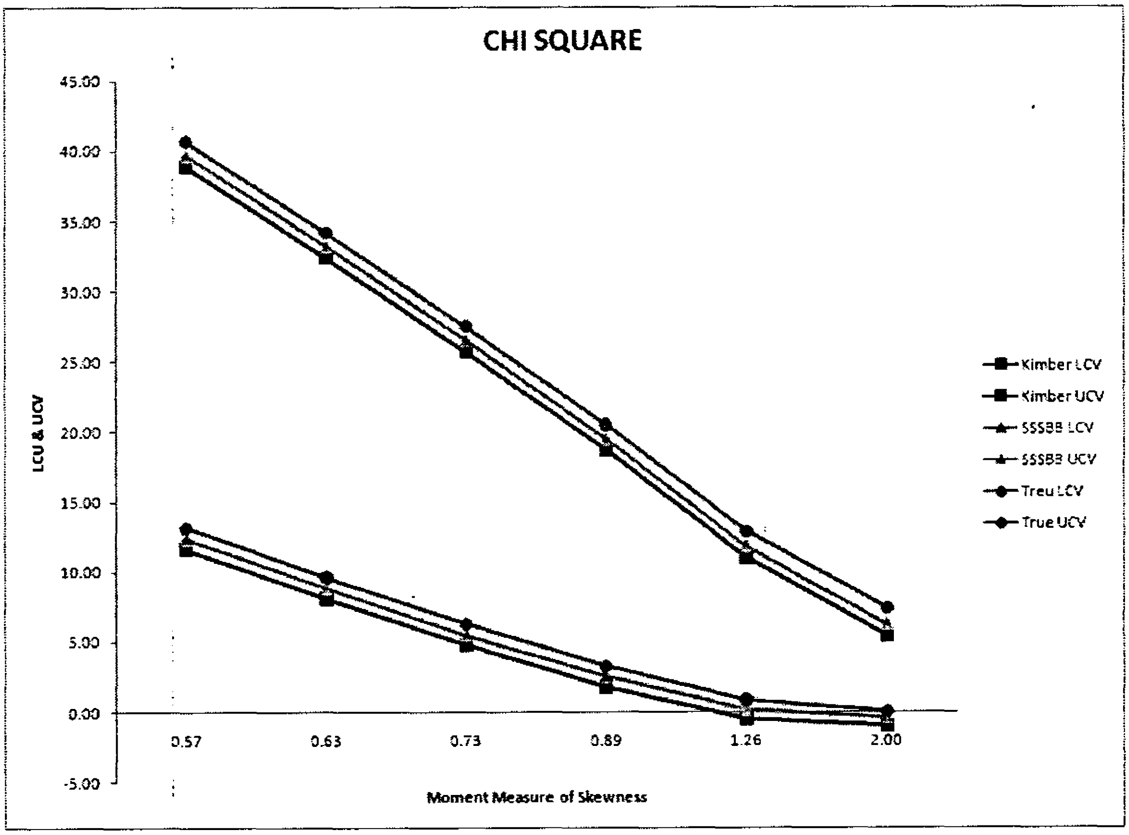


Figure 4.8 reveals that on the lower side of the chi square distribution, lower fence of SSSBB is closer to the true lower fence (2.5%) as compared to the lower fence produced by the Kimber technique. Same situation can be observed on the upper side of the distribution. For both mild and high level of moment measure of skewness, performance of SSSBB is better than Kimber technique on both side of the chi square distribution.

Figure 4.9 Fences of Kimber and SSSBB Technique Matching with True 95% Fence in β Distribution

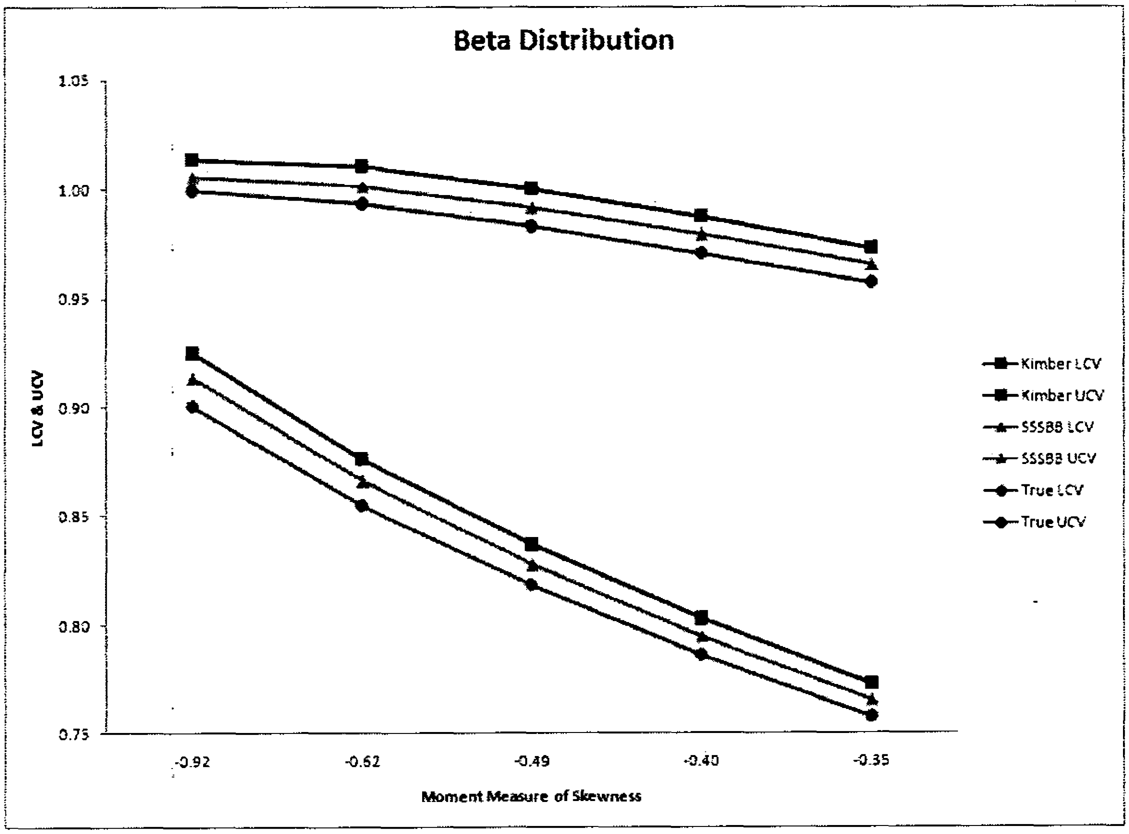
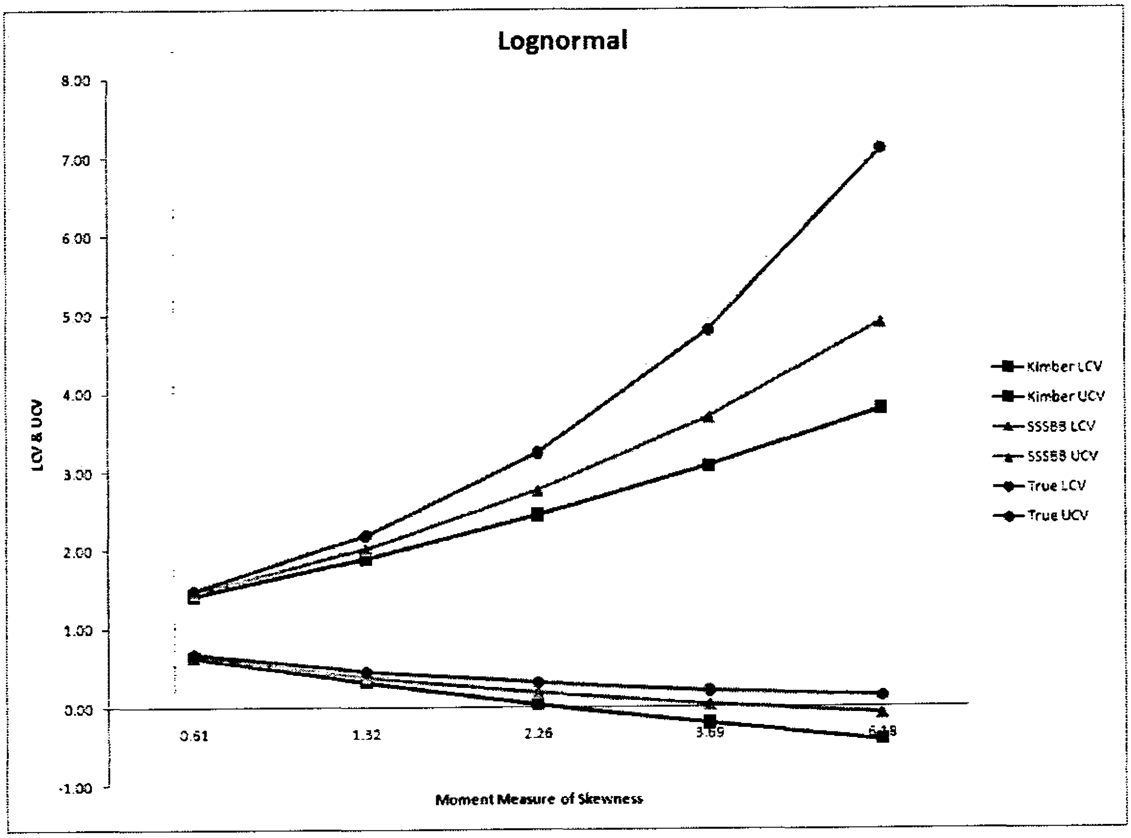


Figure 4.9 shows that for the β distribution, on the left side of the distribution, fence of SSSBB is close to true 95% fence than Kimber's technique fence. Also on the upper side of the distribution, SSSBB fence is close to true upper fence. Hence SSSBB is performing better on both sides of the distribution as compared to Kimber technique in β distribution.

Figure 4.10 shows that on the lower side of the lognormal distribution, SSSBB fence is approaching the true 95% and on the upper side of the distribution, SSSBB fence is closer to true upper fence than Kimber fence. Also it is noticeable that Kimber's interval is on the left side of the true fence on both sides unlike intervals produced by SSSBB.

Figure 4.10 : Fences of Kimber and SSSBB Technique Matching with True 95% Fence in Lognormal Distribution



4.10 Conclusion

From the above discussion it can be concluded that in skewed distributions, Tukey and Kimber techniques constructs interval of critical values wrongly that covers area along the narrow side of the distribution while leave data on the extended side of the distribution. Performance of Tukey's and Kimber techniques falls when the moment measure of skewness increases as compared to SSSBB technique. This phenomenon can be more significant in the lognormal distribution. This newly devised technique constructs fences closer to the true fence than Tukey and Kimber fences in all the distributions and has a clear advantage over Tukey's and Kimber techniques. Fences of SSSBB are always close to the true 95% fence as compared to Tukey and Kimber technique fences. However no technique constructs fence exactly equal to the true 95% fence.

CHAPTER 5

MODIFIED HUBERT VANDERVIEREN BOXPLOT

5.1 Introduction

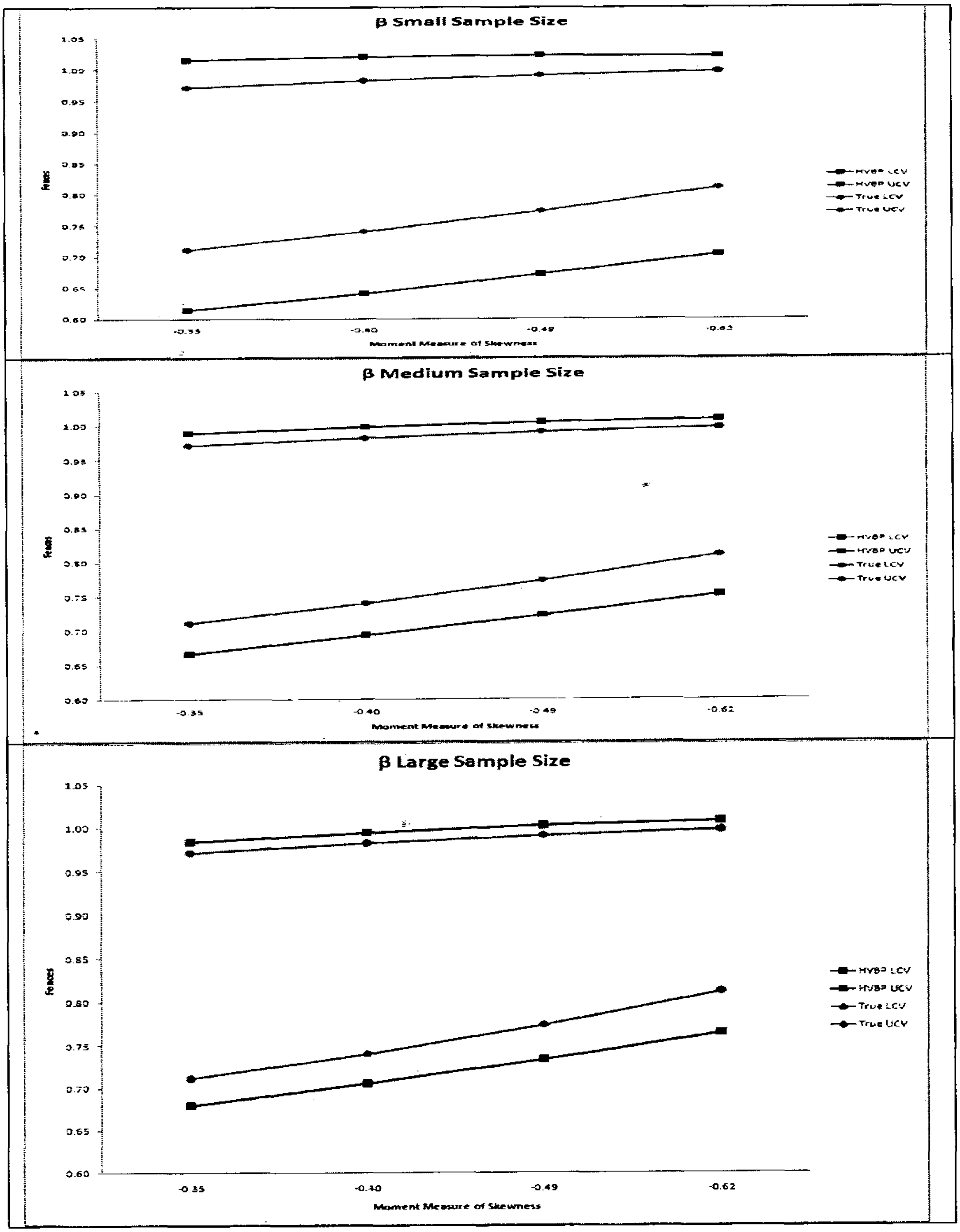
Hubert and Vandervieren (2008) tried to modify Tukey's technique for highly skewed data for detection of outliers in univariate distribution. Here Hubert's Vandervieren boxplot will henceforth be referred as HVBP in this study. A new measure of skewness "Medcouple" was introduced by G. Brys, M. Hubert and A. Struyf (2004) and the medcouple was incorporated in Tukey's technique to address the problem of identifying outliers in skewed distributions. The problem was addressed partially. For skewed distributions and large sample sizes, it performs well but a major problem relates to the construction of fence. HVBP constructs a fence very far from the true 95% boundary of the skewed distribution especially on the extended side of the distribution. Our proposed modification in HVBP performs better in both moderately and highly skewed distributions and more efficiently detects outliers asymptotically. Also it constructs fence closer to the true central 95% boundary of the distribution for most of the distributions. Theoretical approach and simulation studies verify our claim.

5.2 Problem statement

Tukey's technique is used to detect outliers in univariate distributions for symmetric as well as slightly skewed data. As the symmetry of the distribution decreases, its performance worsens and it starts to construct interval of critical values which exceeds the data limit on the one side and leaves some portion on the other side of the data. If the distribution is left skewed and the upper critical value exceeds even the maximum of the data while lower critical value will leave out a lot of data in computer generated distributions.

Hubert and Vandervieren (2008) tried to overcome the problem by incorporating a robust measure of skewness in Tukey's technique. Brys et. al. (2004) introduced "Medcouple" which is a robust measure of skewness and Hubert and Vandervieren incorporated it as a power of exponential times some constant on left and right as -3.5 and 4 changing position depending upon sign of medcouple. Incorporating this function, it condenses the interval from narrow side and extends the interval towards the puffy tail. It functions very well for the distributions which are highly skewed (skewness ≥ 3) and sample size is sufficiently large but fails to work when the skewness is slightly less than 3. For example, when a researcher checks the interval, fitting HVBP technique around the 95% true values of the β distribution, a pattern given in the figure below appears.

Figure 5.1 Fence Construction of HVBP Technique around True 95% Boundary in β Distribution



It constructs the interval of critical values even larger than extremes of the data leaving a great space between true critical values (2.5% and 97.5% of the distribution) and test statistics critical values as shown in the above figure 5.1. Performance of Hubert and Vandervieren Box Plot (HVBP) depends more on the exponential function relative to medcouple. This exponential function is multiplied on both sides with IQR. Medcouple is a small number which remains generally between 0.4 and 0.6 in absolute terms and cannot affect the constant multiplied by it as a power of exponential function. In this way it moves the interval of critical values away from the real position of the data especially in case of skewed data sets.

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

For example if MC = 0.5 that is MC > 0, then $e^{4*0.5} = 7.39$ and $e^{-3.5*0.5} = 0.17$ showing that HVBP technique is extending the upper critical value 7.39 times IQR and compressing the lower critical values 0.17 times IQR respectively even in the distribution which is only slightly skewed in the positive side due to which it extends the interval way above the true upper critical value (97.5% of the distribution) of the data and compresses it even from the true lower critical value (2.5%) of the data. Due to this the range of critical values is incorrectly increased affecting the efficiency of the test. Negatively skewed data sets are mirror images of the positively skewed and the range is now defined as

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

These suffer from the same difficulties.

5.3 Modified Hubert Vandervieren Boxplot (MHVBP)

Hubert and Vandervieren (2008) used constants ($3.5\ and\ 4$) on different sides $[LCV \quad UCV]$ and changed the position of constants with respect to the sign of the medcouple. The problem in using these constants as power of exponential times MC is that it generates a wider fence especially when data are moderately skewed. To overcome this problem of generating large fence for moderately skewed data sets this modification is going to depend on, the compression or expansion of the interval of critical values based on the classical skewness time's medcouple (instead of just sign of classical skewness, constants and medcouple) because by just using the constants HVBP constructs a very large fence even away from the extremes of the data. When data are moderately skewed, it will construct fence closer to the 95% fence and as the skewness is large, the interval will approach to the critical values of HVBP technique. So the main difference between the HVBP and Modified Hubert Vandervieren boxplot (latter on referred as MHVBP) is the use of classical skewness instead of constants.

5.4 Construction of Technique by Proposed Modification

Using similar pattern of Hubert and Vandervieren boxplot, the technique is framed as

$$[L \quad U] = [Q_1 - \ 1.5 * IQR * e^{-SK*|MC|} Q_3 + \ 1.5 * IQR * e^{SK*|MC|}]$$

Here a condition is imposed that if classical skewness is greater than 3.5 then it should be treated as 3.5. The reason to fix maximum level of skewness to 3.5 is to avoid the problem of constructing the large interval of critical values with classical skewness test statistic that might be higher than 3.5. Not allowing the skewness statistic to exceed 3.5

synchronizes the interval of critical value with the data sets as against the adjusted box plot and prevents the interval to be very large in case of highly skewed distributions. It also constructs smaller interval in case of moderately skewed distributions. So, there are clear advantages in making this modification. When the distribution is moderately skewed, HVBP takes into account the constants raised to an exponent and generates an interval large enough that even outliers actually present in the data are not detected and the test commits type II error frequently. By changing the constants with the classical skewness, its performance gets better for small and slightly skewed data sets as we can observe the results from the Monte Carlo simulation study.

5.5 Hypothesis and Methodology

Same methodology will be adopted as we discussed in the chapter 4 for comparison of HVBP technique and MHVBP technique. As both modifications are being made in the Tukey's technique and if the distribution under consideration is fairly symmetric, then both techniques become exactly similar to Tukey's technique. So it can be said that in case of symmetric distributions both techniques with same size and power can be compared at any level of confidence. As the powers of two techniques has been compared in chapter 4 allowing 5% probability of type I error, so in this section the same level for comparison of both modifications in Tukey's technique will be adopted. A comparison of both techniques brings out the following facts:

- Fences of both techniques will be compared separately on both sides of the distribution.

- A technique constructing a both fences closer to the 95% true boundary of the distribution on either side will be treated performing better on that side.
- If a technique is constructing fence close to true boundary on one side and other technique on 2nd side of the distribution then distance of both sides will be compared to access the performance of the technique.
- A technique constructing fence inside will be treated to show a better coverage if the distance of both fences is same as the true fence. If both techniques have fences on opposite side of the true fence there is a chance that some technique might generate a larger fence and at the same time minimize the percentage of outliers and increases chance of Type II error.

5.6 Theoretical Approach and Simulation Study

The study finds the moment measure of skewness of the distribution using various degrees of freedom for chi square distribution and various parameters for the lognormal and β distributions. True boundaries are constructed around 95% central values of the distribution leaving 2.5% on either side and fences of both techniques are taken from the simulated lower and upper critical values. Both upper and lower critical values for both the techniques under discussion are computed through repeated samples. For this purpose simulation study has been done for the distributions discussed above with different number of sample sizes for different levels of skewness. One hundred thousand repetitions have been done for χ^2distribution with 2, 10, 15, 20, and 25 degree of freedom with sample size of 25, 50,100 and 500. Samples from β distribution are taken with similar sample sizes with parameters α and β as β (35, 2), β (35, 3), β (35, 4), β (35, 5).

Correspondingly same sample sizes are taken from lognormal distribution as $ln\mathcal{N}(0, 0.2^2)$, $ln\mathcal{N}(0, 0.4^2)$, $ln\mathcal{N}$ $(0, 0.6^2)$, $ln\mathcal{N}$ $(0, 0.8^2)$,$ln\mathcal{N}(0, 1)$. A total of nine statistics were computed including left outlier, right outlier, total outlier, lower critical value, upper critical value, interval width (constructed by difference of the lower and upper critical values), maximum of the data, minimum of the data and sample skewness of the data for comparison of results obtained from various techniques. But in the methodology discussed above the study is just using LCV and UCV. It is already defined in chapter 4, the three sample sizes 25,100 and 500 as small, medium and large sample sizes respectively. The true boundary of 95% remains the same for the entire sample sizes which are plotted along y-axis and moment measure of skewness along x-axis.

5.7 Size of Tests

As both adjustments are based on Tukey's technique, when the data are symmetric the medcouple equals zero thereby approaching Tukey's technique.

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

It is clear from the above equation that when MC is zero, it will result in the exponent approach the power zero which means the resulting test statistic equals 1. Substituting the value of MC equal to zero will result in the above equation in Tukey technique. Similarly by substituting the value of MC or skewness equal to zero (in case of symmetric distribution), the equation below will also be converted to Tukey technique

$$[L \quad U] = [Q_1 - 1.5 * IQR * e^{-SK*|MC|} Q_3 + 1.5 * IQR * e^{SK*|MC|}]$$

The size of both the techniques is identical as based on the Tukey's technique at any level of significance. Adopting the standard methodology, we compare both the techniques at 95% level of confidence leaving 5% chance for type I error.

5.8 Power of the Test

As the size of both techniques is similar in symmetric distributions, comparison of the powers of both techniques is justified in asymmetric distribution. Power of any technique will depend on constructing the fence around true 95% fence of the distribution. For comparison of powers, chi square, β and lognormal distributions are selected as they are skewed distributions.

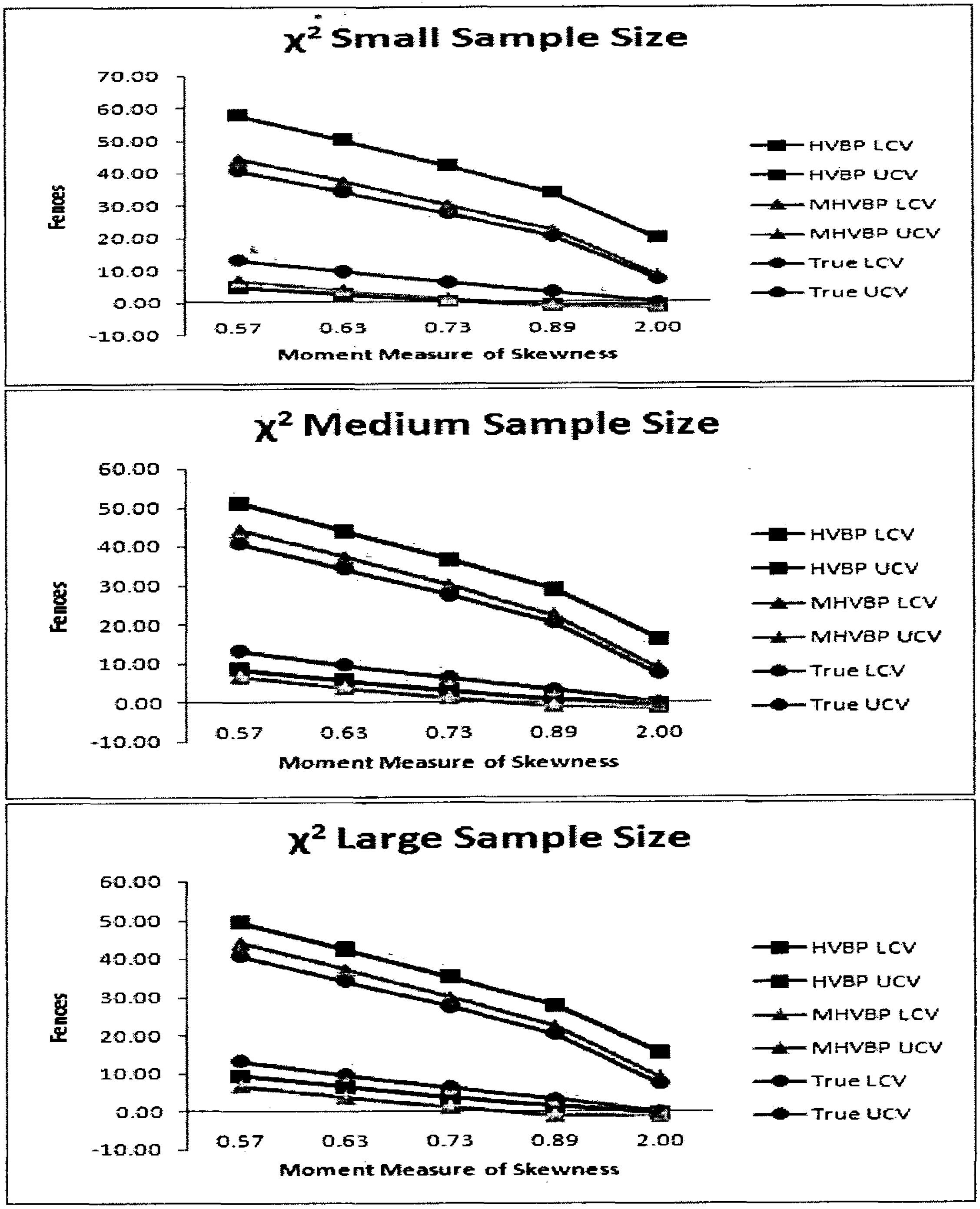
Table 5.1 Fences of HVBP and MHVBP Techniques and 95% True Boundary in χ^2 Distribution

Sample Size	Moment Measure of Skewness		0.57	0.63	0.73	0.89	2.00
True Lower Fence (2.5%)			13.12	9.59	6.26	3.25	0.05
25	Lower Critical Value	HVBP	4.66	2.39	0.38	-1.08	-0.90
		MHVBP	6.70	3.75	1.11	-1.09	-1.64
100		HVBP	8.52	5.63	3.06	0.88	-0.53
		MHVBP	6.56	3.64	1.06	-1.04	-1.32
500		HVBP	9.29	6.29	3.56	1.24	-0.47
		MHVBP	6.51	3.62	1.06	-0.99	-1.17
True Upper Fence (97.5%)			40.65	34.17	27.49	20.48	7.38
25	Upper Critical Value	HVBP	57.72	50.17	42.25	33.96	19.94
		MHVBP	44.34	37.38	30.11	22.50	8.42
100		HVBP	51.03	43.90	36.63	28.97	16.37
		MHVBP	44.18	37.22	30.01	22.40	8.90
500		HVBP	49.50	42.53	35.32	27.86	15.52
		MHVBP	44.14	37.17	29.97	22.39	9.14

Figure 5.1 (top) shows the interval fitting pattern of adjusted boxplot and proposed treatment around the true 95% boundaries in χ^2 distribution for small sample size. It is clear that on the lower side, fences of HVBP and MHVBP overlap and fences of both techniques are at the same distance from the true lower fence implying equal

performance. For the upper fence, it is obvious that true 95% fence and fence of MHVBP overlap while the fence of HVBP is at a large gap from the true upper fence.

Figure 5.2 HVBP and MHVBP Technique Fences Matching with True 95% Boundary in χ^2 Distribution



For medium sample size it can be observed that on the lower side, HVBP fence improved its performance as compared to that of MHVBP while on the upper side it can be noticed that MHVBP has a great advantage over HVBP. Overall comparison of interval shows that MHVBP interval is close to the 95% fence. Nearly same situation of medium sample size can be observed in the large sample size of χ^2 distribution.

Table 5.2 Fences of HVBP and MHVBP Techniques and 95% True Boundary in β Distribution

Sample Size	Moment Measure of Skewness		-0.35	-0.40	-0.49	-0.62
True Lower Fence (2.5%)			**0.76**	**0.79**	**0.82**	**0.85**
25	Lower Critical Value	HVBP	0.61	0.64	0.67	0.70
		MHVBP	0.73	0.76	0.80	0.84
100		HVBP	0.67	0.69	0.72	0.75
		MHVBP	0.73	0.76	0.80	0.84
500		HVBP	0.68	0.71	0.73	0.76
		MHVBP	0.73	0.76	0.80	0.84
True Upper Fence (97.5%)			**0.96**	**0.97**	**0.98**	**0.99**
25	Upper Critical Value	HVBP	1.01	1.02	1.02	1.02
		MHVBP	1.01	1.02	1.03	1.03
100		HVBP	0.99	1.00	1.01	1.01
		MHVBP	1.01	1.02	1.03	1.03
500		HVBP	0.98	0.99	1.00	1.01
		MHVBP	1.01	1.02	1.03	1.03

Figure 5.2 shows the fence construction pattern of HVBP and MHVBP techniques around 95% true fence in β distribution. For small sample it is clear that on the lower side of the distribution, fence of MHVBP is very close to true lower fence (constructed at 2.5% of the distribution) as compared to HVBP while the upper side fences of HVBP and MHVBP overlap which implies that equal performance on the upper side while better performance of MHVBP on the lower side of the distribution. For medium sample size, again fence of HVBP is away (even from the range of the data) from the true lower fence while HVBP has a bit of advantage on the upper side of the distribution. Again overall

fence of MHVBP is close to the true 95% fence in β distribution. Almost similar pattern can be observed in the large sample size.

Figure 5.3 HVBP and MHVBP Technique Fences Matching with True 95% Boundary in β Distribution

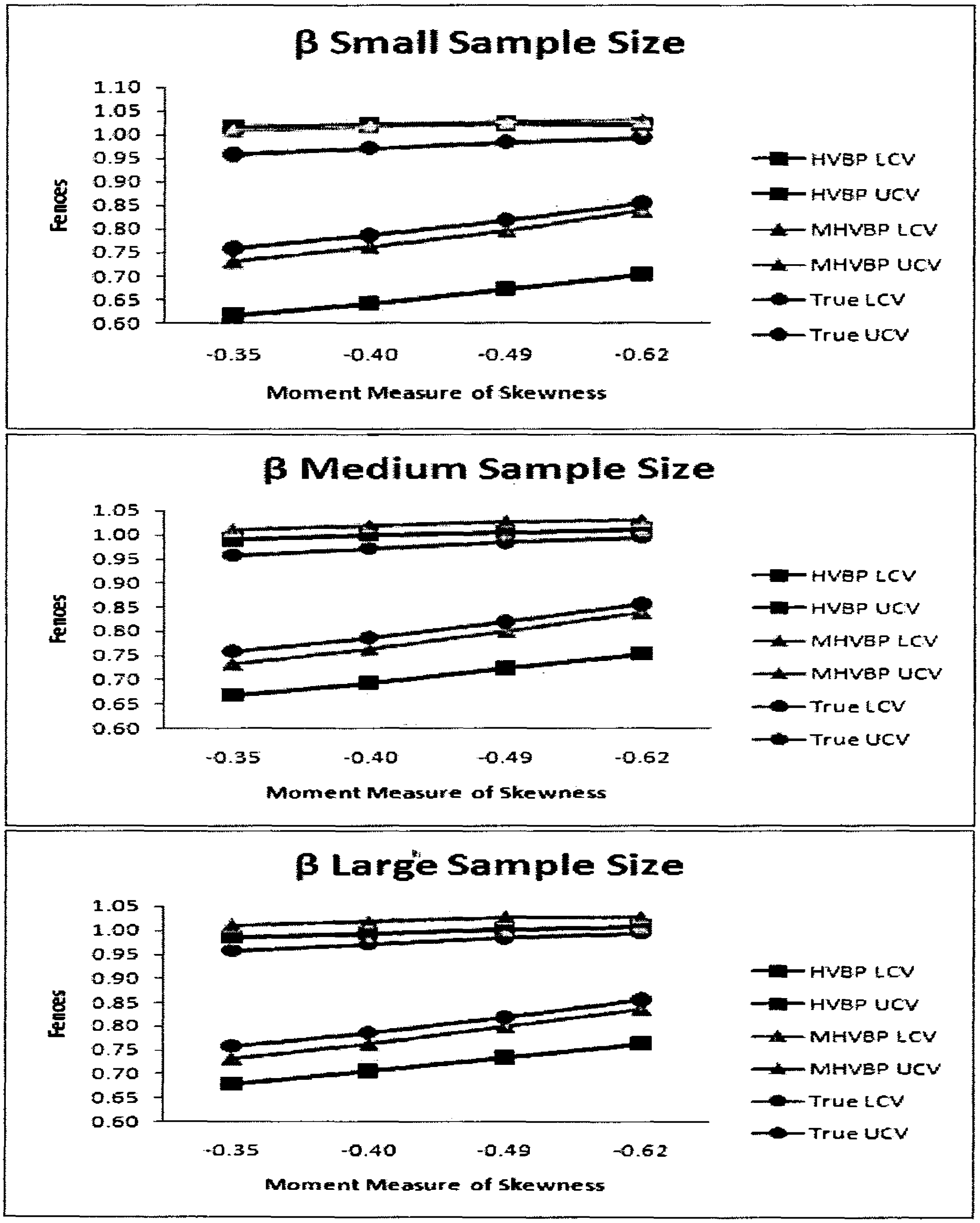
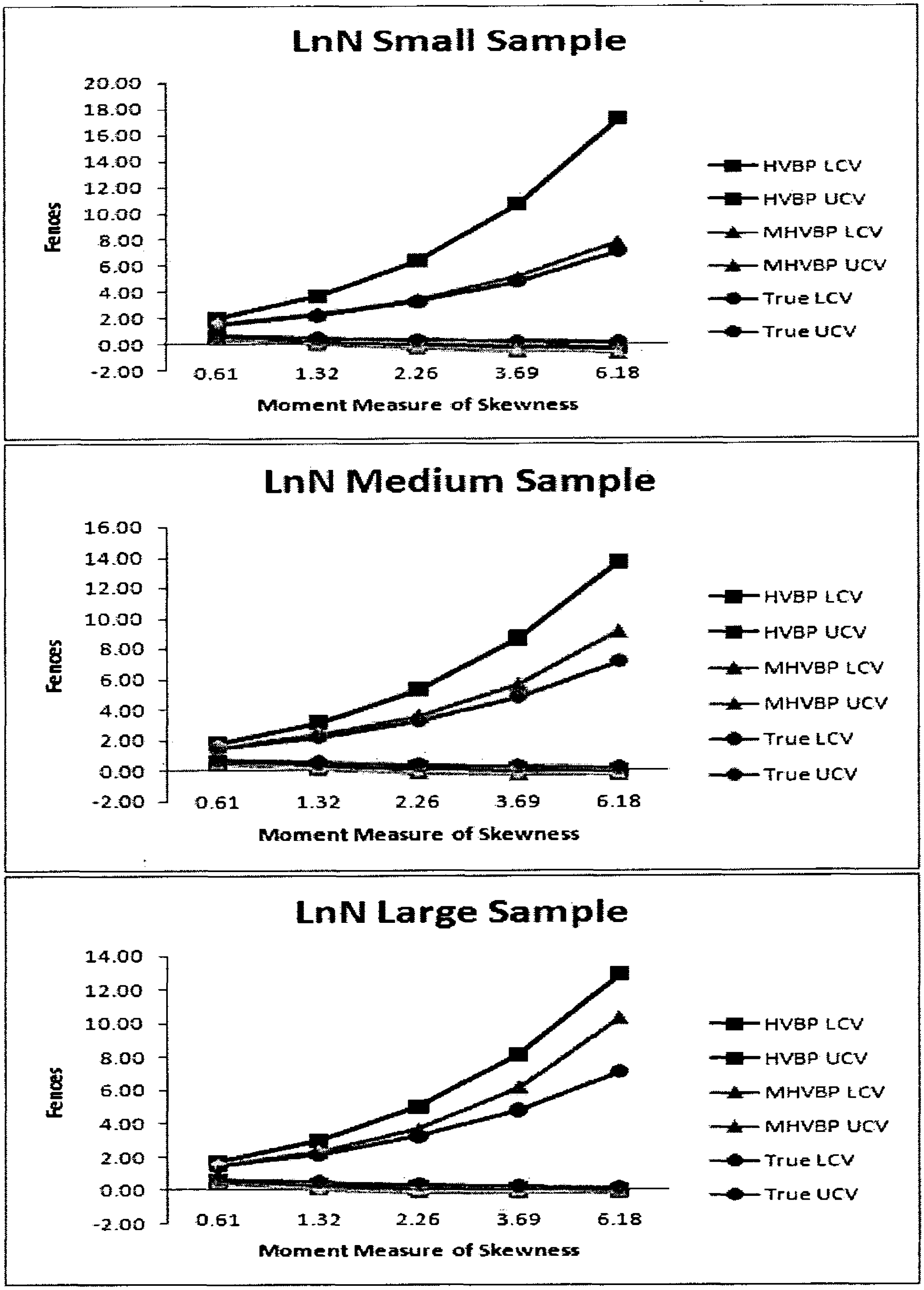


Table 5.3 Fences of HVBP and MHVBP Techniques and 95% True Boundary in Lognormal Distribution

Sample Size	Moment Measure of Skewness		0.61	1.32	2.26	3.69	6.18
True Lower Fence (2.5%)			**0.68**	**0.46**	**0.31**	**0.21**	**0.14**
25	Lower Critical Value	HVBP	0.44	0.12	-0.07	-0.19	-0.27
		MHVBP	0.49	0.06	-0.27	-0.51	-0.66
100		HVBP	0.55	0.28	0.11	0.00	-0.07
		MHVBP	0.49	0.09	-0.15	-0.25	-0.27
500		HVBP	0.57	0.31	0.15	0.04	-0.04
		MHVBP	0.49	0.11	-0.08	-0.09	-0.08
True Upper Fence (97.5%)			**1.48**	**2.19**	**3.24**	**4.80**	**7.10**
25	Upper Critical Value	HVBP	1.98	3.64	6.37	10.67	17.21
		MHVBP	1.58	2.32	3.41	5.11	7.78
100		HVBP	1.78	3.11	5.28	8.66	13.69
		MHVBP	1.57	2.33	3.54	5.64	9.12
500		HVBP	1.73	2.98	5.01	8.17	12.89
		MHVBP	1.57	2.34	3.66	6.20	10.32

Figure 5.3 shows the fence constructing style of the HVBP and MHVBP techniques in lognormal distribution. For small sample size, on the lower side of the distribution, fences of HVBP and MHVBP overlap with the true lower fence while for the upper side of the distribution, true upper fence and fence of MHVBP overlap and HVBP fence has a wide gap from the true fence. For the medium and large sample size, lower fences of HVBP and MHVBP approximate the true lower fence while on the upper tail MHVBP has a significant improvement over HVBP.

Figure 5.4 HVBP and MHVBP Technique Fences Matching with True 95% Boundary in Lognormal Distribution



LnN Small Sample

LnN Medium Sample

LnN Large Sample

4.11 Conventional Approach: Best and Worst Case in Context of Percentage Outliers

As already discussed, detection of percentage of outliers and constructing the fence are nearly the same things but for the sake of percentage of outliers detected, this study presents the comparison of the percentage of outliers detected by both techniques in this section. From the above figures 5.2, 5.3 and 5.4, it is evident that MVHB has constructed fences accurately around the true 95% boundary in chi square distribution with small sample size. So for the best case, chi square distribution with small sample size has been selected while for the worst case, β distribution with medium sample size has been selected. Here both the techniques are based on the Tukey's technique as already discussed and a technique detecting outliers approaching 2.5 will be treated better. It is observed that on the left side of the distribution, MHVBP has percentage much closer to 2.5% as compared to HVBP technique while for the right side of the distribution, at smaller level of skewness both techniques are performing equally and for the high skewness HVBP perform better.

Table 5.4 Percentage Outliers Detected by HVBP and MHVBP Techniques in Chi Square Distribution Small Sample Size

Degree of Freedom	Moment Measure of Skewness	Left Outliers HVBP	Left Outliers MHVBP	Right Outliers HVBP	Right Outliers MHVBP
25	0.57	5.13	1.73	4.31	3.95
20	0.63	5.37	1.63	4.21	4.03
15	0.73	5.28	1.49	4.11	4.26
10	0.89	5.45	1.20	3.82	4.36
2	2	4.97	0.52	2.69	4.75

For the worst case, medium size of the β distribution has been selected and has almost same performance in both medium and large sample sizes. Here in table 5.5, it can be observed that HVBP is performing better as its percentage is approaching 2.5 percent on the left side of the distribution while for right side MHVBP has almost the same performance as HVB has but by looking at total percentage MHVBP looks performing better. So it can be concluded that in worst case MHVBP's performance equals HVBP technique.

Table 5.5 Percentage Outliers Detected by HVBP and MHVBP Techniques in β Distribution Medium Sample Size

Parameters	Moment Measure of Skewness	Left Outliers HVBP	Left Outliers MHVBP	Right Outliers HVBP	Right Outliers MHVBP
(35,2)	-0.62	2.09	4.48	4.11	0.12
(35,3)	-0.49	2.42	4.34	4.19	0.27
(35,4)	-0.40	2.64	4.17	4.25	0.45
(35,5)	-0.35	2.82	4.02	4.26	0.64

5.10 Artificial Outlier Example

Twenty five numbers [10.52, 12.29, 12.75, 13.04, 14.72, 14.84, 15.01, 17.51, 17.87, 18.09, 18.94, 19.15, 19.82, 21.34, 21.54, 23.51, 25.21, 26.51, 27.08, 29.55, 29.73, 30.15, 31.35, 33.13, and 34.01] have been generated from chi square distribution with 20 degree of freedom and last 3observations have been replaced with 3 outliers 40, 55, and 70 on the right side of the distribution. Then by applying both the techniques gave the following results.

Table 5.6 One Sided Artificial Outliers Detected by HVBP and MHVBP

	Left Outliers	Right Outliers	LCV	UCV
HVBP	1	0	11.83	72.80
MHVBP	0	2	9.00	52.49

Here HVBP has wrongly extended the fence towards right side and has detected outliers from the left which is not outlier. On the other hand it could not detect the inserted mild, medium and big outliers in the data on right side of the distribution. Standard deviation of the data (including outliers) is 13.63 and 3SD fence on right side is 65.07 while the fence constructed by HVBP is 72.80 which is even away from 3.5SD. In contrast MHVBP has detected the medium and big outlier while it could not detect the mild outlier and no outlier on the left side of the distribution. Also it is observed that fence of HVBP is close to the real observations.

Again by replacement of the extreme observation on both sides by the outlier -20 and 60 and application of both techniques to detect outliers gave the following results.

Table 5.7 Two Sided Artificial Outliers Detected by HVBP and MHVBP

	Left Outliers	Right Outliers	LCV	UCV
HVBP	1	0	11.38	66.38
MHVBP	1	1	3.56	40.67

Here it can be observed that HVBP has detected left outlier accurately while it could not detect the right outlier which is even bigger than left outlier. Looking at the fence, it is clear that HVBP has erroneously extended the fence on right side while MHVBP has detected outliers accurately on both sides of the data; also its fence seems better than HVBP over the data set.

5.11 Conclusion

On the basis of above discussion it can be concluded that for all sample sizes, HVBP constructs a wider fence outside the true 95% distributional boundary on the extended side of the distribution and chances of Type II error are increased while MHVBP performs better. With the increase in sample size, performance of HVBP improves a bit on the compressed side of the distribution as compared to MHVBP. At all levels of moment measure of skewness, performance of HVBP is not good as compared to MHVBP that performs efficiently in all sample sizes and smaller levels of skewness. Generally HVBP over adjusts the fence while MHVBP constructs smaller fence around the 95% true distributional boundary and shows greater power to construct fence around the true 95% fence. So finally it can be inferred that MHVPBP is a good modification and shows a significant improvement on the HVBP technique.

CHAPTER 6

MEDCOUPLE BASED SPLIT SAMPLE SKEWNESS ADJUSTED TECHNIQUE

6.1 Introduction

The main purpose of this thesis is to introduce a technique which constructs the fence accurately to identify the outliers in the data set efficiently. As it has been already discussed, the technique constructing fence around the 95% true boundary of the distribution should be treated as performing better or in other words it will detect possible outliers in the data set efficiently. A new technique for outlier detection has been devised and discussed in chapter 4 and a modification in HVBP is proposed and discussed in chapter 5. It has been shown that in both chapters 4 & 5, the devised technique and proposed modification has outperformed to the rest of existing techniques. But by deep look at the figures of the fences in chapter 4 and chapter 5, in spite of the fact that our techniques construct fences close to true 95% fence, the constructed fences are still away from the from the target. This study aims at the construction of a technique that overlaps the true fence of 95% of the distribution. In search of our target for the best technique and sophistication of results matching with true fence, medcouple is incorporated in SSSBB technique that is introduced in chapter 4. Although this technique is difficult to apply without computer programming like HVBP and MHVBP techniques, it constructs the fence nicely around the true 95% fence of the distribution. We have incorporated MC in SSSBB technique in the similar fashion as in MHVBP (chapter 5), so it is named split sample skewness and medcouple based boxplot henceforth referred to as MCSSSBB.

6.2 Proposed Modification

The SSSBB technique was designed on the octile basis as

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

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

Where L is the lower critical value and U is upper critical value of the data. An observation outside these boundaries [L U] would be labeled as outlier. The medcouple is the exponential power times the classical skewness with $1.5 * IQR_L$ and $1.5 * IQR_R$ A restriction is imposed rather heuristically that if skewness is greater than 2, it should be treated as 2 selecting this number by hit and trial method because when skewness exceeds 2 it enlarges the interval of critical values and interval width becomes greater leading to an alteration in the parameters and watering down of the efficiency of the test. Mathematically it can be written as

$$[L \quad U] = [Q_{1L} - 1.5 * IQR_L * e^{-|Sk|^{\dagger} * MC} Q_{3R} + 1.5 * IQR_R * e^{|SK|^{\dagger} * MC}] If\ MC \leq 0$$

$$[L \quad U] = [Q_{1L} - 1.5 * IQR_L * e^{|SK|^{\dagger} * MC} Q_{3R} + 1.5 * IQR_R * e^{-|SK|^{\dagger} * MC}] If\ MC \geq 0$$

Where $|SK|^{\dagger} = \begin{cases} 2\ if\ |Sk| \geq 2 \\ |Sk|\ if\ |SK| \leq 2 \end{cases}$

Where SK is the moment measure of skewness and MC is the medcouple.

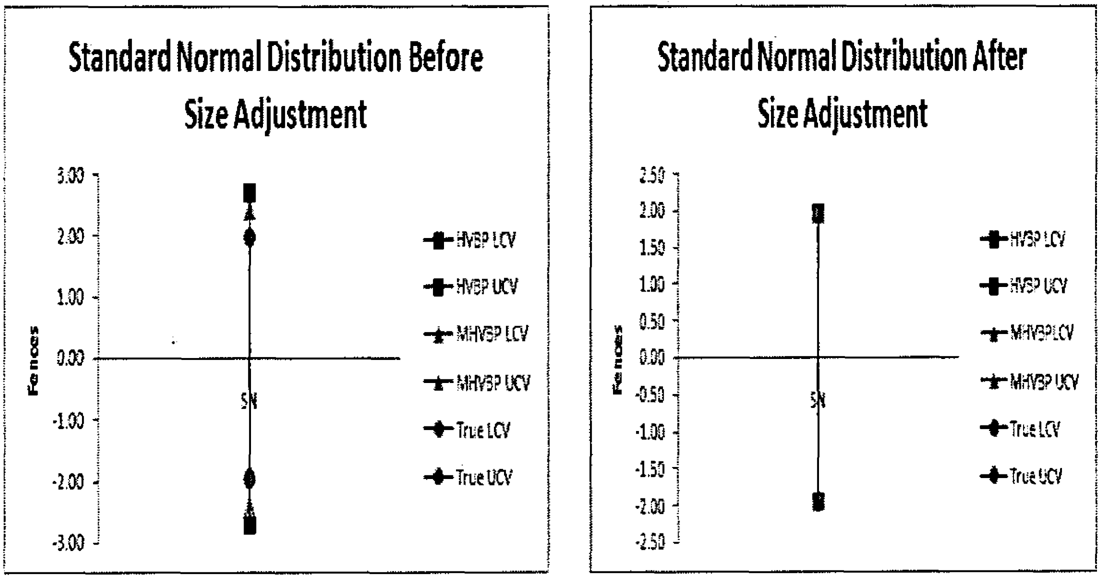
6.3 Monte Carlo Simulation Study

A simulation study has been done for the verification of the claim. The χ^2, β and lognormal distributions are used for this purpose with different sample sizes and different parameters. Sample size has been taken equal to 25, 50 100 500 in all the distributions while experiment has been done on the χ^2 with 2 , 10, 15, 20 and 25 degree of freedom while for the β distribution, the parameters are (35,2), (35,3), (35,4), (35,5) and for the lognormal are (0,0.2), (0,0.4), (0,0.6), (0,0.8), (0,1). One hundred thousand replications have been made for each case that has been done in the Matlab software and a total of nine statistics have been computed as discussed in 5 but just 2 has been used for the analysis purpose.

6.4 Comparison of Fences Produced by MCSSSBB and HVBP Techniques with True 95 Percent Central Boundary

Comparison of these two techniques is reasonable because first both techniques have the same size at 95% level because first HVBP is modification in Tukey's boxplot while MCSSSBB is modification in SSSBB. Both techniques become Tukey's and SSSBB respectively when data are symmetric. We have matched their sizes in chapter 4 section 4.6 at 95% central values of the normal distribution. Since medcouple has been incorporated in the SSSBB technique and HVBP is specially designed for the skewed distributions, these techniques will be compared with respect to fence construction around the true 95 percent central values of the distributions under consideration. Same methodology has been adopted here as discussed in detail in chapter 4 sections 4.5 and chapter 5 and section 5.5 respectively.

Figure 6.1 HVBP and MCSSSBB Technique Fences vs. 95% Boundaries in Standard Normal Distribution



HVBP is based on the Tukey's technique while MCSSSBB is based on SSSBB. When the distribution is fairly symmetric, the HVBP approaches Tukey's technique while MCSSSBB approaches SSSBBB exactly. Following equations clearly show when MC equals zero (in case of symmetric distributions), both equations for HVBP approach Tukey's technique

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

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

As the distribution is symmetric for the MHVBP technique, MC and skewness will be zero approaching the equations below to the SSSBB technique exactly.

$$[L \quad U] = [Q_{1L} - 1.5 * IQR_L * e^{-|Sk|^{\dagger}*MC} Q_{3R} + 1.5 * IQR_R * e^{|SK|^{\dagger}*MC}] If\ MC \leq 0$$

$$[L \quad U] = [Q_{1L} - 1.5 * IQR_L * e^{|SK|^{\dagger}*MC} Q_{3R} + 1.5 * IQR_R * e^{-|SK|^{\dagger}*MC}] If\ MC \geq 0$$

Where $|SK|^{\dagger} = \begin{cases} 2\ if\ |Sk| \geq 2 \\ |Sk|\ if\ |SK| \leq 2 \end{cases}$

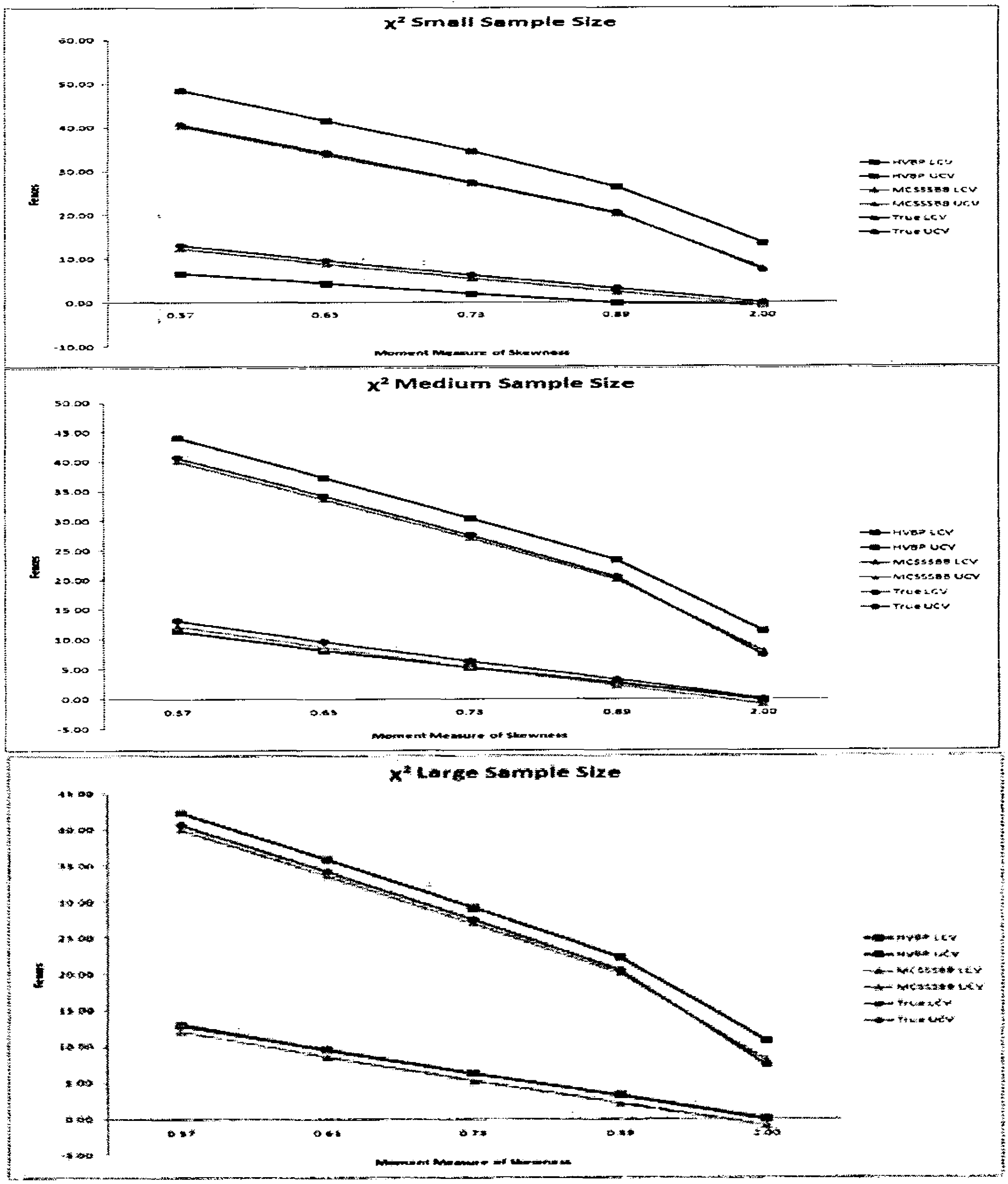
As sizes of both techniques (Tukey and SSSBB) have been matched with 95% true boundary of standard normal distribution, the value of g is taken as 0.95 and 0.97 for Tukey and SSSBB respectively. The size match of both techniques is shown in the above figure 6.1. So for comparison of power of both techniques we shall use the same value of g as mentioned above.

Table 6.1 Fences of HVBP and MCSSSBB Techniques and 95% True Boundary in χ^2 Distribution

Sample Size	Moment Measure of Skewness		0.57	0.63	0.73	0.89	2.00
True Lower Fence (2.5%)			**13.12**	**9.59**	**6.26**	**3.25**	**0.05**
25	Lower Critical Values	HVBP	6.56	4.25	1.93	-0.15	-0.57
		MCSSSBB	12.26	8.68	5.37	2.36	-0.67
100		HVBP	11.34	8.13	5.25	2.64	-0.13
		MCSSSBB	12.15	8.61	5.25	2.21	-0.86
500		HVBP	12.92	9.53	6.23	3.24	-0.08
		MCSSSBB	12.10	8.56	5.22	2.17	-0.94
True Upper fence (97.5%)			**40.65**	**34.17**	**27.49**	**20.48**	**7.38**
25	Upper Critical Values	HVBP	48.43	41.52	34.46	26.41	13.59
		MCSSSBB	40.17	33.60	27.07	20.15	7.76
100		HVBP	44.02	37.34	30.35	23.31	11.45
		MCSSSBB	40.05	33.60	26.97	20.07	7.99
500		HVBP	42.30	35.84	29.11	22.24	10.81
		MCSSSBB	39.99	33.56	26.95	20.06	8.12

Figure 6.1 (top) shows the interval fitting pattern of χ^2 distribution in small sample size around the 95% values. For small sample sizes, it can be observed that on both lower and upper sides of the distribution, fence of MCSSSBB is close to true fence at all levels of skewness. For medium sample sizes, fences of both techniques almost overlap on the lower side and have same distance from the true lower fence and are very close to true lower fence.

Figure 6.2 HVBP and MCSSSBB Technique Fences Matching with True 95% Boundary in χ^2 Distribution



On the upper side, MCSSSBB fence is close to the true fence as compared to HVBP upper fence. For the large sample size, on the lower side of the distribution, fence of HVBP is bit close to the true lower fence as compared to the MCSSSBB techniques; on the upper side MCSSSBB is very close to the true upper fence. The problem of over adjusting of fence outside the data of HVBP technique is clear in all the cases of small, medium and large sample sizes.

Table 6.2 Fences of HVBP and MCSSSBB Techniques and 95% True Boundary in β Distribution

Sample Size	Moment Measure of Skewness		-0.35	-0.40	-0.49	-0.62
True Lower Fence (2.5%)			0.76	0.79	0.82	0.85
25	Lower Critical Values	HVBP	0.68	0.71	0.67	0.70
		MCSSSBB	0.83	0.87	0.79	0.83
100		HVBP	0.72	0.76	0.72	0.75
		MCSSSBB	0.83	0.87	0.79	0.83
500		HVBP	0.73	0.76	0.73	0.76
		MCSSSBB	0.83	0.87	0.79	0.83
True Upper Fence (97.5%)			0.96	0.97	0.98	0.99
25	Upper Critical Values	HVBP	1.01	1.01	1.02	1.02
		MCSSSBB	0.99	1.00	1.01	1.02
100		HVBP	1.00	1.01	1.01	1.01
		MCSSSBB	0.99	1.00	1.01	1.02
500		HVBP	1.00	1.01	1.00	1.01
		MCSSSBB	0.99	1.00	1.01	1.02

Figure 6.2 (top) shows the fence designing pattern of MCSSSBB and HVBP techniques around the true fence for small sample size in β distribution. For the entire sample sizes it is observed that lower fence of MCSSSBB is close to the true fence as compared to HVBP's lower fence. Here the problem of over adjusting the fence by HVBP is solved. Also on the upper side of the β distribution for small sample size, fence of MCSSSBB is close to the true fence as compared to the HVBP upper fence. For medium and large sample sizes, the fences of both techniques show almost same pattern as the HVBP constructs a wider fence as compared to MCSSSBB. In β distribution, overall performance of MCSSSBB is better than HVBP which can be observed from the figure 6.3.

Figure 6.3 HVBP and MCSSSBB Technique Fences Matching with True 95% Boundary in β Distribution

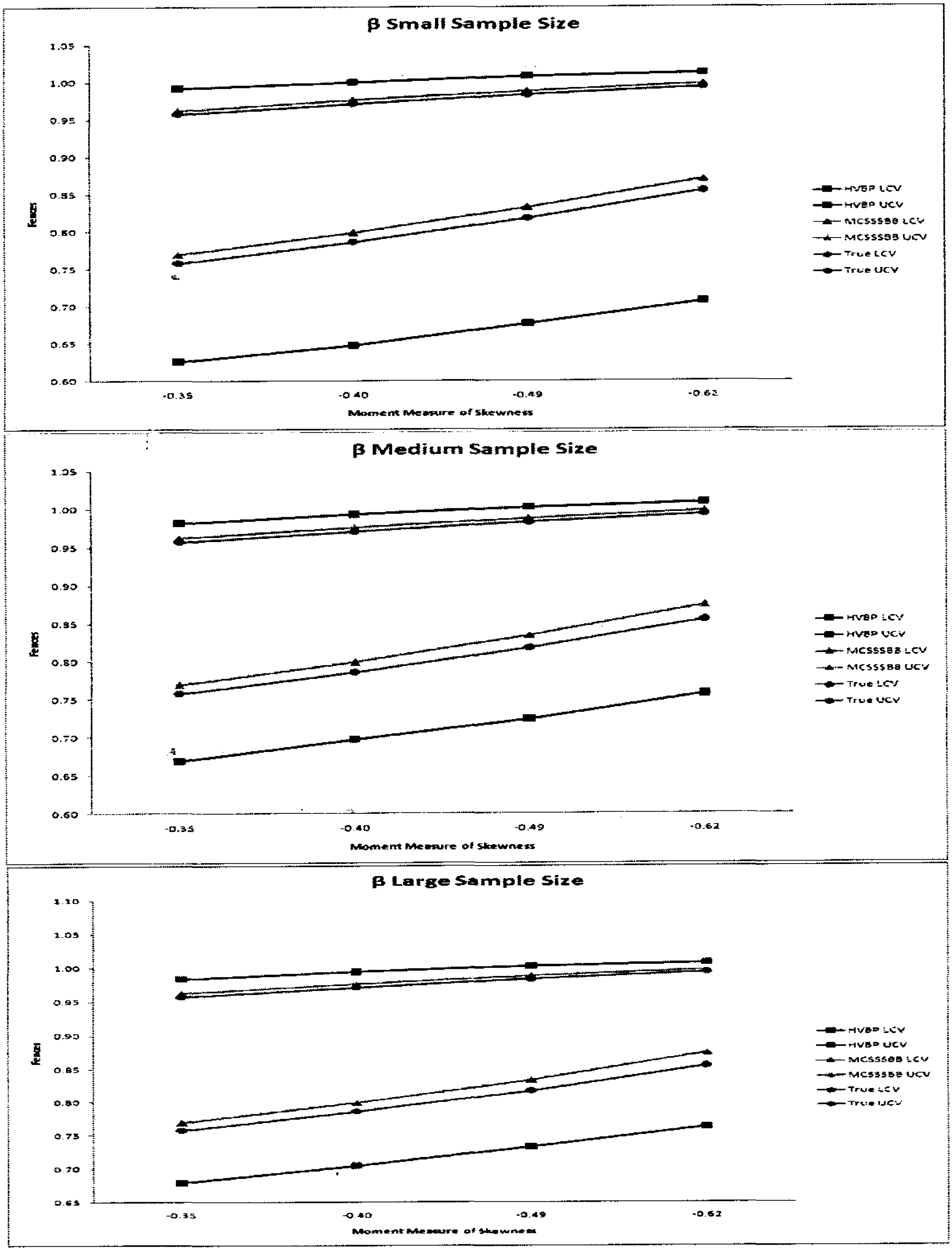
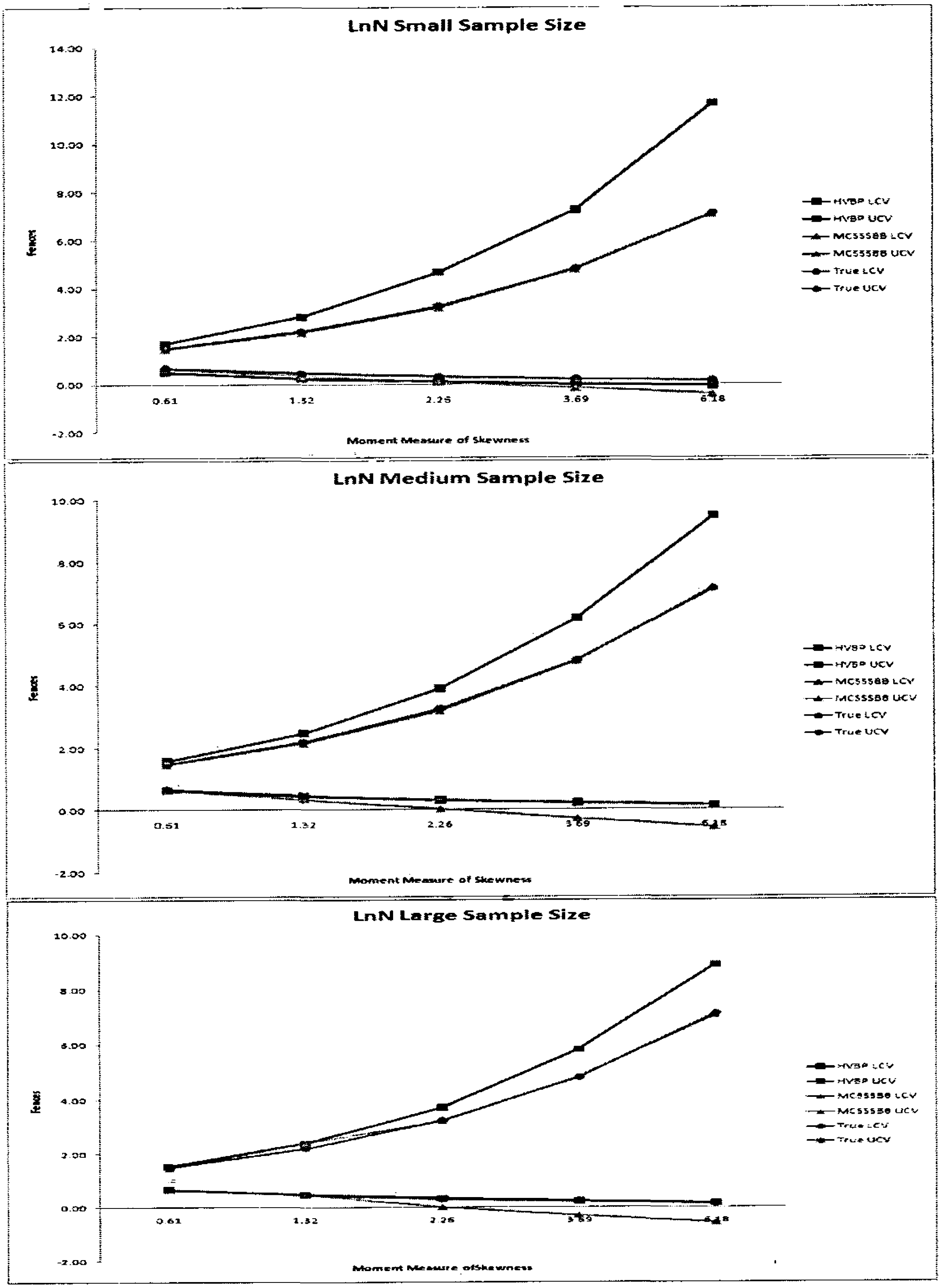


Table 6.3 Fences of HVBP and MHVBP Techniques and 95% True Boundary in Lognormal Distribution

Sample Size	Moment Measure of Skewness		0.61	1.32	2.26	3.69	6.18
True Lower Fence (2.5%)			**0.68**	**0.46**	**0.31**	**0.21**	**0.14**
25	Lower Critical Values	HVBP	0.09	-0.01	-0.08	-0.19	-0.27
		MCSSSBB	0.08	-0.17	-0.41	-0.49	-0.83
100		HVBP	0.31	0.22	0.14	0.00	-0.07
		MCSSSBB	0.03	-0.28	-0.55	-0.64	-1.03
500		HVBP	0.34	0.24	0.16	0.04	-0.04
		MCSSSBB	0.00	-0.30	-0.56	-0.69	-1.05
True Upper Fence (97.5%)			**1.48**	**2.19**	**3.24**	**4.80**	**7.10**
25	Upper Critical Values	HVBP	4.67	7.24	11.64	10.67	17.21
		MCSSSBB	3.17	4.78	7.06	5.98	9.37
100		HVBP	3.89	6.16	9.45	8.66	13.69
		MCSSSBB	3.17	4.78	7.16	6.04	9.30
500		HVBP	3.72	5.80	8.87	8.17	12.89
		MCSSSBB	3.22	4.81	7.04	6.07	9.15

Figure 6.4 shows the fence construction pattern of HVBP and SSSBB techniques for the lognormal distribution for small sample size .On the lower side of the distribution, fences created by both techniques nearly overlap and on the upper side of the distribution, fence of HVBP is farther away from the true upper fence as compared to MCSSSBB. For the medium and large sample sizes it can be observed that upper fences of MCSSSBB are closer to the true upper fence while on the lower side fence of HVBP almost overlap the true fence. As a whole, the upper and lower fence of MCSSSBB seems to be closer to true fence than the fences produced by HVBP.

Figure 6.4 HVBP and MCSSSBB Technique Fences Matching with True 95% Boundary in Lognormal Distribution



For the large sample sizes, lower side of the lognormal distribution in the presence of high level of skewness, MCSSSBB seems to go away from the true fence while it still performs better than the performance of HVBP technique on the upper side of the distribution.

6.5 Conventional Approach: Best and Worst Case in Context of Percentage Outliers

In this section, this study presents the percentage of outliers detected by HVBP and MCSSSBB. As already discussed, comparison of percentage of outliers and fence comparison are the same. Here for the best and worst cases, β distribution with small size has been selected and lognormal distribution with medium sample size. It can be observed that on the left side of the distribution there is bit difference between percentage of outliers detected by both techniques under comparison and with the increase of skewness performance of MCSSSSBB is becoming better as percentage of outliers approach 2.5 percent. On the right side of the β distribution it is clear that MCSSSBB has performed better than HVBP as its percentage of outliers detected are approaching to 2.5 percent.

Table 6.4 Percentage Outliers Detected by HVBP and MCSSSBB Techniques in β Distribution Small Sample Size

Parameter	Moment Measure of Skewness	Left Outliers HVBP	Left Outliers MCSSSBB	Right Outliers HVBP	Right Outliers MCSSSBB
(35,2)	-0.62	2.97	3.70	4.55	1.52
(35,3)	-0.49	3.10	3.37	4.64	1.69
(35,4)	-0.40	3.29	3.14	4.59	1.91
(35,5)	-0.35	3.41	3.07	4.56	1.93

Now looking for the worst case, on left side it can be seen that MCSSSBB has detected percentage of outliers less than 2.5 percent while HVBP has detected more than 2.5percent. By looking at the difference from 2.5 it can be concluded that MCSSSBB is close to the 2.5 percent. For the right side of the distribution it can be seen that for the smaller level of skewness, HVBP is better than MCSSSBB. However by increasing the skewness, percentage of outliers detected by MCSSBB approaches to 2.5 percent.

Table 6.5 Percentage Outliers Detected by HVBP and MCSSSBB Techniques in Lognormal Distribution

Parameter	Moment Measure of Skewness	Left Outliers HVBP	Left Outliers MCSSSBB	Right Outliers HVBP	Right Outliers MCSSSBB
(0,0.2)	0.61	3.63	1.65	2.69	3.04
(0,0.4)	1.32	4.13	0.58	2.32	3.07
(0,0.6)	2.26	4.37	0.08	2.12	2.92
(0,0.8)	3.69	4.59	0.01	2.01	2.79
(0,1)	6.18	4.51	0.00	1.97	2.78

6.6 Conclusion

In χ^2 distribution, MCSSSBB techniques construct fence accurately on the 95% central values of the distribution as compared to HVBP fence. For all the sample sizes of β distribution, MCSSSBB technique performs better on both sides with respect to the fences constructed around the true 95% boundary of the distribution. In lognormal distribution, for small sample size both techniques under comparison perform nearly equal on lower side while for the medium and large sample sizes, HVBP performs better on lower side as compared to MCSSSBB technique. For the upper side of the distribution, in all sample sizes MCSSSBB outperforms as compared to HVBP. Actually in the modification proposed in this chapter we have tried to develop a technique which constructs the fence accurately around the central 95% of the distributions but we observe that some time HVBP performed better while maximum time performance of MCSSSBB performed better.

CHAPTER 7

APPLICATIONS

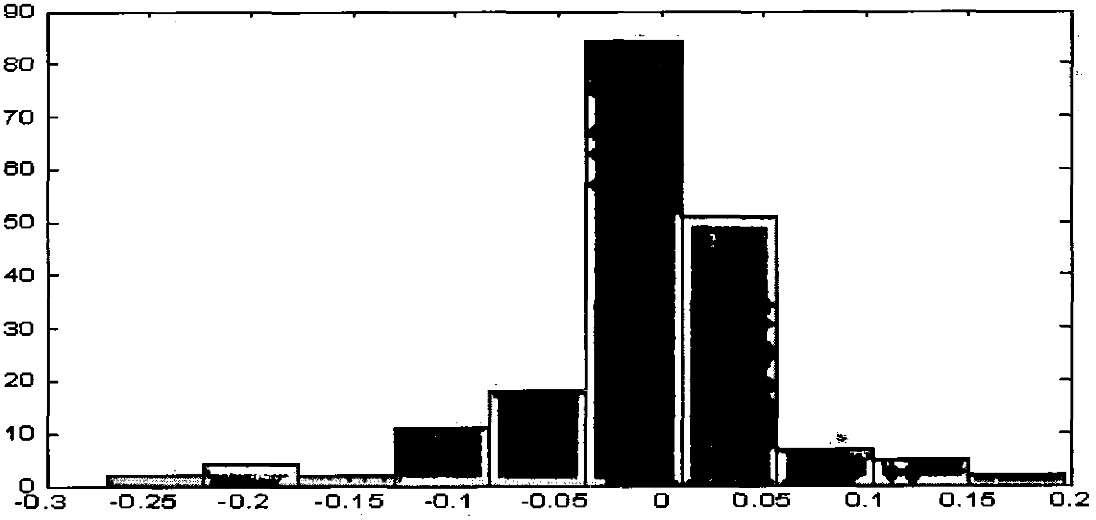
Summary

The newly introduced technique SSSBB in chapter 4, proposed modification MHVBP in HVBP technique in chapter 5, Modification proposed in SSSBB in chapter 6 and the existing techniques (Tukey's technique and HVBP) have been applied to the real data sets. Two skewed data sets have been taken to test the performance of the tests in real life. Section 7.1 deals with the data set of the stock return from Karachi Stock Exchange (KSE) of the United Trust of Pakistan (UTP-2008) for daily return while section 7.2 deals with baby birth weight data collected at Agha Khan Hospital Karachi (Pakistan).

7.1 Stock Return Data Set

Data for daily stock return of United Trust of Pakistan (UTP) Large Cap (2008) from Karachi stock exchange (KSE) are analyzed and both tests are applied for the identification of the outliers. Histogram for the stock returns clearly shows that it is skewed towards left and its classical skewness is nearly -1. Visually it seems that there are no outliers in the data set so that performance of the test detecting less number of extreme observations as outliers will be treated better.

Figure 7.1 **Histogram for the Stock Returns UTPL for the year 2008**



Data consists of a total of 186 observations of the stock prices in the whole year. Tukey's box plot detected 34 outliers (20 from left and 14 from right) constructing the interval of width 0.1563. It can be said that Tukey's method has detected more than 18 percent of observations as outliers. Tukey's has thus detected more observations on the skewed side and lesser on the compressed side as per its nature.

Table 7.1 **Synchronized Left and Right Outliers**

Tukey's Technique				**SSSBB Technique**	
Negative Return Dates		**Positive Return Dates**		**Negative Return Dates**	**Positive Return Dates**
18-Dec-08	23-Jun-08	7-Aug-08	19-Dec-08	18-Dec-08	18-Aug-08
17-Dec-08	16-Jul-08	1-Sep-08	5-Sep-08	17-Dec-08	29-Dec-08
23-Dec-08	23-May-08	23-Sep-08	8-Oct-08	23-Dec-08	22-Jul-08
16-Dec-08	14-Jul-08	27-May-08	26-Dec-08	16-Dec-08	4-Jun-08
30-Dec-08	12-Aug-08	25-Jun-08		30-Dec-08	19-Dec-08
22-Dec-08	28-May-08	24-Jun-08		22-Dec-08	5-Sep-08
9-Oct-08	17-Jul-08	18-Aug-08			8-Oct-08
4-Sep-08	7-Oct-08	29-Dec-08			26-Dec-08
10-Mar-08	19-Sep-08	22-Jul-08			
26-Aug-08	20-Aug-08	4-Jun-08			

It is observed that Tukey's has detected nearly 11% observations as outliers on the left side of the distribution and nearly 7 percent on the right side. Split sample skewness adjusted technique has detected just 14(6 from left and 8 from right) outliers constructing interval of width 0.2639. The left six observations (dates) are from December 16-30 which are the same dates in which Mohtarma Benazir Bhutto (Ex prime minister of Pakistan) came to Pakistan and was assassinated one year earlier. On the right, the outliers comprise 3 dates from the same period and it can be said that the maximum fluctuations are during the period of her 1st death anniversary (9 out of 14 outliers). The rest of the outliers relate to the period when ex- president General Ret. Pervaiz Musharaf resigned from his office and Mr. Zardari became the president and next day after 5th September is "Defense Day". One outlier is from the month of June which is near the annual budget days. Here it is known that these are real observations and being at the extremes tell the story of the assassination of Mohtarma Benazir Bhutto to a researcher who is not so familiar with the history. All the negative returns are from the December which shows violence and agitation following the assassination of Mohtarma Benazir Bhutto.

But Tukey's test detects 34 outliers out of 186 observations (roughly 18% of the data). Only a visual analysis of the data is enough to convince that all the bins of the histogram are joined and no extreme outliers exist. However the SSSBB technique detects nearly 7.5 % of the data as outliers. The UCV and LCV are approaching the maximum and minimum of the data in the SSSBB technique and they extend too far away from the original data in Tukey's method. The below given table 7.2 shows the different statistic for outliers in stock return data.

Table 7.2 Outliers and IW for all Techniques in Stock Return Data

Technique	Left OL*	Right OL	Total OL	LCV**	UCV	Interval Width
Tukey	20	14	34	-0.08	0.08	0.16
SSSBB	6	8	14	-0.16	0.10	0.26
HVBP	19	14	33	-0.08	0.07	0.16
MHVBP	20	14	34	-0.08	0.08	0.16
MCSSSBB	6	8	14	-0.16	0.10	0.26

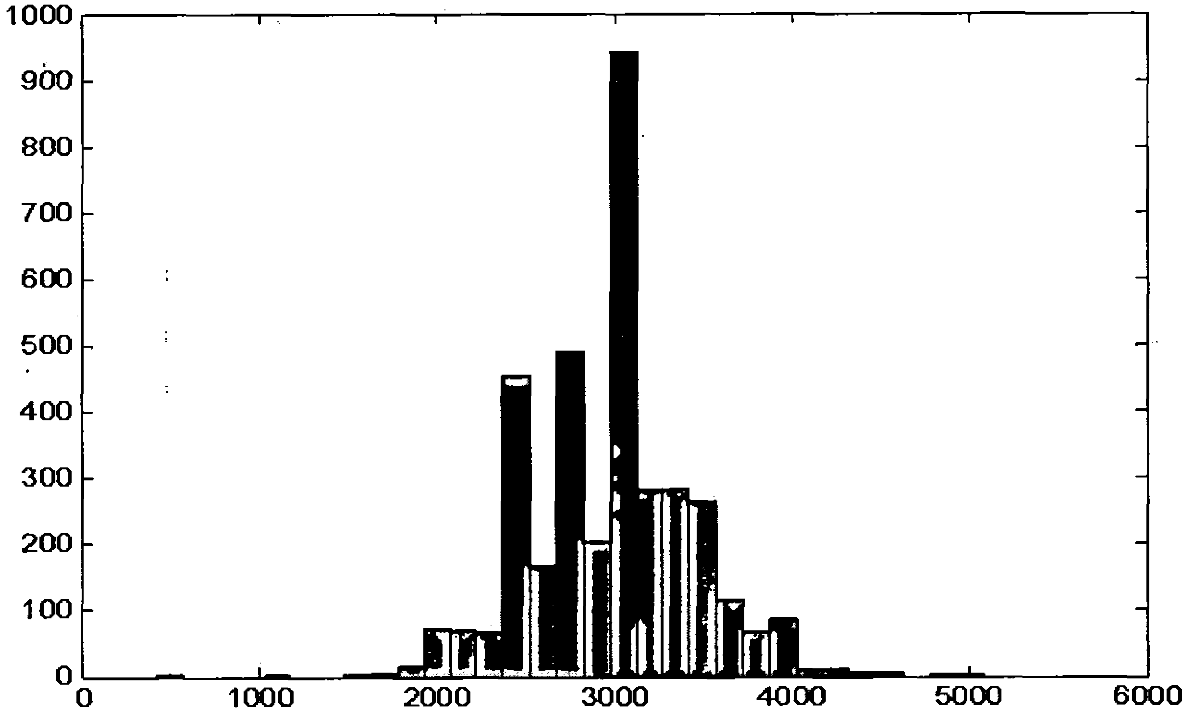
*OL Outliers **LCV Lower Critical Value; UCV Upper Critical Value

7.2 Baby Birth Weight Data

Data for baby birth weight has been taken from Agha Khan Hospital Karachi. Here our assumption is that survival of the baby depends upon his/her birth weight. So an underweight newborn baby is more vulnerable to mortality as compared to a baby with higher birth weight. According to McIntire et al. (1999), infants born with low birth weight are more likely to die or succumb to morbidity. Vangen et.al (2002) proved that heavier is better. Babies with low birth weight, either due to short gestation period or because of fetal growth constraint, are at high risk for short- and long-term disabilities and death (Schieve et al, 2002). Checkup of very low birth weight children points toward increased deaths among all subpopulations.

There is a consensus on the point that socioeconomic conditions of the family and educational background, especially mother's education has a great role in the survival of the infant. Also medical facilities have been so much improved that a baby with very low birth weight might survive by availing these facilities and a baby with relatively higher birth weight from low income family might not due to unavailability of medical facilities. But as already mentioned the data have been taken from the similar income groups (people going to Agha khan hospital are well off and from the educated families and can bear any cost in monetary terms for survival of their baby). Agha Khan Hospital is one of the most efficient hospitals having latest facilities and equipment as compared to public sector hospitals. So it is assumed that data belongs to similar group with respect to income and education and is comparable. Keeping other things constant, the probability of the survival increases as birth weight increases and vice versa.

Figure 7.2 Histogram for Baby Birth Weight



Here data are of 3613 observation of baby birth weight along with their follow up data till 4th week (28th day). Minimum weight is 500grams and the highest weight is 5000garms. Average weight is 2974grams (nearly 3kg) and total deaths up to 4th week are 19.Mortality among the total population is just 0.5%. According to definition of low birth weight, an infant having birth weight less than 2500 grams is treated as low weight. Our data itself proves our claim that low birth weight babies have more chances of mortality, as it is observed that Tukey's technique has detected 26 as left outliers while our proposed technique SSSBB has detected 16 observations on left side as outlier. By mining into data it can be observed that there are five deaths in both cases (either in Tukey's or SSSBB). So it is concluded that just 0.7% data (by Tukey's technique) and 0.4% data (by SSSBB) captures more than 25% of the deaths from the whole data set. This finding corroborates the claim that birth weight has a very close relation with mortality. Secondly it shows the improvement of our test on Tukey's as Tukey's technique detected same number of deaths from 0.7% of the data while SSSBB from 0.4% .Our test is performing more efficiently than Tukey's does.

7.3 Comparison of Tukey's Technique and SSSBB in Baby Birth Weight Data

According to the assumption that birth weight has close relation with the survival, the babies with higher birth weight are more likely to survive than low birth weight babies. For this purpose, left outliers for the mortality should be compared. Summary of the data are as under:

Table 7.3 Summary of Baby Birth Weight Data

Observations	Mean(grams)	SD	Minimum	Maximum	Survivals	Deaths
3613	2974	445	500	5000	3594	19

Left outliers detected by Tukey's technique are 26 while left outliers detected by SSSBB are 16. By analyzing the data with respect to left outliers it can be observed that there are 5 deaths in both cases (in 26 left outliers by Tukey and 16 left outliers by SSSBB). So it can be said that performance of Tukey's technique is 19.23% while the performance of SSSBB is 31.25%. As it can be seen that deaths are also inliers so the total number of deaths with respect to total number of outliers detected are compared. Tukey's has detected 111 outliers in total while SSSBB has detected 29 outliers so the performance of Tukey's as a whole is 17% while performance of SSSBB is 66%.

Table 7.4 Outliers and IW for all Techniques in BBW Data

Technique	Left OL	Right OL	Total OL	LCV	UCV	Interval Width
Tukey	26	85	111	1950	3950	2000
SSSBB	16	13	29	1900	4250	2350
HVBP	5	180	185	1621.10	3745.6	2124.5
MHVBP	26	85	111	1944.00	3944.0	2000.0
MCSSSBB	26	13	39	1904.80	4244.0	2339.2

BBW: Baby Birth Weight

Table 7.5 Performance Comparison in BBW Data

Technique	Left OL	Performance left outliers	Overall Performance
TUKEY	26	19.23%	17.12%
SSSBB	16	31.25%	65.52%
HVBP	5	60.00%	10.27%
MHVBP	26	19.23%	17.12%
MCSSSBB	26	19.23%	48.72%

In comparison of all techniques under consideration, it can be observed that HVBP is performing most efficiently among all the techniques under comparison by detecting just 5 left outliers and two deaths in these 5 outliers performing 60% while SSSBB seems to chase it by 31% performance. Since deaths are also inliers, so looking at total outlier's performance reveals that HVBP have detected 180 right outliers and its performance falls drastically to 10.27% while SSSBB improves its performance from 31.25% to 65.52% by just detecting 13 outliers on the right side leading all the techniques.

7.4 Cost Benefit Analysis

From the above section 7.3, it can be seen that some techniques detected less number of outliers in baby birth weight data while others a greater number of left outliers. It is observed that HVBP detected just 5 left outliers and SSSBB detected 16 left outliers while remaining techniques detected 26 outliers. Also it can be seen that statistically HVBP is more efficient than remaining techniques. This section includes the practical significance of the techniques under consideration.

Let us suppose that underweight babies are advised an intensive care for a week at Agha Khan Hospital and per day treatment expense is Rs.100,000/- per child. Then the total treatment expense of low weight babies detected by Tukey's, MHVBP and MCSSSBB techniques is Rs.18,200,000/- (as 26 outliers detected by all techniques) while for SSSBB and HVBP expenses are Rs. 11,200,000/- and Rs.3,500,000/- respectively. While comparing SSSBB technique with Tukey's, MHVBP and MCSSSBB, it is observed that there are five deaths in both cases and expense on the low weight babies detected by SSSBB is Rs. 7000000/- less than other techniques (excluding the opportunity cost of time of the parents and care takers). On the other hand, it can be seen that HVBP has detected just 5 low weight babies and their expense for one week is Rs.3500000/- (Rs. 7700000/- less than SSSBB) but the main issue arises here is of practical significance. As HVBP has detected 5 low weight babies, so just five babies will be given the intensive care while the remaining 2 babies will be ignored for treatment and are vulnerable to death (as five deaths in low weight babies detected by all other techniques). Keeping all aspects of monetary cost and human life, SSSBB seems to perform better than all other techniques in this data set.

CHAPTER 8

CONCLUSIONS AND RECOMMENDATIONS

8.1 Conclusions

The worth of our proposed technique and modifications made in different techniques for detecting outliers is demonstrated in the previous chapters analytically, by the Monte Carlo simulation and also graphically. It can be seen that there are a lot of problems associated with the Tukey's technique in skewed distributions. The HVBP technique for skewed distribution which based on the medcouple for generating the interval of critical values away from the true 95% fence of the univariate distributions is also not free from problems. The performance of the HV box plot is good for large sample sizes in skewed distribution, but it constructs a very large interval of critical values which takes it away from the true 95% boundary of the distribution. On the other hand, SSSBB performs well in spite of the fact that it is very simple and MHVBP is better than HVBP technique even for the larger samples. Performance of MCSSSBB is maximum time better than HVBP while sometime HVBP perform better than MCSSSBB.

8.2.1 Advantages of Split Sample Skewness Based Boxplot

This study has formulated a directional technique for detecting outliers. Due attention has been given to the shape of underlying distribution, i.e. for the skewed distributions, data on either side of centre is treated separately. Data coverage by this technique is very robust. This technique is very simple as compared to HVBP which uses medcouple that has complicated calculations. This technique is applicable in both large and small sample sizes.

8.3 Advantages of the Modified Hubert Vandervieren Boxplot

This technique detects lesser number of outliers from the random sample than adjusted box plot. Modified Hubert and Vandervieren boxplot generates smaller interval of critical values as compare to HVBP. The proposed modified test is useful for both small and large data sets and its fence for outlier's detection is very close to the true 95% boundary of the distribution for all the sample sizes.

8.4 Recommendations

When a researcher is interested in detecting outliers from a skewed data and also wants to get rid of the messy calculations involving a high computer power, one should use SSSBB instead of Tukey's technique as discussed in chapter 4. But if the researcher is interested in accuracy, MHVBP and MCSSSBB are better alternatives to use because with less skewness it acts like the SSSBB because the performance of the SSSBB is overall better than any other technique. For the real data sets, it is observed that for the left outliers, HVBP is good in its performance but overall performance of SSSBB is best for both left and right outliers. The performance of Tukey and MHVBP techniques are almost the same. Again it is recommended the MCSSSBB technique may be used for detection outliers in skewed distributions when the reader is interested in sophisticated technique and also sometime HVBP performance is better than MCSSSBB but if reader is interested in simple technique the SSSBB is better.

8.5.1 Future Work

Following work is proposed for the future with respect to outlier's detection techniques. These techniques can be designed based on both skewness and sample size. Also contaminated data sets can be analyzed by these techniques and Research can be extended from univariate to bivariate and multivariate. Outlier detection techniques can be designed based on mode instead of mean and median.

Bibliography

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. (1978). The Study of Outliers: Purpose and Model. *Applied Statistics , 27* (3), 242-250.

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data* (3rd ed.). Wiley.

Beckman, R. J., & Cook, R. D. (1983). Outlier..........s. *Technometrics , 25* (2).

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

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).

Choonpradub, C., & McNeil, D. (2005). Can the box plot be improved? *Songklanakarin J. Sci. Technol , 27* (3), 649-657.

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

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

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.

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

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

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

Gibbons, R. D., K.Bhaumic, D., & Aryal, S. (1994). *Statistical Methods for Groundwater Monitoring.* New York: John Wiley & Sons.

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

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

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.

Hair, J. F., Tatham, R. L., Anderson, R. E., & Black, W. (1998). *Multivariate Data Analysis* (5th ed.). Prentice Hall.

Hathcock, A., Silverman, P., Ferré, C., Reynolds, M., Schieve, L., & Drees, M. (2003). *Increasing Infant Mortality Among Very Low Birthweight Infants —Delaware, 1994--2000.* Morbidity and Mortality Weekly Report, Centers for Disease Control and Prevention.

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

Hinkley, D. (1977). On Quick Choice of Power Transformat. *Applied Statistics , 26* (1).

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

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

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

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

Justel, A., & Pena, D. (1996). Gibbs Sampling Will Fail in Outlier Problems with Strong Masking. *Journal of Computational and Graphical Statistics,, 5* (2), 176-189.

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.

Ludbrook, J. (2008). Outlying Observations and Missing Values:How Should They be Handled? *Clinical and Experimental Pharmacology and Physiology , 35*, 670-678.

Maimon, O., Rockach, L., & Bin-Gal, I. (2005). *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers,.* Tel-Aviv University Ramat-Aviv, Tel-Aviv 69978, Israel, Deparhent of Industrial Engineering. Kluwer Academic Publishers.

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

McIntire, D. D., Bloom, S. L., Casey, B. M., & Leveno, K. J. (1999). Birth Weight in Relation to Morbidity and Mortality Among Newborn Infants. *The New England Journal of Medicine , 340* (16), 1234-1238.

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).

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

Schieve, L. A., Meikle, S. F., Ferre, C., Peterson, H. B., Jeng, G., & Wilcox, L. S. (2002). Low and Very Low Birth Weight in Infants Conceived With Use of Assisted Reproductive Technology. *The New England Journal of Medicine , 346* (10), 731-737.

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.

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.

Tsay, R. S., Pena, D., & Pankratz, A. E. (2000). Outliers in Multivariate Time Series. *Biometrika , 87* (4), 789-804.

Tsay, R. S., Pena, D., & Pankratz, A. E. (2000). Outliers in Multivariate Time Series. *Biometrika , 87* (4), 789-804.

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

Vangen, S., Stoltenberg, C., Skjaevern, R., Magnus, P., Harris, J. R., & Stray-Pedersen, B. (2002). The Heaiver the Better: Birth Weight and Perinatal Mortality Different Ethnic Groups. *International Journal of Epidemiology*, *31*, 654-660.

Verma, S. P. (1997). sixteen statistical tests for Outlier Detection and Rejection in Evaluation of International Geochemical Reference Materials. *The Journal of Geostandards and Geoanalysis*, *21* (1), 59-75.

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.

APPENDIX

Previous Techniques

Grubs 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

$$G = \left| \frac{Max(\bar{d} - d)}{SD} \right|$$

Where, $\bar{d}$ 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).

2SD and 3SD Methods

We construct the interval by $\mu \pm 2\sigma$ and $\mu \pm 3\ \sigma$, where μ is the sample mean and σ is the standard deviation of the sample under consideration. The observations that don't lounge in the intervals formed by above statistic are treated as outliers. According to Chebychev's Inequality, For any random variable X with mean μ and variance σ^2, then for any$k > 0$,

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

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

From the inequality $[1-(1/k)^2]$ we are able to determine what portion of our data will be within *k* standard deviations of the mean. (Bain et al. cited by Seo, 2006) For example, we can access that at least 75%, 89%, and 94% of the data are lying within 2, 3, and 4 standard deviations of the mean, respectively. Probabilities of existence of outliers in the data sets can be determined by these results. Although Chebychev's theorem is non parametric and have no distributional assumptions, it has a major drawback that it gives the smallest proportion of observations within *k* standard deviations around the mean (Chebychev's, cited by Seo, 2006). Having prior knowledge about the distribution supports us to guess more efficiently. For example in case of standard 68%,95% and 99.7% data lies within 1,2,3 standard deviations respectively and we consider outlier beyond 2SD or 3SD according to our null.

Dixon's Test

Null hypothesis to apply Dixon test is that data are normally distributed and is based on the statistical distribution of "**sub range ratios**" of ordered data samples, drawn from the same normal population. Along with other demerits one major demerit of this test is that it cannot be applied again on the remainder data set in any case if once observation is detected as outlier or rejected.

Dixon's test is used for small sample sizes to detect outliers when mean of N-1 observations are significantly different from the mean of N observations. The data are arranged in ascending or descending order, when the mean in question is smallest or largest respectively. The critical values depend upon the sample size. Then the test statistic $Q_{\exp} = \left|\frac{X_n - X_{n-1}}{X_n - X_1}\right|$ is computed (e. g., for 3 ≤ N ≤ 7) and decided according to the critical values in the below given table. Null for the static is that there is no significant difference between suspected value and the remaining data.

Critical Values of Dixon Test

N	*CV for 90% confidence level*	*CV for 95% confidence level*	*CV for 99% confidence level*
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

THE Modified Z-Score:

In normal distribution we encounter just with the two parameters mean and standard deviations. These parameters are blessing as they are easy to compute and nearly available in all software's but this blessing becomes a problem if there are some outliers in the sample data because these are highly affected in presence of some outliers even in presence of single outlier mean is affected highly as it has zero break down value (Zaman,1996). To overcome this problem Iglewicz and Hoaglin (1993) proposed to use the median and median of the absolute deviation. The given modified Z-Score (M_i) statistic was computed as

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

Where $E(MAD) = 0.6745\sigma$ for the large normal data sets, Iglewicz and Hoaglin suggested that observations with $Mi > 3.5$ should be labeled as outliers and they verified their claim (suggestion) through simulation technique on the pseudo normal observations for the sample size of 10, 20, and 40.

Leverage Method

Leverage method is based on the following statistics

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{v^2}, where\ v^2 = \sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2$$

$n$is the sample size, $x_i$is the ith observation$\bar{x}$is the mean. Observation can be said as outlier if $h_{ii} > 0.5$. As we know that leverage is the observation that has substantial effect on the regression line. The most common measure of the leverage point is the hat value, contained in the hat matrix. (Hair et al. 1998, Iglewicz and Hoaglin, 1993)

MAD$_E$ METHOD

This method is similar to the Mean±2SD method but robust for detection of outliers and is unaffected by the extreme values. Here median and median absolute deviations are used instead of the mean and standard deviation. This is method is robust as it has break point value of 50%

Critical values for 2MAD$_E$ outlier labeling technique: Median±2MAD$_E$

Critical values for 3MADE outlier labeling method: Median±3MAD_E

Where $\mathrm{MAD}_E = 1.483 \times \mathrm{MAD}$

In this approach two robust estimators (median and median of absolute deviations) are used as in the above test in which robust skewness is used.

Median Rule

(Carling 1998) proposed the statistics for lower and upper critical values [L, U] = Q_2±2.3IQR where Q_2 is the sample median where scale of IQR i.e. 2.3 is not fixed but it depends on target outlier percentage and Generalized Lambda distributions (GLD) are selected.