MODIFICATION OF BOXPLOTS ON THE BASIS OF MODALITY TESTS AND MEASURES OF SKEWNESS

Thesis for Ph.D Econometrics

Researcher:	**Abdul Ghafar Shah**
Registration No:	**23-SE/PhD(Et)/S12**
Supervisor:	**Dr. Abdul Jabbar** **Assistant Professor** **IIU, Islamabad.**
Co-supervisor:	**Dr. Asad Zaman** **Ex-Vice Chancellor** **PIDE, Islamabad**

SHOOL OF ECONOMICS,

INTERNATIONAL INSTITUTE OF ISLAMIC ECONOMICS (IIIE),

INTERNATIONAL ISLAMIC UNIVERSITY ISLAMABAD (IIUI), PAKISTAN.

MODIFICATION OF BOXPLOTS ON THE BASIS OF MODALITY TESTS AND MEASURES OF SKEWNESS

Abdul Ghafar Shah

Registration No. 23-SE/PhD(Et)/S12

Submitted in partial fulfillment of the requirements for the

Doctor of Philosophy in Econometrics

at International Institute of Islamic Economics (IIIE)

International Islamic University

Supervisor: Dr. Abdul Jabbar

Co-supervisor: Dr. Asad Zaman

JUNE 2020

APPROVAL SHEET

MODIFICATION OF BOXPLOTS ON THE BASIS OFMODALITY TESTS AND MEASURES OF SKEWNESS

BY: Abdul Ghafar Shah **Reg. No: 23-SE/PhD(Et)/S12**

Accepted by the International Institute of Islamic Economics (IIIE), International Islamic University Islamabad (IIUI), as partial fulfillment of the requirements for the award of degree of **Doctor of Philosophy in Econometrics.**

Supervisor

Dr. Abdul Jabbar
Assistant Professor, IIIE, IIUI.

Co-supervisor

Prof. Dr. Asad Zaman
Ex-Vice Chancellor, Pakistan Institute of Development Economics (PIDE), Islamabad.

Internal Examiner

Dr. Arshad Ali Bhatti
Assistant Professor, IIIE, IIUI.

External Examiners

Prof. Dr. Eatzaz Ahmad
School of Economics, QAU, Islamabad.

Dr. Zahid Asghar
Director, School of Economics, QAU, Islamabad.

Dr. Hamid Hassan
Head,
School of Economics

Dr. Abdul Rashid
Director General IIIE

International Institute of Islamic Economics (IIIE)

International Islamic University Islamabad (IIUI)

Date of Viva-Voce Examination: **June 22, 2020.**

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.

Abdul Ghafar Shah

June 2020

DEDICATION

To

My Parents

&

My Family

Acknowledgement

Starting with the name of Almighty Allah, Who is most merciful and beneficiary upon His creatures. I am very much thankful to Almighty Allah to give me knowledge and understanding to enhance my skills in the field of Econometrics.

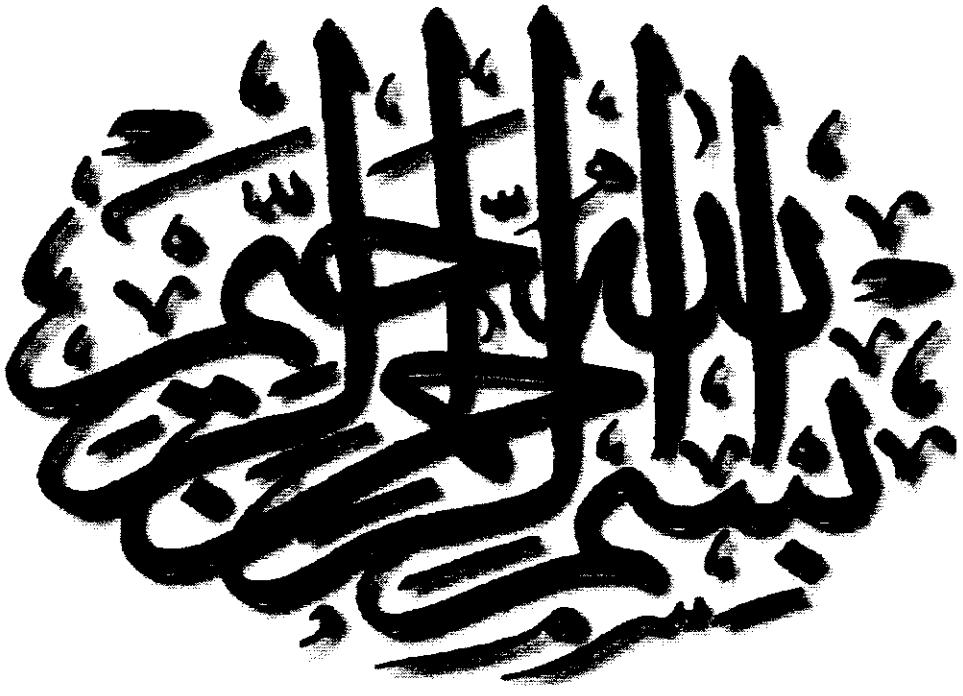
I am thankful to my loving, obligated and respected parents; especially to my mother, father, sisters, brothers, wife and all of my other family members who give me courage and support me at every stage of my study. I am also thankful to my Supervisor Dr. Abdul Jabbar and Co-Supervisor Dr. Asad Zaman to teach me and guide me in a good manner.

Here I would acknowledge the ability and enhance teaching and research methods of my teachers Dr. Asad Zaman, Dr. Arshad Ali Bhatti, Dr. Mumtaz Ahmad, Dr. Iftikhar Hussain Adil and Dr. Atiq-ur-Rehaman. In this regard I am especially thankful to my respected and honorable teachers and all faculty members of IIIE who encouraged and guided me whenever I need them. I am also very much thankful to my valuable, caring and loving friends and class fellows including Asad-ul-Islam, Shahid Akbar, Abdul Waheed, Khan Bahadar, Ali Raza, Muhammad Irfan Malik, Yasir Riaz, Gulfam Haider, Waqar Muhammad Khan and Mehmood-ul-Hussan. I also acknowledge the cooperation of administration staff of IIIE, especially, Syed Niaz Ali Shah Tauqir Ahmad for his support during my stay at IIIE.

I would like to say special thanks to Dr. Muhammad Shafiq, Dr. Habib Nawaz, Prof. Anwar Iqbal Khattak and Prof. Shahid Naveed for their help in drafting and proofreading of this thesis.

Again I do not know how to thank my parents for loving me, believing in me, and empowering my thoughts throughout my life.

Abdul Ghafar Shah



ABSTRACT

The boxplot is proved to be an extremely useful device for displaying general tendencies of the data. However, it has some weaknesses when the data is bimodal or multimodal, and also when the data distribution is not symmetric. There are also evidences that both boxplots and histograms can mislead researchers to understand the nature of the data.

From the literature, it seems that still there exist a huge gap which needs further improvement among the connection of bimodality, asymmetry and existing of outliers with the boxplot. The current study has accumulated all these issues, which is a quite different picture from the existing research of boxplot framework. The study is aimed at remedying these weaknesses via a new and improved approach to the boxplot. This study allows the use of boxplot in a much wider range of situations than the previous methodologies.

In this study, Robertson's and Fryer's (1969) conditions were applied to check the existence of bimodality for various values of parameters in the mixture of normal. For the decision to assess whether the data is unimodal or bimodal, one needs the best modality test and measure of skewness. Therefore, the modality tests are compared through Monte Carlo simulations on the basis of size and power properties. The simulated critical values are used and found that all the modality tests have stable sizes. From the power comparison, it is concluded that the Silverman Bandwidth test is the best test in most aspects. Similarly, various measures and tests of skewness are compared on different generated data. All of these have stable sizes on simulated critical values. But in power comparison, a newly introduced measure P_{norm} leads the performance and high power than other measures and tests. The current study presents a new technique to measure the appropriate degree or size of the bimodality through Trapezoidal and Simpson's rule and also identifies the factors which affect the size.

Before building bimodal boxplot, a necessary cutoff point from Fluss et al. (2005) conditions is also modified and extended for real data. This study introduced a bimodal boxplot, following the idea of Tukey's (1977) boxplot, which shows clear picture and summary statistics of bimodal distribution. Outlier zone around cutoff point is introduced for the detection of outliers in case of bimodal distribution. This study also presents some real data examples, i.e. exchange rates and cricket data for verifying the modality and skewness with their relevant boxplots and detection of outliers. These important results make possible ways and directions in the literature about bimodality, skewness, and boxplot.

Keywords: Skewness, Measures, Tests, Size, Power, Bimodality, Boxplot.

JEL Classification: C10, C12, C15, C16

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYM

Acronym	Full Name
IQR	Inter-Quartile Range
ER	Exchange Rate
Fig	Figure
SSSB	Split sample Skewness Boxplot
KSE	Karachi Stock Exchange
VH	Vandervieren and Hubert
MVH	Modified Vandervieren and Hubert
MC	Med-Couple
Sk	Skewness
SSS	Split sample Skewness
CDF	Cumulative Distribution Function
cv	Critical value
D	Higher Difference
max	Maximum
min	Minimum
MCSS	Monte Carlo sample size
Df	Degree of freedom
GDP	Gross Domestic Product
LRT	Likelihood ratio test
MLRT	Modified Likelihood Ratio Test
DF	Distribution Function
MDF	Monotonically Decreasing Function
inf	Infinite
PM	Proportional Mass test
EM	Excess Mass test
EEM	Empirical Excess Mass

DGP	Data Generating Process
Pdf	Probability distribution function
SD	Standard Deviation
QD	Quartile Deviation
IQM	Inter Quartile Median
OZ	Outlier Zone
IFS	International Financial Statistics
Prs	Pearsonian Coefficient
SM	Standardized Moment
SSSBB	Split sample Skewness Based Boxplot
MSSSBB	Modified Split Sample Skewness Based Boxplot
Skw1	Skewness-1
Skw2	Skewness-2
KS	Kolmogorov Smirnov test
WC	Wilcoxon test
w.r.t	With respect to
SB	Silverman Bandwidth test
PM	Proportional Mass test
UK	United Kingdom
SWR	Sampling With Replacement
ODI	One Day International

CHAPTER 1

INTRODUCTION

1.1 Study Background

Data is the core element of research in statistics and in every field of science. For examining data the graphic representation are more satisfactory than tables in exchanging information along with the groups of data (Gelman, Pasarica and Dodhia, 2002). Furthermore, the graphic presentation provides a sound base to choose more suitable and adequate techniques for parametric inferences. It is necessary to plot and review the data before analysis and inferences, as it describes the structure of the distribution.

Generally, graphic exhibition summarizes and interconnects patterns of the data for better understanding (Cohen, 2006). Most of the graphs that are in common use for the comparison of various data sets use basic measures (i.e. mean and standard deviation, etc.). But these basic statistics are not enough to explain the differences in configuration of the primary data which lead to misleading results (Wildenhain and Rappsilber, 2014).

A number of valuable tools such as graphs, descriptive statistics, and boxplot are used in this connection. However, the graphic state of the boxplot describes useful properties of data in the form of graphical presentation (Cox, 1978). Boxplots handle and signify both basic summary statistics and information about the distribution of the data. The actual graphic design of the boxplot, such as the range-bar, was introduced in the early 1950's (Spear, 1952). The basic purposes of the boxplot consist of efficient extraction of necessary facts and details of the data (figure, position, etc.) along with the existence of outliers, which is a substantial activity for various reasons. Basically, boxplot depends upon median and inter-quartile range 'IQR', which is in most cases more efficient than

the average and the standard deviation, and therefore it gives a more suitable summary of any real data in the majority of cases (Dovoedo, 2011).

Tukey (1977) developed a new version of boxplot in proper shape. This boxplot is a particular graphic procedure that describes data in five numbers that is minimum, first quartile, median, third quartile, and a maximum of any data set. Due to these five numbers quartiles information, Tukey's (1977) boxplot is considered to be robust in skewed data as compared to conventional technique which uses mean and standard deviation. Furthermore, it shows spread and information about skewness (McGill, Tukey and Larsen, 1978).

Similarly, Hubert and Vandervieren (2008) introduced an adjusted boxplot which depends upon the skewness measure called med-couple to find the distance of the whiskers. This modification is made in the whiskers of the boxplot for identifying outliers in the skewed data. However, Schwertman, Owens, and Adnah (2004) introduced a new boxplot which specially modifies the process of calculation of whiskers boundaries according to the normality assumption and a large number of samples (see also Sim, Gan, and Chang, 2005). Numerous techniques have been presented for boxplot so far, some of which can be found in the study of Potter et al. (2006) with their extractions in detail.

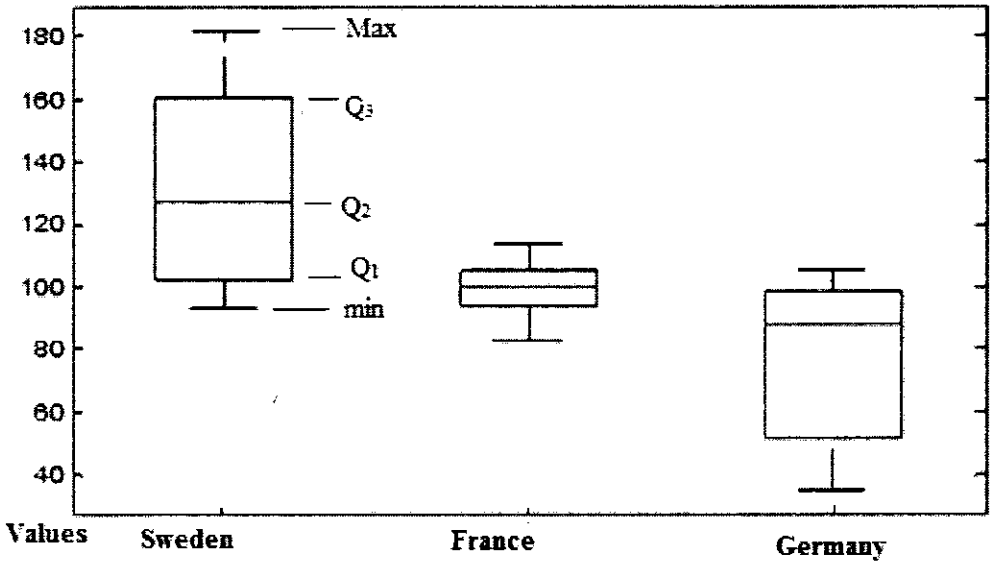
1.1.1 Advantages of Boxplot

Boxplot has numerous advantages over conventional data summaries which makes it popular and important. These advantages are given as follows:

i. Provides Basic Information

Box plot provides basic information about the shape and nature of the data distribution. Comparatively, boxplots are more informative data summaries than the massively used of mean and standard deviation which are applied in most cases. The popularity of boxplots in data analysis is increasing rapidly as their virtues become more widely known, see for instance 'Points of significance: Visualizing samples with boxplots' (Krzywinski and Altman, 2014). Consider an example; here a boxplot is compared with a numerical summary from Exchange Rate (ER) in dollars data series (1961 to 2013) of three countries.

Figure 1.1: Boxplots of 'ER' Data Series of Different Countries



Summary statistics and basic information of the data are presented graphically in boxplots above in Figure1.1, numerically below in Table 1.1.

Table 1.1: Summary Statistics of Different Countries ER

Statistics	Sweden	France	Germany
Skewness	0.155	−0.304	−0.507
Minimum	93.04	82.68	35.11
Q_1	102.88	93.97	52.27
Q_2	126.77	100	88.34
Q_3	160.39	104.74	98.63
Maximum	182.03	113.66	105.76

The above Figure 1.1 and Table 1.1 show the basic information and variation of the data sets for different countries. The distribution of France and Germany is very similar, i.e. negatively skewed and quite different from that of Sweden which is positively skewed.

ii. Low Space and Useful for Comparing Distributions

The boxplot is self-explanatory and easy graphic method of describing one or more data sets. As boxplot occupies low space, it is very easy to show and compare few boxplots in a limited area. According to Carter and Schwertman (2009), boxplot holds small area and hence is especially good for comparing distributions among a few groups or data sets. Boxplot is very useful than tables in exchanging information and comparisons for different groups of data documented by Gelman, Pasarica and Dodhia (2002).

This idea of boxplot shown above in Table 1.1 is also illustrated by Razzaque (2009). Similarly, the data series of different countries in Figure 1.1 are compared which shows valuable information in a short space. Figure 1.1 also describes that each boxplot has low

space with a comparison of distributions. Furthermore, it also shows the deviation of the distributions which is helpful for comparison.

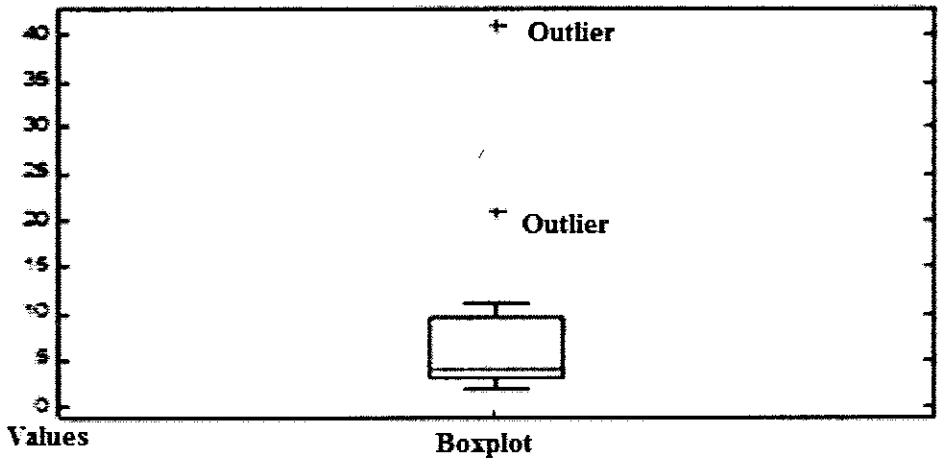
iii. Quickly Computable

The boxplot describes the distributional summary, which generally represents fewer features than a bar chart or kernel density. Basically, boxplots use exact five-point summary that always exists there in original data points which are quickly and easily computable especially by hand discussed Wickham and Stryjewski (2011) in their study.

iv. Useful for Identification of Outliers

It is the quality of boxplot which also shows outliers. Boxplot is good for indicating outliers and the comparison of distributions as stated by M.Lane and Sandor (2009). The boxplot also helps to show the degree of spread and skewness in the data series and identifying the extreme values (Morely, 2014).

Figure 1.2: A Boxplot with Outliers



The above Figure 1.2 represents a boxplot with two outliers denoted by plus '+' signs. Unlike conventional summary statistics, boxplots display outliers clearly.

1.1.2 Disadvantages of Boxplot

Despite their many virtues listed above, boxplots have certain weaknesses which have been documented in the literature as follows:

i. Boxplot does not Work Well in Bimodal Distribution

In the existing research, there is a problem that boxplot does not adequately show bimodality and peakedness. In these particular situations, there are evidences that both bar graphs and boxplots can misguide the observers stated by McNeil (1990). On the basis of its popularity, Choonpradub and McNeil (2005) stated that boxplot is less effective for describing the shapes and properties of existing distributions, especially bimodality.

Figure 1.3: Bimodal Distribution of Germany 'ER' and their Boxplot

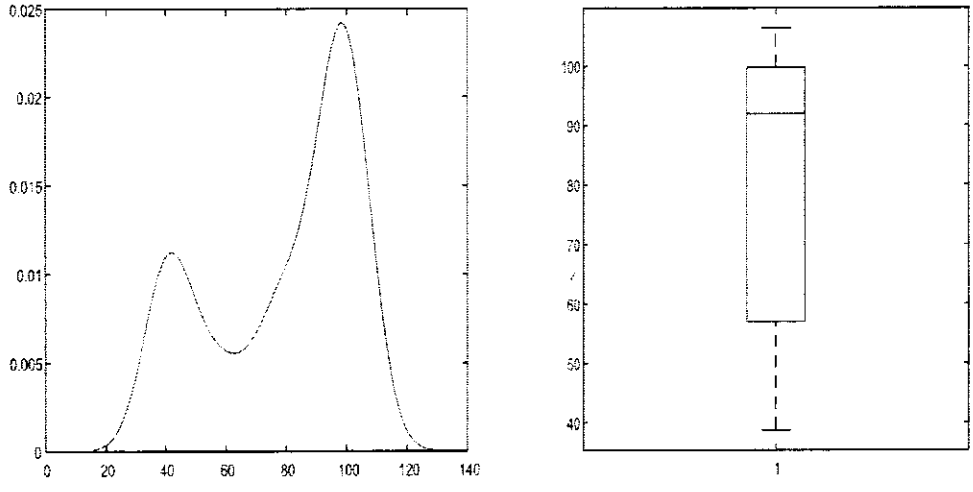


Figure 1.3 shows clearly a bimodal distribution of Germany exchange rate yearly data (1961 to 2013) and right side of this figure describes as unimodal skewed boxplot of the

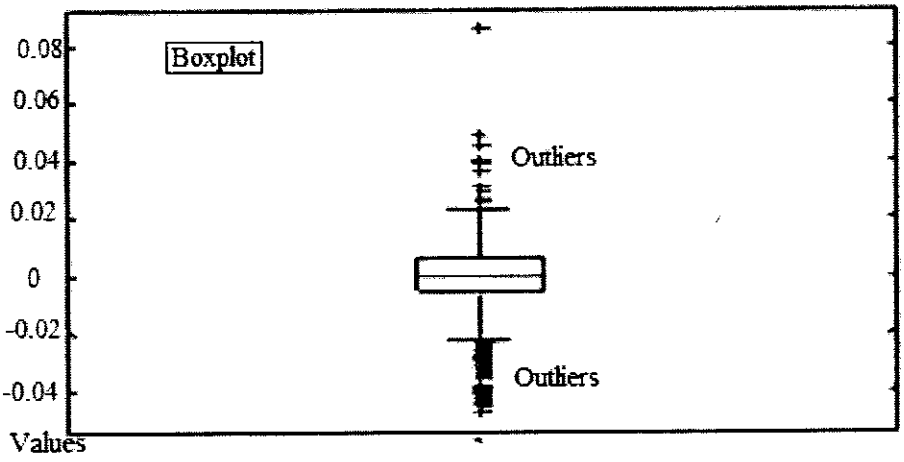
same data. It means that the traditional boxplot cannot specify bimodality clearly and it needs further improvements.

ii. It is Ineffective at Catching Outliers in Skew Distributions

Boxplot is ineffective at catching outliers on the narrow side of the distribution because the formula used for detecting is symmetric, even in skewed distributions. Hence it detects few outliers on the long-tailed side of the asymmetric distribution. The main problem is that it detects several outliers in the short tail and also a number of outliers in the long tail discussed by Adil (2012) in his PhD thesis.

Consider financial data series of KSE-100 index of two years (2013 to 2014) return data. Then two methods namely Tukey Technique and Split Sample Skewed Boxplot (SSSB) are applied on this data for detecting outliers.

Figure 1.4: Boxplot of Return Data (2013 to 2014)



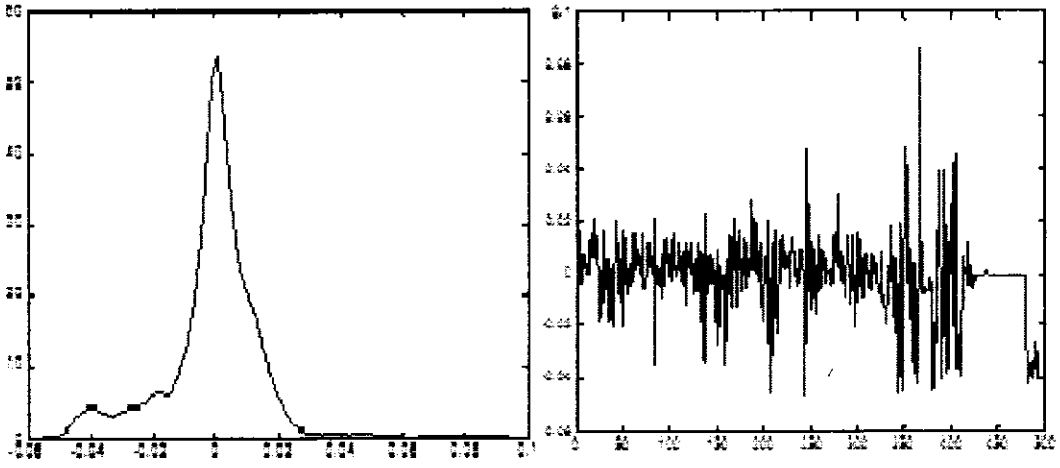
According to Tukey Technique, left outliers are 44 and right outliers are 11. Also, boxplot shows these numbers of outliers on either side respectively. But using SSSB

technique left outliers are 5 and right outliers are 9. It means SSSB technique performed better than Tukey's boxplot.

iii. Boxplot cannot Inform About Clusters of Data

Hintze and Nelson (1998) state that boxplot also fails because it cannot inform about clusters of data. Therefore, they used 'Violin plot' which identifies the existence of clusters in data series and proper shape of the distribution. However, tracking modes may help in finding clusters because clusters are different from skewness. Clusters will be detected well with histograms and density estimates. Using the above KSE return data series, it clearly shows that the distribution has clusters and is negatively skewed.

Figure 1.5: Density of KSE Return Data (2013-2014) and showing Clusters



The boxplot of the data series Figure 1.4 does not show different modes and have no information about clusters. But stare from its density Figure 1.5, the data series shows different modes and clusters.

1.2 Motivation

Boxplots indicate the central point and the variation of data from their central point. Biehler (2004) states that the explanation of variation will lead to the outcome in five various ways, i.e. position information, relating spreads and distributions, overall variation from the median, deviation below and above the median, and group or class information.

Hubert and Vandervieren (2008) stated that Tukey's boxplot provides uncertain results in case of asymmetric distributions. They introduced VH boxplot for the detection of outliers in the data series which was further modified by Akbar and Zaman (2013), called modified VH. Basically, VH boxplot identifies right and left extreme values but fails to show modality. In the existing research, there are many studies which have discussed skewed boxplot. In some situations, boxplot performs well and sometimes it fails to work. Mostly, it shows problems due to outliers and also fails in case of bimodality. Throughout in the literature, no study discusses bimodal boxplot case. From historical backgrounds, it seems that still there exist a huge gap which needs further improvement between the connection of bimodality, asymmetry and existing of outliers with the boxplot. The current study accumulated all these issues, which is quite different from the existing research of boxplot framework.

However, this study is mainly focused on new methodology for making boxplots based on different ideas and improvements (discussed briefly in the next section 1.3). This study uses different modality tests with null hypothesis as 'unimodality' and tests for skewness with a null hypothesis as 'symmetry', discussed later in detail chapter-4 and 6

respectively. This study solves maximum deficiencies existing in skewness and boxplot to fill this gap with some new directions.

1.3 Objectives of the Study

The current study proposed the following objectives and modifications:

i. The first objective of this study is to examine the effectiveness of different types of modality tests in this context. Therefore, this study conducts the Monte-Carlo simulation and finds the best and worst test on the basis of size and power criteria.

ii. The second objective and main is to evaluate with a preliminary tests of modality. If the data displays significant bimodality, then the current study develops an alternative bimodal boxplot. This is intended as one of the main contributions of this thesis.

iii. In the third objective, when the hypothesis of modality is not rejected, then we perform a second test for skewness and different measures of skewness. If the data is significantly skewed, this study plans to use an alternative to the boxplot which adjusts for skewness.

iv. The fourth contribution of this study is to examine these variants and proposes some new ideas for a skewed boxplot. The current study also assesses that how these alternatives perform well at detecting outliers in skewed distribution and comparison of boxplots which adjusts for skewness.

v. A fifth contribution is to evaluate different tests and measures of skewness in this context. This study also applies Monte-Carlo simulation and finds the

performance of various tests and measures on the basis of size and power criteria.

vi. Finally, when the skewness test does not reject symmetry, then the current study proposes the use of standard Tukey's boxplot. This is also new, because when we use Tukey's boxplot which states that the data is unimodal and symmetric according to tests result, so boxplot comes with a certificate of unimodality and symmetry, unlike conventional.

1.4 Significance of the Study

In the existing research, there is a problem that boxplot does not adequately show bimodality and peakedness. There is evidence that both boxplots and histograms can mislead viewers. Therefore, initially, the comparison of the modality tests in this study facilitates the researchers to decide about the best or dominant test in this context.

This study introduces a new measure of skewness, comparing it with the existing measures and tests to clarify their priority for the practitioners. This study will aware the researchers about new and alternate way of the modality test for finding the degree or size of bimodality existence through numerical integrals which also explains the factors that affect the size for generating data. Cutoff or maximum separation is necessary for any economic or other data series before building bimodal boxplot. This study solves the said issue to modify Fluss et al. (2005) conditions which can be easily applied by the researchers and practitioners to find the cutoff point.

This study provides the procedure for the researchers to build bimodal boxplot, which shows a clear picture of bimodal data and visual summary statistics. It implies that the new bimodal boxplot is very important and represents the data in a convenient way. This

study also provides the procedure to detect outliers in the case of bimodality. Lastly, the economic applications and results provide guidelines and assessment to the researchers about the nature of the data, i.e. skewness, modality, and their relevant boxplot.

1.5 Contributions of the Study

This study aims to remove weaknesses through a new and improved approach to the boxplot. The methodology is to allow the use of the boxplot in a much larger range of situations as compared to the past studies. The methodology of this study is an extension to the existing literature and reduces the shortcomings in the subject area.

A small amount of asymmetry is handled well through boxplot. This study used two ideas: The first one is to test for modality before building standard boxplot. The second idea is to explore the test for modality in the context of boxplots. Imran and Zaman (2013) compared tests for unimodality but also by many others. None of these studies have mentioned any connection to boxplots.

The current study, conducted initial tests for modality and symmetry. If both these conditions are satisfied, then original Tukey boxplot is used. When symmetry fails, then the boxplot needs to be modified. There are several proposals on how to build a boxplot for the asymmetric case. The current study examines existing proposals and introduces new ones and it also attempt to find the best alternative for use in the asymmetric case. It is also proposed to develop a new type of boxplot for the bimodal data series named 'bimodal boxplot'. Thus it implies that if the unimodality assumption is tested and failed, then this study proposed a new type of boxplot for the use of the bimodal situation. The proper measure of skewness is also developed in this study.

This thesis is basically organized in ten chapters. Chapter-1 reviews the introductory discussion, motivation, main objectives and significance of this study. Chapter-2 consists of the detailed review of the literature about the boxplot, measures, and tests for skewness, the connection of boxplot with bimodality and modality tests. Chapter-3 explains the methodology of the data generating process, Monte Carlo simulation designs, conditions of bimodality, construction of bimodal boxplot, outliers' detection in a bimodal distribution and information about the usage of data.

Chapter-4 provides the usage of bimodality conditions, the comparison of size and power of the modality tests. Similarly, chapter-5 explains the newly introduced measure of skewness P-norm and their importance. An evaluation of the size and power of the measures and tests for skewness is dealt with in chapter-6. Chapter-7 discusses the size of bimodality, definite integrals and effects of parameters for the case of the mixture of normals. Chapter-8 provides the construction of newly introduced bimodal boxplot, its advantages, and detection of outliers in a bimodal distribution. The main applications of this study on real data have been elaborated in chapter-9. In the last chapter, i.e. chapter-10, conclusions, recommendations and directions for future research are documented in detail.

CHAPTER 2

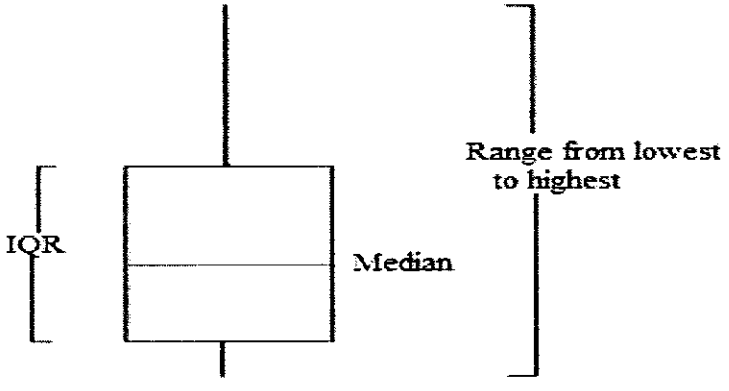
LITERATURE REVIEW

In the light of existing literature, many studies were conducted to polish and modify the boxplot. This study focuses on handling the existing deficiencies and introduces ideas about skewed and bimodal boxplot. A literature review of this study is divided into three sections. Section 2.1 discusses the historical background of boxplot and its modification. Section 2.2 deals with boxplot and skewness connectivity. It also reviews the literature of the presence and detection of outliers in the boxplot. Section 2.3 of this chapter associates the link between boxplot and bimodality. The chapter ends with the gap analysis.

2.1 History of Boxplot and its Modifications

Boxplot was first introduced by Spear (1952), called the range bar. This plot consists of a box within a point or line known as median and two tails (lines) spread within minimum and maximum. Diagrammatically, it can be shown as,

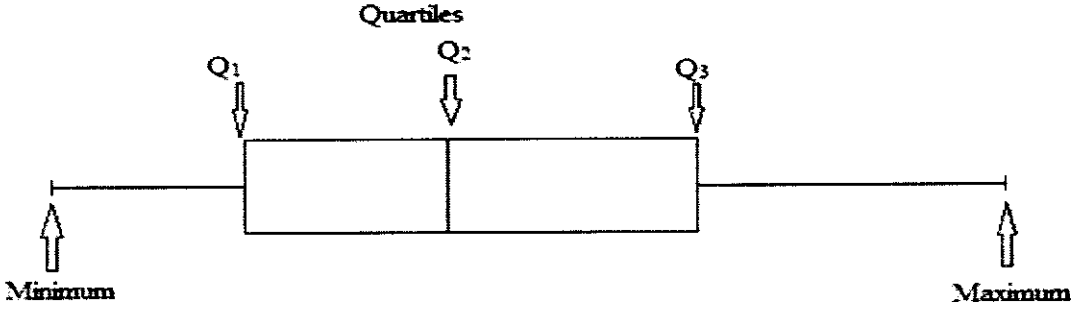
Figure 2.1: Picture of Spear boxplot



Source: Spear, (1952). Charting Statistics. McGraw-Hill, 166.

Tukey (1977) modified the boxplot and was named as 'whiskers and box'. The whiskers showed lower and outer fences, the values outside this range are highlighted as the outliers. Tukey, also developed the use of a five-point summary of data, consisting of minimum, Q_1= First quartile, Q_2= second quartile (median) and Q_3= the third quartile, maximum.

Figure 2.2: Picture of boxplot with a five-point summary



Tukey's boxplot works well when the distribution is symmetric. However, in case of skewed distribution, it fails to detect the outliers. Past researchers (Parzen, 1982; Matthews, 1981; Bibby, J. 1986; Clevand, W.S. 1985; and Parzen, 1979a, 1979b) tried to modify the boxplot. Frigge et al. (1989), introduced the idea of boxplot such as extremes are replaced with constant quantiles (minimum= 2% and maximum= 98%). The same researchers also changed multiplier with any other value except 1.5, within the fences to set the whiskers. For asymmetric distribution inter-quartile range was replaced with '$Q_1 - Q_2$' in the whiskers (Rousseuw et al., 1999). Carr (1994) designed boxplot with separate colors, i.e. red for above and blue for below the median.

2.2 Boxplot and Skewness

In skewed or bimodal distribution, the boxplot can mislead the readers and increase the chances of false results. As a result, its real importance and use are decreased (Tukey, 1977). Tukey boxplot was modified by Hubert and Vandervieren (2008), later renamed as HV, introducing the adjusted boxplot which described the process of changing Tukey's boxplot, for skewed distributions. Furthermore, in their technique, they combined the med-couple (MC) into the conventional boxplot which resulted better for skewed distributions. Modified adjusted boxplot (MAB) introduced by Dovoedo (2011) outweighed both Tukey's and HV techniques. MAB has whiskers which are measured from the median and constructed for using some multiples of the upper and the lower semi inter-quartile ranges.

Some classical tests were introduced and some were modified during research as Doane and Seward (2011) have introduced two modified measures of skewness in the study. They documented that modified standardized moment performed better as compared to other measures of skewness. Adil and Zaman (2012) introduced a new measure of skewness called Split Sample Skewness (SSS), which described that this measure of skewness works better in detecting skewness. Ateeq and Raza (2014), in their study, compared different measures and tests with the hypothesis of skewness. Their study implied that first, one has to measure the skewness of the data and secondly, to test whether the data is significantly skewed or not. If data is skewed then a boxplot can be used to adjust for skewness.

2.2.1 Measures of Skewness

Mathematical formulas and calculating procedures of various measures of skewness considered in this study are given in detail as below:

2.2.1.1 Pearsonian Coefficient of Skewness

Karl Pearson (1905) introduced coefficient of skewness (S_k) and can be computed as,

$$S_k = \frac{(\mu - \text{mode})}{\sigma}$$

Where 'σ' shows the standard deviation of the distribution and μ= mean. Since the mode normally fails to clearly describe the data, therefore, Pearson defined the following alternate measure,

$$S_k = \frac{3(\mu - \text{median})}{\sigma}$$

The magnitude of the Pearsonian coefficient of skewness 'S_k' commonly changes within the interval $(-3, +3)$. In the case of symmetrical distribution, the coefficient $S_k = 0$. For positively skewed distribution, $S_k > 0$, also called right skewed and for negatively skewed distribution, $S_k < 0$, also called left-skewed.

2.2.1.2 Standardized Moment of Skewness

Usually, the third moment is applied for measuring skewness known as the standardized moment (moment ratio) which is given as follows:

$$\gamma = \frac{\mu_3}{\sigma^3}$$

Where μ_3= Third moment.

Similarly, the magnitude of this measure of skewness coefficient 'γ' is within the interval (-2, +2). In this measure, if coefficient 'γ' is higher than zero ($\gamma > 0$), it means positively skewed distribution or rightly skewed. If this coefficient is lower than zero, it is called

negatively skewed or left skewed distribution. If the coefficient value is equal to zero, then the distribution is symmetric (normal).

2.2.1.3 Med-Couple

Standardized moment measure of skewness was affected by the existing outliers in the data. Solving this issue, G. Brys, M. Hubert, and A. Struyf (2004) developed a new measure of skewness called 'Med Couple'. The study described that it was the collection of both measures quartile and octile skewness. It uses two observations: one is before the median; second one after the median and it examines the difference of each one from the median.

Consider a series $x_1, x_2, x_3, \ldots\ldots, x_n$ and the same series can be arranged as $x_1 \leq x_2 \leq x_3 \leq \ldots\ldots\ldots\ldots \leq x_n$. Med Couple is calculated as follows:

$$M_c = med_{x_i \leq m_n \leq x_j} h(x_i, x_j) \qquad (2.1)$$

Where m_n denotes the median, and Med Couple by 'M_c'.

'x_i' shows the value smaller than the median and 'x_j' indicates the value higher than the median 'm_n'. In spite of $x_i \neq x_j$ than Kernel function denoted by 'h' is given below:

$$h(x_i, x_j) = \frac{(x_j - m_n)(m_n - x_i)}{x_j - x_i} \qquad (2.2)$$

The magnitude of this measure of skewness 'M_c' is within the interval $(-1, +1)$. In Equation (2.1), the standardized difference among lengths x_i and x_j from the median is measured. If $M_c > 0$ then distribution is positively skewed and would be negatively skewed if $M_c < 0$. However, if $(x_j - m_n) = (m_n - x_i)$, then $M_c = 0$ implies that the distribution is symmetric.

2.2.1.4 Split Sample Skewness (SSS)

Measures of skewness in the existing literature performed well, but all those measures failed in the presence of outliers. A med-couple measure of skewness somehow performed better in this context but this measure also has some complications in their procedure in case of large observations. To remove this drawback, Adil and Zaman (2012) developed a new measure of skewness, known as Split Sample Skewness (SSS). This measure separates the complete data series in two parts of the same size and calculates five-point summaries, inter-quartile range (IQR) of either side (i.e. before median and after median). Mathematically, it can be shown as:

$$SSS = Ln\left(\frac{IQR_R}{IQR_L}\right) \qquad (2.3)$$

But, $IQR_L = Q_{L3} - Q_{L1}$

And $IQR_R = Q_{R3} - Q_{R1}$

IQR_L is the IQR from the left part means before the median and IQR_R is the IQR from the right part means after the median.

Q_{L1}= 12.5th percentile, Q_{L3}= 37.5th percentile

Q_{R1}= 62.5th percentile, Q_{R3}= 87.5th percentile

If SSS= 0, then it is symmetric distribution. However, if $(IQR_R < IQR_L)$ as SSS <0 negatively skewed and if $(IQR_R > IQR_L)$ as $SSS > 1$, it means positively skewed.

2.2.1.5 Measures Skewness-I and Skewness-II

The concept of skewness-I and skewness-II was introduced by Tabor (2010) and found both the measures working well in the detection of skewness. The study measured skewness-I and skewness-II, by using the values of minimum, 50th percentile or median and maximum and used the following formula:

$$Sk_1 = \frac{\max - \text{median}}{\text{median} - \min}$$

$$Sk_2 = \frac{\frac{1}{2}(\min + \max)}{\text{median}}$$

Where Sk_1= Skewness-I and Sk_2= Skewness-II.

2.2.2 Tests of Skewness

In classical studies, various tests were used to check the skewness with different test statistics of asymptotic distributions. Efron (1979) extended the past studies by introducing a 'Bootstrap Testing Procedure'. This procedure was applied to both symmetric and asymmetric distributions to measure the skewness. Consider random sample $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ selected from 'F' which is unspecified distribution and with median 'v' also unknown as stated. When distribution 'F' is symmetric then $(x - v) = (v - x)$. Here our null hypothesis 'H_0' is given as,

$$H_0: F(x - v) = 1 - F(v - x)$$

Now the test statistics of 'F' is $\theta(x_1, x_2, x_3, \ldots\ldots, x_n)$ used for the measurement of skewness. The test statistics gives significant result when it has a value considerably unusual from zero. For further explanation, they computed the median. Here are some observations which are less than the median in size and some are large than the median. The difference between the median '$2m$' from those values which are larger than the median and the difference between those values from the median '2m' which are less than the median. They got a sample $X' = \{x_1', x_2', x_3', \ldots\ldots, x_n'\}$ and the selection of bootstrap sample is on the basis of standard bootstrap sampling with replacement (SWR) from the set (X, X'). Following are the bootstrap tests procedures for checking the skewness of a distribution with a null and alternate hypothesis as,

H_0: No skewness exist

versus H_1: Skewness exist

i. Student's t-test and ii. Kolmogorov-Smirnov test

2.2.2.1 Student's t-test as a Test of Skewness

Gosset (1908) developed the Student's t-test by considering two independent samples drawn from normal populations, i.e. $X_1 = \{x_{11}, x_{12}, x_{13}, \ldots\ldots, x_{1n_1}\}$ with parameters (μ_1, σ_1^2) and $X_2 = \{x_{21}, x_{22}, x_{23}, \ldots\ldots, x_{2n_2}\}$ with parameters (μ_2, σ_2^2).

The test statistic of this test is given as,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (2.4)$$

Where n_1 is the sample size and $\bar{x}_1$ is sample mean from the first sample while n_2 is the sample size and $\bar{x}_2$ is sample mean from the second sample. The s_p^2 is pooled or common variance and if $\sigma_1^2 = \sigma_2^2$ then it can be estimated as,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Where,

$$s_1^2 = \frac{1}{(n_1 - 1)} \sum (x_1 - \bar{x}_1)^2$$

$$s_2^2 = \frac{1}{(n_2 - 1)} \sum (x_2 - \bar{x}_2)^2$$

In the case of $\nu_1 - \nu_2 = \Delta_0$ such that $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ are the degree of freedoms.

Then test statistics through the null hypothesis as,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

As well as $v = n_1 + n_1 - 2$ degrees of freedom 'df'.

If $(\mu_1 = \mu_2)$, then the test statistics will be as follows,

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

For different variances $(\sigma_1^2 \neq \sigma_2^2)$ then test statistics as,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In this situation, the degree of freedom 'v' for the 't' distribution can be obtained as,

$$v = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad (2.5)$$

The critical value is decided as,

For $t \leq t_{\alpha/2(v)}$ when the alternate hypothesis is $(\mu_1 - \mu_2) \neq \Delta_0$

If $t \leq t_{\alpha(v)}$ when alternate hypothesis is $(\mu_1 - \mu_2) < \Delta_0$

$t \geq t_{\alpha(v)}$ when alternate hypothesis is $(\mu_1 - \mu_2) > \Delta_0$

When 't' calculated value is higher than t-table value, it gives significant result of this test.

2.2.2.2 Kolmogorov-Smirnov Test

A one sample test was introduced by A.N Kolmogorov (1933). But this test is extended to two samples by Smirnov (1939). As both the tests have the same testing methods, therefore, it is called a Kolmogorov-Smirnov test. This test depends on the cumulative distribution functions (CDFs). The null hypothesis H_0 contains that the two samples are selected from the populations having the same cumulative distribution functions, while their alternative hypothesis H_1 is the two samples which are selected from the populations having unequal cumulative distribution functions.

$$H_0: F_1(X_1) = F_2(X_2)$$

$$H_1: F_1(X_1) \neq F_2(X_2)$$

The values of each sample are set in ascending order and cumulative relative frequencies (CRF) are calculated at each value. Then they calculated the differences at each value included in the list. A rejection of the null hypothesis is made if a large difference is identified at any value. The Kolmogorov-Smirnov (KS) test depends upon the higher difference 'D', which is given as,

$$D = \max[S(X_{n_1}) - S(X_{n_2})], \text{ for a one-sided test.}$$

$$D = \max\left|S(X_{n_1}) - S(X_{n_2})\right|, \text{ for the two-sided test.}$$

Where $S(X_{n_1})$ and $S(X_{n_2})$ are respectively cumulative relative frequency distributions of two considerable samples selected from the two populations.

In case of two-sided test, if both of the sample sizes are less than 40 at the selected significance level and the test statistic value of 'D' is larger than critical value 'cv', then H_0 is rejected. Applying one-sided test for sample more than 40, the following test statistics is used:

$$x^2 = 4D^2 \left[\frac{n_1 n_2}{n_1 + n_2}\right]^2 \qquad (2.6)$$

Where 'x^2' is the chi-square distribution along with the degrees of freedom 'df= 2'. If $x^2 \geq x^2_{0.05(2)}$= 5.98, then the test result will be significant.

2.2.3 Outliers in the Boxplot

In the case of symmetric distribution, the Tukey (1977) boxplot is used for detection of outliers and performs well. Kampstra (2008) has found that the identification of outliers is usually unclear, particularly in case of skewed distributions. In the asymmetric distribution, the amount of outliers increases if the number of observations increases. This makes it impossible to see every single outlier. Hubert and Vandervieren (2008) examined the performance of adjusted boxplot on real and generated skewed distributions. A clear contrast is shown between outliers and other observations. Dovoedo (2011) stated that the observations which are outside either the lower or upper fences in the boxplot are called an outlier.

Adil and Zaman (2012) introduced the Split Sample Skewness Based Boxplot (SSSBB) which shows better informative data summary and shows high accuracy in the detection of outliers in case of skewed distributions. They show that the performance of 'SSSBB' procedure is higher in rank to other classical procedures. In their study, Akbar and Zaman (2013) assumed a complete picture of loss functions and it is found that Vandervieren and Hubert Boxplot (VH Boxplot) procedure performs well in financial returns data series as compared to other existing techniques of detecting outliers.

In the current section, all the above studies have no connectivity between outliers detection and bimodal distribution. This area needs a complete setup to detect outliers in case of bimodal distributions.

2.3 Boxplot and Bimodality Link

It is clear from the literature that in case of symmetric distribution, the median is always located in the center of the boxplot. Wainer (1990) implies that at very small whiskered symmetric boxplot with no outlier shown, it is illustrated that the distribution is 'symmetric' or 'short-tailed' but actually such type of distribution sometimes has at least two modes, called multimodal. Boxplot can cover some part of the shape of a distribution making a wrong impression. It implies that by using modality tests, we can check the distribution whether it is unimodal or bimodal. Also, different modality tests are used in this study for the size and power comparison.

Choonpradub and Don McNeil (2005) suggested slight changes in boxplot for showing bimodality by making the two quartiles ends of the box thicker. Such a criterion is based on an exact measure of skewness and kurtosis that enhances the calculations to draw a boxplot. These studies have more information but a lot of issues are still needed to modify a boxplot which provides eminent information such as the shape of the distribution, in case of bimodality.

According to Siva Tian (2010), it is the quality of a boxplot which shows the difference between the shape of skewed and symmetric distribution but fails in case of bimodal distributions. It implies that it is necessary to develop bimodal boxplot with the connection of modality test which is a prior step before making bimodal boxplot.

2.3.1 Detection of Bimodality

Wolfe (1970) applied likelihood ratio in favor of the null hypothesis of normal distribution against a mixture of normal distributions in the alternate hypothesis for different dimensions. Wolfe (1970) results showed that in unimodal distribution, the probability of at least one mode is very high which looks like a mixture of normals.

Engelman and Hartigan (1969) maximize the likelihood ratio on the basis of dividing the data samples into two subparts which have the same sample means selected from the normal distribution in null hypothesis and unequal sample means in the alternate hypothesis. This test was simple in calculation but failed in any bimodal distribution. Silverman (1981, 1986) developed the critical values for bootstrap technique as 'k' number of the bandwidth of kernel density for 'k' number of modes. He found a monotonically diminishing function for the selected bandwidth in the kernel estimate when the number of observations is constant.

The excess mass test was used to test the multimodality by Muller and Sawitzki (1991). They obtained that for k-modes, this test becomes equal to another modality test, i.e. Hartigan Dip test. Bianchi (1997) applied modality tests to examine and detected k-modes consistently by using 119 countries GDP data. Chen et al. (2001) used a Modified Likelihood Ratio Test (MLRT) and a mixture of different models on the basis of parameters to test modality. They concluded that it is not necessary that mixture of any components generate modality, but it is possible only in a mixture of unimodal densities. For this purpose, they also developed a 'Likelihood Ratio Test' (LRT) for bimodality.

Daniel et al. (2008) applied two modality tests, i.e. Silverman test and Hartigan Dip test, for different distributions. After correcting asymptotic levels, they concluded that all the series was multimodal.

Imran and Zaman (2014) used different modality tests for comparison on the basis of size and power in their simulation study. They used different data generating processes as well as real data series and documented that Silverman test performs well as compared to other tests in case of small and large sample size. This study applied the same existing modality tests on the special case of data for the values of parameters where the distribution is bimodal and links it with boxplot.

2.3.2 Modality Tests Comparison

For testing modality, the current study used and discussed four modality tests with their mathematical structure. These basic modality tests are Silverman's Bandwidth test (1981), Hartigan Dip test (1985), Excess Mass Test (1991), and Proportional Mass Test (2011). All of these selected modality tests are usually nonparametric tests having the null hypothesis of unimodality against the alternative hypothesis of bimodality or multimodality. The statistical explanation and mathematical techniques of these modality tests are described as follows:

2.3.2.1 Hartigan Dip Test

This test was introduced by Hartigan (1985), also called Dip test. This test calculates the higher difference among empirical distribution function 'DF' and unimodal (mostly uniform) distribution function for the purpose to reduce the maximum difference. Hence, Dip test assesses the difference among the data distribution and particular theoretical distribution of one mode existing in concerned data series. If $f(x)$ is unimodal

probability density function with 'K' number of modes, the Cumulative Distribution Function (CDF) denoted by $F(x)$ is curved outwards under 'K' and curved inwards above 'K'. For at least two modes, the inwards curve of CDF changes its direction. For positive dip values, Hartigan dip test shows that the distribution is bimodal. [1]Hartigan and Mrs. Hartigan documented that null hypothesis H_0 consists of uniform distribution as asymptotically set the dip values larger for all other distributions. For this reason the power of this test increases. Also in case of small sample size (less than or equal to 100), Dip test works well. Dip test was used for at most two modes.

Let F shows the distribution function and $D(F) = d$ in the case for non-reducing functions G. But, when $X_L \leq X_U$, G is the highly outwards curved minorant of $(F + d)$ in limit $(-\infty, X_L)$, now in variable, G has in a variable much higher gradient of (X_L, X_u), G is a very small quantity of inwards curved majorant of $(F - d)$ in $[X_L, \infty)$, so the procedure as follows as:

(i) Initially consider $X_L = X_I$, $X_u = X_n$, D = 0.

(ii) Find the Greatest Convex (outwards curved) Minorant 'g.c.m' G and Least Concave (inwards curved) Majorant 'l.c.m' L for F in $[X_L, X_u]$, consider the values connecting with F are correspondingly $g_1, g_2, \ldots\ldots, g_k$ and $l_1, l_2, \ldots\ldots, l_m$

(iii) To take $d = \sup |G(g_i) - L(g_i)| > \sup |G(l_i) - L(l_i)|$ and also the Sup exists at $l_i \leq g_i \leq l_{j+1}$ explain furthermore as $x_i^0 = g_i$, $x_u^0 = l_{j+1}$.

[1] The step by step test procedure explained in article "The Hartigan Dip Test of Unimodality" by J. A. Hartigan and P. M. Hartigan (1985)

(iv) Taking $d = \sup|G(l_i) - L(l_i)| \geq \sup|G(g_i) - L(g_i)|$ and also the Sup exists at $g_i \leq l_i \leq g_{i+1}$ explain as $x_l^0 = g_i$, $x_u^0 = l_j$.

(v) When $d \leq D$, finish and put $D(F) = D$

(vi) When $d > D$,

put $D = \sup\{D, sup_{x_l \leq x \leq x_l^0} |G(x) - F(x)|, sup_{x_u^0 \leq x \leq x_u} |L(x) - F(x)|\}$

(vii) Place $x_u^0 = x_u$, $x_l^0 = x_l$ and go back to (ii).

This study used the same method of Hartigan Dip test for any data series and calculated their size and power for comparison with other modality tests.

2.3.2.2 Silverman's Bandwidth or Bump Test

Silverman bandwidth test is also called a bump test or kernel density estimation test. The test statistics consists of kernel density of unimodal distribution estimation by using very small window width. Silverman applied the Gaussian kernel density function in this bump test. The significance level is described by the selection of the single mode density estimation which is described through empirical re-arrangement of the data series.

However, the critical value 'cv' of this test is calculated through the Monte Carlo Simulation method. Silverman test is used for multimodality (i.e. at least two modes) in the alternate hypothesis. This test performs well in case of large sample sizes. For constant observations, Silverman observed that in this computation the bandwidth on many modes is monotonically decreasing function. Applying this terminology, Silverman (1981) explained the k-critical smoothing parameter or bandwidth as the smallest smoothing parameter for h_k of the kernel density computation having k number of

modes. The Gaussian Kernel density creates several modes as the smoothing parameter becomes higher. Smallest quantity of the smoothing parameter is necessary to develop kernel estimation which covers a mode in the test statistics. When the test statistics is large, then the null hypothesis of one mode is rejected. To solve the cause of this extra mode, it is necessary to increase the amount of bandwidth. The main advantage of this test is that it is used for multimodality. The method of Silverman test is described as follows:

The sample $x_1, x_2, \ldots\ldots, x_n$ belong to kernel density estimation with unspecified density function "f".

$$\hat{f}(x,h) = \frac{1}{nh}\sum_{i=1}^{n} \mathrm{k}\left(\frac{\mathrm{xi}-\mathrm{x}}{h}\right)$$

Here, smoothing parameter is denoted by 'h' and 'k' which shows the function of the Gaussian kernel. Silverman (1981) describes that the smoothing parameter 'h' increases the number of modes in $\hat{f}(x,h)$ decreases. As this test applied kernel bandwidth functions to guess the number of modes,

The test statistics of this test are as follows:

$$\hat{h}_{crit} = \inf\{h : \hat{f}(x,h) \text{has '1' mode}\}$$

This lowest quantity smoothing is necessary for the estimated kernel density which has a single mode. For maximum $\hat{h}^1_{crit}$, the hypothesis of unimodality rejects a higher number of smoothing which is essential to vanish extra modes in this test kernel estimation. For every sample, the significant result of $\hat{h}^1_{crit}$ is assessed through smoothing bootstrap procedure. The smallest smoothing parameter $\hat{h}^1_{crit}$ is essential for single mode and the probability $\hat{P}$ as,

$$\hat{P} = P(\hat{h}^{1*}_{crit} \geq \hat{h}^1_{crit})$$

Where the role of $\widehat{P}$ is to get the information about the relative level $\widehat{h}^1_{crit}$. For reasonably large $\hat{h}^1_{crit}$ is examined with results of bootstrap samples, so $\widehat{P}$ has a low value which is powerful evidence of a significant decision. This technique is applied to test for many modes which are normally performed in a particular order, starts from the single mode and begins whenever the test is unsuccessful to reject H_0 of 'k' number of modes.

2.3.2.3 Proportional Mass Test

Proportional Mass (PM) test was introduced by Cavallo and Ringobon (2011) who examined modality in the area for a special number which is nearly zero at both sides. This test calculated mass of prices which vary in the absolute number less than 1%, 2.5%, and 5%. This test is based on the central tendency point (i.e. 0%, mode and mean) of the distribution. The minimum number of modes shows the evidence to reject the null hypothesis of a single mode. For the positive value of the PM test, it means that the distribution is unimodal; and for the negative value, it shows bimodality. But when the distribution is uniform, then this test value is equal to zero when the ranges are positive, greater than 5 or else.

Cavallo and Ringobon (2011) documented that the PM test described the quantity of unimodality on both sides of the main (in center) value which examines the mass of the density around bounds. PM test is dependent on the situation such that a maximum relationship of the mass of unimodal distributions is approximately near the mode.

However, the boundaries of both sides of the mode increase, then the overall mass also increases with a small amount. In case of bimodality around specific value, the mass becomes better and in large amount. For this purpose, the distance of this additional regular increase of mass is applied to decide the amount of unimodality at both sides of a

specific value. Further, they explained the situation of unimodal distribution middle point at 0%. The magnitude of mass between the points (−1%, 1%) should be maximum from the points (−5%, 5%). Therefore it is given as,

$$P\ (|\Delta p| \leq 1) \geq P\ (|\Delta p| \leq 5)/5$$

Proportional mass when i= 1 and j= 5 as follows,

$$PM_{1,5}^{0} = \ln(P\ (|\Delta p| \leq 1))/(\ P\ (|\Delta p| \leq 5)\ /5)$$

This ratio is formalized for the calculation of Proportional mass on both sides of zero as;

$$PM^{0} = \frac{1}{|z|} \sum_{ij \in z} PM_{ij}$$

Set 'Z' consists of the combinations is also i< j

The similar explanation was used when they got interested in testing the amount of unimodality at both sides of mode denoted by m, which is,

$$PM^{m} = \frac{1}{|z|} \sum_{ij \in z} \ln \frac{P\ (|\Delta p - m| \leq i)}{P\ (|\Delta p - m| \leq j/(j\ /i)}$$

The H_0 of this test is also connected with PM^m having positive value, which means the distribution is unimodal. Applying the bootstrap method to check the significance of the test and find the links with a number of positive PM^m. Smaller the link of bootstrap with a number of positive PM^m, leads to the significant result of PM test.

2.3.2.4 Excess Mass Test

Excess Mass 'EM' test was presented by Muller and Sawitzki (1991) for 'm' modes. This test is commonly used for at least one mode i.e. multimodality and for clustering. This test becomes equal to Hartigan Dip test when applied for 'm' modes. This test computed the common difference of a relevant distribution to existing modal, mostly uniform distribution. The Excess Mass Test described a function which performed well in

estimation and can be applied for modality. The specialty of this test is to examine the characteristics of various samples from the variables of uniform distribution. However, they tried towards unexpected changes. An excess (extreme) mass is focused wherever a mode is situated. The excess mass test was applied to determine the common differences of the considerable distribution, that is uniform distribution and the function is computed for testing modality. Muller and Sawitzki (1991) documented the excess mass (EM) procedure which was expected as a common procedure for statistical estimations. This test explained a particular way of analyzing and applying for the representation of modality. The testing method of excess mass test is given as follows:

They supposed that the distribution function is represented by 'F', i.e. same as the sampling density which was indicated by 'f'. The empirical (non-theoretical) distribution function is denoted by $\widehat{F}$ and 'n' sample size selected from 'F'. The mathematical expression of empirical excess mass (EEM) for modes 'm' is as follows:

$$E_{nm}(\lambda) = Sup_{c_1, \ldots, c_n}\left[\sum_{j=1}^{m}(\widehat{F}(C) - \lambda\|C_j\|)\right]$$

Here $\lambda \geq 0$, the supremum (Sup) was selected from the arranged set $\{C_1, C_2, \ldots\ldots, C_m\}$ of disjoint values, where the function $\widehat{F}(C)$ is the $\widehat{F}$ size of C and quantity $\|C_k\|$ is the width of C, and further explanation is,

$$D_{nm}(\lambda) = E_{nm}(\lambda) - E_{n.m-1}(\lambda) \geq 0$$

The null hypothesis H_0 contains that sampling density 'f' has $(m-1)$ modes and alternate hypothesis H_1has 'm' modes, then test statistics is,

$$\Delta_{nm} = Sup_{\lambda>0}\{D_{nm}(\lambda)\}$$

For the higher value of Δ_{nm} mostly the test has significant results. They also introduced empirical procedures for quantity and describe the ideas of higher Δ_{nm}, which emphasizes that mode 'm=1'.

2.4 Gap Analysis

This study applied the existing modality tests on the data (generated data for which the values of parameters, distribution is bimodal). From the test result, when unimodality is rejected, then this study builds bimodal boxplot, which is quite a different presentation from unimodal boxplot. When the null hypothesis H_0 of unimodality is not rejected, then test for symmetry is applied. Also, this study compared and assessed the effectiveness of these modality tests for the purpose of improving boxplots.

A lot of research has been conducted on the classical tests and measures of skewness. In most cases, these measures lead to misleading results. The current study introduced a new measure of skewness and compared with the existing tests and measure of skewness. This study used different measures and Bootstrap tests of skewness for the purpose of checking the symmetry, and according to their results, the most suitable and relevant boxplot is suggested.

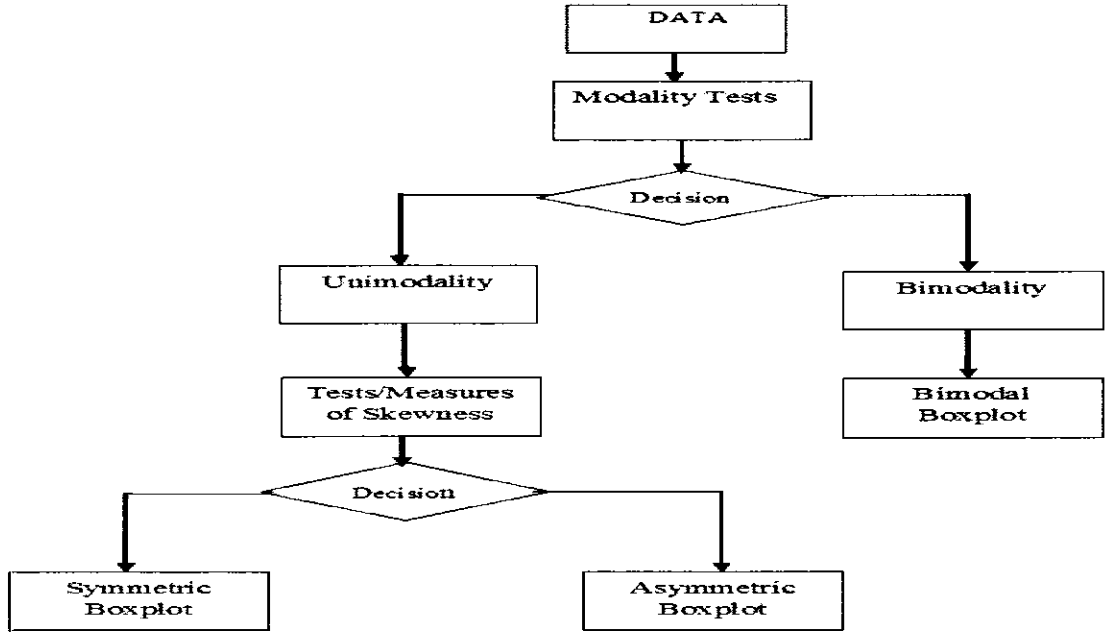
In the existing literature, there is a huge gap to develop a bimodal boxplot and also to detect outliers in the bimodal distribution. This study proposed some new ideas for a bimodal boxplot and assessed how well these alternatives perform at detecting outliers in a bimodal distribution.

CHAPTER 3

METHODOLOGY

The first Section 3.1of this chapter discusses the different data generating processes for unimodal as well as for bimodal distributions. Section 3.2 presented the procedure of Monte-Carlo simulation for modality tests and skewness tests regarding their size and power. Section 3.3 investigated the presence of bimodality and also derived the conditions of bimodality. The current study has also expounded the size of bimodality with the help of numerical methods such as Trapezoidal and Simpson's rules. Section 3.4 deals the outlier detection technique known as Split Sample Skewness Based Boxplot (SSSBB) and its modification in case of bimodality. Lastly, section 3.5 described the information about data used in this study. The following Figure 3.1 summarizes the methodology of this study.

Figure 3.1: Summary of the methodology



3.1 Planned Data Generating Processes (DGPs) for use in Simulations

It was necessary to define and discuss the different DGPs before handling the analysis of the study. Monte Carlo simulations on the basis of these DGPs for testing modality and skewness were performed. Various DGPs were used and the detail discussion is given in the following section.

3.1.1 Data Generating Process-I (DGP-I)

In the subsequent sections of the study, the unimodal distributions are represented with DGP-I. All the considerable distributions of the DGPs are detailed in the subsections.

3.1.1.1 Normal Distribution

The normal distribution is the most popular continuous distribution used in every field. This distribution is usually applied in social and natural sciences in connection with the random variables which have unknown distributions. For comparison, data are generated from the standard normal distribution (which is the simplest case of Gaussian distribution). This study used the Probability Density Function (pdf) for analyzing the size, power and critical value of every test. Let 'X' be the random variable of normal distribution with parameters (i.e. mean μ and variance σ^2) as, $X\sim$ N(μ, σ^2) and their Matlab code 'X= normrnd(μ, σ^2, n, 1)', where 'X' is a standard normal variable with parameters (μ= 0, σ= 1).

3.1.1.2 Log-Normal Distribution

The log-normal distribution is a continuous distribution in which logarithm of random variable belongs to a normal distribution. The process of this distribution is computational realization of multiplication of a lot of positive random variables. This is

approved by the 'central limit theorem' in the logarithmic form. The log-normal distribution was in order to compare the modality and skewness tests by drawing 'n' samples from a log-normal distribution with parameters μ and σ, i.e. X~ Ln-N(μ, σ^2) and their Matlab code 'X= lognrnd(μ, σ^2, n, 1)'.

3.1.1.3 Beta Distribution

This distribution is used to model the characteristics of a random variable restricted to the definite limits or intervals. This distribution is also applied in the Bayesian approach to explaining the starting information regarding the probability of success of an event. The Probability Density Function (pdf) of Beta distribution is also used for analyzing the size, power and critical value of every test. Generally, 'X' is a variable belong to Beta distribution for a random variable X is represented as X~ β(α, β) and their Matlab code 'X= betarnd(α, β, n, 1)', where α and β are two shape parameters of this distribution.

3.1.1.4 Chi-Square Distribution

It is the continuous distribution which is the under the root of the total of the square of random variables, i.e. 'X~ N(0, 1)' or same the random variables have Euclidean distance from the origin. The curve of this distribution is skewed positively and as sample size increases then skewness also decreases. To examine the tests of modality and at various alternatives of skewness, it is selected a random sample from chi-square 'χ^2' distribution with various degree of freedom 'v' such as $X \sim \chi^2_{(v)}$ and their Matlab code 'X= chi2rnd(v, n, 1)'. The Chi-square distribution has one parameter called the degree of freedom 'v'.

3.1.1.5 The Uniform Distribution

Uniform distribution is also called 'rectangular distribution'. All limits of equal length and the distributional attachment have the same probability. The attachment is described through two parameters (i.e. a, b) such that it has their smallest and largest values. If a=0 and b= 1, then this distribution is known as standard uniform distribution. This is the important distribution used in this study for the best and the worst test in case of modality and skewness tests. This density function is called constant or uniform because among the two points with two small intervals of the equal length with the equal probability. Consider 'X' is a random variable selected from Uniform distribution such that X~U(a, b) and their Matlab code 'X= unifrnd(a, b, n, 1)'.

3.1.2 Data Generating Process-II (DGP-II)

For evaluating the bimodality, the current study used the DGP-II, i.e. the mixture of two distributions given as below:

3.1.2.1 Mixture of Two Normal Distributions

In this data generating process, a mixture of two normal distributions was used (i.e. one normal and second one is standard normal distribution) to check the bimodality on the basis of bimodality conditions and tests about modality. The mixtures of two normal distributions are discussed as follows: Consider X_1 as the random sample from the first normal distribution with parameters (i.e. mean= μ_1 and variance= σ_1^2) and X_2 as the random sample from the second normal distribution with parameters (i.e. mean= μ_2 and variance= σ_2^2).

Z= {X_1 with mixing probability α, X_2 with mixing probability $(1-\alpha)$}

Also, it can be written as,

$$Z = \alpha X_1 + (1-\alpha) X_2 \qquad (3.1)$$

Where 'Z' is known as a mixture of normal or bimodal distribution with mixing proportion or probability 'α' of each normal density between or equal the interval (0, 1). This study used Matlab code for this DGP as, and their Matlab code 'Z= dat_genr_mix_norm ($\alpha, \mu_2, \sigma_2^2$, n, 1)'.

3.2 Monte-Carlo Simulation Designs

In this study, the research methodology consists of the two simulation designs separately (i.e. simulation design for modality tests and simulation design for measure and tests of skewness). The step by step methods are given as follows:

3.2.1 Monte-Carlo Simulation Design for Modality tests

This step by step process of simulation design is used for the comparison and evaluation of the modality tests. To calculate the size and power of these tests, here the following procedure is used.

i. The data are generated by proposed DGPs mentioned above in Section 3.1.1.

ii. The test statistics of various modality tests (see Section 2.3.2) are calculated and applied on the selected DGPs.

iii. Size of each modality test are also calculated at 5% significance level fitted for Monte Carlo sample size 'MCSS'= 5000 as,

Size of the modality test = Probability (Reject H_0/ H_0 is true)

Null hypothesis 'H_0' for modality test consists of any unimodal distribution (i.e. normal, Chi-square, uniform or Beta distribution) and our alternate hypothesis (H_1) is the mixture of normal distributions from DGPs.

iv. This study also calculated the power of each modality test at 5% significance level calculated as,

Power of the modality test= Probability (Reject H_0/H_0 is false)

If our null hypothesis is rejected, then we can construct bimodal boxplot discussed in Section 8.3. But if the null hypothesis is accepted, then measures of skewness are applied to decide for skewed or symmetric distribution.

3.2.2 Monte Carlo Simulation Design for Measure and Test of Skewness

This study used different distributions like χ^2, Beta, Uniform and Log-normal distributions to find the best measure of skewness. For any distribution, the intervals build approximately 95% centre values left 2.5% on each side and used both methods for fences from the simulated below and above critical values 'cv'. To find the size and power of measures and tests about skewness, the current study used the procedure as follows:

i. The data is generated using Data generating process 'DGP' mentioned in section (3.1.1).

ii. In the second step, various measures and tests of skewness on these DGPs are applied.

iii. Size of each measure and test of skewness at 5% significance level calculated for Monte Carlo sample size 'MCSS'= 5000 also with various sample size n = 60, 120, 220, 350 as,

Size of the skewness test = Probability (Reject H_0/H_0 is true)

Here null hypothesis for DGP of this study will be normal, and our alternative hypothesis will be anyone from DGP-1 i.e. Beta, Chi-square, Log-normal or uniform distribution.

iv. Power of each measure and skewness test at 5% significance level calculated as,

Power of the skewness test = Probability (Reject H_0/H_0 is false)

3.3 Presence of Bimodality

This section explains the presence of bimodality from the mixture of normals (i.e. [2]N(0,1)+ N(μ_2, σ_2^2)). This study needs only the values of the parameters which represent bimodality denoted by '1' and ignore the values of parameters which show unimodality denoted by '0'. Further, the size of the bimodality is calculated with the help of numerical integrals, i.e. Trapezoidal and Simpson's rules.

3.3.1 Conditions for Bimodality

This study, also applied some important properties about bimodality in the mixture of normal, introduced by Robertson and Fryer (1969). Consider X_1 that comes from the unimodal normal distribution with their parameters X_1~N(μ_1, σ_1^2) and X_2 comes from the unimodal normal distribution with their parameters X_2~ N(μ_2, σ_2^2).

Z= {X_1 with mixing probability α, X_2 with mixing probability $(1-\alpha)$}

[2] This is the mixture of two normal densities N(0,1)= standard normal with μ_1= 0, σ_1^2= 1 and N(μ_2, σ_2^2)= normal with different values of μ_2 and σ_2^2

Or $$Z = \alpha X_1 + (1-\alpha) X_2$$

Whether 'Z' is unimodal or bimodal, it depends upon the ratios of the parameters 'α', $\mu = (\mu_2 - \mu_1)/\sigma_1$ and $\sigma = \sigma_2/\sigma_1$. As from the DGP-II in section (3.1.2), $X_1 \sim$ N (0,1) that is $\mu_1 = 0$ and $\sigma_1^2 = 1$ so $\mu = \mu_2$ and $\sigma = \sigma_2$. In the current study, these values are plug-in in some properties by Robertson and Fryer (1969) and got the following results.

i. Z is unimodal distribution if $0 < \mu \leq \mu_0$, where

$$\mu_0 = \left\{ \frac{2(\sigma^4 - \sigma^2 + 1)^{\frac{3}{2}} - (2\sigma^6 - 3\sigma^4 - 3\sigma^2 + 2)}{\sigma^2} \right\}^{\frac{1}{2}}$$

ii. If $\mu > \mu_0$ then 'Z' is a bimodal distribution also when 'α'lies in the open interval (α_1, α_2) as $\alpha_1 < \alpha < \alpha_2$.

$$(\sigma^2 - 1)Y_l^3 - \mu(\sigma^2 - 2)Y_l^2 - \mu^2 Y_l + \mu\sigma^2 = 0 \qquad (3.2)$$

Where

$$\alpha_l^{-1} = 1 + \frac{\sigma^3}{\mu - Yl} \exp\left\{ -\frac{1}{2} Y_l^2 + \frac{1}{2}\left(\frac{Yl - \mu}{\sigma}\right)^2 \right\} \text{ for l= (1, 2).}$$

$$\alpha_1 = \left[1 + \frac{\sigma^3}{\mu - Y1} \exp\left\{ -\frac{1}{2} Y_1^2 + \frac{1}{2}\left(\frac{Y1 - \mu}{\sigma}\right)^2 \right\} \right]^{-1}$$

$$\alpha_2 = \left[1 + \frac{\sigma^3}{\mu - Y2} \exp\left\{ -\frac{1}{2} Y_2^2 + \frac{1}{2}\left(\frac{Y2 - \mu}{\sigma}\right)^2 \right\} \right]^{-1}$$

Where Y_1 and Y_2 are the roots from the following equation,

With $0 < Y_1 < Y_2 < \mu$, otherwise 'Z' is unimodal distribution.

iii. If $\mu \leq 2$ time minimum of $(1, \sigma)$, 'Z' is unimodal distribution.

Otherwise, $\mu \geq \frac{3\sqrt{3}}{2}$ time minimum of 'Z' is bimodal distribution for $\alpha_1 < \alpha < \alpha_2$.

This study needs any two real roots of (Y_1, Y_2, Y_3) of the 'Y_{th}' cubical equation ignoring the negative and complex roots because according to the conditions restriction $0< Y_1 < Y_2 < \mu$ and set '$\alpha_1 < \alpha < \alpha_2$'. Then on this way, we try to find for which values of 'α, μ_2 and σ_2' the distribution is bimodal.

These three conditions are used for getting different values of three parameters (α, μ_2, σ_2) for mixture of normals of bimodal distribution. On the basis of these parameters, this study further investigates and compares the size and power of four modality tests in Section 2.3.2 along with the tests and measures of skewness in Section 2.2.1.

3.3.2 Determination of the Size of Bimodality

This study calculates the size of bimodality with the help of definite integrals (i.e. Trapezoidal or Simpson's rule). Riemann sums are used to find approximate area divided into rectangles. But there is low accuracy and Trapezoidal or Simpson's rule is applied which divides the area in trapeziums.

3.3.2.1 Trapezoidal Rule

It is difficult to evaluate the integrals through analytical methods. For this purpose, a numerical technique Trapezoidal rule is applied to find the approximate area.

$$\int_{a_0}^{b_0} f(x)dx = \frac{\Delta h}{2}[y_0 + 2(y_1 + y_2 + \ldots\ldots + y_{n-1}) + y_n)] \qquad (3.3)$$

Where a_0, b_0 are given limits, but in this situation $a_0 = \mu_1$ and $b_0 = \mu_2$, $\Delta h = \frac{b_0 - a_0}{n}$

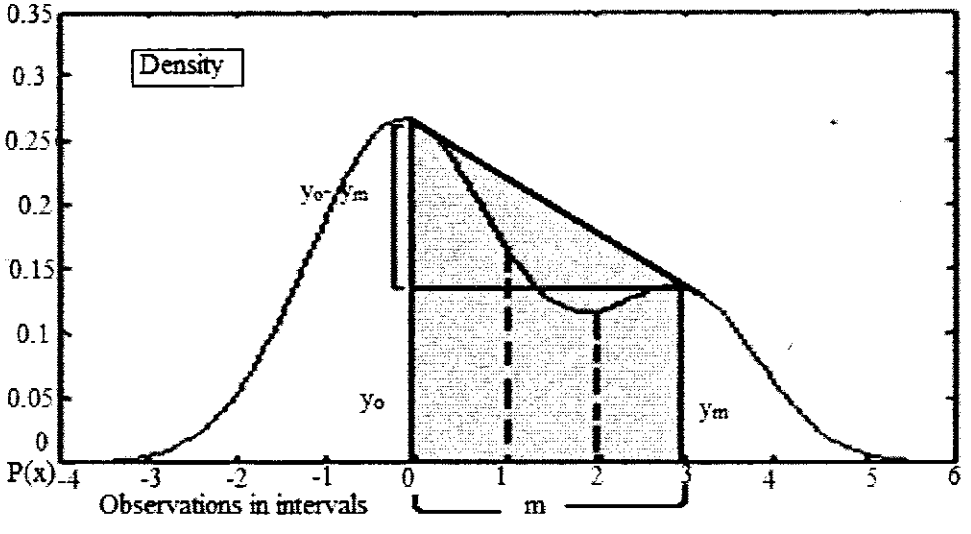
'n' is the number of the sub-intervals or trapeziums of the same length with (n+1) points. The accuracy increases as the value of 'n' increases and 'Δh' decreases.

3.3.2.2 Simpson's Rule

This is a Newton-Cotes formula to estimate the integral of a function $f(x)$, applying the second degree polynomials. It is obtained after the integration for ordering three Lagrange polynomial set the $f(x)$ on three same distance points. These points are y_1, y_2, y_3 and the same distance is denoted by 'h'. Simpson's rule performs better than the trapezoidal rule because it is too close to the exact area. Then according to the Simpson's rule,

$$\int_{a_0}^{b_0} f(x)dx = \frac{\Delta h}{3}[y_1 + 4(y_1 + y_3 + \ldots + y_{2n-1}) + 2(y_2 + y_4 + \ldots + y_{2n}) + y_n] \quad (3.4)$$

Figure 3.2: Size of the Bimodality within Two Modes



The size or degree of bimodality is found by using these two rules. The area that occurs between the two modes was calculated in the distribution through the Simpson's or trapezoidal rule. Similarly, space which lies above the distribution directly links the two modes or peaks as the size of the bimodal distribution. Combination of the area

calculated from trapezoidal or Simpson's Rule and size makes a trapezoid. The height of the first mode= y_0, second mode= y_m and the height $(y_0 - y_m)$ is the perpendicular of the right angle triangle. However, the distance from 'a_0' to 'b_0' is the length and y_m is the width of the rectangle. The size of bimodality is calculated as the difference between the area through Trapezoidal or Simpson's rule and the [3]area of the right angle triangle.

3.4 Split Sample Skewness Based Boxplot (SSSBB) and its Modification in case of Bimodality

This new technique was developed by Adil and Zaman (2012) and compared with the existing outliers detection techniques. They found that the performance of this technique was very well. In case of outliers detection, a basic issue arises in other techniques including Tukey's technique for outliers detection that increases the boundaries of critical values 'cv' where the data values fall minimum and do not consider the area of the data values which is highly skewed. This technique covers all these issues and extends boundaries of the 'cv' up to the original position of the distribution.

According to 'SSSBB' technique, the whole data series is divided into two portions as above median and below the median. The whole data limits are (12.5^{th} percentile, 87.5^{th} percentile) but the Tukey's technique contains the limits which are restricted only from the first quartile to third quartile (i.e. 25^{th}, 75^{th} percentiles). The summary statistics are calculated as, for below median Q_{1L}= first quartile lower (12.5^{th} percentile), Q_{3L}= third quartile (37.5^{th} percentile) and IQR_L= interquartile range lower. In the same way, above median Q_{1R}= first quartile upper (62.5^{th} percentile), Q_{3R}= third quartile upper (87.5^{th}

[3] Area of the right angle triangle 'A'= $\frac{n(y_n - y_0)}{2}$

percentile) and IQR_R= interquartile range upper. Mathematically interquartile ranges of both portions are described as,

$$IQR_L = Q_{3L} - Q_{1L}$$

$$IQR_R = Q_{3R} - Q_{1R}$$

The two boundaries of the data series 'L'= Lower 'cv' and 'U'=upper 'cv' as,

$$(L, U) = (Q_{1L} - 1.5*IQR_L, Q_{3R} + 1.5*IQR_R)$$

The values outside the two boundaries such as (L, R) are called outliers.

Adil and Zaman (2012) also modified 'SSSBB', incorporating in Kimber approach (1990) and Carling (1998).

Modifying SSSBB by using Kimber's Approach ($MSSSBB_k$) in which,

$$IQM_L = M_L - Q_{1L},$$

$$IQM_R = Q_{3R} - M_R,$$

Where 'IQM_L' is inter-quartile median on the left side and 'IQM_R' is inter-quartile median on the right side.

Similarly, $MSSSBB_k$ has the boundaries $(L, U) = (Q_{1L} - 1.5*IQM_L, Q_{3R} + 1.5*IQM_R)$

Modifying SSSBB by incorporating Carling technique ($MSSSBB_c$) as,

$$IQR_L = Q_{3L} - Q_{1L}$$

$$IQR_R = Q_{3R} - Q_{1R}$$

Here the current study divides the whole data series on cutoff point 'C' rather than median. Also finding quartiles of the left side as 'from minimum to C' and due to mixing the quartiles of the right side as "from minimum to maximum" of a bimodal distribution.

This study used Kimber's approach with slight changes in the fences as,

$$(L, U) = (Q_{2L} - 1.5*IQM_L \ , \ Q_{2R} + 1.5*IQM_R)$$

The SSSBB technique is used for detection of outliers in case of skewed distribution and bimodal distribution. This is quite interesting in the case of bimodality.

3.5 Data to be used

In real life, numerous fields such as economics, biological science, social sciences and physical sciences etc., the data shows unimodality or multimodality.

The current study used the KSE-100 Index return data of Pakistan (2013-2014). Exchange Rate (ER) annual data series (1961 to 2013) of Sweden, France, Germany, also consumption and exchange rate quarterly data from 1981-I to 2013-IV of Pakistan, UK, Fiji, and India. All these are taken from International Financial Statistics (IFS). This study used sports data of Pakistani cricket players' career, i.e. Ahmad Shehzad, Shoaib Malik and Umer Akmal, taken from http://www.espncricinfo.com (accessed date: 20th November 2017).

CHAPTER 4

EXISTENCE OF BIMODALITY AND COMPARISON OF MODALITY TESTS

In this chapter, bimodality is detected by using bimodality conditions and different modality tests are compared on the basis of Monte Carlo simulations, size and power.

4.1 Bimodality Conditions and the Existence of Bimodality

This section is based on the DGP-II, i.e. mixture of normal bimodal distribution. Here Robertson and Fryer's (1969) conditions are used for checking the bimodality existence, and for values of the parameters the distribution is bimodal. This study used a mixture of two distributions, i.e. standard normal $X_1 \sim N(\mu_1, \sigma_1^2)$ and normal distribution $X_2 \sim N(\mu_2, \sigma_2^2)$. Putting $\mu_1 = 0$, $\sigma_1^2 = 1$ and different values of other parameters $(\alpha, \mu_2, \sigma_2^2)$ that is $\alpha =$ (0.1, 0.2, 0.3,. . . . . ., 0.9), $\mu_2 =$ (1, 2, 3, ... ,10) and $\sigma_2^2 =$ (0.1, 0.2, 0.3, ... , 0.9). For different values of these three parameters $(\alpha, \mu_2, \sigma_2^2)$, getting the below sample results of unimodality= 0 and bimodality= 1.

Table 4.1: Bimodality Result with Changing of Parameters in a Mixture of Normals

μ_2	Result where $\alpha = 0.1, \sigma_2 = 0.6$	μ_2	Result where $\alpha = 0.1, \sigma_2 = 0.6$
1	0	6	0
2	1	7	0
3	1	8	0
4	1	9	0
5	0	10	0

Table 4.1 shows that when μ_2 are 2, 3, and 4 and α= 0.1 and σ_2^2= 0.6, then the distribution from the mixture of normals is bimodal while for all other μ_2 values with the same value of α and σ_2^2, a unimodal distribution is detected.

After the identification of bimodality in a mixture of normals, the authors get the following important results:

Table 4.2: Result of the Mixture of N (0, 1) + N (μ_2, σ_2)

When	$N(0, 1)+N(\mu_2, \sigma_2)$	Results
$\sigma_2 < \sigma_1$	real and −ve	0, 1
$\sigma_2 > \sigma_1$	complex	0 always

When $\sigma_2 < \sigma_1$ in mixture of normals (i.e. N (0, 1) + N (μ_2, σ_2^2)), then equation (3.2) results negative real roots and the distributions identified either unimodal '0' or bimodal '1'. On the other hand, when $\sigma_2 > \sigma_1$, then Equation (3.2) results complex roots and all results are unimodal.

Table 4.3: Result of the Mixture of Two Normals

When	$N(\mu_1, \sigma_1)+N(\mu_2, \sigma_2)$	Results
$\sigma_2 < \sigma_1$	Real ±ve	0, 1
$\sigma_2 > \sigma_1$	complex	0 always

When $\sigma_2 < \sigma_1$ in a mixture of normals (i.e. N (μ_1, σ_1^2) + N(μ_2, σ_2^2)), getting both positive and negative real roots of Equation (3.2), the resultant distribution is detected either unimodal '0' or bimodal '1'. But when $\sigma_2 > \sigma_1$, then Equation (3.2) results complex

roots and all results show unimodality. Now the current study displays overall parameters values which show a bimodal distribution.

Table 4.4: Summary of Parameters Values for Bimodality

μ_2, α, σ_2				
1, 0.1- 0.6, 0.2	2- 3, 0.1- 0.4, 0.5	2- 9, 0.1- 0.2,0.7	3- 4, 0.6, 0.7	2- 10, 0.1- 0.4, 0.9
1, 0.1- 0.9, 0.3	2, 0.5- 0.7, 0.5- 0.6	2-7, 0.3, 0.7	2- 10, 0.1- 0.6, 0.8	3- 10, 0.5- 0.8, 0.9
1, 0.5- 0.9, 0.4	2- 3, 0.3- 0.4, 0.6	2- 6, 0.4, 0.7	3- 10, 0.7- 0.8, 0.8	4- 10, 0.9, 0.9
1- 2,0.1- 0.4, 0.4	2- 4,0.1- 0.2, 0.6	2- 5, 0.5, 0.7	5- 10, 0.9, 0.8	

Note: 0.1- 0.6 means 0.1, 0.2, 0.3, ..., 0.6

Table 4.4 shows the combination of those parameters which have bimodality in case of a mixture of normals while ignoring the values of the parameters that show unimodality. This study uses all these parameter values to build bimodality distribution to make a comparison of bimodality tests by using the Monte Carlo simulation method on the basis of size and power properties of these tests.

4.2 Simulation Based Comparison of Modality Tests

In this section, various modality tests are compared, (i.e. Dip test, Proportional mass test (PM), Excess Mass Test (EM), and Silverman Bandwidth Test (SB)). This study compares all these four modality tests on the basis of the null hypothesis of unimodality versus alternative hypothesis of bimodality with different sample sizes, (i.e. n= 60, 120, 220, 350). It shows the size and power performances of these tests in different figures. First of all, this study finding the size of the modality tests through Monte-Carlo simulations.

4.2.1 Size of the Modality Tests

As for modality tests the null hypothesis is based upon unimodality and alternate hypothesis on bimodality in a distribution. Here, simulated critical values are calculated corresponding to sample sizes in order to stabilize the size of all tests around the nominal size of 5%.

Figure 4.1: Size of the Modality Tests

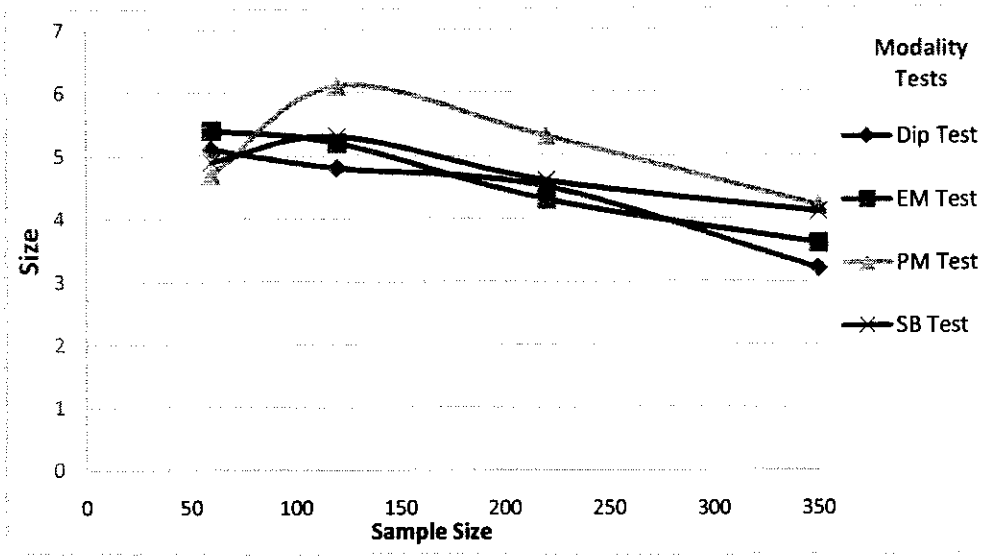


Figure 4.1 describes the Monte Carlo simulation results of the size of modality tests correspond to various sample sizes. At small sample size (n= 60), the size of all tests fluctuates around the nominal size of 5% in which minimum size of 4.7% is identified corresponding to PM test while maximum size of 5.4% is detected for EM test.

As the sample size increases from n= 60 to n= 120, then a similar variation of size for all tests has been observed as has been shown for n= 60, in which Dip test has achieved a minimum size of the size 4.8% while PM test has the highest size of 6.1%. Further, as the sample size gets larger (i.e. n= 220), all four tests (i.e. Dip, EM, PM, and SB) have sizes between 4.3% and 5.3%. For a very high sample size of n= 350, the Dip test has

minimum size of 3.2% while the maximum size is observed 4.2% which is the size of the PM test.

It means that all the modality tests have stable size with 5% nominal simulated critical values. Therefore, the researchers may compare these modality tests further.

4.2.2 Power based Comparison of Modality Tests

The current study compared four modality tests on the basis of power property corresponding to sample sizes n= 60, 120, 220 and 350. Figure 4.2 to Figure 4.9 show the power performance of modality tests with various parameter values.

Figure 4.2: Power of Modality Tests with Parameters (μ_2, α, σ_2)= (1, 0.6, 0.2)

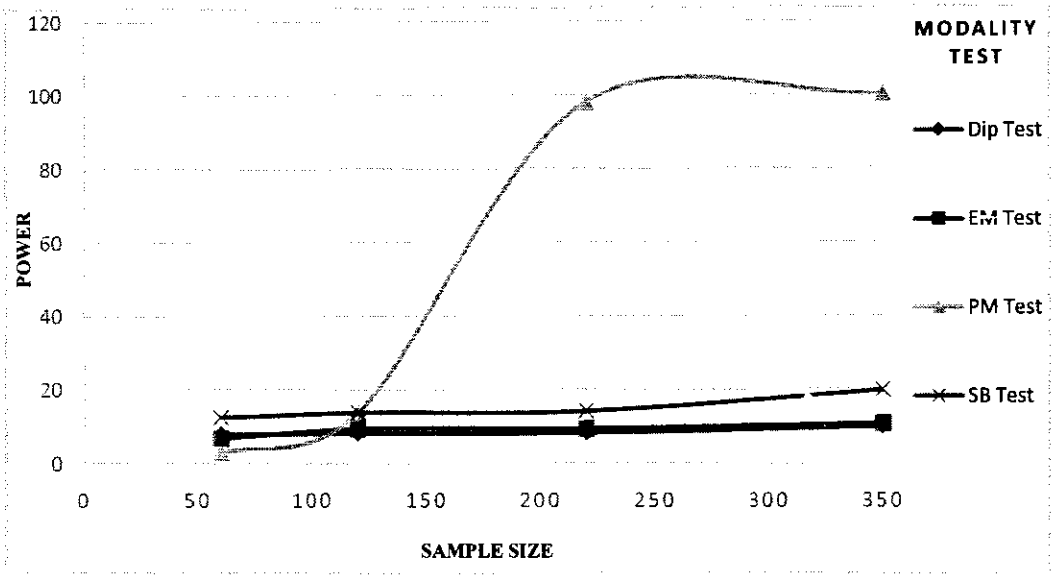


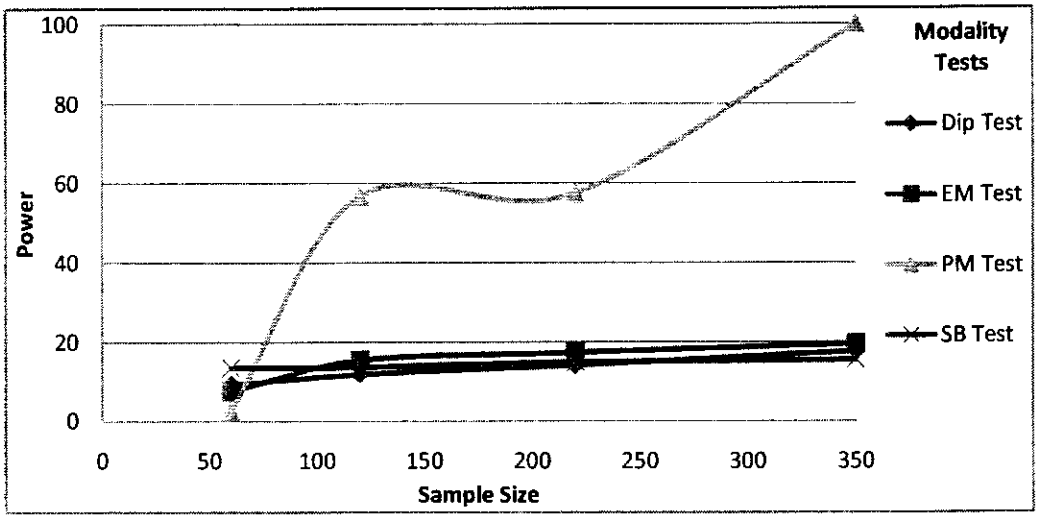
Figure 4.2 describes the power behavior of four modality tests with different sample sizes and parameters values as μ_2= 1, α= 0.6 and σ_2= 0.2. At small sample size (i.e. n= 60), all the four modality tests have gained low power in between 2.8% to 12.4%, in which PM test has got the least power (i.e. 2.8%) while SB test with 12.4% power is identified as the powerful test. While increasing the sample size from n= 60 to 120, the three tests (i.e. Dip, EM, and SB) maintain the same power pattern as has been observed at n= 60. It is

observed that the PM test has gained a little increase in its power behavior while all other three tests have remained with the same power pattern as compared to their results at n= 60. Moreover, as the sample size increases from n= 120 to n= 220 and n= 350, Figure 4.2 indicates that PM test with a rapid increasing pattern in its power is identified as the most powerful test as compared to other three tests. However, among these three least powerful tests, SB test has a little increase in its power behavior while EM and Dip tests have the same power pattern at sample size 220 and 350.

Overall, Figure 4.2 concludes that the PM test is the most powerful test as the sample size increases as compared to all other three tests. SB test with a very little increase in its power is identified as the second best performing test while EM and dip tests with constant behavior for overall sample sizes are recognized as bad performing tests.

Keeping μ_2 and σ_2 constant and by changing the selected α = (0.1, 0.2, 0.3, ... , 0.6), and also for the case where $\mu_2 = 1$, $\sigma_2 =$ 0.4 and α= (0.5, 0.6, ..., 0.9), the simulated power results remain approximately same as shown in the above Figure 4.2.

Figure 4.3: Power of Modality Tests with Parameters (μ_2, α, σ_2) = (1, 0.8, 0.3)



The above Figure 4.3 shows the power results of modality tests with parameters values which are μ_2= 1, α= 0.8 and σ_2= 0.3. At sample size (i.e. n= 60) all the four modality tests have gained low power in between 2.8% to 12.4%, in which PM test has got the least power (i.e. 2.6%) while SB test with 13% power is identified as the powerful test.

When increasing the sample size as n= 120, the three tests (i.e. Dip, EM, and SB) have a minimum increase in the power. It is notified that the PM test has increased its power behavior (i.e. power= 57%). When the sample size increases from n=120 to n=220 and n= 350, Figure 4.2 indicates that the PM test with a rapid increasing pattern in its power (i.e. at n= 220 the power is 58% and at n= 350 the power moves to 100%) has been identified as the most powerful test as compared to other three tests. However, among these three least powerful tests, EM test has a little increase in its power behavior (i.e. Power reaches to 20% on n= 350) while SB and Dip tests have same low power pattern at sample size 220 and 350.

Overall, Figure 4.3 results that the PM test is the most powerful test when the sample size increases as compared to all other three tests. EM test with a very little increase in its power is identified as the second best performing test while SB and dip tests with constant behavior over all sample sizes are recognized as bad performing tests. Similarly, keeping μ_2 and σ_2 constant and by changing the particular α= (0.5, 0.6, 0.7, 0.9), the simulated power results remain approximately the same as shown in the above Figure 4.3.

Figure 4.4: Power of Modality Tests with Parameters (μ_2, α, σ_2)= (9, 0.2, 0.7)

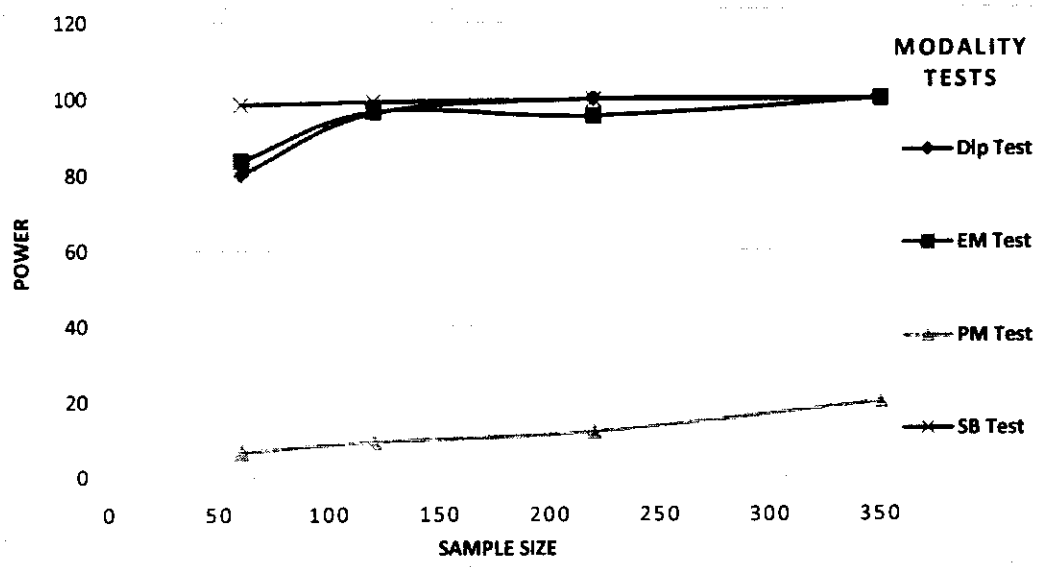


Figure 4.4 describes the power of all the modality tests at different sample sizes and the parameters values (i.e. μ_2= 9, α= 0.2 and σ_2= 0.7). When sample size (i.e. n= 60) all the three modality tests have high powers (i.e. Dip= 80%, SB= 100%, EM= 85%), but PM test has got the low power (i.e. 6.7%). So SB test with high power is identified as the powerful test. At sample size n= 120, the two tests (i.e. Dip and EM) power moves to 99%. Again, SB test is identified as a high powerful test and PM test has low power (i.e. 9%).

As the sample size increases from n= 120 to n= 220 and n= 350, Figure 4.4 indicates that PM test with a slowly increasing pattern in its power (i.e. reaches to 20%) is identified as the lowest power test as compared to other three tests. It is observed that the other three tests (i.e. Dip, SB and EM) maintain the same power pattern as has been detected at n= 120.

It means Figure 4.4 implies that the SB test is the most powerful test as the sample size increases as compared to all other three tests. Moreover, Dip and EM test with an equal increase in its power, are identified as the second best performing tests while PM test with low power behavior at overall sample sizes is recognized as bad performing test.

The results of the above Figure 4.4 remain the same while keeping α= 0.2 and σ_2= 0.7 constant and changing μ_2 as (7 or 8).

Figure 4.5: Power of Modality Tests with Parameters (μ_2, α, σ_2)= (6, 0.4, 0.7)

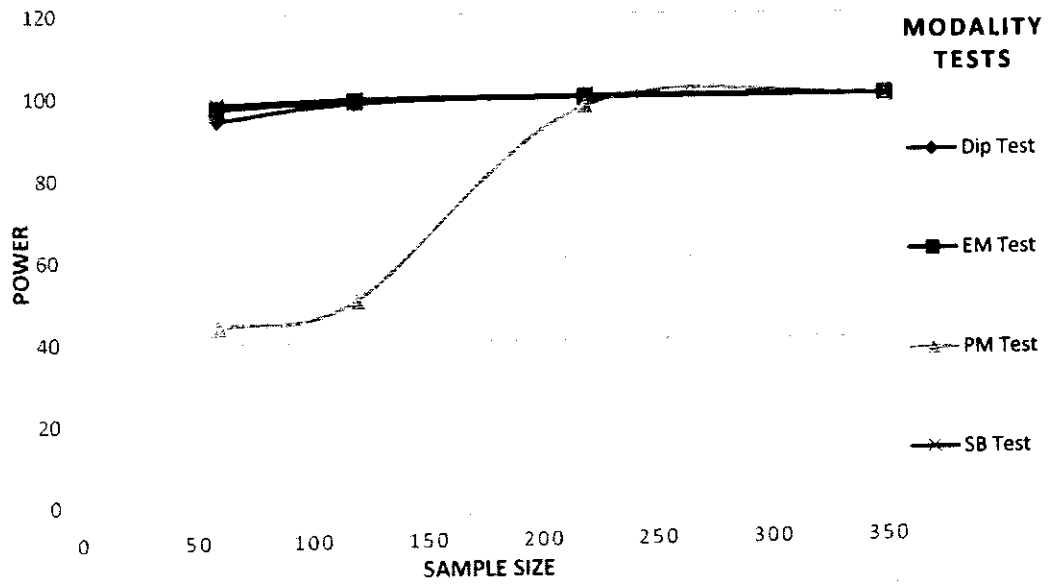


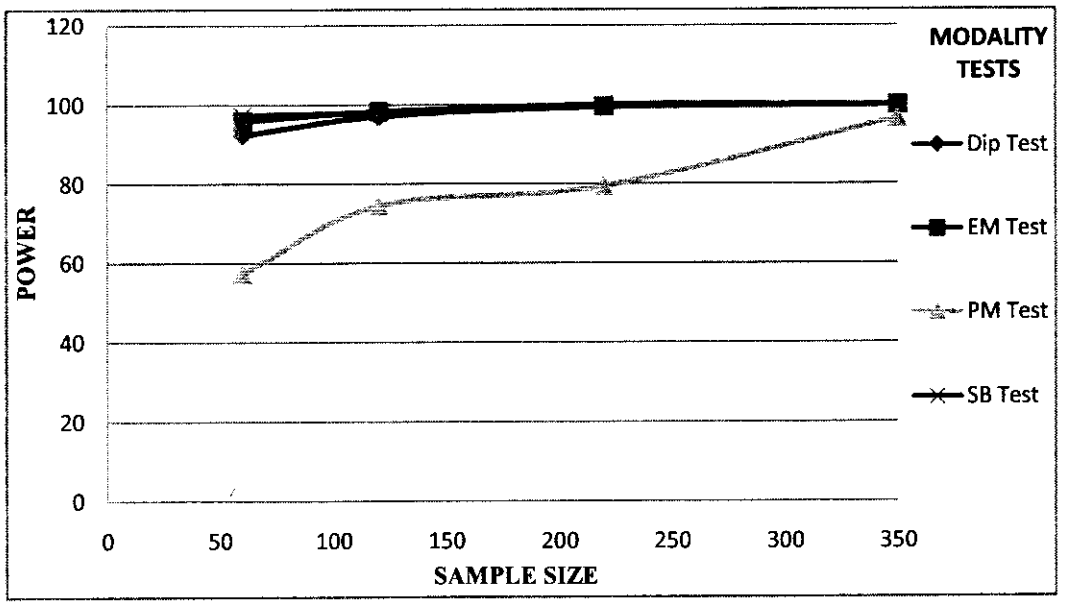
Figure 4.5 describes the power behavior of four modality tests with different sample sizes and parameters values as μ_2= 1, α= 0.6 and σ_2= 0.2. At small sample size (i.e. n= 60), the three modality tests (i.e. Dip, EM, and SB) have gained a high probability of accepting bimodality near 100%. From this result, both SB and EM test are the most powerful tests while the PM test has less power (i.e. 43%).

As increasing the sample size from n= 60 to 120, the three tests (i.e. Dip, EM, and SB) have also increased their power and reached to 100% and it is observed that PM test has

gained a little increase in its power behavior equal to 50%. Moreover, as the sample size increases from n= 120 to n= 220 and n= 350, Figure 4.5 indicates that PM test with rapid increasing pattern in its power is identified as 96% and moves to 100% in case of n= 350 while the other three tests (i.e. Dip, EM, and SB) have high power equal to 100%.

Overall, Figure 4.5 concludes that all the tests have high power at high sample sizes. But at a small sample, the PM test has low power which is bad performing test. In this case, for changing 'α' to 0.5 and keep the same values for other parameters, then the results remain same as shown in the above Figure 4.5.

Figure 4.6: Power of Modality Tests with Parameters (μ_2, α, σ_2)= (8, 0.3, 0.8)



The above Figure 4.6 shows the power result of DGP-II mixture of normal where parameters are μ_2= 8, α= 0.3 and σ_2= 0.8 at various sample sizes. For small sample size n= 60, the three modality tests (i.e. Dip, EM, and SB) have been observed high power. It is identified from the result of this sample size that both SB and EM tests have high power (i.e. 97.3%) while the power of Dip test is detected as 92.3% and PM test has less

power (i.e. 57.3%). At sample size n= 120, the three tests (i.e. Dip, EM, and SB) have high power equal to 98% but the power of EM test increases (i.e. 74.4%). As the sample size increases from n= 120 to n= 220, Figure 4.6 indicates that PM test with increasing pattern in its power from 74.4% to 79.9%, while the power of other three tests (i.e. Dip, EM, and SB) reaches to 100%. At n= 350 the three tests (i.e. Dip, EM, and SB) maintain the same power pattern as has been observed at n= 220 while it is observed that PM test has gained a high increase in its power behavior moves to 97%.

Overall, Figure 4.6 concludes that SB and EM tests are the most powerful tests at all samples as compared to other two tests. Similarly, PM test with a fluctuation in its power is identified as the bad performing test in this type of situation. Keeping α and σ_2 constant and by changing μ_2 as (7 or 9), the results of the above Figure 4.6 remain approximately the same. Also for the case where $\alpha = 0.7$, $\sigma_2 = 0.8$ and $\mu_2 = (5, 6, 7, 8, 9)$, the simulated power results remain unchanged.

Figure 4.7: Power of Modality Tests with Parameters (μ_2, α, σ_2)= (7, 0.9, 0.8)

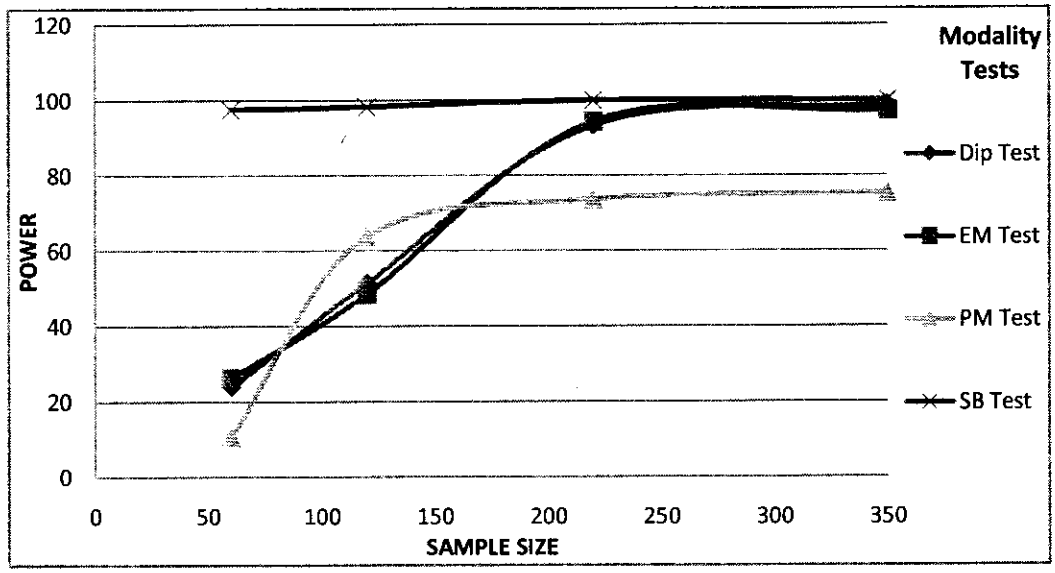


Figure 4.7 describes the power results of four modality tests while changing the parameters μ_2= 10, α= 0.7 and σ_2= 0.8 in DGP-II for various sample sizes. When sample size is small (i.e. n= 60), the three tests (i.e. Dip, PM and EM) have low power 'between 10.7% and 25%' while SB is observed as a most powerful test with the power of 97.5%. At sample size n= 120, the three tests (i.e. Dip, PM, and EM) have power increased and power of PM test is identified as 63.8%. Again, SB is the most powerful test with power 98.3% while EM test with low power at 48.3%. As the sample size increases from n= 120 to n= 220 and 350, Figure 4.7 indicates that SB test maintains its supremacy and its power reaches to 100% while the power of all other three tests also increases (i.e. PM= 63%, Dip and EM= 94% each). Hence, PM test is detected as the least powerful test.

Result of Figure 4.7 implies that SB test is the most powerful test at all sample sizes as compared to other three tests while PM is bad performing test in most sample sizes. The results remain same of the above Figure 4.7 while keeping the parameters values α= 0.9, σ_2= 0.8 constant and changing μ_2= (8, 9, 10). Also, the results remain unchanged when the parameters values are α= 0.9, σ_2= 0.9 fixed and changing μ_2= (7, 8, 9).

Figure 4.8: Power of Modality Tests with Parameters (μ_2, α, σ_2)= (10, 0.2, 0.9)

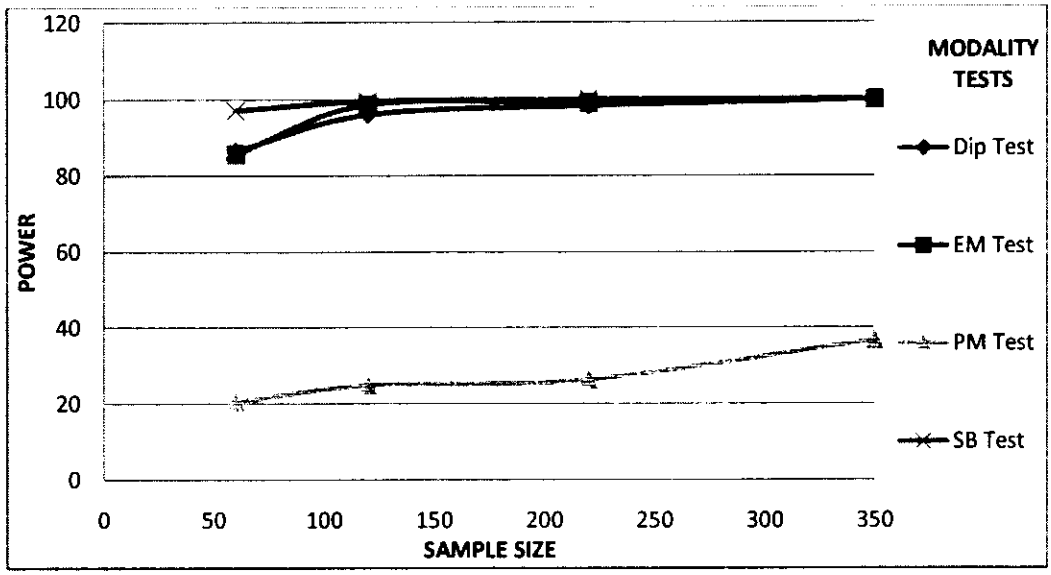


Figure 4.8 describes the power of all the modality tests at different sample sizes and the parameters values (i.e. μ_2= 10, α= 0.2 and σ_2= 0.9). With sample size n= 60 all the three modality tests have high powers (i.e. Dip= 86%, SB= 97.1%, EM= 85.8%), but PM test has got the low power (i.e. 20.4%). So SB test with high power is identified as the powerful test with the power 96%. At sample size n= 120, the two tests (i.e. Dip and EM) power move to 98%. Again SB test is identified as a high powerful test and PM test has low power (i.e. 24.8%). As the sample size increases from n= 120 to n= 220 and n= 350, Figure 4.8 indicates that PM test with a slowly increasing pattern in its power (i.e. reaches to 35%) is identified as the lowest power test as compared to other three tests. It is observed that the other three tests (i.e. Dip, SB and EM) maintain the same power pattern (i.e.100%) as have been detected at n= 120.

Overall, Figure 4.8 concludes that the SB test is the most powerful test as the sample size increases as compared to all other three tests. Moreover, Dip, and EM test with an equal increase in its power, are identified as the second best performing tests while PM test

with low power behavior at overall sample sizes is recognized as bad performing test. Keeping α= 0.2 and σ_2= 0.7 constant and changing μ_2 as (8 or 9), the results of the above Figure 4.8 remain the same.

Figure 4.9: Power of Modality Tests with Parameters (μ_2, α, σ_2)= (7, 0.6, 0.9)

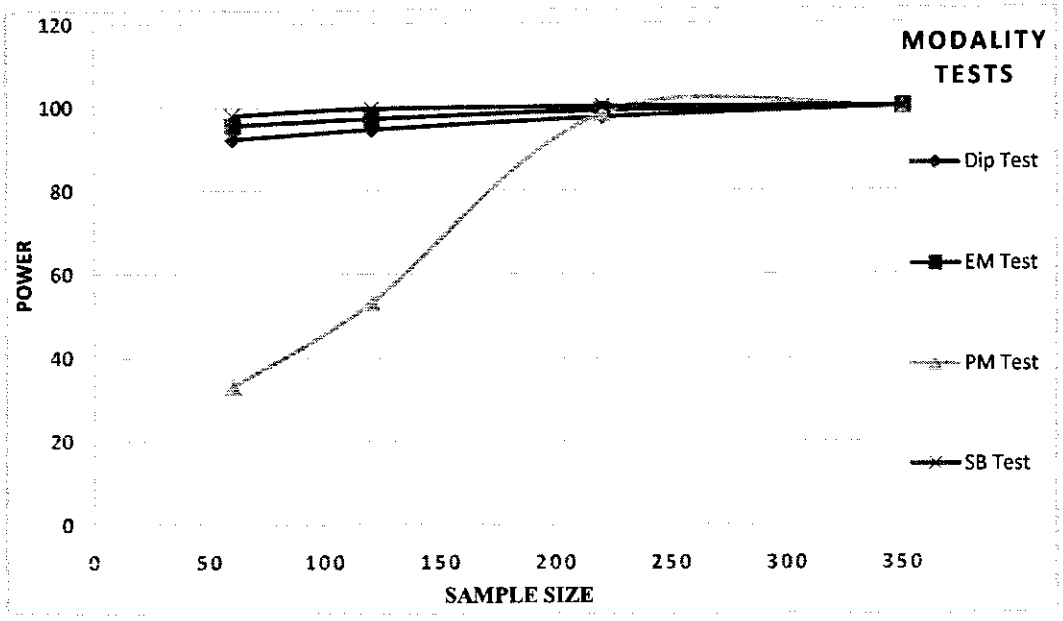


Figure 4.9 shows the power behavior of four modality tests with different sample sizes and parameters values as μ_2= 7, α= 0.6 and σ_2= 0.9. At small sample size (i.e. n= 60), the three modality tests (i.e. Dip, EM, and SB) have gained a high probability of accepting bimodality near 100%. From this result, SB test is the most powerful test while the PM test has less power (i.e. 33.2%). As increasing the sample size from n= 60 to 120, the three tests (i.e. Dip, EM, and SB) have also maintained their power and it is observed that PM test has gained an increase in its power behavior equal to 53.3%. Moreover, as the sample size increases from n= 120 to n= 220 and n= 350, Figure 4.9 indicates that all the modality tests (i.e. Dip, EM, and SB) have high power equal to 100%.

Overall, Figure 4.9 concludes that the SB test has high power at various sample sizes. But at a small sample, the PM test has low power which is bad performing test. Keeping α= 0.6 and σ_2= 0.9 constant and changing μ_2 as (5, 6, 8, 9 or 10), the results of the above Figure 4.9 remain approximately the same.

4.3 Chapter Summary

Using Robertson's and Fryer's (1969) bimodality conditions for the detection of bimodality framework is presented in this chapter. The mixture of two normal distributions such as standard normal $X_1 \sim N(\mu_1, \sigma_1^2)$ and normal distribution $X_2 \sim N(\mu_2, \sigma_2^2)$ with mixing proportion 'α' was used as a DGP.

From the above analysis of Section 4.1, it is concluded that when $\sigma_2 < \sigma_1$ in the mixture (one standard normal and second normal), then the result becomes negative real roots of Equation (3.2) and the distributions detected either unimodal '0' or bimodal '1'. In the situation when $\sigma_2 > \sigma_1$, then Equation (3.2) results as complex roots and all other results show that the mixture is unimodal. When $\sigma_2 < \sigma_1$ in a mixture of two normal distributions, then Equation (3.2) results both positive and negative real roots where the distributions are identified either unimodal '0' or bimodal '1'. But when $\sigma_2 > \sigma_1$, then Equation (3.2) results complex roots and all results show unimodality. However, to get a bimodal distribution the detailed combinations of the parameters are shown in Table 4.4. Using these parameter values, the size and power of modality tests are calculated. According to the 5% nominal simulated critical values, all the selected modality tests have stable sizes.

In power comparison, keeping the mean fixed μ_2= 1 and increasing the other two parameters of the mixture, the PM test has comparatively high power. As increasing the

mean and different values of the other two parameters, all the tests performed well, except PM test which has low power. For further changes in these parameters values, the power of Silverman test is very high which states that this is the robust and powerful test while the PM test is recognized as bad performing test.

CHAPTER 5

NEWLY INTRODUCED MEASURE OF SKEWNESS ON THE BASIS OF P-NORM

Any measure which is equal to zero for a skewed distribution is not considered as a good measure. All of the existing standard measures have this property. There are skew distributions for which the skewness measure is zero. Therefore the measure says there is no skewness when the distribution is, in fact, a skewed distribution. By symmetrizing the data set around the median, then measure the distance between the symmetrized distribution and the original distribution, we get a measure of skewness. Based on this idea, this study introduced and used this new proposed measure.

This chapter explains the procedure of newly introduced measure and its preference over existing measures about the detection of skewness. This chapter also discusses the advantages of measure P-norm (P_{norm}) and highlights the comparison on real data sets of this measure with other measures of skewness.

5.1 Procedure of New Measure of Skewness P-norm

This measure is based on P-norms or 'L_p' and cumulative distribution function '$F(x)$' of a distribution. First of all, a data series have to be symmetrized around the median and then measuring the distance between the symmetric cumulative distribution 'CDF= $F(x)$' and the original data 'CDF= $F(x_1)$'. After calculating the absolute difference of these two CDFs the P-norms (L_0, L_1 and L_2 while p= 0, 1, 2) can be found in the following way;

$$L_0 = ||F(x) - F(x_1)||_0 = Sup_{i=1}^{n} |F(x) - F(x_1)| \qquad (5.2)$$

$$L_1 = \| F(x) - F(x_1)\|_1 = \sum_{i=1}^{n} |F(x) - F(x_1)| \quad (5.3)$$

$$L_2 = \|F(x) - F(x_1)\|_2 = [\sum_{i=1}^{n} |F(x) - F(x_1)|^2]^{\frac{1}{2}} \quad (5.4)$$

Figure 5.1: CDFs of a Data Series

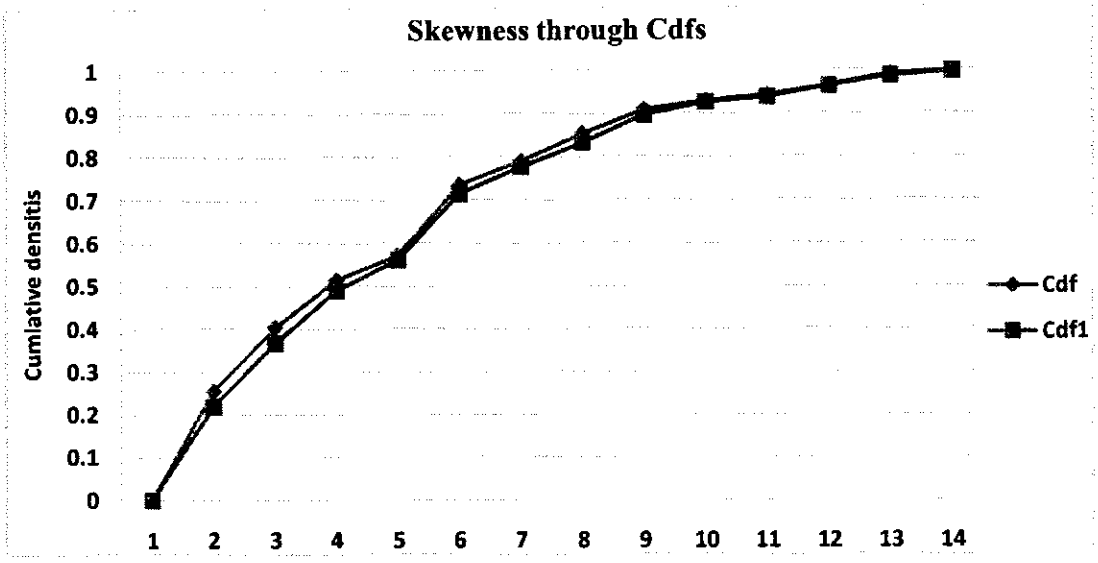


Figure 5.1 shows the two CDFs as the CDF for original data and CDF for symmetrized data. The distance between two CDFs displays the measure of skewness which is measured through P_{norm}. Furthermore, when a line of original CDF is above the symmetrized CDF, this shows the negatively skewed property. But when the line of original CDF is below the symmetrized CDF this means that the distribution is positively skewed. In case of a small or negligible difference, there is no skewness.

There are so many forms of norm and different names such as Euclidean distances and also L_0, L_1 are natural norms and L_2 is called standard norm. These are three measures through which we decide whether the distribution is skewed or not. If all of these or at least two of them will equal to zero then the distribution is symmetric otherwise skewed.

This study introduced and used a new measure, and this is the only measure for which all skewed distributions have non-zero measures.

5.2 Advantages of Measure P-norm over Existing Techniques

This new measure P-norm has several advantages over the existing measures of skewness which are given below:

i. Exact/Correct Measurement in each Case

This measure shows the exact measurement in case of skewed and symmetric distributions. But the existing measures sometimes give a false alarm about the skewness.

ii. Based on each Observation

Most measures have used some specific values of data set to find skewness. This new measure includes all of the observations to detect skewness.

iii. New way/approach with Graphical Manipulation

Along with a numerical measure, this new technique also describes a clear picture on the graphical representation. According to this measure, the difference between the two CDFs shows the skewness of a distribution.

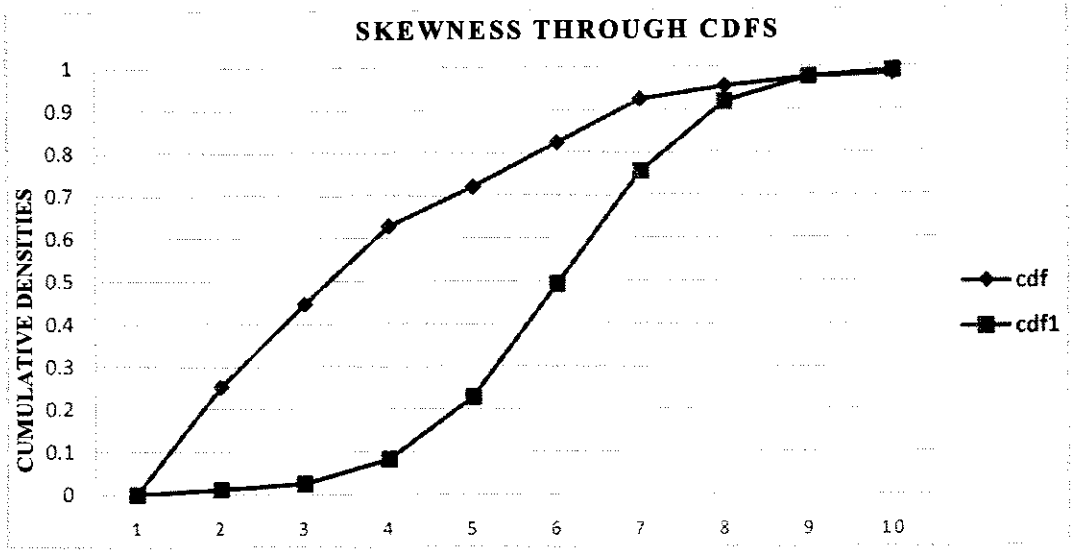
iv. The Efficient Result in the Presence of Outliers

Most measures of skewness divert their result opposite with the addition of outliers in a data. But the advantage of this new measure P_{norm} results separately in original format with or without outliers.

5.3 Highlighting New Measure with the Existing Technique

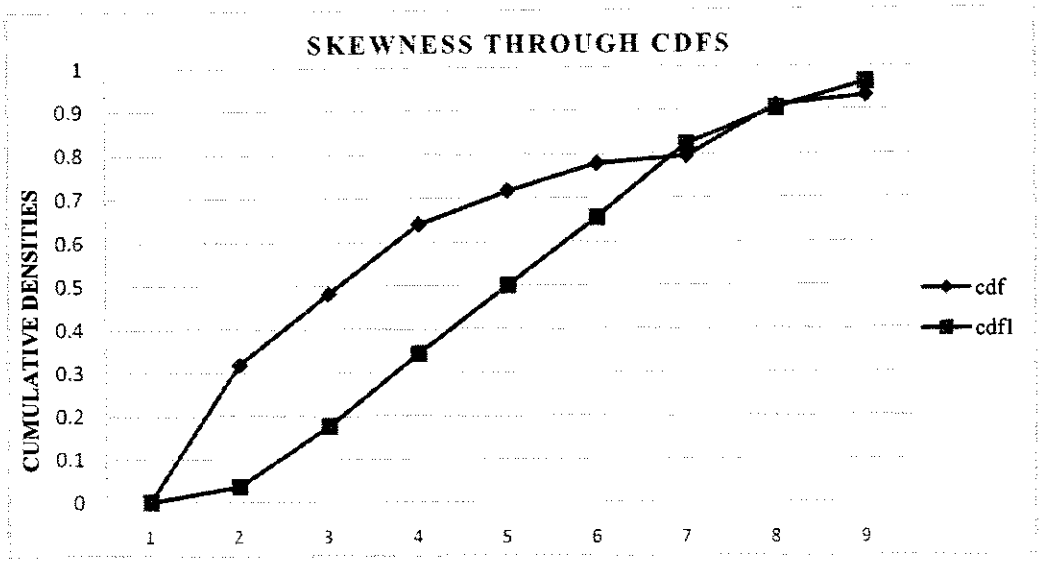
With the help of some real data sets, the properties of newly introduced measure are highlighted with other classical measures of skewness.

Figure 5.2: Skewness Measure of Akmal ODI Scores



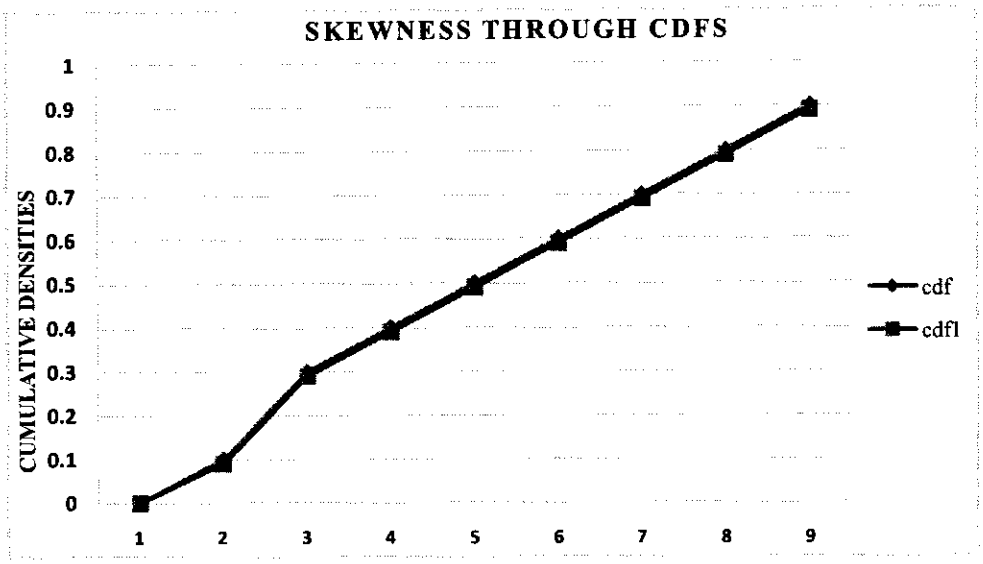
For cricket data of Umer Akmal ODI scores, the measures skewness-1, skewness-2, SSSB and standardized moment show that the series is highly positively skewed. But according to the Pearsonian coefficient and med-couple, the data series is nearly symmetric. This new measure P_{norm} describes that the data series is highly positively skewed. The above Figure 5.2 clearly shows the high difference between the two CDFs, which is calculated through P_{norm}. Therefore, the new measure is also useful to elaborate the skewness graphically.

Figure 5.3: Skewness Measure of Shafiq T20I Scores



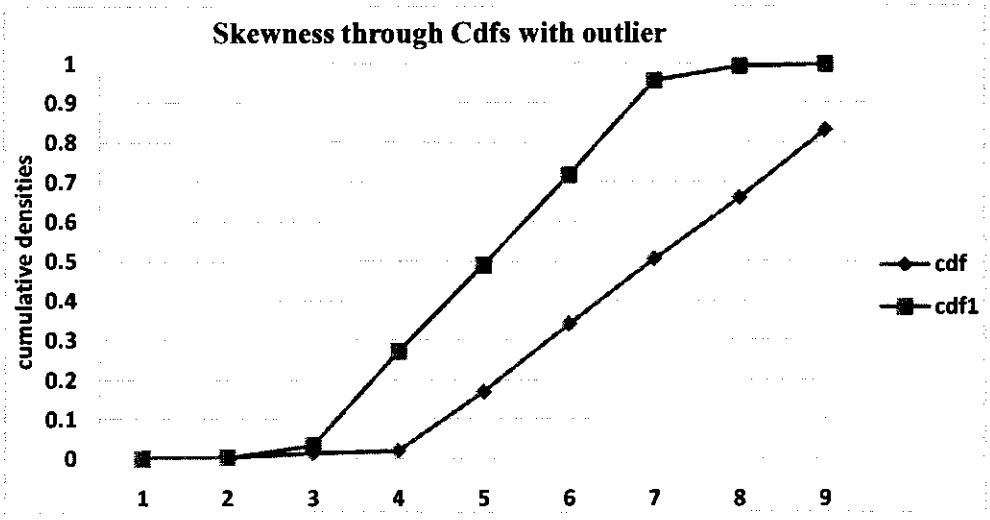
In this example of Asad Shafiq T20I scores, the measures of skewness-I, skewness-II and SSSB show that the series is highly positively skewed. But according to the Pearsonian coefficient, standardized moment and med-couple results, the data series is nearly symmetric. The measure P_{norm} describes that this data series is highly positively skewed. The above Figure 5.3 also shows the high difference between the two CDFs, which is calculated through P_{norm}. It means that the existing measures describe various results in different situations as compared to the proposed measure, i.e. P_{norm}.

Figure 5.4: Skewness Measure of Numbers from (−25 to 25)



The above Figure 5.4 describes the CDFs difference which is approximately zero. It means that the series of 51 observations from (−25 to 25) is symmetric. According to all other measures, the result remains same.

Figure 5.5: Skewness Measure of Numbers with the Addition of Outlier from (−25 to 60)



When replaced a number i.e. '0' with an outlier '60', we get various results in Figure 5.5. The measures SSSB and SM still show the same result that the data is symmetric while Skewness-1 and Skewness-2 show asymmetric. But the measures Pearsonian and Med-couple result that this series is negatively skewed. Actually, the series shifted to the positively skewed with the addition of outliers. When replacing the outlier as '-60', then only the three measures i.e. standardized moment, Pearsonian and P_{norm} results show that this series is negatively skewed. It means that the new measure P_{norm} perform well in the presence of outliers as compared to other measures.

CHAPTER 6

COMPARISON OF VARIOUS MEASURES AND TESTS FOR SKEWNESS

In this chapter, the study compared different measures of skewness (i.e. Pearsonian Coefficient (Prs), Standardized Moment (SM), Med-Couple (MC), Spilt sample Skewness Boxplot (SSSB), Skewness-1 (Skw1), Skewness-2 (Skw2) and newly introduced measure P-Norm (P_{norm})) and skewness tests (i.e. Kolmogorov-Smirnov (KS) test, Student's t-test, Wilcoxon (WC) test). The comparison is based on the testing hypothesis of symmetry and asymmetry with different sample sizes as (i.e. n= 60, 120, 220, 350).

6.1 Size of the Measures and Tests for Skewness

Here, simulated critical values are used to check the size of all considerable measures, and tests of skewness have stable sizes around 5%.

Figure 6.1: Size of Measures and Tests for Skewness

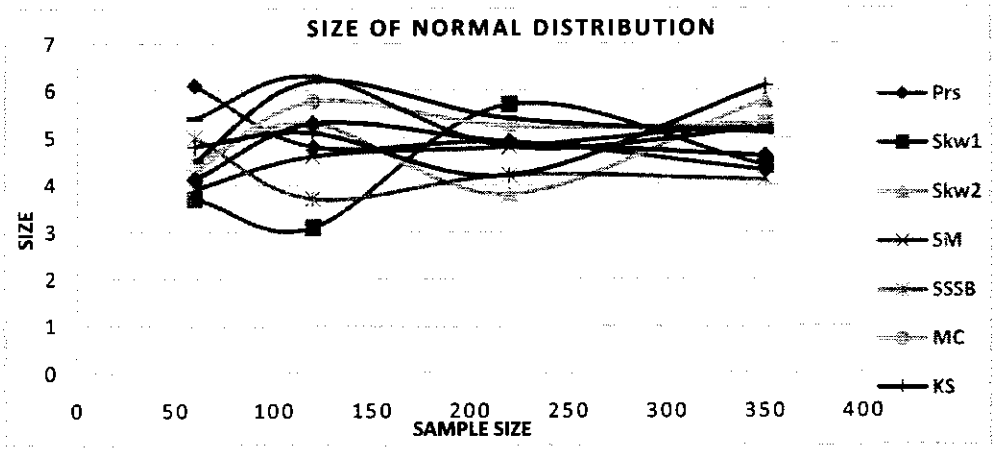


Figure 6.1 describes the Monte Carlo simulation results of the size of measures and tests for skewness corresponding to various sample sizes. At small sample size (n= 60), the

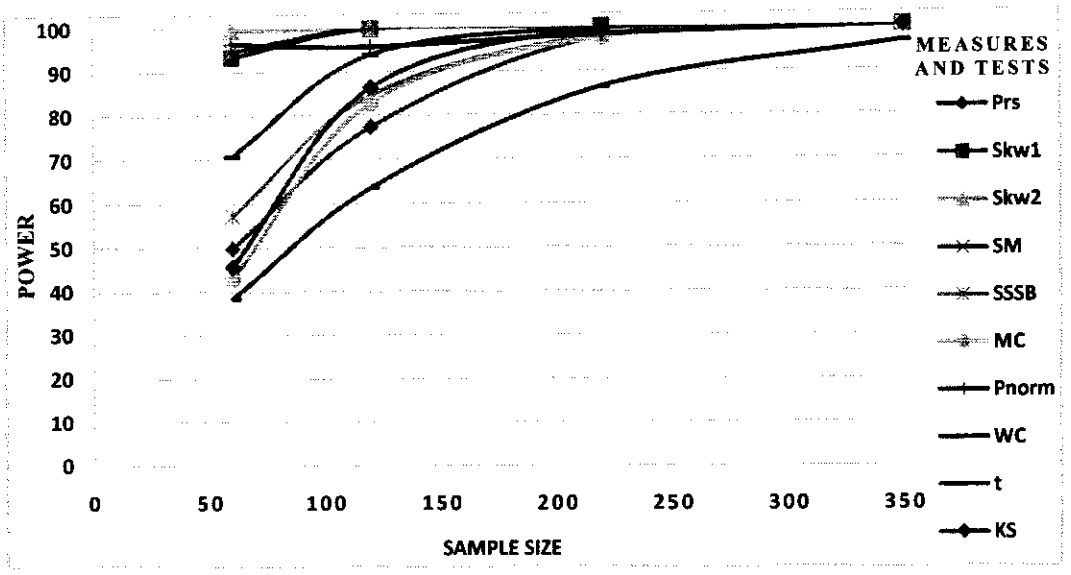
size of all measures and tests fluctuates around the nominal size of 5% in which minimum size of 3.7% is identified corresponding to measure Skw_1 while maximum size of 6.1% is detected for measure Prs. As the sample size increases from n= 60 to n= 120, then a similar variation of size for all tests has been observed as has been shown for n= 60 in which measure Skw_1 has achieved a minimum size of the size 3.1% while t-test has the highest size of 6.3%. Further, as the sample size gets larger (i.e. n= 220), all the measures (i.e. Prs, SM, MC, P_{norm}, SSSB, Skw_1 and Skw_2) and tests (i.e. KS, WC, t-test) have sizes in between 3.8% to 5.7%. For a very high sample size of n= 350, the measure SSSB has minimum size of 4.1% while the maximum size is observed 6.1% which is the size of the KS test.

Overall, Figure 6.1 concludes that all the measures and tests for skewness have stable size with 5% nominal simulated critical values. Therefore, these measures and tests can be compared further on power behavior.

6.2 Power of the Measures and Tests for Skewness

This section contains the power comparison of the various measures and tests for skewness. The comparison is made on the basis of Monte Carlo simulations with their size MCSS= 5000 times. The skewness test consists of the null hypothesis that there is no skewness. In the following Figure 6.2 to Figure 6.11 the power of the tests is plotted on the y-axis and the sample size is adjusted on the x-axis.

Figure 6.2: Power of Log-Normal Distribution with Mean= 0 and SD= 0.5



The above Figure 6.2 shows the power of various measures and tests of skewness using the DGP-I of log-normal distribution with parameters (0, 0.5). At sample size n= 60, the power of measures (i.e. MC, Prs, SSSB) and tests (i.e. WC, KS) are low and below 50%. But other measures (i.e. Skw_1, Skw_2, SM, P_{norm}) and t-test have high power nearly 100%. Also, it is observed that the WC test is detected as the least powerful test while Skw_2 is the most powerful measure among all measures and tests. Moreover, measures P_{norm} and SM with power over 90% are also very close to the power of Skw_2 measure.

As the sample size increases from n= 60 to n= 120, then again WC with 62% power is detected as the least powerful test; while at the same sample size, all measures and tests behave very similarly. Now measure Skw_1 has achieved high power along with P_{norm} and Skw_2. Also, at sample size n= 120, KS test has achieved much better power then these two measures (i.e. Prs and SSSB), because at sample size n= 60, its power was less as compared to these two measures. As the sample size further increases from n= 120 to 220, the majority of the measures and tests achieve maximum power (i.e. 100%) while

WC with increasing pattern again remains the lowest powerful test. At sample size n= 350, all measures and tests achieve approximately 100% power while WC has achieved maximum power (i.e. 96%). But again, WC test remains the least powerful test even though its power is very close to 100%. In all cases for which the sample size is greater than 120, the performance of measures (i.e. Skw_1, Skw_2, SM and P_{norm}) have power approximately 100% on this DGP.

Figure 6.3: Power of Log-Normal Distribution with Mean= 6, SD= 3

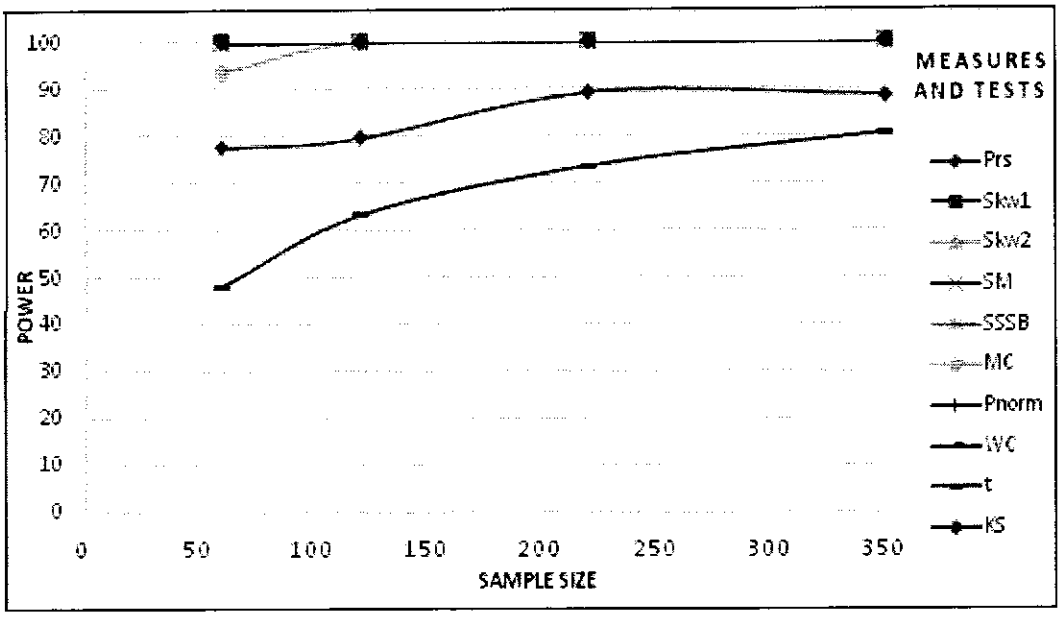


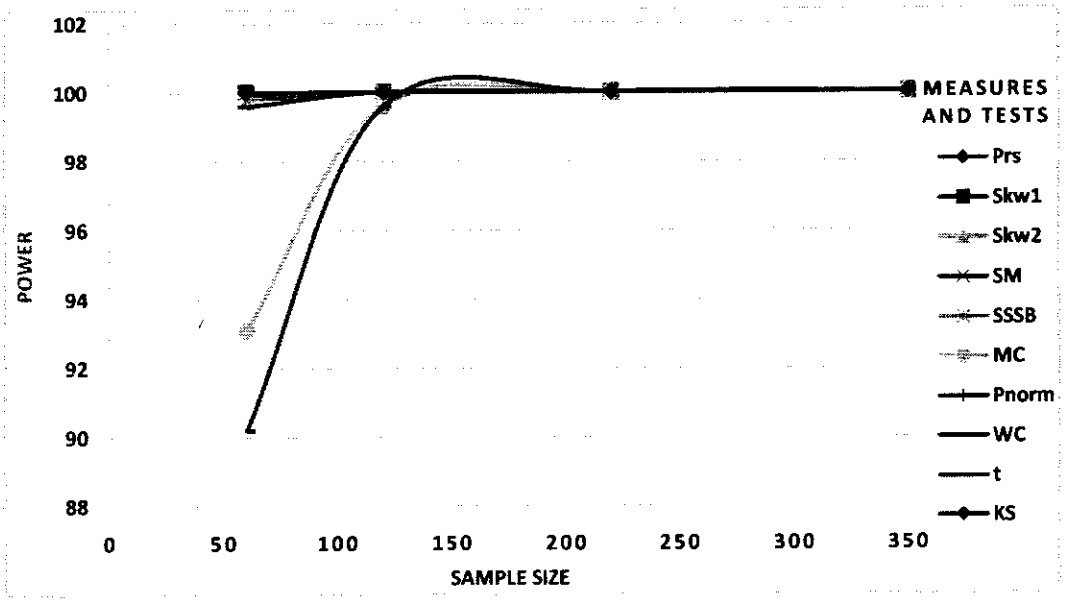
Figure 6.3 describes the power of different measures and tests of skewness using the DGP-I of log-normal distribution with parameters (6, 3). At sample size equal to 60, the measures (i.e. Skw_1, Skw_2, SM, P_{norm}) and tests (i.e. KS, WC) have high power which is 100%, while it is observed that t-test is detected as the least powerful test with 48.1% power.

As the sample size increases from n= 60 to n= 120, then the power of MC increases up to 100%. So again the measures (i.e. Skw_1, Skw_2, SM, MC, P_{norm}) and tests (i.e. KS, WC)

have high power. In this case, the power of t-test also increases to 63.1% but remains the worst test. At sample sizes (i.e. n= 220 and 350), the measures (i.e. Skw_1, Skw_2, SM, MC P_{norm}) and tests (i.e. KS, WC) have maintained the same results and identified as most powerful tests. But t-test with increasing power pattern reaches to 80.6% is still a low power of the test.

Overall, Figure 6.3 concludes that at various sample sizes, it is observed that the measures (i.e. Skw_1, Skw_2, SM, MC, P_{norm}) and tests (i.e. KS, WC) have high power while t-test is identified the least powerful test. Furthermore, for changing the parameters mean and SD that is (2, 1) and (4, 1.5), the result approximately remains the same. It means that in case of log-normal distribution, most measures and tests show that the distribution is skewed with higher frequencies.

Figure 6.4: Power of Chi-Square Distribution with Parameter v= 1



Results of the above Figure 6.4 describe the power of DGP-I using Chi-square distribution with degree of freedom as a parameter (i.e. v= 1). At small sample size n= 60, the measures (i.e. Prs, Skw_1, Skw_2, SM, SSSB, P_{norm}) and tests (i.e. KS, t) have high

power round about 100% while WC test is the least powerful test with 90% power. As sample size increases, all the measures and tests perform well with the highpower of approximately 100%.

Figure 6.5: Power of Chi-Square Distribution 'df' v= 8

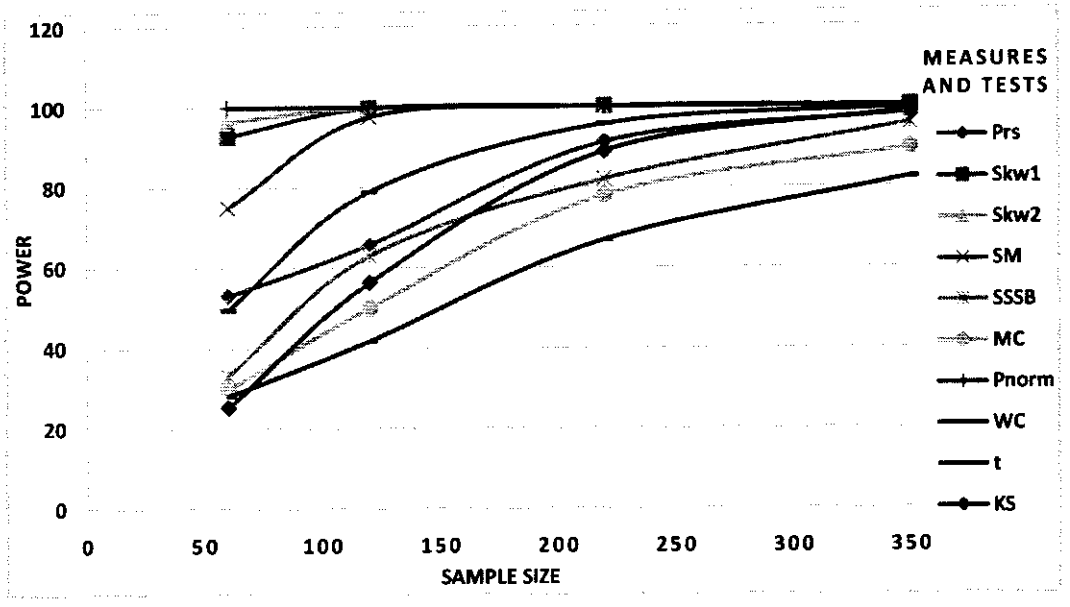


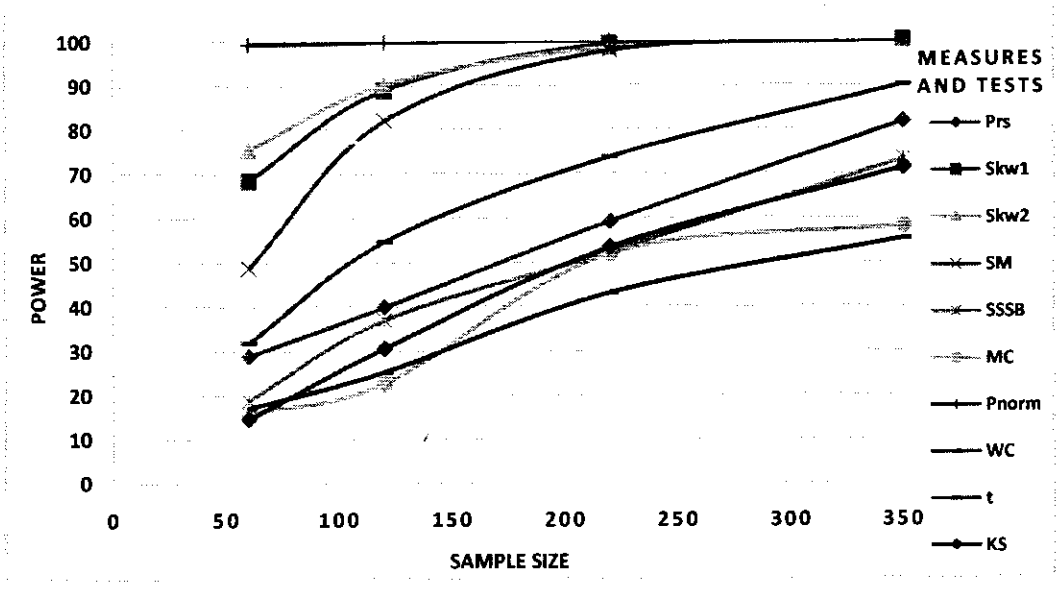
Figure 6.5 shows the power of different measures and tests of skewness using the DGP-I of Chi-square distribution with 'df' v= 8. At small sample size as n= 60, the measure P_{norm} has high power as compared to other measures and tests of skewness while KS test has low power i.e. 25.2%. Also measures Skw_1 and Skw_2 have high powers close to P_{norm}. It is observed that the KS test is detected as the least powerful test while P_{norm} is the most powerful measure among all measures and tests.

As the sample size increase from n= 60 to n= 120, then WC with 41.7% power is detected as the least powerful test; while at the same sample size, all measures and tests have increased power. Now measures Skw_1 and Skw_2 have achieved high power along with P_{norm}. Also, at sample size n= 120 KS test has achieved much better power than

these two measures MC and WC test, because at sample size n= 60, its power was less as compared to these two tests.

As the sample size further increases from n= 120 to 220, the majority of the measures (i.e. SM, P_{norm}, Skw_1, and Skw2) achieve maximum power (i.e. 100%), while WC with increasing pattern again remains the lowest powerful test. At sample size n= 350, the measures (i.e. Prs, SM, P_{norm}, Skw_1, and Skw2) and tests (KS, t-test) have power approximately 100% while WC has achieved maximum power (i.e. 82.2%). But again WC test remains a least powerful test even though its power is increased. In all cases of sample size, the measure P_{norm} has power equal to 100% on this DGP.

Figure 6.6: Power of Chi-Square Distribution with '*df*' v= 16



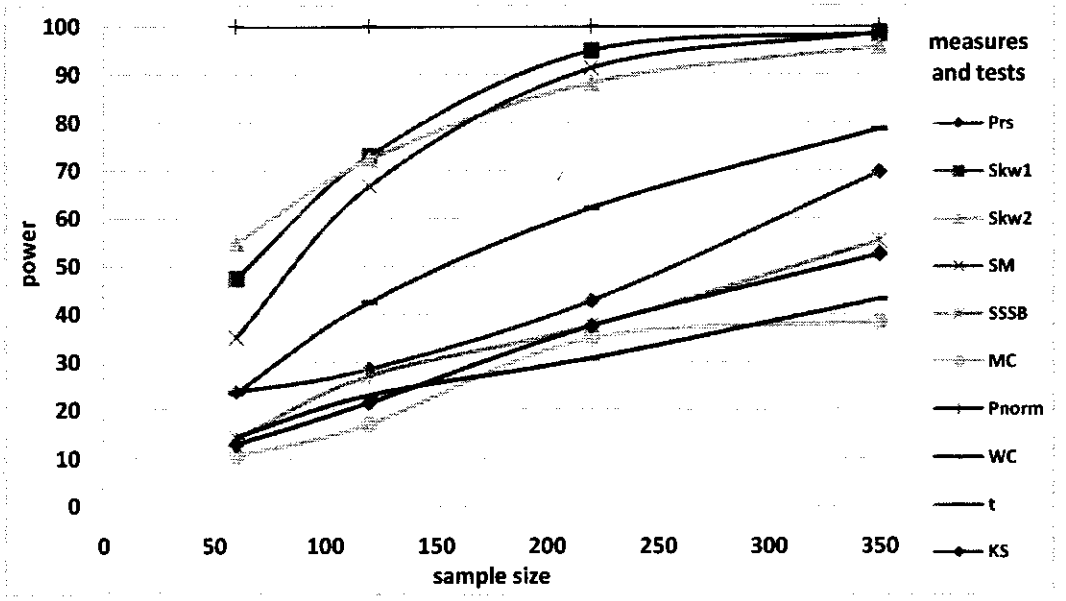
Results of the Figure 6.6 show that for increasing the degree of freedom of Chi-Square distribution in DGP-I, then the power of some measures and tests decreases. When sample size is n= 60, the measures (i.e. SSSB, MC) and tests (i.e. WC, KS) have very low power around 15%, also measure Prs and t-test has power around 30%. Further, it is

observed that the KS test is detected as the least powerful test while P_{norm} is the most powerful measure among all measures and tests.

At sample size n= 120, KS test has achieved much better power than measure MC and WC test, because at sample size n= 60, its power was less as compared to these two tests. Again, measure P_{norm} is detected as the most powerful measure while MC is observed as the least powerful measure. As the sample size further increases from n= 120 to 220, the majority of the measures (i.e. SM, Skw_1, and Skw_2) achieve maximum power nearly 100% while WC with increasing pattern remains the lowest powerful test and both P_{norm} and Skw_1 are most powerful measures. At sample size n= 350, the measures (i.e. SM, Skw_1, and Skw_2) achieve high power equal to 100% while WC has achieved maximum power (i.e. 55.2%) and again this test remains the least powerful test.

Overall, it seems from Figure 6.6 that the measure P_{norm} is most powerful measure while WC is the least powerful test as compared to other measures and tests.

Figure 6.7: Power of Chi-Square Distribution with 'df' v= 24

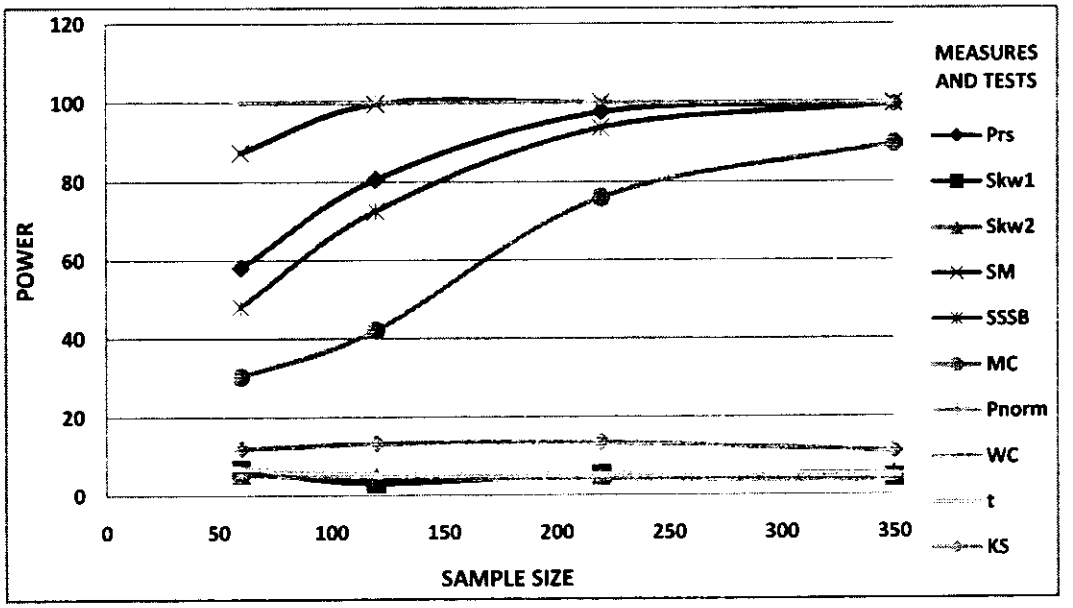


Results of the Figure 6.7 show that for increasing the degree of freedom (i.e. V= 24) of Chi-Square distribution in DGP-I, then the power of some measures and tests decreases as compare to Figure (6.6). At sample size n= 60, the measures (i.e. SSSB, MC) and tests (i.e. WC, KS) have very low power around 13%, also measure Prs and t-test has power around 25%. Also, it is observed that measure MC is detected as the least powerful measure while P_{norm} is the most powerful measure among all measures and tests.

At sample size n= 120, power of most measures and tests increases and again measure P_{norm} is detected as the most powerful measure while MC is observed as the least powerful measure. As the sample size further increases from n= 120 to 220, the majority of the measures (i.e. SM, Skw_1, and Skw_2) achieve maximum power around 90% while WC with increasing pattern remains the lowest powerful test and both P_{norm} is identified as the most powerful measure. At sample size n= 350 the measures (i.e. SM, Skw_1, and Skw_2) achieve high power equal to 99% while MC has achieved maximum power (i.e. 38.3 %) and again this measure remains the least powerful measure.

Overall, Figure 6.7 concludes that the measure P_{norm} is the most powerful measure while MC is the least powerful test as compared to other measures and tests.

Figure 6.8: Power of Beta Distribution with Parameters (a, b)= (2, 15)



Above Figure 6.8 shows the simulation results of DGP-I Beta distribution with parameters (a, b) = (2, 15) for various sample sizes. Clearly looking at the figure that some measures (i.e. Skw_1, Skw_2) and tests (i.e. WC, t, KS) have very small powers as 5% to 10% but the power of measures P_{norm} and SM are so high which is approximately 100% in all cases of sample sizes. At small sample n= 60, it is observed that WC test is detected as the least powerful test while P_{norm} is the most powerful measure among all measures and tests.

As the sample size increases from n= 60 to n= 120, then again WC test is detected as the least powerful test while measures (i.e. SM and P_{norm}) are the most powerful measures. At sample size n= 220, the majority of the measures and tests achieve maximum power (i.e. nearly 100%) while WC test remains the lowest powerful test, and measures (i.e. SM and P_{norm}) have high power as compared to other measures and tests. As the sample size further increases from n= 220 to 350, then measures (i.e. SM, Prs, SSSB and P_{norm}) achieve high power and WC test has low power.

Overall, from Figure 6.8, it is observed that measures SM and P_{norm} have high power while WC test has low power at various sample sizes. Furthermore, the results remain the same as Figure 6.8 by changing the values of parameters (2, 30) and (4, 30) for the same distribution.

Figure 6.9: Power of Beta Distribution with Parameters (a, b)= (4, 15)

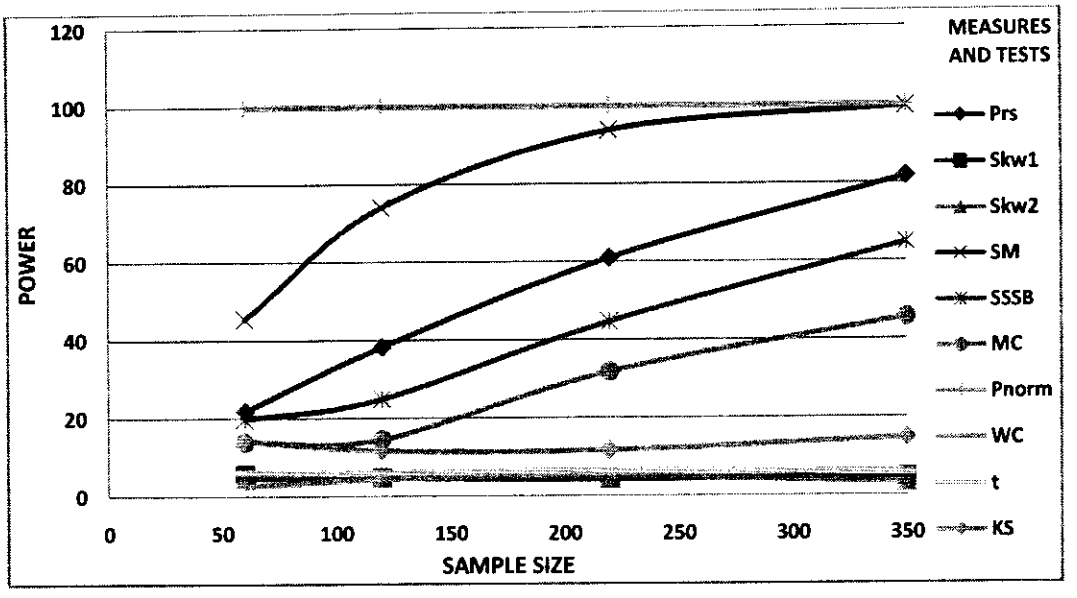


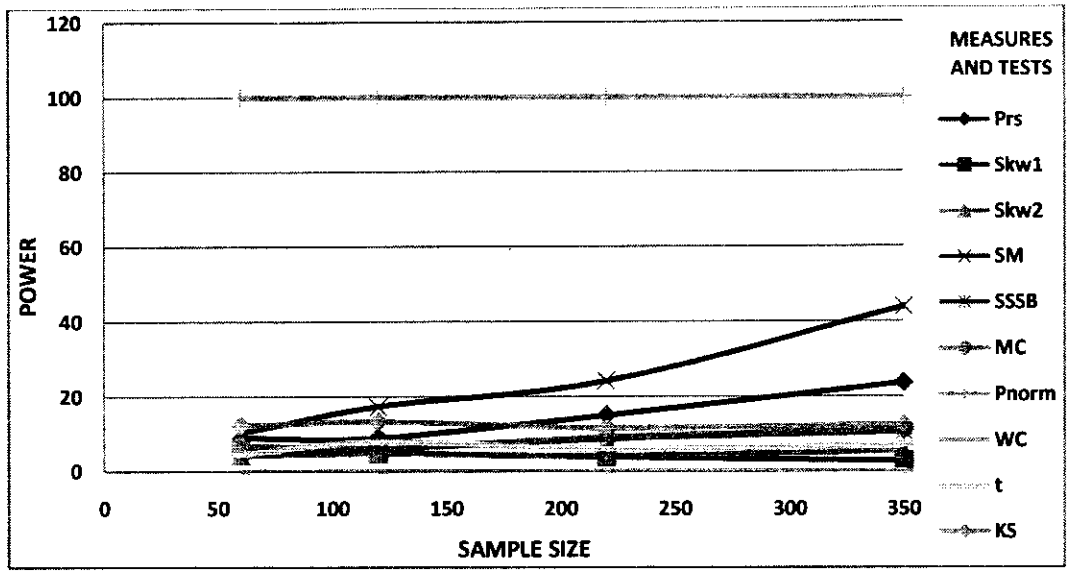
Figure 6.9 describes the simulation result of DGP-I (i.e. Beta distribution with parameters (a, b) = (4, 15)) for various sample sizes. At small sample size (i.e. n= 60), most measures and tests have gained low power in between 5% to 20%, in which WC test has got the least power (i.e. 2.7%), while measure P_{norm} with 100% power is identified as the powerful test.

As increasing the sample size from n= 60 to 120 and 220, the measures (i.e. SM, Prs, MC and SSSB) have increasing power pattern as compared to at n= 60. Again, new measure P_{norm} is observed as a high powerful measure while WC test has low power. Moreover, as the sample size increases from n= 220 to n= 350, Figure 6.9 indicates that measures (i.e. SM, Prs, MC and SSSB) have gained an increase in their power behavior while it is

observed that SM and P_{norm} are more powerful measures and again WC is the least powerful test.

It seems from the Figure 6.9 that power of all the measures increases while all the tests maintain the same power for various sample sizes. Similarly, the simulated power results remain the same approximately as shown in the above Figure 6.9 by changing the parameters (a, b) = (6, 30).

Figure 6.10: Power of Beta Distribution with Parameters (a, b)= (8, 15)



The above Figure 6.10 shows the simulations result that as we increase the parameter 'a' the power of measure P_{norm} remains high as compared to other measures and tests while all other measures and tests have low power between 10% to 40% for various sample sizes. It means that in Figure 6.10, measure P_{norm} is the most powerful measure while WC test is the least powerful test.

Figure 6.11: Power of Beta Distribution with Parameters (a, b)= (6, 15)

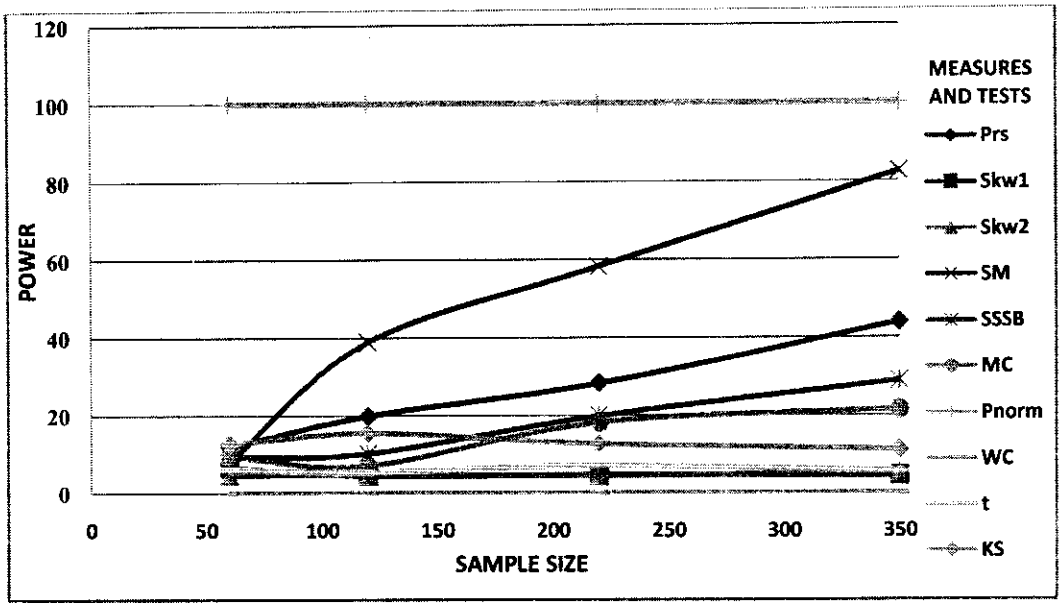


Figure 6.11 describes the simulation results of DGP-I of Beta distribution with parameters (a, b) = (6, 15) for different sample sizes. At small sample size n= 60, all the measures and tests have low powers between 5% and 12.3% except the power of measure P_{norm} which is equal to 100%. It is identified that measure P_{norm} is most powerful measure while WC is the least powerful test. At increasing sample size n= 120 and 220, the power of measures (i.e. SM, MC, SSSB and Prs) have increasing power pattern while other measures and tests maintain the same power as compare to their results at n= 60. At sample size n= 350, the measure Prs has the power 43% and measure SM power is observed as 82%. But again, P_{norm} has high power while WC has low power as compared to all measures and tests of skewness.

Overall, Figure 6.11 concludes that P_{norm} is the most powerful test while WC is the least powerful test. Moreover, Figure 6.11simulated power results remain approximately the same by changing the parameters (a, b) = (4, 30) or (8, 30).

6.3 Chapter Summary

In this chapter, various measures and tests for skewness are compared on the basis of simulation results of size and power. First, to make power comparison logically possible, this study stabilized the size around the nominal size of 5% of all measures and tests by using simulated critical values. Figure 6.1 has concluded that all the measures and tests for skewness including the new measure of skewness P_{norm} have stable sizes. Therefore, these measures and tests for skewness are compared on the basis of power.

It is concluded that when DGP-I of log-normal distribution with parameters mean= 0 and SD= 0.5 is used, and then the measures (i.e. Skw_1, Skw_2 and P_{norm}) have high power as compared to other measures and tests. Similarly, as the parameter values increase in the same DGP, then the power of all the measures and tests of skewness also increases. Figure 6.2 to Figure 6.3 conclude that measures (i.e. Skw_1, Skw_2 and P_{norm}) have high power while WC test has low power.

When DGP-I of Chi-square distribution with parameters v= 1 is used then all the measures and tests of skewness have high power. Similarly, as the parameter values increase (i.e. v= 8) in the same DGP, then the power of the measures (i.e. Skw_1, Skw_2, SM and P_{norm}) also increases. Moreover, if the parameter values increase from v= 1 to 16 and 24, all the measures and tests have low power while the measure P_{norm} has high power approximately 100%.

It is also concluded that when DGP-I of the beta distribution with parameters (a, b)= (2, 15), then the measures Prs, SM and P_{norm} have high power. As parameter value increases from (a, b)= (2, 15) to (a, b)= (4, 15), then only measure P_{norm} has high power as compared to all other measures and tests. Further, as the values of parameter 'a' increase

to 6 or 8, then all the measures and tests have low power but the measure P_{norm} has high power (approximately 100%). Similarly, for increasing the second parameter 'b', the result remains same which has been observed for the change of value corresponding to the parameter 'a'.

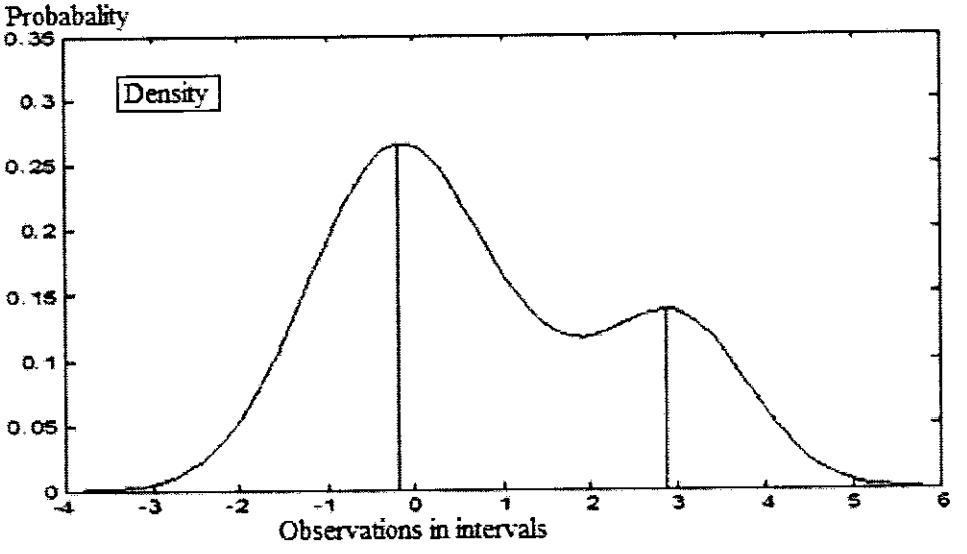
Overall, Figure 6.2 to Figure 6.11results show that the performance of the new measure (i.e. P_{norm}) is obviously very well. Therefore, this measure is best as compare to all other measures and tests of skewness.

CHAPTER 7

SIZE OF BIMODALITY

This chapter has discussed the size of the bimodality. The size of bimodality shows the distance between the two modes or two peaks of a bimodal distribution. This distance is calculated through analytical integrals, i.e. Trapezoidal and Simpson's rules (see details in Section 3.4.2). For this purpose, Table 4.4, the mixture of normal distributions is used with those values of parameters which show bimodality. Here, this chapter checks that on which parameter values the size of bimodality increases or decreases.

Figure 7.1: Modes in a Bimodal Distribution



The above Figure 7.1 describes the two modes or peaks of a bimodal distribution where the distance between these modes is called the size of the bimodality. The size is affected by the values of the three parameters of bimodal distribution which are explained below:

7.1 Size of Bimodality w.r.t Changing Parameters of Bimodal Distribution

This study used mixture of normals with the changing of the values of three parameters to check the variation in size of bimodal distribution.

7.1.1 Changing Mean 'μ_2' in a Mixture of Normals

In this case, mean 'μ_2' is changed while keeping the other parameters (i.e. α, σ_2) constant in DGP-II. The following results of Figure 7.2 to Figure 7.5 show the size of the bimodality.

Figure 7.2: Size of the Bimodality with Various Values of 'μ_2' and Fix α= 0.1, σ_2= 0.7

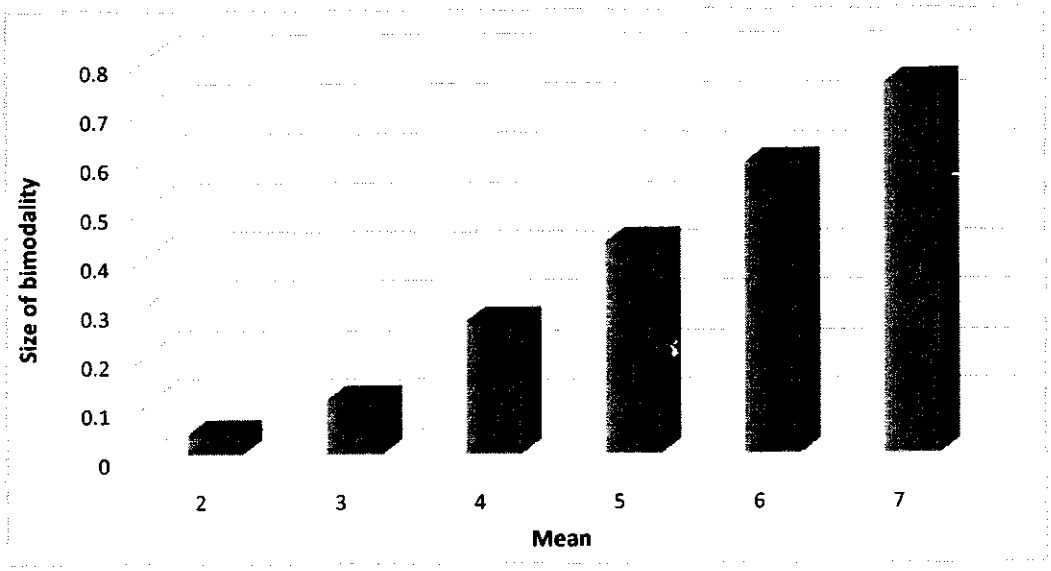
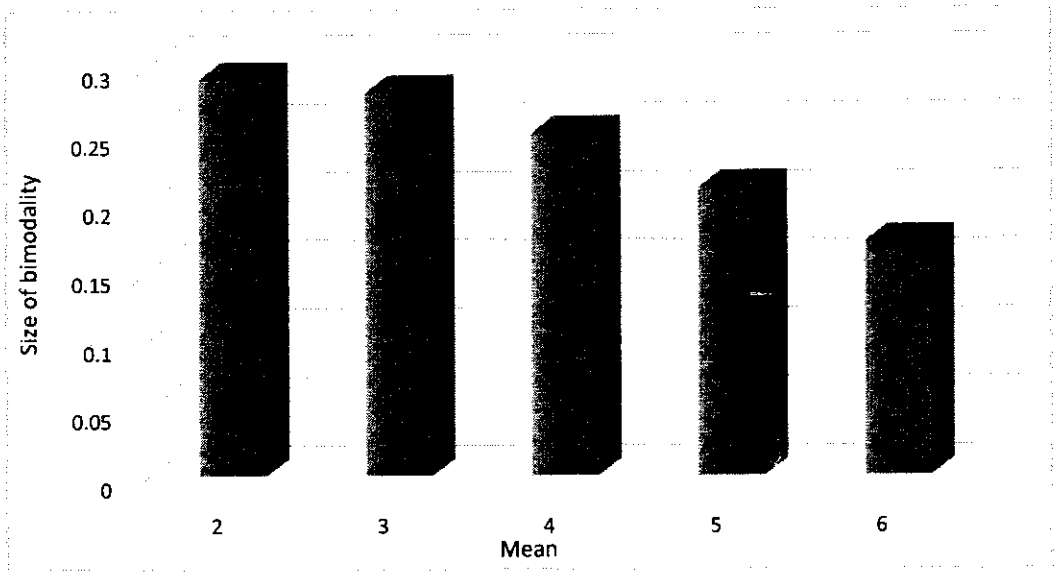


Figure 7.2 describes the size of the bimodality by using Trapezoidal rule for various values of μ_2= (2, 3, ... , 7) and α= 0.1, σ_2 = 0.7. From Figure 7.2, it is observed that at μ_2= 2 the size of the bimodality is 0.04, which is the lowest size. However, as μ_2 value increases to 3, then the size of the bimodality increases to 0.11. Further, as μ_2= 4, then the size of the bimodality moves to 0.27, which is the higher value of the size of

bimodality as compared to the previous two values. At μ_2= 5, the size of the bimodality also increases to 0.43. Moreover, as μ_2 value increase to 6 and 7, then from Figure 7.2, it is observed that the size of the bimodality moves to its highest values (i.e. at μ_2= 6 the value is 0.59 while μ_2= 7 the value is 0.75). Hence, it is observed that as the mean value increases, the size of the bimodality also increases if the values of the other parameters remain the same (i.e. α= 0.1, σ_2 = 0.7) .

Figure 7.3: Size of Bimodality with Different values of 'μ_2' and Constant α= 0.4, σ_2= 0.7



The above Figure 7.3 represents the size of the bimodality by using Trapezoidal rule for the values of parameters α= 0.4, σ_2 = 0.7 and different values of μ_2= (2, 3, ... , 6). From Figure 7.3, it is observed that at μ_2= 2, the size of the bimodality is 0.29, which is the highest size. However, as μ_2 value increases to 3, then the size of the bimodality drops to 0.8. Further, as μ_2 value increases to 4, then the size of the bimodality moves to 0.25. This is the lower value of the size of bimodality as compared to the previous two values. Moreover, as μ_2 value increases to 5 and 6, then from Figure 7.3, it is observed

that the size of the bimodality drops to its lowest values (i.e. at μ_2= 5 the value is 0.21 while μ_2= 6 the value is 0.15). Hence, it is observed that as the mean value increases, the size of the bimodality decreases if the values of the other parameters remain the same (i.e. α= 0.4, σ_2 = 0.7).

Figure 7.4: Size of Bimodality with Different values of 'μ_2' and Fix α= 0.4, σ_2= 0.8

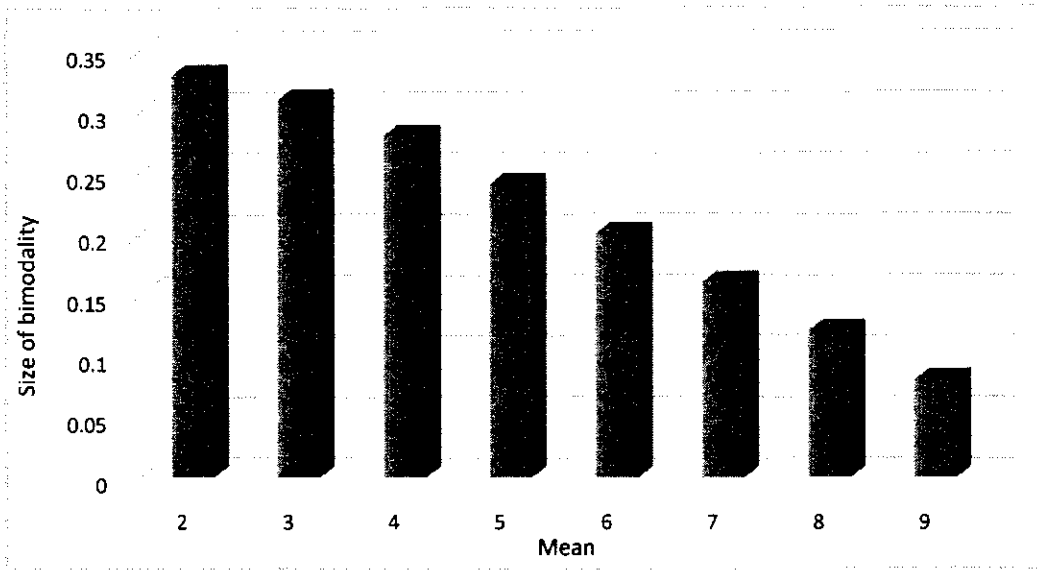
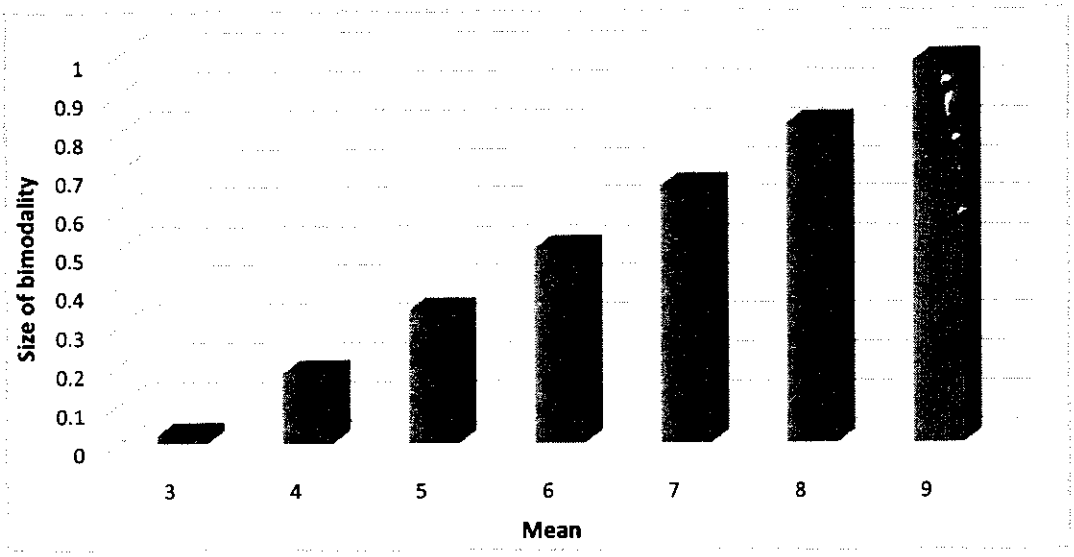


Figure 7.4 describes the size of the bimodality by using various values of μ_2= (2, 3, . . . , 9) and α= 0.4, σ_2 = 0.8 It is detected that at μ_2= 2, the size of the bimodality is 0.33, which is the highest size. As μ_2 value increases to 3, then size of the bimodality decrease to 0.31. However, as μ_2= 4, then the size of the bimodality drops to 0.28. At μ_2= 5, the size of the bimodality also decreases to 0.24. Further, as μ_2 value increases to 6 and 7, then from Figure 7.4, it is observed that the size of the bimodality drops (i.e. at μ_2= 6 the value is 0.2 while μ_2= 7 the value is 0.16). Similarly, as μ_2 value increases to 8 and 9, then it is identified that the size of the bimodality moves to its lowest values (i.e. at μ_2= 8 the value is 0.12 while μ_2= 6 the value is 0.08). Hence, it is observed that as the mean

value increases, the size of the bimodality decreases if the values of the other parameters remain the same (i.e. α= 0.4, σ_2 = 0.8).

Figure 7.5: Size of Bimodality with Different Values of 'μ_2' and Fix α= 0.3, σ_2= 0.9



The above Figure 7.5 shows the size of the bimodality for different values of the parameter μ_2= (3, 4, ... ,9) and constant parameters α= 0.3, σ_2 = 0.9. It is observed that at μ_2= 3, the size of the bimodality is 0.02, which is the lowest size. However, as μ_2 value increases to 4, then the size of the bimodality increases to 0.18. Further, as μ_2 value increases to 5, then the size of the bimodality moves to 0.34. This is the higher value of the size of bimodality as compared to the previous two values. At μ_2= 6, the size of the bimodality increases to 0.5 while μ_2= 7 the size of the bimodality reaches to 0.66. Moreover, as μ_2 value increases to 8 and 9, then from Figure 7.5, it is observed that the size of the bimodality moves to its highest values (i.e. at μ_2= 8 the value is 0.82 while μ_2= 9 the value is 0.98). Hence, it is observed that as the mean value increases, the size of the bimodality also increases if the values of the other parameters remain the same (i.e. α= 0.3, σ_2 = 0.9).

It implies that by using Trapezoidal rule for finding the size of bimodality, as a parameter μ_2 increases, the size of the bimodality also increases except at $\alpha = 0.4$.

7.1.2 Changing Mixing Proportion Alpha 'α' in a Mixture of Normals

In this situation, the mixing proportion 'α' is changed and keeping the other parameters μ_2, σ_2 fixed to check the size of the mixture of normal, i.e. bimodal distribution.

Figure 7.6: Size of Bimodality with Different Values of 'α' and Fix $\mu_2 = 1$, $\sigma_2 = 0.3$

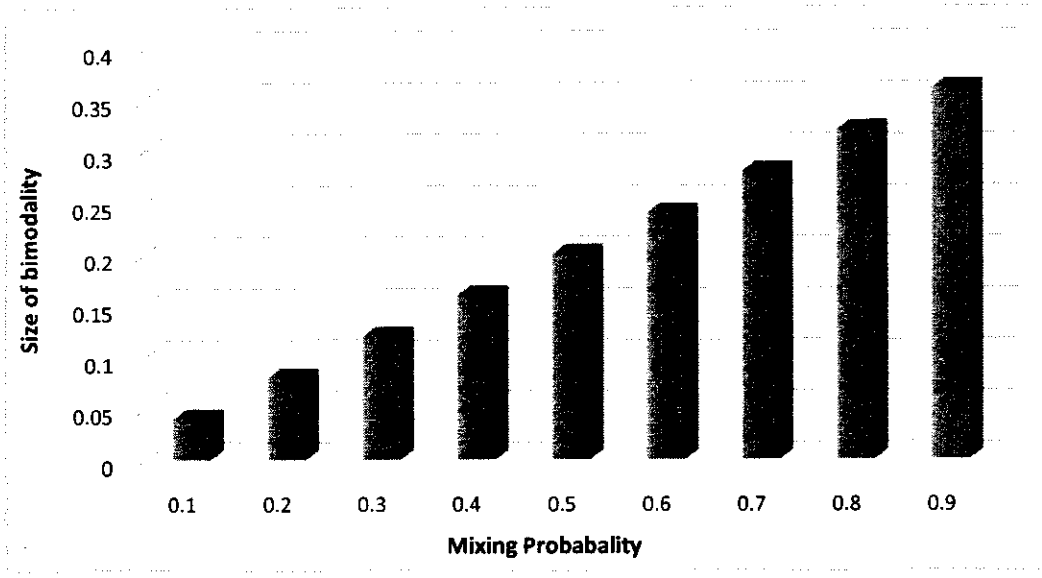
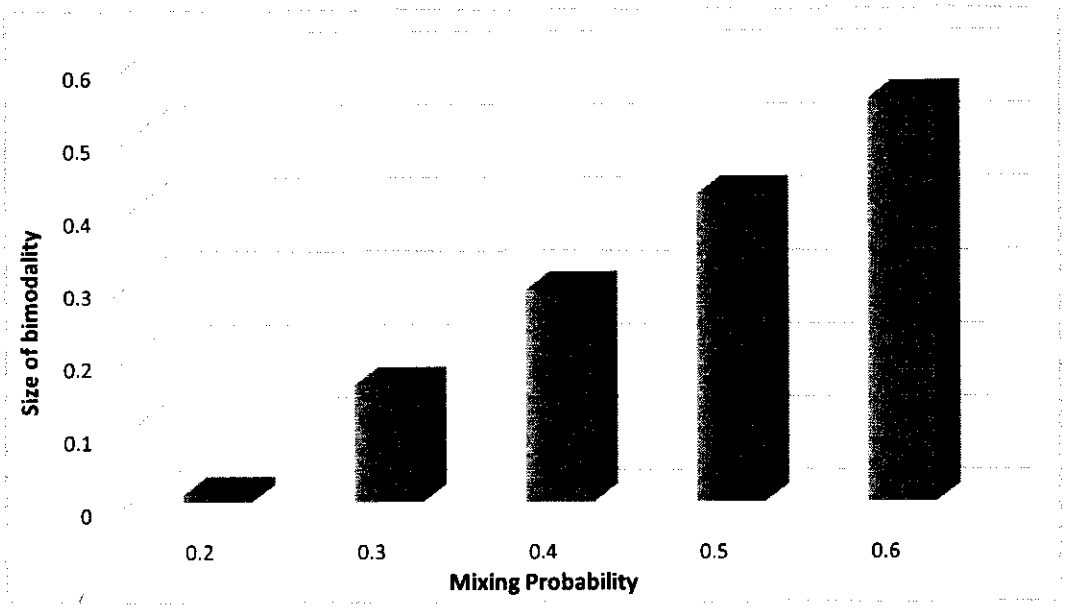


Figure 7.6 shows the size of bimodality where changing the parameter α= (0.1, 0.2, 0.3, . . . , 0.9) and fix the other parameters $\mu_2 = 1$, $\sigma_2 = 0.3$, it is identified that at $\alpha = 0.1$, the size of the bimodality is 0.04, which is the lowest size. However, as $\alpha = 0.2$, then the size of the bimodality increases to 0.08. Further, as α value increases to 0.3, then the size of the bimodality moves to 0.12. This is the higher value of the size of bimodality as compared to the previous two values. At $\alpha = 4$, the size of the bimodality increases to 0.16 while $\alpha = 0.5$ the size of the bimodality reaches to 0.2. Moreover, as α value increases to

0.6 and 0.7, then from Figure 7.6, it is observed that the size of the bimodality moves to its highest values (i.e. at α= 0.8 the value is 0.32 while α= 0.9 the value is 0.36). Hence, it is observed that as the mixing probability increases the size of the bimodality also increases if the values of the other parameters remain the same (i.e. μ_2= 1, σ_2= 0.3). Similarly, as μ_2= 0.1, σ_2= 0.2 and α= (0.1, 0.2, 0.3, ... , 0.6), the result remains approximately same as the size against 'α' is shown in Figure 7.6. It is concluded that when mixing proportion 'α' increases, the size of bimodality also increases.

Figure 7.7: Size of Bimodality with Different Values of 'α' and fix μ_2= 3, σ_2=0.7



The above Figure 7.7 represents the size of bimodality where changing the parameter α= (0.2, 0.3,. . . , 0.6) by fixing the other parameters μ_2= 3, σ_2= 0.7, it is observed that at α= 0.2, the size of the bimodality is 0.01 which is the lowest size. However, as α value increases to 0.3, then the size of the bimodality increases to 0.16. Further, as α= 0.4, then the size of the bimodality moves to 0.29, which is the higher value of the size of bimodality as compared to the previous two values. Moreover, as α value increases to 0.5

and 0.6, then from Figure 7.7, it is observed that the size of the bimodality moves to its highest values (i.e. at α= 0.5 the value is 0.42 while α= 0.6 the value is 0.55). Hence, it is identified that as the mixing probability increases, the size of the bimodality also increases if the values of the other parameters remain the same (i.e. μ_2= 3, σ_2= 0.7).

Figure 7.8: Size of Bimodality with Different Values of 'α' and fix μ_2= 6, σ_2=0.8

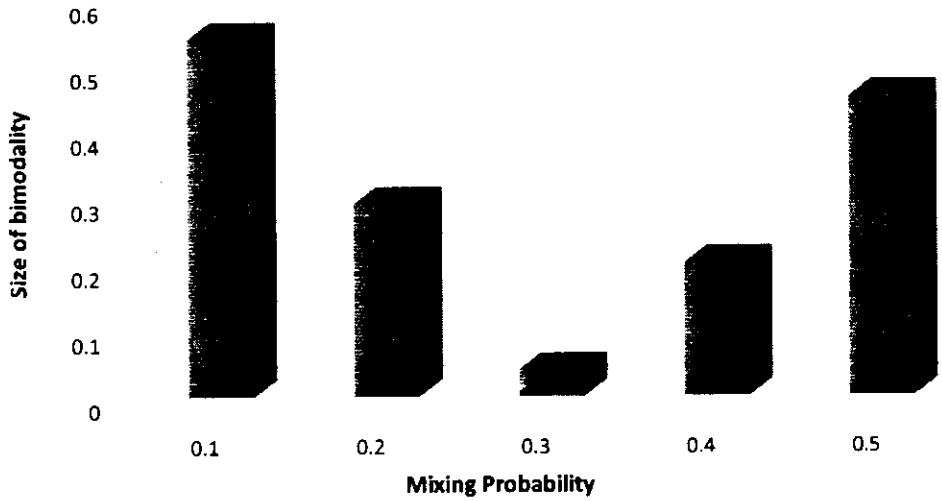
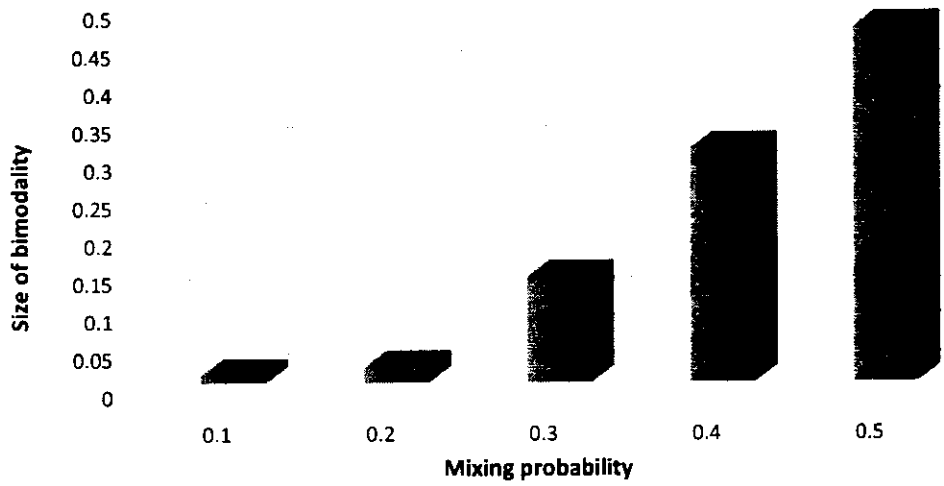


Figure 7.8 describes the size of the bimodality where changing the parameter α= (0.1, 0.2, 0.3, ... , 0.5) and fix the other parameters μ_2= 6, σ_2= 0.8, it is identified that at α= 0.1, the size of the bimodality is 0.54 which is the highest size. However, as α value increases to 0.2, then the size of the bimodality decreases to 0.29. Further, as α value increases to 0.3, then the size of the bimodality moves to 0.04. This is the lower value of the size of bimodality as compared to the previous two values. Moreover, as α value increases to 0.4 and 0.5, then from Figure 7.8, it is observed that the size of the bimodality again increases (i.e. at α= 0.4 the value is 0.2 while α= 0.5 the value is 0.45). Hence, it is observed that as the mixing probability increases, the size of the bimodality

fluctuate, (i.e. first decreases then increases) if the values of the other parameters remains the same (i.e. μ_2= 6, σ_2= 0.8).

Figure 7.9: Size of Bimodality with Different Values of 'α' and fixed μ_2= 4, σ_2= 0.9

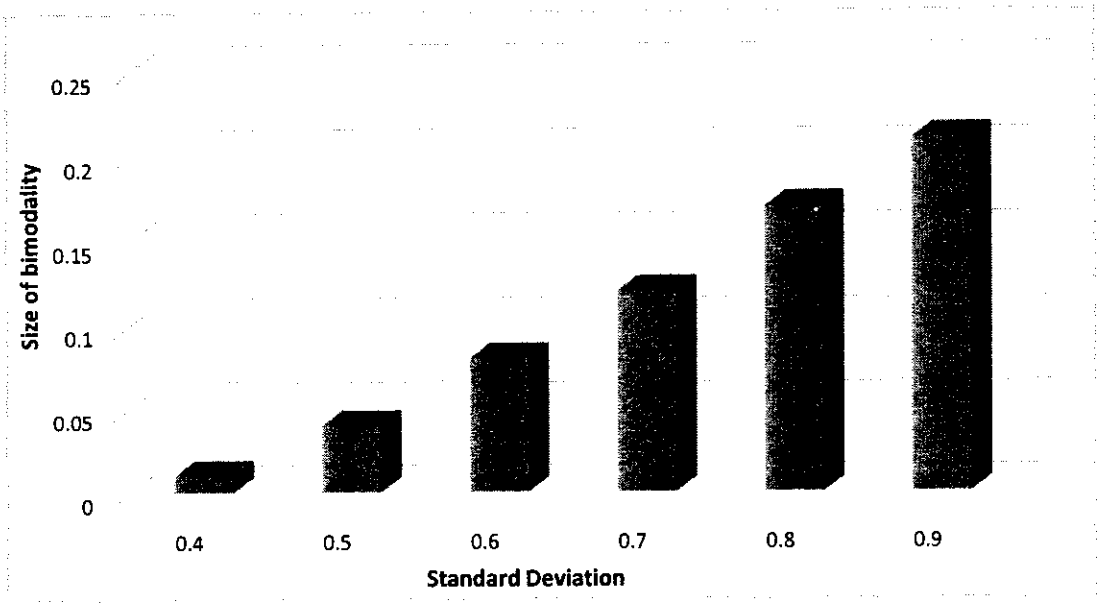


The above Figure 7.9 shows the size of bimodality of the mixture of normal where varying the parameter α= (0.1, 0.2, 0.3, 0.4, 0.5) while fixing the other parameters μ_2= 0.4, σ_2= 0.9, it is observed that at mixing proportion α= 0.1, the size of bimodality is 0.01. At α= 0.2, the size slightly increases to 0.02 while as α= 0.3, then the size of the bimodality moves to 0.14. Moreover, as α value increases to 0.4 and 0.5, then from Figure 7.9, it is observed that the size of the bimodality again increases (i.e. at α= 0.4 the value is 0.31 while α= 0.5 the value is 0.47). Hence, it is identified that as the mixing probability increases, the size of the bimodality also increases if the values of the other parameters remain the same (i.e. μ_2= 4, σ_2= 0.9).

7.1.3 Changing standard deviation 'σ_2' in a mixture of normals

In the third case, changing the standard deviation 'σ_2' and keep the other parameters μ_2 and 'α' fixed to judge the size of DGP-II i.e. bimodal distribution.

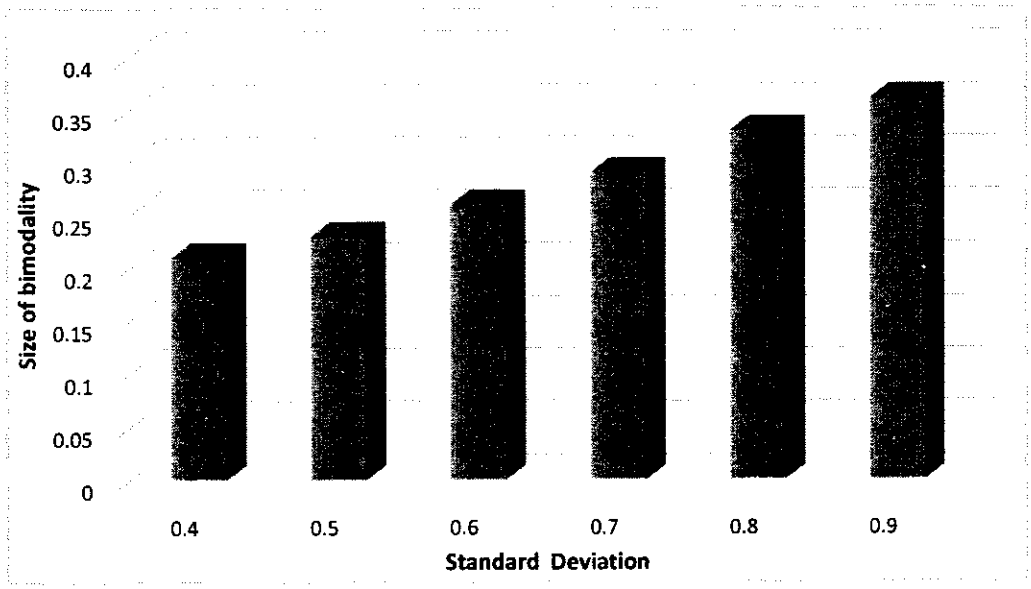
Figure 7.10: Size of Bimodality with Different Values of 'σ_2' and fix μ_2= 2, α= 0.2



The above Figure 7.10 shows the size of bimodality where changing the parameter σ_2= (0.4, 0.5, ... , 0.9) and fixing the other parameters μ_2= 2, α= 0.2, it is identified that at σ_2= 0.4, the size of the bimodality is 0.01 which is the lowest size. However, as σ_2 value increases to 0.5, then the size of the bimodality increases to 0.04. Further, as σ_2 value increases to 0.6 then the size of the bimodality moves to 0.08. This is the higher value of the size of bimodality as compared to the previous two values. At σ_2= 0.7, the size of the bimodality increases to 0.12. Moreover, as σ_2 value increases to 0.8 and 0.9, then from Figure 7.10, it is detected that the size of the bimodality moves to its highest values (i.e. at σ_2= 0.8 the value is 0.17 while σ_2= 0.9 the value is 0.21). Hence, it is observed that as

the standard deviation value increases, the size of the bimodality also increases with little margin if the values of the other parameters remain the same (i.e. α= 0.2, μ_2 = 2).

Figure 7.11: Size of Bimodality with Different Values of 'σ_2' and fix μ_2= 2, α= 0.4



The above Figure 7.11 describes the size of bimodality where changing the parameter σ_2= (0.4, 0.5, ... , 0.9) and fixing the other parameters μ_2= 2, α= 0.4, it is observed that at σ_2= 0.4, the size of the bimodality is 0.21 which is the lowest size. However, as σ_2= 0.5, then the size of the bimodality increases to 0.23. Further, as σ_2= 0.6, then the size of the bimodality moves to 0.26 which is the higher value of the size of bimodality as compared to the previous two values. At σ_2= 0.7, the size of the bimodality increases to 0.29. Moreover, as σ_2 value increases to 0.8 and 0.9, then from Figure 7.11, it is observed that the size of the bimodality moves to its highest values (i.e. at σ_2= 0.8, the value is 0.33 while σ_2= 0.9, the value is 0.36). Hence, it is observed that as the standard deviation value increases, the size of the bimodality also increases gradually if the values of the other parameters remain the same (i.e. α= 0.4, μ_2 = 2).

Similarly, for other combination of parameter values i.e. (μ_2, α)= {(1, 0.1), (2, 0.1), (1, 0.3), (3, 0.3), (3, 0.4), (1, 0.5), (1, 0.6) etc.} and for various considerable values of σ_2, the size of bimodality changes slightly.

7.4 Chapter Summary

This chapter discussed the size of bimodality which shows the distance between the two modes or two peaks of a bimodal distribution. For this purpose, Trapezoidal or Simpson's rules are used on the mixture of normal with the parameter values from Table 4.4. The size of bimodality depends upon the three parameters $(\mu_2, \alpha, \sigma_2)$ in this mixture. From the result of Section 7.1.1, it is concluded that as by increasing mean 'μ_2' while keeping (α, σ_2) fixed, then the size of the bimodality increases and it can be easily visible that the distribution is bimodal. But, only at α= 0.4, in this case, it is identified that the size is in decreasing order. On the other hand, changing the mixing proportion 'α' and keeping constant the other parameters μ_2, σ_2 in Section 7.1.2, it is concluded that for increasing 'α', the size of bimodality also increases. If changing the parameter standard deviation σ_2 and keeping fixed the other parameters in Section 7.1.3, then the size remains slightly changed. It implies that for different values of σ_2, the size of bimodality is affected fractionally.

CHAPTER 8

CONSTRUCTION OF BIMODAL BOXPLOT AND DETECTION OF OUTLIERS

In the case of bimodal distribution, there are two peaks or modes of any distribution. Before building bimodal boxplot, it is necessary to find out the maximum separation of the two peaks or joining point of the two distributions called cutoff point. This chapter discusses the mathematical procedure to calculate the cutoff point in bimodal distribution, the technique to detect outliers and the construction of newly introduced bimodal boxplot.

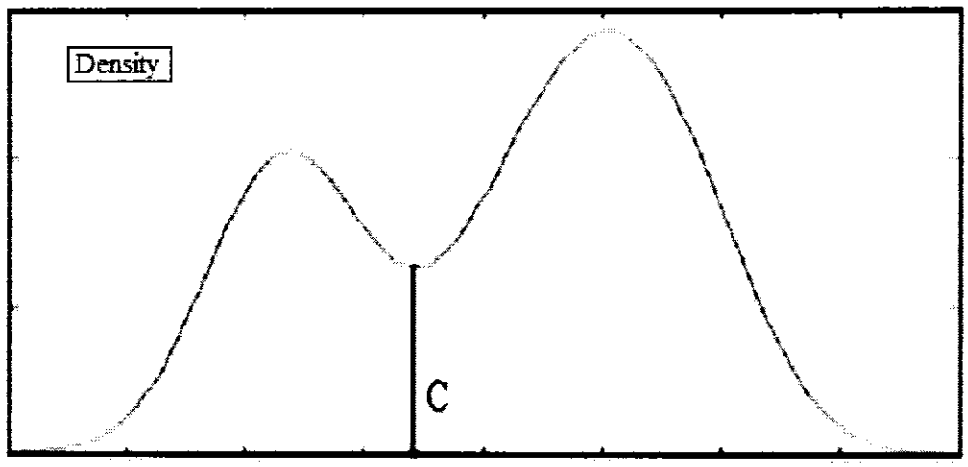
8.1 Procedure of Cutoff Point in Bimodal Distribution

The cutoff is the joining point of two densities or mixing the two tails of the distribution. In other words, it occurs at the maximum separation of a bimodal distribution. For finding the cutoff point, there are several formulas in the literature for the generated data. But the researchers fail to apply and get appropriate results in case of real data.

This study formulated some conditions from the study of [4]Fluss et al. (2005) on the basis of median and IQR (for real case) rather than mean and SD (for DGP case). Because it is difficult to find the mixing probability in real bimodal data series. Formally, the data are split in percentiles (12.5, 25, 37.5, ... , 87.5) in eight parts. Mathematical interpretation of finding a cutoff point is as follows:

[4] See for detail Fluss, Faraggi and Reiser (2005) computational formula of C* for generated data used in their study

Figure 8.1: Cutoff Point in the Bimodal Distribution



Note: The cutoff point is denoted by 'C'= (C_1 or C_2)

$$C_1 = \frac{(Q_1.IQR_R^2 - Q_3.IQR_L^2) - IQR_R.IQR_L\sqrt{(Q_3-Q_1)^2+(IQR_R^2 - IQR_L^2).Log(IQR_R^2/IQR_L^2)}}{IQR_R^2 - IQR_L^2}$$

Where IQR= Inter Quartile Range of whole data series= $Q_3 - Q_1$

IQR_R= Inter Quartile Range on the right side of a Cutoff point = $Q_{3R} - Q_{1R}$

IQR_L= Inter Quartile Range on the left side of a Cutoff point = $Q_{3L} - Q_{1L}$

Also, the second procedure can be used for finding a cutoff point on the basis of quartile deviation.

$$C_2 = \frac{(Q_1.QD_R^2 - Q_3.QD_L^2) - QD_R.QD_L\sqrt{(Q_3-Q_1)^2+(QD_R^2 - QD_L^2).Log(QD_R^2/QD_L^2)}}{QD_R^2 - QD_L^2}$$

Quartile Deviation (QD) of whole data series = $\frac{(Q_1+Q_3)}{2}$

Quartile Deviation on the Right side of Cutoff point (QD_R)= $\frac{(Q_{1R}+Q_{3R})}{2}$

Quartile Deviation on left side of Cutoff point (QD_L)= $\frac{(Q_{1L}+Q_{3L})}{2}$

$C_3 = 0.08 \times C_1$ and $C_4 = 0.125 \times C_2$

The cutoff point 'C' is estimated from the choice of any one of four cases that is C= {C_1, C_2, [5]C_3, [6]C_4} with the comparison of the bimodal graph (ignoring the negative sign and setting with decimal points). After finding cutoff point, it is easy to build boxplot for a bimodal distribution.

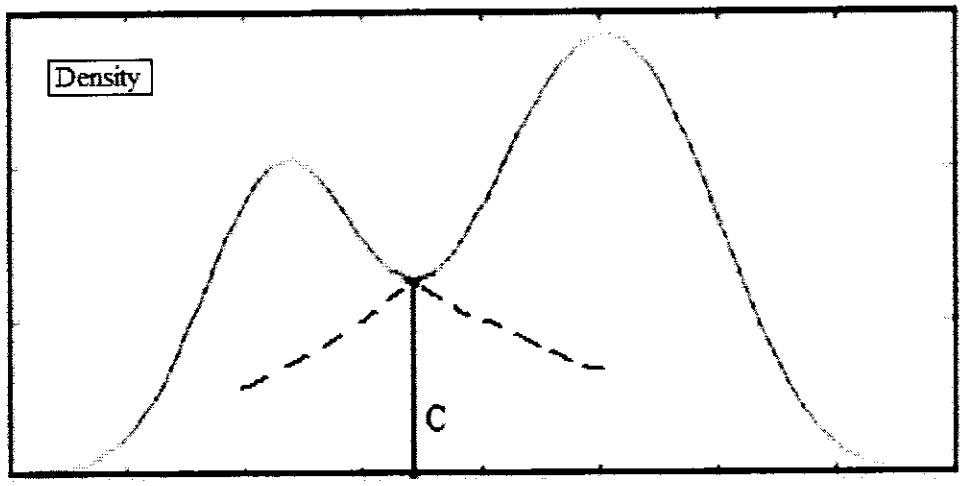
8.2 Detection of Outliers in Bimodal Distribution on the Basis of Cutoff Point

In the literature, a lot of techniques for detection of outliers are found but none can specify the mixing outliers in bimodality. To detect these outliers on either side of the cutoff point, the current study introduced a method for detecting outliers in a bimodal distribution. However, the tails of two densities in bimodal distribution mixed together on cutoff point 'C'. In this area (around 'C') the outliers lie at both densities called outliers zone (OZ). The procedure is mentioned as, first of all, to find percentile rank of each observation. Now checking the percentile rank of 'C' and consider the half quartile (12.5% area) as 'OZ' in which 'C' lies. Also, considering the area around cutoff point i.e. left side (L_S) and right side (R_S),

$$(L_L, U_L) = [M_L - 1.5 \times IQM_L \ , \ M_R + 1.5 \times IQM_R]$$

[5] Multiplication with 8% means to divided the data in eight equal parts
[6] Multiplication with 12.5% means because each part is equal 12.5%

Figure 8.2: Outliers in a Bimodal Distribution



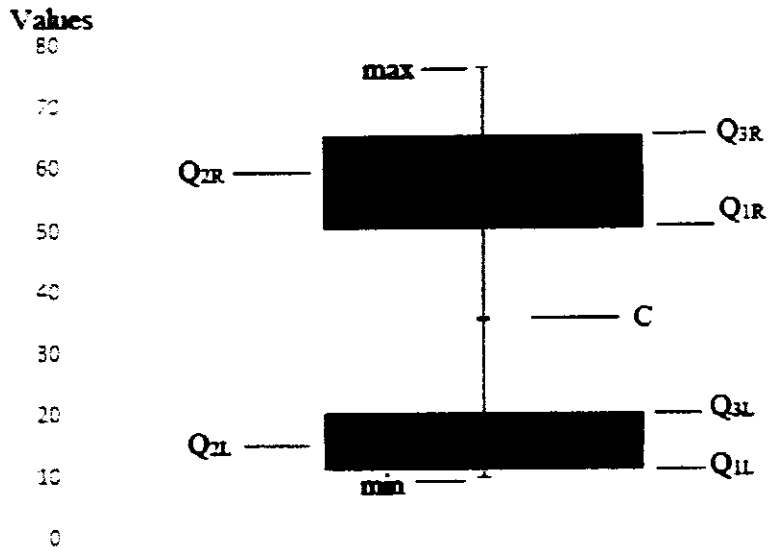
L_S= Number of '$L_L < X_i < C$' and R_S= Number of '$U_L > X_i > C$'

Figure 7.2 shows the tails around cutoff point 'C' where outliers lie. Consider 'X_i' is the set of observations, L_L is lower limit of half quartile and U_L is the upper limit of half quartile. Outlier Zone is any pair of (12.5, 25, 37.5, 50, 62.5, 75 and 87.5) that means half quartile selection depends upon the cutoff point. Outliers set is equal to the combination of L_S and R_S or it can be written as outliers= $\{L_S, R_S\}$. In this way, the outliers are calculated in a bimodal distribution.

8.3 Construction of Bimodal Boxplot

After measuring bimodality and their cutoff point, next goal is to build bimodal boxplot. First, consider the main point of bimodal boxplot which is a cutoff point 'C'. Then build bimodal boxplot on Excel sheet as calculating five-point summary statistics on each side of cutoff point separately. Also, we got two 'five-point summaries' with two medians that separate the median of each part. In the desired boxplot, cutoff point and quartiles are shown clearly.

Figure 8.3: Bimodal Boxplot



The same technique has been followed in this study for boxplot of bimodality as given below:

i. First, to find the main point of bimodal boxplot, i.e. cutoff point 'C'. This is discussed above in Section 8.1.

ii. To calculate summary statistics on each side of cutoff point separately, which have two medians (i.e. separate median of each part).

iii. Then, construct the desired bimodal boxplot on Excel sheet which shows cutoff point and quartiles easily.

8.4 Examples of Various Bimodal Distributions

8.4.1 Bimodal Distributions with their Boxplots

Figures 8.4 to Figure 8.8 contain the bimodal distribution of three countries Pakistan, UK, and India against quarterly data from 1981-I to 2013-IV of various data series with their relevant boxplots.

Figure 8.4: Bimodal Distribution and Bimodal Boxplot of Pakistan Exchange Rate with C= 57.69

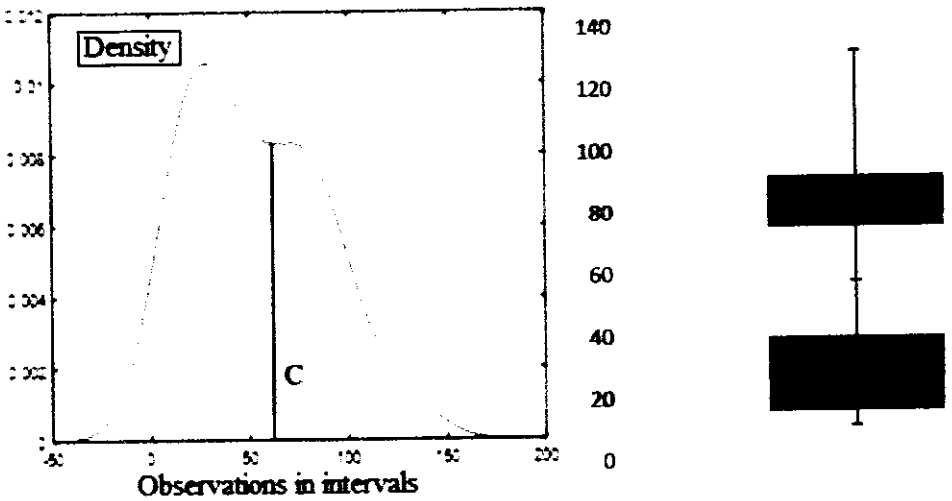
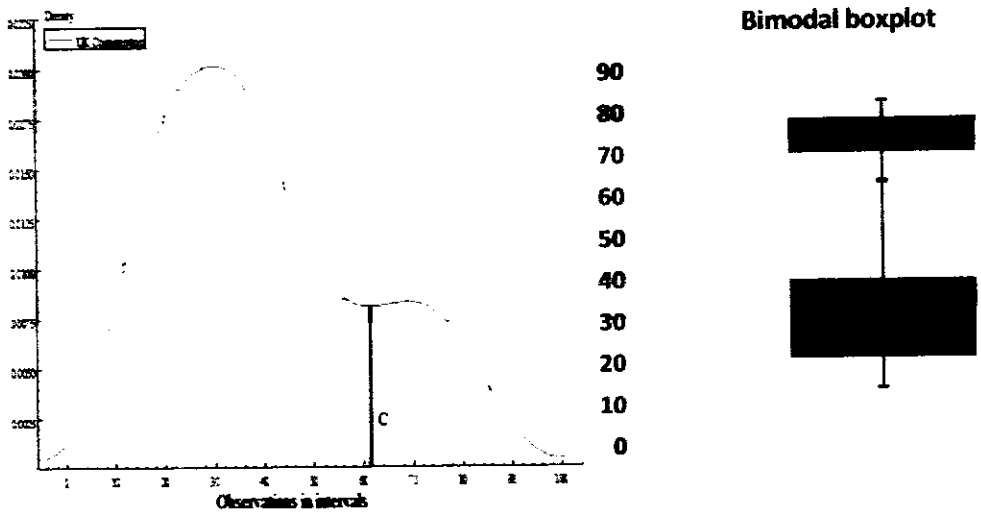


Figure 8.4 shows the bimodal distribution of Pakistan exchange rate. In this distribution, the cutoff point is C= 57.69, and on the right side, its bimodal boxplot clearly displays the summary statistics of each part with a cutoff point.

According to the given bimodal boxplot, the lower part has the summary statistics as, the first quartile of the lower part is 15.86 (i.e. Q_{1L}= 15.86), the median of the lower part is Q_{2L}= 26.3, and the third quartile of the lower part Q_{3L}= 39.74. Similarly, the upper part of the bimodal boxplot has the first quartile of upper part Q_{1U}= 75.13, the median of the upper part is Q_{2U}= 83.9, and the third quartile of upper part Q_{3U}= 91.67.

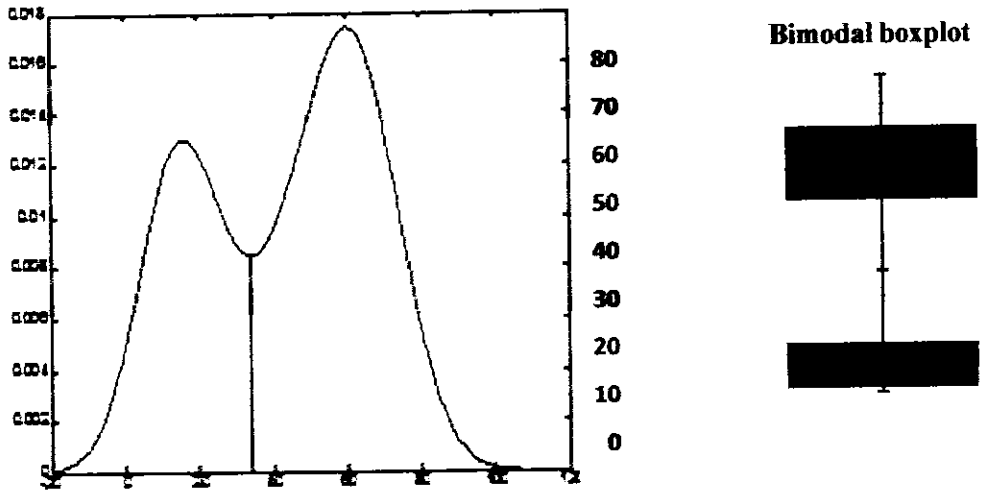
Figure 8.5: Bimodality and Bimodal Boxplot of UK Consumption



The above Figure 8.5 illustrates the bimodal distribution of UK consumption. In this distribution, the cutoff point is C= 63.63 where the two distributions have mixed, and on the right side, a bimodal boxplot is mentioned of this bimodal distribution. Clearly, it describes the summary statistics of each part with a cutoff point.

A bimodal boxplot is the combination of two distribution and the lower part has the summary statistics as, the first quartile of the lower part is Q_{1L}= 20, the median of the lower part is Q_{2L}= 33, and third quartile of the lower part Q_{3L}= 39. Similarly, the upper part of the bimodal boxplot has the first quartile of upper part Q_{1U}= 70, the median of the upper part is Q_{2U}= 73, and the third quartile of upper part Q_{3U}= 78.

Figure 8.6: Bimodality and Bimodal Boxplot of India Exchange Rate



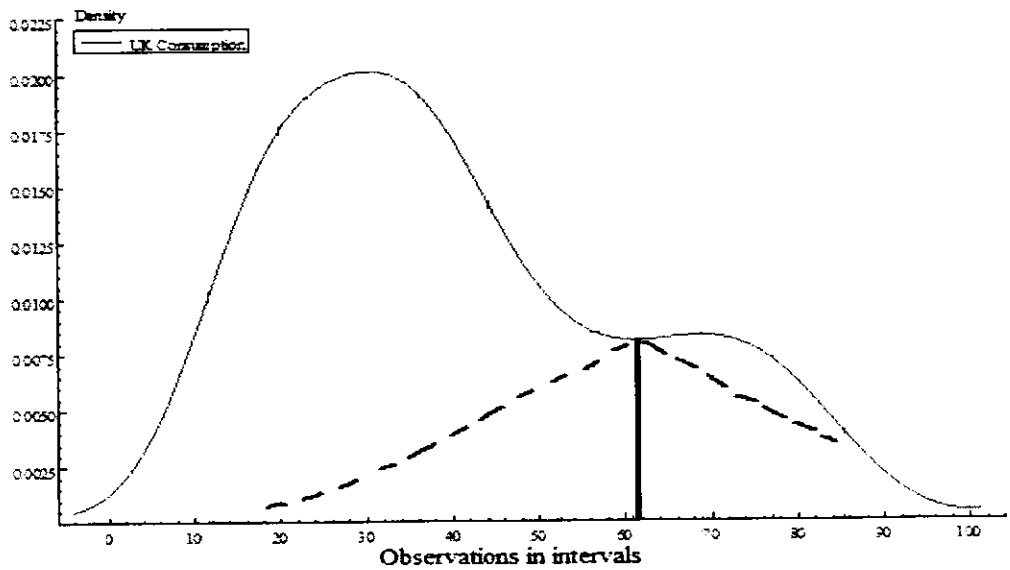
The above Figure 8.6 describes the bimodal distribution of India exchange rate. In this distribution, the cutoff point is C= 35.24 and on the right side, its bimodal boxplot clearly shows the summary statistics of each part with a cutoff point.

According to bimodal boxplot, the lower part has the summary statistics as, the first quartile of the lower part is Q_{1L}= 10, the median of the lower part is Q_{2L}= 14, and the third quartile of the lower part is Q_{3L}= 20. In the same way, the upper part of the bimodal boxplot has the first quartile of upper part Q_{1U}= 50, the median of the upper part is Q_{2U}= 59, and the third quartile of upper part Q_{3U}= 65.

8.4.2 Detection of Outliers in Binomial Distributions

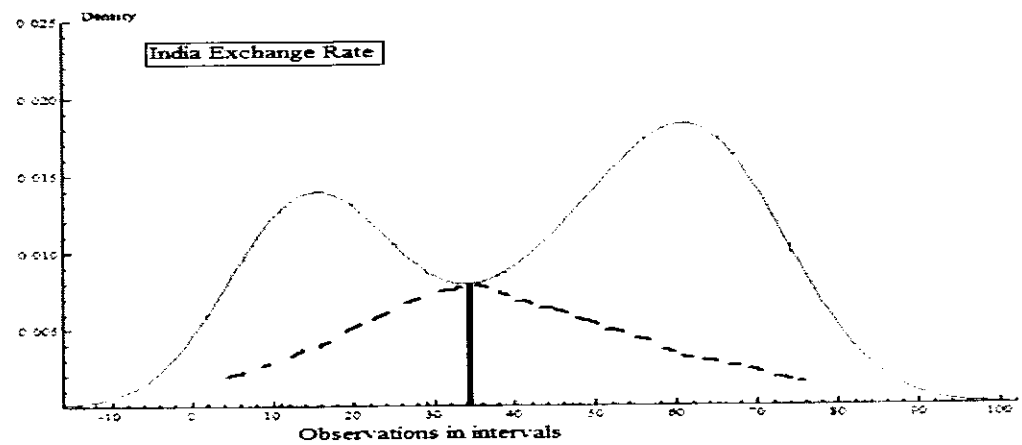
This study also detected outliers in a mixture of two distributions of both sides around cutoff point 'C'. First, percentiles are found (i.e. 12.5, 25 and 37.5) of the left side as 'from minimum to C' and due to a mixture of the two distributions then percentiles of the right side as 'from minimum to maximum' are (i.e. 62.5, 75 and 87.5). Here, the data of the above Section 8.4.1 are used for detection of outliers.

Figure 8.7: Outlier in a Bimodal Distribution of UK Consumption



For UK consumption data series in Figure 8.7, the cutoff point C= 63.63 lies on 81 percent rank. Here, outliers Zone 'OZ' is the pair of (75^{th} to 87.5^{th}) percentiles. It means that half quartile selection depends upon the cutoff point. From 'C' on the left side (L_S) in 'OZ' the numbers of detected outliers are '4', and on the right side (R_S) in 'OZ' detected outliers are '5'.

Figure 8.8: Outliers in a Bimodal Distribution of India Exchange Rate



The above Figure 8.8 shows the India Exchange Rate data series, in which cutoff point is C= 35.24, outliers Zone 'OZ' is the pair of (25^{th} to 37.5^{th}) percentiles. It means that half quartile selection depends upon the cutoff point. It is observed from the left side of 'C' that the number of detected outliers is '2', and on the right side detected outliers are '7'. In this way, outliers around a cutoff point are detected in different bimodal distributions.

8.5 Advantages of the Bimodal Boxplot

Due to its importance and necessity, following are the several main advantages of the newly introduced bimodal boxplot.

i. A clear Picture of the Bimodal Distribution

Conventional boxplot does not describe the bimodality and its properties. Our newly introduced bimodal boxplot shows the features and clear picture of a bimodal distribution.

ii. Summary Statistics of each Side

It is also the main advantage of bimodal boxplot that it evidently shows the information about summary statistics of a bimodal distribution, unlike conventional boxplot. Too many points describe the significant behavior of a data series.

iii. Helpful in the Detection of Outliers

Though bimodal boxplot shows the cutoff point of a distribution which is helpful in the detection of outliers, those outliers are detected on both sides of mixing or cutoff point in outlier zone.

8.6 Chapter Summary

Before building the bimodal boxplot, it is necessary to find out the cutoff point of any bimodal distribution. Fluss et al. (2005) conditions applied on bimodal distribution for cutoff are modified and used for this purpose on real data. The detection of outliers in bimodal distribution is discussed around cutoff point. The newly introduced bimodal boxplot can be constructed to calculate five-point summary statistics on each side of cutoff point separately by using Excel sheet. Figure 8.4 displays the quarterly data (1981-I to 2013-IV) of Pakistan exchange rate, in which first the cutoff point (i.e. C= 57.69) is calculated which is mentioned in the bimodal distribution. Bimodal boxplot has been drawn according to the cutoff point and it observed both of the boxplot (i.e. upper and lower) with summary statistics.

Figure 8.5 shows the quarterly data (1981-I to 2013-IV) of UK consumption with the bimodal boxplot in which cutoff point C= 63.3 also clearly describes the summary statistics. Similarly, India exchange rate of the same period has cutoff point C= 35.24 with the bimodal boxplot which is easy to see the summary statistics in Figure 8.7.

For the same data of UK consumption the outliers zone 'OZ' is the pair of (75^{th} to 87.5^{th}) percentiles. From left side (L_S) on 'C' in 'OZ', the number of detected outliers are '4', and on the right side (R_S) in 'OZ' detected outliers are '5'. In the same way for Indian Exchange Rate, outliers Zone 'OZ' is the pair of (25^{th} to 37.5^{th}) percentiles. It that means that half quartile selection depends upon the cutoff point. From left side on 'C', the number of detected outliers is '2', and on the right side, detected outliers are '7'. On this technique, the outliers are detected in case of bimodal distribution.

CHAPTER 9

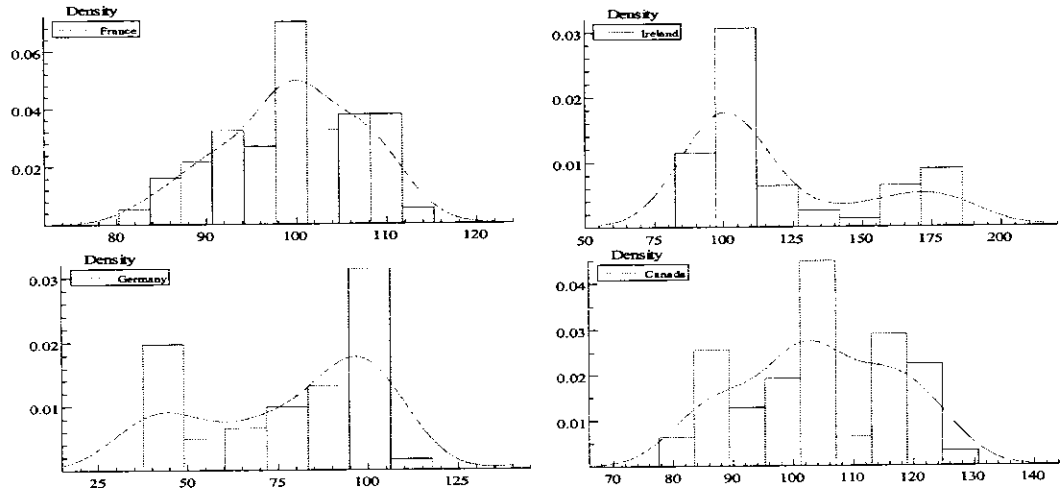
APPLICATIONS OF THE STUDY ON REAL DATA

This chapter deals with the real data set of various countries and cricket players' scores while applying the tools on the basis of the Figure 3.1.

9.1 Presentation of Real Data

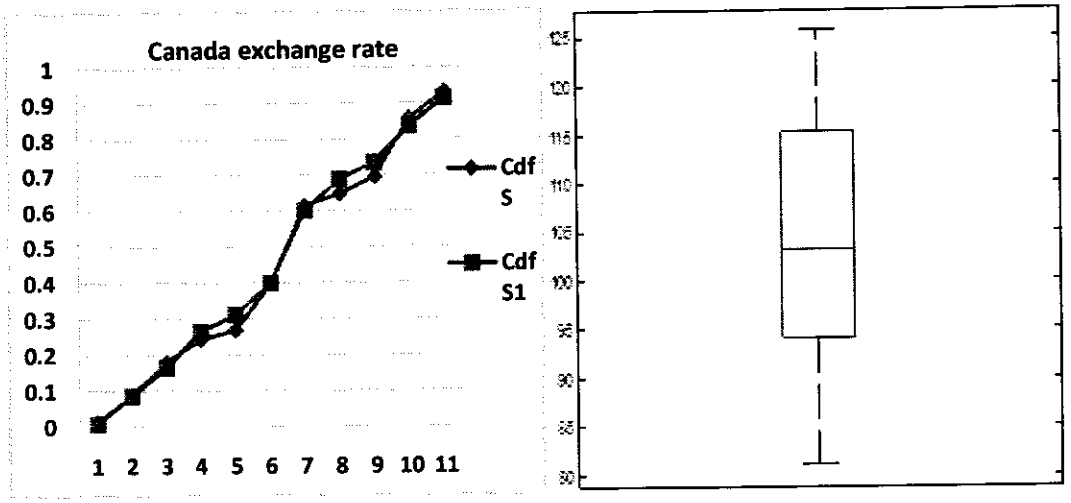
This study included the countries i.e. Canada, France, Germany, and Ireland with annual data of exchange rate for (1961 to 2013). Also, career score data of two Pakistani cricket players namely, Shoaib Malik and Ahmad Shehzad have been taken. First, the SB modality test has been used which is identified as the best test in this study. If data is unimodal, then skewness test is applied to observe the skewness of the data. Those data series which are detected as bimodal, then bimodal boxplots have been drawn for this distribution. According to SB test, Canada and France's data are unimodal while Germany and Ireland's data are bimodal distribution. These results have also been observed in the following Figure 9.1.

Figure 9.1: Specification of Various Distributions for Exchange Rate



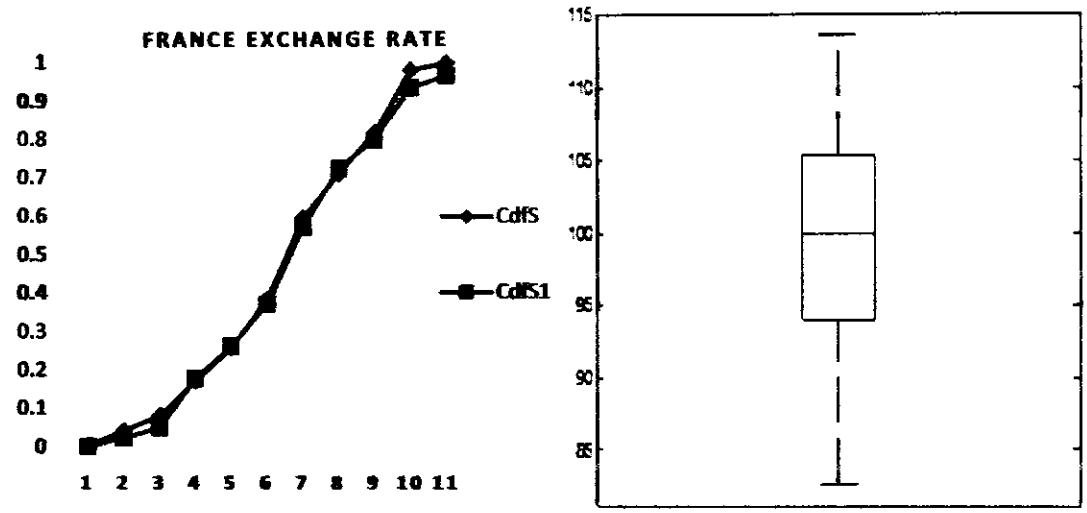
For the unimodal distribution, it may check whether the distribution is skewed or not. Here, measure P_{norm} has been applied because this measure performs well as compared to other measures and tests mentioned in Section 6.3.

Figure 9.2: CDFs and Boxplot of Canada Exchange Rate



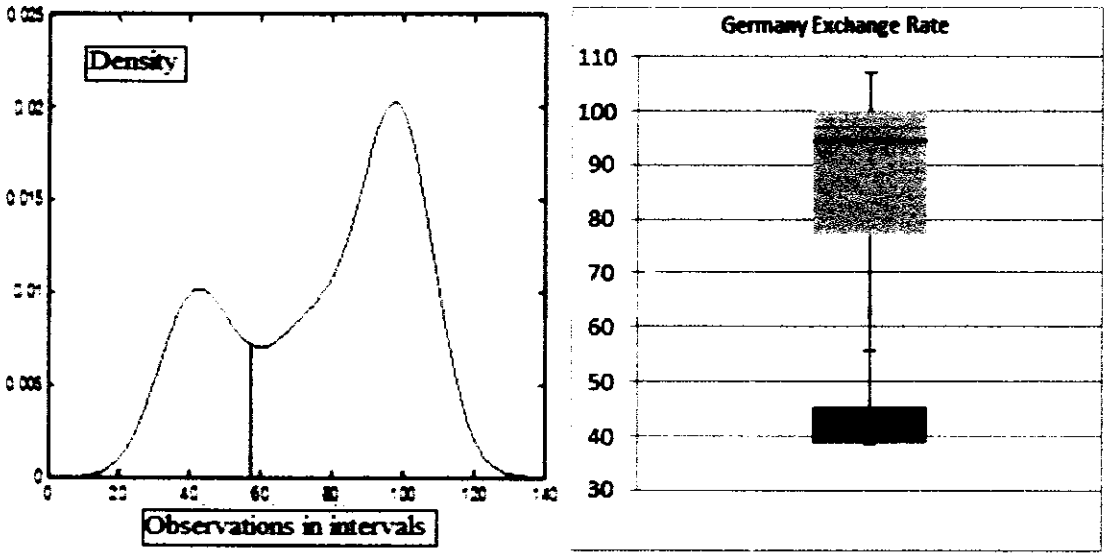
In the above Figure 9.2, the difference between the two CDFs discrepancy shows the skewness of the data. Overall, this picture shows to some extent skewness in the data. Therefore, according to the measure P_{norm}, the exchange rate series of Canada is slightly skewed. This result has also been verified through Tukey's boxplot as shown in the right side of Figure 9.2.

Figure 9.3: CDFs and Boxplot of France Exchange Rate



In the above Figure 9.3, the difference between two CDFs (i.e. CDF of actual data and symmetries CDF) has a very low discrepancy which shows that there exists no skewness. According to measure P_{norm}, the second exchange rate series of France is symmetric. Further, this result has also been verified through Tukey's boxplot as shown in the right side of Figure 9.3.

Figure 9.4: Bimodality and Bimodal Boxplot of Germany Exchange Rate



The above Figure 9.4 shows the Germany exchange rate which is a bimodal distribution and its cutoff point is 55.69. It looks like positively skewed because of a lot of data found on the right side of the cutoff point with a high peak. On the right side of the Figure 9.4, bimodal boxplot of the Germany exchange rate data series is drawn. The point lying on the joining line of the two boxes is actually the cutoff point which is also called the med-whisker with minimum= 38.41, maximum= 106.95 of this data series. Other summary statistics of lower part (before cutoff point) are, first quartile= 38.75, median= 39.49, third quartile= 45.47 and for upper part (after cutoff point) first quartile= 76.95, median= 94.43, third quartile= 100. It means that bimodal boxplot clearly shows the picture of the bimodal data and its summary statistics.

Figure 9.5: Bimodality and Bimodal Boxplot of Ireland Exchange Rate

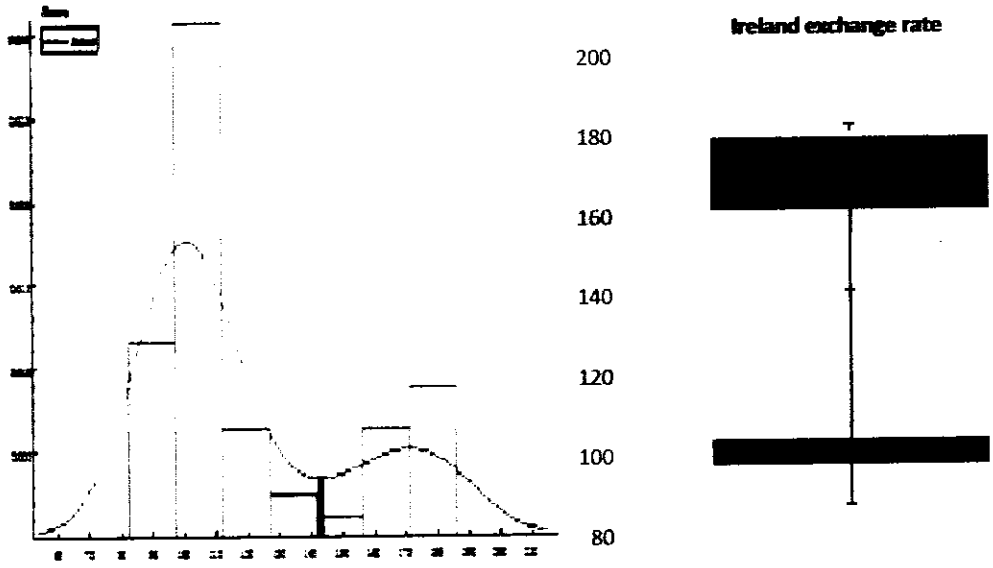


Figure 9.5 describes the Ireland exchange rate which is a bimodal distribution and its cutoff point is 141.3. It looks like positively skewed because a lot of data is found on the Left side of the cutoff point with a high peak. On the right side of the Figure 9.5, bimodal boxplot of the Ireland exchange rate data series is drawn. In this data series the cutoff

point or the med-whisker is shown between the two boxes and minimum= 87.91, maximum= 181.36. Summary statistics of lower part (before cutoff point) are, first quartile= 97.84, median= 100.26, third quartile= 104.18, and for upper part (after cutoff point) first quartile= 161.70, median= 170.38, third quartile= 179.67.

Figure 9.6: Detection of Outliers in Bimodal Distributions of Exchange Rates

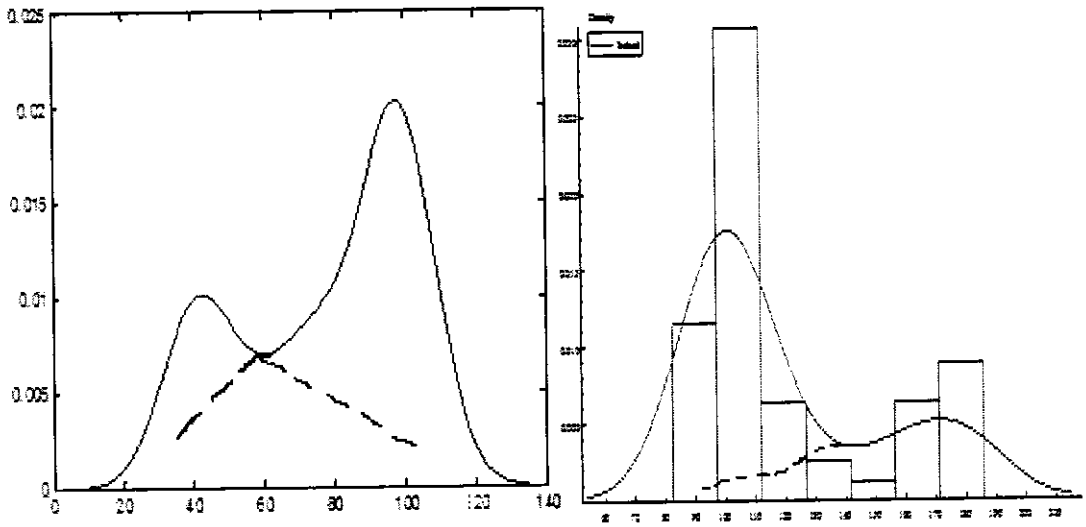
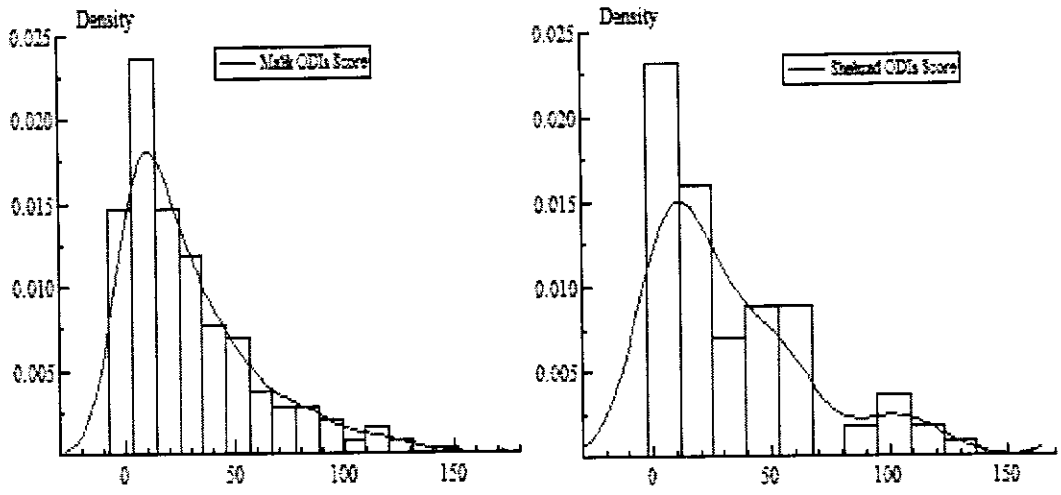


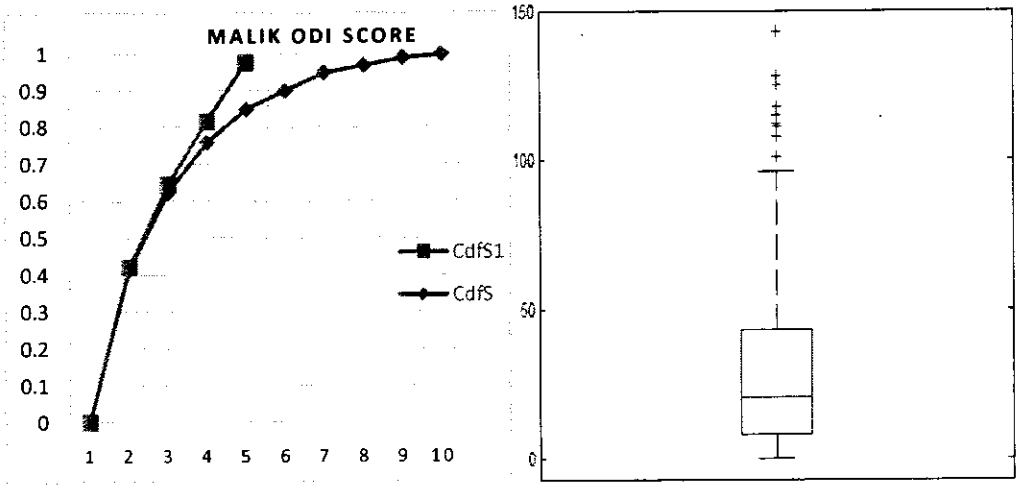
Figure 9.6 describes the detection of outliers in bimodal distribution of Germany and Ireland exchange rates (1961 to 2013). Both, positively and negatively skewed distributions in bimodality form are mixed together to make this bimodal distribution. On the left side, the Germany exchange rate bimodal distribution with the positively skewed distribution has 5 outliers (103.09, 103.17, 103.9, 104.05 and 106.95) while the negatively skewed distribution of Germany exchange rates has 2 outliers (38.41, 38.75). On the right side, the Ireland exchange rate bimodal distribution with the positively skewed distribution has no outlier while negatively skewed distribution has 4 outliers (87.91, 88.55, 90.59 and 92.1).

Figure 9.7: Specification of Two Distributions for Pakistani Cricketer's Scores



The above Figure 9.7 describes the distributions of two cricket players' data. When SB test is applied, it results that Shoaib Malik's ODI (One Day International) career score (1999 to 2017) data is unimodal and Ahmad Shehzad's ODI career (2009 to 2017) score is bimodal. Now P_{norm} is applied on unimodal distribution to check their skewness and also to construct bimodal boxplot for a bimodal distribution.

Figure 9.8: CDFs and Boxplot of Malik ODI Score



The above Figure 9.8 shows the difference between the two CDFs which is very high so it means that high skewness exists. Also, according to the numerical values of measure P_{norm}, Malik ODI score data series is more skewed. This result is also verified through the boxplot shown in the right side of Figure 9.8 which means that this distribution is positively skewed and plus sign represents the outliers.

Figure 9.9: Bimodality and Bimodal Boxplot of Shehzad ODI Score

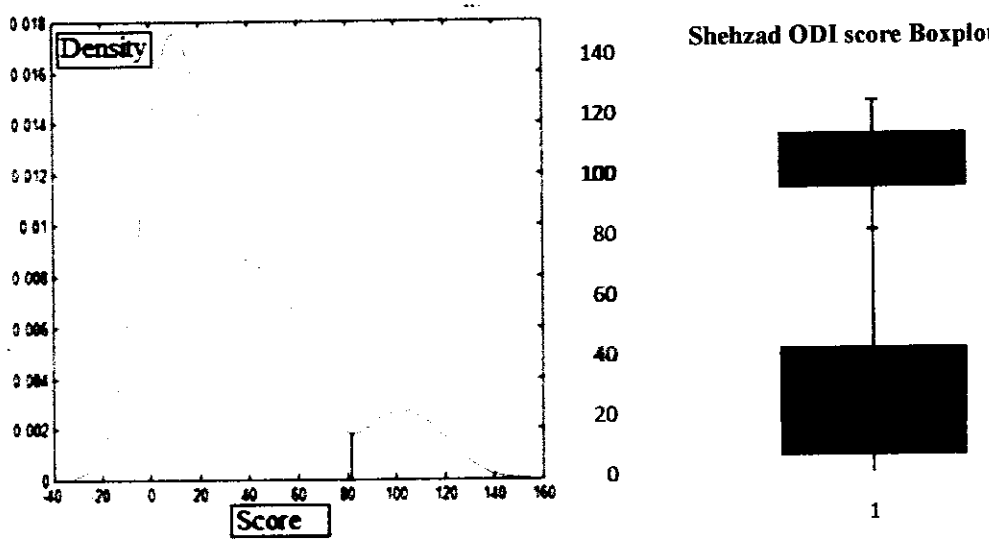


Figure 9.9 shows that the Shehzad's ODIs score is bimodal and its cutoff point is 81. Overall, it looks like it is negatively skewed because of a lot of data found on the left side of the cutoff point with a high peak in the same side. On the right side of the Figure 9.9, a bimodal boxplot of the Shehzad's ODIs score data is mentioned.

The cutoff point or the med-whisker is shown between the two boxes and the minimum and maximum values of this data are 0 and 124 (i.e. minimum= 0, maximum= 124). Other summary statistics are, in the lower part (before cutoff point), first quartile= 6, median= 17, third quartile= 42, and in the upper part (after cutoff point), first quartile=

95, median= 102, third quartile= 113. It means that the bimodal boxplot clearly shows the picture of the bimodal data and its summary statistics.

Figure 9.10: Detection of Outliers in a Bimodal Distribution of Shehzad ODI Score

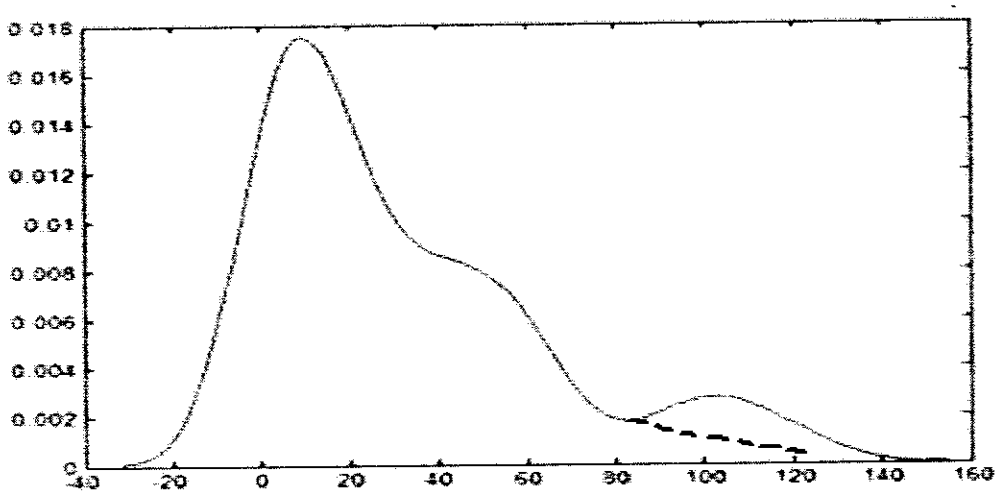


Figure 9.10 describes the detection of outliers in the bimodal distribution of Shehzad's ODI score. In this bimodal distribution, the positively skewed distribution has 9 outliers (81, 93, 95, 102, 103, 113, 115 and 124) while on the right side of the Figure 9.10, small peak skewed distribution has no outlier.

9.2 Chapter Summary

This chapter has demonstrated the existing and newly introduced statistical tool on real data set. Two types of real data sets have been taken here, i.e. exchange rates and cricket players score data. These analyses are extended step by step on the basis of the flowchart in Figure 3.1. First of all, SB test has been used to see whether the data is bimodal or unimodal. According to this test and graph result, the exchange rates of Canada and France are unimodal while Germany and Ireland exchange rates are bimodal. Similarly, for cricket data, Shoaib Malik's ODIs score is detected as unimodal while Ahmad Shehzad's ODIs score is identified as bimodal. For unimodal distributions, measure P_{norm}

has been used to check the symmetry or skewness in the distribution. According to this measure, the exchange rate of France is symmetric while the exchange rate of Canada and Shoaib Malik's ODIs score are skewed distributions. Further, their respective boxplots have been constructed.

Cutoff point or the joining point of a bimodal distribution is necessary before making the bimodal boxplot. Therefore, the Germany exchange rate cutoff point is detected as 55.69, Ireland exchange rate cutoff detected point is 141.3, and for Shehzad's ODIs score C= 81. Their respective bimodal boxplots are shown in above Figures (9.2, 9.4, 9.5, 9.8 and 9.9). From these bimodal boxplots, the cutoff point has been observed and different descriptive statistics which clearly indicate the bimodality nature of these data. Hence, it is observed that this new bimodal boxplot provides useful and accurate information of bimodal distribution.

CHAPTER 10

CONCLUSIONS, RECOMMENDATIONS, AND DIRECTIONS FOR FUTURE RESEARCH

This chapter discusses the main conclusions and results of this study. Also, the recommendations are proposed on the basis of main conclusions. At the end of this chapter, several particular ideas and directions have been suggested for future research work.

10.1 Conclusions

The existence of bimodality and comparison of modality tests have been discussed in chapter-4. Robertson's and Fryer's (1969) conditions have been used for the detection of bimodality in DGP-II, i.e. mixture of normals. The mixture of normals contains one standard normal and the second normal distribution. This study has kept the fix $\mu_1 = 0$, $\sigma_1^2 = 1$ and various values of other parameters (α, μ_2, σ_2) i.e. α = (0.1, 0.2, 0.3,. . . . . ., 0.9), μ_2= (1, 2, 3, ... ,10) and σ_2 = (0.1, 0.2, 0.3, ... , 0.9). When $\sigma_2 < \sigma_1$ in DGP-II, then Equation (3.2) has results negative real roots in which existence of unimodality is denoted by '0' while the existence of bimodality is denoted by '1'. When $\sigma_2 > \sigma_1$, then Equation (3.2) results of complex roots and all the results show that the mixture is unimodal. Similarly, in the mixture of two normal distributions when $\sigma_2 < \sigma_1$, then Equation (3.2) results as both positive and negative real roots and results are either unimodal or bimodal. In the case of $\sigma_2 > \sigma_1$, then Equation (3.2) complex roots and all results show unimodality (see details Table 4.1 to 4.3). These important results and choice of parameters values have shown the bimodality in Table 4.4. Moreover, this study has planned for checking the size and power properties of four modality tests.

This study used simulated critical values around the nominal size of 5% to stabilize all the modality tests. After the simulations result in Figure 4.1, all the four modality tests have stabilized size around the nominal size of 5%. Therefore, the powers of these modality tests have been further investigated.

The simulation results of Figures 4.2 to 4.3 on parameter values $\mu_2 = 1$, $\alpha = (0.5, 0.6, \ldots, 0.9)$ and $\sigma_2 = (0.2, 0.3, 0.4)$ show that the PM test has high power for high sample sizes as compared to the other three tests which have very low power. Also, keeping $\mu_2 = 1$, $\sigma_2 = 0.2$ fix and $\alpha = (0.1, 0.2, \ldots, 0.6)$, then the result remains same as has been observed from Figure (4.2). The power of the PM test increases as the sample size increases against parameter values $\mu_2 = 1$, $\alpha = (0.5, 0.6, 0.7, 0.8)$ and $\sigma_2 = 0.3$. In this situation, the power of Dip test, EM test and SB test is low, i.e. below 20% (see Figure 4.3). According to Figure 4.3, it is observed that the power of the PM test increases as the sample size increases and PM test is the most powerful test as compared to all other three tests. Figure 4.4 with parameters values (i.e. $\mu_2 = 9$, $\alpha = 0.2$ and $\sigma_2 = 0.7$) has identified that SB test is the most powerful test while PM test with low power behavior at overall sample sizes is recognized as bad performing test.

Results of Figure 4.5 with parameters combinations $\mu_2 = (5, 6, 8, 9, 10)$, $\alpha = (0.4, 0.5, 0.6)$ and $\sigma_2 = (0.7, 0.9)$ conclude that all the tests have high power at high sample sizes; while at small sample, the PM test has low power which is bad performing test. For the combination of the parameters $\mu_2 = (8, 9, 10)$, $\alpha = (0.7, 0.9)$ and $\sigma_2 = (0.8, 0.9)$ in DGP-II Figure 4.7 to Figure 4.9 indicate that SB test is the most powerful test in all sample sizes. However, the power of the EM test and Dip test also increases as the sample size

increases where the PM test has worse performance. Overall, it is observed that SB test is the robust and best test among all the modality tests.

This study has also introduced a new measure of skewness 'P_{norm}' based on CDFs which has been discussed in Chapter-5. This measure has many advantages over existing measures of skewness. Section 5.2 highlights the contrast of results among other measures of skewness and P_{norm}. These results appreciate the performance of P_{norm} in various cases.

In chapter-6, various measures and tests about skewness are compared on the basis of the Monte Carlo simulation results of size and power. All of the measures, that is, Prs, SM, MC, SSSB, Skw_1, Skw_2, and new measure P_{norm}, and tests, that is, KS test, t-test, and WC test have stable sizes on simulated critical values approximately 5% nominal size (see Figure 6.1). Due to the stable size, all these measures and tests for skewness are further compared on the basis of their power. In Section 6.2, Figure 6.2 to Figure 6.11 explains the power performance of all the measures and tests.

The simulations results of DGP-I lognormal distribution with parameters (mean=0, SD= 0.5) in Figure 6.2 describe that at small sample size (i.e. n= 60) the power of Prs, SSSB, MC and tests KS, WC are very low. But when the sample size increases, the power of all these measures and tests also increases. Changing the parameter values (μ= 6, σ= 3) for the same DGP, the simulation results in Figure 6.3 show that all the measures and tests have high power between 70%to 100%. But for this DGP with various parameter values, the new skewness measure, P_{norm}, has high power, approximately 100%, as compared to other measures and tests. Similarly, when the parameter values are changed, then results approximately remain same.

In the DGP-I, if Chi-square distribution with parameter v=1 is used, then all the measures and tests for skewness have high power according to Figure 6.4. As the parameter value increases to 8 (i.e. v= 8), then the measures Skw_1, Skw_2, SM, and P_{norm} have high power as compared to other measures and tests. Further, increase in the parameter value (v= 16 or 24), it is observed that all the measures and tests have low power except the measure P_{norm} which has high power, nearly equal to 100%, for all considerable sample sizes (See Figures 6.5 to Figure 6.7).

For more clarification, another skewed beta distribution has also been included for comparison, based on power. Figure 6.8 represents the simulations results for DGP-I of the beta distribution with parameter values (a, b)= (2, 15), where it is identified that the measures Prs, SM and P_{norm} have high powers. As parameter values increase to (a, b)= (4, 15), then only measure P_{norm} has high power as compared to all other measures and tests. When changing a= 6 or 8, then all the measures and tests have low powers but P_{norm} has a high probability of asymmetry, approximately 100%. Similarly, increasing the second parameter 'b', in the results the same variations and changes have been observed with the parameter 'a' (See Figure 6.9 to Figure 6.11). So, it is concluded that a new measure P_{norm} is a robust measure of skewness in the choice of symmetry and asymmetry.

Chapter-7 has discussed the size of bimodality in detail. For the size of bimodality, Trapezoidal and Simpson's rules have been used on the DGP-II with different values of three parameters (i.e. μ_2, α and σ_2), (see details in Table 4.4). It is found that the size of bimodality is affected due to change in parameters values. In case of DGP-II, it is identified that by increasing mean value (i.e. μ_2) and keeping 'α', σ_2 constant, the size of the bimodality increases, while α= 0.4 is the only case where the size is in decreasing

order (See Figures 7.2 to Figure 7.5). For changing mixing proportion, when alpha 'α' is increased and keeping μ_2 and σ_2 values constant, then the size of bimodality also increases (See Figures 7.6 to Figure 7.9). But, as the value of the standard deviation 'σ_2' changes, the size of bimodality also changes with very little margin (see Figure 7.10 to Figure7.11). So, it is concluded that the size of bimodality affected more when changes the values of parameters 'α' and μ_2.

In Chapter-8, newly introduced bimodal boxplot have been constructed which need cutoff point taken from Fluss et al. (2005) conditions which are then modified and used in this study for real data. Figure 8.5 describes the UK consumption of quarterly data from 1981-I to 2009-IV having cutoff point C= 63.63 and it has observed a clear picture of bimodal boxplot, where summary statistics are easily shown. Also, India quarterly exchange rate data from 1981-I to 2009-IV has detected cutoff point C= 35.24 with bimodal boxplot (see Figure 8.6), which clearly indicates a true picture of the data with five-point summary statistics on each side of cutoff point separately.

For the similar bimodal data of UK consumption and India exchange rate, outliers have been detected through selecting the specific outlier areas. Figure 8.7 sketches the UK consumption data, in which outliers Zone 'OZ' is the pair of (75^{th} to 87.5^{th}) percentiles showing that the mean half quartile selection depends upon the cutoff point. From 'C' to the left side (L_S) in 'OZ' where the number of detected outliers are '4' while on the right side (R_S) in 'OZ' number of detected outliers are '5'.

For India Exchange Rate data in Figure 8.8, outliers Zone 'OZ' is the pair of (25^{th} to 37.5^{th}) percentiles around cutoff point. From the left side on 'C', the number of detected

outliers are '2', and on the right side detected outliers are '7'. In this way, outliers are detected around a cutoff point in bimodal distributions.

In Chapter-9, this study used our methodology and applications on different real data series which are also explained through flowchart in Figure 3.1. The selected data is divided into two different categories, that is, exchange rate annual data (1961 to 2013) of Canada, France, Germany and Ireland, and cricket data of three Pakistani players of Shoaib Malik, Umer Akmal and Ahmad Shehzad's career score. According to Silverman modality test, the data of Canada and France are identified as unimodal while Germany and Ireland data are detected as bimodal (see Figure 9.1).

For a measure of skewness through P_{norm}, the exchange rate series of Canada is slightly skewed while exchange rate series of France is symmetric. The CDFs and Tukey's boxplots are shown in Figures 9.2 to Figure 9.3. Also, the Germany exchange rate is a bimodal distribution with cutoff point 55.69 while its respective bimodal boxplot is shown in Figure 9.4. Similarly, Figure 9.5 describes that the Ireland exchange rate is a bimodal distribution with cutoff point 141.3 and its bimodal boxplot. The Germany exchange rate and Ireland exchange rate have a number of outliers, i.e. 7 and 4, respectively (see Figure 9.6).

For Shoaib Malik's cricket data of ODI career scores (1999 to 2017) shows unimodality while Ahmad Shehzad's ODI career (2009 to 2017) scores is bimodal, observed in Figure 9.7. According to the numerical values of measure P_{norm}, Malik's ODI scores data series is highly skewed. The CDFs difference and its boxplot are shown in Figure 9.8. Shehzad's ODI scores is a bimodal distribution having cutoff point 81 along with its bimodal boxplot shown in Figure 9.9. This distribution has '9' outliers (see Figure 9.10).

As a result, it can be said that for both unimodal and bimodal distributions, the researchers may draw boxplot and summary statistics.

10.2 Recommendations

In this study, Robertson and Fryer's (1969) conditions have been used for checking the existence of bimodality, and for which values of the parameters the distribution is bimodal. In this aspect, it is recommended that for $\sigma_2 < \sigma_1$ in DGP-II, the result remains unimodality or bimodality and, for $\sigma_2 > \sigma_1$ the result states that the mixture is unimodal by using these conditions. Similarly, for $\sigma_2 < \sigma_1$ in a mixture of two normal distributions, the results have unimodality or bimodality, and for $\sigma_2 > \sigma_1$, all the results show unimodality. From our analysis, it is concluded that using analytical critical values, the size of several tests is not good and facing over rejection problems.

However, if simulated critical values are used, then the sizes of all tests are stable. So, it is suggested that using 5% nominal simulated critical values 'cv' rather than analytical critical values to avoid over-rejection problem. From power comparison analysis in case of mixture of normals, it is recommended to use Silverman bandwidth test as it performs well as compared to other modality tests. Also, according to simulations results for measures and tests of skewness in this study, it is recommended that the newly introduced measure of skewness, P_{norm}, performs well as compared to other measures and tests for skewness. This is the correct measure based on the difference between cumulative distribution functions. Further, when parameter in DGP of Chi-square distribution increases then the power of mostly measures and tests decreased. Also for DGP of Beta distribution as the values of first parameter 'a' increases the power of most

of the measures and tests decreased but the powers has ineffective mostly when changed the values of second parameter 'b'.

A new technique is introduced in this study to find out the size of the bimodality (the distance between the two modes or two peaks of a bimodal distribution) through Trapezoidal and Simpson's rules. It is also recommended from this technique that the size of bimodality depends upon the parameter values of mean 'μ' and mixing probability 'α' in bimodal data generating process.

For building bimodal boxplot in bimodal distribution, it is necessary to diagnose the cutoff point or the maximum separation point. The construction of newly introduced bimodal boxplot is recommended along with modification and extension of the Fluss et al. (2005) conditions for real data. In the last, 'OZ' (outlier detection area) around cutoff point is recommended for the detection of outliers in case of bimodal distribution.

The analysis of this study clearly shows the direction and behaviors of the data which will help the economists to improve their analysis. Our finding suggests, checking the bimodality and skewness of the data and then detecting the outliers to fit any modal which will give valid inferences about the data under consideration.

10.3 Directions for Future Research

This study is limited only for bimodal distributions and data is generated through unimodal and a mixture of two distributions. This study can be extended for multimodality and data can also be generated through a mixture of three or more distributions. Similarly, bimodal boxplot can be modified for multimodal boxplot through multiple cutoff points in the multimodal distribution. Detection of outliers for bimodality around a cutoff point can be extended to various 'OZ' for multimodal distributions.

Different outlier detection techniques can also be analyzed in this context. Analytical integrals can also be compared on the basis of the size of bimodality and extended to multimodality.

REFERENCES

Adil I. H. (2012). Robust outlier detection techniques for skewed distributions and applications to real data. Ph.D. Thesis. *International Institute of Islamic Economics, International Islamic University Islamabad, Pakistan.*

Adil, I. H. (2015). A Modified Approach for Detection of Outliers. *Pakistan Journal of Statistics and Operation Research*, *11*(1), 91-102.

Benjamini, Y. (1988). Opening the box of a boxplot. *The American Statistician*, *42*(4), 257-262.

Berens, W. (1988). The suitability of the weighted lp-norm in estimating actual road distances. *European Journal of Operational Research*, *34*(1), 39-43.

Bianchi, M. (1997). Testing for convergence: evidence from non-parametric multimodality tests. *Journal of Applied Econometrics*, *12*(4), 393-409.

Biehler, R. (2004, July). Variation, co-variation, and statistical group comparison: Some results from epistemological and empirical research on technology supported statistics education. In *Tenth International Congress on Mathematics Education, Copenhagen.*

Brimberg, J., Kakhki, H. T., & Wesolowsky, G. O. (2003). Location among regions with varying norms. *Annals of Operations Research*, *122*(1-4), 87-102.

Brys, M. H. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics, 13*(4), 996-1017.

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*(6), 604-621.

Chen, H., Chen, J., & Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(1), 19-29.

Cheng, M. Y., & Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(3), 579-589.

Choonpradub, C., & McNeil, D. (2005). Can the box plot be improved. *Songklanakarin Journal of Science and Technology*, *27*(3), 649-657.

Cohen, D. J., & Cohen, J. (2006). The sectioned density plot. *The American Statistician*, *60*(2), 167-174.

Doane, D. P., & Seward, L. E. (2011). Measuring skewness: a forgotten statistic?. *Journal of Statistics Education*, *19*(2).

Dovoedo, Y. H., & Chakraborti, S. (2015). Boxplot-based outlier detection for the location-scale family. *Communications in statistics-simulation and computation*, *44*(6), 1492-1513.

Dümbgen, L., & Riedwyl, H. (2007). On fences and asymmetry in box-and-whiskers plots. *The American Statistician*, *61*(4), 356-359.

Fisher, N. I., & Marron, J. S. (2001). Mode testing via the excess mass estimate. *Biometrika*, *88*(2), 499-517.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *47*(4), 458-472.

Frankland, B. W., & Zumbo, B. D. (2002). Quantifying bimodality Part I: an easily implemented method using SPSS. *Journal of Modern Applied Statistical Methods*, *1*(1), 22.

Frigge, M., Hoaglin, D., & Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician, 43*(1), 50-54.

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's practice what we preach: turning tables into graphs. *The American Statistician*, *56*(2), 121-130.

Groeneveld, R. A., & Meeden, G. (1984). Measuring skewness and kurtosis. *Journal of the Royal Statistical Society, 33*(4), 391-401.

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*(2), 181-184.

Hubert, M., & Veeken, S. V. (2007). Outlier detection for skewed data. Technical Report. *Katholieke Universiteit Leuven, department of mathematics.*

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

Hubert, M., Raymaekers, J., Rousseeuw, P. J., & Segaert, P. (2016). Finding outliers in surface data and video. *arXiv preprint arXiv:1601.08133.*

Imran, M. R. (2014). Testing for unimodality. MS. Thesis. *International Institute of Islamic Economics, International Islamic University Islamabad, Pakistan.*

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*, 21-30.

Kress, R. (1998). Numerical Analysis. *Springer, New York.*

Krzywinski, M., & Altman, N. (2014). Visualizing samples with box plots: use box plots to illustrate the spread and differences of samples. *Nature Methods, 11*(2), 119-121.

Lane, D. M., & Sándor, A. (2009). Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological methods, 14*(3), 239.

Love, R. F., & Walker, J. H. (1994). An empirical comparison of block and round norms for modeling actual distances. *Computers & Operations Research.*

Marmolejo-Ramos, F., & Tian, T. S. (2010). The shifting boxplot. A boxplot based on essential summary statistics around the mean. *International Journal of Psychological Research, 3*(1), 37-45.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician, 32*(1), 12-16.

Müller, D. W., & Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association, 86*(415), 738-746.

Potter, K., Hagen, H., Kerren, A., & Dannenmann, P. (2006). Methods for presenting statistical information: The box plot. *Visualization of large and unstructured data sets*, *4*, 97-106.

Raza, A. (2014). Comparison of classical tests of skewness versus bootstrap tests of skewness. MS. Thesis. *International Institute of Islamic Economics, International Islamic University Islamabad, Pakistan.*

Razzaque, S. (2009). The ultimatum game and gender effect: Experimental evidence from pakistan. *The Pakistan Development Review*, 23-46.

Rousseuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, *53*(4), 382-387.

Robertson, C. A., & Fryer, J. G. (1969). Some descriptive properties of normal mixtures. *Scandinavian Actuarial Journal*, *1969*(3-4), 137-146.

Schwertman, N. C., Owens, M. A., & Adnan, R. (2004). A simple more general boxplot method for identifying outliers. *Computational statistics & data analysis*, *47*(1), 165-174.

Shyu, C., & Ytreberg, F. M. (2008). Use of polynomial interpolation to reduce bias and uncertainty of free energy estimates via thermodynamic integration. *arXiv preprint arXiv:0809.0882.*

Spear, M. E. (1952). Charting Statistics. *McGraw-Hill,* 166.

Suli, E., & Mayers, D. F. (2003). An introduction to numerical analysis. *Cambridge University Press, Cambridge, UK.*

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

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

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

Wainer, H. (1990). Graphical Visions from William Playfair to John Tukey. *Statistical Science*, 340-346.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *The journal of general psychology*, *121*(4), 391-401.

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

APPENDIX

A.1 Mixture of a Normal and Two Uniform Distributions

The mixture of one uniform and one normal distribution is designed as. Let X_1 is uniformly distributed with parameters 'a' and 'b', X_2 is uniformly distributed with parameters a_1 and b_1 and X_3 is normally distributed with parameters μ and σ^2.

Let $X_1 \sim U(a, b)$, $X_2 \sim U(a_1, b_1)$ and $X_3 \sim N(\mu, \sigma^2)$

Then the mixture of two uniforms and normal is,

$$Z = \alpha X_2 + \beta X_2 + (1 - \alpha - \beta)X_3$$

Where $\alpha + \beta = 1$, these are the probabilities of a mixture of two uniforms and normal distribution.

Figure A.1: Bimodal Distribution with Parameters (μ_2, α, σ_2)= (3, 0.6, 0.7) and C= 2.43

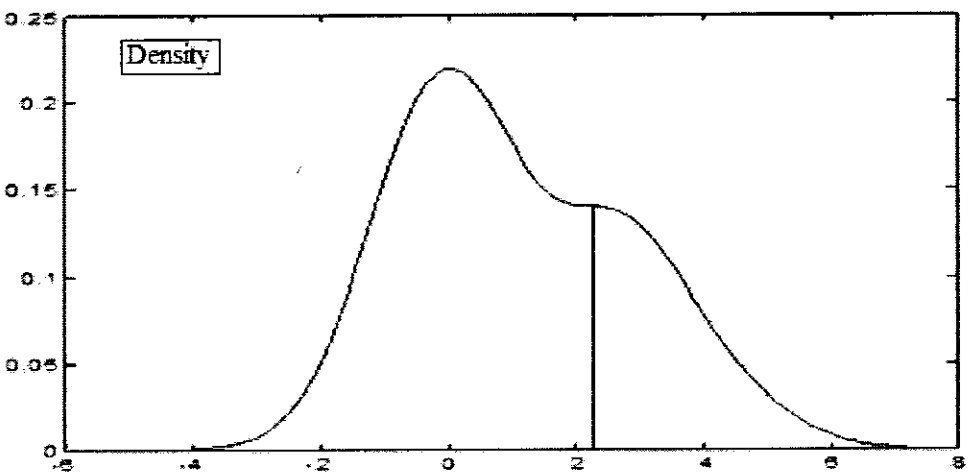


Figure A.2: Bimodal Distribution with Parameters (μ_2, α, σ_2)= (2, 0.5, 0.6) and C= 0.75

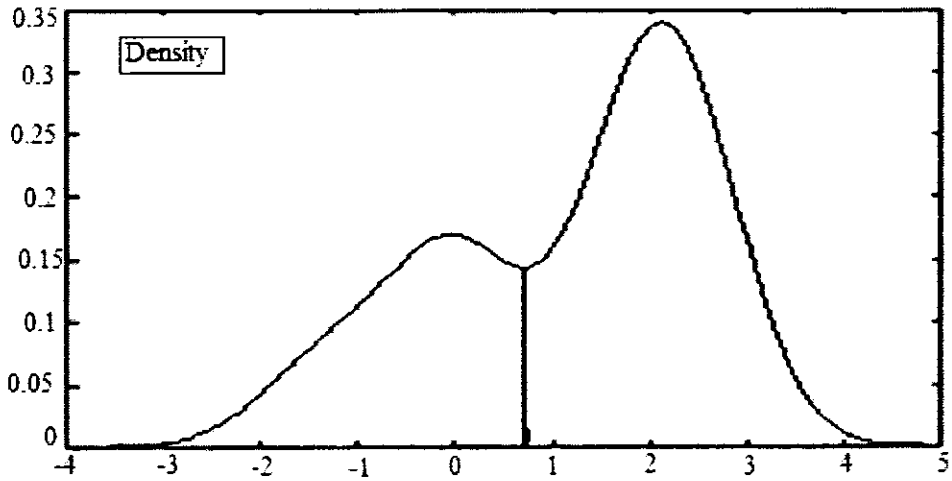


Figure A.3: Bimodal Distribution with Parameters (μ_2, α, σ_2) = (4, 0.5, 0.8) and C= 2.1

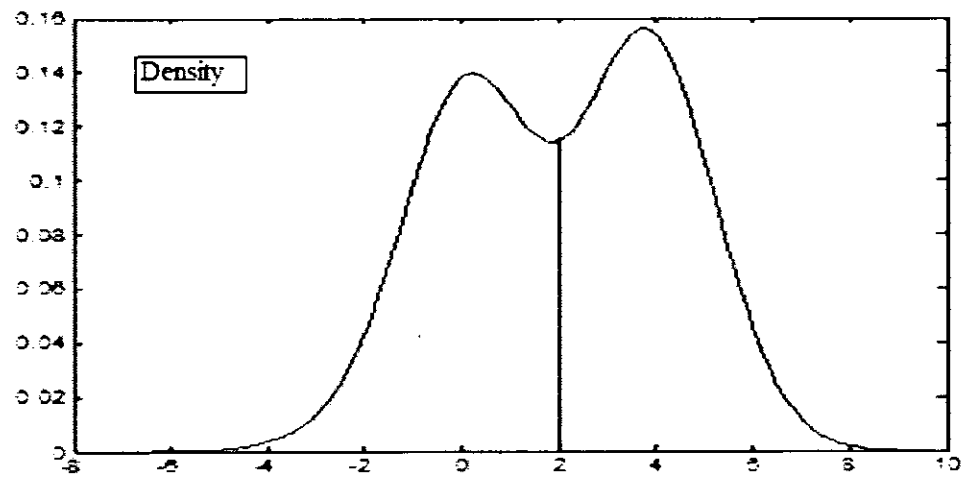


Figure A.4: Bimodal Distribution with Parameters (μ_2, α, σ_2)= (4, 0.6, 0.7) and C= 2.22

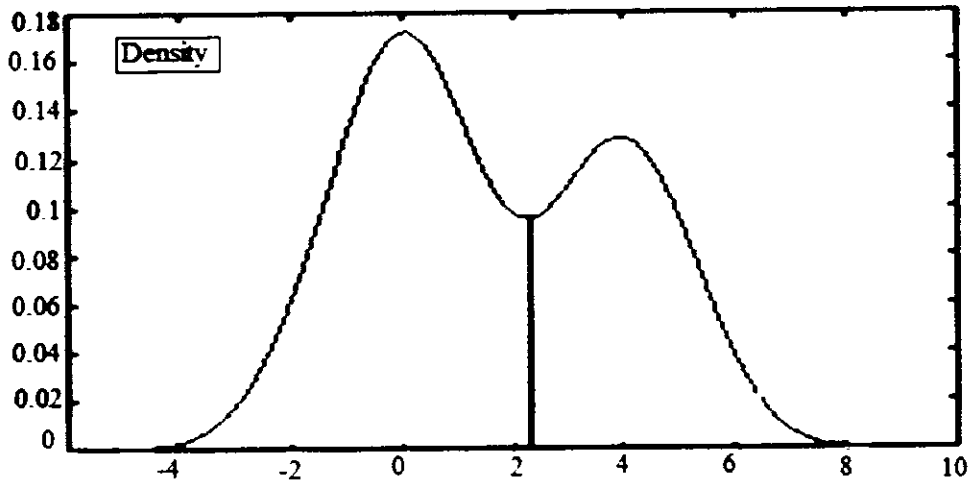


Figure A.5: Bimodal Distribution and Bimodal Boxplot of Belgium Export Rate with C= 2651.3

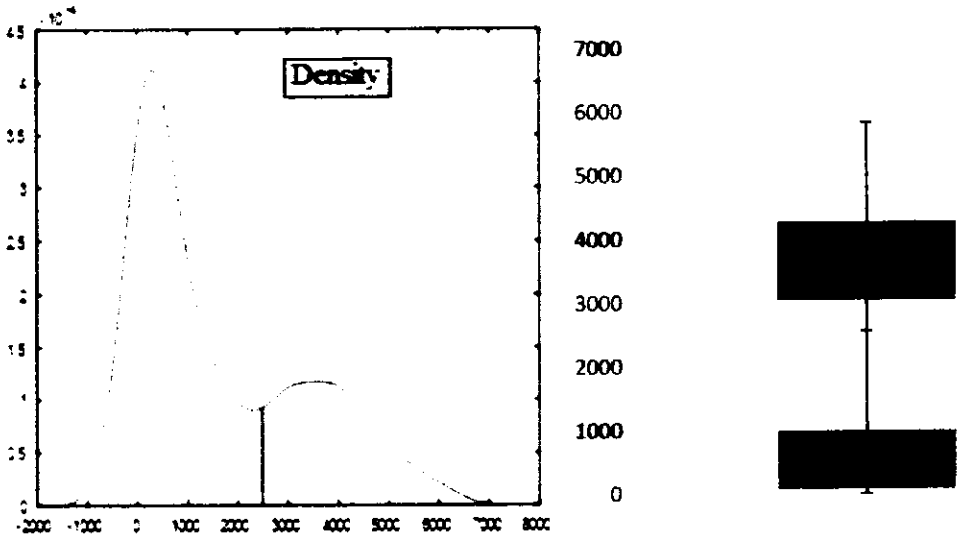


Figure A.6: Bimodal Distribution and Bimodal Boxplot of Philippine Export Rate with C= 378.11

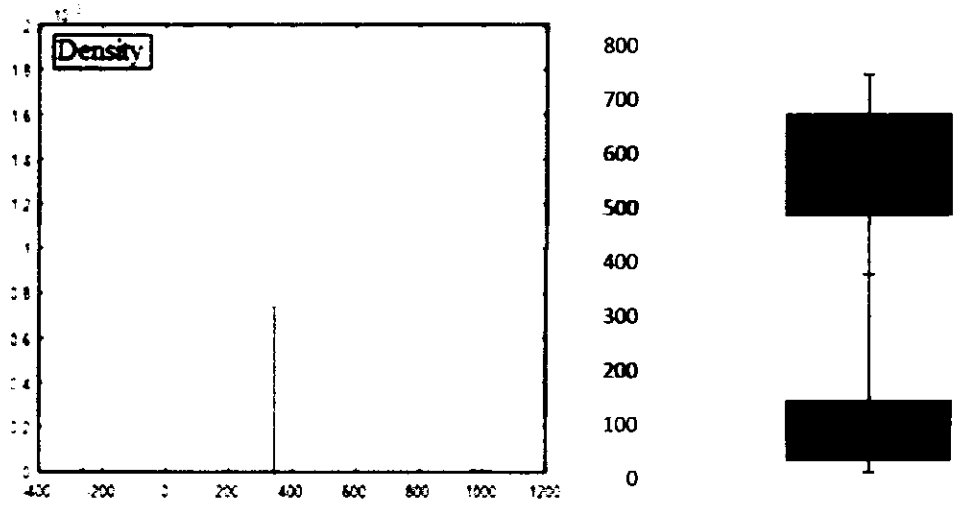


Figure A.7: Bimodal distribution and Bimodal Boxplot of Fiji Exchange Rate with C= 148.19