A New Approach to Determine the Coefficient of Skewness and An Alternative Form of Boxplot

Dr. Md. Forhad Hossain
Former Professor, Department of Statistics, Jahangirnagar University, Savar, Dhaka
Former Vice-Chancellor, Mawlana Bhashani Science and Technology University, Santosh,Tangail
E-mail: forhad.ju88@yahoo.com

Introduction

- Asymmetric dataset with rare statistical analysis.
- Equality of mean and variance may imply different shape characteristics.
- Skewness measures the degree of asymmetry of a dataset.
- There are many forms for measuring the coefficient of skewness. Such as,

  1. Fishers moment skewness.
  2. Pearson’s first and second coefficient of skewness.
  3. Yule’s coefficient of skewness.
  4. Forhad Adnan coefficient of skewness.

Problems and Motivation

1. The existence of one or more extreme value(s) may change the size of the tail of a dataset wrongly.

2. Irregular distance of the observations from the middle point of a dataset compared to average distance.

3. Problems of measures of central tendency and distance between two consecutive numbers.

4. a. Fisher's moment method of skewness is based on moments.

b. Pearson's both method is based on mean and standard deviation.

5. a. Yule's coefficient of skewness is based on three quartile and consider middle most 50% observation.
b. Forhad-Adnan's method is based on sample median.

6. Problems with traditional boxplot that indicates the degree of skewness through the spacing of different parts It
.

Objective of the Study

1. To determine the direction and value of the coefficient of sample skewness.

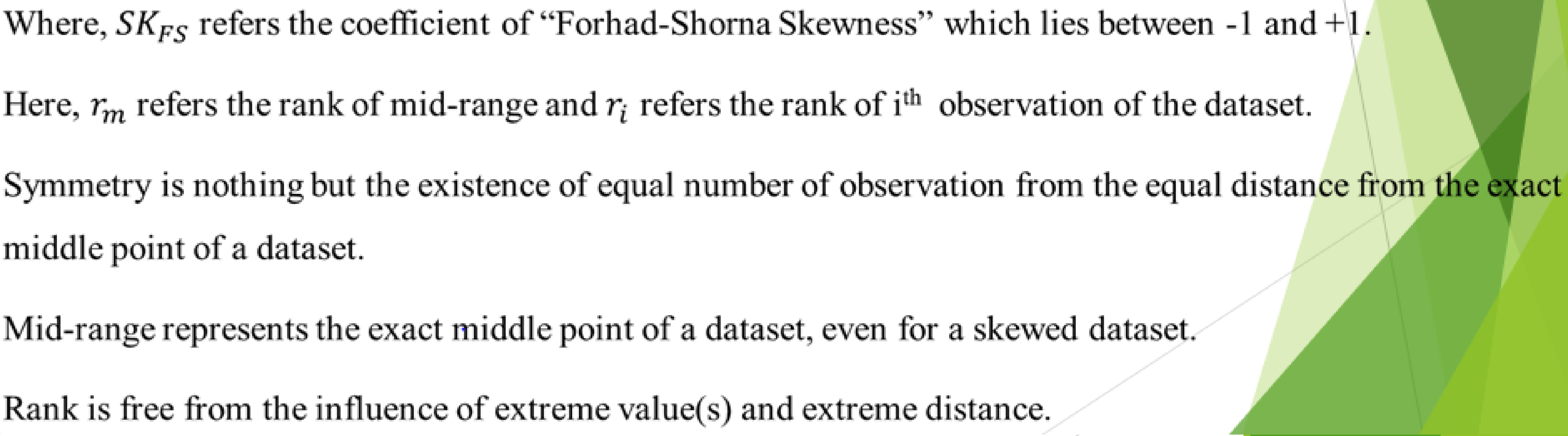
2. To establish a new form of measuring the coefficient of skewness.

3. To provide a better alternative to boxplot.

“F-S Skewness”

- A new approach of measuring the coefficient of skewness, termed as “F-S Skewness” has been suggested to solve the above mentioned problems,
- “F-S Skewness is a robust measure that may be defined as,

$$SK_{FS} = \frac{\sum_{i=1}^{n}(r_m - r_i)}{\sum_{i=1}^{n}|r_m - r_i|}$$



Where, SK_{FS} refers the coefficient of “Forhad-Shorna Skewness” which lies between -1 and +1.

Here, r_m refers the rank of mid-range and r_i refers the rank of ith observation of the dataset.

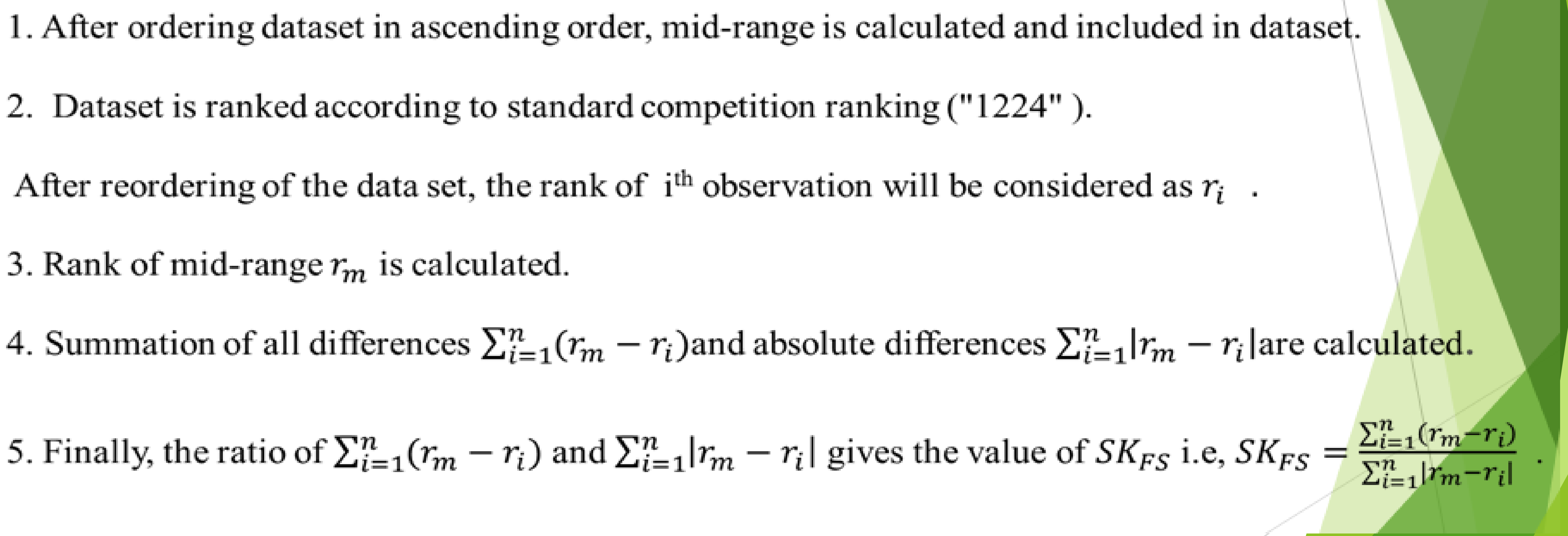
Symmetry is nothing but the existence of equal number of observation from the equal distance from the exact middle point of a dataset.

Mid-range represents the exact middle point of a dataset, even for a skewed dataset.

Rank is free from the influence of extreme value(s) and extreme distance.

Calculation of “F-S Skewness”

- In calculating “F-S Skewnes”, the following steps have been taken account,



1. After ordering dataset in ascending order, mid-range is calculated and included in dataset.

2. Dataset is ranked according to standard competition ranking ("1224").

After reordering of the data set, the rank of ith observation will be considered as r_i .

3. Rank of mid-range r_m is calculated.

4. Summation of all differences $\sum_{i=1}^{n}(r_m - r_i)$and absolute differences $\sum_{i=1}^{n}|r_m - r_i|$are calculated.

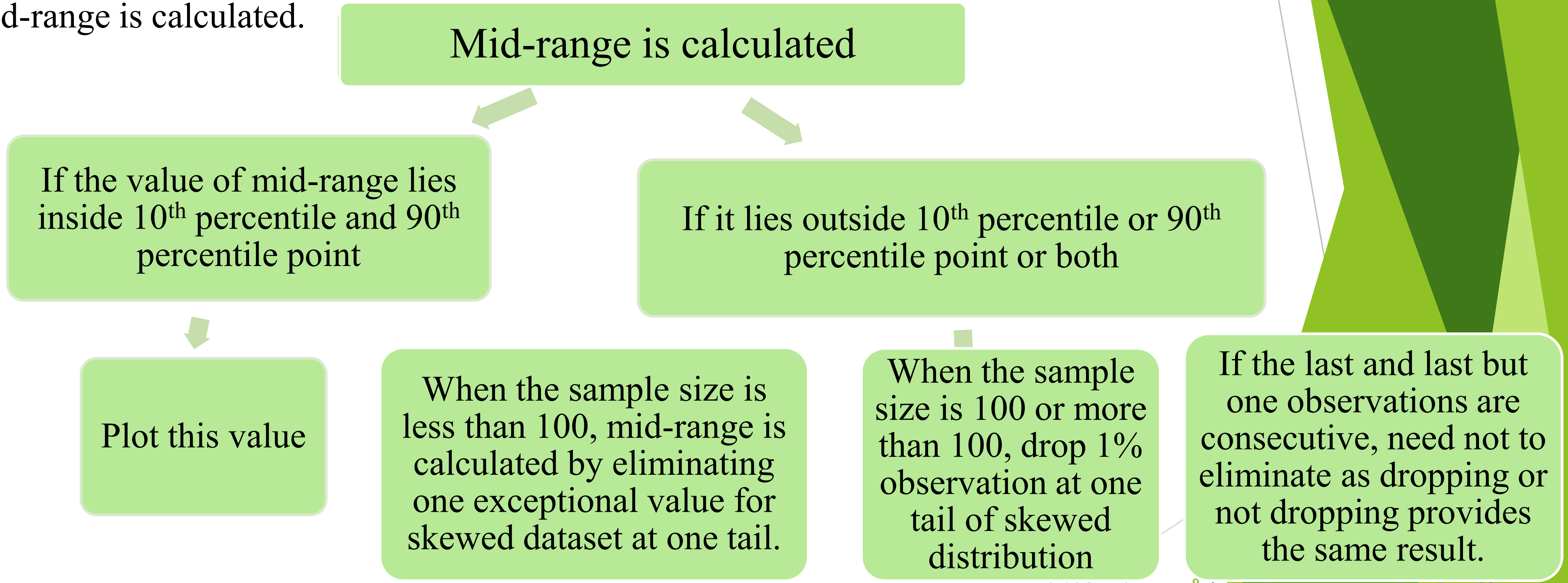
5. Finally, the ratio of $\sum_{i=1}^{n}(r_m - r_i)$ and $\sum_{i=1}^{n}|r_m - r_i|$ gives the value of SK_{FS} i.e, $SK_{FS} = \frac{\sum_{i=1}^{n}(r_m - r_i)}{\sum_{i=1}^{n}|r_m - r_i|}$.

“Modified Boxplot”

- Traditional boxplot represents scatteredness of observation, check skewness and detect outlier of a dataset.
- Due to some limitation of boxplot as mentioned, a better alternate termed as “Modified Boxplot” has been suggested here.
- “Modified Boxplot” is based on hundred percent observation of a dataset.
- It is based on six point of a dataset, which are minimum, maximum, median, mid-range, 10th percentile point and 90th percentile point. So, we may term it as six point summary.
- Midrange is the exact middle point of a dataset.
- Median represents the fifty percent observation of a dataset to its left and the rest fifty percent observation to its right side.
- So, the position of median compared to mid-range is the salient concept.

Construction of “Modified Boxplot”

- In the construction of “Modified Boxplot”, the following steps are used,
- The values of a dataset are plotted horizontally and a horizontal line is drawn.
- After that, minimum, maximum, median, 10th percentile and 90th percentile point are plotted on this line.
- Mid-range is calculated.



Mid-range is calculated

If the value of mid-range lies inside 10th percentile and 90th percentile point

Plot this value

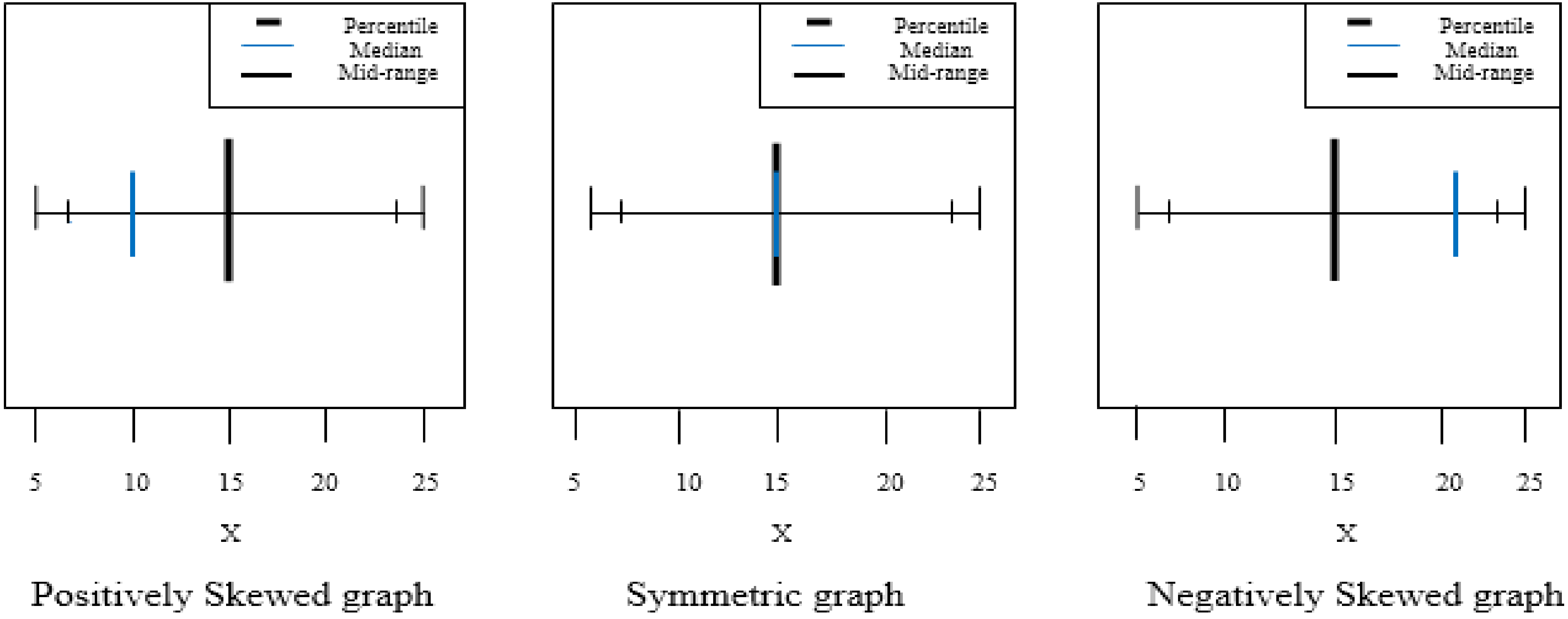
If it lies outside 10th percentile or 90th percentile point or both

When the sample size is less than 100, mid-range is calculated by eliminating one exceptional value for skewed dataset at one tail.

When the sample size is 100 or more than 100, drop 1% observation at one tail of skewed distribution

If the last and last but one observations are consecutive, need not to eliminate as dropping or not dropping provides the same result.

Template of “Modified Boxplot”



Positively Skewed graph

Symmetric graph

Negatively Skewed graph

Figure 1: Template of “Modified Boxplot”.

Interpretation: If the median lies at the left side of mid-range, the dataset may be considered as positively skewed and if it falls at the right side of mid-range, the dataset is negatively skewed, and if median falls on the mid-range, the dataset may be considered as a symmetric one.

Boundary of “Modified Boxplot”

- The fence (boundary of modified boxplot) for detecting outlier is,

$$[P_{10} - .9375PR;\ P_{90} + .9375PR]$$

- Where P_{10} is 10thpercentile point, $P_{90}$is 90th percentile point and PRrefers to the difference between $P_{10}\ and\ P_{90}$.
- Have considered 0.9375 instead of 1.5 as 1.5IQR and .9375PR measures the same.
- PR is based on middle 80% observation of a dataset whereas IQR is based on middle 50% observation.
- So, PR will provide much better information than IQR.
- Any Observation lying outside this fence may be considered as outlier.

Analysis Related to "Modified Boxplot"

- Have considered Normal, positively skewed Gamma and negatively skewed Gamma, positively skewed Weibull and negatively skewed Weibull distribution.
- Generated random number and obtained dataset of sample sizes (30, 50, 70, 90, 100, 110, 130, and 150).
- Then, we have applied the proposed modified boxplot and traditional boxplot method.

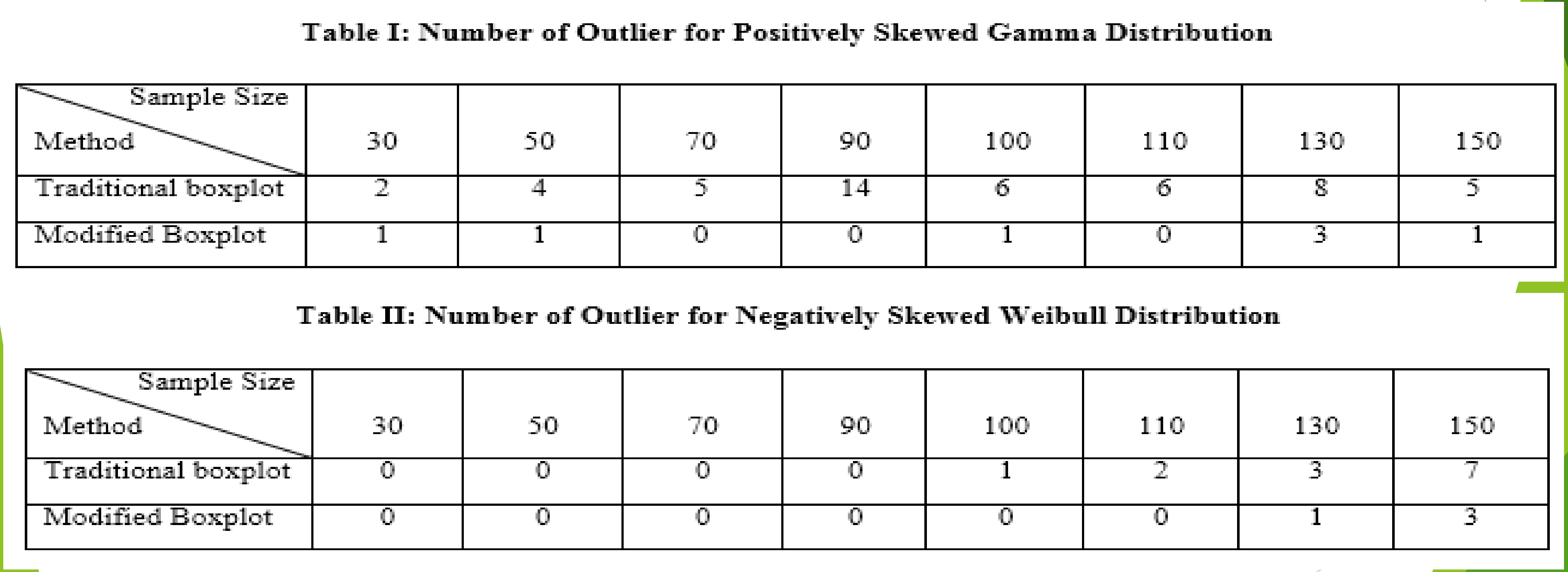


Table I: Number of Outlier for Positively Skewed Gamma Distribution

Sample Size Method	30	50	70	90	100	110	130	150
Traditional boxplot	2	4	5	14	6	6	8	5
Modified Boxplot	1	1	0	0	1	0	3	1

Table II: Number of Outlier for Negatively Skewed Weibull Distribution

Sample Size Method	30	50	70	90	100	110	130	150
Traditional boxplot	0	0	0	0	1	2	3	7
Modified Boxplot	0	0	0	0	0	0	1	3

These analyses support the modified boxplot and has provided better result compared to traditional boxplot method.

Analysis for “F-S Skewness” (Methodology)

1. Construction of dataset of different sample size.

2. Determination of coefficient of sample skewness for the constructed datasets.

3. Comparison of new formula with the existing formulae of measuring the coefficient of skewness.

4. Construction of modified boxplot plot as an alternate to traditional boxplot.

Considered Distributions

Normal distribution

Lognormal distribution

Positively skewed Gamma distribution

Negatively skewed Weibull distribution

Negatively skewed Weibull distribution

Considered Forms of Measuring the Coefficient of Skewness

1. Pearson’s second coefficient of skewness using mean and median(Pearson).

2. Fisher’s moment coefficient of skewness (Moment).

3. Yule’s coefficient of skewness (Bowley).

4. Forhad-Adnan skewness using median (FA).

5. The proposed F-S Skewness (FS).

Determination of the coefficient of Sample Skewness

1. Using a random seed (large prime number), we have made data bank of size 2000000 by Monte Carlo Simulation technique from each of the considered distribution.
2. Through bootstrap method, we have resampled the data bank and obtained 500000 datasets of different sample size(10, 20, 30, 40, 50, 60 and 100).
3. Applied the considered 5 forms to each of these 500000 datasets for the considered size and for considered distributions and obtained corresponding coefficient of sample skewness.

Thus we have obtained datasets of the coefficient of sample skewness each of size 500000.

Comparison of Different Coefficient of Skewness

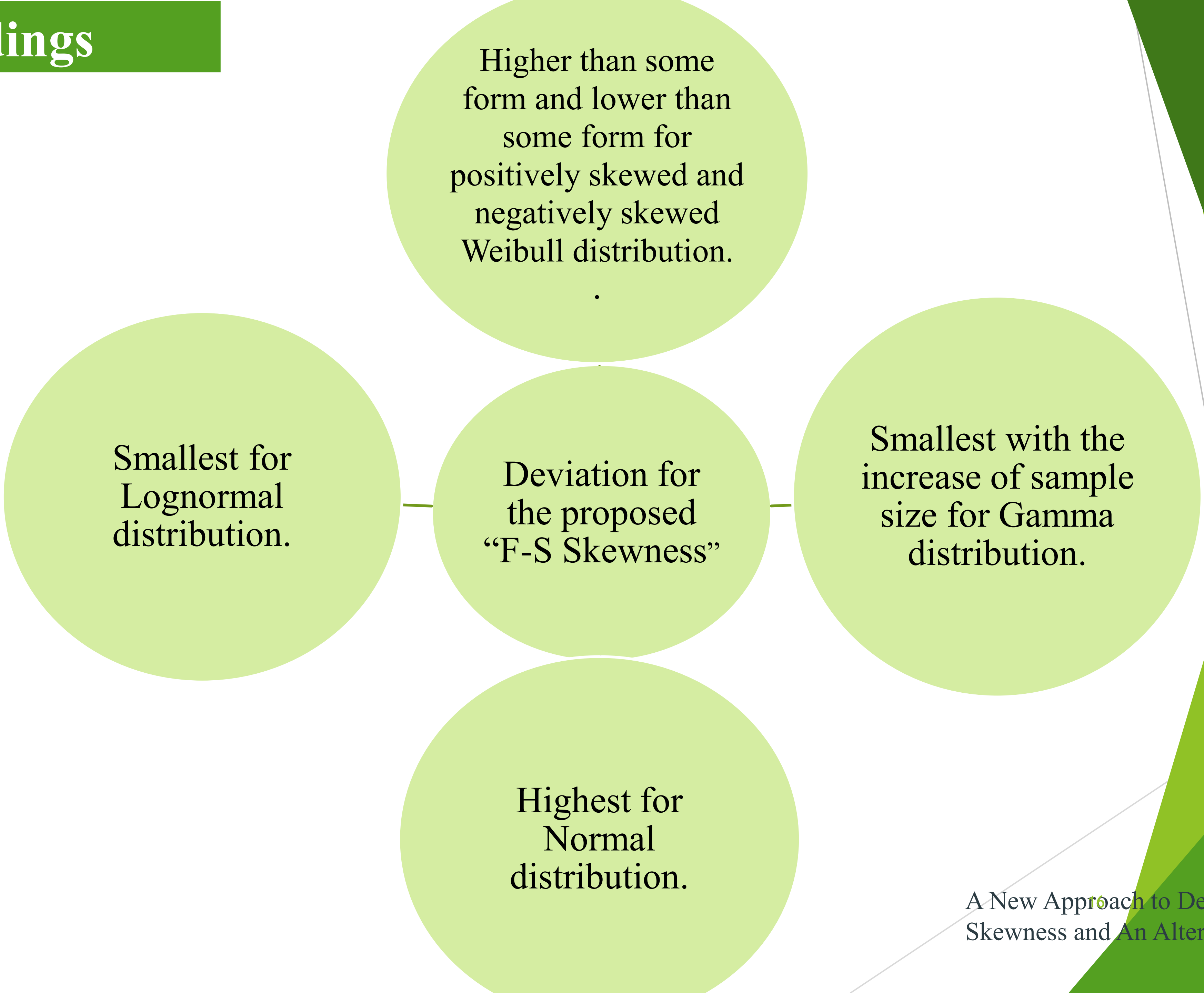
First method

- We have made comparison of “F-S Skewness” with the existing forms of measuring the coefficient of skewness.
- For comparison, firstly, we have considered
  1. Standard deviation
  2. Mean deviation from mean
  3. Mean deviation from median.

for all the datasets of coefficient of sample skewness for all sample sizes, all forms and all distributions.

- The smaller the deviation, the better the form.

Findings



Deviation for the proposed “F-S Skewness”

Higher than some form and lower than some form for positively skewed and negatively skewed Weibull distribution.
.

Smallest with the increase of sample size for Gamma distribution.

Highest for Normal distribution.

Smallest for Lognormal distribution.

Second Method

- We have considered the form of the coefficient of population skewness for all the five distributions.
- Applied the above-mentioned formulae on the previously obtained data bank.
- Then we have determined the coefficient of skewness .
- The closer the value of the coefficient of skewness (using different formulae) to the value of the coefficient of population skewness, the better the formula.
- **Findings:**
- Fishers skewness is the most closer, next is the proposed “F-S Skewness”.

Application in Real Life Problem

- Have considered three real life problems taken from Daniel (2007) and Devore (2000).
- Graphical representation of all the datasets provide better results in detecting outlier and showing the direction of skewness for proposed modified boxplot.
- By comparing with all the above specified forms, the proposed “F-S Skewness” performs well in measuring the coefficient of skewness for these datasets.
- Graphical representation and different method of measuring the coefficient of skewness for these three datasets are as follows,

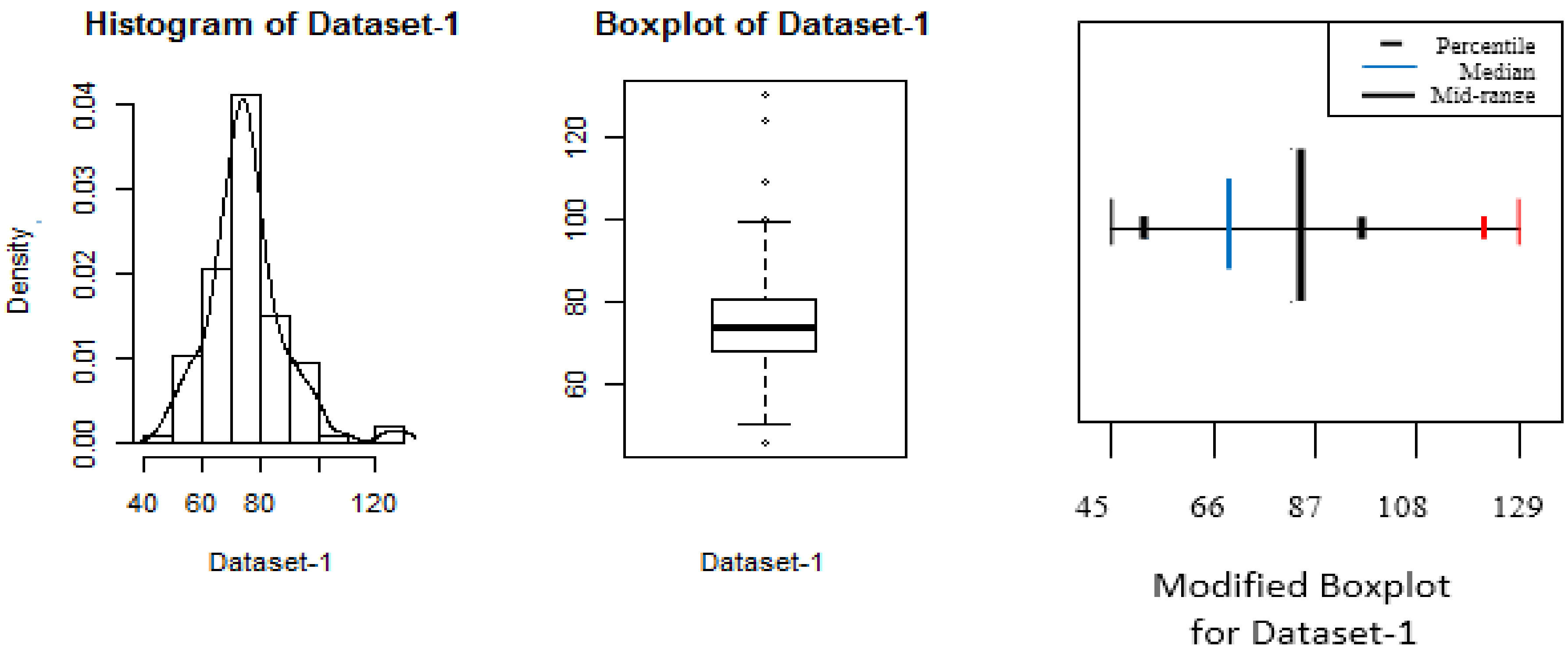


Figure 2: Histogram, Boxplot and Modified Boxplot of Dataset-1.

Table III: Value of the coefficient of skewness.

Method Dataset	Pearson	Moment	Bowley	FA	FS
	(-3 to +3)	(-3 to +3)	(-1 to +1)	(-1 to +1)	(-1 to +1)
Dataset-1	0.35118	0.993362	0.042471	0.16678	0.90880

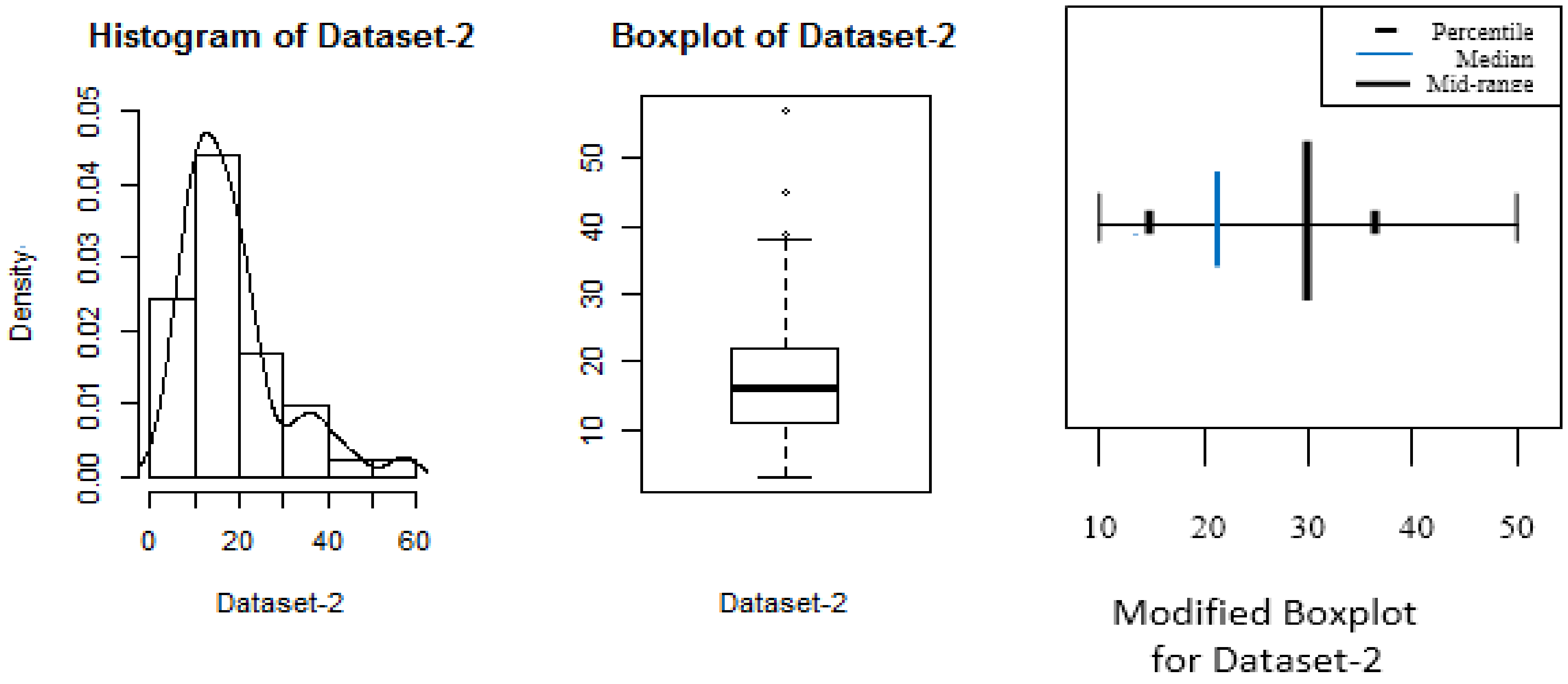


Figure 3: Histogram, Boxplot and Modified Boxplot of Dataset-2.

Table IV: Value of the coefficient of skewness.

Method	Pearson	Moment	Bowley	FA	FS
Dataset	(-3 to +3)	(-3 to +3)	(-1 to +1)	(-1 to +1)	(-1 to +1)
Dataset-2	0.591003	1.428262	0.0909091	0.283951	0.937685

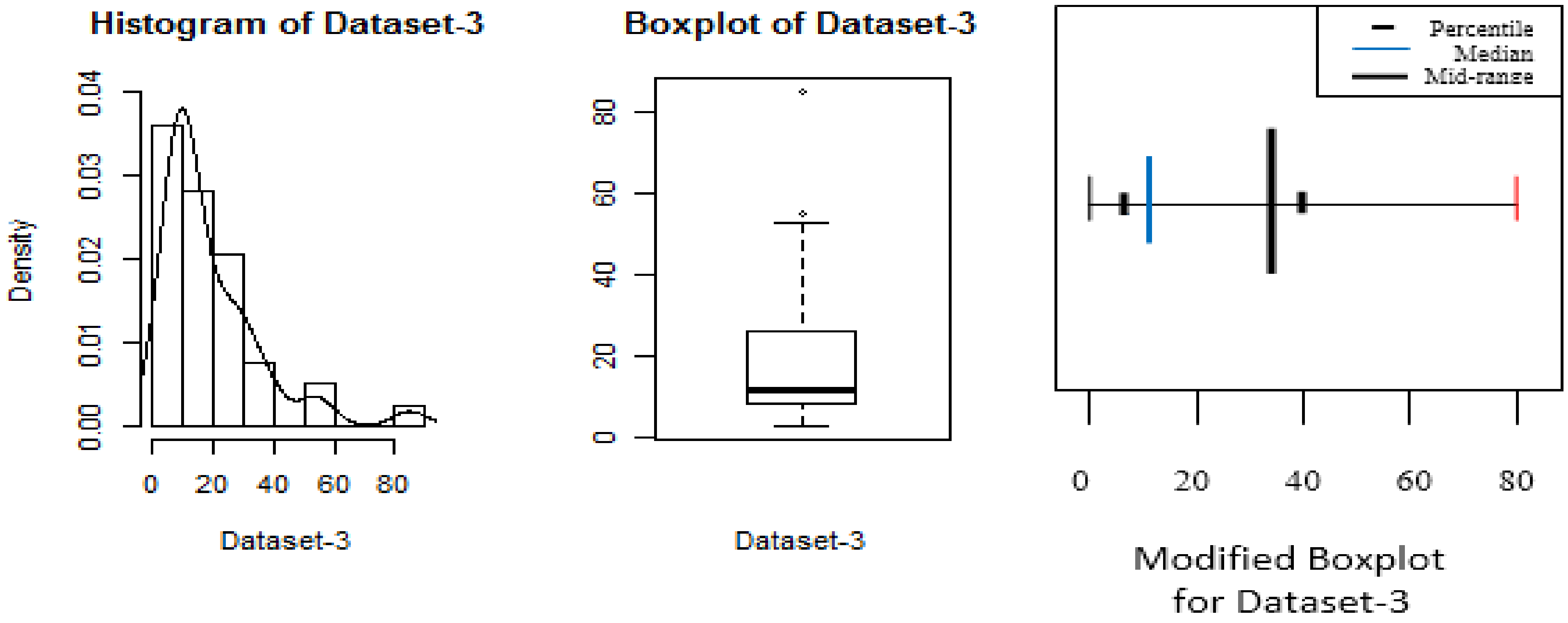
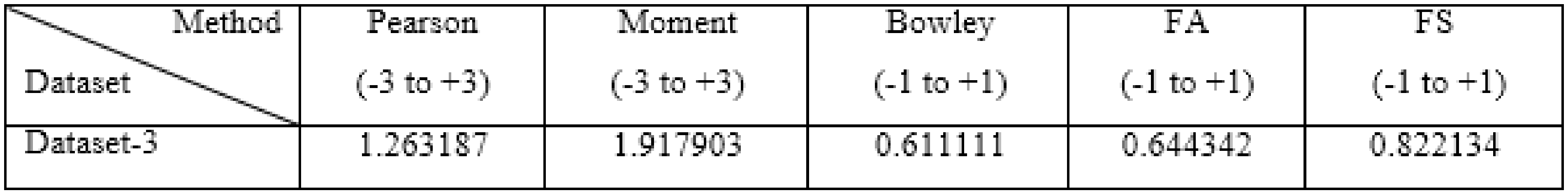


Figure 4: Histogram, Boxplot and Modified Boxplot of Dataset-3.

Table V: Value of the coefficient of skewness.



Method Dataset	Pearson (-3 to +3)	Moment (-3 to +3)	Bowley (-1 to +1)	FA (-1 to +1)	FS (-1 to +1)
Dataset-3	1.263187	1.917903	0.611111	0.644342	0.822134

Advantages of “F-S Skewness”

- After ranking, observations are not affected by extreme value or outlier.
- It only considers the number of observation left or right from the middle point of a dataset.
- It is free from the influence of extreme distance from the middle point of a dataset.
- It is free from the effect of irregular distance of two consecutive numbers.
- It is based on all the observations of a dataset through rank.

Conclusion

A new measure “F-S Skewness” has been suggested to solve some problems associated with the existing forms of measuring the coefficient of skewness of a data set.

The proposed “F-S Skewness” performs better than other forms specially for skewed distribution.

To overcome problems in traditional boxplot, “Modified Boxplot” has been suggested.
It is simpler, based on 100% observation of a data set, not affected by extreme value(s) and provide better results than traditional boxplot in detecting skewness and outlier.

References

- Bowley, A. L. (1901): Elements of Statistics, P.S. King & Son, London.
- Daniel, W.W. (2007): “Biostatistics: A Foundation for Analysis in the Health Sciences”, 7th edition, John Wiley &
- sons, Inc, pp.55 .
- Devore, J. L. (2000): “Probability and Statistics for Engineering and Sciences”, 5th edition, Dusbury Press, Boston,
- pp. 43-44.
- Gibbons, J.D. and Chakraborti, S. (2003): “Nonparametric Statistical Inference”, 4th edition, Marcel Dekkar, Inc.
- Hossain, M.F. and Adnan, M.A.S.A (2007): “A New Approach to Determine the Asymmetry of a Distribution,” Journal of Applied Statistical Science, Vol.15, pp. 127-134.
- Pearson, K. (1894): Contributions to the mathematical theory of evolution. I. In: Karl Pearson's Early Statistical Papers, Cambridge University Press, Cambridge, pp. 1–40.
- Pearson, K. (1895): Contributions to the mathematical theory of evolution. II: Skew variation in homogeneous material. In: Karl Pearson's Early Statistical Papers, Cambridge University Press, Cambridge, pp. 41–112.

Pearson, E.S. and Hartley, H.O. (1966): “Biometrika Tables for Statisticians”,vols. I and II. Cambridge University Press, Cambridge.

THANK
You!