An Asymmetrically Modified Boxplot for Exploratory Data Analysis

Michael Walker* S. Chakraborti†

Abstract

The boxplot, formalized by John Tukey, is a simple and effective graphical tool in many fields and disciplines. This paper highlights the origins and progression of the boxplot that is now widely used as an industry standard as well as its inherent limitations in outlier detection when dealing with asymmetric data. This background is necessary in understanding the ultimate aim of the paper, which is to present a new modification to the boxplot, the Ratio-Skewed boxplot, for use with any univariate data set, symmetric or skewed, regardless of the sample size. By incorporating an additional term to account for underlying skewness observed within the quartiles, the proposed methodology adjusts the boxplot fences in order to improve the effectiveness of the detection of outliers. Further, this additional term is shown to be highly related to the nonparametric measure of skewness known as Bowley's Coefficient. Through simulation studies this modification of the boxplot is shown to be simple and effective, as well as very consistent in outlier detection for several known distributions.

Key Words: Boxplot, Order Statistics, Outlier Detection, Box Plot, Nonparametric Statistics, Graphics

1. Introduction

The boxplot, first introduced in 1952 by Mary Eleanor Spear and later formalized by John Tukey, has grown into common use as a quick and effective graphical tool in many fields and disciplines that rely on the analysis of data. This paper highlights the origins and progression of the boxplot that is now widely used as an industry standard as well as its inherent limitations in outlier detection when dealing with asymmetric data. This background is necessary in understanding the ultimate aim of the paper, which is to present a new modification to the boxplot, the Ratio-Skewed boxplot, for use with any univariate data set, symmetric or skewed, regardless of sample size. This modification accounts for underlying skewness within the sample data and extends or retracts the fences of the boxplot accordingly in order to improve its ability to detect outliers. Further, the proposed modification is shown to have a close relation to Bowley's Coefficient for skewness. Through simulation testing the Ratio-Skewed boxplot is shown to be simple yet effective as well as very consistent in outlier detection for several known distributions.

2. Background

The essential features of what we now know as the boxplot were first introduced by Mary Eleanor Spear [13] sixty years ago. Though only briefly discussed, the book shows potential variations of what was then introduced as the range bar chart. The final modification clearly shows an early version of the boxplot later popularized by Tukey, complete with box and whiskers highlighting the points of the five number summary.

*Department of Information Systems, Statistics and Management Science, The University of Alabama, Tuscaloosa, AL 35487, mlwalker3@crimson.ua.edu

†Department of Information Systems, Statistics and Management Science, The University of Alabama, Tuscaloosa, AL 35487, schakrab@cba.ua.edu

Twenty years later, John Tukey [14] introduced a modified version of the range bar chart, known then as the schematic plot, that could highlight potential outliers in the sample data. In this early incarnation of the now widely used boxplot, the distribution of sample data was defined as being within the "hinges", the interquartile range, within the "sides", the inner fences, and being within the "corners", the outer fences. The "sides" were determined as being one interquartile range length beyond each quartile with the "corners" being two interquartile range lengths beyond each quartile. Thus, the boxplot as we know it was born as the schematic plot, with the box representing the interquartile range, dashed lines on either side leading to the most extreme values within the "sides", with points beyond the sides, but within the corners, plotted individually as being "outside", and extreme data points beyond the corners plotted separately as being "detached". In this early version of the boxplot, Tukey essentially used $k = 1.0$ as the inner fence constant to highlight potential outliers and $k = 2.0$ as the outer fence constant to highlight extreme outliers. Through repetition and experience, Tukey had altered and finalized these constants to $k = 1.5$ and $k = 3.0$ for the inner and outer fence constants, respectively, by the time he released his book, Exploratory Data Analysis [15].

Within a decade, the boxplot had become a widely used tool for EDA and had been implemented into several statistical software packages. However, while Tukey had settled on inner and outer fence constants of $k = 1.5$ and $k = 3.0$, respectively, these were not universally accepted and many software packages were using different constants to construct boxplots in the software as well as slightly differing definitions in finding the quartiles. The inner and outer fence constants differed considerably from $k = 1.0$ and 1.5 to $k = 1.0$ and 2.0, $k = 1.5$ and 3.0, and even $k = 2.0$ for the inner fences. While an inner fence using $k = 1.0$ flags potential outliers at a very high rate, in 50% of Gaussian samples, an inner fence using $k = 2.0$ only does so in approximately 10% of samples, while an inner fence using $k = 1.5$ flags potential outliers in 25% of samples [7]. The question of which fence constants to use in boxplot construction seemed to be more of an arbitrary selection than one based on mathematical or statistical considerations, relying on the type of data and needs of the analysis. This may be best summarized as an idea that it is more important not to miss any potential outlier than to avoid casting doubt on a good observation [8]. The aim of Frigge, Hoaglin, and Iglewicz in 1989 was to bring the differing fence constant selection and methods of boxplot construction into a standardized definition used across statistical software platforms, now universally using Tukeys fence constants of $k = 1.5$ and 3.0 as the default values. While these constants are now the standard in the construction of boxplots, their properties have continued to receive little formal appraisal [3].

3. Fairly Symmetric Assumptions and the Fence Constant

While the boxplot is extremely useful for quickly and efficiently visualizing distributions of sample data and highlighting potential outliers, its usefulness can be hindered by an underlying assumption of symmetry in the data. For data that are fairly symmetrically distributed, a boxplot based detection method works very well. When applying Tukeys boxplot to the standard Normal distribution, the inner fences are located at approximately ± 2.7, leaving an area of 0.0035 in each tail of the distribution. So, for a Gaussian sample, we would expect that only 0.7% of the data would be flagged as potential outliers. While there is no formal distinct definition of an outlier, this seems to be an appropriate representation of the idea of extreme value detection. For instance, another rule of thumb for outlier detection is the three-sigma rule, which states that outliers are values that are more extreme than three standard deviations away from the mean of a distribution. When applying this rule to the standard Normal distribution, we have an area of 0.00135 in each

tail, with the expectation that 0.27% of data in a Gaussian sample should be considered as extreme values compared to the characteristics of the distribution. Had the three-sigma rule been used in Tukeys formulation of boxplot outlier rules, the necessary constant to align the boxplots inner fences at these points would have resulted in a constant of $k = 1.723903$. Although not formally stated, Tukey may have had this idea in mind when finalizing the fence constants, possibly settling on $k = 1.5$ for intuitive reasons. In his book Exploratory Data Analysis, Tukey asserts that everything illustrated in the book can be done with pencil and paper and the only tools the illustrator used were a pen and a straightedge. He also maintains that absolute precision and extended decimal places are not entirely necessary in the book while addressing rounding, cutting, and decimal points at the very beginning of the first chapter. It is entirely possible that he was aware of the resulting three-sigma constant of $k = 1.723903$, but chose $k = 1.5$ not only to pull the fences in slightly, adjusting for sampling error witnessed through repetition and experience, but also for its inherent ease of use in calculation, ultimately given by

$$f_L^T = q_1 - 1.5IQR \qquad \text{and} \qquad f_U^T = q_3 + 1.5IQR,$$

with IQR denoting the interquartile range, and q_1 and q_3 denoting the first and third quartiles, respectively.

3.1 A Possible Formal Justification of 1.5

Fairly recently, a justification of the use of $k = 1.5$ as the inner fence constant was presented by Dumbgen and Riedwyl [6]. They present the idea that, presumably, everyone would suspect a data set to contain outliers or be highly skewed if the sample mean falls outside of the interval set by the interquartile range. Defining the five number summary as in Table 1 below,

Table 1: Five Number Summary

Minimum	First Quartile	Median	Third Quartile	Maximum
q_0	q_1	q_2	q_3	q_4

the authors begin with a proof that the sample mean will always fall between

$$\frac{q_0+q_1+q_2+q_3}{4} \qquad \text{and} \qquad \frac{q_1+q_2+q_3+q_4}{4},$$

and subsequently set bounds for the fences defined by

$$\frac{q_0+q_1+q_2+q_3}{4} \geq q_1 \qquad \text{and} \qquad \frac{q_1+q_2+q_3+q_4}{4} \leq q_3.$$

By rearranging these inequalities, the lower and upper fences are defined, respectively, as

$$f_L^{new} = q_1 - IQR - (q_2 - q_1) \qquad \text{and} \qquad f_U^{new} = q_3 + IQR + (q_3 - q_2).$$

It is noted that if the median is in the center of the box, i.e. the data are symmetric, then these fences coincide with those of Tukey with a fence constant of $k = 1.5$.

4. Skewed Distribution Failure

The above modification of the boxplot does seem to present a justification for the use of 1.5 as the fence constant for symmetric data, however its characteristics stem from an inherent failure in Tukeys boxplot to accurately flag outliers in data that are asymmetric. While Tukeys boxplot is an extremely useful tool for quickly visualizing distributions of one or several univariate samples, the outlier detection rules are only effective for data that are at least fairly symmetric. When applied to asymmetric, highly skewed data, the rate of flagged outliers tends to increase beyond acceptable bounds. For instance, while only 0.7% of the area under the standard Normal distribution falls beyond the established fences, the same boxplot applied to a Chi-squared distribution, with one degree of freedom, results in a lower fence at $x = -1.73$ that extends beyond the range of the distribution and an upper fence at $x = 3.156$ that leaves 7.57% of the distribution beyond the upper fence. So, for data from highly skewed distributions, a traditional boxplot flags an alarmingly high number of potential outliers, most of which should be expected to occur naturally in a highly skewed distribution.

Much research has been dedicated in recent years to modifying the boxplot to account for data from skewed distribution, with the earliest notable contribution from Kimber in 1990. Kimber introduced a modification to account for skewness by using the upper and lower parts of the interquartile range by splitting the interquartile range at the median. Data from a right-skewed distribution should naturally have right-skewed quartiles respective to the median, such that $(q_3 - q_2)$ is larger than $(q_2 q_1)$, with the opposite holding true for left-skewed distributions. It is this reasoning that led to Kimber's modified fences,

$$f_L^K = q_1 - c[2(q_2 - q_1)] \qquad \text{and} \qquad f_U^K = q_3 + c[2(q_3 - q_2)],$$

with c generally taken to be $c = 1.5$ [11]. By splitting the interquartile range, these fences are created by adjusting for asymmetry observed within the interquartile range to account for overall skewness in the data. Also, if the data are symmetric, these fences are identical to those obtained from Tukey's boxplot. While this modification does increase the performance of outlier detection for skewed distributions, the increase is only slight. Applying Kimber's boxplot to the same Chi-squared distribution, with one degree of freedom, as above, the lower fence is located at $x = -0.959$, still slightly below the range of the distribution, and the upper fence is located at $x = 3.928$, leaving 4.75% of the distribution beyond the upper fence. A slight improvement, yet still seemingly unacceptable, Kimber's idea of splitting the interquartile range has been a basis for much of the ensuing research on the subject, including Dumbgen and Riedwyl's modification highlighted earlier.

Further research into refining the boxplot for use with skewed distributions, many using Kimber's ideas as a basis or comparison, seem to overcomplicate the issue by introducing complex formulas for the fences, with or without added dependent variables, and making assumptions on the underlying distribution. Barnett and Cohen introduced two new modifications to the formulations of the fences, labeled as the Weibull fences and the Lognormal fences [1]. Both of these modifications rely on an assumption of the underlying distribution of the data. Since they were meant to be used with a very specific type of data, lifetime data, the distribution assumption is justified for their purposes. In general, however, assumptions on the underlying distribution of the sample data are rarely justified. The comparison of these modifications used on the lifetime data showed each to perform better than both the Tukey and Kimber methods. The Weibull fences are more complicated computationally, but the Lognormal actually performed better, despite the simplicity of calculation of the fences as follows,

$$f_L^{LN} = q_2\left(\frac{q_1}{q_3}\right)^2 \qquad \text{and} \qquad f_U^{LN} = q_2\left(\frac{q_3}{q_1}\right)^2.$$

While this modification of the boxplot performs well with the lifetime data, when used with typical skewed or symmetric data, these formulations may overextend the upper fence while underestimating the lower fence by a vast margin and are unlikely to highlight potential outliers of any kind.

Another recent modification of the boxplot for skewed data seems to perform fairly well by making use of a recently introduced robust measure of skewness known as the medcouple [4]. The subsequent adjusted boxplot includes the dependent variable MC, the medcouple, in the formulations of the fences, with additional differences in these formulations depending on whether $MC \geq 0$ or $MC < 0$ [10]. This modification, along with many others that include, among other things, using variable fence constants [5], sequential fence constants [12], and empirical distributions all attempt to create boxplots that may be more effectively used with asymmetric data. While such ideas are progressive and enlightening, they many times seem to push Tukeys original intentions for the simple use of EDA as the foundation or preliminary first step before confirmatory data analysis into complexity.

5. A New Approach for Skewed Distributions

The idea of splitting the interquartile range introduced by Kimber is very important in adjusting the boxplot for use with skewed distributions. Since any inherent skewness in distribution will be at least somewhat visible throughout the data, a skewed distribution will result in differing values for the lower and upper parts of the interquartile range. Specifically, let these interquartile splits be denoted as

$$SIQR_L = (q_2 - q_1) \qquad \text{and} \qquad SIQR_U = (q_3 - q_2).$$

For a symmetric distribution, the values of $SIQR_L$, the lower semi-interquartile range, and $SIQR_U$, the upper semi-interquartile range, will be equal. However for a positively, or right-skewed distribution, $SIQR_U$ will typically be greater than $SIQR_L$, and for a negatively, or left-skewed distribution, $SIQR_L$ will similarly be greater than $SIQR_U$. While Kimber suggested replacing the IQR with $2(SIQR_L)$ in creating the lower fence, and $2(SIQR_U)$ in creating the upper fence, this adjustment only slightly accounts for the inherent skewness in the use of boxplots.

The skewness observed within the interquartile range should not be seen as absolute, but relative. For a highly right-skewed distribution, $SIQR_U$ will be significantly larger in relation to $SIQR_L$, with the opposite holding true for left-skewed distributions. In this sense, for a right-skewed distribution, the distance or value of $SIQR_U$ would be less of a measure of skewness than the ratio of $SIQR_U$ to $SIQR_L$. So, using Tukey's original definition of the fences, with Kimber's idea as a foundation, we can utilize these ratios to more accurately adjust the upper and lower fences to account for underlying skewness in the data, resulting in the fences being defined as

$$f_L^{RS} = q_1 - 1.5IQR\left(\frac{SIQR_L}{SIQR_U}\right) \qquad \text{and} \qquad f_U^{RS} = q_3 + 1.5IQR\left(\frac{SIQR_U}{SIQR_L}\right).$$

Not only does this modification more accurately account for skewness than previous methods, but, like Kimber's modification, as well as Dumbgen and Riedwyl's modification, when the distribution is symmetric, these fences become identical to those intended by Tukey as both of these proposed ratio multipliers equal one.

As an example, a random sample from the Chi-squared distribution, with one degree of freedom, $n = 99$, was created. The plots in Figure 1 illustrate the differences in fence creation and the detection of potential outliers among the Tukey, Kimber, and Ratio-Skewed boxplots. Tukey's version flags an extraordinary number of outliers just beyond the upper

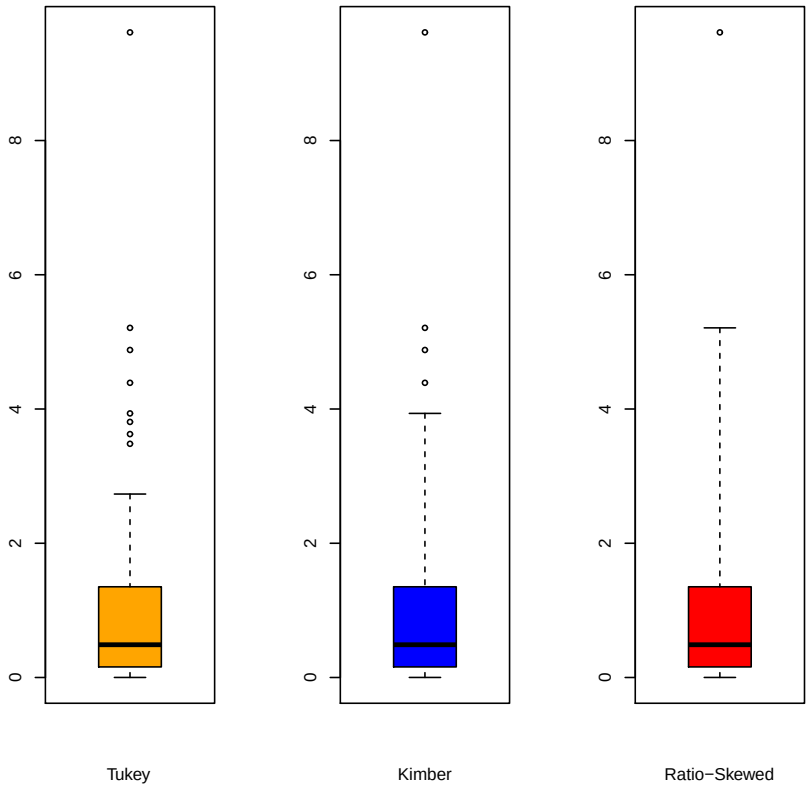

Figure 1: Comparison using positively skewed sample data, $n = 99$

fence, which in this Chi-squared distribution should by no means be considered outliers. Kimber's adjustment moves the upper fence farther out to account for skew, yet still leaves a number of data points collectively just outside the upper fence. The Ratio-Skewed modification moves the upper fence even farther due to the relative skew witnessed within the interquartile range, leaving just one extreme data point as a potential outlier. With a sample of $n = 99$, it seems that leaving one data point as a potential outlier could reasonably be expected regardless of the underlying and often unknown distribution of the sample.

6. Simulation Study

In order to compare the effectiveness of the proposed Ratio-Skewed modification against that of both the traditional Tukey method and Kimber's method, a simulation study was performed. Using randomly generated data from various known distributions, with varying levels of asymmetry, 1000 simulations were run for each sample size n, within each distribution. The results, seen in Table 2 represent the mean proportion of data points flagged as a potential outlier using the three methods. As n increases, this percentage essentially approaches the probability of a Type I error for each respective distribution. For the symmetric or fairly symmetric distributions, the three methods behave very similarly. While Tukeys method seems to perform slightly better in general for these distributions, the difference in the three is negligible. For the more skewed distributions, the difference in the three methods is easily seen, with the Ratio-Skewed approach outperforming both the Tukey and the Kimber methods. The most striking result however is that, regardless of the shape of the underlying distribution, the Ratio-Skewed method results in an almost constant percentage of data points flagged as outliers for similarly sized samples of n. As n increases, these percentages of potential outliers approach values at or near the 0.01 to 0.02 mark throughout, a percentage that may be expected in any distribution, not too extreme, nor too forgiving.

Table 2: Mean Proportion of Data Points Flagged as Outliers

Distribution	n	Tukey	Kimber	Ratio-Skewed
Normal(0,1)	10	0.0423	0.0581	0.0685
	20	0.0238	0.0332	0.0452
	30	0.0169	0.0259	0.0353
	50	0.0137	0.0182	0.0251
	100	0.0104	0.0132	0.0176
	500	0.0075	0.0079	0.0087
	1000	0.0073	0.0076	0.0081
Chi-Squared(1)	10	0.0819	0.0673	0.0542
	20	0.0809	0.0609	0.0389
	30	0.0809	0.0584	0.0347
	50	0.0782	0.0546	0.0284
	100	0.0767	0.0497	0.0211
	500	0.0764	0.0484	0.0171
	1000	0.0753	0.0476	0.0165
Chi-Squared(20)	10	0.0400	0.0548	0.0651
	20	0.0260	0.0341	0.0438
	30	0.0241	0.0284	0.0367
	50	0.0187	0.0200	0.0241
	100	0.0159	0.0150	0.0168
	500	0.0144	0.0116	0.0097
	1000	0.0140	0.0111	0.0090
Gamma(0.5,0.1)	10	0.0823	0.0696	0.0567
	20	0.0775	0.0598	0.0392
	30	0.0800	0.0569	0.0321
	50	0.0789	0.0534	0.0275
	100	0.0761	0.0490	0.0203
	500	0.0753	0.0474	0.0167
	1000	0.0757	0.0477	0.0165
F(90,10)	10	0.0665	0.0715	0.0743
	20	0.0583	0.0522	0.0515
	30	0.0552	0.0485	0.0449
	50	0.0542	0.0445	0.0374
	100	0.0531	0.0416	0.0316
	500	0.0521	0.0395	0.0267
	1000	0.0516	0.0394	0.0268

7. The Ratio as a Measure of Skewness and Bowley's Coefficient

The introduction of the ratios of the *IQR* splits into Tukey's original boxplot formulation have so far been shown to be a simple and effective method to adjust the fences accordingly due to the underlying skewness in the data. In fact these lower and upper ratios, defined as

$$R_L = \frac{SIQR_L}{SIQR_U} \qquad \text{and} \qquad R_U = \frac{SIQR_U}{SIQR_L},$$

respectively, can be shown to be measures of skewness themselves, centered around a value of 1. For a perfectly symmetric distribution, these two splits will be equal, leading to the ratio multipliers in the proposed boxplot formulation being equal to one, which then leaves

Tukey's original formulas for the fences. Likewise, differences in the splits will increase or decrease the ratios slightly above or below one, extending or retracting the fences in order to control unnecessary flagging of outliers, or false positives.

7.1 The Sum of the Ratios

In simulations it was seen that the sum of these upper and lower ratio multipliers, $R_U + R_L$, always had a minimum value of 2. For a perfectly symmetric underlying distribution, each of the ratios would be equal to one, so the sum of the upper and lower ratios would equal 2. However, deviations from symmetry inflate this value, especially noticed in skewed underlying distributions, and it seems that the sum of these two ratios can be a nonparametric measure of skewness in the data. In fact, this sum reduces down to a measure that will always be greater than or equal to 2. Defining the upper ratio as R_U and the lower ratio as R_L, the sum of these ratios can be shown to reduce to the following,

$$R_U + R_L = \frac{(IQR)^2}{(IQR_U)(IQR_L)} - 2.$$

It should be noted that if the data are perfectly symmetric, then $IQR_U = IQR_L = \frac{IQR}{2}$, leading to a minimum value of this sum, which is a value of 2. As these IQR splits begin to differ, this sum increases to values greater than 2. Keeping in mind that $IQR_U + IQR_L$ always equals IQR, if one of these semi-interquartile ranges increases, the other decreases, leading to larger overall values for this sum of the ratios. Thus, the higher the value of this sum, the more skewness is being observed in the data.

7.2 The Difference of the Ratios

Similar to the case of the sum of the ratios, the difference of these ratios can also be seen as a measure of skewness. While the sum of these ratios equal 2 in the perfectly symmetric case, the difference of these ratios will equal 0 in the perfectly symmetric case and this difference can be shown to reduce to following formulation,

$$R_U - R_L = \frac{IQR(q_3 + q_1 - 2q_2)}{(IQR_U)(IQR_L)}.$$

An interesting result here arises when recognizing the form of Bowley's coefficient [2], in itself a nonparametric measure of skewness defined by $B_c = \frac{q_3+q_1-2q_2}{q_3-q_1}$, or equivalently $B_c = \frac{q_3+q_1-2q_2}{IQR}$, centered around 0 and taking values between -1 and 1 to indicate underlying left and right skewness, respectively. With this in mind, the previous formulation of the difference in the ratios can be shown to be equivalent to

$$R_U - R_L = (R_U + R_L + 2)B_c.$$

So, the difference in the ratios yields Bowley's coefficient for skewness multiplied by a shifted sum of the ratios found earlier. As the sum of the ratios was found in itself to be a measure of skewness with a minimum value of 2, the first term here is a measure of skewness with a minimum value of 4. As the Bowley coefficient is known to be centered at 0, this first term multiplier is affecting the spread of the Bowley coefficient such that, under normal conditions, the difference of the two ratios is centered at 0 with a spread of at least 4 times that of Bowley's coefficient alone. A graphical comparison follows, using a relatively large sample size of $n = 999$ for each of 10000 iterations, each random sample of Uniform(0,1) variables yielded a value for Bowley's coefficient and a value for

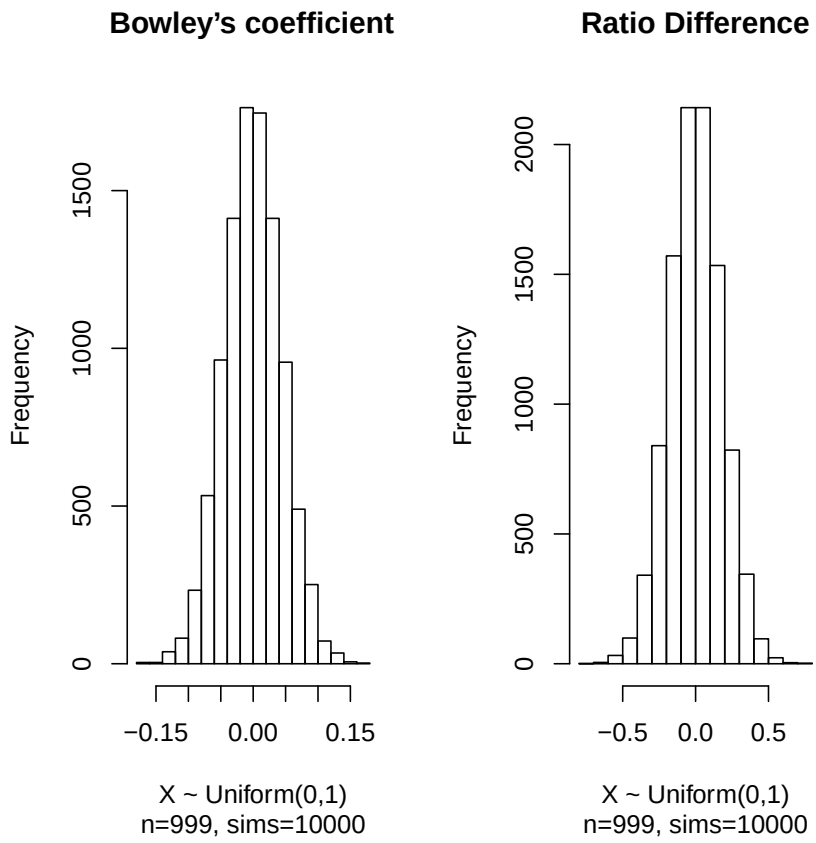

Figure 2: Comparison of Bowley's Coefficient and Ratio Difference, $n = 999$

the difference of the ratios. The results are shown in Figure 2. This difference in the ratios is inherently more sensitive to deviations from symmetry than Bowley's coefficient due to the presence of the sum of the ratios, along with Bowley's coefficient, embedded within this difference.

7.3 Redefining the Ratios

Using the previous result from the difference of the ratios, these ratios can then each be redefined in terms of Bowley's coefficient,

$$R_L = \frac{1 - B_c}{1 + B_c} \qquad \text{and} \qquad R_U = \frac{1 + B_c}{1 - B_c}.$$

Redefining the ratios in this way, the proposed boxplot fences are equivalent to

$$f_L^{RS} = Q_1 - 1.5IQR\frac{1 - B_c}{1 + B_c} \qquad \text{and} \qquad f_U^{RS} = Q_3 + 1.5IQR\frac{1 + B_c}{1 - B_c}.$$

This equivalent form of the proposed Ratio-Skewed fences further show that the ratio multipliers of the IQR splits are indeed incorporating a measure of skewness into the adjustment of the upper and lower fences of the boxplot. For a perfectly symmetric distribution, the value for Bowley's Coefficient, B_c, equals 0, leading to these ratio adjusted multipliers being equal to 1 and reverting the fence formulation back to Tukey's intended fences.

8. Conclusion

In the thirty-five years since Tukey's Exploratory Data Analysis, the use of the boxplot as a simple yet effective method of graphical analysis and comparison has become commonplace throughout various fields of study involving data. While there are today more sophisticated methods to graphically explore sets of data, the boxplot remains relevant due to its simplicity, ease of interpretation, and relative effectiveness. However, it is not without its limitations, specifically its inability to effectively point out potential outliers in data that are inherently skewed. This issue has been researched thoroughly for many years, with possible solutions to the problem becoming ever more complex. Many of these highly complex solutions have definite applications when dealing with specific types and sizes of data due to underlying assumptions, but the recent complex approaches to finding a single definitive modification of the boxplot for general use seem to have increasingly lost touch with Tukeys original intentions of simplicity and ease of interpretation in boxplots and EDA in general.

The Ratio-Skewed modification presented in this paper is a potential method of creating boxplots that work effectively regardless of underlying skewness or sample size. Through simulation and practice this modification has been shown to be quite useful. Not only is it shown to identify outliers at roughly the same rate throughout varying distributions, but is also as simple and easy to interpret as the standard boxplot used today. By retaining the basic structure of Tukey's boxplot, the addition of the $SIQR$ ratios in order to measure and adjust for underlying skewness is an enhancement with an apparently useful impact that can immediately be implemented throughout research and industry, as well as in the classroom, with ease.

References

[1] Barnett, O., and Cohen, A. (2000), "The Histogram and Boxplot for the Display of Lifetime Data," Journal of Computational and Graphical Statistics, 9:4, 759-778.

[2] Bowley, A. L. (1920), *Elements of Statistics* (4th ed.), New York, NY: Charles Scribner's Sons.

[3] Brant, R. (1990), "Comparing Classical and Resistant Outlier Rules," Journal of the American Statistical Association, 85:412, 1083-1090.

[4] Brys, G., Hubert, M., and Struyf, A. (2004), "A Robust Measure of Skewness," Journal of Computational and Graphical Statistics, 13, 996-1017.

[5] Carling, K. (2000), "Resistant Outlier Rules and the Non-Gaussian Case," Computational Statistics and Data Analysis, 33, 249-258.

[6] Dumbgen, L., and Riedwyl, H. (2007), "On Fences and Asymmetry in Box-and-Whiskers Plots," The American Statistician, 61:4, 356-359.

[7] Frigge, M., Hoaglin, D. C., and Iglewicz, B. (1989), "Some Implementation of the Boxplot," The American Statistician, 43:1, 50-54.

[8] Hampel, F. R. (1985), "The Breakdown Points of the Mean Combined With Some Rejection Rules," Technometrics, 27, 95-107.

[9] Hoaglin, D. C., and Iglewicz, B. (1987), "Fine-Tuning Some Resistant Rules for Outlier Labeling," Journal of the American Statistical Association, 82:400, 1147-1149.

[10] Hubert, M., and Vandervieren, E. (2008), "An Adjusted Boxplot for Skewed Distributions," Computational Statistics and Data Analysis, 52, 5186-5201.

[11] Kimber, A. C. (1990), "Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions," Applied Statistics, 39, 21-30.

[12] Schwertman, N. C., and de Silva, R. (2006), "Identifying Outliers With Sequential Fences," Computational Statistics and Data Analysis, 51, 3800-3810.

[13] Spear, M. E. (1952), *Charting Statistics*, New York, NY: McGraw-Hill.

[14] Tukey, J. W. (1972), "Statistical Papers in Honor of George W. Snedecor," Iowa State University Press, ed. T. A. Bancroft, Ames, Iowa, 293-316.

[15] Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.