Can the Box Plot be Improved?

Chamnein Choonpradub and Don McNeil*

Invented by Spear in 1952 and popularized by Tukey in 1977, the box plot is widely used for displaying and comparing samples of continuous observations. Despite its popularity, it is less effective for showing shape behaviour of distributions, particularly bimodality. Using robust estimators of data skewness and kurtosis to classify the distribution into categories, we suggest a simple enhancement for indicating bimodality, central peakedness, and skewness. We also suggest a new graphical method for displaying confidence intervals when comparing several samples of continuous data.

KEY WORDS: Box plot; Bimodality; Peakedness; Skewness; Kurtosis; Graphing confidence intervals; Multiple comparisons.

1. Introduction

The essential features of the box plot, called the range plot by Spear (1952) and popularized by Tukey (1977), are (a) a rectangular box extending from the lower quartile to the upper quartile of the data sample with a central dot or dividing line denoting the position of the median, and (b) additional lines called whiskers extending from each end of the box. In Spear's original definition the whiskers extend all the way to the minimum and maximum values, while in Tukey's modification each whisker extends no further than a fixed multiple of the interquartile range, with more extreme data (outliers) individually plotted.

Various modifications have been suggested, some purely cosmetic, some designed to better reveal the distribution of the data, and others to include confidence interval information. Using his principle of maximizing the data-ink ratio, Tufte (1983: 124-125) proposed that the box be entirely removed, but Benjamini (1988) rejected this idea on the grounds that it "gives the strange impression of seeing no data where the data are actually mostly concentrated". A recommendation by Frigge, Hoaglin and Iglewicz (1989) that the whiskers have length 1.5 times the interquartile range is now commonly accepted (see, for example, Cleveland 1994).

To some extent the box plot can show both skewness and bimodality in a distribution. Clearly, if the distribution is symmetric the symbol denoting the median is located at the centre of the box. Moreover, as Wainer (1990) pointed out, if the whiskers are sufficiently short relative to the interquartile range the distribution cannot be unimodal. But the reverse statements are not true. Like regression analyses that don't show residuals (Anscombe, 1973), box plots can mask the shape of a distribution, giving a misleading impression. Figure 1 displays histograms of four rather different sets of data each of size 100 and having the same range, and their common box plot. Each histogram has 15 bins of width 1.2 starting at 1.0.

---

*Chamnein Choonpradub is Lecturer, Department of Applied Mathematics and Computer Science, Prince of Songkla University, Pattani, Thailand, 94000. Don McNeil is Emeritus Professor, Department of Statistics, Macquarie University, Sydney, Australia, 2109.

The first sample comprises the normal scores for a sample of this size, scaled to range from 1.0 to 19.0. Sample 2 is a mixture of two identical symmetric clusters of data each of size 49 and centered at 7.4 and 12.6, respectively, together with isolated values at the ends of the range. Sample 3 is a mixture of 70 values spaced evenly over the range, 15 values at 9.5, and 15 values at 10.5. Sample 4 comprises a value at 1.0, 24 values at 7.4, 50 approximately evenly spaced values ranging from 7.4 to 12.6, and 25 approximately evenly spaced values ranging from 12.6 to 19.0.
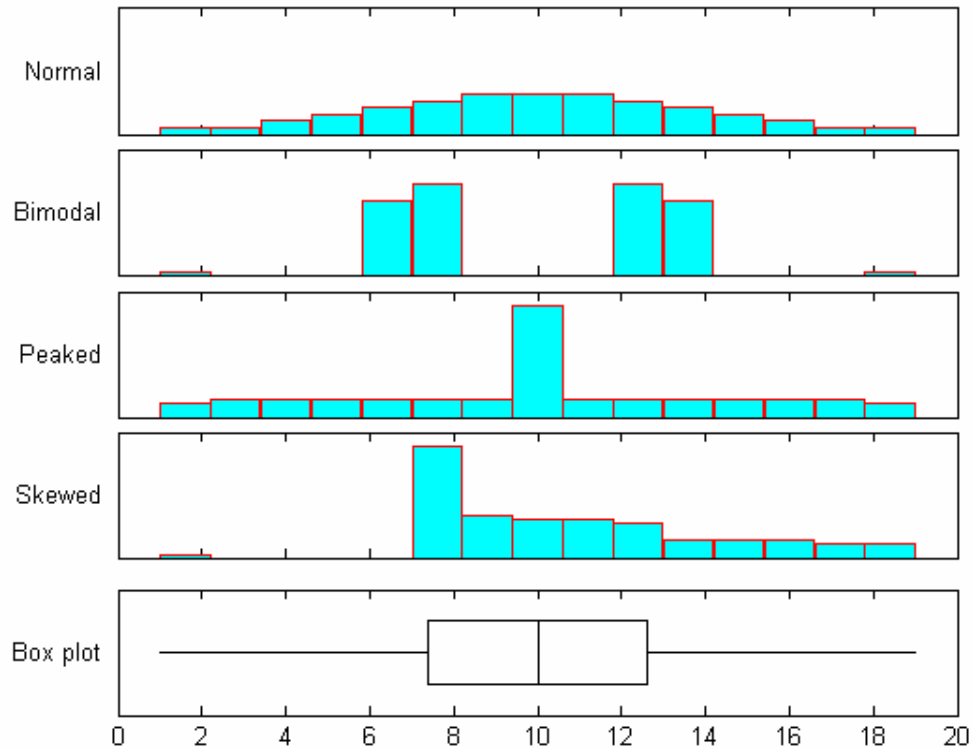

Figure 1: Histograms and box plot: four samples each of size 100

In an attempt to improve the box plot to show shape information, Benjamini (1988) suggested a "histplot", obtained by varying the width of the box according to the density of the data at the median and quartiles, where these densities are estimated from a histogram with a small number of bins. Benjamini (1988) also suggested a variation called a "vase plot", in which the linear segments in the histplot are replaced by smooth curves based on a kernel density estimate. Hintze and Nelson (1998) suggested a further modification called a "violin plot", which is essentially the same as the vase plot, except that it extends to cover the whole range of the data.

While these methods provide informative and useful displays, in essence they just replace the box plot by a kind of histogram, rather than modifying it. The problem remains to choose the extent of smoothing, which in turn should depend on the sample size. The box plot has become popular largely because of its simplicity. This raises the question: Is there a simple modification of the box plot that provides better information about the shape of the distribution, especially bimodality?

2

2. Showing skewness and kurtosis in a box plot

A possible approach is to thicken appropriate vertical lines in the box. Thus, if a distribution is right skewed, replace the edge of the box denoting the lower quartile by a thick line. If it is left skewed, thicken the edge corresponding to the upper quartile. If it is bimodal, thicken both edges. Similarly, if the distribution is peaked in the middle, thicken the line denoting the median. Figure 2 shows these possibilities for some typical samples.

An allocation rule is needed. Choonpradub (2003) did a study of viewers' choices when asked to classify sets of histograms into six classes as follows: (1) bell-shaped, (2) right-skewed, (3) left-skewed, (4) bimodal, (5) symmetric & long-tailed, or (6) other shape. The study involved 334 undergraduate and graduate students from Australia and Thailand separated into six groups, with the subjects in each group shown histograms of 16 samples with different shapes, so there were 96 samples in all. Each histogram was labeled with its sample size (50, 100 or 200).
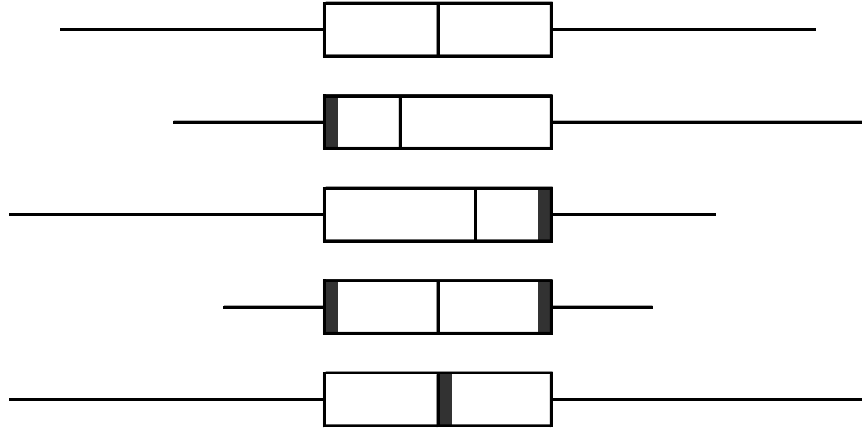

Figure 2. Box plot shapes: (from top) normal, right-skewed, left-skewed, bimodal, centrally peaked

Since bimodality corresponds to a low value of the kurtosis (scaled fourth moment), it is reasonable to use the sample skewness and kurtosis coefficients to allocate the distribution to one of the five classes. But Choonpradub's subjects placed undue attention on outliers, and she advocated the use of robust measures of skewness (γ) and kurtosis (κ), based on interquantile ranges of the sample distribution F as follows.

$$\gamma = c_1 \frac{F^{-1}(\alpha) + F^{-1}(1-\alpha) - 2F^{-1}(0.5)}{F^{-1}(1-\alpha) - F^{-1}(\alpha)}, \tag{1}$$

$$\kappa = c_2 - c_3 \frac{F^{-1}(1-\beta) - F^{-1}(\beta)}{F^{-1}(1-\alpha) - F^{-1}(\alpha)}. \tag{2}$$

The robust skewness is thus defined in terms of the extent to which the median, $F^{-1}(0.5)$, is displaced from the interval $F^{-1}(1-\alpha)$–$F^{-1}(\alpha)$ spanning the area between the two α-tails, while the robust kurtosis is a linear function of the ratio of the widths of two similar intervals with tail areas β and α, respectively, where $\beta > \alpha$. Note that if α

3

is 0.25, γ can be computed directly from the box plot because $F^{-1}(1-\alpha)-F^{-1}(\alpha)$ is then the interquartile range.

When choosing the parameter α it is important to bear in mind that box plots already show outliers quite well, as well as skewness within the central half of the distribution. These considerations dictate that α should be sufficiently large to make γ resistant against the outliers already shown, but substantially smaller than 0.25. A reasonable range might be 0.05 to 0.1. The parameter β should be at least 0.25 because the robust kurtosis should focus on peakedness or emptiness in the middle of the distribution, and to achieve this, the inner interval should be enclosed between the quartiles.

The constants c_1, c_2 and c_3 could be selected to make the robust measures agree with the conventional coefficients of skewness and kurtosis when there are no outliers. The standard outlier-free distribution is clearly the normal distribution with kurtosis 0. Also the minimum kurtosis (–2) occurs for a symmetric binary distribution. Matching these requirements, Equation (2) gives

$$c_2 = 2\frac{\Phi^{-1}(1-\beta)}{\Phi^{-1}(1-\alpha)-\Phi^{-1}(1-\beta)}, \tag{3}$$

$$c_3 = c_2+2, \tag{4}$$

where Φ is the standardized normal distribution function.

A reasonable choice for the pivotal skewed distribution might be the half-normal distribution, for which the coefficient of skewness is

$$\gamma = \frac{\left(\frac{4}{\pi}-2\right)\sqrt{\frac{2}{\pi}}}{\left(1-\frac{2}{\pi}\right)\sqrt{1-\frac{2}{\pi}}} = 0.9953.$$

Thus, using Equation (1) where F is the standardized half-normal distribution, we get

$$c_1 = 0.9953\frac{\Phi^{-1}(1-\alpha/2)-\Phi^{-1}(0.5+\alpha/2)}{\Phi^{-1}(\alpha/2)+\Phi^{-1}(1-\alpha/2)-2\Phi^{-1}(0.75)}. \tag{5}$$

For $\alpha = 0.1$ and $\beta = 0.35$, equations (3)-(5) give $c_1 = 3.587$, $c_2 = 0.860$ and $c_3 = 2.860$.

Figure 3 shows a scatter plot of the robust skewness and kurtosis coefficients for the 96 samples Choonpradub used. The plotting symbols are circles for samples seen as bell-shaped, triangles for samples perceived to be skewed, squares for samples seen as bimodal or short-tailed, and horizontal bars for samples seen as long-tailed.

The graph also shows regions that could be used to allocate samples to distributional shapes based on the robust skewness and kurtosis. Based on the subjects' allocations in Choonpradub's (2003) study the following classification rule could be used.

      1: Normal if $|\gamma| \leq 0.4$ and $|\kappa| \leq 0.2$;
      2: Centrally peaked if $\kappa > \max(0.2, |\gamma|/2)$;
      3: Right-skewed if $\gamma > 0.4$ and $-0.2 \leq \kappa < 2\gamma$;
      4: Left-skewed if $\gamma < -0.4$ and $-0.2 \leq \kappa < 2|\gamma|$;
      5: Short-tailed (possibly bimodal) if $\kappa < -0.2$.

In Figure 3 the samples misclassified by the viewers according to this rule are plotted as filled symbols or, in the case of samples seen as long-tailed, by plus signs. In the next section we examine the more discrepant anomalies in detail.

Note that the skewed sample shown in Figure 1 has robust skewness 1.40 and robust skewness –0.19, which places it only slightly above the lower boundary of the region classified as right-skewed. This suggests that the horizontal boundary between the "skewed" and "short-tailed" regions should be replaced by a flat-topped hill.
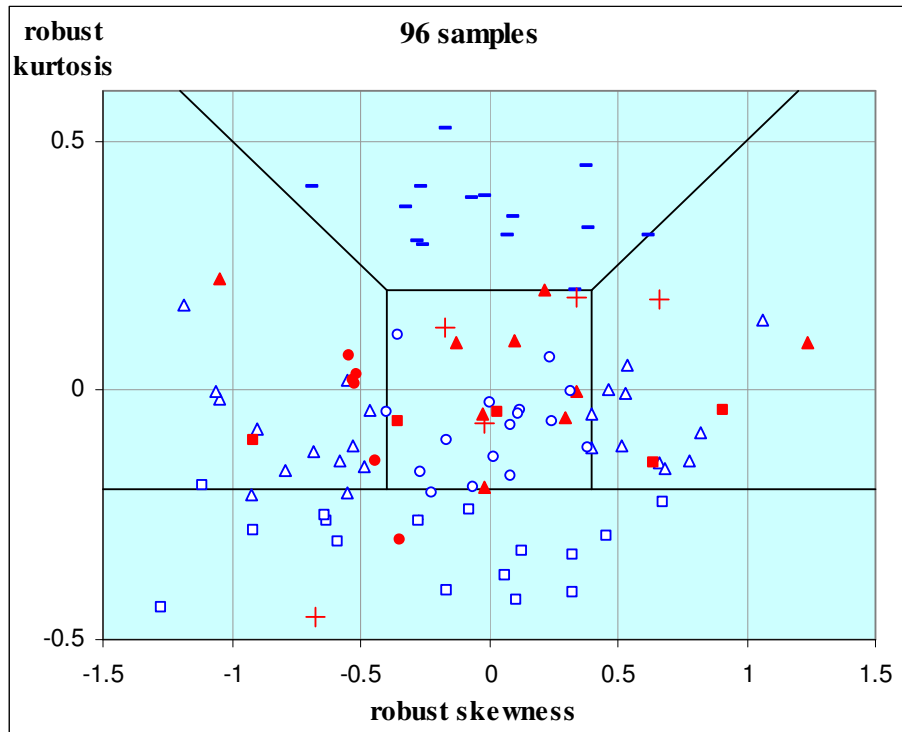

Figure 3: Robust skewness and kurtosis for 96 samples and allocation regions

3. Anomalous samples

Figure 4 shows histograms and modified box plots for the eight samples where there was greatest disagreement between the viewers' perceptions and the rule. The samples are labeled as S1 to S96, and the viewers in Group 1 were shown sample S1-S16, those in Group 2 were shown S17-S32, etc. The sample sizes are shown next to the sample identifier.

The four samples shown in the left panel (S49, S57, S61 and S38) have robust skewness and kurtosis coefficients (1.24, 0.10), (–0.34, –0.30), (–0.67, –0.45) and (–0.92, –0.10), respectively, and are thus classified as non-normal by the rule. Those graphed in the right panel (S20, S59, S60 and S72) have coefficients (0.030, –0.04), (–0.02, –0.07), (–0.03, –0.05) and (0.10, 0.10), respectively, and are thus classified as normal by the rule.

S49 (size 50) was seen by 34 viewers as left-skewed, by 17 as long-tailed, and by the remaining 6 as bell-shaped. None said it was right-skewed. It has (Fisher) coefficients of skewness and kurtosis –0.17 and 3.90. It appears that many viewers chose to ignore the outlier on the right but not the two on the left when making their decisions. The same viewers had divided opinions about the shape of S57 (size 100): 27 said it was bell-shaped, 21 left-skewed and seven bimodal, and these results are understandable, given that the sample would be classified as normal under a more liberal rule.

S61 (size 200) gave a surprising result. In this case 28 of the 57 viewers (the same ones who looked at S49 and S57) said it was long-tailed, 19 said it was bimodal, and seven saw it as bell-shaped.
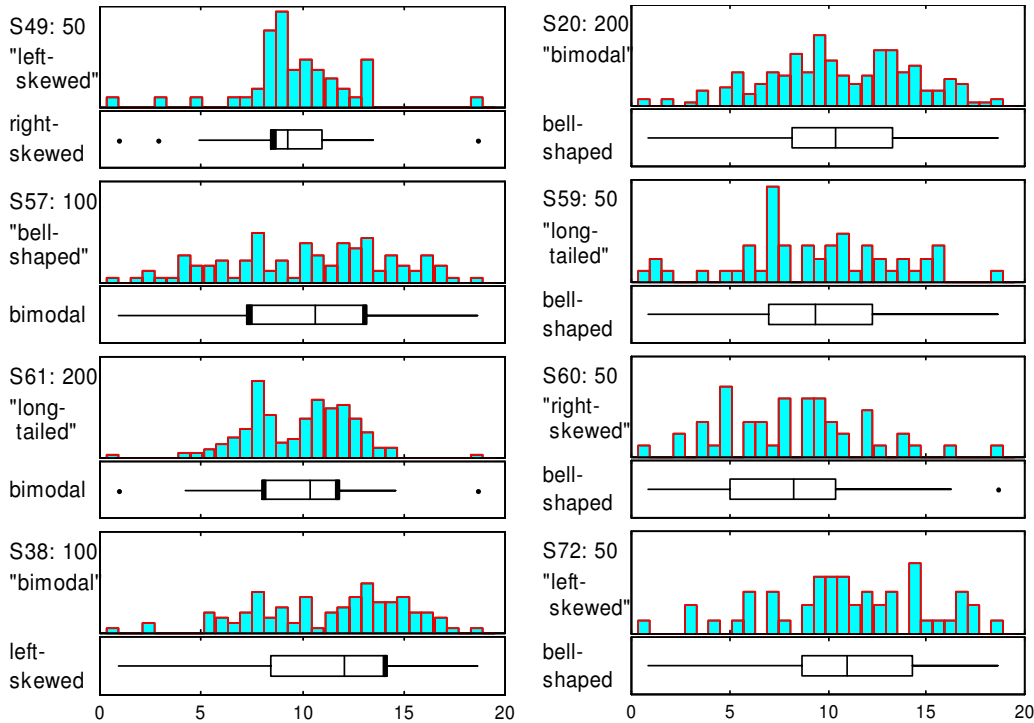
.

Figure 4: Samples where viewers saw differently to the allocation rule

The results for S38 came from a different group of viewers, and were not remarkable. Of the 53 viewers, 30 saw bimodality and the others were evenly divided between bell-shaped, left-skewed, or other shape.

The four samples graphed in the right panel of Figure 4 are all classified as normal by the rule. Experienced teachers of Statistics know that students attribute non-normal features to sampling variability, so none of these results is surprising

4. Graphing confidence intervals

Turning to modifications of the box plot to include confidence intervals, McGill, Tukey, and Larsen (1978) suggested using the width of the box to represent the sample size, and/or using notches to denote a confidence interval for the median. As an alternative to notches to represent the confidence interval, Benjamini (1988) preferred a shaded bar centered at the median.

These alternatives provide additional information to the viewer in a single graphical component, but given that box plots show spread of data and confidence intervals get shorter when sample sizes increase, for comparing several populations it seems better to separate the confidence intervals from the box plots. For example, one could show box plots of one or more samples in the upper panel of a graph with confidence intervals for the corresponding population medians (or means, if preferred) in the lower panel.

While it is useful to show confidence intervals when comparing location parameters, a confidence interval for the disparity between these parameters is more informative. For two samples, student's t-test gives a confidence interval for the difference between the population means, which can be graphed on a horizontal axis together with a vertical line passing through the origin. Thus, if the difference between the two sample means is statistically significant, the confidence interval does not intersect the vertical line, and vice versa. The graph could be placed below the conventional graph showing confidence intervals for the individual population means, using the same scale and with the point denoting the absolute value of the difference between the means aligned with the mean of the combined samples.

This graph can be extended to more than two samples, using a measure of the disparity between the means of several populations, such as the root-mean squared difference. Figure 5 shows such a comparison using four samples of blood lead levels of schoolchildren attending schools at different locations on the Pattani River (polluted due to historical tin mining near two villages and a boat repair yard near the other two, see Geater et al 2000). The blood lead concentrations were measured in micrograms/deciliter and then log-transformed. Pairs of points denoting the means are joined if the corresponding Kramer-Tukey pairwise test based on the studentized range (see, for example, Cheung and Cheng 1996) is not statistically significant.
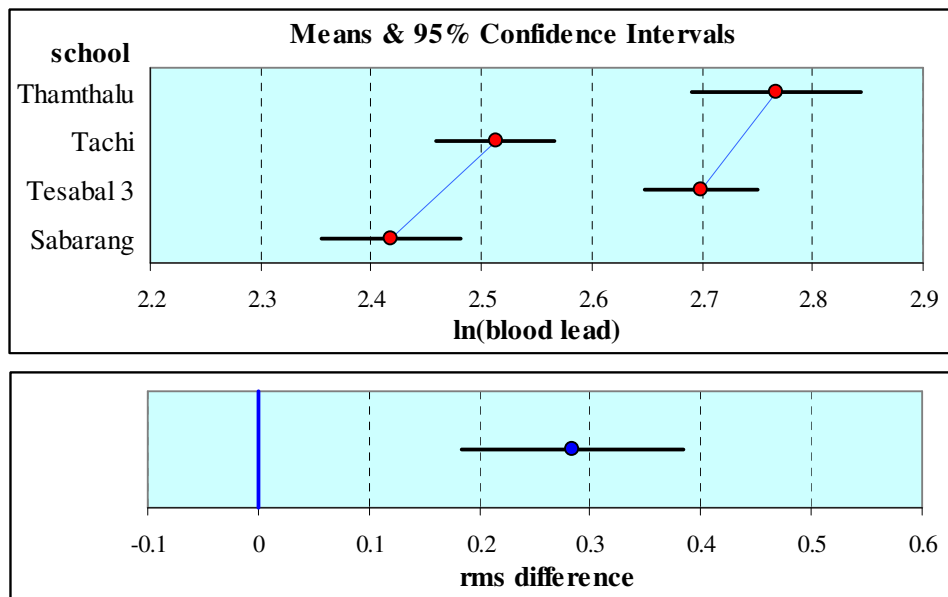

Figure 5: A graphical method showing confidence intervals for comparing means

5. Conclusions and discussion

Tukey's box plot highlights outliers and shows skewness in the central half of the distribution. By masking data between the outliers, it does not give the viewer much opportunity to be unduly swayed by sampling fluctuations. We believe that previous attempts to improve the box plot by showing shape information have not become popular for at least three reasons as follows.

First, including shape information makes the graph more complicated and more difficult to interpret, particularly if one is primarily interested in comparing several samples. Second, shape variations have relied on density estimation requiring an

additional smoothness parameter to be estimated. Third, and perhaps most important, showing too much shape information when graphing the distribution of a sample of data can mislead viewers into making erroneous conclusions.

Given that the box plot does not adequately show bimodality and central peakedness, we have suggested a very minor enhancement that does not change the essential shape of the box plot, namely, to show bimodality by thickening the ends of the box denoting the quartiles, to show central peakedness by thickening the dividing line denoting the median, and to highlight skewness by thickening just one end of the box. The criteria for making such enhancements is based on robust estimators of skewness and kurtosis that extend the statistics used to create the box plot (the median and quartiles) to include two further quantiles, and the sample size is not an issue.

While it could be argued that trying to improve the box plot is like gilding the lily, and that if more data need to be displayed histograms are adequate, there is evidence that both box plots and histograms can mislead viewers. Most statistically literate viewers, when shown a very short-whiskered symmetric box plot with no outliers, describe the underlying distribution as "symmetric" or "short-tailed" but overlook the fact that such a distribution must have two or more modes (Wainer, 1990). Choonpradub's (2003) study based on over 300 university students studying Statistics provided evidence of their inability to make correct conclusions about distributional shape from histograms.

Attempts to improve the box plot to include information about sample size (and thus, indirectly, confidence intervals of location parameters) have not been widely adopted, we believe, again because they detract from the box plot's basic simplicity. We argue that such information is more effectively shown in a separate graph, or at least in a separate panel of the graph showing the box plot(s). When several populations are to be compared, we recommend graphing confidence intervals for the means, with pairs of points denoting the sample means joined by dotted lines whenever the corresponding pairwise multiple comparison tests are not statistically significant, and with a separate confidence interval for the root-mean-squared difference in the population means.

References

Anscombe, F.J. (1973), "Graphs in Statistical Analysis," *The American Statistician*, 27(1): 19-

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

Cheung, S., and Chen, W. (1996), "Simultaneous Confidence Intervals for Pairwise Multiple Comparisons in a Two-way Unbalanced Design," *Biometrics*, 52, 463-472.

Choonpradub, C. (2003), *Improved Graphical Methods for Displaying Single and Paired Samples of Continuous Data*. PhD Dissertation. Macquarie University, Sydney, Australia.

Cleveland, W.S. (1994), *The Elements of Graphing Data (revised ed)*. AT&T Bell Laboratories. Murray Hill, NJ.

Frigge, M., Hoaglin, D.C., and Iglewicz, B. (1989), "Some Implementations of the Boxplot," *The American Statistician*, 43(1), 50-54.

Geater, A., Duerawee, M., Chompikul, J., Chairatanamanokorn, S., Pongsuwan, N., Chongsuvivatwong, V., and McNeil, D. (2000), "Blood Lead Levels among Schoolchildren living in the Pattani River Basin: Two Contamination Scenarios?" *Journal of Environmental Medicine*, 2(1), 11-16.

Hintze, J.L., and Nelson R.D. (1998), "Violin Plots: a Box Plot-Density Trace Synergism," *The American Statistician*, 52(2): 181-184.

McGill, R., Tukey, J.W., and Larsen, W.A. (1978), "Variations of Box Plots," *The American Statistician*, 32(1): 12-16.

Tufte, E.R. (1983), *The Visual Display of Quantitative Information*. Graphics Press. Cheshire, Connecticut.

Tukey, J.W. (1977), *Exploratory Data Analysis*. McGraw-Hill. Reading, MA.

Wainer, H. (1990), "Graphical visions from William Playfair to John Tukey," *Statistical Science*, 5: 340-346.