To Boxplot or Not to Boxplot?

Gary Kader and Mike Perry
Appalachian State University, North Carolina, USA.

◆ INTRODUCTION ◆

A boxplot is a mathematically simple technique for giving a graphical representation to a set of data values. It is intended to display not only the location of the data values but also the variation, with the box portion of the plot showing the variation in the middle half of the values, and the "whiskers" showing the variation in the extreme quarters. Boxplots are especially useful for showing comparisons of two or more sets of data values.

The boxplot has achieved widespread use in the mathematics curriculum, partly due to its simplicity but perhaps more so due to its applicability to a variety of statistical problems of genuine interest. There is, however, a real danger of misusing the boxplot, and some of these misuses are beginning to appear in the teaching literature. The construction of a boxplot may be mathematically naive, but appropriate application is statistically rather more sophisticated. The object of this paper is to point out a misapplication of boxplots, to suggest why the misuse occurs, and to suggest how we might adjust our teaching to correct it.

◆ WHAT IS A BOXPLOT? ◆

The basic boxplot is a graph which displays the location and ranges of each quarter of the distribution for a set of univariate data. These ranges are determined by a Five-Number Summary: Minimum, Lower Quartile, Median, Upper Quartile, and Maximum. An outlier analysis is often included as part of the boxplot representation. Texts such as Moore (1995) give details for constructing boxplots.

Professional statisticians consider the boxplot to be an informal technique and for this reason may be rather casual about definitions and terminology related to both its construction and use. Freund and Perles (1987) discuss three different commonly used methods for calculating quartiles. The method used in this article is employed by Landwehr and Watkins(1986) and Moore (1995):

The lower quartile is the median of all of the values to the left of the median of the whole set of data; the upper quartile is the median of all of the values to the right of the median of the whole set of data.

◆ WHEN ARE BOXPLOTS OKAY? ◆

Data analysts speak of using a boxplot to graphically display a *batch* of data, where a batch is a set of measurements on one variable.

The boxplot should only be used when the group of numbers being summarised consists of measurements on the same variable. Boxplots should only be used for comparisons when the groups of data being compared consist of observations from the same variable. The four boxplots in Figure 1 are appropriate for comparing the mileage efficiency of four groups of automobiles. The data consists of miles per gallon (MPG) ratings of 51 automobile models. Each of the boxplots is the summary of a set of measurements on the same variable, MPG.

Landwehr and Watkins (1986) emphasise these two cases: (1) the one group/one-variable case and (2) the many-group/one-variable situation.

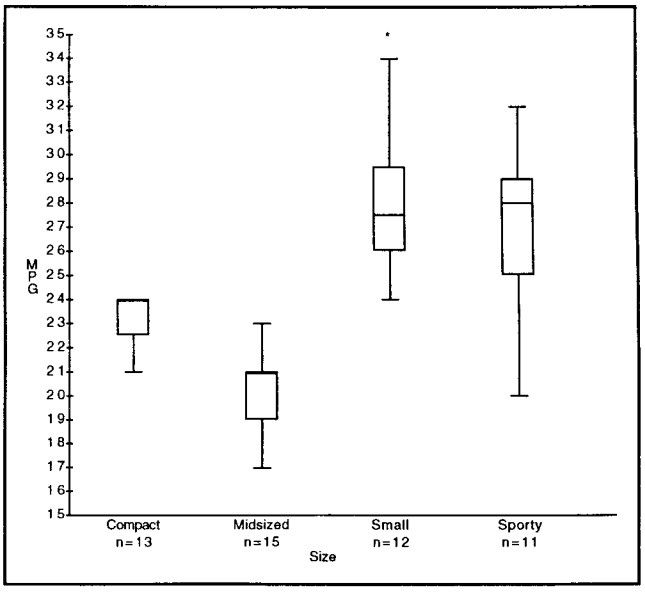

Fig. 1. Miles per Gallon by Auto Size.

Example 1

The following is an example of an improper analysis given in Hirsch (1992). A similar analysis using auto reliability data is given by Bryan (1988). Consider the data on automobiles presented in Table 1. Eight cars, all priced under $10,000, were rated by experts in each of eleven categories to determine the "best" car. Each column contains a set of observations from the same variable. Each row is a particular case and contains a rating for each of the eleven variables for one automobile. Boxplots of the eleven ratings for each car can be constructed as shown in Figure 2. Is this appropriate? All of the eleven variables use the same type of rating. In spite of this, it is not appropriate to boxplot the ratings for a particular car.

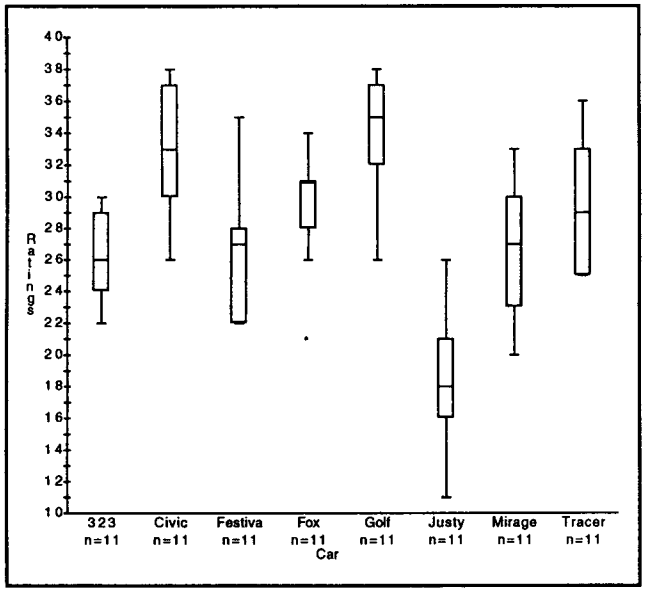

Fig. 2. Ratings of Automobiles.

The motivation for using a boxplot for each car might be to make comparisons between the cars. At first glance this seems to make sense, but a closer look will reveal a problem. To emphasise why this is an inadequate and potentially misleading representation for comparing the ratings of two cars, consider the two boxplots for the X-car and Y-car shown in Figure 3. Any reasonable interpretation would suggest that the ratings for the two cars are the same, whether you are comparing the variation of ratings or the location of ratings. These boxplots represent the hypothetical data given in Table 2. Take a closer look. The medians are the "same", but they are also "different." The medians are both equal to 20. But the median for X-car is the *Comfort* rating, and the median for Y-car is the *Fun to Drive* rating. There are similar differences for the quartiles and maximums and minimums in spite of the fact that the respective values are equal. A closer look at the ratings in Table 2 reveals that for any type of rating, there is a ten or twelve point difference between the X-car and the Y-car. For

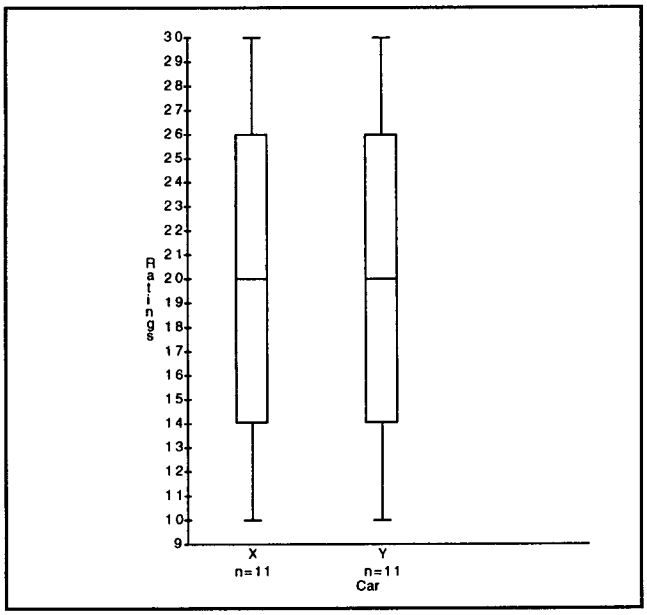

Fig. 3. Ratings of Two Automobiles.

instance, the X-car is 10 points higher on the *Style* rating, the Y-car is 12 points higher on the *Brakes* rating.

What would be an appropriate display to show the variation of ratings and to provide a comparison of two or more types of cars? The profile plots in Figure 4 give a comparison of Car 1 and Car 2 (Table 1).

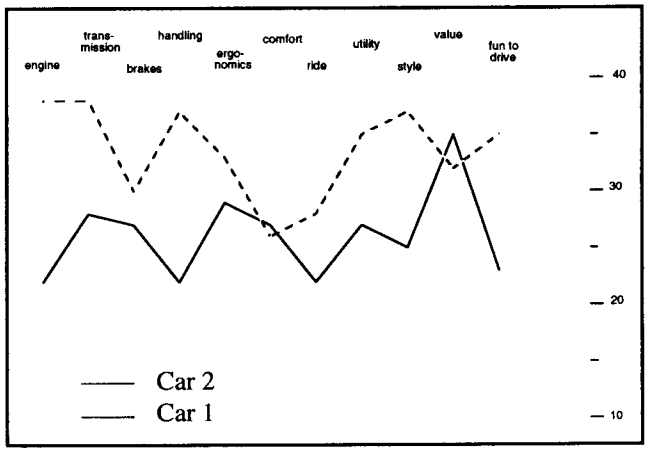

Fig. 4. Profile Plots.

Example 2

Next, consider the data on National Basketball Association (NBA) players presented in Table 3. Data of this type are examined in the article by Hilgert (1995) which updates a similar activity described in Burrill (1988). They address the question "Who is the best center in the NBA?" Their data consists of subjective rankings on five characteristics (aggressiveness, shooting range, teamwork, offence, and defence) for ten players. The rankings are on a 1 to 10 scale, with 1 being the best. Two of the players' characteristics and their ratings are given in Table 3. They suggest comparing the 10 players based on the boxplot for each player's 5 scores (Figure 5). As before, it is not proper to boxplot a case, and a profile plot as used in Example 1 would be useful for making the desired comparisons.

Table 1. Ratings of Automobiles.

Car	Engine	Transmission	Brakes	Handling	Ergonomics	Comfort	Ride	Utility	Style	Value	Fun to Drive
1	22	28	27	22	29	27	22	27	25	35	23
2	38	38	30	37	33	26	28	33	37	32	35
3	28	26	22	26	30	28	30	29	23	26	24
4	27	29	25	25	32	36	35	33	28	33	25
5	27	33	25	20	30	27	29	31	29	23	23
6	13	21	21	17	23	18	16	26	18	21	11
7	26	33	29	34	31	31	31	31	21	28	29
8	35	29	32	38	37	34	32	37	26	35	36

Table 2. Ratings of Two Automobiles.

Car	Engine	Transmission	Brakes	Handling	Ergonomics	Comfort	Ride	Utility	Style	Value	Fun to Drive
X-Car	10	12	14	16	18	20	22	24	26	28	30
Y-Car	22	24	26	28	30	10	12	14	16	18	20

Table 3. Ratings of NBA Centers.

			Ranks		
Player	AG	SR	TW	OF	DF
Brad Daugherty (Cleveland Cavaliers)	8	5	7	5	9
Patrick Ewing (New York Knicks)	5	3	1	4	5

◆ THE ROLE ◆ OF COMPUTER SOFTWARE

Data analysis software encourages you to think about the structure of the data, especially if you enter the data yourself. For instance, *MINITAB* handles data as a table (matrix). The user is encouraged (Farber and Schaefer, 1992) to use the columns of the table for values of the same variable, the rows as the cases. Once you have entered your data, you can only produce boxplots of the columns. To produce two or more boxplots together for the purpose of comparison with *MINITAB* you must put all of the data in the same column and code the groups in another column. Thus you are only allowed to compare boxplots which are summaries of observations from the same column.

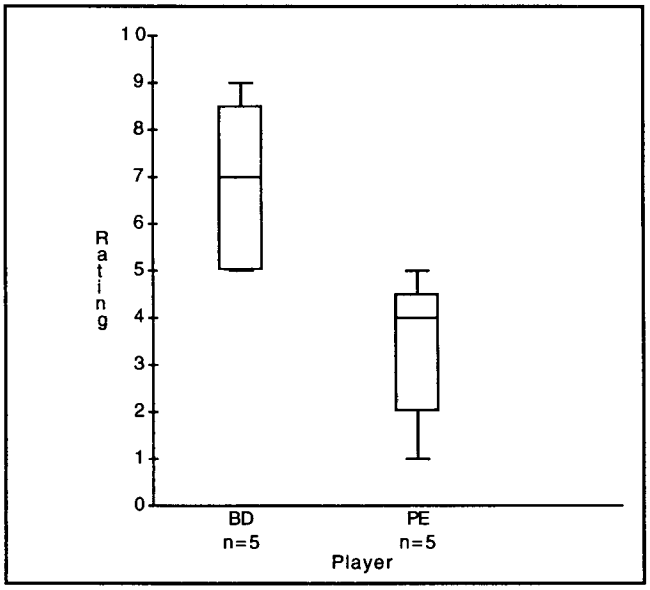

Fig. 5. Comparison of NBA Players.

◆ CONCLUSION ◆

How do we respond to this potential misapplication in our teaching? We must present data analysis as a whole problem solving process. This process includes a properly posed question, relevant data, appropriate analysis, and careful interpretation of results. Appropriate analysis of data depends on a full understanding of the data. As teachers of statistics, we should put as much, or more, emphasis on the data as we do on the analysis techniques. This should include an understanding of how the data is collected and the nature of the measurement being used. We must also emphasise the variables being observed and the structure of the data. Only then can we expect to avoid the types of misapplications of techniques that have been discussed here.

References

Bryan, E.H. (1988). "Exploring Data with Box Plots". Mathematics Teacher, **81**(8), National Council of Teachers of Mathematics: Reston, Virginia.

Burrill, G. (1988). *Statistical Decision Making.* NCTM Student Math Notes, National Council of Teachers of Mathematics: Reston, Virginia.

Farber, E. and Schaefer, R.L. (1992). *User's Manual for the Student Edition of MINITAB.* Addison-Wesley Pub. Co.: New York.

Freund, J.E. and Perles, B.M. (1987). "A New Look at Quartiles of Ungrouped Data". The American Statistician, **41**(3), American Statistical Association: Alexandria, Virginia.

Hilgert, F. (1995). "Who is the Best Center in the NBA?" Wisconsin Teacher of Mathematics, **46**(1), Wisconsin Mathematics Council: Milwaukee.

Hirsch, C., series editor (1992). *Curriculum and Evaluation Standards for School Mathematics / Addenda Series Grades 9-12.* National Council of Teachers of Mathematics: Reston, Virginia.

Landwehr, J.M. and Watkins, A.E (1986). *Exploring Data.* The Quantitative Literacy Series, Dale Seymour Publications: Palo Alto, California.

Moore, D.S. (1995). *The Basic Practice of Statistics.* W.H. Freeman & Co.: New York.

Kader, Gary and Perry, Mike (Summer, 1996). To Boxplot or Not to Boxplot? *Teaching Statistics*, 18 (2), 39-41.