PM R 8 (2016) 268-272

www.pmrjournal.org

Statistically Speaking

The Box Plots Alternative for Visualizing Quantitative Data

Regina L. Nuzzo, PhD

Introduction

Although bar charts are popular among researchers and are ubiquitous in quantitative software packages, they do not always provide the best visualization for a dataset. This column discusses a simple, alternative graphical method that is often underappreciated: the box plot, also known as the box-and-whisker plot. The basic elements of the box plot are presented, along with how to correctly interpret box plots, variations that are available to provide more information, and free online software that researchers can use to create box plots for publication.

Weaknesses of Using Bar Charts

A bar chart is one of the simplest of all data visualizations and is included in every quantitative software package. The bar chart is a good method for summarizing counts or proportions with categorical data, yet it is not always the best option for summarizing or comparing numeric responses in samples. For example, suppose that a clinical team wanted to summarize the Berg Balance Scale scores of 3 groups of patients before and after a therapy, and again after 1 year (Figure 1). Using a bar chart for this display results in several potential problems or weaknesses:

1. The value of interest is the position of the mean, which is represented in the bar chart only with the top line of the bar. The bar itself is unnecessary and conveys no information, which can be considered a waste of space and ink [1,2].
2. The bar leads the viewer's eye to believe that length is an important dimension. Most bar charts start the vertical axis at 0, yet this choice is arbitrary and often misleading, because the sample data may not necessarily include 0. In fact, the minimum value in the sample may be negative, or it may be orders of magnitude greater than 0, and thus the length of the bar is purely arbitrary.
3. Bar charts display the sample mean for a set of data, yet the mean is not a robust summary of the "average" central tendency of a population. The mean is highly sensitive to extreme values, and so for skewed data or in the presence of outliers, the mean may not lie near the center bulk of the data. When working with small samples, samples with outliers, or populations that are not known to be symmetric, a better measure of average is the median, which is not displayed in bar charts.
4. Error bars are often added to bar charts, but it has been shown that viewers have a difficult time judging and interpreting overlaps in error bars [3]. Plots are not always clearly labeled to convey whether the error bar represents one standard deviation, one standard error of the mean, or one half-width of a 95% confidence interval. Furthermore, confidence intervals require assumptions about the data and can be misleading for small samples or skewed distributions [4].

Box Plots: Styles, Strengths, and Value

Princeton statistician John Tukey designed the box plot as an easy-to-draw data visualization as part of exploratory data analysis [5]. The box plot has persisted into the computer age as an information-rich graphic that conveys key features of a numeric dataset at a glance. Unlike the bar chart, it uses statistical summaries (median and interquartile range) that are robust in the presence of skewness and outliers and require no assumptions about the population. It shows the full range of the sample data, provides information about the tails, and indicates the shape of the data. It can be used for samples as small as $n = 5$ and allows for quick side-by-side comparisons between groups.

The components of a box plot are as follows (Figure 2):

1. The box plot divides the sample data into fourths, or quartiles: 2 box panels and 2 whiskers (plus any outlying values beyond the whiskers).

1934-1482/$ - see front matter
http://dx.doi.org/10.1016/j.pmrj.2016.02.001

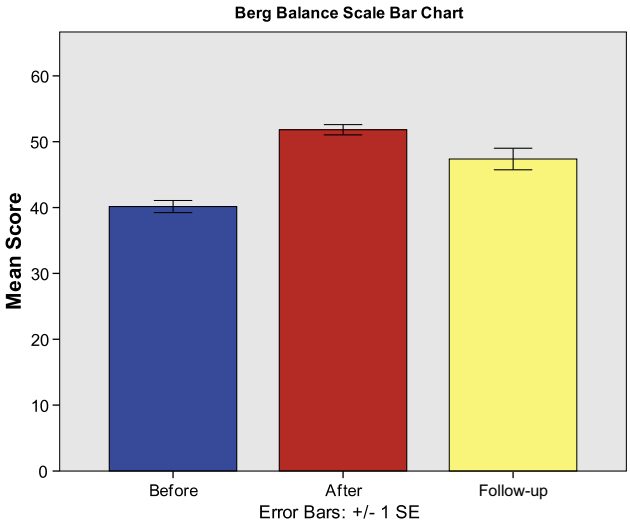

Figure 1. Typical bar chart for the balance scores of a group of patients before therapy, after therapy, and at 1-year follow-up. Mean scores for each group are shown with the top line of the bar only. Error bars show 1 standard error (SE) of the mean. N = 80, 40, and 10 as a result of patients lost to follow-up.

2. The box spans the middle 50% of the data. The outer edges of the box, often called the "hinges," indicate the first quartile (the 25th percentile, or the value at which 25% of the data fall below) and the third quartile (the 75th percentile, or the value at which 25% of the data fall above).

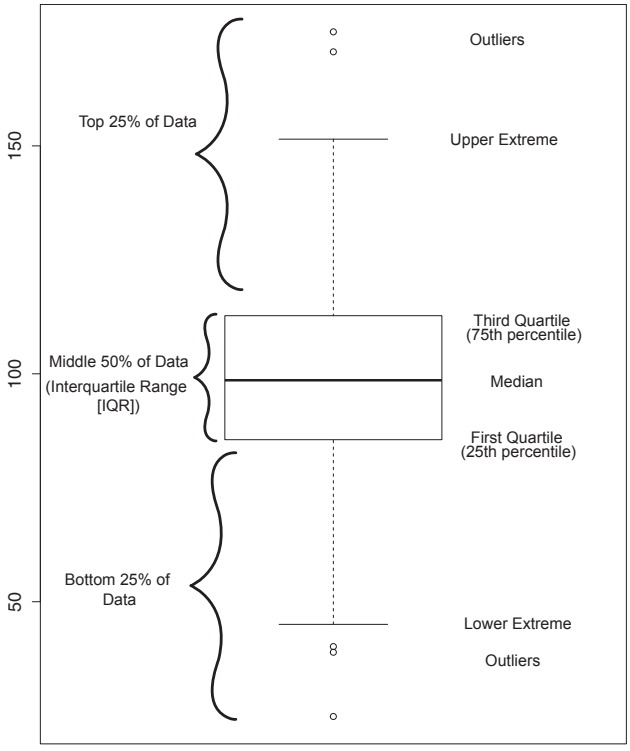

Figure 2. Annotated box plot of 1000 points from a normal distribution with a mean of 100 and a standard deviation of 20.

3. The middle line of the box indicates the median (the second quartile, or the 50th percentile).
4. The length of the box is the interquartile range (IQR), which is a measure of spread similar to the standard deviation.
5. The whiskers show the extent of the data range for the other 50% of the data. The whiskers start at the edges of the center bulk (the first and third quartiles) and extend to what is considered "extreme" values in the data, typically taken to be a distance of 1.5 IQR beyond the first and third quartiles, although other variations are possible (described in a subsequent section).
6. Data values beyond the extremes are considered outliers or potential outliers and are marked as individual points.

Box Plot Demonstration Using Dataset From Figure 1

The data in Figure 1 yield more information when expressed as box plots (Figure 3). Here the viewer can compare the medians among the groups, and other information is also apparent. Patients before therapy had a median balance score of about 39, which rose to about 53 after therapy and dropped to about 48 a year later. The positive skew of the patients' scores before therapy can be seen from the long right (upper) whisker, revealing that whereas most patients initially tended to score fairly low, a few (including 2 high outliers) scored much higher. Immediately after therapy, the patients' scores were much higher, showing almost no overlap with scores before therapy. The data had a negative skew, seen in the longer left (lower) whisker and the one moderately low outlier. The maximum possible score of the Berg Balance Scale is 56, would explain the ceiling effect in this group of data. One year after

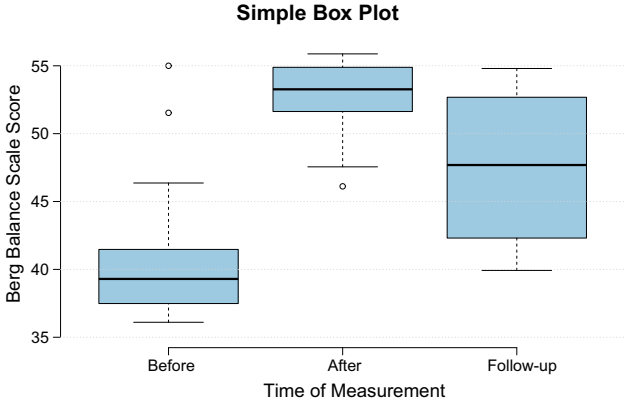

Figure 3. Simple side-by-side box plots for the data in Figure 1. Center lines show the medians, box limits indicate the 25th and 75th percentiles as determined by R software, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, and outliers are represented by dots. n = 80, 40, and 10 sample points.