Box Plots

In 1977, John Tukey published an efficient method for displaying a five-number data summary. The graph is called a boxplot (also known as a box and whisker plot).

In descriptive statistics, a boxplot (also known as a box-and-whisker diagram or plot) is a convenient way of graphically depicting groups of numerical data through their five-number summaries (the smallest observation, lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation). A boxplot may also indicate which observations, if any, might be considered outliers. The boxplot was invented in 1977 by the American statistician John Tukey.

Boxplots can be useful to display differences between populations without making any assumptions of the underlying statistical distribution. The spacings between the different parts of the box help indicate the degree of dispersion (spread) and skewness in the data, and identify outliers. Boxplots can be drawn either horizontally or vertically.

Construction

For a data set, one constructs a horizontal box plot in the following manner:

1. Calculate the first quartile (x.25), the median (x.50) and third quartile (x.75)
2. Calculate the interquartile range (IQR) by subtracting the first quartile from the third quartile. (x.75 − x.25)
3. Construct a box above the number line bounded on the left by the first quartile (x.25) and on the right by the third quartile (x.75).
4. Indicate where the median lies inside of the box with the presence of a symbol or a line dividing the box at the median value.
5. The mean value of the data can also be labeled with a point.

Any data observation which lies more than lower than the first quartile or higher than the third quartile is considered an outlier. Indicate where the smallest value that is not an outlier is by connecting it to the box with a horizontal line or "whisker". Optionally, also mark the position of this value more clearly using a small vertical line. Likewise, connect the largest value that is not an outlier to the box by a "whisker" (and optionally mark it with another small vertical line).

Advantages of Boxplots

1. Graphically display a variable's location and spread at a glance.
2. Provide some indication of the data's symmetry and skewness.
3. Unlike many other methods of data display, boxplots show outliers.
4. By using a boxplot for each categorical variable side-by-side on the same graph, one quickly can compare data sets.

Disadvantages of Boxplots

One drawback of boxplots is that they tend to emphasize the tails of a distribution, which are the least certain points in the data set. They also hide many of the details of the distribution. Displaying a histogram in conjunction with the boxplot helps in this regard, and both are important tools for exploratory data analysis.

The boxplot is interpreted as follows:

The box itself contains the middle 50% of the data. The upper edge (hinge) of the box indicates the 75th percentile of the data set, and the lower hinge indicates the 25th percentile. The range of the middle two quartiles is known as the inter-quartile range.

The line in the box indicates the median value of the data.

If the median line within the box is not equidistant from the hinges, then the data is skewed.

The ends of the vertical lines or "whiskers" indicate the minimum and maximum data values, unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range.

The points outside the ends of the whiskers are outliers or suspected outliers.
Boxplot Enhancements

Beyond the basic information, boxplots sometimes are enhanced to convey additional information:

The mean and its confidence interval can be shown using a diamond shape in the box.

The expected range of the median can be shown using notches in the box.

Indicate outliers by open and closed dots. "Extreme" outliers, or those which lie more than three times the IQR () to the left and right from the first and third quartiles respectively, are indicated by the presence of a closed dot. "Mild" outliers - that is, those observations which lie more than 1.5 times the IQR from the first and third quartile but are not also extreme outliers are indicated by the presence of a open dot. (Sometimes no distinction is made between "mild" and "extreme" outliers.)

Add an appropriate label to the number line and title the boxplot.

A boxplot may be constructed in a similar manner vertically as opposed to horizontally by merely interchanging "bottom" for "left", "top" for "right" and "vertical" for "horizontal" in the above description.