A box-and-whisker plot, often referred to as a box plot, was developed by John Tukey. It is a convenient graphic tool in descriptive analysis to display a group or groups of numerical data through their medians, means, quartiles, and minimum and maximum observations. A box plot is useful to display the distribution of data, examine symmetry, and indicate potential outliers and can also be used to compare parallel groups of data.

There are two typical types of box plots: skeletal and schematic and the latter one was described by Tukey in the 1977 book. The box is identical for both types of box plots. A box is constructed to represent 50% of the data: the upper boundary of the box representing the upper quartile (Q3) of the data and lower boundary representing the lower quartile (Q1). The length of the box presents the interquartile range (IQR = Q3 – Q1). The median is identified by a horizontal line dawn inside the box and the mean is marked as a plus (+) symbol (Fig. 1).

The whisker part differs between two types of plots. In a skeletal box plot, whiskers are drawn from the upper and lower edges of the box to the maximum and minimum values of the data, respectively. However, in a schematic box plot, whiskers are drawn from the upper and lower edges of the box to the largest and smallest observations, respectively, within the upper and lower fence. The upper fence is defined as 1.5*IQR higher than Q3 and the lower fence is defined as 1.5*IQR lower than Q1. The fences are not drawn in plots. Observations outside the fences are regarded as outliers, which are labeled with little squares.

Box plots are available in many statistical software packages, such as SAS (SAS Institute, Inc., Cary, NC), S-PLUS (Insightful Corp., Seattle, WA), and SPSS (SPSS, Inc., Chicago, IL). However, there is no a standard way for computing quartiles and fences. Different computation methods are used with packages, and even several methods are used within a package. For example, SAS offers five definitions for computing quartile statistics. Details about these definitions are beyond the content of this article.

A box plot is a quick tool to study the distribution of a group of numeric data. One can easily observe locations of quartiles, mean, median, and extreme values. Normality of the data can also be examined. A box plot for normally distributed data should be symmetric, which means the mean is close to the median line, the median line roughly evenly divides the box, and length of the two whiskers are roughly equal. For a nonnormally distribution, skewness can be detected from a box plot. If the median line deviates from the center of the box downward, and the upper whisker is longer than the lower one, that is, the right tail is longer and the distribution is right skewed; a left-skewed distribution is present in the opposite way. The box plot in Fig. 1 indicates that the data are not normally distributed but left skewed. In addition, a schematic box plot will provide detailed information about outliers.

Another particular advantage of box plots is to easily compare distributions of many parallel groups of data through side-by-side box plots without making any distribution assumption. Figure 2 is an example of the two-type box plots for real data of maximum aggregation results of platelets (PLTs) among six groups of individuals. Examining these plots one can immediately see that the data are not normally distributed, the distributions of the results vary between some groups, and the variability within groups is also present. A box takes small space, so box plots have a big advantage to compare many populations in a plot over other distribution graphs such as histograms and estimated density function plots.

A researcher can make box plots using a statistical package or draw a box plot himself with computing the statistics by Excel (Microsoft Corp., Redmond, WA). The plotting job is small but the amount of information displayed is large.
Fig. 1. Summary of the information provided by skeletal and schematic box plots.

Fig. 2. Examples of the skeletal (left) and schematic (right) box plots showing PLT aggregation data among six different groups.

REFERENCES
