SEJODR

StaTips Part VII: Anatomy of a Boxplot

Perinetti, Giuseppe *

* *Private practice, Nocciano (PE), Italy*

ABSTRACT

Often researchers have the need to show graphically the behaviour of ordinal data or continuous data with great asymmetrical distribution. In all of these situations, a boxplot may be an appropriate way of data presentation. Because of its non-parametric nature, this kind of diagram is particularly useful when dealing with asymmetrical distribution. The different components of the most common variant of boxplot are herein detailed.

Perinetti G. StaTips Part VII: Anatomy of a Boxplot. South Eur J Orthod Dentofac Res. 2019;6(2):25-26.

FRAMING OF THE PROBLEM

Often researchers have the need to show graphically the behaviour of ordinal data or continuous data with great asymmetrical distribution. These situations make the mean and standard deviation poorly descriptive of the data sets from which they are derived. A typical example is represented by microbiologic data, i.e. bacterial counts, where data sets may have values falling within several orders of magnitude. In all of these situations, a boxplot may be an appropriate way of data presentation.

COMPONENTS OF A BOXPLOT

A boxplot (also referred to as box-and-whisker plot) is a histogram-like graphical representation of a data set through interquartile ranges (IQRs), initially proposed by Tukey.[1] Because of its non-parametric nature, this kind of diagram is particularly useful when dealing with asymmetrical distribution (eventually on a logarithmic scale). Most common statistical packages include functions to construct a boxplot.

Among the several variants of boxplots (even for 2D data), the most common features of a 1D boxplot (Figure 1) are detailed as follows:

1. The bottom of the box indicates the 25th percentile (25% of cases have values below this threshold). The top of the box represents the 75th percentile (25% of cases have values above this threshold). As a consequence, 50% of cases have values that fall within the box. The bottom and top of the box are also referred to as 'hinges'.
2. The line within the box represents the median value of the whole data set. As a consequence, cases are equally distributed above and below this line.
3. The cross within the box, occasionally reported, represents the mean of the whole data set.
4. The T-bars that extend from the boxes are referred to as 'inner fences' or 'whiskers'. These extends cover up to ±1.5x IQR (which is also 1.5x height of the box). Sometimes, if no value fall within the T-bars ranges, inner fences may represent the minimum and maximum values of the whole data set. More the data distribution is symmetrical, more cases fall within the T-bars ranges.
5. The circles (or points) are referred to as 'outlier' values. These are defined as values that fall beyond the T-bar ranges but are also within the ±3x IQR range and are plotted as individual values.
6. The asterisks are referred to as 'extreme' values. These are defined as values that are beyond the ±3x IQR range and are plotted as individual values.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest regarding the publication of this paper.

Corresponding Author:
Perinetti Giuseppe
Via San Lorenzo 69/1,
65010 Nocciano (PE), Italy.
e-mail: G.Perinetti@yahoo.com

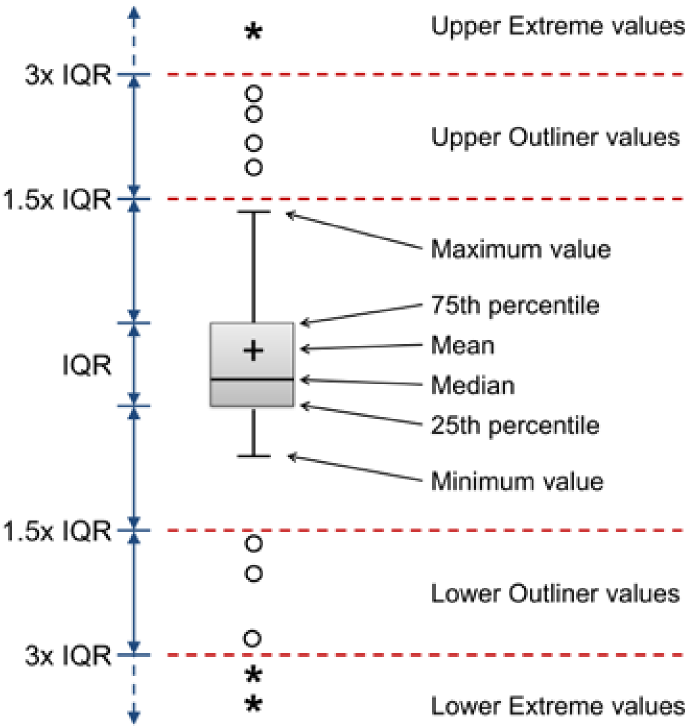

Figure 1. *Diagram illustrating the components of typical boxplot (see text for details).*

REFERENCE

1. Tukey JW. Box-and-Whisker Plots. Exploratory Data Analysis. Reading, MA: Addison Wesley; 1977. p. 39-43.