JUST AGRICULTURE
multidisciplinary e-Newsletter

Box and Whisker Plot: An Outline

S. Vishnu Shankar
Research scholar, Department of Basic Sciences, Dr. Y. S. Parmar University of Horticulture & Forestry, Solan, Himachal Pradesh - 173 230

ARTICLE ID: 11

Box plots

A box plot, also known as a box-and-whisker plot, was first given by John Turkey in 1977. It is a graphical representation of a dataset that provides a summary of its distribution and key statistical measures. It shows the minimum, first quartile, median, third quartile, and maximum values of the data. The data should be in metric scale to perform the box plot. The lines extending parallel from the boxes are known as the "whiskers", which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally. They are widely used in data analysis and statistical visualization for various purposes.

Common uses of box plots:

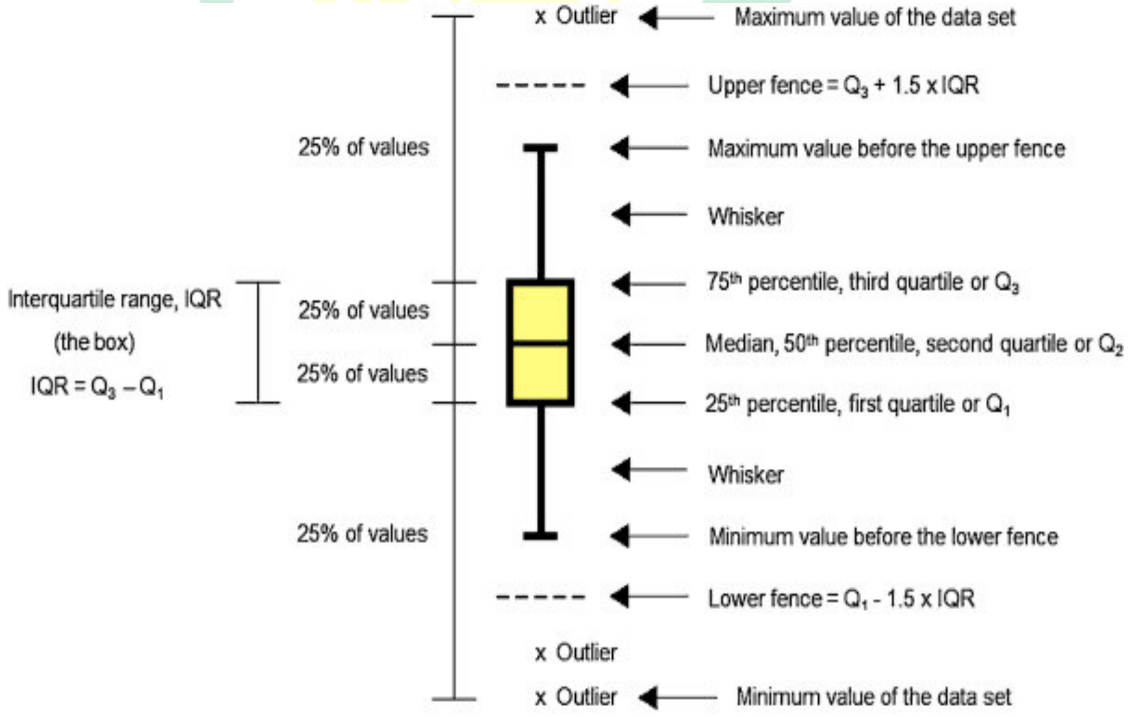
x Outlier
Maximum value of the data set
Upper fence = Q_3 + 1.5 x IQR
25% of values
Maximum value before the upper fence
Whisker
Interquartile range, IQR
(the box)
IQR = $Q_3 - Q_1$
25% of values
25% of values
75th percentile, third quartile or Q_3
Median, 50th percentile, second quartile or Q_2
25th percentile, first quartile or Q_1
Whisker
25% of values
Minimum value before the lower fence
Lower fence = Q_1 - 1.5 x IQR
x Outlier
x Outlier
Minimum value of the data set

Fig 1. Anatomy of Box and Whisker Plot

- **Summary of data distribution:** Box plots provide a concise summary of the distribution of a dataset, including measures such as the median, quartiles, and range. They allow you to quickly understand the spread and central tendency of the data.
- **Comparison of groups or categories:** Box plots are useful for comparing the distributions of multiple groups or categories. By placing multiple box plots side by side, you can visually compare their medians, quartiles, and ranges. This helps identify differences or similarities between groups and detect any outliers.
- **Identification of outliers:** Box plots are effective for identifying outliers, which are data points that significantly deviate from the rest of the dataset. Outliers appear as individual points outside the whiskers, making them easy to spot and investigate further. Removing or addressing outliers appropriately can improve the accuracy and reliability of statistical analysis.
- **Detection of skewness and symmetry:** Box plots can indicate the skewness (asymmetry) of a distribution. If one whisker is noticeably longer than the other, or if the median is not centred in the box, it suggests a skewed distribution. Symmetric distributions have approximately equal whisker lengths and a median in the center of the box.
- **Comparison of data across time or conditions:** Box plots can be used to compare data across different time periods, experimental conditions, or any other categorical variable. By plotting the box plots together, you can observe changes in the distribution over time or differences between conditions.
- **Assessment of data variability:** Box plots provide insights into the spread of the data by showing the interquartile range (IQR), which represents the middle 50% of the data. A narrow IQR indicates low variability, while a wide IQR suggests high variability.

Types of box plots

Notched box plots and Variable-width box plots are the two types of box plots that show variations of the traditional box plot by providing additional information about the dataset.

1. **Notched box plots:** Notched box plots incorporate additional information by adding notches to the boxes. The notches provide a graphical representation of the confidence interval around the median. The width of the notches corresponds to the variability or uncertainty in the median estimation. The notches in a box plot are constructed using a method called the Wilk's confidence interval. If the notches of the two box plots do not overlap, it suggests that there is a significant difference between the medians of the two groups at a chosen significance level (typically 95%)

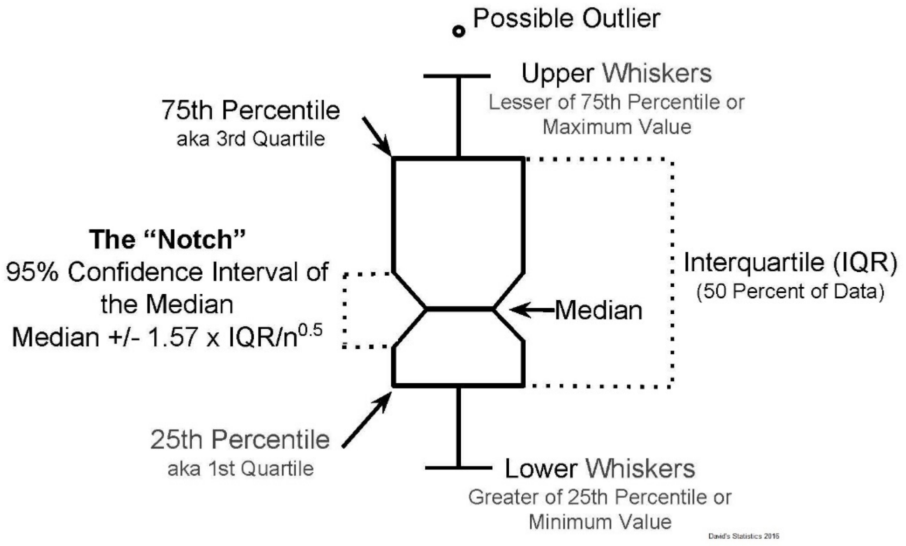

Fig 2. Anatomy of Notched Box plot

2. **Variable-width box plots:** In a traditional box plot, the width of the box is typically constant across all categories or groups being compared. However, in variable-width box plots, the width of the box is proportional to the sample size or the frequency of data points in each category. By using variable-width boxes, these plots allow for a more accurate visual representation of the distribution, considering the varying sample sizes. This can be particularly useful when comparing groups with significantly different sample sizes, as it emphasizes the precision of estimates based on larger samples. The layout of variable-width box plots will be similar to conventional box plots.

Constraints in box plots

- **Ignores individual data points:** Box plots primarily focus on summary statistics and do not display individual data points within each group.

- **Limited information on shape:** Although box plots provide information about quartiles and median, they do not provide a detailed picture of the shape of the distribution. They can miss important details about modes, multimodality, or other nuances in the data.
- **May hide small sample sizes:** When constructing box plots, the sample size of each group may not be evident. If the sample size is small, the variability and uncertainty of the estimate may not be apparent from the plot alone.
- **Not suitable for certain data types:** Box plots are most useful for numerical data. They are less suitable for categorical or ordinal data, as they don't provide an intuitive representation of the distribution for such variables.