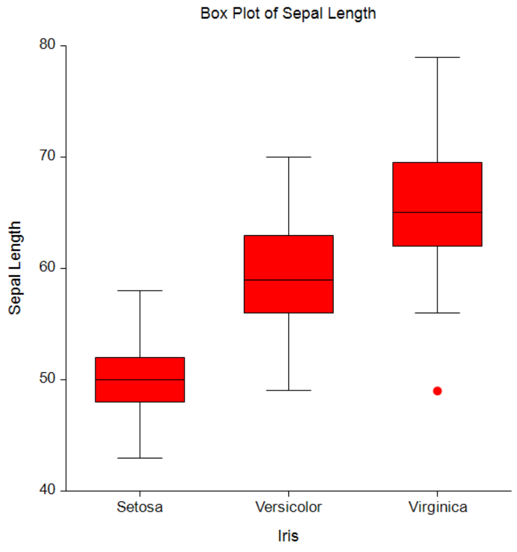
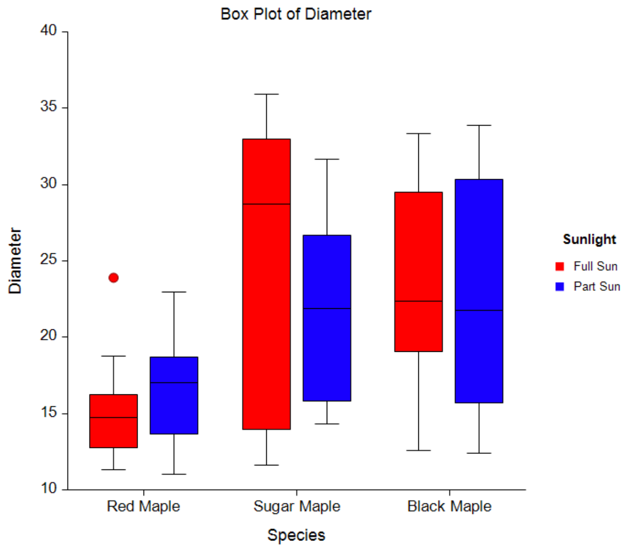
Chapter 152

Box Plots

Introduction

When analyzing data, you often need to study the characteristics of a single group of numbers, observations, or measurements. You might want to know the center and the spread about this central value. You might want to investigate extreme values (referred to as outliers) or study the distribution or pattern of the data values. Several plots are available to allow you to study the distribution. One such plot is the box plot.





Box Plot Definition

The box plot is defined by five data-summary values and also shows the outliers.

Median and Box

The box portion of the box plot is defined by two lines at the 25th percentile and 75th percentile. The 25th percentile is the value at which 25% of the data values are below this value. Thus, the middle 50% of the data values fall between the 25th percentile and the 75th percentile. The distance between the upper (75th percentile) and lower (25th percentile) lines of the box is called the inter-quartile range (IQR). IQR is a popular measure of spread.

A line is drawn inside the box at the median (the 50th percentile). The median is a popular measure of the variable’s location (center).

Whisker and Outlier Boundaries

A box plot whisker is a line that goes out from the box to the whisker boundaries. Often a crossbar line is drawn at the whisker boundary. Points outside the whisker boundaries are considered outliers. An additional boundary is sometimes used for severe outliers, although there is no line drawn at the severe outlier boundaries.

In **NCSS** there are two ways to define these boundaries. One way uses a multiplier of the inter-quartile range. The other uses percentiles.

Boundaries using the Inter-Quartile Range (IQR)

This is the traditional method for determining the boundaries. In this method, the whisker boundary is found by multiplying a value (usually 1.5) times the IQR, and then going out that distance from the edge of the box. The whisker boundary is then brought back in to the first data value that is reached. In technical terms (for the upper whisker boundary), it is the largest observation that is less than or equal to the upper edge of the box plus the multiplier times IQR.

The severe outlier boundary is defined similarly, but the multiplier is larger (usually 3).

Boundaries using Percentiles

The whisker boundary may also be defined in terms of percentiles, similarly to the edges of the box. For example, the two whisker boundaries might be the 10th percentile and 90th percentile (or 5th and 95th).

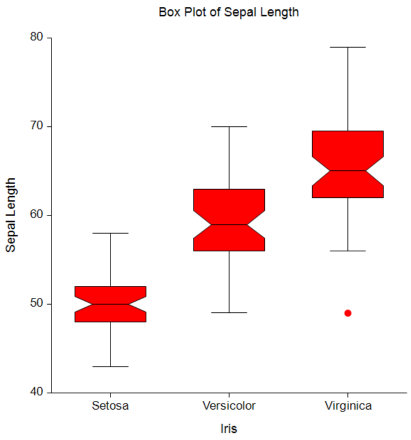
The severe outlier boundaries might be the 1st and 99th percentiles, for example.

Multiple Comparisons

Box plots are often used for comparing the distributions of several groups of data, since they summarize the center and spread of the data very nicely. When making comparisons among the locations (medians) of various batches, a modified box plot called the *notched box plot* is useful. The notches are constructed using the formula:

$$Median \pm 1.57 \times (IQR)/\sqrt{n}$$

Notched box plots are used to make multiple comparisons among the batches. If the notches of two boxes do not overlap, we may assume that the medians are significantly different (the centers are statistically significant). The 1.57 is selected for the 95% level of significance. The box plot below is an example of a notched box plot.



Note that when making comparisons among several batches, the notched box plots do not make any adjustment for the multiplicity of tests being conducted. Numeric testing with multiple comparison adjustments is recommended when multiple tests occur.

Data Structure

A box plot is constructed from a numeric variable. A second variable may be used to divide the first variable into groups (e.g., age group or gender). In the two-factor procedure, a third variable may be used to divide the groups into subgroups.

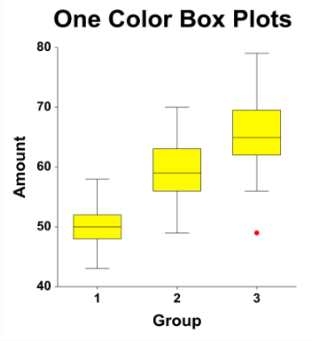
Box Plot Window Options

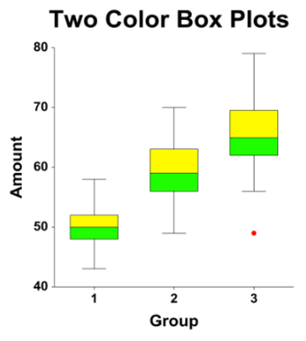
This section describes the specific options available on the Box Plot window, which is displayed when the Box Plot button is clicked. Common options, such as axes, labels, legends, and titles are documented in the Graphics Components chapter.

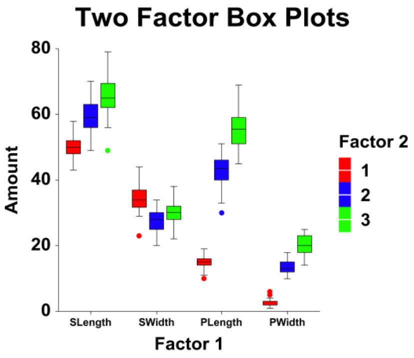
Box Plot Tab

General Section

This option specifies whether the same box colors are used above and below the median.



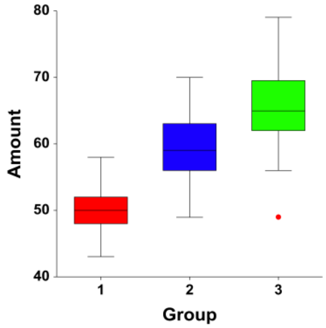




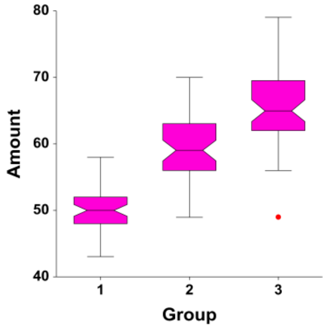
Boxes Section

You can modify the colors of the boxes and their outline using the options in this section. You can also add special notches to the box so visual multiple comparisons to be made.

Multiple Colors



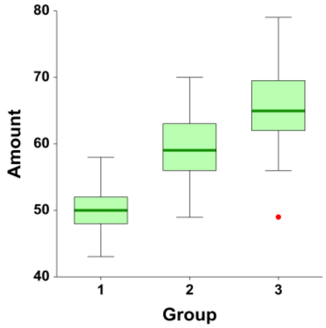
Comparison Notches



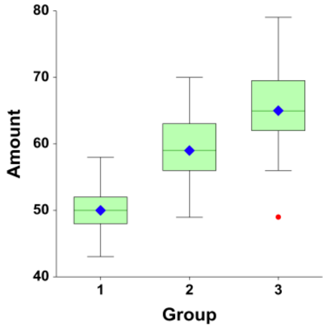
Median Section

You can modify the color of the median line and/or symbol using the options in this section.

Median Line



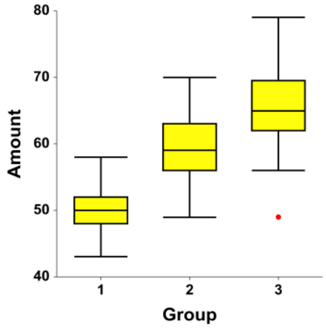
Median Symbol



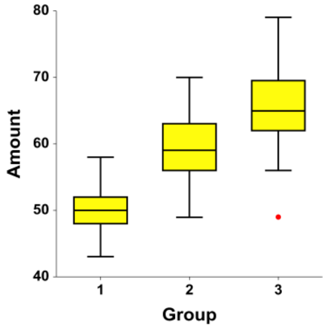
Whiskers Section

You can modify the format of the whiskers (the lines extending from the boxes) and the crossbars using the options in this section.

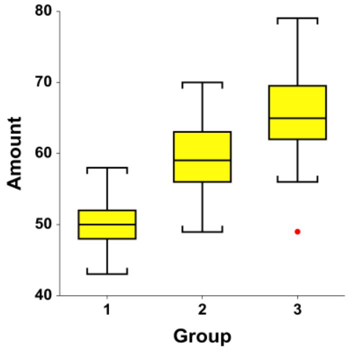
Crossbar = 100%



Crossbar = 50%

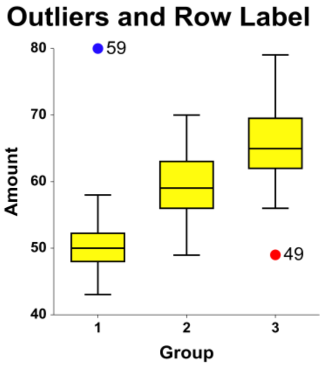


Crossbar with Ticks



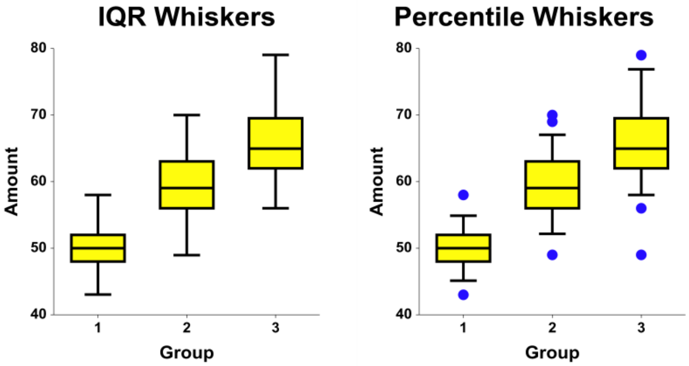
Outliers and Outlier Labels Sections

You can modify the outlier symbols and labels using the options in these sections.



Whisker and Outlier Boundaries Section

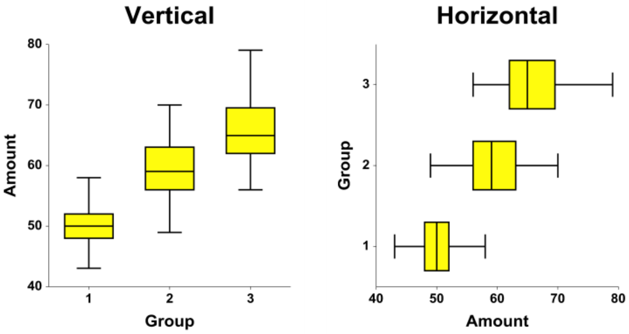
You can modify the way in which the box and whisker boundaries are calculated using the option in this section. You can also change the multipliers used to calculate the regions for the outliers. The technical details are given above.



Layout Tab

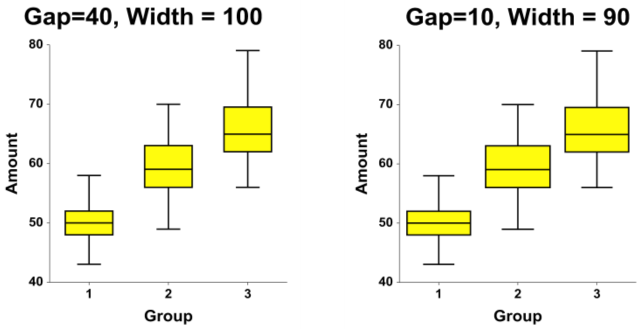
Orientation Section

You can orient the box plots horizontally or vertically.



Object Spacing and Size Section

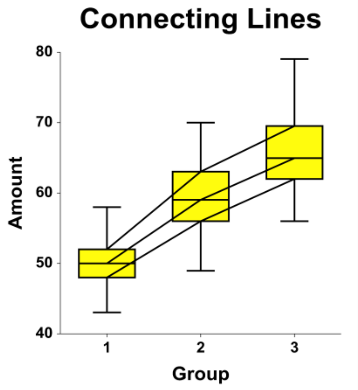
You can change the size of, and the gap between, individual box plots.



Connecting Lines Tab

Connect Between Groups Section

You can add reference lines at group means and percentiles.



Titles, Legend, Numeric Axis, Group Axis, Grid Lines, and Background Tabs

Details on setting the options in these tabs are given in the Graphics Components chapter.

Example 1 – Creating a Box Plot

This section presents an example of how to generate a box plot. The data used are from the Fisher dataset. We will create box plots of the *SepalLength* variable, grouping on the type of iris.

Setup

To run this example, complete the following steps:

1 Open the Fisher example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Fisher** and click **OK**.

2 Specify the Box Plots procedure options

- Find and open the **Box Plots** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Data Variable(s)..**SepalLength**
Horizontal (Group) Variable ....**Iris**

Reports Options (*in the Toolbar*)

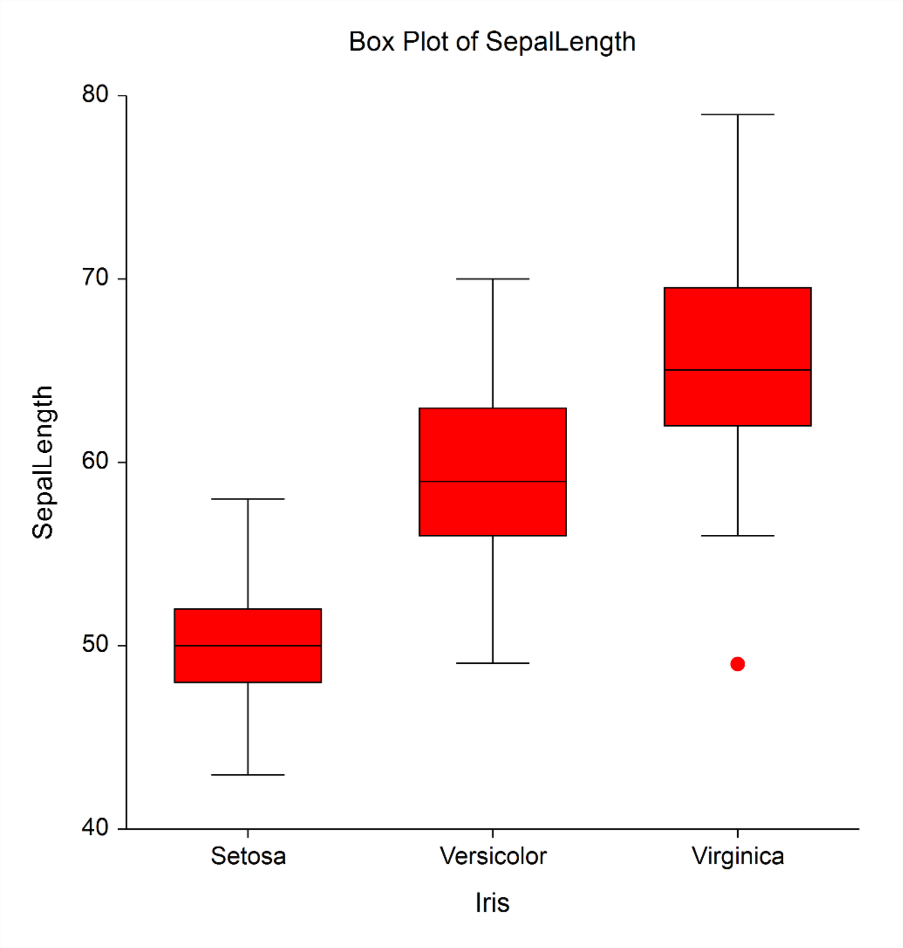
Variable Labels...**Column Names**
Data Labels...**Value Labels**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Box Plot Output

Box Plots



Example 2 – Creating a Box Plot with Subgroups

This section presents an example of how to generate a box plot with subgroups. The data used are from the fictitious Tree dataset. We will create box plots of the *Diameter* variable, grouping on *Species*, with subgroups according to *Sunlight*.

Setup

To run this example, complete the following steps:

1 Open the Tree example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Tree** and click **OK**.

2 Specify the Box Plots (2 Factors) procedure options

- Find and open the **Box Plots (2 Factors)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Data Variable(s)..**Diameter**
Horizontal (Group) Variable ....**Species**
Legend (Subgroup) Variable............................**Sunlight**

Report Options (*in the Toolbar*)

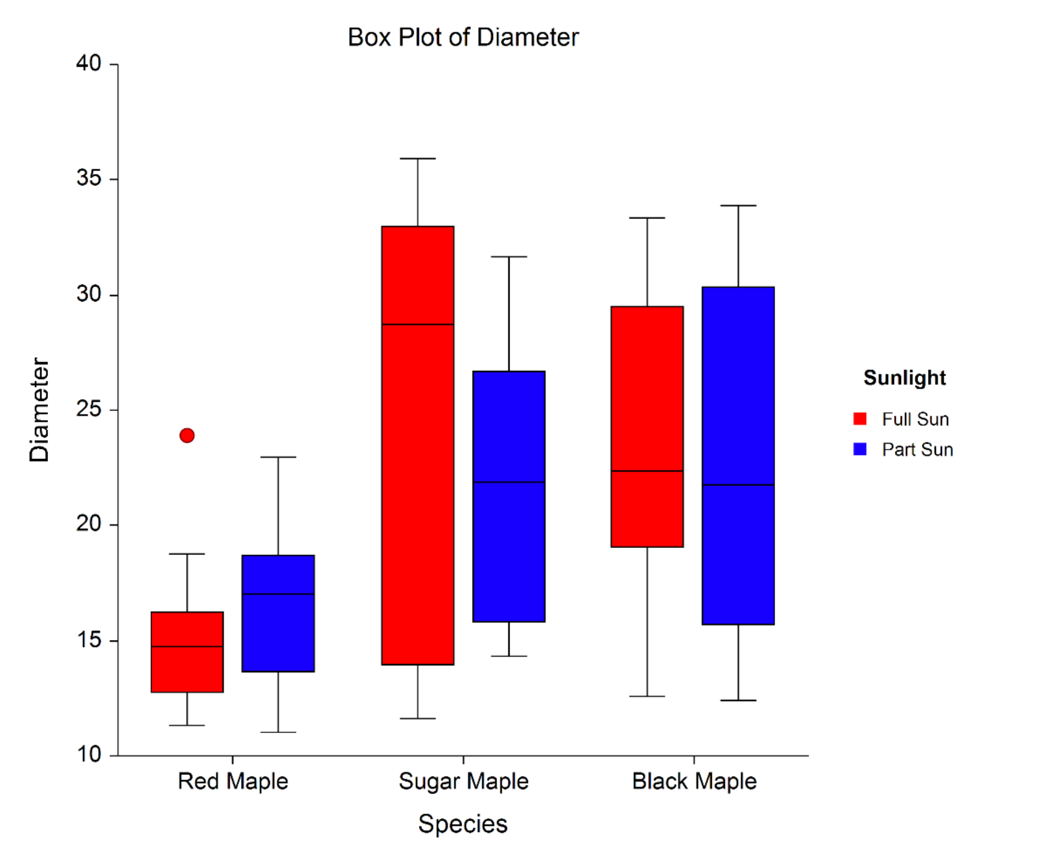
Data Labels...**Value Labels**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Box Plot Output

Box Plots


Box Plot of Diameter
Diameter
Red Maple
Sugar Maple
Black Maple
Species
Sunlight
🟥 Full Sun
🟦 Part Sun