Box-and-Whisker Plots

Introduction to Box-and-Whisker Plots – Erin Norton, Stephen Savoia & Elizabeth Hayden

In cases where there are two variables, the kind of plot you choose is dependent upon the nature of your explanatory variable on the x axis (your response variable is located on the y axis). When the explanatory variable is categorical rather than continuous, such as species or gender, then the appropriate plot is a **box-and-whisker plot**. Categorical variables are factors with two or more levels.

The standard command for a box-and-whisker plot is:

> **plot(factor,y)** which gives a box-and-whisker plot of y at levels of factor x.

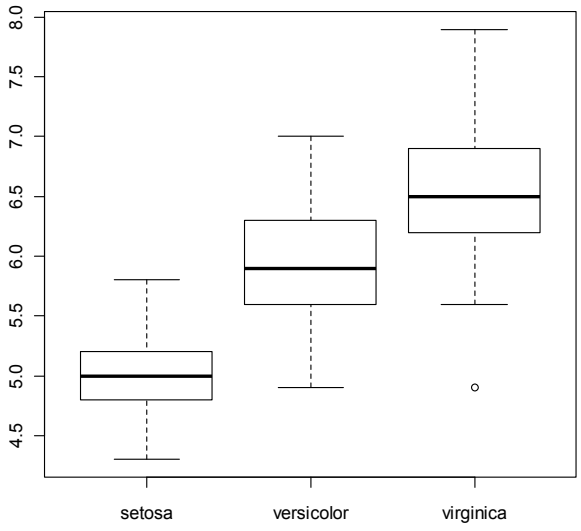
In certain instances when the x variable is numerical, you must declare the variable a factor using the following commands:

> **dataset** where "dataset" depends upon the dataset you are using.
> **attach(dataset)**
> **names(dataset)** which lists the possible variables of the dataset.
> **yourxvariable=factor(yourxvariable)** where "yourxvariable" depends upon the variable you are using.

Using the iris data set, the typically desired x variable, Species, is not numeric. Therefore, declaring Species a factor is not necessary. However, for the sake of this guide, I will include it. The response variable in iris can be one of the following: Sepal.Length, Sepal.Width, Petal.Length, or Petal.Width. The commands for plotting Species (x) versus Sepal.Length (y) are:

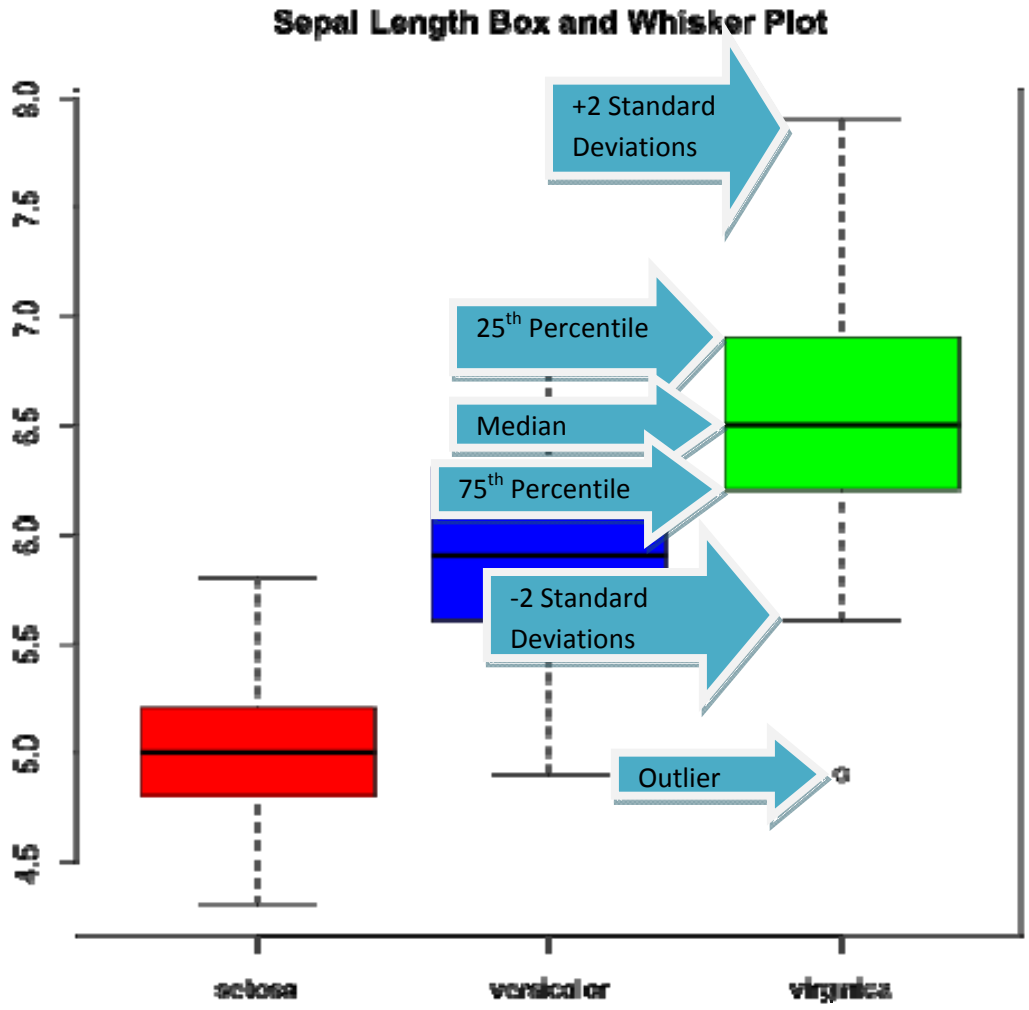
> **iris**
> **attach(iris)**
> **names(iris)**
> **Species=factor(Species)**
> **plot(Species, Sepal.Length)**

> This results in the following:



The horizontal line shows the **median** Sepal.Length measured for each species of iris flower. The box itself represents the **interquartile** range (middle 50% of the data) or all of the values that fall between the 1st and 3rd **quartiles**. The "whiskers" of the plot are 1.5 times the interquartile value and represent roughly 2 standard deviations or the maximum/minimum value whichever is greater/lesser. Every point outside of the whiskers is considered an **outlier** because it falls outside of 2 standard deviations of the median. Draw a dot to represent these points. If two dots have the same value, draw them side by side

The following graph represents the specific descriptive statistical terms addressed above:



To verify the accuracy of the graph, a fairly complicated partitioning of data is necessary. If you want to find the mean, median, maximum, minimum, variance, and/or standard deviation of a particular iris species within each variable, the following command is required:

xbar=tapply(Sepal.Length,Species,mean) Enter key
n=tapply(Sepal.Length,Species,length) Enter key
mn=tapply(Sepal.Length,Species,min) Enter key
mx=tapply(Sepal.Length,Species,max) Enter key
s=tapply(Sepal.Length,Species,sd) Enter key
v=tapply(Sepal.Length,Species,var) Enter key
cbind("NUMBER"=n,"MINIMUM"=mn,"MAXIMUM"=mx,"MEAN"=xbar,"STD DEV"=s,"VARIANCE"=v) Enter key

A table will appear:

Species	Number	Minimum	Maximum	Mean	Std Dev	Variance
Setosa	50	4.3	5.8	5.006	0.3524897	0.1242490
Versicolor	50	4.9	7.0	5.936	0.5161711	0.2664327
Virginica	50	4.9	7.9	6.588	0.6358796	0.4043439

From this table, it is clear that the median Sepal.Length for species setosa, for example, is indeed at 5. The maximum and minimum Sepal.Length values for setosa, 5.8 and 4.3 respectively, are displayed in the graph as the end of the "whiskers." To check if the first and third quartiles in the graph are correct, another series of steps must be taken involving the minimum, maximum, mean/median, and standard deviation values. The commands are as follows:

x=seq(4.3,5.8,0.1) which declares your minimum and maximum setosa Sepal.Length x values , and your incremental value.
mu=5.006 which is your setosa Sepal.Length median.
sigma=0.3524897 which is your setosa Sepal.Length standard deviation.
SL.setosa=Sepal.Length[Species=="setosa"]
summary(SL.setosa) which gives you the following table:

0%	25%	50%	75%	100%
4.300	4.800	5.006	5.200	5.800

From this table, the first and third quartile values, 25% and 75% respectively, match the bottom and top of the box in the plot.

Adding Labels and Colors to Your Graph

In order to inform the reader of what your graph is displaying, the axes should have meaningful labels, and there should always be a title at the top.

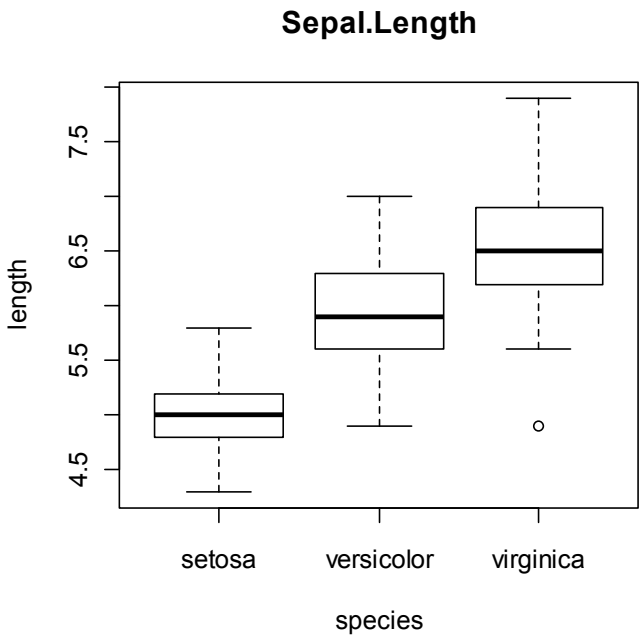
Labeling the Axes

Within the function, after specifying the parameters of the plot, specify the y axis label as ylab="*your choice of label*", the x axis label as xlab="*your choice of label*", and the main title as main="*your choice of title*". Separate these commands with commas then end the command with parentheses.

For example, to label a box and whisker plot of the sepal lengths of the three species of the data set iris, you might type:

plot(Species,Sepal.Length,ylab="length",xlab="species",main="Sepal.Length")

This results in the following:



You may also add labels or titles after the graph is generated. You can do so by typing:

```
plot(Species, Sepal.Length)
title(main="Sepal.Length")
title(xlab="species")
title(ylab="length")
```

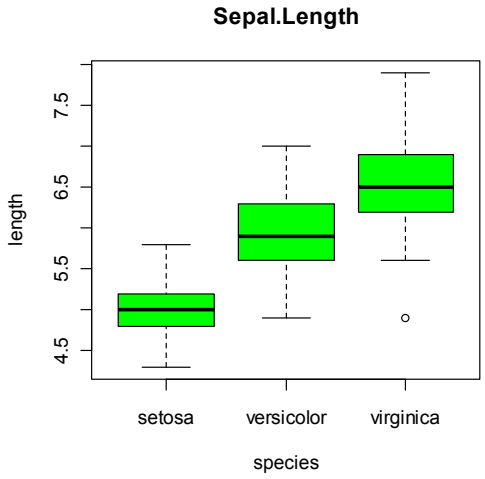
Adding Color to Plots

Sometimes you may want to add color to make distinctions between multiple box-and-whisker plots on one graph.

To add the same color to every box on the graph, follow the commands within the parentheses with col=”green”

```
plot(Species,Sepal.Length,ylab="length",xlab="species",main="Sepal.Length",col="green")
```

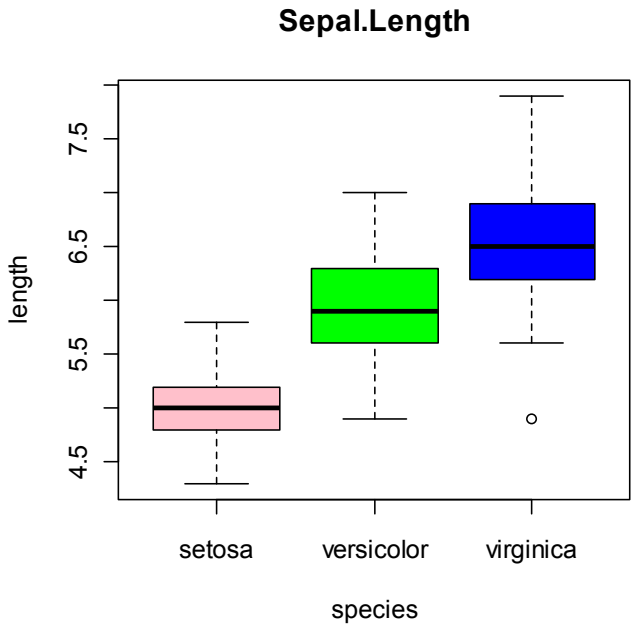
This results in the following:



To make each box a different color, specify this at the end of the command with col=c("*color for box 1", "color for box 2", "color for box 3", etc.)*

plot(Species,Sepal.Length,ylab="length",xlab="species",main="Sepal.Length", col=c("pink","blue","green"))

This results in the following graph:



Representing a Confidence Interval

If in a given data set you want to test whether or not the medians are significantly different from each other, you can look to the notches in a box and whisker plot to identify how significant the differences are within a set of data. The notches represent roughly a 95% confidence interval for the differences in the medians. If the notches do not overlap, a test would likely prove that the medians are significantly different. Add *notch=T* to your input to generate notches.

plot(Species,Sepal.Length,main="Sepal Length Box and Whisker Plot",col=c("red","blue","green"),notch=T)

This results in the following:

In the example dataset, none of the notches overlap, and it is therefore likely that under a significance test, the medians would prove significantly different from one another. If the notches extend past the box itself, it is likely that the sample variance is high and points to the invalidity of the test.