R Tutorial for STAT 350 Lab 2

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

1. Numerical Summaries

Example 1. Time to Start a Business (**Data: eg01-23time24.txt**)

An entrepreneur faces many bureaucratic and legal hurdles when starting a new business. The World Bank collects information about starting businesses throughout the world. It has determined the time, in days, to complete all of the procedures required to start a business. Data for 195 countries are included in the data set. For this section we will examine data for a sample of 24 of these countries. Here are the data:

13	66	36	12	8	27	6	7	5	7	52	48
15	7	12	94	28	5	13	60	5	5	18	18

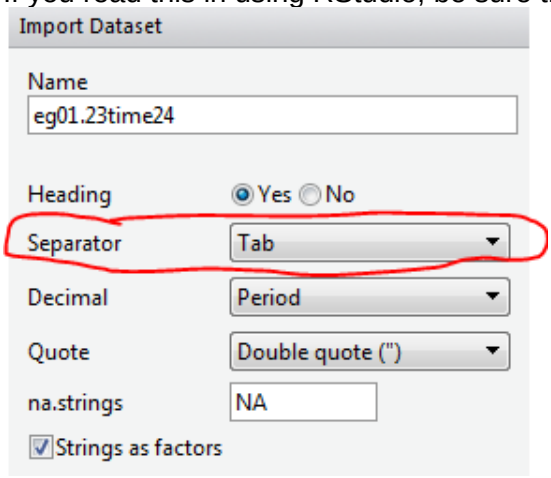
a) Find the mean and the standard deviation for the time to start for the new businesses.
b) Find the five number summary for the time to start for the new businesses.
c) Create a histogram for the length of all of the flowers.
d) Create a boxplot (modified) for the new businesses.

Solution:

reading in the data:

```
> TimeStart <- read.table("eg01-23time24.txt", header = T,
    sep= "\t")
> #In this data file, there are both spaces and tabs. The
> # sep command tells R that you only want to use tabs (\t) as
> # separators. This occurs in some of the data sets so if you
> # notice that there are spaces in either the variable names or
> # the data itself, you need to add this addition keyword in
> # the command.
> View(TimeStart)
```

If you read this in using RStudio, be sure that the Separator is set to tab



You will need to tell R the variable of interest from the table that was read in. There are three ways to do this:

R Tutorial for STAT 350 Lab 2

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

1) `>attach(TimeStart)`
   Then cite by the variable name alone as listed in the R object TimeStart itself. Note when using the attach command, the name of the R object has to be different than the name of any of the variables in the object.
   This is the method that I prefer.
2) Use the order in which the variables occur in square brackets after the R object
3) Use the variable name following a '$' sign after the R object. This is the method that we used in Lab 1.

In this example, we are interested in the variable TimeToStart which is the second variable. Therefore, you would indicate it by "TimeToStart", "TimeStart[, 2]", or "TimeStart$TimeToStart", respectively. I personally prefer the first method, so long as it is the main R object of interest as only one R object can be attached at a time. If this is not the case, I prefer the third method. You can use the output `head(TimeStart)` or look at the data table to determine what the variable names are (or from the .txt file itself).

For this part, I will show you how to use all three of the methods. For the rest of this tutorial, I will only be using method 1.

a) Find the mean and the standard deviation for the time to start for the new businesses.

Solution:

1)
```
> attach(TimeStart)
> mean(TimeToStart)
[1] 23.625
> sd(TimeToStart)
[1] 23.82876
```
2)
```
> mean(TimeStart[,2])
[1] 23.625
> sd(TimeStart[,2])
[1] 23.82876
```

3)
```
> mean(TimeStart$TimeToStart)
[1] 23.625
> sd(TimeStart$TimeToStart)
[1] 23.82876
```

From the R output above,
mean = 23.625, standard deviation =23.82876

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

b) Find the five number summary for the time to start for new businesses.

Solution:

```
> fivenum(TimeToStart)
[1]  5  7 13 32 94
```

From the R output above,
Min = 5, Q_1 = 7, Median = 13, Q_3 = 32, Max = 94

Note: There are other ways of obtaining the five number summary. However, this function will generate the values of the quartiles that are similar to the values that are obtained in our textbook. In R, always define which number is which when using this command.

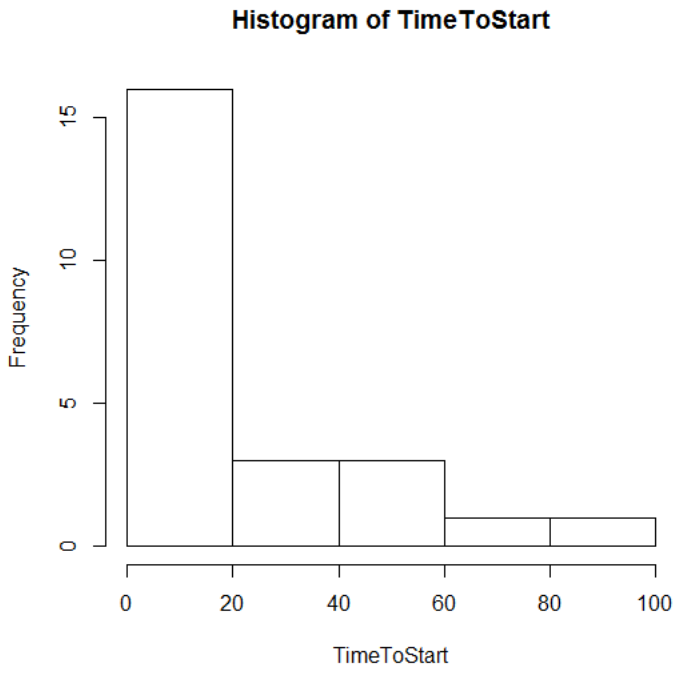
2. Creating Histograms

c) Create a histogram for the length of all of the flowers.

Solution:

R usually does a good job of generating histograms with an appropriate number of bins.
$number\ of\ binx\ \approx \sqrt{number\ of\ data\ points}$

```
> hist(TimeToStart)
```



R Tutorial for STAT 350 Lab 2

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

I strongly recommend that you change the size of the graph so that it fits better on the page.

(OPTIONAL) We will be discussing additional options for creating histograms in later tutorials, such as utilizing the "lattice" graphics package that can be installed from within an R session:

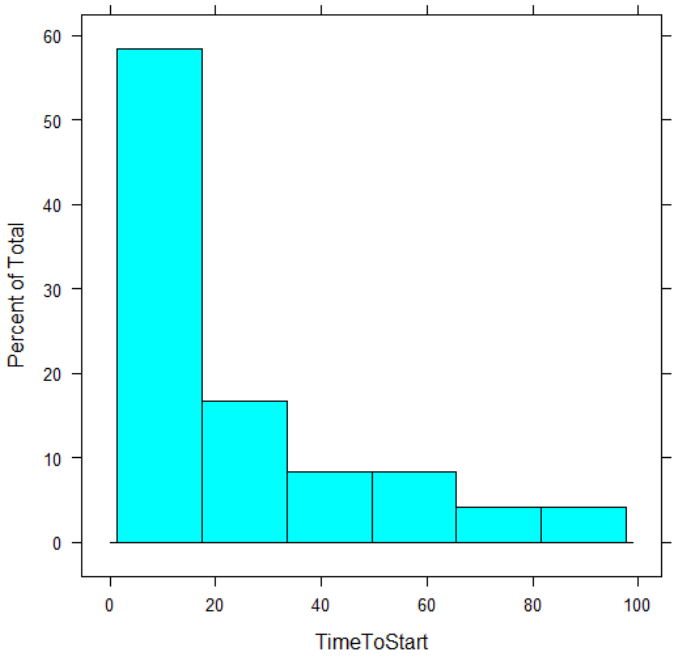
```
>install.packages("lattice") #Only needs to be run once
```

R will ask you from which mirror to install the package. Please choose the one that is closest to your location.

```
>library(lattice) #Needs to be run for each and every R session
```

Please ignore any warning messages.

```
>histogram(~TimeToStart)
```



3. Boxplots

d) Create a boxplot (modified) for the new businesses.

Solution:

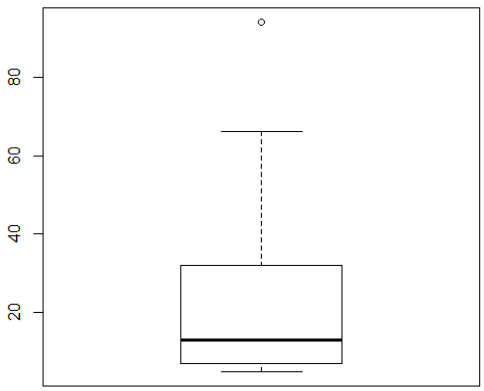
The following is the procedure for one variable. This will generate the modified boxplot that is the boxplot in which the outliers are explicitly plotted. The procedure is different if you have more than one variable.

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

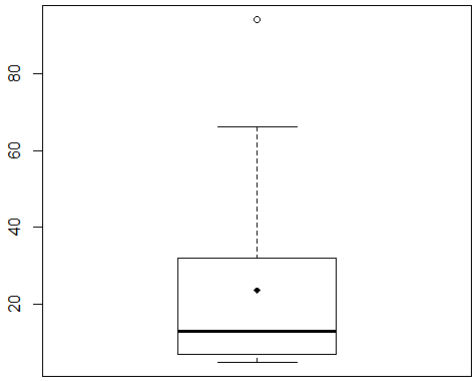
```
> boxplot(TimeToStart)
```

If you want to add the location of the mean (I find this very useful), also include the following two lines:

```
> means = mean(TimeToStart)
> points(means, pch = 18)
```



without mean



with mean

(OPTIONAL) The function for boxplot in lattice is bwplot(). Feel free to play with this function. We will include further options on this function later in the semester.

```
> bwplot(TimeToStart)
```