STAT 371: Discussion 2: Summarizing Data

1. Create a new R Markdown file via the menu choice "File > New File > R Markdown". Or open and edit this R Markdown file. Save it as "Discusssion2.Rmd" in your class folder.
2. If you created a new file, delete "##R Markdown" and everything below it. Knit your document and confirm it is in your class folder.
3. Save the sloth.csv data file from the Data Sets Page in the Administration Module on Canvas to the class folder (or a subfolder of it) on your computer.

We will be looking at a couple of different ways to *get* data to analyze.

a. You can enter vectors manually by typing it in with the concatenate "c" function.

```
WeightKg<-c(4,5,4.6,3.5,3.8,4,4.7,4.6,5,6.3,5,5.3)
RecoverTime<-c(10,12.5,12.3,8.75,10,10,11.8,11.5,12.5,16,12.5,13.25)
```

b. There are some data sets that are available by default after installing R (like "iris" from lecture)

```
library(datasets)
data(iris)

# For a full list of these datasets, type library(help = "datasets")
```

c. You can import the data from a text or csv (or other similar file). In order to do this, you need to set your working directory to where you have the data csv or text file saved with the menu selections: "Session>Set Working Directory>Choose". You can also manually type in a 'setwd" command, but that usually takes me forever.

4. Save the sloth data into a new variable named sloth data using the read.csv function. You may need to look at ?read.csv to see what other parameters you will need to specify.

```
#for csv file
slothdata <- read.csv('sloth.csv')
#We specify header=TRUE since the first row has names that describe the columns
```

5. Copy and past the code below to recreate the sloth data by hand from the csv file. This method is not recommended because it is easier to make typos, however it is good practice with the function rep which can make entering repetative data (HWK 2 Q4) much easier.

```
Species<-rep(rep(c("ThreeToed", "TwoToed"), each=3), times=2)
WeightKg<-c(4.0,5.0, 4.6, 3.5, 3.8, 4.0, 4.7, 4.6, 5.0, 6.3, 5.0, 5.3)
Gender<-c("M", "F", "F", "M", "M", "F","M", "M", "F", "M", "F", "M")
Treatment<-rep(c("High", "Low"), each=6)
RecoverTime<-c(10.00, 12.50, 12.30, 8.75, 10.00, 10.00, 11.80, 11.50, 12.50, 16.00, 12.50, 13.25)

slothdata.manual<-data.frame(Species, WeightKg, Gender, Treatment, RecoverTime)
```

6. Use the View(), head(), and str() functions on the imported "slothdata" and the manually entered "slothdata.manual" to confirm that the data looks like we want it to. Only use View() in the console as it doesn't behave well when we go to knit our document.

```
#View(slothdata)
head(slothdata)
```

```
##     Species WeightKg Gender Treatment RecoverTime
## 1 ThreeToed      4.0      M      High       10.00
## 2 ThreeToed      5.0      F      High       12.50
## 3 ThreeToed      4.6      F      High       12.30
## 4   TwoToed      3.5      M      High        8.75
## 5   TwoToed      3.8      M      High       10.00
## 6   TwoToed      4.0      F      High       10.00
```

```
str(slothdata)
```

```
## 'data.frame':    12 obs. of  5 variables:
##  $ Species    : Factor w/ 2 levels "ThreeToed","TwoToed": 1 1 1 2 2 2 1 1 1 2 ...
##  $ WeightKg   : num  4 5 4.6 3.5 3.8 4 4.7 4.6 5 6.3 ...
##  $ Gender     : Factor w/ 2 levels "F","M": 2 1 1 2 2 1 2 2 1 2 ...
##  $ Treatment  : Factor w/ 2 levels "High","Low": 1 1 1 1 1 1 2 2 2 2 ...
##  $ RecoverTime: num  10 12.5 12.3 8.75 10 10 11.8 11.5 12.5 16 ...
```

```
head(slothdata.manual)
```

```
##     Species WeightKg Gender Treatment RecoverTime
## 1 ThreeToed      4.0      M      High       10.00
## 2 ThreeToed      5.0      F      High       12.50
## 3 ThreeToed      4.6      F      High       12.30
## 4   TwoToed      3.5      M      High        8.75
## 5   TwoToed      3.8      M      High       10.00
## 6   TwoToed      4.0      F      High       10.00
```

```
str(slothdata.manual)
```

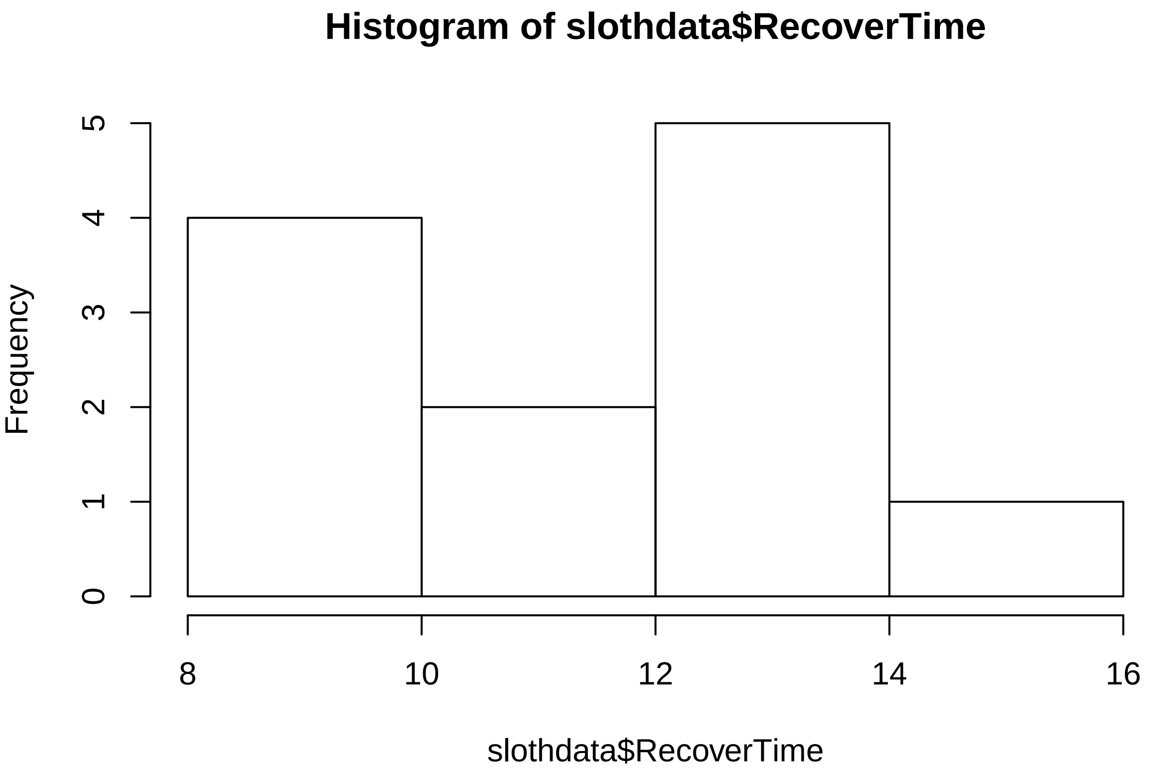
```
## 'data.frame':    12 obs. of  5 variables:
##  $ Species    : Factor w/ 2 levels "ThreeToed","TwoToed": 1 1 1 2 2 2 1 1 1 2 ...
##  $ WeightKg   : num  4 5 4.6 3.5 3.8 4 4.7 4.6 5 6.3 ...
##  $ Gender     : Factor w/ 2 levels "F","M": 2 1 1 2 2 1 2 2 1 2 ...
##  $ Treatment  : Factor w/ 2 levels "High","Low": 1 1 1 1 1 1 2 2 2 2 ...
##  $ RecoverTime: num  10 12.5 12.3 8.75 10 10 11.8 11.5 12.5 16 ...
```

7. Describe the data (number of observations, number of variables, type of variables). Did R correctly recognize the types of variables when we imported? When we manually created data frame?

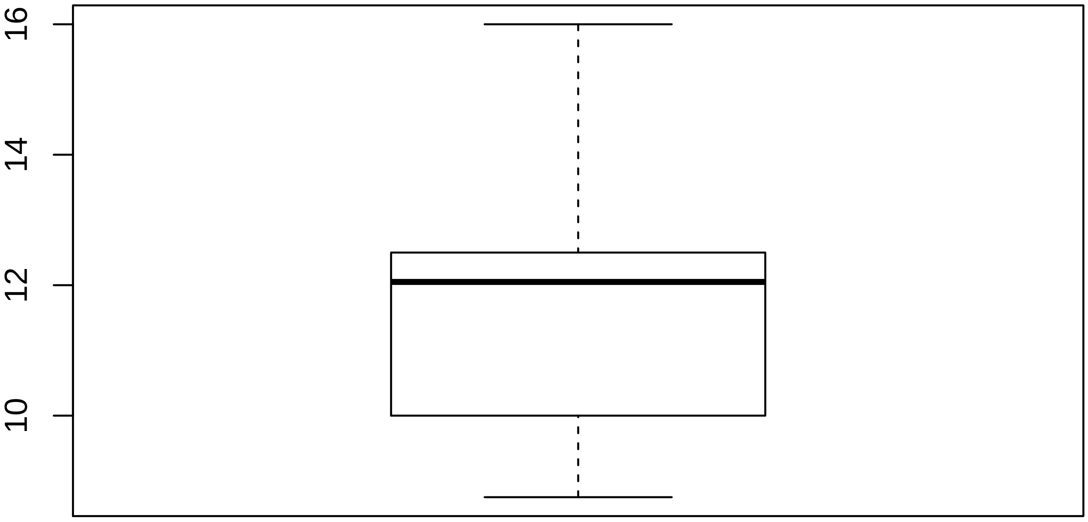
There are 12 observations of 5 variables. Species, Gender and Treatment are all categorical and Weight and recovery time are numeric. Notice, if treatment had been entered as 1, 2 instead of high, low, R would have defaulted those values to numeric instead of categorical

8. We can use $ to refer to a single column within a data frame. Summarize the recovery time for all sloths observed with a histogram, boxplot, and appropriate numerical summaries. Use either dataframe.

```
hist(slothdata$RecoverTime)
```



```
boxplot(slothdata$RecoverTime)
```



```
summary(slothdata$RecoverTime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.75   10.00   12.05   11.76   12.50   16.00
```

```
sd(slothdata$RecoverTime)
```

```
## [1] 1.920089
```

9. Use the subset function to define new dataframes for sloths of the same species. That is, create one for only the two-toed sloths (two.toed.sloths) and another for the three-toed sloths (three.toed.sloths). View your new dataframes to make sure they contain the data you want.

```
two.toed.sloths<-subset(slothdata, Species=="TwoToed")
three.toed.sloths<-subset(slothdata, Species=="ThreeToed")
```

10. Compute and compare the mean, sd, and 5 number summary for the recovery times of the two species seperately.

```
summary(two.toed.sloths$RecoverTime); sd(two.toed.sloths$RecoverTime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.75   10.00   11.25   11.75   13.06   16.00
```

```
## [1] 2.683282
```

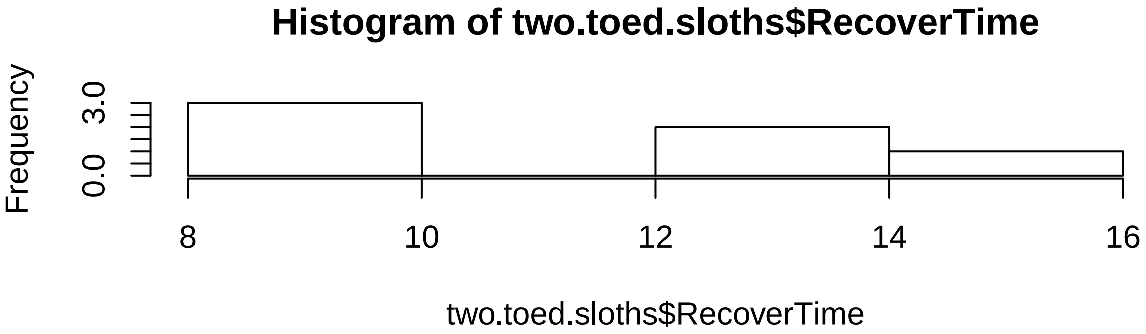
```
summary(three.toed.sloths$RecoverTime); sd(three.toed.sloths$RecoverTime)
```

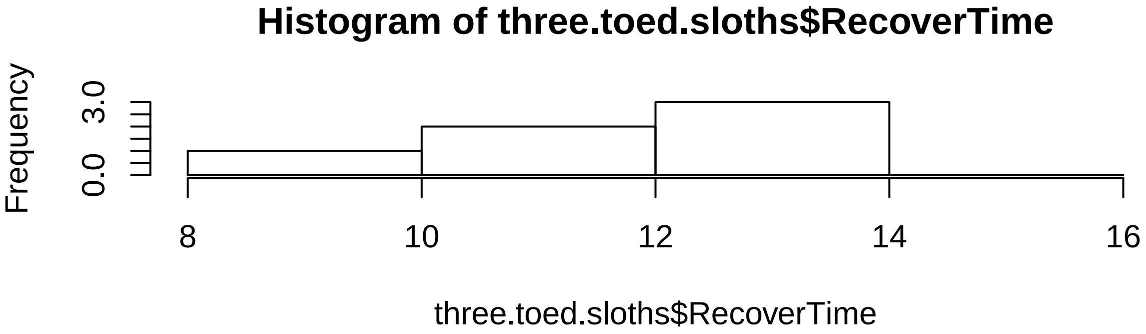
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   11.57   12.05   11.77   12.45   12.50
```

```
## [1] 0.9542886
```

11. Create stacked histograms so it is easier to compare the recovery times of the two species. Describe how the relative size of the values computed in 10 are reflected in the histograms. Note any interesting characteristics the graph shows or hides.

```
par(mfrow=c(2,1))
hist(two.toed.sloths$RecoverTime, breaks=c(seq(8,16,2)))
hist(three.toed.sloths$RecoverTime, breaks=c(seq(8,16,2)))
```



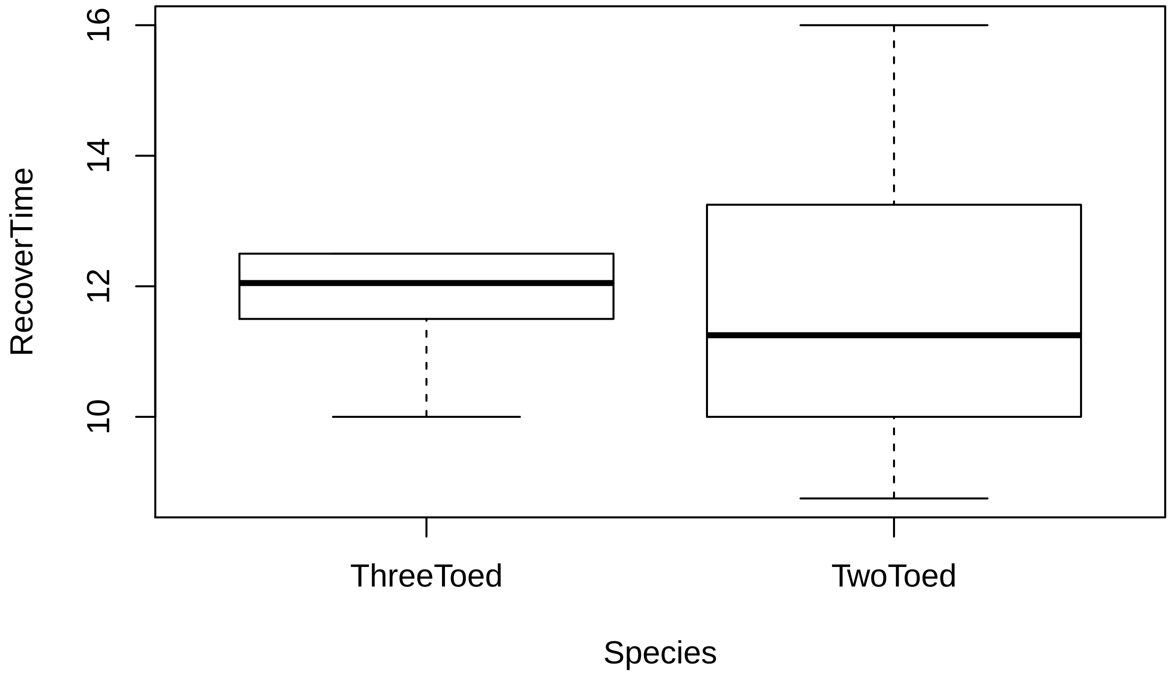


```r
par(mfrow=c(1,1))
```

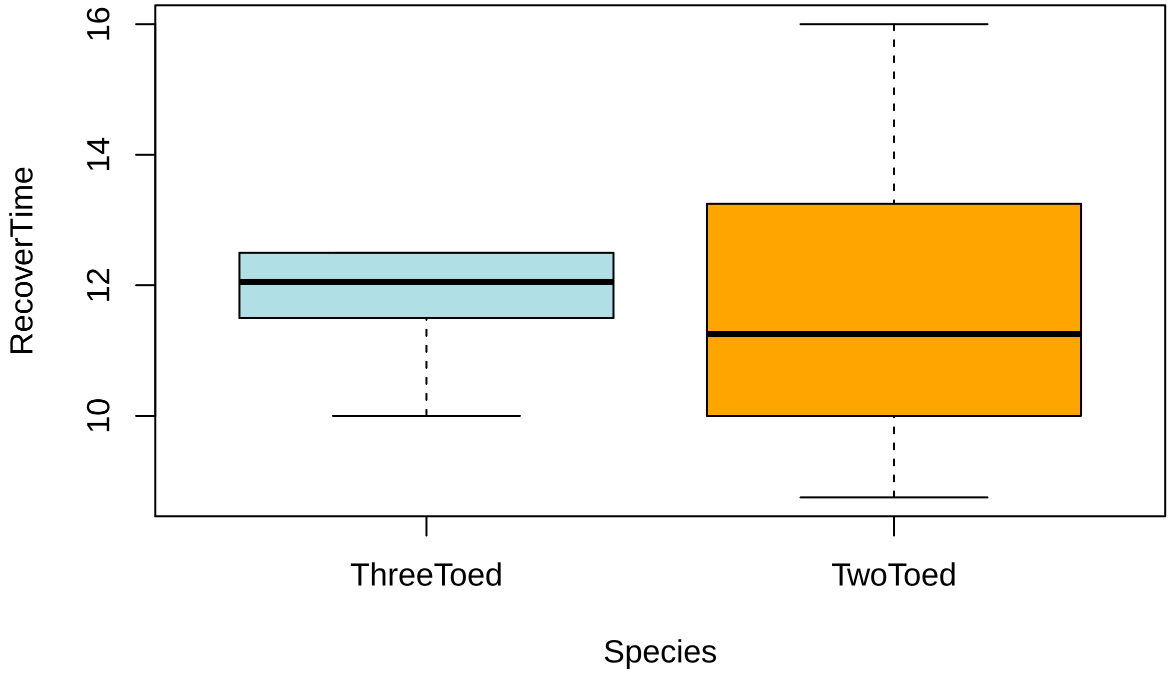
We see that the center values (mean and median) are pretty similar within each set and between the two sets. This is not surprising as they are positioned similarly on the x axis. We do see that the three toed sloth recovery times had much smaller sd than the two toed and this is apparent because the values are more concentrated around the center of the data in the three toed histogram. It is difficult to estimate the exact mean or median value from the graph, but we can see where there are empty ranges of values.

12. Create side-by-side box plots to compare the recovery times of the two species. Describe how the relative size of the values computed in 10 are reflected in the boxplots. Note any interesting characteristics the graph shows or hides.

```r
boxplot(RecoverTime~Species, data=slothdata)
```



```
boxplot(RecoverTime~Species, data=slothdata, col=c('powderblue', 'orange'))
```



We can more clearly compare the median values and see the median of the three toed recovery time is longer than that of the two toed. We also see the increased spread in the two toed recovery time. We cannot see where there are empty ranges of values nor how many actual data values there are.