Lab 2 Worksheet

Kyle Payne

September 1, 2015

# Fun with Boxplots

Download the 'data\_cortex' data, import the dataset including the headers, and run the following code.

## this will be where the code is actually located on the   
## computer, i.e., this code is already run when using  
## the import dataset button.  
  
data\_cortex <- read.csv("~/Desktop/data\_cortex.csv")  
  
## on your computer '~/Desktop/data\_cortex.csv' will be the  
## path to the file.  
  
## grab only rows that have no missing data  
## and reassigns this to another data.frame called data\_cortex  
## this essentially just re-writes the data\_cortex data.frame  
  
data\_cortex <- data\_cortex[complete.cases(data\_cortex),]

1. Create a boxplot of the variable 'NR1\_N'. Last lab session we talked about selecting variables from R data.frames. So the following should look familiar

NR1\_N <- data\_cortex[, "NR1\_N"]  
NR1\_N\_2 <- data\_cortex[,5]  
  
  
library(knitr) ## you dont have to run this line  
kable(head(NR1\_N)) ## or this one

|  |
| --- |
| 2.921435 |
| 2.862575 |
| 2.968155 |
| 2.624901 |
| 2.634509 |
| 2.598085 |

kable(head(NR1\_N\_2)) ## or this one

|  |
| --- |
| 2.921435 |
| 2.862575 |
| 2.968155 |
| 2.624901 |
| 2.634509 |
| 2.598085 |

See?, the numbers are the same, so that means that NR1\_N and NR1\_N\_2 are the same variable. One way, I pulled the variable name from the right hand side of the comma in the bracket notation, and the second way I just put the number of the column that it is in counting from the left. There is one more way be can reference columns in the data.frame, and this is to use the $.

NR1\_N\_3 <- data\_cortex$NR1\_N  
  
kable(head(NR1\_N))

|  |
| --- |
| 2.921435 |
| 2.862575 |
| 2.968155 |
| 2.624901 |
| 2.634509 |
| 2.598085 |

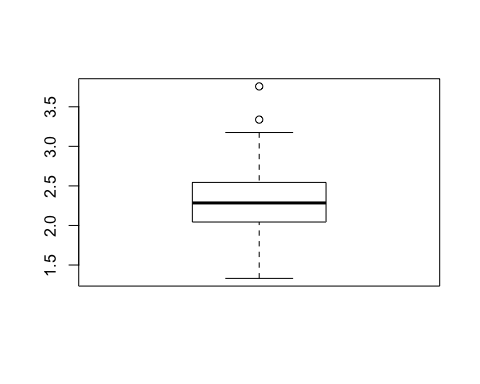
That is, to use the data.frame$column\_name format, where you put the name of the data.frame on the left hand of the $ and the column\_name you'd like to use on the right hand side.

mouse\_IDs <- data\_cortex$MouseID  
head(mouse\_IDs)

## [1] 3415\_1 3415\_2 3415\_3 3415\_4 3415\_5 3415\_6  
## 1080 Levels: 18899\_1 18899\_10 18899\_11 18899\_12 18899\_13 ... J3295\_9

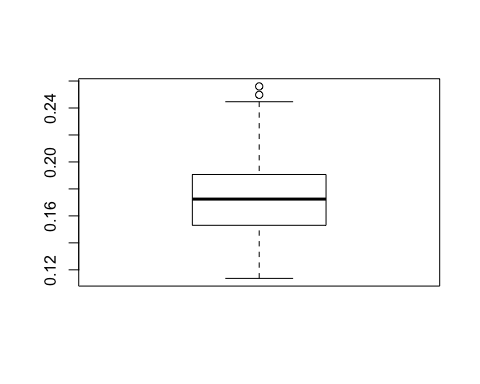
We can now also create a boxplot for NR1\_N.

boxplot(NR1\_N)

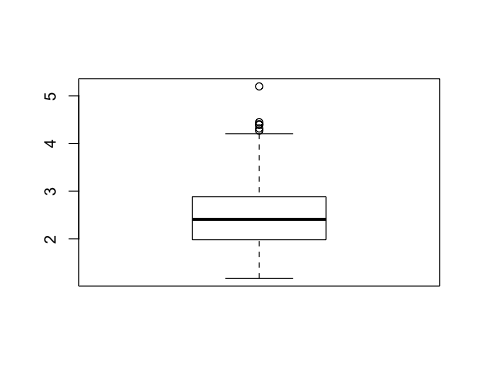


Of course we can do this for multiple variables.

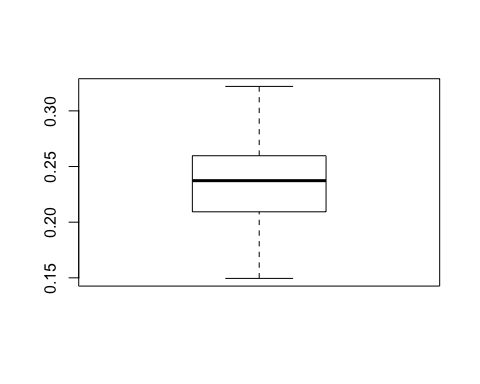
### these are the proteins I'm grabbing 'CREB\_N', 'ERK\_N', 'JNK\_N'  
  
###  
boxplot(data\_cortex[, 'CREB\_N'])



boxplot(data\_cortex[, 'ERK\_N'])



boxplot(data\_cortex[, 'JNK\_N'])



We can also get a list of the means for each of these variables

## grab the variables that we made the boxplots with.  
prots <- c('CREB\_N', 'ERK\_N', 'JNK\_N')   
## grab those variables out of the data.frame and  
## reassign them to a new data.frame  
dc\_prots <- data\_cortex[,prots]  
  
## colMeans calculates the mean for each column  
kable(colMeans(dc\_prots))

|  |  |
| --- | --- |
| CREB\_N | 0.1731612 |
| ERK\_N | 2.4823928 |
| JNK\_N | 0.2350940 |

## standard deviation  
## the FUN part mean 'function', i.e. which function  
## you would like to apply to each column  
kable(sapply(dc\_prots, FUN=sd))

|  |  |
| --- | --- |
| CREB\_N | 0.0256362 |
| ERK\_N | 0.6644644 |
| JNK\_N | 0.0323462 |

kable(sapply(dc\_prots, FUN=quantile))

|  |  |  |  |
| --- | --- | --- | --- |
|  | CREB\_N | ERK\_N | JNK\_N |
| 0% | 0.1136364 | 1.169514 | 0.1494872 |
| 25% | 0.1530874 | 1.982082 | 0.2093339 |
| 50% | 0.1724507 | 2.406569 | 0.2372972 |
| 75% | 0.1906849 | 2.883884 | 0.2596802 |
| 100% | 0.2560119 | 5.198404 | 0.3219850 |

sapply stands for 'simple' + 'apply', which means roughly, 'apply this function to each column in my data.frame, then output the results in a simple format'. This is a nice function to use if you have to perform the same task over several variables. In fact the colMeans function is just a convinent wrapper for

kable(sapply(dc\_prots, FUN=mean))

|  |  |
| --- | --- |
| CREB\_N | 0.1731612 |
| ERK\_N | 2.4823928 |
| JNK\_N | 0.2350940 |

## Activity

* Grab the variables 'SOD1\_N' , 'pGSK3B\_N' , 'AcetylH3K9\_N' from data\_cortex.
* Create a boxplot for each.
* Note any possible outliers in each variable.
* Calculate the means, standard deviation, and Q1, Q3 or the 25th and 75th quantile. of each of the variables.