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),]
```

    a. 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

$$\underline{2.598085}$$

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.
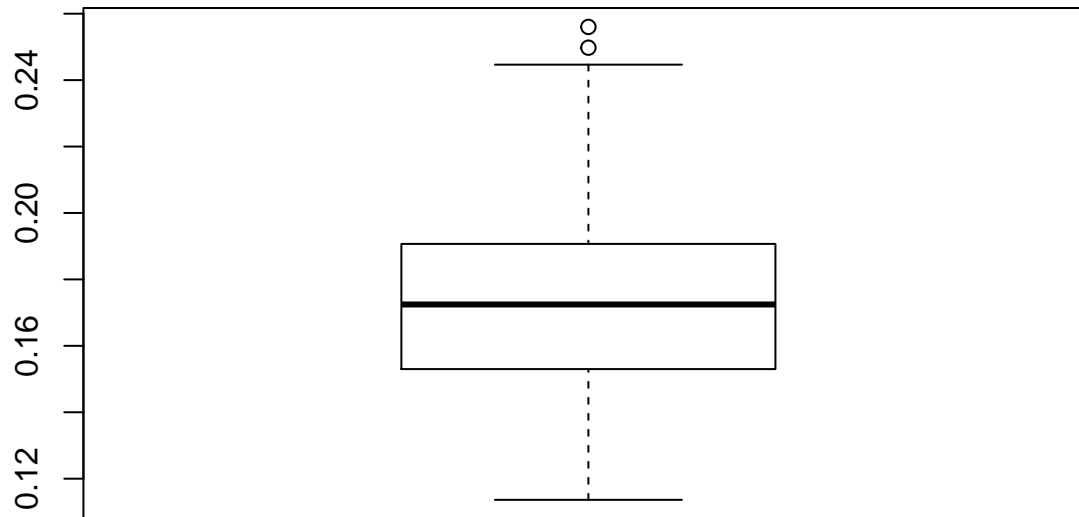
```
boxplot(NR1_N)
```
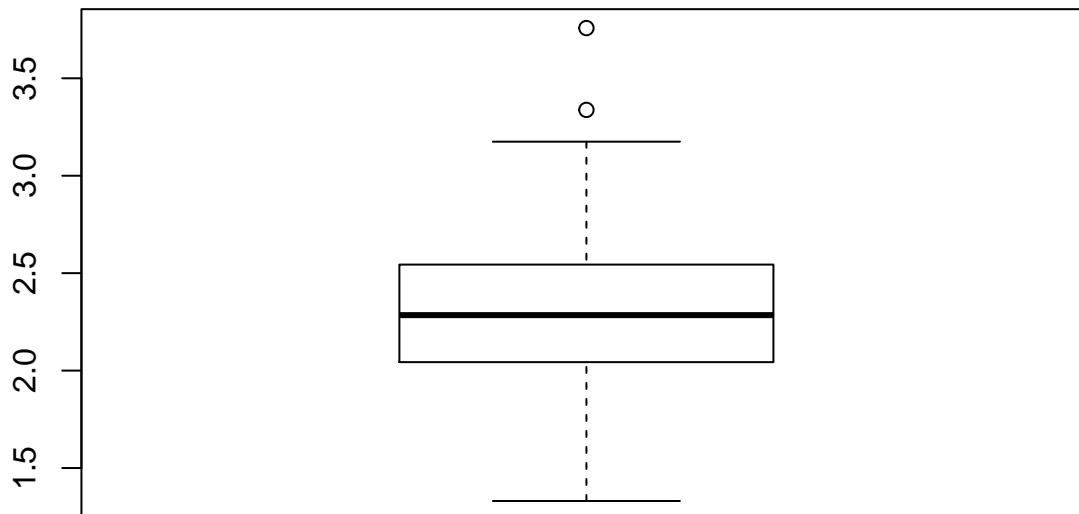

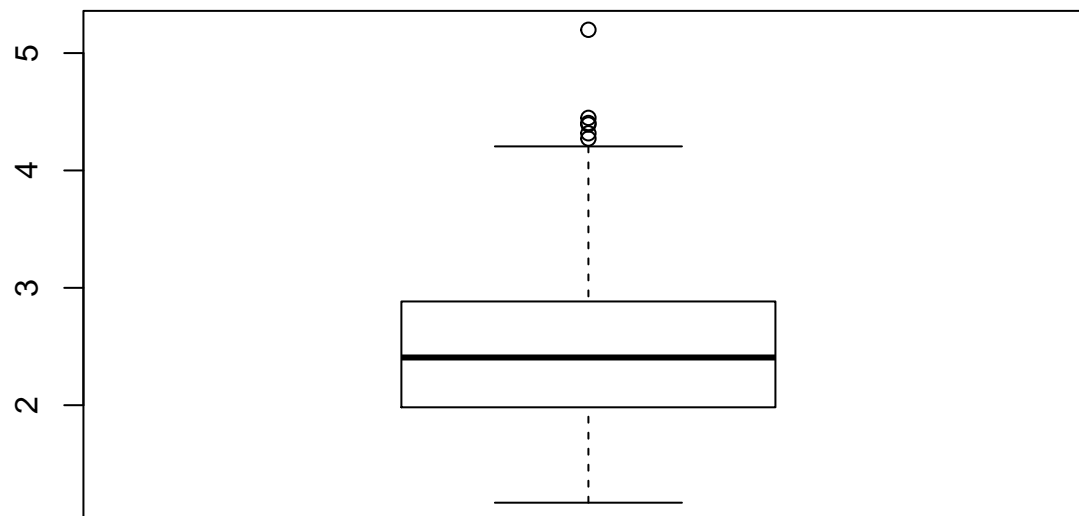Of course we can do this for multiple variables.

```
### these are the proteins I'm grabbing 'CREB_N', 'ERK_N', 'JNK_N'

### start of the boxplot code
boxplot(data_cortex[, 'CREB_N'])
```
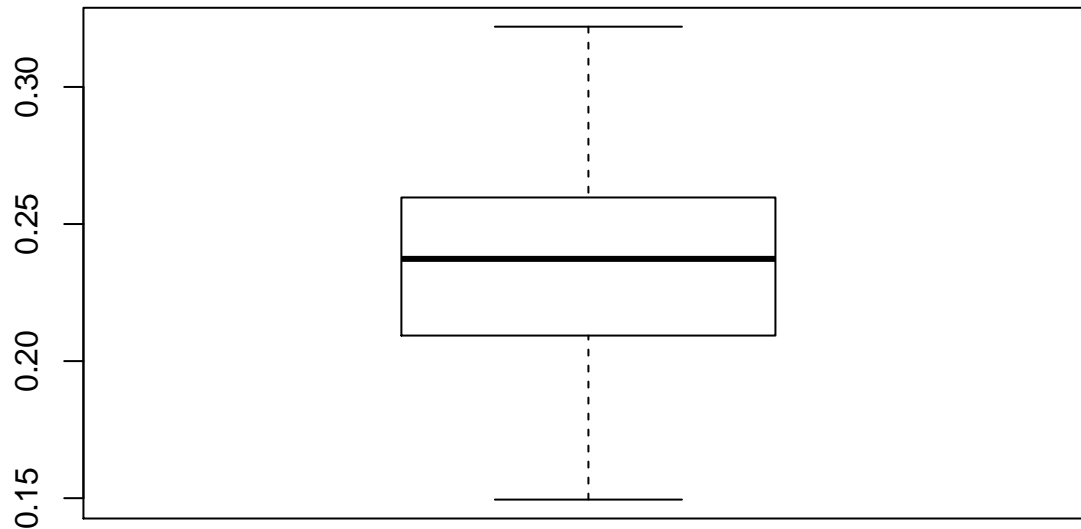


```
boxplot(data_cortex[, 'ERK_N'])
```



```
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

```r
kable(sapply(dc_prots, FUN=mean))
```

CREB_N	0.1731612
ERK_N	2.4823928
JNK_N	0.2350940

Activity

1. Grab the variables 'SOD1_N' , 'pGSK3B_N' , 'AcetylH3K9_N' from data_cortex.
2. Create a boxplot for each.
3. Note any possible outliers in each variable.
4. Calculate the means, standard deviation, and Q1, Q3 or the 25th and 75th quantile of each of the variables.

Probability Examples

1. An XYZ cell phone is made with 55 components. Each of the components works independently of each other, and each has a 0.004 probability of being defective. What is the probability that this phone won't work completely?

First we need to figure out what probability distribution best describes this problem. Let's see,

1. The experiment consists of n identical trials.
2. Each trial results in one of two outcomes (defective or not defective).
3. The probability of success on a single trial is equal to π and π remains the same from trial to trial. $\pi = .004$ in this case.
4. The trials are independent.
5. The random variable Y is the number of defective observed during the n trails. This value will be from 0 to n.

So it looks like this is binomial distributed random variable. The probability mass function looks like:

$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

2. From a group of friends that consists of 7 men and 5 women, what's the probability of getting 4 men and 2 women togehter for dinner?

Obviously, if you select a friend to go to the dinner party, he's not going to be placed back into the general sample space. i.e. if you go to dinner with your buddy Mark, you can't then select Mark again and then go to dinner with two Marks. This is an example of sampling without replacement.

Thus, sampling from two groups without replacement is a hypergeometric problem, with p.m.f.

$$P(X = k) = \frac{\binom{K}{k}\binom{N-k}{n-k}}{\binom{N}{n}}$$

Thus for this example, $N = 12$, $K = 7$, $N - K = 5$, $n = 6$.

Answers

Please don't read until you've given this at least one attempt.

1. The event A being that the phone won't work completely is equivalent to the event that at least 1 of the components is defective. If X is the number of defective components in the phone, then at least one of the components being defective is $X \geq 1 = \overline{\{X = 0\}}$, i.e. the complement that none are defective. Thus we are looking at the probability of a complement event, $P(\overline{\{X = 0\}}) = 1 - P(X = 0)$. Then $P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{55}{0}(0.004)^0(1 - 0.004)^55 = 1 - 1 \times 1 \times 0.802 = .197$

```
1-pbinom(0, size=55, prob=0.004)
```

```
## [1] 0.1978352
```

2. $P(X = 4) = \frac{\binom{7}{4}\binom{5}{2}}{\binom{12}{6}}$