Basic Inferential Data Analysis

Michael Kroog

August 25, 2017

```
## looking at the structure of the data frame
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
## looking at a summary of the data frame
summary(ToothGrowth)
```

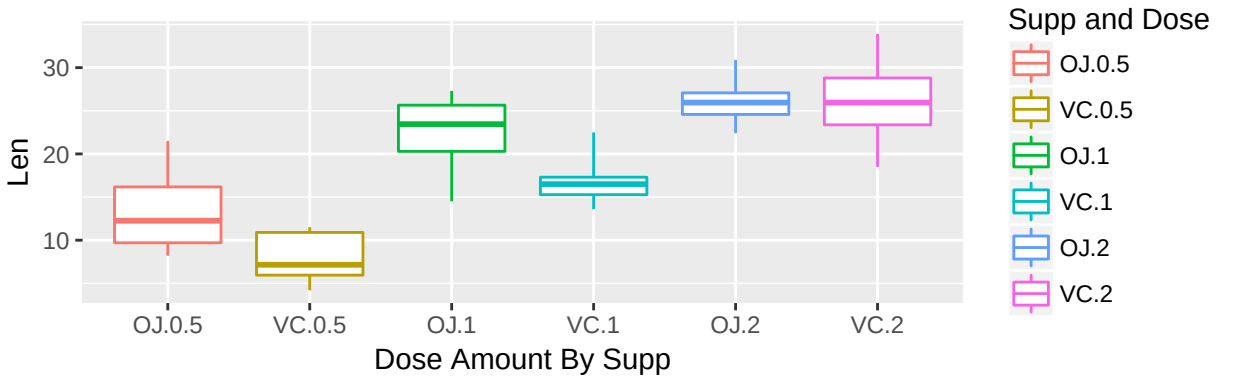
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
tooth <- ToothGrowth
tooth$suppdose <- interaction(ToothGrowth$supp, ToothGrowth$dose)

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.1
```

```
ggplot(tooth, aes(x = suppdose, y = len)) + geom_boxplot(coef = 3, aes(color = suppdose)) +
                                            xlab("Dose Amount By Supp") + ylab("Len") +
                                            labs(color = "Supp and Dose")
```



A little EDA of the structure and summary show we are looking at a data frame of 60 observation of 3 variables, len and dose are numeric while supp is a Factor with 2 levels. The summary is broken down by supp and dose and ploted above. The plot shows that for dose of 0.5 and 1 the median of OJ shows longer tooth growth, while for a dose of 2 the median of both are about equal, but VC has a longer range.

```
## splitting the dataframe by dose
toothall <- split.data.frame(ToothGrowth, f = ToothGrowth$dose)
list2env(toothall, envir = .GlobalEnv)
```

```
## <environment: R_GlobalEnv>
```

```
tooth.5 <- `0.5`
tooth1 <- `1`
tooth2 <- `2`

## finding the variance of the len by dose
var(tooth.5[1:10, 1])
```

```
## [1] 7.544
```

```
var(tooth.5[11:20, 1])
```

```
## [1] 19.889
```

```
var(tooth1[1:10, 1])
```

```
## [1] 6.326778
```

```
var(tooth1[11:20, 1])
```

```
## [1] 15.29556
```

```
var(tooth2[1:10, 1])
```

```
## [1] 23.01822
```

```
var(tooth2[11:20, 1])
```

```
## [1] 7.049333
```

```
## preforming a t test by dose
t.test(len ~ supp, data = tooth.5, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.003179
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.34604     Inf
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
t.test(len ~ supp, data = tooth1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70             16.77
```

```
t.test(len ~ supp, data = tooth2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

In order to perform our test and draw conclsions we need to assume that the data is IID and approximatley normally distributed, or at least not skewed. After conducting EDA we will compare the variances to determine if they are equal or unequal. Since for all three dose measurments the variances are unequal we will use the defualt var.equal = FALSE for our t test. Also since the data is not paired we will use the defualt paired = FALSE.

We will state the Null Hypothesis for each dose (0.5, 1 and 2) is the difference of mean len is equal to 0 and the Alternative Hypothesis is the difference of mean len is not equal to 0. As the results above show for the dose of 0.5 and 1 the p-values are both less than alpha of 0.05 and therefore we would reject the Null and for the dose of 2 the p-value is greater than alpha so we would accept the null. Looking at the boxplot below we can see this visually by looking at the mean which is represented by the black dashed line.

```
ggplot(tooth, aes(x = suppdose, y = len)) + geom_boxplot(coef = 3, aes(color = suppdose)) +
                                            xlab("Dose Amount By Supp") + ylab("Len") +
                                            labs(color = "Supp and Dose") +
                                            stat_summary(fun.y = mean, geom = "errorbar",
                                                    aes(ymax = ..y.., ymin = ..y.., width = .75),
                                                    linetype = "dashed")
```