Statistical Inference Week 4 Assignment Part 2

Kaitlyn McGrew

1/16/2018

Part 2: Basic Inferential Data Analysis

This is part 2 of the final assignment for the coursera course statistical inference part of the data science specialization. We will compare tooth growth by supp and dose from the Tooth Growth data provided in the R datasets Package.

1. Load the ToothGrowth data and perform some basic exploratory data analyses

```
library(datasets)
library(ggplot2)
colnames(ToothGrowth)
```

```
## [1] "len"  "supp" "dose"
```

```
head(ToothGrowth)
```

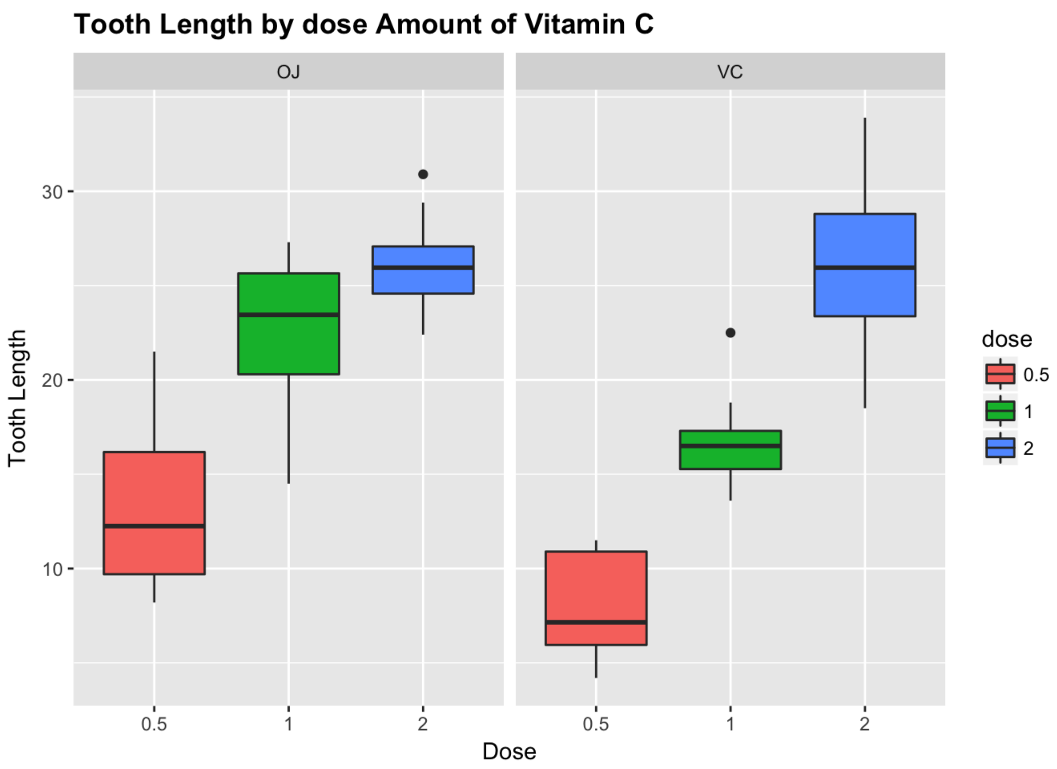
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

2. Provide a basic summary of the data

```
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  
```

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(aes(x=dose, y=len), data = ToothGrowth) +
        geom_boxplot(aes(fill=dose)) +
        ggtitle("Tooth Length by dose Amount of Vitamin C") +
        xlab("Dose") +
        ylab("Tooth Length") +
        facet_grid(~supp) +
        theme(plot.title = element_text(lineheight = .9, face = "bold"))
```



3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (only use the techniques from class, even if there’s other approaches worth considering)

perform an ANOVA

```
anova <- aov(len ~ supp * dose, data = ToothGrowth)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *  
## Residuals   54  712.1    13.2                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the results show all 3 categories including interaction between variables as having a P-value of >0.05 then a tukeyHSD test is required.

```
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp * dose, data = ToothGrowth)
##
## $supp
##       diff       lwr       upr     p adj
## VC-OJ -3.7 -5.579828 -1.820172 0.0002312
##
## $dose
##         diff       lwr       upr   p adj
## 1-0.5  9.130  6.362488 11.897512 0.0e+00
## 2-0.5 15.495 12.727488 18.262512 0.0e+00
## 2-1    6.365  3.597488  9.132512 2.7e-06
##
## $`supp:dose`
##                diff        lwr        upr     p adj
## VC:0.5-OJ:0.5 -5.25 -10.048124 -0.4518762 0.0242521
## OJ:1-OJ:0.5    9.47   4.671876 14.2681238 0.0000046
## VC:1-OJ:0.5    3.54  -1.258124  8.3381238 0.2640208
## OJ:2-OJ:0.5   12.83   8.031876 17.6281238 0.0000000
## VC:2-OJ:0.5   12.91   8.111876 17.7081238 0.0000000
## OJ:1-VC:0.5   14.72   9.921876 19.5181238 0.0000000
## VC:1-VC:0.5    8.79   3.991876 13.5881238 0.0000210
## OJ:2-VC:0.5   18.08  13.281876 22.8781238 0.0000000
## VC:2-VC:0.5   18.16  13.361876 22.9581238 0.0000000
## VC:1-OJ:1     -5.93 -10.728124 -1.1318762 0.0073930
## OJ:2-OJ:1      3.36  -1.438124  8.1581238 0.3187361
## VC:2-OJ:1      3.44  -1.358124  8.2381238 0.2936430
## OJ:2-VC:1      9.29   4.491876 14.0881238 0.0000069
## VC:2-VC:1      9.37   4.571876 14.1681238 0.0000058
## VC:2-OJ:2      0.08  -4.718124  4.8781238 1.0000000
```

The tukey test shows a significant difference between supp and dose only all other interactions are not significant.

State your conclusions and the assumptions needed for your conclusions. There is clearly a correlation between tooth growth and an increase in vitamin c. However the difference between dosing method, either supplement or through orange juice isn’t as clear from the graphs but was found to not be significant. Assumptions made are that this sample is a proper representation of the population in question, the assignment for categories was random and that the distribution of the means is normal.