Report for basic inferential data analysis

NOTE: I understand that "Each pdf report should be no more than 3 pages with 3 pages of supportimg appendix material if needed (code, figures, etcetera).". However, I find it is easier to follow the report if all the appendix material is integrated with the conclusions, since I'm "kniting" the R markdown file. Thanks for understanding, hope you enjoy :)

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

The data "ToothGrowth" was loaded into the R environment

```
data(ToothGrowth)
```

I started by looking at the data structure

```
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 ...
```

So the data is a data.frame with 60 observations and 3 variables. These variables are len, supp, dose.

I had a quick look to the data to see its organization

```
head(ToothGrowth,3)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
```

```
tail(ToothGrowth,3)
```

```
##     len supp dose
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

After having a general look at the data, I asked if there are any many missing values in the data, which could influence the analysis

```
sum(is.na(ToothGrowth))
```

```
## [1] 0
```

With no mising values, I started to explore the statistics summary

```
summary(ToothGrowth)
```

```
##       len        supp         dose      
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00  
##  Mean   :18.8           Mean   :1.17  
##  3rd Qu.:25.3           3rd Qu.:2.00  
##  Max.   :33.9           Max.   :2.00
```

And as a visual representation visual representation

```
par(mfrow=c(1,2))
boxplot(ToothGrowth$len,ylab="Tooth length")
hist(ToothGrowth$len,xlab="Tooth length",main="")
```

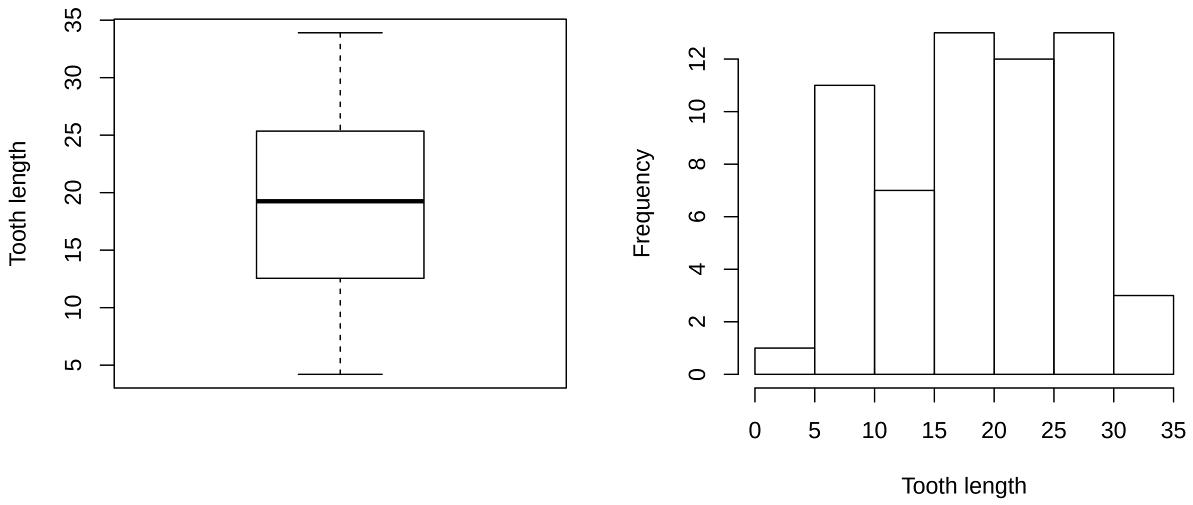

FIGURE 1: Exploratory representation of the teeth length

The data seems to be fairly normally distributed. However, the collection of lengths is not meaningful by itself, since it could depend on a variety of other measured variables, namely the supplement and its dose

```
par(mfrow=c(1,1))
boxplot(ToothGrowth$len~ToothGrowth$dose+ToothGrowth$supp,xlab="Dose/Supplement",ylab="Tooth length")
```

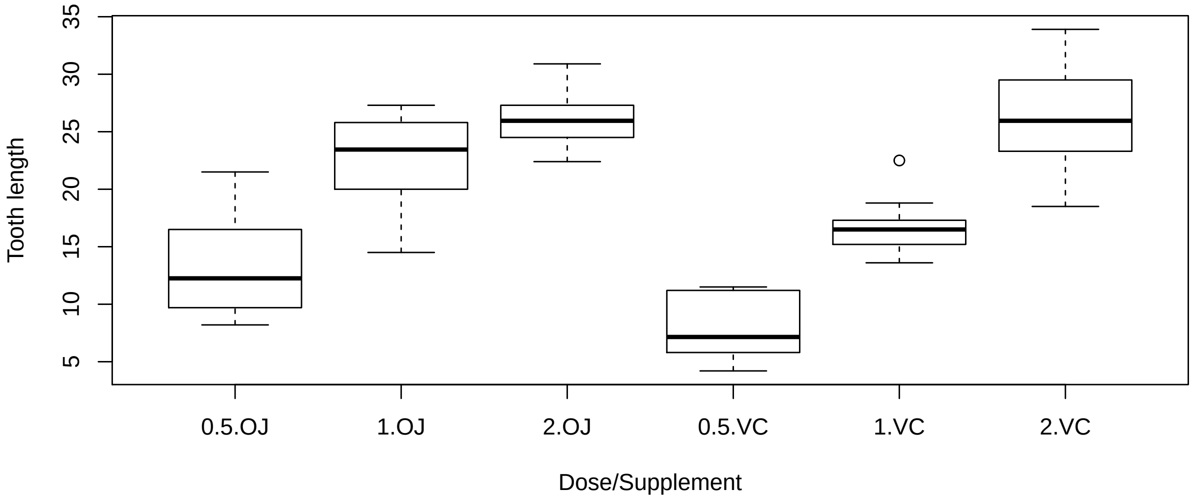

FIGURE 2: Representation of the teeth length by the various variables studied

It looks like with an increase of dosage of the supplement is associated with higher tooth length.

2. Provide a basic summary of the data.

To compare this different conditions, its important to know what are the summaries for each

```
by(ToothGrowth$len,list(ToothGrowth$supp,ToothGrowth$dose),summary)
```

```
## : OJ
## : 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.2     9.7    12.2    13.2    16.2    21.5
## --------------------------------------------------------
## : VC
## : 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20    5.95    7.15    7.98   10.90   11.50
## --------------------------------------------------------
## : OJ
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    14.5    20.3    23.5    22.7    25.6    27.3
## --------------------------------------------------------
## : VC
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    13.6    15.3    16.5    16.8    17.3    22.5
## --------------------------------------------------------
## : OJ
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    22.4    24.6    26.0    26.1    27.1    30.9
## --------------------------------------------------------
## : VC
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.5    23.4    26.0    26.1    28.8    33.9
```

3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

To statistically infer from the exploratory analysis, I started by doing a T-test on the various conditions. This test assumes that the populations estimated are normally distributed and that the collected samples are representative independent and identically distributed random variables from these populations. In this particular case, it seems plausible that the population of teeth lengths have a the same variace.

So, I'm interested in understaning if the mean values estimated for each supplement are statistically different

```
#Test OJ vs VC
t.test(ToothGrowth[ToothGrowth$supp=="OJ",1],ToothGrowth[ToothGrowth$supp=="VC",1],var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$supp == "OJ", 1] and ToothGrowth[ToothGrowth$supp == "VC", 1]
## t = 1.915, df = 58, p-value = 0.06039
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.167  7.567
## sample estimates:
## mean of x mean of y
##     20.66     16.96
```

It seems that the difference for supplements, regardless of the used dosage, is not significant at a level of alpha=0.05. This is confirmed by the **p.value 0.0604**, which is greater than alpha, and thus I fail to reject the null hypothesis that the mean of the tooth length is the same for both supplements. **The confidence interval obtained is from -0.167 to 7.567**, which contains the value 0. Altoghter, this suggests that there is a higher than 5% change of getting a mean value of **20.6633**, if the true population mean is **16.9633** (also indicated by the t-statistic **1.9153**, which shows that 20.6633 is less than 2 standard deviations away from the mean)

However, as I noticed before, it seems from the boxplots that higher doses result in higher tooth length. As such, I tested the hypotheses that differente doages result in differente tooth lengths.

```
#Test 1 vs 0.5
t.test(ToothGrowth[ToothGrowth$dose==1,1],ToothGrowth[ToothGrowth$dose==0.5,1],var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$dose == 1, 1] and ToothGrowth[ToothGrowth$dose == 0.5, 1]
## t = 6.477, df = 38, p-value = 1.266e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.276 11.984
## sample estimates:
## mean of x mean of y
##     19.73     10.61
```

```r
#Test 2 vs 0.5
t.test(ToothGrowth[ToothGrowth$dose==2,1],ToothGrowth[ToothGrowth$dose==0.5,1],var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$dose == 2, 1] and ToothGrowth[ToothGrowth$dose == 0.5, 1]
## t = 11.8, df = 38, p-value = 2.838e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.84 18.15
## sample estimates:
## mean of x mean of y
##     26.10     10.61
```

We can see that both p-values are very small, indicating that **doses of 1 or 2 result in increased effect on tooth length compared to dose of 0.5**, regardeless of the supplement.

It is clear that multiple comparisons are required to explore all the possible effects. As such, simply adding extra t-tests is going to lead to an increase in false positives. Consequently, I did a multivariable analysis by taking the False Discovery Rate (FDR) *(this is explained in week 4 of the course so I'm not sure if should be included in the report)*. This method was chosen instead of Bonferroni because it is less conservative, so more relationships can be infered from the data, which can be used in further studies

```r
#Take the length for every dose/supplement combination
oj05<-ToothGrowth[ToothGrowth$supp=="OJ"&ToothGrowth$dose=="0.5",1]
oj1<-ToothGrowth[ToothGrowth$supp=="OJ"&ToothGrowth$dose=="1",1]
oj2<-ToothGrowth[ToothGrowth$supp=="OJ"&ToothGrowth$dose=="2",1]
vc05<-ToothGrowth[ToothGrowth$supp=="VC"&ToothGrowth$dose=="0.5",1]
vc1<-ToothGrowth[ToothGrowth$supp=="VC"&ToothGrowth$dose=="1",1]
vc2<-ToothGrowth[ToothGrowth$supp=="VC"&ToothGrowth$dose=="2",1]

#Individually test all conditions
tests<-data.frame()
tests<-rbind(tests,"oj05 - oj1"=t.test(oj05,oj1,var.equal = TRUE)$p.value)
tests<-rbind(tests,"oj05 - oj2"=t.test(oj05,oj2,var.equal = TRUE)$p.value)
tests<-rbind(tests,"oj1 - oj2"=t.test(oj1,oj2,var.equal = TRUE)$p.value)
tests<-rbind(tests,"vc05 - vc1"=t.test(vc05,vc1,var.equal = TRUE)$p.value)
tests<-rbind(tests,"vc05 - vc2"=t.test(vc05,vc2,var.equal = TRUE)$p.value)
tests<-rbind(tests,"vc1 - vc2"=t.test(vc1,vc2,var.equal = TRUE)$p.value)
tests<-rbind(tests,"vc05 - oc05"=t.test(vc05,oj05,var.equal = TRUE)$p.value)
tests<-rbind(tests,"vc1 - oc1"=t.test(vc1,oj1,var.equal = TRUE)$p.value)
tests<-rbind(tests,"vc2 - oc2"=t.test(vc2,oj2,var.equal = TRUE)$p.value)

#Organize the data
tests<-cbind(tests,rownames(tests))
tests<-tests[,c(2,1)]
```

```
rownames(tests)<-NULL
names(tests)<-c("Comparison","p-value")
tests
```

```
##    Comparison   p-value
## 1  oj05 - oj1 8.358e-05
## 2  oj05 - oj2 3.402e-07
## 3   oj1 - oj2 3.736e-02
## 4  vc05 - vc1 6.492e-07
## 5  vc05 - vc2 4.957e-09
## 6   vc1 - vc2 3.398e-05
## 7 vc05 - oc05 5.304e-03
## 8   vc1 - oc1 7.807e-04
## 9   vc2 - oc2 9.637e-01
```

```
#Find the FDR
m=nrow(tests)
alpha=0.05

tests<-tests[order(tests[,2]),]
tests<-cbind(tests,"i"=1:m) # Attribute the rank
tests<-cbind(tests,"i/m"=tests[,3]/m)# calculate i/m
tests<-cbind(tests,"i/m*alpha"=tests[,4]*alpha)# Calculate i/m*a
tests<-cbind(tests,"FDR"=tests[,2]<tests[,5]) #Check if p<i/m*a
tests
```

```
##    Comparison   p-value i    i/m i/m*alpha   FDR
## 5  vc05 - vc2 4.957e-09 1 0.1111  0.005556  TRUE
## 2  oj05 - oj2 3.402e-07 2 0.2222  0.011111  TRUE
## 4  vc05 - vc1 6.492e-07 3 0.3333  0.016667  TRUE
## 6   vc1 - vc2 3.398e-05 4 0.4444  0.022222  TRUE
## 1  oj05 - oj1 8.358e-05 5 0.5556  0.027778  TRUE
## 8   vc1 - oc1 7.807e-04 6 0.6667  0.033333  TRUE
## 7 vc05 - oc05 5.304e-03 7 0.7778  0.038889  TRUE
## 3   oj1 - oj2 3.736e-02 8 0.8889  0.044444  TRUE
## 9   vc2 - oc2 9.637e-01 9 1.0000  0.050000 FALSE
```

4. State your conclusions and the assumptions needed for your conclusions.

The assumptions needed for the conclusions are:

- The populations estimated are normally distributed, regardeless of supplement and dose
- The collected samples are representative independent and identically distributed random variables from these populations.
- The population of teeth lengths have a the same variance, regardeless of supplement and dose

The conclusions are as follow:

- Increasing the dose of any supplement results in increase teeth length
- At doses 0.5 and 1, OC supplement is more efficient than VC supplement, resulting in an average 1.7 and 1.4 times longer teeth, respectively
- At dose 2, both treatments are equally efficient