Statistical inference assignment: part 2

MJM Beuken

5 november 2017

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

```
## load packages and ToothGrowth dataset.
library(datasets)
library(ggplot2)
```

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

```
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 ...
```

```
data <- ToothGrowth
```

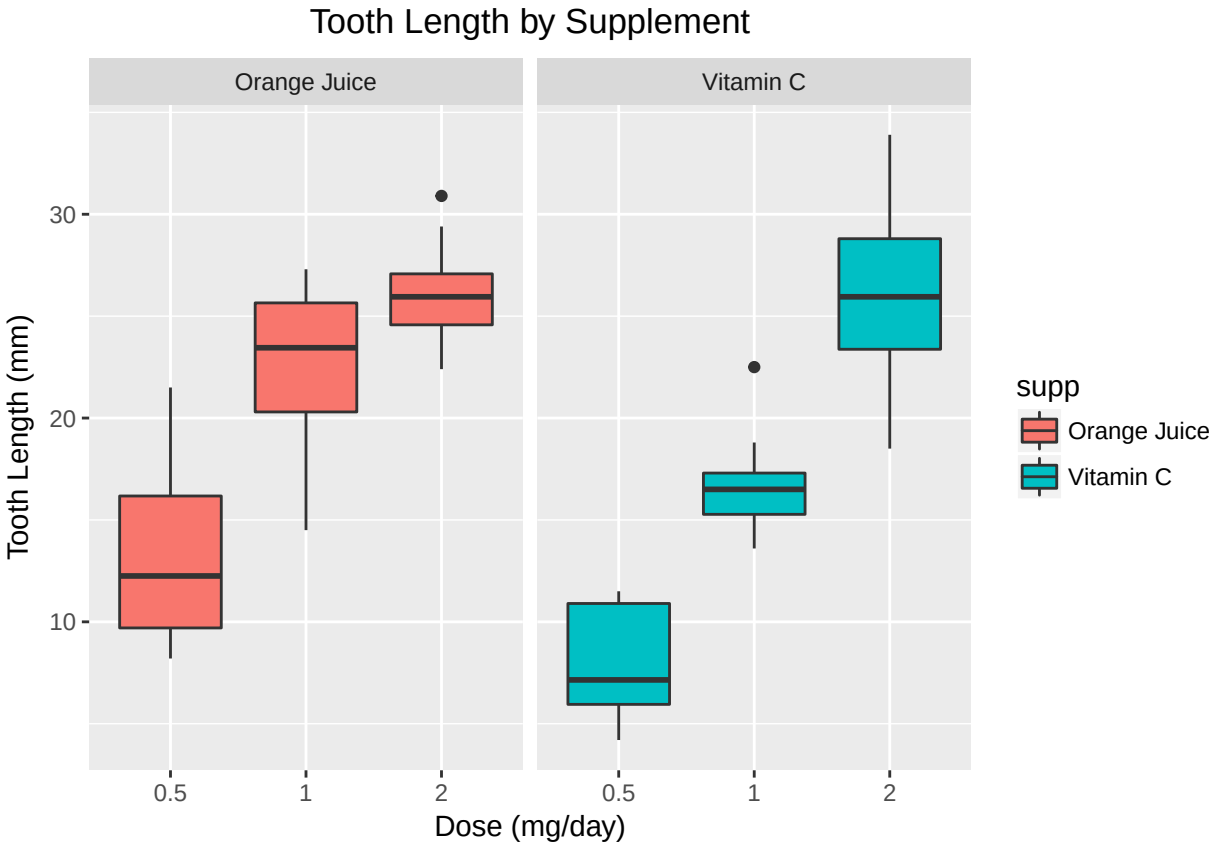
Provide a basic summary of the data.

```
summary(data)
```

```
##       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
```

```
## explore data with boxplots

levels(data$supp) <- c("Orange Juice", "Vitamin C")
g <- ggplot(data, aes(x = factor(dose), y = len))+
                facet_grid(.~supp)+
                geom_boxplot(aes(fill = supp))+
                labs(title = "Tooth Length by Supplement")+
                labs(x = "Dose (mg/day)", y = "Tooth Length (mm)")+
                theme(plot.title = element_text(hjust = 0.5))
print(g)
```



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).

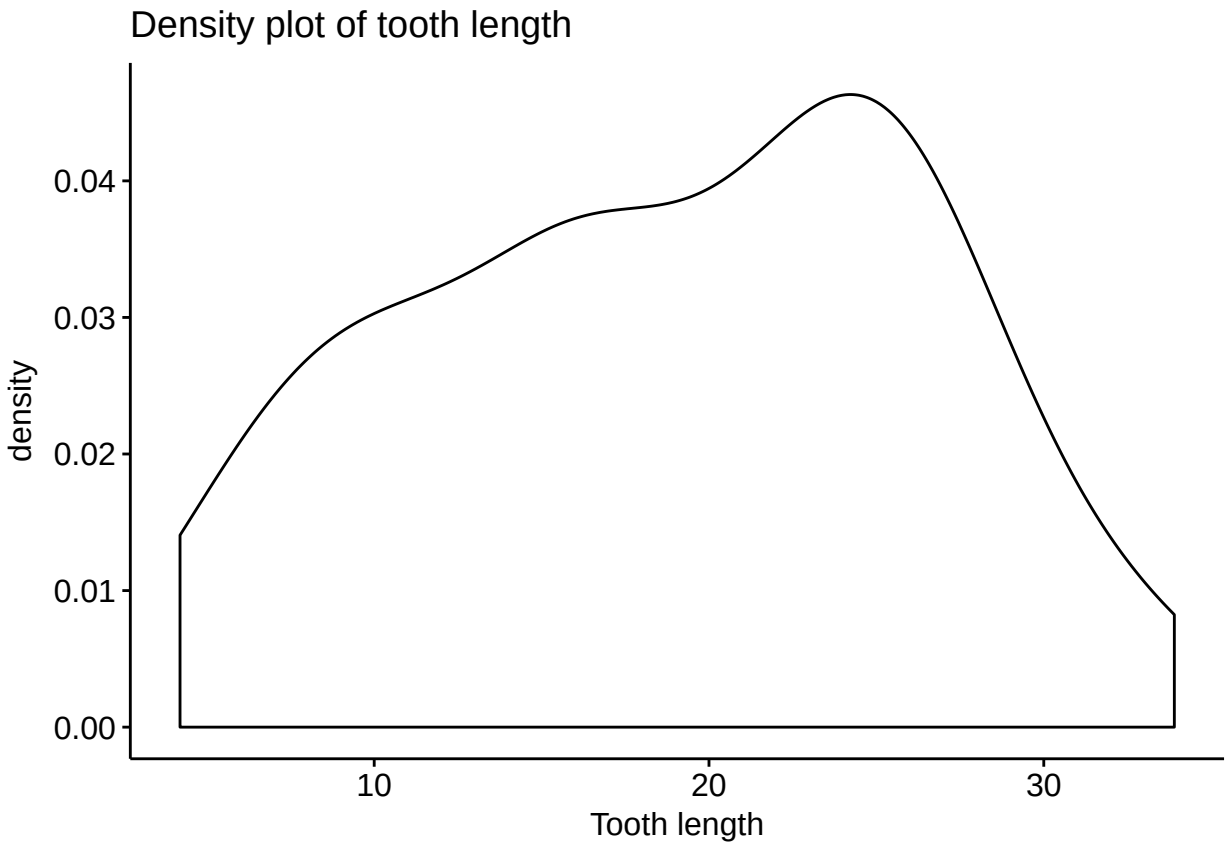
Checking if data Toothgrowth follows a Normal distribution before using the t-distribution.

```
library("ggpubr")
```

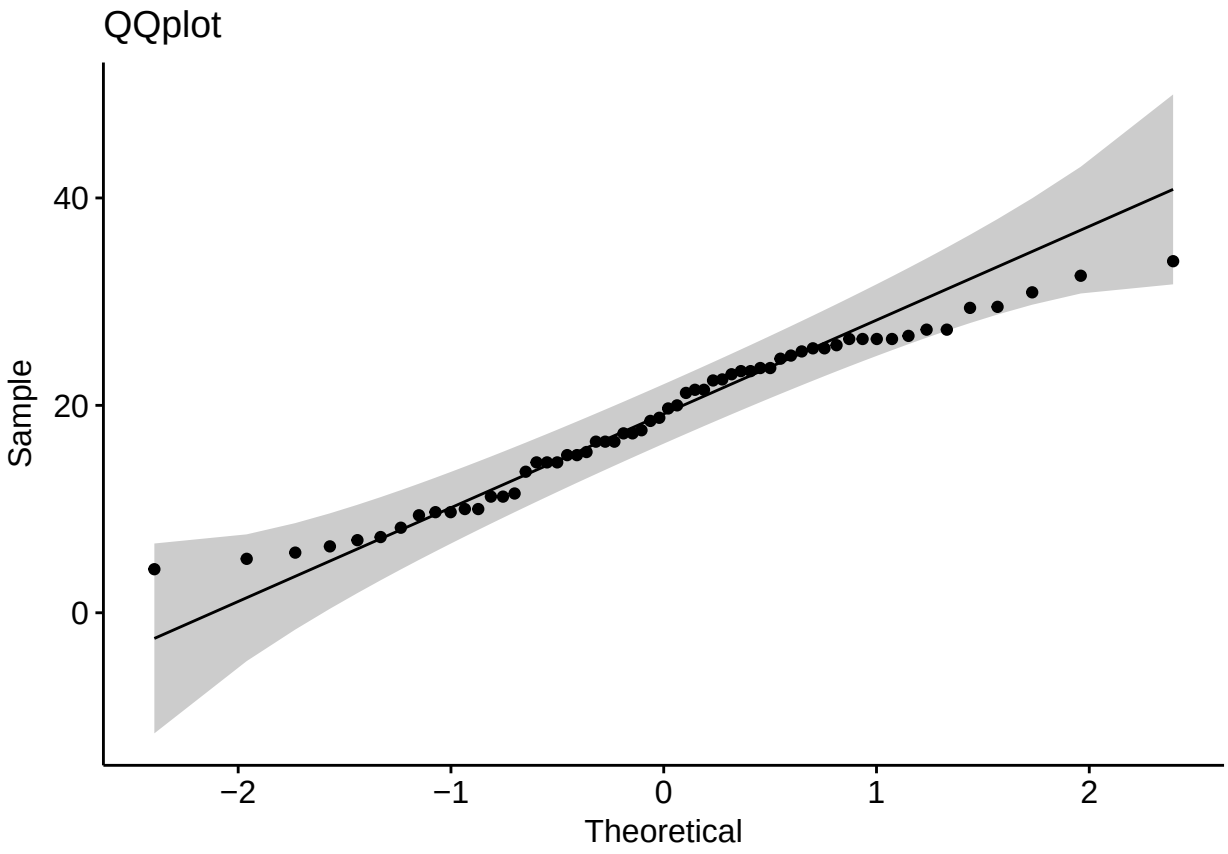
```
## Warning: package 'ggpubr' was built under R version 3.4.2
```

```
## Loading required package: magrittr
```

```
ggdensity(data$len, main = "Density plot of tooth length", xlab = "Tooth length")
```




```
ggqqplot(data$len) + ggtitle("QQplot")
```



As all the points lay approximately along the reference line, we can assume normality hence we can use the two sample t-test.

Hypothesis testing

Testing average tooth growth overall (without looking at different dosage).Considering alpha being 5%

Ho: mu_oj - mu_VC = 0

Ha: mu_oj - mu_VC ??? 0

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

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group Orange Juice    mean in group Vitamin C
##                   20.66333                   16.96333
```

the 95% confidence interval does contain nill. So the Ho will be failed to be rejected. With a 5%-probability of making a error of type 1 we could say that the average tooth growth with Orange juice is equal to the average tooth growth with vitamin C.

Testing three different hypothesis stating for a dose of x mg/day both supplements deliver same average growth with two sample t-test considering alpha being 5%.

For dose = 0.5 mg a day

Ho: mu_oj - mu_VC = 0

Ha: mu_oj - mu_VC ??? 0

```
t.test(len ~ supp, data = subset(data, dose == 0.5))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group Orange Juice    mean in group Vitamin C
##                      13.23                       7.98
```

the 95% confidence interval does not contain nill. So the Ho wil be rejected in favor of the Ha. With a 5%-probability of making a error of type 1 we could say that the average tooth growth due to Orange juice is significant larger than the average tooth growth due to vitamin C.

For dose = 1 mg a day

Ho: mu_oj - mu_VC = 0

Ha: mu_oj - mu_VC ??? 0

```
t.test(len ~ supp, data = subset(data, dose == 1))
```

```
##
##  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 Orange Juice    mean in group Vitamin C
##                      22.70                      16.77
```

the 95% confidence interval does not contain nill. So the Ho wil be rejected in favor of the Ha. With a 5%-probability of making a error of type 1 we could say that the average tooth growth due to Orange juice is significant larger than the average tooth growth due to vitamin C.

For dose = 2 mg a day

Ho: mu_oj - mu_VC = 0

Ha: mu_oj - mu_VC ??? 0

```
t.test(len ~ supp, data = subset(data, dose == 2))
```

```
##
##  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 Orange Juice    mean in group Vitamin C
##                      26.06                      26.14
```

the 95% confidence interval does contain nill. So the Ho will be failed to be rejected. With a 5%-probability of making a error of type 1 we could say that the average tooth growth due to Orange juice is equal to the average tooth growth due to vitamin C.

Conclusion:

We could state that, with smaller dosage (0.5 or 1 mg per day), Orange juice leads to a significant larger average tooth growth in comparison to vitamin C. For the larger dosage (2 mg per day) this difference can not be pointed out. We can not say that the average tooth growth is larger for Orange juice in comparison to vitamin C. Overall (despite different dosage) we can not speak of a significant larger average tooth growth due to Orange juice in comparison to vitamin C.

Assuming that:

*** The sample data set is representative for the population dataset.**

*** Dosage and supplement were randomly assigned.**

*** The distribution of the means is actually normal distributed.**