Statistical inference, Course Project, Part 2

Assignment text in *italic*.

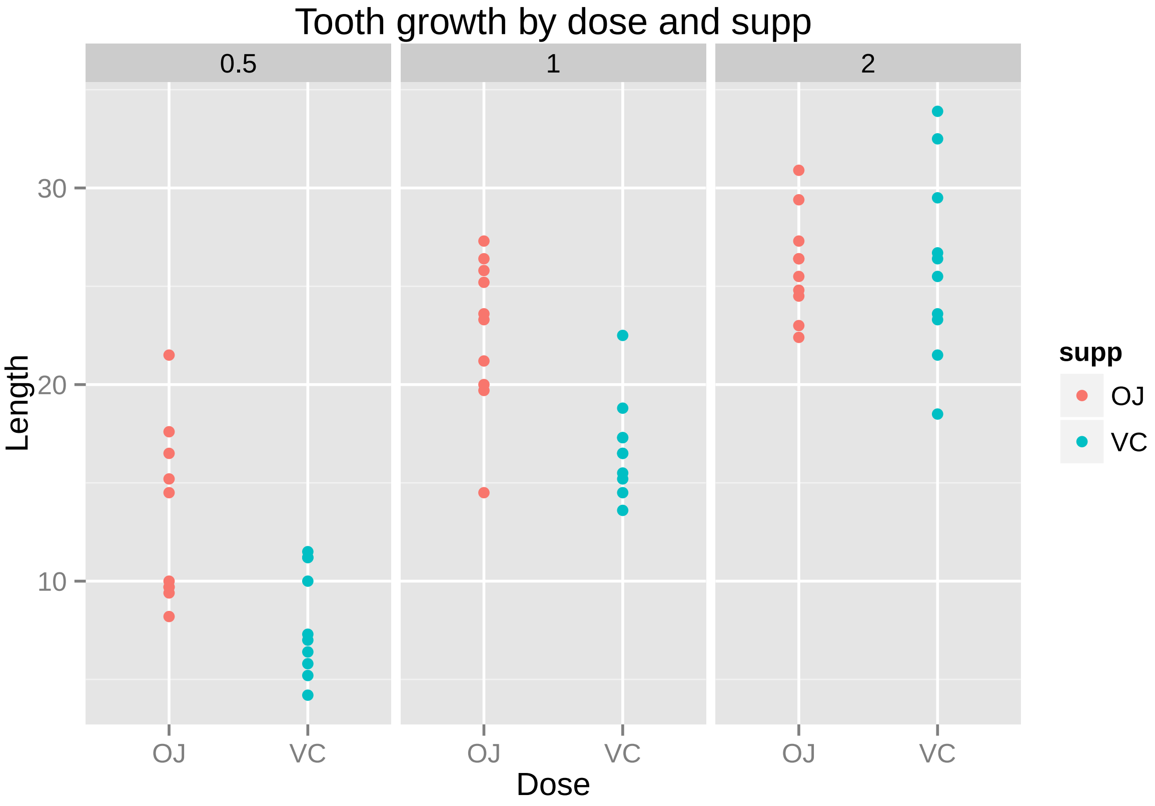
Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.

Question 1

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

```r
# Load data set containing the ToothGrowth data
library(datasets)
# Load graphics package
library(ggplot2)

# Draw plot to graphically explore the data
ggplot(data = ToothGrowth, aes(x=supp, y=len, group=supp, color=supp)) +
    ggtitle("Tooth growth by dose and supp") +
    xlab("Dose") +
    ylab("Length") +
    geom_point() +
    facet_grid(. ~ dose)
```



Based on the plot, we can draw the following conclusios:
- Dose is positively correlated with tooth growth for both supp types
- OJ has a stronger effect on growth than VC for lower doses 0.5 and 1.0, but not (necessarily) with dose 2.0
- The variance of OJ decreases with dose size, while the variance of VC increases

Question 2

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  
```

The basic data summary shows that
- the data set contains 60 observations of tooth length
- it is grouped by supp type and dose
- supp is either OJ och VC
- dose is 0.5, 1.0 or 2.0

The conclusions from question 1 about the distributions can also be drawn from the data summary below.

```
means <- aggregate(ToothGrowth$len, by=list(ToothGrowth$dose, ToothGrowth$supp), FUN = mean)
sds <- aggregate(ToothGrowth$len, by=list(ToothGrowth$dose, ToothGrowth$supp), FUN = sd)
summary <- cbind(means, sds$x)
colnames(summary) <- c("dose", "supp", "mean", "sd")
summary
```

```
##   dose supp  mean       sd
## 1  0.5   OJ 13.23 4.459709
## 2  1.0   OJ 22.70 3.910953
## 3  2.0   OJ 26.06 2.655058
## 4  0.5   VC  7.98 2.746634
## 5  1.0   VC 16.77 2.515309
## 6  2.0   VC 26.14 4.797731
```

Question 3

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)

We run two sample (non-paired) t-tests comparing the tooth length by supp and dose. A t-test is better than a z-test as we have small sample sizes when we group by supp (30) or dose (20).

Difference by supp

T-test comparing the difference in length by supp:

```
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ"], ToothGrowth$len[ToothGrowth$supp == "VC"])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ"] and ToothGrowth$len[ToothGrowth$supp == "VC"]
## 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 of x mean of y
##  20.66333  16.96333
```

The p-value of this test is 0.06063, which by commonly used rules for p-value interpretation would imply that there is a "low presumption" against the null hypothesis, which in this case is that there is no difference in the lengths between the data sets. We hence have support for the argument that OJ (mean length 20.66) has a stronger impact on growth than VC (mean length 16.96). It should be noted that this test does not consider the impact of the dose. As we concluded earlier, dose has an impact on the variance of both OJ and VC sample means, but also on the relative differens on means.

T-test comparing the difference in length by supp, grouped by dose:

```r
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 0.5],
       ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose == 0.5])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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 of x mean of y
##     13.23      7.98
```

```r
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 1.0],
       ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose == 1.0])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## 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 of x mean of y
##     22.70     16.77
```

```r
t.test(ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 2.0],
       ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose == 2.0])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose ==  and ToothGrowth$len[ToothGrow
## t = -0.0461, 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 of x mean of y
##     26.06     26.14
```

These tests support our previous conclusions:
- Dose 0.5 : p-value 0.006 = weak support for a statistically significant difference in growth
- Dose 1.0 : p-value 0.001 = strong support for a statistically significant difference in growth
- Dose 2.0 : p-value 0.96 = no support for a statistically significant difference in growth

Difference by dose

T-test comparing the difference in length by dose:

```r
t.test(ToothGrowth$len[ToothGrowth$dose == 0.5], ToothGrowth$len[ToothGrowth$dose == 1.0])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 0.5] and ToothGrowth$len[ToothGrowth$dose == 1]
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

```r
t.test(ToothGrowth$len[ToothGrowth$dose == 1.0], ToothGrowth$len[ToothGrowth$dose == 2.0])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 1] and ToothGrowth$len[ToothGrowth$dose == 2]
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

```r
t.test(ToothGrowth$len[ToothGrowth$dose == 0.5], ToothGrowth$len[ToothGrowth$dose == 2.0])
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 0.5] and ToothGrowth$len[ToothGrowth$dose == 2]
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

The t-test results show:
- Dose 0.5 vs dose 1.0 : p-value 1.3e-07
- Dose 0.5 vs dose 1.0 : p-value 1.9e-05
- Dose 0.5 vs dose 1.0 : p-value 4.4e-14

All three p-values, by commonly used rules for p-value interpretation, would imply that there is a "very strong presumption" against the null hypothesis, which in this case is that there is no difference in lengths between the data sets. We hence have support for the argument that dose size is positively correlated with growth.

Question 4

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

Conslusions:
- Dose size is strongly correlated with tooth length
- The supp OJ has a higher impact on topth length than VC with dose sizes 0.5 and 1.0, but is not superior for dose size 2.0
- The variance of the supp VC increases with dose size while the variance of supp OJ decreases with dose size

Assumptions:
- Samples are approximately normal distributed, or t-distributed
- Observations are independent from each other
- Observations are free from observational error from potential confounding effects