Tooth Growth analysis

Xingmin Aaron Zhang

In this project, we will investigate the ToothGrowth dataset in R.

First, we load this datset to our working directory.

```
library(datasets)
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
```

We perform some exploratory analysis to get a better idea of what this dataset looks like. len is a numeric measurement. There are two supplements, VC and OJ, and three doeses, 0.5, 1 and 2.

```
ToothGrowth$dose = as.factor(ToothGrowth$dose)
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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

We also look at what combinations of supplements and doses were used in this experiment. Each supplement was used at three doses, 0.5, 1 and 2, so there were a total of 6 conditions.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ToothGrowth %>% select(supp, dose) %>% distinct()
```

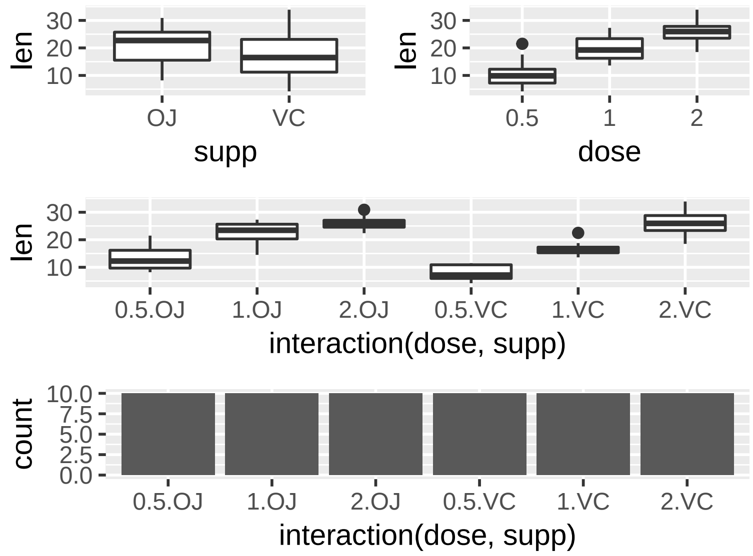

Figure 1: Figure 1 The lengh of tooth growth in response to different supplements at increasing doses

```
##   supp dose
## 1   VC  0.5
## 2   VC    1
## 3   VC    2
## 4   OJ  0.5
## 5   OJ    1
## 6   OJ    2
```

We use histograms to show how does length changes with supplement, dose, or the combination.

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
p1 <- ggplot(ToothGrowth) + geom_boxplot(aes(x = supp, y = len))
p2 <- ggplot(ToothGrowth) + geom_boxplot(aes(x = dose, y = len))
p3 <- ggplot(ToothGrowth) + geom_boxplot(aes(x = interaction(dose, supp), y = len))
p4 <- ggplot(ToothGrowth) + geom_histogram(aes(x = interaction(dose, supp)), stat = "count")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
grid.arrange(p1, p2, p3, p4, ncol = 2, layout_matrix = rbind(c(1, 2), c(3, 3), c(4,4)))
```

From the above analysis, we can summarize our findings as follows: 1. There are two suppelements, OJ and VC, and three doses, 0.5, 1 and 2. 2. There are 6 combinations of treatments, each supplement was used at three doses. Each combination had 10 replicates. 3. For different supplements, growth from OJ seems higher than VC. 4. growth increased with increasing doses.

We want to perform Student's T test to test whether tooth growth was significantly different in response to different supplements, or at different doses. First we fix the variable dose and only assess the effect of different supplements.

```r
s <- split(ToothGrowth$len, interaction(ToothGrowth$supp, ToothGrowth$dose))
names(s)
```

```
## [1] "OJ.0.5" "VC.0.5" "OJ.1"   "VC.1"   "OJ.2"   "VC.2"
```

```r
t.test(s$OJ.0.5, s$VC.0.5, paired = FALSE, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  s$OJ.0.5 and s$VC.0.5
## 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 of x mean of y
##     13.23      7.98
```

Result: The p value for observing such a t statistic or more extreme is only 0.003. So under 95% confidence we reject NULL hypothesis and accept alternative.

```r
t.test(s$OJ.1, s$VC.1, paired = FALSE, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  s$OJ.1 and s$VC.1
## t = 4.0328, df = 15.358, p-value = 0.0005192
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.356158      Inf
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

Result: The t statistic is in the 95 percent rejection interval (one-sided, actual $p < 0.01$). So reject NULL hypothesis and accept alternative.

```r
t.test(s$OJ.2, s$VC.2, paired = FALSE, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  s$OJ.2 and s$VC.2
## 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 of x mean of y
##     26.06     26.14
```

Result: The t statistic is not in the 95 percent confidence interval (two-sided, actual $p < 0.01$). So accept NULL hypothesis.

Next, we perform T test to test the effect of doses on tooth growth.

```
t.test(s$OJ.0.5, s$OJ.1, alternative = "less", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  s$OJ.0.5 and s$OJ.1
## t = -5.0486, df = 17.698, p-value = 4.392e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.214316
## sample estimates:
## mean of x mean of y
##     13.23     22.70
```

Result: ?

```
t.test(s$OJ.1, s$OJ.2, alternative = "less", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  s$OJ.1 and s$OJ.2
## t = -2.2478, df = 15.842, p-value = 0.0196
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.7486236
## sample estimates:
## mean of x mean of y
##     22.70     26.06
```

Result: The t statistic is in the 95 percent confidence interval (two-sided, actual $p < 0.01$). So accept NULL hypothesis.

```
t.test(s$OJ.0.5, s$OJ.2, alternative = "less", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  s$OJ.0.5 and s$OJ.2
## t = -7.817, df = 14.668, p-value = 6.619e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -9.94845
## sample estimates:
## mean of x mean of y
##     13.23     26.06
```

Result: The t statistic is not in the 95 percent confidence interval (two-sided, actual p < 0.01). So accept NULL hypothesis.

```r
t.test(s$VC.0.5, s$VC.1, alternative = "less", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  s$VC.0.5 and s$VC.1
## t = -7.4634, df = 17.862, p-value = 3.406e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.746867
## sample estimates:
## mean of x mean of y
##      7.98     16.77
```

Result: The t statistic is not in the 95 percent confidence interval (two-sided, actual p < 0.01). So accept NULL hypothesis.

```r
t.test(s$VC.1, s$VC.2, alternative = "less", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  s$VC.1 and s$VC.2
## t = -5.4698, df = 13.6, p-value = 4.578e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.346525
## sample estimates:
## mean of x mean of y
##     16.77     26.14
```

Result: The t statistic is not in the 95 percent confidence interval (two-sided, actual p < 0.01). So accept NULL hypothesis.

```r
t.test(s$VC.0.5, s$VC.2, alternative = "less", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  s$VC.0.5 and s$VC.2
## t = -10.388, df = 14.327, p-value = 2.341e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -15.08583
## sample estimates:
## mean of x mean of y
##      7.98     26.14
```

Result: The t statistic is not in the 95 percent confidence interval (two-sided, actual $p < 0.01$). So accept NULL hypothesis.