Hypothesis Analysis on ToothGrowth data set

Coursera Statistical Inference Course Project II

Xinyu W

7/27/2020

Overview

This project will investigate the significance of different dose and/or different delivery methods on the length of odontoblasts (cells responsible for tooth growth). There are three levels of doses of vitamin C (0.5, 1, and 2 mg/day), and two delivery methods (orage juice and ascorbic acid). This project involves t-test for hypothesis analysis.

Set Global Environment

```
library(knitr)
opts_chunk$set(warning = FALSE, message = FALSE,
               echo = TRUE)
library(ggplot2)
```

Load the ToothGrowth data and perform some basic exploratory data analyses

After loading the *ToothGrowth* dataset, we see it contains 60 observations of 3 variables - len, supp and doses. Looking into the dataset, we find first 30 observations from VC and the rest from OJ. Right now the dose column is numeric. But for the convenience of later comparison, we will convert it to factor.

```
#load the dataset ToothGrowth and show the basic info
library(datasets)
data("ToothGrowth")
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 ...
```

```
head(ToothGrowth,6)
```

```
##    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
```

```r
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

Provide a basic summary of the data

Again, by looking into the summary of the data set, we can see how many observations there are for each level of doses and each supplement type.

```r
#give a summary of ToothGrowth
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

```r
#explore the exact data of column dose and supp
table(ToothGrowth$dose)
```

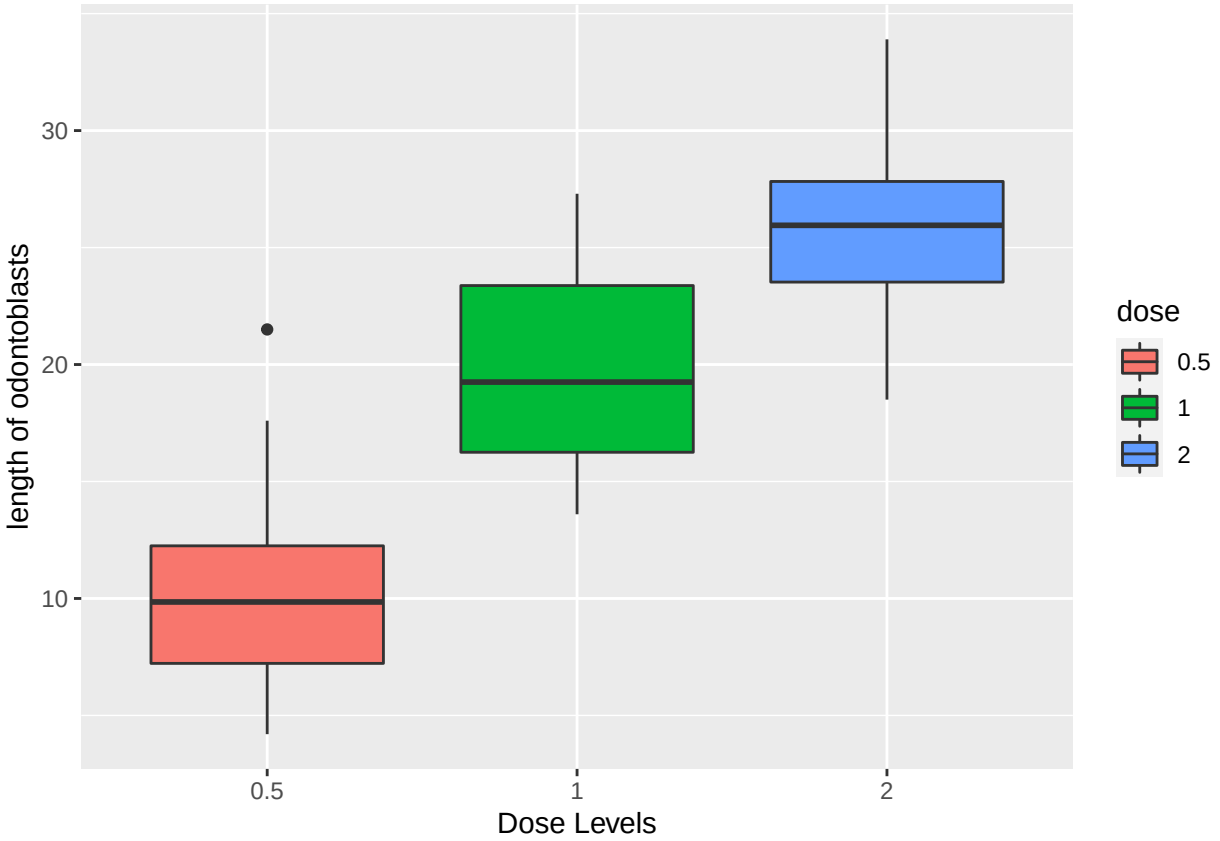
```
##
## 0.5   1   2
##  20  20  20
```

```r
table(ToothGrowth$supp)
```

```
##
## OJ VC
## 30 30
```

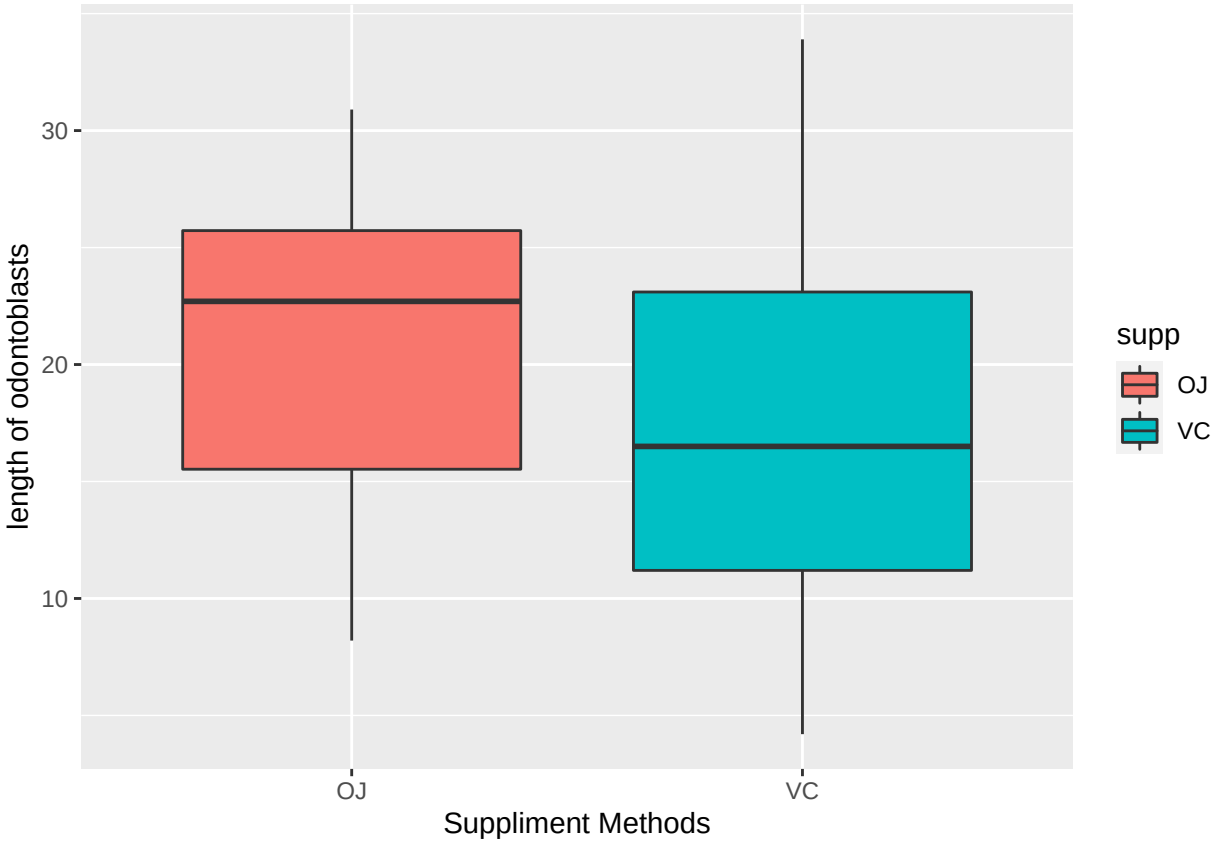
By plotting the *len* data according to different dose level, we can have a sense of how it distributes and changes. Here, we can approximately see that as dose increases, the length also increases. There is a relatively wider range in dose 1.0 than dose 0.5 and 2.0

```r
#plot the dose data according to len
library(ggplot2)
ggplot(aes(x=as.factor(dose), y = len), data = ToothGrowth) + geom_boxplot(aes(fill=dose)) + labs(x = "
```



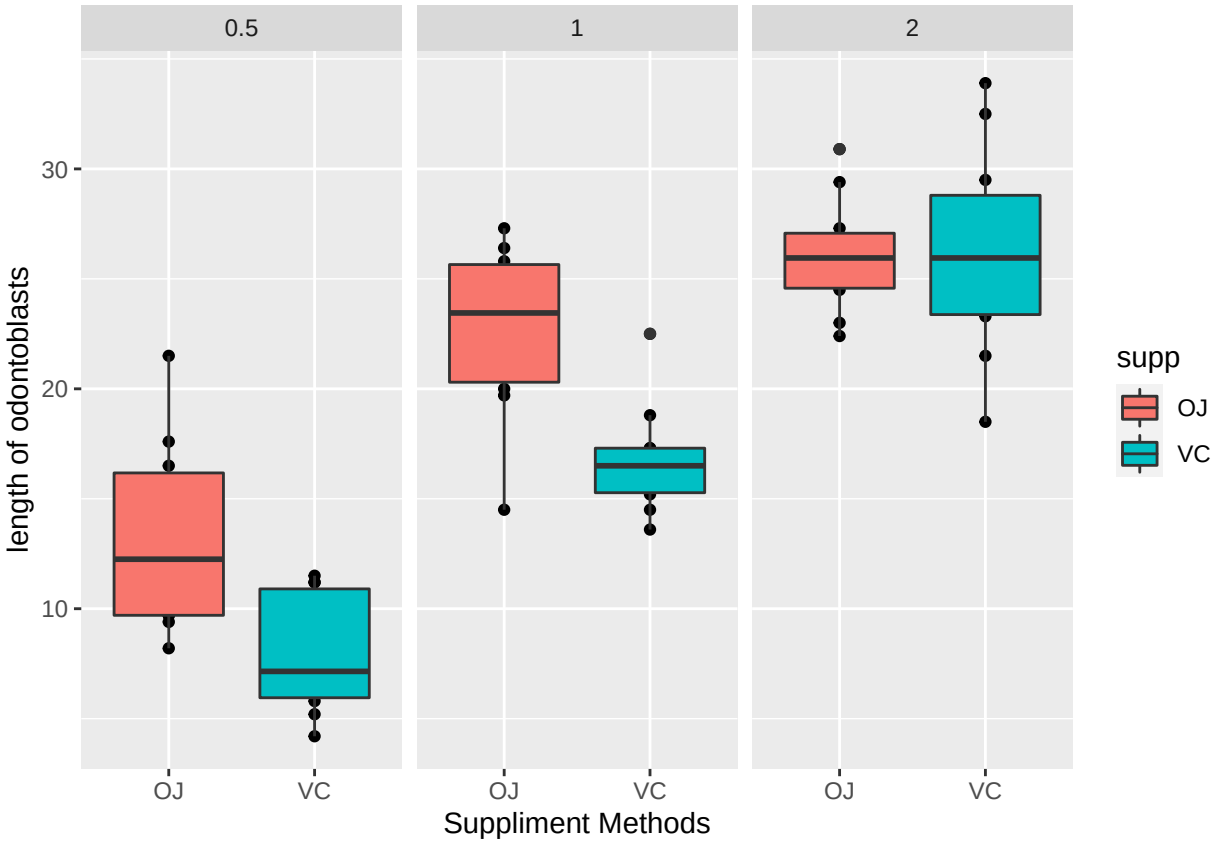
With the plot of suppliment methods, we see there's a great overlapping area between OJ and VC, which gives us a sense that these two methods *might* have lower significance in the length variable. We can later try to confirm this through t-test.

```
#plot the supp data according to len
ggplot(aes(x=supp, y = len), data = ToothGrowth) + geom_boxplot(aes(fill=supp)) + labs(x = "Suppliment
```



Here we generate the plot through different facets by dose. What's useful in showing these details is that we see different sizes of overlapping areas in the supp plots if we divide overall data by dose levels. This gives us a hint that the significance of supplicant methods might not be clear until we pay attention to the combining effect with dose levels.

```
#show an overview of both supp and dose data we got.
qplot(x = supp, y = len, data = ToothGrowth, facets = ~dose) + geom_boxplot(aes(fill=supp)) + labs(x =
```



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

In this project, we are going to take t-test and analyze through confidence intervals and p-values, since the number of observations of each variable we are going to compare for is under or equal to 30.

Compare between different supp

H0: The supplement type (VC or OJ) does NOT have any influence on tooth growth
Ha: The supplement type (VC or OJ) DOES have influence on tooth growth

```
t.test(len~supp, data = ToothGrowth, paired = FALSE, var.equal = FALSE)
```

```
##
##  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 OJ mean in group VC
##         20.66333         16.96333
```

Here, we see:

- The p-value is 0.06, which is larger than the significance level alpha (0.05)
- However, the confidence interval cross from -0.17 to 7.57 (from negative to positive), thus NOT clear whether we could reject **H0**

Compare among different dose

H0: The different dose levels(0.5,1,2) does NOT have any influence on tooth growth
Ha: The different dose levels(0.5,1,2) DOES have influence on tooth growth

```
dose <- ToothGrowth$dose
supp <- ToothGrowth$supp
len <- ToothGrowth$len
t.test(len[dose==0.5],len[dose==1], data=ToothGrowth,
       paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  len[dose == 0.5] and len[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
```

```
t.test(len[dose==0.5],len[dose==2], data=ToothGrowth,
       paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  len[dose == 0.5] and len[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
```

```
t.test(len[dose==1],len[dose==2], data=ToothGrowth,
       paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  len[dose == 1] and len[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
```

We clearly see that whether it's from 0.5 to 1, 1 to 2, or 0.5 to 2:

- The confidence interval is consistently negative.
- The p-value is so small (close to 0) and much smaller than alpha (0.05)

Thus, we can be pretty sure about the significance of dose levels on teeth growth, which means that we are going to REJECT our **H0**.

Further compare the supp accoding to different dose

Recalling what we saw on the overall plot of supplement methods by different dose levels, we need to take a further step in "stripping off" the effect of dose levels and exploring the effect from types of supplement.

```
TG_0.5 <- subset(ToothGrowth, dose == 0.5)
TG_1.0 <- subset(ToothGrowth, dose == 1)
TG_2.0 <- subset(ToothGrowth, dose == 2)
t.test(len~supp, data = TG_0.5, paired = FALSE, var.equal = FALSE)
```

```
##
##  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 OJ mean in group VC
##            13.23             7.98
```

```
t.test(len~supp, data = TG_1.0, paired = FALSE, var.equal = FALSE)
```

```
##
##  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 OJ mean in group VC
##            22.70            16.77
```

```r
t.test(len~supp, data = TG_2.0, paired = FALSE, var.equal = FALSE)
```

```
##
##  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 OJ mean in group VC
##            26.06            26.14
```

Here, we see:

- In dose levels 0.5 and 1, both confidence intervals are positive, but the p-value is too small that we would take this as extreme.
- In dose level 2.0, although the p-value is surprisingly 0.96, the confidence interval across from negative to positive. Thus, we can NOT accept **H0**.

State your conclusions and the assumptions needed for your conclusions

From all the plots, tests and analysis above, we here generate our conclusions:

- The supplement type has little or **NO** effect on the length of odontoblasts (tooth growth).
- The different dose levels DOES have effect on the length of odontoblasts (tooth growth). As dose level increases, the tooth growth increases.

And, we state our assumptions here

- This sample of guinea pigs is representative of the population.
- For each t-test, the variance is treated as different.
- The experiment along with its design and method is well-controled without any other significant factors to our analysis.