Statistical Inference Course Project - Part 2 & Appendix

Roshan Riazi

Overview

Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package. So first we read the help page for ToothGrowth data and then we will perform the 4 specified steps.

ToothGrowth Data Description

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

Step 1

In step 1, we should load the ToothGrowth data and perform some basic exploratory data analyses.

First we load ToothGrowth dataset:

```
library(datasets)
data(ToothGrowth)
```

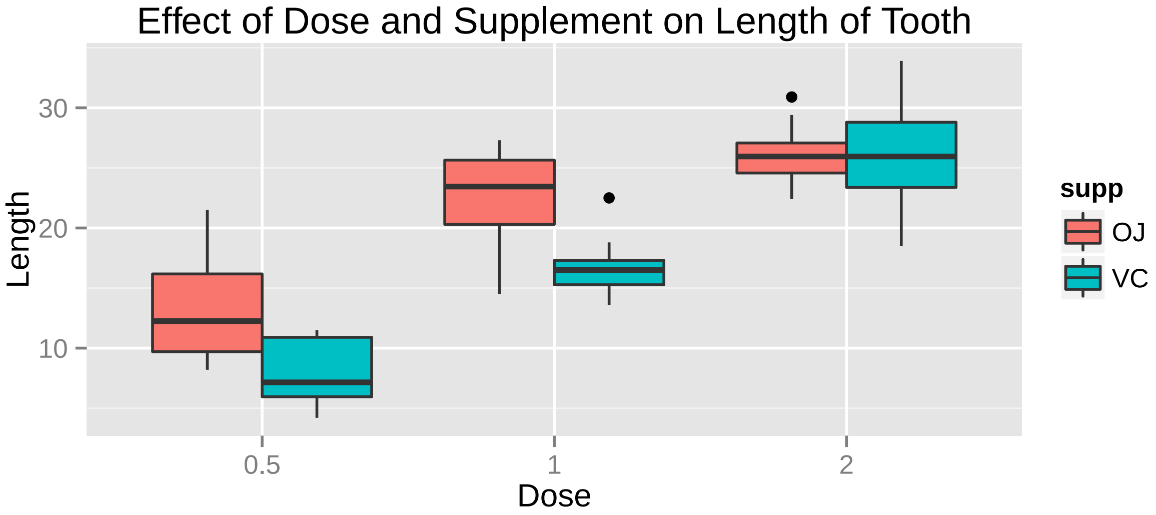
Then we can look at its structure, and change dose variable to factor.

```
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 ...
```

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

So we make a baxplot to better see the effect of dose and supplement on length of tooth:



There seems to be a relationship between dose and length, and some relationships between supplement and length for dose equal to 0.5 and 1.

Step 2

In step 2, we should provide a basic summary of the data.

We will look at its summary, and then calculate length's mean and standard deviation for the whole dataset to get a feel of its properties.

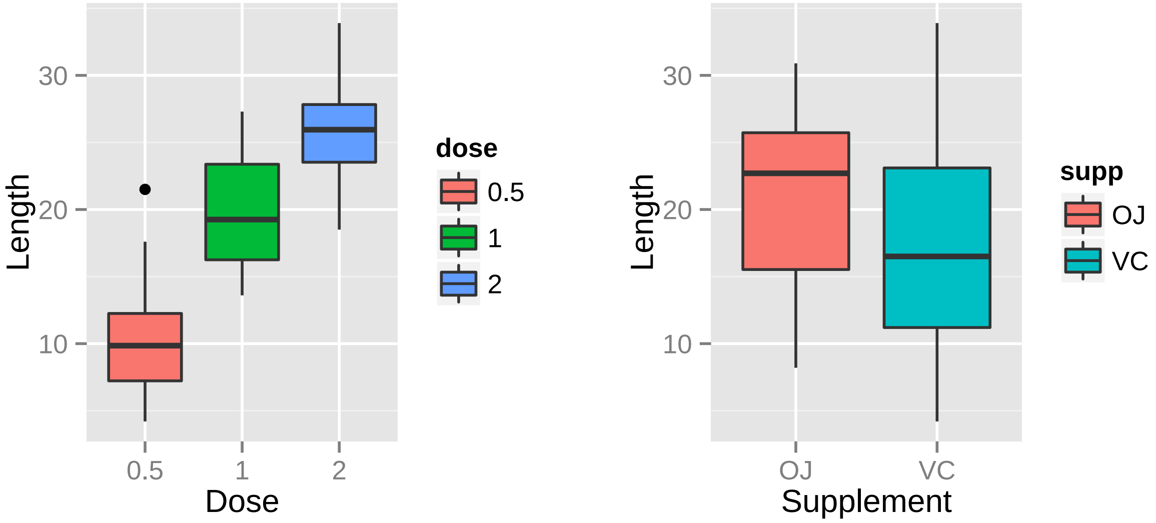
```
summary(ToothGrowth)
```

```
##       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
```

```
sd(ToothGrowth$len)
```

```
## [1] 7.649315
```

We can plot the seperate effect of dose on length, and the seperate effect of supplement on length.



As this plot shows the seperate effect of dose and supplement on length, it will help us in developing hypothesis for step 3.

Lets see the mean of length for different dose groups:

```
tapply(ToothGrowth$len, INDEX = list(ToothGrowth$dose), FUN = mean)
```

```
##    0.5      1      2
## 10.605 19.735 26.100
```

In step 3 we will test to see if the difference between these means are significant or not.

Lets see the mean of length for different supplement groups:

```
tapply(ToothGrowth$len, INDEX = list(ToothGrowth$supp), FUN = mean)
```

```
##       OJ       VC
## 20.66333 16.96333
```

Again, in step 3 we will test to see if the difference between these means are significant or not.

Step 3

In step 3, we should use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (We should only use the techniques from class, even if there's other approaches worth considering)

In each of our hypothesis tests, null hypothesis is that difference between mean of two groups is zero and alternative hypothesis is that difference between mean of two groups isn't zero.

First, lets look at p-values of hypothesis tests for different dose groups.

```
## dose0.5-dose1 dose0.5-dose2   dose1-dose2
##  1.268301e-07  4.397525e-14  1.906430e-05
```

Now, lets look at p-values of hypothesis test for different supplement groups.

```
t2 <- t.test(len ~ supp, data = ToothGrowth)
t2$p.value
```

```
## [1] 0.06063451
```

Step 4

In step 4, we should state our conclusions and the assumptions needed for our conclusions.

Our assumptions are:

- Every group follows a normal distribution.
- The data used to carry out the test should be sampled independently from the two populations being compared.
- Observations in different groups are not paired.
- Different groups can have unequal variances.

Our conclusions are:

- We can reject null hypothesis for the relationship between length and dose (p-value < 0.05). So there is a significant (and positive) relationship between length and dose.
- We cannot reject null hypothesis for the relationship between length and supplement (p-value > 0.05). So there isn't a significant relationship between length and supplement in general.

Appendix

Codes

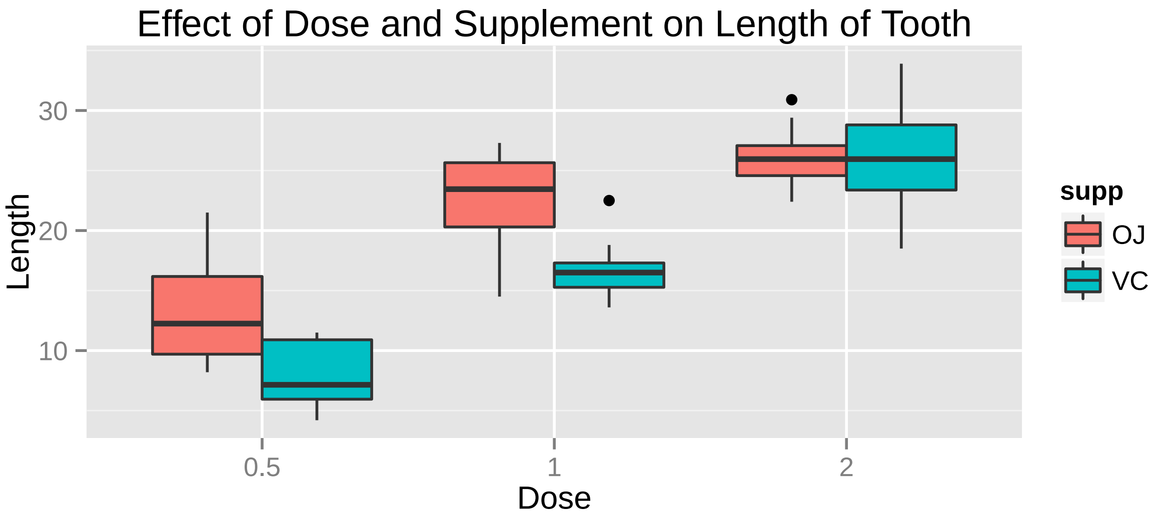
Looking at head of ToothGrow data in step 1

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

Code of first plot

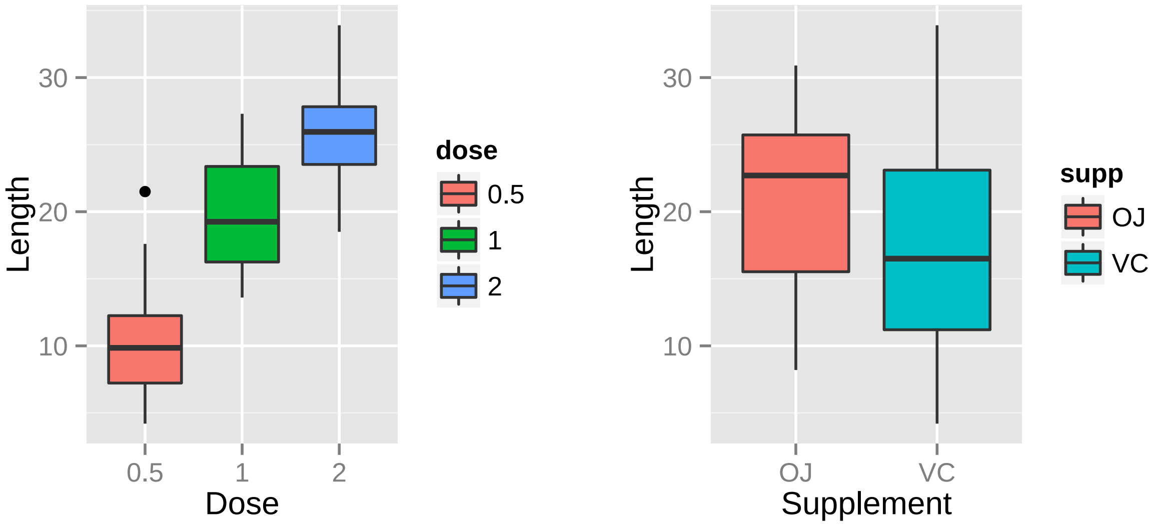
```
library(ggplot2)
plot2 <- ggplot(ToothGrowth, aes(x = dose, y = len, fill = supp))
plot2 <- plot2 + geom_boxplot()
plot2 <- plot2 + labs(title = "Effect of Dose and Supplement on Length of Tooth", x = "Dose", y = "Leng
plot2
```



Code of second plot

```
suppressMessages(library(gridExtra))
plot2.1 <- ggplot(ToothGrowth, aes(x = dose, y = len, fill = dose))
plot2.1 <- plot2.1 + geom_boxplot(aes(fill = dose))
```

```r
plot2.1 <- plot2.1 + labs(x = "Dose", y = "Length")
plot2.2 <- ggplot(ToothGrowth, aes(x = supp, y = len, fill = supp))
plot2.2 <- plot2.2 + geom_boxplot(aes(fill = supp))
plot2.2 <- plot2.2 + labs(x = "Supplement", y = "Length")
grid.arrange(plot2.1, plot2.2, ncol = 2)
```



Code of p-values of hypothesis tests for different dose groups

```r
suppressMessages(library(dplyr))
toothDose0.5 <- filter(ToothGrowth, dose == 0.5)
toothDose1 <- filter(ToothGrowth, dose == 1)
toothDose2 <- filter(ToothGrowth, dose == 2)
t1.1 <- t.test(toothDose0.5$len, toothDose1$len)
t1.2 <-t.test(toothDose0.5$len, toothDose2$len)
t1.3 <- t.test(toothDose1$len, toothDose2$len)
c("dose0.5-dose1" = t1.1$p.value, "dose0.5-dose2" = t1.2$p.value, "dose1-dose2" = t1.3$p.value)
```

```
## dose0.5-dose1 dose0.5-dose2   dose1-dose2
##  1.268301e-07  4.397525e-14  1.906430e-05
```

Complete results of hypothesis tests

```r
t1.1
```

```
##
##  Welch Two Sample t-test
##
## data:  toothDose0.5$len and toothDose1$len
## 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
```

```
t1.2
```

```
##
##  Welch Two Sample t-test
##
## data:  toothDose0.5$len and toothDose2$len
## 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
```

```
t1.3
```

```
##
##  Welch Two Sample t-test
##
## data:  toothDose1$len and toothDose2$len
## 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
```

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

```
##
##  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
```