Statistical Inference Course Project Part 2

Ziyao Gao

10/30/2017

Assignment Description

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

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
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)
4. State your conclusions and the assumptions needed for your conclusions.

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

```
# install the pacakes needed
library(datasets)
library(ggplot2)

# load the data
data(ToothGrowth)

# explore structure of the data
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 ...
```

```
# summarize 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
```

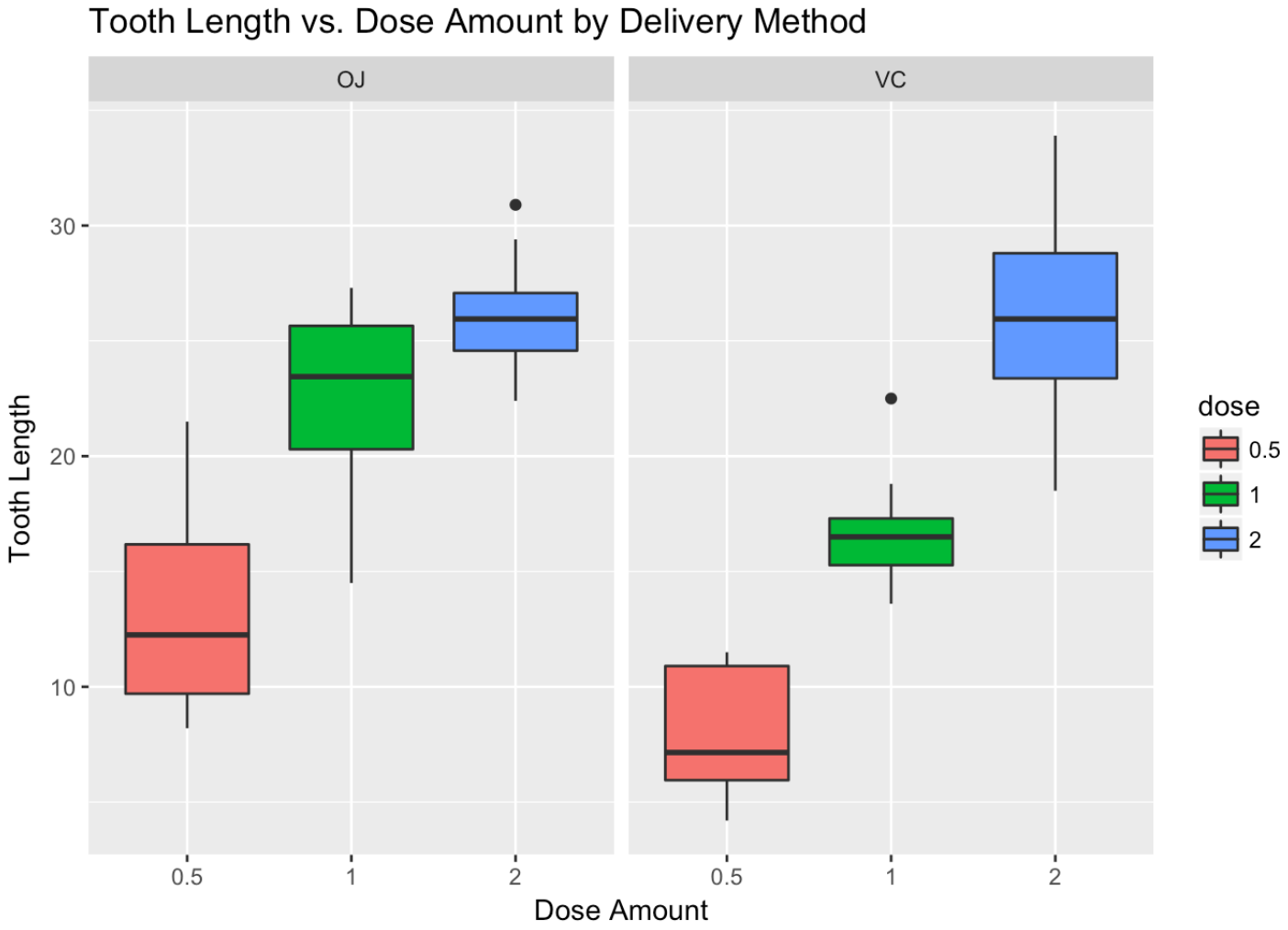
```
# take a look of the first few rows of data
head(ToothGrowth)
```

```
##    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
```

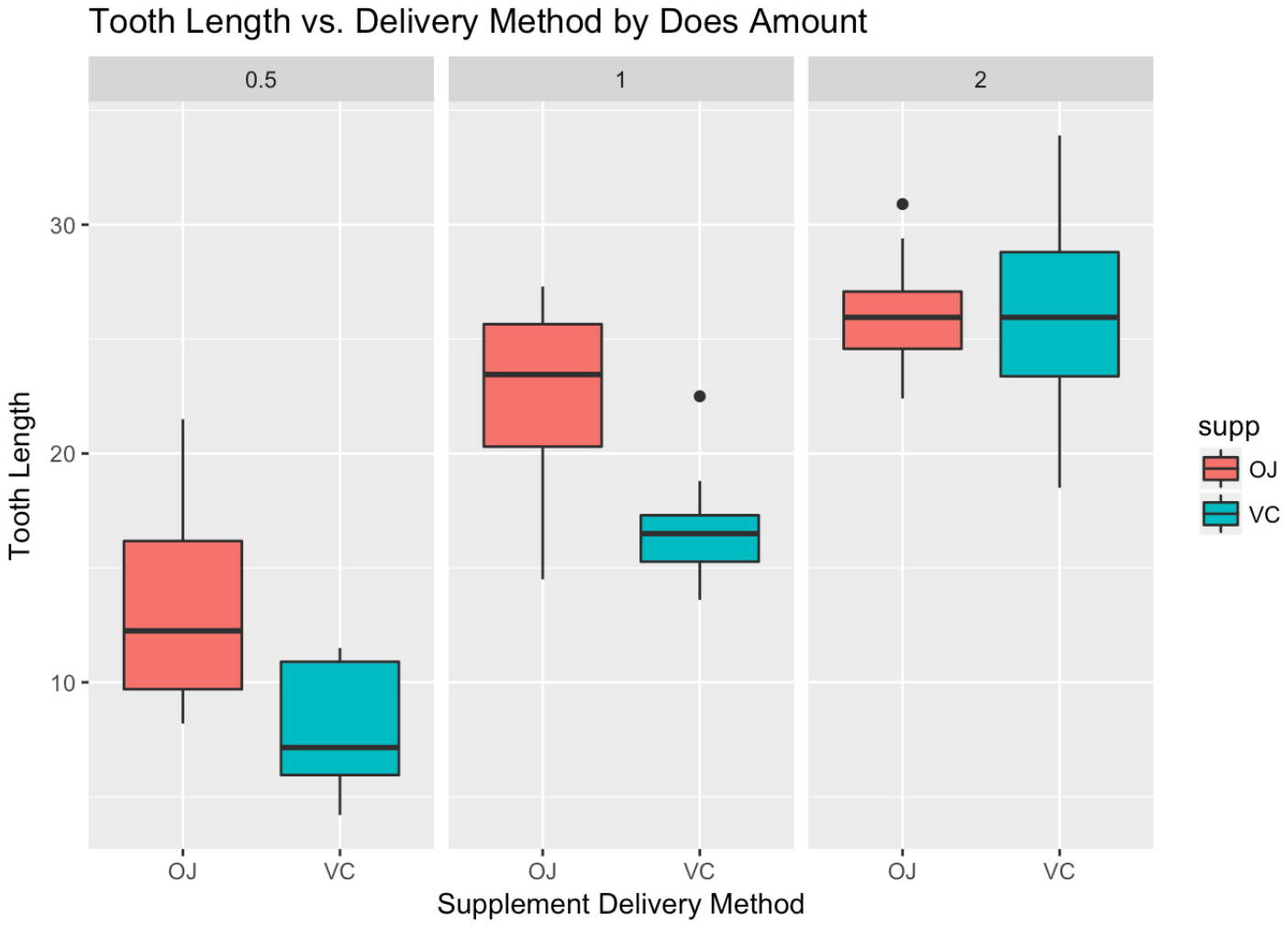
2. Provide a summary of the data through plots

```
# convert dose to factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

# plot length by dose amount, broken by supplement delivery method
ggplot(aes(x = dose, y = len), data = ToothGrowth) +
        geom_boxplot(aes(fill = dose)) +
        xlab("Dose Amount") +
        ylab("Tooth Length") +
        facet_grid(~ supp) +
        ggtitle("Tooth Length vs. Dose Amount by Delivery Method")
```




```
# plot length by supplement delivery method, broken by dose amount
ggplot(aes(x = supp, y = len), data = ToothGrowth) +
        geom_boxplot(aes(fill = supp)) +
        xlab("Supplement Delivery Method") +
        ylab("Tooth Length") +
        facet_grid(~ dose) +
        ggtitle("Tooth Length vs. Delivery Method by Does Amount")
```



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

Let's first check the tooth length by supplement using t-test

```
# run t-test
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
```

As the p-value is 0.06 and the confidernce interval contains 0 as well, we fail to reject the null hypothisis that supplement types have no effect on the tooth length

Now let's compare the tooth length by dose amount using t-test

```
# subset data per dose amount level 0.5 and 1.0
sub1 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
t.test(len~dose, data = sub1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 0.5   mean in group 1
##            10.605            19.735
```

```
# subset data per dose amount level 0.5 and 2.0
sub2 <- subset(ToothGrowth, dose %in% c(0.5, 2.0))
t.test(len~dose, data = sub2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 0.5   mean in group 2
##            10.605            26.100
```

```
# subset data per dose amount level 1.0 and 2.0
sub3 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))
t.test(len~dose, data = sub3)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## 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 in group 1 mean in group 2
##          19.735          26.100
```

As we can see, the p-values for all of those three t-tests are fairly small and their confidence intervals do not contain 0 as well, so we can reject the null hypothesis that the dose amount has no effect on the tooth length.

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

Conslusions:

1. Supplement delivery method has no effect on tooth growth.
2. Tooth growth increases with increased dose amount.

Assumptions:

1. The sample is representitve of the entire population.
2. The distribution of the sample means follows the Central Limit Theorem.