Statistical Inference Course Project: Part 2

lcombs

02/05/2016

Investigating Tooth Growth:

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

```
library(magrittr)

library(datasets)
data(ToothGrowth)

library(data.table)
ToothGrowth <-
        ToothGrowth %>%
        as.data.table()

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
```

```
str(ToothGrowth)
```

```
## Classes 'data.table' and '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 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

2. Provide a basic summary of the data.

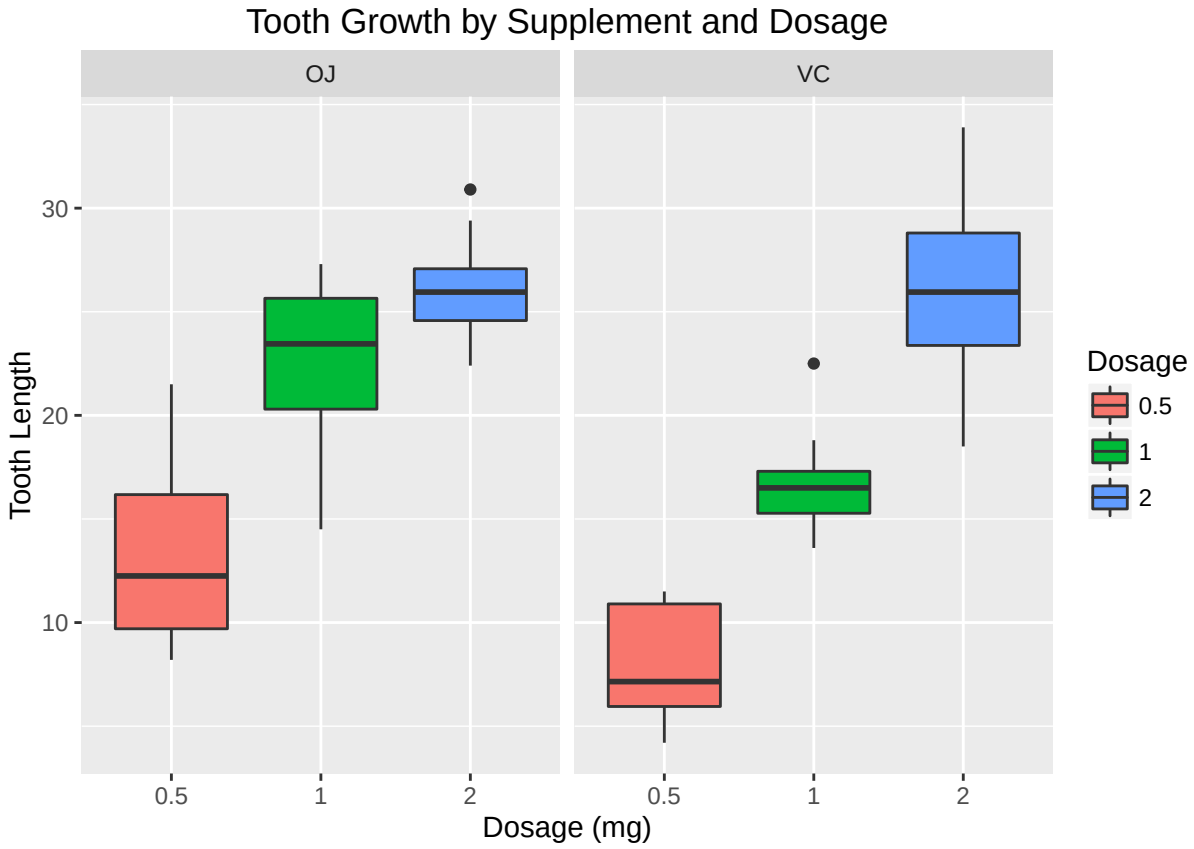
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```
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
```

```
ggplot(ToothGrowth,
       aes(x = dose  %>% as.factor(),
           y = len,
           fill = dose  %>% as.factor())) +
        geom_boxplot() +
        facet_grid(. ~ supp) +
        ggtitle("Tooth Growth by Supplement and Dosage") +
        xlab("Dosage (mg)") +
        ylab("Tooth Length") +
        guides(fill = guide_legend(title = "Dosage"))
```



3. Compare tooth growth by supp and dose.

We will assume equal variance because we have an equal number of data points within each subgroup.

```
t_test_function <-
        function(dataset = ToothGrowth){
                t_test <-
                        t.test(len ~ supp,
                               paired = FALSE,
                               var.equal = TRUE,
                               data = dataset)
                paste0("The p-value is ",
                       t_test$p.value  %>%
                               round(., 3),
                       ". The 95% confidence interval is between ",
                       t_test$conf.int[1] %>%
                               round(., 2),
                       " and ",
                       t_test$conf.int[2] %>%
                               round(., 2),
                       ".")

        }
```

Regardless of dose: The p-value is 0.06. The 95% confidence interval is between -0.17 and 7.57.

Overall, there does not seem to be a significant difference, but we should try comparing within different doses.

Assume the dose is 0.5: The p-value is 0.005. The 95% confidence interval is between 1.77 and 8.73.

Assume the dose is 1: The p-value is 0.001. The 95% confidence interval is between 2.84 and 9.02.

Assume the dose is 2: The p-value is 0.964. The 95% confidence interval is between -3.72 and 3.56.

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

In conclusion, the first t-test shows that if we disregard dose, there is no difference between the two supplements. However, in the graph, we saw there appeared to be a rather large difference between the two populations when we take into account dose. In particular, we find significant differences between those within the population who took a dose of .5 or 1, but no significant difference for those who took a dose of 2. In computing these intervals, we assume that the variances are equal because the subpopulations within each test group are the same. We perform a t-test so we also assume the population distributions are approximately normal and that samples are independent of each other. We assume the experiment was carried out and the data was gathered so these assumptions would be met.