Statistical Inference - Course Project 2

Tuesday, November 18 2014

Summary

In this project, we will analyze the ToothGrowth data in the R datasets package. The following are the objectives of this project:

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.

1. Exploratory Data Analysis

We first load the datasets package and examine the structure of the ToothGrowth dataset.

```
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 ...
```

Now we look at some basic statistics of this dataset.

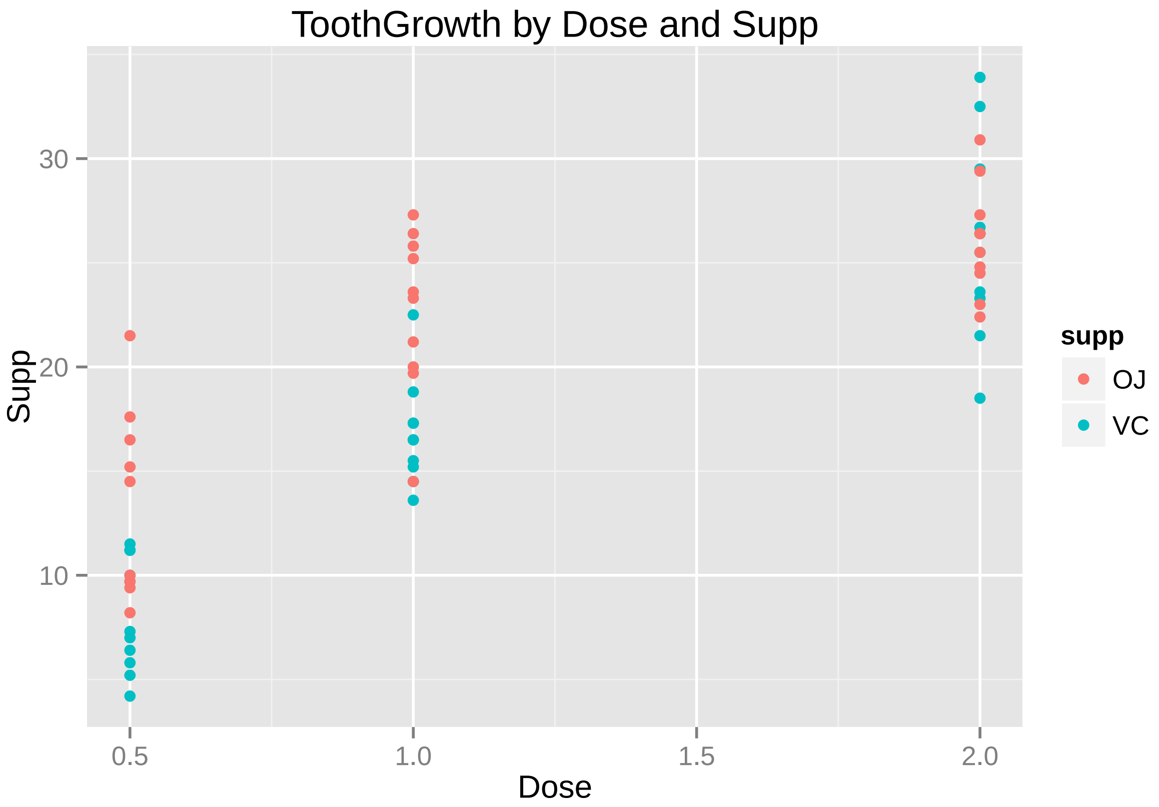
```
summary (ToothGrowth)
```

```
##       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  
```

2. Summary and Graphical Analysis

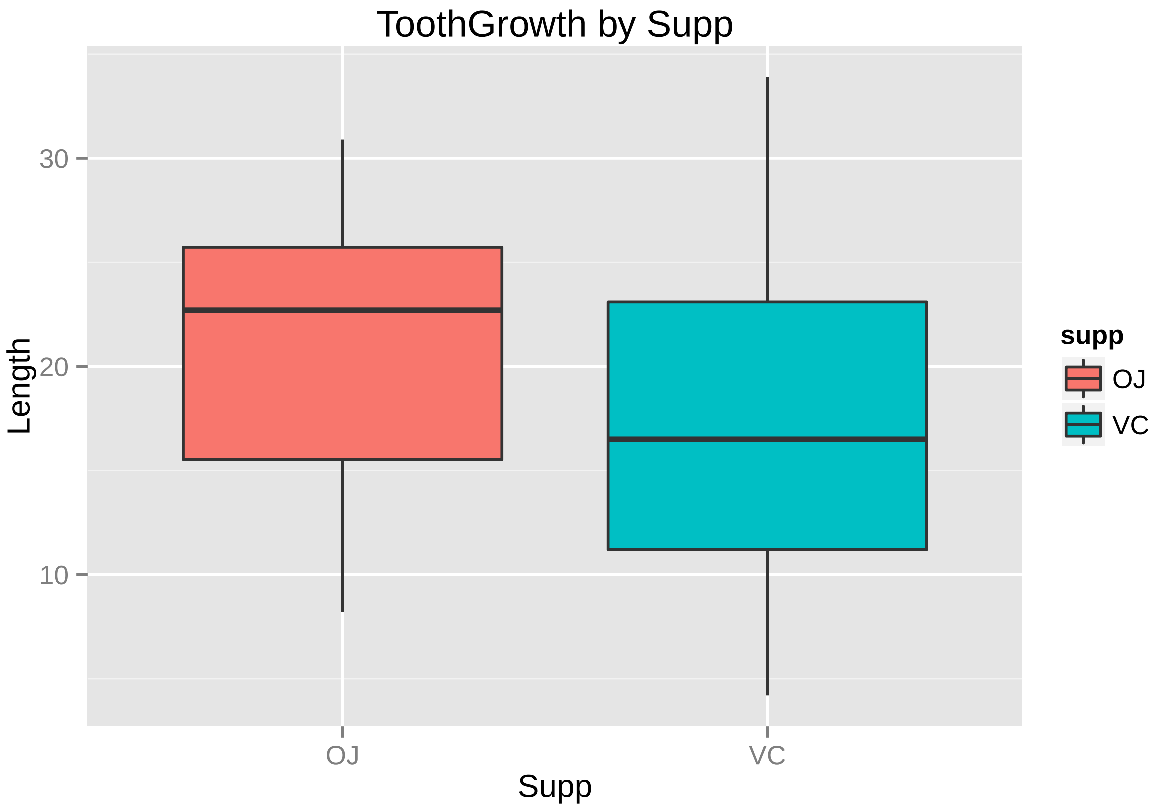
Now we examine the data graphically. The plot below shows the `len` plotted against `dose` color conditioned on `supp`. We see that as the `dose` increases, the `len` increases as well. However, it is not clear whether the `supp` has any effect.

```
library (ggplot2)
par (mfrow = c(1,1))
ggplot (aes(x = dose, y = len), data = ToothGrowth) +
    geom_point(aes (color = supp)) +
    labs (list(title = "ToothGrowth by Dose and Supp", x = "Dose", y = "Supp"))
```



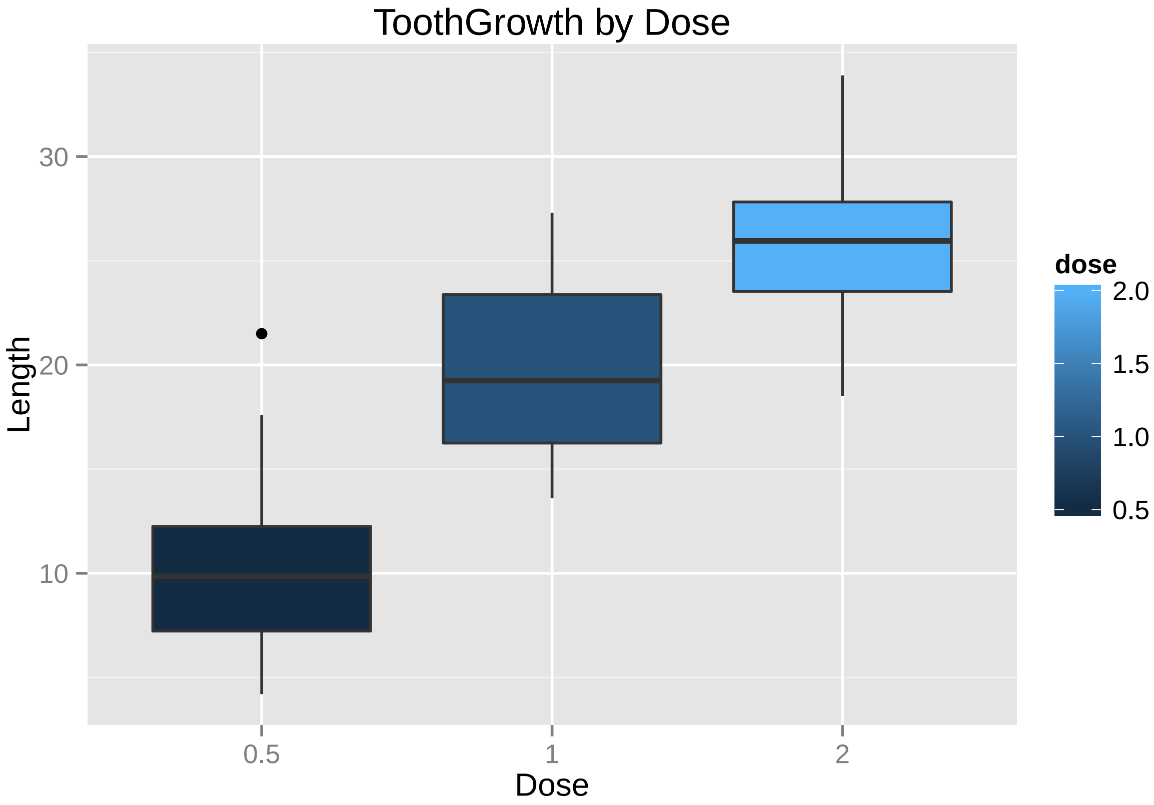
Now, we look at a boxplot of `supp` against `len` to check whether we can see any discernable difference or pattern. However, the plot shows that there is a lot of overlap between the two delivery methods i.e. `supp`. So we still do not know whether `supp` has a significant effect on `len` or not.

```
par (mfrow = c(1,1))
ggplot(aes(x=supp, y=len), data=ToothGrowth) + geom_boxplot(aes(fill=supp)) +
    labs (list(title = "ToothGrowth by Supp", x = "Supp", y = "Length"))
```



Now, we look at a boxplot of `dose` against `len` to check whether we can see any pattern. The plot shows that as `dose` increases, the `len` increases as well. It is possible that `dose` has a significant effect on `len`.

```
par (mfrow = c(1,1))
ggplot(aes(x=as.factor(dose), y=len), data=ToothGrowth) + geom_boxplot(aes(fill=dose)) +
    labs (list(title = "ToothGrowth by Dose", x = "Dose", y = "Length"))
```



3. Confidence Intervals & Hypothesis Tests

First, we use the t-test to check whether the `supp` is statistically significant.

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

The p-value is 0.0606345, which is relatively high. The 95% confidence interval contains zero. This indicates that we can not reject the null hypothesis that the two supplement types have no effect on tooth length `len`.

We will now test whether the dose has a significant effect on the tooth length. We know that the `dose` variable has three distinct values. We cannot use a t-test to test for the significance of all three values. So we have to divide them into separate sub-datasets. Then we use t-tests to check whether the `dose` has significant effect on the tooth length `len`.

```
# Sub-groups based on dose pairs.
dose051 <- subset (ToothGrowth, dose %in% c(0.5, 1.0))
dose052 <- subset (ToothGrowth, dose %in% c(0.5, 2.0))
dose12 <- subset (ToothGrowth, dose %in% c(1.0, 2.0))

# Check for differences in length due to change in dose levels: (0.5, 1.0)
t.test(len ~ dose, data = dose051)
```

```
##
##  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
```

```
# Check for differences in length due to change in dose levels: (0.5, 2.0)
t.test(len ~ dose, data = dose052)
```

```
##
##  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
```

```
# Check for differences in length due to change in dose levels: (1.0, 2.0)
t.test(len ~ dose, data = dose12)
```

```
##
##  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
```

In all three dose t-tests, the p-value is less than the fischer p-value of 0.05. Also, none of the 95% confidence interval contains zero. We see that the mean tooth length `len` increases with doseage `dose`. So we should reject the null hypothesis concluding that an increase in the dose level leads to an increase in tooth length.

4. Conslusions

1. Inreasing the doseage `dose` increases tooth growth `len`.
2. The type of supplement used `supp`, whether its orange juice or vitamin C, has no discernable effect on tooth growth.

Assumptions:

1. Variances between the different populations of guinea pigs are not the same.
2. The populations are independent.
3. Experiment was conducted by randomizing the sample guinea pig selection.