Statistical Inference assignment - part2

Saturday, November 22, 2014

Subject: Evaluate the "ToothGrowth" Dataset

To prepare for the analysis, the following library and packages are load to RStudio ggplot2, reshape2, datasets.

```
## Warning: package 'ggplot2' was built under R version 3.1.2

## Warning: package 'reshape2' was built under R version 3.1.2
```

1. load 'ToothGrowth' Dataset

```
##       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

##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

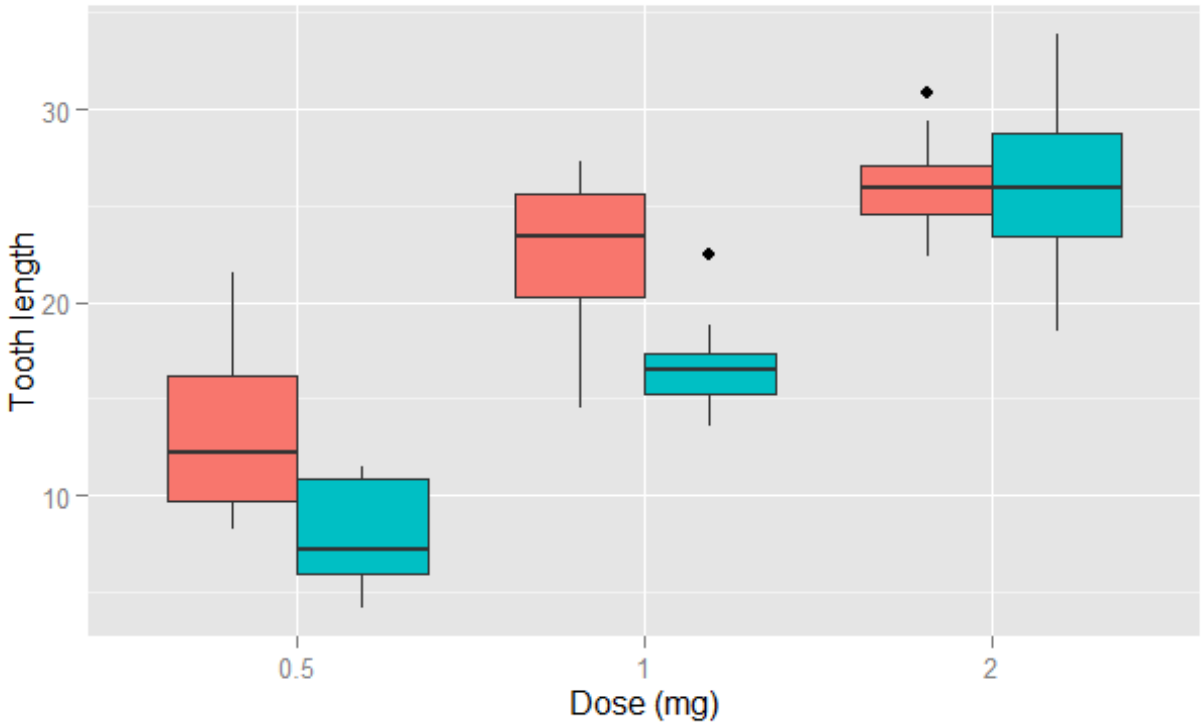
The dataset has 3 variables of 60 observations. The response output - tooth's length(len) versus two types deliver methods (the orange juice/ascorbic acid) and three levels of doses(0.5, 1, 2) test among 10 guinea pigs.

The question is, do either the delivery method or the dose size significantly impact the toothgrowth ? A box plot by dose level and delivery methods can tell the story.

2. Provide a basic summary of the data.

A good way summarize the data is a clustered boxplot.

```
ggplot(ToothGrowth, aes(x = factor(dose), y = len, fill = supp)) +
    xlab("Dose (mg)") +
    ylab("Tooth length") +
    scale_fill_discrete(name="Delivery method",
                        breaks=c("OJ", "VC"),
                        labels=c("Orange juice", "Ascorbic acid")) +
geom_boxplot()
```



Based on the box plot above, it seems the higher doses has higher impact on the toothgrowth and the Orange Juice has higher impact at 0.5 and 1.0 mg level compare to Ascorbic acid, but the impact is about the same at 2.0 mg level.

3. Two Factor ANOVA and confidence interval evaluation

In the case of ToothGrowth data, it has two factors with 6(2x3) combinations of treatments for each of the subject. In stead of evaluating each factor at time. ANOVA test is better method.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## dose         1 2224.3  2224.3 133.415  < 2e-16 ***
## supp         1  205.3   205.3  12.317 0.000894 ***
## dose:supp    1   88.9    88.9   5.333 0.024631 *
## Residuals   56  933.6    16.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The situation is similar for the orange juice case with the difference that the confidence intervalls for the doses 1mg and 2mg overlap. Lets perform an explicit hypothesis test for this doses:

Now we should take a look on the differences between the delivery methods for each dose:

```
##   dose supp  mean confidence.intervall1 confidence.intervall2
## 1  0.5   OJ 13.23             10.039717             16.420283
## 2  0.5   VC  7.98              6.015176              9.944824
## 3  1.0   OJ 22.70             19.902273             25.497727
```

```
## 4  1.0   VC 16.77              14.970657              18.569343
## 5  2.0   OJ 26.06              24.160686              27.959314
## 6  2.0   VC 26.14              22.707910              29.572090
```

We see that the confidence intervalls are pairwise disjoint for the delivery methods, so that we can conclude with high confidence that the mean of the tooth growth is higher for orange juice for doses of 0.5mg and 1.0mg. But we can see that the intervalls for the 2.0mg doses overlap. Lets do a t.test to decide if the means are different:

4. What is your conclusion

Based on the p-value, we can accept that both dose and deliver mehtod have significan impact to the toothgrowth. There is very litte interaction between the level of dose and deliver method. It can be possiblly ignored at high level dose.