Statistical Inference Project Pt.2: Tests and Inference

Valentin Goverdovsky

24 December 2015

Overview

This report report briefly explores the ToothGrowth dataset from the datasets library. First we do a simple exploratory analysis to establish couple of interesting questions that could be answered with this dataset using techniques taught in the class. Subsequently we try to answer these questions.

Exporatory analysis

Initially we have to load the data and produce its summary

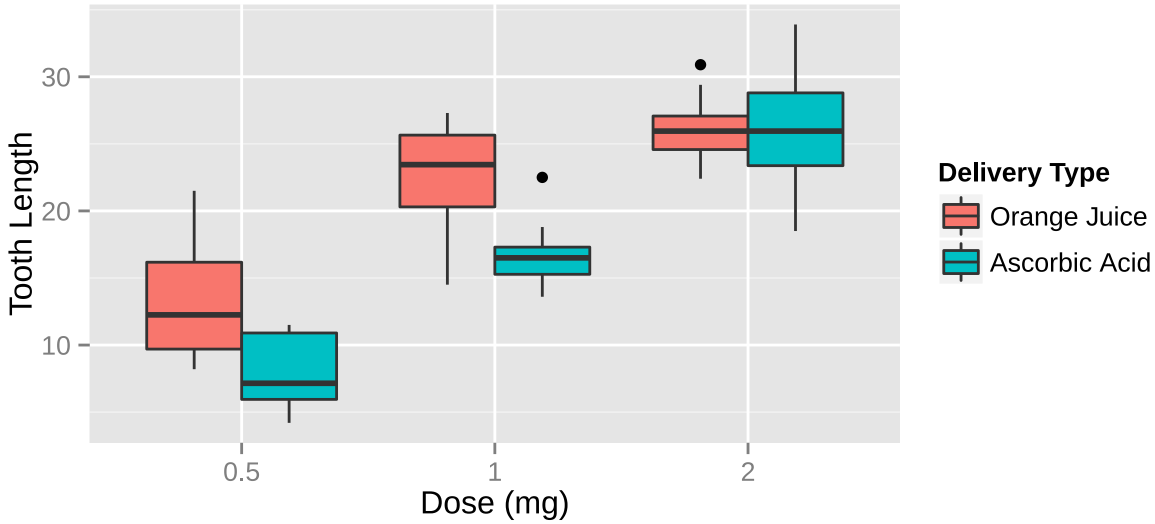
```
summary(ToothGrowth)
```

```
##       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  
```

The summary reveals that there are only three columns in the dataset and most likely the first column, designated as 'len' is the response variable, while the other two ('supp', 'dose') are essentially factor columns. From the description of the dataset we find that the the response variable is the length of teeth in each of the 10 guinea pigs in response to different delivery methods and dosages of Vitamin C. Delivery methods are: OJ - Orange Juice and VC - Ascrobic Acid, while dosages are split into three amounts: 0.5, 1 and 2 mg.

At this point it is clear that we can make a boxplot which can nicely illustrate the data and help us find some interesting questions to answer with this dataset. The figure with the boxplot is shown below.

```
g <- ggplot(ToothGrowth,aes(x=factor(dose),y=len)) + geom_boxplot(aes(fill=factor(supp))) +
     labs(x = 'Dose (mg)', y = 'Tooth Length') +
     scale_fill_discrete(name="Delivery Type",
                         breaks=c("OJ", "VC"),
                         labels=c("Orange Juice", 'Ascorbic Acid'))
g
```



This figure shows some interesting patterns, for example it seems that supplementation of Vitamin C via the orange juice has higher impact on tooth length than supplementation via the ascorbic acid - this will be the first question we will investigate. On the other hand at dosage level of 2 mg, there's a large overlap between the responses to different delivery methods, thus another question worth investigating is if there's any statistically significant difference at this dosage level. Finally for orange juice delivery methods there doesn't seem to be a very significant difference in tooth length as the dosage is doubled from 1 mg to 2 mg and establishing if there is a statistically significant increase will be the final question we will attempt to answer in this report.

Statistical analysis

Orange juice vs. Ascorbic Acid

As outlined earlier, first we would like to establish if the teeth are longer when Vitamin C is supplemented with orange juice compared to supplementation via the ascorbic acid. The appropriate method to establish if such a difference exists would be a one-sided t test at 95% confidence. We assume that the underlying data is Gaussian and the samples are IID. The test we carry out is not paired, since it's not clear how exactly the experiment was conducted, even though the description mentions that there are 10 guinea pigs in the trial, we lack knowledge of animal biology (Do the theeth of the guinea pigs grow back if taken out? Is the growth rate affected by the number of times the teeth are taken out? etc.) to comfortably assume paired data. Additionally we conservatively assume non-equal variance of the data in two groups of delivery methods, since no additional information is provide with regards to the underlying distributions.

```
OJ <- ToothGrowth[ToothGrowth$supp=='OJ','len']
VC <- ToothGrowth[ToothGrowth$supp=='VC','len']
delivTest <- t.test(OJ, VC, alternative='greater', paired = FALSE, var.equal = FALSE, conf.level = 0.95
delivTest$p.value
```

```
## [1] 0.03031725
```

Based on the results of this test with the p-value of 0.03 which is less than 0.05, we reject the null hypothesis that the means of tooth length are the same between different delivery types in favor of an alternative hypothesis which states that the mean tooth length is greater with Vitamin C delivery via the orange juice than the mean tooth length with the delivery via the ascorbic acid. It must be noted that if we were to carry out a two.sided test to check for the 'difference' in means, then we would have failed to reject the null. Similarly we would NOT reject the null at the confidence of 97.5%.

Orange Juice vs. Ascorbic Acid at 2 mg dosage

From the boxplot we have presented earlier it is clear that there is a large overlap between the values of the response with different delivery methods at 2 mg dosage level. Here we investigate if in fact there is any difference. First, a non-paired t test at 95% confidence with non-equal variances is carried out.

```
OJ2 <- ToothGrowth[ToothGrowth$supp=='OJ' & ToothGrowth$dose==2.0,'len']
VC2 <- ToothGrowth[ToothGrowth$supp=='VC' & ToothGrowth$dose==2.0,'len']
dose2 <- t.test(OJ2, VC2, alternative='two.sided', paired = FALSE, var.equal = FALSE, conf.level = 0.95
dose2$p.value
```

```
## [1] 0.9638516
```

In this case the p-value is 0.9639 which is a lot higher than 0.05, thus at 95% confidence we CANNOT reject the null stating that at 2 mg dosage there's no difference in mean tooth length with different delivery methods. We can also check what effect the assumption of data to be non-paired has by running a paired test as follows.

```
dose2Pair <- t.test(OJ2, VC2, alternative='two.sided', paired = TRUE, var.equal = FALSE, conf.level = 0
dose2Pair$p.value
```

```
## [1] 0.9669567
```

We still fail to reject the null at 95% confidence, since the p-value is still very high.

Dosage effects with Orange Juice delivery method

The final test we would like to carry out is to investigate if doubling the Vitamin C dosage from 1 mg to 2 mg with orange juice delivery method is associated with statistically significant increase in the tooth length. We can carry out this test as follows.

```
OJ1 <- ToothGrowth[ToothGrowth$supp=='OJ' & ToothGrowth$dose==1.0,'len']
OJ2 <- ToothGrowth[ToothGrowth$supp=='OJ' & ToothGrowth$dose==2.0,'len']
dos <- t.test(OJ2, OJ1, alternative='greater', paired = FALSE, var.equal = FALSE, conf.level = 0.95)
dos$p.value
```

```
## [1] 0.01959757
```

The resulting p-value is 0.0196 which is lower than 0.05, thus we reject the null in favour of the alternative hypothesis at 95% confidence. Once again the data is assumed to be non-paired with different variance and the underlying distributions are Gaussian.

Conclusion

Based on the statistical analysis carried out in the previous sections we can make the following conclusions:

1) We conclude at 95% confidence that orange juice delivery method is associated with longer tooth length compared to the ascrobic acid delivery method, disregarding the dosage amount.
2) There is no statistically significant difference at 95% confidence, in the teeth length between the two delivery methods at 2 mg dosage of Vitamin C.
3) Doubling the dosage of Vitamin C from 1 mg to 2 mg with orange juice delivery method is associated with statistically significant increase in the teeth length at 95% confidence.

Appendix

Code for generating the boxplot figure:

```
g <- ggplot(ToothGrowth,aes(x=factor(dose),y=len)) + geom_boxplot(aes(fill=factor(supp))) +
     labs(x = 'Dose (mg)', y = 'Tooth Length') +
     scale_fill_discrete(name="Delivery Type",
                         breaks=c("OJ", "VC"),
                         labels=c("Orange Juice", 'Ascorbic Acid'))
g
```

Full results of the t test comparing orange juice vs. ascorbic acid:

```
##
##  Welch Two Sample t-test
##
## data:  OJ and VC
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

Full results of the t test comparing effects of dosage change from 1 mg to 2 mg for orange juice delivery method:

```
##
##  Welch Two Sample t-test
##
## data:  OJ2 and VC2
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

Full results of the paired t test comparing effects of dosage change from 1 mg to 2 mg for orange juice delivery method:

```
##
##  Paired t-test
##
## data:  OJ2 and VC2
## t = -0.042592, df = 9, p-value = 0.967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of the differences
##                   -0.08
```

Full results of the t test comparing arange juice vs. ascorbic acid at 2 mg dosage:

```
## 
##  Welch Two Sample t-test
## 
## data:  OJ2 and OJ1
## t = 2.2478, df = 15.842, p-value = 0.0196
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7486236       Inf
## sample estimates:
## mean of x mean of y 
##     26.06     22.70
```