Statistical Inference Course Project Part 2

Lee 25 Jan 2015

Summary

This report uses "ToothGrowth" data set and studies the variables influencing tooth length. Box plots and t-tests are used for the comparisons of the variables.

Data processing

```
# load the data
data(ToothGrowth) # Data details in Appendix I
ToothGrowth$dose<-as.factor(ToothGrowth$dose) # factor the variable
```

`supp` is a categorical variable and expressed as factor. `dose` can be handled as categorical variable as there are only 3 empirical values.

Exploratory data analysis

```
# Get the data summary
summary(ToothGrowth)
```

```
##       len        supp     dose   
##  Min.   : 4.2   OJ:30   0.5:20  
##  1st Qu.:13.1   VC:30   1  :20  
##  Median :19.2           2  :20  
##  Mean   :18.8                   
##  3rd Qu.:25.3                   
##  Max.   :33.9                   
```

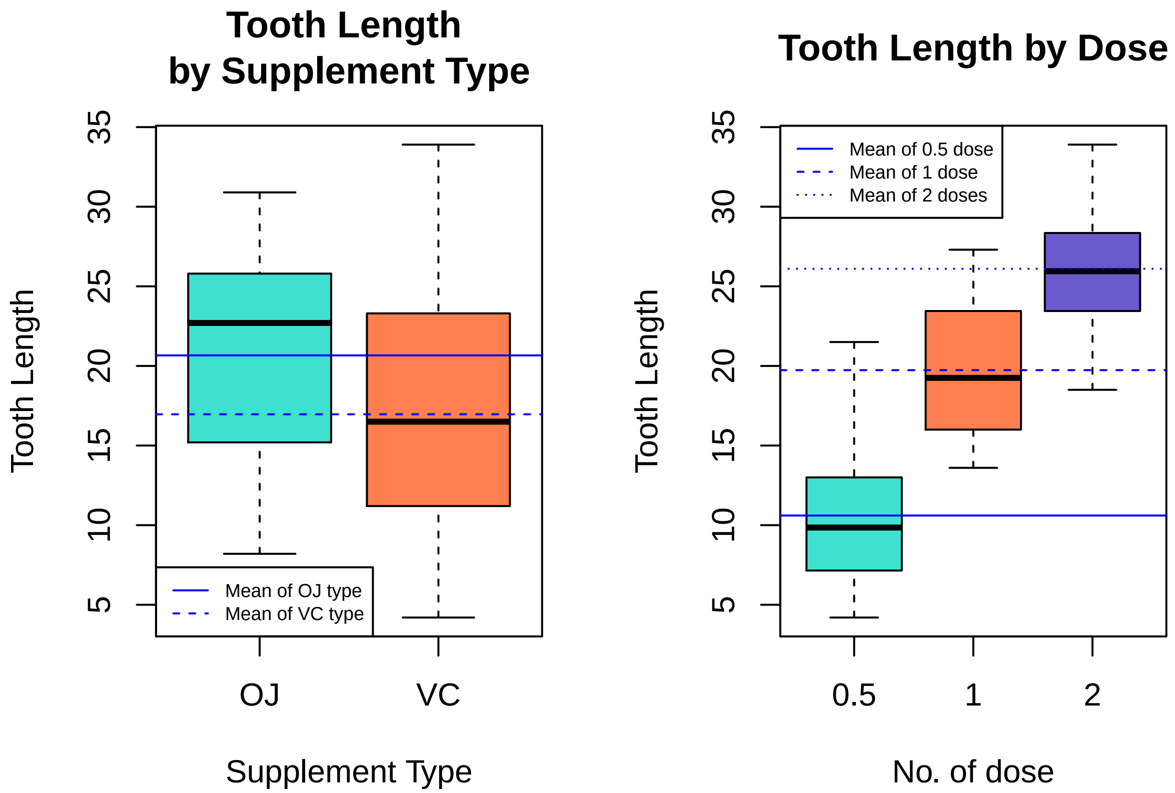
```
# View the len mean by supp type
len.mean.bysupp <- aggregate(len ~  supp, data = ToothGrowth, mean);len.mean.bysupp
```

```
##   supp   len
## 1   OJ 20.66
## 2   VC 16.96
```

```
# View the len mean by dose
len.mean.bydose <- aggregate(len ~  dose, data = ToothGrowth, mean);len.mean.bydose
```

```
##   dose   len
## 1  0.5 10.61
## 2    1 19.73
## 3    2 26.10
```

In the left box plot below, OJ teeth shows a narrower distribution with a higher mean of tooth length. Although VC teeth has a lower mean length, the mean is slightly higher than the median.



From the right box plot, the more doses is taken, the higher mean tooth length is. The mean tooth length is closer to the median as more doses were taken. With incomplete dose, the distribution of the tooth length is more spreaded. It can be implied that a higher number of doses leverage longer tooth length with less variation.

Confidence intervals and/or tests

T-test is performed to test the relationship between `len` and `supp` (Steps in Appendix II). The p-value is greater than 0.05 and the confidence interval contains zero. The null hypothesis cannot be rejected and thus supplement type has no significant effect on tooth length.

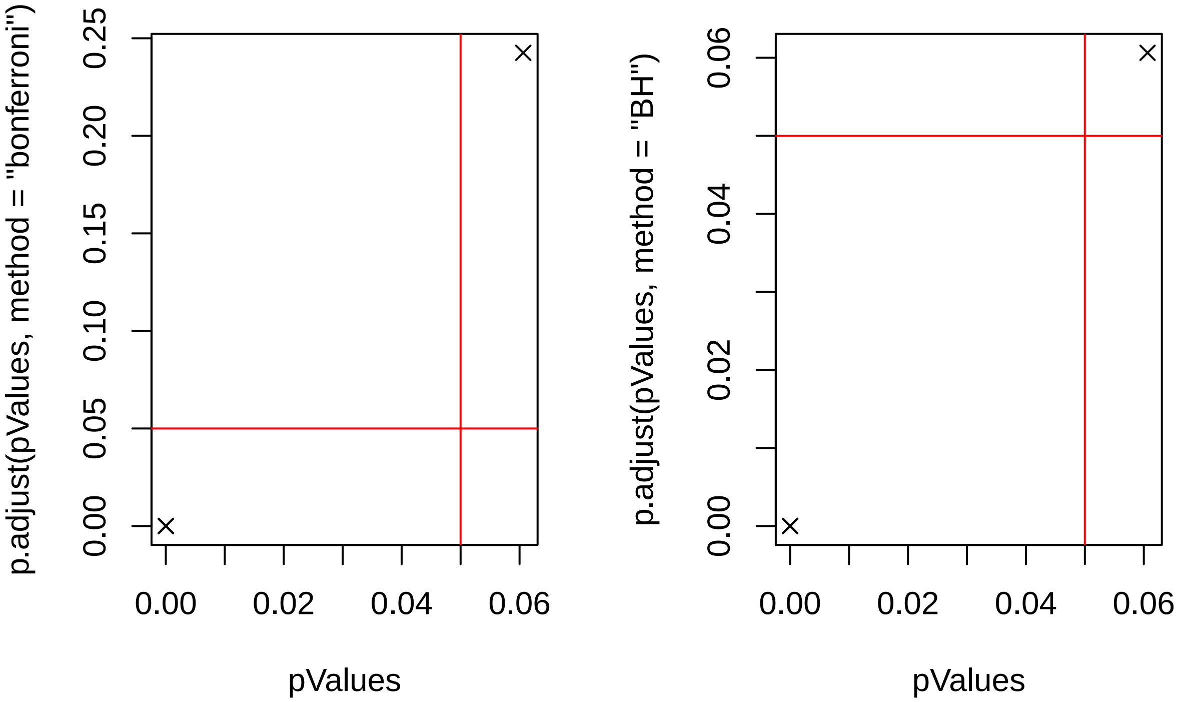
T-test is performed to test the relationship between `len` and different values of `dose` (Steps in Appendix II). All the p-values are much smaller than than 0.05 and the confidence interval does not contain zero. The null hypothesis can be rejected and thus the number of dose has significant effect on tooth length.

Let's check the p-values against Family-wise error rate and False discovery rate

```
# Collect the p-value of the tests
pValues <- c(t.supp$p.value,t.dose.exludeshalf$p.value,
             t.dose.exludesone$p.value,t.dose.exludestwo$p.value)
pValues;par(mfrow = c(1, 2))
```

```
## [1] 6.063e-02 1.906e-05 4.398e-14 1.268e-07
```

```r
# Adjust p-values Family-wise error rate
plot(pValues, p.adjust(pValues, method = "bonferroni"), pch = 4)
abline(v=0.05, col="red");abline(h=0.05, col="red")
# Adjust p-values False discovery rate
plot(pValues, p.adjust(pValues, method = "BH"), pch = 4)
abline(v=0.05, col="red");abline(h=0.05, col="red")
```



The adjusted p-values do not contradict with the result of the t-tests.

Conclusion

1. The change in tooth length caused by supplement type is not significant.
2. The number of doses has significant impact on tooth length. The more doses it takes, the longer tooth length is expected.

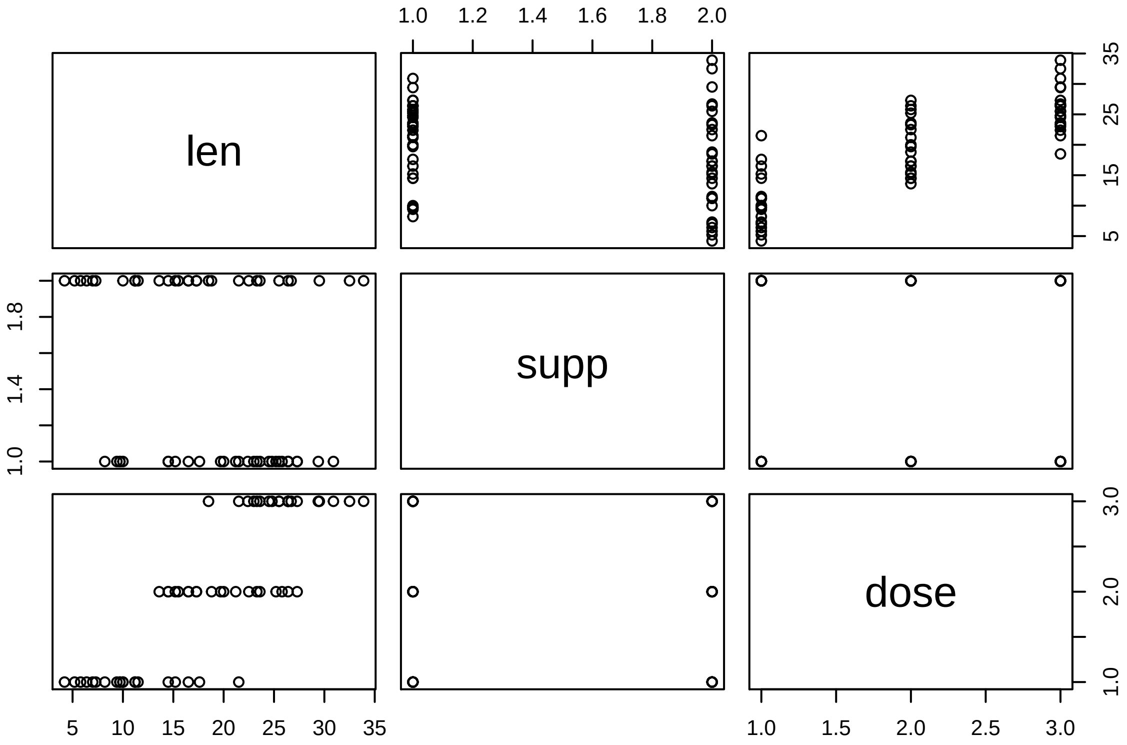
Assumptions

1. The samples are randomly collected and independent of each other.
2. The sample size is sufficiently large enought to represent the population.
3. The supplement type and the number of dose are independent of each other.
4. For the t-tests, the variances are assumed to be different for the two groups being compared. This assumption is less stronger than the case in which the variances are assumed to be equal.

Appendix I: Data overview

```
# get information about the data
?ToothGrowth;str(ToothGrowth);pairs(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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```



Appendix II : Test details

```
# test the relationship between len and supp
t.supp <- t.test(len ~ supp, data = ToothGrowth); t.supp
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.915, df = 55.31, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.171  7.571
## sample estimates:
## mean in group OJ mean in group VC
##            20.66            16.96
```

```
# test the relationship between len and different numbers of doses
ToothGrowth.exclude.half <- ToothGrowth[ToothGrowth$dose != 0.5,]
ToothGrowth.exclude.one <- ToothGrowth[ToothGrowth$dose != 1,]
ToothGrowth.exclude.two <- ToothGrowth[ToothGrowth$dose != 2,]
t.dose.exludeshalf <- t.test(len ~ dose, data = ToothGrowth.exclude.half)
t.dose.exludesone <- t.test(len ~ dose, data = ToothGrowth.exclude.one)
t.dose.exludestwo <- t.test(len ~ dose, data = ToothGrowth.exclude.two)
t.dose.exludeshalf;t.dose.exludesone;t.dose.exludestwo
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by dose
## t = -4.901, df = 37.1, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996 -3.734
## sample estimates:
## mean in group 1 mean in group 2
##           19.73           26.10
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by dose
## t = -11.8, df = 36.88, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.16 -12.83
## sample estimates:
## mean in group 0.5   mean in group 2
##             10.61             26.10
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by dose
## t = -6.477, df = 37.99, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.984  -6.276
## sample estimates:
## mean in group 0.5   mean in group 1
##             10.61             19.73
```