Course Project, Statistical Inference: Question 2

Felipe Campelo, Ph.D.

December 19, 2014

Abstract

This document presents the solution for Course Project Question 2, of the Statistical Inference course on Coursera.

Problem Statement

Now in the second portion of the class, we're going to analyze the *ToothGrowth* data in the *R datasets* package.
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*. (**Only use the techniques from class, even if there's other approaches worth considering**)
4. State your conclusions and the assumptions needed for your conclusions.

Solution

First, lets take a look and summarize the data:

```
# load and summarize data
library(datasets)
data(ToothGrowth)
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
```

We have basically 3 variables in this dataset. After consulting the relevant documentation (available in R by typing **?ToothGrowth**) we see that these three variables refer to:

- **len**: Tooth length (*numeric*; units not provided);
- **supp**: Supplement type (*factor* with levels {VC;OJ});
- **dose**: Dose of vitamin C (*numeric*; in milligrams; levels: {0.5,1,2}).

Before we proceed with the analysis, it is interesting to get a feel for the data. One way of achieving this is to employ boxplots to investigate the location and spread of the observations. But first we'll need to convert the **dose** variable from a *numeric* to a *factor* variable:

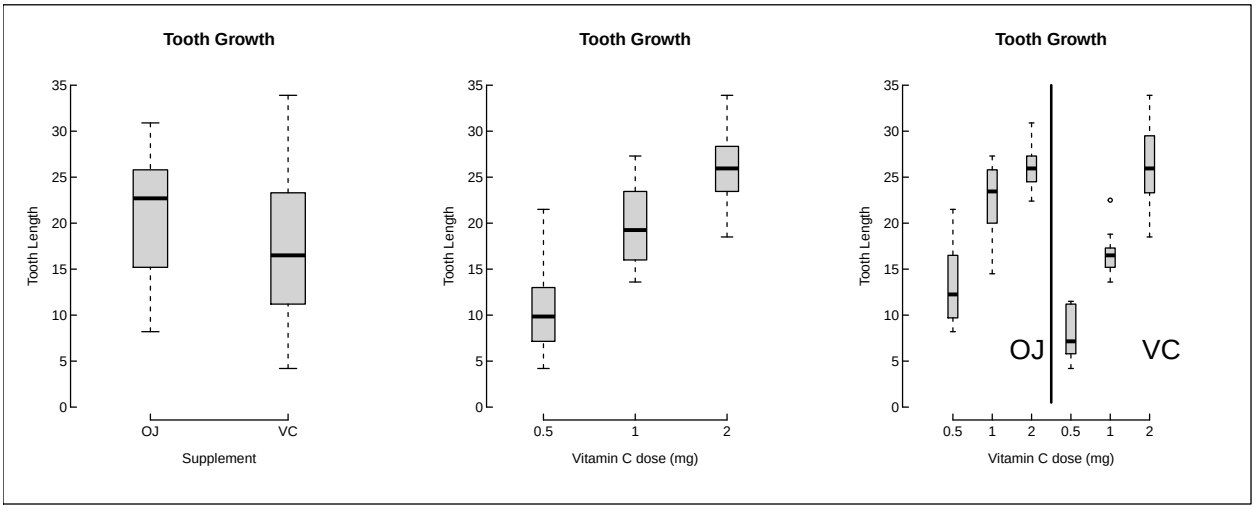
```r
# Convert dose to a factor variable
ToothGrowth$dose<-as.factor(ToothGrowth$dose)

# Set graphical parameters
par(xpd=FALSE,mfrow=c(1,3),mar=c(5,4,4,3)+0.1,mgp=c(1.8,.2,0),tck=0.02,las=1,bty="n")

# Boxplots 1: len by supp
with(ToothGrowth,
     boxplot(len~supp,boxwex=0.25,col="lightgray",ylim=c(0,35),
             ylab="Tooth Length",xlab="Supplement",main="Tooth Growth"))

# Boxplots 2: len by dose
with(ToothGrowth,
     boxplot(len~dose,boxwex=0.25,col="lightgray",ylim=c(0,35),
             ylab="Tooth Length",xlab="Vitamin C dose (mg)",main="Tooth Growth"))

# Boxplots 3: len by supp:dose
with(ToothGrowth,
     boxplot(len~dose*supp,boxwex = 0.25,col="lightgray",xlim=c(0,7),ylim=c(0,35),
             ylab="Tooth Length",xlab="Vitamin C dose (mg)",main="Tooth Growth",
             names=rep(levels(dose),2)))
points(x=c(3.5,3.5),y=c(0.5,35),type="l",lwd=2,lty=1)
text(x=c(3.55,7),y=c(6,6),labels=c("OJ","VC"),pos=2,cex=2)
box(which="outer")
```



From the exploratory plots, it seems like **dose** has a strong effect on tooth growth. On the other hand, **supp** does not seem to have much influence, with at most a subtle effect. To evaluate whether there are any statistically significant differences in mean tooth growth, we can employ 2 independent-samples t-tests (with Welch's correction for unequal variances). The validity of these tests is based on the following assumptions:

- There is no significant interaction between the **dose** and **supp** factor (which is apparent from the rightmost plot above);
- The data is normal, or at the very least bell-shaped;
- The observations are independent;

No significance correction for multiple hypotheses testing will be performed. The results of the significance tests are shown below. All tests are performed at the 0.05 significance level.

```r
# Hypothesis test regarding differences between the two levels of supp:
with(ToothGrowth,t.test(len~supp))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

```r
# Hypothesis tests regarding differences between the levels of dose, pairwise:
with(ToothGrowth,pairwise.t.test(len,dose,p.adjust="none"))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  len and dose
##
##   0.5     1
## 1 6.7e-09 -
## 2 < 2e-16 1.4e-05
##
## P value adjustment method: none
```

From the results shown above we can derive the following conclusions:

- There is not sufficient evidence to reject the null hypothesis of no difference in mean tooth length in subjects receiving either the *VC* or *OJ* supplementation method at the significance level of 0.05;
- There is sufficient evidence to reject the null hypotheses in all comparisons of levels of the factor **dose**, at the 0.05 significance level. This result, coupled with the observed distribution from the boxplots, suggests that the mean tooth growth is smallest for the 0.5mg dose, intermediate for the 1mg dose, and largest for the 2mg dose.

The scientific interpretation of these findings is beyond the scope of this report.