Statistical Inference Project 2

Igor Hut

October 25, 2015

Contents

Overview of the project assignment

This report written for the second part of the course project assignment for the Coursera course "Statistical Inference" which is a part of "Data Science" specialization. In this part of the project, we are supposed to perform basic inferential analyses using the ToothGrowth data in the R datasets package. According to acompaying help file the given dataset consists of 60 observations, namely lengths of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1 and 2 mg) with each of two delivery methods (orange juice or ascorbic acid). The overview of general steps that should be performed is given below:

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.

1. Load data and necessary libraries

```
# load neccesary libraries
library(datasets)
library(dplyr)

# load data
data(ToothGrowth)
toothGrowth<-ToothGrowth
```

2. Provide basic summary of the data and perform basic exploratory data analysis

```
# look at the dataset structure
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 ...
```

```
# convert variable dose from numeric to factor
toothGrowth$dose <- factor(ToothGrowth$dose, labels = c("0,5mg", "1mg", "2mg"))

# look at the dataset structure after conversion
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: Factor w/ 3 levels "0,5mg","1mg",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# basic summary statistics
summary(toothGrowth)
```

```
##       len        supp       dose
##  Min.   : 4.20   OJ:30   0,5mg:20
##  1st Qu.:13.07   VC:30   1mg  :20
##  Median :19.25           2mg  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

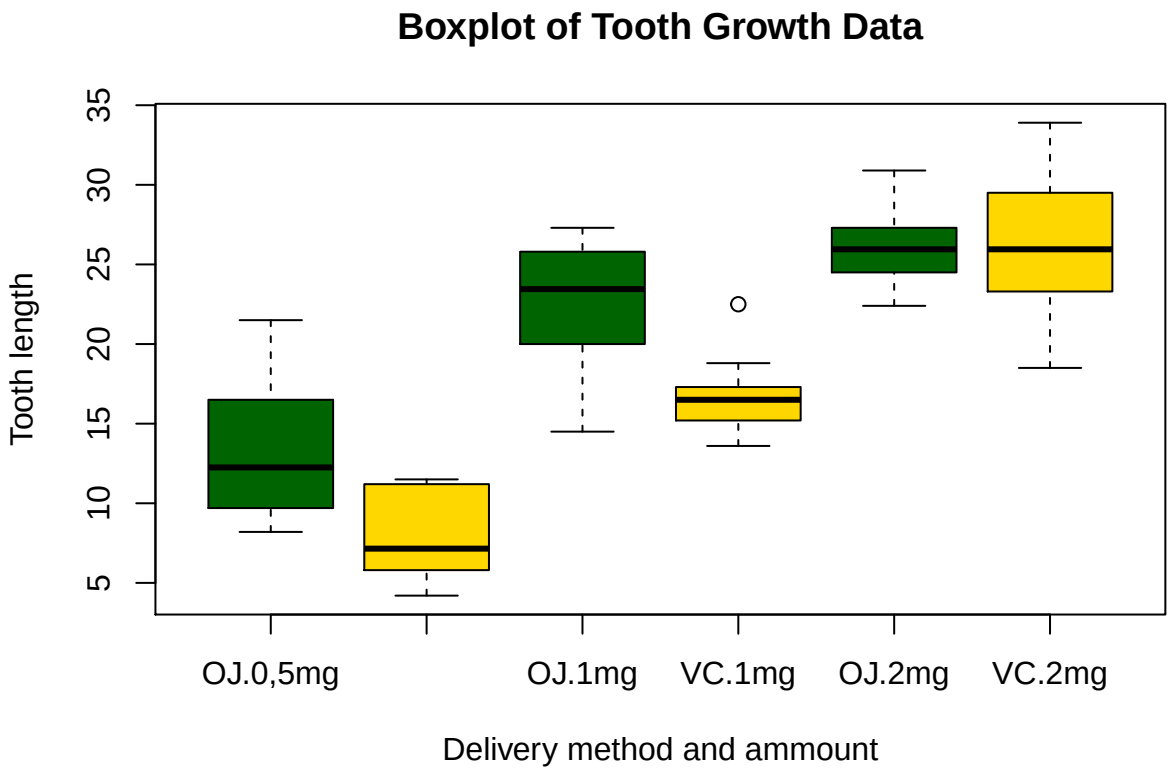
```
# summary statistics within each combination of dose level and delivery method

by(toothGrowth$len, INDICES = list(toothGrowth$supp, toothGrowth$dose), summary)
```

```
## : OJ
## : 0,5mg
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20    9.70   12.25   13.23   16.18   21.50
## ------------------------------------------------------------
## : VC
## : 0,5mg
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20    5.95    7.15    7.98   10.90   11.50
## ------------------------------------------------------------
## : OJ
## : 1mg
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.50   20.30   23.45   22.70   25.65   27.30
## ------------------------------------------------------------
## : VC
## : 1mg
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   15.27   16.50   16.77   17.30   22.50
## ------------------------------------------------------------
## : OJ
```

```
## : 2mg
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.40   24.58   25.95   26.06   27.08   30.90
## ------------------------------------------------------------
## : VC
## : 2mg
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.50   23.38   25.95   26.14   28.80   33.90
```

```
boxplot(len ~ supp * dose, data=toothGrowth, xlab = 'Delivery method and ammount', ylab= 'Tooth length'
```



Box plot clearly indicates that on average the length of the teeth increases with the increase of the vitamin C dose, i.e. positive correlation between these two seems to be evident. On the other hand it is not clear whether there is any correlation between the teeth length and the form in which C vitamin is administred (orange juice-OJ vs supplement-VC). Next we will use confidence intervals and hypothesis tests to compare tooth growth by delivery method and dosage.

3. Confidence intervals and hypothesis tests - comparing tooth growth by delivery method and dosage

Checking for correlation between the delivery method and change in tooth growth:

```
t.test(len ~ supp, paired = F, var.equal = F, data = ToothGrowth)
```

```
##
##  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
```

A confidence interval of [-0.171, 7.571] does not allow for rejecton of the null hypothesis.

Checking for correlation between the dose level and change in tooth growth:

```
dose1 <- filter(ToothGrowth, dose==0.5|dose==1.0)
dose2 <- filter(ToothGrowth, dose==0.5|dose==2.0)
dose3 <- filter(ToothGrowth, dose==1.0|dose==2.0)
t.test(len ~ dose, paired = F, var.equal = F, data = dose1)
```

```
##
##  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
```

```
t.test(len ~ dose, paired = F, var.equal = F, data = dose2)
```

```
##
##  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
```

```
t.test(len ~ dose, paired = F, var.equal = F, data = dose3)
```

```
##
```

```
##  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
```

As indicated by confidence intervals [-11.98, -6.276] for doses 0.5 and 1.0, [-18.16, -12.83] for doses 0.5 and 2.0, and [-8.996, -3.734] for doses 1.0 and 2.0, the null hypothesis can be rejected which means that there is a significant correlation between tooth length and dose levels.

Checking the data for correlation between dose level and change in tooth growth within each dose level group:

```
dose1 <- filter(ToothGrowth, dose == 0.5)
dose2 <- filter(ToothGrowth, dose == 1.0)
dose3 <- filter(ToothGrowth, dose == 2.0)
t.test(len ~ supp, paired = F, var.equal = F, data = dose1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
t.test(len ~ supp, paired = F, var.equal = F, data = dose2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70             16.77
```

```r
t.test(len ~ supp, paired = F, var.equal = F, data = dose3)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##            26.06            26.14
```

The confidence intervals for dose levels 0.5mg and 1.0mg ([1.72, 8,78] and [2.80, 9.06]), allow for the rejection of the null hypothesis, which indicates that there is a significant correlation between tooth length and dose levels. However, the confidence interval for dose level 2.0 ([-3.80, 3.64]) is not enough to reject the null hypothesis.

4. Assumptions and consequent conclusions

Assumptions:

- The poplulations are independent
- The variances between populations are different
- Random population was used, comprised of similar guinea pigs
- It was a double blind study.

For the populations to be independent, 60 guinea pigs would have to be used in such a manner that each combination of dose level and delivery method was not affected by the other methods.

On the account that given assumptions hold, it may be infered that there is a significant difference between tooth length and dose levels according to the used delivery method. **A higher dose level consistently leeds to longer teeth.** Initially it appeared that the delivery method had no significant impact on tooth length, but when controlling for dose level it emerged that there is a significant difference for 0.5mg and 1.0mg doses, but not for 2.0mg. **Hence it appears that orange juice is a better delivery method with a larger impact on tooth length for doses of 0.5 and 1.0 mg of Vitamin C, but above a certain dose lower than 2.0 mg there is no significant difference.**