Statistical Inference Course Project Part 2: Guinea pig ToothGrowth EDA

Edgar Bahilo Rodríguez

December 12 of 2018

Contents

1 Introduction

This R Markdown notebook develops part 2 of the statistical inference project for the Statistical Inferece part of the Data Science Specialization by Coursera and John Hopkins University. As it is described in the instructions of the assignment the purpose of this document is:

1. Perform some basic statistics for the ToothGrowth dataset.
2. Illustrate via exploratory data analysis the difference between the variables of the dataset.
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. Stating the conclussions of the assignment and the assumptions that you have made.

2 Data loading and summary statistics

The code below displays the basic information and summary statistics for the ToothGrowth dataset.

```
data(ToothGrowth)
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 ...
```

```
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 can look into specif statistics using by function or dplyr syntax:

```
ToothGrowth %>% filter(supp=="VC" & dose==0.5) %>% summary()
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ: 0   Min.   :0.5
##  1st Qu.: 5.95   VC:10   1st Qu.:0.5
##  Median : 7.15           Median :0.5
##  Mean   : 7.98           Mean   :0.5
##  3rd Qu.:10.90           3rd Qu.:0.5
##  Max.   :11.50           Max.   :0.5
```

```
ToothGrowth %>% filter(supp=="OJ" & dose==0.5) %>% summary()
```

```
##       len        supp         dose
##  Min.   : 8.20   OJ:10   Min.   :0.5
##  1st Qu.: 9.70   VC: 0   1st Qu.:0.5
##  Median :12.25           Median :0.5
##  Mean   :13.23           Mean   :0.5
##  3rd Qu.:16.18           3rd Qu.:0.5
##  Max.   :21.50           Max.   :0.5
```

```
ToothGrowth %>% filter(supp=="VC" & dose==1) %>% summary()
```

```
##       len        supp         dose
##  Min.   :13.60   OJ: 0   Min.   :1
##  1st Qu.:15.28   VC:10   1st Qu.:1
##  Median :16.50           Median :1
##  Mean   :16.77           Mean   :1
##  3rd Qu.:17.30           3rd Qu.:1
##  Max.   :22.50           Max.   :1
```

```
ToothGrowth %>% filter(supp=="OJ" & dose==1) %>% summary()
```

```
##       len        supp         dose
##  Min.   :14.50   OJ:10   Min.   :1
##  1st Qu.:20.30   VC: 0   1st Qu.:1
##  Median :23.45           Median :1
##  Mean   :22.70           Mean   :1
##  3rd Qu.:25.65           3rd Qu.:1
##  Max.   :27.30           Max.   :1
```

```
ToothGrowth %>% filter(supp=="VC" & dose==2) %>% summary()
```

```
##       len        supp         dose
##  Min.   :18.50   OJ: 0   Min.   :2
##  1st Qu.:23.38   VC:10   1st Qu.:2
##  Median :25.95           Median :2
##  Mean   :26.14           Mean   :2
##  3rd Qu.:28.80           3rd Qu.:2
##  Max.   :33.90           Max.   :2
```

```
ToothGrowth %>% filter(supp=="OJ" & dose==2) %>% summary()
```

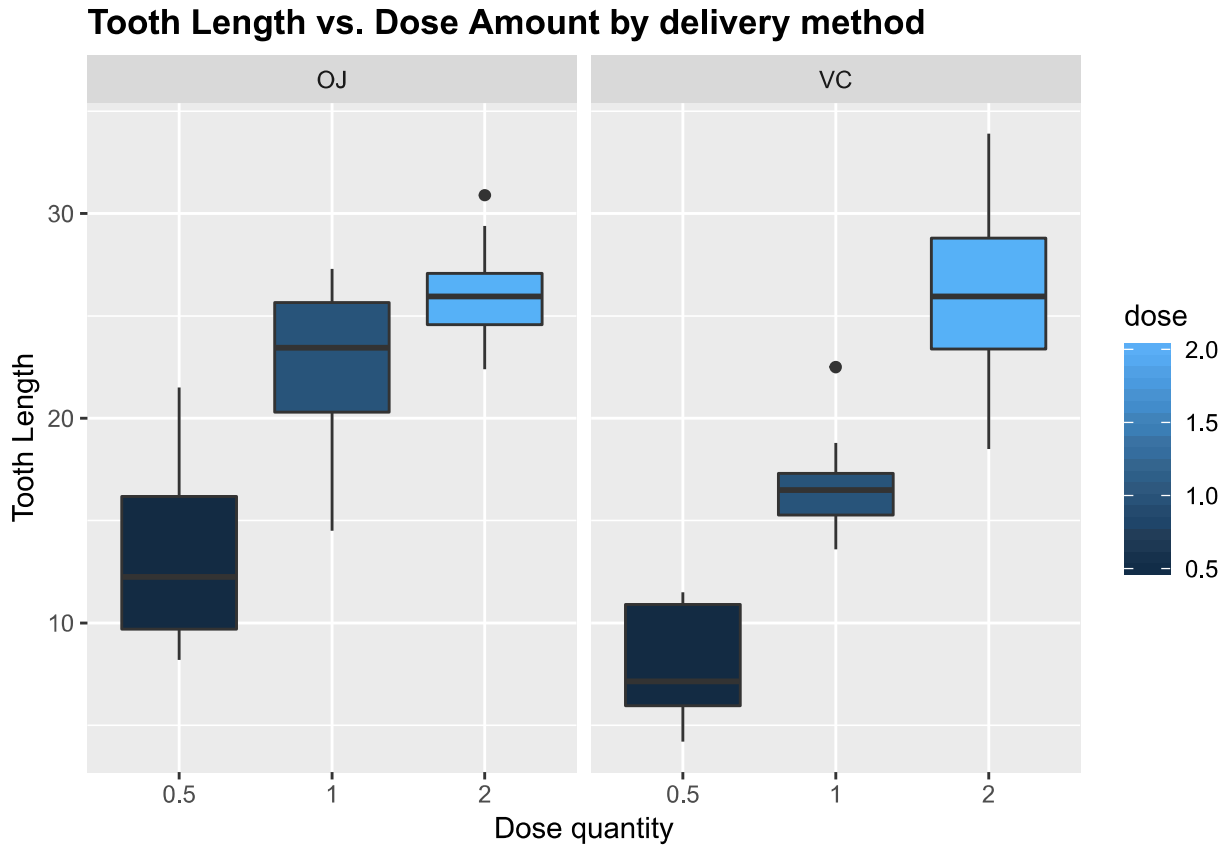
```
##       len        supp         dose
##  Min.   :22.40   OJ:10   Min.   :2
##  1st Qu.:24.57   VC: 0   1st Qu.:2
```

```
##  Median :25.95          Median :2
##  Mean   :26.06          Mean   :2
##  3rd Qu.:27.07          3rd Qu.:2
##  Max.   :30.90          Max.   :2
```

3 Data Visualization

We can use ggplot to visualise boxplots for the different suplements type (VC and OJ) and for the different doses.

```
ggplot(aes(x=as.factor(dose), y=len), data=ToothGrowth) + geom_boxplot(aes(fill=dose))+
xlab("Dose quantity") + ylab("Tooth Length") + facet_grid(~ supp) +
ggtitle("Tooth Length vs. Dose Amount by delivery method") +
theme(plot.title = element_text(lineheight=.8, face="bold"))
```



Now we compare separating the dose:

```
ggplot(aes(x=supp, y=len), data=ToothGrowth) + geom_boxplot(aes(fill=supp)) +
xlab("Supplement Delivery") + ylab("Tooth Length") + facet_grid(~ dose) +
ggtitle("Tooth Length vs. Delivery Method by Dose Amount") +
theme(plot.title = element_text(lineheight=.8, face="bold"))
```

4 Statistical tests

We check now the variances of the 2 suplement groups before performing any statistical test.

```
meansupp = split(ToothGrowth$len, ToothGrowth$supp)
sapply(meansupp, var)
```

```
##       OJ       VC
## 43.63344 68.32723
```

The difference seems considerable therefore we will assume unequal variance between the two groups. Now we try to find if the tooth length is afected by the supplement type:

```
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
```

P-value is greater than 0.05 and the confidence interval contains zero. This indicates that we can not reject

the null hypothesis that the different supplement types have no effect on tooth length. Now we try to find if the dose is the significant parameter that drives the tooth lenght, we create three different subsets with difference doses and we run the t-student test for all of the subsets.

```
dose1 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
dose2 <- subset(ToothGrowth, dose %in% c(0.5, 2.0))
dose3 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))
all_1<-list(dose1, dose2, dose3)
for (i in 1:length(all_1)){
  print(t.test(len ~ dose, paired = F, var.equal = F, data = all_1[[i]]))
}
```

```
##
##  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
##
##
##  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
##
##
##  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
```

And now we do the same but analyzing the data for dose level and change in tooth growth within for each specific dose level:

```
dose4 <- subset(ToothGrowth, dose==0.5)
dose5 <- subset(ToothGrowth, dose==1.0)
dose6 <- subset(ToothGrowth, dose==2.0)
all_2<-list(dose4, dose5, dose6)
for (i in 1:length(all_2)){
```

```
  print(t.test(len ~ supp, paired = F, var.equal = TRUE, data = all_2[[i]]))
}
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.770262 8.729738
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
##
##
##  Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 18, p-value = 0.0007807
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.840692 9.019308
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
##
##
##  Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 18, p-value = 0.9637
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.722999  3.562999
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

For dose levels $0.5mg$ and $1.0mg$ the confidence intervals are as follows: $[1.72, 8, 78]_{0.5mg}$ and $[2.80, 9.06]_{1.0mg}$. Looking at the p-values of each test ($p - value_{0.5mg} = 0.006$ and $p - value_{1.0mg} = 0.001$) we can reject the null hypothesis and state the significant correlation between tooth length and dose levels. On the other hand, the confidence interval for dose level 2.0 is $[-3.80, 3.64]_{2.0mg}$ and$p - value_{2.0mg} = 0.9639$ so it is most likely that for this dose level the lenght is not correlated with the dose.

5 Conclusions

1. Supplement type has no effect on tooth growth.
2. Increasing the dose level leads to increased tooth growth.
3. It seems that orange juice is better than pure vitamin C with larger impact in tooth lenght. However as it was stated above, at higher doses the suplement type seems irrelevant ($\mu_{OJ-2.0MG} = 26.06$ and $\mu_{VC-2.0MG} = 26.14$)

6 Assumptions

1. The experiment is assuming that each guinea pig was randomly assigned to a combination of dosage and supplement type, therefore each sample was independent of each other.
2. The sample of 60 Guinea pigs is assumed to be representative of all Guinea pigs, therefore it is assumed that we can extrapolate these results to the whole population.
3. For t-test regarding tooth length per supplement type, the variances are assumed to be different for the two groups being compared. For t-tests regarding tooth length per dosage level, the variances are assumed to be equal for the three combinations of the two groups being compared.