Basic Inferential Data Analysis: Tooth Growth

Gloria Q

2/2/2020

Introduction

This assignment was conducted as a requirement for the John Hopkins *Statistical Inference* course offered via Coursera, with the intent to perform exploratory and statistical analysis on the ToothGrowth data set in the datasets R package.

Basic summary of data

Importing data and understand general structure of data set:

```
#load the data set to R and save it into a data frame variable
library(datasets)
ToothGrowth <- ToothGrowth
#to read background info on the data set, will appear in Help window of Rstudio
?ToothGrowth
#to get a first original picture of the data set
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 ...
```

```
#to get a feel of the number distribtuion in ToothGrowth$len
summary(ToothGrowth$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.07   19.25   18.81   25.27   33.90
```

```
# to get an idea on the different doses administered
table(ToothGrowth$dose)
```

```
##
## 0.5   1   2
##  20  20  20
```

Based on the background info provided by the help viewer and a quick look at the data set structure, this data set consists of odontoblast (cells involved in tooth growth) length quantification in 60 guinea pigs who were administered vitamin C (VitC) at various doses (0.5, 1 or 2 mg/day) via orange juice (OJ) or ascorbic acid (VC). This information indicates that there are 6 test groups in this data set, 3 groups for each dose times 2 different sources of VitC. The next interesting point to examine would be how VitC dosage versus source impacts tooth length.

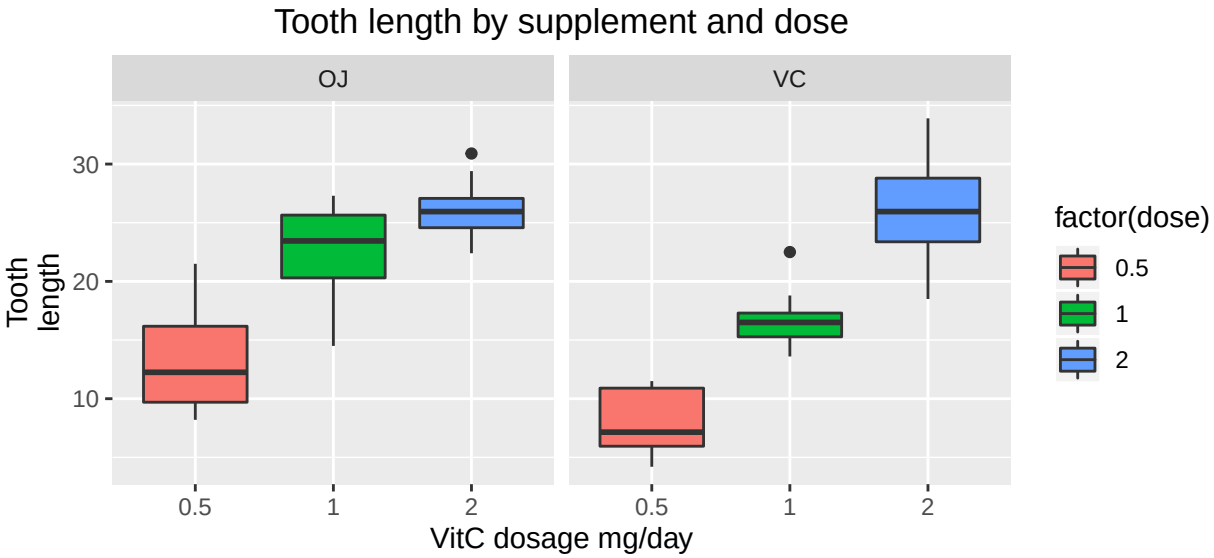
Basic exploratory analysis

The next step is to visualize tooth length distribution and the impact VitC with various doses plays on it:

```r
#load necessary package to R
library(dplyr); library(knitr); library(kableExtra)
#perform a summary statistic (mean and variance) for each group on tooth length
Stats <- ToothGrowth %>% group_by(supp, dose) %>% summarize(meanLength=mean(len), varLength=var(len))
kable(Stats, "latex", booktabs=T) %>% kable_styling(latex_options = "striped", position="center")
```

supp	dose	meanLength	varLength
OJ	0.5	13.23	19.889000
OJ	1.0	22.70	15.295556
OJ	2.0	26.06	7.049333
VC	0.5	7.98	7.544000
VC	1.0	16.77	6.326778
VC	2.0	26.14	23.018222

```r
library(ggplot2)
#plot VitC and dosage influence on tooth growth
tBox <- ggplot(ToothGrowth, aes(x=factor(dose), y=len, fill=factor(dose))) +
geom_boxplot() + facet_grid(. ~ supp) + labs(x="VitC dosage mg/day", y="Tooth
length", title="Tooth length by supplement and dose") + theme(plot.title = element_text(hjust=0.5))
print(tBox)
```



Based on summary statistics and graph above, there are different degrees of impact VitC source has on tooth length which appears to be dependent on dosage amount. To better evaluate this, more in depth hypothesis testing on these 6 groups will be conducted.

Statistical Analysis

Hypothesis t-testing will be used here for the small n for each group and it will be assumed that the guinea pigs used for data collection originated from a normal population and were randomly selected for each of the test groups.

1) Vitc source comparison for each dose: $H_o : \mu_{OJ} = \mu_{VC}$ and $H_\alpha : \mu_{OJ} > \mu_{VC}$ is tested for each group per dose. The summary table below:

```r
#load necessary packages to R
library(tidyr); library(broom)
#t-test OJ vs VC for each dose
stats <- ToothGrowth %>% group_by(dose) %>% do(tidy(t.test(len ~ supp,
```

```
paired=FALSE, var.equal=FALSE, conf.level=0.95, data=.))) %>% rename(OJmean =
estimate1,VCmean = estimate2,tStat = statistic,df = parameter)
#print only necessary columns:
kable(stats[,c(1,3:9)], "latex", booktabs=T, align="c") %>%
kable_styling(latex_options = "striped", position="center")
```

dose	OJmean	VCmean	tStat	p.value	df	conf.low	conf.high
0.5	13.23	7.98	3.1697328	0.0063586	14.96875	1.719057	8.780943
1.0	22.70	16.77	4.0327696	0.0010384	15.35767	2.802148	9.057852
2.0	26.06	26.14	-0.0461361	0.9638516	14.03982	-3.798071	3.638070

For comparing tooth length OJ vs VC for 0.5 and 1 mg/day dose, the h_o is rejected, but not for 2 mg/day. For this comparison, the 0.5 and 1 mg/day dose play a role in tooth length when comaring VitC sources.

2) VitC source comparison: $H_o : \mu_{OJ} = \mu_{VC}$ and $H_\alpha : \mu_{OJ} > \mu_{VC}$ is tested. R code as follows:

```
#t-test OJ vs VC
suppStat <- ToothGrowth %>% do(tidy(t.test(len ~ supp,
paired=FALSE, var.equal=FALSE, conf.level=0.95, data=.))) %>% rename(OJmean =
estimate1,VCmean = estimate2,tStat = statistic,df = parameter)
#print only necessary columns
kable(suppStat[,c(2:8)], "latex", booktabs=T, align="c") %>%
kable_styling(latex_options = "striped", position="center")
```

OJmean	VCmean	tStat	p.value	df	conf.low	conf.high
20.66333	16.96333	1.915268	0.0606345	55.30943	-0.1710156	7.571016

Based on the p-value here, the H_o cannot be rejected, there it cannot be assumed that VitC source alone plays a role in tooth length.

4) Validate the correct p-value was used for hypothesis testing:

```
#collect all previously calculate p-values into 1 vector then adjust using BH method
pVal <- c(pull(stats, var=p.value), pull(suppStat, var=p.value))
pValadj <- p.adjust(pVal, method="BH"); pTable <- rbind(OriginalpVal=pVal, NewPval=pValadj)
colnames(pTable) <- c("0.5", "1", "2", "OJ vs VC")
#print table
kable(pTable, "latex", booktabs=T, align="c") %>% kable_styling(latex_options = "striped", position="c")
```

	0.5	1	2	OJ vs VC
OriginalpVal	0.0063586	0.0010384	0.9638516	0.0606345
NewPval	0.0127172	0.0041535	0.9638516	0.0808460

The number of test that fail to reject the null hypothesis prior to the correct is the same as post-correction.

Conclusion

Based on the different hypothesis testing conducted and those that were able to reject H_o, VitC source alone does not have an effect on tooth length, but the lower dosages, 0.5 and 1 md/day, has an effect on tooth length. The effect of dosage alone wasn't tested because it seemed repetitive, takes up too much space in the report without an appendix and it appeared evident in the exploratory analysis already that there would be an obvious impact of dosage on tooth length.