Inferential Data Analysis

Alberto Macías

9/29/2020

Synopsis

We realize an inferential data analysis on the ToothGrowth data set in order to study how vitamin C provided via orange juice or via ascorbic acid affects in the tooth length in 60 guinea pigs.

Loading the data and performing the EDA

```
library(ggplot2)
library(dplyr)
library(datasets)
data("ToothGrowth")
```

According to R documentation, Toothgrowth data set contains data about the "length of odontoblast (which are cells responsible for tooth growth) in 60 guinea pigs". It is a data frame with 3 variables and 60 observations, one for each guinea pig. The variables are len, a numeric vector that contains the lengths of the cells in unspecified units; supp, a two level factor that indicates the supplement type of vitamin C: orange juice (OJ) or ascorbic acid (VC); and dose, a numeric vector that contains the dosage in which vitamin C was provided, in 3 dosages, 0.5, 1 and 2 milligrams per day (mg/day).

We divide ToothGrowth data set into to data frames, vc and oj, divided by the supplement type. We compute mean, standard deviation and variance of the cell lengths

```
vc <- subset(ToothGrowth, supp=="VC")
oj <- subset(ToothGrowth, supp=="OJ")
mnvc <- mean(vc$len); varvc <- var(vc$len)
mnoj <- mean(oj$len); varoj <- var(oj$len)
mnvc; varvc;
```

```
## [1] 16.96333
```

```
## [1] 68.32723
```

```
mnoj; varoj
```

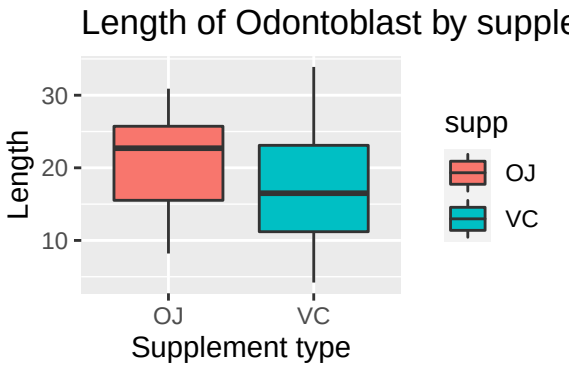
```
## [1] 20.66333
```

```
## [1] 43.63344
```

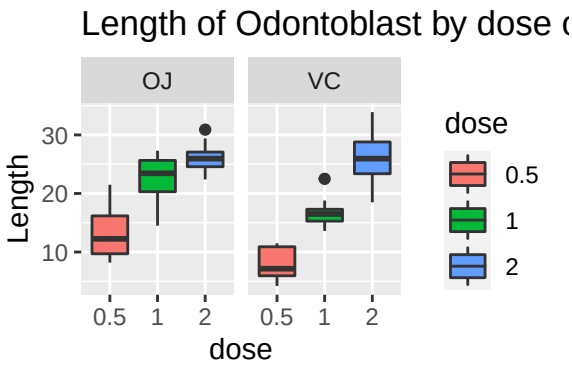
We note that the mean of length of cells in oj data is greater than mean of length in vc, which could indicate that the orange juice supplement of vitamin c affects more to cell length than ascorbic acid. Also that data in the orange juice set is less variable.

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x=supp, y=len, fill= supp))+
            geom_boxplot()+
            ggtitle("Length of Odontoblast by supplement")+
```

```
            ylab("Length")+
            xlab("Supplement type")
```




```
ggplot(ToothGrowth, aes(x=dose, y=len, fill=dose, group = dose))+
            geom_boxplot()+
            facet_wrap(~supp)+
            ggtitle("Length of Odontoblast by dose of ascorbic acid.")+
            ylab("Length")
```



If we observe the data in oj and vc,specially in oj, we can see also that with greater dose it seems that there is greater lenght and is les variable.

Inferential Analysis

We want to analyse statistically first if providing vitamin C by orange juice is more effective than by ascorbic acid and we perform a paired t test to do that. Since we are not beet told if the variance in booth groups is equal, we don not suppose it.

```
t.test(oj$len, vc$len, alternative = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj$len and vc$len
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
```

```
##  20.66333  16.96333
```

From the t test we see that the difference in means is greater than 0, which we already suspect and the p-value in this case is 0.03032 so at a 5% type error I rate we would not reject the null hypthesis, that is, we see evidence that orange juice supplement has more effect on the growth of odontoblast.

At last, we analyse if the 2 mg/day dose has more effect in the growth of odontoblast.

```
oj2 <- filter(oj, dose==2)
vc2 <- filter(vc, dose==2)
t.test(oj2$len, vc2$len, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj2$len and vc2$len
## 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 of x mean of y
##     26.06     26.14
```

So, the t statistic is pretty close to 0, the p-value is close to 1 and the confidences interval contains 0, so it does not seem that the dosage is significant in the estudy

Conclusion

From the inferential analysis of the data, it seem likely probable that providing vitimin C via orange juice has more effect in the lenght of odontoblast that via ascorbic acid has. Also, it seem that there is no much difference on the 2 mg/day dosage between booth supplements.

References

- Statistical inference for data science by Brian Caffo
- Basic R Guide for NSC Statistics by Deanna Li