Statistical Inference Course Project - Basic Inferential Data Analysis

Prykhodko Pavel

June 9, 2018

Overview

In this portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.
Our dataset represents the response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs.
Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods: orange juice (OJ) or ascorbic acid (a form of vitamin C and coded as VC).
As a potential data scientists we will try to decide, which way of receiving vitamin C is more efficient.

To reproduce this exercise you will need *datasets*, *gridExtra* and *ggplot2* library.

```
library(ggplot2)
library(gridExtra)
library(datasets)
```

Data summary

Before analyzing our data let's first load it and give it a quick look:

```
toothGrowthData <- datasets::ToothGrowth
head(toothGrowthData)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
summary(toothGrowthData)
```

```
##       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
```

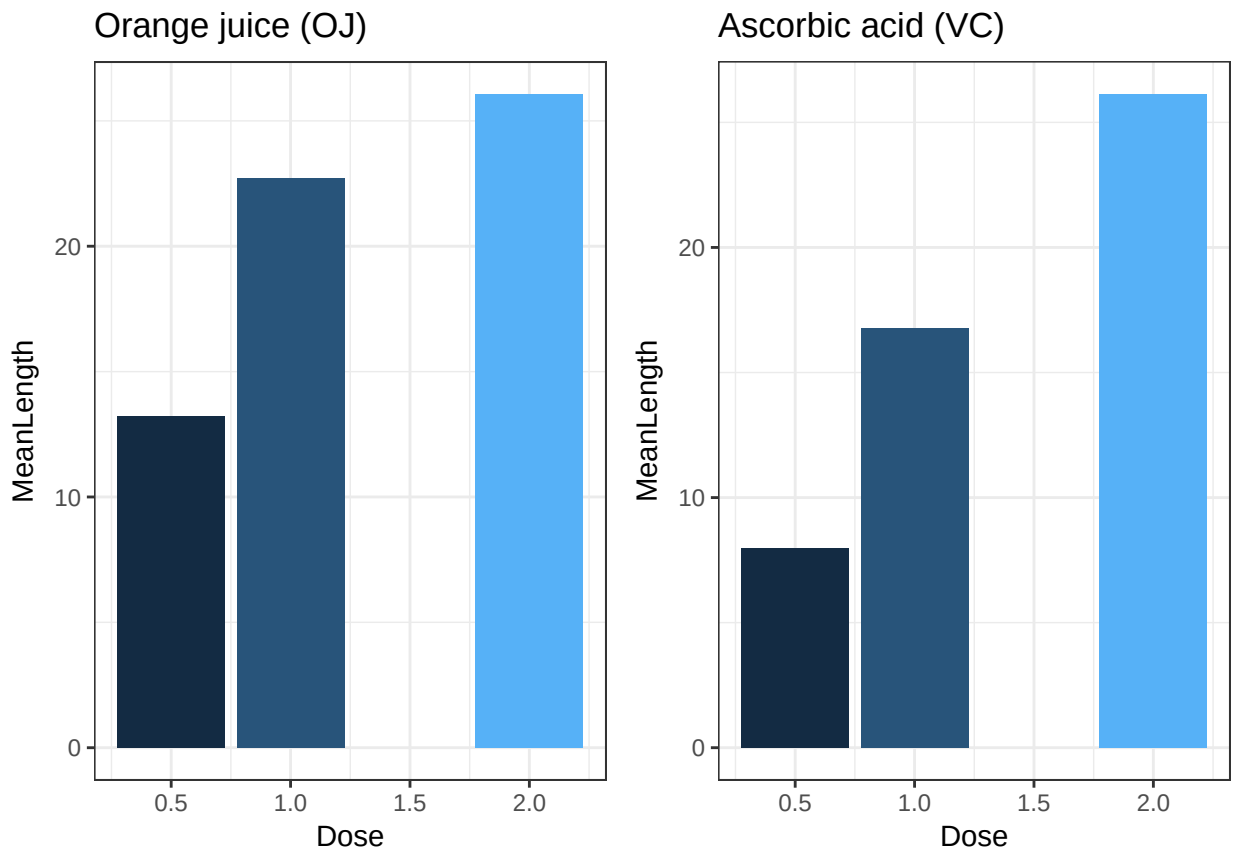
In this dataset we have 2 supplements (VC and OJ) with 30 observations each. Each 30 observations of supplement have 3 different doses administered (0.5, 1.0 and 2.0). So 10 guinea pigs got 0.5 VC, 10 - 0.5 OJ, 10 - 1.0 VC, 10 - 1.0 OJ and so on.

Data visualisation

Let's explore the data visually:
We will break our dataset in two by supplements, calculate the mean growth by doses and display that in two geom_bar plots.

```
growthDataBySupplement <- split(toothGrowthData, toothGrowthData$supp)
growthDataOJ <- growthDataBySupplement$OJ
meanOJGrowthByDose <- aggregate(growthDataOJ$len, by=list(growthDataOJ$dose), FUN=mean)
names(meanOJGrowthByDose) <- c('Dose', 'MeanLength')
growthDataVC <- growthDataBySupplement$VC
meanVCGrowthByDose <- aggregate(growthDataVC$len, by=list(growthDataVC$dose), FUN=mean)
names(meanVCGrowthByDose) <- c('Dose', 'MeanLength')

grid.arrange(
    ggplot(meanOJGrowthByDose, aes(Dose, MeanLength, fill = Dose)) +
      geom_bar(stat = 'identity') + ggtitle("Orange juice (OJ)") +
      theme_bw() + guides(fill = FALSE),
    ggplot(meanVCGrowthByDose, aes(Dose, MeanLength, fill = Dose)) +
      geom_bar(stat = 'identity') + ggtitle("Ascorbic acid (VC)") +
      theme_bw() + guides(fill = FALSE),
    ncol=2
)
```



Which supplement is more effective

By the first look orange juice seems to be more efficient when administered in low doses, but let's give the data a more detailed view and try to make some statistical inference out of it.

First theory

Our first null hypothesis will be the statement, that both supplements are equal in all dose range.

```
hypotesisOneTestResult <- t.test(len ~ supp, data = toothGrowthData)
hypotesisOneTestResult$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

The P value (`hypotesisOneTestResult$p.value`) is bigger, than significance level: 0.0606345, so we can not reject our null hypothesis.

Second theory

But as you have already guessed from the plot above, we are more interested to compare our supplements on a low doses.

So, our second null hypothesis is that orange juice (OJ) is just as efficient supplement for the tooth growth as ascorbic acid (VC) when they are administered by 0.5 or 1 mg / day.
The alternative hypothesis will be that 0.5 or 1 mg/day dosage of orange juice (OJ) is more efficient that ascorbic acid (VC). Let's run our next test:

```
hypotesisTwoTestResult <- t.test(len ~ supp, data = subset(toothGrowthData, dose == list(0.5, 1)))
hypotesisTwoTestResult$conf.int
```

```
## [1]  2.584956 13.415044
## attr(,"conf.level")
## [1] 0.95
```

```
hypotesisTwoTestResult$p.value
```

```
## [1] 0.006181295
```

This test's P value 0.0061813 is less than the significance level, so the null hypothesis can be rejected and we accept the alternative theory.

Conclusion

I am not sure why would you stimulate the tooth growth for guinea pigs (may be you are creating a guinea pig army?), but for low doses such as 0.5 or 1 mg / day you should use orange juice (OG), that is more efficient for your evil purpose that ascorbic acid.