Course Project Statistical Inference, Part 2

Author: Hendrik L.

August 2015

Overview

This project part applies methods from the JHU/Coursera Statistical inference course in order to investigate the ToothGrowth data set. This dataset resulted from a study done by E. W. Crampton in 1947 on the growth of odontoblast cells (sic!) of the incisor teeth as a criterion of vitamin C intake of the guinea pig during a feeding period of 42 days. The study was published here: The Journal of Nutrition 33(5): 491-504.

1. Retrieval and exploratory analysis of the ToothGrowth data

The ToothGrowth dataset is part of the R datasets library. Firstly, the visualization of the data structure and a basic summary of the statistical properties of the dataset are applied:

```
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)
```

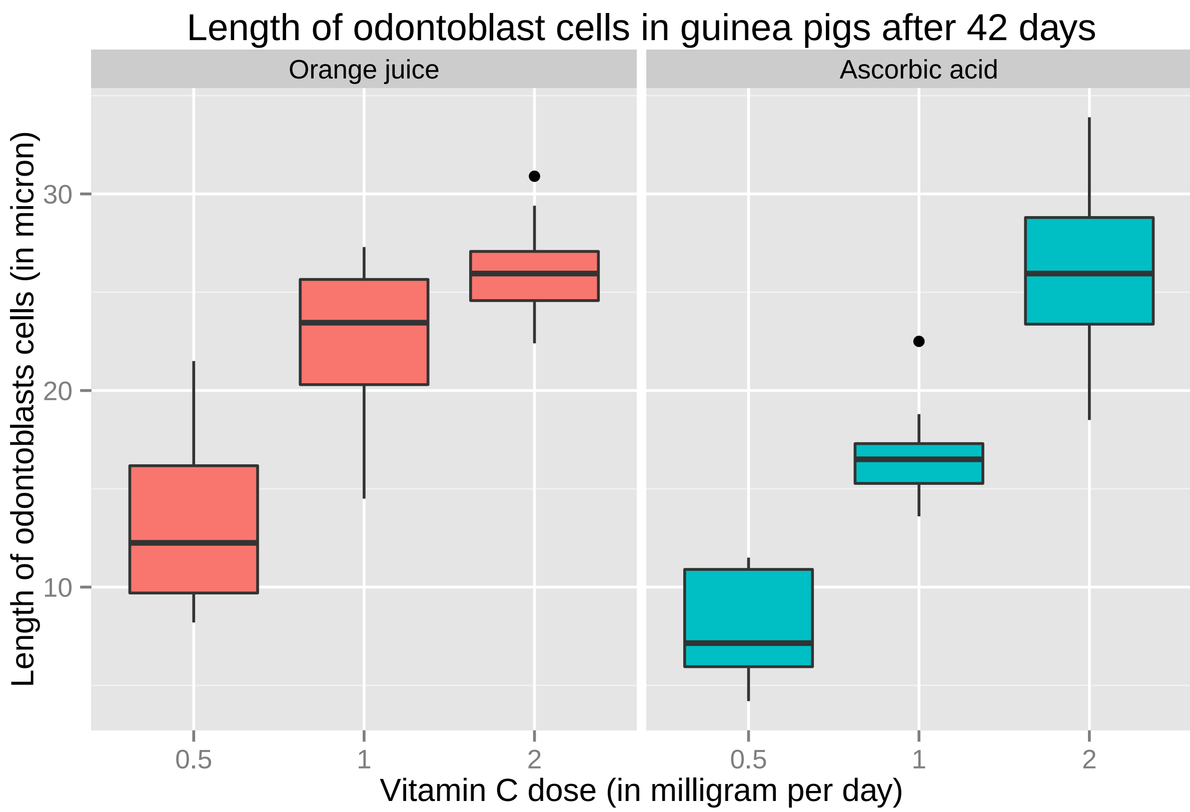
```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000
```

```
nlevels(as.factor(ToothGrowth$dose))
```

```
## [1] 3
```

The analysis above suggests to transform the numeric `dose` variable into a categorical variable, because there are just 3 different dose levels in the whole dataset. The following box plot shows how much the continuous variable `length of odontoblasts cells` (measured in microns) varies with the levels of the two categorical variables `supplement` and `dose`.

```
tgData <- ToothGrowth
tgData$dose <- as.factor(tgData$dose)
levels(tgData$supp) <- c("Orange juice", "Ascorbic acid")
ggplot(tgData, aes(x = dose, y = len)) +
  facet_grid(. ~ supp) +
  geom_boxplot(aes(fill = supp), show_guide = FALSE) +
  labs(title = "Length of odontoblast cells in guinea pigs after 42 days",
    x = "Vitamin C dose (in milligram per day)",
    y = "Length of odontoblasts cells (in micron)")
```



2. Basic data summary

The box plot of the data shows that odontoblast cells grow more with increasing intake dosage of vitamin C. This holds true for both orange juice and ascorbic acid. With a lower dosage, orange juice appears to be more effective than ascorbic acid. However, this difference seems to vanish with an increased dosage.

3. Confidence interval & hypothesis tests

3.1. Test no. 1

Let us assume that vitamin C in orange juice and ascorbic acid leads to the same growth of odontoblast cells across the data set.

```
hypothesisTest1 <- t.test(len ~ supp, data = tgData)
hypothesisTest1$conf.int
```

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

```
hypothesisTest1$p.value
```

```
## [1] 0.06063451
```

The zero lies within the confidence interval.
The p-value is greater than the threshold of 0.05.
The hypothesis cannot be rejected.

3.2. Test no. 2

Let us assume that vitamin C in orange juice and ascorbic acid leads to the same growth of odontoblast cells for a dosage of 0.5mg per day.

```
hypothesisTest2 <- t.test(len ~ supp, data = subset(tgData, dose == 0.5))
hypothesisTest2$conf.int
```

```
## [1] 1.719057 8.780943
## attr(,"conf.level")
## [1] 0.95
```

```
hypothesisTest2$p.value
```

```
## [1] 0.006358607
```

The zero lies outside of the confidence interval.
The p-value lies below the threshold of 0.05.
The hypothesis can be rejected, which suggests that the alternative shall be accepted:
A dosage of 0.5mg vitamin C in orange juice leads to a higher growth of odontoblast cells than the same amount of ascorbic acid.

3.3. Test no. 3

Let us assume that a dosage of 1mg vitamin C per day (via orange juice and ascorbic acid) leads to the same growth of odontoblast cells.

```r
hypothesisTest3 <- t.test(len ~ supp, data = subset(tgData, dose == 1))
hypothesisTest3$conf.int
```

```
## [1] 2.802148 9.057852
## attr(,"conf.level")
## [1] 0.95
```

```r
hypothesisTest3$p.value
```

```
## [1] 0.001038376
```

The zero lies outside of the confidence interval.
The p-value lies below the threshold of 0.05.
The hypothesis can be rejected, which suggests that the alternative shall be accepted:
A dosage of 1mg vitamin C in orange juice leads to a higher growth of odontoblast cells than the same amount of ascorbic acid.

3.4. Test no. 4

Let us assume that a dosage of 2mg vitamin C per day (via orange juice and ascorbic acid) leads to the same growth of odontoblast cells.

```r
hypothesisTest4 <- t.test(len ~ supp, data = subset(tgData, dose == 2))
hypothesisTest4$conf.int
```

```
## [1] -3.79807  3.63807
## attr(,"conf.level")
## [1] 0.95
```

```r
hypothesisTest4$p.value
```

```
## [1] 0.9638516
```

The zero lies within the confidence interval.
The p-value is greater than the threshold of 0.05.
The hypothesis cannot be rejected.

4. Conclusions

For dosages of 0.5mg and 1mg per day, the ToothGrowth dataset suggests that vitamin C intake from orange juice leads a higher growth of odontoblast cells in guinea pigs than from ascorbic acid. However, this effect vanishes with the higher dosage of 2mg per day. Therefore, the accuracy of the assertions we derived from our inital analysis and from the box plot could be confirmed by means of the 4 hypothesis tests.

The underlying assumptions of this work are:

- the phenomenon of odontoblast cell growth follows a normal distribution
- there were no other factors that influenced odontoblast growth significantly (during Crampton's study)