Statistical Inference Course Project

Balint SZEBENYI

2015-07-11

Overview

This document is my course project submission for the coursera course “Statistical Inference”, which is part of the Data Science courses.

The project has got two objectives. One montre carlo simulation to estimate the basic characteristics of the exponential distribution. The second part analyzes the ToothGrowth data of R to illustrate an inference example.

As I am trying to learn dplyr and ggplot2 I have tried to answer the questions using these libraries.

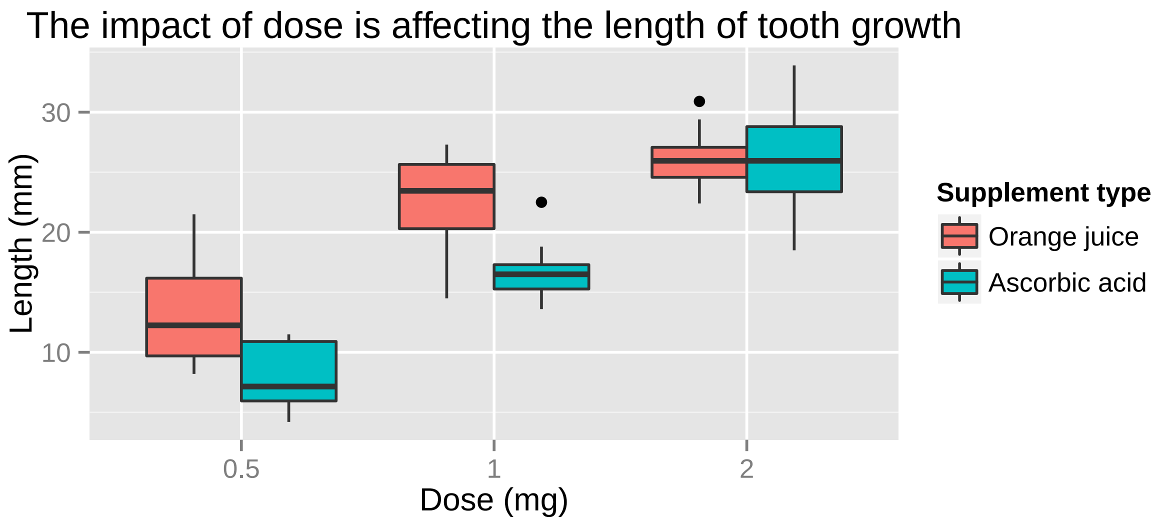
The inference example

The inference exercise is based on the ToothGrowth dataset, which summarizes the results of an experiment with 10 guineapigs, who received VitaminC at three dose levels (0.5, 1, 2 mgs) by two methods (orange juice - OJ, and asorbic acid - VC). The data consists of 30 observations altogether.

Basic summary of the data

```r
data(ToothGrowth)
t <- tbl_df(ToothGrowth)

ggplot(data = t,
       aes(x = factor(dose),
           y = len,
           fill = supp)) +
  geom_boxplot(position = "dodge") +
  labs(title = "The impact of dose is affecting the length of tooth growth",
       x = "Dose (mg)",
       y = "Length (mm)") +
  scale_fill_discrete(name = "Supplement type",
                      labels = c("Orange juice",
                                 "Ascorbic acid"))
```



As it can be seen in the figure above the method how a certain dose is delivered affects its effect on teeth length. A straightforward tendency is shown on the graph, the higher the dose the longer the teeth will become. This relationship is however not linear, doubling the dose does not result in double teeth length. A jump of 0.5 milligrams from 0.5 to 1 makes a larger difference than a jump from 1 to 2 milligrams, at least in the case of ascorbic acid. Orange juice seems to have a positive effect on intake as it results in longer teeth in smaller doses than the same amount delivered via ascorbic acid. When the dose is 2 milligrams the teeth lengths are equal if only the average is observed, but a higher variance is visible in case os ascorbic acid. However, it cannot be announced that ascorbic acid is more unreliable, because in other cases the range of teeth length does not differ as much as in the case of orange juice.

```
t %>%
     group_by(dose, supp) %>%
     summarise(mean(len), sd(len))
```

```
## Source: local data frame [6 x 4]
## Groups: dose
##
##   dose supp mean(len)  sd(len)
## 1  0.5   OJ     13.23 4.459709
## 2  0.5   VC      7.98 2.746634
## 3  1.0   OJ     22.70 3.910953
## 4  1.0   VC     16.77 2.515309
## 5  2.0   OJ     26.06 2.655058
## 6  2.0   VC     26.14 4.797731
```

The boxplots show why it is important to observe not just the mean statistics of data but at least the standard deviation as well. If one observed only the mean then the differences in length spread would not be visible. With boxplots this can be visualized in a concise way. Considering only the averages one could say that orange juice is always better when it comes to smaller doses, but taking into consideration the standard deviation more detail comes into the sunlight.

I have decided against analyzing histograms since the dataset would have been too small for that.

Does the intake method affect tooth growth?

```r
res1 <- t.test(x = ToothGrowth$len[ToothGrowth$supp == "OJ"],
               y = ToothGrowth$len[ToothGrowth$supp == "VC"],
               paired = FALSE,
               var.equal = FALSE)
```

According to the output, it does not, since the p value is 0.0606345. The H_0 hypothesis would be that the there is no difference between the two groups. At the standard 5% significance level one cannot reject the hypothesis, the two groups do not show enough difference which would underpin that the input method affects the experiment. It should be noted however that the sample size is rather small.

Assumptions: I have made the test using unequal variances, to be on the safe side since there was no information about it in the codebook. The test was not a paired one as it has not been stated anywhere. If it would have been a paired test, then the p value would have been 0.00255 supporting a rejection of H_0.

Do higher doses lead to an increase in tooth growth?

```r
res2 <- t.test(x = ToothGrowth$len[ToothGrowth$dose == 0.5],
               y = ToothGrowth$len[ToothGrowth$dose == 1.0],
               paired = FALSE,
               var.equal = FALSE,
               alternative = "less")
```

Using a t test it can be stated that dose indeed affects tooth growth, since the p value is . The same can be read from the confidenece interval. Stating that the confidence interval ranges from minus infinity to r -6.7533227 the test says that the difference between the two groups is at least 7.735 in the dimension of tooth length.

```r
res3 <- t.test(x = ToothGrowth$len[ToothGrowth$dose == 1.0],
               y = ToothGrowth$len[ToothGrowth$dose == 2.0],
               paired = FALSE,
               var.equal = FALSE,
               alternative = "less")
```

Observing the output of test 3 the same is true what we have found before: the group that received a dose of 2 mgs should have a mean that has a mean which is not more than 19.735 plus 4.174.

Knowing from the plots that there was a straight tendency without u or inverted u relations between the variables it can be stated that even the doses 0.5 and 2.0 do not lead to equal tooth growth.

My assumptions in case test 2 and 3 are identical to that of test 1 with the addition of using one-sided tests because it was known which direction to test for.

Appendix

Test results

```r
print(res1)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ"] and ToothGrowth$len[ToothGrowth$supp == "VC"]
## 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 of x mean of y
##  20.66333  16.96333
```

```r
print(res2)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 0.5] and ToothGrowth$len[ToothGrowth$dose == 1]
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

```r
print(res3)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 1] and ToothGrowth$len[ToothGrowth$dose == 2]
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```