Statistical Inference

Part 2: Basic Inferential Data Analysis about tooth growth

MATIAS, V. A.

07/04/2021

Overview

The second part of the project analyzes the dataset ToothGrowth and does some exploratory data analysis, and applies basic statistical inference to bring relevant information

Data

The dataset has 60 rows and 3 columns (appendix, chunk code 1). Each line has the length of the odontoblasts in the study (cell responsible for the growth of the teeth), the supplement used and the size of the dose (0.5, 1 or 2mg / day). The study was carried out on guinea pigs. Let's look at a random sample of 6 tuples.

```
##    len supp dose
## 1 29.4   OJ  2.0
## 2 23.6   OJ  1.0
## 3 21.2   OJ  1.0
## 4 25.8   OJ  1.0
## 5  7.3   VC  0.5
## 6 16.5   VC  1.0
```

Where:

- len: numeric Tooth length
- supp: Supplement type. OJ is Orange Juice, and VC is ascorbic acid (a form of Vitamin C)
- dose: Dose in milligrams/day

There are 30 observations for each of the supplements, as well as 20 for each of the doses. This allows you to analyze the distribution of the data using the histogram below (Figure 1).

The graph shows that there is not a big difference between supplies when the dose is equal to 2mg / day. For the other doses, however, greater growth is noticeable when the supply is orange juice. Even so, the data show that there is an expected trend that increasing the dose leads to greater length of odontoblasts.

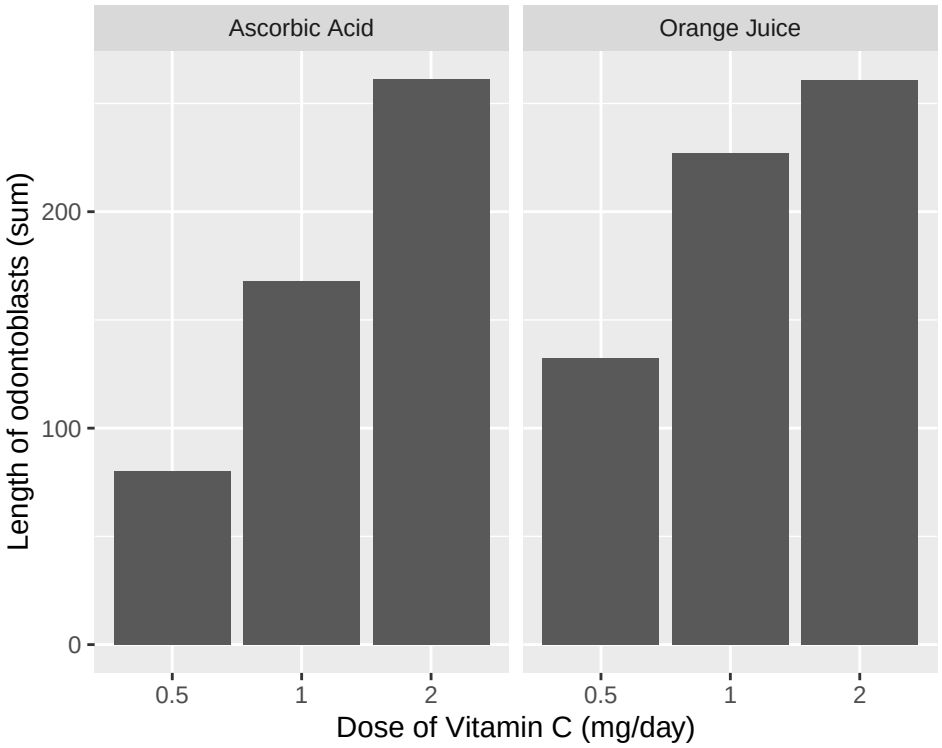

Figure 1: Tooth growth (in odontoblasts) for different doses and supplements of Vitamin C

Inference

To begin, let's analyze the length distribution of odontoblasts looking only at supplements. The blue lines in figure 2 represent the 99% confidence interval. Mean values and confidence intervals are also reported.

```
## Average length for Orange Juice: 20.66
## Average length for Ascorbic Acid: 17
## 99% confidence interval for Orange Juice: 11.79 - 29.54
## 99% confidence interval for Ascorbic Acid: 5.86 - 28.07
```

As there is not much data (20 for each distribution), it is not possible to assume and generalize a normality, however, there is a large concentration of data close to the average of the distributions and few values beyond the level of significance, indicating a possible trend. Still, lower values of length are more common in the distribution of orange juice, considering its average and, mainly, the lower limit of the confidence interval. The distribution for Ascorbic Acid contrasts with the former for the higher density (consequently, probability) of values close to a high average.

To analyze the influence of the dose regardless of the method used, we will use hypothesis tests. Looking at the histogram in figure 1, we can see that there is a tendency for a larger dose to positively influence the length of the odontoblasts, so we are assuming a null hypothesis that there is no influence on the dose, against an alternative hypothesis that the length of the odontoblasts the 2mg dose is higher than the other two.

The other hypothesis test also comes from the analysis of the histogram in figure 1. We can see that there does not seem to be a significant difference between the methods for assessing the length of odontoblasts

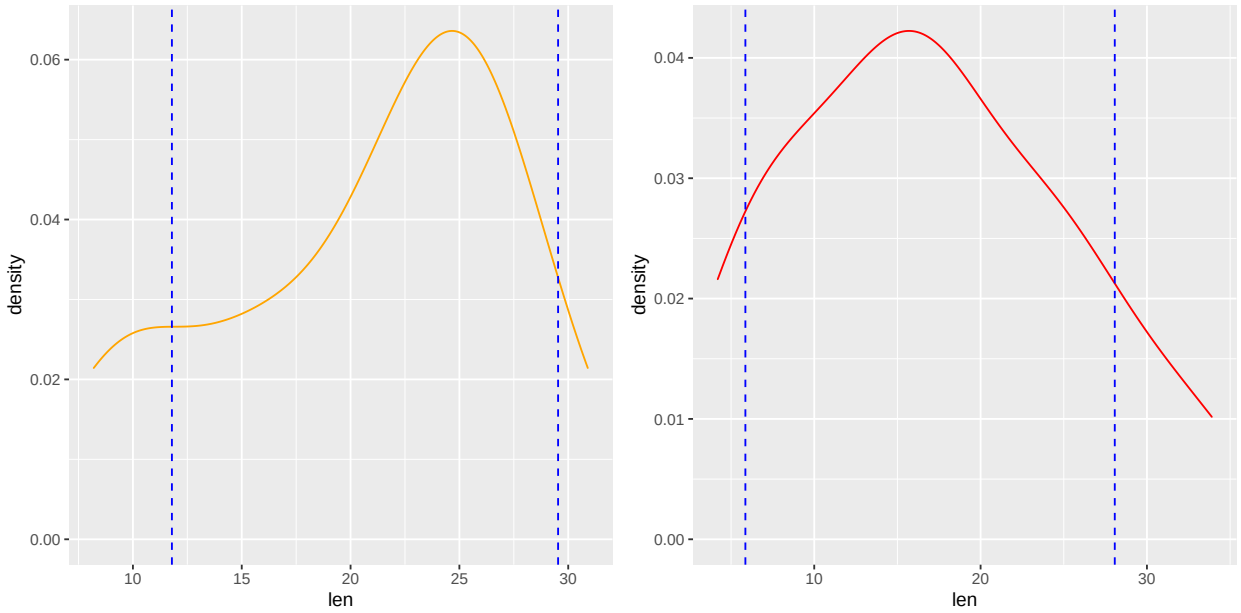

Figure 2: Length distribution of odontoblasts for orange juice supplement (orange line) and Ascorbic Acid (red line)

when the dose is equal to 2. We will test with a null hypothesis that there is no significant difference, against one of which there is a significant difference (greater or lesser). The p-values have been rounded to the fifth decimal place.

```
## P-value for 0.5mg dose is more effective than 2mg dose: 0
## P-value for 1mg dose is more effective than 2mg dose: 0.00001
## P-value for any significant difference between the two methods at
##         a dose of 2mg: 0.06063
```

The p-values for the first two T tests resulted in values very close to zero, demonstrating that, for our data, a dose of 2mg / day is indeed more effective for the growth of odontoblasts compared to the other two.

On the other hand, the p-value of the last T test is not as close to zero as the other two (0.06). As there are approximately 6% of the data varying in the t-distribution tails, there will be an even greater need to measure the result with the level of significance α determined. If $\alpha = 0.05$, we reject the null hypothesis (that there is no significant variation between the treatment methods for 2mg / day), but if we choose a $\alpha = 0.1$, this null hypothesis cannot be refuted. Since we don't have a lot of data, it is acceptable to choose a higher level of significance instead of rejecting the hypothesis right away. To reduce α, more data will be needed.

Conclusions

As was analyzed in the previous item, despite the small amount of data, these seem to tend to a normal one within their confidence intervals. The distribution for orange juices has a distribution center close to 20, while the distribution center for Ascorbic Acid is facing 17 (length of odontoblasts). The 99% confidence intervals show that there is a greater proportion of low length values for the distribution of Ascorbic Acid.

The hypothesis tests showed that the dose of 2mg/day is more effective and that it is very likely (less than 10% probability given the data) that the supplement influences the length of odontoblasts for a dose equal to 2mg/day

Appendix

```
# Chunk code 1: getting the data
library('dplyr')

data('ToothGrowth')

dim(ToothGrowth) # 60 rows X 3 columns

sample_n(ToothGrowth, size = 6) # sample of six tuples

?ToothGrowth # about the data
```

```
# Figure 1 code
library(ggplot2)

table(ToothGrowth$supp) # both supplements have 30 observations
table(ToothGrowth$dose) # 20 observations for each of the three doses


# Changing OJ and VC to orange juice and ascorbic acid
ToothGrowth <- ToothGrowth %>%
  mutate(supp = if_else(supp == 'OJ',
                        true = 'Orange Juice',
                        false = 'Ascorbic Acid'))


#plot
ggplot(ToothGrowth) +
  geom_bar(aes(x = as.factor(dose), y = len), stat = 'identity') +
  facet_wrap(~supp) +
  labs(x = 'Level of Vitamin C (mg/day)',
       y = 'Length of odontoblasts (sum)')
```

```
# Figure 2 code
library(gridExtra)

vc <- ToothGrowth %>%
  filter(supp == 'Ascorbic Acid')

oj <- ToothGrowth %>%
  filter(supp == 'Orange Juice')

stats_oj <- summary(oj$len)
stats_vc <- summary(vc$len)

conf_oj <- mean(oj$len) + c(-1, 1) * qnorm(0.99) * sd(oj$len) / sqrt(length(oj))
conf_vc <- mean(vc$len) + c(-1, 1) * qnorm(0.99) * sd(vc$len) / sqrt(length(oj))

plot_oj <- ggplot() +
  geom_density(aes(len), oj, color = 'orange') +
  geom_vline(xintercept = conf_oj, color = 'blue', linetype ='dashed')
```

```
plot_vc <-ggplot() +
  geom_density(aes(len), vc, color = 'red') +
  geom_vline(xintercept = conf_vc, color = 'blue', linetype ='dashed')


grid.arrange(plot_oj, plot_vc, nrow=1)
```

```
# Chunk code 2: Statistics
message("Average length for Orange Juice: ", round(mean(oj$len), digits = 2),
        "\nAverage length for Ascorbic Acid: ", round(mean(vc$len, digits = 2)),
        "\n99% confidence interval for Orange Juice: ", round(conf_oj[1], digits = 2),
        " - ", round(conf_oj[2], digits = 2),
        "\n99% confidence interval for Ascorbic Acid: ", round(conf_vc[1], digits = 2),
        " - ", round(conf_vc[2], digits = 2))
```

```
#Chunck code 3: P-values

dose_half <- ToothGrowth %>%
  filter(dose == 0.5)
dose_one <- ToothGrowth %>%
  filter(dose == 1.0)
dose_two <- ToothGrowth %>%
  filter(dose == 2.0)


p_value_two_half <- t.test(dose_two$len, dose_half$len,
       alternative = 'greater')$p.value
p_value_two_one <- t.test(dose_two$len, dose_one$len,
       alternative = 'greater')$p.value
p_value_oj_vc <- t.test(oj$len, vc$len)$p.value

message('P-value for 0.5mg dose is more effective than 2mg dose: ',
        round(p_value_two_half, digits = 5),
        '\nP-value for 1mg dose is more effective than 2mg dose: ',
        format(round(p_value_two_one, digits = 5), scientific = FALSE),
        '\nP-value for any significant difference between the two methods at
        a dose of 2mg: ',
        round(p_value_oj_vc, digits = 5))
```