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 some 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 27.3   OJ  2.0
## 2  5.8   VC  0.5
## 3 25.2   OJ  1.0
## 4 22.5   VC  1.0
## 5  6.4   VC  0.5
## 6  9.7   OJ  0.5
```

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.

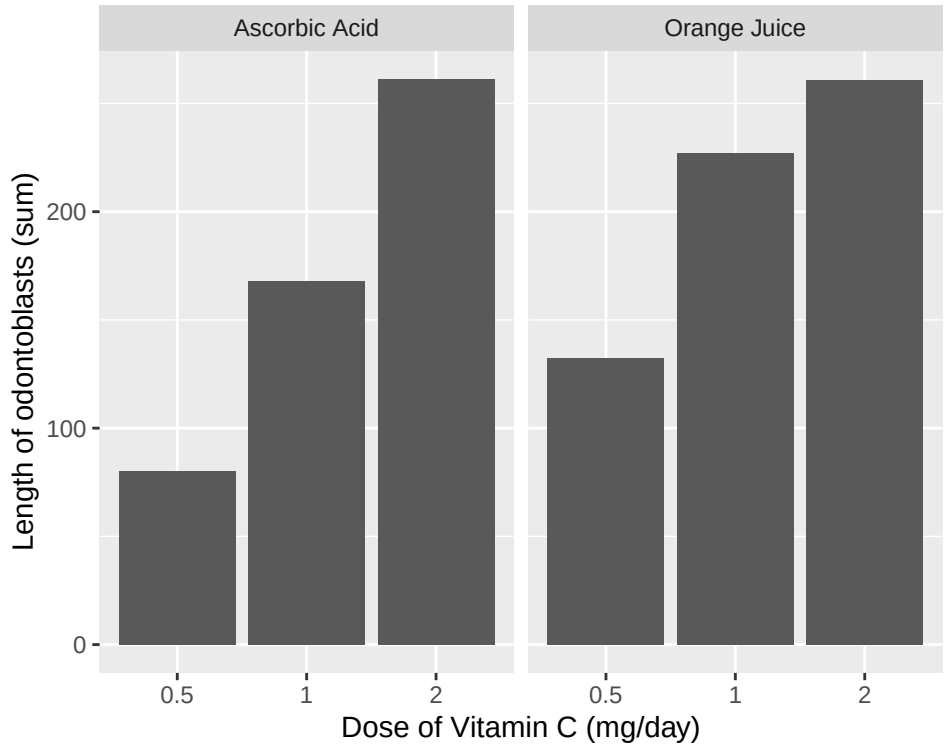

Figure 1: Tooth growth (in odontoblasts) for different doses and supplements of Vitamin C

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.

```
t.test(vc$len, oj$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  vc$len and oj$len
## t = -1.9153, df = 55.309, p-value = 0.06063
```

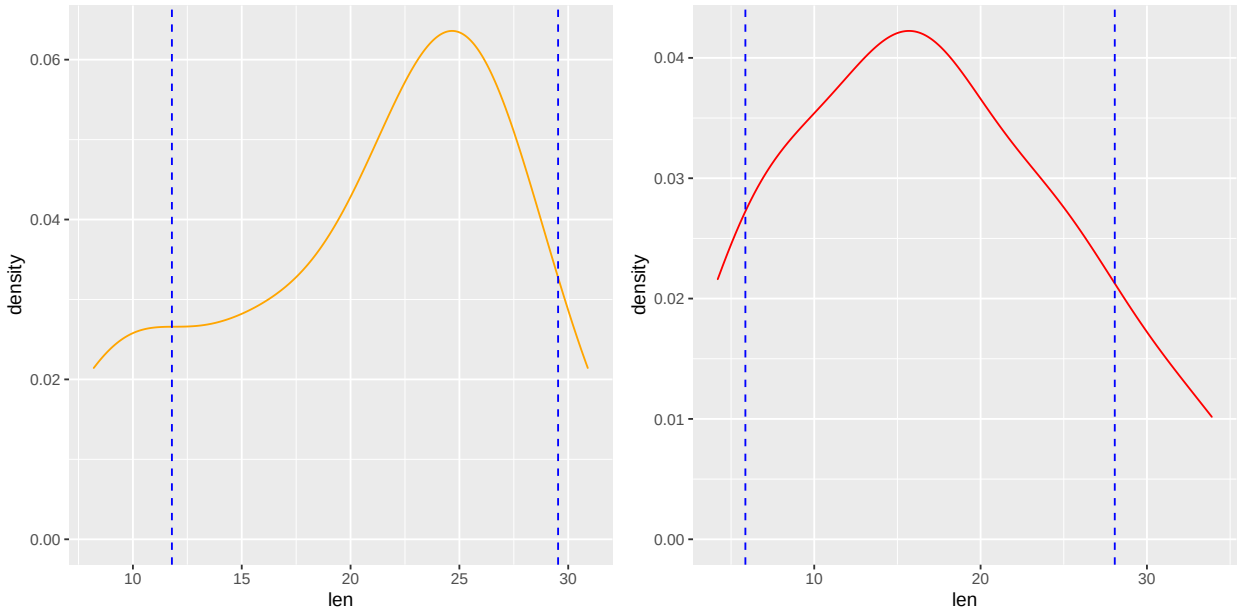

Figure 2: Length distribution of odontoblasts for orange juice supplement (orange line) and Ascorbic Acid (red line)

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

Conclusions

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
```

```r
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)')
```

```r
# 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)
```

```r
# 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))
```