Statistical Inference Course Project - Part 2

Anastasia Kuprina

7/5/2019

Overview

The dataset provides data about the effect of Vitamin C on tooth growth in Guinea pigs. 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 by one of the two delivery methods.

Analysis

Step 1: load data & explore

```
library(datasets)
library(tidyverse)
library(magrittr)
library(boot)
library(glue)

data("ToothGrowth")
df <- ToothGrowth %>%
  as_tibble()

names(df) <- c('tooth_length', 'supplement_type', 'daily_dose')
# I personally dislike names like 'len' and 'supp' as it's easy to forget what they mean

glimpse(df)
```

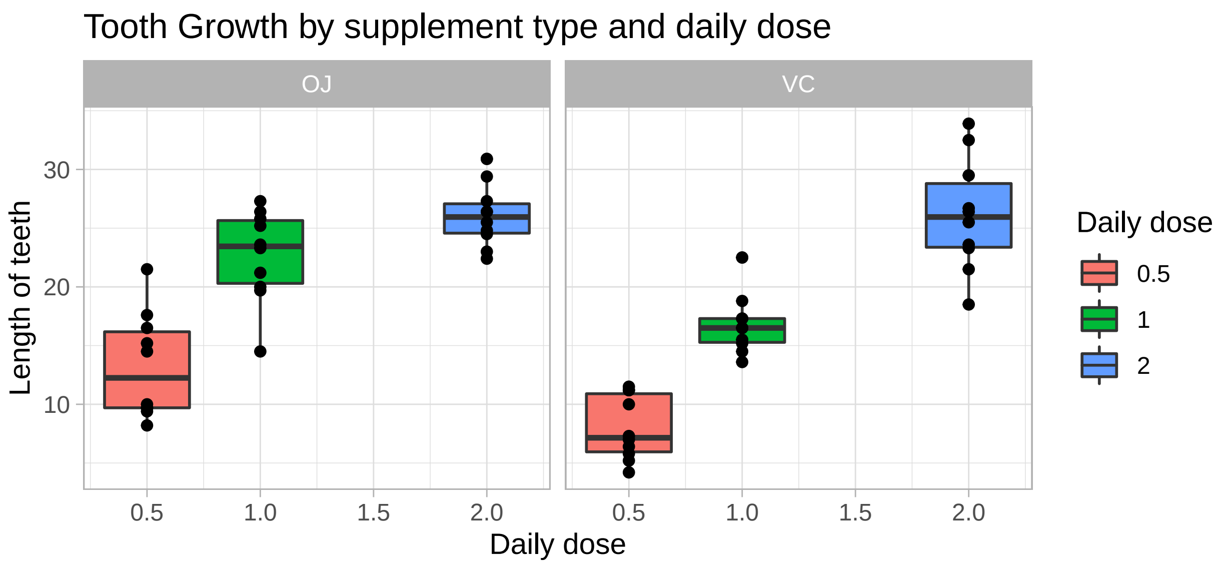
```
## Observations: 60
## Variables: 3
## $ tooth_length    <dbl> 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5....
## $ supplement_type <fct> VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC...
## $ daily_dose      <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0...
```

```
summary(df)
```

```
##   tooth_length   supplement_type   daily_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  
```

Step 2: dig around and plot

We can try to detect visually if any differences are present with faceting that ggplot offers



Clearly there are a few differences for the groups. Daily dose of 2mg is resulting in higher tooth growth (especially when not giben with Orange Juice - OJ). Overall, however, Orange juice (OJ) seems to have more positive effects compared to VC method.

Hypothesis testing

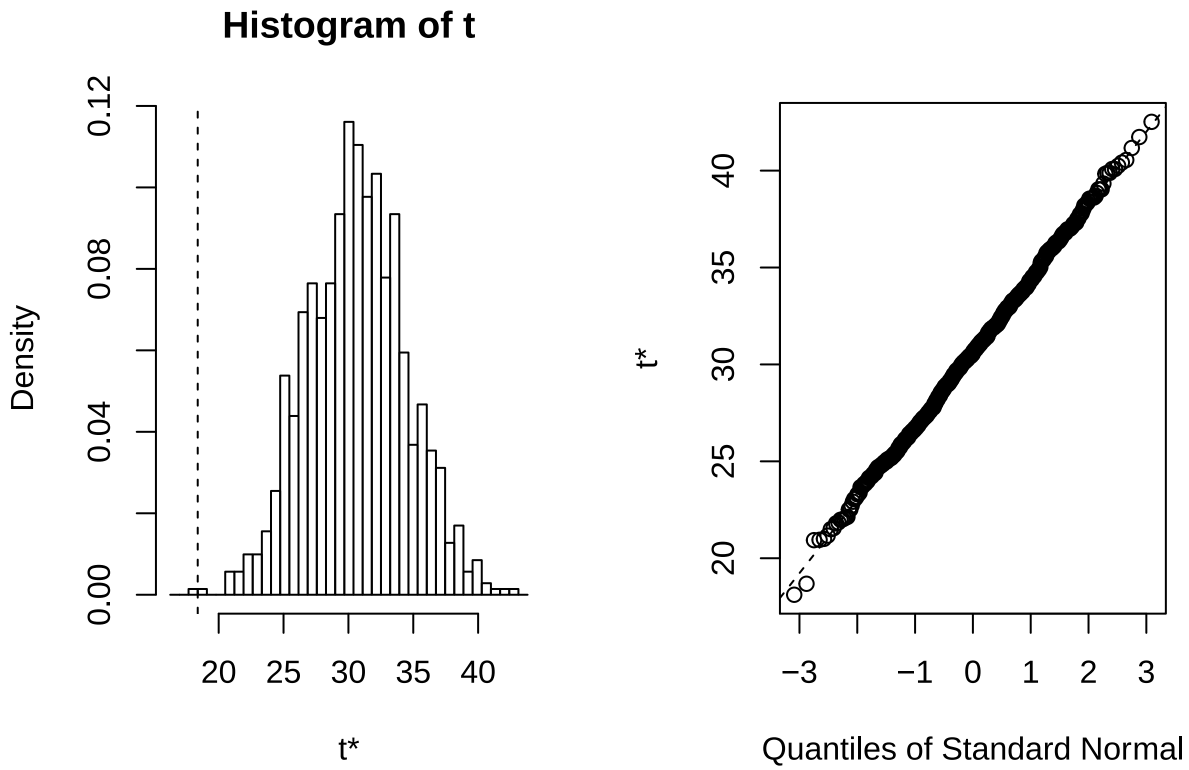
My null hypothesis: supplement type does not have any effect on tooth growth (i.e. tooth growth is the same between supplement types).

In order to run a t-test, we should ensure that means of the data are normally distributed.

```
mean_func <- function(i, x) {
  mean(x[i])
}

is_normal <- function(vector, reps = 1000) {
  boot_mean <- boot(vector, mean_func, reps)
  plot(boot_mean)
}

is_normal(df$tooth_length)
```



Visually checking, the data looks close to normal after 1000 resamples. Now onto the t-test!

```
t.test(tooth_length ~ supplement_type, data = df, var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tooth_length by supplement_type
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

The results of the t-test show no difference between the groups (confidence interval includes 0 for the means). We saw that 2mg dose had higher effect for VC supplement method and thus may be increasing the overall mean. We can try and run the test for data without dose of 2mg in the data. This has drawbacks such as (1) biasing the data, (2) decreasing the sample size and degrees of freedom.

```
t.test(tooth_length ~ supplement_type, data = df %>% filter(daily_dose < 2), var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tooth_length by supplement_type
## t = 3.0503, df = 36.553, p-value = 0.004239
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.875234 9.304766
## sample estimates:
## mean in group OJ mean in group VC
##           17.965           12.375
```

With selectively chosen data, confidence intervals no longer include 0. Based on this, we can conclude that:

- means of our data are normally distributed due to the Central Limit Theorem and thus we have used t-test for hypothesis testing
- we assume variances of the two groups are not equal due to not knowing otherwise

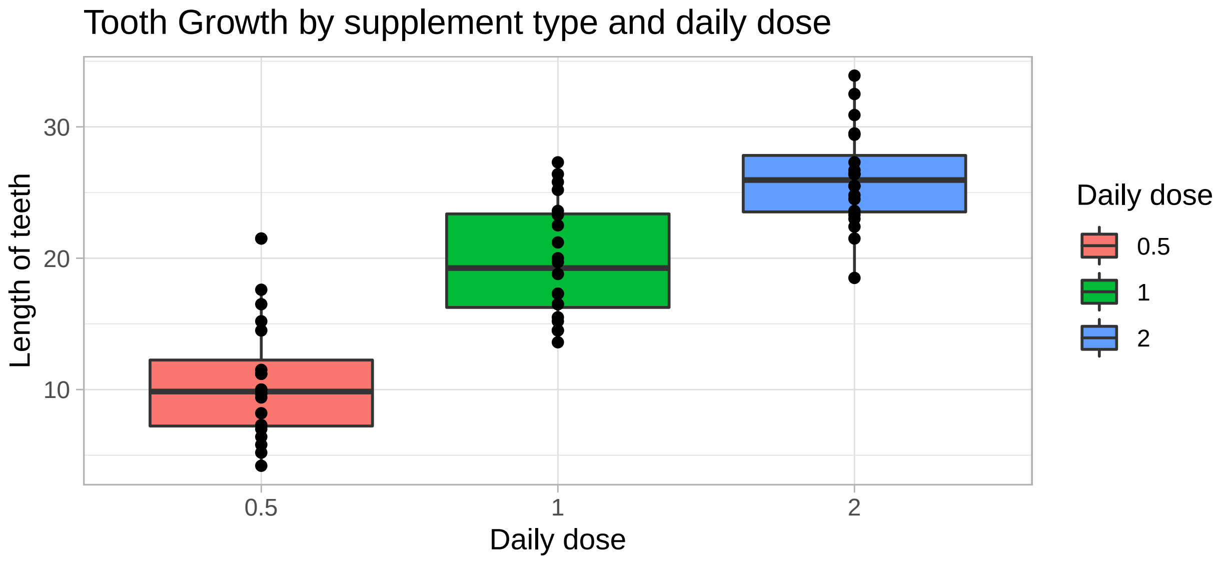
In data terms:

- if we do not take 2mg dose into account, with alpha of 0.05 we can reject null hypothesis and conclude that the **two supplement methods do have effect on tooth growth** (or we have seen an unfortunately extreme sample)
- **if we do take 2mg dose into account, with alpha of 0.05 we accept the null hypothesis and conclude there is no difference in supplement methods**

In practical actionable terms, if we have to stick with VC, we should give a dose of 2mg. If we can give lower doses, we could go with OJ instead as its efficiency at lower doses is higher.

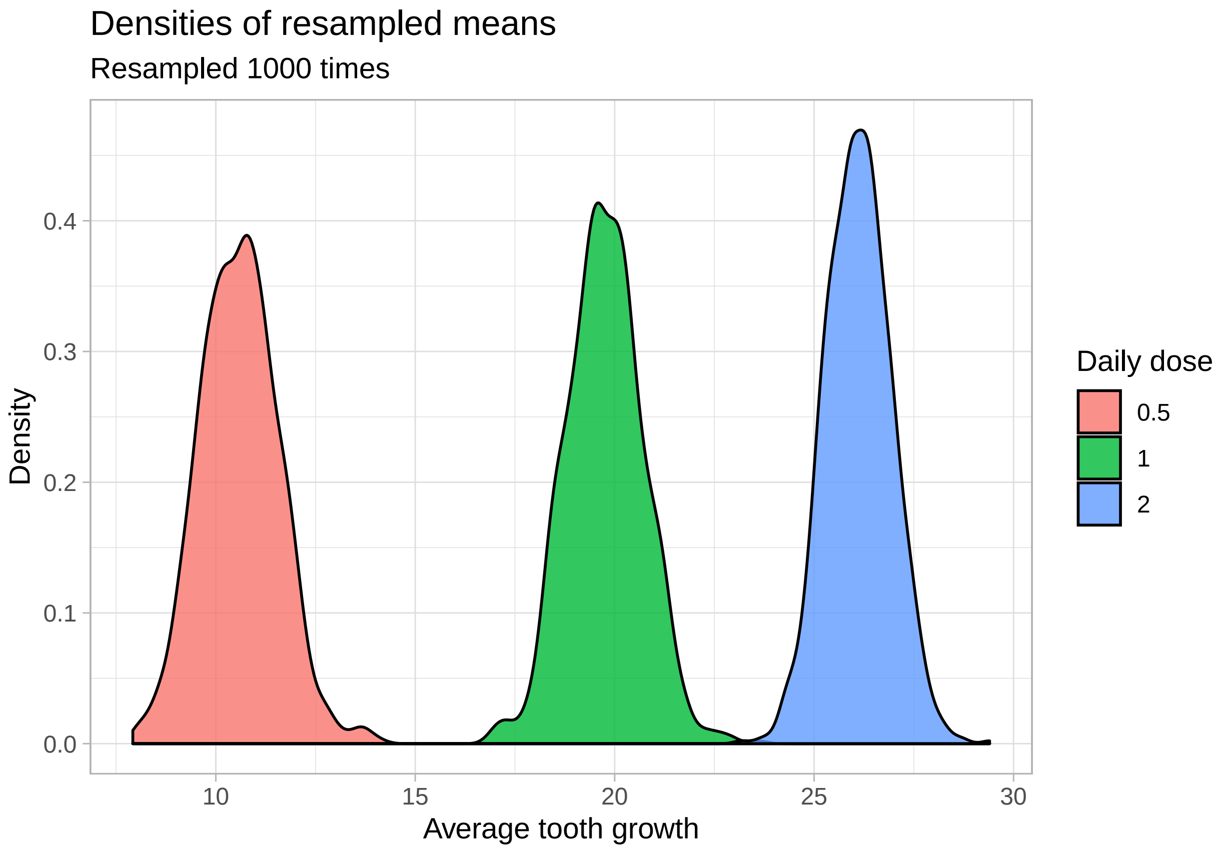
Bootstrapping means by group

We have seen the groups have difference based on the supplement method. However, what about the dose?



There is quite a lot of overlap in single observations but the groups look different. Could the difference be a result of chance?

```
number_of_resamples = 1000
data.frame(boot = 1:number_of_resamples) %>% # how many repetitions are needed?
    group_by(boot) %>%
    do(sample_n(df, nrow(df), replace = T)) %>% # resample with replacement
    mutate(daily_dose = as_factor(daily_dose)) %>% # turn daily dose into factor
    group_by(boot, daily_dose) %>% # group by iteration number and daily dose
    summarise(avg = mean(tooth_length), # find avg
              stdev = sd(tooth_length),
              n_samples = n()) %>%
  ggplot() + # plot
  geom_density(aes(x = avg, fill = daily_dose),
               alpha = 0.8) +
  theme_light() +
  labs(title = 'Densities of resampled means',
       subtitle = glue('Resampled {number_of_resamples} times'),
       x = 'Average tooth growth',
       y = 'Density',
       fill = 'Daily dose')
```



This shows that the means are quite different from each other based on the daily dose amount (very little overlap in the densities). A conclusion we can draw from this, that based on all data (**irrespective of the method) higher doses of Vitamin C result in faster tooth growth**.