Statistical inference - Course Project: Part 2

- Author: Anderson Hitoshi Uyekita
- Date: Monday, 04 July 2022

Synopsis

The Part 2 of Course Project aims to analyze the ToothGrowth database using confidence intervals and/or tests. This dataset has 60 observations and 3 variables, and a summary was provided with a brief of exploratory analysis. As a results of this project, supplement type has no effect on tooth growth and increasing the dose level leads to increased tooth growth.

1. Objectives

- Task 1: Load the `ToothGrowth` data and perform some basic exploratory data analyses
- Task 2: Provide a basic summary of the data.
- Task 3: Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
- Task 4: State your conclusions and the assumptions needed for your conclusions.

2. Requeriments, Settings, and Load Data

Please find the Requirements and Settings to reproduce this experiment in the APPENDIX section or Forking the Github Repository.

3. Loading Data and EDA

> Task 1: Load the `ToothGrowth` data and perform some basic exploratory data analyses

The `ToothGrowth` dataset is part of the `datasets` package.

```
# Loading the ToothGrowth dataset as a tibble.
dataset_tg <- dplyr::as_tibble(datasets::ToothGrowth)
```

According to the `str()` function, the Tooth Growth dataset has 60 observations and 3 variables.

```
## tibble [60 x 3] (S3: tbl_df/tbl/data.frame)
##  $ len : num [1:60] 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 [1:60] 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Finally, the Figure 1 synthesizes the `ToothGrowth` dataset in a box plot.

4. Basic summary of the data

> Task 2: Provide a basic summary of the data.

Following the Course Project instruction, the `summary()` function will provide the basic summary of the `ToothGrowth` dataset.

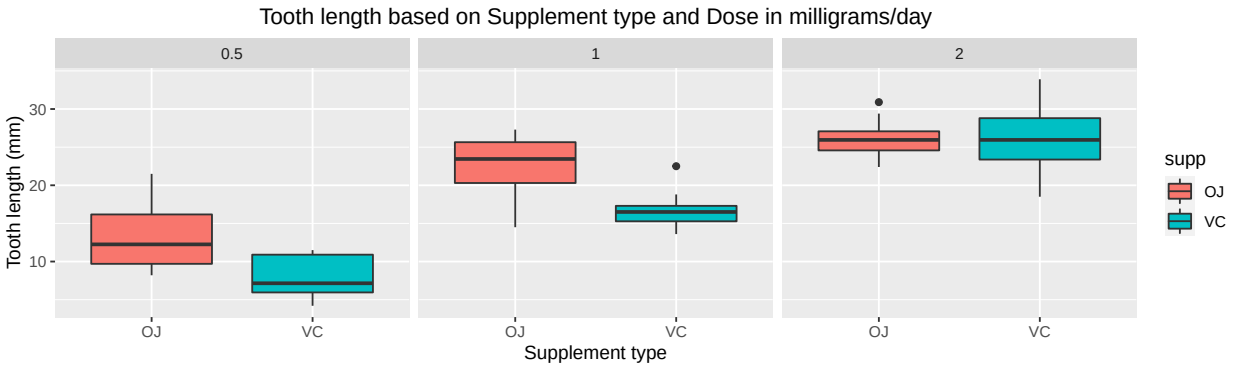

Figure 1: Data Visualization to aid the Exploratory Data Analysis. Graph Source Code in Appendix.

```
##       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
```

For further information about the ToothGrowth dataset, please read the description in R Documentation.

5. Compare tooth growth by supplement and dose

There are more than one comparison of tooth growth by supplement (`OJ` and `VC`) and dose. Thus, to turn this study much clearly we divided this section into 3 parts: Comparison between supplements, 1mg and 0.5 mg dose, and 2 and 1 mg dose.

5.1. Test 1: Growth Tooth Differences between supplements OJ and VC

We are testing if exist some differences between those supplements. It means, we are looking for a p value greater then 0.05. Thus, we need to assume two hypotesis: `Ho` equals means and `H1` means are differents.

First of all, we need to check the `len` variance between `OJ` and `VC` supplement.

Those variance are far different so the `var.equal` should be set to `FALSE` (`Len` Variance using OJ =43.6334368 and `Len` Variance using VC = 68.3272299). Now, we can use de `t.test` to compare the supplements performances are the same.

```
t.test(dataset_tg_OJ$len, dataset_tg_VC$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dataset_tg_OJ$len and dataset_tg_VC$len
## 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
```

The p-value of the test is 0.06. It means we do not have evidence to reject the null hypothesis. Supplement types seems to have no impact on Tooth growth.

5.2. Test 2: Growth Tooth Differences by dosages 2 and 1 mg/day

For this test we define the `Ho` as the null hypotheses of equal means between the two groups, versus the alternative hypothesis (`H1`) that the two means are different.

```
t.test(filter(dataset_tg,dose==2)$len, filter(dataset_tg,dose==1)$len, paired = FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  filter(dataset_tg, dose == 2)$len and filter(dataset_tg, dose == 1)$len
## t = 4.9005, df = 38, p-value = 1.811e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.735613 8.994387
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

As results of the t.test we have enough evidence to reject the null hypothesis (Ho). It means if I increase the dosage from 1mg to 2mg creates an positive effect on tooth length.

Test 3: Growth Tooth Differences between supplements OJ and VC For this test we define the `Ho` as the null hypotheses of equal means between the two groups, versus the alternative hypothesis (`H1`) that the two means are different.

```
t.test(filter(dataset_tg,dose==1)$len, filter(dataset_tg,dose==0.5)$len, paired = FALSE, var.equal = TRU
```

```
##
##  Two Sample t-test
##
## data:  filter(dataset_tg, dose == 1)$len and filter(dataset_tg, dose == 0.5)$len
## t = 6.4766, df = 38, p-value = 1.266e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.276252 11.983748
## sample estimates:
## mean of x mean of y
##    19.735    10.605
```

As results of the t.test we have enough evidence to reject the null hypothesis (Ho). It means if I increase the dosage from 0.5mg to 1mg creates an positive effect on tooth length.

6. Conclusions

- By the Test 1 we can conclude that supplement has no effect on tooth growth.
- By the Test 2 and Test 3 we can conclude that increasing the dose level leads to increased tooth growth.

Assumptions

- For t-tests regarding tooth length per dosage level, the variances are assumed to be equal for the three combinations of the two groups being compared.

APPENDIX

In order to reproduce this Course Project in any environment, please find below the Packages, Seed definition and `SessionInfo()`.

Requirements

- Requirements to reproduce this exercise: `ggplot2`, `dplyr`, and `datasets`.
- Make a copy of the original dataset and converting into a `dplyr` table.

```
# Loading libraries
library(ggplot2)
library(dplyr)
library(datasets)

# Force results to be in English
Sys.setlocale("LC_ALL", "English.utf8")

# Set seed
set.seed(2022)
```

Session Info

```
## R version 4.2.0 (2022-04-22 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_1.0.9   ggplot2_3.3.6
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9        pillar_1.7.0     compiler_4.2.0   tools_4.2.0
##  [5] digest_0.6.29    lubridate_1.8.0  evaluate_0.15    lifecycle_1.0.1
##  [9] tibble_3.1.7     gtable_0.3.0     pkgconfig_2.0.3  rlang_1.0.3
## [13] cli_3.3.0        DBI_1.1.3        rstudioapi_0.13  yaml_2.3.5
## [17] xfun_0.31        fastmap_1.1.0    withr_2.5.0      stringr_1.4.0
## [21] knitr_1.39.3     generics_0.1.2   vctrs_0.4.1      grid_4.2.0
## [25] tidyselect_1.1.2 glue_1.6.2       R6_2.5.1         fansi_1.0.3
## [29] rmarkdown_2.14   farver_2.1.0     purrr_0.3.4      magrittr_2.0.3
## [33] scales_1.2.0     ellipsis_0.3.2   htmltools_0.5.2  assertthat_0.2.1
```

```
## [37] colorspace_2.0-3 labeling_0.4.2   utf8_1.2.2       stringi_1.7.6
## [41] munsell_0.5.0    crayon_1.5.1
```