Course Project – Part 2: Basic Inferential Data Analysis

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

Synopsis

Part 2 of the Course Project aims to analyze the `ToothGrowth` database using confidence intervals and tests. This dataset has 60 observations on 3 (three) variables and describes the tooth growth in guinea pigs in respect of a vitamin C supplement by two delivery methods. According to the results, it is possible to identify there is no evidence to affirm Orange Juice (`OJ`) and Ascorbic Acid (`VC`) have different performance outcomes. However, there is strong evidence that increasing the vitamin C dosage increases 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 and Settings

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 `Tooth Growth` dataset is part of the `datasets` package. It is about the experiments in guinea pigs *(Cavia porcellus)* feeding with different levels of vitamin C from 2 delivery methods (Orange Juice – `OJ` – and Ascorbic Acid – `VC`).

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

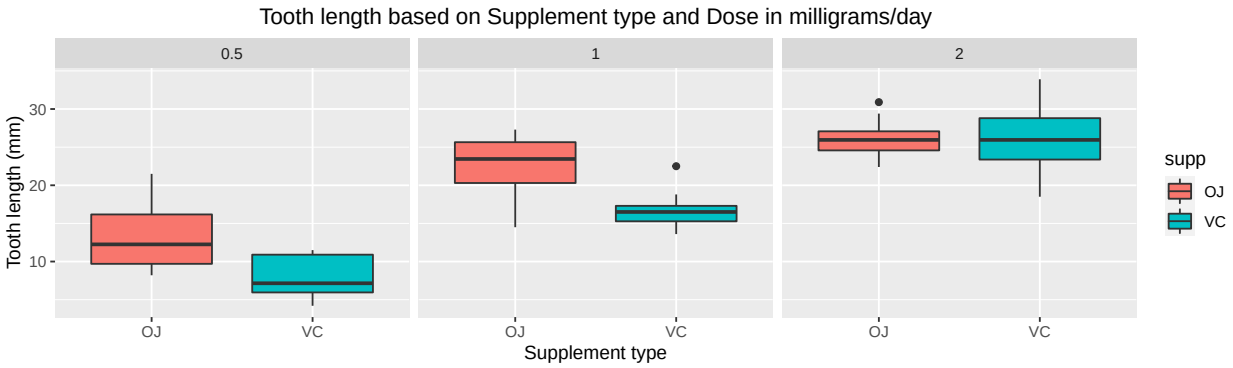

Figure 1: Data Visualization to aid the Exploratory Data Analysis. Graph Source Code in Appendix.

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

```
##       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
```

5. Results

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)

Based on Figure 1, it is possible to state two main hypotheses, which this document will cover in 5.1. and 5.2..

5.1. Hypothesis 1 (Code Source in APPENDIX)

Have the supplements the same performance in tooth growth?

According to the question above, I have formulated the following Hypothesis.

- **The Null hypothesis** ($H_0 : \mu_{(len,OJ)} = \mu_{(len,VC)}$): Both supplement outcomes have an equal average of odontoblasts' length growth.
- **The Alternative hypothesis** ($H_1 : \mu_{(len,OJ)} \neq \mu_{(len,VC)}$): Orange Juice and Ascorbic Acid result in different averages of odontoblasts' length growth.

The `p-value` from "Hypothesis 1" is 0.06, which is greater than $\alpha = 0.05$. Also, the Confidence interval, [-0.17, 7.57], contains zero. For those reasons, there is no evidence to reject the null hypothesis (**Failed to Reject** H_0), so there is **no evidence** to affirm that OJ or VC supplement has different results in tooth growth.

5.2. Hypothesis 2 (Code Source in APPENDIX)

Has the Dose affected Tooth Growth?

Due to the `Dose` variable having three levels (2, 1, and 0.5 mg/day), Hypothesis 2 was divided into three. Therefore, the following Hypothesis statement synthesizes the process of evaluation.

- **The Null hypothesis** ($H_0 : \mu_{(len,X)} = \mu_{(len,Y)}$): X and Y mg/day Doses outcomes have an equal average of odontoblasts' length growth.
- **The Alternative hypothesis** ($H_1 : \mu_{(len,X)} \neq \mu_{(len,Y)}$): X mg/day and Y mg/day Doses result in different averages of odontoblasts' length growth.

Table 1 summarizes the t-test performed to evaluate each pair.

Table 1 – `p-values` and Confidence Intervals of the Hypothesis tests using different `Dose`.

Hipothesis	p-value	Confidence Interval	Decision
$H_0 : \mu_{(len,2.0)} = \mu_{(len,1.0)}$ and $H_1 : \mu_{(len,2.0)} \neq \mu_{(len,1.0)}$	0.00002	[3.74, 8.99]	Reject H_0
$H_0 : \mu_{(len,1.0)} = \mu_{(len,0.5)}$ and $H_1 : \mu_{(len,1.0)} \neq \mu_{(len,0.5)}$	0.0000001	[6.28, 11.98]	Reject H_0
$H_0 : \mu_{(len,2.0)} = \mu_{(len,0.5)}$ and $H_1 : \mu_{(len,2.0)} \neq \mu_{(len,0.5)}$	0.00000000000003	[12.8, 18.2]	Reject H_0

All 3 (three) tests in Table 1 show **strong evidence** that increasing dosage produces higher tooth growth.

6. Conclusions

Task 4: State your conclusions and the assumptions needed for your conclusions.

Finally, by the outcome of Hypothesis 1, there is no evidence to say that Orange Juice performance is different than Ascorbic Acid, so it is possible to conclude both supplements have equal performance. However, according to Hypothesis 2, there is strong evidence that increasing the dosage will boost tooth growth.

Assumptions

Please, find below the assumption adopted in this study.

1. Due to the few observations, the data follows a T-distribution.
2. The data gathered is independent and identically distributed
3. The Variances are considered to be unequal
4. The dataset is considered a representative sample of the population.

APPENDIX

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

A1. 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)
```

A2. 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
```

A3. Figure 1 - Source Code

```
# Plotting the Box-plot using the ToothGrowth dataset.
ggplot(data = dataset_tg,

       # Supplement on x-axis and Tooth Length in y-axis.
       aes(x = supp, y = len)) +

    # Creating the box-plot colored by supplement.
    geom_boxplot(aes(fill = supp)) +

    # Adding title.
    ggtitle(label = "Tooth length based on Supplement type and Dose in milligrams/day") +

    # Defining x-axis label.
    xlab("Supplement type") +

    # Defining y-axis label.
    ylab("Tooth length (mm)") +

    # Dividing into facets.
    facet_grid(cols = vars(dose)) +

    # Adjusting the title position.
    theme(plot.title = element_text(hjust = 0.5))
```

A4. Hypothesis 1

Comparison Orange Juice and Ascorbic Acid

```
##
##  Two Sample t-test
##
## data:  base::subset(dataset_tg, supp == "OJ")$len and base::subset(dataset_tg, supp == "VC")$len
## t = 1.9153, df = 58, p-value = 0.06039
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1670064  7.5670064
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

A5. Hypothesis 2

Comparison Dosage 2mg/day and 1mg/day

```
##
##  Two Sample t-test
##
## data:  base::subset(dataset_tg, dose == 2)$len and base::subset(dataset_tg, dose == 1)$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
```

Comparison Dosage 1mg/day and 0.5mg/day

```
##
##  Two Sample t-test
##
## data:  base::subset(dataset_tg, dose == 1)$len and base::subset(dataset_tg, dose == 0.5)$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
```

Comparison Dosage 2mg/day and 0.5mg/day

```
##
##  Two Sample t-test
##
## data:  base::subset(dataset_tg, dose == 2)$len and base::subset(dataset_tg, dose == 0.5)$len
## t = 11.799, df = 38, p-value = 2.838e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.83648 18.15352
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```