Basic Inferential Data Analysis

JZstats
Course 6: Statistical Inference
Data Science Specialization, Coursera

OVERVIEW

Based on the 'ToothGrowth' data by Galton, a comparison was conducted for the effects of Vitamin C in tooth growth of Guinea Pigs between the groups defined by two factors, supply method and dose as well as their interaction.

1. EXPLORATORY DATA ANALYSIS

The 'ToothGrowth' data consists of 60 independent observations for the length of tooth in Guinea Pigs over two factors, the supply method and the dose of Vitamin C.

1.1 By Supply Method

Vitamin C was supplied either as *Orange Juice* (OJ) or as *Ascorbic Acid* (VC). For the groups defined by the supply method, 1 Research Question was formed:

RQ-01: Is length of tooth bigger, when Vitamin C is supplied as orange juice instead of ascorbic acid?

Figure 1

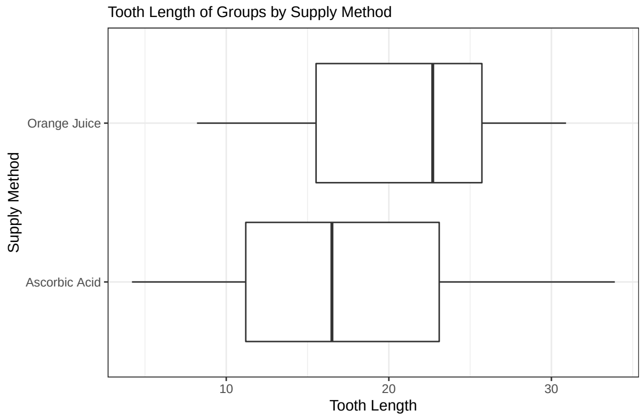

Table 1: Statistics by Supply Method

Supply Method	n	mean	sd
Ascorbic Acid	30	16.96333	8.266029
Orange Juice	30	20.66333	6.605561

SUPPLEMENTARY INFORMATION
Further details can be found at the GitHub repository:
`https://github.com/jzstats/Statistical-Inference-Course-Project/tree/master/basic_inferential_data_analysis`

1.2 By Dose

Vitamin C was supplied in doses of 0.5, 1 or 2 mg/day. For the groups defined by dose, 3 Research Question were formed:

RQ-02: Is length of tooth bigger, when 2 instead of 0.5 mg/day of Vitamin C is supplied?
RQ-03: Is length of tooth bigger, when 2 instead of 1 mg/day of Vitamin C is supplied?
RQ-04: Is length of tooth bigger, when 1 instead of 0.5 mg/day of Vitamin C is supplied?

Figure 2

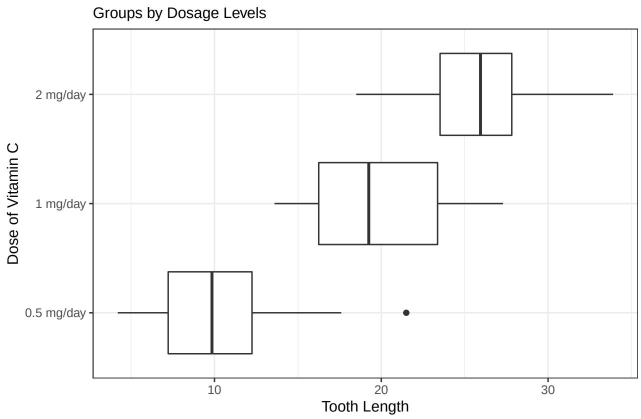

Table 2: Statistics by Dose

Dose of Vitamin C	n	mean	sd
0.5 mg/day	20	10.605	4.499763
1 mg/day	20	19.735	4.415436
2 mg/day	20	26.100	3.774150

1.3 By Supply Method and Dose

For the groups defined by the interaction of supply method and dose, 9 Research Questions were formed:

For constant dose:
RQ-05: Is length of tooth different, when 2 mg/day of Vitamin C is supplied as orange juice instead of ascorbic acid?
RQ-06: Is length of tooth bigger, when 1 mg/day of Vitamin C is supplied as orange juice instead of ascorbic acid?
RQ-07: Is length of tooth bigger, when 0.5 mg/day of Vitamin C is supplied as orange juice instead of ascorbic acid?

For constant supply method:
RQ-08: Is length of tooth bigger, when 2 instead of 0.5 mg/day of Vitamin C is supplied as orange juice?
RQ-09: Is length of tooth bigger, when 2 instead of 1 mg/day of Vitamin C is supplied as orange juice?
RQ-10: Is length of tooth bigger, when 1 instead of 0.5 mg/day of Vitamin C is supplied as orange juice?
RQ-11: Is length of tooth bigger, when 2 instead of 0.5 mg/day of Vitamin C is supplied as ascorbic acid?
RQ-12: Is length of tooth bigger, when 2 instead of 1 mg/day of Vitamin C is supplied as ascorbic acid?
RQ-13: Is length of tooth bigger, when 1 instead of 0.5 mg/day of Vitamin C is supplied as ascorbic acid?

Figure 3
Tooth Length of Groups by Supply Method and Dose

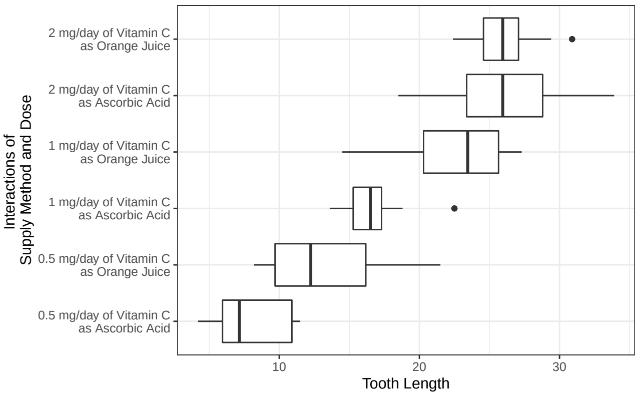

Table 3: Statistics by Supply Method and Dose

Interactions of Supply Method and Dose	n	mean	sd
0.5 mg/day of Vitamin C as Ascorbic Acid	10	7.98	2.746634
0.5 mg/day of Vitamin C as Orange Juice	10	13.23	4.459708
1 mg/day of Vitamin C as Ascorbic Acid	10	16.77	2.515309
1 mg/day of Vitamin C as Orange Juice	10	22.70	3.910953
2 mg/day of Vitamin C as Ascorbic Acid	10	26.14	4.797731
2 mg/day of Vitamin C as Orange Juice	10	26.06	2.655058

2. STATISTICAL ANALYSIS

2.1 Assumptions

A major assumption was made, to restrict the methodology only in the approaches that had been discussed in the course:

(A1): The length of tooth for all groups, follows Normal distribution with unknown expected value and variance.

2.2 Multiple t-tests

Each of the Research Questions, was appropriately translated into a statistical hypothesis test:

Hypothesis test for (RQ-01):
$[H_0{:}\mu_{OJ} \leq \mu_{VC}]$ VS $[H_a{:}\mu_{OJ} > \mu_{VC}]$

Hypothesis test for (RQ-02):
$[H_0{:}\mu_{2mg} \leq \mu_{0.5mg}]$ VS $[H_a{:}\mu_{2mg} > \mu_{0.5mg}]$

Hypothesis test for (RQ-03):
$[H_0{:}\mu_{2mg} \leq \mu_{1mg}]$ VS $[H_a{:}\mu_{2mg} > \mu_{1mg}]$

Hypothesis test for (RQ-04):
$[H_0{:}\mu_{1mg} \leq \mu_{0.5mg}]$ VS $[H_a{:}\mu_{1mg} > \mu_{0.5mg}]$

Hypothesis test for (RQ-05):
$[H_0{:}\mu_{OJ:2mg} = \mu_{VC:2mg}]$ VS $[H_a{:}\mu_{OJ:2mg} \neq \mu_{VC:2mg}]$

Hypothesis test for (RQ-06):
$[H_0{:}\mu_{OJ:1mg} \leq \mu_{VC:1mg}]$ VS $[H_a{:}\mu_{OJ:1mg} > \mu_{VC:1mg}]$

Hypothesis test for (RQ-07):
$[H_0{:}\mu_{OJ:0.5mg} \leq \mu_{VC:0.5mg}]$ VS $[H_a{:}\mu_{OJ:0.5mg} > \mu_{VC:0.5mg}]$

Hypothesis test for (RQ-08):
$[H_0{:}\mu_{OJ:2mg} \leq \mu_{OJ:0.5mg}]$ VS $[H_a{:}\mu_{OJ:2mg} > \mu_{OJ:0.5mg}]$

Hypothesis test for (RQ-09):
$[H_0{:}\mu_{OJ:2mg} \leq \mu_{OJ:1mg}]$ VS $[H_a{:}\mu_{OJ:2mg} > \mu_{OJ:1mg}]$

Hypothesis test for (RQ-10):
$[H_0{:}\mu_{OJ:1mg} \leq \mu_{OJ:0.5mg}]$ VS $[H_a{:}\mu_{OJ:1mg} > \mu_{OJ:0.5mg}]$

Hypothesis test for (RQ-11):
$[H_0{:}\mu_{VC:2mg} \leq \mu_{VC:0.5mg}]$ VS $[H_a{:}\mu_{VC:2mg} > \mu_{VC:0.5mg}]$

Hypothesis test for (RQ-12):
$[H_0{:}\mu_{VC:2mg} \leq \mu_{VC:1mg}]$ VS $[H_a{:}\mu_{VC:2mg} > \mu_{VC:1mg}]$

Hypothesis test for (RQ-13):
$[H_0{:}\mu_{VC:1mg} \leq \mu_{VC:0.5mg}]$ VS $[H_a{:}\mu_{VC:1mg} > \mu_{VC:0.5mg}]$

Under the assumption (A1), for the hypothesis above, 13 Welch two sample t-tests were conducted.

2.3 Adjust p-values

The p-values that were originally obtained, were adjusted (to compensate for the multiple tests) by the Benjamini–Hochberg procedure so that the False Discovery Rate (FDR) was bounded to be at most 0.05.

2.4 Results

For all hypothesis tests, except one for the (RQ-05), there were enough evidence to reject the NULL hypothesis H_0 in favor of the alternative H_a.

Table 4: Results

RQ	x	y	p_adj	is_sig
01	OJ	VC	0.0328437	Yes
02	2mg	0.5mg	0.0000000	Yes
03	2mg	1mg	0.0000207	Yes
04	1mg	0.5mg	0.0000003	Yes
05	OJ:2mg	VC:2mg	0.9638516	No
06	OJ:1mg	VC:1mg	0.0007499	Yes
07	OJ:0.5mg	VC:0.5mg	0.0041331	Yes
08	OJ:2mg	OJ:0.5mg	0.0000017	Yes
09	OJ:2mg	OJ:1mg	0.0231608	Yes
10	OJ:1mg	OJ:0.5mg	0.0000744	Yes
11	VC:2mg	VC:0.5mg	0.0000002	Yes
12	VC:2mg	VC:1mg	0.0000744	Yes
13	VC:1mg	VC:0.5mg	0.0000011	Yes

3. CONCLUSIONS

From the results of the Statistical Analysis (displayed in Table 4), the following conclusions were drawn for the 13 Research Questions of interest:

For (RQ-01), the data provides substantial evidence ($p_{(RQ-01)} < 0.0329$) to reject the NULL hypothesis H_0 in favor of the alternative H_a, according to which the expected tooth length is bigger when Vitamin C is supplied as orange juice instead of ascorbic acid.

For (RQ-02), (RQ_03) and (RQ-04), the data provides substantial evidence ($p_{(RQ-02)} < 0.0001$, $p_{(RQ-03)} < 0.0001$ and $p_{(RQ-04)} < 0.0001$ respectively) to reject the NULL hypothesis H_0 in favor of the alternatives H_a, according to which the expected tooth length is bigger when the dose of Vitamin C is 2 instead of 0.5 mg/day, 2 instead of 1 mg/day and 1 instead of 0.5 mg/day respectively (independently of the supply method).

For (RQ-05), the data does NOT provide substantial evidence ($p_{(RQ-05)} > 0.9638$) to reject the NULL hypothesis H_0, according to which the expected tooth length is the same when dose of 2 mg/day of Vitamin C is supplied either as orange juice or as ascorbic acid.

For (RQ-06) and (RQ-07), the data provides substantial evidence ($p_{(RQ-06)} < 0.0008$ and $p_{(RQ-07)} < 0.0042$ respectively) to reject the NULL hypothesis H_0 in favor of the alternative H_a, according to which the expected tooth length is bigger when dose of either 1 or 0.5 mg/day of Vitamin C is supplied as orange juice instead of ascorbic acid.

For (RQ-08), (RQ_09) and (RQ-10), the data provides substantial evidence ($p_{(RQ-08)} < 0.0001$, $p_{(RQ-09)} < 0.02317$ and $p_{(RQ-10)} < 0.0001$ respectively) to reject the NULL hypothesis H_0 in favor of the alternative H_a, according to which the expected tooth growth is bigger when Vitamin C is supplied as orange juice in dose of 2 instead of 0.5 mg/day, 2 instead of 1 mg/day and 1 instead of 0.5 mg/day respectively.

For (RQ-11), (RQ_12) and (RQ-13), the data provides substantial evidence ($p_{(RQ-11)} < 0.0001$, $p_{(RQ-12)} < 0.0001$ and $p_{(RQ-13)} < 0.0001$ respectively) to reject the NULL hypothesis H_0 in favor of the alternative H_a, according to which the expected tooth growth is bigger when Vitamin C is supplied as ascorbic acid in dose of 2 instead of 0.5 mg/day, 2 instead of 1 mg/day and 1 instead of 0.5 mg/day respectively.

4. REFERENCES

Caffo, B. (2016). Statistical inference for data science. Retrieved from `https://leanpub.com/LittleInferenceBook`

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289–300. `http://www.jstor.org/stable/2346101`.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL `http://www.rstudio.com/`.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, `https://doi.org/10.21105/joss.01686`

Hao Zhu (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. `https://CRAN.R-project.org/package=kableExtra`

JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL `https://rmarkdown.rstudio.com`.

Yihui Xie and J.J. Allaire and Garrett Grolemund (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. URL `https://bookdown.org/yihui/rmarkdown`.

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

JJ Allaire, Yihui Xie, R Foundation, Hadley Wickham, Journal of Statistical Software, Ramnath Vaidyanathan, Association for Computing Machinery, Carl Boettiger, Elsevier, Karl Broman, Kirill Mueller, Bastiaan Quast, Randall Pruim, Ben Marwick, Charlotte Wickham, Oliver Keyes, Miao Yu, Daniel Emaasit, Thierry Onkelinx, Alessandro Gasparini, Marc-Andre Desautels, Dominik Leutnant, MDPI, Taylor and Francis, Oğuzhan Öğreden, Dalton Hance, Daniel Nüst, Petter Uvesten, Elio Campitelli, John Muschelli, Zhian N. Kamvar, Noam Ross, Robrecht Cannoodt, Duncan Luguern and David M. Kaplan (2020). rticles: Article Formats for R Markdown. R package version 0.14. `https://CRAN.R-project.org/package=rticles`

5. APPENDIX

All the code that were used for this assignment has been included in the APPENDIX.

5.1 Load The Required Libraries

```
library(tidyverse)
library(kableExtra)
```

5.2 Data Processing

Minor data processing was conducted to set the data table in an appropriate format for the needs of this assignment.

```
tooth_growth <- bind_rows(
  ToothGrowth %>%
    transmute(
      "factor" = "supp",
      "group_abbr" = as.character(supp),
      "group" = str_replace_all(
        string = group_abbr,
        c("OJ" = "Orange Juice",
          "VC" = "Ascorbic Acid")),
      "length" = len
    ),
  ToothGrowth %>%
    transmute(
      "factor" = "dose",
      "group_abbr" = paste0(dose, "mg"),
      "group" = str_replace_all(
        string = group_abbr,
        c("0.5mg" = "0.5 mg/day",
          "1mg" = "1 mg/day",
          "2mg" = "2 mg/day")),
      "length" = len
    ),
  ToothGrowth %>%
    transmute(
      "factor" = "supp_and_dose",
      "group_abbr" = paste0(supp, ":", dose, "mg"),
      "group" = str_replace_all(
        string = group_abbr,
        c("OJ:0.5mg" = paste0(
          "0.5 mg/day of Vitamin C", "\n",
          "as Orange Juice"),
          "OJ:1mg" = paste0(
            "1 mg/day of Vitamin C", "\n",
            "as Orange Juice"),
          "OJ:2mg" = paste0(
            "2 mg/day of Vitamin C", "\n",
            "as Orange Juice"),
          "VC:0.5mg" = paste0(
            "0.5 mg/day of Vitamin C", "\n",
            "as Ascorbic Acid"),
          "VC:1mg" = paste0(
            "1 mg/day of Vitamin C", "\n",
            "as Ascorbic Acid"),
          "VC:2mg" = paste0(
            "2 mg/day of Vitamin C", "\n",
            "as Ascorbic Acid"))),
      "length" = len
    )
)
```

5.3 Exploratory Data Analysis

Descriptive statistics were examined through figures and tables in order to identify some useful aspects of the data that helped to form the Research Questions.

5.3.1 *Code for Figure 1*

The following code was used to create Figure 1:

```
figure_1 <- ggplot(
  filter(tooth_growth, factor == "supp"),
  aes(x = length, y = group)
) +
  geom_boxplot() +
  labs(
    title = "Figure 1",
    subtitle =
      "Tooth Length of Groups by Supply Method",
    x = "Tooth Length",
    y = "Supply Method"
  ) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

5.3.2 *Code for Table 1*

The following code was used to create Table 1:

```
table_1 <- kable(
  x = tooth_growth %>%
    filter(factor == "supp") %>%
    group_by(group) %>%
    summarise("n" = n(),
              "mean" = mean(length),
              "sd" = sd(length)
    ) %>%
    rename("Supply Method" = group),
  booktabs = TRUE,
  caption =
    "Statistics by Supply Method"
) %>%
  kable_styling(
    latex_options = c("striped", "HOLD_position")
  )
```

5.3.3 *Code for Figure 2*

The following code was used to create Figure 2:

```
figure_2 <- ggplot(
  filter(tooth_growth, factor == "dose"),
  aes(x = length, y = group)
) +
  geom_boxplot() +
  labs(
    title = "Figure 2",
    subtitle = "Groups by Dosage Levels",
    x = "Tooth Length",
    y = "Dose of Vitamin C"
  ) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

5.3.4 Code for Table 2

The following code was used to create Table 2:

```
table_2 <- kable(
  x = tooth_growth %>%
    filter(factor == "dose") %>%
    group_by(group) %>%
    summarise("n" = n(),
              "mean" = mean(length),
              "sd" = sd(length)
    ) %>%
    rename("Dose of Vitamin C" = group),
  booktabs = TRUE,
  caption = "Statistics by Dose"
) %>%
  kable_styling(
    latex_options = c("striped", "hold_position")
  )
```

5.3.5 Code for Figure 3

The following code was used to create Figure 3:

```
figure_3 <- ggplot(
  filter(tooth_growth, factor == "supp_and_dose"),
  aes(x = length, y = group)
) +
  geom_boxplot(show.legend = FALSE) +
  labs(
    title = "Figure 3",
    subtitle = paste0("Tooth Length of Groups ",
                      "by Supply Method and Dose"),
    x = "Tooth Length",
    y = "Interactions of \nSupply Method and Dose"
  ) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

5.3.6 Code for Table 3

The following code was used to create Table 3:

```
table_3 <- kable(
  x = tooth_growth %>%
    filter(factor == "supp_and_dose") %>%
    mutate(group = str_replace(group, "\n", " ")
    ) %>%
    group_by(group) %>%
    summarise("n" = n(),
              "mean" = mean(length),
              "sd" = sd(length)
    ) %>%
    rename(
      "Interactions of Supply Method and Dose" =
        group
    ),
  booktabs = TRUE,
  caption = "Statistics by Supply Method and Dose"
) %>%
  kable_styling(
    latex_options = c("striped", "HOLD_position",
                      "scale_down")
  )
```

5.4 Statistical Analysis

5.4.1 Multiple t tests

A list was created with pairs of abbreviated labels, for the groups that should be compared.

```
group_comparisons <- list(
  "RQ-01" = c("OJ","VC"),
  "RQ-02" = c("2mg","0.5mg"),
  "RQ-03" = c("2mg","1mg"),
  "RQ-04" = c("1mg","0.5mg"),
  "RQ-05" = c("OJ:2mg","VC:2mg"),
  "RQ-06" = c("OJ:1mg","VC:1mg"),
  "RQ-07" = c("OJ:0.5mg","VC:0.5mg"),
  "RQ-08" = c("OJ:2mg","OJ:0.5mg"),
  "RQ-09" = c("OJ:2mg","OJ:1mg"),
  "RQ-10" = c("OJ:1mg","OJ:0.5mg"),
  "RQ-11" = c("VC:2mg","VC:0.5mg"),
  "RQ-12" = c("VC:2mg","VC:1mg"),
  "RQ-13" = c("VC:1mg","VC:0.5mg")
)
```

Based on the list with the group comparisons the 13 Welch t-test were conducted. Notice that for the hypothesis test of (RQ-05) the alternative hypothesis is two sided.

```
multiple_t_tests <- Map(
  f = function(groups, Ha) {
    t.test(
      x = tooth_growth %>%
        filter(group_abbr == groups[[1]]) %>%
        select(length) %>%
        `[[`(1),
      y = tooth_growth %>%
        filter(group_abbr == groups[[2]]) %>%
        select(length) %>%
        `[[`(1),
      alternative = Ha
    )
  },
  "groups" = group_comparisons,
  "Ha" = c(
    rep("greater", 4),
    # Only for the hypothesis test of (RQ-05)
    # a 'two.sided' test was conducted.
    "two.sided",
    rep("greater", 8)
  )
)
```

5.4.2 Adjust p-values

The following code was used to adjust the original p-values obtained for the multiple tests:

```
adjusted_p_values <- p.adjust(
  p = map_dbl(.x = multiple_t_tests,
              .f = ~.x[["p.value"]]),
  method = "fdr"
)
```

5.4.3 *Results*

From the list with the results of the multiple Welch t-tests, a tibble was created with the values of the adjusted p-values for each of 13 comparisons.

```
results <- tibble(
  "RQ" = str_pad(1:13, width = 2, pad = "0"),
  "x" = map_chr(group_comparisons, ~.x[[1]]),
  "y" = map_chr(group_comparisons, ~.x[[2]]),
  "p_adj" = adjusted_p_values,
  "is_sig" = ifelse(p_adj < 0.05, "Yes", "No")
)
```

Code for Table 4

The following code was used to create Table 4:

```
table_4 <- kable(
  x = results,
  caption = "Results",
  booktabs = TRUE,
) %>%
  kable_styling(
    latex_options = c("striped", "HOLD_position")
  )
```