Statistical Inference Course Project - Part 2: Inferential Data Analysis

Sebastian Jojoa

Overview

The following report is part of the final project for the Statistical Inference course, given by John Hopkins Bloomberg School of Public Health through Coursera. The report presents a short analysis of the ToothGrowth dataset with the objective of comparing the effect of different treatments in tooth length. Information about the dataset can be optained in R, after loading the `datasets` library, by typing `?ToothGrowth` in the command line.

Dataset

The `ToothGrowth` dataset contains the length of teeth (variable "len" in dataset) in each guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg; variable "dose") with each of two delivery methods (orange juice or ascorbic acid; variable "supp").

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  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  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

We can see there are a total of 60 measurements in the dataset. Let's see how they are divided:

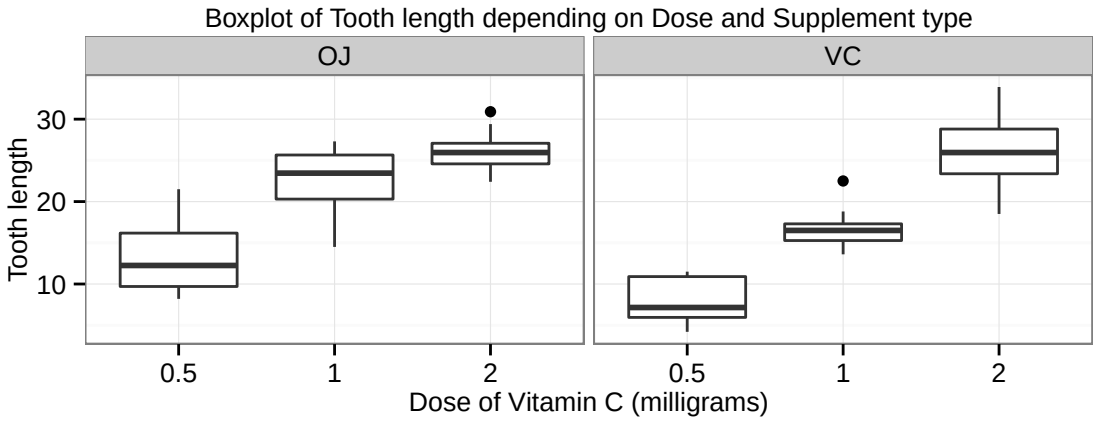
Table 1: Sample size per treatment combination. Dose levels: 0.5, 1 or 2 milligrams. Supplement type: OJ = Orange juice, VC = Ascorbic acid

	OJ	VC
0.5	10	10
1	10	10
2	10	10

The data is divided in 10 samples per combination of dose level of Vitamin C and supplement type. So there are a total of 20 measurements for each dose level and a a total of 30 for each supplement type:

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

This last summary also gives as a slight glimpse of how the teeth lenght data is distributed, but let's look at it more closely based on the combination of treatments each guinea pig was submitted to:

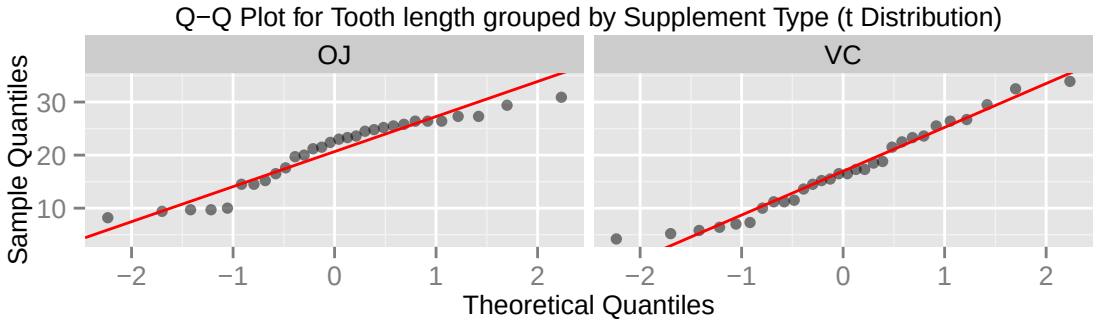


From this plot, we can see that it appears to be an effect of the does of Vitamin C in tooth length, but whether there is an effect of supplement type is not completely clear. At least it would appear to be sp for low dose levels (0.5 and 1 mg) but for high levels (2 mg) it appears to be the same.

Hypothesis testing

Now that we have made some hypothesis is time to test them. We want to know if the dose of Vitamin C and the supplement type of this vitamin affects the teeth lenght. In order to decide which kind of test we should use, let's remeber that in the assignment we were specifically ask to only "use the techniques from class, even if there's other approaches worth considering". We know that teeth length is a continuos variable, so we cannot use binomial or Poisson distribution because these are for descrete variables. This left us with either normal or Student's t distribution, and because we have a rather small sample size (30 at best when comparing supplement types), it is better to use the later.

Now for the hypothesis testing process, we will first check if the data follows a t distribution using a quantile-quantile plot, and if it does we will proceed to perform a t test, assuming unpaired samples and unequal variances, and calculate 95% confidence intervals.



The Q-Q plot for supplement type shows that although the distribution has deviations in the tails, the center tends to be close to the t distribution. However, since the distribution is close enough to the t distribution, we will proceed with the t test.

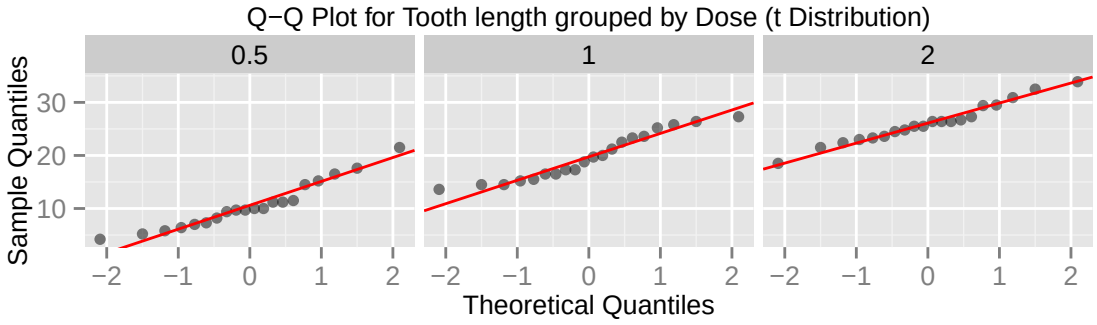
Table 2: Student's t test result for effect of Supplement type in Tooth length

P value	Confidence interval lower limit	Confidence interval upper limit
0.06	-0.17	7.57

In table 2 we can see that the p-value obtained is greater than 0.05, which means that we cannot say that there is an effect of supplement type in teeth length with a 95% confidence level (failed to reject the null).

This is further assured when we look at the limits for the 95% confidence interval for the difference in means between the two groups (Orange juice vs Ascorbic acid), because the interval includes 0, hence there is not sufficiently strong evidence to state that the supplement type affects teeth growth (see Boxplot of Tooth length depending Supplement type in Appendix materials).

To evaluate if dose level of Vitamin C has an effect in teeth length we must first check if the distribution of the data follows a t distribtuion.



The Q-Q plot shows that the data closely follows a t distribution. In order to do the hypothesis testing, since the t test only compares two groups at a time and we have 3 different dose levels, we will perform 3 separate t tests.

Table 3: Student's t test results for effect of Dose levels in tooth length

	P value	Confidence interval lower limit	Confidence interval upper limit
0.5 vs 1	0.00	-11.98	-6.28
0.5 vs 2	0.00	-18.16	-12.83
1 vs 2	0.00	-9.00	-3.73

The p values for all the test was smaller than 0.05, which means that, at a 95% confidence level, there is an effect of Vitamin C dose in teeth lenght (see Boxplot of Tooth length depending on Dose in Appendix materials). Taking into account our previous boxplot and the confidence interval for the difference in means obtained (negative values mean that the second mean was higher than the first one), we can see that the higher the dose was, the greater the tooth length.

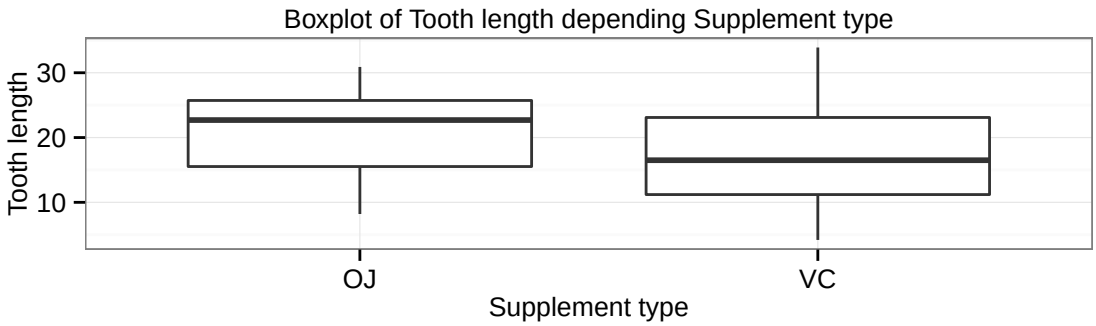
Conclusion

In conclusion, we can state that based on our analyses, the teeth length in guinea pigs is affected by the dose of Vitamin C they receive, presenting longer teeth with higher doses. On the other hand, whether the supplement method for this vitamin is orange juice or ascorbic acid does not play a significant effect in tooth growth.

To reach this conclusion, we made some assumpitons. First, we assumed that we were dealing with unpaired samples and unequal variances. They were unpaired samples because all measurements were takien from different individuals, and in tables 4 and 5 (Appendix materials) we can see that the varainces are in fact different. The only case the different variance assumption does not hold is for the test between 0.5 and 1 mg dose levels, but in this case, the t test was also perform for equal variances (not shown in report) and the results were the same. We also assumed that there is no interaction between dose level and supplement type, but this might not be true, if this interaction were to exist, it could possibly explain the deviations of the data from the t distribution. For this reason we might want to make a more complete analysis in the future, one in which we can discriminate the effects of each level of each treatment and their interaction (see Q-Q plot in Appendix materials).

Appendix materials

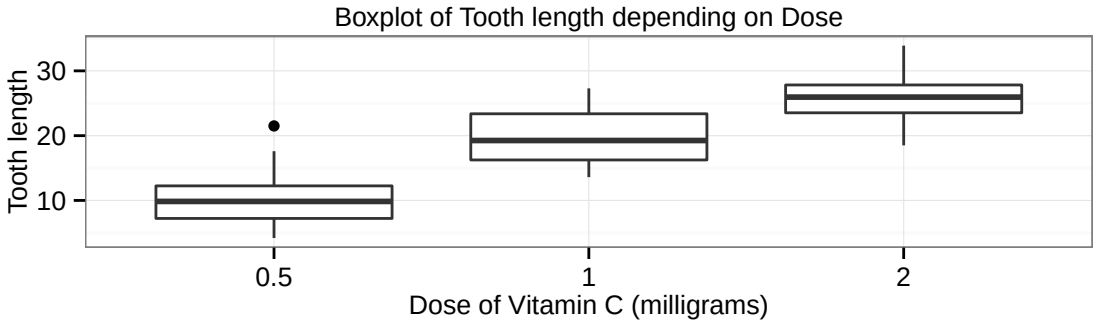
Supplementary plots and tables



The boxplot shows the distribution of the data grouped by supplement type and table 4 confirms the difference in variances for this test.

Table 4: Summary values for Tooth length grouped by Supplement type

Supplement type	Average	Standard deviation	Variance
OJ	20.66	6.61	43.63
VC	16.96	8.27	68.33



The boxplot shows the distribution of the data grouped by supplement type and table 5 confirms the difference in variances for most of the tests. For the t test of dose levels 0.5 vs 1, where the variances are almost equal, assuming equal or unequal variances did not affected the outcome of the test significantly (results not shown).

Table 5: Summary values for Tooth length grouped by Dose

Dose	Average	Standard deviation	Variance
0.5	10.61	4.50	20.25
1	19.73	4.42	19.50
2	26.10	3.77	14.24

This Q-Q plot shows that separating the data based on the two treatments each subject received renders a closer behavior to the t distribution. This is further reason for considering doing a more complex test that treats data this way.

Code chunks used to generate report

```
# Loading packages
{r loading_packages_and_setup, echo = FALSE, warning = FALSE,results = 'hold'}
for (package in c("ggplot2", "dplyr", "xtable")) {
    if (!(require(package, character.only = TRUE, quietly = TRUE, warn.conflicts = FALSE))) {
        install.packages(package)
        library(package, character.only = TRUE, quietly = TRUE, warn.conflicts = FALSE)
    }
}
options(xtable.comment = FALSE)

# Loading data and inital summary
{r loading_data, echo = FALSE}
library(datasets)
data(ToothGrowth)
str(ToothGrowth)

# Table 1
{r xtable_treatments, results='asis', echo = FALSE}
print(xtable(table(ToothGrowth$dose, ToothGrowth$supp),
             caption = "Sample size per treatment combination.
             Dose levels: 0.5, 1 or 2 milligrams.
             Supplement type: OJ = Orange juice, VC = Ascorbic acid"),
      caption.placement="top")

# Converts dose to factor and summary of data
{r summary, echo = FALSE}
ToothGrowth$dose <- as.factor(as.character(ToothGrowth$dose))
summary(ToothGrowth)

# Create boxplot
```

```
{r boxplot, echo = FALSE, fig.height = 2.5, fig.width = 6, fig.align = 'center'}
ggplot(data = ToothGrowth, aes(x = dose, y = len)) +
    geom_boxplot(aes(group = dose)) +
    facet_grid(. ~ supp) + theme_bw() +
    xlab("Dose of Vitamin C (milligrams)") +
    ylab("Tooth length") +
    ggtitle("Boxplot of Tooth length depending on Dose and Supplement type") +
    theme(plot.title = element_text(size=10), axis.title = element_text(size=10))

# First Q-Q plot and t test for supplement type
{r t_test_supp, echo = FALSE, fig.height = 2, fig.width = 6, fig.align = 'center'}
linesSupp <- as.data.frame(ToothGrowth %>%
                               group_by(supp) %>%
                               summarize(average = mean(len),
                                         standDev = sd(len),
                                         variance = var(len)))

ggplot(ToothGrowth, aes(sample = len)) +
    stat_qq(distribution = qt, dparams = list(df = 29), alpha = 0.5) +
    labs(title = "Q-Q Plot for Tooth length grouped by Supplement Type (t Distribution)",
         y = "Sample Quantiles",
         x = "Theoretical Quantiles") +
    facet_grid(. ~ supp) +
    geom_abline(data = linesSupp,
                aes(intercept = average, slope = standDev),
                color="red") +
    theme(plot.title = element_text(size=10),
          axis.title = element_text(size=10))

test <- t.test(len ~ supp, data = ToothGrowth, var.equal = FALSE)

extractStat <- function(test) {
    val <- test[[3]]
    val <- c(val, test[[4]][1])
    val <- c(val, test[[4]][2])
    return(val)
}

OJvsVC <- extractStat(test)
data <- data.frame(rbind(OJvsVC))
names(data) <- c("P value", "Confidence interval lower limit", "Confidence interval upper limit")

# Table 2
{r table_supp_test, results='asis', echo = FALSE}
print(xtable(data, "Student's t test result for effect of Supplement type in Tooth length"),
      include.rownames = FALSE, caption.placement="top")

# Second Q-Q plot and t tests for dose
{r t_test_dose, echo = FALSE, fig.height = 2, fig.width = 6, fig.align = 'center'}
linesDose <- as.data.frame(ToothGrowth %>%
                               group_by(dose) %>%
                               summarize(average = mean(len),
                                         standDev = sd(len),
```

```
                                        variance = var(len)))

ggplot(ToothGrowth, aes(sample = len)) +
    stat_qq(distribution = qt, dparams = list(df = 19), alpha = 0.5) +
    labs(title = "Q-Q Plot for Tooth length grouped by Dose (t Distribution)",
         y = "Sample Quantiles",
         x = "Theoretical Quantiles") +
    facet_grid(. ~ dose) +
    geom_abline(data = linesDose,
                aes(intercept = average, slope = standDev),
                color="red") +
    theme(plot.title = element_text(size=10),
          axis.title = element_text(size=10))

test05_1 <- with(ToothGrowth, t.test(len[dose == 0.5], len[dose == 1], var.equal = FALSE))
test05_2 <- with(ToothGrowth, t.test(len[dose == 0.5], len[dose == 2], var.equal = FALSE))
test1_2 <- with(ToothGrowth, t.test(len[dose == 1], len[dose == 2], var.equal = FALSE))

stat05_1 <- extractStat(test05_1)
stat05_2 <- extractStat(test05_2)
stat1_2 <- extractStat(test1_2)

stats <- data.frame(rbind(stat05_1, stat05_2, stat1_2),
                    row.names = c("0.5 vs 1", "0.5 vs 2", "1 vs 2"))

names(stats) <- c("P value", "Confidence interval lower limit", "Confidence interval upper limit")

# Table 3
{r table_dose_test, results='asis', echo = FALSE}
print(xtable(stats,
             "Student's t test results for effect of Dose levels in tooth length"),
      caption.placement="top")

# First supplementary plot
{r additional_plots1, echo = FALSE, fig.height = 2, fig.width = 6, fig.align = 'center'}
ggplot(data = ToothGrowth, aes(x = supp, y = len)) +
    geom_boxplot() +
    theme_bw() +
    xlab("Supplement type") +
    ylab("Tooth length") +
    ggtitle("Boxplot of Tooth length depending Supplement type") +
    theme(plot.title = element_text(size=10),
          axis.title = element_text(size=10))

# Table 4
{r variancesSupp, results = 'asis', echo = FALSE}
names(linesSupp) <- c("Supplement type", "Average", "Standard deviation", "Variance")
print(xtable(linesSupp, "Summary values for Tooth length grouped by Supplement type"),
      include.rownames = FALSE, caption.placement="top")

# Second supplementary plot
{r additional_plots2, echo = FALSE, fig.height = 2, fig.width = 6, fig.align = 'center'}
ggplot(data = ToothGrowth, aes(x = dose, y = len)) +
```

```
    geom_boxplot() +
    theme_bw() +
    xlab("Dose of Vitamin C (milligrams)") +
    ylab("Tooth length") +
    ggtitle("Boxplot of Tooth length depending on Dose") +
    theme(plot.title = element_text(size=10),
          axis.title = element_text(size=10))

# Table 5
names(linesDose) <- c("Dose", "Average", "Standard deviation", "Variance")
print(xtable(linesDose, "Summary values for Tooth length grouped by Dose"),
      include.rownames = FALSE, caption.placement="top")

# Third supplementary plot
{r additional_plots3, echo = FALSE, fig.height = 3, fig.width = 6}
# Divided by treatment - t distribution
lines <- ToothGrowth %>%
    group_by(dose, supp) %>%
    summarize(average = mean(len), standDev = sd(len))

ggplot(ToothGrowth, aes(sample = len)) +
    stat_qq(distribution = qt, dparams = list(df = 9), alpha = 0.5) +
    labs(title = "Q-Q Plot for Tooth lenght (t Distribution)",
         y = "Sample Quantiles",
         x = "Theoretical Quantiles") +
    facet_grid(supp ~ dose) +
    geom_abline(data = lines,
                aes(intercept = average, slope = standDev),
                color="red") +
    theme(plot.title = element_text(size=10),
          axis.title = element_text(size=10))
```