Basic inferential data analysis of ToothGrowth dataset

Thomas Fischer

May 30, 2018

Abstract

This document provides the assignment 'Course Project Part 2' for Coursera's Statistical Inference Class in the Coursera Data Science series. Replication files are available on the author's Github account (https://github.com/tomfischersz).

1. Synopsis

In this report we aim to conduct some basic inferential data analysis on the ToothGrowth dataset of the R library 'datasets'. We aim to answer the question, if dosage and/or delivery method of vitamin C affects tooth growth in guinea pigs. We therefore observe patterns from the data, formulate hypotheses and then use statistical tests like confident intervals or student's t-test to validate these hypotheses.

2. The ToothGrowth Data Set

The data consists of 60 observations with 3 variables, here the first few observations:

Table 1: The first few observations of the data set ToothGrowth

len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5

The help page[1] for the data set ToothGrowth gives following description:

> 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 (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC).

Our data are results from a study performed on guinea pigs to determine the effect of vitamin C on tooth growth. The data contains 3 variables:

- **len:** The response (dependent) variable for the experiment measured for 60 guinea pigs is the tooth length.
- **supp** and **dose** Two factors (independent variables), the delivery method of the vitamin C (supplement type) and the dose levels of vitamin C in mg/day. We are interested in the effect of these two factors on the response.

Table 3 depicts a aggregated summary of our data. We can see that there are 6 factor-level combinations and each of these 6 combinations were applied to 10 guinea pigs each. We hereafter call this different combinations just treatment (and also added a new column), e.g. "OJ_0.5" just denotes the treatment with the factors 'Orange Juice' with a dose level of 0.5 mg/day.

[1] Use R command help(ToothGrowth) to get further information.

3.Exploratory Data Analysis

We now visualize the means and spread of tooth growth for our six distinct treatment groups (Code):

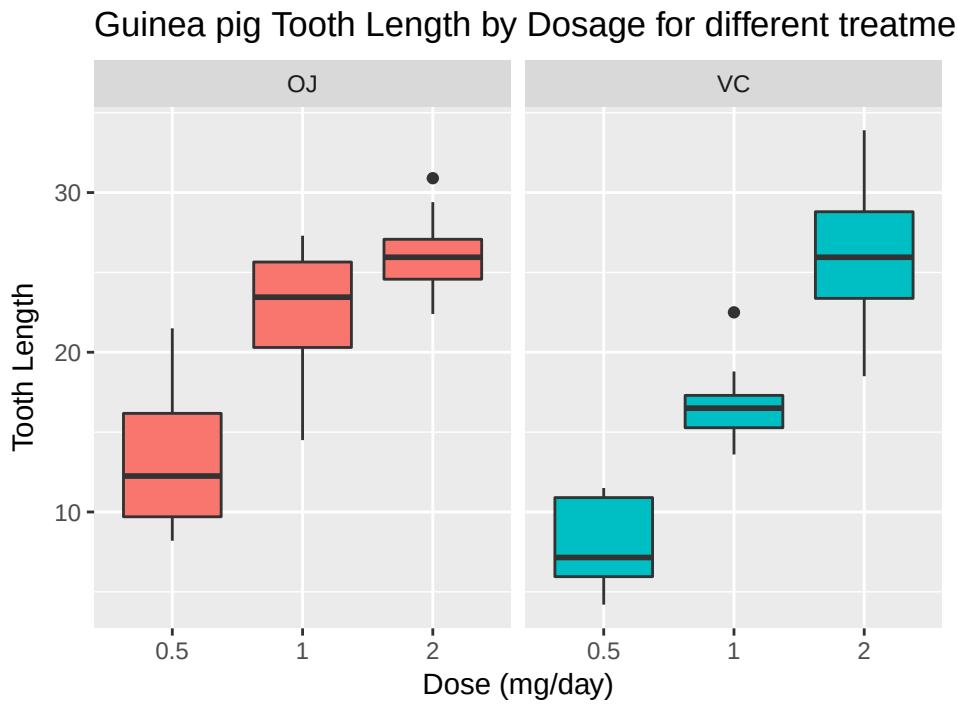

Figure 1: Comparing the possible effects of three varying doses of vitamin C for the two different supplement types (Orange Juice and Vitamin C).

Figure 1 suggests that the dose and the delivery method both have some effect on the tooth growth. It appears that the average tooth growth increases with the dose levels and that orange juice might have higher growth rates than Vitamin C except for dose levels of 2 mg.

4. Basic Inference Analysis (hypothesis tests)

We are now testing several hypotheses. Our significance level (i.e. the risk of getting a Type I error) for all tests will be $\alpha = 0.05$. We strictly only use student t-tests as required in the assignment (disregarding regression analysis and anova test).

4.1 Assumptions

Before proceeding in our analysis it is important to assure certain assumptions necessary to apply student's t-test, so we must be sure that following assumptions are not violated:

- Independent and identically distributed: We are assuming that the process of choosing 60 guinea pigs for the experiment was independed and that they are drawn from the same population. Otherwise our results would be not reliable, e.g. if the guinea pigs origin from two different breeders, or there are differences in male and female populations our conclusions could be flawed.
- The probability distributions of the measured tooth length for each treatment are normal. Depicting Figure 2 it seems that this assumption appears to be reasonably satisfied.

4.2 Hypothesis Test I

We want to test the null hypothesis that the mean tooth length for the two delivery methods are equal against the alternative hypothesis that they differ:

$H_0 : \mu_{OJ} = \mu_{VC}$
$H_a : \mu_{OJ} \neq \mu_{VC}$

Stated the relevant null and alternative hypotheses, we then conduct a two-tailed t-test (Code):

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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
```

As the obtained p-value of 0.061 is greater than the significance level of 0.05 (and the confidence interval at 95% contains 0) we cannot reject our null hypothesis. Looking at figure 1 again, failing to reject the null hypothesis is likely due to the similar results in tooth length for a vitamin C dose of 2 mg/day.

4.2 Hypothesis Test II

Our next hypothesis test will be examining if, for orange juice only, higher doses of vitamin C are significantly associated with higher tooth length. We are conducting two one-tailed t-tests and therefore need to adjust our confidence intervals. We adjust the original confidence level of our tests of 95% using Bonferroni correction to $1 - \frac{\alpha}{m} = 0.975$, where m is the number of hypotheses. Our new significance level is $\alpha = 0.025$.

$H_0 : \mu_{OJ\_0.5} = \mu_{OJ\_1} = \mu_{OJ\_2}$
$H_a : \mu_{OJ\_0.5} <= \mu_{OJ\_1} <= \mu_{OJ\_2}$

Conducted the relevant t-test (Code) we get following results:

Table 2: Summary of t-tests for different levels of doses (Orange Juice)

Sample Groups	p-values	Lower Conf.Interval	Upper Conf.Interval
OJ_0.5 versus OJ_1	0.00	-Inf	-5.52
OJ_0.5 versus OJ_1	0.02	-Inf	-0.19

As we can see, both p-values are below our significance level $\alpha = 0.025$ and both confidence intervals for the difference of means for the treatments are below zeros. We therefore can conclude to reject the null hypothesis, i.e. for orange juice we examine different effects depending on the dose of vitamin C.

5. Conclusion

- No evidence for the hypothesis that tooth length differs for different delivery methods.
- Strong evidence that tooth length varies for different doses given the delivery method orange juice.

Appendix I: Figures and Tables

Table 3: Summary of the different treatments for the guinea pigs with their associated average tooth length and the corresponding standard deviation

Supplement	Dose (mg/day)	Treatment	N (number of pigs)	Mean	Standard Deviation
OJ	0.5	OJ_0.5	10	13.23	4.46
OJ	1.0	OJ_1	10	22.70	3.91
OJ	2.0	OJ_2	10	26.06	2.66
VC	0.5	VC_0.5	10	7.98	2.75
VC	1.0	VC_1	10	16.77	2.52
VC	2.0	VC_2	10	26.14	4.80

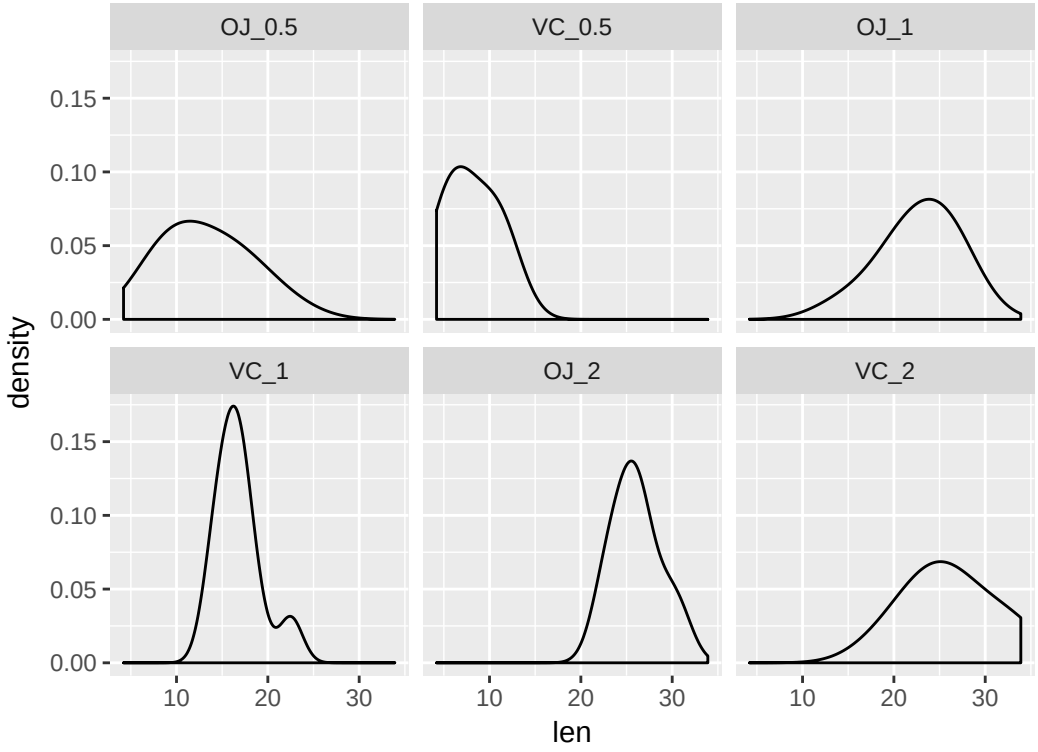

Figure 2: Density distributions for all treatment groups.

Appendix II: R Source Code

1. Load required libraries:

```
require(knitr)
require(kableExtra)
require(datasets)
require(ggplot2)
require(dplyr)
```

2. Load data:

```
data(ToothGrowth)
# names(ToothGrowth) <- c('length', 'supplement', 'dose')
```

3. Add new variable treatment:

```
ToothGrowth$treatment=with(ToothGrowth,interaction(supp,dose, sep = '_'))
```

4. First few observations:

```
kable(head(ToothGrowth[, 1:3], n=3),
      format = 'latex',
      booktabs = TRUE,
      caption = "The first few observations of the data set
      ToothGrowth\\label{tab:show_obs}")  %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

5. Aggregating data in data.frame:

```
df_summary <-
    ToothGrowth %>%
    group_by(supp, dose, treatment) %>%
    summarise(N = n(),
              mean_len = mean(len),
              sd_len = sd(len)) %>%
    as.data.frame()
```

6. Boxplots for different treatments:

```
fig_1 <- ggplot(ToothGrowth, aes(x=factor(dose), y=len)) +
    facet_grid(.~supp) +
    geom_boxplot(aes(fill = supp), show.legend = FALSE) +
    labs(title = "Guinea pig Tooth Length by Dosage for different treatments",
         x = "Dose (mg/day)",
         y = "Tooth Length")
```

7. Distribution of Tooth Length for different treatments:

```
fig_2 <- ggplot(ToothGrowth, aes(x = len)) +
    geom_density(adjust = 1.5) +
    facet_wrap(~ treatment)
```

8. Hypothesis Test I

```
t_01 <- t.test(len~supp,data=ToothGrowth, paired = FALSE, var.equal = FALSE, alternative = 'two.sided')
```

9. Hypothesis Test II

```
t_02_1 <-
    t.test(len~dose,
           data = ToothGrowth[ToothGrowth$treatment %in% c('OJ_0.5', 'OJ_1'),],
           paired = FALSE, var.equal = FALSE,
           alternative = 'less', conf.level = 0.975)
t_02_2 <-
    t.test(len~dose,
           data = ToothGrowth[ToothGrowth$treatment %in% c('OJ_1', 'OJ_2'),],
           paired = FALSE, var.equal = FALSE,
           alternative = 'less', conf.level = 0.975)
sum_ttests <-
    data.frame(sample_group = c('OJ_0.5 versus OJ_1', 'OJ_0.5 versus OJ_1'),
               p_value = c(round(t_02_1$p.value,4), round(t_02_2$p.value,4)),
               confint_lower = c(t_02_1$conf.int[[1]], t_02_2$conf.int[[1]]),
               confint_upper = c(t_02_1$conf.int[[2]], t_02_2$conf.int[[2]]))
```