Basic Inferential Data Analysis

Fritz Lin

1 Overview

In this project we are going to analyze the ToothGrowth data in the R `datasets` package. We will investigate the data using basic exploratory data analysis as well as basic inferential analysis.

2 Basic Exploratory Data Analysis

The data of interest is the ToothGrowth data, which reports the effect of vitamin C on tooth growth in guinea pigs. Accoring to the documentation "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))."
We will now have a look at the data.

```
# Initial inspection of data
str(ToothGrowth, give.attr = FALSE)
```

```
## '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 ...
```

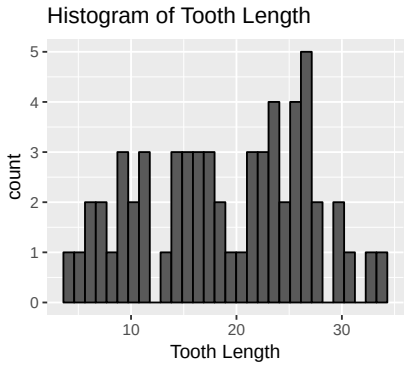
```
# Convert "dose" to factor rather than numeric
ToothGrowth$dose <- factor(ToothGrowth$dose)
table(ToothGrowth$supp, ToothGrowth$dose)
```

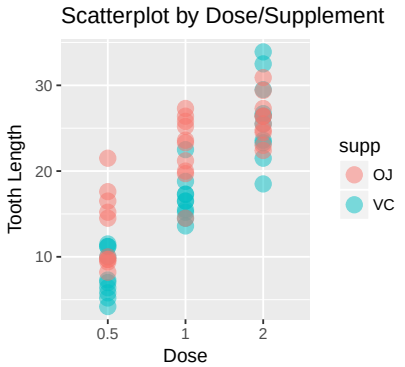
```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

As can be seen, the data includes 60 observations on 3 variables:

- **len**: a numeric variable measuring the *Tooth length*
- **supp**: a factor variable indicating the *Supplement type*, either via **VC** (vitamin C) or **OJ** (orange juice)
- **dose**: a numeric variable indicating the *Dose* in milligrams/day, either **0.5mg**, **1.0mg** or **2.0mg**

Furthermore, it can be seen that each group (grouped by supplement type and dose) consists of 10 subjects.





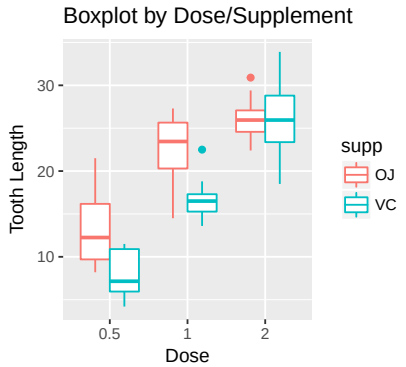


Table 1: Summary statistics Overall

n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
60	18.813	7.649	19.25	18.946	9.044	4.2	33.9	29.7	-0.143	-1.043	0.988

We will now examine whether the data is normally distributed, a required assumption for hypothesis testing with t-tests, by first drawing a histogram of the tooth lengths and providing a summary of the statistics. The histogram does not give a clear picture of whether the data is normally distributed. If we look at the statistics in Table 1, the median is slightly higher than the mean, suggesting it is slightly left-skewed. The higher standard deviation in the data set would also suggest that the distribution is platykurtic. This is all confirmed by the negative values of skew and kurtosis.

The Shapiro-Wilk test for normality (H_0: data is normally distributed) will give us more certainty.

```
# test for normality
shapiro.test(ToothGrowth$len)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ToothGrowth$len
## W = 0.96743, p-value = 0.1091
```

Given a p-value of 0.1091 we fail to reject the null at the 5% significance level, so we assume the data is normally distributed.

With this, we now look further into the data by breaking it down by dose and supplement. If we look at the scatterplot and boxplot above, we make the following observations:

- **OJ** has a greater effect on tooth growth with dosis of 0.5mg or 1.0mg, but not 2.0mg than **VJ**
- the amount of dosis has a greater effect on tooth growth via **VC**
- an amount of 1.0mg has a greater effect on tooth growth than 0.5mg via **OJ**
- an amount of 2.0mg or 1.0mg via **OJ** has a similar effect, although the data for the latter appears to inhibit more variability.

Next, we have a look at some summary statistics of the groups, and see if we can confirm our initial observations. More graphs can be found in Appendix B

Table 2: Summary Statistics Broken

supp	dose	mean	median	sd
OJ	0.5	13.230	12.25	4.460
OJ	1	22.700	23.45	3.911
OJ	2	26.060	25.95	2.655
VC	0.5	7.980	7.15	2.747
VC	1	16.770	16.50	2.515
VC	2	26.140	25.95	4.798
OJ	-	20.663	22.70	6.606
VC	-	16.963	16.50	8.266
-	0.5	10.605	9.85	4.500
-	1	19.735	19.25	4.415
-	2	26.100	25.95	3.774

3 Basic Summary of the Data

As can be seen from Table 2, the statistics confirm mostly of our initial observations from the graphs. **OJ** has a greater impact than **VC**, except if the dose is 2.0mg. Although the mean for both groups are very close (26.06 vs 26.14), **VC** does inhibit more variability.
Next, the statistics also confirm that with larger dosis, the effect on tooth growth will be larger. However, we note that the difference between an administration of 2.0mg and 1.0mg dose via **OJ** is rather small (26.06 vs 22.70), although the latter inhibits more variability (2.66 vs 3.91).
If we take a look at the standard deviations between **OJ** and **VC**, we note for the 0.5mg and 1.0mg dose that **OJ** inhibits more variability than **VC**, while the opposite is true for the

2.0mg group, as mentioned already above.
We also see that in general with higher dose, the tooth length will be longer, and that **OJ** has greater effect than **VC** overall, as depicted in the last five rows of the table.

In this regard, we would like to conduct several hypothesis tests to see if there is a statistically significant difference between the groups.

4 Hypothesis Testing

We will perform several two-group intervals testing at the 95% confidence interval. In particular, hypothesis tests of interests with their outcome are listed in Table 3. Since we perform seven tests in total, we need to control our error rates by adjusting for p-values.

Table 3: Results of Hypothesis Tests

Alternative	t statistic	p-value	p-adjusted	lower CI	higher CI
OJ > VC	1.915	0.030	0.035	0.471	Inf
1.0mg > 0.5mg	6.477	0.000	0.000	6.753	Inf
2.0mg > 1.0mg	4.900	0.000	0.000	4.175	Inf
OJ 0.5mg > VC 0.5mg	3.170	0.003	0.005	2.378	Inf
OJ 1.0mg > VC 1.0mg	4.033	0.000	0.001	3.380	Inf
OJ 2.0mg != VC 2.0mg	-0.046	0.964	0.964	-3.798	3.638
OJ 2.0mg > OJ 1.0mg	2.248	0.020	0.027	0.749	Inf

At the 5% significance level we would reject all null hypotheses except for the 6th test, that is H_a: OJ 2.0mg $\neq$ VC 2.0mg. The result of our tests suggests that **OJ** is more effective than **VC** at lower to mid dosis, but fairly equivalent at a higher amount.

5 Conclusion

We have analyzed the TootGrowth data in the R `datasets` package using basic exploratory data analysis as well as basic inferential analysis. With regard to the latter we have performed seven hypothesis tests with the following assumptions:

- Given a negative Shapiro-Wilk test, we assume the data is normally distributed
- H_0: there is no difference in the means of the two groups compared
- Equal variance of the two groups compared, since it is assumed that we deal with guinea pigs that should come from a same population
- Rather than performing a paired test, we assume that the groups are independent, since it is assumed that each guinea pig received an independent treatment

After performing our analysis we come to the conclusion that:

- Higher dosis of Vitamin C, either in the form of orange juice or ascorbic acid, will lead to a higher tooth length
- Orange juice has a bigger effect than ascorbic acid in lower to mid dosis
- There is no significant difference between orange juice or ascorbic acid, when prescribed in higher dosis

For the sake of reproducibility all code to perform our analysis can be found in Appendix A.

6 Appendices

6.1 Appendix A

```
library(datasets)
# Load data and provide basic info on data
data(ToothGrowth)
```

```
# Initial inspection of data
str(ToothGrowth, give.attr = FALSE)
# Convert "dose" to factor rather than numeric
ToothGrowth$dose <- factor(ToothGrowth$dose)
table(ToothGrowth$supp, ToothGrowth$dose)
```

```
library(psych)
library(knitr)
library(kableExtra)
# summary statistics (Table 1)
stats <- describe(ToothGrowth$len)
rownames(stats) <- NULL
stats <- within(stats, rm("vars"))
kable(stats, booktabs = T, caption = "Summary statistics Overall", digits = 3)
```

```
# test for normality
shapiro.test(ToothGrowth$len)
```

```
library(ggplot2)
library(grid)
library(gridExtra)
# plotting
g1 <- ggplot(ToothGrowth, aes(x = len)) +
  geom_histogram(color = "black") +
  labs(x = "Tooth Length", title = "Histogram of Tooth Length")
g2 <- ggplot(ToothGrowth, aes(dose, len)) +
  geom_point(aes(color = supp), size = 4, alpha = 0.5) +
  labs(x = "Dose", y = "Tooth Length", title = "Scatterplot by Dose/Supplement")
g3 <- ggplot(ToothGrowth, aes(dose, len)) + geom_boxplot(aes(color = supp)) +
  labs(x = "Dose", y = "Tooth Length", title = "Boxplot by Dose/Supplement")

grid.arrange(g1, g2, g3, ncol = 3)
```

```
library(dplyr)
# summary statistics by supplement and dose (Table 2)
s1 <- ToothGrowth.summary1 <- ToothGrowth %>% group_by(supp, dose) %>%
  summarize(mean = mean(len), median = median(len), sd = sd(len))

s2 <- ToothGrowth.summary2 <- ToothGrowth %>% group_by(supp) %>%
  summarize(mean = mean(len), median = median(len), sd = sd(len))

s3 <- ToothGrowth.summary3 <- ToothGrowth %>% group_by(dose) %>%
  summarize(mean = mean(len), median = median(len), sd = sd(len))

s <- full_join(full_join(s1,s2), s3)
s$supp <- as.character(s$supp)
```

```r
s$dose <- as.character(s$dose)
s[is.na(s)] <- "-"

kable(s, booktabs = T, caption = "Summary Statistics Broken",
      digits = 3, linesep = c("")) %>%
  kable_styling(latex_options = c("striped"), position = "float_right")
```

```r
# Reorder "dose"
ToothGrowth$dose <- factor(ToothGrowth$dose, levels=c("2", "1", "0.5"))

# Hypothesis tests
t.supp <- t.test(len ~ supp, paired=F, var.equal=T,
                 ToothGrowth, alternative = "greater")

t.10vs05 <- t.test(len ~ dose, paired=F, var.equal=T,
                   filter(ToothGrowth, dose %in% c(1, 0.5)),
                   alternative = "greater")

t.20vs10 <- t.test(len ~ dose, paired=F, var.equal=T,
                   filter(ToothGrowth, dose %in% c(2, 1)),
                   alternative = "greater")

t.OJ05vsVC05  <- t.test(len ~ supp, paired=F, var.equal=T,
                        filter(ToothGrowth, dose %in% c(0.5)),
                        alternative = "greater")

t.OJ10vsVC10  <- t.test(len ~ supp, paired=F, var.equal=T,
                        filter(ToothGrowth, dose %in% c(1)),
                        alternative = "greater")

t.OJ20vsVC20  <- t.test(len ~ supp, paired=F, var.equal=F,
                        filter(ToothGrowth, dose %in% c(2)))

t.OJ20vsOJ10  <- t.test(len ~ dose, paired=F, var.equal=F,
                        filter(ToothGrowth, supp %in% c("OJ") & dose %in% c(2, 1)),
                        alternative = "greater")

# Summary of the tests (Table 3)
t.statistic <- c(t.supp$statistic, t.10vs05$statistic, t.20vs10$statistic,
                 t.OJ05vsVC05$statistic, t.OJ10vsVC10$statistic,
                 t.OJ20vsVC20$statistic, t.OJ20vsOJ10$statistic)

t.p.value <- c(t.supp$p.value, t.10vs05$p.value, t.20vs10$p.value,
                 t.OJ05vsVC05$p.value, t.OJ10vsVC10$p.value,
                 t.OJ20vsVC20$p.value, t.OJ20vsOJ10$p.value)

t.p.adj <- p.adjust(t.p.value, method="BH")

t.lower <- c(t.supp$conf.int[1], t.10vs05$conf.int[1], t.20vs10$conf.int[1],
                 t.OJ05vsVC05$conf.int[1], t.OJ10vsVC10$conf.int[1],
                 t.OJ20vsVC20$conf.int[1], t.OJ20vsOJ10$conf.int[1])

t.upper <- c(t.supp$conf.int[2], t.10vs05$conf.int[2], t.20vs10$conf.int[2],
                 t.OJ05vsVC05$conf.int[2], t.OJ10vsVC10$conf.int[2],
```

```
                t.OJ20vsVC20$conf.int[2], t.OJ20vsOJ10$conf.int[2])

t.Ha <- c("OJ > VC", "1.0mg > 0.5mg", "2.0mg > 1.0mg",
          "OJ 0.5mg > VC 0.5mg", "OJ 1.0mg > VC 1.0mg",
          "OJ 2.0mg != VC 2.0mg", "OJ 2.0mg > OJ 1.0mg")

t.summary <- data.frame(t.Ha, t.statistic, t.p.value, t.p.adj, t.lower, t.upper)
names(t.summary) <- c("Alternative", "t statistic",
                      "p-value", "p-adjusted",
                      "lower CI", "higher CI")

kable(t.summary, booktabs = T, caption = "Results of Hypothesis Tests",
      digits = 3, linesep = c("")) %>%
  kable_styling(latex_options = c("hold_position", "striped"), position = "center")
```

6.2 Appendix B

```
# plotting
g3 <- ggplot(ToothGrowth, aes(supp, len)) + geom_boxplot(aes(color = supp)) +
  labs(x = "Supplement") +
  theme(axis.title.y=element_blank())

g4 <- ggplot(ToothGrowth, aes(dose, len)) + geom_boxplot(aes(color = dose)) +
  labs(x = "Dose") +
  theme(axis.title.y=element_blank())

grid.arrange(g3, g4, ncol = 2, left = textGrob("Tooth Length", rot = 90, vjust = 1))
```