SI - Part 2: Inferential Data Analysis

Pascal P

11 August 2018

1 Overview

The aim of this project is to do an analysis on the Tooth Growth data, part of the `R datasets` package.
In the following we will load the data, present some basic summary, then define some plots which will allow us to make some observations. We will proceed with some hypothesis tests, stating our assumptions, to compare tooth growth versus supplements and doses. Finally we will state our conclusions.

2 Exploratory Data

2.1 Load the data and basic summary

According to `help page`, the `ToothGrowth` data-set is a data frame of 60 observations on 3 variables:

- `len`: numeric, tooth length (a measure for the tooth growth)
- `supp`: factor, supplement type (VC [ascorbic acid, a form of vitamin C] or OJ [Orange Juice]).
- `dose`: numeric, dose in milligrams/day

```
str(ToothGrowth)
```

```
# '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 ...
```

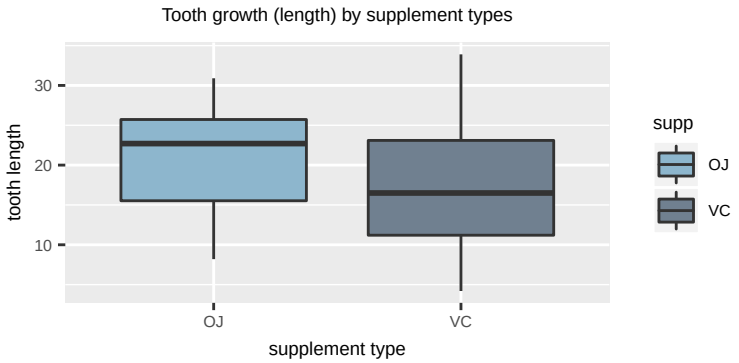
A basic summary is as follows:

Table 1: summary per supp and dose

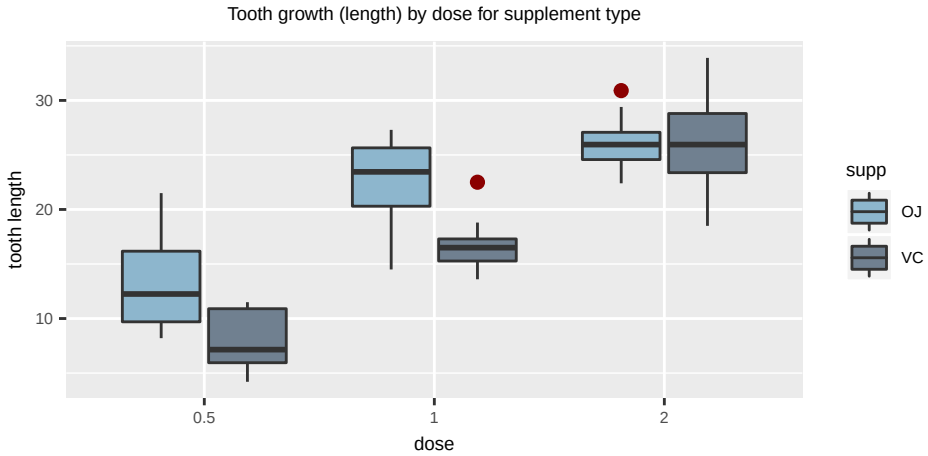
supp	dose	mean	std_dev
OJ	0.5	13.23	4.460
VC	0.5	7.98	2.747
OJ	1.0	22.70	3.911
VC	1.0	16.77	2.515
OJ	2.0	26.06	2.655
VC	2.0	26.14	4.798

2.2 Plots and observations

We will start our exploratory data analysis with some basic plots.



It appears from the plot above that supplement type OJ is more effective for tooth growth than VC.
Let see what is the effect of the dose, for this we need to convert the 3 different dose levels into a factor (additional column `dose_f`).



This tells us:

- that as we increase the dose (regardless of the supplement type) we have an increase in the tooth growth and
- for dose `0.5 mg/day` and `1 mg/day`, OJ appears to be more effective than VC in the increase of tooth growth
- at `2 mg/day` OJ and VC appears to have similar effects on tooth growth

3 Comparing tooth growth by supplement and dose

First, let me states my assumptions for the rest of this study:

1. independence of the groups, which means that I will use the `R options paired=FALSE`
2. unequal variance, and
3. data is iid (independent, identically distributed) normal.

Now I will test the alternative hypothesis that the mean length (tooth growth, guinea pigs) for OJ and VC differs significantly for dosage less than `2mg/day` (in this instance the null hypothesis is that there is no significant difference).

```
t00$p.value; t00$conf.int
```

```
# [1] 0.00423861
```

```
# [1] -9.304766 -1.875234
# attr(,"conf.level")
# [1] 0.95
```

We can see that the `p-value (= 0.004239)` $< \alpha (= 0.05)$ and therefore reject the null hypothesis (for full details, cf. appendix).

Next, I will test the alternative hypothesis that the mean length (of the guinea pigs) for OJ and VC differs significantly for dosage at `2mg/day`. In this case the expectation is that we should not reject the null hypothesis (cf. plot and statements in previous part, above).

```
t01$p.value; t01$conf.int
```

```
# [1] 0.9638516
```

```
# [1] -3.63807  3.79807
# attr(,"conf.level")
# [1] 0.95
```

Clearly this time (and as expected), we cannot reject the null hypothesis as `p-value (= 0.9639)` $> \alpha(= 0.05)$. Also note that the confidence interval contains 0.

Next we will proceed with 3 t-tests (number of ways to compare two given doses among three) for the alternative hypothesis that the mean length (tooth growth, guinea pigs) for different doses differs significantly.

```
knitr::kable(sum_df, caption='t-test summary')
```

Table 2: t-test summary

name	p.value	t.stat	df	low.conf.int	up.conf.int	mean.x	mean.y
dose_0_5mg and dose_1_0mg	1.00e-07	-6.477	38.0	-11.984	-6.276	10.605	19.735
dose_0_5mg and dose_2_0mg	0.00e+00	-11.799	36.9	-18.156	-12.834	10.605	26.100
dose_1_0mg and dose_2_0mg	1.91e-05	-4.900	37.1	-8.996	-3.734	19.735	26.100

In each case, we found that:

- the `p-value` is less than the given $\alpha(= 0.05)$ and
- `0` is not in the confidence interval.

This means that we can reject the null-hypothesis and retain the alternative which states that the mean in growth tooth differs significantly with different doses.

4 Conclusions

Based on our exploratory analysis and comparison for tooth growth presented above we can say that:

- for doses less than `2mg/day` the OJ supplement is more effective than VC fore tooth growth (for guinea pigs).
- at higher dosage (here `2.0mg`) there is no significant difference between the two supplements.
- increasing dosage (regardless of the supplement type) leads to an apparent increase in tooth growth (*I did not present any evidence beyond a plot for this though*).

5 Appendix

5.1 Code for basic summary table

```
# summary table
sum_df0 <- aggregate(ToothGrowth$len,
                     by=list(ToothGrowth$supp, ToothGrowth$dose), FUN=mean)
colnames(sum_df0) <- c("supp", "dose", "mean")
std_dev <- aggregate(ToothGrowth$len,
                     by=list(ToothGrowth$supp, ToothGrowth$dose),
                     FUN=function(x) { round(sd(x), 3) })$x

knitr::kable(cbind(sum_df0, std_dev), caption='summary per supp and dose')
```

5.2 Code for plots

```
library(ggplot2)

ggplot(ToothGrowth, aes(x=supp, y=len, fill=supp)) +
  geom_boxplot(outlier.colour="darkred", outlier.size=2) +
  scale_fill_manual(values=c("lightskyblue3", "slategrey")) +
  ggtitle("Tooth growth (length) by supplement types") +
  ylab("tooth length") + xlab("supplement type") +
  theme(plot.title=element_text(hjust = 0.5, size=7),
        text=element_text(size=7))
```

```
# create a new column for holding dose as a factor
ToothGrowth$dose_f <- as.factor(ToothGrowth$dose)

ggplot(ToothGrowth, aes(x=dose_f, y=len, fill=supp)) +
  geom_boxplot(outlier.colour="darkred", outlier.size=2) +
  scale_fill_manual(values=c("lightskyblue3", "slategrey")) +
  ggtitle("Tooth growth (length) by dose for supplement type") +
  ylab("tooth length") + xlab("dose") +
  theme(plot.title=element_text(hjust = 0.5, size=7),
        text=element_text(size=7))```
```

5.3 Code for t-tests

5.3.1

```
# select supplement with dose < 2.0 mg/day
vc_gp_lt_2mg <- subset(ToothGrowth, supp == 'VC' & dose < 2.0, select=c(1, 3))
oj_gp_lt_2mg <- subset(ToothGrowth, supp == 'OJ' & dose < 2.0, select=c(1, 3))
```

```
t.test(vc_gp_lt_2mg$len, oj_gp_lt_2mg$len, paired=FALSE, var.equal=FALSE)
```

```
#
#   Welch Two Sample t-test
#
# data:  vc_gp_lt_2mg$len and oj_gp_lt_2mg$len
# t = -3.0503, df = 36.553, p-value = 0.004239
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -9.304766 -1.875234
```

```
# sample estimates:
# mean of x mean of y
#    12.375    17.965
```

5.3.2

```
# select supplement with dose == 2.0 mg/day
vc_gp_2mg <- subset(ToothGrowth, supp == 'VC' & dose >= 2.0, select=c(1, 3))
oj_gp_2mg <- subset(ToothGrowth, supp == 'OJ' & dose >= 2.0, select=c(1, 3))
```

```
t.test(vc_gp_2mg$len, oj_gp_2mg$len, paired=FALSE, var.equal=FALSE)
```

```
#
#   Welch Two Sample t-test
#
# data:  vc_gp_2mg$len and oj_gp_2mg$len
# t = 0.046136, df = 14.04, p-value = 0.9639
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -3.63807  3.79807
# sample estimates:
# mean of x mean of y
#     26.14     26.06
```

5.3.3

```
# define 3 groups according to doses
dose_0_5mg <- subset(ToothGrowth, dose == 0.5, len)
dose_1_0mg <- subset(ToothGrowth, dose == 1.0, len)
dose_2_0mg <- subset(ToothGrowth, dose == 2.0, len)

# perform t-test for each combination
t1 <- t.test(dose_0_5mg, dose_1_0mg, paired=FALSE, var.equal=FALSE)
t2 <- t.test(dose_0_5mg, dose_2_0mg, paired=FALSE, var.equal=FALSE)
t3 <- t.test(dose_1_0mg, dose_2_0mg, paired=FALSE, var.equal=FALSE)
lt <- list(t1, t2, t3)

# gather the results
sum_df <- data.frame(
  'name' = sapply(lt, function(x) { x$data.name}),
  'p-value'  = sapply(lt, function(x) { x$p.value }),
  't-stat' = sapply(lt, function(x) { round(x$statistic[[1]], 3) }),
  'df' = sapply(lt, function(x) { round(x$parameter[[1]], 1) }),
  'low-conf-int' = sapply(lt, function(x) { round(x$conf.int[1], 3) }),
  'up-conf-int' = sapply(lt, function(x) { round(x$conf.int[2], 3) }),
  'mean-x' = sapply(lt, function(x) { x$estimate[1] }),
  'mean-y' = sapply(lt, function(x) { x$estimate[2] })
)
rownames(sum_df) <- NULL

# show results
knitr::kable(sum_df, caption='t-test summary')
```

5.4 References

- Coursera “Statistical Inference” course, and in particular week 3 and 4.

- R documentation, *ToothGrowth datasets*