Statistical Inference: ToothGrowth Inferential Analysis

Marcio Gualtieri

Contents

Installing the Required Packages

You might need to install the following packages if you don't already have them:

```
install.packages("xtable")
install.packages("ggplot2")
install.packages("gridExtra")
```

Importing the Required Packages

Once the libraries are installed, they need to be loaded as follows:

```
suppressMessages(library(xtable))                          # Pretty printing dataframes
suppressMessages(library(ggplot2))                         # Plotting
suppressMessages(library(gridExtra, warn.conflicts = FALSE))
```

ToothGrowth Inferential Analysis

Data Loading

Our data-set consists of the ToothGrowth data-set, which comes with R. You will find more information about this data-set here.

```
data(ToothGrowth)
```

Exploratory Data Analysis

Data Types

From R's documentation:

Variable	Type	Meaning
len	numeric	Tooth growth.
supp	factor	Supplement type: VC (ascorbic acid) or OJ (Orange Juice).
dose	numeric	Dose in milligrams/day: 0.5, 1 or 2

But we also can take a look at the schema by ourselves:

```
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 ...
```

Let's also take a sample to see how the data-set looks like:

```
render_table <- function(data) {
  options(xtable.comment = FALSE)
  print(xtable(data))
}

sample_data_frame <- function(data, size) {
  sample_index <- sample(1:nrow(data), size)
  return(data[sample_index, ])
}

render_table(sample_data_frame(ToothGrowth, 6))
```

	len	supp	dose
36	10.00	OJ	0.50
18	14.50	VC	1.00
24	25.50	VC	2.00
52	26.40	OJ	2.00
37	8.20	OJ	0.50
48	21.20	OJ	1.00

Record Count

Given the different combinations for "supp" and "dose", we have the following sub-sets of data:

```
VC_data <- ToothGrowth[ToothGrowth$supp == 'VC', ]
OJ_data <- ToothGrowth[ToothGrowth$supp == 'OJ', ]
VC_0.5_data <- ToothGrowth[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 0.5, ]
OJ_0.5_data <- ToothGrowth[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 0.5, ]
```

```
VC_1_data <- ToothGrowth[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 1, ]
OJ_1_data <- ToothGrowth[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 1, ]
VC_2_data <- ToothGrowth[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 2, ]
OJ_2_data <- ToothGrowth[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 2, ]
```

I'm going to do this now since these will be useful when we do inferential analysis.

The record count summary is therefore:

```
record_count <- data.frame(record = c("total",
                                      "supp VC", "supp OJ",
                                      "dose 0.5", "dose 1", "dose 2",
                                      "0.5 VC", "0.5 OJ",
                                      "1 VC", "1 OJ",
                                      "2 VC", "2 OJ"),
                           count =  c(nrow(ToothGrowth),
                                      nrow(VC_data), nrow(OJ_data),
                                      nrow(ToothGrowth[ToothGrowth$dose == 0.5, ]),
                                      nrow(ToothGrowth[ToothGrowth$dose == 1, ]),
                                      nrow(ToothGrowth[ToothGrowth$dose == 2, ]),
                                      nrow(VC_0.5_data), nrow(OJ_0.5_data),
                                      nrow(VC_1_data), nrow(OJ_1_data),
                                      nrow(VC_2_data), nrow(OJ_2_data)))

render_table(record_count)
```

	record	count
1	total	60
2	supp VC	30
3	supp OJ	30
4	dose 0.5	20
5	dose 1	20
6	dose 2	20
7	0.5 VC	10
8	0.5 OJ	10
9	1 VC	10
10	1 OJ	10
11	2 VC	10
12	2 OJ	10

Data Visualization

Let's visualize how the data is segmented through box plots. Given that "dose" is actually categorical, we will convert it to factor before producing box plots. Plots will behave differently if they interpret a column to be continuous.

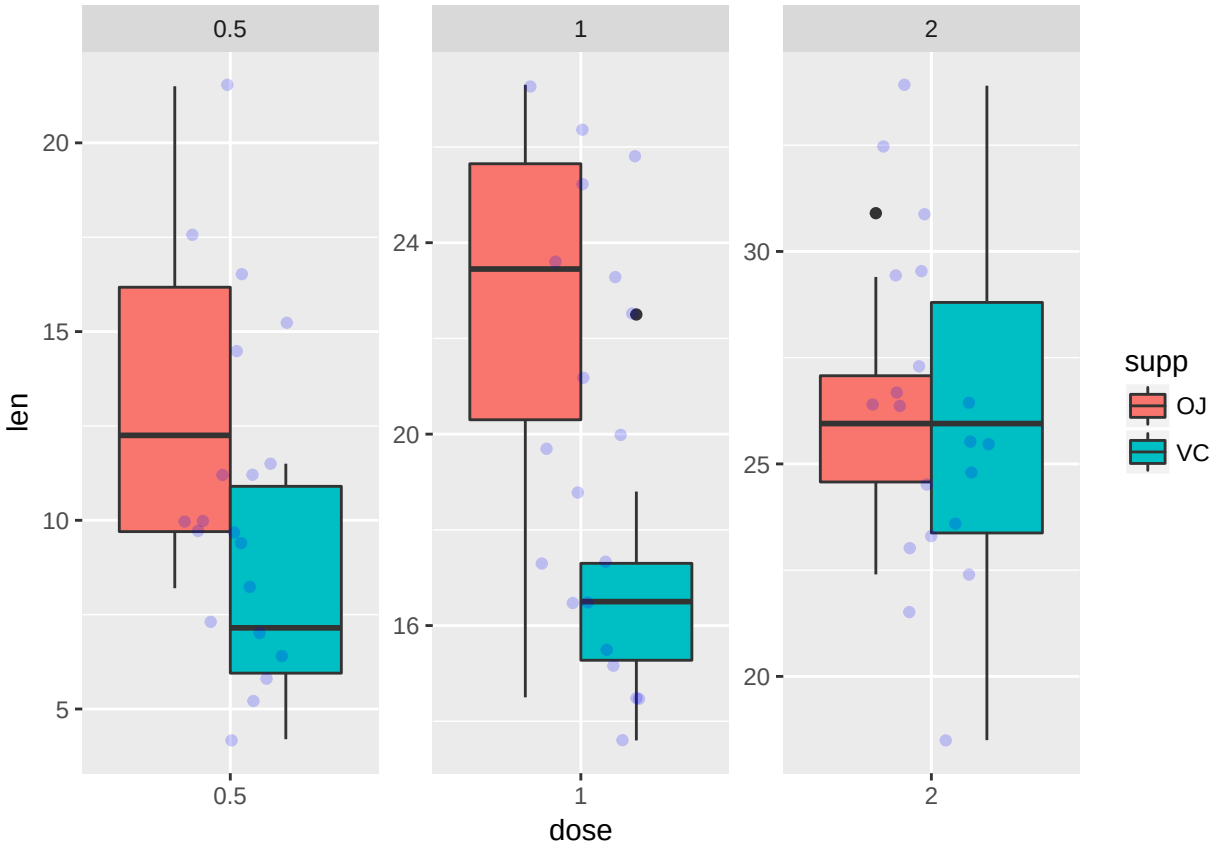
```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

Here's the box plot for the data:

```
box_plot <- function(data, x_column, y_column, color_column) {
  ggplot(data, aes_string(x = x_column, y = y_column)) +
  geom_boxplot(aes_string(fill = color_column)) +
  geom_point(position = position_jitter(width = 0.2), color = "blue", alpha = 0.2) +
  facet_wrap(as.formula(paste("~", x_column)), scales="free")
```

```
}

box_plot(ToothGrowth, "dose", "len", "supp")
```



Visual inspection seems to suggest that "OJ" does better than "VC" for all doses but 2.0, for which they seem to do only equally well.

Comparing Tooth Growth by "dose" and "supp"

Given that we don't know the standard deviation for this distribution and that the number of records is rather small, we are going to use a t-distribution.

Hypothesis Test

Test's Configuration

Let's first define the hypothesis that the treatment had no effect at all, that is:

$$H_0 : len = 0$$

$$H_a : len > 0$$

We also define $\alpha = 0.05$ (a 95% confidence interval).

P-values

```
t.test(ToothGrowth$len, y = NULL, alternative = c("greater"), mu = 0,
       paired = FALSE, var.equal = FALSE, conf.level = 0.95)$p.value
```

```
## [1] 3.470317e-27
```

With such P-Value smaller, much smaller than $\alpha = 0.05$, we thus reject the null hypothesis. We state, with 95% confidence, that there is growth. Let's apply the same hypothesis for each one of sub-sets for each different combination of "supp" and "dose":

```
t.test(VC_data$len, y = NULL, alternative = c("greater"),
       mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)$p.value
```

```
## [1] 2.181391e-12
```

```
t.test(OJ_data$len, y = NULL, alternative = c("greater"), mu = 0,
       paired = FALSE, var.equal = FALSE, conf.level = 0.95)$p.value
```

```
## [1] 5.186731e-17
```

The same goes for different treatments: They all do seem to have an affect.

```
p_value <- function(data)
  t.test(data$len, y = NULL,
         alternative = c("greater"), mu = 0,
         paired = FALSE, var.equal = FALSE,
         conf.level = 0.95)$p.value

p_values <- c(p_value(VC_0.5_data), p_value(OJ_0.5_data),
              p_value(VC_1_data), p_value(OJ_1_data),
              p_value(VC_2_data), p_value(OJ_2_data))

mean(p_values > 0.05)
```

```
## [1] 0
```

Thus the null hypothesis is rejected for any combination of "supp" and "dose". Let's try to find out if any of the treatments is superior to the other:

```
t.test(OJ_0.5_data$len, VC_0.5_data$len, alternative = "greater",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95)$p.value
```

```
## [1] 0.003179303
```

```
t.test(OJ_1_data$len, VC_1_data$len, alternative = "greater",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95)$p.value
```

```
## [1] 0.0005191879
```

```
t.test(OJ_2_data$len, VC_2_data$len, alternative = "greater",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95)$p.value
```

```
## [1] 0.5180742
```

We thus reject the hypothesis that "OJ" is equal to "VC" for doses equal 0.5 and 1, but we can't reject the null hypothesis for a dose equal to 2.

Conclusions

Treatments for all possible configurations of "dose" and "supp" seem to show growth. For small doses (0.5 and 1), "OJ" seems superior to "VC", but for a greater dose (2), "OJ" could be as effective as "VC". You might remember that this conclusions are in agreement with the intuition we got from the box plots in a previous section.

Permutation Test

Just for the sake of fun, let's try to apply permutation test to this data-set.

Test's Configuration

Here's our statistic function, the difference between the means:

```
testStat <- function(y, g) mean(y[g == "OJ"]) - mean(y[g == "VC"])
```

We are trying to determine if "OJ" and "VC" are interchangeable (and therefore equivalent) regarding tooth growth:

```
y <- ToothGrowth$len
group <- ToothGrowth$supp
```

We are also defining $\alpha = 0.05$ for this test.

P-value

Let's compute the P-value for the permutations:

```
observedStat <- testStat(y, group)
permutations <- sapply(1 : 10000, function(i) testStat(y, sample(group)))
mean(permutations > observedStat)
```

```
## [1] 0.0295
```

Conclusion

We obtained a P-value smaller than $\alpha = 0.05$, thus we reject the hypothesis that "OJ" and "VC" are interchangeable.