Statistical Inference Course Project: Part 2

Bruno

2023-08-04

Summary

In this work, we use the dataset ToothGrowth from the package *datasets*. As explained in the RDocumentation (click here for more details), this dataset describes the response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs, where each individual received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice (OJ) or ascorbic acid (VC).

The following packages were used in this analysis.

```
library(datasets)
library(tidyverse)
library(reshape2)
library(ggplot2)
```

Exploratory Analysis

Our dataset consists of three variables: **len**, **supp**, and **dose**. Such variables represent the length of odontoblasts, supplement type, and dosage, respectively. We format the variable len as numeric, and supp and dose as factor variables.

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

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(df)
```

```
##       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
```

Let us group the data by supp and dose in order to explore the possible dependence of the tooth length on the supplement and on the dosage. In the following table, we summarize the average tooth length and standard deviation grouped by the different dosages and supplements.

```
df %>% group_by(supp, dose) %>% summarise(mean = mean(len), sd = sd(len))
```

```
## # A tibble: 6 x 4
## # Groups:   supp [2]
##   supp  dose   mean    sd
##   <fct> <fct> <dbl> <dbl>
## 1 OJ    0.5   13.2   4.46
## 2 OJ    1     22.7   3.91
## 3 OJ    2     26.1   2.66
## 4 VC    0.5    7.98  2.75
## 5 VC    1     16.8   2.52
## 6 VC    2     26.1   4.80
```

In Figure 1, for each supplement (OJ or VC), we show the boxplots containing tooth lengths corresponding to different dosages. In these plots, we can identify a trend of growth of tooth length with respect to the dosage for both supplements. In order to verify whether the difference of mean tooth lengths corresponding to different dosages is statistically significant, we need to perform some hypothesis testing.

```
ggplot(df, aes(x = dose, y = len, fill = dose)) +
    geom_boxplot() +
    facet_grid(. ~ supp)
```

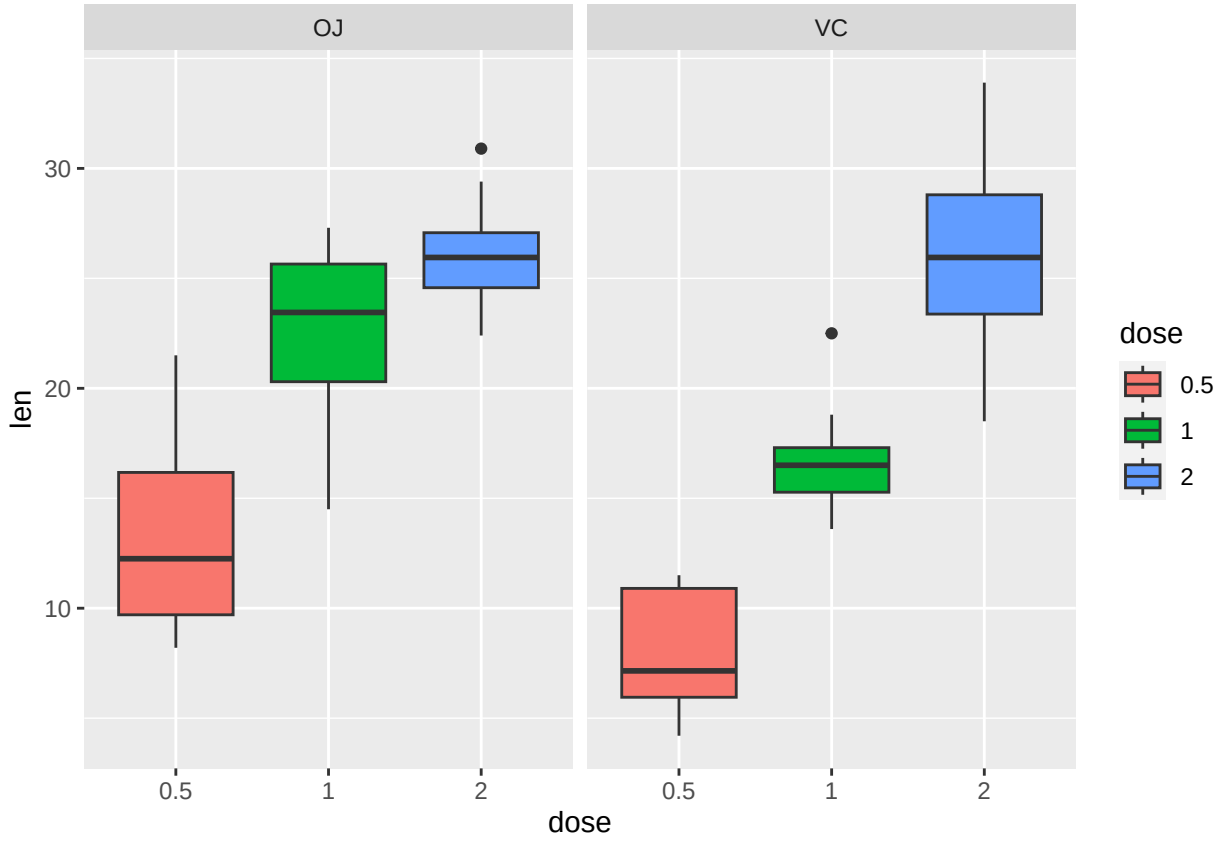

Figure 1: Boxplots of tooth lengths versus the dosage corresponding to the two supplements used.

We also show in Figure 2 the boxplots comparing the tooth lengths corresponding to different supplements separated by different dosages. In these plots, the average tooth length corresponding to the OJ supplement appears distinct from the average corresponding to VC for dosages equal to 0.5 and 1 mg/day. Now, such a distinction is not so apparent for dosages equal to 2 mg/day. Thus, we need to perform hypothesis testing to confirm whether our hypotheses are true.

```r
ggplot(df, aes(x = supp, y = len, fill = dose)) +
    geom_boxplot() +
    facet_grid(. ~ dose)
```

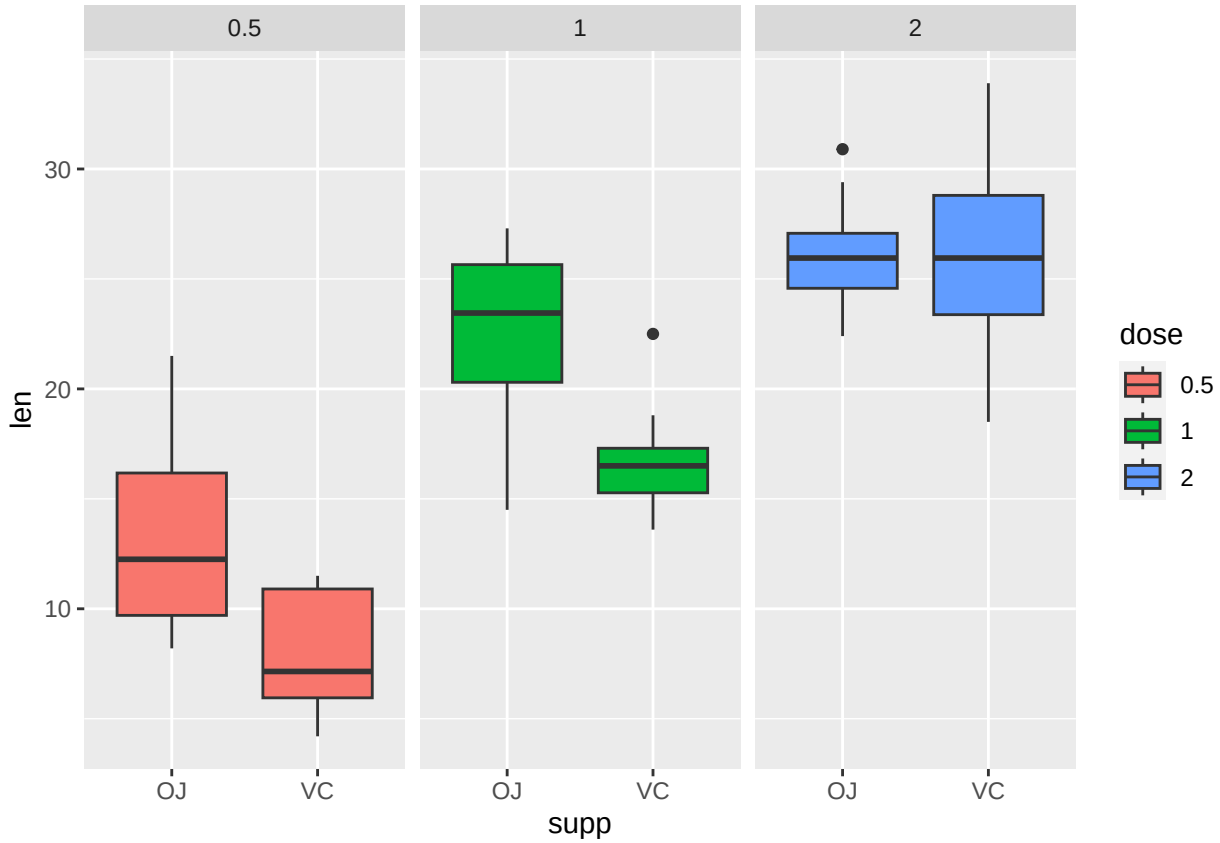

Figure 2: Box plots of tooth lengths versus supplements corresponding to tree different dosages.

Hypotheses tests

Assumptions

In this section, we assume that the underlying probability distributions of tooth length measurements are normally distributed. Moreover, we also assume that the lengths for different dosages and supplements have different variances. For our analysis, we consider $\alpha = 0.05$.

Fixed supplement

In this first part, given a supplement (OJ or VC), and doses $i < j$, we consider the following null and alternative hypotheses:

- H_0: The difference $\mu_i - \mu_j$ between the mean tooth length for dosages i and j of the given supplement is equal zero
- H_a: The difference $\mu_i - \mu_j$ between the mean tooth length for dosages i and j of the given supplement is less than zero

Due to our assumptions, the most appropriate test to perform in our analysis is the t-test. The code below performs the hypothesis testing using the function t.test.

```
tab1 <- NULL
Dose <- c(0.5, 1.0, 2.0)
for (k in c("OJ", "VC"))
    for (i in Dose)
        for (j in Dose)
            if (i < j) {
                tab1 <- rbind(tab1,
                              data.frame(
                                  supp = k,
                                  dose.1 = i,
                                  dose.2 = j,
                                  p = with(filter(df, supp == k),
                                           t.test(len[dose == i],
                                                  len[dose == j],
                                                  alternative = "less"))$p.value
                              ))
            }

print(tab1)
```

```
##   supp dose.1 dose.2           p
## 1   OJ    0.5      1 4.392460e-05
## 2   OJ    0.5      2 6.618919e-07
## 3   OJ    1.0      2 1.959757e-02
## 4   VC    0.5      1 3.405509e-07
## 5   VC    0.5      2 2.340789e-08
## 6   VC    1.0      2 4.577802e-05
```

The table above presents the p-values for the t-tests comparing the means for doses i and j, $i < j$, for each supplement. We can conclude from the p-values that the null hypotheses can be rejected; therefore, we can conclude that, in fact, the tooth length increases with the increase of supplement dose.

Fixed dose

Now let us consider the case where the dose is fixed and we compare the tooth growth for the two supplements. Given a dose of i, let us consider the following hypotheses:

- H_0: The difference $\mu_{\text{OJ}} - \mu_{\text{VC}}$ between the means for OJ and VC is equal zero
- H_a: The difference $\mu_{\text{OJ}} - \mu_{\text{VC}}$ between the means for OJ and VC is different from zero

Similarly as before, we perform the hypotheses tests and report the p-values in a table format.

```
tab2 <- NULL
for (i in Dose)
    tab2 <- rbind(
                tab2,
                data.frame(dose = i,
                           p = with(filter(df, dose == i), t.test(len ~ supp))$p.value)
            )

print(tab2)
```

```
##   dose           p
## 1  0.5 0.006358607
## 2  1.0 0.001038376
## 3  2.0 0.963851589
```

From the table above, we conclude that, in fact, there is a statistically significant difference between the means of tooth growth for OJ and VC for doses equal $i = 0.5$ and $i = 1.0$, whereas for $i = 2.0$ we fail to reject the null hypothesis.

Adjusted p-values

Since we performed a series of hypothesis tests, let us adjust the p-values obtained to confirm whether the conclusions from the previous section still hold. In the following , we adjust the p-values using two methods: Bonferroni and Benjamini-Hochberg.

```
ptab <-
    data.frame(p_value = c(tab1$p, tab2$p)) %>%
    mutate(
        Bonferroni = p.adjust(p_value, method = "bonferroni"),
        BH = p.adjust(p_value, method = "BH")
    )

print(ptab)
```

```
##        p_value   Bonferroni           BH
## 1 4.392460e-05 3.953214e-04 8.240043e-05
## 2 6.618919e-07 5.957027e-06 1.985676e-06
## 3 1.959757e-02 1.763781e-01 2.204727e-02
## 4 3.405509e-07 3.064958e-06 1.532479e-06
## 5 2.340789e-08 2.106710e-07 2.106710e-07
## 6 4.577802e-05 4.120021e-04 8.240043e-05
## 7 6.358607e-03 5.722746e-02 8.175352e-03
## 8 1.038376e-03 9.345383e-03 1.557564e-03
## 9 9.638516e-01 1.000000e+00 9.638516e-01
```

To compare the adjusted p-values with the original values, we plot them in the following graph, see Figure 3. Since Bonferroni's method tends to be very conservative, let us take into account only the results obtained through the Benjamini–Hochberg method. Therefore, we can conclude that, although we cannot say that there is a difference in the means corresponding to OJ and VC when the dosage is equal to 2 mg/day, the difference of all other means we compared is statistically significant.

```
ptab <-
    melt(
        data = cbind(index = 1:9, ptab),
        id.vars = 1,
        variable.name = 'Method',
        value.name = "p"
    )

ggplot(data = ptab, aes(x = index, y = p, color = Method, shape = Method)) +
    geom_point() +
    geom_hline(yintercept = 0.05, color = "red")
```

Conclusion

We conclude that given a supplement (OJ or VC), the mean tooth lengths increase with the increase in the dosage ($p < 0.05$). Moreover, when the dosage is fixed equal to 0.5 or 1.0 mg/day, there is a statistically significant ($p < 0.05$) difference between the mean tooth growth for OJ and VC; however, the means cannot be distinguished for doses equal to 2.0 mg/day.

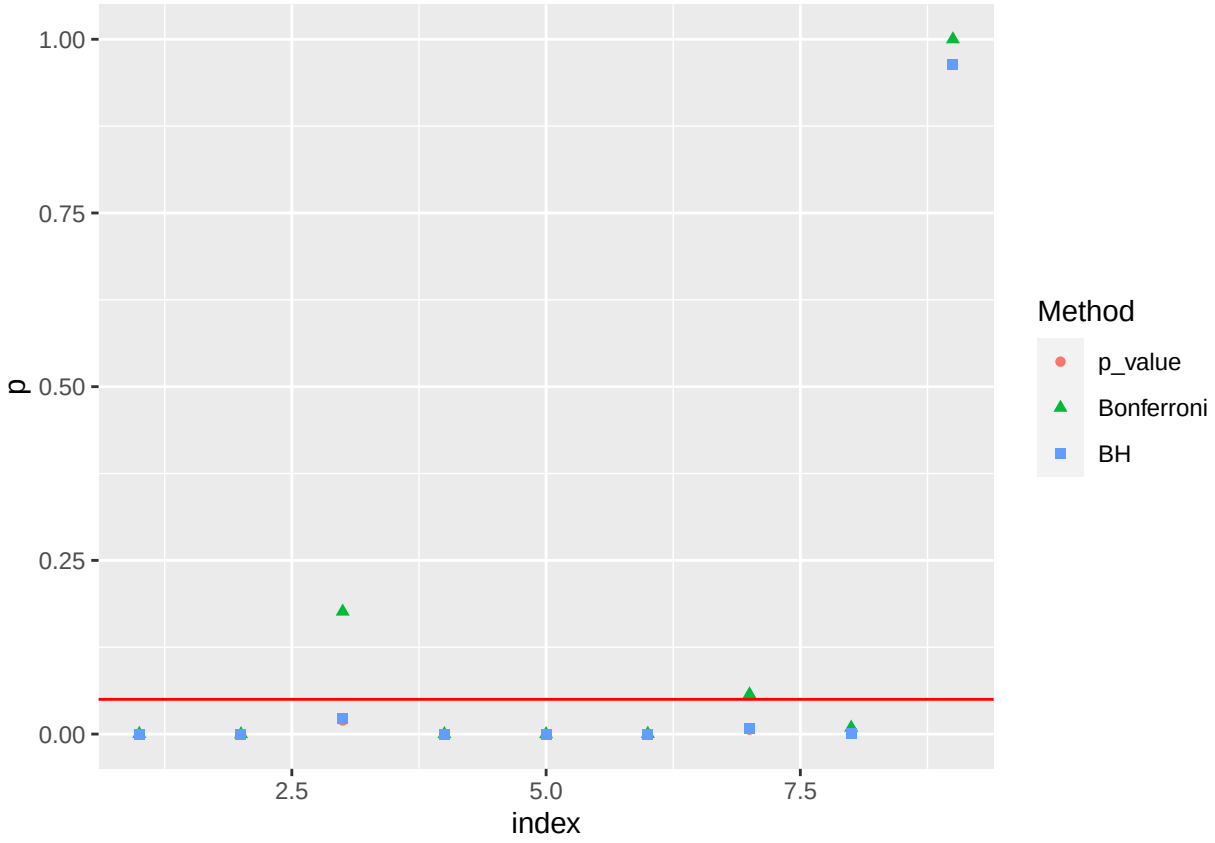

Figure 3: Comparison between the p-values obtained with their adjusted values by using Bonferoni and Benjamini–Hochberg methods. The red horizontal line represents the level 0.05.