Basic Inferential Data Analysis

Adam Reeder

Overview

This document is my submission for the second part of the final assignment of the Statistical Inference class in the Data Science specialization by the John Hopkins University on Coursera.
In this document we will analyze the ToothGrowth data from the R datasets package.

Basic exploratory data analysis and summary

We first load the data and perform some basic exploratory data analysis.

```
data(ToothGrowth)
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 ...
```

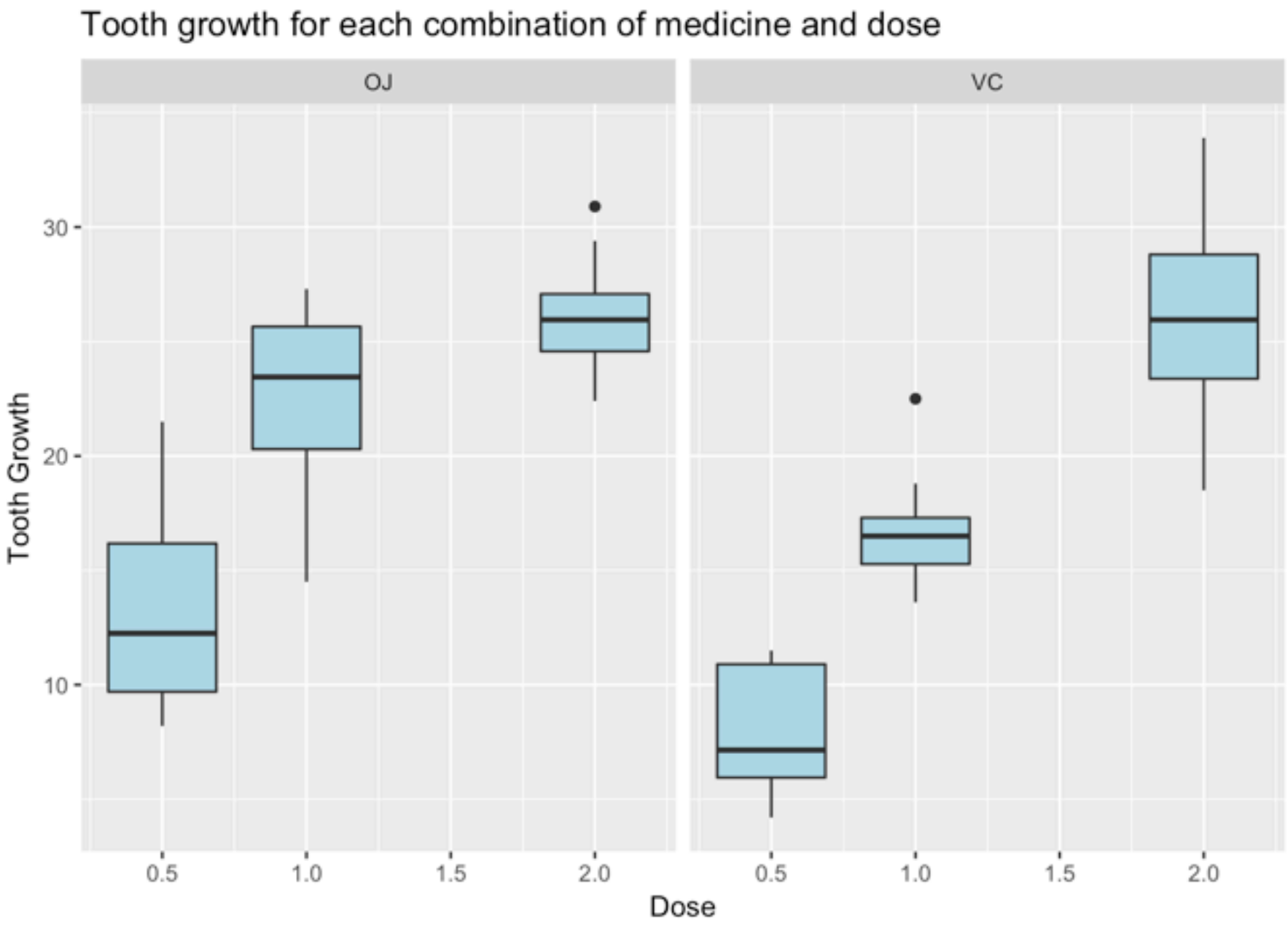
```
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
table(ToothGrowth$supp, ToothGrowth$dose)
```

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

The dataframe has three columns. It seems to be data collected from a test measuring tooth growth as a function of the medicine provided in a given dose. The first, `len`, is the tooth growth for each individual, probably in millimiters. The second one, `supp`, is a factor with two levels, each one of which probably is a given medicine. And the third one, `dose`, seems to be the quantity of medicine provided to the individual in a unknown unit.
So, two medicines were tested with three types of doses each : 0.5, 1.0 and 2.0. Each combination of medicine and dose was tested 10 times. We print below a graph representing the tooth growth for each medicine and each dose using a boxplot.



Hypothesis testing

For each dose, we want to do a hypothesis test where the null hypothesis H0 is “on average, medicine OJ causes the same tooth growth as medicine VC”. Since we will do 3 tests, we will use the Bonferroni correction in order to avoid making false discoveries. Since each tests only has a sample size of 10, we will use t-testing as it is more conservative and thus more reliable for small sample sizes.

First, we store each sample we want to test in a dedicated variables.

```
oj0.5 <- filter(ToothGrowth, supp == "OJ" & dose == .5)$len
oj1 <- filter(ToothGrowth, supp == "OJ" & dose == 1)$len
oj2 <- filter(ToothGrowth, supp == "OJ" & dose == 2)$len
vc0.5 <- filter(ToothGrowth, supp == "VC" & dose == .5)$len
vc1 <- filter(ToothGrowth, supp == "VC" & dose == 1)$len
vc2 <- filter(ToothGrowth, supp == "VC" & dose == 2)$len
```

Then we use the function `t.test` and fetch the p-values for each test. For doses 0.5 and 1, however, we chose to conduct one-sided tests, since the data seems to imply that medicine OJ induces higher tooth growth than medicine VC, that's the hypothesis we want to test. Therefore, we divide the corresponding p-values by 2.

```
p0.5 <- t.test(oj0.5, vc0.5)$p.value/2
p1 <- t.test(oj1, vc1)$p.value/2
p2 <- t.test(oj2, vc2)$p.value
pvalues <- c(p0.5, p1, p2)
```

Finally, we apply the Bonferroni correction to those p-values.

```
pvalues <- p.adjust(pvalues, method = "bonferroni")
pvalues
```

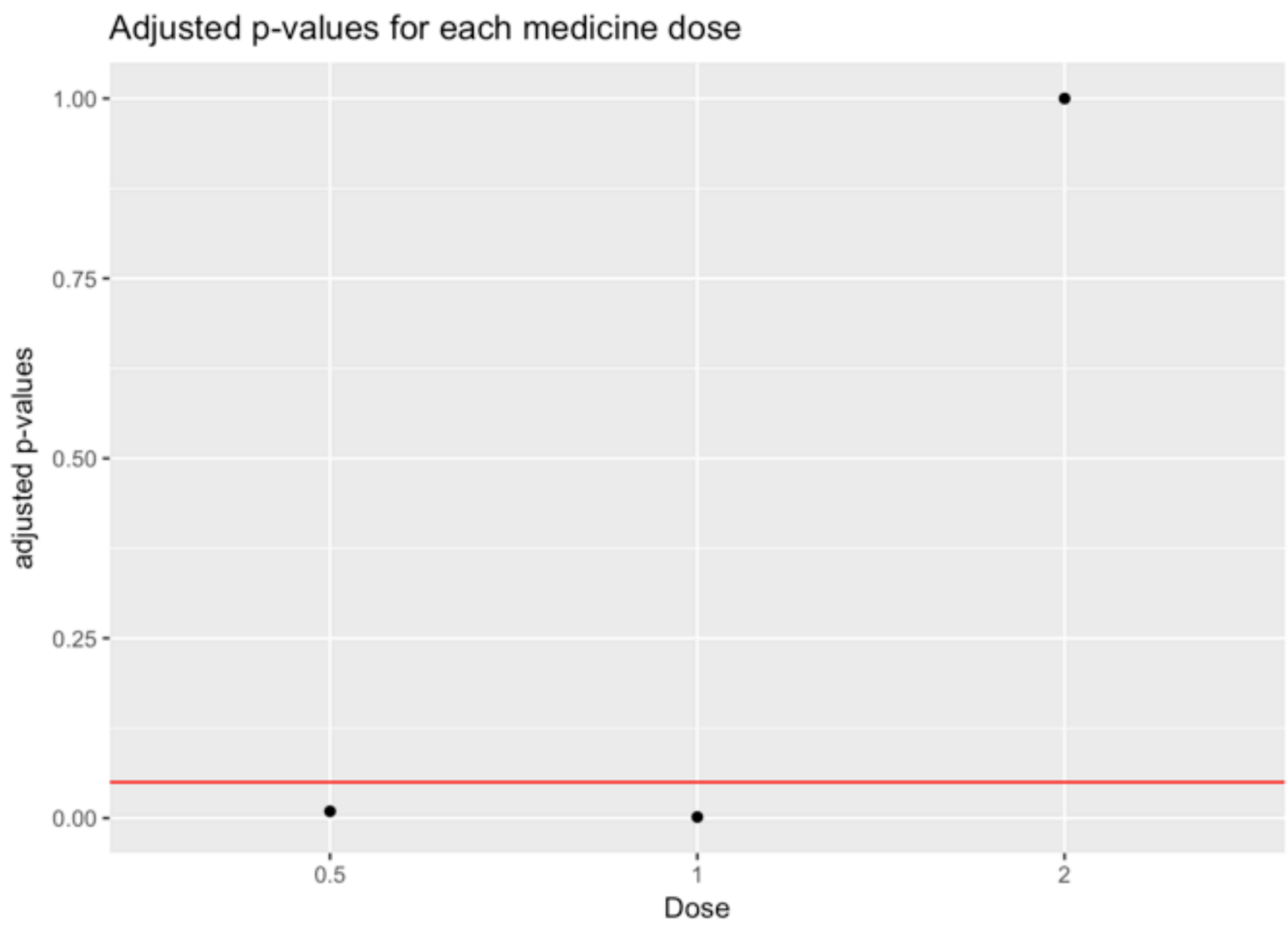
```
## [1] 0.009537910 0.001557564 1.000000000
```

Conclusion

From the adjusted p-values we obtain, we conclude using `alpha = 0.05` that : * For doses 0.5 and 1, we **reject** the null hypothesis in favor of the alternative. * For doses 2, we **fail to reject** the null hypothesis.

Concretely, that means : * For doses 0.5 and 1, we conclude that medicine OJ **induces higher growth** than medicine VC. * For dose 2, **we cannot conclude** that either medicine induces higher growth than the other.

The following plot represents the rejecting process. The dots are the adjusted p-values and the red horizontal line is `alpha = 0.05` . A null hypothesis is rejected is the corresponding adjusted p-value is below `alpha` .



Appendices

Code for Tooth growth for each combination of medicine and dose plot

```
g_boxplot <- ggplot(ToothGrowth, aes(dose, len)) +
    facet_grid(. ~ supp) +
    geom_boxplot(aes(group = dose), fill = "lightblue") +
    labs(x = "Dose", y = "Tooth Growth",
         title = "Tooth growth for each combination of medicine and dose")

suppressMessages(print(g_boxplot))
```

Code for Adjusted p-values for each medicine dose plot

```
g_accept <- ggplot(data.frame(cbind(dose = c(.5, 1, 2), pvalues)),
                   aes(factor(dose), pvalues)) +
    geom_point() +
    geom_hline(yintercept = 0.05, colour = "red") +
    labs(x = "Dose", y = "adjusted p-values",
         title = "Adjusted p-values for each medicine dose")

suppressMessages(print(g_accept))
```