Statistical Inference Course Project - Part 2

Alexander M Fisher

November 17 2020

Introduction: This is part 2 of the course project that goes along with the statistical inference course a part of the data science specialization that is run by John Hopkins University on Coursera. The project consists of two parts:

- A simulation exercise.
- Basic inferential data analysis.

Basic inferential data analysis will be covered in this report, which involves analysing the ToothGrowth data in the R datasets package. For the this part there are four main instructions listed below.

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
4. State your conclusions and the assumptions needed for your conclusions.

Part 2: Basic inferential data analysis

Initial Exploratory Analysis: Lets load the data into R and take a quick look at the data.

```
data("ToothGrowth")
data <- ToothGrowth
head(data)
```

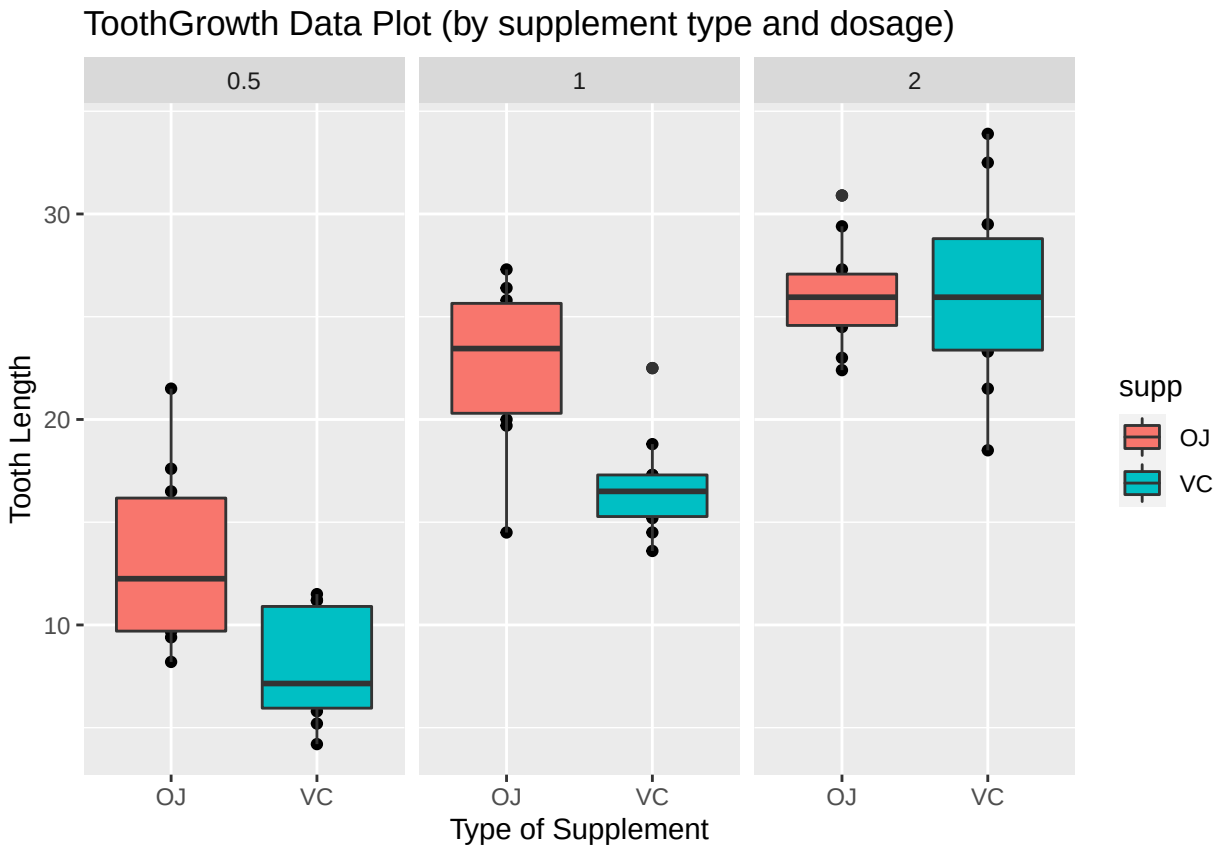
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
str(data)
```

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

From this we can see there are 2 numeric columns, and a factor column with two levels. The dimension is 60 by 3 hence 60 observations. Note also if we run `unique(ToothGrowth$does)` we can see there are only three unique values of dose. Either a dose of 0.5, 1, or 2 units was given. Lets go a head and make dose variable also a factor and then make a boxplot.

```
library(ggplot2)
plot <- qplot(data=data, supp,len, facets=~dose,main="ToothGrowth Data Plot (by supplement type and dosa
        xlab="Type of Supplement", ylab="Tooth Length") +
        geom_boxplot(aes(fill = supp))
print(plot)
```



We can see from this initial plot that there is a linear relationship between amount of dosage administered and tooth length. That is to say the more supplement given then the bigger tooth length as a response. It can also be seen that for all dosages OJ result in larger tooth length, although only very narrowly for the dose of 2.

Hypothesis Tests: From the above graph it can be seen there are perhaps some questions we would like to answer. These will include,

- Is orange juice OJ a better or more effective form of administering vitamin C compared with ascorbic acid (VC). I.e. does OJ promote more tooth growth.
- In addition we may like to look more specifically at each does level and ask for that level of dose is there a significant difference in tooth length between supp types.
- Lastly I am also interested if (only for supp OJ) tooth length is statistically bigger for dose 2mg vs does 1mg.

These test will be done with a two sample t test where we take sub groups of data from the main data frame and run `t.test()`. Below I listed a few assumptions that are taken with regards to the data and tests.

- The variables must be independent and identically distributed
- Tooth growth follows a normal distribution.
- For all tests alpha level is 0.05

OJ vs VC:

- H_0: there is no difference in tooth length means between each supp OJ and VC.
- H_a: the alternative is supp OJ mean tooth length > supp VC mean tooth length

Lest get the data and run the test.

```
OJ <- data[data$supp == "OJ", "len"]
VC <- data[data$supp == "VC", "len"]

test_1 <- t.test(x = OJ, y = VC, alternative = "greater" , paired = FALSE, var.equal = FALSE, conf.level
print(test_1)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ and VC
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

The null is rejected as **0.0302<0.05**. That is to say that OJ corresponds to greater tooth lengths (across all doses) than VC.

Dose 0.5 OJ vs VC:

- H_0: there is no difference in tooth length means between each OJ and VC.
- H_a: the alternative is OJ mean tooth length > VC mean tooth length

Lest get the data and run the test.

```
OJ <- data[data$supp == "OJ" & data$dose == "0.5", "len"]
VC <- data[data$supp == "VC" & data$dose == "0.5", "len"]
test_2 <- t.test(x = OJ, y = VC, alternative = "greater" , paired = FALSE, var.equal = FALSE)
print(test_2$p.value)
```

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

So reject null, accept alternative. So OJ outperforms VC with regards to mean tooth length for doses of 0.5mg.

Dose 1 OJ vs VC:

- H_0: there is no difference in tooth length means between each OJ and VC.
- H_a: the alternative is OJ mean tooth length > VC mean tooth length

Lest get the data and run the test.

```
OJ <- data[data$supp == "OJ" & data$dose == "1", "len"]
VC <- data[data$supp == "VC" & data$dose == "1", "len"]
test_3 <- t.test(x = OJ, y = VC, alternative = "greater" , paired = FALSE, var.equal = FALSE)
print(test_3$p.value)
```

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

So reject null, accept alternative. So OJ outperforms VC with regards to mean tooth length for doses of 1mg as well. These last two results were expected as the boxplot indicates a reasonable difference between OJ and VC for both 0.5mg and 1mg doses. This is reflected in the p-values, especially in the test just done which has a p-value approx equal to 0%. The next will be more interesting, however. Most likely there will not be enough evidence to suggest OJ outperforms VC for doses of 2mg. Lets see!

Dose 2 OJ vs VC:

- H_0: there is no difference in tooth length means between OJ and VC.
- H_a: the alternative is OJ mean tooth length > VC mean tooth length

```
OJ <- data[data$supp == "OJ" & data$dose == "2", "len"]
VC <- data[data$supp == "VC" & data$dose == "2", "len"]
test_4 <- t.test(x = OJ, y = VC, alternative = "greater" , paired = FALSE, var.equal = FALSE)
print(test_4$p.value)
```

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

As expected, as indicated visually in the boxplot, for doses of 2mg, OJ and VC perform the same on average. It can be noted that there is more variance in VC so for consistency OJ is probably the better supplement to administer still. The final test I will complete is probably the most interesting as it is difficult to tell directly by the graph. The question is on average does administering 2mg instead of 1mg of OJ produce better, i.e. higher tooth lengths.

Dose OJ 2mg vs 1mg:

- H_0: there is no difference in tooth length means between 2mg and 1mg for OJ
- H_a: the alternative is mean tooth length for 2mg > mean tooth length for 1mg

```
OJ_2 <- data[data$supp == "OJ" & data$dose == "2", "len"]
OJ_1 <- data[data$supp == "OJ" & data$dose == "1", "len"]
test_4 <- t.test(x = OJ_2, y = OJ_1, alternative = "greater" , paired = FALSE, var.equal = FALSE)
print(test_4)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ_2 and OJ_1
## t = 2.2478, df = 15.842, p-value = 0.0196
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7486236       Inf
## sample estimates:
## mean of x mean of y
##     26.06     22.70
```

The null is again rejected and alternate accepted. To recap the findings then,

- OJ doses produces longer tooth lengths on average compared with VC.
- For doses 0.5mg and 1mg there is a statistical difference between OJ and VC, that is to say again OJ produces longer tooth lengths.
- For dose 2mg there is no statistical difference between OJ and VC.
- Although there appears visually a slight increase in OJ for doses 1mg and 2mg, it is statistically significant and does 2mg therefore produce on average larger tooth lengths.

It can hopefully be seen that on average ignoring supplement types that 1mg doses produce larger tooth lengths than 0.5mg doses. This appears to also be the case between 2mg and 1mg doses although not as severe.

This could be worth while testing directly. Indirectly it has however, been shown via the tests completed that there is indeed a positive relationship between dose and tooth length. It can be seen for VC that each dosage increase produces unequivocally increases in tooth length. This increase is the same clearly for OJ for doses 0.5mg and 1mg, and shown statistically significant also for doses 1mg to 2mg (OJ). On a final note perhaps a linear regression/model can be done to further investigate the relationship between dose, and tooth length.
