Part 2: Basic Inferential Data Analysis

Yohance Nicholas

3/15/2020

Overview

The second part of the coursework project requires further application of the concepts taught throughout the course in Statistical Inference. This is done with the assistance of the base Tooth Growth Dataset which captures The Effect of Vitamin C on Tooth Growth in Guinea Pigs. In the experiment, each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC). As such, The dataset contains 60 observations of three variables, namely:

1. Tooth Length - **len**
2. Supplement used (Vitamin C or Orange Juice) - **supp**
3. The dose in milligrams/day - **dose**

The coursework project requires basic exploratory data analysis in the form of:

1. A summary of the data
2. Confidence Intervals and/or Hypothesis Tests to compare tooth growth by supplement and dose (Using strictly those techniques covered in the online course)
3. Conclusions based on the information and an identification of all assumptions required for these conclusions

Step 1: Load Required Packages and perform Exploratory Data Analysis

Load required data

```
library(datasets)
data(ToothGrowth)
head(ToothGrowth)
```

```
##    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
```

Basic Exploratory Data Analysis

```
library(psych)
describe(ToothGrowth)
```

```
##       vars  n  mean   sd median trimmed  mad min  max range  skew kurtosis   se
## len      1 60 18.81 7.65  19.25   18.95 9.04 4.2 33.9  29.7 -0.14    -1.04 0.99
## supp*    2 60  1.50 0.50   1.50    1.50 0.74 1.0  2.0   1.0  0.00    -2.03 0.07
## dose     3 60  1.17 0.63   1.00    1.15 0.74 0.5  2.0   1.5  0.37    -1.55 0.08
```

Step 2: Provide a Basic Summary and Visual Inspection of the dataset

As was mentioned previously, the dataset contains 3 variables - *len, supp and dose*. A summary of these three variables can be found below:

```
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
```

In order to visualise the potential impact of the efficacy of different delivery methods of vitamin C on tooth growth in guinea pigs, Figure 1 provides a box plot comparison of the relationship between Tooth Length and Vitamin C Doses by Supplement Type:

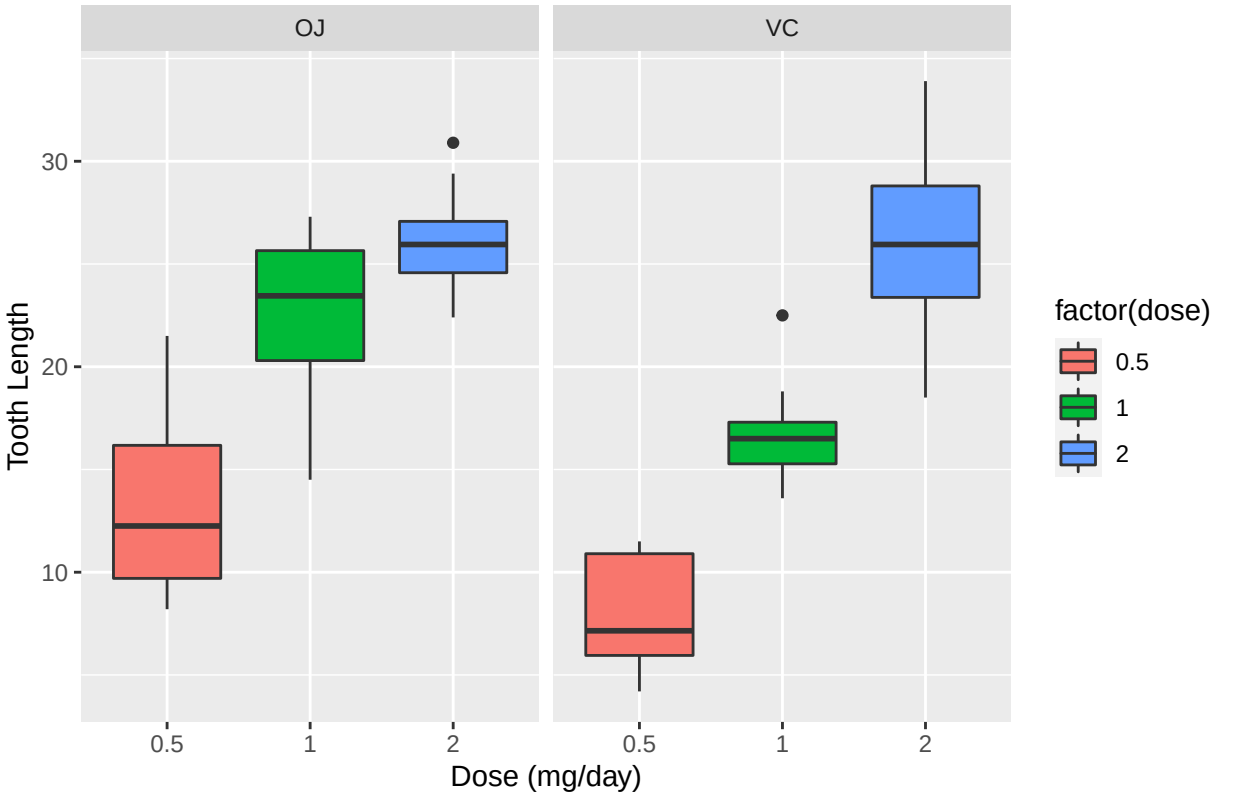
```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
gg_boxplot <- ggplot(ToothGrowth,
                     aes(x=factor(dose),
                         y=len,
                         fill=factor(dose))) +
  geom_boxplot() +
  facet_grid(.~supp) +
  labs(title="Figure 1: Box-plot Comparison of Different Supplement Type for Different Vitamin Dose",
       y="Tooth Length",
       x="Dose (mg/day)")
gg_boxplot
```

Figure 1: Box−plot Comparison of Different Supplement Type for Different Vi



A visual inspection of the box plots seem to suggest that increasing the dosage of vitamin C increases tooth growth in the sample of Guinea Pigs. The side by side comparison of these box plots also seems to suggest that Orange juice is more effective than ascorbic acid for tooth growth when in the 0.5 to 1.0 mg/day dosage range. Both types of supplements seem to be equally effective when the dosage is 2.0 mg/day. In the subsequent section, these hypotheses will be explored further.

Step 3: Confidence Intervals and Hypothesis Testing

Student's t-test or t-test is a simple yet very useful statistical test. The basic idea behind t-test is the inference problem from a small sample size data set to test whether its sample mean may have large deviation from the true population mean.

In t-test, the null hypothesis is that the mean of the two samples is equal. This means that the alternative hypothesis for the test is that the difference of the mean is not equal to zero. In a hypothesis test, we want to reject or accept the null hypothesis with some confidence interval. Since we test the difference between the two means, the confidence interval in this case specifies the range of values within which the difference may lie.

The t-test will also produce the p-value, which is the probability of wrongly rejecting the null hypothesis. The p-value is always compared with the significance level of the test. For instances, at 95% level of confidence, the significant level is 5% and the p-value is reported as $p<0.05$. Small p-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more confident we can reject the null hypothesis.

Calculating Confidence Intervals

Calculate Confidence Interval for Dose=0.5 mg/day.

```
dose_0.5 <- subset(ToothGrowth,
                   dose == 0.5)
test_0.5 <- t.test(len ~ supp,
                    paired=F,
                    var.equal=F,
                    data=dose_0.5)
```

Calculate Confidence Interval for Dose=0.5 mg/day.

```
dose_1 <- subset(ToothGrowth,
                 dose==1)
test_1 <- t.test(len ~ supp,
                 paired=F,
                 var.equal=F,
                 data=dose_1)
```

Calculate Confidence Interval for Dose=0.5 mg/day.

```
dose_2 <- subset(ToothGrowth,
                 dose==2)
test_2 <- t.test(len ~ supp,
                 paired=F,
                 var.equal=F,
                 data=dose_2)
```

Summarizing p-value and Confidence Interval in a table

```
table1 <- data.frame(
  "p.value"=c(test_0.5$p.value, test_1$p.value, test_2$p.value),
  "Conf.Low"=c(test_0.5$conf.int[1], test_1$conf.int[1], test_2$conf.int[1]),
  "Conf.High"=c(test_0.5$conf.int[2], test_1$conf.int[2], test_2$conf.int[2]),
  row.names=c("0.5 mg/day", "1.0 mg/day", "2.0 mg/day")
)
table1
```

```
##                 p.value  Conf.Low Conf.High
## 0.5 mg/day 0.006358607  1.719057  8.780943
## 1.0 mg/day 0.001038376  2.802148  9.057852
## 2.0 mg/day 0.963851589 -3.798070  3.638070
```

Step 4: Conclusions and the assumptions needed for your conclusions.

Based on the p-value, for Dose=0.05mg/day & 1mg/day, since the p-value is less than 5%, we reject the null hypothesis and conclude there is significant difference between Orange Juice and Ascorbic Acid at low dose (less than or equal 1mg/day). For Dose=2mg/day, since the p-value is greater than 5%, we failed to reject the null hypothesis and conclude there is no significant difference between Orange Juice and Ascorbic Acid at high dose (2mg/day).

In summary, the hypothesis tests demonstrate that orange juice delivers more tooth growth than ascorbic acid for dosages 0.5 & 1.0. Orange juice and ascorbic acid deliver the same amount of tooth growth for dose amount 2.0 mg/day. For the entire data set we cannot conclude orange juice is more effective that ascorbic acid.

Assumptions

- Sample of Guinea Pigs is random
- Guinea Pigs selected are homogeneous (of the same species and balanced between the sexes)
- Normal distribution of the tooth lengths
- No other unmeasured factors are affecting tooth length