Statistical Inference Course Project Part 2

Author: YZ

Part 2: Basic Inferential Data Analysis

This part of the analysis is focused on the ToothGrowth dataset from the R datasets package

Step 1. Load the ToothGrowth data and perform some basic exploratory data analyses

Before doing any analysis we need to get the description of the dataset with "?ToothGrowth" command.

Data description:

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. 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)). Available variables were "len" (Tooth length), "supp" (factor Supplement type (VC or OJ), "dose" (numeric Dose in milligrams/day)

After loading the data, we can see the following: there are 30 pigs in each "Supplement" goup; there are 20 pigs in each "Dose" group; in total there are 6 cells 10 pigs in each split by Supplement+Dose

```
data(ToothGrowth)
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
```

Step 3. Provide a basic summary of the data.

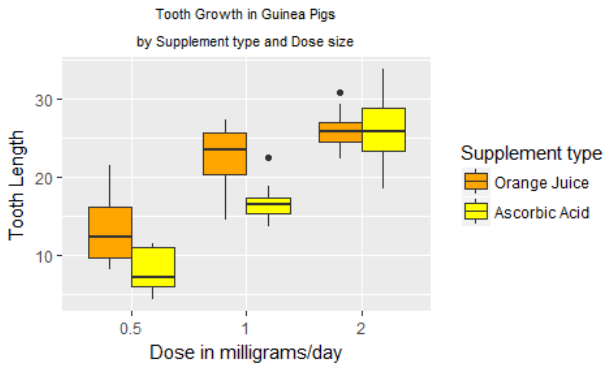
Based on the boxplot displayed below, we can see that dosage is clearly a differentiating factor: the higher the dose, the longer the teeth Distinction by supplement type is not as clear across all three groups, because in Dose = 2 group, the mean length appears to be the same between supplement types.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
t <- ggplot(ToothGrowth, aes(factor(dose), len, fill = supp)) +
    geom_boxplot() +
    labs(title = "Tooth Growth in Guinea Pigs",
         subtitle = "by Supplement type and Dose size",
         x = "Dose in milligrams/day",
         y = "Tooth Length") +
    theme(plot.title = element_text(size = rel(0.75), hjust = 0.5), plot.subtitle = element_text(size = rel(0.75),
hjust = 0.5)) +
    scale_fill_manual(name = "Supplement type",
                      values = c("orange", "yellow"),
                      labels = c("OJ" = "Orange Juice", "VC" = "Ascorbic Acid"))
t
```



Step 4. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose

Analysis assumptions:

- Based on data description, the data is not paired (60 distince pigs)
- We have no information about variance, hence, we'll compare results for equal and equal variance
- Given low number of obersvations, we'll calculate t-test based confidence intervals (95%)

First, we will calculate confidence intervals by supplement type.

```
oj <- ToothGrowth$len[ToothGrowth$supp=='OJ']
vc <- ToothGrowth$len[ToothGrowth$supp=='VC']

t.test(oj,vc,paired = FALSE, var.equal = TRUE)$conf
```

```
## [1] -0.1670064  7.5670064
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(oj,vc,paired = FALSE, var.equal = FALSE)$conf
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

Next, we'll compare average teeth length by dose. Considering we have three types of doses, we'll need to perform six comparisons (with variance equal/unequal). The comparison will be performed for lower vs higher dose groups. Hence, we would expect negative interval boudaries if dose is significant.

```
dose05 <- ToothGrowth$len[ToothGrowth$dose==0.5]
dose10 <- ToothGrowth$len[ToothGrowth$dose==1]
dose20 <- ToothGrowth$len[ToothGrowth$dose==2]

t_dose <- rbind(
    as.vector(t.test(dose05,dose10,paired = FALSE, var.equal = TRUE)$conf.int),
    as.vector(t.test(dose05,dose10,paired = FALSE, var.equal = FALSE)$conf.int),
    as.vector(t.test(dose10,dose20,paired = FALSE, var.equal = TRUE)$conf.int),
    as.vector(t.test(dose10,dose20,paired = FALSE, var.equal = FALSE)$conf.int),
    as.vector(t.test(dose05,dose20,paired = FALSE, var.equal = TRUE)$conf.int),
    as.vector(t.test(dose05,dose20,paired = FALSE, var.equal = FALSE)$conf.int)
)

t_dose_df <- data.frame(cbind(
                c("0.5 vs 1.0", "0.5 vs 1.0", "1.0 vs 2.0", "1.0 vs 2.0", "0.5 vs 2.0", "0.5 vs 2.0"),
                c("TRUE", "FALSE", "TRUE", "FALSE", "TRUE", "FALSE"),
                round(as.numeric(t_dose[,1]),2),
                round(as.numeric(t_dose[,2]),2)))

colnames(t_dose_df) <- c("Dose", "var.equal", "Lower limit", "Upper limit")
t_dose_df$`Lower limit` <- as.numeric(as.character(t_dose_df$`Lower limit`))
t_dose_df$`Upper limit` <- as.numeric(as.character(t_dose_df$`Upper limit`))
t_dose_df$`Interval length` <- t_dose_df$`Upper limit` - t_dose_df$`Lower limit`

t_dose_df
```

```
##         Dose var.equal Lower limit Upper limit Interval length
## 1 0.5 vs 1.0      TRUE      -11.98       -6.28            5.70
## 2 0.5 vs 1.0     FALSE      -11.98       -6.28            5.70
## 3 1.0 vs 2.0      TRUE       -8.99       -3.74            5.25
## 4 1.0 vs 2.0     FALSE       -9.00       -3.73            5.27
## 5 0.5 vs 2.0      TRUE      -18.15      -12.84            5.31
## 6 0.5 vs 2.0     FALSE      -18.16      -12.83            5.33
```

Step 5. State your conclusions and the assumptions needed for your conclusions.

Supplement does not appear to be a significant factor in tooth growth

- 95% confidence interval includes zero in both versions of the test (equal/unequal variance), hence, we conclude that there is no significant difference in average teeth length across two groups.

Dose size is a significant factor in tooth growth

- Variance equal/unequal factor is not relevant as the intervals are almost identical across two groups
- In all six test the intervals are well below zero, which means that dose matters for tooth growth: the higher the dose, the longer the teeth TRUE