Analysis of Tooth Growth Dataset in R

Overview

This report is comprised of two parts. In the first part, I investigated the distribution of averages of 40 exponentials by performing 1000 simulations. I compared the simulated sample mean to the theoretical mean(`1 / lambda`), and the simulated sample variance to the theoretical variance (`1 / lambda`). In part two, I briefly analyzed, summarized, and performed hypothesis tests of a dataset related to tooth growth.

Part 2

Goal: Explore, analyze, and perform a hypothesis test of the `ToothGrowth` dataset in R.

Exploratory Analysis and Data Summary

Here is a summary of the `ToothGrowth` dataset from the Help section in R.

> "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)."

The dataset is a dataframe with 60 observations and 3 variables: `len`, `supp`, and `dose`.

```
str(ToothGrowth)
```

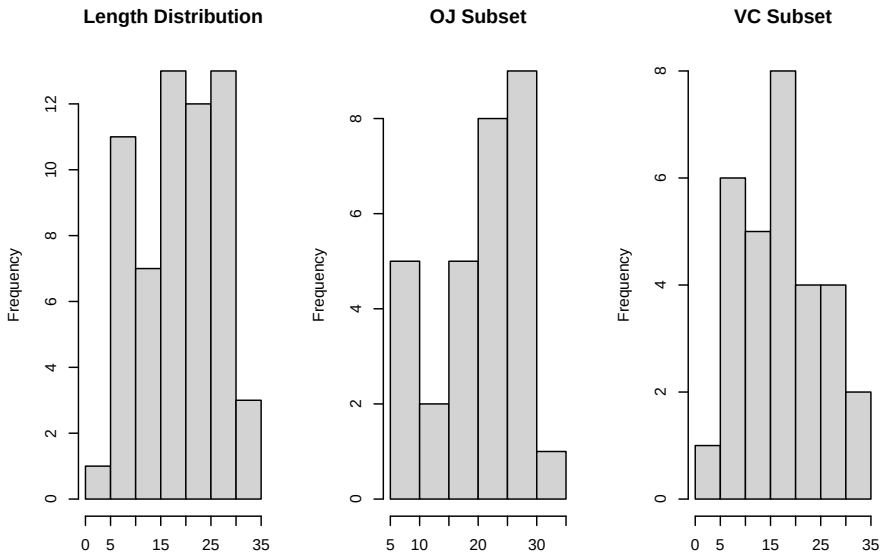
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
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
```

The outcome variable of interest is `len`, so I want to know what the distribution of this variable looks like.

```
par(mfrow = c(1, 3))
hist(ToothGrowth$len, main = "Length Distribution", xlab = "")
hist(subset(ToothGrowth, supp == "OJ")[,1], main = "OJ Subset", xlab = "")
hist(subset(ToothGrowth, supp == "VC")[,1], main = "VC Subset", xlab = "")
```



The data seems fairly normal and does not appear to be skewed in either direction, so I'll assume the normal distribution assumption has not been violated.

Now I'll make graphs to visualize any difference between groups; but first, I need to convert the `dose` column to a factor.

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

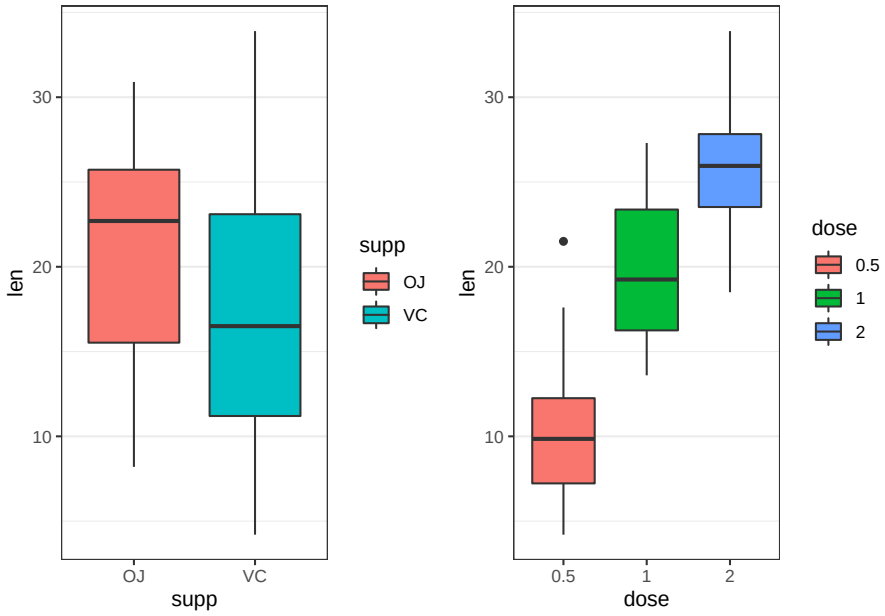
The data is now ready for plotting. Let's see what the group differences look like for `supp` and `dose`.

```
library(ggplot2)
library(gridExtra)

p3 <-
ggplot(ToothGrowth, aes(supp, len, fill = supp)) +
  geom_boxplot() +
  theme_bw() +
  theme(panel.grid.major.x = element_blank())

p4 <-
ggplot(ToothGrowth, aes(dose, len, fill = dose)) +
  geom_boxplot() +
  theme_bw() +
  theme(panel.grid.major.x = element_blank())

grid.arrange(p3, p4, ncol = 2)
```



Before performing the hypothesis tests, I created subsets of the data to make the code a little more clean and readable.

```
OJ <- subset(ToothGrowth, supp == "OJ")[1]
VC <- subset(ToothGrowth, supp == "VC")[1]

low <- subset(ToothGrowth, dose == 0.5)[1]
med <- subset(ToothGrowth, dose == 1)[1]
high <- subset(ToothGrowth, dose == 2)[1]
```

Differences in Tooth Growth by Supplement Type

I used a Two-Sample t-test to assess the difference between `supp` groups.

```
t.test(OJ, VC, alternative = "two.sided", paired = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ and VC
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

Although it's close, the p-value is not significant, which suggests there is not a significant difference between groups.

Differences in Tooth Growth by Dose

Generally an ANOVA model would be used for this analysis, but the assignment instructed us to use only techniques from class; so, I'll use 3 Two-Sample t-tests to compare the `dose` levels.

```r
(low_med <- t.test(low, med, alternative = "two.sided", paired = F))
```

```
##
##  Welch Two Sample t-test
##
## data:  low and med
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

```r
(low_high <- t.test(low, high, alternative = "two.sided", paired = F))
```

```
##
##  Welch Two Sample t-test
##
## data:  low and high
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

```r
(med_high <- t.test(med, high, alternative = "two.sided", paired = F))
```

```
##
##  Welch Two Sample t-test
##
## data:  med and high
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

The p-values are significant for all three group comparisons.

Controlling Family-Wise Error Rate

Since we performed multiple comparisons with the `dose` variable, we need to control for the FWER.

```r
p.adjust(c(low_med$p.value,
           low_high$p.value,
           med_high$p.value),
         method = "bonferroni")
```

```
## [1] 3.804902e-07 1.319257e-13 5.719289e-05
```

Even after adjusting, the p-values are significant, which suggests that each group is significantly different from the other two.