Statistical Inference: Course Project

Basic Inferential Data Analysis:

Analysis of ToothGrowth Data:

Load ToothGrowth Data:

```r
data(ToothGrowth)
```

Exploratory Data Analysis:

```r
str(ToothGrowth)  # Review Data Structure
```

```r
summary(ToothGrowth)  # Review Data Statistics
```

```r
head(ToothGrowth)  # Review some of the actual data
```

```r
unique(ToothGrowth$dose)  # Review unique values of dose control variable.
```

Plots to Evaluate relation of “len” with “supp” and “dose”:

```r
par(mfrow=c(3,2),mar=c(4,4,2,0),oma=c(0,0,2,0))
plot(aggregate(len-supp,ToothGrowth,mean),ylab="len mean")
boxplot(len-supp,ToothGrowth,xlab="supp",ylab="len")
plot(aggregate(len-dose,ToothGrowth,mean),pch=19, ylab="len mean")
lines(aggregate(len-dose,ToothGrowth,mean))
boxplot(len-dose,ToothGrowth,xlab="dose",ylab="len")
boxplot(len-supp+dose,ToothGrowth,xlab="supp+dose",ylab="len")
title(main="Evaluation Of len Vs supp, dose, supp+dose",outer=T)
```
Evaluation Of len Vs supp, dose, supp+dose

“ToothGrowth” data structure and summary overview shows that the data set has 60 observations of 3 variables, “len”, “supp” and “dose”. “len” and “dose” are numeric, while “supp” is a factor variable. Summary statistics show that variable “len” has a max value - 33.9, min value - 4.2 and mean - 18.8133. Variable “supp” has only two unique values with 30 observations each. Variable dose has a max value - 2, min value - 0.5 and mean - 1.1667. Further review of variable “dose” shows that it has only three unique
values, 0.5, 1 & 2. If necessary, we can always convert “dose” to a factor variable.

Objective of this data analysis is to evaluate the impact of control variables “supp” and “dose” on the target variable “len”, individually or together. Assuming that a higher “len” value indicates a higher impact and a higher value of “dose” indicates a higher dose, a first evaluation of the above plots, yields the following hypotheses:

1. For impact of control variable “supp” only on target variable “len”, “OJ” has a higher impact on target variable “len”.
2. For impact of control variable “dose” only on target variable “len”, higher the “dose”, higher is the impact.
3. For combined impact of control variables “supp” and “dose”, “OJ” has higher impact on target variable “len” for “dose” 0.5 & 1.
4. For combined impact of control variables “supp” and “dose”, “OJ” and “VC” have same impact on target variable “len” for “dose” 2.

Hypothesis Testing:

Hypothesis #1 - For impact of control variable “supp” only on target variable “len”, “OJ” has a higher impact on target variable “len”:

```r
G <- t.test(len ~ supp, paired = F, var.equal = F, data = ToothGrowth);
print(G)
```

```r
##
## Welch Two Sample t-test
##
## t = 1.915, df = 55.31, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    -0.171    7.571
## sample estimates:
## mean in group OJ mean in group VC
##     20.66     16.96
```

P-Value, 0.0606 is greater than α=0.05 (α for confidence interval of 95%), confidence interval, (-0.171, 7.571) for the difference of the means of each group spans 0, hence null hypothesis is Failed to Reject, hence Hypothesis #1 is Rejected.

Hypothesis #2 - For impact of control variable “dose” only on target variable “len”, Higher the “dose”, higher is the impact:

2a : dose 1 has higher impact than dose 0.5
```r
Ga = t.test(len ~ dose, paired = F, var.equal = F, data = ToothGrowth[ToothGrowth$dose %in% c(0.5, 1)],)
print(Ga)
```

```
## Welch Two Sample t-test
## data:  len by dose
## t = -6.477, df = 37.99, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.984 -6.276
## sample estimates:
## mean in group 0.5 mean in group 1
## 10.61 19.73
```

**P-Value:** $1.2683 \times 10^{-7}$ is less than $\alpha=0.05$ ($\alpha$ for confidence interval of 95%), hence null hypothesis is **Rejected**, hence Hypothesis #2a is **Failed to Reject**.

**Hypothesis 2a:** dose 2 has higher impact than dose 1

```
Gb = t.test(len ~ dose, paired = F, var.equal = F, data = ToothGrowth[ToothGrowth$dose %in% c(1, 2)],)
print(Gb)
```

```
## Welch Two Sample t-test
## data:  len by dose
## t = -4.901, df = 37.1, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996 -3.734
## sample estimates:
## mean in group 1 mean in group 2
## 19.73 26.10
```

**P-Value:** $1.9064 \times 10^{-5}$ is less than $\alpha=0.05$ ($\alpha$ for confidence interval of 95%), hence null hypothesis is **Rejected**, hence Hypothesis #2b is **Failed to Reject**.

Hypothesis 2 is **Failed to Reject**, based on above two evaluations.

---

**Hypothesis #3:** For combined impact of control variable “supp” and “dose”, “OJ” has higher impact on target variable “len” for “dose” 0.5 & 1 :

**3a:** “OJ” has higher impact for dose 0.5

```
Ga = t.test(len ~ supp, paired = F, var.equal = F, data = ToothGrowth[ToothGrowth$dose == 0.5],)
print(Ga)
```

```
## Welch Two Sample t-test
## data:  len by supp
## t = -3.078, df = 18.2, p-value = 0.004463
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.289 -1.724
## sample estimates:
## mean in group OJ mean in group control
## 14.10 17.66
```

**P-Value:** $0.004463$ is less than $\alpha=0.05$ ($\alpha$ for confidence interval of 95%), hence null hypothesis is **Rejected**, hence Hypothesis #3a is **Failed to Reject**.
## data: len by supp
t = 3.17, df = 14.97, p-value = 0.006359
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
1.719 8.781
mean in group OJ mean in group VC
13.23 7.98

P-Value, 0.0064 is less than $\alpha=0.05$ ($\alpha$ for confidence interval of 95%), confidence interval, (1.7191, 8.7809) for the difference of the means of each group does not span 0, hence null hypothesis is Rejected, hence Hypothesis #3a is Failed to Reject.

3b: “OJ” has higher impact for dose 1

G = t.test(len~supp,paired=F,var.equal=F,data=ToothGrowth[ToothGrowth$dose==1,]); print(G)

## Welch Two Sample t-test
data: len by supp
t = 4.033, df = 15.36, p-value = 0.001038
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
2.802 9.058
mean in group OJ mean in group VC
22.70 16.77

P-Value, 0.001 is less than $\alpha=0.05$ ($\alpha$ for confidence interval of 95%), confidence interval, (2.8021, 9.0579) for the difference of the means of each group does not span 0, hence null hypothesis is Rejected, hence Hypothesis #3b is Failed to Reject.

Hypothesis 3 is Failed to Reject, based on above two evaluations.

Hypothesis #4 - For combined impact of control variables “supp” and “dose”, “OJ” and “VC” have same impact on target variable “len” for “dose” 2:

G = t.test(len~supp,paired=F,var.equal=F,data=ToothGrowth[ToothGrowth$dose==2,]); print(G)

## Welch Two Sample t-test
data: len by supp
t = -0.0461, df = 14.04, p-value = 0.9639
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.798 3.638
mean in group OJ mean in group VC
26.06 26.14

Hypothesis #4 - For combined impact of control variables “supp” and “dose”, “OJ” and “VC” have same impact on target varibale “len” for “dose” 2:
P-Value, 0.9639 is greater than $\alpha=0.05$ (\(\alpha\) for confidence interval of 95%), confidence interval, (-3.7981, 3.6381) for the difference of the means of each group spans 0, hence null hypothesis is Failed to Reject, hence Hypothesis #4 is Failed to Reject.

Conclusions & Assumptions :

Conclusions: Based on the above evaluation of the four hypothesis, following are the conclusions:

1. For impact of control variable “supp” only, there is no significant difference on target variable “len” for different values of “supp”.
2. For impact of control variable “dose” only, higher the dose, higher is the impact on target variable “len”.
3. For combined impact of control variables, there is significant difference on target variable “len” for different values of “supp” for “dose 0.5 and 1”. There is no significant difference for different values of “supp” for “dose 2”.

Assumptions:

1. A higher value of “len” indicates a higher impact.
2. Higher value of “dose” indicates increased dosages.
3. Data provided is independently distributed.
4. Data follows T distribution as the observations are limited.
5. Data is derived from samples representative of the population.
6. Variances are considered to be unequal.