Statistical Inference Course Project

Rustam Mansyrov

The principal objective of the project is to analyze "Tooth Growth" dataset from the R "datasets" package. The dataset represents the length of odontoblasts(teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C(0.5, 1 and 2 mg) having two delivery methods(orange juice and ascorbic acid). Therefore, the data comprises three features: length of the tooth, type of the supplement and dose level, respectively.

Exploratory Data Analysis

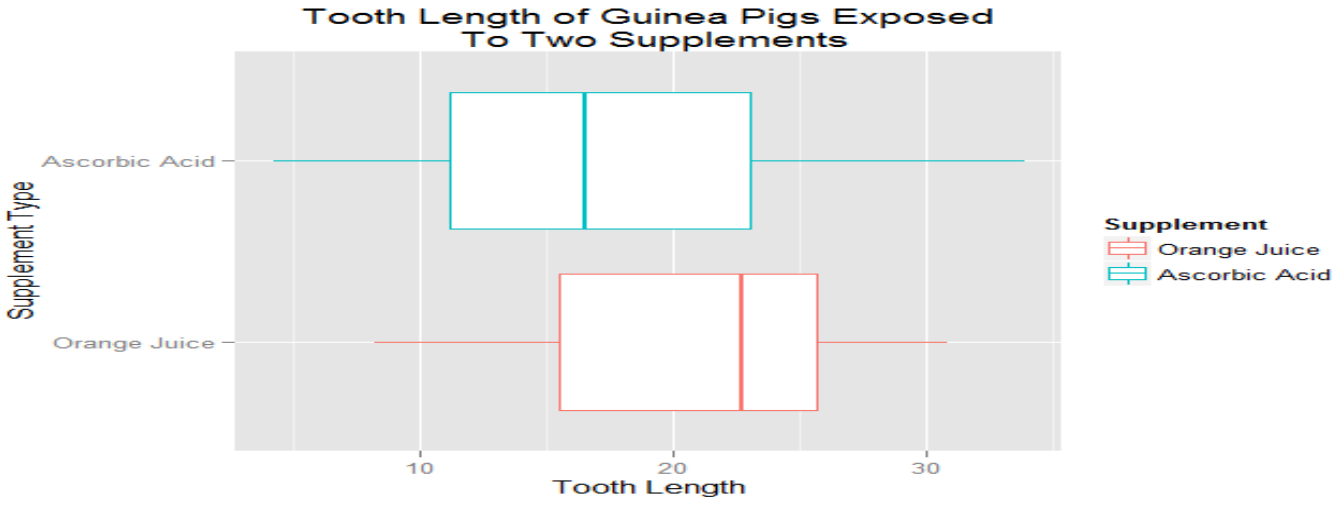
The brief summary and structure of the data are as follows:

```
   Tooth_length           Supplement      Dose
 Min.   : 4.20   Orange Juice :30   Min.   :0.500
 1st Qu.:13.07   Ascorbic Acid: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
```

It is seen from the summary that data has 60 rows and 3 colomns. The lengths of teeth are equally divided into two groups: the first - having the orange juice delivery method and second - ascorbic acid. The median and mean of the first feature is quite close - 19.25 and 18.81. More explicitly, let`s look at the means and standard deviations dividing the lengths into two delivery method groups:

```
        Supplement     Mean      Std
1  Orange Juice 20.66333 6.605561
2 Ascorbic Acid 16.96333 8.266029
```

Previously, without separating the data, the mean was 18.81. However, when the lengths were separted according to the supplement type that pigs were exposed to, means are a little bit far from 18.81. That is, the pigs exposed to orange juice have mean length of teeth of 20.66 and the rest pigs who were exposed to ascorbic acid have mean of 16.9. The difference in means is significant enought to be considered for the comparison. The standard deviations for the lengths are 6 and 8, which also implies that the group where pigs had ascorbic acid as supplement, has bigger variation. To be able to observe the stated explicitly, let us pay attention at the following boxplots below:



Apparently, the variation in the group where pigs were exposed to Ascorbic Acid is bigger, meaning that the length of tooth is more uncertain in that group.

Another way to compare the teeth growth of Guinea Pigs is by dividing them into dose amounts. Briefly, the summary of these grouping is as follows:

```
  dose_factor   Mean       Sd
1         0.5 10.605 4.499763
2           1 19.735 4.415436
3           2 26.100 3.774150
```

Obviously, it is not difficult to notice that those pigs exposed to 0.5 mg of dose amount has mean length of teeth of 10.65, which is much smaller than other means. Pigs with 1 mg dosage has mean of 19.7 and with 2 mg dosage - 26.1. So, we might conclude that growth of teeth is firmly associated with the dosage of supplement pigs get. As for standard deviations they are not significantly different, approximately close to each other.

Inferential Data Analysis

Hypothesis Testing & Confidence Intervals

Two possible questions of interest might be posed in order to investigate the inferential characteristics of teeth of Guinea Pigs.

- Is the mean for length of teeth of Guinea pigs being exposed to orange juice different from that of ascorbic acid?
- Are the means for length of teeth of Guinea pigs being exposed to different dose amounts of supplement different from each other?

To be able to answer the first question, we have to state the null and alternative hypotheses.

H0: mean for pigs with orange juice supplement is equal to mean for pigs with ascorbic acid supplement
H1: means are not equal to each other

Lower Bound of 95% C.I	Upper Bound of 95% C.I	P - value	Type I Error
-0.1710156	7.571	0.06063	0.05

Confidence interval of (-0.17, 7.57) tells us that the difference between two means contains 0 meaning that under 5% Type I Error level one has no sufficient evidence to reject null hypothesis. That is, means for pigs with orange juice as supplement and pigs exposed to ascorbic acid are equal. P-value of 0.06(greater than 0.05) also implies that we fail to reject null hypothesis.

As for second question, there will be three different hypotheses:

H01: mean length for pigs with 0.5 mg dosage is less than or equal to mean for pigs with 1 mg dosage

H11: mean length for pigs with 0.5 mg dosage is greater than mean for pigs with 1 mg dosage

Lower Bound of 95% C.I	Upper Bound of 95% C.I	P - value	Type I Error
-11.50	Inf	1	0.05

Confidence interval of (-11.50, Inf) contains zero meaning that again one has no sufficient evidence to reject null hypothesis and, in fact, looking at p-value it is very obvious that we fail to reject null hypothesis. The pigs exposed to 0.5 mg dosage tend to have smaller mean growth of teeth than that having 1 mg dosage.

H02: mean length for pigs with 0.5 mg dosage is less than or equal to mean for pigs with 2 mg dosage

H12: mean length for pigs with 0.5 mg dosage is greater than mean for pigs with 2 mg dosage

Lower Bound of 95% C.I	Upper Bound of 95% C.I	P - value	Type I Error
-17.71074	Inf	1	0.05

As for the previous case, the confidence interval of (-17.71, Inf) again contains zero meaning that one cannot reject null hypothesis and p - value is equal to 1 and it is much greater than 0.05, which tells that we fail to reject null hypothesis. The pigs exposed to 0.5 mg dosage tend to have smaller mean of growth of teeth than that having 2 mg dosage.

H03: mean length for pigs with 1 mg dosage is less than or equal to mean for pigs with 2 mg dosage

H13: mean length for pigs with 1 mg dosage is greater than mean for pigs with 2 mg dosage

Lower Bound of 95% C.I	Upper Bound of 95% C.I	P - value	Type I Error
-8.55613	Inf	1	0.05

Finally, the last interval (-8.55, Inf) has zero in it and means that we cannot reject null hypothesis. Again, p-value is 1, which is much greater than 0.05. Under 5% chance of Type I Error, we fail to reject null hypothesis and state that mean length of teeth of pigs exposed to 1 mg dosage is less that of 2mg dosage.

Assumptions

Speaking of assumptions, the main aspect which should be considered is underlying . All the data upon which analysis was made are assumed to be follow t-Student`s distribution. The second assumption made is that sample size is not sufficient enough to apply Central Limit Theorem(subjective opinion). And, finally, the population variances are assummed to be unequal to each other, except only the case when comparing means for length of teeth with 0.5 mg and 1 mg dosage.

Conlusion

As aforementioned above, there are two main questions of interest for the length of teeth of the pigs in the Inferential Part of the project. The first one is whether the means length of teeth for pigs having different supplement types are equal or not. This question is answered using hypothesis testing. Confidence interval and p-value yieled the result in favor of null hypothesis implying that the mean length of teeth is not affected by the type of supplement pigs get. Secondly, it is interested whether pigs having different dosage have the same mean length of teeth or not. It is found by hypothesis testing that the bigger the dosage, the bigger the mean length of teeth.

Appendix

R codes:

```
data(ToothGrowth)
data <- ToothGrowth
library(ggplot2)
library(dplyr)
names(data) <- c("Tooth_length", "Supplement", "Dose")
data <- mutate(data, Supplement = factor(Supplement, levels = c("OJ", "VC"), labels =
c("Orange Juice", "Ascorbic Acid")))
summary(data)
result1 <- group_by(data, Supplement) %>%
        summarize(Mean = mean(Tooth_length), Std = sd(Tooth_length))
result1
g <- ggplot(data, aes(y = Tooth_length, x = factor(Supplement), color = Supplement))
g <- g + geom_boxplot() + coord_flip() + labs(y = "Tooth Length", x = "Supplement Type")
+ ggtitle("Tooth Length of Guinea Pigs Exposed\n To Two Supplements")
print(g)
factor_dose <- mutate(data, dose_factor = factor(Dose))
result2 <- group_by(factor_dose, dose_factor) %>%
        summarize(Mean = mean(Tooth_length), Sd = sd(Tooth_length))
result2
x <- filter(data, Supplement == "Orange Juice")[,1]
y <- filter(data, Supplement != "Orange Juice")[,1]
t.test(x, y, alternative = "two.sided", var.equal = FALSE, paired = FALSE)
x <- filter(factor_dose, dose_factor == 0.5)[,1]
y <- filter(factor_dose, dose_factor == 1)[,1]
t.test(x, y, alternative = "greater", var.equal = TRUE, paired = FALSE)
x <- filter(factor_dose, dose_factor == 0.5)[,1]
y <- filter(factor_dose, dose_factor == 2)[,1]
t.test(x, y, alternative = "greater", var.equal = FALSE, paired = FALSE)
x <- filter(factor_dose, dose_factor == 1)[,1]
y <- filter(factor_dose, dose_factor == 2)[,1]
t.test(x, y, alternative = "greater", var.equal = FALSE, paired = FALSE)
```