Statistical Inference Course Project Part 2

Sarthak

```
# loading the required libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Basic Inferential Data Analysis

1. Overview

In this analysis, the ToothGrowth data will be analyzed. This shows the effect of vitamin C on teeth-growth in guinea pigs. Two vitamin C supplements are used, with varying dose levels.

2. Loading the Dataset and basic Exploratory Data Analysis

The ToothGrowth data set has to be loaded from the `datasets` package in R.

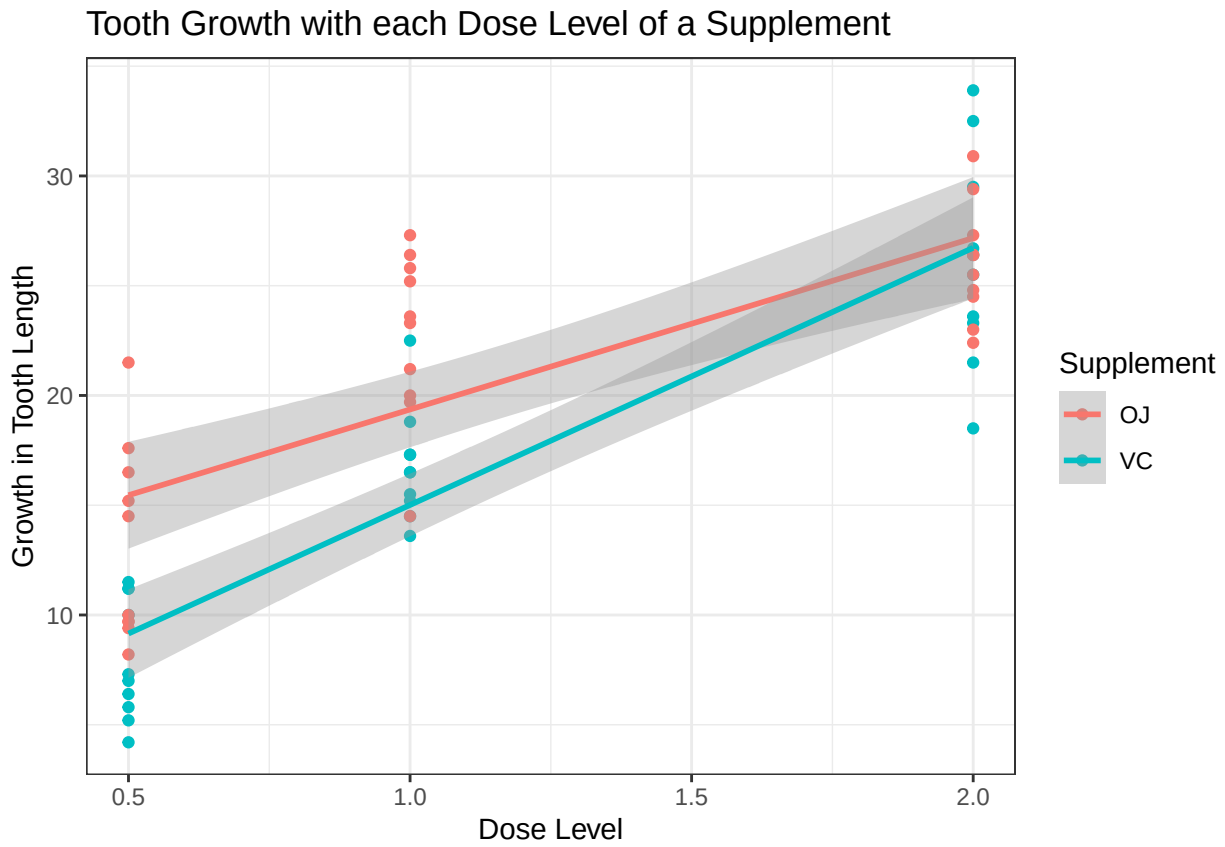
```
# loading the datasets package
library(datasets)

# loading the ToothGrowth data set
data("ToothGrowth")
```

Next, a plot will be created showing the trends in tooth-growth with respect to each supplement

```
# creating a point plot of the variation of teeth growth with respect to each dose level
# of a supplement, and fitting a linear line over those points
qplot(x = dose, y = len, data = ToothGrowth, color = supp, geom = "point",
      main = "Tooth Growth with each Dose Level of a Supplement",
      xlab = "Dose Level", ylab = "Growth in Tooth Length") +
geom_smooth(method = "lm") + theme_bw() + labs(colour = "Supplement")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From the plot, it is clear that the tooth length increases with increasing levels of dose of both supplements, and the *OJ* supplement has higher tooth length growths than the *VC* supplement. The growth in tooth lengths for the two supplements merge at dose levels of 2 mg/ml, whereas there is some difference in growth levels in the lower doses for the two supplements. Also, the growth rate between different dose levels of *VC* is higher than that of *OJ*.

3. Basic Data Summary

- **A summary of the whole dataset has been shown below**

```
# generating a summary of all variables in the dataset
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
```

```
# mean tooth growth length
mean(ToothGrowth$len)
```

```
## [1] 18.81333
```

```
# standard deviation of the length of tooth growth
sd(ToothGrowth$len)
```

```
## [1] 7.649315
```

- **Next, the growth in length data is summarized for the *OJ* supplement**

```
# summarizing the growth in length data for the 'OJ' supplement
summary(ToothGrowth[ToothGrowth$supp == "OJ",]$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20   15.53   22.70   20.66   25.73   30.90
```

```
# mean tooth growth length with the 'OJ' supplement
mean(ToothGrowth[ToothGrowth$supp == "OJ",]$len)
```

```
## [1] 20.66333
```

```
# standard deviation of the length of tooth growth with the 'OJ' supplement
sd(ToothGrowth[ToothGrowth$supp == "OJ",]$len)
```

```
## [1] 6.605561
```

- **Finally, the growth in length data is summarized for the *VC* supplement**

```
# summarizing the growth in length data for the 'VC' supplement
summary(ToothGrowth[ToothGrowth$supp == "VC",]$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   11.20   16.50   16.96   23.10   33.90
```

```
# mean tooth growth length with the 'VC' supplement
mean(ToothGrowth[ToothGrowth$supp == "VC",]$len)
```

```
## [1] 16.96333
```

```
# standard deviation of the length of tooth growth with the 'VC' supplement
sd(ToothGrowth[ToothGrowth$supp == "VC",]$len)
```

```
## [1] 8.266029
```

4. Performing Hypothesis Tests:

In this section, the growth in tooth length with respect to each supplement and each dose level will be compared, by using hypothesis tests.

First, the hypothesis of the difference in means of teeth growth under each supplement will be tested, for all doses. Two-sample t-test will be performed, with the null hypothesis being there is no difference in means (*which means a two-sided t test is performed*). (*Also, since 2 random samples of growth in tooth length, for supplements* OJ *and* VC*, are compared, we are performing a two-sample t test*)

```
# performing a two-sided, two-sample t-test comparing samples for supplements
# 'OJ' and 'VC'
t.test(len ~ supp, data = ToothGrowth, paired = FALSE, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

From the t-test, it is clear that the p-value is slightly above 0.05, which means that the null hypothesis of no difference between means cannot be rejected.

Next, lets test the hypothesis of the difference in means of teeth growth length under each supplement for each dose level. Just as before, a two-sample, two-sided t-test will be performed, along with the same null hypothesis.

```
# performing a two-sided, two-sample t-test comparing samples for supplements
# 'OJ' and 'VC', for the dose level of 0.5 mg/ml
t.test(len ~ supp, data = ToothGrowth[ToothGrowth$dose == 0.5,],
       paired = FALSE, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

From the t-test, it is clear that the p-value is way below 0.05, which means that the null hypothesis is rejected, and there is a significant difference between the means of the samples for the dose level of 0.5 mg/ml.

```
# performing a two-sided, two-sample t-test comparing samples for supplements
# 'OJ' and 'VC', for the dose level of 1.0 mg/ml
t.test(len ~ supp, data = ToothGrowth[ToothGrowth$dose == 1.0,],
       paired = FALSE, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

From the t-test, it is clear that the p-value is way below 0.05, which means that the null hypothesis is rejected, and there is a significant difference between the means of the samples for the dose level of 1.0 mg/ml.

```
# performing a two-sided, two-sample t-test comparing samples for supplements
# 'OJ' and 'VC', for the dose level of 2.0 mg/ml
t.test(len ~ supp, data = ToothGrowth[ToothGrowth$dose == 2.0,],
       paired = FALSE, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

From the t-test, it is clear that the p-value is way above 0.05, which means that the null hypothesis cannot be rejected, and there isn't any significant difference between the means of the samples for the dose level of 2.0 mg/ml.

5. Assumptions:

The assumptions used in the hypothesis tests using two-sided, two-sample t-test are:
a) The data approximately follows the normal distribution.
b) The variances of the two samples are not the same.
c) The two samples are independent.
d) The samples are random.