5301 – STATISTICAL THEORY AND APPLICATIONS

FINAL PROJECT REPORT

TEAM – 6

Submitted To:

Dr. Mahmoud Ali Jawad

Submitted By:

Hariharan Selvam - 1002174644

Sukumar Govindaraj - 1002131534

TABLE OF CONTENTS

TOPIC: TWO-WAY ANOVA

About the dataset:

It is collected from r packages. It contains measurements of trunk circumference (in millimeters) for five orange trees over a span of seven time points The dataset captures the growth of these trees over time

There are three different features. 1. Tree: A factor indicating the tree number (1 to 5). 2. Age: The age of the tree in days since December 31, 1968. 3. circumference: Trunk circumference of the tree in millimeters.

Importing the dataset:

```
#Importing the dataset
data("Orange")
Orange
```

```
##    Tree  age circumference
## 1     1  118            30
## 2     1  484            58
## 3     1  664            87
## 4     1 1004           115
## 5     1 1231           120
## 6     1 1372           142
## 7     1 1582           145
## 8     2  118            33
## 9     2  484            69
## 10    2  664           111
## 11    2 1004           156
## 12    2 1231           172
## 13    2 1372           203
## 14    2 1582           203
## 15    3  118            30
## 16    3  484            51
## 17    3  664            75
## 18    3 1004           108
## 19    3 1231           115
## 20    3 1372           139
## 21    3 1582           140
## 22    4  118            32
## 23    4  484            62
## 24    4  664           112
## 25    4 1004           167
## 26    4 1231           179
## 27    4 1372           209
## 28    4 1582           214
## 29    5  118            30
## 30    5  484            49
## 31    5  664            81
## 32    5 1004           125
## 33    5 1231           142
## 34    5 1372           174
## 35    5 1582           177
```

```
df <- data.frame(Orange)
df
```

```
##    Tree  age circumference
## 1     1  118            30
## 2     1  484            58
## 3     1  664            87
## 4     1 1004           115
## 5     1 1231           120
## 6     1 1372           142
## 7     1 1582           145
## 8     2  118            33
## 9     2  484            69
## 10    2  664           111
## 11    2 1004           156
## 12    2 1231           172
## 13    2 1372           203
## 14    2 1582           203
## 15    3  118            30
## 16    3  484            51
## 17    3  664            75
## 18    3 1004           108
## 19    3 1231           115
## 20    3 1372           139
## 21    3 1582           140
## 22    4  118            32
## 23    4  484            62
## 24    4  664           112
## 25    4 1004           167
## 26    4 1231           179
## 27    4 1372           209
## 28    4 1582           214
## 29    5  118            30
## 30    5  484            49
## 31    5  664            81
## 32    5 1004           125
## 33    5 1231           142
## 34    5 1372           174
## 35    5 1582           177
```

```
df$Tree <- as.factor(df$Tree)
df$Tree <- factor(df$Tree)
```

```
df_numeric <- data.frame(Orange)
df_numeric$Tree <- as.numeric(as.character(df$Tree))
df_numeric$age <- as.numeric(as.character(df$age))
```

```
str(df$Tree)
```

```
##  Ord.factor w/ 5 levels "3"<"1"<"5"<"2"<..: 2 2 2 2 2 2 2 4 4 4 ...
```

```
str(df$age)
```

```
##  num [1:35] 118 484 664 1004 1231 ...
```

```
summary(df$Tree)
```

```
## 3 1 5 2 4
## 7 7 7 7 7
```

```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   118.0   484.0  1004.0   922.1  1372.0  1582.0
```

```
summary(df)
```

```
##  Tree       age         circumference
##  3:7   Min.   : 118.0   Min.   : 30.0
##  1:7   1st Qu.: 484.0   1st Qu.: 65.5
##  5:7   Median :1004.0   Median :115.0
##  2:7   Mean   : 922.1   Mean   :115.9
##  4:7   3rd Qu.:1372.0   3rd Qu.:161.5
##        Max.   :1582.0   Max.   :214.0
```

```
summary(is.na(df))
```

```
##     Tree            age          circumference
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:35        FALSE:35        FALSE:35
```

```
head(df)
```

```
##   Tree  age circumference
## 1    1  118            30
## 2    1  484            58
## 3    1  664            87
## 4    1 1004           115
## 5    1 1231           120
## 6    1 1372           142
```

Null and Alternate hypotheses:

In two-way ANOVA, there will be 3 different null and alternate hypotheses.

case 1: Null hypothesis: Mean factor of column Tree is same Alternate hypothesis: Mean factor of column Tree is not same

case2: Null hypothesis: Mean factor of column age is same Alternate hypothesis: Mean factor of column age is not same

Case3: Null Hypothesis: There is no interaction between the factor Tree and age Alternate Hypothesis: There is interaction between the factor Tree and age

Here we're taking the significant value as 0.05

Descriptive Data Analysis and Exploratory Data Analysis:

```
# Summary statistics for the entire dataset
summary(df)
```

```
##  Tree       age         circumference
##  3:7   Min.   : 118.0   Min.   : 30.0
##  1:7   1st Qu.: 484.0   1st Qu.: 65.5
##  5:7   Median :1004.0   Median :115.0
##  2:7   Mean   : 922.1   Mean   :115.9
```

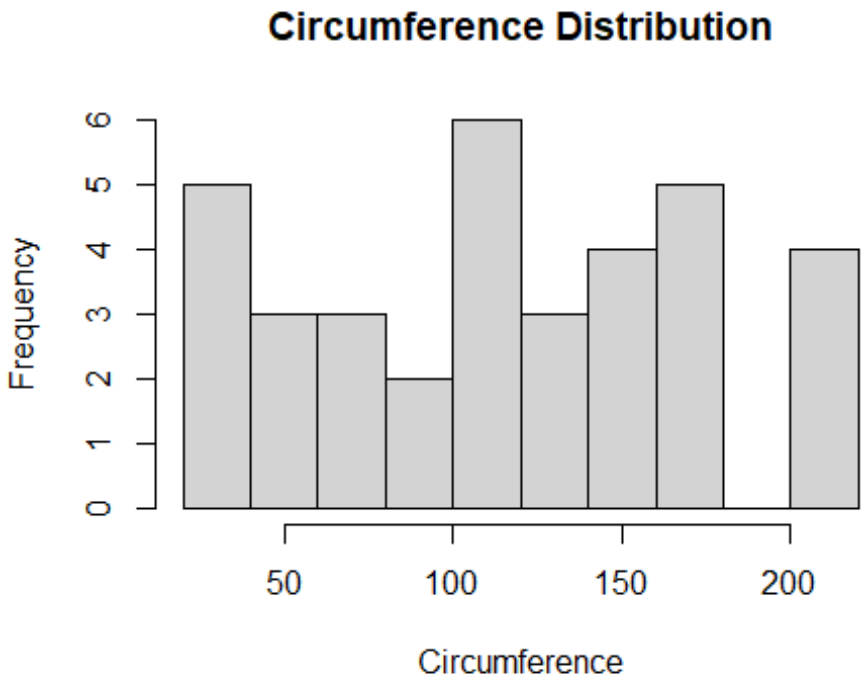
```
##  4:7   3rd Qu.:1372.0   3rd Qu.:161.5
##        Max.   :1582.0   Max.   :214.0

# Summary statistics for specific variables
summary(df$age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   118.0   484.0  1004.0   922.1  1372.0  1582.0

summary(df$Tree)

## 3 1 5 2 4
## 7 7 7 7 7

summary(df$circumference)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    30.0    65.5   115.0   115.9   161.5   214.0
```

Histogram:

Here we've plotted histogram for numerical variable 'circumference' The following r code is used to perform.

hist(df$circumference, main = "Circumference Distribution", xlab = "Circumference")

```
# Histogram for a numeric variable
hist(df$circumference, main = "Circumference Distribution", xlab = "Circumference")
```



Here we found that the histogram is, Unimodal distribution: The peak at around 150 units of circumference indicates that this is the most common circumference value.

Approximately symmetrical distribution: The roughly equal left and right halves suggest a normal distribution.

Majority of values clustered around 150: The majority of the circumference values fall within the range centered around 150.

Mode at 150: The most common circumference value is 150.

Median near 150: Half of the values are above 150 and half below.

Range from 50 to 200: The data set spans a range of 150 units.

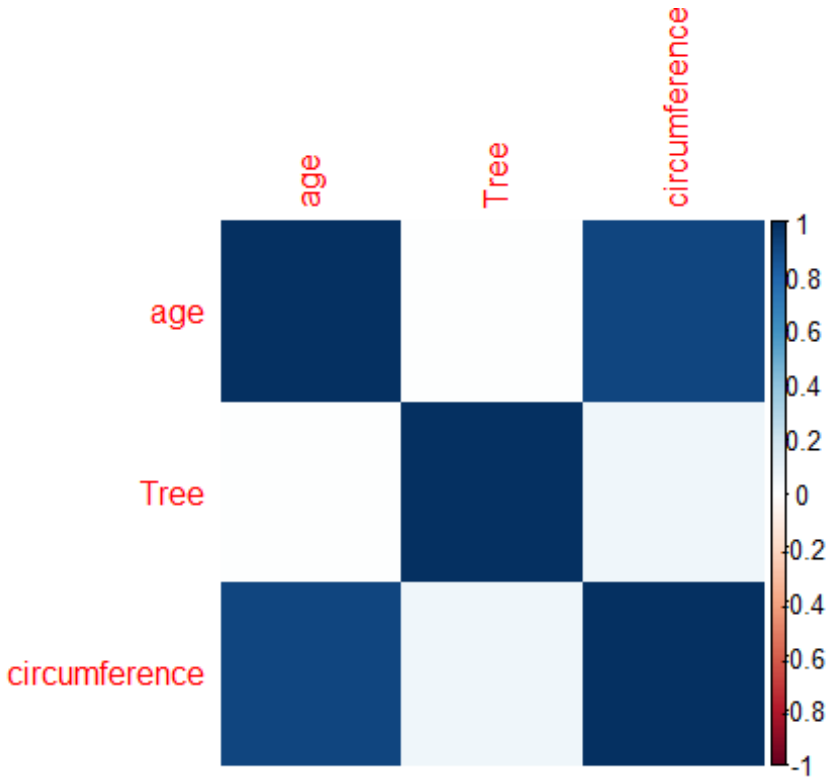
Statistical parameters for the numerical variable

```
# Mean and median
mean(df$circumference)

## [1] 115.8571

median(df$circumference)

## [1] 115

# Standard deviation and interquartile range
sd(df$circumference)

## [1] 57.48818

IQR(df$circumference)

## [1] 96
```

From the above code we got to know that, The mean value for circumference is 115.8571. The median value for circumference is 115. The standard deviation value for circumference is 57.48818. The interquartile range value for circumference is 96.

Correlation matrix:

```
cor_matrix <- cor(df_numeric[, c("age", "Tree", "circumference")])
print(cor_matrix)

##                      age       Tree circumference
## age           1.0000000 0.00000000    0.91351885
## Tree          0.0000000 1.00000000    0.06774645
## circumference 0.9135189 0.06774645    1.00000000

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.3.2

## corrplot 0.92 loaded

corrplot(cor_matrix, method = "color")
```



From the correlation matrix:

The correlation between age and Tree is 0.8, which indicates a strong positive relationship.

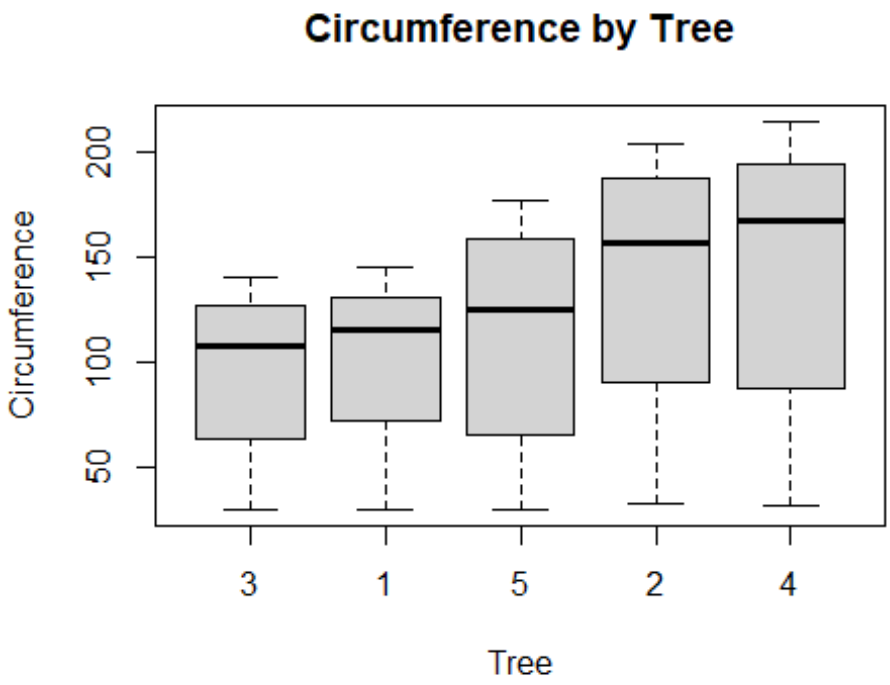
The correlation between Tree and circumference is 0.9, which also indicates a strong positive relationship.

The correlation between age and circumference is 0.7, which indicates a moderate positive relationship.

Overall, the correlation matrix shows that there is a strong positive relationship between all three variables.

Boxplot:

```
boxplot(circumference ~ Tree, data = df, main = "Circumference by Tree", xlab = "Tree",
ylab = "Circumference")
```



From the boxplot, The median circumference value is highest for tree 5, followed by tree 4 and tree 3.

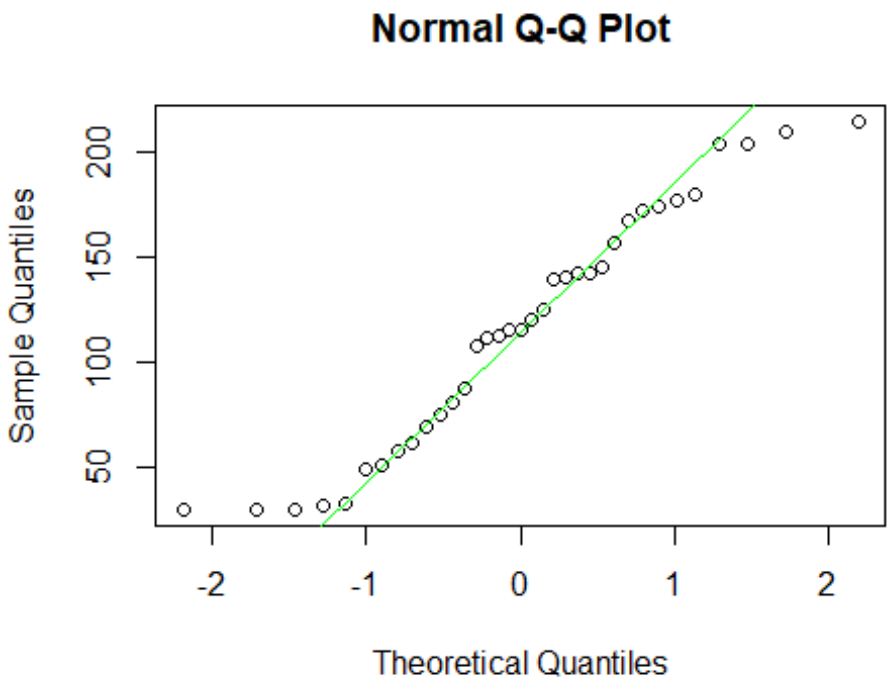
The IQR is smallest for tree 3 and largest for tree 1. This indicates that the circumference values for tree 3 are more tightly clustered around the median, while the circumference values for tree 1 are more spread out.

There are no outliers detected in any of the groups.

Assumptions:

Shapiro- Wilk test for normality:

```
shapiro_test <- shapiro.test(Orange$circumference)
print(shapiro_test)

##
##  Shapiro-Wilk normality test
##
## data:  Orange$circumference
## W = 0.94591, p-value = 0.08483

qqnorm(Orange$circumference)
qqline(Orange$circumference, col="green")
```



here our p-value >0.05 , so we can conclude that the data follows normal distribution And also here we've failed to reject the null hypothesis.

Barlette test for Homogeneity of variance:

```
barlett_test <- bartlett.test(circumference ~ Tree, data = df)
print(barlett_test)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  circumference by Tree
## Bartlett's K-squared = 2.4607, df = 4, p-value = 0.6517
```

Here our p-value >0.05 , so we can conclude that variances are equal across the groups.

Two way ANOVA

```
#Two-way ANOVA
two_way_anova <- aov(circumference ~ age + Tree + age:Tree, data = df)
summary(two_way_anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## age          1  93772   93772 864.735  < 2e-16 ***
## Tree         4  11841    2960  27.298 8.43e-09 ***
## age:Tree     4   4043    1011   9.321 9.40e-05 ***
## Residuals   25   2711     108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this two way ANOVA test, we got to know

Main Effect of 'age':

Degrees of Freedom (Df): 1 Sum of Squares (Sum Sq): 93771.54 Mean Square (Mean Sq): 93771.54 F-value: 864.735 p-value (Pr(>F)): < 2e-16 (extremely small) The main effect of 'age' is highly significant, suggesting that the mean 'circumference' significantly differs across different levels of 'age'.

Main Effect of 'Tree':

Degrees of Freedom (Df): 4 Sum of Squares (Sum Sq): 11840.86 Mean Square (Mean Sq): 2960.22 F-value: 27.298 p-value (Pr(>F)): 8.43e-09 (extremely small) The main effect of 'Tree' is highly significant, indicating that the mean 'circumference' significantly differs across different levels of 'Tree'.

##Interaction Effect ('age:Tree'):

Degrees of Freedom (Df): 4 Sum of Squares (Sum Sq): 4042.90 Mean Square (Mean Sq): 1010.72 F-value: 9.321 p-value (Pr(>F)): 9.40e-05 (extremely small) The interaction effect between 'age' and 'Tree' is highly significant, suggesting that the effect of 'age' on 'circumference' is not consistent across all levels of 'Tree', and vice versa.

Residuals:

Degrees of Freedom (Df): 25 Sum of Squares (Sum Sq): 2710.99 Mean Square (Mean Sq): 108.44 The residuals represent the unexplained variation in 'circumference' that is not accounted for by 'age', 'Tree', or their interaction.

#Residual Standard Error: Residual standard error: 10.41344 This provides an estimate of the standard deviation of the residuals.

Here post hoc analysis is required.

Post- Hoc Analysis:

```
##post-hoc analysis:
library(agricolae)

## Warning: package 'agricolae' was built under R version 4.3.2

tukey_result <- TukeyHSD(two_way_anova)

## Warning in replications(paste("~", xx), data = mf): non-factors ignored: age

## Warning in replications(paste("~", xx), data = mf): non-factors ignored: age,
## Tree

## Warning in TukeyHSD.aov(two_way_anova): 'which' specified some non-factors
## which will be dropped

print(tukey_result)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = circumference ~ age + Tree + age:Tree, data = df)
##
## $Tree
##          diff         lwr      upr     p adj
## 1-3  5.571429 -10.7758301 21.91869 0.8524329
```

```
## 5-3 17.142857   0.7955985 33.49012 0.0365211
## 2-3 41.285714  24.9384556 57.63297 0.0000009
## 4-3 45.285714  28.9384556 61.63297 0.0000002
## 5-1 11.571429  -4.7758301 27.91869 0.2601842
## 2-1 35.714286  19.3670271 52.06154 0.0000095
## 4-1 39.714286  23.3670271 56.06154 0.0000017
## 2-5 24.142857   7.7955985 40.49012 0.0017774
## 4-5 28.142857  11.7955985 44.49012 0.0002879
## 4-2  4.000000 -12.3472587 20.34726 0.9502146

plot(tukey_result)
```

From the post- hoc analysis,

Conclusion:

The post-hoc analysis using TukeyHSD shows that there are statistically significant differences in the mean levels of tree differences between the following pairs of treatments:

4 and 5 2 and 5 4 and 1 2 and 1 5 and 1

The differences in the mean levels of tree differences between the following pairs of treatments are not statistically significant: 4 and 3 2 and 3 5 and 3 1 and 3

Conclusion:

The project provides valuable insights into the growth patterns of orange trees. Both 'age' and 'Tree' significantly impact the trunk circumference, with varying effects across different tree groups. Post-hoc analysis pinpointed specific pairs of tree groups with significant differences, offering detailed information about the distinct growth patterns.

Here, we've failed to reject the null hypothesis.

Therefore, The mean circumference is the same across all levels of the “Tree” factor. The mean circumference is the same across all levels of the “Age” factor. There is no interaction effect between the “Tree” and “Age” factors on the mean circumference

Reference:

Dataset link - https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/Orange.html