Final Assignment 2

Andrei Udriste

5/12/2021

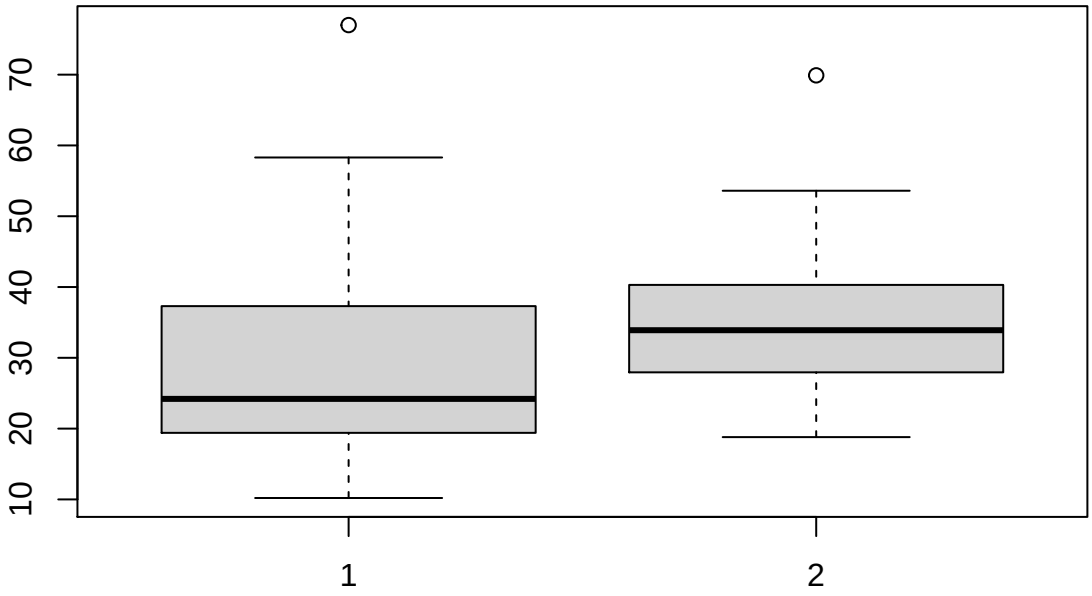
```
data = read.table("treeVolume.txt", header = TRUE)

beech_tree = data$volume[data$type == "beech"]
oak_tree = data$volume[data$type == "oak"]
```

```
length(beech_tree)
```

a)

```
## [1] 31
```

```
boxplot(beech_tree, oak_tree)
```




```
data$type = factor(data$type)
model = lm(data$volume ~ data$type)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: data$volume
##           Df  Sum Sq Mean Sq F value Pr(>F)
## data$type  1   379.5  379.52  1.8984 0.1736
## Residuals 57 11394.8  199.91
```

After creating a ANOVA model and running it we obtain that the p-value is equal with 0.1736, which is much more bigger than 0.05 this means that we do not reject the null hypotheses H0. Because the null hypotheses H0 has not been rejected the tree type does not have any influence on the volume of the tree.

```
cat("\n beech prediction =", predict(model, data.frame(type="beech"), type="response")[1], "\n")
```

```
## Warning: 'newdata' had 1 row but variables found have 59 rows
```

```
##
##  beech prediction = 30.17097
```

```
cat("mean of the volume for the beech trees", mean(beech_tree), "\n")
```

```
## mean of the volume for the beech trees 30.17097
```

```
cat("\n oak prediction =", predict(model, data.frame(type="oak"), type="response")[50], "\n")
```

```
## Warning: 'newdata' had 1 row but variables found have 59 rows
```

```
##
##  oak prediction = 35.25
```

```
cat("mean of the volume for the oak trees", mean(oak_tree))
```

```
## mean of the volume for the oak trees 35.25
```

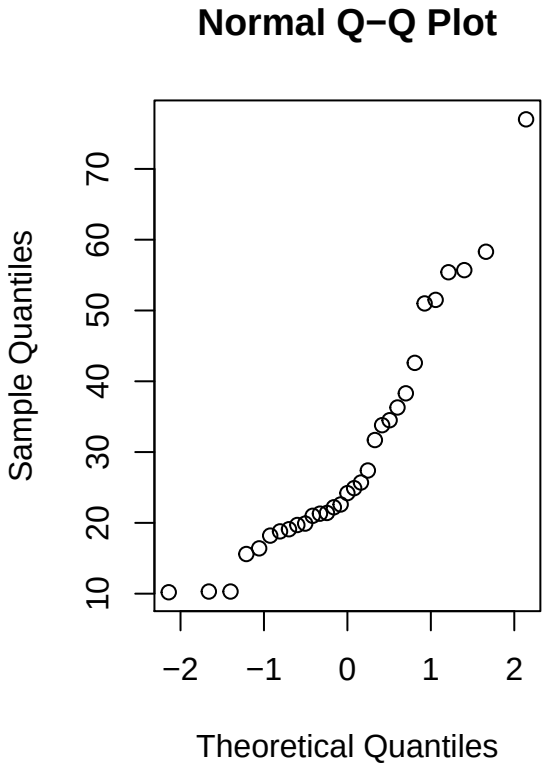
After prediction the volume for the beech tree and the oak tree using the function predict, we obtain the following values:

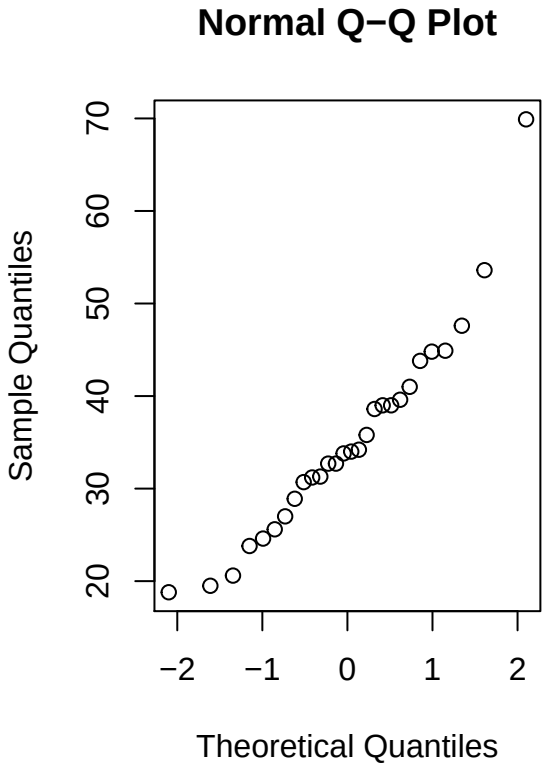
- beech tree volume = 30.1709
- oak tree volume = 35.25

One observation would be that the predicted value are the same as the mean of the volume for each tree type, this could be seen in the print statement from above.

b) If we want to use the t.test in this case we need to check whatever the data is from a normal distribution, because the t.test can be used only on data that comes from a normal distribution.

```
par(mfrow=c(1,2))
qqnorm(beech_tree)
qqnorm(oak_tree)
```





The first method to check whatever the data comes from a normal distribution would be to plot a QQ-plot using the data and see if it creates a straight line. If we observe the two QQ-plots for the volume of beech trees and oak trees, we can observe that the QQ-plot for the oak trees has somewhat of a straight line, but the QQ-plot for the beech trees is not straight. This would indicate that the data does not come from a normal distribution so we can not preform a t.test on this type of data. Another way to extra check if the data comes from a normal distribution is to perform a Shapiro test, and to obtain a p-value bigger than 0.05.

```
shapiro.test(beech_tree)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beech_tree
## W = 0.88757, p-value = 0.003579
```

```
shapiro.test(oak_tree)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  oak_tree
## W = 0.9349, p-value = 0.08199
```

After performing the Shapiro test on both datasets we can observe the same result as the one obtained from the QQ-plots. The beech tree data has a p-value of $0.003579 < 0.05$ this would mean that we reject the null hypotheses H0, so the data does not come from a normal distribution. If we look at the p-value for the oak trees we obtain a value of $0.0819 > 0.05$ this means that we do not reject the null hypotheses H0, so the data comes from a normal distribution, but the value of the p-value is not that bigger than 0.05 so is not that wise to work with this assumption.

Because the data does not come from a normal distribution is it much more appropriate t use test that do not rely on normality like Mann-Whitney and Kolmogorov-Smirnov. Those test rely on ranks so there is no

need for normality.

```
wilcox.test(beech_tree, oak_tree)
```

```
## Warning in wilcox.test.default(beech_tree, oak_tree): cannot compute exact p-
## value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  beech_tree and oak_tree
## W = 300, p-value = 0.0427
## alternative hypothesis: true location shift is not equal to 0
```

After performing the Mann-Whitney we obtain a p-value of $0.0427 < 0.05$, this means that we have to reject the null hypotheses H0, so there is no correlation between the volume of beech and the volume of oak tree types, but again the value of the p-value is relatively close to the threshold of 0.05 so is not wise to completely trust the test.

```
ks.test(beech_tree, oak_tree)
```

```
## Warning in ks.test(beech_tree, oak_tree): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  beech_tree and oak_tree
## D = 0.37673, p-value = 0.03072
## alternative hypothesis: two-sided
```

After performing the Kolmogorov-Smirnov we obtain a p-value of $0.03072 < 0.05$, this means that we have to reject the null hypotheses H0, so there is no correlation between the volume of beech and the volume of oak tree types. In this case the p-value is much more further than the Mann-Whitney so there is much more confidence in reject the null hypotheses H0.

On thing that is common in both test is the fact that the we obtain values similar to the ANOVA model created at point a), but in that case we did not reject the null hypotheses because the p-value was bigger than 0.05, but in this case we had to reject the null hypotheses for both tests.

```
data$type = factor(data$type)
model_c = lm(data$volume ~ data$type + data$diameter + data$height)
anova(model_c)
```

c)

```
## Analysis of Variance Table
##
## Response: data$volume
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## data$type      1   379.5   379.5  36.089 1.565e-07 ***
## data$diameter  1 10492.3 10492.3 997.728 < 2.2e-16 ***
## data$height    1   324.2   324.2  30.825 8.416e-07 ***
## Residuals     55   578.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we perform an ANOVA test using also the diameter and the height we obtain that the p-value for the type has now change and is equal with 1.56e-07 < 0.05, this means that now we reject the null hypotheses H0, so

there is an influence between the tree type and the volume of the tree. This is completely different from point a where we found that there the type of the tree does not have any influence on the volume.

```
predicted_values = predict(model_c, data.frame(type="beech", diameter=mean(data$diameter), height=mean(
```

```
## Warning: 'newdata' had 1 row but variables found have 59 rows
```

```
beech_pred = 0
for(i in 1:length(beech_tree)){
  beech_pred = beech_pred + predicted_values[i]
}

oak_pred = 0
for(i in length(beech_tree):length(predicted_values)){
  oak_pred = oak_pred + predicted_values[i]
}

cat("\n beech prediction =", beech_pred / length(beech_tree))
```

```
##
##  beech prediction = 30.17097
```

```
cat("\n oak prediction =", oak_pred / length(oak_tree))
```

```
##
##  oak prediction = 37.72496
```

The predicted value for the two tree types are:

- beech tree volume = 30.17097 30.1709
- oak tree volume = 37.72496 35.25

In the case of the beech trees we can see that there is no difference in the prediction, we obtain the same value as the prediction made at point a, 30.1709. In the case of the oak trees we obtain a different value from the prediction made at point a. The prediction made at point a was 35.25, while the prediction made with the new model is 37.724, which is with two point higher than the old prediction. This means that the height and the diameter of the tree does influence the model.

```
beech_diam = data$diameter[data$type == "beech"]
oak_diam = data$diameter[data$type == "oak"]
```

```
model_d_d = lm(data$volume ~ data$diameter)
anova(model_d_d)
```

d)

```
## Analysis of Variance Table
##
## Response: data$volume
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## data$diameter  1 10826.5 10826.5   651.1 < 2.2e-16 ***
## Residuals     57   947.8    16.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we create an ANOVA model only using the volume on the diameter of the tree we can observe the influence of the diameter over the volume. After running the model we obtain a p-value of 2.2e-16 < 0.05, this means

that we reject the null hypotheses H0, so the diameter of the tree has very big influence over the volume of the tree.

```
model_d_b = lm(beech_tree ~ beech_diam)
anova(model_d_b)
```

```
## Analysis of Variance Table
##
## Response: beech_tree
##            Df Sum Sq Mean Sq F value    Pr(>F)
## beech_diam  1 7581.8  7581.8  419.36 < 2.2e-16 ***
## Residuals  29  524.3    18.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we run the ANOVA model but this time using only the data from the beech trees we obtain a p-value = 2.2e-16 $<$ 0.05, so this means that we reject the null hypotheses HO, so the diameter has a very big influence over the volume of the beech trees.

```
model_d_o = lm(oak_tree ~ oak_diam)
anova(model_d_o)
```

```
## Analysis of Variance Table
##
## Response: oak_tree
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## oak_diam   1 2920.02 2920.02  205.91 7.238e-14 ***
## Residuals 26  368.71   14.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

But if we run the same test but in this case we use the data for the oak trees we obtain a p-value equal with 7.238e-14 $<$ 0.05 so we still reject the null hypotheses H0, but we can see that this p-value is higher than the one obtained from the beech trees so the diameter doesn't have such a big influence over the volume as in the case of the beech trees.

```
beech_height = data$height[data$type == "beech"]
oak_height = data$height[data$type == "oak"]
```

```
model_d_h = lm(data$volume ~ data$height)
anova(model_d_h)
```

```
## Analysis of Variance Table
##
## Response: data$volume
##             Df Sum Sq Mean Sq F value   Pr(>F)
## data$height  1 2187.6 2187.63  13.007 0.000654 ***
## Residuals   57 9586.7  168.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can perform the ANOVA test the same way for the height of the trees an we obtain a p-value = 0.000654 $<$ 0.05 this means that we reject the null hypotheses H0, so the height of the tree has an influence over the volume of the tree, but we can see that the p-value in this case is much more higher than the p-value obtained when we used the diameter of the tree. This means that the diameter of the tree has a bigger influence over the volume of the tree than the height.

```r
model_d_b_h = lm(beech_tree ~ beech_height)
anova(model_d_b_h)
```

```
## Analysis of Variance Table
##
## Response: beech_tree
##              Df Sum Sq Mean Sq F value    Pr(>F)
## beech_height  1 2901.2 2901.19  16.165 0.0003784 ***
## Residuals    29 5204.9  179.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we perform an ANOVA test using only the data from the beech trees we obtain a p-value $= 0.000378 < 0.05$ so we reject the null hypotheses H0, this means that the height of the tree has an influence over the volume of the beech trees.

```r
model_d_o_h = lm(oak_tree ~ oak_height)
anova(model_d_o_h)
```

```
## Analysis of Variance Table
##
## Response: oak_tree
##            Df Sum Sq Mean Sq F value Pr(>F)
## oak_height  1   80.3  80.255  0.6504 0.4273
## Residuals  26 3208.5 123.403
```

If we perform an ANOVA test using the oak data, we obtain a p-value of $0.427 > 0.05$, this means that we do not reject he null hypotheses H0, this means that there is no influence between the height of the oak trees and their volume. Which is the exact opposite case from the beech trees where the height of the trees had an influence over the volume.