Statistics Rapport

Study about oak trees

Table of Contents

[Dataset 3](#_Toc137649374)

[Questions 3](#_Toc137649375)

[Question 1: *Study and discuss the distribution of the variables Volume and Width. To do this, discuss appropriate graphical representations. Also, formally determine whether the data is normally distributed. If this is not the case, in what way does the data deviate from normally distributed data? Discuss.* 3](#_Toc137649376)

[Answer: 3](#_Toc137649377)

[Question 2: *Investigate whether there is a correlation between "thick acorns," which are oak trees whose acorn volume is at least 3 cm3, and the area in which the tree occurs. To do this, create a new variable called "thick acorn." Then, perform an appropriate test to determine if there is a significant correlation between thick acorns and the tree's geographic location.* 4](#_Toc137649378)

[Answer: 4](#_Toc137649379)

[Question 3: *Can you predict Height from log(Volume)? Answer this question thoroughly and as completely as possible.* 4](#_Toc137649380)

[Answer: 4](#_Toc137649381)

# Dataset

We make use of a custom dataset, founded in the eik.csv file.

Here we can find five variables:

* Boom = tree: the id of the tree
* Regio = region: the region the tree was measured at
* Grootte = width: the width of the area the tree grows in (100 km2)
* Volume = volume: the volume of the tree (cm3)
* Hoogte = height: the height of the tree (m)

Following columns are deleted from my dataset: {1, 3, 9, 11, 17}

Student number: 20213082

# Questions

## Question 1: *Study and discuss the distribution of the variables Volume and Width. To do this, discuss appropriate graphical representations. Also, formally determine whether the data is normally distributed. If this is not the case, in what way does the data deviate from normally distributed data? Discuss.*

### Answer:

We make use of histograms to check whether our data is normally distributed.

If we look at both variables, we can see that both are positive skewed graphs.

A picture containing text, screenshot, diagram, design

Description automatically generated A picture containing text, screenshot, diagram, multimedia software

Description automatically generated

The data deviates more to the right if we compare it to a normal distribution.

Now if we discuss about both histograms, we can conclude the following:

* We can see that there are more trees that grow in smaller areas.
* We can also see that more trees have a smaller volume.

## Question 2: *Investigate whether there is a correlation between "thick acorns," which are oak trees whose acorn volume is at least 3 cm3, and the area in which the tree occurs. To do this, create a new variable called "thick acorn." Then, perform an appropriate test to determine if there is a significant correlation between thick acorns and the tree's geographic location.*

### Answer:

In this question we look at the corelation between the thickness of the acorn and the region they appear at.

We make an extra variable called “dikke\_eikels”, it is one if it is bigger than 3 cubic centimetres and zero if not.

We make use of a chi-squared test to determine if there is a correlation between how thick the acorn is and what region it comes from. Our null hypothesis is that there is no correlation between how thick the acorns are and where they come from. We choose this as our null hypothesis because we want to prove that there is a correlation between these two variables.

After performing our chi-squared test we get the following results:   
 X-squared = 1.7702, df = 1, p-value = 0.1834

We can see that our p-value is higher than 0.05. This suggests that there is not enough evidence to conclude a significant association between "dikke eikels" (thick acorns) and the region where the tree occurs.

But if we look at bar plot, we can see that there is more thick acorns in Atlantic.

A screenshot of a computer

Description automatically generated with medium confidence

## Question 3: *Can you predict Height from log(Volume)? Answer this question thoroughly and as completely as possible.*

### Answer:

To determine if there is a relation between height and log(Volume), we make use of a scatter plot.

By doing so we can see if our data is a linear line, implying that there is a correlation between log(Volume) and height.

A screen shot of a graph

Description automatically generated with medium confidence

We can see that they don’t have any correlations with each other.

If we look at the plot more in detail:

* We can see that there is a linear line forming. Some points are on the line.
* More outliers than linear line
* Line is slightly slanted to the top, so small correlation.

From these three points we mostly conclude that it’s not correlated with each other.

Let’s do another test, we can also calculate the correlation between these two with one function.  
After executing this function we can get a result of 0.1445543, we know that there is a correlation if and only if this value is near 1. When it’s positive near zero, we can conclude we can see small correlation but not enough to determine that there is a correlation.

Okay now we know there is not really a correlation but a small one, but what does the small correlation do for us. Let’s build a linear regression model and see what it gives us as results.

We will train our linear regression model with 70% of our data and 30% we use for testing. After doing that we get the following data “16.2951932 0.6998405”. This implies that at near zero, the height is 16.2951932 meters, and that for every one-unit change the height increases with approximately 0.7 meters.

Conclusion: Not much of a correlation between height and log(Volume). Small changes when the log(Volume) increases.