Elementary Statistics Assignment

Pim Van den Bosch[1]

[1]University of Antwerp

June 15, 2023

Abstract

Using the eik.csv dataset, the report focuses on understanding the distribution of the variables 'Volume' and 'Size', examining the correlation between large acorns and the area in which the tree is found, and investigating the predictive power of 'Volume' on 'Height'. Different statistical techniques, including graphical analysis, empirical distribution function, quantile function, Q-Q plot, Shapiro-Wilk test, chi-square test of independence, and ordinary least squares regression were employed to answer these research questions. The results indicated the non-normal distribution of both variables 'Volume' and 'Size', no significant association between the region and the presence of large acorns, and a limited capability of the logarithm of 'Volume' to predict 'Height'.

1 Introduction

This study utilizes a dataset that contains information about various oak species found across the United States. The dataset, named `eik.csv`, can be found at `https://github.com/Nabla7/intro_stats` (together with the jupyter notebook) and includes the following variables:

1. **Tree:** The sequence number of the considered tree species.
2. **Region:** The region where the tree is found, either 'Atlantic' or 'California'.
3. **Size:** The size of the area where the species is found, in units of 100 km^2.
4. **Volume:** The volume of the acorn, in cm^3.
5. **Height:** The height of the tree, in meters.

Further examination of these variables provides insights into the geographic distribution and physical characteristics of the oak species under consideration.

In this report we will consider the following three questions:

1. We discuss the distribution of the variables 'Volume' and 'Size'. To this end, we consider suitable graphical representations. We will also formally verify whether the data follows a normal distribution. If this is not the case, we investigate how the data deviates from a normal distribution.

2. Is there a correlation between large acorns, defined as oaks whose acorn volume is at least 3 cm^3, and the area in which the tree is found? To answer this, we create a new variable 'large acorn' and then carry out the appropriate tests.

3. We investigate whether we can predict the 'Height' from the logarithm of the 'Volume'

2 Methods

To address the first research question concerning the distribution of the 'Volume' and 'Size' variables, we employ both graphical and formal statistical methods.

2.1 Graphical Analysis

For a comprehensive understanding of the data, we first generate histograms and boxplots. Histograms provide insight into the data distribution and allow us to visually assess the central tendency, variability, and skewness. They can also highlight any obvious outliers or gaps in the data. Boxplots, on the other hand, offer a more succinct statistical summary of the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values.

2.2 Empirical Distribution Function (EDF)

We utilize the Empirical Distribution Function (EDF) to assess the cumulative distribution of the data, defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{1}$$

where I is the indicator function, equal to 1 if $X_i \leq x$ and 0 otherwise, n is the number of data points, and X_i is the i-th data point.

2.3 Quantile Function and Q-Q plot

Next, we investigate the quantile function, also known as the inverse of the cumulative distribution function (CDF). This function tells us the value below which a given percentage of the data falls.

A Q-Q (Quantile-Quantile) plot is then employed to compare the quantiles of our data's distribution to the quantiles of a standard normal distribution. If our data follows a normal distribution, the points in the Q-Q plot should approximately lie on the 45-degree reference line.

2.4 Formal Normality Test: Shapiro-Wilk Test

After graphical assessment, we formally test for normality of the data using the Shapiro-Wilk test. The null hypothesis H_0 states that the data follows a normal distribution. The test statistic is given as:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

where $x_{(i)}$ are the ordered sample values, a_i are the constants generated from the covariances, variances, and means of the sample size n from a normal distribution, and $\bar{x}$ is the sample mean.

2.5 Transformations

To further assess normality, common transformations such as logarithmic and square root transformations are applied to the data. Histograms, Q-Q plots, and EDFs for these transformed variables are examined, and the Shapiro-Wilk test is performed to formally assess the normality of the transformed data.

2.6 Chi-Square Test of Independence

For the second research question, we explore the relationship between large acorns and the areas where the trees are found. To determine whether the occurrence of large acorns is dependent on the area, we construct a contingency table and perform a Chi-square test of independence.

The Chi-square test of independence compares the observed frequencies in each category of a contingency table with the frequencies we would expect to see if the variables were independent of each other. The test statistic is calculated as follows:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3}$$

where O_{ij} is the observed frequency in each cell and E_{ij} is the expected frequency under the assumption of independence.

The null hypothesis H_0 of the Chi-square test is that the two variables are independent. If the p-value is less than the chosen significance level (e.g., $\alpha = 0.05$), then we reject the null hypothesis and conclude that the variables are dependent.

2.7 Visual Representation of Proportions

To further visualize the relationship between large acorns and their regions, we normalize the contingency table to get the proportions of large acorns in each region. We then create a stacked bar chart of the proportions for each region.

2.8 Ordinary Least Squares Regression

The third research question concerns whether we can predict the 'Height' from the logarithm of the 'Volume'. This is a regression problem, and we address it using an Ordinary Least Squares (OLS) regression model.

The model assumes that the relationship between the predictor variable (in this case, 'log_Volume') and the response variable ('Height') can be approximated by a linear equation:

$$\text{Height} = \beta_0 + \beta_1 \times \text{log\_Volume} + \epsilon \tag{4}$$

where β_0 and β_1 are the parameters to be estimated and ϵ represents the residuals or error term. The OLS method finds the estimates of β_0 and β_1 that minimize the sum of the squared residuals, $\sum \epsilon^2$.

2.9 Model Evaluation

After fitting the OLS regression model, we examine the following statistics to evaluate the performance and validity of the model:

1. **R-squared**: This is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. It ranges from 0 to 1, where 1 indicates that the independent variables perfectly predict the dependent variable, and 0 indicates that the independent variables do not predict the dependent variable at all.

2. **Adjusted R-squared**: This is a modified version of R-squared that adjusts for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.

3. **F-statistic and corresponding p-value**: The F-statistic is used to test the overall significance of the model. The null hypothesis under the F-test is that the model with no independent variables fits the data as well as our model. If the p-value corresponding to the F-statistic is less than our significance level (e.g., 0.05), we can reject the null hypothesis and conclude that the model provides a better fit than the intercept-only model.

4. **T-statistics and corresponding p-values for each coefficient**: These are used to test the hypothesis that each coefficient is different from zero. If the p-value is less than the significance level, we can reject the null hypothesis and conclude that the predictor is making a significant contribution to the model.

2.10 Residual Analysis

The residuals of the model, defined as the difference between the observed and predicted values, are analyzed to validate the assumptions of the OLS regression:

1. **Linearity**: The relationship between predictors and the response variable is linear.
2. **Independence**: The residuals are independent.
3. **Homoscedasticity**: The variance of the residuals is constant.
4. **Normality**: The residuals are normally distributed.

We use various diagnostic plots to inspect these assumptions:

1. **Residual vs. Fitted Values Plot**: This is used to check the linearity and homoscedasticity assumptions. We expect to see no clear patterns and equal variance across the entire range of fitted values.
2. **Q-Q Plot of Residuals**: This is used to check the normality assumption. We expect the points to fall approximately along the 45-degree reference line.
3. **Histogram of Residuals**: This can also be used to check the normality assumption. We expect the histogram to resemble a bell-shaped curve.

We further perform the Shapiro-Wilk test on the residuals to formally test for normality. We also compute the Durbin-Watson statistic to test the independence assumption. A value close to 2 suggests that there is no autocorrelation in the residuals.

3 Results

3.1 Graphical Analysis

The distribution of both Volume' and Size' variables was examined through histograms and boxplots respectively in (Figures 1 and 2)

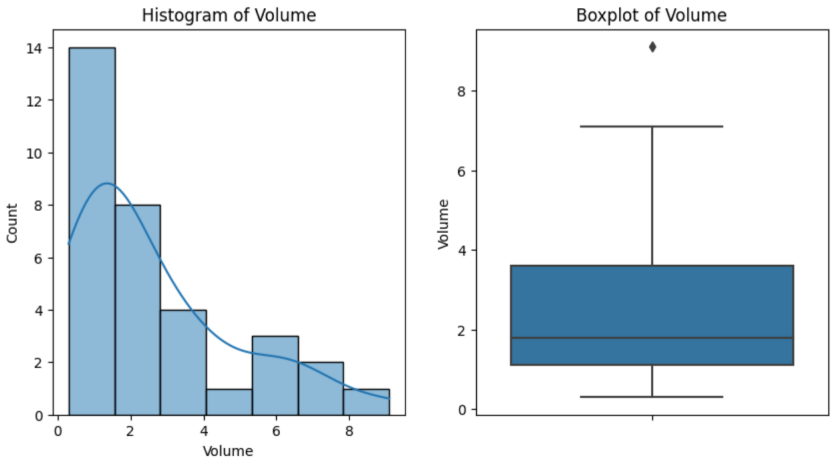

Figure 1: Histogram and Boxplot of 'Volume'

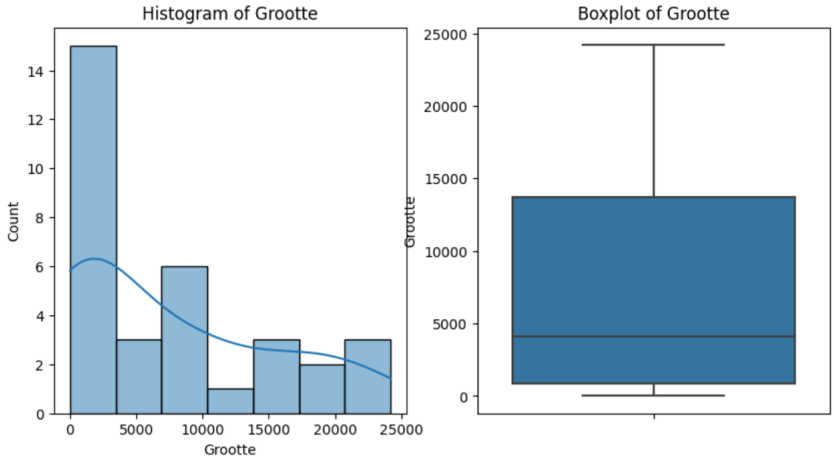

Figure 2: Histogram and Boxplot of 'Size'

3.2 Empirical Distribution Function (EDF)

The cumulative distribution of the data was observed with the EDF plots, we also plot the associated PDF on a scaled histogram. (Figures 3 and 4).

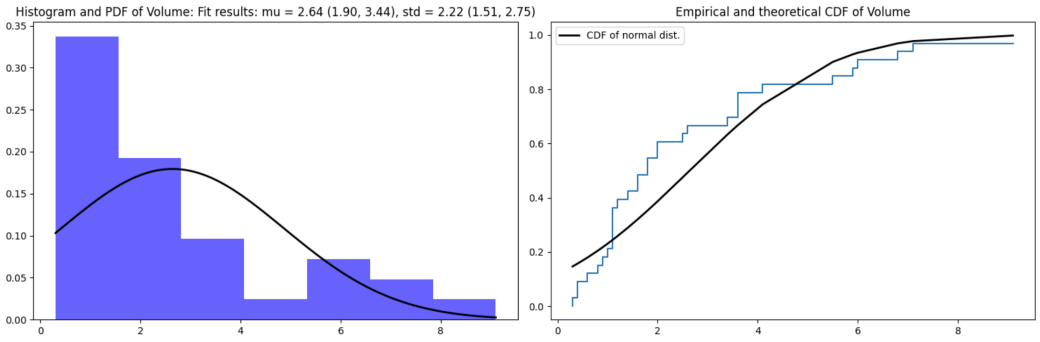

Figure 3: PDF and EDF of 'Volume'

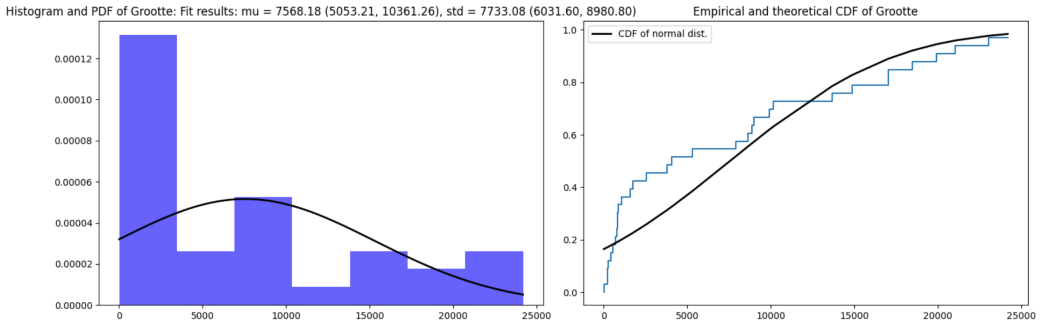

Figure 4: PDF and EDF of 'Size'

3.3 Quantile Function and Q-Q plot

The Q-Q plots were generated for both 'Volume' and 'Size' variables (Figures 5.

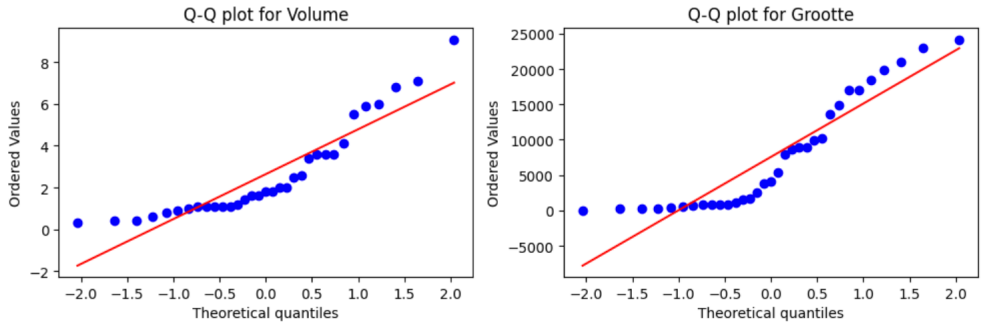

Figure 5: Q-Q plot for 'Volume' and 'Size'

3.4 Shapiro-Wilk Test

The results of the Shapiro-Wilk test are presented in Table 1, descriptive statistics for the log and square root transformations have also been provided.

Variable	Mean	Median	Skewness	Kurtosis	Shapiro-Wilk
Volume	2.64	1.8	1.29	0.93	W=0.84, p=0.0002
Grootte	7568.18	4082.0	0.79	-0.75	W=0.84, p=0.0003
Log_Volume	0.62	0.59	-0.09	-0.61	W=0.97, p=0.53
Log_Grootte	7.96	8.31	-0.89	0.72	W=0.91, p=0.007
Square root_Volume	1.49	1.34	0.69	-0.41	W=0.93, p=0.043
Square root_Grootte	72.73	63.89	0.28	-1.40	W=0.91, p=0.009

Table 1: Summary Statistics and Shapiro-Wilk Test Results

3.5 Proportion, Chi-Square Test of Independence

We visualize the proportion in Figure 6. The Chi-square test results are summarized in Table 2.

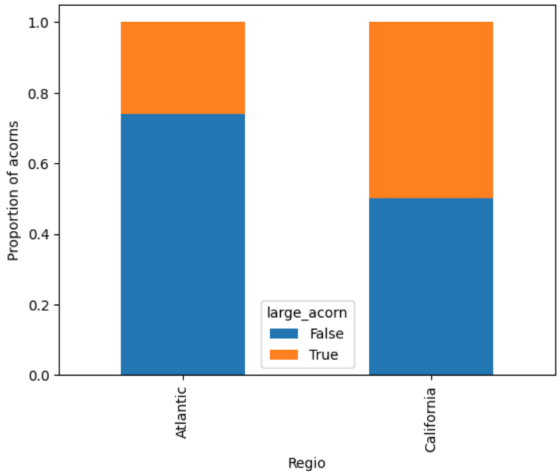

Figure 6: Proportion for 'Regio' and 'Large Acorn'

Table 2: Results of Chi-Square Test of Independence

Variable	Chi-square statistic	p-value
Area and Size	0.879	0.349

3.6 Ordinary Least Squares Regression

The results of the OLS regression are presented in Figure 7. Residuals were evaluated via various diagnostic plots including the residual plot and Q-Q plot of residuals. The Shapiro-Wilk test is also calculated for the residuals and are reported in Table 3. Finally, various statistics are presented in 8

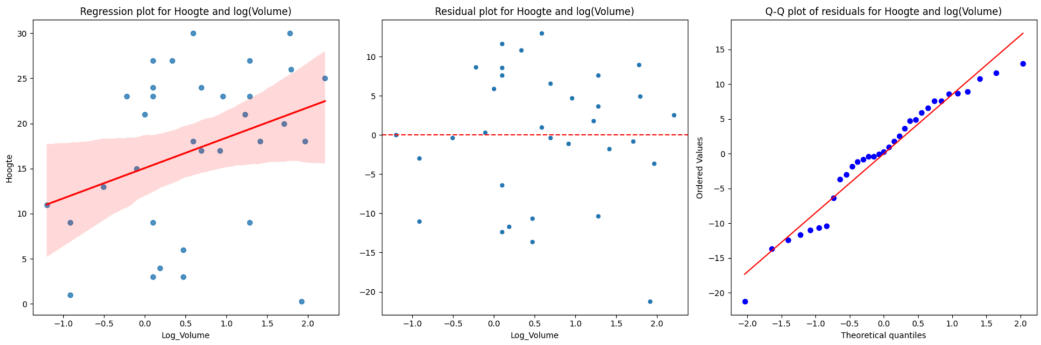

Figure 7: Regression and Residual analysis plot

Table 3: Results of Residual Tests

Test	Test Statistic	p-value
Shapiro-Wilk	0.953	0.159

Regression Results for Hoogte and log(Volume)

OLS Regression Results

Dep. Variable:	Hoogte	R-squared:	0.112
Model:	OLS	Adj. R-squared:	0.083
Method:	Least Squares	F-statistic:	3.892
Date:	Thu, 15 Jun 2023	Prob (F-statistic):	0.0575
Time:	19:12:16	Log-Likelihood:	-116.50
No. Observations:	33	AIC:	237.0
Df Residuals:	31	BIC:	240.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	15.0609	1.817	8.290	0.000	11.356	18.766
Log_Volume	3.3595	1.703	1.973	0.057	-0.113	6.833

Omnibus:	2.302	Durbin-Watson:	1.674
Prob(Omnibus):	0.316	Jarque-Bera (JB):	2.024
Skew:	-0.583	Prob(JB):	0.363
Kurtosis:	2.662	Cond. No.	1.94

Figure 8: Statistics for the regression

4 Discussion

4.1 Distribution of Volume' and Size'

Our analysis showed that both Volume' and Size' variables do not follow a normal distribution. This conclusion is backed by the results from the Q-Q plots, EDFs, histograms, and the Shapiro-Wilk test. However, after applying logarithmic transformations to both variables, the distributions approached normality, particularly for Volume'. This is suggested by the much larger p-value (0.53) from the Shapiro-Wilk test for Log-Volume' compared to the original Volume' variable (p=0.0002). This transformation indicates that the logarithm of the Volume' may better suit analytical methods that assume normality, such as parametric statistical tests or linear regression.

For 'Size', even after applying the logarithmic transformation, the distribution did not completely conform to normality (Shapiro-Wilk p=0.007). Nevertheless, the transformed variable had a better approximation to a normal distribution compared to the original variable.

4.2 Correlation Between Large Acorns and the Area

The Chi-square test of independence revealed no significant association between the region (Atlantic or California) and the presence of large acorns. The Chi-square statistic was 0.879 with a p-value of 0.349, which is greater than the usual significance level of 0.05. This implies that we do not have enough evidence to reject the null hypothesis that the two variables are independent.

The visual representation of proportions corroborates this finding. The proportion of large acorns is nearly the same in both regions. Hence, we can say that large acorns are not region-specific and the environment of both regions (Atlantic and California) are equally likely to produce oaks with large acorns.

4.3 Predicting 'Height' from the Logarithm of the 'Volume'

The Ordinary Least Squares (OLS) regression model was employed to analyze the relationship between the 'Height' (Hoogte) and the natural logarithm of the 'Volume' (Log_Volume). The results of the regression are presented in Figures 7 and 8. The model accounts for about 11.2% of the variance in the 'Height', as indicated by the R-squared value of 0.112. The adjusted R-squared, which takes into account the degrees of freedom, slightly decreases to 0.083.

The coefficient of Log_Volume is 3.3595, suggesting that a unit increase in the logarithm of 'Volume' is associated with an increase in 'Height' by approximately 3.36 units. However, this relationship is marginally significant at the conventional 0.05 level (p-value = 0.057).

The lack of a clear line of fit in the regression analysis implies that the logarithm of 'Volume' may not be a strong predictor for 'Height'. Despite this, the residuals seem to follow a normal distribution as there is no discernable pattern in the residuals plot. Although the Q-Q plot deviates slightly at the extremes, it generally conforms to a straight line, indicating that the residuals may follow a normal distribution.

This is further supported by the Shapiro-Wilk test (Table 3), which yields a test statistic of 0.953 and a p-value of 0.159. Since the p-value is greater than the typical alpha level of 0.05, we fail to reject the null hypothesis of normal distribution of residuals.

In summary, while the logarithm of 'Volume' is somewhat associated with 'Height', the relationship is weak and marginally significant. Despite this, the residuals appear to be normally distributed, suggesting that the assumptions of OLS regression are largely met. Therefore, further research with additional variables may be necessary to improve the prediction of 'Height'.

5 Conclusion

The investigation revealed that neither the 'Volume' nor the 'Size' variable follows a normal distribution, even though applying a logarithmic transformation helped these variables approach normality. A chi-square test of independence revealed that the size of the acorns is not region-specific, suggesting that large acorns are not exclusive to either the Atlantic or California region. Finally, ordinary least squares regression demonstrated that the logarithm of 'Volume' is only marginally significant in predicting the 'Height', explaining approximately 11.2 percent of the variability. The residuals from this model seem to follow a normal distribution.