Assignment 3: Italian Olives

Abstract

All work present is my (Emmanuel Olowe)'s original work.

Does the `palmitoleic` reading in a `olive` dataset depend on the region of Italy it was extracted from? A One-Way Anova and Kruskal-Wallis Rank Sum tests were used to determine this. A simulation test was used to determine suitability of these tests to answering the question. In the simulation study the One-Way Anova was found to have size of 0.76% whereas the Kruskal-Wallis tests were found to have a size of 1.46%. The One-Way Anova was found to have a power of 98.42% whereas the Kruskal-Wallis Rank Sum was found to have power of 98.27%. The overall accuracy of each test was 98.7% and 98.36%, respectively. Both tests rejected the null hypothesis that the mean `palmitoleic` of all the regions were the same.

Introduction

The dataset chosen was the `olive` dataset from within the `dslabs` library. From this the research question of "Does the palmitoleic reading in a olive sample of the dataset depend on the region of Italy it was extracted from?" was devised. There are three regions in the data set (Northern Italy, Southern Italy, Sardinia). Various sorts of initial analyses were performed upon the data to better understand its structure.

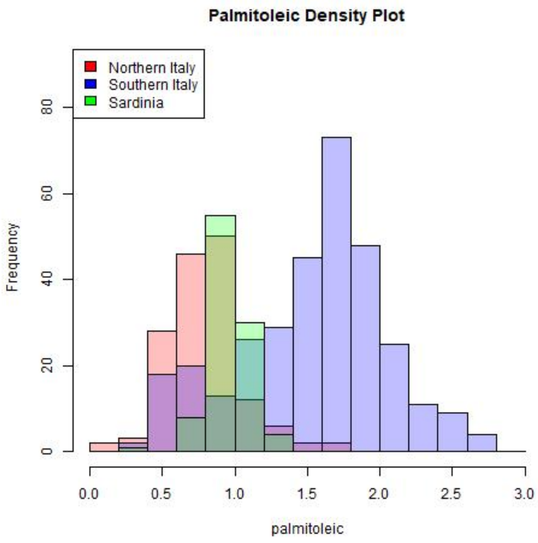

Figure 1.1

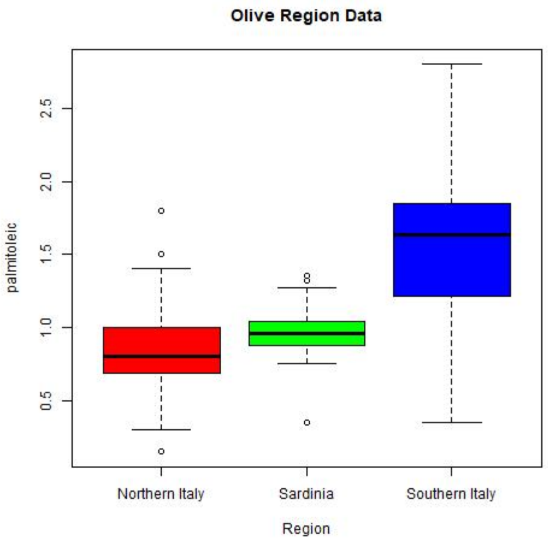

Figure 1.2

It could be seen from the sample data that the sample densities from the olive data [Figure 1.1] that the each region data came from difference distributions are they were all most dense in difference locations and all seemed to have difference levels of deviation away from their most dense locations. From the box plot [Figure 1.2] it can be seen clearly that the sample means of the regions are not all equal and it is clearly evident that all sample distributions have different standard deviations. Southern Italy here has the largest mean and the largest deviation away from mean.

Given there was evidence to believe that the means were different appropriate statistical tests were selected. The selected parametric test was a One-Way Anova. This was as the response variable (palmitoleic reading) was dependent on a categorical value (region). Checks were made to ensure that the assumptions of Anova test were held.

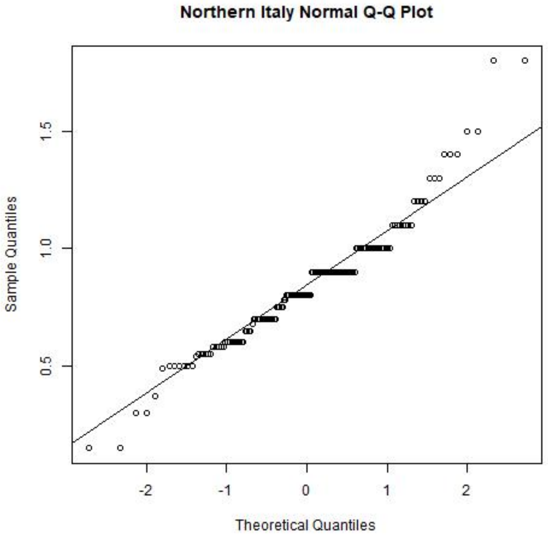

Figure 2.1

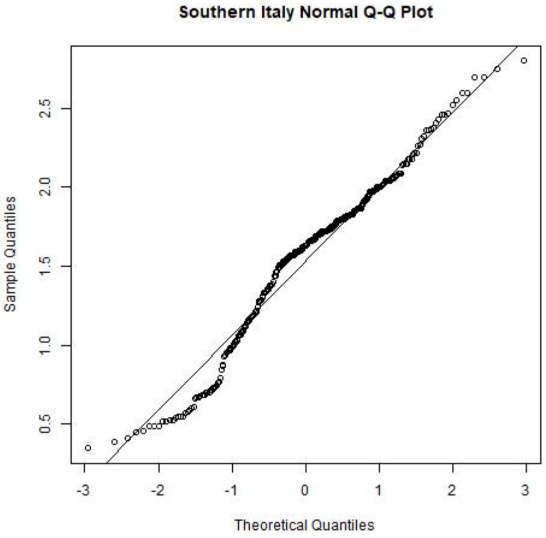

Figure 2.2

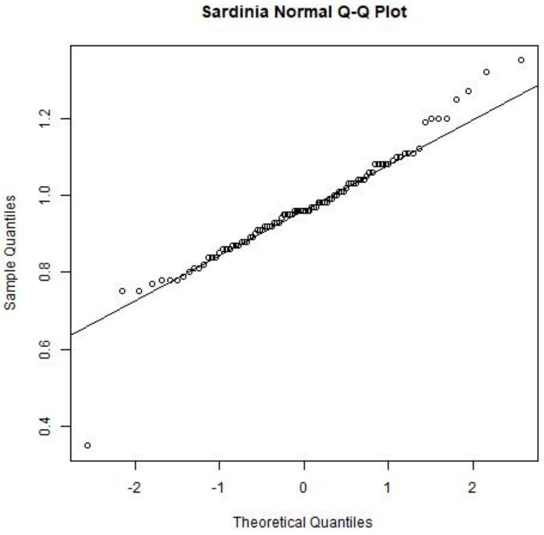

Figure 2.3

The assumptions of the One-Way Anova test are:

1. Observations are independent
2. No Significant outliers
3. Each Category is Normally distributed
4. Each Category has equal variance

Given that each sample only has one region the independence is a reasonable assumption. Given density plot [Figure 1.1] it can be seen that the shape of the regions resemble normal distributions. QQ plots were also created to determine the accuracy of the normality assumption [Figure 2.1-2.3] It could be seen that most of the regions did approximate the normal distribution but North did show unusual behaviour as well as south having great deviation away of lower end of the distribution. The assumption with no evidence for was that each category had equal variance.

If these assumptions did not hold it was determined that the optimal alternative test would be the non-parametric Kruskal-Wallis Test.

Hypothesis Test

This was the hypothesis test devised.

$$
\begin{gathered}
H_0 : \ \mu_{\text{North}} = \mu_{\text{South}} = \mu_{\text{Sardinia}} \\
H_1 : \text{Not All } \mu \text{ Are Equal}
\end{gathered}
$$

Running the Program

All instructions for running any element of this research are contained in the `README.md`

Methodology

Implementation of Statistical Tests

`R` comes with standard library for both chosen statistical tests

```
anova <- aov(palmitoleic ~ region, data = data)
kruskal_wallis <- kruskal.test(palmitoleic ~ region, data = data)
```

To ensure a standardised output from both functions they wrapped by functions to handle the data insertion and hypothesis testing. It was decided that a p value boundary to reject null hypothesis should be determined via simulation. So the significant levels would be:

$$
\begin{gathered}
p_1 = 0.05 \\
p_2 = 0.01
\end{gathered}
$$

Both wrapped functions would return TRUE if Null Hypothesis if was rejected and FALSE otherwise.

Testing Correctness

To determine the correctness of the implemented functions assertion tests were used to verify the desired functionality of each methods operations

Simulation Study

A study on simulated data was devised in order to determine the suitability of each model to the given data. Simulated data with known true means were generated and the simulation would then run the simulated data on the model and determine the correctness of the output. 3 possible configurations of dataset were developed:

1. All of the groups being from the same distribution
2. 2 Groups belong to the same distribution and 1 is from another distribution
3. All groups belong to different distributions

Other factors also need to be considered in the generation of data

1. All data in dataset is only given to 2dp. (effect of measurement error and rounding on results will need to be considered on generated samples)
2. The difference in means between the groups (what is the minimum difference in mean that the statistical test can differentiate); this is significant as value of palmitoleic are small and in the usually all to 3sf.

3. The difference in standard deviations between the groups (the standard deviation of the regions is not fixed)

In order to choose reasonable starting values for μ and σ of the simulated data. Values from the dataset were examined. It was decided that starting choices for μ would be generated from out with the region of the maximum and minimum values of the `palmitoleic` reading as these were seasonable values future readings could tend towards. To select region to generate standard deviation values from the minimum standard deviation of the regions were use and the standard deviation of all `palmitoleic` readings.

$$\mu_{\text{simulation}} \in [0, 3]$$
$$\sigma_{\text{simulation}} \in [0.1, 0.26]$$

600 samples were generated per simulation. This number was chosen because its approximately how many samples were present in the dataset. 5000 simulations were conducted with random generation of parameters and test type.

$$\text{Number of Samples} = 600$$
$$\text{Simulations} = 5000$$

The three types of test devised came from the 3 possible scenarios of data configuration.

As could be seen in the dataset. The number of observations belonging to each region were not evenly split so this was factored into the configuration of the simulations.

Type 1: All means the same

$$\mu_{\text{north}} = \mu_{\text{south}} = \mu_{\text{sardinia}}$$

$$35_{\text{north}} : 45_{\text{south}} : 25_{\text{sardinia}}$$

Type 2: One Mean different

$$\mu_{\text{north}} \neq \mu_{\text{south}} = \mu_{\text{sardinia}}$$

$$30_{\text{north}} : 45_{\text{south}} : 25_{\text{sardinia}}$$

Type 3: All means different

$$\mu_{\text{north}} \neq \mu_{\text{south}} \neq \mu_{\text{sardinia}}$$

$$35_{\text{north}} : 40_{\text{south}} : 25_{\text{sardinia}}$$

A flowchart was used to design the data flow and control structures which would be in use to develop the study. It is available in the directory: `design/flowchart.pdf`

Results

α is the significant level of the test.

Simulation Study

All results given to 2 decimal places.

`palmitoleic` Rounded to 2DP and $\alpha = 0.05$

Statistical Test	Size (%)	Power (%)	Accuracy (%)
One Way Anova	5.18	98.46	97.22
Kruskal-Wallis Rank Sum	4.49	98.65	97.6

Rounding `palmitoleic` to 2DP and $\alpha = 0.01$

Statistical Test	Size (%)	Power (%)	Accuracy (%)
One Way Anova	0.76	98.42	98.7
Kruskal-Wallis Rank Sum	1.46	98.27	98.36

`palmitoleic` Not Rounded to 2DP and $\alpha = 0.05$

Statistical Test	Size (%)	Power (%)	Accuracy (%)
One Way Anova	4.58	98.55	97.5
Kruskal-Wallis Rank Sum	5.02	98.91	97.6

`palmitoleic` Not Rounded to 2DP and $\alpha = 0.01$

Statistical Test	Size (%)	Power (%)	Accuracy (%)
One Way Anova	0.98	98.56	98.72
Kruskal-Wallis Rank Sum	1.13	98.01	98.3

Hypothesis Test

Statistical Test	H_0 when $\alpha = 0.05$	H_0 when $\alpha = 0.01$
One Way Anova	Rejected	Rejected
Kruskal-Wallis Rank Sum	Rejected	Rejected

Please provide up to 5 minutes for the completion of running of the simulation study.

Discussion

Both Statistical tests have high statistical power ($> 98\%$) regardless of rounding. This indicated that there will be small risk to Type II statistical errors. Even with increased significance level it could still be seen that the power of both statistical tests remained high ($> 98\%$).

As expected a higher significant level increase the probability of Type I statistical errors, by over 600% in the rounded version. It can be seen that the lower significant level does not severely affect power (less than 1% in all cases).

There is reasonable evidence to conclude that $\alpha = 0.01$ is the optimal choice for significance level which minimises Type I errors and maximises accuracy of the statistical test.

When significance level was $\alpha = 0.05$, the One Way Anova is outperformed by the Kruskal-Wallis Rank Sum by 0.38%. Given this is also seen without round with the outperformance by 0.1%. Though these numbers may be marginal, these difference could be caused by the underlying assumptions of One Way Anova not being met. Whereas when significance level was $\alpha = 0.01$, the Anova outperformed the Kruskal-Wallis Rank Sum by 0.34%, this could be due to the assumptions of the Anova being reasonable.

Finally it was selected that $\alpha = 0.01$ and rounded 2 Decimal Places were the most valid representation of dataset to the greatest degree of accuracy.

To improve significance of findings, more experimentation could have been conducted into the effects of the ratio observations in regions effect on the statistical tests.

Both tests at both significance levels reject the null hypothesis indicating that there is significant evidence against the null hypothesis leading to the conclusion alternative hypothesis must be accepted.

$$H_1 : \text{Not All } \mu \text{ Are Equal}$$

Conclusion

To conclude, a investigation of the whether the palmitoleic reading of an olive is dependent on the region of Italy is was collected from was undergone. A One Way Anova and Kruskal-Wallis Rank Sum were statistical tests used to test the Null Hypothesis. A simulation study was created and undergone to test the validity of the chosen statistical tests. The simulation study found the size, power, accuracy:

$$\{[0.76, 1.46], [98.42, 98.27], [98.7, 98.36]\}$$

It was determined that the null Hypothesis was rejected by both tests and determined that there was significant evidences against the null hypothesis that the mean palmitoleic was the same across all the regions.

References

[1] https://www.datanovia.com/en/lessons/anova-in-r/ - How to do ANOVA tests

[2] https://www.researchgate.net/post/Three-means-comparison-by-t-test-or-ANOVA - Why ANOVA over t-test

[3] https://bookdown.org/BaktiSiregar/data-science-for-beginners-part-2/11-ANOVA.html running Anova Tests