University of St Andrews | FOUNDED 1413

Essay cover sheet to be completed by student and attached to the front of essay.

STUDENT ID : **_____180009169_____**

COURSE : **_____MT5762 Introductory Data Analysis___**

ESSAY/PROJECT TITLE : **_____Basic tests and inference_ _____**

NAME OF TUTOR(S) : **_____ Carl Donovan _____**

SUBMISSION DEADLINE : **_____2018/10/11_____**

DATE SUBMITTED : **_____2018/10/11_____**

DECLARATION

- I confirm that I have read and understood the University's policy on plagiarism and I agree to participate in the courses provided by ELT to develop good academic practice and ways of citing and referencing the work of others.
- I confirm that this assignment is all my own work and that I have only had help from ELT staff when preparing it.
- I confirm that in preparing this piece of work I have not copied any other person's work, or any other piece of my own work.
- I confirm that this piece of work has not previously been submitted for assessment on another course.

_____180009169_____ (Student ID in place of signature)

_____2018/10/11_____ (date of submission)

1. Introduction

Statistics is not only a useful tool for scientific studies but also contribute to the process of prosecution. For instance, someone has explored the origin determination of cannabis plants with statistics, which can help police when prosecuting growers (Smith, 2000). This report is also focus on the data about cannabis plants. The data set used in this report contains information about the levels of thirty-eight elements in cannabis leaves which are grouped by four soil types which are potting mix (pm) and three locations about New Zealand. Three locations are bhb, mb, and nth, being: Blockhouse Bay (Auckland suburb), Mission Bay (Auckland suburb) and Northland (a northern region). It will first show there are differences in the elemental composition between Cannabis leaves in four groups, then prove that some elements are related in terms of their content levels in sampled leaves. Finally, this report will discuss whether the origin determination of cannabis can be made just from the elemental composition data. This investigation will use RStudio as the analysis package

The analysis process begins with summarizing the means and standard deviations, then using Analysis of Variance (ANOVA) and Tukey's Honest Significant Difference (TukeyHSD) to achieve more details about the elemental composition differences. For one test that has assumptions badly violated, non-parametric tests are used as substitutions. As for the relationships between elements, the Pearson correlation coefficient and scatter plots are used to analyze. Finally, the origin determination question will be discussed with some boxplots.

2. Exploratory analysis of the data

2a) Different elemental composition between soil types

This investigation will not include analysis for all the elements. Instead, five elements are chosen, which are Ti, Ca, Ga, Ba and Zn. There are significant differences in the elemental composition of Cannabis leaves grown in different soil types. The table1 below shows the means and standard deviations for every element in each group. Intuitively, the elemental composition is different between groups and the difference is significant. However, this only gives people a sense but no strong evidence. Further on, the difference can be verified by using Analysis of Variance (ANOVA). Before implementing the test, three assumptions should be checked. The first assumption is that the sample data should be independent. The last two is that the data of each group should come from one normal distribution and have equal standard deviation. However, these assumptions can be combined as one that the residuals of all data should obey one normal distribution. To test this, qqnorm, qqline and shapiro.test functions are adopted. The null hypothesis in this test is that the residuals are from one normal distribution. The Q-Q plots (Figure1) show that the residuals of Ti and Ca are likely to be normal distributed, but other elements data are likely to have leptokurtosis. Furthermore, the same conclusion can be made from the p-values of Shapiro-Wilk normality test (Table2). Using a type-1 error of 5%, the p-values of Ti and Ca are larger than 0.05. Therefore, the null hypothesis cannot be rejected, which means these two elements pass the test and suitable for ANOVA. However, other elements do not pass the test because the p-values of them are much less than the benchmark, which means there is a strong evidence to reject the null hypothesis. The ANOVA then conducted on Ti and Ca. The null hypothesis of ANOVA is that the means of all group are equal. Compared with 5%, the p-values of Ti and Ca are both much smaller, which is a strong evidence to reject the null hypothesis. In other words, there is at least one group has different elemental composition with another groups.

Table1. Summary of means and standard deviations

Elements / Group	Ti (mean/sd)	Ca (mean/sd)	Ga (mean/sd)	Ba (mean/sd)	Zn (mean/sd)

bnb	5.10/1.32	63615.38/11586.91	11.84/2.02	486.15/69.59	1323.08/329.53
mb	0.95/1.20	86.70/48.26	0.15/0.22	0.78/0.36	4.56/1.35
nth	7.58/1.62	554111.11/9184.83	2.62/0.38	112.44/13.45	456.67/54.77
pm	8.85/1.45	30208.33/9069.68	1.72/0.59	71.79/15.62	330.83/61.71

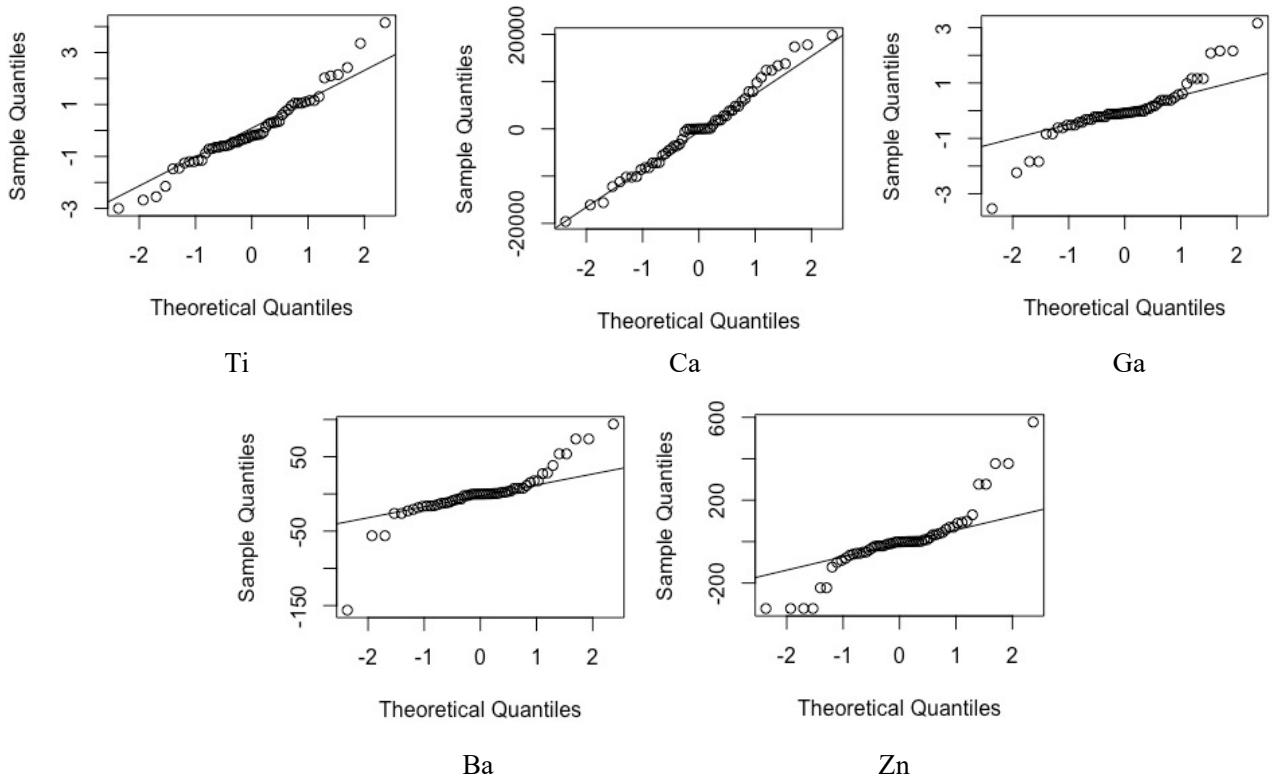


Figure1. Q-Q plots

Table2. Shapiro-Wilk normality test

Elements	Ti	Ca	Ga	Ba	Zn
P-values	0.1043	0.7058	<0.01	<0.01	<0.01

Although the ANOVA verified that there are at least one group has different level of elemental content compared to another group, it cannot indicate which pairs are different and how different they are. Multiple comparisons such as Tukey's Honest Significant Difference (TukeyHSD) can help to achieve

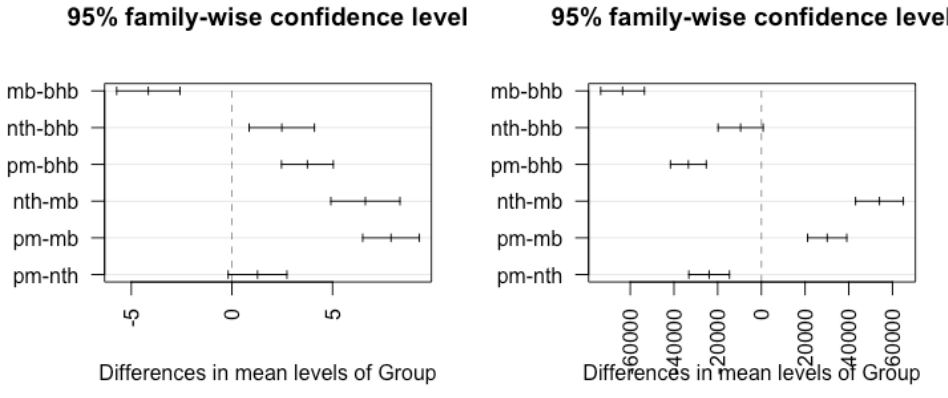
details like these. The null hypothesis of TukeyHSD is that the means of the specific pairs are equal, which is focus on each pair and more specific than ANOVA. The table3 shows the TukeyHSD results for Ti that all p-values are much smaller than 0.05, except the pair "pm-nth" which is over 0.1. Therefore, there is no evidence to reject the null hypothesis for the pair "pm-nth". On the contrary, there is a strong evidence to against null hypothesis for other pairs and say they have different average levels of Ti. The same analysis process can apply to element Ca. The table4 shows TukeyHSD results for Ca. All pairs except "nth-bhb" are likely to have different content levels of Ca. Furthermore, the figure2 proves the confidence intervals of the differences between groups with 95% confidence level.

Table3. TukeyHSD results for Ti

Pairs	mb-bhb	nth-bhb	pm-bhb	nth-mb	pm-mb	pm-nth
P-values	<0.01	<0.01	<0.01	<0.01	<0.01	0.11

Table4. TukeyHSD results for Ca

Pairs	mb-bhb	nth-bhb	pm-bhb	nth-mb	pm-mb	pm-nth
P-values	<0.01	0.08	<0.01	<0.01	<0.01	<0.01



Ti Ca

Figure2. confidence intervals with 95% confidence level

For those elements which do not meet assumptions, using ANOVA as well as TukeyHSD can lead to wrong conclusions such as differences that ANOVA does not detect. Instead, Kruskal-Wallis test is more suitable in this condition. The null hypothesis of Kruskal-Wallis test is the same as ANOVA, the means of all group are equal. Taking Zn as an example, the p-value of Kruskal test is much less than 0.05, which indicates a strong evidence to against null hypothesis and prove at least one group has different level of Zn to another group. To compare specific pairs, Wilcoxon test can be used. However, it can only test one pair at one time. According to Kabacoff (2015), wmc( ) function that written by himself can compare all groups in the dataframe when the possibility of type-1 error is controlled. The null hypothesis of this test is same as TukeyHSD, the means of the specific pairs are equal. Applying this function to Zn, the results are showed in table5. All p-values are much less than 0.05, which is a strong evidence to reject the null hypothesis. Therefore, it is believed that the average levels of Zn are differences in all pairs of groups.

Table5. wmc( ) results for Zn

Pairs	mb-pm	mb-nth	mb-bhb	pm-nth	pm-bhb	nth-bhb
P-values	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

2b) Related content levels between elements

Although the elemental composition is different between groups, there are relationships between elements. Preliminarily, Pearson correlation coefficients are calculated to estimate which pairs are possible to be related (Table6). According to the results, there are three pairs of elements (Ga ~ Ba, Ga ~ Zn, Ba ~ Zn) have large correlation coefficients which are over 0.9. Therefore, they seem to be related in terms of their levels in the sampled leaves. This phase then followed by using the cor.test( ) function in order to verify their relationships. The null hypothesis of this test is that there is no correlation between two elements. The test results show that the p-values of all three pairs are much less than 0.05, which is a strong evidence to reject the null hypothesis. Therefore, there are strong correlations between them. Furthermore,

this is Pearson correlation coefficient and they are all positive, so the relationships are positive linear relationships. The Figure3 gives some plots to visualize.

Table6. Pearson Correlation coefficients between elements

	Ti	Ca	Ga	Ba	Zn
Ti	1.00	0.27	-0.10	-0.09	0.02
Ca	0.27	1.00	0.77	0.79	0.82
Ga	-0.10	0.77	1.00	0.99	0.93
Ba	-0.09	0.79	0.99	1.00	0.96
Zn	0.02	0.82	0.93	0.96	1.00

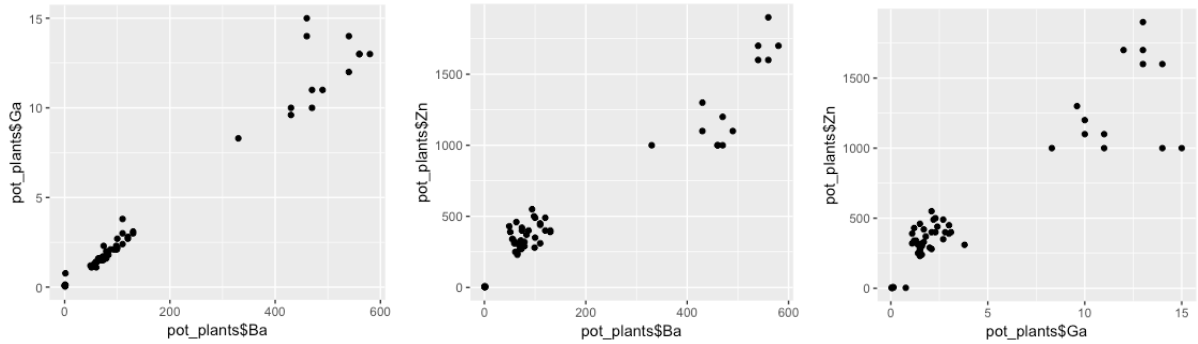


Figure3. Related elements

2c) Determination of the type of soil from the elemental composition

It might not very cautious to determinate the soil types that the plants were grown in just from the elemental composition of the leaves. Although it has been proved that the average levels of Ti and Ca are significantly different between Cannabis leaves grown in four soil types, there could still be variability in different units. In other words, the ranges of elements content for different groups can overlap. By making boxplots of Ti and Ca, it is clear that only the level of Ca in mb group is not overlapped with other groups (Figure4). Finding some unique elements in plants which were grown in particular soil types

can be a strong evidence to make determination. Moreover, finding elements which have significant different ranges might also help to achieve the goal.

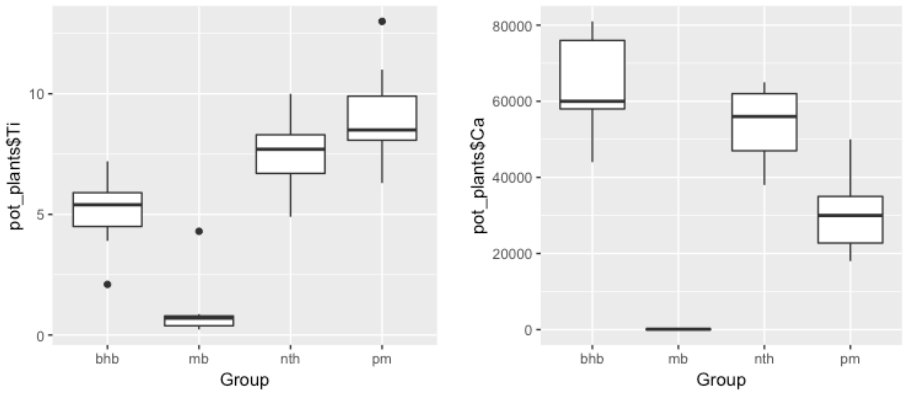


Figure4. Boxplots of Ti and Ca

3. Conclusion

To sum up, after cleaning and doing exploratory analysis of the data, three conclusions can be made. First, there are strong evidences indicate that the levels of elements such as Ti and Ca are significantly different between Cannabis leaves grown in four soil types. Second, the content levels of some pairs of elements, such as Ga and Ba, have closely positive linear relationships. Finally, it is not very cautious to make determinations about which places the plants were grown in just from the elemental composition of the leaves mainly because the ranges of elements content for different groups are overlapped. Finding some elements which are unique or have significant different ranges between four groups might be helpful to solve the origin determination problem.

Reference:

Simth, N. (2000) *Criminals may rue pot from this plot.* Available at:

https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=142910 (Accessed: 10th Oct. 2018)

Kabacoff, R. (2015) *R in action.* Manning Publications