Major Assignment

Student Number: 1004452
Name: Mu Tong
Email: mtm@student.unimelb.edu.au

Case1: Study of graft arteries – Professor Brian Buxton

a) Table 1: Summary table for risk factors of the study participants

Risk Factors	Number of 0	Number of 1	Proportion of 1
Age in years	Mean: 65.773 Min:42 Max:81 Median: 68.5		
Sex	12 (female)	98 (male)	89.09%
Presence of diabetes	83 (no)	27 (yes)	24.55%
History of cigarette smoking	37 (never)	73 (ever)	66.36%
Presence of peripheral vascular disease	91 (no)	19 (yes)	17.27%
Presence of cerebrovascular disease	99 (no)	11 (yes)	10.00%
Presence of hypercholesterolemia	57 (no)	53 (yes)	48.18%

Definition of variables in data files is already defined in the website on the Data option. There are 7 risk factors, and the number of 0s and 1s are collected in the table. We can find that the mean and median of age from 110 participants is 65.773 and 68.5 respectively, and the minimum and maximum age are 42 and 81 respectively. The proportion of male from 110 participants is 89.09%, and the presence of diabetes, smoking, peripheral vascular disease, cerebrovascular disease and hypercholesterolemia are 24.55%, 66.36%, 17.27%, 10.00%, 48.18% respectively.

b)

The 3 main indices used were:

1 Percentage of luminal narrowing

2 Intimal thickness index

3 Intima-to-media ratio

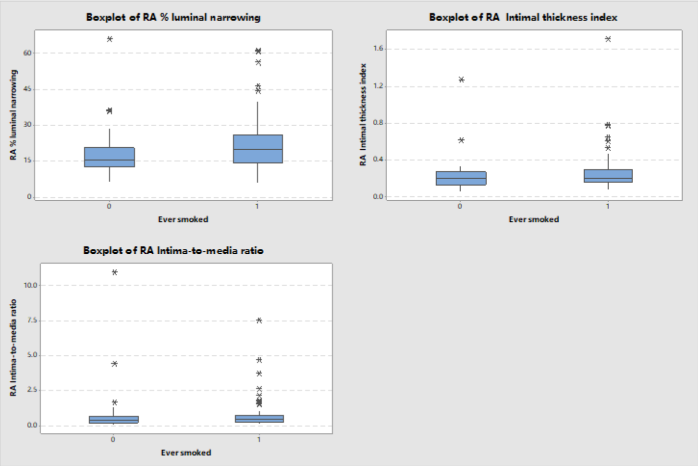


Figure 1: Boxplot of three main indices used for RA

c) No solution

d) There are three assumptions:

1. the variances of distribution are assumed to be the same.

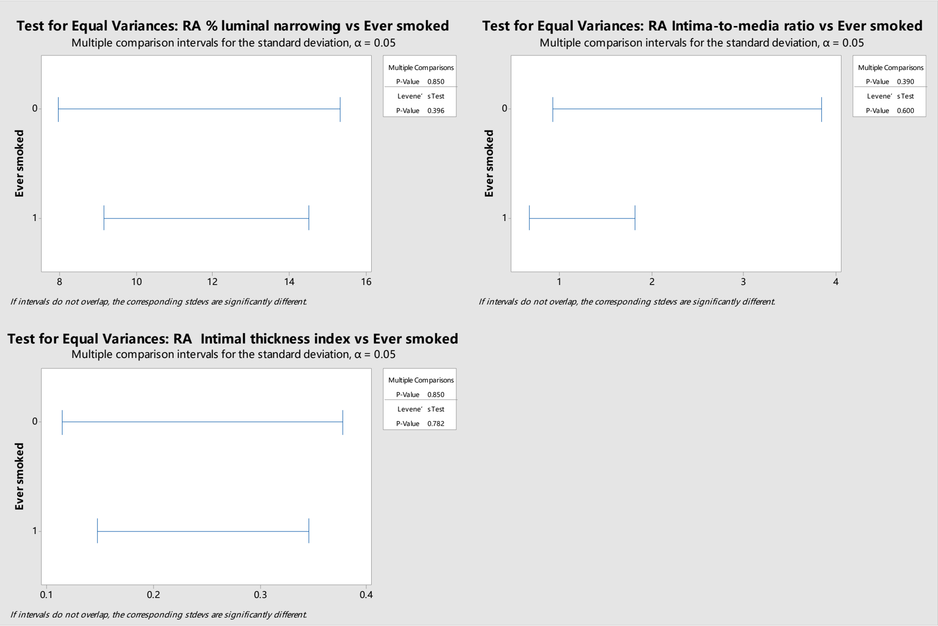


Figure 2: Test for Equal Variance for three main indices for RA

We can find that the p-value for three indices are all much greater than 0.05, which means that this gives us no reason to doubt the null hypothesis. In other words, we can accept the null hypothesis, which means that we can assume that the distribution has the same variance.

2. The distributions are assumed to be Normal.

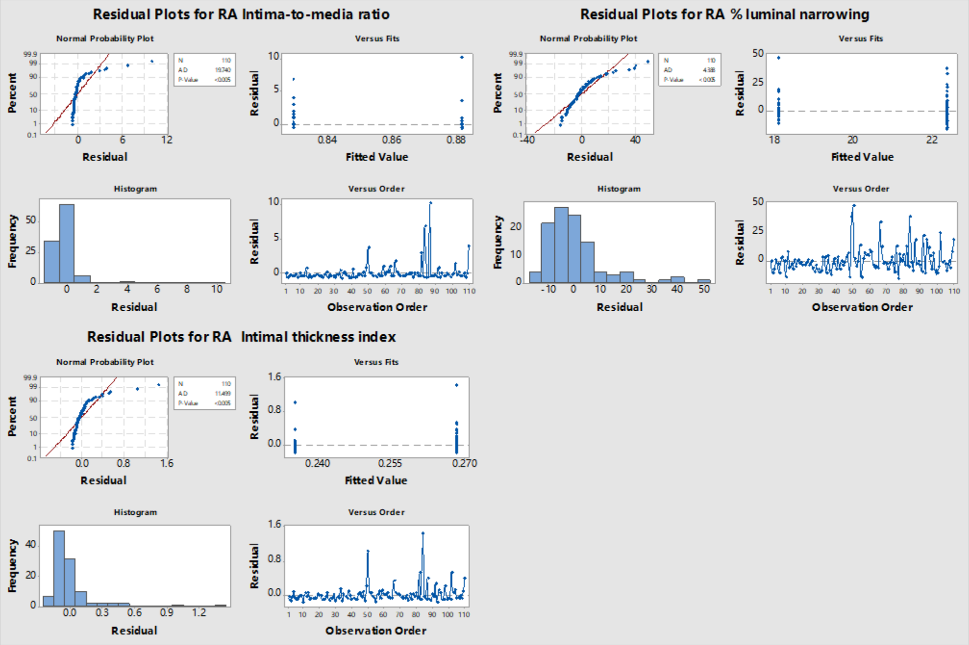


Figure 3: Residual Plots for Three Main Indices for RA

We can find that three p-values are all smaller than 0.005, which is relatively small, which means that this is the evidence against the null hypothesis. In other words, the results show that it is likely that the distributions for three indices are all not normal.

3. All observations are assumed to be independent. From the study design, we can find that the data are collected from 110 patients (40 patients are excluded because risk factors details were incomplete), and the patients are independent from each other, so the observations should be independent.

e) Summary Statistics

Sample	N	Mean	StDev	Variance	Median	IQR
RA % luminal narrowing	110	20.93	11.30	127.62	18.58	10.88
RA Intimal thickness index	110	0.2572	0.2153	0.0464	0.2096	0.1197
RA Intima-to-media ratio	110	0.847	1.380	1.906	0.48	0.466

Inferential Statistics

Sample	Estimated Difference	95% CI	T-Value	P-Value
RA % Luminal Narrowing	-4.30	(-8.77, 0.16)	-1.91	0.059
RA Intimal thickness	-0.0329	(-0.1192, 0.0534)	-0.75	0.452
RA Intima-to-Media Ratio	0.053	(-0.502, 0.607)	0.19	0.851

Method: difference is (the mean when ever smoked = 0) – (the mean when ever smoked = 1)

f)

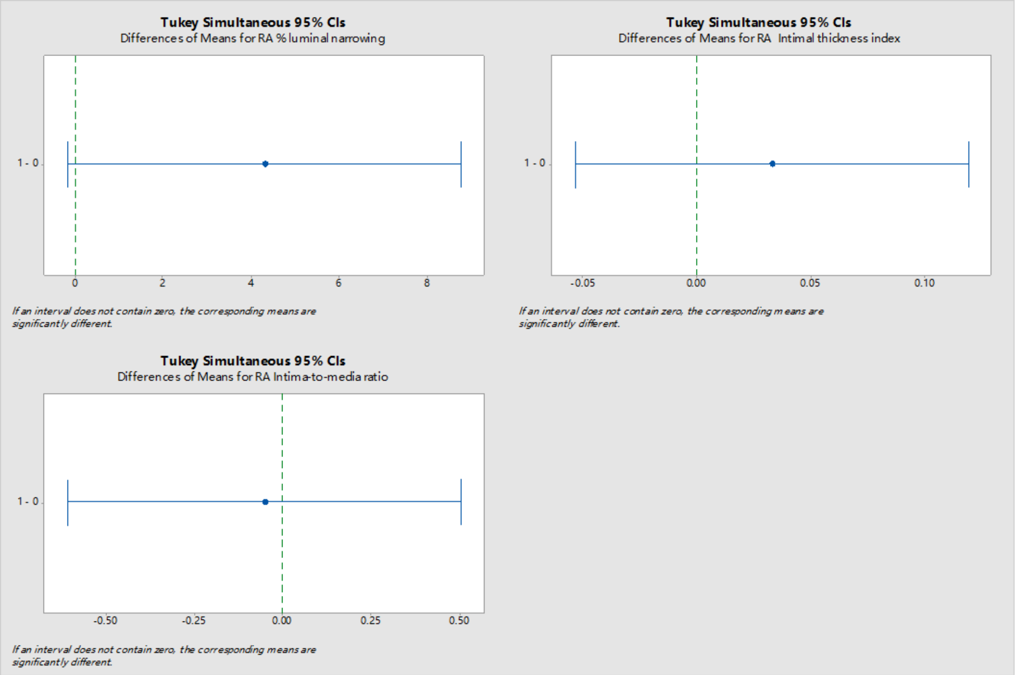


Figure 4: Tukey 95% CI for the Mean Differences of Three Main Indices

g) We can find that the last two p-values are much greater than 0.05, and the difference is close to 0, which means that this gives us no reason to doubt the null hypothesis. In other words, we can accept the null hypothesis, which means that it is more likely that there is no difference for RA Intimal thickness and RA Intima-to-Media Ratio when the individual has ever smoked or not. For RA Luminal Narrowing, the estimated difference is 4.30, and the p-value is 0.059, which is relatively small, but also greater than 0.05, so we can say that the data is not statistically significant. However, the difference is not close to 0, which is relatively large; therefore, smoking people may have higher percentage of luminal narrowing, which indicates more severe disease, in other words, the blood flow may be impaired. We can get the information that there is 66.36% of patients has ever smoked, so it is quite dangerous that we use radial artery for CABGs.

h)

Inferential Statistics

Sample	Estimated Difference	95% CI	T-Value	P-Value
Intimal Abnormality	-0.1588	(-0.3413, 0.0237)	-1.73	0.087

Methods: difference is (the mean when smoked = 0) – (the mean when smoked = 1)

We can find that the difference is smaller than 0, which means smoking people has higher intimal abnormality in this sample, but the p-value is 0.087, which is greater than 0.05, so the data is not statistically significant.

We can get the information that all p-values for RA are greater than 0.05, which means that the data are not statistically significant, but we can’t say that the findings are not useless.

Case2

a)

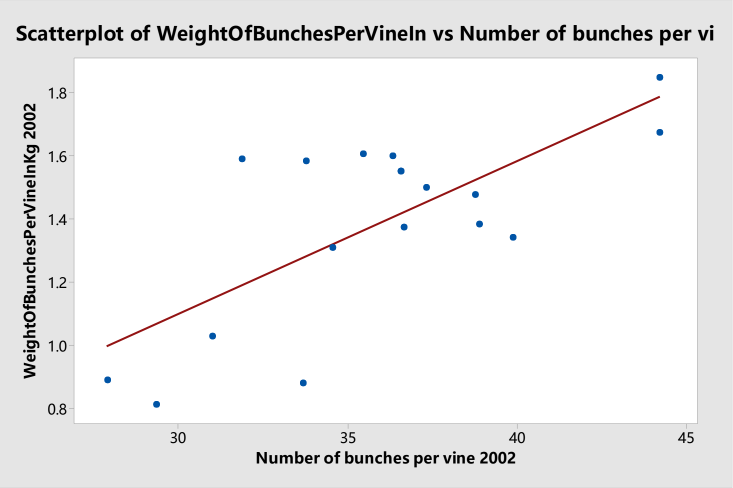


Figure 5: Scatterplot of average number of bunches harvested and the average weight harvested for 2002

b)

We can get the Pearson correlation is 0.718, and the p-value is 0.001 from Minitab, which means that there is a relatively strong positive relationship between these two variables.

c)

Summary Table for Analysis of Variance

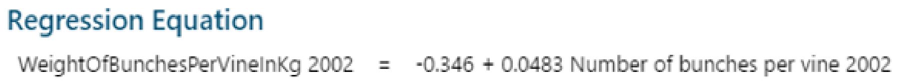
Source	DF	Seq SS	Contribution	F-Value	P-Value
Number of bunches per vine 2002	1	0.77906	51.59%	17.05	0.001
Error	16	0.73090	48.41%		
Total	17	1.50995	100%		

Summary Table for model

S	R-sq	R-sq(adj)	RPESS	R-sq(pred)
0.213731	51.59%	48.57%	0.905145	40.05%

Summary Table for coefficients

Term	Coef	95% CI	T-Value	P-Value
Constant	-0.346	(-1.243,0.551)	-0.82	0.426
Number of bunches per vine 2002	0.0483	(0.0235,0.0731)	4.13	0.001



Regression Equation

WeightOfBunchesPerVineInKg 2002 = -0.346 + 0.0483 Number of bunches per vine 2002

Figure 6: Regression Equation for the number of bunches and the average weight for 2002

We can use regression model to find the coefficients, the estimated coefficients are showed in the table, therefore, we can get the regression equation. -0.346 means the constant, and 0.0483 means that if add 1 bunches per vine, then the weight of bunches per vine will increase 0.0483 kg.

d)

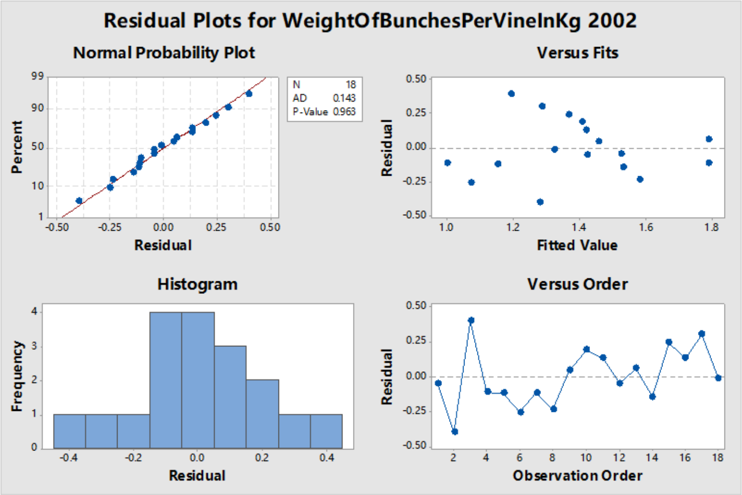


Figure 7: Residual Plots for Wight of Bunches per Vine in 2002

We can get that the p-value is 0.963, which is relatively large, so it gives us no reason to doubt the null hypothesis, which means that the distribution is more likely normal, but we can find that the sample size is 18, which is relatively small, therefore, we can't guarantee that the distribution is definitely normal.

e) Result Table

Average number of bunches	Fit	95% PI
35	1.34391	(0.877814,1.81000)

95% prediction interval means that if the average number of bunches is 35, we are 95% confident that the mean average weight of bunches for 2002 will lie between 0.877814 and 1.81.

f) The predicted value of the average weight when the average number of bunches is 0 is -0.346.

g) The regression model is the model that predict the value of an outcome variable from some suitable explanatory variables, so the model doesn't make sense when the explanatory variable is too small or too big. Therefore, it doesn't make sense when the explanatory variable is 0.

h)

We can find that the sample size is only 18, which is relatively small, it may cause some unexpected outliers, so increasing the number of sample sizes is a possible approach.

i)

I expect that it may not be similar, that is because that if we analyze based on individuals, some factors will affect the result of analysis, for example, the soil will be different if we don't use block, therefore, using block will reduce the variance of the random error, and greatly increase the precision of the experiment.

Case3

a) Treatment: 1: herbicide, 2: compost, 3: straw

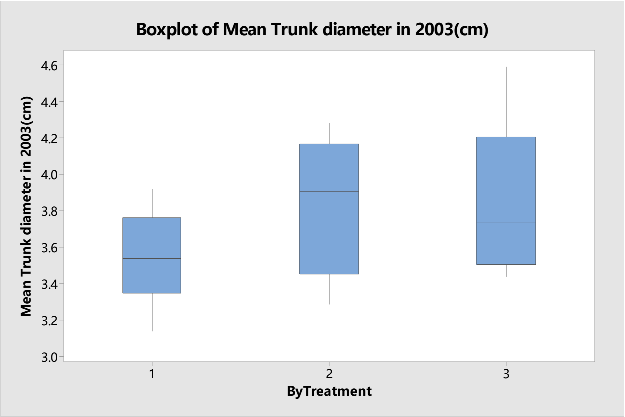


Figure 8: Boxplot of Mean Trunk Diameter in 2003

b)
We can find that compost treatment has the highest mean from three treatments, and we can see that compost and straw have similar variance because the range of box is similar. As for herbicide, the overall mean is the lowest, but the variance may be the smallest.
2 mean higher

c)

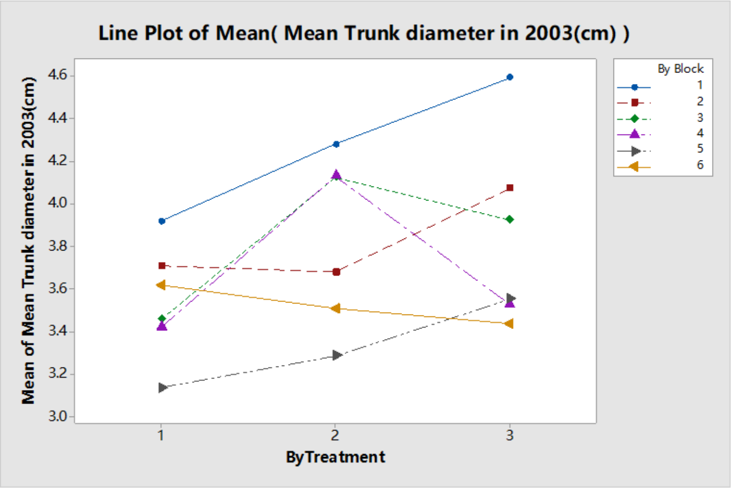


Figure 9: Line Plot for Mean Trunk diameter in 2003

I will include treatment and block as factors, that is because we can find that there appear to have a block effect, so it is likely that block is sensitive. Therefore, we use ANOVA to analyze, and we found that the p-value for block is 0.014, which is smaller than 0.05. In other words, this gives the evidence that we should against the null hypothesis, which means that there will be some differences when we use block factor, and treatment is the main interest, so treatment and block will be the factors. According to EWblock and Aspect, we can know that block is the combination of EWblock and Aspect, so we don't need to add EWblock and Aspect as factors.

d) Assuming that the variances are the same.
The variances of each of these three distributions for treatments (herbicide, compost, straw) are assumed to be the same.

e)
Summary Table for Analysis of Variance

Source	DF	Seq SS	Contribution	F-Value	P-Value
By Treatment	2	0.3623	14.51%	3.03	0.094
By Block	5	1.5362	61.53%	5.14	0.014
Error	10	0.5981	23.96%		
Total	17	2.4967	100%		

We can find that the p-value for the treatment is 0.094, which is greater than 0.05, and it is relatively large, so it gives us no reason to doubt the null hypothesis. In other words, it is more likely that treatments will not affect the trunk diameter. As for block, the p-value is 0.014, which is relatively small, and it is smaller than 0.05, which means that this is the evidence against the null hypothesis. In other words, it is more likely that block will affect the trunk diameter.

f) Table for the Mean by Treatment

By Treatment	Mean
herbicide	3.546
compost	3.838
straw	3.854

We can find that the means by compost and straw treatment are higher and similar, and the mean by herbicide treatment is about 0.3 smaller than other two treatments.

g)

We can use test for equal variance in ANOVA, and we find that the p-value for Levene is 0.372, which is much greater than 0.05, and relatively large, so it gives evidence that we have no reason to doubt the null hypothesis, which means that we accept the null hypothesis that all variance are the same.

h)

Table of 95% CI for comparing the mean by treatments

Difference of levels	Difference of Means	95% CI	T-Value	Adjusted P-value
Compost - herbicide	0.293	(-0.273,0.858)	1.34	0.394
Straw - herbicide	0.309	(-0.256,0.874)	1.42	0.357
Straw - compost	0.016	(-0.549,0.581)	0.07	0.997

We can find that all p-value for comparisions are much greater than 0.05, which means this gives us no reason to doubt the null hypothesis. In other words, it is more likely that there will be no difference between treatments, the data is not statistically signifant.

i)

We can get the information that it is more likely that the treatment will not affect the trunk diameter, the data is not statistically significant, but we can find the information that compost and straw will have a better and similar result according to the difference of mean, therefore, compost and straw will be both fine for vine growth.

j)

We can use Friedman test to analyze the data without the distribution assumptions, because we only have 6 blocks for each treatment, the sample size is relatively small, which may cause in some unexpected outliers, so the distribution may not be normal, so Friedman test may be a good choice.