UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

UNIVERSITI TEKNOLOGI MALAYSIA SCHOOL OF COMPUTING SESSION 2020/2021 SEMESTER 2

SECI 2143-04-PROBABILITY & STATISTICAL DATA ANALYSIS

LINK TO VIDEO PRESENTATION:
https://drive.google.com/file/d/1EuUOoFDg-7NSJrQGMj-9GnYgSIG8zLN5/view?usp=sharing

GROUP MEMBER

Name	Matric Number
Nabil Alkahar	A20EC0281
Bintang Prakasa Antovie	A20EC0295
Christian Dimas Budiyono Ramadhani	A20EC0296
Rayhan Rafi Arviandy	A20EC0329

I. Introduction and Background

This report is going to discuss the inferential statistics on the Iris plant dataset. The purpose of this study is to see the inferences on the population. The dataset is a public dataset that is the property of UCL machine learning.

This case report utilizes the inferential statistics concepts to identify the relationship between petal and sepal data in the dataset. Using tools such as R Studio, the group will then try to conduct tests on the data such as Linear Regression, Hypothesis testing, Correlation testing etc.

II. Dataset

The Iris dataset was used in R.A. Fisher's classic 1936 paper, The Use of Multiple Measurements in Taxonomic Problems, consists of 150 data of 3 species of irises. With numerical data of the sepal length, sepal width, petal length and petal width.

III. Data Analysis

A. Compulsory Tests

[1] Hypothesis or Sample Test

We use data 10-60 as a sample

$H_0 = \mu = 5.843$

$H_1 = \mu \neq 5.843$

$\alpha = 0.05$

R Code:

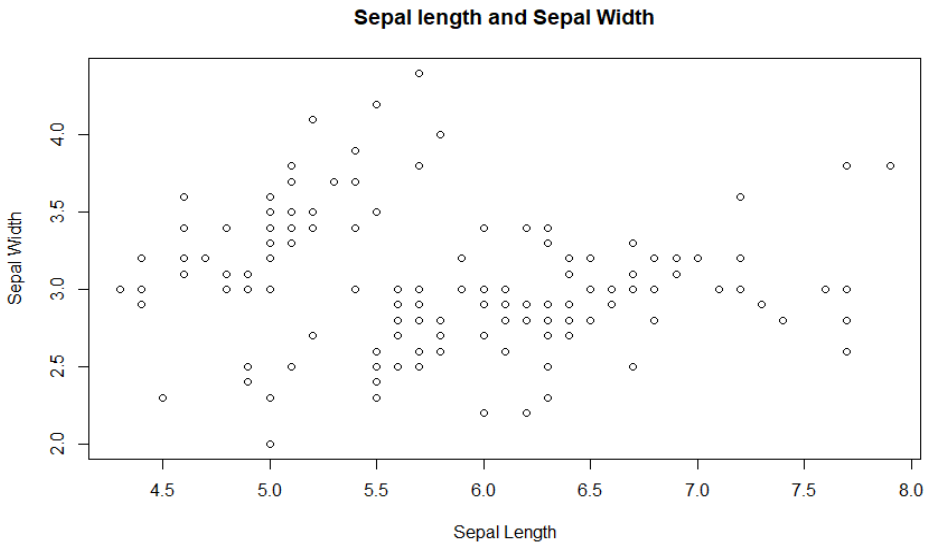
```
x = Iris$SepalLengthCm[10:60];
t.test(x, alternative = 'two.sided',mu=5.843);
```

Output:

```
    One Sample t-test

data:  x
t = -6.9245, df = 50, p-value = 7.876e-09
alternative hypothesis: true mean is not equal to 5.843
95 percent confidence interval:
 5.074198 5.419920
sample estimates:
mean of x
 5.247059
```

Since we have p-value less than α, we Reject H_0 because we don't have enough evidence to prove that $\mu = 5.843$

[2] Correlation Test



R Code:

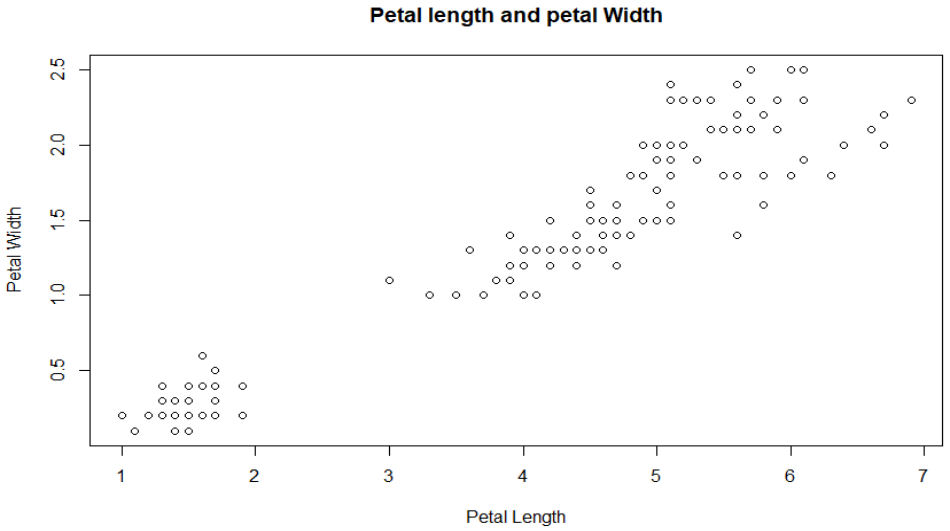
```
cor.test(data$SepalLengthCm,data$SepalWidthCm)
```

Output:

```
Pearson's product-moment correlation
```

data: data$SepalLengthCm and data$SepalWidthCm
t = -1.3386, df = 148, p-value = 0.1828
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.26498618 0.05180021
sample estimates:
cor
-0.1093692

Thus, negative correlation between Sepal Length and Sepal Width.



R Code:

```
cor.test(data$PetalLengthCm,data$PetalWidthCm)
```

Output:

Pearson's product-moment correlation

data: data$PetalLengthCm and data$PetalWidthCm
t = 43.32, df = 148, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9489049 0.9729061
sample estimates:
      cor
0.9627571

Thus, Positive correlation and a near perfect correlation between petal length and width.

[3] Regression Test

Variable = Sepal Length & Sepal Width

Type of Regression: Simple Regression

i. Hypothesis Statement

H_0: $\beta = 0$ (no linear relationship)

H_1: $\beta \neq 0$ (linear relationship exist)

ii. Execution

n =150
df = 148
α = 0,05

Min = -1.1023
Max = 1.33779

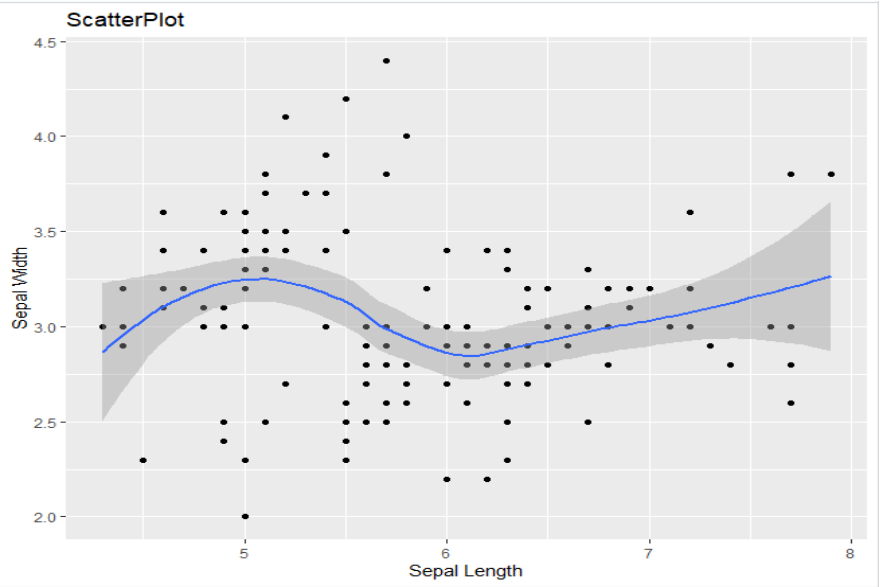
F-Statistic = 1.792
p-value = 0.1828

Standard error of estimate = 0.4324

$\hat{y} = 3.38864 - 0.05727x$

```
Call:
lm(formula = Iris$SepalWidthCm ~ Iris$SepalLengthCm)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10230 -0.23930 -0.01639  0.27414  1.33779

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.38864    0.25248  13.421   <2e-16 ***
Iris$SepalLengthCm -0.05727    0.04278  -1.339    0.183
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4324 on 148 degrees of freedom
Multiple R-squared:  0.01196,   Adjusted R-squared:  0.005286
F-statistic: 1.792 on 1 and 148 DF,  p-value: 0.1828
```

```
Call:
lm(formula = Iris$SepalWidthCm ~ Iris$SepalLengthCm)

Coefficients:
       (Intercept)  Iris$SepalLengthCm
           3.38864            -0.05727
```



iii. Conclusion

Since p-value = 0.1828 > α = 0.05; We reject H_0. This means a relationship exist between Sepal Length and Sepal Width.

Variable = Petal Length & Petal Width

Type of Regression: Simple Regression

1. Hypothesis Statement

   H_0: β = 0 (no linear relationship)

   H_1: β ≠ 0 (linear relationship exist)

2. Execution

   n =150

   df = 148

   α = 0,05

   Min = -0.56543

   Max = 0.64278

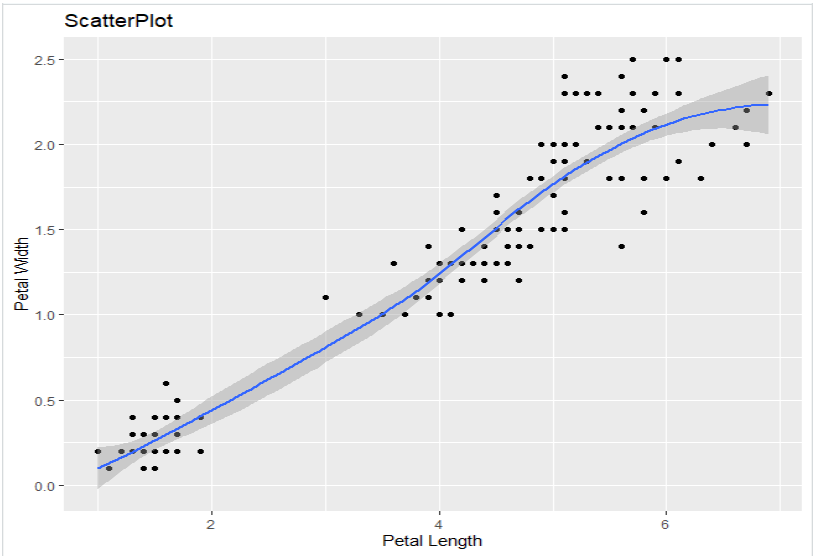
F-Statistic = 1877
p-value = 2.2e-16

Standard error of estimate = 0.207

$\hat{y} = -0.3665 + 0.4164x$

```
Call:
lm(formula = Iris$PetalWidthCm ~ Iris$PetalLengthCm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56543 -0.12409 -0.01647  0.13251  0.64278

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.366514   0.039889  -9.188 3.35e-16 ***
Iris$PetalLengthCm  0.416419   0.009613  43.320  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.207 on 148 degrees of freedom
Multiple R-squared:  0.9269,    Adjusted R-squared:  0.9264
F-statistic:  1877 on 1 and 148 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Iris$PetalWidthCm ~ Iris$PetalLengthCm)

Coefficients:
       (Intercept)  Iris$PetalLengthCm
           -0.3665              0.4164
```



3. Conclusion

Since p-value = 2.2e-16 < α = 0.05; We fail reject H_0. This means a relationship does not exist between Petal Length and Petal Width.

B. Optional Test - ANOVA

1. Hypothesis Statement

   $H_0 = \mu1 = \mu2 = \mu3 = \mu4$

   H_1 = at least one mean is different

2. Execution

   $\alpha = 0.05$

   n = 150

   k = 4

   R code: dataResult <- aov(SepalLengthCm ~ PetalWidthCm, data = Iris)

   summary(dataResult)

```
              Df Sum Sq Mean Sq F value Pr(>F)
PetalWidthCm   1  68.36   68.36   299.2 <2e-16 ***
Residuals    148  33.81    0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   F-Statistic = 299.2
   p-value = 2e-16

3. Conclusion

   Since, the F-statistic value is less than the P-value (299.2 < 2e-16) we fail to reject the null hypothesis. There is sufficient evidence to claim that the different types of sepal and petal have the same mean for iris. There are no significant differences between the mean of sepal and petal. Therefore, all does have same mean

Conclusion

From the test conducted on the data set, we can conclude that both petal and sepal length and width have relations. This is proven by seeing the correlation and regression testing above, and seeing the ANOVA test result, we can conclude that the null hypothesis that there are no significant differences between the mean of sepal and petal.