UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI 2143: PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2 REPORT

SECTION : 04 – 1SECR

COURSE NAME : BACHELOR OF COMPUTER SCIENCE – COMPUTER NETWORKS & SECURITY

NO.	NAME	STUDENT ID
1	AMEENUDDIN BIN ISMAIL	A19EC0016

LECTURER'S NAME: DR SUHAILA BINTI MOHAMAD YUSUF

DATE OF SUBMISSION: 14/6/2020

Table of Contents

Introduction

In this project, I will practice and use all of the knowledge about inferential statistic that I learn in class. There are four inferential statistic that will be use in this project which is hypothesis testing on two sample, correlation, regression and analysis of variance (Anova). Other than that, all test will be calculate using R studio since the objective of this project is to apply the uses of R studio in calculation of four inferential statistic.

The secondary data set that I choose is about the record of the marks secured by the students after a monthly test in one school in United State of America. This data set includes scores from three subjects and a variety of personal, social, and economic factor that will affect the test scores. The data set has 1000 students as a population. The three test is Mathematic, reading, writing test. Firstly, we will study about different of proportion between Mathematics' score of male and female students for hypothesis testing on two sample. Second, we will find the correlation coefficient between the variable writing score and reading score of students. The linear regression model will be apply in the scatter plot. For anova, the different of means of three test subject will be test.

Inferential Analysis

Hypothesis Testing on Two Sample

Question:

In August 2014, American Psychological Association (APA) suggested that men and women actually has an equal aptitude for learning math and science (Association, 2014). To test this claim, we will choose a simple random sample of 100 male students and 150 female students from a population of 1000 students in school in United State.

After a monthly test, 30% of male students score A (marks ≥ 80) in Mathematic subject and 17.3% of female students score A in Mathematic subject. Based on these data set, can we reject APA's claim that men and women has an equal aptitude for learning Mathematic? Use a 0.05 level of significance.

Solution:

	Male	Female
Score A in math, x (marks ≥ 80)	$x_1 = 30$	$x_2 = 26$
Sample, n	$n_1 = 100$	$n_2 = 150$
Percentage	30%	17.30%

Table 1: hypothesis testing

Information:

- Significance level = 0.05

Hypothesis:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Formula:

- Pooled Sample Proportion:

$$\bar{p} = \frac{x_1 + x_2}{n_1 = n_2}$$

$$\bar{q} = 1 - \bar{p}$$

- Test Statistic, z

$$z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

R Studio:

```
x1 = 30
x2 = 26
n1 = 100
n2 = 150

phat1 <- x1/n1
phat2 <- x2/n2

pbar = (x1+x2)/(n1+n2)
qbar = 1-pbar

z = ((phat1-phat2)-0)/sqrt((pbar*qbar/n1)+(pbar*qbar/n2))

alpha = 0.05
z.alpha = qnorm(alpha/2)
```

Figure 1: hypothesis testing

z	2.35333144197704
z.alpha	-1.95996398454005

Figure 2: test statistic & critical value

Test Statistic, z = 2.35

Critical Value = 1.95, -1.95

Conclusion:

Test Statistic, z ≥ Critical Value

$2.35 \geq 1.95$

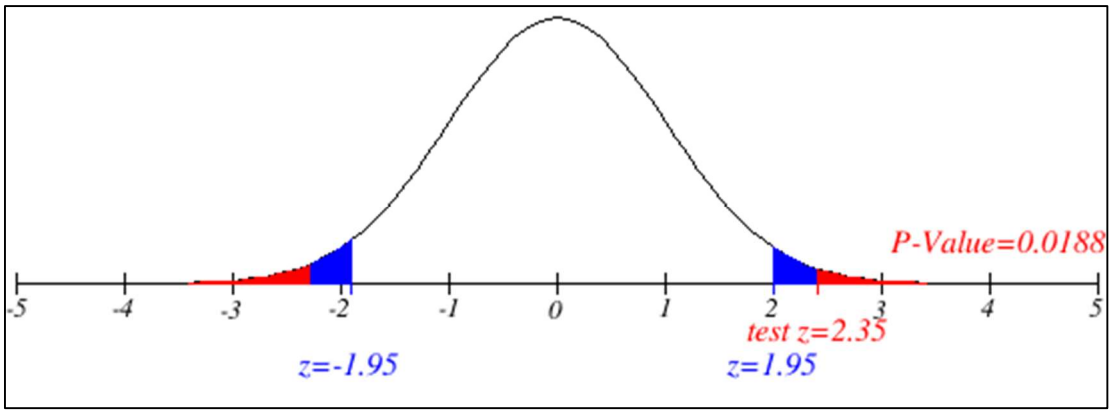

Figure 3: Graph for hypothesis testing

$\therefore reject\ null\ hypothesis, H_0$

Discussion:

We not reject $H_1 : p_1 \neq p_2$ because it is enough evidence to reject the claim that men and women has an equal aptitude for learning Mathematic.

Correlation

Task:

What is the relationship between reading score, x and writing score, y from the sample of nine students?

reading score, x	writing score, y
72	74
90	88
95	93
57	44
78	75
83	78
95	92
43	39
64	67

Table 2: Correlation (x,y)

- Construct a scatter plot and label it.
- Compute the value of the correlation coefficient, r between x and y.

Solution:

Hypothesis:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

$$test\ statistic, t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad , \alpha = 0.05$$

R Studio:

- Calculate the correlation coefficient, r:

```
1  x<- c(72,90,95,57,78,83,95,43,64)
2  y<- c(74,88,93,44,75,78,92,39,67)
3
4
5  cor(x,y)
```

Figure 4: calculation of r

```
> cor(x,y)
[1] 0.9734843
>
```

Figure 5:r = 0.97

$r = 0.97$

- Calculate test statistic, t and critical value, t(α⁄2,n-2):

```
1  n=9
2  r=0.97
3
4  t=(r)/sqrt((1-(r*r))/(n-2))
5
6  alpha = 0.05
7  t.alpha = qt(1-alpha/2,n-2)
```

Figure 6: calculation of test statistic, t

t	10.556671654118
t.alpha	2.36462425159278

Figure 7: test statistic, t and critical value

Conclusion:

$\therefore test\ statistic, t = 10.56 > 2.36$, we reject null hypothesis, H_0.

Discussion:

Because there is sufficient evidence that linear correlation exist between x and y.

- Construct a scatter plot and label it:

```
1  x<- c(72,90,95,57,78,83,95,43,64)
2  y<- c(74,88,93,44,75,78,92,39,67)
3
4  plot(x,y)
5
```

Figure 8:plotting a scatter graph

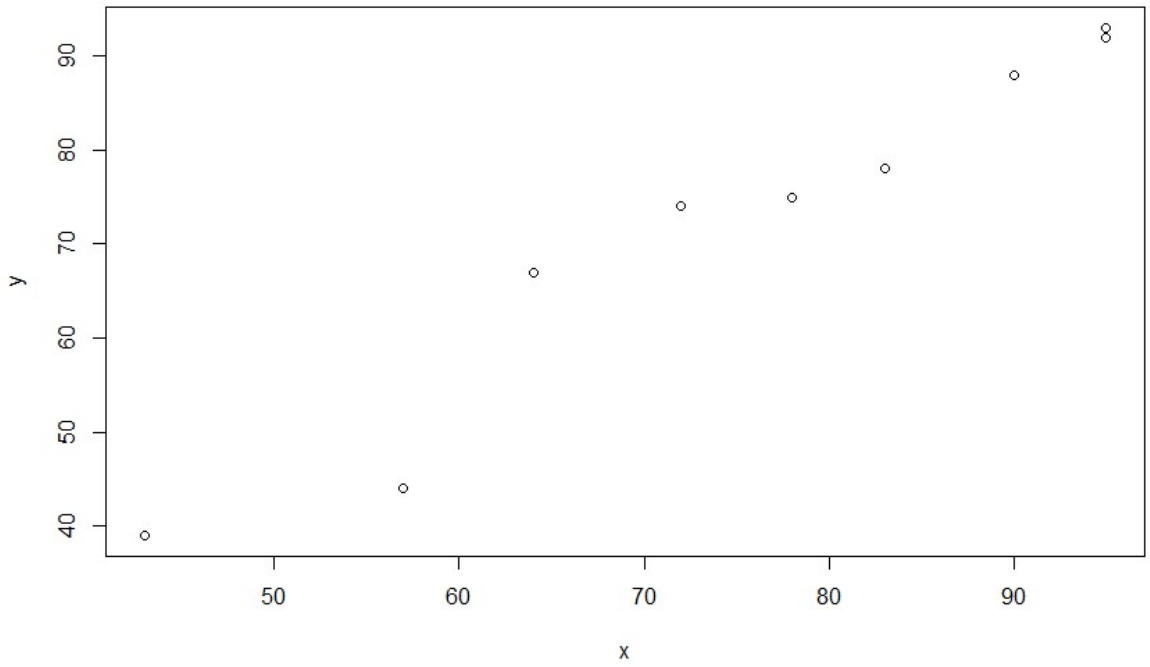

Figure 9: scatter graph

Conclusion:

$\therefore r > 0.8$, the relationship between x and y is strong positive linear relationship. It is means that the students that score excellent in reading test, will have excellent writing score. We can conclude that people that have good reading skill will improve their writing skill.

Regression

Task:

- Perform linear regression and build linear regression model.
- Plot a scatter graph and add linear regression model into the plot.

Solution:

R Studio:

- Perform linear regression and build linear regression model:

```
x<- c(72,90,95,57,78,83,95,43,64)
y<- c(74,88,93,44,75,78,92,39,67)

plot(x,y)


cor(x,y)

model <- lm(y~x)
model
summary(model)

```

Figure 10: build linear regression model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.38093    7.24660  -1.019    0.342
x            1.05824    0.09399  11.259 9.74e-06 **
---
```

Figure 11: coefficients

$$\hat{y} = -7.381 + 1.058x$$

- Plot a scatter graph and add linear regression model into the plot:

```
x<- c(72,90,95,57,78,83,95,43,64)
y<- c(74,88,93,44,75,78,92,39,67)

plot(x,y)


cor(x,y)

model <- lm(y~x)
model
summary(model)



plot(x,y)
abline(model)
```

Figure 12:adding linear regression model into the plot

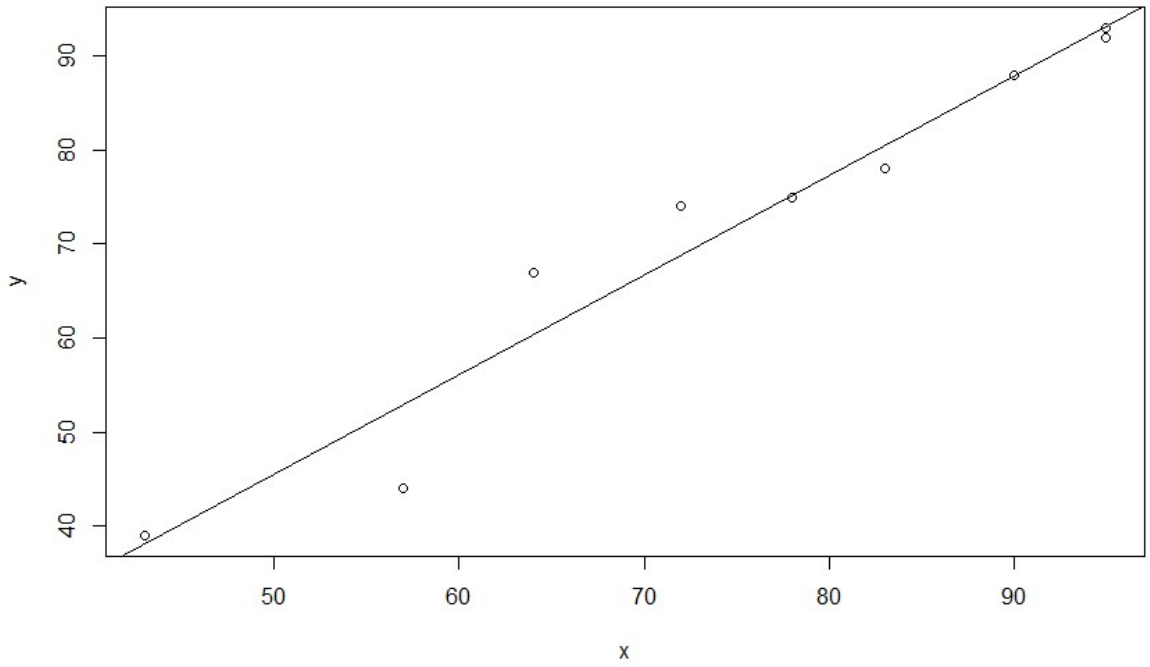

Figure 13: scatter graph with linear regression model

Conclusion:

1. This linear regression model is simple regression model because it only an independent variable, x (reading score).
2. The linear regression model:

$$\hat{y} = -7.381 + 1.058x$$

$$b_0 = -7.381$$

$\therefore b_0$ is estimated average value of writing score, y when the reading score, x is zero. Because there is no -7.381 mark in test, we assume student score zero mark. Student is estimate to score zero mark in writing test when they score zero mark in reading test.

$$b_1 = 1.058$$

$\therefore b_1$ is estimate change in average value of writing score, y as a result of one-unit change in reading score, x. As a conclusion, the average value of writing score, y is increase by 1.058 for each additional one mark of reading score, x.

Analysis of Variance (ANOVA)

Task:

Test Score		
math score	**reading score**	**writing score**
73	86	82
65	72	74
27	34	36
71	79	71
43	45	50
79	86	92
78	81	82
65	66	62
63	72	70
58	67	62

Table 3: test score of three subjects

Table 3 lists the test score of three subject, Mathematic, reading and writing from three random sample of ten students. We will use a 0.05 significance level to test either different subject test has the same mean or at least one mean is different.

Solution:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: at\ least\ one\ mean, \mu\ is\ different$$

- Find n, $\bar{x}$, s for each test subject:

R Studio:

```r
n=10
# count standard deviation
A<-sd(c(73,65,27,71,43,79,78,65,63,58))
B<-sd(c(86,72,34,79,45,86,81,66,72,67))
C<-sd(c(82,74,36,71,50,92,82,62,70,62))


#count mean
x<-c(73,65,27,71,43,79,78,65,63,58)
a <- mean(x)

x<-c(86,72,34,79,45,86,81,66,72,67)
b <- mean(x)

x<-c(82,74,36,71,50,92,82,62,70,62)
c <- mean(x)
```

Figure 14: calculate n, mean and s

	test score		
	math score, A	**reading score, B**	**writing score, C**
n	10	10	10
$\overline{x}$	62.2	68.8	68.1
s	16.2	17.2	16.5

Table 4: n, mean and s

- Find variance between samples:
    - Find mean between samples, $\bar{\bar{x}}$
    - Find standard deviation between samples, $s_{\bar{x}}$
    - Find variance between samples, $ns_{\bar{x}}^2$

R Studio:

```
#count mean between sample
D <- mean(c(a,b,c))
```

Figure 15: calculate mean between sample

```
D    66.3666666666667
```

Figure 16: value of mean between sample

```
#count s.d between sample
E<-sd(c(a,b,c))
```

Figure 17:calculate standard deviation between sample

```
E    3.62537354397217
```

Figure 18:value s. d between sample

```
#count variance between samples
f = n*(E)*(E)
```

Figure 19:calculate variance between sample

```
f    131.433333333333
```

Figure 20:value of variance between sample

$\bar{\bar{x}}$	66.4
$s_{\bar{x}}$	3.6
$ns_{\bar{x}}^2$	131.4

Table 5:value of mean, s.d, variance

- Find variance within sample, s_p^2:

```
#count variance within samples
G = ((A*A)+(B*B)+(C*C))/k
```

Figure 21:calculate variance within samples

G	276.225925925926

Figure 22: value of variance within sample

s_p^2	276.2

Table 6: value of variance within samples

- Calculate test statistic, F:

```
#calculate test statistic
H = f/G
```

Figure 23: calculate test statistic

H	0.475818237889005

Figure 24: value of test statistic

- Calculate Critical Value:
    - Find numerator and denominator:

```
#numerator&denominator
num = k-1
den = k*(n-1)
```

Figure 25:calculate numerator and denominator

num	2

Figure 26:value of numerator

den	27

Figure 27:value of denominator

    - Find F critical value:

```
#count F crtitical value
qf(.95,df1 = num,df2 = den)
```

Figure 28:calculate F-critical value

```
> qf(.95,df1 = num,df2 = den)
[1] 3.354131
>
```

Figure 29; value of critical point

Conclusion:

Since $F_{test\ statistic} < F_{critical\ value}(0.48 < 3.35)$, we fail to reject the null hypothesis, H_0.

Discussion:

There is sufficient evidence that the three test subjects have the same mean for mark score.

Conclusion

There are many valuable experience that I got from this project that can never be learn in the theoretically class only. For example, my skill in using R Studio to do inferential analysis improve massively while working in this project. My understanding about inferential analysis become more excellent because I need to apply it to the data set that I choose myself. Any deficiency in this project mainly came form me, because my knowledge about R Studio is limited as this program rarely used by me in daily life. So, I hope in the future, I can use R Studio program more skillfully.

References

- Association, A. P. (2014, August). *Psychology: Science in Action*. Retrieved from Research debunks myths about cognitive difference: https://www.apa.org/action/resources/research-in-action/share