UTM
UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF
CIVIL ENGINEERING

SCHOOL OF COMPUTING

PROJECT 2:
ASSESSMENT

Code & Subject **:** SECI2143 PROBABILITY & STATISTICAL DATA ANALYSIS

Section **:** 04

Name of Lecturer: Dr. Suhaila Mohamad Yusuf

NO	NAME	MATRIC NO
1.	NG PEI WEN	A19EC0117

Table of Content

No	Content	Page
1	Introduction	3-4
2	Focus on topic	5-11
3	Conclusion	11-12
4	Reference	12

Introduction

LGBT is a word that formed from 4 word that are L for lesbian, G for gay, B for bisexual and T for transgender. In today's world, more and more people have become one of the members of this group. It has become a trend where members group LGBT will share the facts and promote that they are LGBT in social media such as Facebook, Instagram, YouTube and many more. LGBT group have fight for their right to live in this world such as same-sex marriage which is against the religion and rules in some countries. Malaysia is an Islamic country where does not go along with this LGBT trend. All the religions in Malaysia such as Islam, Buddha, Hindu, Christian and others do not promote LGBT while religions told the community not to do so. Therefore, we as Malaysian rarely see posts or people that are promoting LGBT trend in our country.

At the beginning of the project, I have difficulty on finding secondary data about LGBT. This is because LGBT are not acceptable in mostly countries or community. Racist and discrimination often occurred toward this group of people, hence the data collected regarding this field is lesser compared to another topic. Therefore, I used the data set from U.S as we all know people who live in western countries are more open-minded compare to Asian. We can also say that western countries have started the trend earlier than Asia countries.

In this project, I have taken secondary data from the internet where it collected data of LGBT in United State for research purposes. The website provides various data set on different topic of United State and also from all over the world. Since, we can choose on the topic we want to carry out statistical data analysis so I had chosen a data set which mainly on lesbian and gay that is same-sex couples in United State. The reason why I choose this topic is I want to test on the proportion and the factors that affect the number of same-sex couples.

Four tests have been carried out to test on the data set for further analysis that are hypothesis testing 1 sample, correlation analysis, regression analysis and also lastly Chi-squared test of independence. In the first test, I will test on the proportion of female-female couples among the same-sex couples. The population of female is wide more than the population of male in this world, therefore I have stated a claim that the proportion of female-female couples is more than 0.5 with 0.01 significance level. In the second test, I have carried out correlation analysis to test on the relationship between age group and the number of same-sex couples at 0.05 significance level. Through test 2, I can know that is the age group one of the factors that will affects the increase of

same-sex couples. In test 3, a regression model has been made for the variables income (yearly) and the number of same-sex couples. Regression analysis enable us to find out is there any relationship on two variables. Thus, I will get to know if the income of household will affect the number of same-sex couples. In the last test that is test 4, I will test on the independence of races and gender of same-sex couples with 0.05 significance level. This test will show us that does races actually is one of the factors that affect an individual to be lesbian or gay?

Support of topic

Test 1: Hypothesis testing 1 sample

In this test, I am going to test the claim that the proportion of lesbian among LGBT group is more than 0.5. The claim is made based on the total number of females is more than total number of males in this world. Hence, we carried out this test with $\alpha = 0.01$ to test on the claim.

H0: $p = 0.5$

H1: $p > 0.5$

Critical region: At 0.01 significance level, Z (0.01) = 2.33

Test Statistic:

```
1  n=995420
2  k=510355
3  p=0.5
4  pbar=k/n
5  alpha=0.01
6
7  //test statistics
8   z= (pbar-p)/sqrt((p*(1-p))/n)
9
10
11 pbar
12 z
```

```
> pbar
[1] 0.5127032
> z
[1] 25.34811
```

Pbar (sample proportion) = 0.5127 (4 d.p)

Z test = 25.34811 > 2.33

Decision: Reject H0.

Conclusion:

At 0.01 significance level, there is sufficient evidence to support the claim that the proportion of lesbian among LGBT group is more than 0.5.

Test 2: Correlation Analysis

Test 2 is carried out to determine is there a linear relationship between the age group and the number of same-sex couples at 0.05 significance level?

H0: $\rho = 0$

H1: $\rho \neq 0$

Critical Region:

At $\alpha = 0.05$, T (0.05/2,4) = 2.776 or -2.776 ; degree freedom = n-2 = 6-2 = 4

Test Statistics:

Age group	15-24	25-34	35-44	45-54	55-64	Above 65
Frequency	40812	209038	190125	209038	191121	154290

```
x.age<-c(19.5,29.5,39.5,49.5,59.5,69.5)
y.number<-c(40812,209038,190125,209038,191121,154290)

#calculate r,corr coefficient
cor(x.age,y.number)
plot(x.age,y.number)
```

```
> x.age<-c(19.5,29.5,39.5,49.5,59.5,69.5)
> y.number<-c(40812,209038,190125,209038,191121,154290)
> #calculate r,corr coefficient
> cor(x.age,y.number)
[1] 0.4421315
> plot(x.age,y.number)
```

Answer get using R studio: Pearson's technique

r = 0.44213 (5 d.p)

```
n1 <-6
r <- cor(x.age,y.number)

t1 <- r/sqrt((1-(r*r))/(n1-2))
t1
```

T test = 0.986 (3 d.p) < 2.776

Decision: Fail to reject H0.

Conclusion:

At 0.05 significance level, there is insufficient evidence to reject H0. We can conclude that there is no linear relationship between the variables age group and number of same-sex couples.

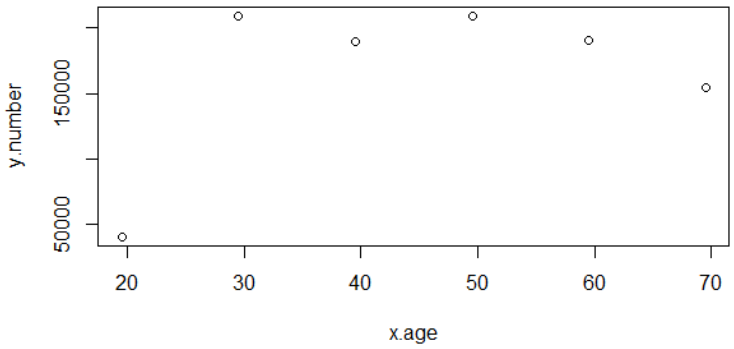


Diagram 1 shows the scatter plot of y.number (number of same-sex couples) against the x.age (age group)

Test 3: Regression Analysis

H0: $\beta = 0$ (No linear relationship)

H1: $\beta \neq 0$ (Linear relationship does exist)

Critical Region:

At 0.10 significance level, T (0.10/2, 3) = 2.353 or -2.353 ; degree freedom = n-2 = 3

Test Statistic:

Total number of same-sex couples = 995420

Income (’499.5)	22	42	62	87	112
Number of same-sex couples (%)	12.1	9.6	16.4	15.3	46.7

```
#linear regression model
income<-c(42,62,87,112,22)
numberp<-c(9.6,16.4,15.3,46.7,12.1)
model1<- lm(numberp~income)
model1

#draw graph with linear regression model
plot(income,numberp,xlab = "income ( '499.5)",ylab = "Number of couples(percentage
abline(model1)
```

```
Call:
lm(formula = numberp ~ income)

Coefficients:
(Intercept)       income
    -2.2591       0.3428
```

The regression equation is $\hat{y}$= -2.2591+ 0.3428 x

Hence, the slope of this model is 0.3428. b1= 0.3428

By using formula S_ε and S_{b1}

$$S_\varepsilon = \sqrt{\frac{SSE}{n-k-1}} \qquad S_{b1} = \frac{S_\varepsilon}{\sqrt{\sum x^2 - \frac{\sum x^2}{n}}}$$

$S\varepsilon = \sqrt{\frac{321.7015}{3}} = 10.355$ (3d.p) while Sb1 = 0.1453 (4d.p)

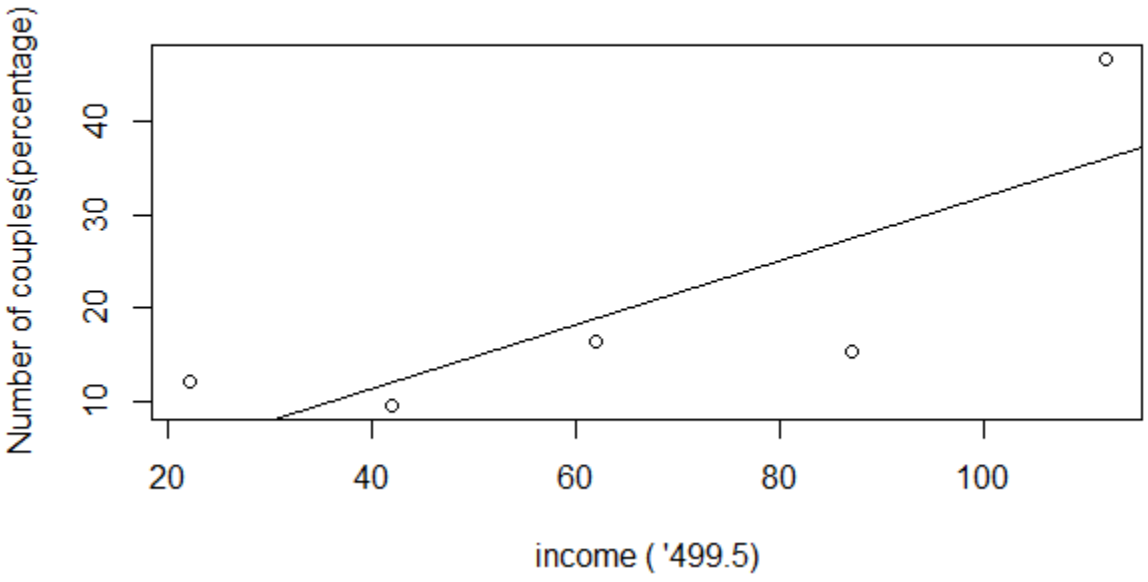
```
> se <- sqrt(sse/(n-k-1))
> sb<- se/ sqrt(sum(income^2)- ((sum(income)^2)/n))
> se
[1] 10.35538
> sb
[1] 0.1452894
```

T test = $\frac{b1-B1}{sb1} = \frac{0.3428}{0.1453} = 2.359$ (3 d.p) > 2.353

Decision: Reject H0

Conclusion:

At 90% confidence interval, there are sufficient evidence that the income affect the number of same-sex couples.



Test 4: Chi-squared test on Independence

H0: Gender of same-sex couples is independent of race.

H1: Gender of same-sex couples is dependent of race.

Critical Region:

```
> alpha <-0.05
> x2.alpha<-qchisq(alpha,df=6,lower.tail = FALSE)
> x2.alpha
[1] 12.59159
```

At α=0.05, X-squared (0.05,6) = 12.592 (3d.p)

Test Statistic:

Race	Male-male couple	Female-female couple
White	401149	406243
Black or African American	28619	53587
American Indian or Alaska Native	4366	4083
Asian	19888	13269
Native Hawaiian or Pacific Islander	970	1021
Some other race	15522	14800
Two or more races	14551	17352
Total	485065	510355

```
#perform chi-square test on data table
chisq.test(d,correct = FALSE)
alpha <-0.05
x2.alpha<-qchisq(alpha,df=6,lower.tail = FALSE)
```

```
        Pearson's Chi-squared test

data:  d
X-squared = 8573.8, df = 6, p-value < 2.2e-16
```

From R studio, we calculate that X-squared test = 8573.8 > 12.592

Decision: Reject H0

Conclusion:

At 0.05 significance level, there is sufficient evidence to conclude that gender of same-sex couples is dependent of race.

Conclusion

After carry out 4 tests, we can conclude that the proportion of female-female couples (lesbian) is more than 0.5 which also means that the proportion of male-male couples (gay) is less than 0.5 where by the proportion is calculated among LGBT groups. This is because the sum of proportion of female-female couples and male-male couples have to be 1. This claim also supported by the evidence that is videos and posts about lesbian are more commonly seen compared to gay. Secondly, I found out that there is no relationship between age group and the number of same-sex couples. The reason I carried out this test is because there is rumor state that when a single individual grows older, he or she has the higher probability to become gay or lesbian. Through this test, there is insufficient evidence to support the claim.

Thirdly, there is evidence to state that household income will affect the number of same-sex couples. An assumption can be made from this test that is with higher income, the higher the number of same-sex couples. This is because not every family can accept or stay with the family member who is LGBT, hence with a high income the individual (who has a same- sex couple) is able to earn a living for himself where he does not need to rely on his parents for living. Lastly, I found that gender of same-sex couple is dependent of race. Since I am a Malaysian and also an Asian too, I have made the assumption for the data in test 4 for Asian. Refer to the table, there is an abnormal situation for race Asian which the number of male-male couples is more than the female-female couples where for other races female-female couples are normally more than male-male couples. The assumption that I have make is that females in Asia is more shy, closed-minded and educated since young that they have to be polite and follow the rules so as when they grow up, they will be married to a nice guy and treats their husband politely. The opposite situation is the males in Asia have given more freedom to do what they want since young and parents want them to be brave so as to overcome the problem in their life as they will be the one who protect their

family members one day. Therefore, males in Asia will be brave to do what they want while female rarely do. Hence, this may be one of the reasons why Asian male-male couples is more than female-female couples.

Through this project I have learn that research can further interpreting data and make inference on the data by using hypothesis testing, correlation and regression analysis, Chi-squared test and many more. These analyses provide a more systematic and a clearer view for the research to interpret the data collected. Moreover, conclusion can be draw for the population based on the sample data collected by using hypothesis testing too. I am a computer science student where logical thinking and mathematical skill are both important to me as I have to carry out more analysis when I start to work. I hope that I can contributed to country and community by applying the knowledge learn in this course and project in the future.

Reference

1. https://www.census.gov/topics/families/same-sex-couples.html

2. https://www.bloomberg.com/news/articles/2018-10-25/malaysia-s-mahathir-says-asia-won-t-follow-west-on-lgbt-rights

3. https://analyticsindiamag.com/importance-of-hypothesis-testing-in-data-science/

4. https://news.gallup.com/poll/234863/estimate-lgbt-population-rises.aspx