UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

UNIVERSITI TEKNOLOGI MALAYSIA
SCHOOL OF COMPUTING
SESSION 2019/2020 SEMESTER 2

COURSE CODE

SECI 2143 – PROBABILITY & STATISTICAL DATA ANALYSIS

LECTURE'S NAME
DR CHAN WENG HOWE

PROJECT 2 – Heart Disease

STUDENT'S NAME
Shakirah Binti Mohd Shukur

MATRIC NO
A19EC0159

SECTION
02

Table of Contents

INTRODUCTION

Heart disease or cardiovascular disease (CVD) is a condition where the heart or blood vessels was affected due to certain condition. There are a few types of heart disease which are arrhythmia (abnormal heart rhythm), atherosclerosis (hardening of arteries), cardiomyopathy (heart's muscles become harden or weak), congenital heart defects (heart irregular since birth), coronary artery disease @ ischemic heart disease (plaque in the heart's arteries) and lastly heart infection (occur due to bacteria, viruses or parasites).

The symptoms usually different for each type. But mostly had a patient that had diabetes, high blood pressure, high cholesterol and had bad physical activity, such as smoking are easy to get heart disease. Sometimes, due to the age and family history also lead the patient towards to cardiovascular disease. According to the latest Statistics on Cause of Death which was released by the Department of Statistics Malaysia in 2018 about 18,267 deaths approximately 15.6% from the total of deaths.

For this case study, I used data from https://www.kaggle.com/volodymyrgavrysh/heart-disease that collected by VolodymyrGavrysh. This dataset provides information on the risk factors for heart disease. Experiments with the Cleveland database have concentrated on attempting to distinguish presence (value 1) or absence (value 0) of heart disease in the patient. There are 303 data collected (population) and I use only 165 (sample) from the data. I picked the data that patient has heart disease.

All the test and analysis will be conducted by using only a few variables from the data. For hypothesis testing 1-sample I used max heart rate. Age and gender will be used for a chi square test of independent. Meanwhile, for regression and correlation I used max heart rate and cholesterol level. Lastly, I tested rest blood pressure by using goodness of fit test.

CONTENT

Hypothesis 1 sample test

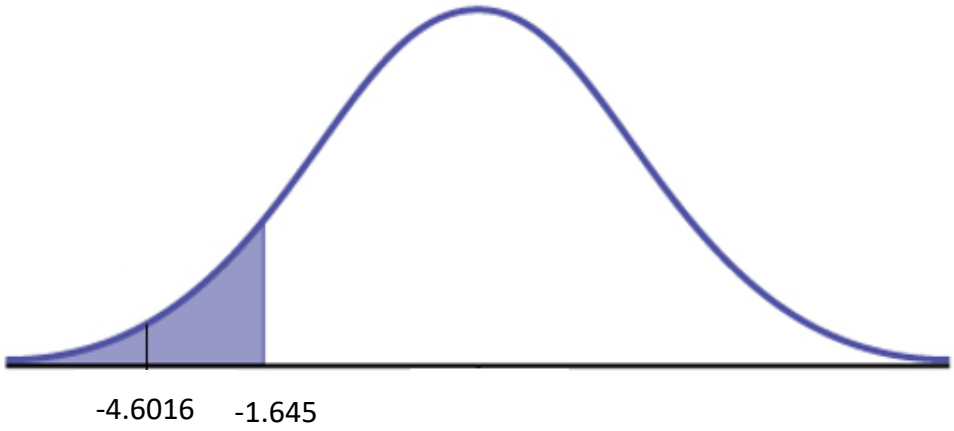
Variable: Max heart rate

Ho: $\mu = 3.29$

H1: $\mu < 3.29$

$\alpha = 0.05$

critical value = $Z_{(0.05)} = -1.645$ (Left-tailed test)



Conclusion:

Based on the R script, the test statistic value (z = -4.601597) < than the critical value (z = -1.645) and did not fall in the critical region. Hence reject Ho. There are sufficient evidences that the mean of max heart rate is less than 3.29.

Chi square

Variable: Age and gender

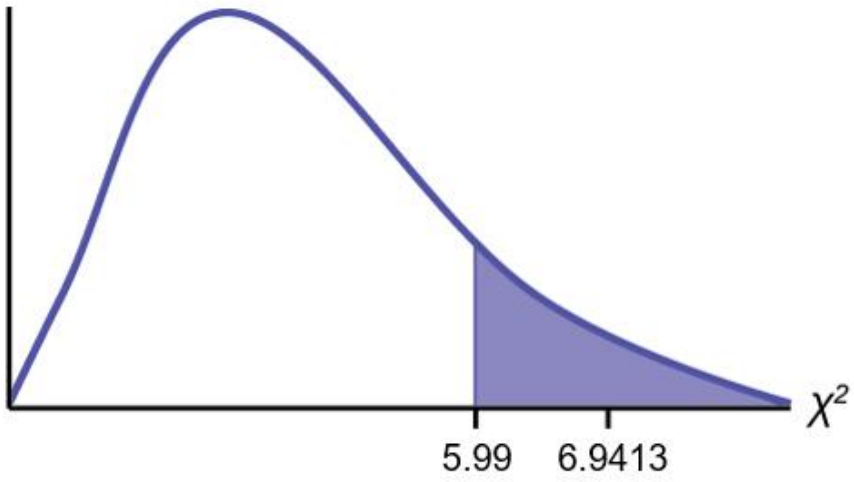
Ho: There are no relationship between the age and gender

H1: There are relationship between the age and gender

$\alpha = 0.05$

df = (3 -1)(2 – 1) = 2

critical value = $x^2(2, 0.05) = 5.99$

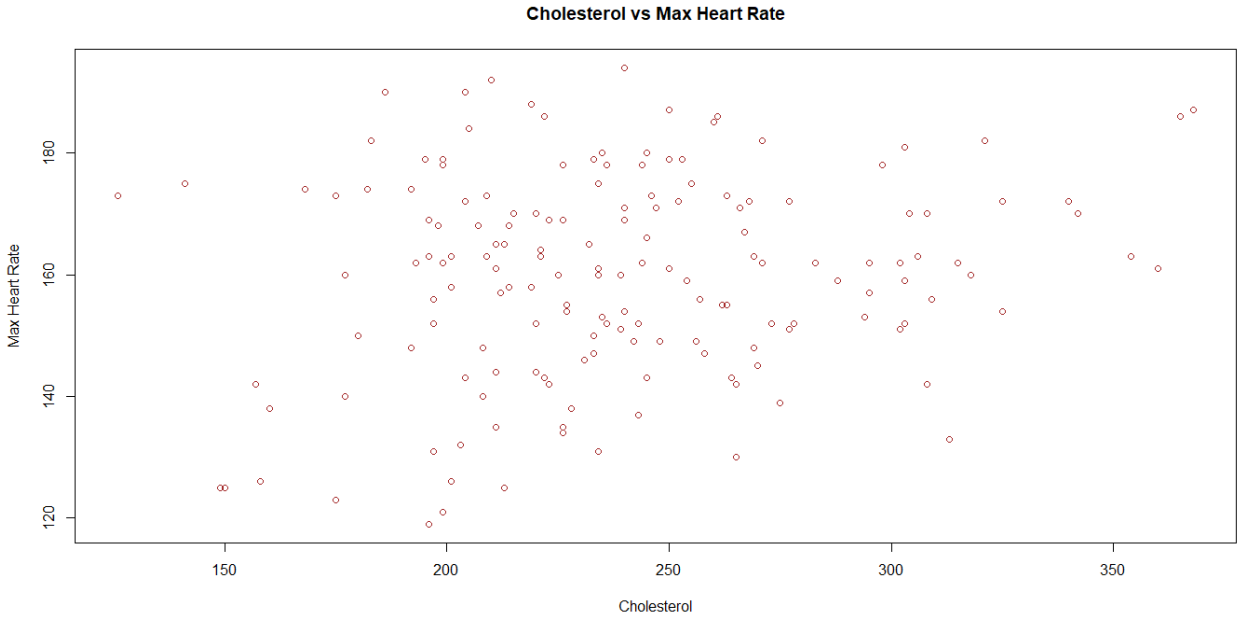


Conclusion:

Based on the R script, the test statistic value ($x^2 = 6.9413$) > than the critical value ($x^2 = 5.99$) and did fall in the critical region. Hence reject Ho. There are sufficient evidences that there are relationship between the age and gender.

Correlation

Variable: Cholesterol level and max heart rate



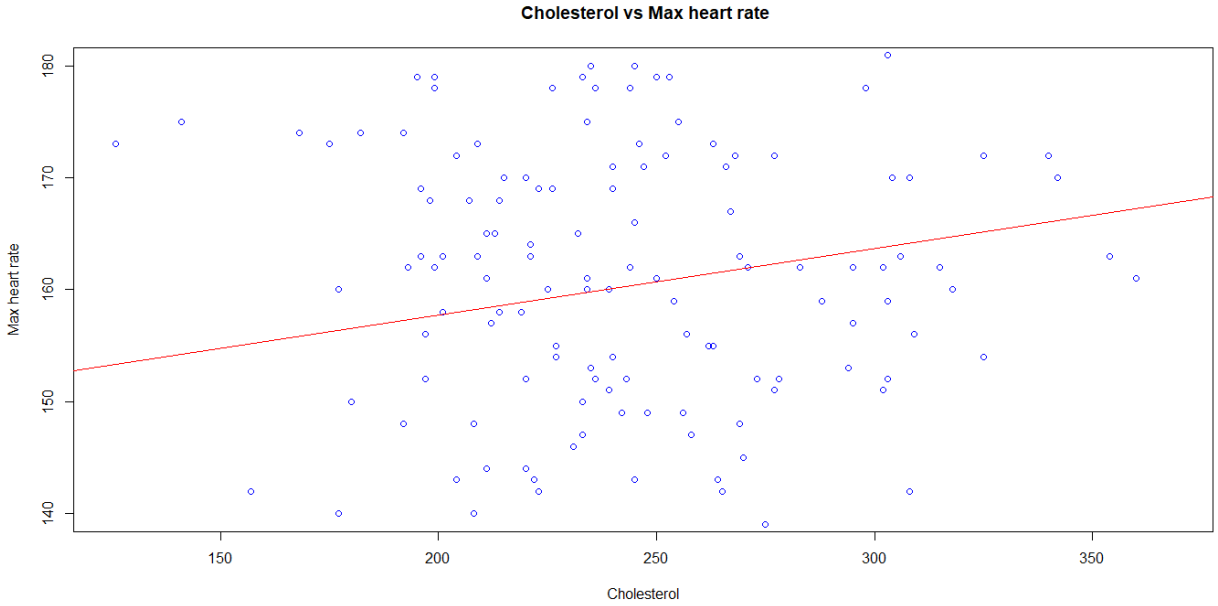
Conclusion:

Based on the graph, there is positive linear and weak relationship between the cholesterol level and max heart rate. The value of linear regression model is $\hat{y}$ = 145.85084 + 0.05931x. The relationship between the variable is 0.1654.

```
> cor(x, y)
[1] 0.1653741
> model <- lm(y ~ x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-38.476 -11.281   0.227  11.550  33.914

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 145.85084    6.73834  21.645   <2e-16 ***
x             0.05931    0.02771   2.141   0.0338 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.49 on 163 degrees of freedom
Multiple R-squared:  0.02735,   Adjusted R-squared:  0.02138
F-statistic: 4.583 on 1 and 163 DF,  p-value: 0.03377
```

Regression

Variable: Cholesterol level and max heart rate



Conclusion:

Based on the graph, there is a positive linear relationship. The cholesterol level will affect the max heart rate of the patient. Some but not all of the variation in max heart rate is explained by the variation in cholesterol level. This relationship has a large standard error.

Goodness of fit test

Variable: Resting blood pressure

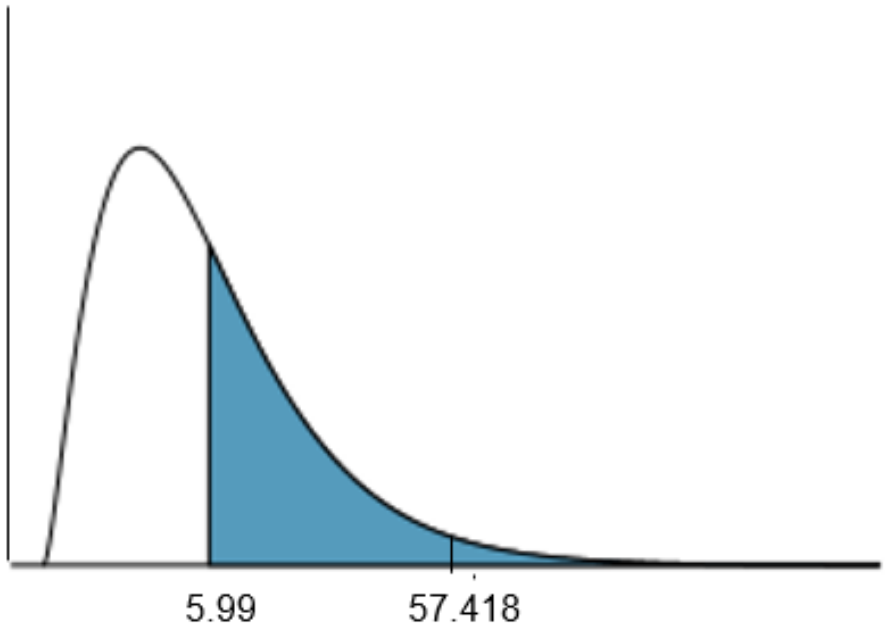
Ho: $P1 = P2 = P3$

H1: At least one of the 3 proportion is different from the others.

$\alpha = 0.05$

df = 3-1 = 2

critical value = $x^2(2, 0.05) = 5.99$



Conclusion:

Based on the R script, the test statistic value ($x^2 = 57.418$) > than the critical value ($x^2 = 5.99$) and fall in the critical region. Hence reject Ho. There are sufficient evidences that at least one of the 3 proportion is different from others.

CONCLUSION

Based on the hypothesis test that's been carried out. There are a many aspects that lead to the heart disease. All the hypothesis that been testes had come to the conclusion which had relationship between x variable and y variable. For correlation and regression, we can see from the graph that both is depending each other. The patient will have a high max heart rate if they had higher cholesterol level. But, the data lead to a weak relationship that can cause a large standard error. Some will not obey to the hypothesis. The value of linear regression model ($\hat{y}$ = 145.85084 + 0.05931x) show that 145.85084 is the estimated value if the x is equal to 0.

REFERENCES

1. https://www.thestar.com.my/news/nation/2019/10/31/heart-attack-leading-cause-of-death#:~:text=Ischaemic%20heart%20diseases%2C%20which%20occur,of%20Death%20in%20Malaysia%20released

2. https://www.youtube.com/watch?v=aHzdUFbMfMo&t=1s

3. https://www.youtube.com/watch?v=VNG7MtXidrg&t=374s