UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF ENGINEERING

SCHOOL OF COMPUTING

SEMESTER 2/20212022

SECI2143 KEBARANGKALIAN STATISTIK & ANALISIS DATA (PROBABILITY & STATISTICAL DATA ANALYSIS)

SECTION 06

PROJECT 2

Project Video Link

https://youtu.be/_NtKEqQP8rM

LECTURER: DR ARYATI BINTI BAKRI

GROUP NO: 4

NAME	MATRIC NO
FIKRI AKMAL AIZUDDIN BIN BAHRIM	A21EC0025
NURAIN NAJWA BUKARI	A21EC0117
MUHAMMAD FAHMI BIN ROSLEE	A21EC0285
IQMAL AIZAT BIN MOHD ZAMRI	A21EC0032

TABLE OF CONTENTS

1. INTRODUCTION

Heart attacks have often been common events caused by a clog that prevents blood from flowing to the heart or brain. The most frequent cause is an accumulation of fatty deposits on the inner walls of blood arteries that supply the heart or brain. Strokes can be caused by blood clots or bleeding from a blood artery in the brain. In 2019, an estimated 17.9 million individuals died from Cardiovascular, accounting for 32% of all global fatalities. 85 percent of these fatalities were caused by a heart attack or a stroke. We want to identify how to overcome this health problem by finding the factors to help them. Factors we can observe from age, chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved and heart attack. To accomplish the goal, we have chosen a few variables and some test studies are being conducted.

The objective of this case study is for gaining knowledge on how to use RStudio and R programming language for statistical purposes. In this case study, we want to obtain the data of the Heart Attack Analysis & Prediction Dataset through secondary data collection from Kaggle. There are about 96 females with age range of 34 to 76 and 207 male with age range of 29 until 77 who have been diagnosed to have heart attacks. We use this dataset to describe the pattern or relationship between different variables. We use 303 samples from respondents in the dataset for our project to draw inferences, hypotheses and identify characteristic and population patterns. On top of that , we also want to determine whether there is a significant relationship between chest pain type (Cp) and gender using Chi Square test of independence.

2. DATASET

In this project, the data of Heart Attack Analysis & Prediction Dataset was obtained through secondary data collection from a trustworthy website. This dataset aims to predict various factors that caused heart attack disease among respondents. This study was important for the people who have similar factors of heart attack disease to take some precautions for their health.

The data which are respective to the 303 respondents are selected from the database and used for hypothesis testing to determine whether there is enough evidence to support the null hypothesis for hypothesis testing, correlation, regression and chi square. The sample is normally distributed and plotted with RStudio.

Variables	Type of Variable	Level of Measurement
Age	Quantitative	Ratio
Sex	Qualitative	Nominal
Chest Pain	Qualitative	Ordinal
Resting blood pressure	Quantitative	Ratio
Cholesterol in mg/dl fetched via BMI sensor	Quantitative	Ratio
Fasting blood sugar > 120 mg/dl	Qualitative	Nominal
Resting electrocardiographic result	Qualitative	Nominal
Maximum heart rate achieved	Quantitative	Ratio
Heart attack	Qualitative	Nominal

1. Age

Age	Frequency	Age	Frequency	Age	Frequency
29	1	48	7	62	11
34	2	49	5	63	9
35	4	50	7	64	10
37	2	51	12	65	8
38	3	52	13	66	7
39	4	53	8	67	9
40	3	54	16	68	4
41	10	55	8	69	3
42	8	56	11	70	4
43	8	57	17	71	3
44	11	58	19	74	1
45	8	59	14	76	1
46	7	60	11	77	1
47	5	61	8		

2. Gender

Gender	Frequency
Male	207
Female	96

3. Chest Pain

Chest Pain	Frequency
Typical angina	143
Atypical angina	50
Non-anginal pain	87
Asymptomatic	23

4. Resting Blood Pressure

Resting Blood Pressure	Frequency
81 - 100	6
101 - 120	91
121 - 140	141
141 - 160	50
161 - 180	13
181 - 200	2

5. Cholesterol

Cholesterol	Frequency
101 - 200	51
201 - 300	209
301 - 400	39
401 - 500	3
501 - 600	1

6. Fasting Blood Sugar

Fasting Blood Sugar	Frequency
True	258
False	45

7. Resting Electrocardiographic Results

Resting Electrocardiographic Results	Frequency
Normal	147
Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)	152
Showing probable or definite left ventricular hypertrophy by Estes' criteria	4

8. Maximum Heart Rate Achieved

Maximum Heart Rate Achieved	Frequency
61 - 80	1
81 - 100	7
101 -120	29
121 - 140	54
141 - 160	102
161 - 180	92
181 - 200	17
201 - 220	1

9. Heart Attack

Heart Attack	Frequency
More Chance	165
Less Chance	138

3. DATA ANALYSIS

Part A

One Sample Hypothesis Testing

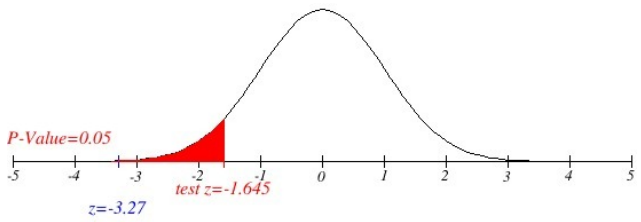
A group of researchers conducted a survey to study the relation between resting blood pressure and cholesterol. A random sample of 303 respondent data has been obtained and measured. 0.05 significance level is used to test the claim that the sample mean of resting blood pressure is lesser than the population mean which is 135.

1. Hypothesis Statement
   H_0 : $\mu = 135$
   H_1 : $\mu < 135$

2. The z test statistics, Z_0 with n > 30
   $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

   $\bar{X} = 131.62$
   $\mu = 135$
   $\sigma = 17.51$
   n = 303
   Test statistics, $Z_0 = -3.270$

3. Critical value, Z_α with 95% confidence level :
   Critical value, $Z_\alpha = -1.645$



4. Conclusion :
   Since Test statistics, $Z_0 = -3.270$ < Critical value, $Z_\alpha = -1.645$, thus we do not reject H_0
   There is not sufficient evidence that the mean of resting blood pressure is less than 135.

Correlation test

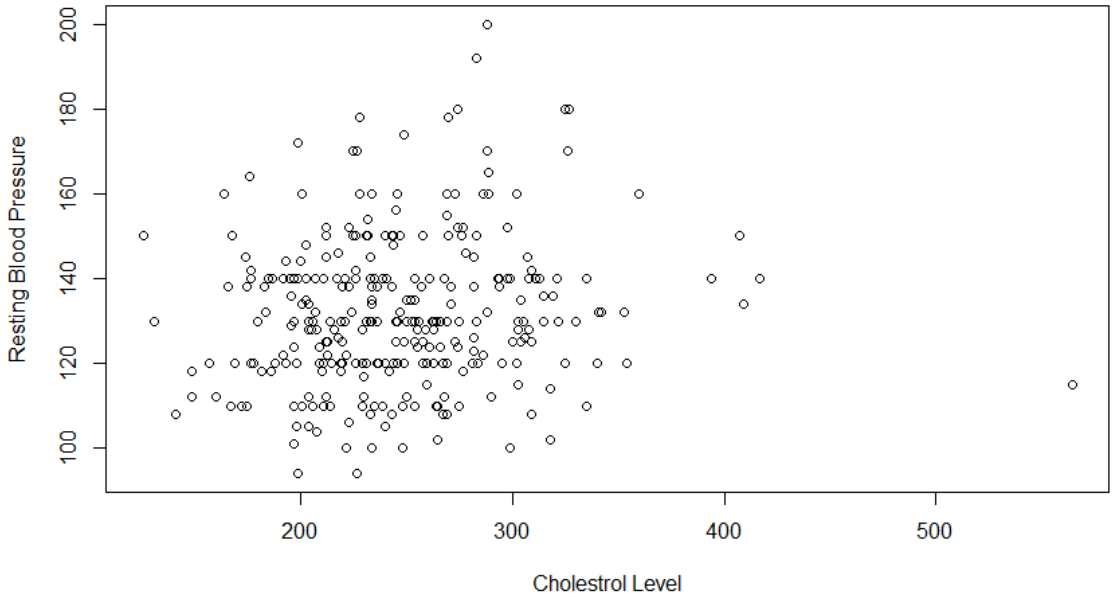
$H_0 : p = 0$ (no linear correlation exist)

$H_1 : p \neq 0$ (linear correlation exist)

X - variable : Cholesterol
Y - variable : Resting blood pressure

R = 0.1232 (relatively weak positive linear association between x and y)
Test statistic, t = 2.154
T critical value = 1.968

Reject H_0. Test statistic > T critical value (2.154 > 1.968). There is sufficient evidence of a linear relationship between cholesterol level and resting blood pressure at the 5% level of significance.

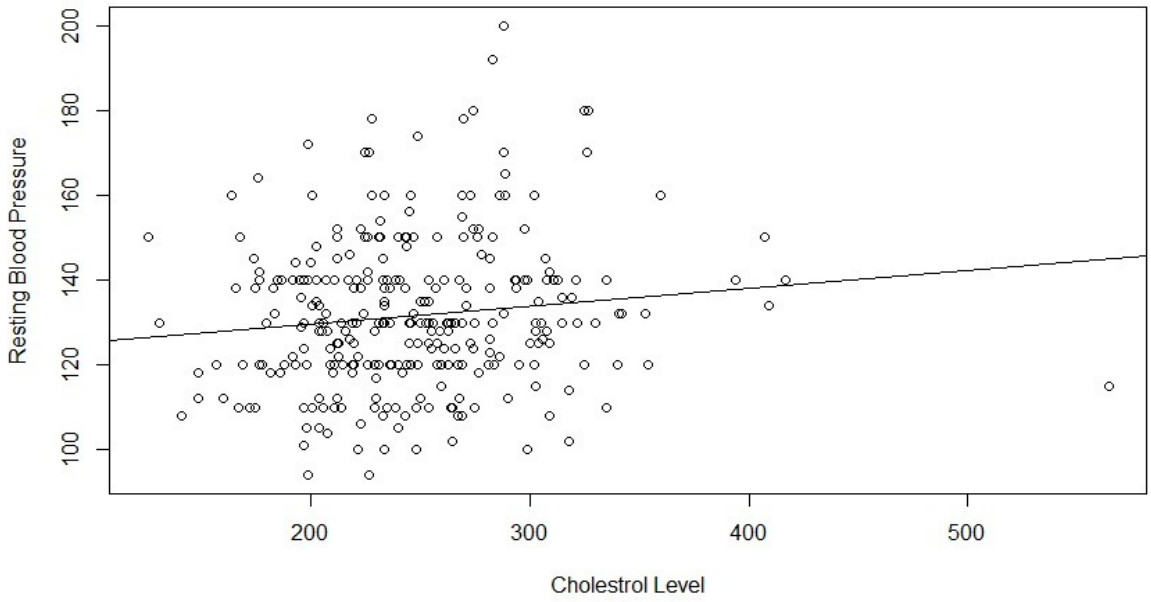


Regression test

For a regression test, we want to see if there is a relationship between the Resting Blood Pressure and Cholesterol level.

The line we received is
$\hat{y} = 121.35976 + 0.04168x$, where
Intercept(b_0) = 121.35976 and slope(b_1) = 0.04168

b_0 is the estimated average value of resting blood pressure when the value of Cholesterol level is 0, but in our data, there is no sample with value of cholesterol level is 0, therefore 121.35976 is the portion of blood pressure that is not explained by the cholesterol level. b_1 value tells us that for every increase of resting blood pressure, there will be an increase of 0.04168 on the level of cholesterol.



$R = 0.1231742$
The correlation value indicates that there is a positive relationship between cholesterol level and resting blood pressure, that would mean that the higher the cholesterol level, the higher the value of resting blood pressure.

$R^2 = 0.0151719$
This value we obtained would mean that the relation between resting blood pressure and cholesterol level is weak, as it is far from 1.

Inference about the Slope: t Test

We want to test the claim that there is no regression relationship between resting blood pressure and cholesterol level, at 95% of confidence level.

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

$\alpha = 0.05, n = 303, \hat{y} = 121.35976 + 0.04168x, df = 301$

Critical value,
$-\ t_{0.025,\,301} = -\ 1.967877$
$t_{0.025,\,301} = 1.967877$

$t = \frac{b_1 - \beta_1}{S_{b1}}$

By using Rstudio,
$S_\varepsilon = 17.43$
$S_{b1} = 0.01935$
$t_0 = 2.154$

Since $t_0 = 2.154 > t_\alpha = 1.967877$, we reject H_0

Thus, there is enough evidence that a linear regression exists between resting blood pressure and cholesterol level.
There is sufficient evidence that resting blood pressure affects cholesterol level.

Part B

Chi-squared test

For chi-square, we're looking to see if there is a connection between two qualitative variables which are chest pain and gender among respondents. The data sample of 303 from chest pain was divided into respective categories based on the frequency of gender between male and female. We claim that, at 95% of significant levels, the chest pain is independent of gender.

H_0 = Chest pain is independent of gender

H_1 = Chest pain is not independent of gender

We are using this formula to calculate expected count, $e_{ij} = \frac{(i^{th}\ Row\ Total)(j^{th}\ Column\ Total)}{Total\ Sample\ Size}$

And for test static, x^2 we are using this formula and apply it into Rstudio: $x^2 = \frac{\Sigma(O_{ij} - E_{ij})^2}{E_{ij}}$

Chest pain	Gender		TOTAL
	Male	Female	
Typical angina	104	39	143
Atypical angina	32	18	50
Non-anginal pain	52	35	87
Asymptomatic	19	4	23
TOTAL	207	96	303

Table 1

From the RStudio, we obtained, $x^2 = 6.8221$, the degree of freedom, $df = 3$, and the $p - value = 0.07779$. Next, by referring to the Chi-Square Distribution table, provided with $\alpha = 0.05$, $df = 3$, we achieve the value of 7.814728.

From the values that we have received, we can confirm that the critical value is larger than the test value ($7.814728 > 6.8221$), since the value of test value is smaller than the critical value, thus we do not reject H_0 and there is sufficient evidence that supports that chest pain is independent of gender.

4. CONCLUSION

In one sample hypothesis testing, we measured the cholesterol level and the sample mean of resting blood pressure which is 135. In our test, we have conclude that test statistics, Z_0 = -3.270 < Critical value, Z_α = -1.645, thus we do not reject H_0 because there is not sufficient evidence that the mean of resting blood pressure is less than 135.

Next, in the correlation test, we measured to see if a linear correlation exists between cholesterol levels and resting blood pressure. In our testing, there indeed exists a relatively weak linear correlation with the correlation coefficient, r - 0.1232. Thus, we reject the null hypothesis that indicates no linear correlation exists between the two variables.

In the regression test, the line that we received is $\hat{y} = 121.35976 + 0.04168x$, where Intercept($b_0$) = 121.35976 and slope($b_1$) = 0.04168. By using this value, we can find the value of r = 0.1231742 . The correlation value indicates that there is a positive relationship between cholesterol level and resting blood pressure, that would mean that the higher the cholesterol level, the higher the value of resting blood pressure.

Lastly, in the chi- square test, we're looking to see if there is a connection between two qualitative variables which are chest pain and gender among respondents. After testing, we get that the critical value is larger than the test value ($7.814728 > 6.8221$), which means that we reject H_0.

To conclude, this project has enhanced our abilities to use different kinds of formulas and techniques that have been taught by our Lecturer, Dr. Aryati Binti Bakri to analyze data in a population or a large sample and make conclusions from it.

5. APPENDIX

Rahman Rashik. “Heart Attack Analysis & Prediction Dataset.” *Kaggle*, 2021, https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

World Health Organization. “Cardiovascular diseases (CVDs).” 11 June 2021, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

e-Portfolio Reflection

NAME & MATRIC NO	LINK
FIKRI AKMAL AIZUDDIN BIN BAHRIM A21EC0025	https://eportfolio.utm.my/user/fikri-akmal-aizuddin-bin-bahri/seci2143-probability-statistical-data-analysis-1
NURAIN NAJWA BUKARI A21EC0117	https://eportfolio.utm.my/user/nurain-najwa-bukari/seci2143-06-probability-statistical-data-analysis
MUHAMMAD FAHMI BIN ROSLEE A21EC0285	https://eportfolio.utm.my/user/muhammad-fahmi-bin-roslee/seci-2143-probability-statistical-data-analysis
IQMAL AIZAT BIN MOHD ZAMRI A21EC0032	https://eportfolio.utm.my/user/iqmal-aizat-bin-mohd-zamri/psda-probability-and-statistic-data-analysis