UTM
UNIVERSITI TEKNOLOGI MALAYSIA

School of Computing
Faculty of Engineering

Semester 2
Session 2019/2020

CODE & SUBJECT :

SECI2143

PROBABILITY & STATISTICAL DATA ANALYSIS

Project 2

NAME : LOH YEW CHONG

MATRIC NUMBER : A19EC0076

TITLE : PREDICTION OF HEART ATTACK AND THE STUDY OF THE FACTORS THAT MAY CAUSE HEART ATTACK

NAME OF LECTURER : DR. CHAN WENG HOWE

SECTION : 02

CONTENTS

Introduction

The data for this dataset was collected by Robert Detrano V.A.Medical Center,Long Beach and Cleveland Clinic Foundation, Andas Janosi from Hungrian Institute of Cardiology,Budapest , William Steinbrunn from University Hospital, Zurich, Switzerland , and Matthias Pfisterer from University Hospital , Basel,Switzerland. The dataset is for prediction of heart attack.

Some dataset's data is missing and it is in “?” symbol in the dataset. Hence I will delete the data that is represented with the “/” symbol when I would like to use the variable for my inferential statistic . For example, i will remove the row of data that consists of the ‘?” symbol when the symbol is located at any column of variable that I might use for my hypothesis testing and others.

This study is for the prediction of the heart attack. According to an internet article entitled “Heart attack” by MAYO CLINIC , it states that the men with the age of 45 or older and women with the age of 55 or older are more likely to have a heart attack than are younger men and women. Hence , I would like to test whether the age for men to get heart attack is 45 or older and the mean age for women to get heart attack is 55 or older . Besides that , I would like to study how the serum cholesterol in mg/dl (chol) affects the trestbps(resting blood pressure) which is by correlation analysis. Undeniably , I would also like to study the relationship between the resting blood pressure(trestbps) with the thalach maximum heart rate achieved (thalach) which is by regression analysis . On top of that , I would also like to determine whether there is a significant relationship between chest pain type (Cp) and resting electrocardiographic results (Restecg) resting electrocardiographic results using Chi Square test of independence .

Hypothesis testing

Test 1 : 1 sample test to test the mean age of men getting heart attack

This 1 sample testing is to test whether the statement is true that the mean age of men getting heart attack is 45 years old or older . Assume the confidence level to be 95%, significant level , α = 0.05 . Let the population mean age of men getting heart attack be μ .

H_0: μ = 44

H_1: μ > 44

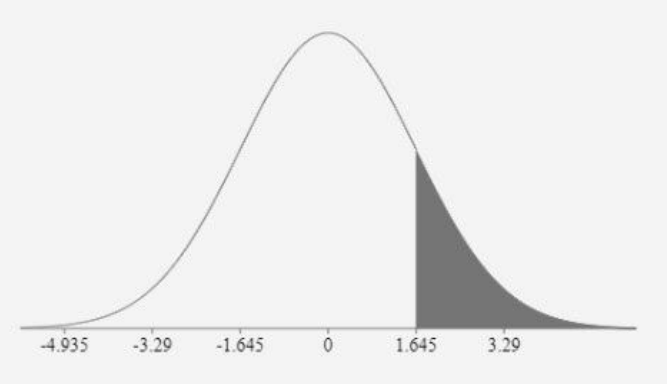
α = 0.05

Using the Rstudio ,

Sample size , n = 196

Sample mean , $\bar{X}$ =47.78571

Sample Standard deviation, σ = 7.881494



The shaded region is the rejection region.

Test statistic, $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$, z = 6.724613

Critical value , c.v = $z_{0.05}$ = 1.644854 = 1.645

Decision making - using critical region :

Since the test statistic value = 6.724613 is greater than the critical value = 1.644854 which falls within the critical region. Hence, we reject the null hypothesis.

Conclusion :

There is sufficient evidence to prove that the mean age of men getting heart attack is greater than 44 which is in other words the mean age of men getting heart attack is 45 or older.

Test 2 : 1 Sample testing to test the mean age of women getting heart attack

This 1 sample testing is to test whether the statement is true that the mean age of women getting heart attack is 55 years old or older . Assume the confidence level to be 95%, significant level , α = 0.05 . Let the population mean age of women getting heart attack be μ.

H_0: μ = 54

H_1: μ > 54

α = 0.05

Using the Rstudio ,

Sample size , n = 73

Sample mean , $\bar{X}$ = 47.65753

Sample Standard deviation, σ = 7.652558

Test statistic, $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$, z = -7.081299

P-Value, P(-7.081299) = 7.14047e-13

Decision making - P-value Method :

Since the p-value which equals 7.14047e-13 is smaller or lesser than significance level , α = 0.05. Hence , the null hypothesis is rejected.

Conclusion :

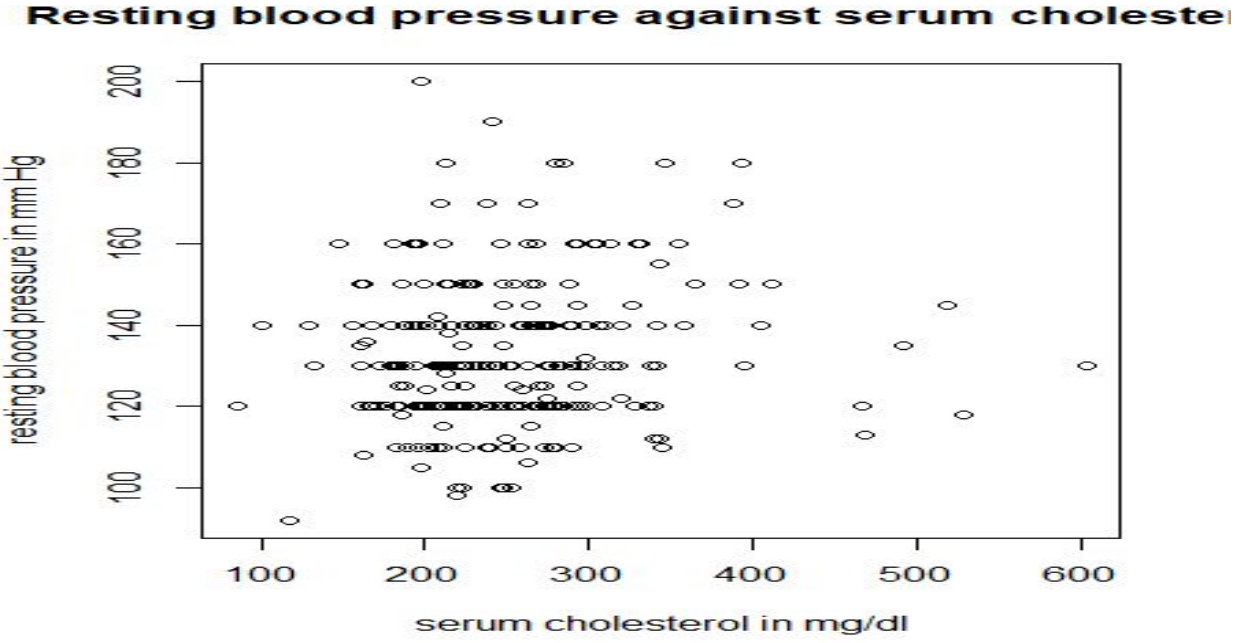
There is sufficient evidence to prove the mean age of women getting heart attack is greater than 54 or in other words the mean age of women getting heart attack is 55 years old or older.

Test 3 : Correlation Analysis to investigate the relationship between the serum cholesterol mg/dl and the resting blood pressure in mmHg.

This test is to measure the strength of the relationship between the serum cholesterol mg/dl and the resting blood pressure in mmHg. Assume the confidence level to be 95%, significant level , α = 0.05 .

H0: ρ = 0 (no linear correlation between the serum cholesterol mg/dl and the resting blood pressure in mmHg.)

H1: ρ ≠ 0 (linear correlation exists between the serum cholesterol mg/dl and the resting blood pressure in mmHg.)



The independent variable is the serum cholesterol in mg/dl while the dependent variable is the resting blood pressure in mmHg. Since both the variables are ratio scale data , hence I can use Person's product-moment correlation using cor.test() function in R to obtain the correlation efficient (r).

α = 0.05
α = 0.0 as it is 2 tail test

Using Rstudio,

Correlation coefficient, r = 0.833139.

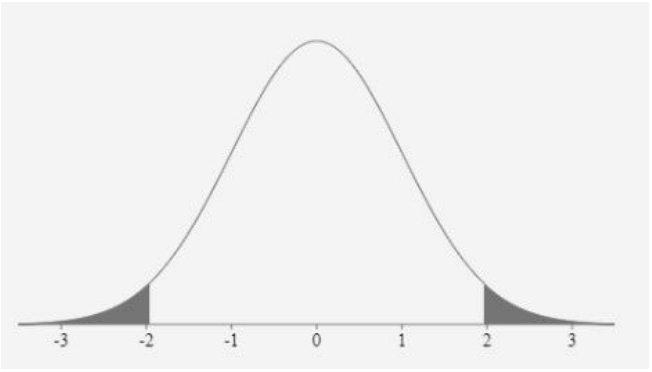
Sample size = 269

Degree of freedom , df = 267

Test statistic ,

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

t = 1.43352



The shaded region is the rejection region.

Critical value ,
$-t_{0.025,267}$ = -1.968066
$t_{0.025,267}$ = 1.968066

The program round up the value , hence the critical shown in the normal distribution is -2 and 2 instead of -1.968066 and 1.968066

Decision :
Since the test statistic , t = 1.43352 is in between the $-t_{0.025,267}$ = -1.968066 and $t_{0.025,267}$ = 1.968066. It does not fall within the rejection region. Hence we failed to reject the null hypothesis.

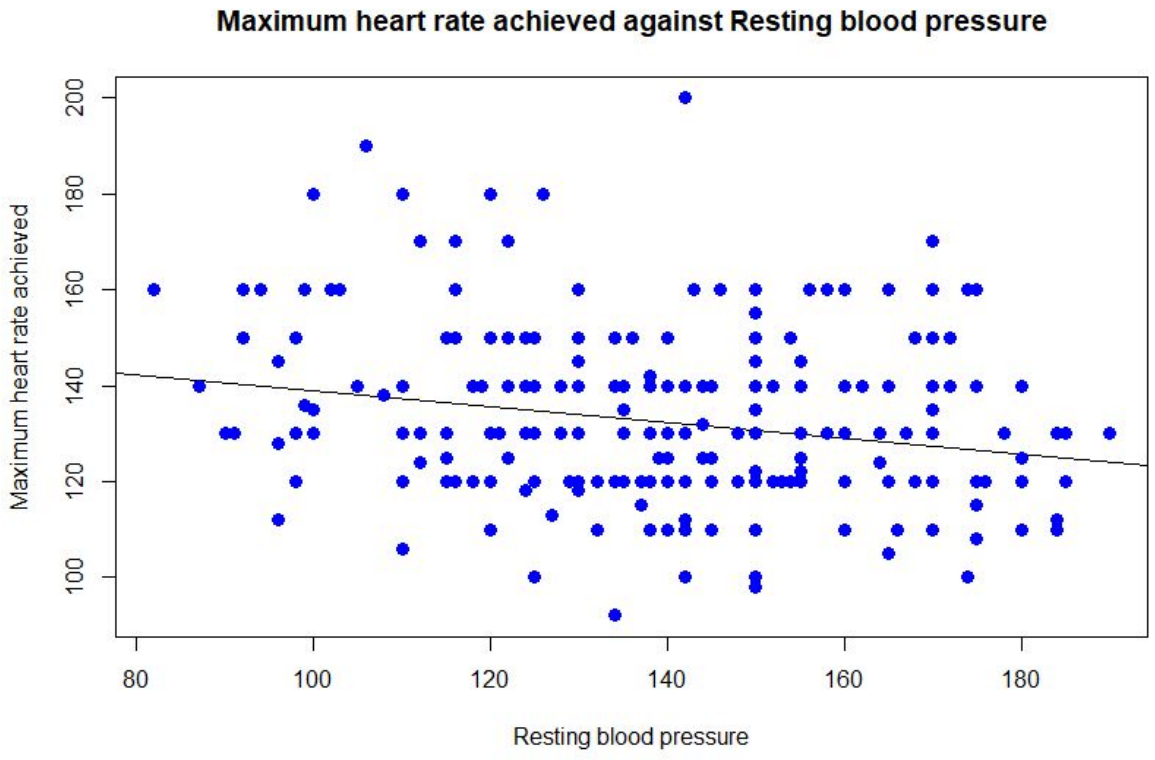
Conclusion
Since the correlation coefficient, r = 0.833139 which is positive and falls within 0.8 and1 , hence it has a strong linear relationship between the serum cholesterol mg/dl and the resting blood pressure in mmHg. There is sufficient evidence to prove that there is no linear correlation between the serum cholesterol mg/dl and the resting blood pressure in mmHg.

Test 4 : Regression analysis to investigate the relationship between the resting blood pressure with the maximum heart rate achieved

Assume the confidence level to be 95%, significant level , α = 0.05 .

H_0: $\beta 1 = 0$ (no linear regression between resting blood pressure with the maximum heart rate achieved)

H_1: $\beta 1 \neq 0$ (linear regression exists between the resting blood pressure with the maximum heart rate achieved)



The independent variable (variable that used to explain the dependent variable) is resting blood pressure while the dependent variable (variable that i wish to explain) is maximum heart rate achieved .

The r-squared value of regression, R^2 is 0.04797.

This shows that there is only 4.797% of the variation maximum heart rate achieved is explained by the resting blood pressure.

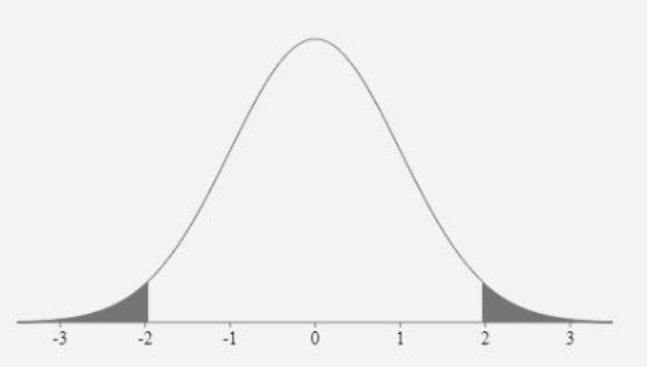
α = 0.05
Sample size , n = 269

Degree of freedom , dof = 267

The regression line ,

$\hat{y}$ = 177.95 - 0.293x.

Test statistic
Degree of freedom , df = 267



Critical value ,
$-t_{0.025,267}$ = -1.968066
The shaded region is the rejection region.
$t_{0.025,267}$ = 1.968066

The shaded region is rejection area
The program round up the value , hence the critical shown in the normal distribution is -2 and 2 instead of -1.968066 and 1.968066

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

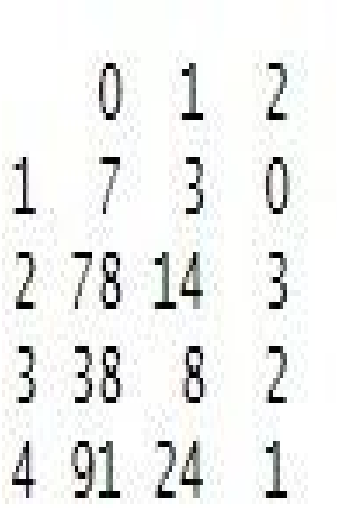
Test statistic ,
t = -3.668461

Decision :
Since the test statistic , t = -3.668461 is smaller than $-t_{0.025,267}$ = -1.968066 and $t_{0.025,267}$ = 1.968066. It falls within the rejection region. Hence, we reject the null hypothesis.

Conclusion :

There is sufficient evidence that linear regression regression exists between the serum cholesterol mg/dl and the resting blood pressure in mmHg.There is enough evidence that resting blood pressure affect the maximum heart rate achieved.

Test 5 : Chi Square test of independence to determine whether there is a significant relationship between chest pain type and resting electrocardiographic results.



	0	1	2
1	7	3	0
2	78	14	3
3	38	8	2
4	91	24	1

Assume the confidence level to be 95%, significant level , α = 0.05 .

H_0 : The chest pain type and the resting electrocardiographic results are independent.

H_1: The chest pain type and the resting electrocardiographic results are dependent.

← contingency table obtained in Rstudio at first

The contingency table obtained at first

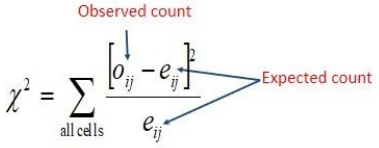
	Resting electrocardiographic results		
	normal	having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)	showing probable or definite left ventricular hypertrophy
typical angina	7	3	0
typical angina	78	14	3
non-anginal pain	38	8	2
asymptomatic	91	24	1

Contingency table that used for test statistic :

```
   [,1]
1    10
2    92
3    46
4   115
```

```
        Chi-squared test for given probabilities

data:  tbl3
X-squared = 101.04, df = 3, p-value < 2.2e-16
```



Test statistic ,

χ^2= 101.04

The degree of freedom is 3 .
α = 0.05

Critical value = 7.814728

The p-value is 2.2e-16.

Decision:

Since the test statistic ,χ^2= 101.04 is greater than the critical value which is 7.814728 . It falls within the critical region . Besides that the p-value obtained is 2.2e-16 is smaller than 0.05 . Hence, we reject the null hypothesis.

Conclusion :There is enough evidence that the chest pain type and the resting electrocardiographic results are dependent.

Discussion

The one sample testing of the mean age of men getting age proves that the mean age of men getting heart attack is greater than 44 or in other it is 45 years old and older. While for the one sample testing of the mean age of women getting age, it proves that the mean age of women getting heart attack is greater than 54 or in other words it is 55 years old and older.

Conclusion

In conclusion, the statement that state that “Men age 45 or older and women age 55 or older are more likely to have a heart attack than are younger men and women” are true and this already been supported with one sample testing. There are strong relationship between between serum cholesterol in mg/dl)and resting blood pressure .There is a positive linear relationship between the resting blood pressure with the maximum heart rate achieved. The chest pain type and resting electrocardiographic results are dependent . Hence they affect each other.

References

Source of data :

https://www.kaggle.com/imnikhilanand/heart-attack-prediction

To get the population mean age of men and women getting heart attack :

(Unknown).(n.d.).*Heart attack* .Retrieved from MAYO CLINIC : https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106