SEMESTER 2

SESSION 2019/2020

PROBABILITY & STATISTICAL DATA ANALYSIS

SECI2143-02

PROJECT 2:

FRAMINGHAM HEART STUDY

STUDENT'S NAME

TEE HUI YOU (A19EC0170)

LECTURER'S NAME

DR CHAN WENG HOWE

SUBMISSION DATE: **27TH JUNE 2020**

Table of Contents

1.0 INTRODUCTION

1.1 DESCRIPTION OF DATASET

In project 2, we were required to apply inferential statistics on a certain secondary data obtained from the open source. Various hypothesis tests were required to be performed in the inferential statistics. Hence, a dataset with many numerical variables are needed. After searching through dataset from websites, I decided to use dataset of the Framingham Heart Study. The image attached below is a preview of my dataset.

GENDER	AGE	EDUCATIO	Current Sn	Cigarettes	Blood Pres	Stroke	Hypertens	Diabetes	Total Chol	systolic BP	diastolic B	BMI	Heartbeat	glucose	TenYearCHD
Male	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
Female	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
Male	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
Female	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
Female	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
Female	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
Female	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
Female	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
Male	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0
Male	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0
Female	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0
Female	43	2	0	0	0	0	0	0	247	131	88	27.64	72	61	0
Male	46	1	1	15	0	0	1	0	294	142	94	26.31	98	64	0
Female	41	3	0	0	1	0	1	0	332	124	88	31.31	65	84	0
Female	39	2	1	9	0	0	0	0	226	114	64	22.35	85	NA	0
Female	38	2	1	20	0	0	1	0	221	140	90	21.35	95	70	1
Male	48	3	1	10	0	0	1	0	232	138	90	22.37	64	72	0
Female	46	2	1	20	0	0	0	0	291	112	78	23.38	80	89	1
Female	38	2	1	5	0	0	0	0	195	122	84.5	23.24	75	78	0
Male	41	2	0	0	0	0	0	0	195	139	88	26.88	85	65	0
Female	42	2	1	30	0	0	0	0	190	108	70.5	21.59	72	85	0
Female	43	1	0	0	0	0	0	0	185	123.5	77.5	29.89	70	NA	0
Female	52	1	0	0	0	0	0	0	234	148	78	34.17	70	113	0
Female	52	3	1	20	0	0	0	0	215	132	82	25.11	71	75	0
Male	44	2	1	30	0	0	1	0	270	137.5	90	21.96	75	83	0
Male	47	4	1	20	0	0	0	0	294	102	68	24.18	62	66	1
Female	60	1	0	0	0	0	0	0	260	110	72.5	26.59	65	NA	0
Male	35	2	1	20	0	0	1	0	225	132	91	26.09	73	83	0

This dataset is collected by NHLBI, US to conduct Framingham Heart Study. The respondents of this datasets were the residents of Framingham and surrounding towns.

In my study, I did not use all of the variables and columns given. The variables selected in my analysis and its descriptions were listed in the following table.

Variables (responses)	**Data Type**	**Descriptions**
Gender (male, female)	Nominal	Gender of respondents
Age (numeric data)	Interval	Ages of respondents
Stroke (0, 1)	Ordinal	0 represents no stroke 1 represents has stroke
Total Cholesterol (numeric data)	Interval	Total cholesterol of respondents measured in unit of mg/dL
**Systolic BP (numeric data)	Interval	Systolic blood pressure of respondents measured in unit of mm/Hg

**Diastolic BP (numeric data)	Interval	Diastolic blood pressure of respondents measured in unit of mm/Hg
BMI (numeric data)	Interval	BMI of respondents calculated from his/her height and weight
Heartbeat Rate (numeric data)	Interval	Heartbeat Rate of the respondents

*** represents data which is not included in proposal of project*

After filtering the responses, some of the data of respondents which consist of "NA" and will affect the study were removed. Finally, a total of 4142 out of 4241 responses were used in my Framingham Heart Study.

Link: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset/data

1.2 AIM OF STUDY

This study is to identify common factors or characteristics that contribute to cardiovascular disease. In this study, heartbeat rate is compared and tested with variables such as gender, age, cholesterol level, blood pressure, and BMI, each using suitable type of test to test on the hypotheses made and the relationship between each other. Relationship between heartbeat rate and those factors mentioned above are being analysed.

List of studies:

1. Test whether the population heartbeat rate mean is equal to a human's normal heartbeat rate mean. If hypothesis is rejected, the population heartbeat rate mean is not equals to a human's normal heartbeat rate mean assumed.
2. Test on the population systolic and diastolic BP mean with normal human's systolic BP mean. If hypothesis is rejected, population systolic and diastolic BP mean is not equals to normal human's systolic BP mean.
3. Testing whether heartbeat rate is dependent on gender and what is the relationship between heartbeat rate with BMI and total cholesterol level.
4. Examine whether proportion of stroke patient for different age group is the same. Hence, a test is also done to analysis the proportion of stroke patients in different age group.

2.0 HYPOTHESIS TESTING

2.1 1-SAMPLE HYPOTHESIS TEST

2.1.1 HEARTBEAT RATE

Test whether the population heartbeat rate mean is equal to a human's normal heartbeat rate mean with population variance unknown. "Normal heart rate varies from person to person, but a normal range for adults is 60 to 100 beats per minute, according to the Mayo Clinic." (Gholipour, 2018). Hence, the normal heartbeat mean is assumed as the median value of the range which is 80 beats per second.

$$H_0: \mu = 80$$

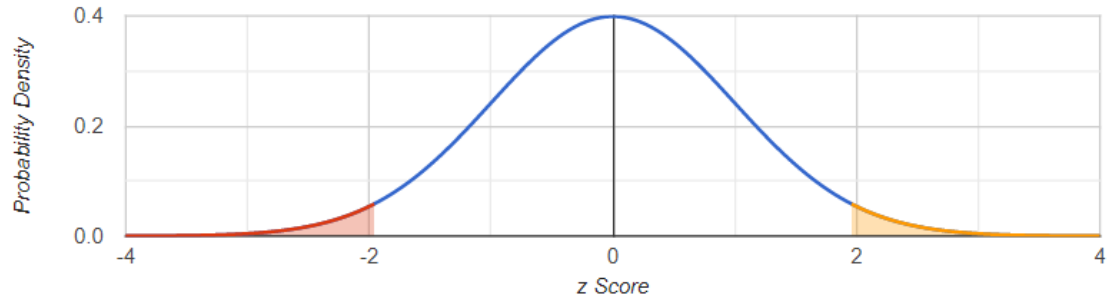
$$H_1: \mu \neq 80$$

$$\bar{x} = 75.837, s = 8.708$$

$$\alpha = 0.05$$

Test statistic, $Z = \frac{\bar{x}-u}{s/\sqrt{n}}$ = -22.27

Critical value, $z_{0.05}$ = 1.960



Decision:

Test statistic < left tail critical value. Test statistic lies in the critical region. H_0 is rejected.

Conclusion:

There is sufficient evidence to conclude that the population mean of heartbeat rate of residents from Framingham and surrounding towns is not equal to 80 beats per second.

2.1.2 SYSTOLIC BP

Test whether the population systolic BP mean is equal to a normal human's systolic BP mean with population variance unknown. "Blood pressure numbers of less than 120/80 mm Hg are considered within the normal range." (Association, 2020) Hence in this test, normal human's systolic BP is assumed as 120mm/Hg.

$$H_0: \mu = 120$$

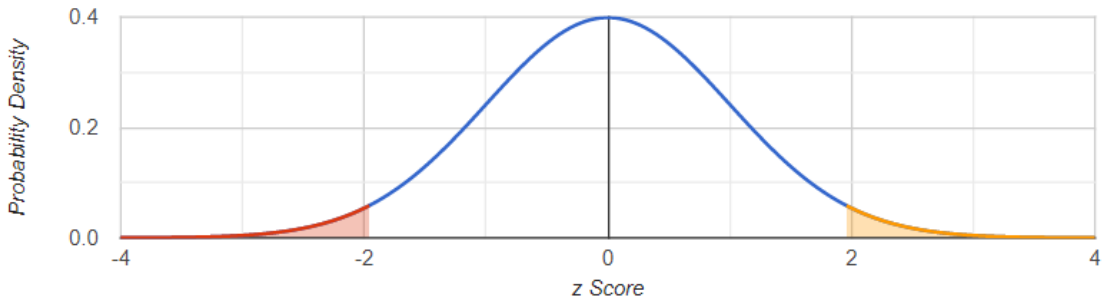
$$H_1: \mu \neq 120$$

$$\bar{x} = 132.295, s = 21.956$$

$$\alpha = 0.05$$

$$\text{Test statistic, } z = \frac{\bar{x}-u}{s/\sqrt{n}} = 36.038$$

$$\text{Critical value, } z_{0.05} = 1.960$$



Decision:

Test statistic > right tail critical value. Test statistic lies in the critical region. H_0 is rejected.

Conclusion:

There is sufficient evidence to conclude that the population mean of systolic BP of residents from Framingham and surrounding towns is not equal to 120 mm/Hg.

2.1.3 DIASTOLIC BP

Test whether the population diastolic BP mean is equal to a normal human's diastolic BP mean with population variance unknown. "Blood pressure numbers of less than 120/80 mm Hg are considered within the normal range." (Association, 2020) Hence in this test, normal human's diastolic BP is assumed as 80mm/Hg.

$$H_0: \mu = 80$$

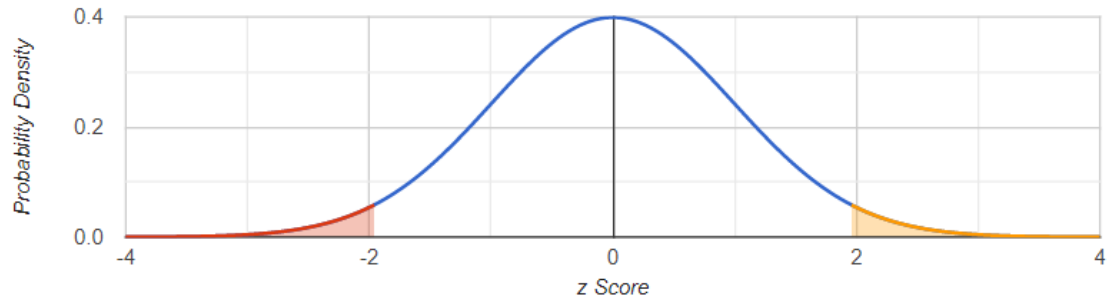
$$H_1: \mu \neq 80$$

$$\bar{x} = 82.903, \quad s = 11.872$$

$$\alpha = 0.05$$

$$\text{Test statistic, } z = \frac{\bar{x} - u}{s/\sqrt{n}} = 15.74$$

$$\text{Critical value, } z_{0.05} = 1.960$$



Decision:

Test statistic > right tail critical value. Test statistic lies in the critical region. H_0 is rejected.

Conclusion:

There is sufficient evidence to conclude that the population mean of diastolic BP of residents from Framingham and surrounding towns is not equal to 80 mm/Hg.

2.2 GOODNESS-OF-FIT TEST

2.2.1 GROUP OF AGE AND STROKE

Test whether proportion of stroke patient for different age group is the same. In this test, the respondents who has stroke are further categorized into 3 different groups according to their age.

Group (number of responses)	Age
Young age stroke patients (1)	< 41
Middle age stroke patients (17)	40 < age < 61
Old age stroke patients (5)	60 < age < 81

$$H_0: p_{young} = p_{middle} = p_{old}$$

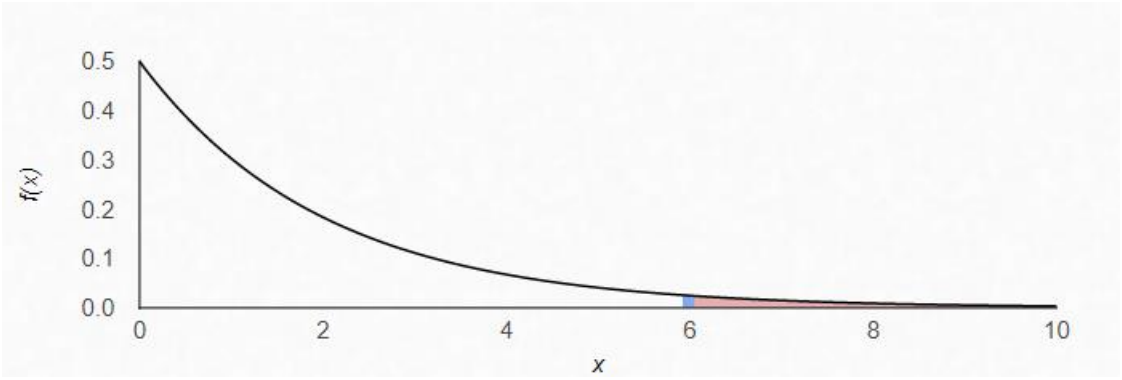
$$H_1: p_{young} \neq p_{middle} \neq p_{old}$$

$$\alpha = 0.05$$

k = 3-1 = 2

Test statistic, $x^2 = \sum \frac{(O-E)^2}{E}$ = 18.087

Critical value, $x^2_{2,\ 0.05}$ = 5.991



Decision:

Test statistic > critical value. Test statistic lies in the critical region. H_0 is rejected.

Conclusion:

There is sufficient evidence to conclude that the proportion of stroke patient for different age group of residents from Framingham and surrounding towns is not equal.

2.3 CHI-SQUARE TEST OF INDEPENDENCE

2.3.1 GENDER AND HEARTBEAT RATE

Test whether heartbeat rate is dependent on gender.

H_0: Heartbeat rate is independent on the gender

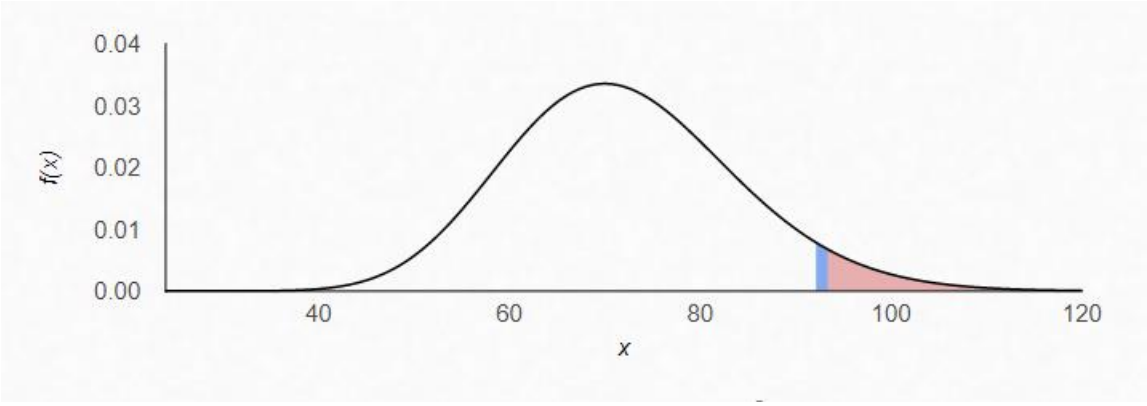
H_1 : Heartbeat rate is dependent on the gender

$$\alpha = 0.05$$

$$df = 72$$

$$\text{Test statistic, } x^2 = \sum \frac{(O\_ij - E\_ij)^2}{E_{ij}} = 129.39$$

Critical value, $x^2{}_{0.05}$ = 92.808



Decision:

Test statistic > critical value. Test statistic lies in the critical region. H_0 is rejected.

Conclusion:

There is sufficient evidence to conclude that heartbeat rate is dependent on gender.

2.4 CORRELATION

2.4.1 HEARTBEAT RATE AND BMI

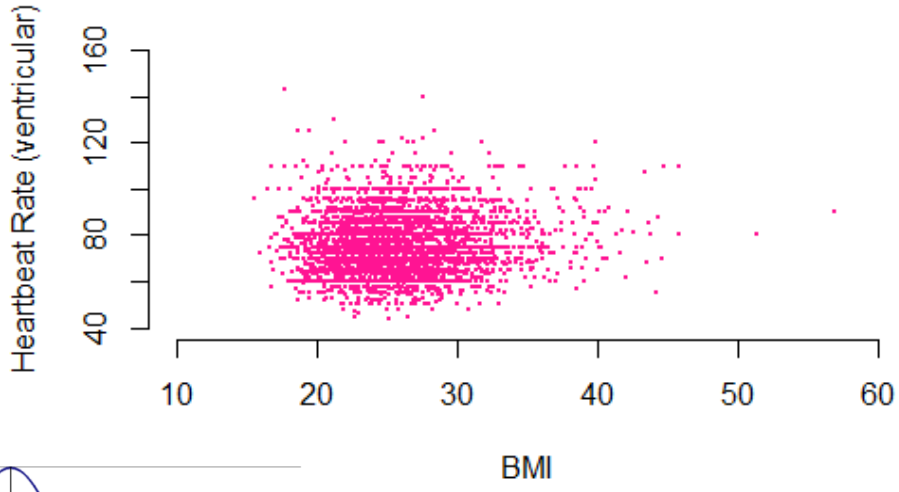
Test whether the linear relation exist between BMI and heartbeat rate. This test will measure the strength of the linear relationship between BMI and heartbeat rate. Pearson's Product-Moment Correlation Coefficient technique is used to find sample correlation coefficient because the variables chosen are interval variable.

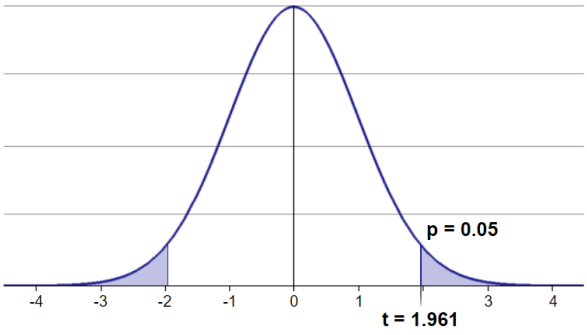
X: BMI

Y: Heartbeat rate

H_0: $\rho = 0$

H_1 $\rho \neq 0$





r = 0.0706

$\alpha = 0.05$, df = n-2 = 4140

Test statistic, $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = 4.555$

Critical value, $t_{0.05\,/2,\ 4140} = 1.961$

Decision:

Test statistic > right tail critical value. Test statistic lies in the critical region. H_0 is rejected.

Conclusion:

There is sufficient evidence of a linear relationship between BMI and heartbeat rate. However, the sample correlation coefficient, $r \approx 0$. Hence, we can conclude that the linear relationship between BMI and heartbeat rate is weak.

2.5 REGRESSION

2.5.1 BMI RELATED TO HEARTBEAT RATE

Test on how BMI is related to heartbeat rate.

Dependent: Heartbeat Rate, Independent: BMI

$$H_0 : \beta_1 = 0$$

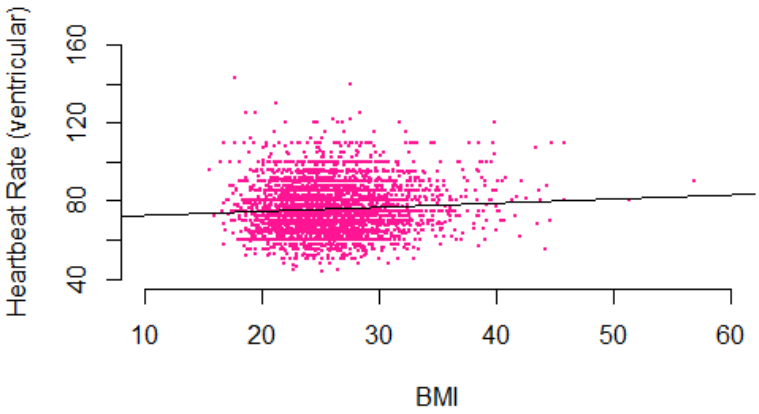
$$H_1 : \beta_1 \neq 0$$

$b_0 = 23.9773$

$b_1 = 0.02394$

$\hat{y} = 723.9773 + 0.02394x$

$R^2 = 0.00499$



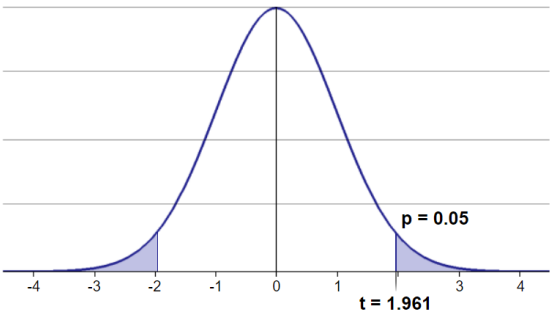
$\alpha = 0.05$

df = n-2 = 4140

S_{b_1} = 0.00526

Test statistic, $t = \frac{b_1 - \beta_1}{S_{b_1}} = 4.55452$

Critical value, $t_{0.05/2,\ 4140} = 1.96054$



Decision:

Test statistic > right tail critical value. Test statistic lies in the critical region. H_0 is rejected.

Conclusion:

There is sufficient evidence that BMI affects heartbeat rate in a positive linear relationship with S_{b_1} = 0.00526.

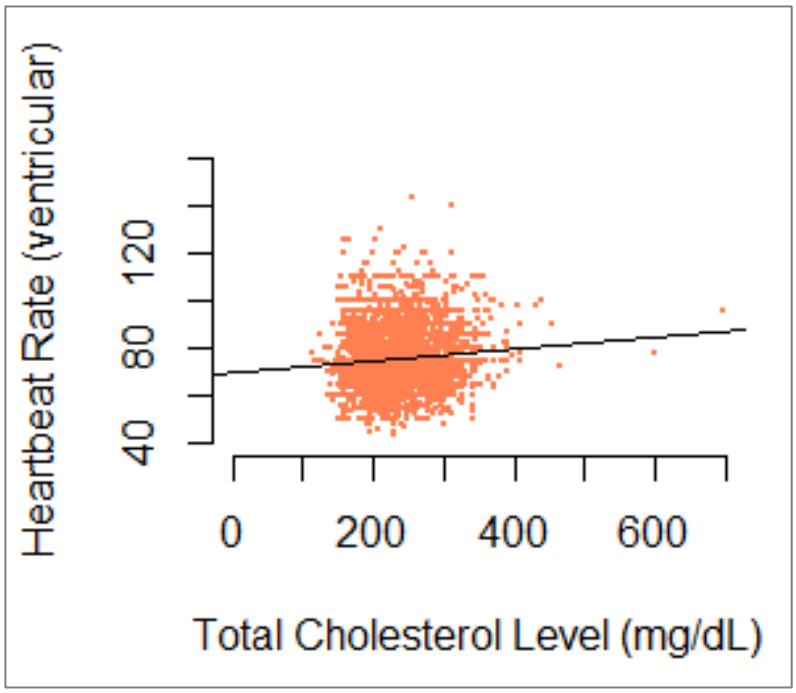
2.5.2 TOTAL CHOLESTEROL LEVEL RELATED TO HEARBEAT RATE

Test on how total cholesterol level is related to heartbeat rate.

Dependent: Heartbeat Rate, Independent: Total Cholesterol

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$



$b_0 = 70.16479$

$b_1 = 0.02396$

$\hat{y} = 70.16479 + 0.02396x$

$R^2 = 0.00788$

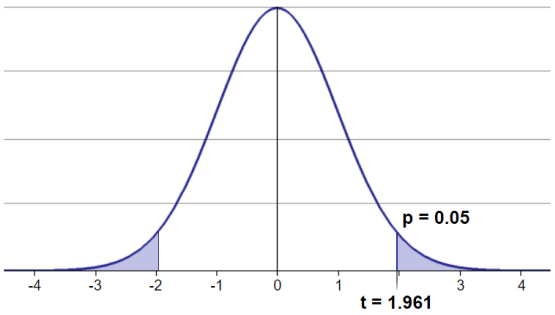
$\alpha = 0.05$

df = n-2 = 4140

$S_{b_1} = 0.00418$

Test statistic, $t = \frac{b_1 - \beta_1}{S_{b_1}} = 57.32057$

Critical value, $t_{0.05/2,\ 4140} = 1.96054$



Decision:

Test statistic > right tail critical value. Test statistic lies in the critical region. H_0 is rejected.

Conclusion:

There is sufficient evidence that total cholesterol affects heartbeat rate in a positive linear rela tionship with S_{b_1} = 0.00418.

3.0 DISCUSSION

After conducting the hypothesis tests, there are several findings to be discussed. First, in 1-sample hypothesis tests on blood pressure and heartbeat rate, analyses were done by comparing the mean calculated from dataset with a normal human's heartbeat rate and blood pressure. We can see that although blood pressure and heartbeat rate of residents from Framingham and surrounding towns is not equal to the assumed value, the test statistic values calculated is close to the critical value. This shows that the mean of heartbeat rate and blood pressure is close to the assumed values with small standard deviations. We get to know that most of the residents from Framingham and surrounding towns have good heart conditions when the survey was done.

I was also interested in whether heartbeat rate is dependent on gender. Hence, a Chi-Square Test on Independence was conducted to test it. In result, we successfully proved that heartbeat rate, indeed is dependent on gender. In my opinion, this may due to the difference in metabolic rate of male and female. However, this may also due to the activities done by both genders too. Hence, further analysis is needed to determine the reasons of heartbeat rate is dependent to gender.

Other than gender, I hope to relate BMI with the heartbeat rate too. A correlation test is conducted. Pearson's Product-Moment Correlation Coefficient technique is used to find sample correlation coefficient to determine the strength of relationship of both variables. Although the existence of linear relationship between BMI and heartbeat rate was proved, we couldn't identify how BMI will affect heartbeat rate. Hence, a regression test is done to determine the R^2 value. The value obtained was 0.00499 which shows a positive linear relationship between BMI and heartbeat rate. When BMI of respondents increase, the heartbeat rate will also increase. This may due to the fats in the blood capillaries are causing resistance of blood flow and causes the heartbeat rate to increase where further analysis is needed.

Besides BMI, regression test is done to test the effect of increase in cholesterol level of an individual to the value of heartbeat rate. Results showing that increase in cholesterol level in the body will causes the heartbeat rate to increase. Scientific explanation given by American Heart Association News states that LDL cholesterol can clog up the arteries and causes increase in heart attack and stroke risk while HDL cholesterol can help eliminate the LDL (News, 2019). Hence, we can predict that there is higher composition of LDL in the

cholesterol level of respondents which causes the blood vessels to be clogged and increase the heartbeat.

In both of the regression test, standard deviation of regression slope, S_{b_1} are equal to 0.00526 and 0.00418 respectively. These standard errors are considered as small and indicates that the variation in slope of regression lines from different possible samples would be almost the same. Hence, the estimated regression model is trustworthy and can be used to predict the heartbeat rate of a resident in Framingham and surrounding towns when cholesterol level is provided.

Lastly, the Goodness-Of-Fit test shows that proportion of stroke patient in Framingham and surrounding towns are different for each age group. Throughout the 4142 respondents, there are only 23 respondents having stroke. There are only 1 young age stroke patient is found, showing that stroke patients of young age patients are rarely found. Most of the stroke patients of stroke patients are from middle age (17 patients) while there are only 5 old age stroke patients. This sample taken is too small and couldn't represents the population accurately. However, based on the result of hypothesis test we obtained, the test statistic value is 3 times larger than the critical value and lies in the critical region. Hence, we can conclude that proportion of stroke patients from different age group is different.

5.0 CONCLUSION

A summary of conclusions of hypothesis tests at significance level of 95%:

a. The population mean of heartbeat rate is not equal to 80 beats per second.
b. The population mean of systolic BP is not equal to 120 mm/Hg.
c. The population mean of diastolic BP is not equal to 80 mm/Hg.
d. The proportion of stroke patient for different age group is not equal.
e. Heartbeat rate is dependent on gender.
f. There is weak linear relationship between BMI and heartbeat rate.
g. BMI and Cholesterol level affects heartbeat rate in a positive linear relationship.

The residents from Framingham and surrounding towns have good heart conditions when the survey was done. Heartbeat rate is dependent to gender and is also proved to be increased when BMI and cholesterol level increased. Stroke patients' proportion for different age group is different.

4.0 REFERENCE

Association, A. H. (2020). *Understanding Blood Pressure Readings*. Retrieved from https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Bognar, M. (2020). *Chi-Square Distribution*. Retrieved from https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html

Editor, M. B. (30 May, 2013). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?* Retrieved from The Minitab Blog: https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit#:~:text=R%2Dsquared%20is%20a%20statistical,multiple%20determination%20for%20multiple%20regression.&text=0%25%20indicat

Gholipour, B. (12 Jan, 2018). *What Is a Normal Heart Rate?* Retrieved from https://www.livescience.com/42081-normal-heart-rate.html#:~:text=For%20adults%2018%20and%20older,bpm%2C%20according%20to%20the%20AHA.

Johnston, N. (n.d.). Retrieved from StatDistributions.com: http://www.statdistributions.com/t/

kassambara. (25 Dec, 2019). *PCH of R Best Tips*. Retrieved from Datanovia: https://www.datanovia.com/en/blog/pch-in-r-best-tips/

News, A. H. (1 Feb, 2019). *8 things that can affect your heart – and what to do about them*. Retrieved from https://www.heart.org/en/news/2019/02/01/8-things-that-can-affect-your-heart-and-what-to-do-about-them

Raymond, J. (n.d.). *z-Score Calculator*. Retrieved from https://www.zscorecalculator.com/

Subset Data Frame Rows in R. (n.d.). Retrieved from Datanovia: https://www.datanovia.com/en/lessons/subset-data-frame-rows-in-r/