UTM

UNIVERSITI TEKNOLOGI MALAYSIA

PROBABILITY & STATISTICAL DATA ANALYSIS

SEMESTER II 2020/2021

SECI2143-05

GROUP PROJECT 2 – Team KFC

HEART FAILURE PREDICTION RECORDS

NAME OF STUDENTS	MATRIC NO.
BRENDAN DYLAN GAMPA ANAK JOSEPH DUSIT @ DUSIT	A20EC0021
HAFIY HAKIMI BIN SAIFUL REDZUAN	A20EC0040
MOHAMAD HAZIQ ZIKRY BIN MOHAMMAD RAZAK	A20EC0079
AUM JEEVAN A/L AUM NIRANGKAR	A20EC0017

LECTURER: DR. NOR AZIZAH BINTI ALI

SUBMISSION DATE: 30TH JUNE 2021

TABLE OF CONTENTS

INTRODUCTION

Heart failure is the inability of the heart to supply satisfactory blood stream and therefore oxygen conveyance to peripheral tissues and organs. Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure. Under perfusion of organs leads to reduced exercise capacity, fatigue, and shortness of breath.

The aim of this study is to **predict the rate of survivability** among the respondents with heart failure by considering selective variables that possibly has a relationship with another variable or independent by themselves. By conducting this project, we are able to know what the relationship between these variables are and how they affect each other. The outcomes obtained from this project can be used for the research on heart failure in the future.

DATASET

In project 2, Team KFC was required to apply inferential statistics on a specific secondary data obtained from the open source. We were required to carry out various hypothesis tests on the inferential statistics. Hence, a dataset with many numerical variables were needed. After searching dataset from the world wide web, Team KFC decided to use the dataset of survivability of patients with heart failure from serum creatinine and ejection fraction alone.

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

Variables	Data Type	Description	Method of analysis
age	Quantitative	Age of patients	Correlation, ANOVA
creatinine_phosphate	Quantitative	Level of the CPK enzyme in the blood (mcg/L)	ANOVA
diabetes	Qualitative	If the patient has diabetes (Boolean)	Chi-Square test for independence
serum_creatinine	Quantitative	Level of serum creatinine in the blood (mg/dL)	Correlation
sex	Qualitative	Woman or man (Boolean)	Chi-Square test for independence
DEATH_EVENT	Qualitative	If patient passed away (Boolean)	Hypothesis Testing for 1 sample

Table 1: Variables used in the analysis of hypothesis testing.

DATA ANALYSIS

HYPOTHESIS TESTING –1 SAMPLE

The purpose of hypothesis testing is to test whether null hypothesis can be rejected or approved. This test will investigate whether **the average death event among the first 100 respondents is 50% and above with population variance unknown**. Hence, the 50% death event as the median value of the data is 50 deaths. (95% significance level)

*H*0: μ = 0.50 (equal to 50 deaths)

*H*1: μ > 0.50 (more than 50 deaths)

$$\bar{X} = \frac{66(1)+34(0)}{100}$$

$$s = \sqrt{\frac{66(1-0.66)^2+34(0-0.66)^2}{100-1}}$$

$\bar{x}$ = 0.66, s =0.48, α = 0.05

Test statistic, $Z = \frac{\bar{x}-u}{s/\sqrt{n}} = 3.33$

Critical value, $z_{0.05} = 1.960$

P-Value= **0.000868**

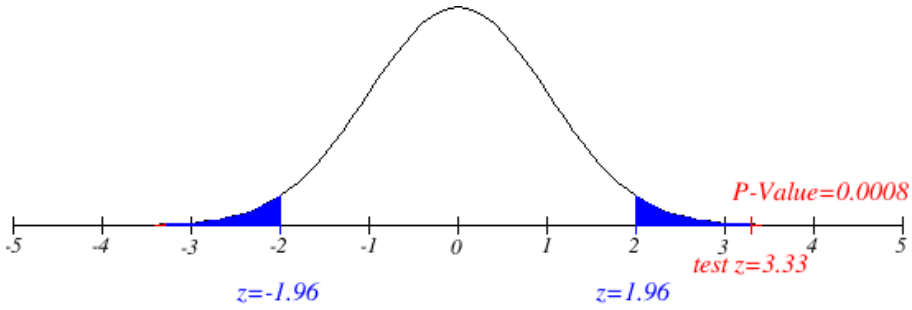

Diagram 1: Normal distribution graph of z

Decision:

By referring to the figure above, test statistics 3.33 falls within the critical region. Since the P-value, 0.000868 < 0.05 and Test statistic 3.33 > Critical value 1.96, **therefore rejecting H_0.**

Conclusion:

There is **sufficient evidence** to conclude that average death event among the first 100 respondents is 50% and higher.

CORRELATION

This study uses correlation to measure the strength of linear relationship between age and the level of serum creatinine in the blood (mg/dL). Pearson's Product Moment Correlation Coefficient technique is used in R Studio to find the sample correlation coefficient, r since the variables chosen are both ratios. The first 99 out of the total 300 responses are taken as the sample of the study.

The hypothesis statement is based on the following:

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

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

Diagrams 2 and 3 shows the screenshot of the code made in R Studio with the scatterplot, respectively.

```
> #scatterplot
> plot(agecrea$age,agecrea$serum_creatinine, xlab = 'Age',ylab='Serum Creatinine',
+      main='Scatterplot age vs serum creatinine',xlim=c(40,100),
+      ylim=c(0,10),col='blue')
> model<-lm(agecrea$serum_creatinine~agecrea$age)
> abline(model)
> #calculate correlation coefficient
> cor(agecrea$age,agecrea$serum_creatinine)
[1] 0.1894759
> summary(model)

Call:
lm(formula = agecrea$serum_creatinine ~ agecrea$age)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1931 -0.5966 -0.3213  0.0242  7.5140

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.39902    0.63533   0.628   0.5315
agecrea$age  0.01859    0.00978   1.901   0.0603 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.256 on 97 degrees of freedom
Multiple R-squared:  0.0359,    Adjusted R-squared:  0.02596
F-statistic: 3.612 on 1 and 97 DF,  p-value: 0.06033
```

Diagram 2: Code for the correlation analysis in R Studio

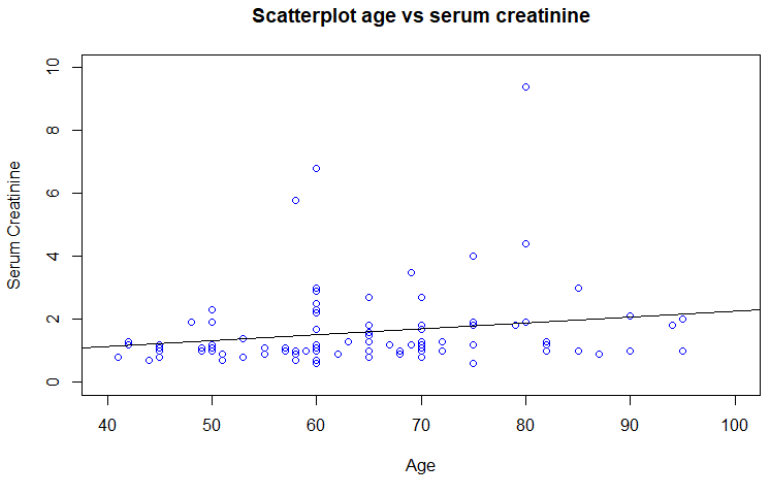

Diagram 3: Scatterplot age vs serum creatinine

The sample correlation coefficient, r obtained is 0.1894759 showing a **positive correlation**. Since the r value is between -0.5 and 0.5, the **strength of linear relationship is weak**.

This test is **a two-tailed test**:

$\alpha = 0.05, \quad \alpha / 2 = 0.025$

$degree\ of\ freedom\ (d.f) = 99 - 2 = 97$

$critical\ value\ t_{0.05,97} = 1.984723$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.1894759}{\sqrt{\frac{1-(0.1894759)^2}{100-2}}} = 1.9103$$

Since P-value = 0.59018 > 0.05 and $1.9103 < 1.9847$, **fail to reject H_0 .** There is sufficient evidence that there is **no linear correlation** between age and serum creatinine.

CHI-SQUARE TEST OF INDEPENDENCE

Chi-square test of independence is used to test if a relationship exists between two qualitative variables and two-way contingency table is used.

In this project, we have 100 data as the sample. The two variables chosen are whether the patients have diabetes and what sex is the patient. We need to test if there is evidence of the relationship between them at the significance level of 0.05.

The hypothesis statement is shown below:

$$H_0: Diabetes\ is\ independent\ of\ sex$$

$$H_1: Diabetes\ is\ not\ independent\ of\ sex$$

A two-ways contingency table and chi-square test of independence implementation are formed by using R Studio.

```
> # get the contingency table
> table1<- table(diabetes, sex)
> table1
        sex
diabetes Female Male
     No      15   40
     Yes     19   26
> # perform chi-square test on the data table
> chisq.test(table1, correct=FALSE)

        Pearson's Chi-squared test

data:  table1
X-squared = 2.4649, df = 1, p-value = 0.1164

> #Expected Frequency
> chisq.test(table1, correct=FALSE)$expected
        sex
diabetes Female Male
     No    18.7 36.3
     Yes   15.3 29.7
> #Critical Value
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=1, lower.tail=FALSE)
> x2.alpha
[1] 3.841459
```

Diagram 4: Code for the chi-square test of independence analysis in R Studio

The test statistic of chi-square test for independence is calculated by using this formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

By referring to the figure, the test statistic calculated, χ^2 is 2.4649 and its P-value is 0.1164. At 0.05 significance level and with the degree of freedom of 1, the critical value is 3.841459 and thus:

Critical region: $\chi^2 = 3.841459$.
Since P-value = 0.1164 > 0.05 or $\chi^2 = 2.4649 < 3.841459$, $\boldsymbol{H_0}$ **fail to be rejected.** Therefore, there is sufficient evidence to conclude that the diabetic status of the patient is independent of the sex of the patient at the 0.05 significance level.

ANOVA

ANOVA test is used to test for significant differences between means.

In this project, there are 100 samples and the variables chosen are age and creatinine phosphate (Level of the Creatinine Phosphokinase (CPK) enzyme in the blood (mcg/L)). In this test, only 30 samples were taken. The age is divided into three groups which are 40-59, 60-79, and 80-99. Each age group consists of 10 samples. The test was run to see if there are any significant differences between the age groups at significance level of 0.05.

The hypothesis statement is as below:

$$H_0: \mu 1 = \mu 2 = \mu 3$$

$$H_1: At\ least\ one\ mean\ is\ different.$$

One-way ANOVA test with equal sample sizes were conducted by R Studio to obtain the test statistics.

```
> heart <- read_excel("D:/Work/UTM/Semester 2/Probability Data and Statistical Analysis/heart.xlsx")
> View(heart)
> attach(heart)
> aov(`cpk enzyme`~age)
Call:
   aov(formula = `cpk enzyme` ~ age)

Terms:
                      age Residuals
Sum of Squares    5221323  52511463
Deg. of Freedom         2        27

Residual standard error: 1394.586
Estimated effects may be unbalanced
> summary(aov(`cpk enzyme`~age))
            Df   Sum Sq Mean Sq F value Pr(>F)
age          2  5221323 2610662   1.342  0.278
Residuals   27 52511463 1944869
```

Diagram 5: Code for the ANOVA analysis in R Studio

Referring to the F-table, with numerator of Degree of Freedom is 3 and denominator of Degree of Freedom is 16, with 0.05 significance level, the critical value is 3.24.

According to the test performed in R Studio, F – value is 1.342.

P-value is calculated referring to F-value and degree of freedom.

P-value = 0.29593

Since P-value 0.29593 > significance level 0.05 and F-value statistics 1.342 < F-critical value 3.24, we **fail to reject H_0**. There is sufficient evidence to claim that the different age groups have the same mean for the level of CPK enzyme in the blood(mcg/L).

CONCLUSION

From the hypothesis test which we have conducted on different pairs of data from the dataset, our most interesting finding could be that most of tests conducted are independent or failed to reject except the average death among the respondent. We can see that the average death of those diagnosed with heart failure increases from time to time due to the rejection of null hypothesis. The level of serum creatinine in respondent's blood, which contributes to heart failure differs and can increase or decrease at any age. From this, we can conclude that heart failure can occur at any age means aging issue is not the factor. Regular exercise could be the key to maintain the level of serum creatinine in our blood which also reduce the risk towards heart failure.

In terms of patient who diagnosed with diabetes in correlation with the potential of having a heart failure, gender is independent to both diseases. This means gender has no effects whether man or female will diagnose heart failure the most. Moreover, when we test the level of creatinine phosphate enzymes in respondent's blood with their age, the result obtained that the average mean among 3 divided groups are equal means both variables are not affecting one another. Balanced diet intake helps in reducing the glucose level that can cope diabetes at the same time with heart failure. Despites that, steroid consumption must be avoided so that our body mass does not increase abnormally and cause the enzyme of creatinine phosphate to create excessive chemical reactions in our body.

What we have learned from all the activities conducted is that the importance of having hypothesis test as to predict flow of the data mostly between two opposite variables. This is because without conducting any hypothesis tests, we could not predict either there is any relationship available between the variables. These activities also help us to strengthen our concepts regarding the hypothesis test and how they were applied in real life situation. Hence, we can conclude that our objective for this project is achieved whereby the rate of survivability among patients with heart failure is low based on the hypothesis test conducted.

REFERENCES

Stacy Sampson. (16 January, 2019). *Why Are Enzymes Important?* Retrieved from healthline: https://www.healthline.com/health/why-are-enzymes-important

Zia Sherell. (26 February, 2021). *What to know about high creatinine levels*. Retrieved from Medical News Today: https://www.medicalnewstoday.com/articles/when-to-worry-about-creatinine-levels

APPENDIX

Dataset link: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

A	B	C	D	E	F	G	H	I	J	K	L	M	N
age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT	
75	0	582	0	20	1	265000	1.9	130	1	0	4	1	
55	0	7861	0	38	0	263358	1.1	136	1	0	6	1	
65	0	146	0	20	0	162000	1.3	129	1	1	7	1	
50	1	111	0	20	0	210000	1.9	137	1	0	7	1	
65	1	160	1	20	0	327000	2.7	116	0	0	8	1	
90	1	47	0	40	1	204000	2.1	132	1	1	8	1	
75	1	246	0	15	0	127000	1.2	137	1	0	10	1	
60	1	315	1	60	0	454000	1.1	131	1	1	10	1	
65	0	157	0	65	0	263358	1.5	138	0	0	10	1	
80	1	123	0	35	1	388000	9.4	133	1	1	10	1	
75	1	81	0	38	1	368000	4	131	1	1	10	1	
62	0	231	0	25	1	253000	0.9	140	1	1	10	1	
45	1	981	0	30	0	136000	1.1	137	1	0	11	1	
50	1	168	0	38	1	276000	1.1	137	1	0	11	1	
49	1	80	0	30	1	427000	1	138	0	0	12	0	
82	1	379	0	50	0	47000	1.3	136	1	0	13	1	
87	1	149	0	38	0	262000	0.9	140	1	0	14	1	
45	0	582	0	14	0	166000	0.8	127	1	0	14	1	
70	1	125	0	25	1	237000	1	140	0	0	15	1	
48	1	582	1	55	0	87000	1.9	121	0	0	15	1	
65	1	52	0	25	1	276000	1.3	137	0	0	16	0	
65	1	128	1	30	1	297000	1.6	136	0	0	20	1	
68	1	220	0	35	1	289000	0.9	140	1	1	20	1	
53	0	63	1	60	0	368000	0.8	135	1	0	22	0	
75	0	582	1	30	1	263358	1.83	134	0	0	23	1	
80	0	148	1	38	0	149000	1.9	144	1	1	23	1	
95	1	112	0	40	1	196000	1	138	0	0	24	1	
70	0	122	1	45	1	284000	1.3	136	1	1	26	1	
58	1	60	0	38	0	153000	5.8	134	1	0	26	1	
82	0	70	1	30	0	200000	1.2	132	1	1	26	1	
94	0	582	1	38	1	263358	1.83	134	1	0	27	1	
85	0	23	0	45	0	360000	3	132	1	0	28	1	
50	1	249	1	35	1	319000	1	128	0	0	28	1	
50	1	159	1	30	0	302000	1.2	138	0	0	29	0	
65	0	94	1	50	1	188000	1	140	1	0	29	1	

Diagram 6: Heart failure clinical records dataset