Heart Failure Prediction

C. D. V. P. Basnayake

S15025

—

ST 3010

—

Introduction to Health Statistics

Contents

1. An Introduction/Background of the study

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age. (World Health Organization, 2022)

In this report, it is intended to analyze the heart failure clinical records of 299 patients. Status of death event is the dependent variable while it has been studying by using several independent variables namely age of the patient, status of anemia, level of creatinine phosphokinase, status of diabetes, ejection fraction, status of high blood pressure, platelets count, amount of serum creatinine, gender, status of smoking and time.

All these data are analyze using R studio and it would be using one sample t-test, two sample t-test, Relative Risk and Odd Ration for analytical purpose. One sample t-test is used to compare the mean of one sample to a known/unknown standard (or theoretical/hypothetical) mean (μ). In this report, it will be tested the platelet count of dead person is equal to 100000.

The two-sample t-test (also known as the independent samples t-test) is a method used to test whether the unknown population means of two groups are equal or not. In this report, will be used to test whether there is a statistically significant difference among dead patients and alive patients.

Other statistical tools are Relative risk. The relative risk (also known as risk ratio [RR]) is the ratio of risk of an event in one group (e.g., exposed group) versus the risk of the event in the other group (e.g., nonexposed group). By using this method, it will be test whether there is a significant influence from being exposed to smoking to death of the patient.

These statistic tools are commonly used in health sector and in this report, it will be making use of them for analytical purpose using R studio.

2. *Objectives of the study*

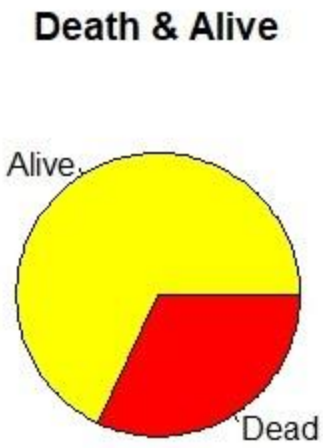
1. Identify what are the factors cause death of patients and relative risk of them.
2. Identify what are variables which do not have any impact on death of a patient.
3. Identify what are critical levels in the studied variables which cause a person to die.

3. *Data Description*

- Anemia states in which blood hemoglobin level is below the normal range for the patient's age and sex. It was measured by nominal scale and 1 indicates lower than normal and 0 indicates normal.
- Creatine phosphokinase (CPK) is an enzyme in the body which is found mainly in the heart, brain, and skeletal muscle. And it also measured using interval scale. Normal CPK ranges 10 to 120 micrograms per liter (mcg/L).
- Status of diabetes measured using nominal scale and 0 indicates negative for diabetes and 1 indicates being positive.
- Ejection fraction is ventricular and diastolic volume (EDV) which is ejected with each stroke and it is measured using interval scale.
- High blood pressure is measured and it express as nominal scale where 0 is used to indicates having high blood pressure while 1 indicates not having high blood pressure.
- Platelets count is measured by using interval scale (kilo platelets/mL)
- Level of serum creatinine is reported as milligrams of creatinine to a deciliter of blood (mg/dL) which is interval scale.
- A sodium blood test (also called a serum sodium test) is a way to measure the amount of sodium in blood (mEq/L) and here it is measured by interval scale.
- Gender measured by using nominal scale. In variable of gender, 0 indicates female while 1 is for male patients.
- Status of smoking is measured by using nominal scale 0 and 1 are used to indicate not exposure to smoking and exposure to smoking respectively.
- Time column show the number of days before the dead event happened, or the patient decided not to continue with the experimental
- Event of death measured using nominal scale and 0 indicates being alive and 1 indicates death
- Age indicates the age of the patients (measured in years).

4 *Data analysis and interpretation*

1.



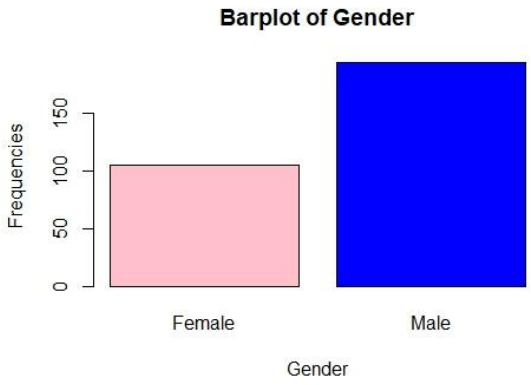

```
DEATH_EVENT
  0   1
203  96
```

In this data set data was collected from 299 persons and,

- We can see, from those 96 died in the end &
- 203 are alive

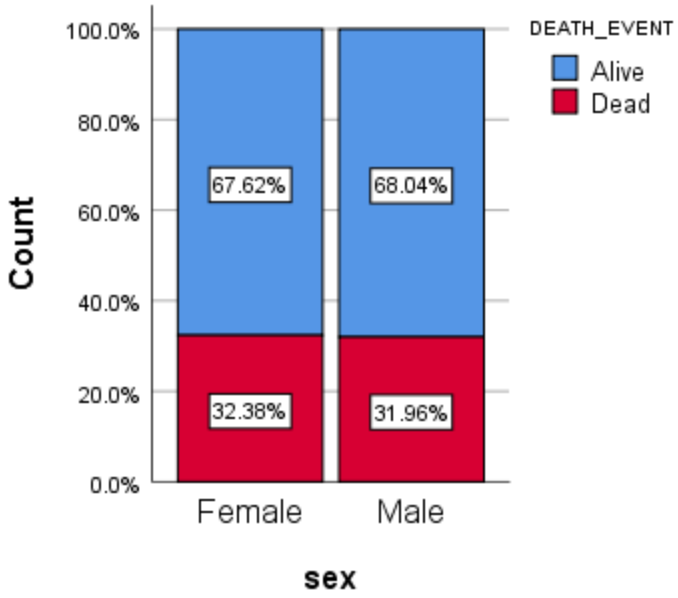
2.

```
sex
  0   1
105 194
```



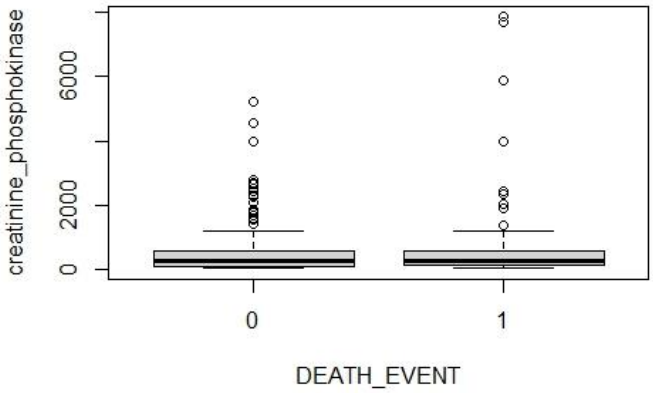
Initially there were 105 females and 194 males.

3.



Here the male alive and dead percentages and female alive and dead percentages are almost same.

4.



This boxplot of creatinine phosphokinase vs death event shows that there is not much of a difference between the distribution of creatinine phosphokinase amount in dead and alive people. But in dead people there can be seen some extreme values of creatinine phosphokinase amounts. So, to be sure a t test was conducted at 5% significant level,

```
        Welch Two Sample t-test

data:  new_data2$creatinine_phosphokinase and new_data$crea
tinine_phosphokinase
t = -0.90119, df = 125.32, p-value = 0.3692
alternative hypothesis: true difference in means is not equ
al to 0
95 percent confidence interval:
 -415.9482  155.6608
sample estimates:
mean of x mean of y
 540.0542  670.1979
```

H0 : Mean values of creatinine phosphokinase amount in dead and alive people are same
H1 : Mean values of creatinine phosphokinase amount in dead and alive people are different

This is the results of conducted t test, the p value here is 0.3692, it is larger than 0.05. So, we fail to reject the null hypothesis.

So, we can conclude that there is not enough evidence under 95% confidence level to say that mean values of creatinine phosphokinase amount in dead and alive people are not significantly different.

5.

This is the frequency table for death event with smoking factor. Using this table and calculating relative risk we can measure how the smoking factor affect the death event.

	DEATH_EVENT	
smoking	Dead	Alive
Yes	30	66
No	66	137

The relative risk we got here is 0.96 which is very close to 1. Which means when compared to a non-smoker there is very little different in risk of dying for a person who smokes.

6.

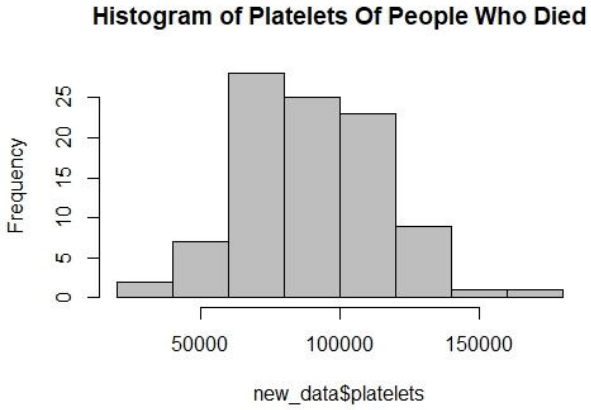
This is the frequency table for death event with diabetes factor. Using this table and calculating relative risk we can measure how the diabetes factor affect the death event.

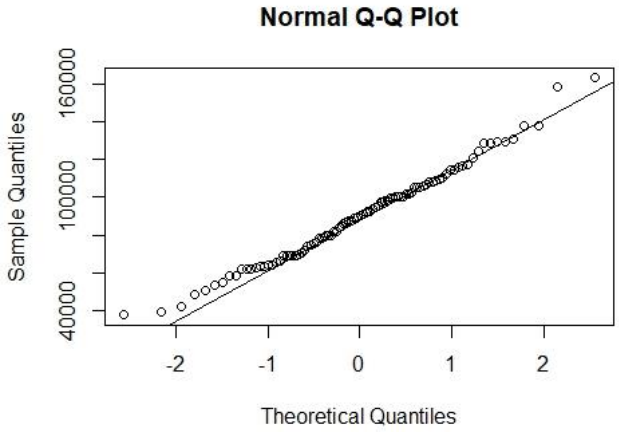
	DEATH_EVENT	
diabetes	Dead	Alive
Yes	40	85
No	56	118

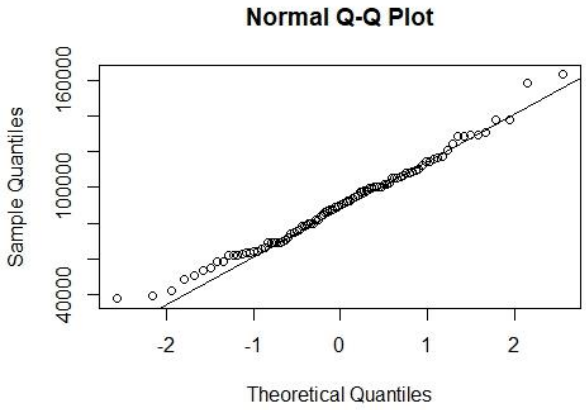
The relative risk we got here is 0.99 which is very close to 1. Which means when compared to a non-diabetic patient there is very little different in risk of dying for a person who has diabetic.

7.

The following graphs and tables show the distribution of platelet amount of people who are dead.








```
        Shapiro-Wilk normality test

data:  new_data$platelets
W = 0.98705, p-value = 0.4707
```

By looking at the histogram, q-q plots and the p-value of the shapiro test we can say that the platelets count of people who dies is normally distributed.

It is known that if the platelet count is less than 100000 of a heart patient it is a critical situation. So, a t test was done to check weather the mean platelets count of dead people is significantly lesser that the critical platelets count.

H0 : Mean values of platelets amount in dead people are greater than or equal to 100000

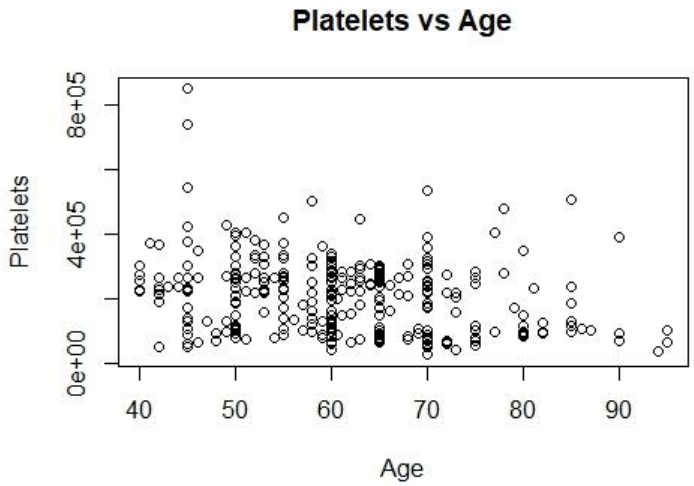
H1 : Mean values of platelets amount in dead people are less than 100000

The p-value we got from the test is 0.0001 which is lesser than 0.05. And the test statistic is -3.83 where critical value is -1.66. By that we can reject the null hypothesis.

So, we can conclude that under 95% confidence level there is enough evidence to say that mean values of platelet amount in dead is significantly lesser than 100000.

8.

Also, by analyzing the following plot it can be certain that there is no any relationship between Platelets count and Age.



9.

The creatinine phosphokinase of a healthy person should be within the 10 – 120 region getting it bellow or over that range increase the risk of death. Up or bellow that region is critical. So, t tests were done,

H0 : Mean values of creatinine phosphokinase in dead people are less than or equal to 120
H1 : Mean values of platelets amount in dead people are greater than 120

```
        One Sample t-test

data:  new_data$creatinine_phosphokinase
t = 4.0946, df = 95, p-value = 4.443e-05
alternative hypothesis: true mean is greater than 120
```

The p-value we got from the test is $4.43*10^{-5}$ which is lesser than 0.05. So, we can reject the null hypothesis.
Then it's can be concluded that that there is enough evidence at 5% significance level to say that the mean value of creatinine phosphokinase amount in dead people is greater than 120

H0 : Mean values of creatinine phosphokinase in dead people are greater than or equal to 10
H1 : Mean values of platelets amount in dead people are less than 10

```
        One Sample t-test

data:  new_data$creatinine_phosphokinase
t = 4.9132, df = 95, p-value = 1
```

The p-value we got from the test is 1 which is greater than 0.05. So, we fail to reject null hypothesis. So, at 5% significance level there is no enough evidence to say the mean value of creatinine phosphokinase amount in dead people is less than 10.

By above 2 t tests, it can be concluded that its critical to have creatinine phosphokinase level more than 120, since it can be cause to a death.

10.

To check whether there is a difference between Serum Sodium levels in dead and alive patients a 2-sample t test was conducted.

H0 : Mean values of Serum sodium level in blood in dead and alive people are same
H1 : Mean values of Serum sodium level in blood in dead and alive people are different

```
        Welch Two Sample t-test

data:  new_data2$serum_sodium and new_data$serum_sodium
t = 3.1645, df = 154.01, p-value = 0.001872
alternative hypothesis: true difference in means is not equ
al to 0
```

The p-value we got from the t test is 0.002 which is less than 0.05. So, we can reject null hypothesis. So, at 5% significance level there is enough evidence to say the mean values of serum sodium level in blood in dead and alive people are different.

11.

To check whether there is a difference between Serum Creatinine levels in dead and alive patients a 2-sample t test was conducted.

H0 : Mean values of serum creatinine level in blood in dead and alive people are same
H1 : Mean values of serum creatinine level in blood in dead and alive people are different

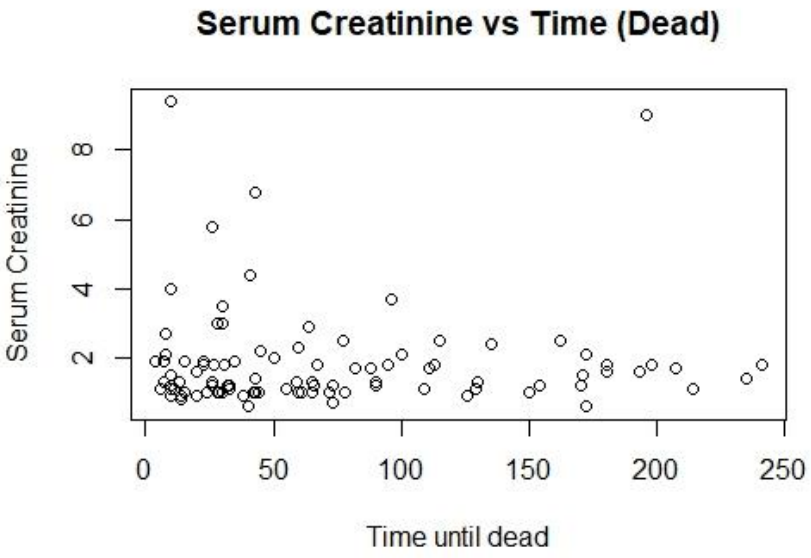
```
        Welch Two Sample t-test

data:  new_data2$serum_creatinine and new_data$serum_creatinine
t = -4.1526, df = 113.19, p-value = 6.399e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9615153 -0.3403977
sample estimates:
mean of x mean of y
 1.184877  1.835833
```

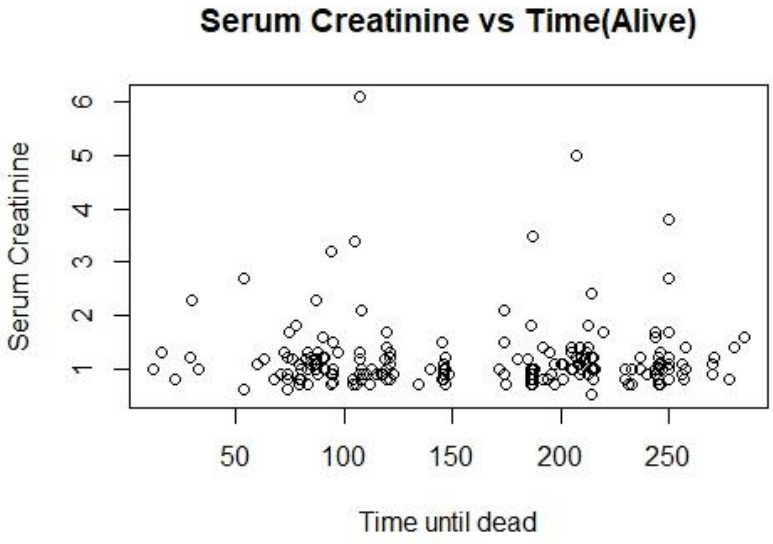
The p-value we got from the t test is $6.39*10^{-5}$ which is less than 0.05. So, we can reject null hypothesis. So, at 5% significance level there is enough evidence to say the mean values of serum creatinine level in blood in dead and alive people are different.

12.

This is the scatterplot of Serum Creatinine vs Time of died people



This is the scatterplot of Serum Creatinine vs Time of alive people



Here also we can see the difference of Serum Creatinine levels in dead and alive.

13.

This is the frequency table for death event with anaemia factor. Using this table and calculating relative risk we can measure how the anaemia factor affect the death event.

	DEATH_EVENT	
anaemia	Dead	Alive
Low	46	83
Normal	50	120

The relative risk we got here is 1.21 which is greater than 1. Which means when compared to a person who has a normal anaemia level there is 1.21 higher risk of dying for a person who has a lower anaemia level.

5 Conclusion

- In this data set there are details about 299 heart patients. 96 of them died.
  So, the death proportion of the study population is 0.32. It is almost same as in the target population which is the death proportion is 0.33. So, this dataset represent the real life situation well.

- By looking at the count vs sex stacked bar chart we can conclude that being a male or a female doesn't make any impact on being alive or dead.

- Mean values of creatinine phosphokinase amount in dead and alive people are not significantly different, so it can't be used to identify whether a person is going to be dead or not

- According to the data set we can conclude that smoking doesn't make any difference in life or death of CVD patients.

- As same as in smoking, having diabetes or not, doesn't make any difference in life or death of CVD patients.

- But having lower level of anaemeia increases the risk of death among CVD patients.

- To identify a patient is at a risk of death can be identified using their platelets count. Because in dead patients it is lower than 100000 (kilo platelets/mL) , also it can be seen that platelets count has no relationship with age.

- From the results of two sample t-test It can be identified that if a CVD patient has creatinine phosphokinase level more than 120, he/she at risk of death from CVD.

- Serum sodium and serum creatine can also be used to identify the death risk of a patient, since the dead patients had different amount of serum sodium and serum creatine levels than the patients who survived.

6 R Code

```
setwd("H:\\V SEM\\ST 3010 –Introduction to Health Statistics\\assignment")
data1=read.csv("heart_failure_clinical_records_dataset.csv")
attach(data1)
#to get overview
summary(data1)
str(data1)
#to get total observations(since there are no any missing values)
l=length(age)
l

#install.packages("descriptr")
library(descriptr)
sex_count=table(sex)
sex_count
dim(sex_count)
death_count=table(DEATH_EVENT)
death_count

#charts

pie(death_count,main = "Death & Alive", col = c("yellow","red"),labels = c("Alive","Dead"))
barplot(sex_count,xlab = "Gender",ylab = "Frequencies",col = c("pink","blue"),names.arg =
c("Female","Male"),main = "Barplot of Gender")
boxplot(creatinine_phosphokinase~DEATH_EVENT)
t.test(new_data2$creatinine_phosphokinase,new_data$creatinine_phosphokinase,alternative =
"two.sided",var.equal = FALSE)

#Dead vs smoking
table_Smoking_Dead=table(smoking=factor(smoking,levels=c("1","0")),DEATH_EVENT=factor(DEATH
_EVENT ,levels=c("1","0")))
table_Smoking_Dead
colnames(table_Smoking_Dead)=c("Dead","Alive")
rownames(table_Smoking_Dead)=c("Yes","No")
table_Smoking_Dead

#install.packages("epiR")
library(epiR)

epi.2by2(table_Smoking_Dead,method="cohort.count",conf.level = 0.95)
percentage_relativerisk=(0.961-1)*100
percentage_relativerisk

#Dead vs diabetes
table(diabetes,DEATH_EVENT)
table_diabetes_Dead=table(diabetes =factor(diabetes
,levels=c("1","0")),DEATH_EVENT=factor(DEATH_EVENT ,levels=c("1","0")))
table_diabetes_Dead
```

```
colnames(table_diabetes_Dead)=c("Dead","Alive")
rownames(table_diabetes_Dead)=c("Yes","No")
table_diabetes_Dead

epi.2by2(table_diabetes_Dead,method="cohort.count",conf.level = 0.95)
percentage_relativerisk1=(0.99-1)*100
percentage_relativerisk1


#new_data = data1[which(DEATH_EVENT=="1")]
new_data=subset(data1,DEATH_EVENT=="1")
new_count=table(new_data$sex)
new_count
qqnorm(new_data$platelets)
qqline(new_data$platelets)
shapiro.test(new_data$platelets)
hist(new_data$platelets,col = "grey",main = "Histogram of Platelets Of People Who Died")
sactterplot=plot(age,platelets,xlab = "Age",ylab = "Platelets",col=c("black"),main = "Platelets vs Age")

#ttest
t_test_platelets=t.test(new_data$platelets,alternative = "less",conf.level = 0.95,mu=100000)
t_test_platelets$statistic
t_test_platelets
qt(0.05,95)

#creatinine level (should between 10-120)
t_test_creatinine=t.test(new_data$creatinine_phosphokinase,alternative = "two.sided",conf.level =
0.95,mu=65)
t_test_creatinine
t_test_creatinine1=t.test(new_data$creatinine_phosphokinase,alternative = "greater",conf.level =
0.95,mu=120)
t_test_creatinine1
t_test_creatinine2=t.test(new_data$creatinine_phosphokinase,alternative = "less",conf.level = 0.95,mu=10)
t_test_creatinine2

#serum sodium level with dead
new_data2=subset(data1,DEATH_EVENT=="0")
new_data2_count=table(new_data2$sex)
new_data2_count

#t test for two independent samples
#serum sodium
t_test_sodium = t.test(new_data2$serum_sodium,new_data$serum_sodium,alternative =
"two.sided",var.equal = FALSE)
t_test_sodium
df=((var(new_data2$serum_sodium)/203 +
var(new_data$serum_sodium)/96)^2)/(((var(new_data2$serum_sodium)/203)^2/202)+((var(new_data$seru
m_sodium)/96)^2 /95))
df
qt(0.975,df)
```

```
t_test_sodium$statistic
(1-pt(3.1645,df=df))*2

#serum_creatinine
t_test_serum_creatinine = t.test(new_data2$serum_creatinine,new_data$serum_creatinine,alternative =
"two.sided",var.equal = FALSE)
t_test_sodium
df=((var(new_data2$serum_creatinine)/203 +
var(new_data$serum_creatinine)/96)^2)/(((var(new_data2$serum_creatinine)/203)^2/202)+((var(new_data$
serum_creatinine)/96)^2 /95))
df
qt(0.025,df)
t_test_sodium$statistic
(1-pt(3.1645,df=df))*2

#time and serum_creatinine
sactterplot1=plot(new_data$time,new_data$serum_creatinine,xlab = "Time until dead",ylab = "Serum
Creatinine",col=c("black"),main = "Serum Creatinine vs Time (Dead)")
sactterplot1=plot(new_data2$time,new_data2$serum_creatinine,xlab = "Time until dead",ylab = "Serum
Creatinine",col=c("black"),main = "Serum Creatinine vs Time(Alive)")

#Dead vs anaemia
table_anaemia_Dead=table(anaemia=factor(anaemia,levels=c("1","0")),DEATH_EVENT=factor(DEATH_
EVENT ,levels=c("1","0")))
table_anaemia_Dead
colnames(table_anaemia_Dead)=c("Dead","Alive")
rownames(table_anaemia_Dead)=c("Low","Normal")
table_anaemia_Dead


epi.2by2(table_anaemia_Dead,method="cohort.count",conf.level = 0.95)
percentage_relativerisk=(1.21-1)*100
percentage_relativerisk
```