Applied Analytics Assignment-2

Cholestrol and exercise induce angina association to heart diseases.

Sai Vamsi Chunduru - S3884753, Pragati Patidar – S3858702, Kyron Reshi – S3920193, Arjun Padmanabha Pillai – S3887231

last updated on 17 October, 2021

RPubs link information

-Cholestrol and exercise induce angina association to heart diseases.

- https://rpubs.com/Pragati/823209

Introduction

- An estimated 17.9 million people died from Cardiovascular diseases (CVDs) in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.
- Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2019, 38% were caused by CVDs.
- Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol.
- It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin.
- It is imperative to determine the key biological and environmental indicators that affect Cardiovascular disease to adequately allocate resources in society to combat this disease. -The dataset used in this project is UCI Heart Disease dataset,the original database includes 76 attributes, we are using a subset of 14 of them.
- In this project we performed a statistical analysis on likelihood of heart disease. So, we focused on 2 of these indicators (cholesterol and Exercise induced angina) to determine whether they had an effect on Cardiovascular disease.

Problem Statement

- Do individuals with Exercise induced angina have a higher likelihood of a heart attack?
- Does having high cholesterol increases the likelihood of a heart attack?
- Summary: The cholesterol,Exercise_induce_angia and target variable are used in this investigation.A histogram and bar plot is plotted based on Exercise_induce_angia and target variable, a chi-square association test is performed to test the association.
- A normality test is used to ensure that the data is is drawn from a normal population distribution. After assumptions are tested, a two sample t tests is used to evaluate if there is a difference in the mean of cholesterol levels in heart diseases and no heart diseases.
- The following hypothesis have been set out and tested through this investigation:
- There is association between exercise induce angina and risk of heart attack of an individual.
- The mean of cholesterol levels are higher in heart disease individuals.

Data

- Our data “heart.csv” was open source retrieved from 4 creators across Cleveland Hospital, University Hospital Zurich, University Hospital Basel and the Hungarian Institute of Cardiology. Source: https://www.kaggle.com/ronitf/heart-disease-uci/activity

- Description of Variables age: The person’s age in years, sex: The person’s sex (1 = male, 0 = female), cp: The chest pain experienced (Value 0: asymptomatic, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: typical angina), trestbps: The person’s resting blood pressure (mm Hg on admission to the hospital), fbs: The person’s fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false), restecg: Resting electrocardiograph measurement (Value 0: showing probable or definite left ventricular hypertrophy by Estes’ criteria, Value 1: normal, Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)), thalach: The person’s maximum heart rate achieved, oldpeak: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot. See more here), slope: the slope of the peak exercise ST segment, ca: The number of major vessels (0-3), thal: A blood disorder called thalassemia (1 = fixed defect; 2 = normal; 7 = reversible defect).

- The data set contains 14 attributes from 303 patients, however we focused on below three attributes

- Cholesterol: Serum Cholesterol level of patient in a continuous variable denoted in mg/dl from 126mg/dl to 564mg/dl.

- Exercise_Induced_Angina: Whether patient has exercise induced angina as a discrete variable originally denoted as [1,0], however we factorized and assigned labels 1 = yes, 0 = no.

- target: Whether the patient had cardiovascular disease as a discrete variable originally denoted as [1,0], however we factorized and assigned labels 1 = heart disease, 0 = no heart disease.

- By using both descriptive and continuous variables, we were able to use various statistical methods to determine relationships between them.

Descriptive Statistics and Visualisation

```
hd<-read_csv("heart.csv")
#renaming variables
names(hd)[5] <- 'Cholesterol'
names(hd)[9] <-'Exercise_Induced_Angina'
#factorizing variables
#target variable as factor variable 1 for  having disease and 0 for not having heart disease
hd$target <- hd$target %>% factor(levels=c(0,1),
                                        labels=c("no heart disease","heart disease"))
#Exercise_Induced_Angina variable as factor variable 1 for yes and 0 for no.
hd$Exercise_Induced_Angina<- hd$Exercise_Induced_Angina%>% factor(levels=c(1,0) , labels=c("Yes","No"))
#summarizing required variables
summary(hd$Cholesterol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   126.0   211.0   240.0   246.3   274.5   564.0
```

```
summary(hd$Exercise_Induced_Angina)
```

```
## Yes  No
##  99 204
```

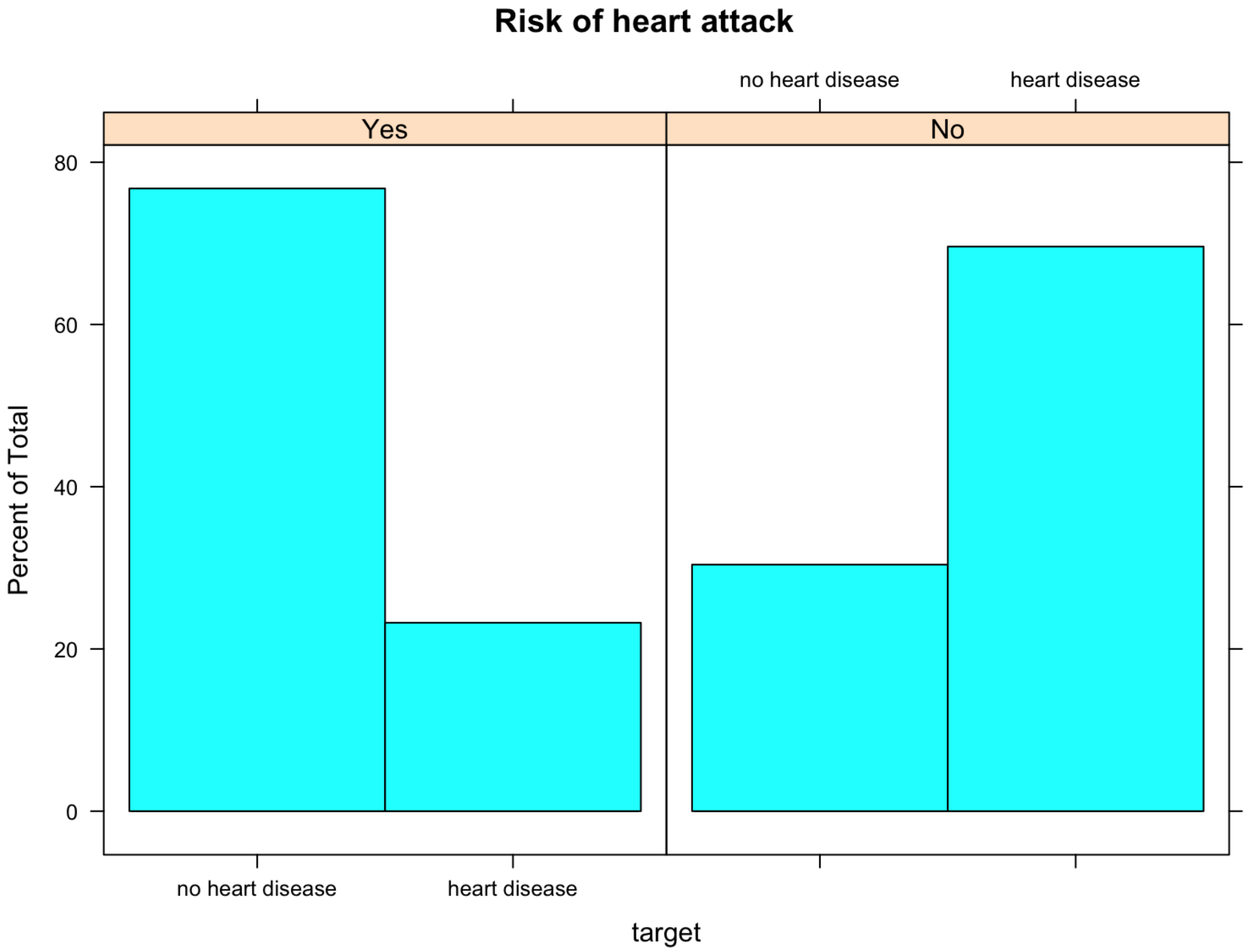
```
summary(hd$target)
```

```
## no heart disease    heart disease
##              138              165
```

Decsriptive Statistics Cont.

- By observing the histogram, exercise induced angina-no appears to have a higher likelihood of a heart attack as the proportion is much bigger compared to the exercise induced angina-yes whose likelihood of heart attack looks far too less.

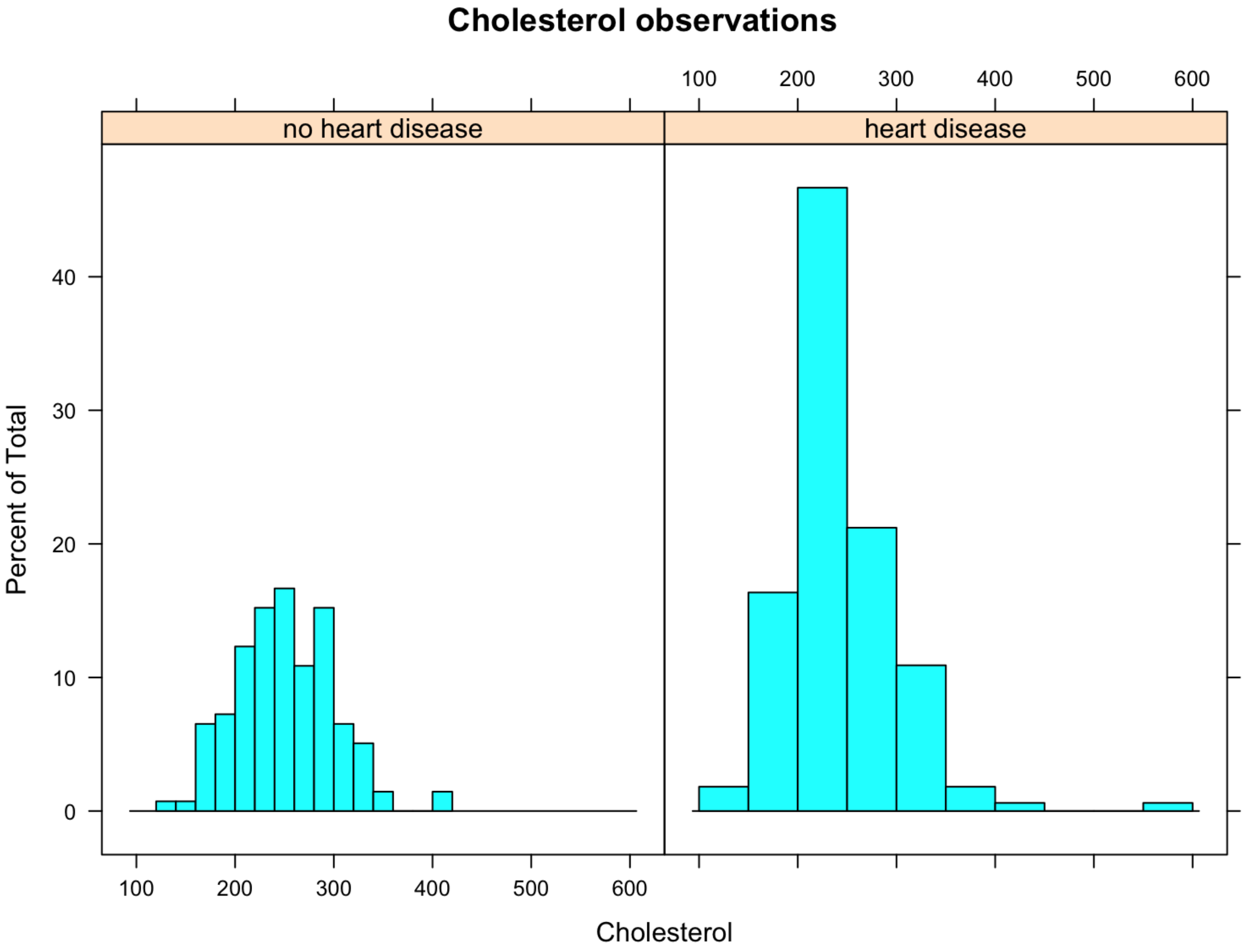
```
hd %>% histogram(~target | Exercise_Induced_Angina, data= ., main = "Risk of heart attack")
```



Decsriptive Statistics Cont1

-In terms of the Cholesterol of individuals, the no heart disease sample appears to have the highest percent of total in the range of 220 to 240 cholesterol levels followed by 280 to 300 cholesterol levels. While the heart disease sample appears to have the highest percent of total in the range of 200-250 cholesterol levels.

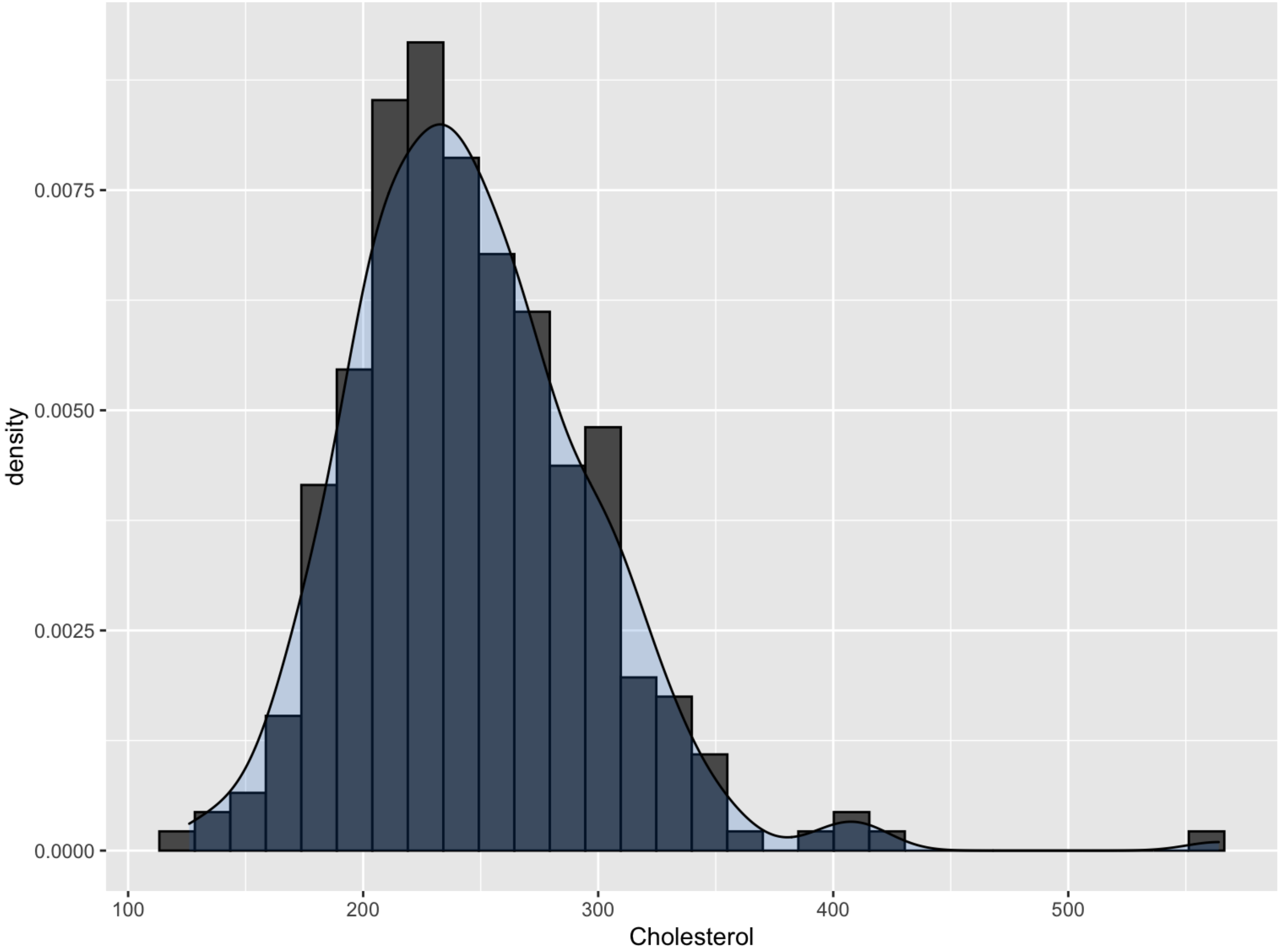
```
hd %>% histogram(~Cholesterol | target, data= ., main = "Cholesterol observations", breaks=10)
```



Decsriptive Statistics Cont2

-The plot below shows the distribution in the cholesterol levels of the individual in this investigation. The curve appears to be more positively skewed.

```
#code for checking  distribution of data -bins=30
hd %>% ggplot(aes(x=Cholesterol)) + geom_histogram(aes(y=..density..), colour="black")+
        geom_density(alpha=.2, fill="dodgerblue3")
```



Hypothesis Testing- Pearson’s Chi-squared test of association

- H0: Likelihood that heart attack and Exercise_Induced_Angina are not associated
- H1: Likelihood that heart attack and Exercise_Induced_Angina are associated- Assumption is that no more than 25% of expected counts are less than 5 and all individual counts are 1 or greater.
- table_ang_target1 shows that Exercise_Induced_Angina-‘yes’ have a 0.13 chance of high likelihood of heart attack while Exercise_Induced_Angina-‘no’ have a 0.86 chance of high likelihood of heart attack.

```
table_ang_target <- table(hd$Exercise_Induced_Angina , hd$target)
table_ang_target %>% addmargins()
```

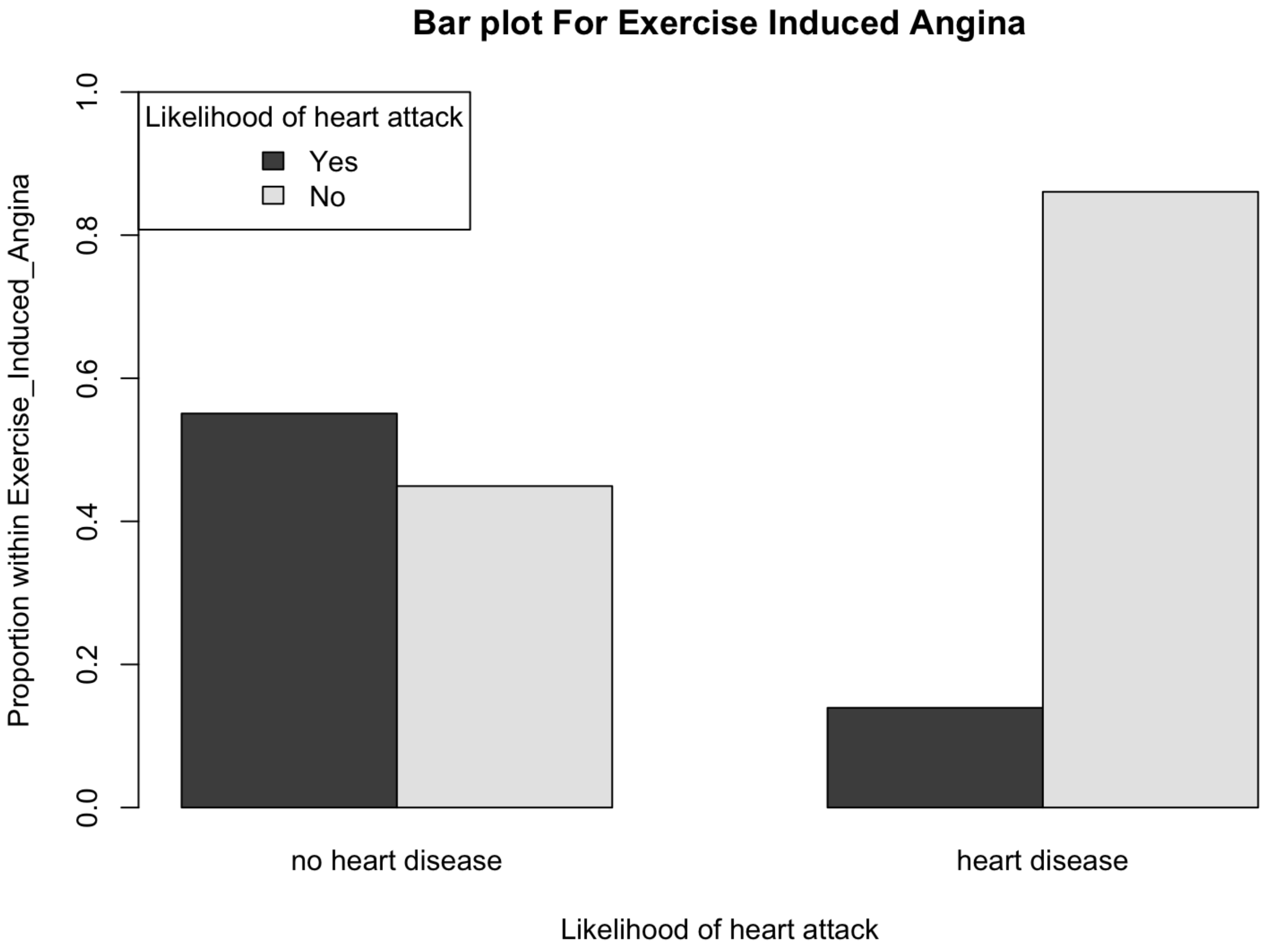
```
##
##      no heart disease heart disease Sum
##   Yes               76            23  99
##   No                62           142 204
##   Sum              138           165 303
```

```
table_ang_target1 <- table_ang_target %>% prop.table(margin=2) #proportions
table_ang_target1
```

```
##
##      no heart disease heart disease
##   Yes        0.5507246     0.1393939
##   No         0.4492754     0.8606061
```

Hypothesis Testing cont.

```
barplot(table_ang_target1, main="Bar plot For Exercise Induced Angina",
        ylab="Proportion within Exercise_Induced_Angina", xlab="Likelihood of heart attack",
        ylim=c(0,1),legend=row.names(table_ang_target1), beside=TRUE,
        args.legend=c(x="topleft",horiz=FALSE,title="Likelihood of heart attack"))
```



Hypothesis Testing cont1

```
chi_ang_target <- chisq.test(table_ang_target)
chi_ang_target
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_ang_target
## X-squared = 55.945, df = 1, p-value = 7.454e-14
```

```
chi_ang_target$expected ##  Pearson's Chi-squared test
```

```
##
##       no heart disease heart disease
##   Yes         45.08911      53.91089
##   No          92.91089     111.08911
```

- The chi-square test of association for the Exercise_Induced_Angina and (target)likelihood of heart attack gives a p-value = 7.454e-14 which is less than the 0.001. Therefore the null hypothesis, H0 is rejected and the chi-square test of association is statistically significant. The results tell us that that is Exercise_Induced_Angina-‘yes’ leads to no heart disease in most cases.Therefore the likelihood of a heart disease/no heart disease is dependent on Exercise_Induced_Angina of an individual.

Hypothesis Testing- Independent Two sample t-Test :

- 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 the person's cholesterol measurement (in mg/dl) and likelihood of heart disease are equal or not. We want to know if the mean score for the cholesterol measurement of person having heart disease is different from the person does not have heart disease.
- Using two variables, One variable Target defines the two groups heart disease and no heart disease. The second variable is the measurement of interest that is cholesterol which continuous variable. -The data values are independent. The cholesterol measurement for any one person does not depend on the cholesterol measurement for another person. -The summary statistics table shows the sample mean for the groups (no heart disease= 131 or heart disease= 126) compared with person's cholesterol measurement.
- Above mentioned histogram compares the distribution of data. -By plotting QQ Plot a simple random sample from the population. We assume the data are normally distributed, as the sample size in both groups are greater than 30, sampling distribution will approximate a normal distribution.. -The data values are cholesterol measurements. The measurements are continuous by above histogram. -The variances for heart disease and no heart disease are equal, by using Homogeneity of variances by levene Test.
- As p(p-value = 0.1388)>0.05, population variances are homogeneous. Now we can apply the two-sample t-test. b- State the Null and Alternate hypothesis for the appropriate hypothesis test. c- Report the test statistic, , -value and 95% CI of the mean difference from the results of the hypothesis test.
- H0:M1=M2 (mean score for the cholesterol measurement of person having heart disease is equal from the person does not having heart disease.)
- HA:M1!=M2 (mean score for the cholesterol measurement of person having heart disease is equal from the person does not having heart disease).

Two sample t-test hypothesis cont.

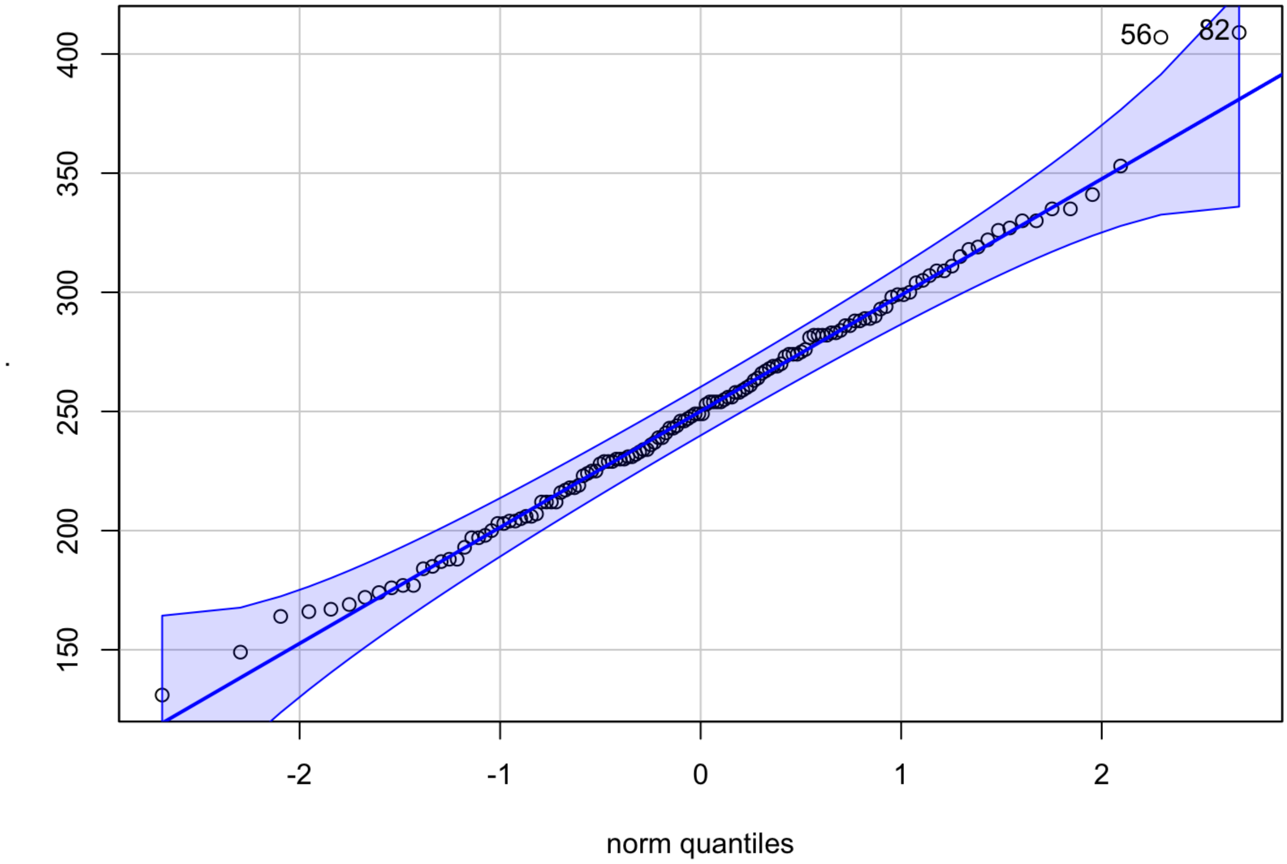
```r
hd %>% group_by(target) %>% summarise(Min = min(Cholesterol,na.rm = TRUE),
                                           Q1 = quantile(Cholesterol,probs = .25,na.rm = TRUE),
                                           Median = median(Cholesterol, na.rm = TRUE),
                                           Q3 = quantile(Cholesterol,probs = .75,na.rm = TRUE),
                                           Max = max(Cholesterol,na.rm = TRUE),
                                           Mean = mean(Cholesterol, na.rm = TRUE),
                                           SD = sd(Cholesterol, na.rm = TRUE),
                                           n = n(),
                                           Missing = sum(is.na(Cholesterol))) -> table1

knitr::kable(table1)
```

target	Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
no heart disease	131	217.25	249	283	409	251.0870	49.45461	138	0
heart disease	126	208.00	234	267	564	242.2303	53.55287	165	0

Two sample t-test hypothesis cont1

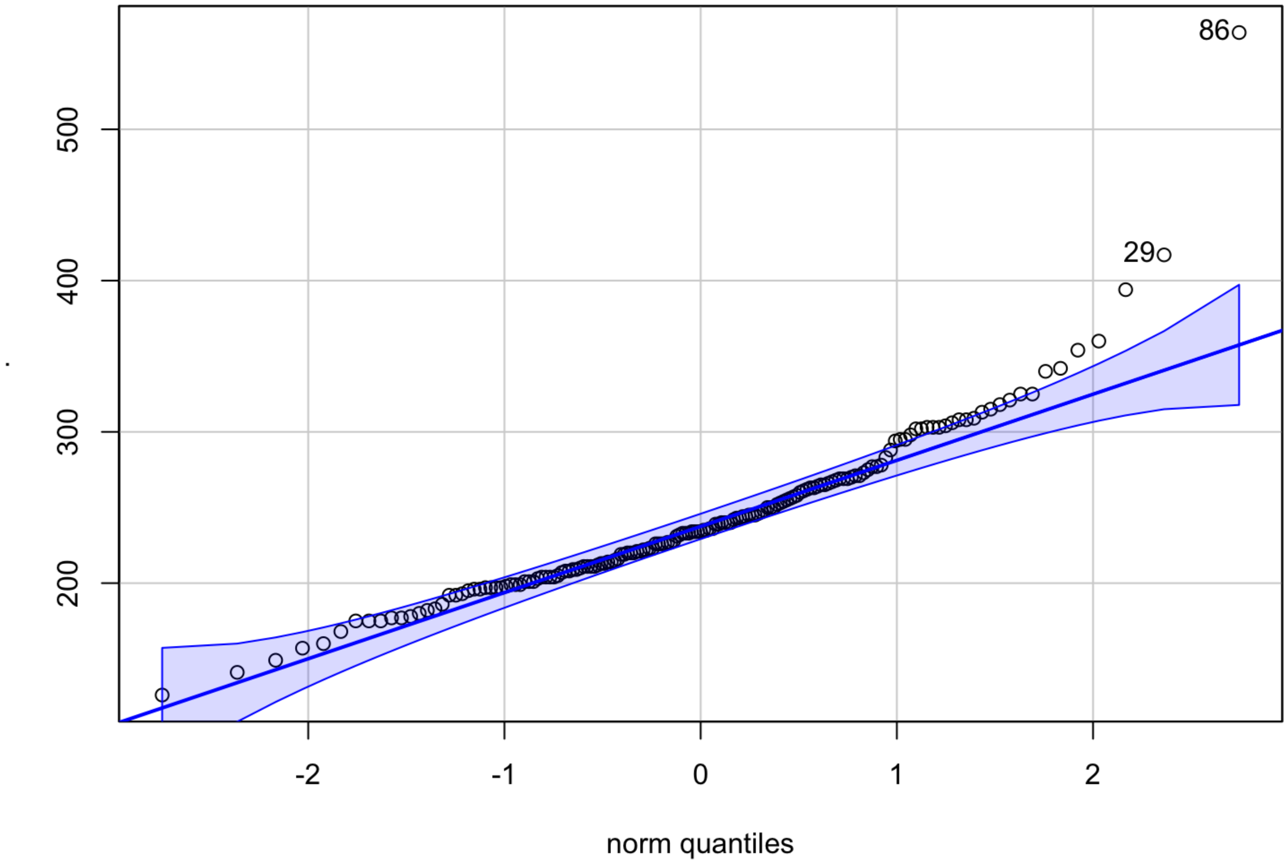
```
# QQ plot for target == "no heart disease" for showing distribution
target_no <- hd %>% filter(target == "no heart disease")
target_no$Cholesterol%>% qqPlot(dist="norm")
```



```
## [1] 82 56
```

Two sample t-test hypothesis cont2

```
# QQ plot for target == "no heart disease" for showing distribution
target_yes<- hd %>% filter(target == "heart disease")
target_yes$Cholesterol%>% qqPlot(dist="norm")
```



```
## [1] 86 29
```

Two sample t-test hypothesis cont4

```
Ltest<-leveneTest( Cholesterol ~ target, data = hd) # Homogeneity of variances using leven test
knitr::kable(Ltest)
```

	Df	F value	Pr(>F)
group	1	0.1014635	0.7503013
	301	NA	NA

```
test_result<- t.test(Cholesterol ~ target,
                data = hd,
                var.equal = TRUE, alternative = "two.sided" ) #Independent two sample t-Test
test_result
```

```
##
##  Two Sample t-test
##
## data:  Cholesterol by target
## t = 1.4842, df = 301, p-value = 0.1388
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.885882 20.599189
## sample estimates:
## mean in group no heart disease    mean in group heart disease
##                       251.0870                       242.2303
```

Two sample t-test hypothesis cont5

```
#Use the p-value and CI of the mean to make a decision about the null hypothesis.
test_result$p.value
```

```
## [1] 0.1387903
```

```
test_result$conf.int
```

```
## [1] -2.885882 20.599189
## attr(,"conf.level")
## [1] 0.95
```

- To Conclude whether or not the results of the hypothesis test were statistically not significant and draw a conclusion in the context of the example. Our decision should be to reject H0: µ1 = µ1(null hypothesis) as the p > .05 and the 95% CI of the estimated population difference [-0.60, -0.13], which did not capture H0: µ1 - µ1 = 0. The results of the two-sample t-test were therefore statistically significant. This meant that the mean cholesterol level for people with heart disease was almost same ones who did not. Or it can be concluded there is no association between cholesterol level and heart disease.

Discussion

- The first analysis was based on a categorical association if exercise induced angina plays a role in the likelihood of a heart attack. Based on the visualization, we can observe that there is less likelihood of heart disease in exercise induce angina individuals compared to individuals with no exercise induce angina.
- Using the chi square test of association between exercise induce angina and likelihood, the p-valye is les than 0.001 and is statistically significant. We can conclude likelihood of having or no having a heart disease is dependent on the exercise induced angina of the individual.
- In the second analysis, it can be noted that the average level of cholesterol for people with heart disease is the same as the average cholesterol level for people without heart disease, this clearly shows that there is no association between cholesterol and our target heart disease.
- A normal distribution plot of cholesterol variable indicates a very clear positive skew a large number of individuals have their cholesterol point between the range of 200 and 300 which is considered high. To further strengthen and reinforce our conclusion a QQ plot was done for the same assuming normally distributed data and homogeneity of variances was tested using levene test which came out to be true, upon this two sample t tests were conducted we got a p value greater than 0.05 and hence rejected the null hypothesis.
- The results of the two sample t tests were statistically significant which pointed to the fact that the average cholesterol level in both people with and without heart disease was the same.
- Hence, we can conclusively say that there is there is no statistically significant difference between the person's cholesterol levels regardless of the presence or absence of heart disease. Nonetheless, there is an association between exercise induced angina of the individual and heart disease.

References

- Dataset Source: "Heart Disease UCI, heart.csv https://archive.ics.uci.edu/ml/datasets/Heart+Disease created by:

  1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
  2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
  3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
  4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

- Centers for Disease Control and Prevention, National Center for Health Statistics. “About Multiple Cause of Death, 1999–2019”. CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2019. Accessed February 1, 2021. https://wonder.cdc.gov/mcd-icd10.html"
- McLeod, S. A. (2019, May 20). What a p-value tells you about statistical significance. Simply Psychology. https://www.simplypsychology.org/p-value.html
- https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)