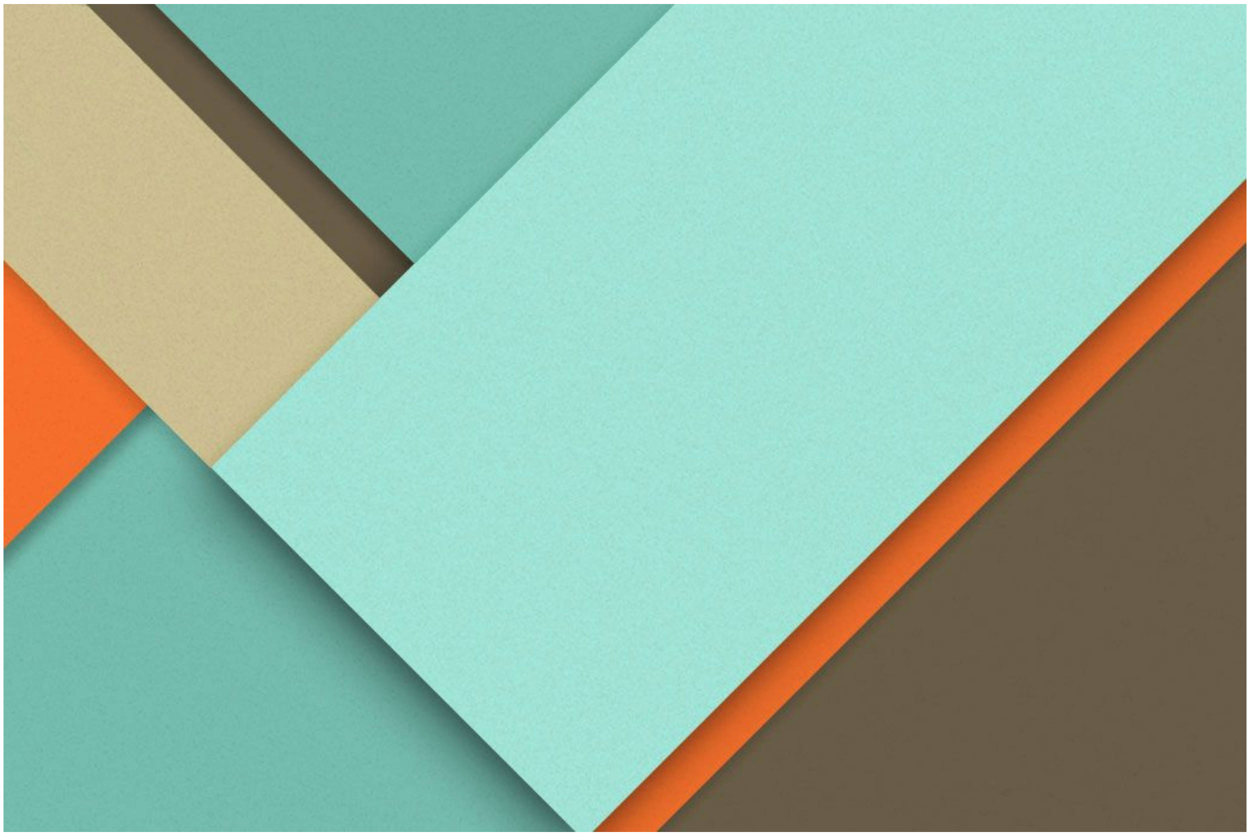


What are the factors that contribute to Heart Diseases?

06.12.2019

Phani Sai Kamal Lingam
NUID: 001063813
MS in Data Analytics and Engineering
Northeastern University - Seattle

Introduction

The Framingham Heart Study Dataset is part of the prestigious project of The National Heart, Lung & Blood Institute and Boston University, which started in 1948 and committed to identifying the common factors that contribute to cardiovascular disease. The study is conducted by recruiting more than 5,000 people, especially between the ages of 30 and 62 from the town of Framingham, Massachusetts who had not yet developed any overt symptoms of stroke or suffered from a heart attack. Over the years, the Framingham Study has proven the major CVD risk factors and the effects of these factors such as total cholesterol, age, gender, etc.. Since the study has proven itself to be reliable has drawn my interest to make inferences about the factors responsible for such heart diseases.

The dependence and correlation between the factors of study and risk of developing symptoms of heart disease over time due to these factors made me frame my question in the most possible way to utilize the data to its fullest potential and draw meaningful insights from it. With each question in each test, I tried to conclude upon the effects of the factors with reasonable statements.

The Framingham Heart Study Dataset came with 8 categorical variables like gender, smoking behavior, diabetic condition etc., and 8 quantitative variables like age, heart rate, number of cigarettes consumed per day etc., with few missing entries.

Variable Name	Variable Type	Description	Units
Gender	Categorical / Nominal	Sexual orientation with levels as Male and Female.	NA
Age	Quantitative / Discrete	Calendar years between the age of 32 and 70.	Years
Education	Categorical / Ordinal	Represents the Highest level of education received by individuals with the	NA

		level being High School, General Education Development, Vocational School and College.	
Smoking Behavior	Categorical / Nominal	A flag denoting whether the individual has a smoking habit or not.	NA
Cigarettes Per Day	Quantitative / Discrete	The count of cigarettes consumed per day by a smoker.	Count
Blood Pressure Medication	Categorical / Nominal	Identifies whether the individual is receiving any medication for Blood Pressure control or not.	NA
Prevalent Stroke	Categorical / Nominal	A flag to represent any strokes in the past.	NA
Prevalent Hypertension	Categorical / Nominal	Binary variable denoting whether the person suffered from hypertension or not.	NA
Diabetic Condition	Categorical / Nominal	Status flag to show whether the person of interest is suffering from diabetes condition or not.	NA
Total Cholesterol	Quantitative / Discrete	The level of cholesterol in the human body.	mg/dL
Systolic Blood Pressure	Quantitative / Discrete	The top reading of your Blood Pressure level, generally 120mmHg.	mmHg
Diastolic Blood Pressure	Quantitative / Discrete	The bottom reading of your Blood Pressure	mmHg

		level, generally 80mmHg.	
Body Mass Index	Quantitative / Continuous	A factor that is helpful in identifying obesity levels.	kg/m 2
Heart Rate	Quantitative / Discrete	The number of times one's heartbeats per minute.	BPM
Glucose	Quantitative / Discrete	The body glucose level lying in the range of 40 to 400.	mg/dL
Ten Year Coronary Heart Disease Risk	Categorical / Nominal	A predicted outcome of whether the individual is Vulnerable or Immune to Heart Diseases in the future.	NA

Sampling Strategy and Concerns

The Framingham Heart Study is conducted specifically in the town of Framingham which raises concerns because of possible sampling bias or healthy user bias due to constraint of data being from a specific location. But the Study throughout all these years has proven this wrong and created its own name in Healthcare Research by reaching several milestones. Since the data is being used to understand the different factors and their effects on humans, the impact of bias is quite negligible and the profound results made it sound and reliable to use. To introduce randomization to the system, further subsampling and bootstrapping techniques are used to improve the data quality and attain trustworthy results.

Exploratory Data Analysis

In this section, the major factors that are responsible for heart diseases are identified and visualized in the form of clean and elegant graphs using the GGPlot2 library in R. The effects of confounding variables and the explanatory variables are depicted with the help of these graphs.

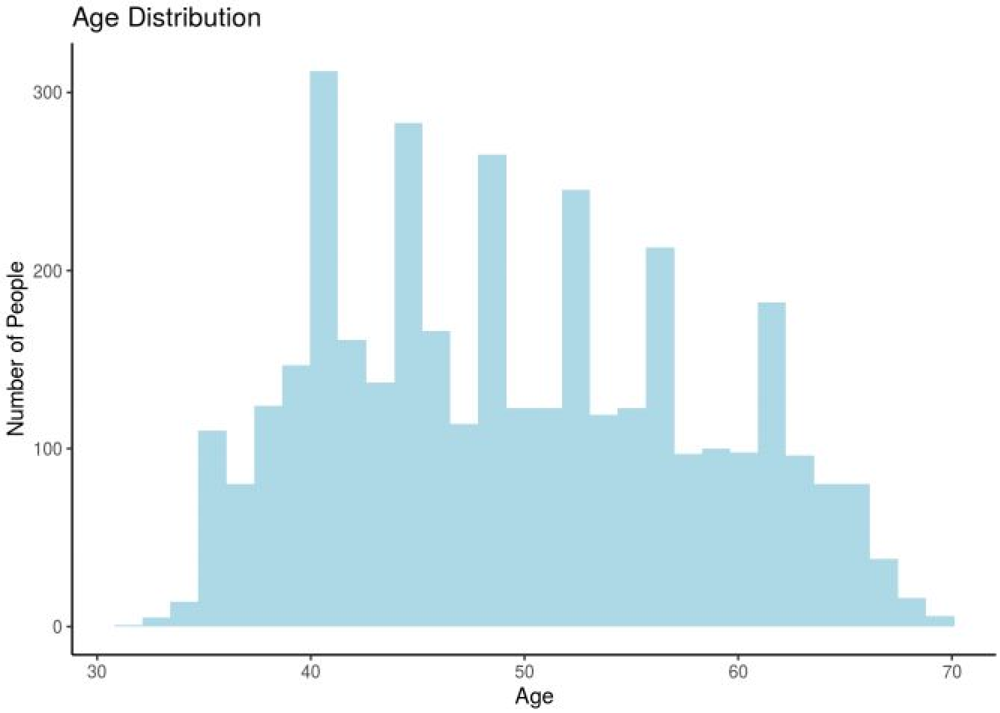


Fig.1. Histogram denoting Age Distribution.

The graph depicts the distribution of age of the population in the dataset. It denotes the interest age group for the study as between the ages of 30 and 70. We can also observe that most of the population belongs to the age interval between 35 and 55. Age is considered to be a major factor which dictates the risk of getting heart diseases. It is told than men after the age of 40 and women after the age of 45 are having more chances of suffering from CVDs.

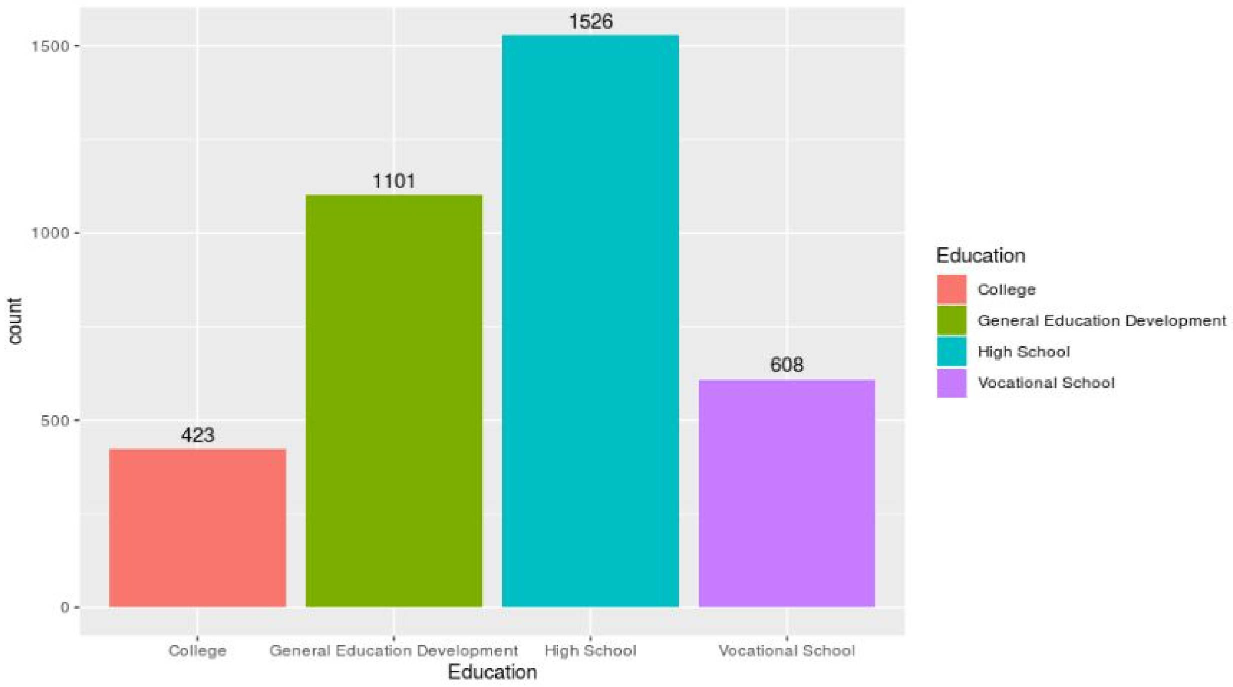


Fig. 2. Histogram depicting Level of Education of the population.

The histogram tells you about the education literacy of the participants in the study. Each bar represents a different level of education. We can infer that most of the participants received education up to High School only.

The education levels are ordered with the lowest being the High School and the highest being the College with General Education Development and Vocational School in the middle. Although literacy is not a factor causing heart diseases it helps us to understand the different backgrounds the population comes from and belongs to. But it is also considered that education helps to improve Health Cautiousness and also the standard of living of an individual. This will also be helpful during Chi-square Distribution Test as a requirement for the test is to have a categorical variable with more than 2 levels.

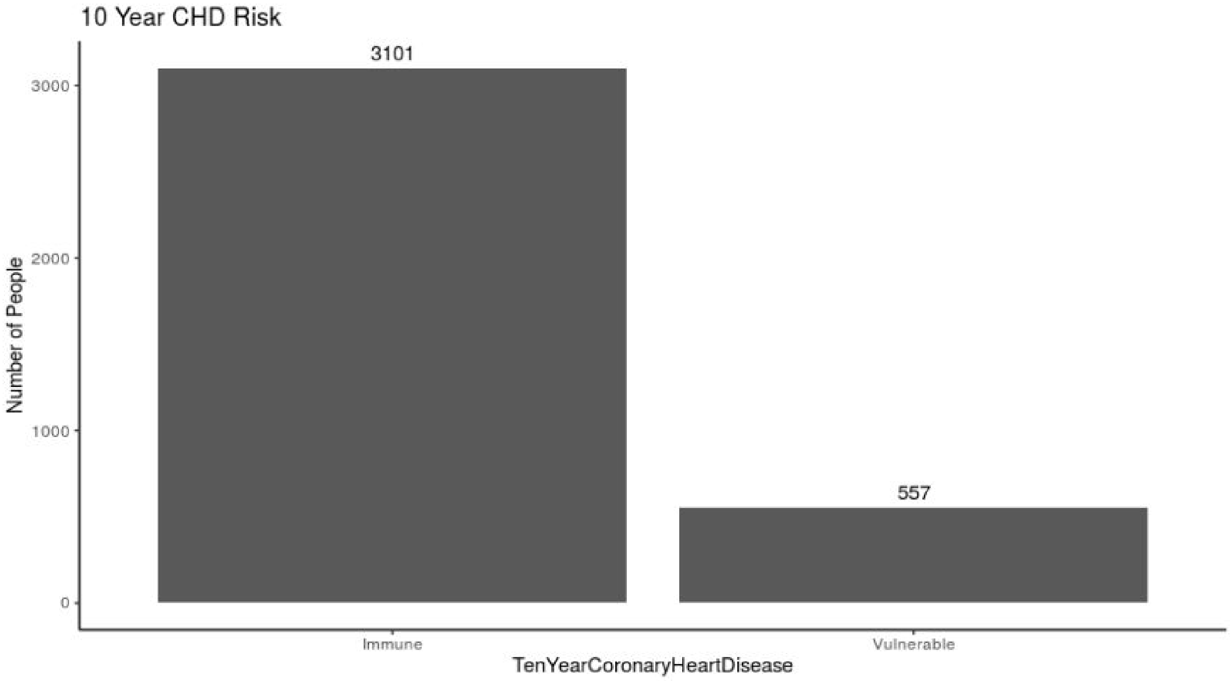


Fig. 3. Histogram depicting proportions of 10 Year CHD Risk.

The above histogram represents the proportions of the population who are Vulnerable to Coronary Heart Diseases and who are Immune to Coronary Heart Diseases in the next 10 years. In our data, most of the people are immune to CHD. But we do not know what makes them vulnerable to CHD yet. In order to draw some conclusions, the below graphs will provide a clear and board explanation about the factors and their effects on the population.

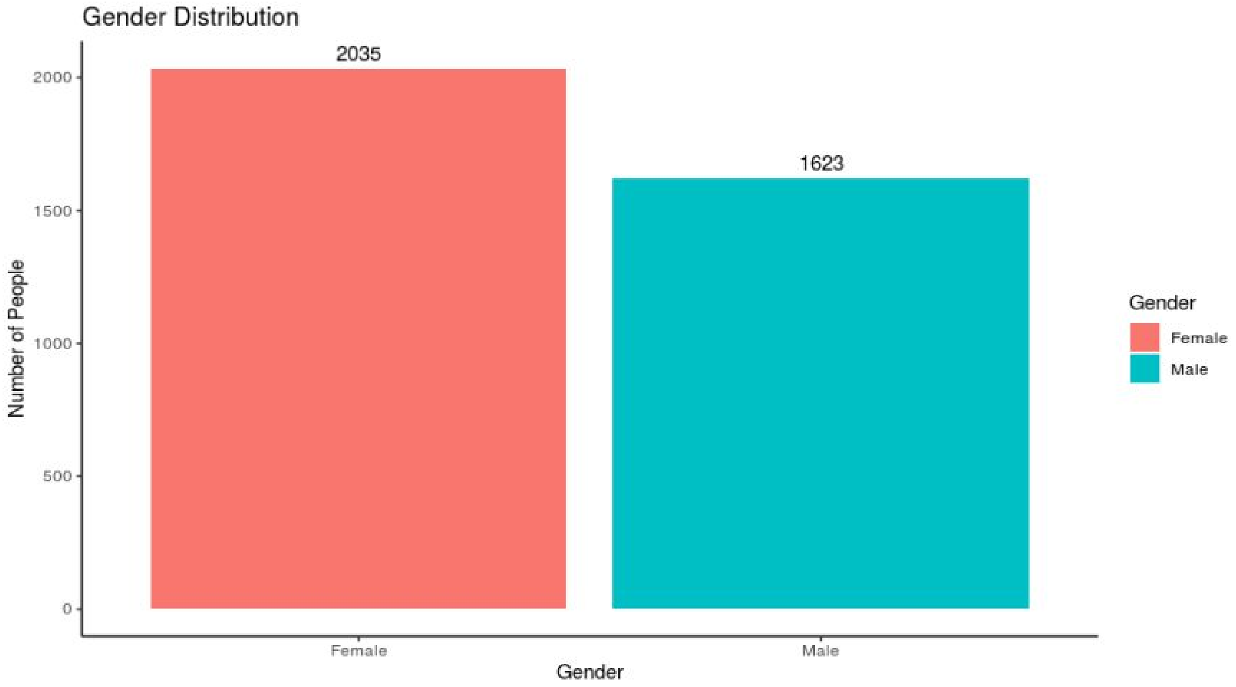


Fig. 4. Bar graph depicting Gender Distribution.

The above graph shows the proportions of Male and Female in population data. Gender have proven to be a crucial factor in Framingham Heart Study. It was observed that Male candidate are more likely to suffer from heart diseases than their Female counterparts, which will be proven later in the Statistical Analysis Section. In our data there is a higher proportion of Females than Males. Even the larger population of Females won't be able to cover up the differences using the testing. The below graph would provide a more detailed explanation visually about the heart disease risks.

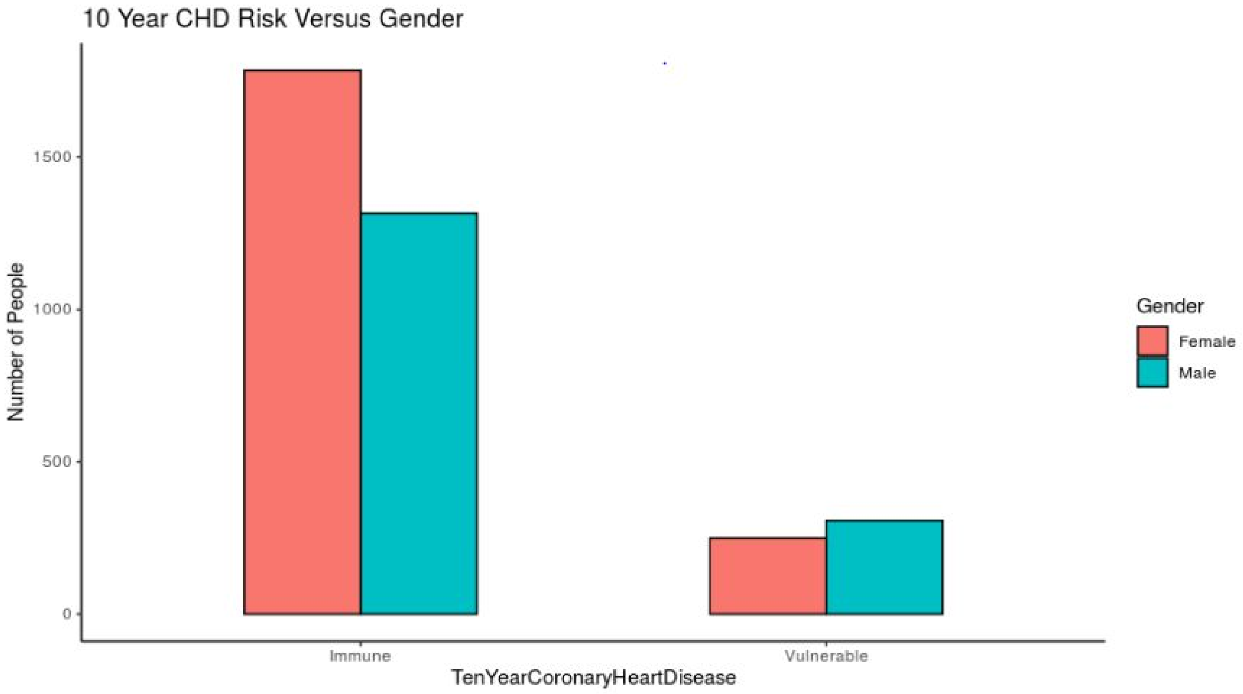


Fig. 5. Bar graph depicting Gender Versus Ten Year CHD Risk.

The graph creates a clear view about the dependence of risk of suffering from heart diseases on gender. The first pair of bars on the left side depicts the male and female who are immune to heart diseases. We can observe that less male are immune to heart diseases than that of females, this may be accepted due to the low population proportion of male category. But on the right hand side of the pair of bars depicts the population vulnerable to heart diseases. Here, even though the population of females is higher than the male, the proportion of male being vulnerable to heart diseases is higher than the female. This in a way proves the relation between heart diseases and gender.

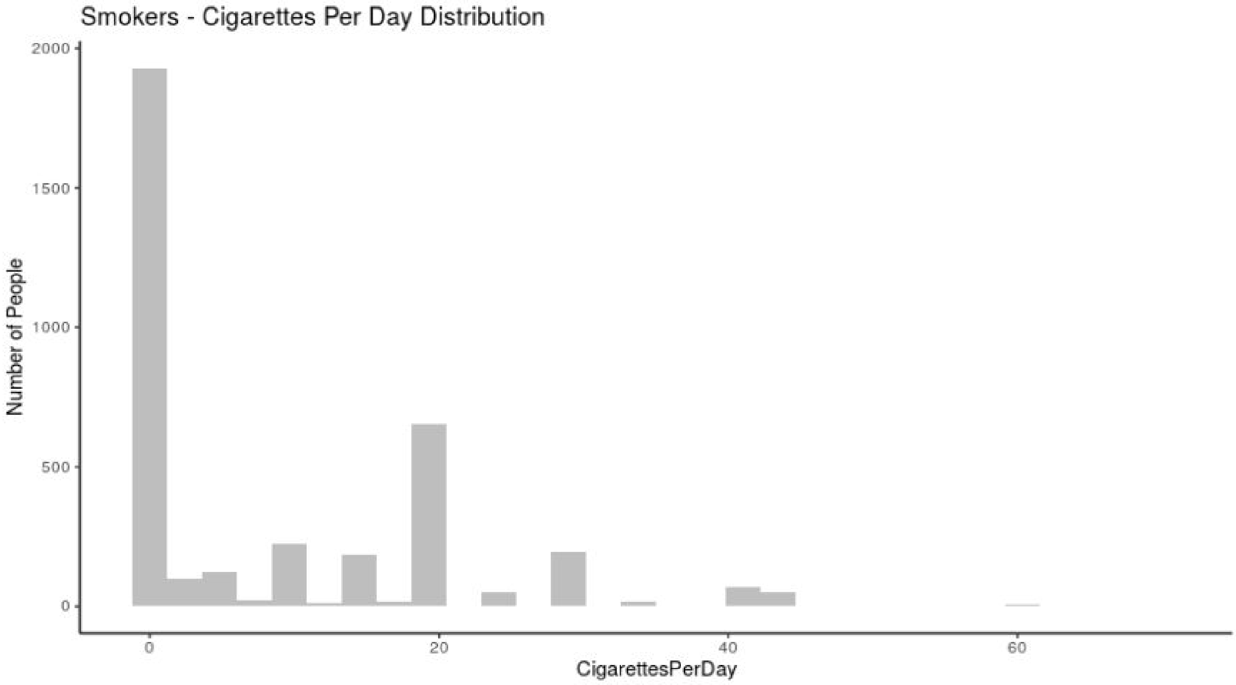


Fig. 6. Histogram depicting the Distribution of Cigarettes consumed per day.

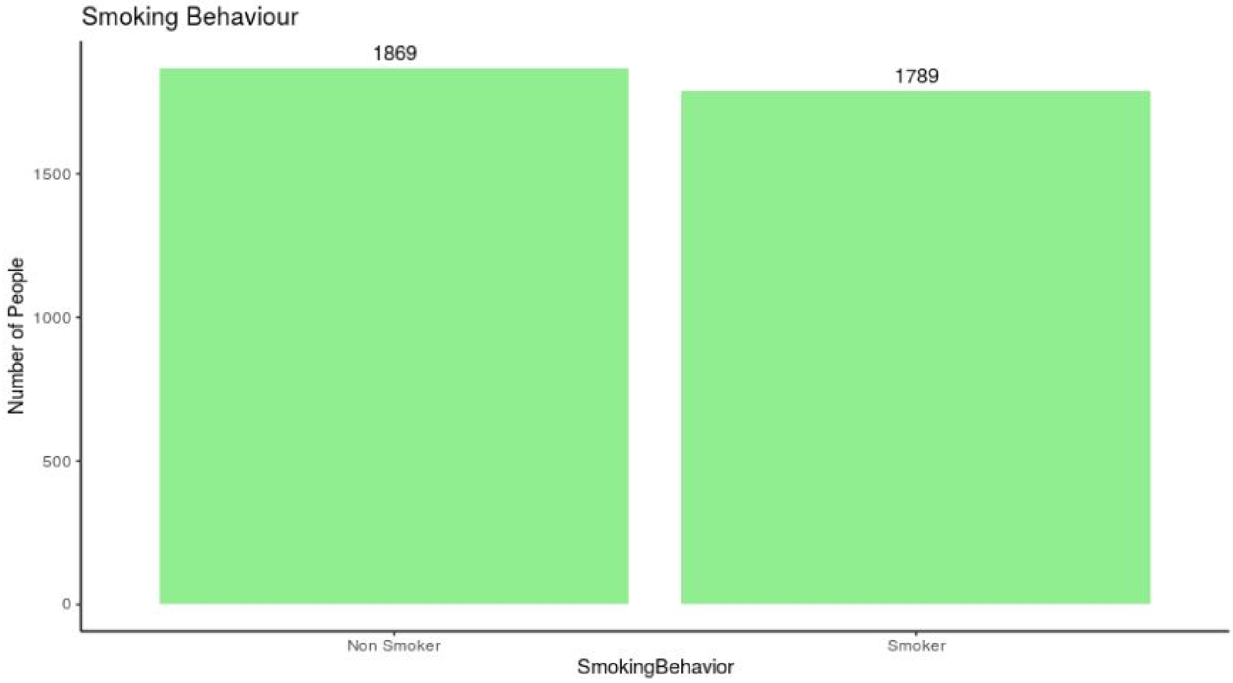


Fig. 7. Bar graph depicting the proportion based on Smoking Behavior.

The above graphs represents the Distribution of population based on the Cigarettes consumed per day. The upper graph is having a single monolithic bar on the left hand side at x = 0. This means that most of the population does not consume any cigarettes at all. This is also proven in the bottom graph. Here we can clearly see that the proportion of Non Smokers is higher than that of the Smokers in the population data.

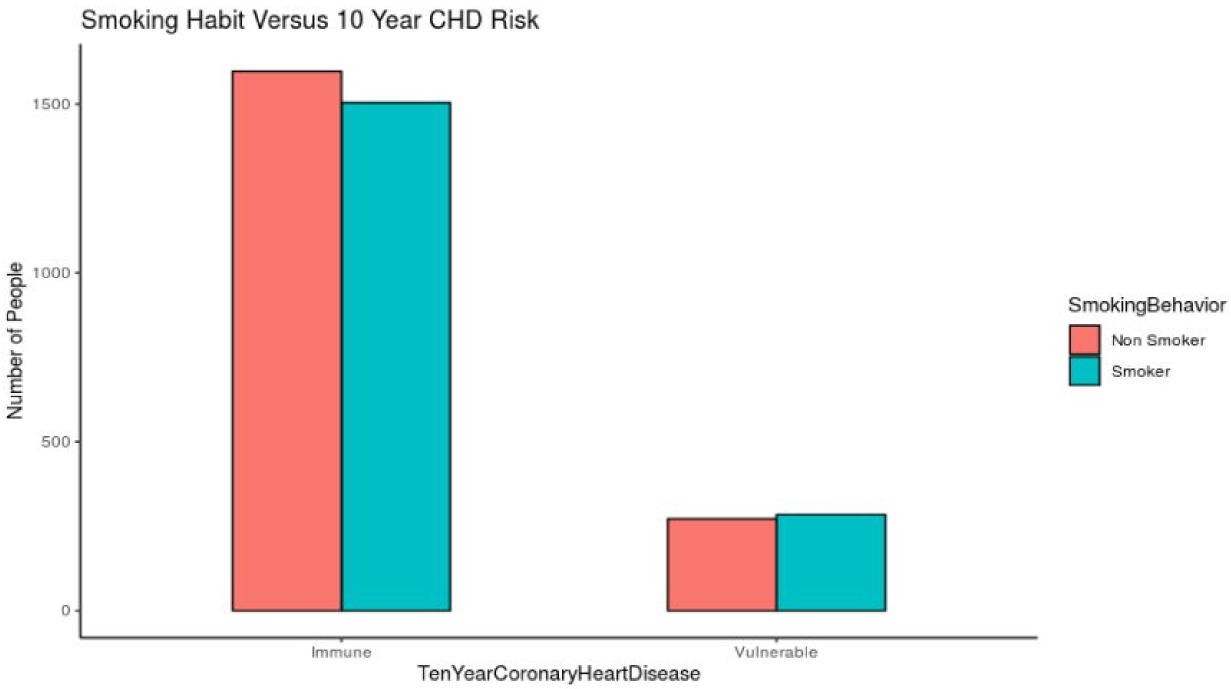


Fig. 8. Bar graph depicting the effect of Smoking Behaviour.

The above shows how strongly Smoking impacts the risk of suffering from heart disease. The smokers are more vulnerable to heart diseases than that of non smoker even if the total proportion of smokers is less in the population data. This is quite similar to that of what happened in the case of gender. It is also proven in the Framingham Study that even if a smoker quits smoking the impact will reside for the next 25 years making the smoker still risky to heart diseases.

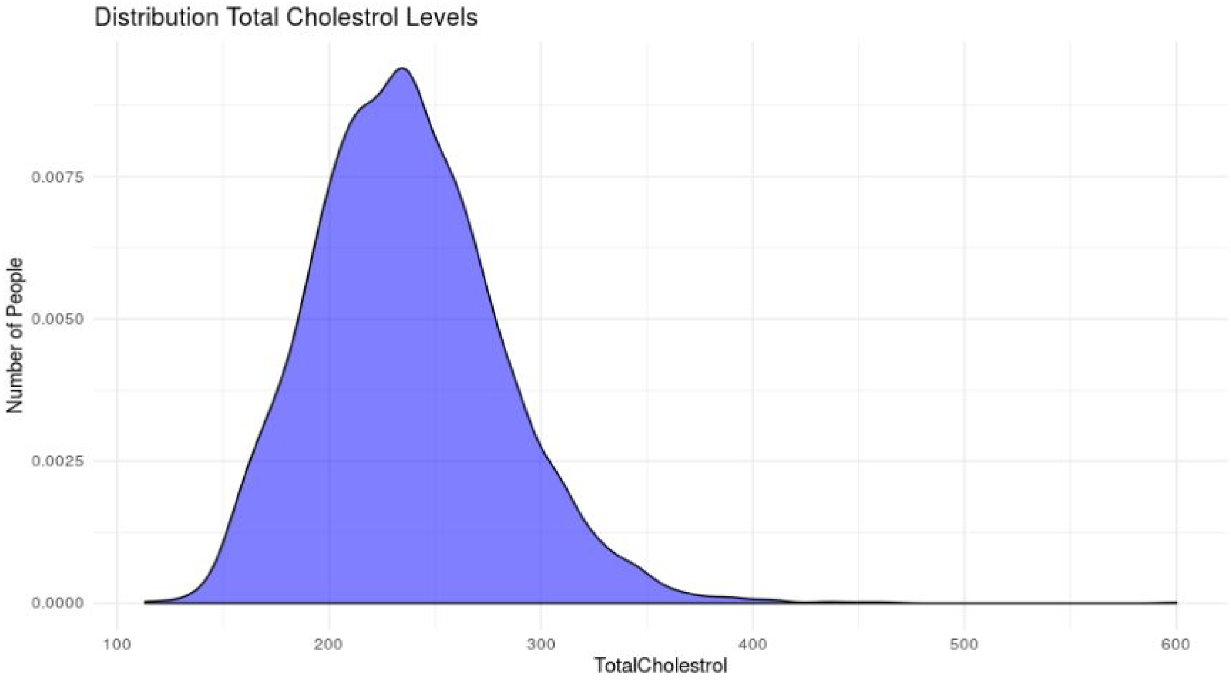


Fig. 9. Density Plot for Distribution of Total Cholesterol Levels of Population.

The amount of Cholesterol in a human body plays a critical role in heart diseases. It is said that when there is too much cholesterol in your blood, it builds up in the walls of your arteries, causing a process called atherosclerosis, a form of heart disease. The higher the total cholesterol in your body, the higher are the chances of you suffering from a heart disease. A value of 200 mg/dL is considered to be a good value, but from our distribution graph we can understand that most of the population is above this 200 mark and lying between then 200 to 300 interval.

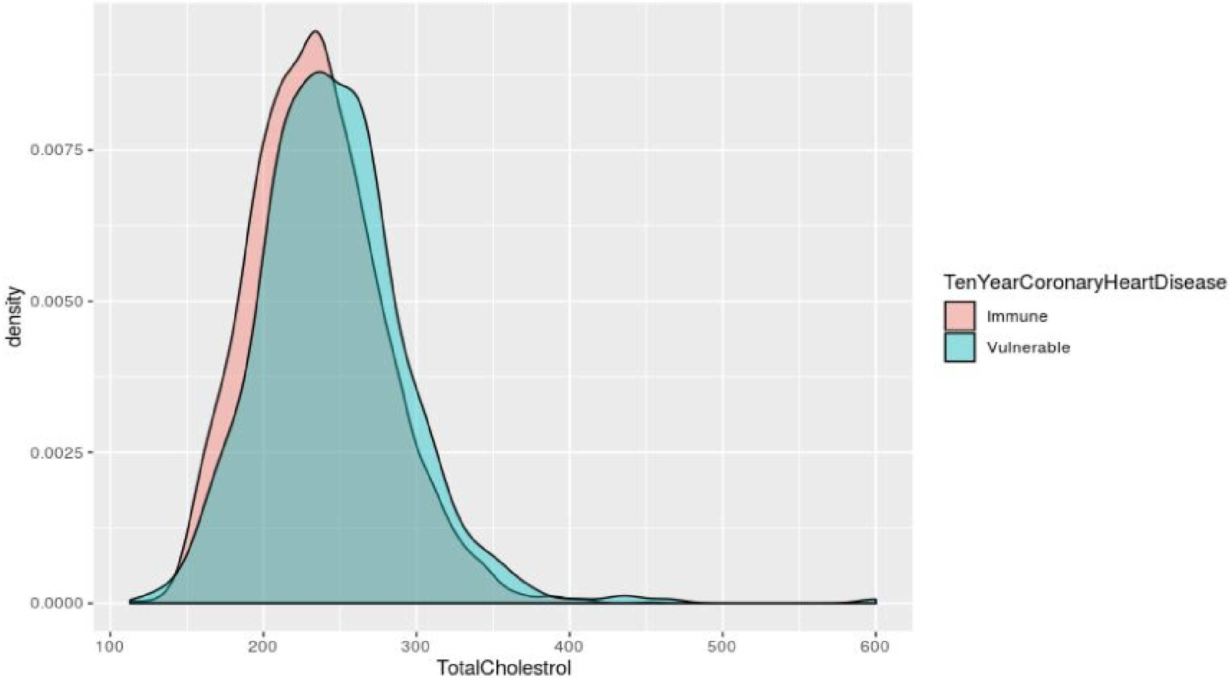


Fig. 10. Density Plot depicting the Cholesterol Levels

The above graph suggest the Total Cholesterol of people who are vulnerable to heart diseases have a higher value (skewed towards right) than that of the Total Cholesterol level of people who are immune to heart diseases. This proves the above statement we mentioned about the effect of cholesterol and its impact on heart diseases.

This graph helps to understand that there is a chance of possible difference in means between the cholesterol levels of those who are vulnerable and those who are immune to heart diseases.

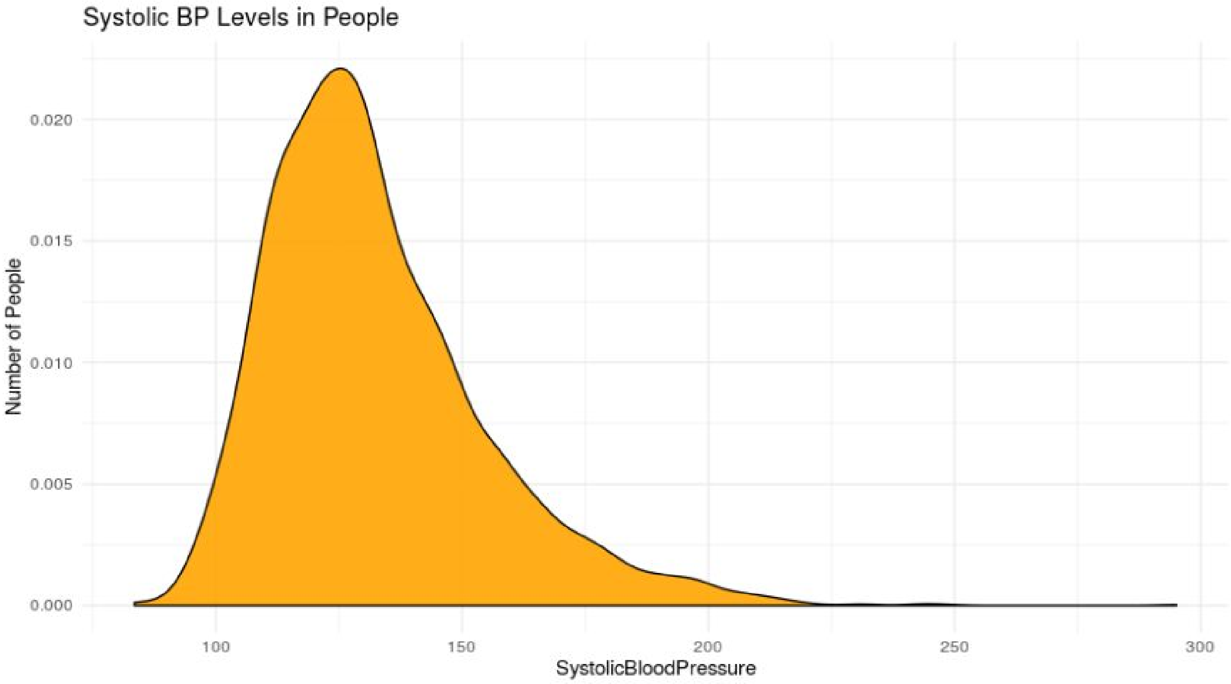


Fig. 11. Density Plot for Distribution of Systolic Blood Pressure of the Population.

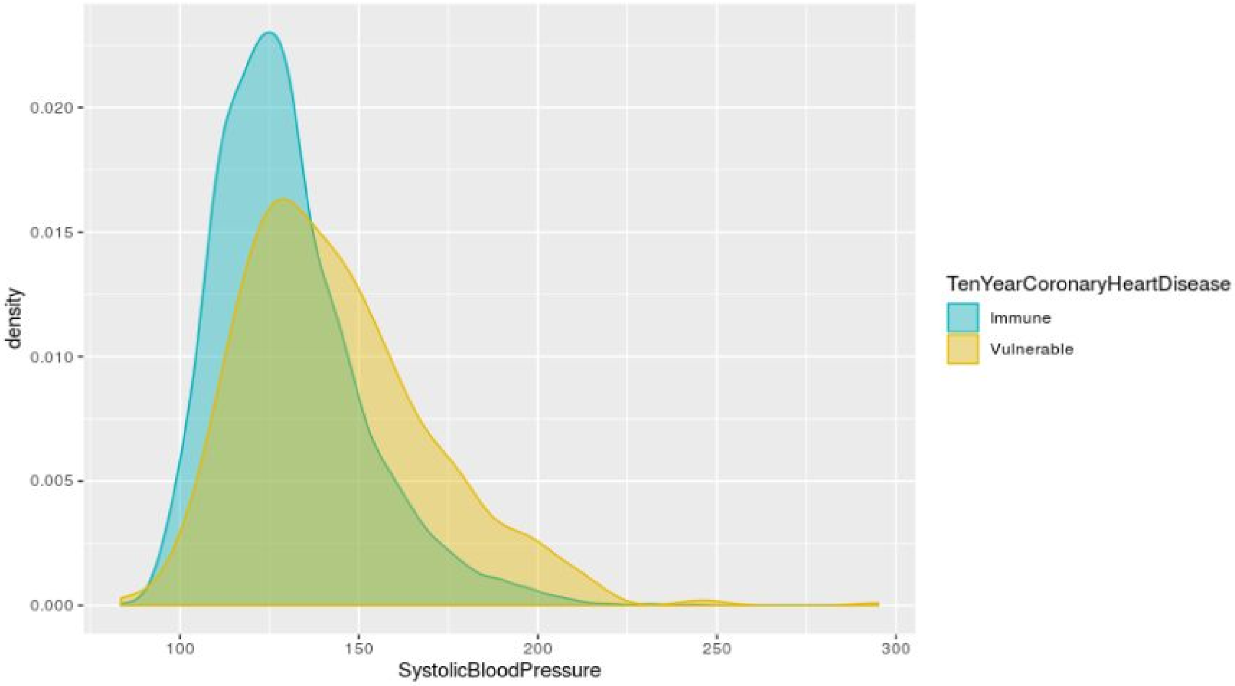


Fig. 12. Density Plot depicting the Systolic Blood Pressure Versus Ten Year CHD Risk.

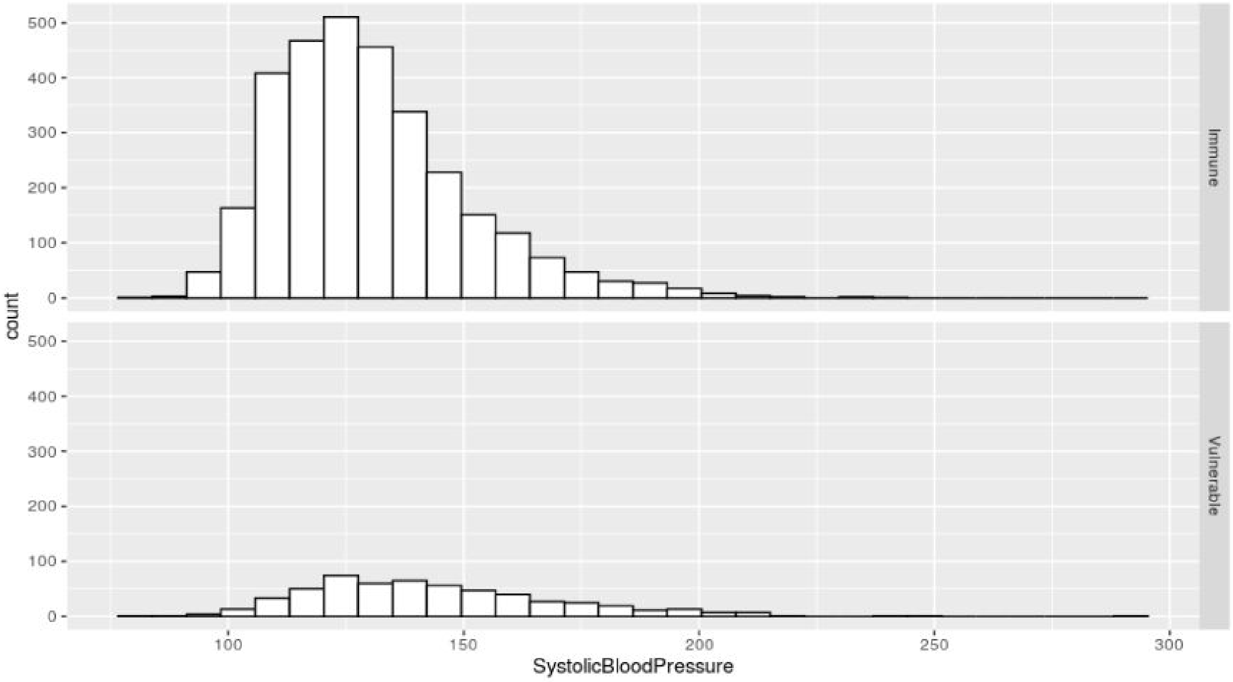


Fig. 13. Distribution of Systolic Blood Pressure Versus Vulnerable and Immune.

The above graphs show the relationship between Systolic Blood Pressure and Ten Year Coronary Heart Disease risk. Similar to that of the Total Cholesterol graph, the graph in Fig. 12 shows that the Average Systolic Blood Pressure in people Vulnerable to heart diseases than the Average Systolic Blood Pressure in people Immune to heart diseases.

In Fig. 12 we can see that Systolic Blood Pressure of people Vulnerable to CHD is skewed towards the right representing its higher mean value than that of the Systolic Blood Pressure of people Immune to CHD. This can be helpful in doing our statistical analysis, as higher levels of Systolic Blood Pressure is proven to be associated with heart diseases.

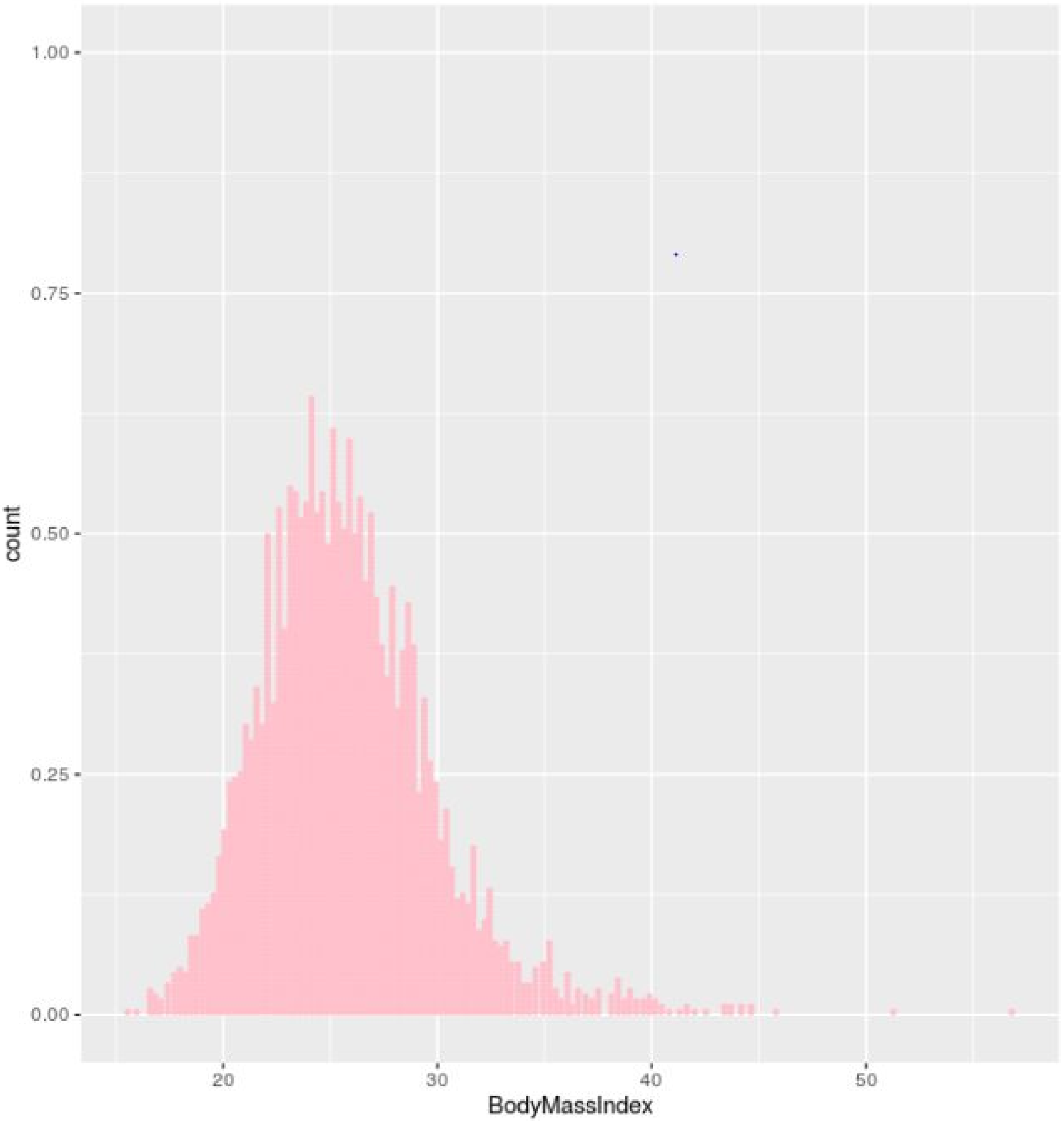


Fig. 14. Distribution of Body Mass Index of the population.

The graph represents the distribution BMI within the population data. BMI of 18 to 24 is considered perfect or fit. But in our population data most of people fall in the BMI interval between 25 and 35, which represents the fat and obese regions on the BMI chart. BMI is also proven to be a reliable factor for estimating 10 Year CHD Risk as the higher BMI is linked to higher Cholesterol level and high Blood Pressure.

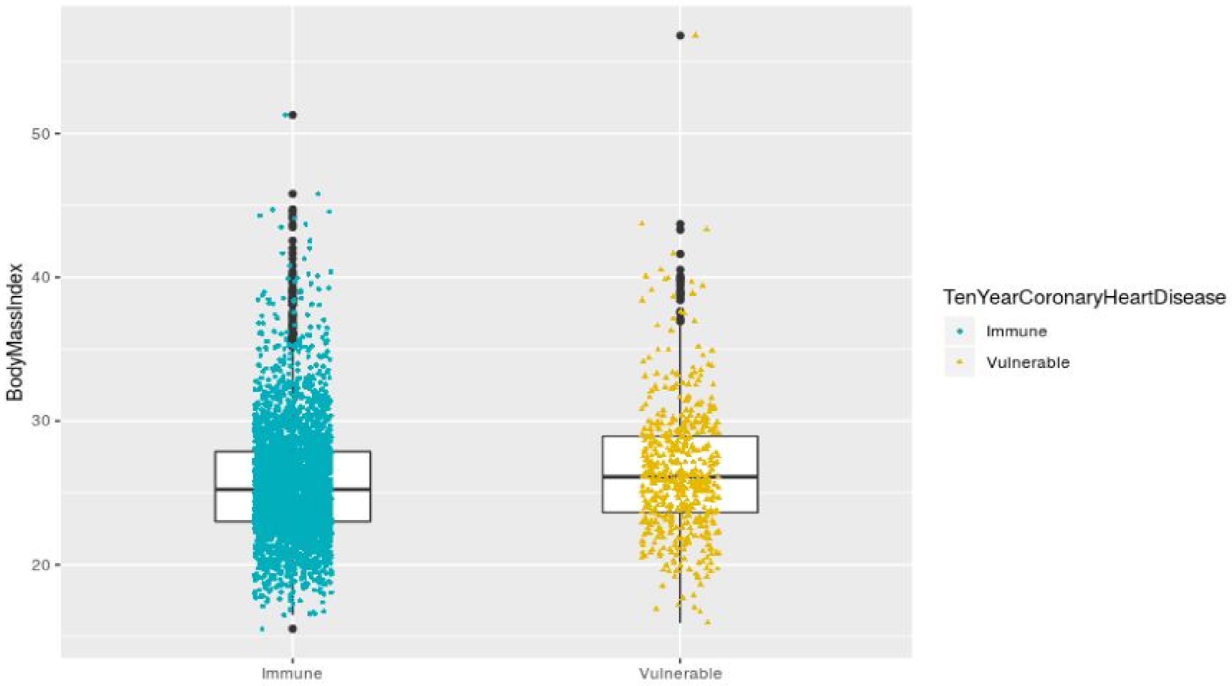


Fig. 15. Boxplot depicting the BMI Distribution in Categorical Manner.

The boxplot shows the concentrated regions where most of the population fall in their BMI levels. The blue dots represent the Immune Category and the yellow triangle represents the Vulnerable Category. We can also observe that there are a lot of outliers in both the category. The difference in the mean BMI of both categories is similar, which make this variable not of such significant role in the statistical testing phase.

With the help of Exploratory Data Analysis, we are able to grasp some of the insights lying under the data. These insights will help us choose the variables during the statistical testing phase. To further gain insights from the data lets move to the statistical analysis.

Statistical Analysis

1. One Sample t-Test

a. Traditional Statistical Tools

Hypothesis:

Null Hypothesis: The true mean Systolic Blood Pressure Level of people is 120 mmHg

$$H_0 \ : \ \mu = 120 \, mmHg$$

Alternate Hypothesis: The true mean Systolic Blood Pressure Level of people is different than 120 mmHg

$$H_A \ : \ \mu \neq 120 \, mmHg$$

Parameter of Interest:

The population parameter we want to make inference to is

$$\mu$$

Note: The chosen variables for the test satisfies all the necessary conditions and checks required for the test to give reliable outcomes.

Sample Statistic:

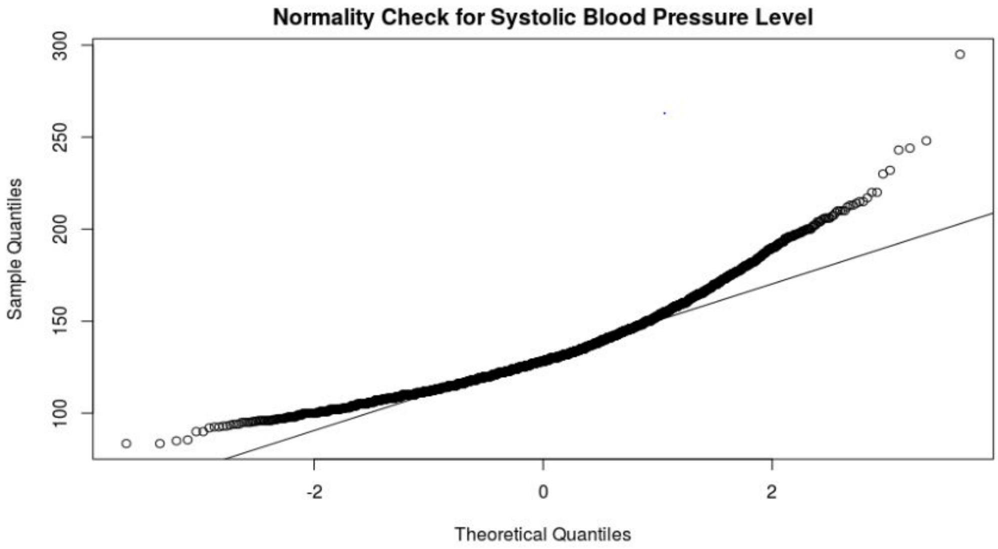
The sample statistic is the sample mean Systolic Blood Pressure

$$\overline{x}$$

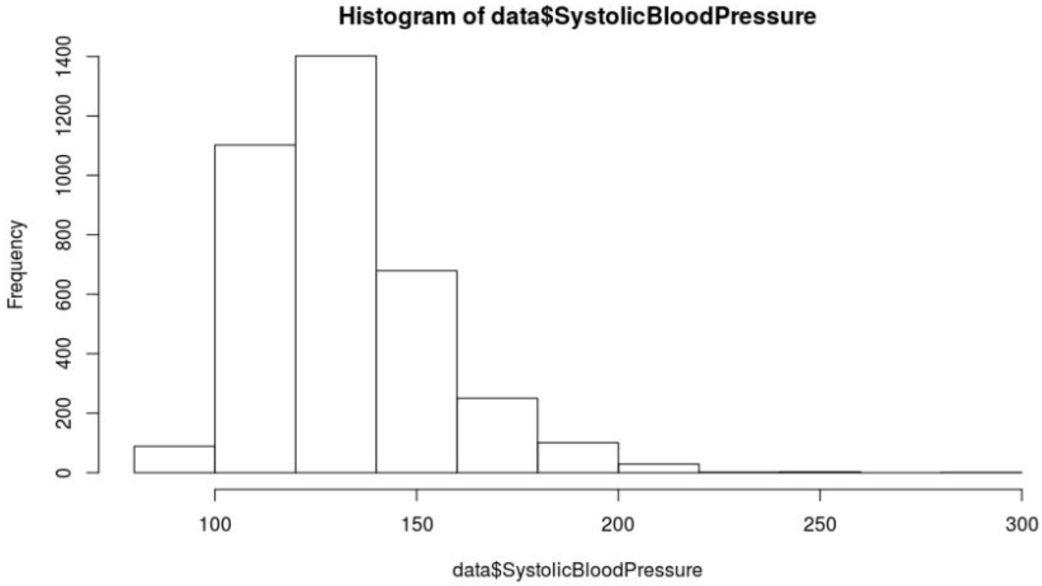
Test Statistic:

$$t_{n-1} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Normal Q-Q Plot

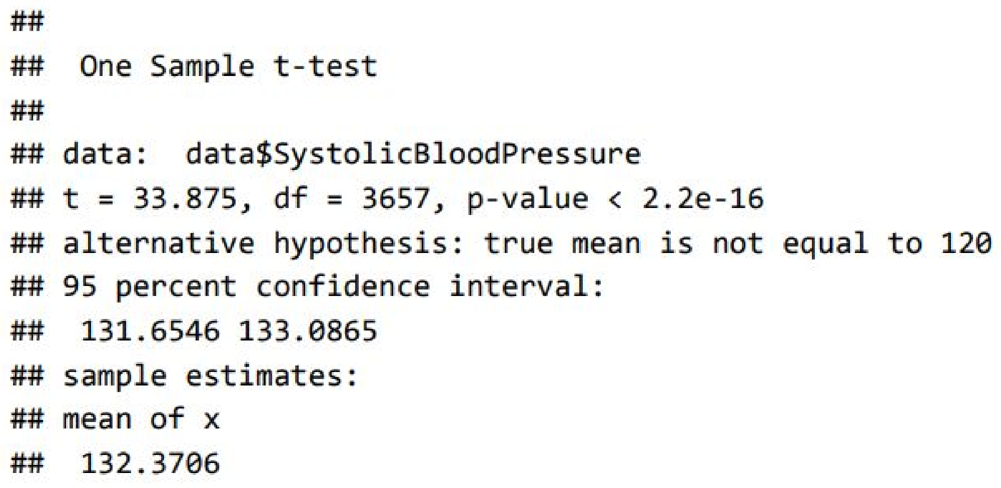


Histogram of Sample Distribution



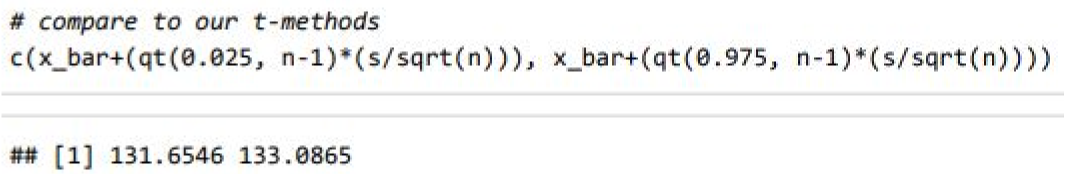
P - Value

Analysis involves comparing the P-Value with the significance levelα = 0.05 . If P-Value is < α then our null hypothesis is rejected whereas if P-value is > α there exist a weak evidence and we will not be able to reject the null hypothesis



```
##
##  One Sample t-test
##
## data:  data$SystolicBloodPressure
## t = 33.875, df = 3657, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 120
## 95 percent confidence interval:
##  131.6546 133.0865
## sample estimates:
## mean of x
##  132.3706
```

Confidence Intervals



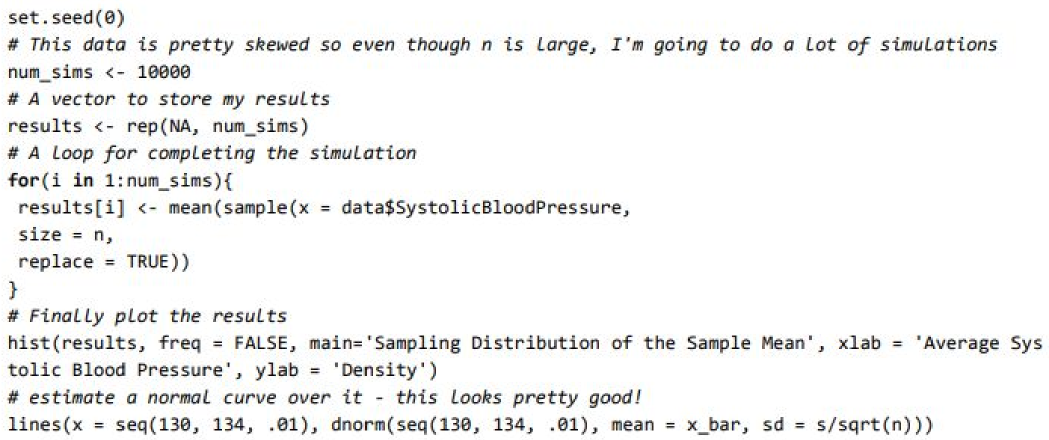
```
# compare to our t-methods
c(x_bar+(qt(0.025, n-1)*(s/sqrt(n))), x_bar+(qt(0.975, n-1)*(s/sqrt(n))))
```

```
## [1] 131.6546 133.0865
```

Interpretation

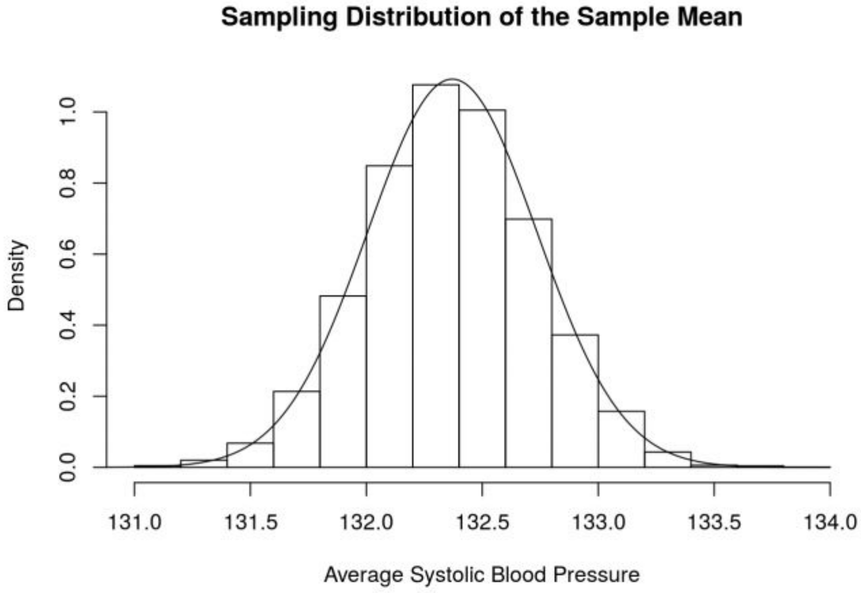
There is strong evidence (p-value = 5.135369e-219) to suggest that the true mean Systolic Blood Pressure Level of people is different from the given mean of 120 mmHg. We reject the null hypothesis that the true mean Systolic Blood Pressure Level of people is 26 minutes at the level. With 95% confidence, the true mean Systolic Blood Pressure Level is between 131.6546 mmHg and 133.0865 mmHg which suggests that the true mean commute time is greater than 120 mmHg.

b. Bootstrap Methods

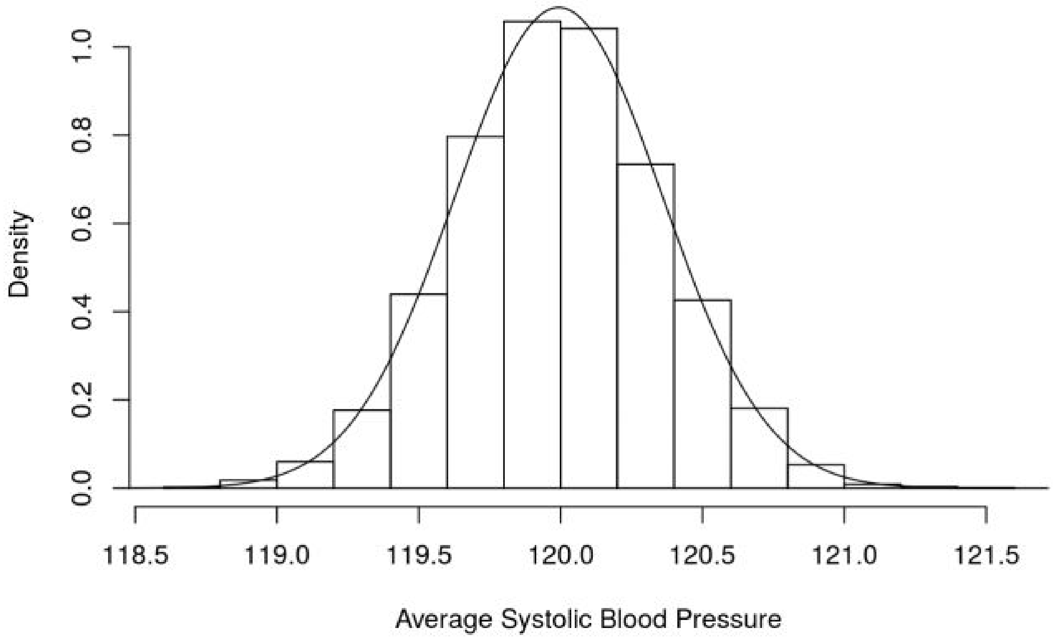


```
set.seed(0)
# This data is pretty skewed so even though n is large, I'm going to do a lot of simulations
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(sample(x = data$SystolicBloodPressure,
 size = n,
 replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Average Sys
tolic Blood Pressure', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(130, 134, .01), dnorm(seq(130, 134, .01), mean = x_bar, sd = s/sqrt(n)))
```

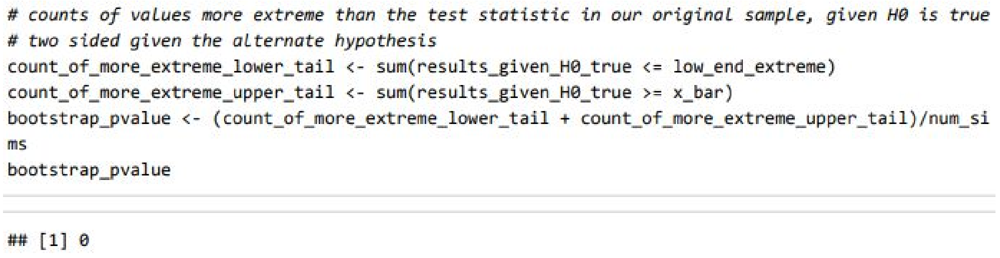
Histogram of Sampling Distribution:



Histogram of Null Hypotheses:



Bootstrap P - Value



```
# counts of values more extreme than the test statistic in our original sample, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= low_end_extreme)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= x_bar)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_si
ms
bootstrap_pvalue
```

```
## [1] 0
```

Comparison:

The interval using the empirical methods, especially the quantile method, is wider which agrees with our p-value being a bit more conservative. In this case the traditional t-tools were making our results a bit more significant and our confidence intervals narrower than what we found using the empirical method.

2. One SampleTest of Proportion

a. Traditional Statistical Tools

Hypothesis:

Null Hypothesis: The true proportion of Female in the population is 48%

$$H_0 \ : \ p_F = 0.48$$

Alternate Hypothesis: The true proportion of Female in the population is greater than 48%

$$H_A \ : \ p_R > 0.48$$

Parameter of Interest:

The population parameter we want to make inferences to is the population proportion females in the given population

$$p_F$$

Note: The chosen variables for the test satisfies all the necessary conditions and checks required for the test to give reliable outcomes.

Sample Statistic:

The sample statistic is the

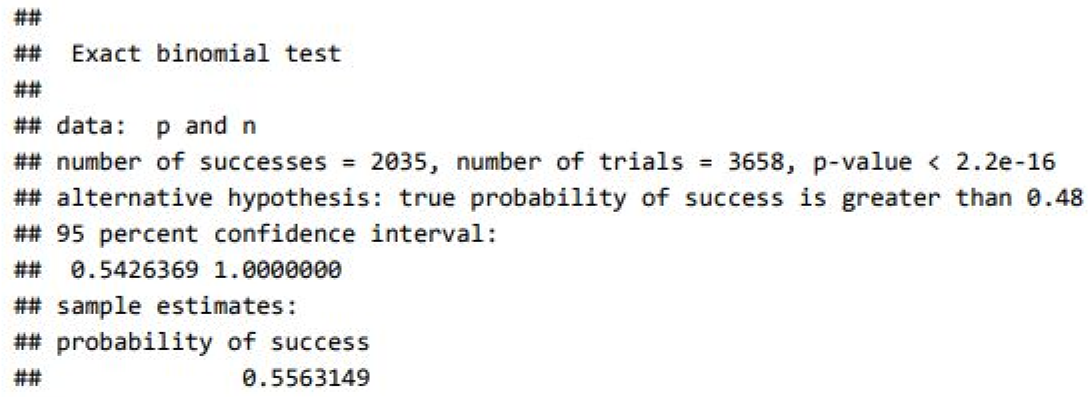
$$The\ sample\ statistic\ is\ \hat{p} = \frac{2035}{3658} = 0.5563149$$

Test Statistic:

$$z = \frac{p - p_0}{\sqrt{\frac{p_0 \times (1-p_0)}{n}}}$$

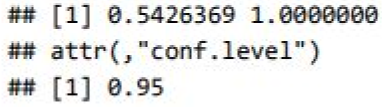
P - Value

Analysis involves comparing the P-Value with the significance levelα = 0.05 . If P-Value is < α then our null hypothesis is rejected whereas if P-value is > α there exist a weak evidence and we will not be able to reject the null hypothesis



```
##
##  Exact binomial test
##
## data:  p and n
## number of successes = 2035, number of trials = 3658, p-value < 2.2e-16
## alternative hypothesis: true probability of success is greater than 0.48
## 95 percent confidence interval:
##  0.5426369 1.0000000
## sample estimates:
## probability of success
##              0.5563149
```

Confidence Intervals

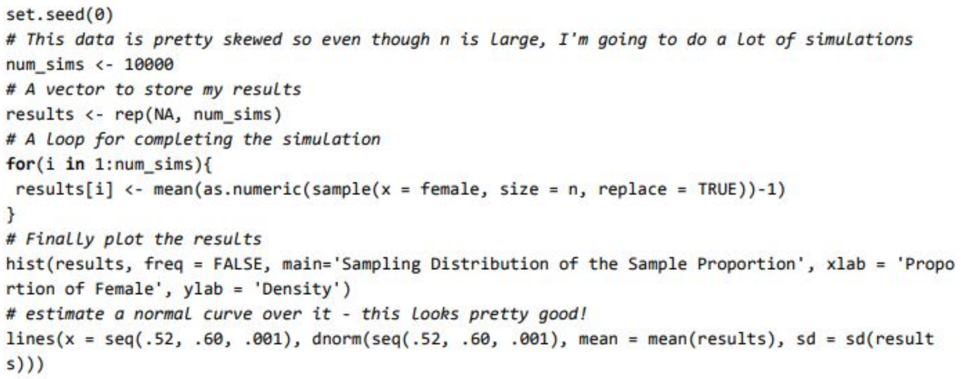


```
## [1] 0.5426369 1.0000000
## attr(,"conf.level")
## [1] 0.95
```

Interpretation

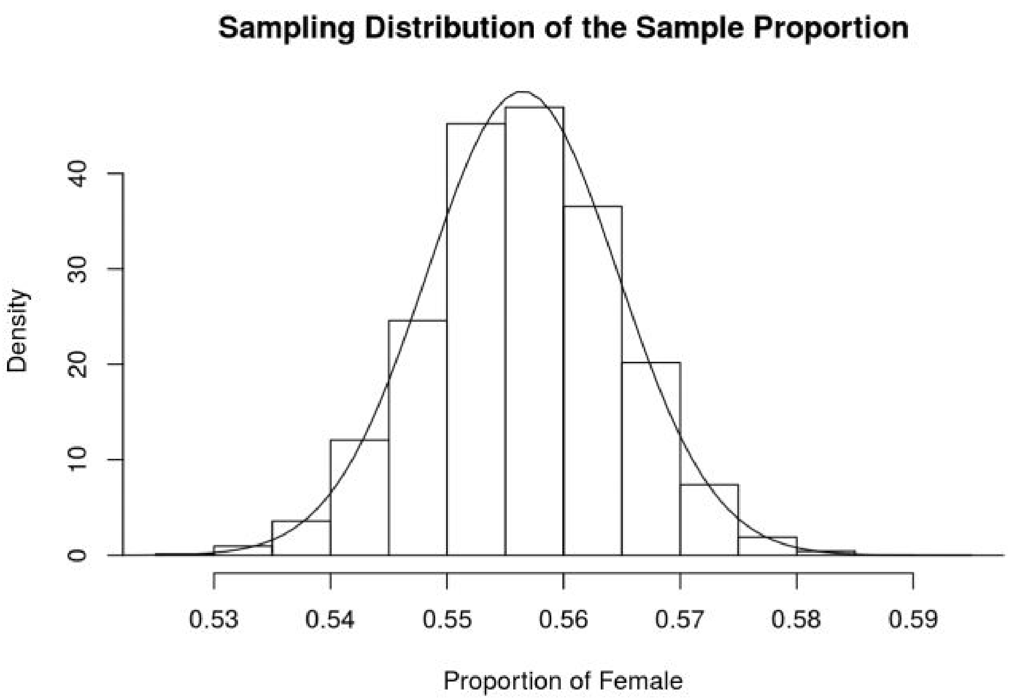
Using the exact binomial methods for a one-sample test of proportion, there is strong evidence (p-value = 5.141e-12) to suggest that the true proportion of Female in the population is greater than 48%. We can successfully reject the null hypothesis that the true proportion of male in the population is equal to 48% at the level. The true proportion of male in the population is between 0.5428433 and 1.0000000.

b. Bootstrap Methods

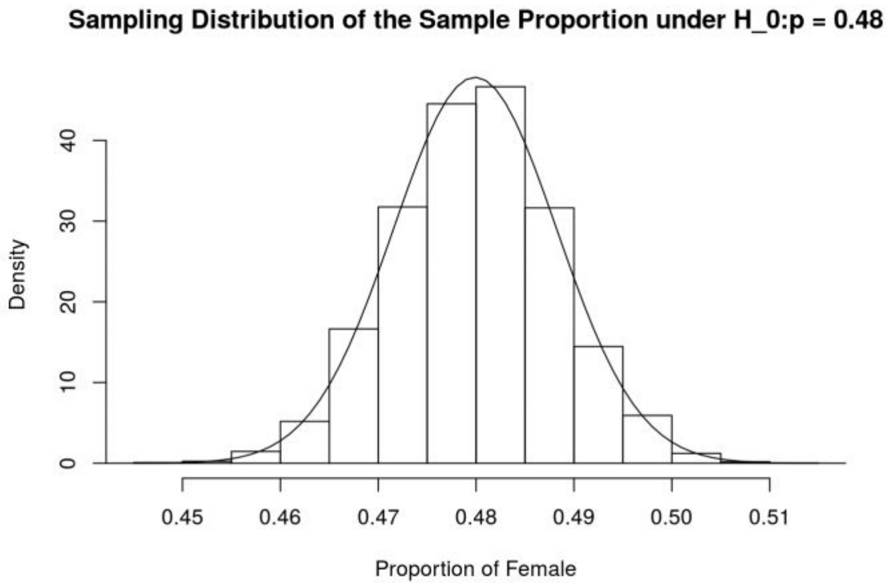


```
set.seed(0)
# This data is pretty skewed so even though n is large, I'm going to do a lot of simulations
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(as.numeric(sample(x = female, size = n, replace = TRUE))-1)
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Proportion', xlab = 'Propo
rtion of Female', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(.52, .60, .001), dnorm(seq(.52, .60, .001), mean = mean(results), sd = sd(result
s)))
```

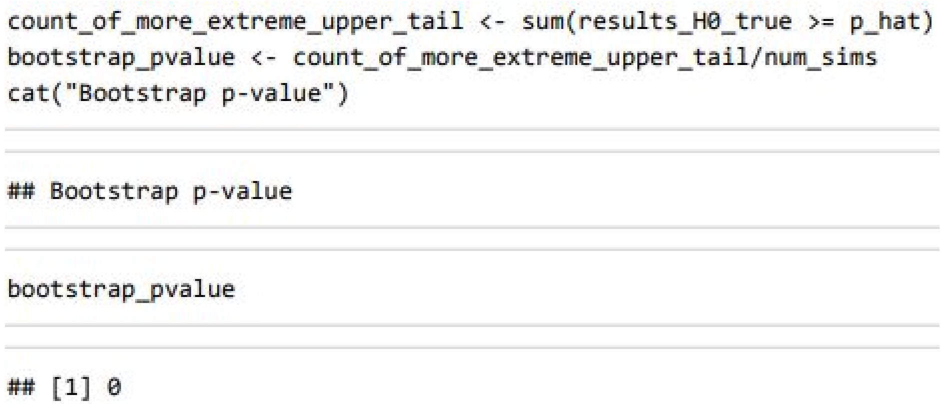
Histogram of Sampling Distribution:



Histogram of Null Hypotheses:



Bootstrap P - Value



```
count_of_more_extreme_upper_tail <- sum(results_H0_true >= p_hat)
bootstrap_pvalue <- count_of_more_extreme_upper_tail/num_sims
cat("Bootstrap p-value")
```

```
## Bootstrap p-value
```

```
bootstrap_pvalue
```

```
## [1] 0
```

Comparison:

In this case our bootstrapped p-value is closer to the exact binomial and it is more conservative than the normal approximation. Remember though that this is just one simulation of 10000 - so everytime you run this simulation you will get a different result.

3. Two sample t-Test for Difference in Means

a. Traditional Statistical Tools

Hypothesis:

Null Hypothesis: The true population mean Systolic Blood Pressure of Patients Vulnerable to Heart Disease is equal to the true population mean Systolic Blood Pressure of Patients Immune to Heart Disease.

$$H_0 : \mu_v - \mu_i = 0 \; or \; \mu_v = \mu_i$$

Alternate Hypothesis: The true population mean Systolic Blood Pressure of Patients Vulnerable to Heart Disease is not equal to the true population mean Systolic Blood Pressure of Patients Immune to Heart Disease.

$$H_{A1} : \mu_v - \mu_i \neq 0 \; or \; \mu_v \neq \mu_i$$

Parameter of Interest:

We are interested in the true population mean difference in Systolic Blood Pressure Levels between those who are Vulnerable to Heart Disease and those who are Immune to Heart Disease.

$$\overline{\mu}_v - \overline{\mu}_i$$

Note: The chosen variables for the test satisfies all the necessary conditions and checks required for the test to give reliable outcomes.

Sample Statistic:

The sample statistic is the Difference in Means

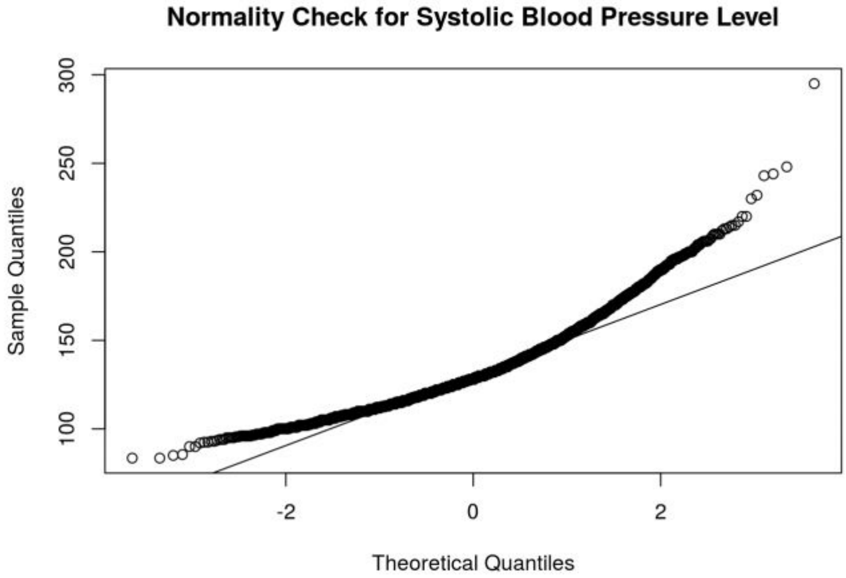
$$\overline{x}_v - \overline{x}_i$$

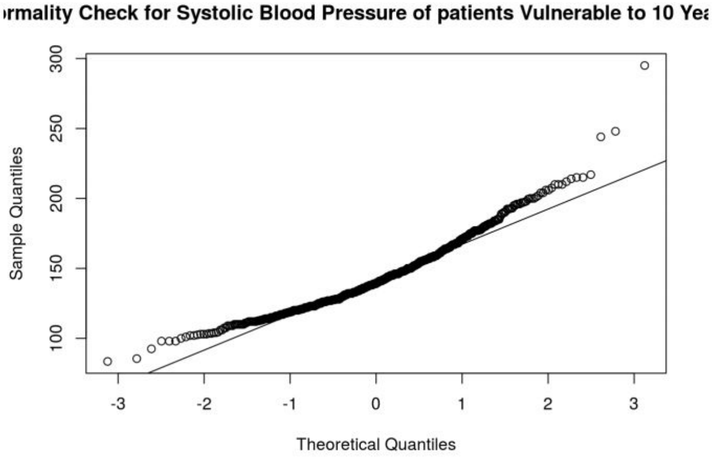
Test Statistic:

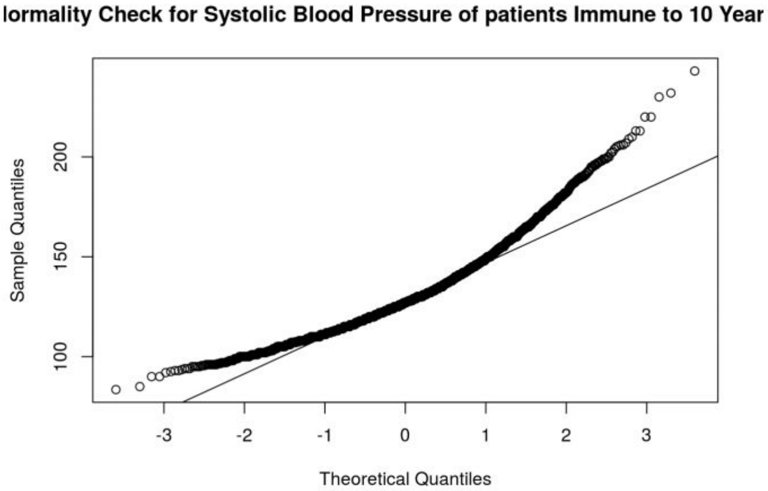
$$t = \frac{(\overline{x}_v - \overline{x}_i) - (\mu_v - \mu_i)}{\sqrt{\frac{\sigma_v^2}{n_v} + \frac{\sigma_i^2}{n_i}}}$$

$$\mu_0 = \mu_v - \mu_i = 0$$

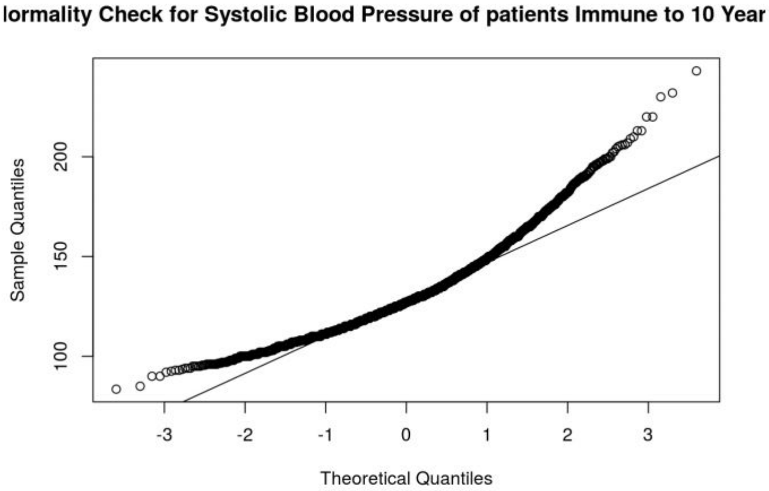
Normal Q-Q Plot of Population Data

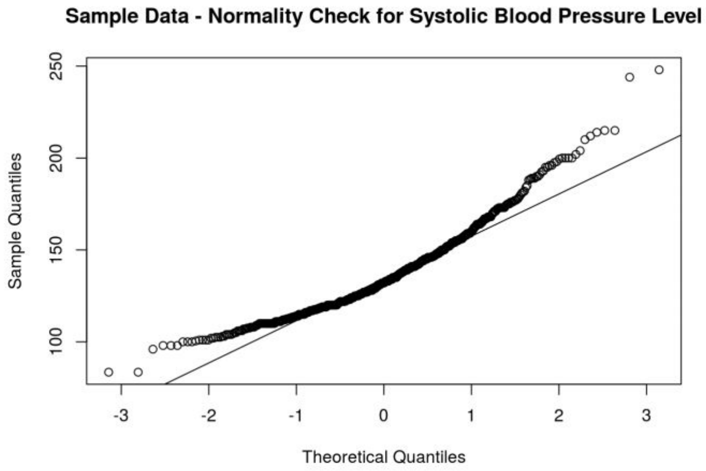


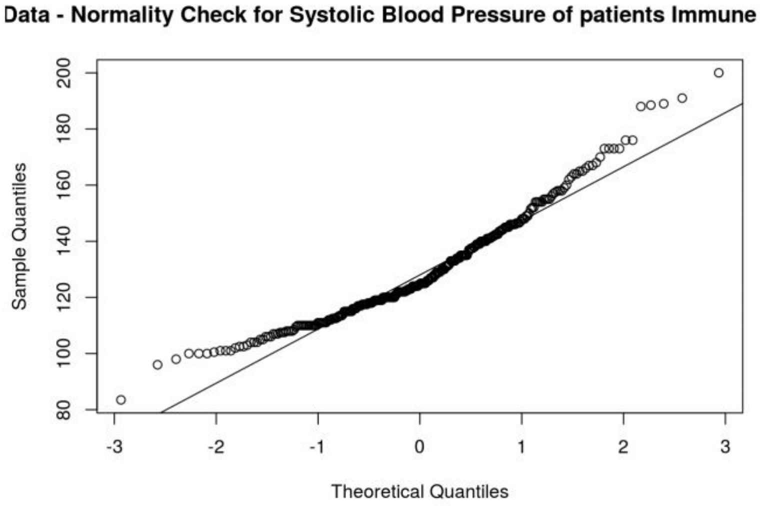




Normal Q-Q Plot of Sampled Data

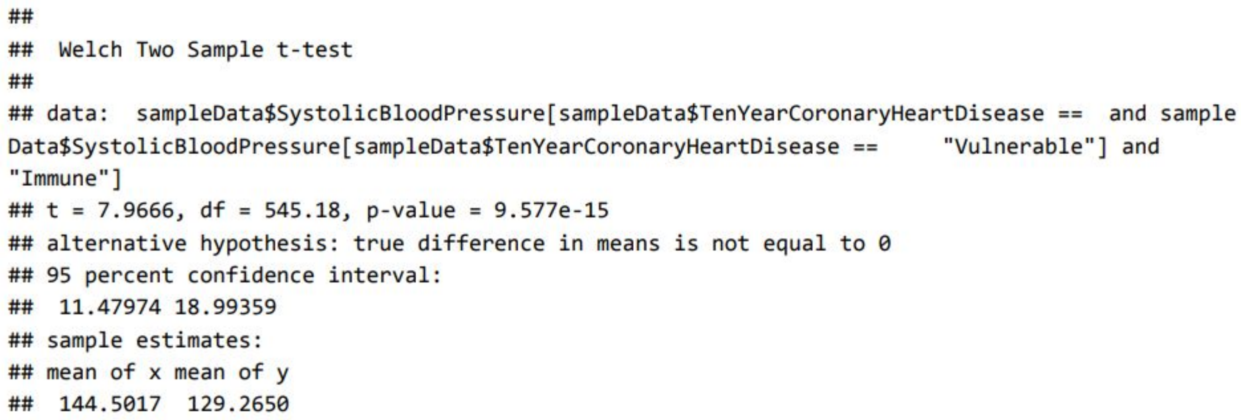






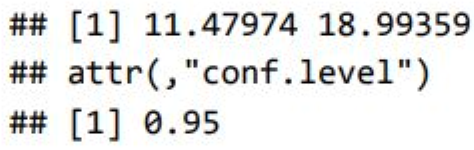
P - Value

Analysis involves comparing the P-Value with the significance levelα = 0.05 . If P-Value is < α then our null hypothesis is rejected whereas if P-value is > α there exist a weak evidence and we will not be able to reject the null hypothesis



```
##
##  Welch Two Sample t-test
##
## data:  sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease ==  and sample
Data$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease ==     "Vulnerable"] and
"Immune"]
## t = 7.9666, df = 545.18, p-value = 9.577e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.47974 18.99359
## sample estimates:
## mean of x mean of y
##  144.5017  129.2650
```

Confidence Intervals

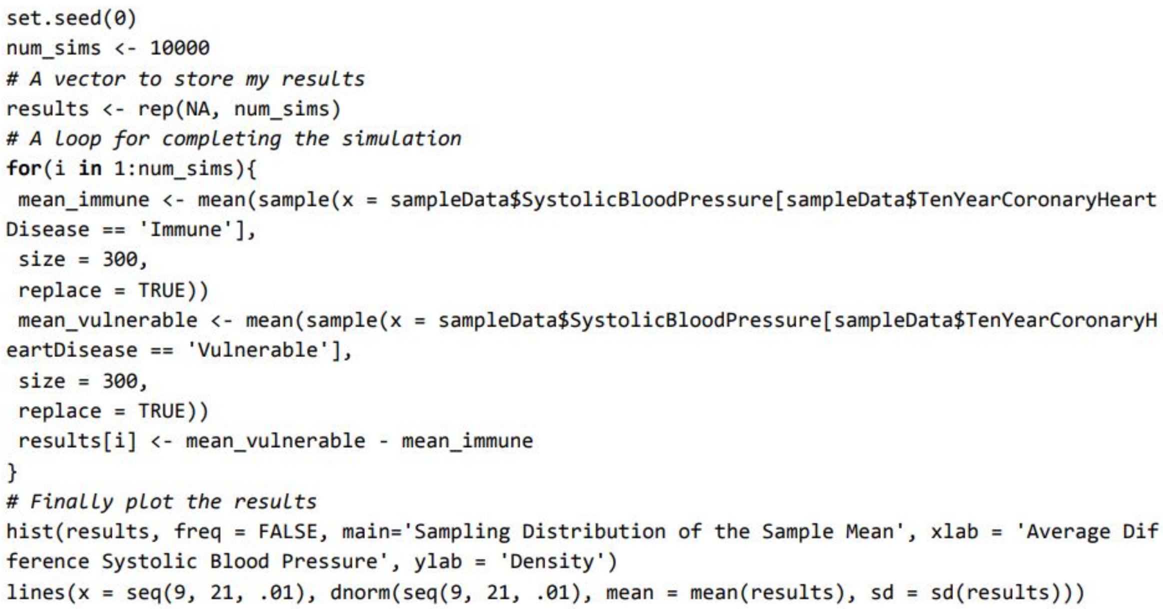


```
## [1] 11.47974 18.99359
## attr(,"conf.level")
## [1] 0.95
```

Interpretation

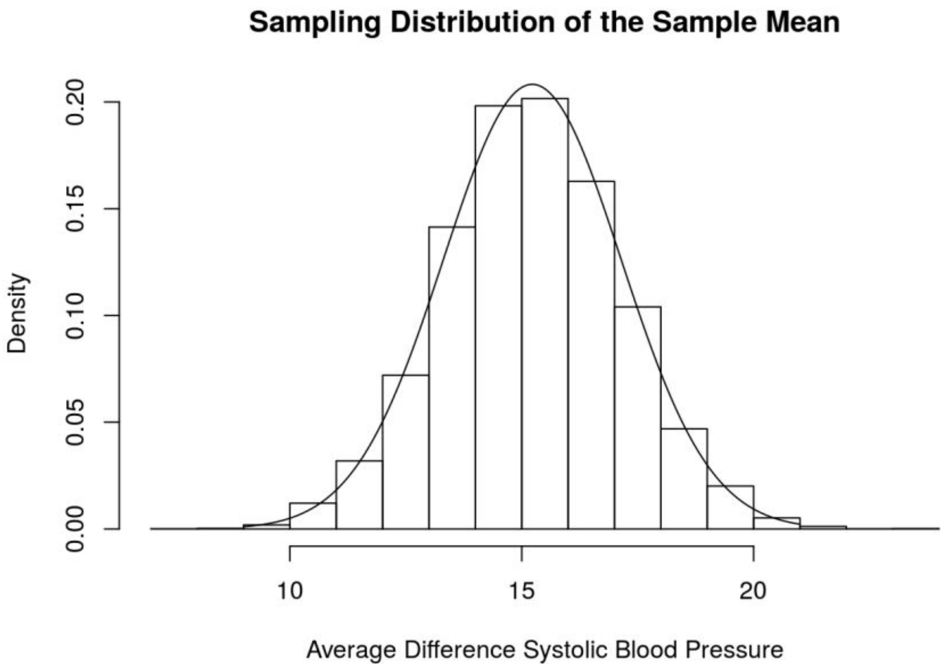
There is strong evidence (p-value=0.056) to suggest that the true population mean Systolic Blood Pressure for Patients Vulnerable to Heart Disease is different than those who are Immune to Heart Disease. We succeed to reject the null hypothesis that there is no difference between the mean Systolic Blood Pressure between the Vulnerable and Immune groups at the level. With 95% confidence, the true difference between the mean Systolic Blood Pressure between those who are Vulnerable and those who are Immune is between 11.47974 and 18.99359. The null hypothesized difference between the mean Systolic Blood Pressure is zero and zero is not in the 95% confidence interval - this result is not consistent with the results of our hypothesis test but it is possible to have this type of inconsistency when using bootstrap methods. The values of the confidence interval suggest that on average those who are Immune to Heart Disease have a lower Systolic Blood Pressure Level than those who are Vulnerable to Heart Disease.

b. Bootstrap Methods

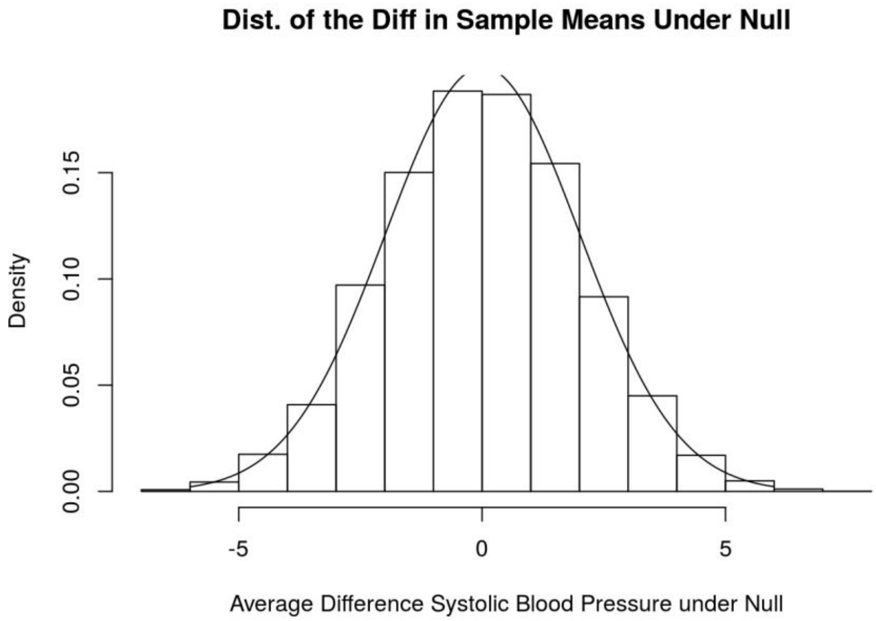


```
set.seed(0)
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 mean_immune <- mean(sample(x = sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeart
Disease == 'Immune'],
 size = 300,
 replace = TRUE))
 mean_vulnerable <- mean(sample(x = sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryH
eartDisease == 'Vulnerable'],
 size = 300,
 replace = TRUE))
 results[i] <- mean_vulnerable - mean_immune
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Average Dif
ference Systolic Blood Pressure', ylab = 'Density')
lines(x = seq(9, 21, .01), dnorm(seq(9, 21, .01), mean = mean(results), sd = sd(results)))
```

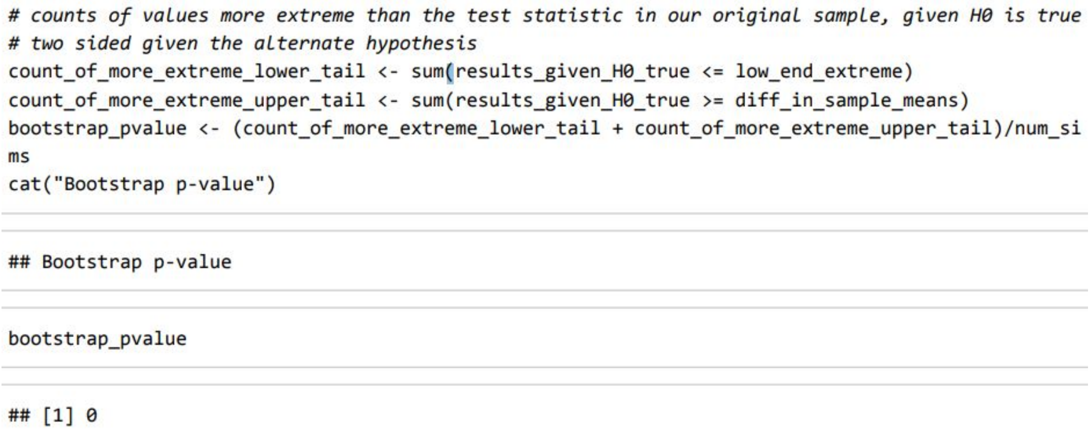
Histogram of Sampling Distribution:



Histogram of Null Hypotheses:



Bootstrap P - Value



```
# counts of values more extreme than the test statistic in our original sample, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= low_end_extreme)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= diff_in_sample_means)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_si
ms
cat("Bootstrap p-value")
```

```
## Bootstrap p-value
```

```
bootstrap_pvalue
```

```
## [1] 0
```

Comparison:

There is strong evidence (p-value=0) to suggest that the true population mean Systolic Blood Pressure for Patients Vulnerable to Heart Disease is different than those who are Immune to Heart Disease. We succeed to reject the null hypothesis that there is no difference between the mean Systolic Blood Pressure between the Vulnerable and Immune groups at the level. With 95% confidence, the true difference between the mean Systolic Blood Pressure between those who are Vulnerable and those who are Immune is between 11.47974 and 18.99359. The null hypothesized difference between the mean Systolic Blood Pressure is zero and zero is not in the 95% confidence interval - this result is not consistent with the results of our hypothesis test but it is possible to have this type of inconsistency when using bootstrap methods. The values of the confidence interval suggest that on average those who are Immune to Heart Disease have a lower Systolic Blood Pressure Level than those who are Vulnerable to Heart Disease.

4. Two sample test for Difference in Proportions

a. Traditional Statistical Tools

Hypothesis:

Null Hypothesis: There is no difference between the true population proportion of Male Patients Vulnerable to Heart Disease and the true population proportion of Female Patients Vulnerable to Heart Disease.

$$H_0 \ : \ p_M - p_F = 0$$

Alternate Hypothesis: There is a difference between the true population proportion of Male Patients Vulnerable to Heart Disease and the true population proportion of Female Patients Vulnerable to Heart Disease.

$$H_A \ : \ p_M - p_F \neq 0$$

Note: The chosen variables for the test satisfies all the necessary conditions and checks required for the test to give reliable outcomes.

Parameter of Interest:

We are interested in the difference between the true population proportion of Male who are Vulnerable to Heart Disease and true population proportion of Female who are Vulnerable to Heart Disease.

$$p_M - p_F$$

Sample Statistic:

The sample statistic is

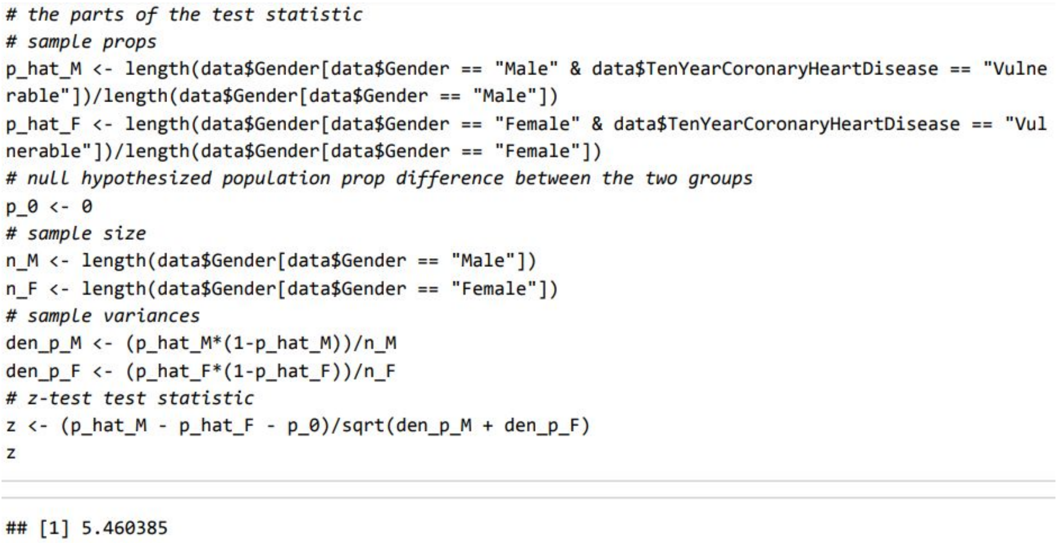
$$\hat{p_M} - \hat{p_F}$$

Test Statistic:

$$z = \frac{(\hat{p_M} - \hat{p_F}) - (p_M - p_F)}{\sqrt{\frac{\hat{p_M}(1-\hat{p_M})}{n_M} + \frac{\hat{p_F}(1-\hat{p_F})}{n_F}}}$$

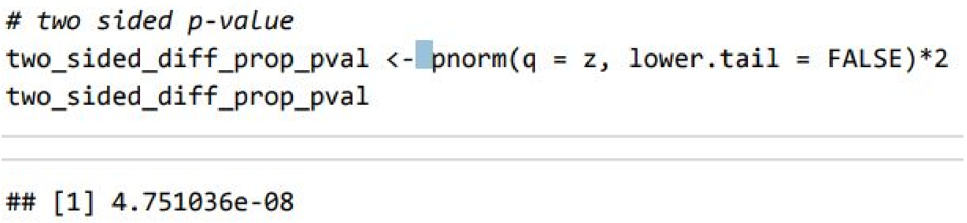
P - Value

Analysis involves comparing the P-Value with the significance levelα = 0.05 . If P-Value is < α then our null hypothesis is rejected whereas if P-value is > α there exist a weak evidence and we will not be able to reject the null hypothesis



```
# the parts of the test statistic
# sample props
p_hat_M <- length(data$Gender[data$Gender == "Male" & data$TenYearCoronaryHeartDisease == "Vulne
rable"])/length(data$Gender[data$Gender == "Male"])
p_hat_F <- length(data$Gender[data$Gender == "Female" & data$TenYearCoronaryHeartDisease == "Vul
nerable"])/length(data$Gender[data$Gender == "Female"])
# null hypothesized population prop difference between the two groups
p_0 <- 0
# sample size
n_M <- length(data$Gender[data$Gender == "Male"])
n_F <- length(data$Gender[data$Gender == "Female"])
# sample variances
den_p_M <- (p_hat_M*(1-p_hat_M))/n_M
den_p_F <- (p_hat_F*(1-p_hat_F))/n_F
# z-test test statistic
z <- (p_hat_M - p_hat_F - p_0)/sqrt(den_p_M + den_p_F)
z
```

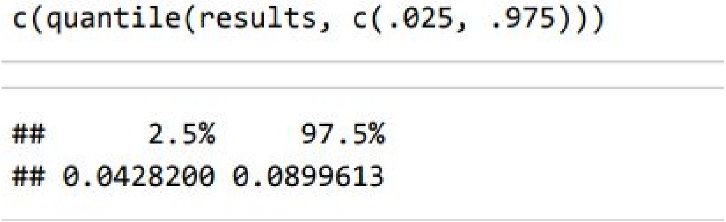
```
## [1] 5.460385
```



```
# two sided p-value
two_sided_diff_prop_pval <- pnorm(q = z, lower.tail = FALSE)*2
two_sided_diff_prop_pval
```

```
## [1] 4.751036e-08
```

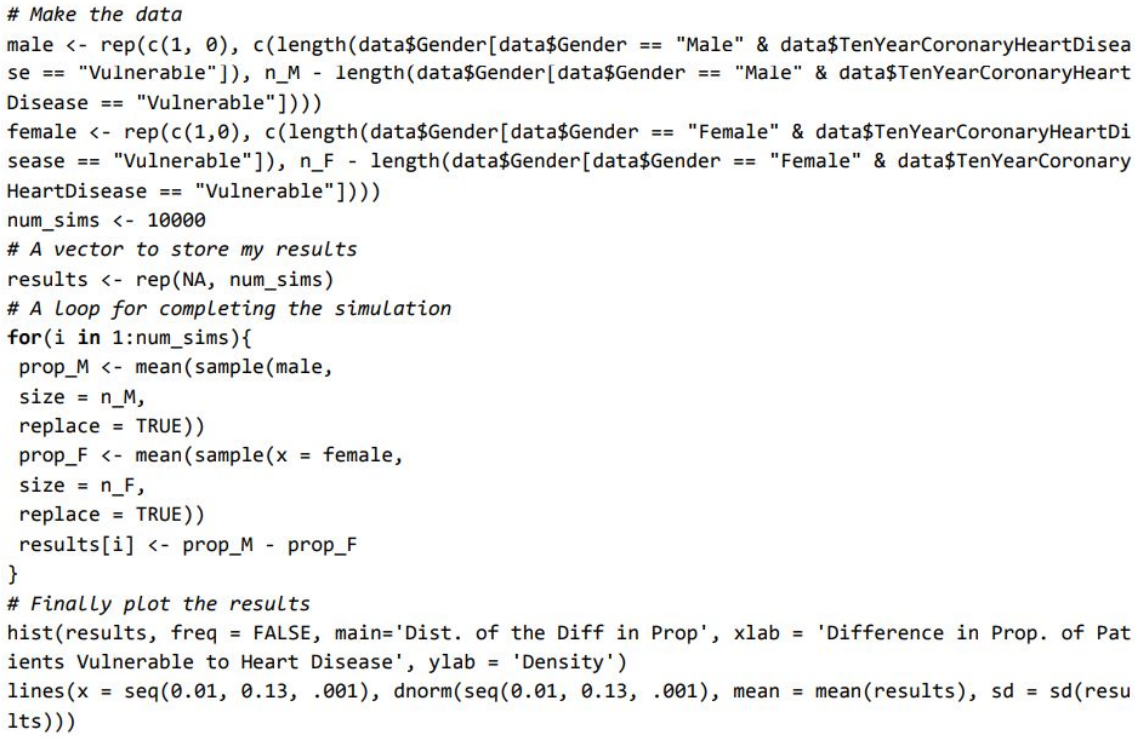
Confidence Intervals



```
c(quantile(results, c(.025, .975)))
```

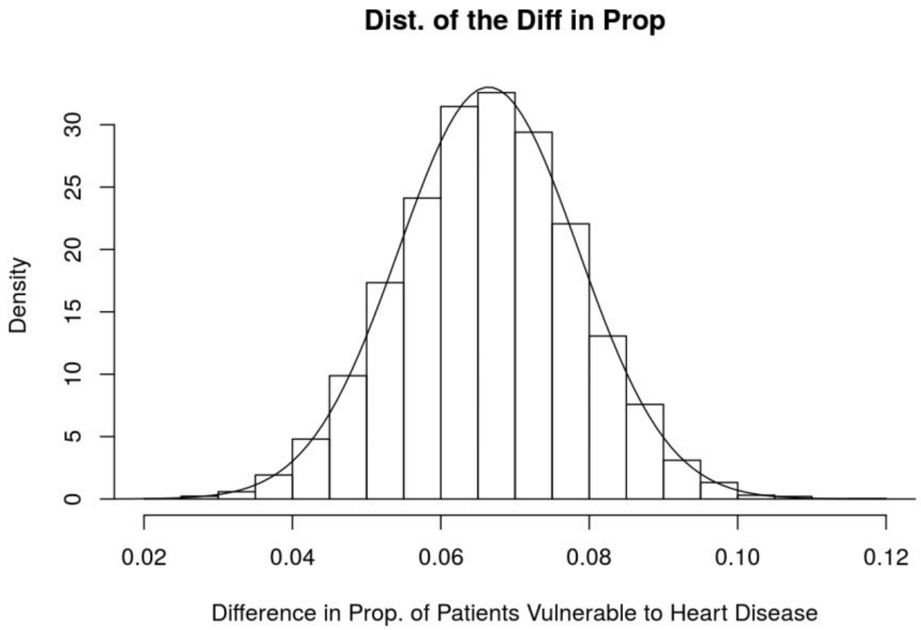
```
##      2.5%     97.5%
## 0.0428200 0.0899613
```

b. Bootstrap Methods

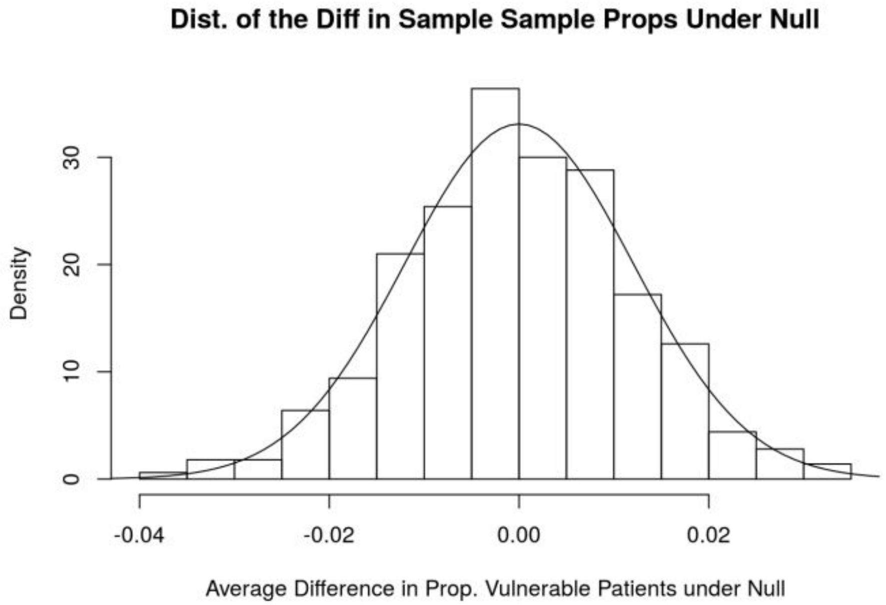


```
# Make the data
male <- rep(c(1, 0), c(length(data$Gender[data$Gender == "Male" & data$TenYearCoronaryHeartDisea
se == "Vulnerable"]), n_M - length(data$Gender[data$Gender == "Male" & data$TenYearCoronaryHeart
Disease == "Vulnerable"])))
female <- rep(c(1,0), c(length(data$Gender[data$Gender == "Female" & data$TenYearCoronaryHeartDi
sease == "Vulnerable"]), n_F - length(data$Gender[data$Gender == "Female" & data$TenYearCoronary
HeartDisease == "Vulnerable"])))
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 prop_M <- mean(sample(male,
 size = n_M,
 replace = TRUE))
 prop_F <- mean(sample(x = female,
 size = n_F,
 replace = TRUE))
 results[i] <- prop_M - prop_F
}
# Finally plot the results
hist(results, freq = FALSE, main='Dist. of the Diff in Prop', xlab = 'Difference in Prop. of Pat
ients Vulnerable to Heart Disease', ylab = 'Density')
lines(x = seq(0.01, 0.13, .001), dnorm(seq(0.01, 0.13, .001), mean = mean(results), sd = sd(resu
lts)))
```

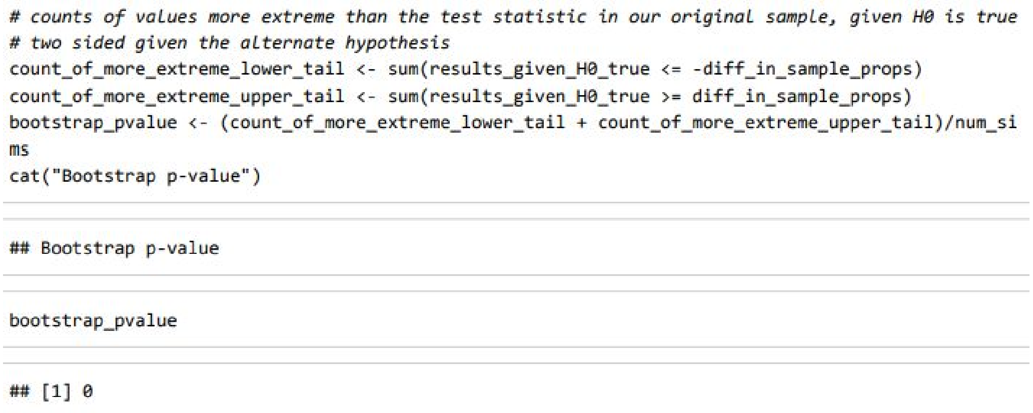
Histogram of Sampling Distribution:



Histogram of Null Hypotheses:



Bootstrap P - Value



```
# counts of values more extreme than the test statistic in our original sample, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= -diff_in_sample_props)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= diff_in_sample_props)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_si
ms
cat("Bootstrap p-value")
```

```
## Bootstrap p-value
```

```
bootstrap_pvalue
```

```
## [1] 0
```

Interpretation and Comparison:

Using randomization methods, there is strong evidence (p-value = 0) to suggest that there is a difference between the true proportion of Male Vulnerable to Heart Disease compared to their Female Counterparts. We successfully reject the null hypothesis that the true proportion of fMale Vulnerable to Heart Disease is equal to the true proportion of Female Vulnerable to Heart Disease at the level. Using confidence intervals created by the bootstrap method, we can say with 95% confidence that the true population proportion difference lies between 4.2% to 9.0% which means Male are more vulnerable to heart disease than the Female. The null hypothesized difference of 0 is outside the confidence interval which agrees with our rejection of the null hypothesis.

5. .Chi-square goodness of fit

a. Traditional Statistical Tools

Hypothesis:

The proportion of each level of education is the same and is equal to 0.25.

$$H_0 \ : \ p_C = p_{GED} = p_{HS} = p_{VS} = 0.25$$

Alternate Hypothesis: At least one of the proportions is not equal to 0.25.

$$H_A \ : \ Some \ p_i \neq 0.25$$

Parameter of Interest:

We are interested in the true proportions of people in each level of education

$$p_C, \ p_{GED}, \ p_{HS}, \ p_{VS}$$

Note: The chosen variables for the test satisfies all the necessary conditions and checks required for the test to give reliable outcomes.

Sample Statistics:

The sample statistics are

$$\hat{p_C}, \ \hat{p_{GED}}, \ \hat{p_{HS}}, \ \hat{p_{VS}}$$

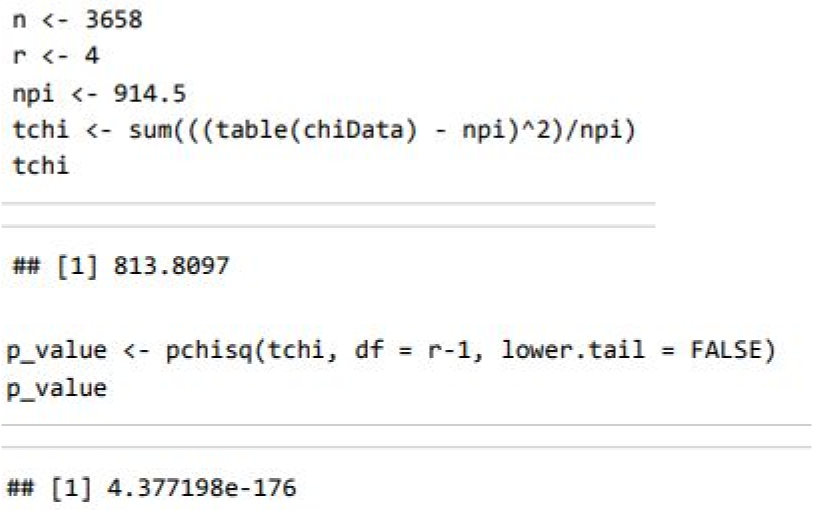
Test Statistic and Distribution:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E)^2}{E} \sim \chi^2_{k-1}$$

$$n = 3658$$
$$p_i = 0.25$$
$$expected\ count,\ np_i = 3658 \times 0.25$$
$$np_i = 914.5$$

P - Value

Analysis involves comparing the P-Value with the significance levelα = 0.05 . If P-Value is < α then our null hypothesis is rejected whereas if P-value is > α there exist a weak evidence and we will not be able to reject the null hypothesis



```
n <- 3658
r <- 4
npi <- 914.5
tchi <- sum(((table(chiData) - npi)^2)/npi)
tchi
```

```
## [1] 813.8097
```

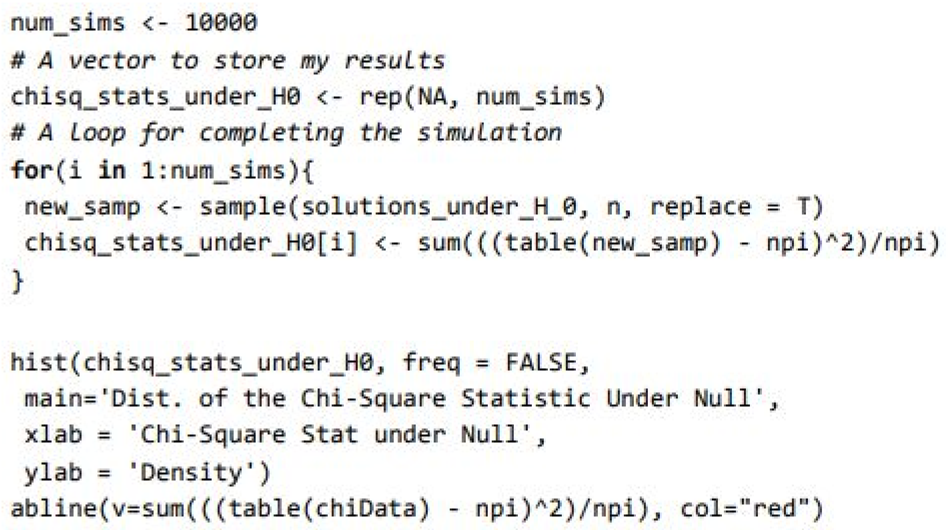
```
p_value <- pchisq(tchi, df = r-1, lower.tail = FALSE)
p_value
```

```
## [1] 4.377198e-176
```

Confidence Intervals

There is no confidence interval for a goodness of fit test.

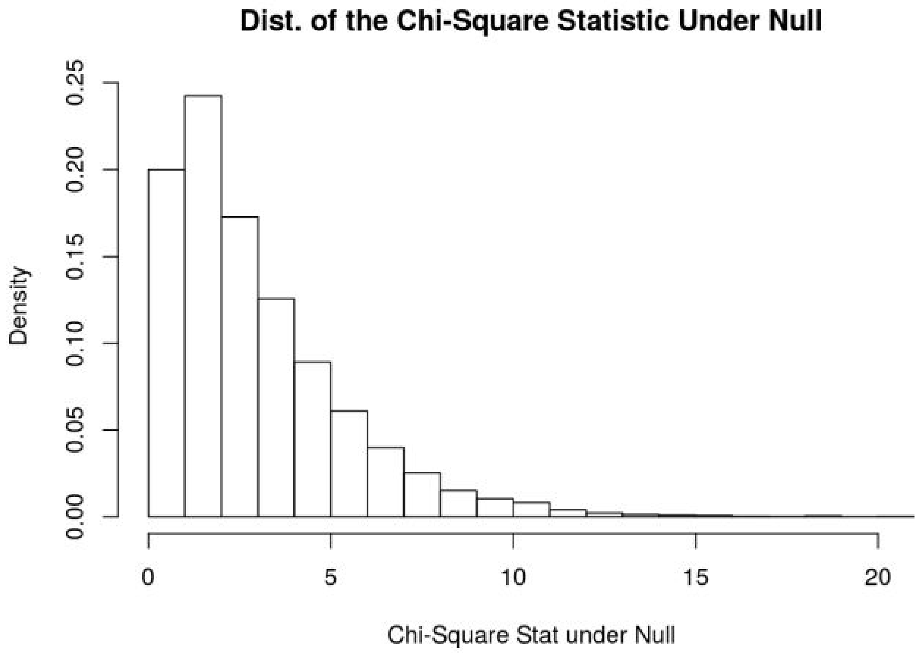
b. Randomization Approach



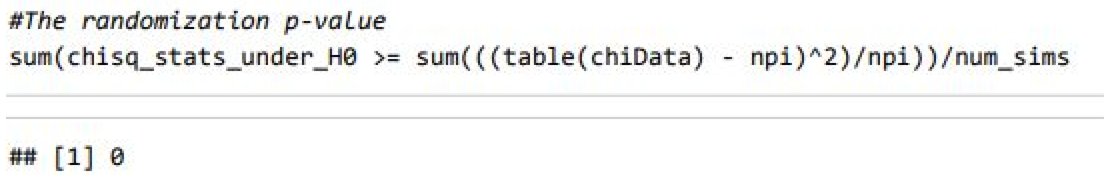
```
num_sims <- 10000
# A vector to store my results
chisq_stats_under_H0 <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 new_samp <- sample(solutions_under_H_0, n, replace = T)
 chisq_stats_under_H0[i] <- sum(((table(new_samp) - npi)^2)/npi)
}

hist(chisq_stats_under_H0, freq = FALSE,
 main='Dist. of the Chi-Square Statistic Under Null',
 xlab = 'Chi-Square Stat under Null',
 ylab = 'Density')
abline(v=sum(((table(chiData) - npi)^2)/npi), col="red")
```

Histogram of Null Hypothesis Distribution:



Chi-square P - Value



```
#The randomization p-value
sum(chisq_stats_under_H0 >= sum((table(chiData) - npi)^2)/npi))/num_sims
```

```
## [1] 0
```

Interpretation:

The data provides strong evidence that one or more of the proportions of solutions is different than .25. We successfully reject the null hypothesis that the proportions of the solutions are all equal to .25 at the level.

Discussion

I. Summary

Coronary conduit infection, likewise called coronary illness, causes about 735,000 respiratory failures every year in the U.S. What's more, murders in excess of 630,000 Americans every year. As per the American Heart Association, more than 7 million Americans have endured a respiratory failure in their lifetime. Since coronary illness is so normal and regularly is quiet until it strikes, it is essential to perceive the elements that put you in danger.

Smokers have more than double the hazard for cardiovascular failure as nonsmokers and are substantially more prone to kick the bucket on the off chance that they endure a coronary failure. Smoking is likewise the most preventable hazard factor. In the event that you smoke, quit. Even better, never start smoking by any means. Nonsmokers who are presented to consistent smoke additionally have an expanded hazard.

Improve cholesterol levels. The hazard for coronary illness increments as your aggregate sum of cholesterol increments. Your complete cholesterol objective ought to be under 200 mg/dl; HDL, the great cholesterol, higher than 40 mg/dl in men and 50 mg/dl in ladies (and the higher the better); and LDL ought to be under 130 mg/dl in solid grown-ups. For those with diabetes or different hazard factors for coronary illness, the LDL objective ought to be under 100 mg/dl (a few specialists prescribe under 70 mg/dl on the off chance that you are high hazard). Translation and treatment of cholesterol esteems must be individualized, considering the entirety of your hazard factors for coronary illness. An eating routine low in cholesterol and soaked and trans fats and high in complex starches and great fats (omega 3s) will help lower cholesterol levels and decrease your hazard for coronary illness. Customary exercise will likewise help lower "awful" cholesterol and raise "great" cholesterol. Regularly, drugs are expected to arrive at cholesterol objectives.

Control hypertension, around 67 million individuals in the U.S. have hypertension, or hypertension, making it the most widely recognized coronary illness hazard factor. About one out of three grown-ups has systolic circulatory strain (the upper number) more than 130, or potentially diastolic pulse (the lower number) more than 80, which is the meaning of hypertension. Like cholesterol, circulatory strain understanding and treatment ought to be individualized, considering your whole hazard profile. Control circulatory strain through diet, work out, weight the board, watching your salt, and if necessary, prescriptions.

Control diabetes, If not appropriately controlled, diabetes can prompt huge heart harm, including coronary episodes and demise. Control diabetes through a sound eating routine, working out, keeping up a solid weight, and accepting prescriptions as endorsed by your primary care physician.

A considerable lot of us have stationary existences, practicing rarely or not in the slightest degree. Individuals who don't practice have higher paces of death and coronary illness contrasted with individuals who perform even gentle to direct measures of physical action. Indeed, even recreation time exercises like planting or strolling can bring down your danger of coronary illness. A great many people should practice 30 minutes per day, at moderate power, on most days. Increasingly lively exercises are related with more advantages. Exercise ought to be vigorous, including the huge muscle gatherings.

II. Limitations

The Framingham Risk Score predicts just future coronary illness (CHD) occasions, nonetheless, it doesn't foresee future all out cardiovascular occasions, implying that it doesn't anticipate hazard for stroke, transient ischemic attack (TIA), and cardiovascular breakdown.

Appendix

Dataset: Framingham Heart Study Dataset

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

```r
knitr::opts_chunk$set(echo = TRUE)

#install.packages("units")

install.packages("ggplot2")

#library("units")

library("ggplot2")

dataSet <- read.csv(file = "framingham.csv")

head(dataSet)

summary(dataSet)

data <- dataSet[complete.cases(dataSet), ]

head(data)

summary(data)

names(data) <- c("Gender", "Age", "Education", "SmokingBehavior", "CigarettesPerDay",
"BloodPressureMedication", "PrevalentStroke", "PrevalentHypertension", "DiabeticCondition",
"TotalCholesterol", "SystolicBloodPressure", "DiastolicBloodPressure", "BodyMassIndex",
"HeartRate", "GlucoseLevel", "TenYearCoronaryHeartDisease")

head(data)

data$Gender[data$Gender == 0] <- "Female"

data$Gender[data$Gender == 1] <- "Male"

data$Education[data$Education == 1] <- "High School"

data$Education[data$Education == 2] <- "General Education Development"

data$Education[data$Education == 3] <- "Vocational School"

data$Education[data$Education == 4] <- "College"

data$SmokingBehavior[data$SmokingBehavior == 0] <- "Non Smoker"

data$SmokingBehavior[data$SmokingBehavior == 1] <- "Smoker"

data$BloodPressureMedication[data$BloodPressureMedication == 0] <- "Not Under BP Medication"

data$BloodPressureMedication[data$BloodPressureMedication == 1] <- "Under BP Medication"

data$PrevalentStroke[data$PrevalentStroke == 0] <- "No"

data$PrevalentStroke[data$PrevalentStroke == 1] <- "Yes"
```

```
data$PrevalentHypertension[data$PrevalentHypertension == 0] <- "No"
data$PrevalentHypertension[data$PrevalentHypertension == 1] <- "Yes"
data$DiabeticCondition[data$DiabeticCondition == 0] <- "Non Diabetic"
data$DiabeticCondition[data$DiabeticCondition == 1] <- "Diabetic"
data$TenYearCoronaryHeartDisease[data$TenYearCoronaryHeartDisease == 0] <- "Immune"
data$TenYearCoronaryHeartDisease[data$TenYearCoronaryHeartDisease == 1] <- "Vulnerable"
data$Gender <- as.factor(data$Gender)
data$Education <- as.factor(data$Education)
data$SmokingBehavior <- as.factor(data$SmokingBehavior)
data$BloodPressureMedication <- as.factor(data$BloodPressureMedication)
data$PrevalentStroke <- as.factor(data$PrevalentStroke)
data$PrevalentHypertension <- as.factor(data$PrevalentHypertension)
data$DiabeticCondition <- as.factor(data$DiabeticCondition)
data$TenYearCoronaryHeartDisease <- as.factor(data$TenYearCoronaryHeartDisease)
head(data)
#units(data$Age) <- "years"
#units(data$TotalCholesterol) <- "mg/dL"
#units(data$SystolicBloodPressure) <- "mmHg"
#units(data$DiastolicBloodPressure) <- "mmHg"
#units(data$BodyMassIndex) <- "kg/m^2"
#units(data$GlucoseLevel) <- "mg/dL"
#head(data)
summary(data)
ggplot(data, aes(x = Age)) +
 geom_histogram(bins = 30, fill = "lightblue") +
 theme_bw() + theme_classic() +
 ggtitle("Age Distribution") + ylab("Number of People")
ggplot(data, aes(x = Education, fill = Education)) +
 geom_bar() + ggtitle("Received Level of Education") +
 geom_text(stat = 'count', aes(label =..count..), vjust = -0.5)
ggplot(data, aes(x = Gender, fill = Gender)) +
 geom_bar() +
 geom_text(stat = 'count', aes(label =..count..), vjust = -0.5) +
 theme_bw() + theme_classic() +
 ggtitle("Gender Distribution") + ylab("Number of People")
ggplot(data, aes(x = TenYearCoronaryHeartDisease)) +
 geom_bar(aes(fill = Gender), position = 'dodge', width = 0.5, color='black') +
 theme_bw() + theme_classic() +
 ylab("Number of People") + ggtitle("10 Year CHD Risk Versus Gender")
ggplot(data, aes(x = CigarettesPerDay)) +
 geom_histogram(bins = 30, fill = "gray") +
 theme_bw() + theme_classic() +
 ggtitle("Smokers - Cigarettes Per Day Distribution") + ylab("Number of People")
ggplot(data, aes(x = SmokingBehavior)) +
 geom_bar(fill = "lightgreen") +
 geom_text(stat = 'count', aes(label =..count..), vjust = -0.5) +
 theme_bw() + theme_classic() +
 ggtitle("Smoking Behaviour") + ylab("Number of People")
ggplot(data, aes(x = TenYearCoronaryHeartDisease)) +
 geom_bar(aes(fill = SmokingBehavior), position = 'dodge', width = 0.5, color= 'black') +
 theme_bw() + theme_classic() +
 ylab("Number of People") + ggtitle("Smoking Habit Versus 10 Year CHD Risk")
ggplot(data, aes(x = TotalCholesterol)) +
 geom_density(fill = "blue", alpha = 0.5) +
 theme_minimal() +
 ggtitle("Distribution Total Cholesterol Levels") + ylab("Number of People")
ggplot(data, aes(x = TotalCholesterol)) +
 geom_density(aes(fill = TenYearCoronaryHeartDisease), alpha = 0.4)
ggplot(data, aes(x = SystolicBloodPressure)) +
 geom_density(fill ="orange", alpha = 0.9) +
 theme_minimal() +
 ggtitle("Systolic BP Levels in People") + ylab("Number of People")
ggplot(data, aes(x = SystolicBloodPressure)) +
 geom_density(aes(color = TenYearCoronaryHeartDisease, fill = TenYearCoronaryHeartDisease),
alpha = 0.4, position = "identity") +
 scale_fill_manual(values = c("#00AFBB", "#E7B800")) +
 scale_color_manual(values = c("#00AFBB", "#E7B800"))
# Systolic Blood Pressure vs Ten Year Coronary Heart Disease
ggplot(data, aes(x = SystolicBloodPressure))+
 geom_histogram(bins = 30, color="black", fill="white")+
 facet_grid(TenYearCoronaryHeartDisease ~ .)
ggplot(data, aes(x =BodyMassIndex)) +
 geom_dotplot(color = "pink", fill = "pink", binwidth = 1/4)
ggplot(data, aes(x = TenYearCoronaryHeartDisease, y = BodyMassIndex)) +
 geom_boxplot(width = 0.4, fill = "white") +
 geom_jitter(aes(color = TenYearCoronaryHeartDisease, shape = TenYearCoronaryHeartDisease),
       width = 0.1, size = 1) +
 scale_color_manual(values = c("#00AFBB", "#E7B800")) +
 labs(x = NULL)
ggplot(data, aes(x = TenYearCoronaryHeartDisease)) +
 geom_bar() +
 geom_text(stat = 'count', aes(label =..count..), vjust = -0.5) +
 theme_bw() + theme_classic() +
 ggtitle("10 Year CHD Risk") + ylab("Number of People")
qqnorm(data$SystolicBloodPressure, main ="Normality Check for Systolic Blood Pressure Level")
qqline(data$SystolicBloodPressure)
hist(data$SystolicBloodPressure)
# the parts of the test statistic
# sample mean
x_bar <- mean(data$SystolicBloodPressure)
# null hypothesized population mean
mu_0 <- 120
# sample st. dev
s <- sd(data$SystolicBloodPressure)
# sample size
```

```r
n <- length(data$SystolicBloodPressure)

# t-test test statistic

t <- (x_bar - mu_0)/(s/sqrt(n))

t

# two-sided p-value so multiply by 2

two_sided_t_pval <- pt(q = t, df = n-1, lower.tail = FALSE)*2

two_sided_t_pval

qt(0.025, n-1)

# lower bound

x_bar + (qt(0.025, n-1)*(s/sqrt(n))) # alternately you can use x_bar-(qt(0.975, n-1)*(s/sqrt(n)))

# upper bound

x_bar + (qt(0.975, n-1)*(s/sqrt(n))) # alternately you can use x_bar-(qt(0.025, n-1)*(s/sqrt(n)))

t.test(data$SystolicBloodPressure, alternative = "two.sided", mu = 120)

set.seed(0)

# This data is pretty skewed so even though n is large, I'm going to do a lot of simulations

num_sims <- 10000

# A vector to store my results

results <- rep(NA, num_sims)

# A loop for completing the simulation

for(i in 1:num_sims){

 results[i] <- mean(sample(x = data$SystolicBloodPressure,

 size = n,

 replace = TRUE))

}

# Finally plot the results

hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Average Systolic Blood Pressure', ylab = 'Density')

# estimate a normal curve over it - this looks pretty good!

lines(x = seq(130, 134, .01), dnorm(seq(130, 134, .01), mean = x_bar, sd = s/sqrt(n)))

set.seed(0)

# Shift the sample so that the null hypothesis is true

bp_given_H0_true <- data$SystolicBloodPressure - mean(data$SystolicBloodPressure) + mu_0
```

```
# This data is pretty skewed so even though n is large, I'm going to do a lot of simulations

num_sims <- 10000

# A vector to store my results

results_given_H0_true <- rep(NA, num_sims)

# A loop for completing the simulation

for(i in 1:num_sims){

 results_given_H0_true[i] <- mean(sample(x = bp_given_H0_true,

 size = n,

 replace = TRUE))

}

# Finally plot the results

hist(results_given_H0_true, freq = FALSE, main='Sampling Distribution of the Sample Mean, Given
Null Hypothesis is True', xlab = 'Average Systolic Blood Pressure', ylab = 'Density')

# add line to show values more extreme on upper end

abline(v=x_bar, col = "red")

# add line to show values more extreme on lower end

low_end_extreme <- mean(results_given_H0_true)+(mean(results_given_H0_true)-x_bar)

lines(x = seq(117, 122, .01), dnorm(seq(117, 122, .01), mean = mean(results_given_H0_true), sd =
sd(results_given_H0_true)))

abline(v=low_end_extreme, col="red")

# counts of values more extreme than the test statistic in our original sample, given H0 is true

# two sided given the alternate hypothesis

count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= low_end_extreme)

count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= x_bar)

bootstrap_pvalue <- (count_of_more_extreme_lower_tail +
count_of_more_extreme_upper_tail)/num_sims

bootstrap_pvalue

# two sided t p-value

two_sided_t_pval

# need the standard error which is the standard deviation of the results

bootstrap_SE_X_bar <- sd(results)

# an estimate is to use the formula statistic +/- 2*SE
```

```
c(x_bar - 2*bootstrap_SE_X_bar, x_bar + 2*bootstrap_SE_X_bar)
# you can also use the 5th and 95th quantiles to determine the bounds:
c(quantile(results, c(.025, .975)))
# compare to our t-methods
c(x_bar+(qt(0.025, n-1)*(s/sqrt(n))), x_bar+(qt(0.975, n-1)*(s/sqrt(n))))
p_0 <- 0.48
p_0
p <- length(data$Gender[data$Gender == "Female"])
p
n <- length(data$Gender)
n
p_hat <- p/n
p_hat
z <- (p_hat - p_0) / sqrt((p_0*(1-p_0)) / n)
z
binom.test(x = p, n = n, p = p_0, alternative = "greater")
pnorm(z, lower.tail = FALSE)
cat("Exact Binomial Test")
binom.test(x = p, n = n, p = p_0, alternative = "greater")$conf.int
cat("Normal Approx")
c(p_hat - (1.64)*sqrt(((p_hat)*(1-p_hat))/n), 1)
female <- data$Gender
female
female <- relevel(female, "Male")
levels(female) <- c(0, 1)
female
table(female)
set.seed(0)
# This data is pretty skewed so even though n is large, I'm going to do a lot of simulations
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
```

```
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(as.numeric(sample(x = female, size = n, replace = TRUE))-1)
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Proportion', xlab = 'Proportion of Female', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(.52, .60, .001), dnorm(seq(.52, .60, .001), mean = mean(results), sd = sd(results)))
cat("Bootstrap Confidence Interval")
c(quantile(results, c(0.05,1)))
cat("exact binomial test")
binom.test(x = p, n = n, p = p_0, alternative = "greater")$conf.int
cat("normal approx")
c(p_hat - (1.64)*sqrt(((p_hat)*(1-p_hat))/n),1)
# Under the assumption that the null hypothesis is true, we have 48% female
female_sim <- rep(c(1, 0), c(.48*n, (1-.48)*n))
num_sims <- 10000
# A vector to store my results
results_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results_H0_true[i] <- mean(sample(x = female_sim,
 size = n,
 replace = TRUE))
}
# Finally plot the results
hist(results_H0_true, freq = FALSE, main='Sampling Distribution of the Sample Proportion under H_0:p = 0.48', xlab = 'Proportion of Female', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(.30, .65, .001), dnorm(seq(.30, .65, .001), mean = mean(results_H0_true), sd = sd(results_H0_true)))
```

```
abline(v=p_hat, col="red")
count_of_more_extreme_upper_tail <- sum(results_H0_true >= p_hat)
bootstrap_pvalue <- count_of_more_extreme_upper_tail/num_sims
cat("Bootstrap p-value")
bootstrap_pvalue
cat("Exact Binomial p-value")
binom.test(x = p, n = n, p = p_0, alternative = "greater")$p.value
cat("Normal Approximation p-value")
pnorm(z, lower.tail = FALSE)
qqnorm(data$SystolicBloodPressure, main ="Normality Check for Systolic Blood Pressure Level")
qqline(data$SystolicBloodPressure)
qqnorm(data$SystolicBloodPressure[data$TenYearCoronaryHeartDisease == "Vulnerable"], main ="Normality Check for Systolic Blood Pressure of patients Vulnerable to 10 Year CHD")
qqline(data$SystolicBloodPressure[data$TenYearCoronaryHeartDisease == "Vulnerable"])
qqnorm(data$SystolicBloodPressure[data$TenYearCoronaryHeartDisease == "Immune"], main ="Normality Check for Systolic Blood Pressure of patients Immune to 10 Year CHD")
qqline(data$SystolicBloodPressure[data$TenYearCoronaryHeartDisease == "Immune"])
set.seed(0)
immunePatientsData <- subset(data, data$TenYearCoronaryHeartDisease == "Immune")
immuneDataSample <- immunePatientsData[sample(nrow(immunePatientsData), 300), ]
head(immuneDataSample)
set.seed(0)
vulnerablePatientsData <- subset(data, data$TenYearCoronaryHeartDisease == "Vulnerable")
vulnerableDataSample <- vulnerablePatientsData[sample(nrow(vulnerablePatientsData), 300), ]
head(vulnerableDataSample)
sampleData <- rbind(immuneDataSample, vulnerableDataSample)
head(sampleData)
summary(sampleData)
qqnorm(sampleData$SystolicBloodPressure, main ="Sample Data - Normality Check for Systolic Blood Pressure Level")
qqline(sampleData$SystolicBloodPressure)
```

```
qqnorm(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Vulnerable"], main ="Sample - Data Normality Check for Systolic Blood Pressure of patients Vulnerable to 10 Year CHD")

qqline(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Vulnerable"])

qqnorm(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Immune"], main ="Sample Data - Normality Check for Systolic Blood Pressure of patients Immune to 10 Year CHD")

qqline(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Immune"])

t.test(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Vulnerable"], sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Immune"])

# Mean Systolic Blood Pressure of Vulnerable Patients

mu_v <- mean(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == 'Vulnerable'])

mu_v

# Mean Systolic Blood Pressure of Immune Patients

mu_i <- mean(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == 'Immune'])

mu_i

# Null Hypothesis

mu_0 <- 0

# Variance of Systolic Blood Pressure of Vulnerable Patients

var_v <- var(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == 'Vulnerable'])

var_v

# Variance of Systolic Blood Pressure of Vulnerable Patients

var_i <- var(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == 'Immune'])

var_i

# Sample Size of Systolic Blood Pressure of Vulnerable Patients

n_v <- length(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == 'Vulnerable'])

n_v
```

```
# Sample Size of Systolic Blood Pressure of Vulnerable Patients

n_i <- length(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == 'Immune'])

n_i

# t-value (test statistic)

t <- (mu_v - mu_i - mu_0)/sqrt(var_v/n_v + var_i/n_i)

t

# p-value for 2 sided t-test

p_value <- pt(q = t, df = min(n_v, n_i) - 1, lower.tail = FALSE)*2

p_value

# Lower Boundary of Confidence Interval

lowerBound <- mu_v - mu_i + qt(0.05, min(n_v, n_i) - 1)*sqrt(var_v/n_v + var_i/n_i)

lowerBound

# Upper Boundary of Confidence Interval

upperBound <- mu_v - mu_i + qt(0.95, min(n_v, n_i) - 1)*sqrt(var_v/n_v + var_i/n_i)

upperBound

set.seed(0)

num_sims <- 10000

# A vector to store my results

results <- rep(NA, num_sims)

# A loop for completing the simulation

for(i in 1:num_sims){

 mean_immune <- mean(sample(x = sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == 'Immune'],

 size = 300,

 replace = TRUE))

 mean_vulnerable <- mean(sample(x = sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == 'Vulnerable'],

 size = 300,

 replace = TRUE))

 results[i] <- mean_vulnerable - mean_immune

}
```

```
# Finally plot the results

hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', xlab = 'Average Difference Systolic Blood Pressure', ylab = 'Density')

lines(x = seq(9, 21, .01), dnorm(seq(9, 21, .01), mean = mean(results), sd = sd(results)))

# Bootstrap one-sided CI

c(quantile(results, c(.025, .975)))

t.test(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Vulnerable"], sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Immune"])$conf.int

set.seed(0)

transform(sampleData,Group=sample(TenYearCoronaryHeartDisease))

set.seed(0)

num_sims <- 10000

# A vector to store my results

results_given_H0_true <- rep(NA, num_sims)

# A loop for completing the simulation

for(i in 1:num_sims){

 # the idea here is if there is no relationship we should be able to shuffle the groups

  shuffled_groups <- transform(sampleData,Group=sample(TenYearCoronaryHeartDisease))

  mean_immune <- mean(shuffled_groups$SystolicBloodPressure[shuffled_groups$Group=="Immune"])

  mean_vulnerable <- mean(shuffled_groups$SystolicBloodPressure[shuffled_groups$Group=="Vulnerable"])

  results_given_H0_true[i] <- mean_vulnerable - mean_immune

}

# Finally plot the results

hist(results_given_H0_true, freq = FALSE, main='Dist. of the Diff in Sample Means Under Null', xlab = 'Average Difference Systolic Blood Pressure under Null', ylab = 'Density')

diff_in_sample_means <- mean(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease=="Vulnerable"]) - mean(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease=="Immune"])

low_end_extreme <- mean(results_given_H0_true) + (mean(results_given_H0_true) - diff_in_sample_means)
```

```
lines(x = seq(-6, 6, .01), dnorm(seq(-6, 6, .01), mean = mean(results_given_H0_true), sd = sd(results_given_H0_true)))

abline(v=diff_in_sample_means, col = "blue")

abline(v=abs(diff_in_sample_means), col = "red")

# counts of values more extreme than the test statistic in our original sample, given H0 is true

# two sided given the alternate hypothesis

count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= low_end_extreme)

count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= diff_in_sample_means)

bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims

cat("Bootstrap p-value")

bootstrap_pvalue

cat("t-test p-value")

t.test(sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Vulnerable"], sampleData$SystolicBloodPressure[sampleData$TenYearCoronaryHeartDisease == "Immune"])$p.value

# the parts of the test statistic

# sample props

p_hat_M <- length(data$Gender[data$Gender == "Male" & data$TenYearCoronaryHeartDisease == "Vulnerable"])/length(data$Gender[data$Gender == "Male"])

p_hat_F <- length(data$Gender[data$Gender == "Female" & data$TenYearCoronaryHeartDisease == "Vulnerable"])/length(data$Gender[data$Gender == "Female"])

# null hypothesized population prop difference between the two groups

p_0 <- 0

# sample size

n_M <- length(data$Gender[data$Gender == "Male"])

n_F <- length(data$Gender[data$Gender == "Female"])

# sample variances

den_p_M <- (p_hat_M*(1-p_hat_M))/n_M

den_p_F <- (p_hat_F*(1-p_hat_F))/n_F

# z-test test statistic

z <- (p_hat_M - p_hat_F - p_0)/sqrt(den_p_M + den_p_F)

z
```

```
# two sided p-value

two_sided_diff_prop_pval <- pnorm(q = z, lower.tail = FALSE)*2

two_sided_diff_prop_pval

# lower bound

(p_hat_M - p_hat_F)+(qnorm(0.025)*sqrt(den_p_M + den_p_F))

# upper bound

(p_hat_M - p_hat_F)+(qnorm(0.975)*sqrt(den_p_M + den_p_F))

# Make the data

male <- rep(c(1, 0), c(length(data$Gender[data$Gender == "Male" &
data$TenYearCoronaryHeartDisease == "Vulnerable"]), n_M - length(data$Gender[data$Gender ==
"Male" & data$TenYearCoronaryHeartDisease == "Vulnerable"])))

female <- rep(c(1,0), c(length(data$Gender[data$Gender == "Female" &
data$TenYearCoronaryHeartDisease == "Vulnerable"]), n_F - length(data$Gender[data$Gender ==
"Female" & data$TenYearCoronaryHeartDisease == "Vulnerable"])))

num_sims <- 10000

# A vector to store my results

results <- rep(NA, num_sims)

# A loop for completing the simulation

for(i in 1:num_sims){

 prop_M <- mean(sample(male,

 size = n_M,

 replace = TRUE))

 prop_F <- mean(sample(x = female,

 size = n_F,

 replace = TRUE))

 results[i] <- prop_M - prop_F

}

# Finally plot the results

hist(results, freq = FALSE, main='Dist. of the Diff in Prop', xlab = 'Difference in Prop. of Patients
Vulnerable to Heart Disease', ylab = 'Density')

lines(x = seq(0.01, 0.13, .001), dnorm(seq(0.01, 0.13, .001), mean = mean(results), sd = sd(results)))

cat("Bootstrap")

c(quantile(results, c(.025, .975)))
```

```
cat("Normal Approximation")

c((p_hat_M - p_hat_F)+(qnorm(0.025)*sqrt(den_p_M + den_p_F)), (p_hat_M -
p_hat_F)+(qnorm(0.975)*sqrt(den_p_M + den_p_F)))

# Make the data

df_combined <- data.frame("vulnerable_patients" = c(male, female), "gender" = rep(c("male",
"female"), c(n_M, n_F)))

# Sanity checks

summary(df_combined$gender)

mean(df_combined$vulnerable_patients[df_combined$gender=="male"]) == p_hat_M

mean(df_combined$vulnerable_patients[df_combined$gender=="female"]) == p_hat_F

num_sims <- 1000

# A vector to store my results

results_given_H0_true <- rep(NA, num_sims)

# A loop for completing the simulation

for(i in 1:num_sims){

 # the idea here is if there is no relationship we should be able to shuffle the groups

 shuffled_groups <- transform(df_combined, gender=sample(gender))

 prop_M <- mean(shuffled_groups$vulnerable_patients[shuffled_groups$gender=="male"

])

 prop_F <- mean(shuffled_groups$vulnerable_patients[shuffled_groups$gender=="female"

])

 results_given_H0_true[i] <- prop_M - prop_F

}

# Finally plot the results

hist(results_given_H0_true, freq = FALSE,

 main='Dist. of the Diff in Sample Sample Props Under Null',

 xlab = 'Average Difference in Prop. Vulnerable Patients under Null',

 ylab = 'Density')

diff_in_sample_props <- p_hat_M - p_hat_F

lines(x = seq(-0.05, 0.05, .001), dnorm(seq(-0.05, 0.05, .001), mean = mean(results_given_H0_true), sd
= sd(results_given_H0_true)))

abline(v=diff_in_sample_props, col = "blue")
abline(v=-diff_in_sample_props, col = "red")
# counts of values more extreme than the test statistic in our original sample, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= -diff_in_sample_props)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= diff_in_sample_props)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail +
count_of_more_extreme_upper_tail)/num_sims
cat("Bootstrap p-value")
bootstrap_pvalue
cat("Normal Approx p-value")
two_sided_diff_prop_pval
chiData <- data$Education
head(chiData)
table(chiData)
prop.table(table(chiData))
n <- 3658
r <- 4
npi <- 914.5
tchi <- sum(((table(chiData) - npi)^2)/npi)
tchi
p_value <- pchisq(tchi, df = r-1, lower.tail = FALSE)
p_value
# Create our data under the assumption that H_0 is true
solutions_under_H_0 <- rep(c("C", "GED", "HS", "VS"), npi)
# Sanity Check
table(solutions_under_H_0)
num_sims <- 10000
# A vector to store my results
chisq_stats_under_H0 <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 new_samp <- sample(solutions_under_H_0, n, replace = T)
 chisq_stats_under_H0[i] <- sum(((table(new_samp) - npi)^2)/npi)
}
hist(chisq_stats_under_H0, freq = FALSE,
 main='Dist. of the Chi-Square Statistic Under Null',
 xlab = 'Chi-Square Stat under Null',
 ylab = 'Density')
abline(v=sum(((table(chiData) - npi)^2)/npi), col="red")
#The randomization p-value
sum(chisq_stats_under_H0 >= sum(((table(chiData) - npi)^2)/npi))/num_sims
```