Project Part 1

Introduction

The number one cause of deaths in the US is coronary heart disease. I wanted to research what factors have a correlation to coronary heart disease to better understand and target the problem. Once I began researching, I learned that one of the most critical problems that affects coronary heart disease in the US is the obesity epidemic. Obese individuals require more blood to supply oxygen and nutrients to their bodies, which causes an increase in blood pressure (Penn 2019). This increase in blood pressure is one of the main causes of heart attacks and other cardiovascular diseases. Obesity can also cause a spike in bad cholesterol and triglyceride levels as well as lower good cholesterol, which is important to reduce the risk for heart disease. Sleep is another factor that affects CHD. According to the CDC, adults who sleep less than 7 hours each night are more likely to say they have had health problems, including heart attack, asthma, and depression. Some of these health problems raise the risk for heart disease, heart attack, and stroke (CDC 2021). I want to be able to analyze the trend in heart disease deaths over the past 2 decades in the US overall and specifically in Virginia in 2020. With the rise of medical advancements, it's important to further study these correlations because we will better be able to target solutions for CHD knowing the range of influence between the multiple factors stated above.

The questions I am trying to answer are:

- Do counties in Virginia in 2020 that have a higher percent of the population that sleep less than 7 hours per night have a higher percent of people with CHD (coronary heart disease)?
- Do counties in Virginia in 2020 that have a higher percent of obesity in the population have a higher percent of people with CHD (coronary heart disease)?
- Is there a change in deaths due to heart diseases in the US from 1999 and 2020?

Data Summary

The data for the first table was found using the CDC's Environmental National Public Health Data Explorer which is a collaboration containing multiple sources of data, including the U.S. Census Bureau and the National Health Service so I believe it is credible. The response variable for the first two research questions is the percent of adults in Virginia in 2020 who have coronary heart disease. The explanatory variables for the first two questions are percent of the adult population in Virginia in 2020 that got more than 7 hours of sleep and percent of the adult population in Virginia in 2020 that have obesity. The population for the first dataset are adults over 18 in counties in Virginia and each observation in the data set is one county, with the merged dataset containing 133 counties. The people in the counties were surveyed and the data was collected to predict geographic distribution of health behaviors of individuals.

The data for the second table was found using the Center for Disease Control and Prevention which gathers data from multiple primary sources so I believe it is credible. The response variable for the third research question is the number of deaths of heart disease in the US. The explanatory variable for the third question is the years 1999-2020. The population for the second dataset are adults in counties in the USA (all states). The merged dataset contains the number of heart disease deaths in years 1999-2020. The data is collected from health records for the purpose of determining the prevalence of heart disease in the US population.

There were some counties in Virginia that did not have any data so I had to exclude those counties from the first data table. A potential issue I am worried about in the data is the fact that the population of people in the US has mostly increased each year from 2010 to 2020, which could affect the results of the graphical summaries related to the number of people who died from heart diseases. I also thought there would be a correlation between the average added sugar consumption per year in the US adult population diet and heart disease deaths, but when I created a scatter plot there seemed to be no correlation so I didn't include that variable in my analysis however it is still included in my second dataset (data2).

Data Dictionary

Table 1 (Research Questions 1 and 2)

Variable	Type	Units	Measurement
CHD (Coronary Heart Disease)	Response	Quantitative (% of population)	Age-adjusted prevalence of current coronary heart disease among adults >= 18 (Virginia, 2020)
Obesity	Explanatory	Quantitative (% of population)	Age-adjusted prevalence of Obesity among adults >= 18 (Virginia, 2020)
Sleep	Explanatory	Quantitative (% of population)	Age-adjusted prevalence of sleeping less than 7 hours among adults >= 18 (Virginia, 2020)

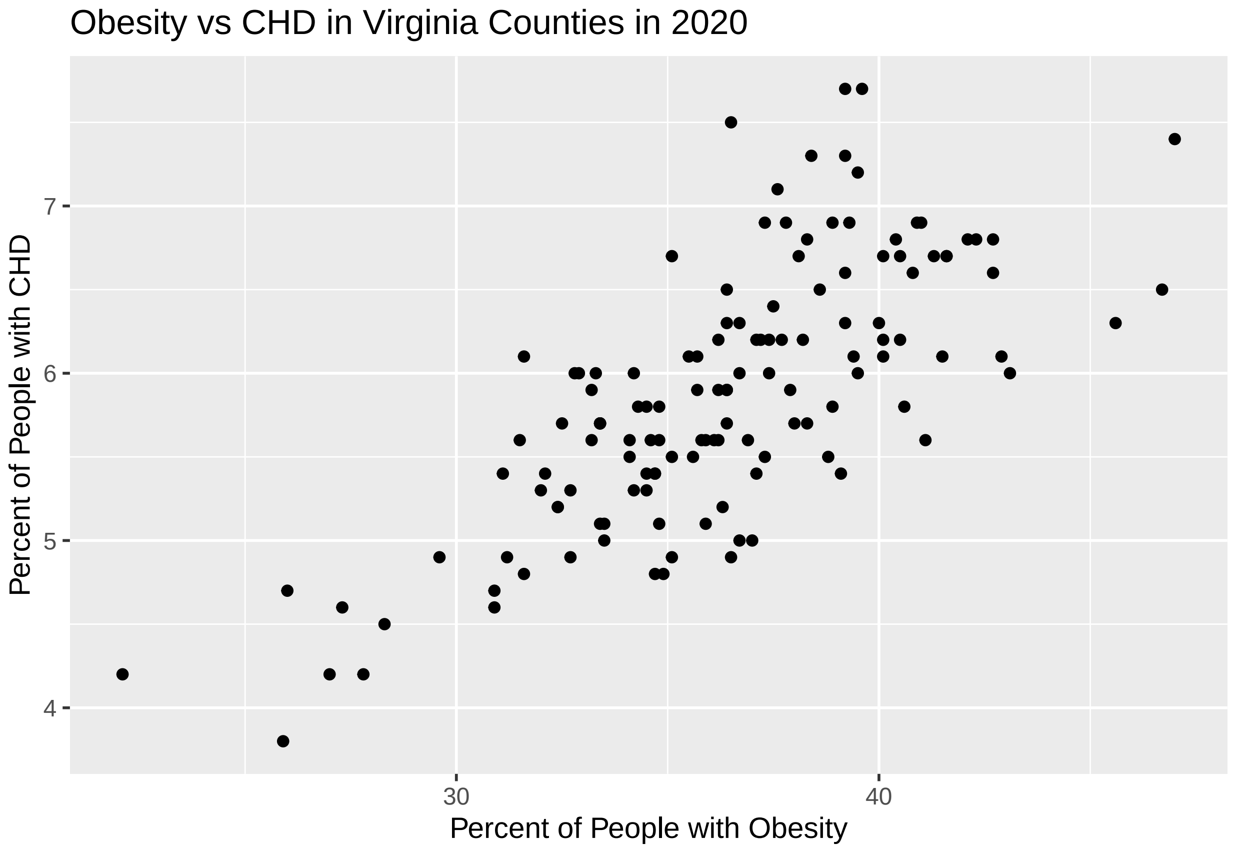
Table 2 (Research Question 3)

Variable	Type	Category	Measurement
Heart.Disease.Deaths	Response	Quantitative (# of people)	Deaths in the US
Time	Explanatory	Quantitative (% of population)	1999-2020

EDA

Research Question 1 - Scatter Plot of Obesity vs CHD (Cardiovascular Heart Disease)

```
graph1<- ggplot(data1, aes(x=Obesity, y=CHD)) + geom_point() +
  labs(title="Obesity vs CHD in Virginia Counties in 2020",
       x= "Percent of People with Obesity", y="Percent of People with CHD")
graph1
```

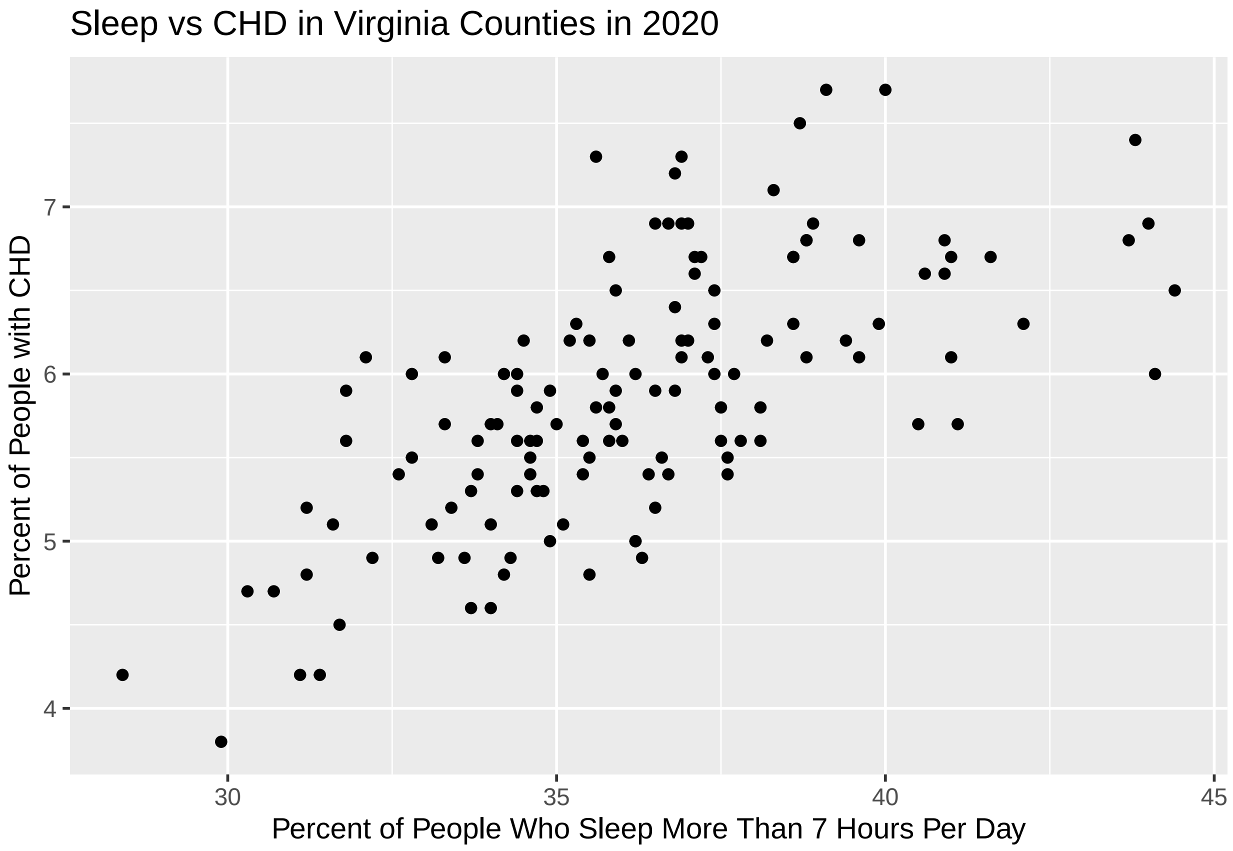


Correlation Coefficent of Obesity vs CHD

```
[1] 0.7364742
```

Research Question 2 - Scatter Plot of Sleep vs CHD (Cardiovascular Heart Disease)

```
graph2<- ggplot(data1, aes(x=Sleep, y=CHD)) + geom_point() +
  labs(title="Sleep vs CHD in Virginia Counties in 2020",
       x= "Percent of People Who Sleep More Than 7 Hours Per Day",
       y="Percent of People with CHD")
graph2
```

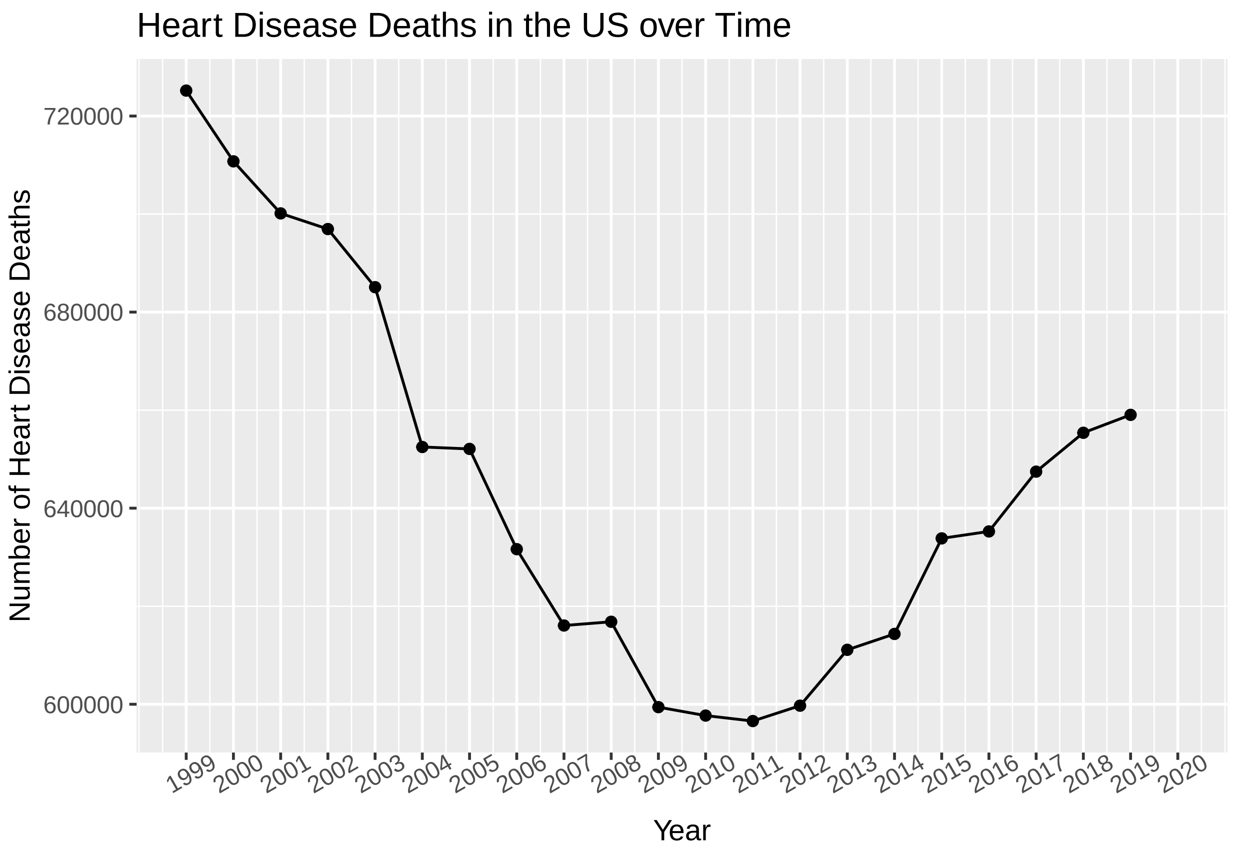


Correlation Coefficent of Sleep vs CHD

```
[1] 0.6735712
```

Research Question 3 - Line plot of Heart Disease Deaths in the US over Time

```
graph3 <- ggplot(data2, aes(x=Time, y=Heart.Disease.Deaths)) + geom_line() +
  scale_x_continuous(breaks = seq(1999,2020), limits=c(1999,2020))+
  labs(title="Heart Disease Deaths in the US over Time",
       x="Year", y="Number of Heart Disease Deaths") + geom_point() +
  theme(axis.text.x = element_text(angle=30))
graph3
```



5 Number Summary of the Number of Heart Disease Deaths

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s 
 596577  614348  635260  644622  659041  725192      12 
```

Conclusion

The first question I wanted to research was: do counties in Virginia in 2020 that have a higher percent of obesity in the population have a higher percent of people with CHD(coronary heart disease)? According to the first scatter plot (graph1) we can see that there is a positive linear correlation between percent of people with Obesity and percent of people with CHD. This is also shown in the numerical summary of the correlation coefficient between Obesity and CHD. The correlation coefficient of 0.7364742 indicates a moderately strong positive relationship between Obesity and CHD. This indicates that as the percent of people with obesity increases, the percentage of people with CHD also increases. Further study can be done within obesity to determine what factors such as poor diet or amount of excess fat contribute to heart diseases.

The second question I wanted to research was: do counties in Virginia in 2020 that have a higher percent of the population that sleep less than 7 hours per night have a higher percent of people with CHD(coronary heart disease)? According to the second scatter plot (graph2) there is a moderately strong positive linear relationship between the percent of people who sleep more than 7 hours per day and the percent of people with CHD. The correlation coefficient of 0.6735712 indicates a moderately strong positive relationship between Sleep and CHD. Sleeping more than 7 hours every night might help people prevent heart diseases in the future since there is a moderately strong association between sleep and CHD.

The third question I wanted to research was: is there a change of deaths due to heart diseases in the US from 1990 to 2020? According to the line plot (graph3), the number of people who died from heart disease decreased between 1999 and 2010, where it was at its lowest, then began to slowly increase until 2020. The number of heart disease deaths was at its highest in 1999. I conducted a 5 number summary on the Heart.Disease.Deaths variable to see the spread of data and how big of a problem heart disease is in the US. The minimum number of deaths in a given year was 596577 and the maximum was 725192 with a mean of 644622. The data is slightly skewed to the right which could mean that there are some outliers in the data that are causing the median to be slightly less than the mean. This indicates that overall there were on average more deaths in the first decade (1999-2009) than the second decade (2010-2020).

Overall, the two factors that have the strongest correlation to CHD is Obesity and Sleep according to the EDA. The decrease in overall heart disease deaths from 1999-2010 could be due to the rapid progress of treatment and prevention of heart disease and the increase from 2010-2020 could be due to the increase of high fat and sugar diets in the population leading to obesity. Medically, since excess fat can lead to build up in arteries and the correlation coefficient of Obesity is larger than Sleep, for the second part of this project, I want to statistically analyze my second research question on the correlation of obesity and CHD in Virginia counties in 2020.

Appendix of Data

```
   StateFIPS    State CountyFIPS      County Year CHD Sleep Obesity
1         51 Virginia      51001    Accomack 2020 6.8  38.8    38.3
2         51 Virginia      51003   Albemarle 2020 4.7  30.3    26.0
3         51 Virginia      51510  Alexandria 2020 4.6  33.7    27.3
4         51 Virginia      51005   Alleghany 2020 6.3  37.4    36.7
5         51 Virginia      51007      Amelia 2020 6.1  36.9    35.7
6         51 Virginia      51009     Amherst 2020 5.9  36.8    36.4
7         51 Virginia      51011  Appomattox 2020 6.0  37.4    36.7
8         51 Virginia      51013   Arlington 2020 4.2  28.4    27.8
9         51 Virginia      51015     Augusta 2020 5.8  34.7    34.8
10        51 Virginia      51017        Bath 2020 6.9  36.5    38.9
11        51 Virginia      51019     Bedford 2020 5.4  33.8    32.1
12        51 Virginia      51021       Bland 2020 6.2  36.1    37.4
13        51 Virginia      51023   Botetourt 2020 5.4  34.6    34.5
14        51 Virginia      51520     Bristol 2020 6.9  36.7    37.3
15        51 Virginia      51025   Brunswick 2020 6.8  40.9    42.3
16        51 Virginia      51027    Buchanan 2020 7.7  40.0    39.6
17        51 Virginia      51029  Buckingham 2020 6.8  38.8    42.7
18        51 Virginia      51530 Buena Vista 2020 6.9  36.9    37.8
19        51 Virginia      51031    Campbell 2020 6.0  37.7    39.5
20        51 Virginia      51033    Caroline 2020 5.5  37.6    38.8
```

```
   Time Heart.Disease.Deaths Added.Sugar
1  1999               725192        22.0
2  2000               710760        21.8
3  2001               700142        21.4
4  2002               696947        21.0
5  2003               685089        20.2
6  2004               652486        20.5
7  2005               652091        20.9
8  2006               631636        20.6
9  2007               616067        20.3
10 2008               616828        21.6
11 2009               599413        21.1
```

```
12 2010               597689        21.9
13 2011               596577        21.9
14 2012               599711        22.1
15 2013               611105        22.6
16 2014               614348        22.7
17 2015               633842        23.0
18 2016               635260        23.2
19 2017               647457        23.0
20 2018               655381        22.8
```

Works Cited

Centers for Disease Control and Prevention. (2021, January 4). How does sleep affect your heart health? Centers for Disease Control and Prevention. Retrieved March 29, 2023, from https://www.cdc.gov/bloodpressure/sleep.htm#:~:text=Adults%20who%20sleep%20less%20than,attack%2C%20asthma%2C%20and%20depression.&text=Some%20of%20these%20health%20problems,%2C%20heart%20attack%2C%20and%20stroke

Centers for Disease Control and Prevention. (n.d.). National Environmental Public Health Tracking Network Data explorer. Centers for Disease Control and Prevention. Retrieved March 29, 2023, from https://ephtracking.cdc.gov/DataExplorer/

Pennmedicine.org. (n.d.). Retrieved March 29, 2023, from https://www.pennmedicine.org/updates/blogs/metabolic-and-bariatric-surgery-blog/2019/march/obesity-and-heart-disease#:~:text=Obese%20individuals%20require%20more%20blood,more%20common%20for%20obese%20individuals.

Shibboleth authentication request. (n.d.). Retrieved March 29, 2023, from https://dataplanet-sagepub-com.proxy1.library.virginia.edu/dataset?view=AA0BXQAAgACVAQAAAAAAAAAA3_zMslwIJ8Ve1X%24GAc7FAUjaq8ZQ7yYRHqC8LGvYqP15pysWNlrQCufoUNQMJ7R3UTBp5yIOc_fYFnceD16Naff3lI8liJBQFlzPfJevz6L6y14Ene_DI5oEBjWGC4MrsT7yVm84B8PuxQzi1hevsoAyz8xF5Pft%24NX8acf4gtfHUnQ8P9HXDkoQT_eJew2VJ%24lmDKESj4ixskhvH9YOTjROzSXNPiZOE_DWb08SYjyo448L6gTITKJmouuiXIoBby%24cY1o

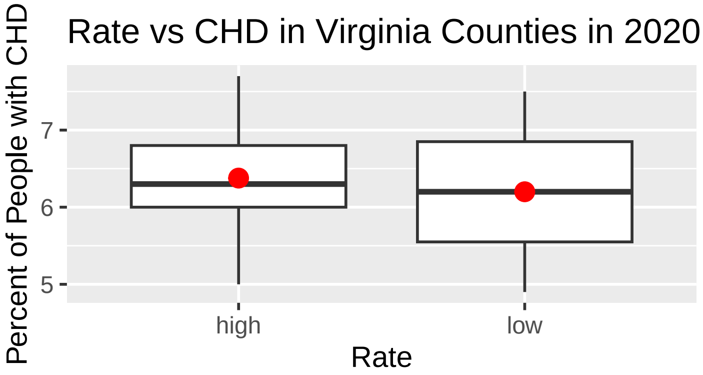
Project Part 2

Introduction

The number one cause of deaths in the US is coronary heart disease. I wanted to research what factors have a correlation to coronary heart disease to better understand and target the problem. From part 1 of the project I found that obesity has the highest correlation to CHD so I wanted to further analyze this data. Obese individuals require more blood to supply oxygen and nutrients to their bodies, which causes an increase in blood pressure (Penn 2019). This increase in blood pressure is one of the main causes of heart attacks and other cardiovascular diseases. Obesity can also cause a spike in bad cholesterol and triglyceride levels as well as lower good cholesterol, which is important to reduce the risk for heart disease. According to the NIH, my choosen value of 36.5 is the closest to the average percent of obese people in a county so I split the obesity column into high and low based on that value. The question I am trying to answer is: Do counties in Virginia in 2020 that have a higher percent of obesity in the population have a higher percent of people with CHD (coronary heart disease)?

EDA

I analyzed two boxplots for the high and low obesity populations and found their mean to be slightly different so I decided to conduct a two sample t test for means to further explore if this difference is significant.



Methods/Analysis

I first assessed my question to decide what hypothesis test I should conduct. Since I am comparing average CHD in two different samples (high and low obesity) I decided to conduct the two sample t test for means since the population standard deviation is not known. The independence assumption is met because the data is collected through random sampling and the counties are independent of each other. From the QQ plots shown in the appendix (left = high, right = low) since the residuals follow the normal distribution line we can conclude that the normality assumption is met for both populations. The constant variance assumption also seems to be met as the residual plot of both populations seems to have not much fanning out pattern and the values fall roughly between the horizontal bands.

The hypothesis that I will be testing are: Ho: μ(CHD in high obesity) = μ(CHD in low obesity) Ha: μ(CHD in high obesity) > μ(CHD in low obesity)

Results

```
t.test(highO$CHD, lowO$CHD, mu=0, alternative="greater")
```

```
    Welch Two Sample t-test

data:  highO$CHD and lowO$CHD
t = 9.147, df = 127.02, p-value = 6.115e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.7910769       Inf
sample estimates:
mean of x mean of y
 6.371875  5.405797
```

After conducting the two sample t test we found our p value to be 6.115e-16 and our test statistics to be 9.147. Using an alpha of 0.05 since the p-value is less than 0.05 we can reject the null hypothesis that the mean of CHD in the high obesity population is equal to the mean of CHD in the low population. We can conclude that mean of CHD in the high obesity population is greater than the mean of CHD in the low population.

Conclusion

```
library(pwr)
pwr.t2n.test(d=0.5, n1=64, n2=69, sig.level=0.05, alternative="greater")
```

```
     t test power calculation

             n1 = 64
             n2 = 69
              d = 0.5
      sig.level = 0.05
          power = 0.8890175
    alternative = greater
```

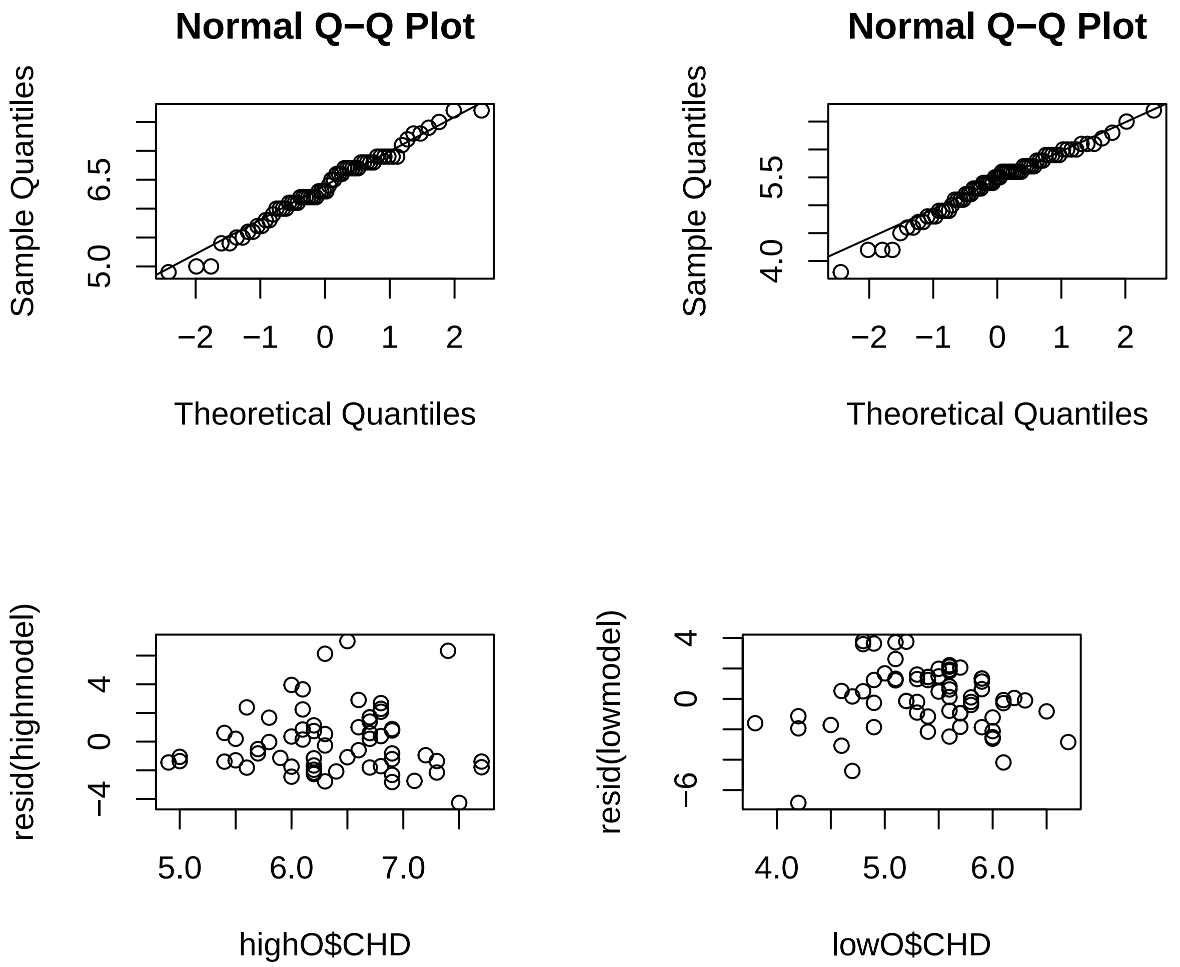
I choose a medium effect size because the resulting power is above 0.8 which supports the validity of the test. Increasing the power beyond 0.8 would require a large sample size. The power of the test is 0.8890175 which means the probability that the null hypothesis will be rejected when the alternative hypothesis is true is high.

A two sample t test requires two independent normal distributions which can limit the type of populations selected for this hypothesis test. I can further expand this research by conducting ANOVA or multilinear regression analysis to compare multiple factors to CHD levels, not just obesity.

Based on the conclusions from the hypothesis test we can assume that there is a correlation between obesity percentage and CHD levels. Populations with a high obesity percentage are more likely to have high CHD levels than lower obesity populations which is consistent with my research as obesity can cause a rise in cholesterol levels which can lead to heart strokes.

We can generalize this conclusion to counties in Virginia in 2020 but not for all population or years as it is possible that the correlation may not hold true for other populations. Correlation does not imply causation. The conclusion that high obesity populations are more likely to have high CHD levels does not necessarily mean that obesity causes CHD. There may be other factors that are contributing to the correlation, and additional research would be necessary to determine the causal relationship between obesity and CHD.

Appendix


Normal Q−Q Plot
Sample Quantiles
5.0
6.5
−2
−1
0
1
2
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
4.0
5.5
−2
−1
0
1
2
Theoretical Quantiles
resid(highmodel)
−4
0
4
5.0
6.0
7.0
highO$CHD
resid(lowmodel)
−6
0
4
4.0
5.0
6.0
lowO$CHD

References

1. Centers for Disease Control and Prevention. (n.d.). National Environmental Public Health Tracking Network Data explorer. Centers for Disease Control and Prevention. Retrieved March 29, 2023, from https://ephtracking.cdc.gov/DataExplorer/

2. Pennmedicine.org. (n.d.). Retrieved March 29, 2023, from <https://www.pennmedicine.org/updates/blogs/metabolic-and-bariatric-su rgery-blog/2019/march/obesity-and-heart-disease#:~:text=Obese%20indivi duals%20require%20more%20blood,more%20common%20for%20obese%20individuals.>

3. Shibboleth authentication request. (n.d.). Retrieved March 29, 2023, from https://dataplanet-sagepub-com.proxy1.library.virginia.edu/dataset?view=AA0BXQAAgACVAQAAAAAAAAAA3_zMslwIJ8Ve1X%24GAc7FAUjaq8ZQ7yYRHqC8LGvYqP15pysWNlrQCufoUNQMJ7R3UTBp5yIOc_fYFnceD16Naff3lI8liJBQFlzPfJevz6L6y14Ene_DI5oEBjWGC4MrsT7yVm84B8PuxQzi1hevsoAyz8xF5Pft%24NX8acf4gtfHUnQ8P9HXDkoQT_eJew2VJ%24lmDKESj4ixskhvH9YOTjROzSXNPiZOE_DWb08SYjyo448L6gTITKJmouuiXIoBby%24cY1o

4. U.S. Department of Health and Human Services. (n.d.). Overweight & Obesity Statistics - Niddk. National Institute of Diabetes and Digestive and Kidney Diseases. Retrieved May 6, 2023, from https://www.niddk.nih.gov/health-information/health-statistics/overwei ght-obesity#:~:text=including%20severe%20obesity).-,About%201%20in%201 1%20adults%20(9.2%25)%20have%20severe%20obesity,who%20are%20overweight%20(27.5%25).