Phase 3

Saaketh Chenna

31/03/2022

Problem

Dataset Overview

This dataset contains information regarding lifestyle choices, demographic, physical characteristics of the individuals in this study. In particular, this dataset is focused on an individuals heart health and the presense of heart disease and factors that may impact cardiovascular health.

Research Objective

To understand what factors (lifestyle choices, demographic, physical characteristics etc.) impact cardiovascular health.

Research Questions

1. What is the mean weight of someone who has cardiovascular disease?
2. Is a normal cholesterol level more prevalent among those who smoke compared to those who do not?
3. Is there a positive correlation between weight and systolic blood pressure?
4. How does diastolic blood pressure vary between patients who have normal, above normal, and well above normal glucose levels, respectively?
5. Is cardiovascular disease more common in males?

Plan

Population of Interest

The population of interest for this study is adults as heart disease is not prevalent in children.

Sampling Frame

The sampling frame consists of females and males ages 19 and up in Canada. One of the recorded characteristics for this study is alcohol intake so we choose the sampling frame as 19 and above as 19 is the oldest legal drinking age in Canada.

Sample

Our sample consists of 70,000 individuals randomly selected from the sampling frame. The sample consisted of half men and half women.

Sampling Design

To get participants for our sample we contacted doctors across Canada, in every province and territory. The patients were contacted by email and the doctors supplied us information regarding patient health and lifestyle habits from individuals who consented to having their information used. This study implemented **stratified sampling** and **random sampling**. The individuals were then split into two **strata** based on gender. This information was placed into a computer algorithm which randomly selected individuals from both strata to create our sample of 70,000 individuals which composed of half females and half males.

Study Design

The study we are conducting is an **observational study**. The study design we are implementing is a **case-control study**: a sample of **controls** who do not have the condition (do not have heart disease) were compared to cases with the outcome of interest (have heart disease). The **case-control study** allows us to match the cases with the **controls** and compare the two. The variable being monitored to indicate poor heart health is the presence or absence of cardiovascular disease.

Disclaimer

The data we are analyzing is real and the dataset is composed of 70,000 individuals; however, the method of collection and analysis of the data is made up. For example we mentioned that the study was conducted in Canada and the sample is composed of half females and half males. In reality, the actual dataset is comprised of mostly females and there is no indication of where and when the study took place.

Data

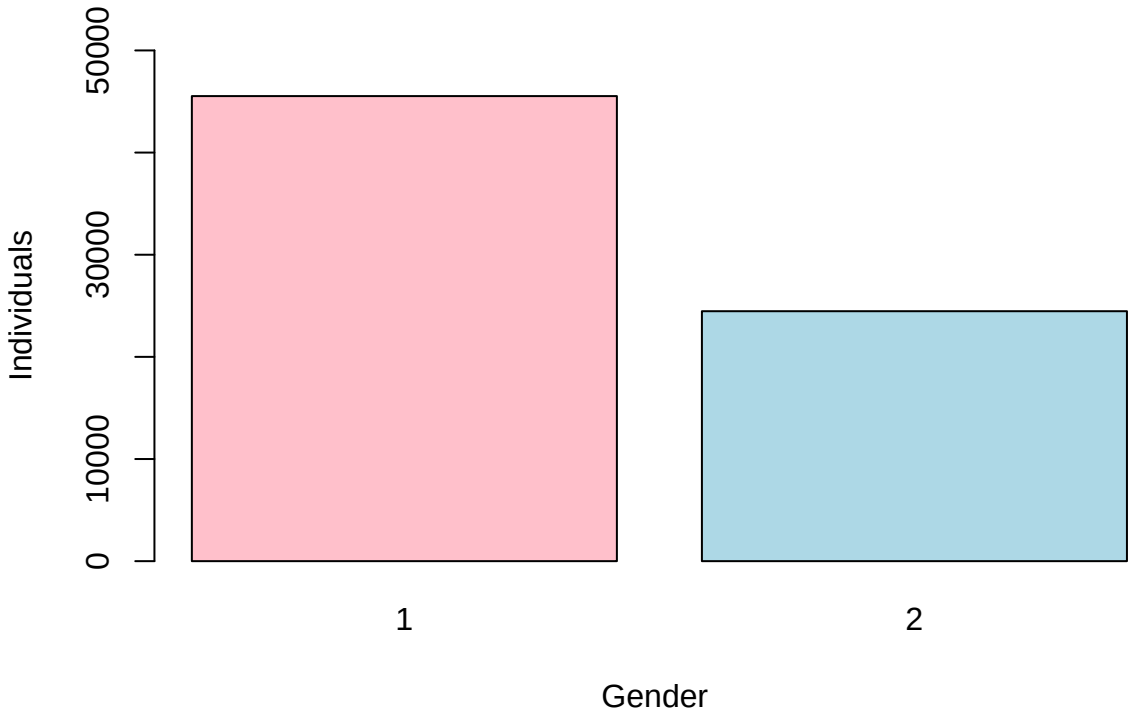
Graphical Summaries

Bar Plot

Bar plotss are used to present one or more categorical variables. Bar plots can be used for univariate, bivariate and multivariate data. A bar graph can be oriented horizontally or vertically. To create a bar graph we use the barplot() function.

For example, a bar plot can be used to demonstrate the frequency of females and males within the sample.

```
genderdata <- table(table$gender)
barplot(genderdata, beside = TRUE, xlab = "Gender ",
ylab = "Individuals", col = c("pink","light blue"), ylim = c(0,50000))
```



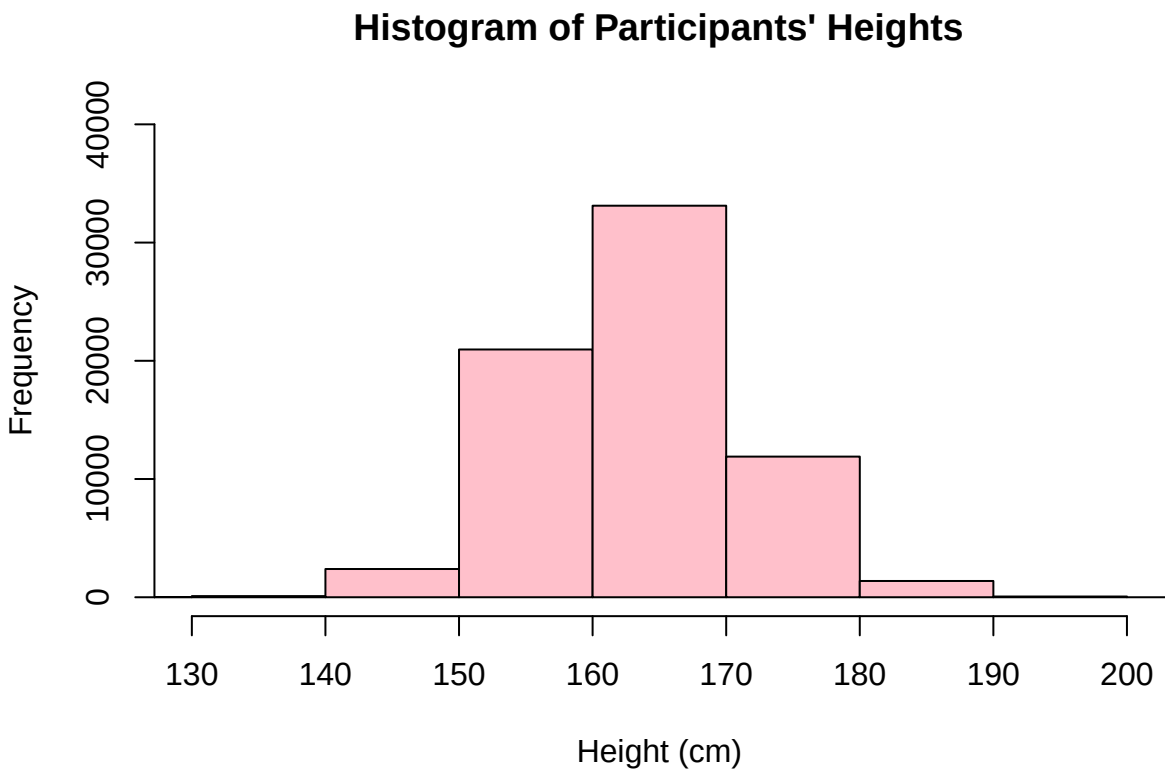

```
  # In the bar graph 1 represents the females and 2 represents the males.
```

Histograms/ Dot plots

Histograms and dot plots gives a visual representation of the frequency distributions for a single quantitative variable. A histogram is typically use when the data set is large whereas a dot plot is used for smaller sets of data. To create a histogram we use he hist() function.

For example, we can use a histogram (as our data set is large with 70,000 individuals) to represent the range of participants' heights.

```
hist(table$height,xlab = "Height (cm)",xlim = c(130,200), ylim=c(0,40000),
    col = "pink",  main = "Histogram of Participants' Heights")
```

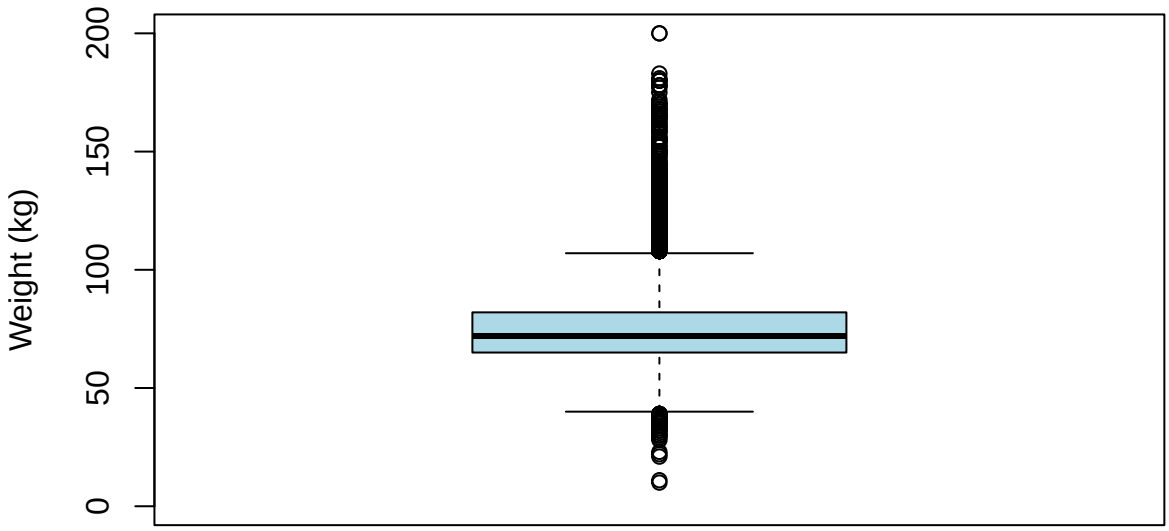


Boxplot

A boxplot is a visual representation of the 5 number summary. The five number summary is composed of the minimum value, first quartile value, median value, third quartile value, and the maximum value. A boxplot is used for a single quantitative variable. A boxplot can also be used for a single quantitative variable across categorical groups. Additionally, box plots help with separating outliers in the data set. To create a boxplot we use the use the boxplot() function provided in R.

For example, a box plot can be use for the quantitatve varible such as the participants' weights.

```
boxplot(table$weight, ylab = "Weight (kg)", col= "light blue",  ylim = c(0,200))
```



Stripchart

A stripchart plots a quantitative variable across categorical groups. Stripcharts can use either bivariate or multivariate data. We can also apply jitter to make the data more visible in the graph as when there us a large dataset the data points tend, making the values harder to visualize. To create a stripchart we use the stripchart() function provided in R.

For example, we can graph the participants' systolic blood pressure (mmHg) against their consumption of alcohol. The figure below is an example of a good figure.

```
stripchart(table$ap_hi~table$alco, vertical = TRUE, method = "jitter", jitter = 0.40,
           pch = 1, col = c("pink"),
           #group.names = c("Does not consume","Does consume"),
          xlab = "Alcohol Consumption",
           ylab = "Systolic Blood Pressure (mmHg)", ylim = c(10,250),
           main = " Figure 1")
```

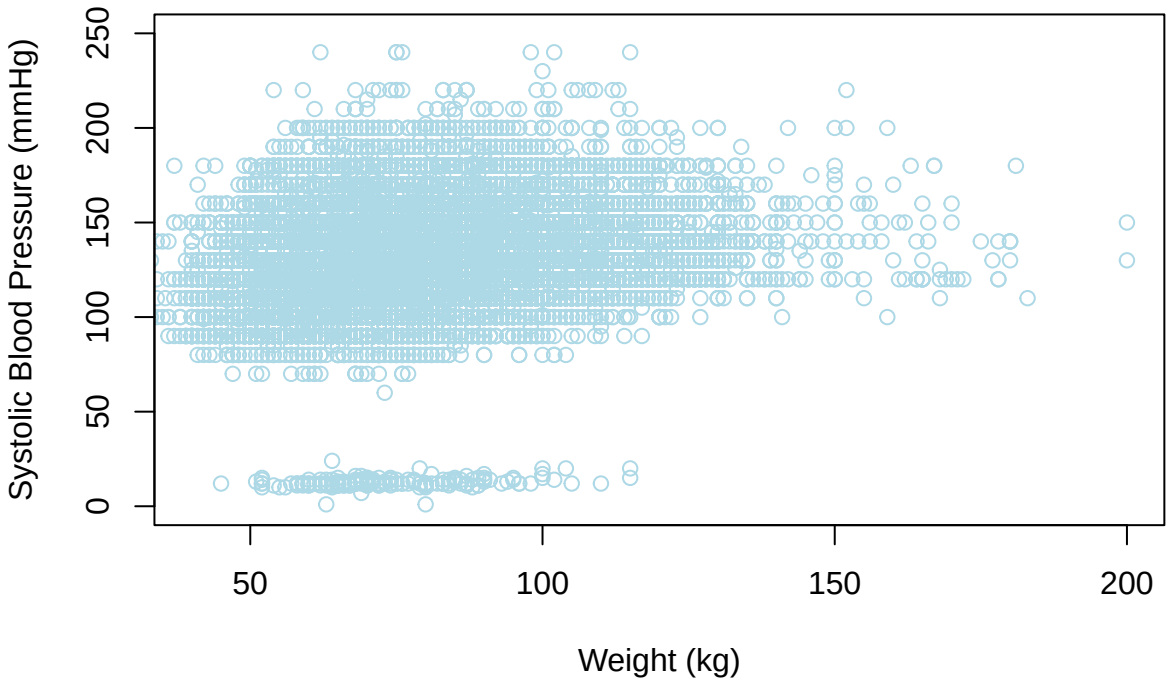
Figure 1. The relationship between Alcohol Consumption and Systolic Blood Pressure (mmHg). The x axis represents if the participants consume alcohol where 0 represents does not consume alcohol and 1 represents consumes alcohol. This data was taken from a sample of 70,000 individuals, both females and males, above the age o f 19.

Scatterplot

A scatterplot uses Cartesian coordinates to display two quantitative variables as coordinates. A scatterplot is typically used for bivariate data however multivariate data can be used when introducing a categorical variable. The categorical variable can be represented as a different shape or colour on the scatterplot. To create a scatterplot we use the scatterplot() function provided in R.

For example, a stripchart can be used to visualize the relationship between weight and systolic blood pressure.

```
plot(table$weight,table$ap_hi, xlim = c(40,200), ylim = c(0,250), xlab = "Weight (kg)",
ylab = "Systolic Blood Pressure (mmHg)", col=c("light blue"))
```

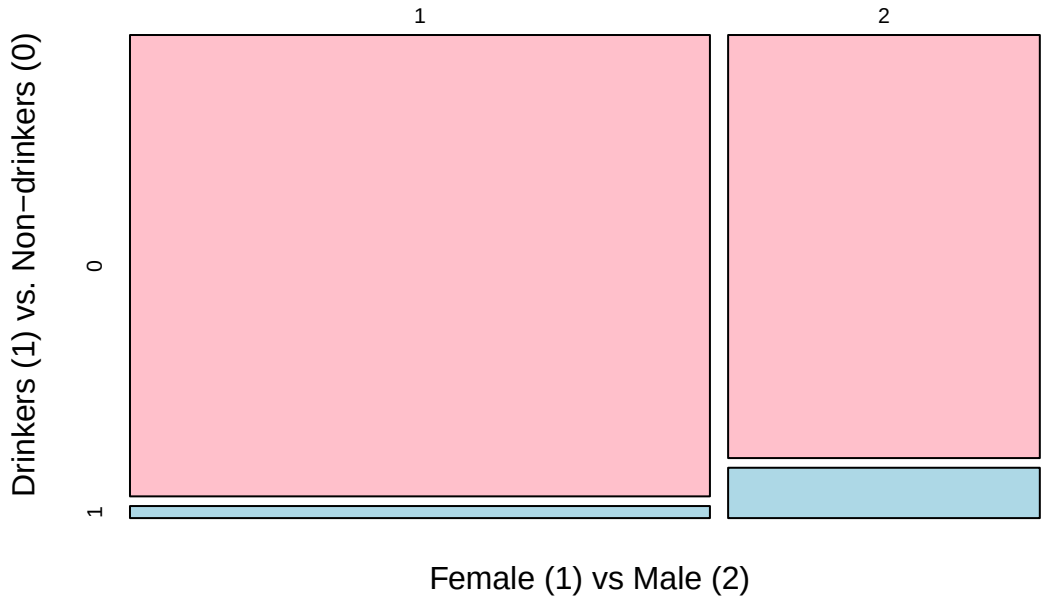


Mosaic Plot

Mosaic plots uses the relative sizes of rectangles to represent the relative frequencies of two or more categorical variables. Mosaic plots can use bivariate or multivariate data. To create a mosaic plot we use the use the mosaicplot() function provided in R.

For example, we can use a Mosaic plot to visualize the relationship between gender and alcohol consumption within our sample size.

```
mosaicplot(table(table$gender,table$alco), xlab = " Female (1) vs Male (2)",
ylab = "Drinkers (1) vs. Non-drinkers (0)", main = NULL, col=c("pink", "light blue"))
```



Numerical Summaries

Two-Way Table

The Two-way tables use bivariate data to summarize the relationship between two categorical variables. A two-way table can summarize the data as raw counts or as proportions.

For example, we can use a two-way table to compare the prevalence of heart disease in males versus the prevalence of heart disease in females (for this example we are using a two-way table that summarizes the data as proportions from the original sample).

```
gctable <- prop.table(table(table$gender,table$cardio))
colnames(gctable) <- c(" Doest Not Have Heart Disease",
                       "Has Heart Disease")
rownames(gctable) <- c("Female","Male")
gctable <- as.table(gctable)
gctable
```

```
##
##           Doest Not Have Heart Disease Has Heart Disease
##   Female                     0.3273429         0.3230857
##   Male                       0.1729571         0.1766143
```

Frequency Distribution

A frequency distribution is a table that summarizes that the frequency of categorical variables as raw counts.

For example, we can use a frequency distribution table to see how many individuals in our sample have heart disease, where 0 indicates does not have heart disease and 1 indicates has heart disease.

```
table(table$cardio)
```

```
##
##     0     1
## 35021 34979
```

Relative Frequency Distribution

A relative frequency distribution is a table that summarizes that the frequency of categorical variables as proportions.

For example, a frequency distribution would be useful to categorize the number of individuals with normal cholesterol, above normal cholesterol levels, and well above normal cholesterol levels. .

```
prop.table(table(table$cholesterol))
```

```
##
##         1         2         3
## 0.7483571 0.1364143 0.1152286
```

```
#normal cholesterol=1, above normal cholesterol levels=2, well above normal cholesterol levels=3
```

Mean

The mean represents the average of all values involved (i.e., it incorporates all the values from the distribution.)

For example, we can use the mean to find the average weight (kg) of the participants.

```
mean(table$weight)
```

```
## [1] 74.20569
```

Median

The median is the center-most value in the data, and is resistant to outliers and skewing.

We can use the median function to find the median age (in days) of the participants in the sample.

```
median(table$age)
```

```
## [1] 19703
```

Range The range is the maximum value subtracted by the minimum value. The range can be inflated by outliers.

We can use the range function to calculate the highest and lowest weight (kg).

```
range(table$weight)
```

```
## [1]  10 200
```

5-Number Summary

The 5-number summary is composed of: the minimum value, the first quartile (25% value), the median, the third quartile (75% value), and the maximum value.

The five number summary can be used to give a summary of the diastolic blood pressures (mmHg) of the participants.

```
summary(table$ap_lo)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   -70.00    80.00    80.00    96.63    90.00 11000.00
```

Interquartile Range (IQR)

The IQR is the difference between the third quartile and the first quartile from the four quartiles the distribution is divided into.

For example, we can use the IQR function to find the IQR of the participants' systolic blood pressure (mmHg).

```
IQR(table$ap_hi)
```

```
## [1] 20
```

Variance

Variance is the measure of dispersion, it indicates the distance between individuals and the mean, (the variance is most suitable for for distributions without major outliers).

For example, we can calculate the variance of the diastolic blood presure (mmHg) of the participants.

```
var(table$ap_lo)
```

```
## [1] 35521.89
```

Standard Deviation

The standard deviation is the square root of the variance.

For example, we can use the standard deviation function to to find the standard deviation of the diastolic blood pressure of the patients.

```
sd(table$ap_lo)
```

```
## [1] 188.4725
```

Analysis

Example 1

T confidence interval for mean

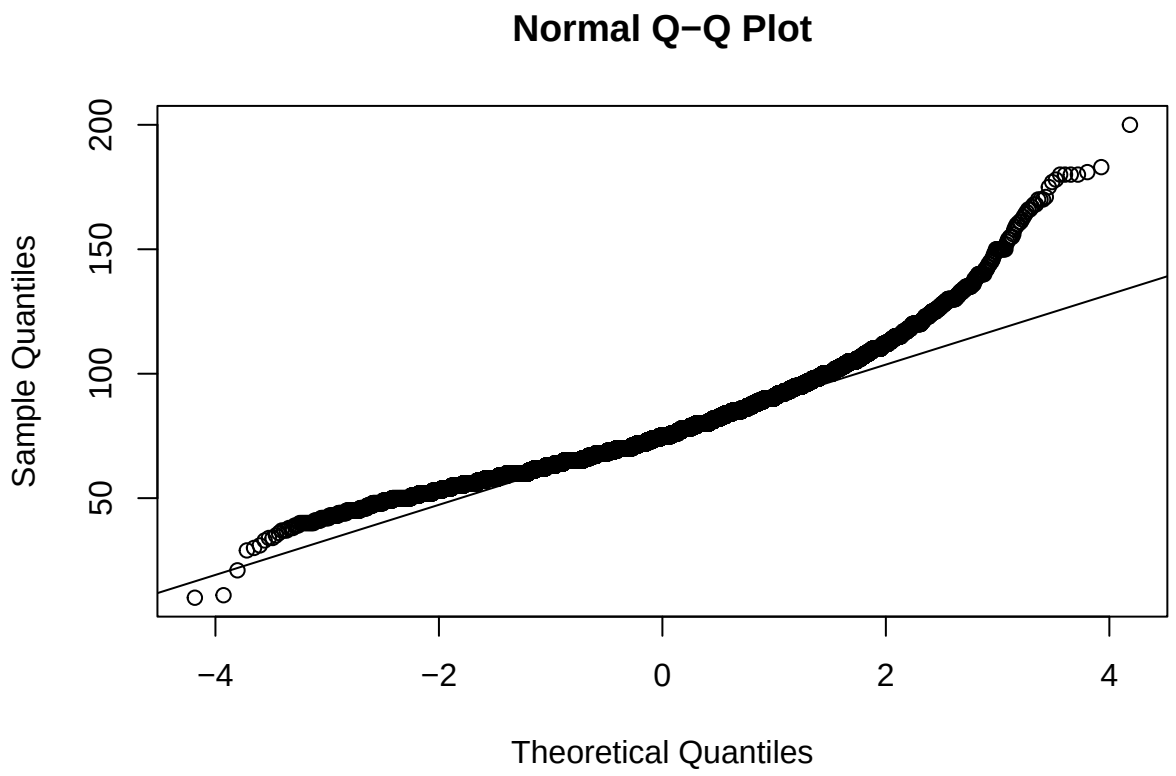
Research Question: What is the mean weight of an individual who has cardiovascular disease?

The T confidence interval for mean is used to estimate a population parameter using sample data. The parameter of interest is the mean weight.

Conditions 1.The sample must be obtained using a simple random sampling - In our study we utilized a stratified and random sampling study design. Additionally, our patients were contacted via email and asked to participate therefore there is non response bias, so our sample does not meet this criteria.

2. The second condition for a t confidence interval for mean is that the sample data must come from a normal distribution. The normality condition can be confirmed using a Quantile-Quantile Plot. The "weight" sample data for individuals with cardiovascular disease does not follow th patter of a normal distribution however our sample of 34979 individuals with heart disease is large enough(n>50) to assume it follows a normal distribution .

```
qqnorm(y=table$weight[table$cardio==1])
qqline(y=table$weight[table$cardio==1])
```




```
#determining sample size
length(table$weight[table$cardio==1])
```

```
## [1] 34979
```

```
# conducting t confidence interval for mean
t.test(table$weight[table$cardio==1], conf.level = 0.92)
```

```
##
##  One Sample t-test
##
## data:  table$weight[table$cardio == 1]
## t = 960.41, df = 34978, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 92 percent confidence interval:
##  76.68233 76.96241
## sample estimates:
## mean of x
##  76.82237
```

Example 2

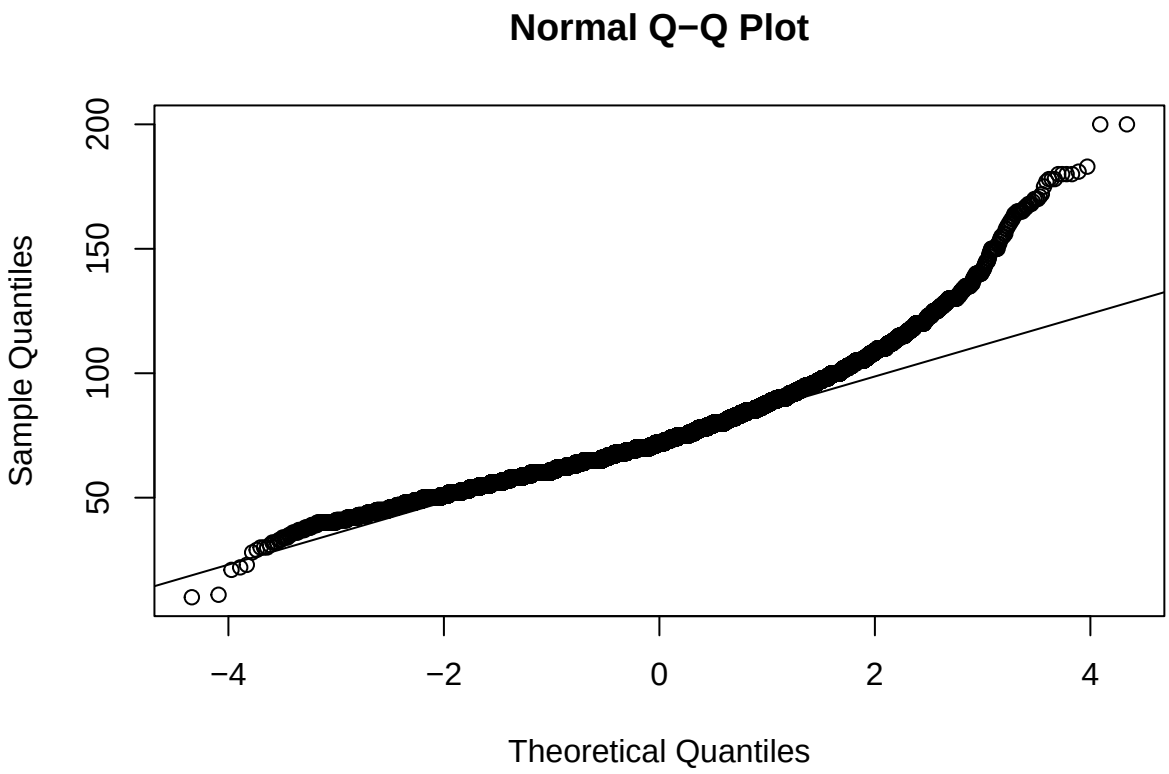
T Test for Difference in Means

Research Question: Is there a differnce between mean weitght for individuals with heart disease and individuals without heart disease?

Conditions: 1. Both samples must be obtained using a simple random sampling - In our study we utilized a stratified and random sampling study design. Additionally, our patients were contacted via email and asked to participate therefore there is non response bias, so our sample does not meet this criteria.

2. The samples must be independent This is true for our data as a repeated measures study was not used and the data is not paired.
3. The sample data must come from a normal distribution. The normality condition can be confirmed using a Quantile-Quantile Plot.The "weight" sample data for individuals with cardiovascular disease does not follow the patter of a normal distribution however both of our samples, individuals with heart disease(n=34979) and individuals without heart disease (n=35021), are large enough (n>50) to assume that it it follows a normal distribution.

```
qqnorm(table$weight)
qqline(table$weight)
```




```
table(table$cardio ==1)
```

```
##
## FALSE  TRUE
## 35021 34979
```

Statistical Hypotheses

Null Hypothesis: The weight of an individual does not impact the whether an individual gets heart disease.

$\#(H_0)$: $(mu_weightheartdisease)$ -$(mu_weightnoheartdisease) = 0$

Alternative Hypothesis: The weight of an individual has an impact on whether an individual gets heart disease. (HA) : μpresent - μnone = 0

```
Heartdisease <- (table$weight[table$cardio == "1"])
Noheartdisease <- (table$weight[table$cardio == "0"])
t.test(x=Heartdisease, y=Noheartdisease, mu=0, alternative = "two.sided", conf.level=0.92)
```

```
##
##  Welch Two Sample t-test
##
## data:  Heartdisease and Noheartdisease
## t = 48.872, df = 69038, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 92 percent confidence interval:
```

```
##  5.042857 5.417577
## sample estimates:
## mean of x mean of y
##  76.82237  71.59215
```

Example 3

Simple linear regression

Research Question : Is there a relationship between age and heart disease?

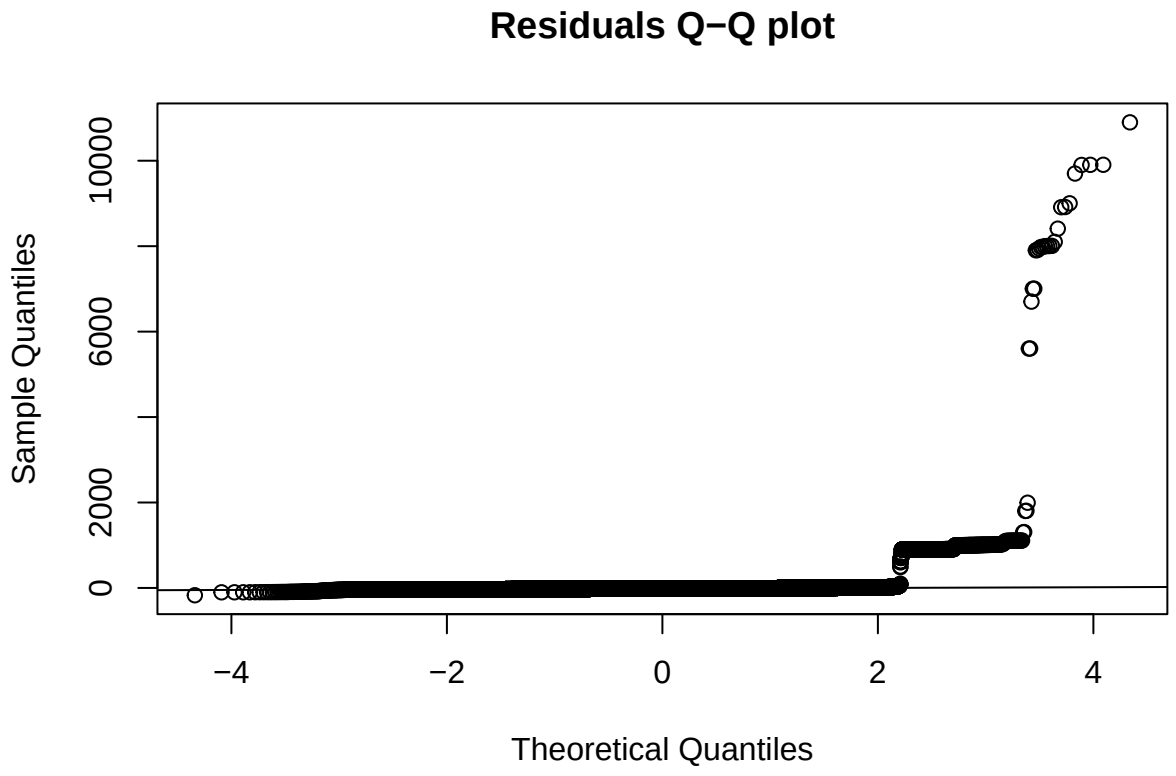
Conditions:

1.There must be a linear relationship between the explanatory (x), age, and response (y) variables, resting blood pressure. This condition is verified using a scatter plot and line of best fit.

2. The response variable observations (y) are independent. This condition is met as one individual's resting blood pressure does not impact another individual's blood pressure. Therefore, the response variables are independent.

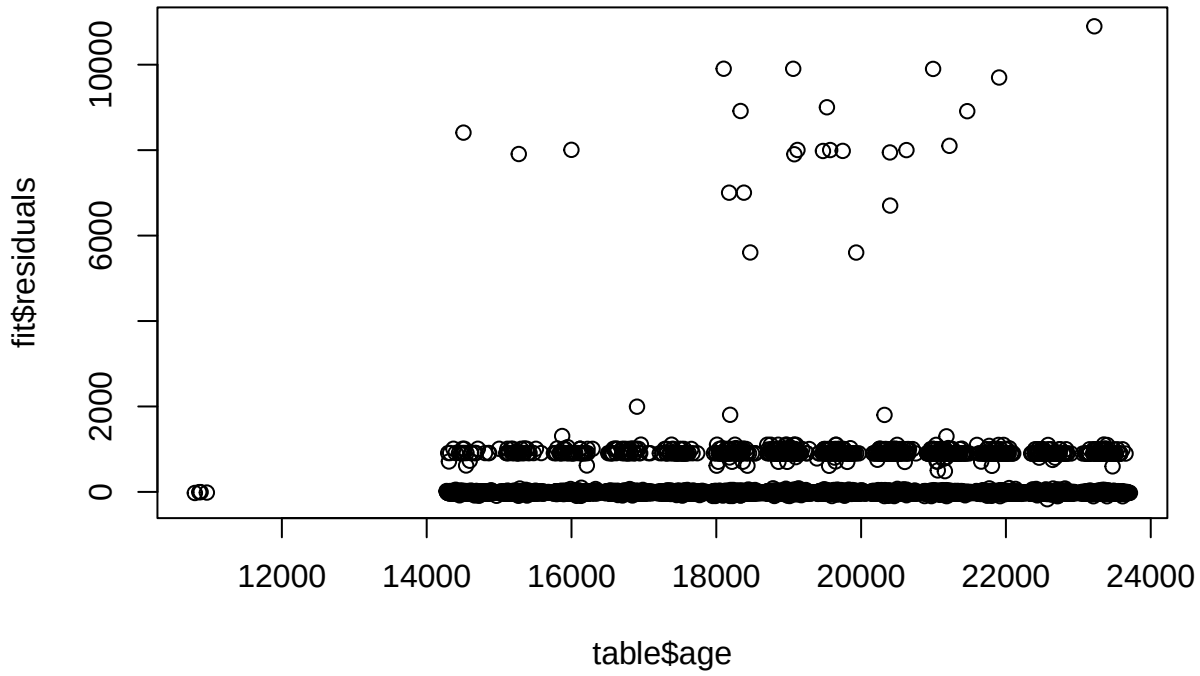
3. The next condition is that the population distribution of the response variable (systolic blood pressure) is normally distributed for each value of the explanatory variable (age). This can be confirmed using a QQ plot.

```
#Checking to see if the population distribution of y for each value of x is
#normally distributed by using a residuals QQ plot
qqnorm(lm(table$ap_lo~table$age)$residuals, main = "Residuals Q-Q plot")
qqline(lm(table$ap_lo~table$age)$residuals)
```

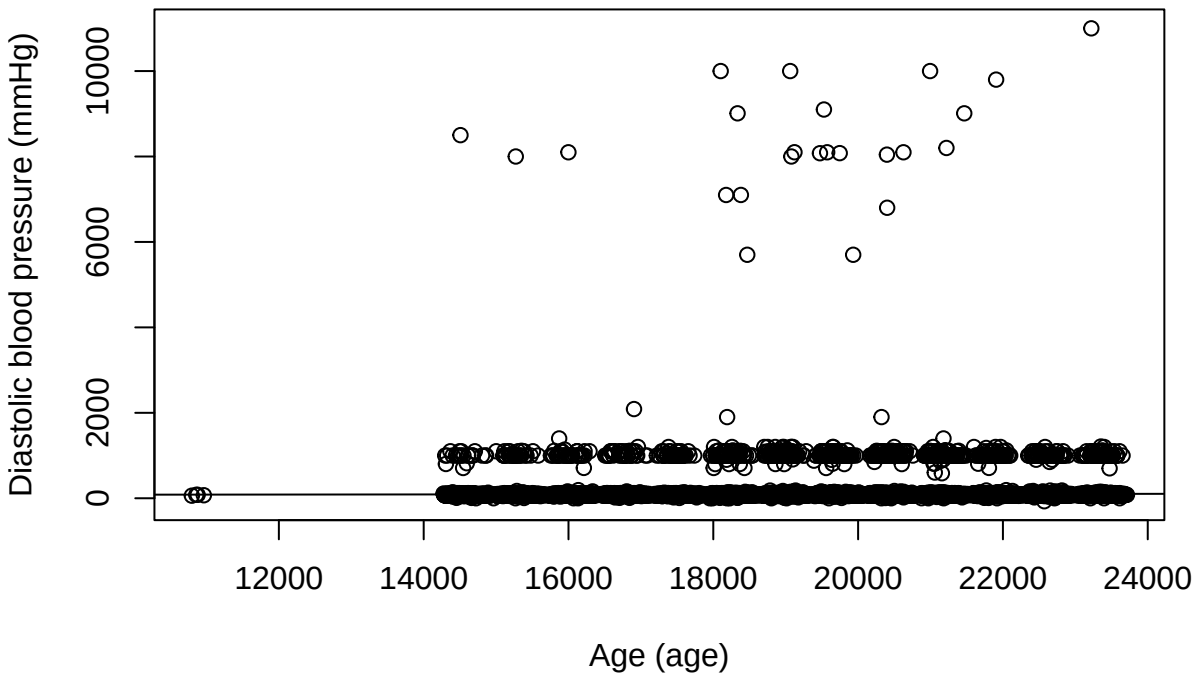


4. The standard deviation of the population distribution of response variable (y) is the same for all values of 'y' at 'x' - This condition can be proven through the Residual plot below. As long as the spread seems constant across the plot, this condition can be assumed to be met.

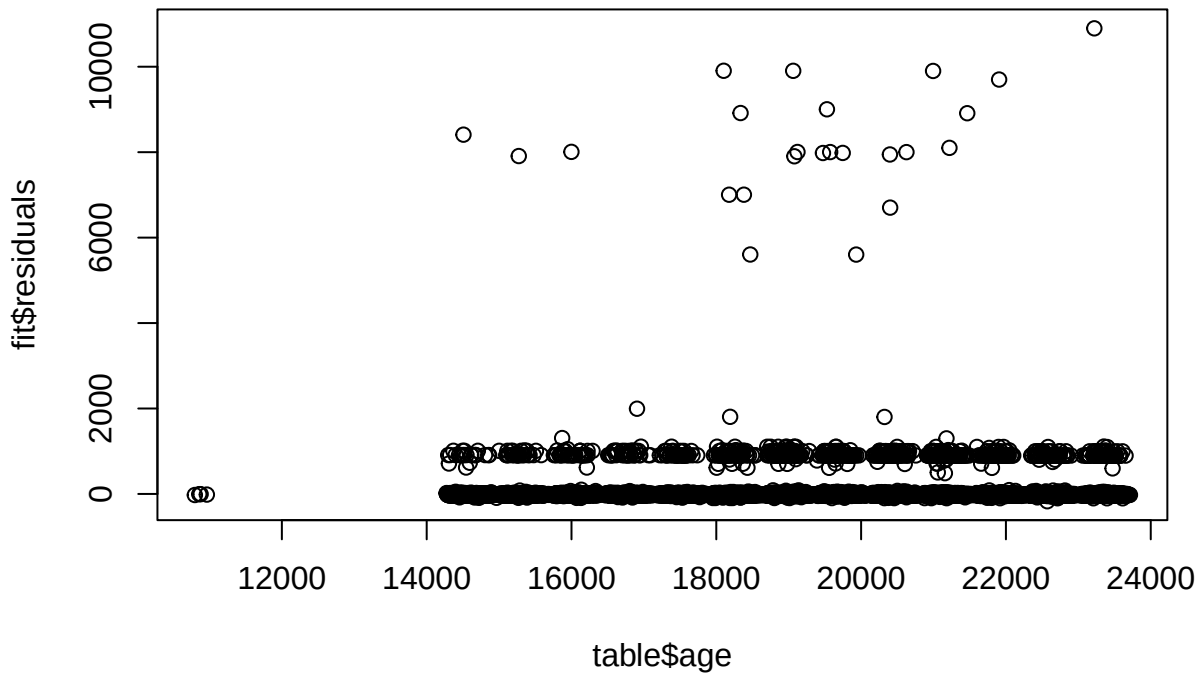
```
#Checking to see whether the sd of population distribution of y is the same
#at all y at x using a residual plot
#Define new variable for use in conditions
fit <- lm(table$ap_lo~table$age)
plot(table$age,fit$residuals)
```




```
# Checking if relationship between age (in days) and heart disease is  linear
plot(table$ap_lo~table$age, xlab = "Age (age)", ylab = "Diastolic blood pressure (mmHg)")
abline(lm(table$ap_lo~table$age))
```




```
#Checking to see whether the sd of population distribution of y is the same
#at all y at x using a residual plot
#Define new variable for use in conditions
fit <- lm(table$ap_lo~table$age)
plot(table$age,fit$residuals)
```



• Large sample confidence interval for proportion • T test for mean • Large sample test for proportion • T confidence interval for the difference in means • T test for the difference in means • Large sample confidence interval for difference in proportions • Large sample test for differences in proportions • T test for slope • One-factor ANOVA plus follow up analyses

Example 4

Large sample confidence interval for proportion

Research Question: What proportion of the population consumes alcohol?

Conditions: 1.The sample data must be obtained using a simple random sampling - In our study we utilized a stratified and random sampling study design. Additionally, our patients were contacted via email and asked to participate therefore there is non response bias, so our sample does not meet this criteria.

2.The counts of succeses (ie individuals who drink) is described by a binomial modelsample data

3.A normal approximation of binomial is reasonable. np>=15 and nq>=15

To conduct a Large sample confidence interval for proportion use the prop.test( ) function. Prop.test is used when trying to calculate the confidence interval for a large sample, where n= number of individuals in sample and x = number of successes (as integers). In this example it would be number of individuals who consume alcohol. The conf.level argument is the confidence level of the returned confidence interval, this number has to be between the number 0 and 1 as it is a percentage written in decimal form.

Example 5

T test for mean

Research Question: Does mean diastolic influence whether an individual develop heart disease?

Conditions:

1.The sample data must be obtained using a simple random sampling. In our study we utilized a stratified and random sampling study design. Additionally, our patients were contacted via email and asked to participate therefore there is non response bias, so our sample does not meet this criteria.

2. The "ap_hi" sample data of individuals with heart disease must come from a normal distribution. The normality condition can be confirmed using a QQ Plot.

t.test() function is used for the T test for mean. The x arguement in the t.test function represents a numeric vector of data, which in this case would be the corelation between the diastolic influence. The conf.level arguement on this funnction represents the confidence level of the returned confidence interval, the value in this arguement has to be within the range of 0 to 1 as it is a percentage written in decimal form.

$$\mu$$

(mu) as the arguement takes in the value of the mean of the dataset.

Example 6

Large sample test for proportion

Reasearch question: What proportion of males are smokers? Conditions 1.The sample must be obtained using a simple random sampling. In our study we utilized a stratified and random sampling study design. Additionally, our patients were contacted via email and asked to participate therefore there is non response bias, so our sample does not meet this criteria.

2. The sample data is to follow a binomial distribution.This is true for our study as there is a fixed numher of observations, n=7000. Each observation can be determined as a sucess or failure: a male who smokes or a male who does not smoke. The observations are independent from one another, an individual being a male smoker does not impact if another individual is a male smoker. The probability of each male individual being a smoker is constant.

3. A normal approximation of the binomial model is reasonable. To confirm this condition a QQ plot can be used or we can assume our population follows a normal model as it is large n>5o and the number of sucesses is greater than 15 and the number of failures is greater than 15. (check these numbers later)

Prop.test() function is used for Large sample test for proportion. The arguements used for the function prop.test are x, n, and conf.level. The x arguement takes in all "successes" are the male individuals who smoke. The n arguement is used for our sample are all male individuals. Conf.level takes in the confidence level of the returned confidence interval, this arguement has to be within the range of 0 to 1 ass it is a percentage written in decimal form.

Example 7 **T confidence interval for the difference in means.** Research question: Do males or females have a higher mean weight? Conditions:

1. Both samples must be obtained using a simple random sampling. In our study we utilized a stratified and random sampling study design. Additionally, our patients were contacted via email and asked to participate therefore there is non response bias, so our sample does not meet this criteria.

2. The samples must be independent This is true for our data as a repeated measures study was not used and the data is not paired.

3. The "weight" sample data must come from a normal distribution. The normality condition can be confirmed using a QQ Plot. Additionally our sample is large enough (n>50) to assume that it it follows a normal distribution.

t.test() function is used for T confidence interval for the difference in means. The x argument in the t.test function represents a numeric vector of data, which in this case would be the correlation between higher mean weight between male and female. The conf.level argument on this function represents the confidence level of the returned confidence interval, the value in this argument has to be within the range of 0 to 1 as it is a percentage written in decimal form.

$$\mu$$

(mu) as the argument takes in the value of the mean of the dataset.

Example 8 **Large sample confidence interval for difference in means**

Research Question: Does age influence the risk of heart disease? Conditions: 1.The sample must be obtained using a simple random sampling. In our study we utilized a stratified and random sampling study design. Additionally, our patients were contacted via email and asked to participate therefore there is non response bias, so our sample does not meet this criteria.

2. The sample data is to follow a binomial distribution.This is true for our study as there is a fixed number of observations, n=7000. Each observation can be determined as a success or failure: a male who smokes or a male who does not smoke. The observations are independent from one another, an individual being a male smoker does not impact if another individual is a male smoker. The probability of each male individual being a smoker is constant.

3. A normal approximation of the binomial model is reasonable. To confirm this condition a QQ plot can be used or we can assume our population follows a normal model as it is large n>5o and the number of successes is greater than 15 and the number of failures is greater than 15. (check these numbers later (fix this)

prop.test() function is used for Large sample confidence interval for difference in means.

Example 9 **Large sample test for differences in proportions.**

prop.test() function is used for Large sample confidence interval for difference in proportions.

Example 10 One factor ANOVA plus follow up analyses

Conclusion

Hints and Reminders when using R

1. Use Logical Operators when trying to manipulate data in R, it is important to use certain logical operators such as "!"(Not), "=="(Equals), ">="(Greater Than or Equals To), "<="(Smaller Than or Equals To), ">"(Greater Than), "<"(Smaller Than) to be able to compare data and datasets. Logical Operators can greatly help the user to be able to compare and sort data in a meaningful way.

2. A helpful hint when working with R studio is to add comments using "#" because it is a good way to explain what you are doing on a given line of code. You can also use comments to help organize your thoughts and jot down any notes or reminders you want to make for your reference while working. In addition, comments are useful when you are comparing two similar lines of code, as you can simply comment one out and run the other one to see how the results differ.

3. Utilizing the knitting feature when creating a new graph or equation is beneficial as it is easier to deal with problems and errors as they occur. It is better and more efficient to knit the Rmd file more frequently as it helps save time, rather than knitting once at the end when correcting all the errors at once will become overwhelming and tedious.

4. Using symbols in R can be handy; however, for them to be used, it is confusing to put/use them in the R markdown file. Moreover, to add symbols in the R markdown, LaTex (a software system for document preparation that allows for mathematical and scientific symbols) must be used. R needs to know that we will be using LaTex symbols, and for that to happen, we need to add a '$' at the beginning and end of the expression, i.e. *expression*. Example 1: (H_A) represents the symbol for the Alternative Hypothesis. Example 2: (Δ) represents the symbol for Delta. Examining example 2, we can see a backslash before the expression, which allows the word (Delta) to become a mathematical/scientific symbol. Further, it can be used for a variety of symbols by substituting the term "Delta" with another word for the symbol of interest.

5.sz Phase 3 necessitates the frequent use of inference processes, which includes determining if the data at hand meets the requirements required for these inference methods to be used. It's important to remember that while it's ideal if the data matches the assumptions, it's not the end if they don't. It is preferable to express unequivocally that the data does not correspond to the conditions, rather than manipulating or playing with the data until it complies with the inference procedure' conditions.

Group Reflection

For phase 3, our group underestimated the time and focus required to organize and prepare all aspects of the research process while working on this project. Because of the phased structure, needing to account for several various requirements before realizing the significance and relevance of these standards for the latter portions of the project meant that a lot of little but crucial things were brushed over in the beginning. The impact of the planning process on the rest of the stages of research in a research project was demonstrated by having a more thorough and concrete understanding of the research goal and the tools of analysis allows for a much smoother transition between the progressive stages of research, as demonstrated by the PPDAC framework. By the end of the project, we were able to apply and see how many of these concepts work in tandem to reach meaningful conclusions with the data we were working with, thanks to our prior experiences (phase 1 and 2) working with statistical concepts and using R chunks to create numerical and graphical summaries to derive simple conclusions.

With a new point of view, we're able to comprehend how to practically apply many of the principles covered in class to a scenario that may happen to any of us in the workplace in the future. For example, it is arguable that R's greatest strength is its ability to perform a wide range of operations and arguments, allowing us to extract a wide range of information from the data at hand. Furthermore, unlike Excel or Sheets, which only allow for numbers in their cells, rather than a combination of numbers and text, the R Markdown format allows for the compilation of both numerical information from datasets and extensive chunks of text (like conclusions) into one document. The fact that our group has a diverse range of potential vocations (from medicine to computer science) might lead one to expect that R would be unimportant in some cases and essential in others. However, we all believe that the elementary understanding of R that we have gained during this project will be useful in our future jobs. For a start, we believe that most of us (60% chance) will replace Excel or Sheets as our primary data analysis tool in favour of R and its extensive set of capabilities.

DataSet Information

The following data represents the variable names, descriptions, and types of variables in the cardio_train.csv file.

Variable Names | Description | Type of Variable | | ID | ID number of the participant | Categorical, Nominal Age | Individuals' ages in days | Quantitative, Discrete
Gender | Individuals' gender (1 = women, 2 = men) | Categorical, Nominal Height | Individuals' height in centimeters | Quantitative, Discrete Weight | Individuals' weight in kilograms | Quantitative, Discrete Ap_hi | Individuals' systolic blood pressure | Quantitative, Discrete Ap_lo| Individuals' diastolic blood pressure | Quantitative, Discrete Cholesterol | Cholesterol level: 1 = normal, 2 = above normal, 3 = well above normal | Categorical, Ordinal Gluc | Glucose level: 1 = normal, 2 = above normal, 3 = well above normal | Categorical, Ordinal Smoke | Whether individuals have smokes or not (1 = yes, 0 = no) | Categorical, Nominal Alco | Whether the individual consumes alcohol or not (1 = yes, 0 = no) | Categorical, Nominal Active | Whether the individual does Physical Activities (1 = yes, 0 = no) | Categorical, Nominal Cardio | The Presence or Absence of cardiovascular diseases (1 = yes, 0 = no) | Categorical, Nominal

Number of columns: 13

Number of rows: 70,000

The first row is the header

Dataset Reference

Ulianova, S. (2019, January 20). Cardiovascular disease dataset. Kaggle. Retrieved February 17, 2022, from https://www.kaggle.com/sulianova/cardiovascular-disease-dataset