Final Project

Part 1

Introduction

Education is the most powerful weapon which one can use to change the world. It is the core domain of almost each and everything that exists in this world.

The dataset “Students Performance in Exams” is obtained from Kaggle. This is an educational dataset and contains information about the marks scored by high school students in the United States. It consists of information about the students’ marks in Math, Writing, and Reading. The other factors included in the dataset are students’ gender, race or ethnicity, their parental level of education, lunch and test preparation course status.

Some of the initial observations of the dataset are:

1. There are more females in our sample than males.
2. The score for the three tests: Math, Reading, and Writing are on a scale of 0 to 100.
3. There are more students from race ‘Group C’ in the dataset.
4. Most of the students have not completed the test preparation course.
5. More than half of the students avail standard lunch at school.
6. Parents of very few students have a master’s level of education.

This study aims at identifying the factors and the extent to which these factors affect the academic performance of high school students. The dataset has a total of 8 features which covers the following categories:

1. Demographic features: Race/ethnicity and gender
2. Academics: Parental level of education, students score in Math, Writing and Reading, test preparation course status

Through this study, I will try to explore a lot of different questions that can help us understand this dataset and bring some useful insights for the readers. The questions covered in our study are:

1. What is the average reading score of high school students in the United States?
2. What is the proportion of females in high school in the United States?
3. Is there a difference in the average math scores between the students who complete the test preparation course and those who don’t?
4. Does the proportion of males who avail standard lunch at school differ from the proportion of females who avail standard lunch at school?
5. Do the high schools in the United States have students of all races in equal proportion?

In order to solve these questions I have used the following testing strategies and found results using both the traditional statistical method and bootstrapping method:

1. One sample t-test
2. One sample test of proportion

3. Two sample t-test for difference in means
4. Two sample test for difference in proportions
5. Chi-square goodness of fit test

After researching a lot about this dataset, I found some sources which claim that the data set is fictional and has been made for the educational purpose. Many studies have been done using this dataset to find out how effective the test preparation course is, what are the major factors that contribute to the test outcomes and what can be some of the best ways to improve the scores of students in these tests.

It seems like stratified sampling has been done in order to collect the data. In stratified sampling, the population is generally divided into groups called strata by using some characteristics. Here, the strata seem to be gender: females and males. It might be possible that the sample is taken from each of these strata using either random sampling or systematic sampling or convenience sampling.

Despite having an approximately equal number of males and females in the dataset, there exist few concerns. The collected data has more students from race ‘Group C’ and has very few students who have completed the test preparation course. Also, there does not seem to be much difference between the summary of scores of students in Math, Reading and Writing which is giving a feeling that either the sampling strategy was not good or the data had been manipulated to bring good statistical results. Also, it is unknown whether the dataset belongs to one high school or has been collected from different high schools of different regions of the United States.

Information about the variables involved in the dataset:

1. Gender - Catgeorcial variable - Nominal scale
2. Race/Ethnicity - Catgeorcial variable - Nominal scale
3. Parental level of education - Catgeorcial variable - Nominal scale
4. Lunch - Catgeorcial variable - Nominal scale
5. Test preparation course - Catgeorcial variable - Nominal scale
6. Math score - Quantitiave variable (Discrete) - Interval scale
7. Reading score - Quantitiave variable (Discrete) - Interval scale
8. Writing score - Quantitiave variable (Discrete) - Interval scale

The dataset is of interest to me because I have a subject matter expertise in the education sector. I have worked for more than 7 years in the education domain for various edutech companies and during my tenure, I have explored curriculums of various countries including India, Singapore, and the United States. Education, its techniques, methodologies, and measures are close to my heart and I always remain willing to find out interesting insights from various studies that involve education as the main subject. It would be really fascinating for me to know the trends of high school students in the United States based on many factors.

Part 2

Exploratory Analysis

Let us load our CSV file to see the summary of our dataset.

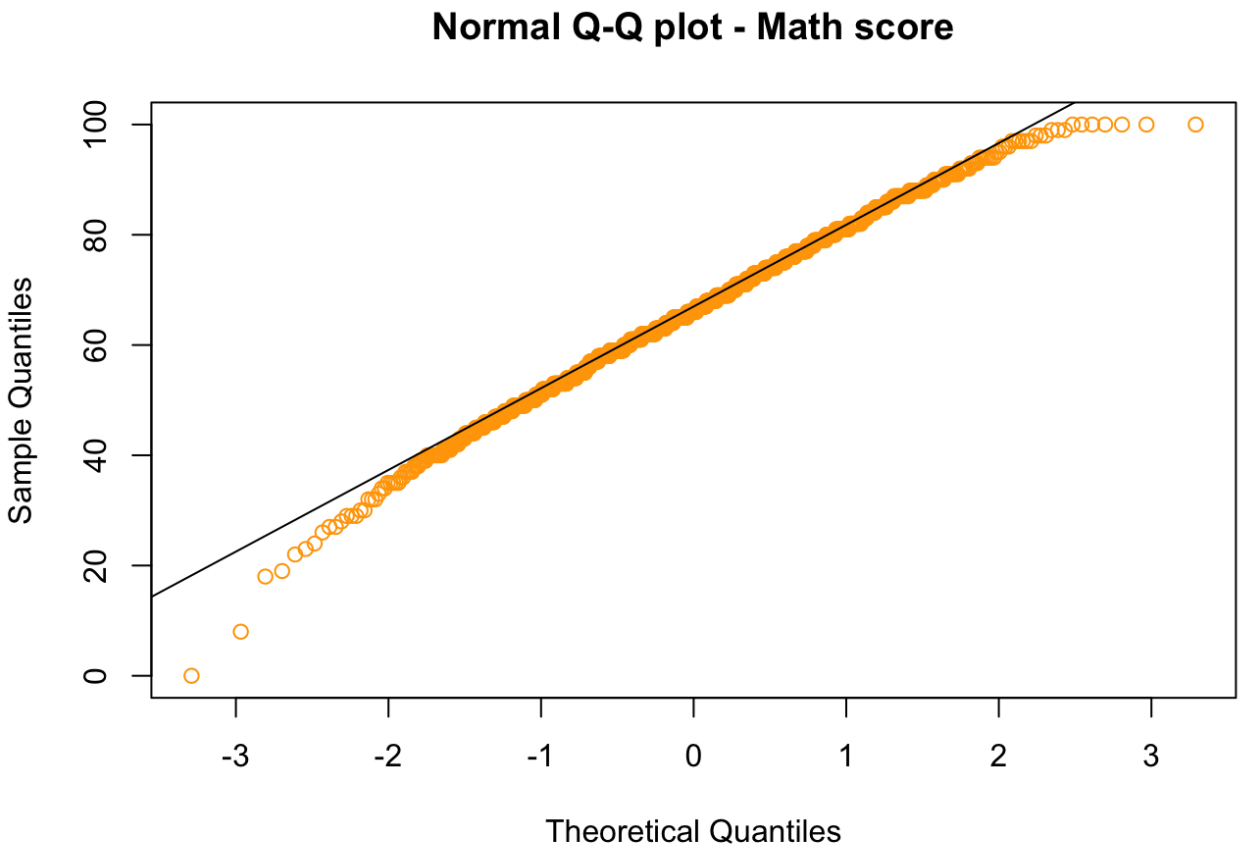
```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master's degree     standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate's degree     standard
##   test.preparation.course math.score reading.score writing.score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

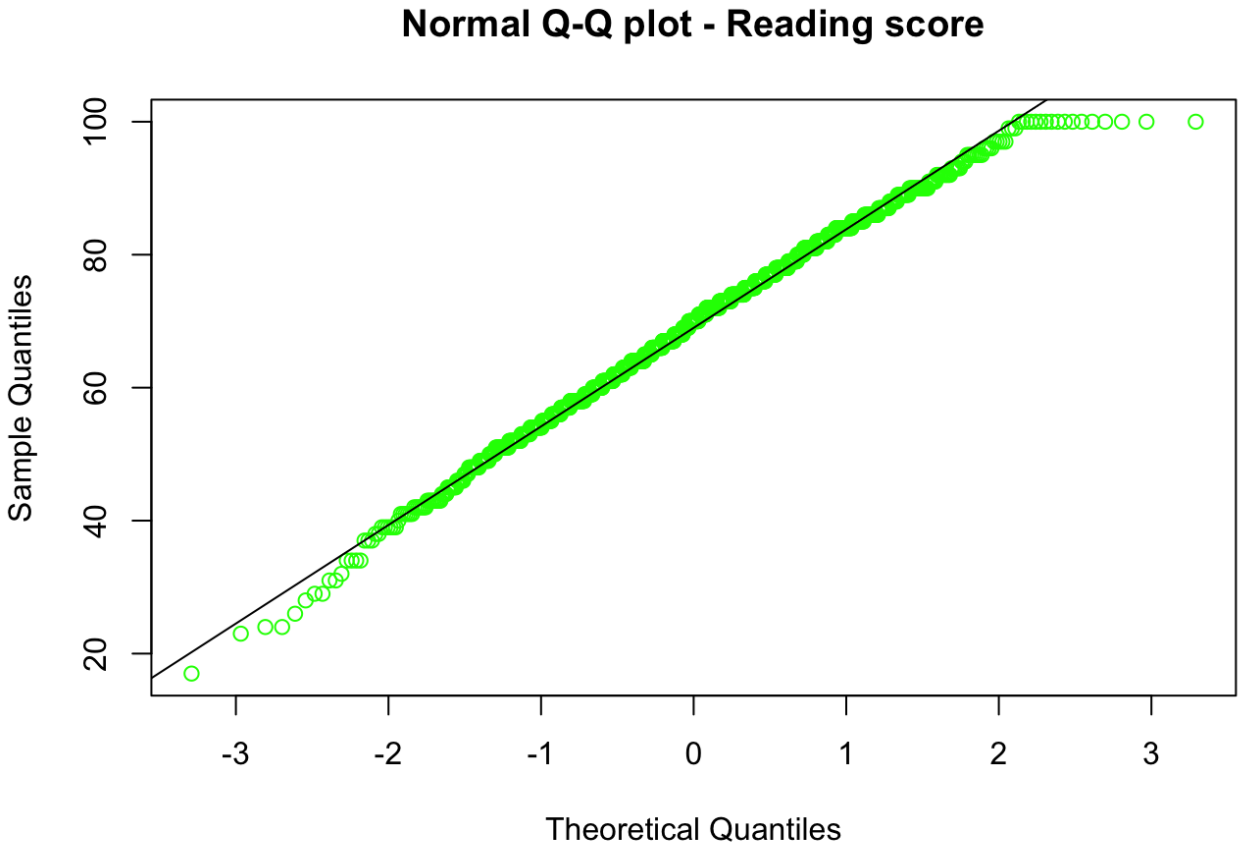
```
##     gender    race.ethnicity     parental.level.of.education          lunch
##  female:518   group A: 89    associate's degree:222          free/reduced:355
##  male  :482   group B:190    bachelor's degree :118          standard    :645
##               group C:319    high school       :196
##               group D:262    master's degree   : 59
##               group E:140    some college      :226
##                              some high school  :179
##  test.preparation.course   math.score     reading.score    writing.score
##  completed:358           Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
##  none     :642           1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
##                          Median : 66.00   Median : 70.00   Median : 69.00
##                          Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
##                          3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
##                          Max.   :100.00   Max.   :100.00   Max.   :100.00
```

Let us try to understand the relationship between the various categorical and quantitative variables present in our dataset by doing some exploratory analysis.

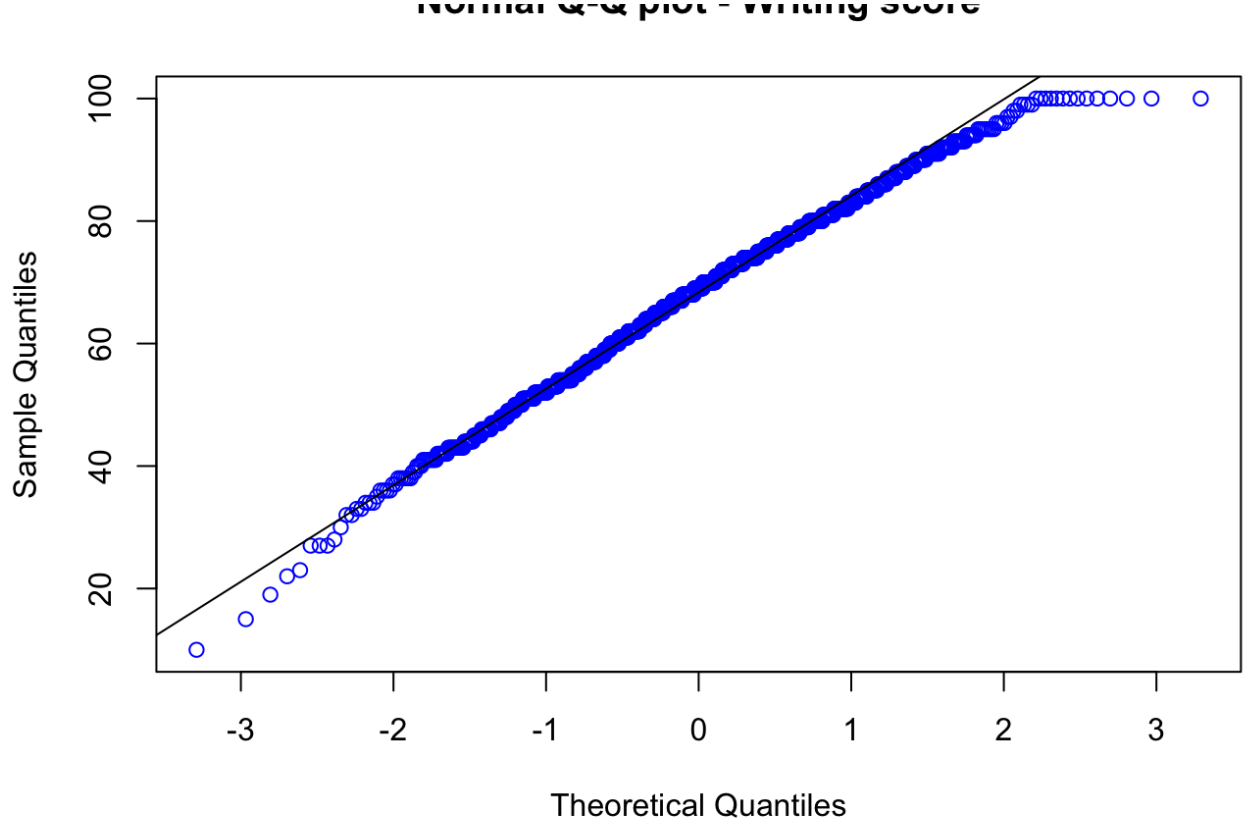
In order to do our statistical analysis, it is mandatory to check whether the population data is normally distributed or not. If the data is normally distributed, the t-test can be used as the statistical method for one-sample test of mean and two-sample test of difference in means in order to perform our analysis as it is already known that we do not have population variance.

Let us first try to understand our sample by plotting the Q-Q plot.



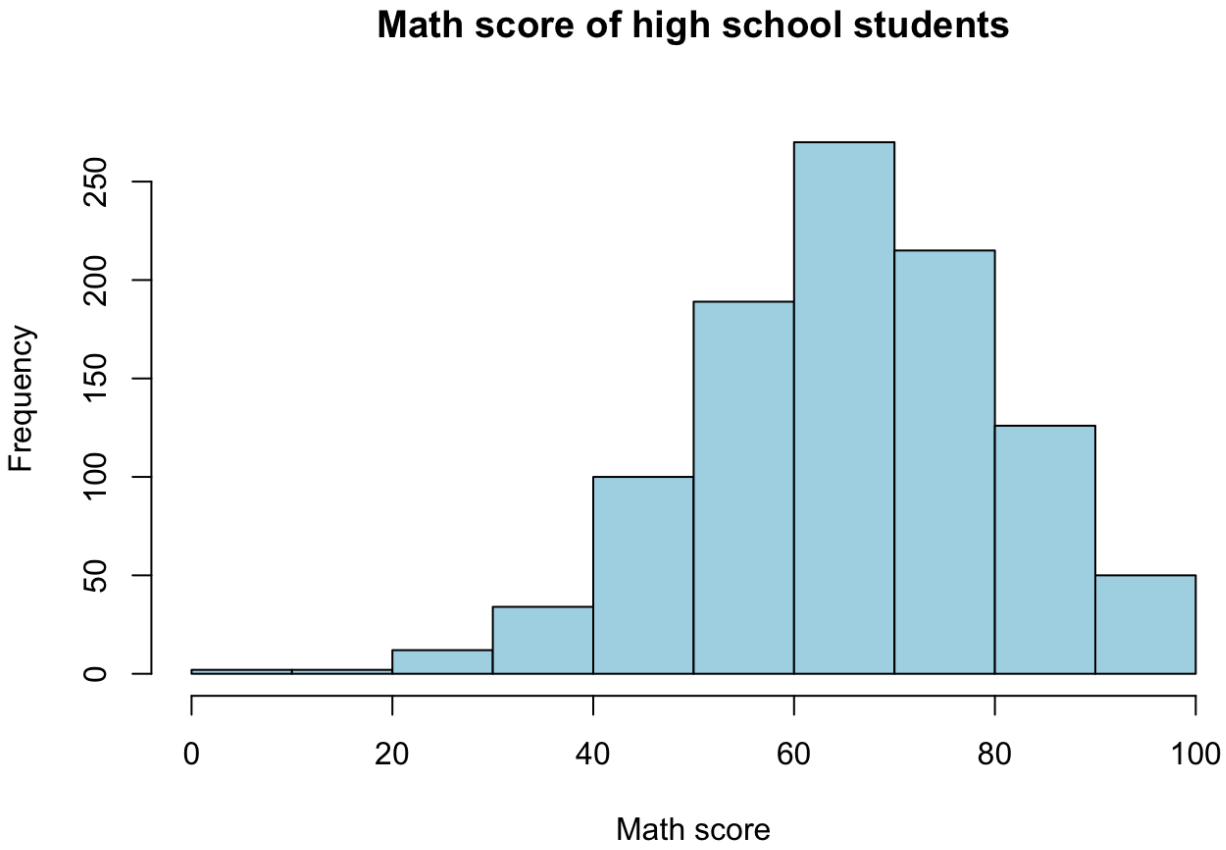


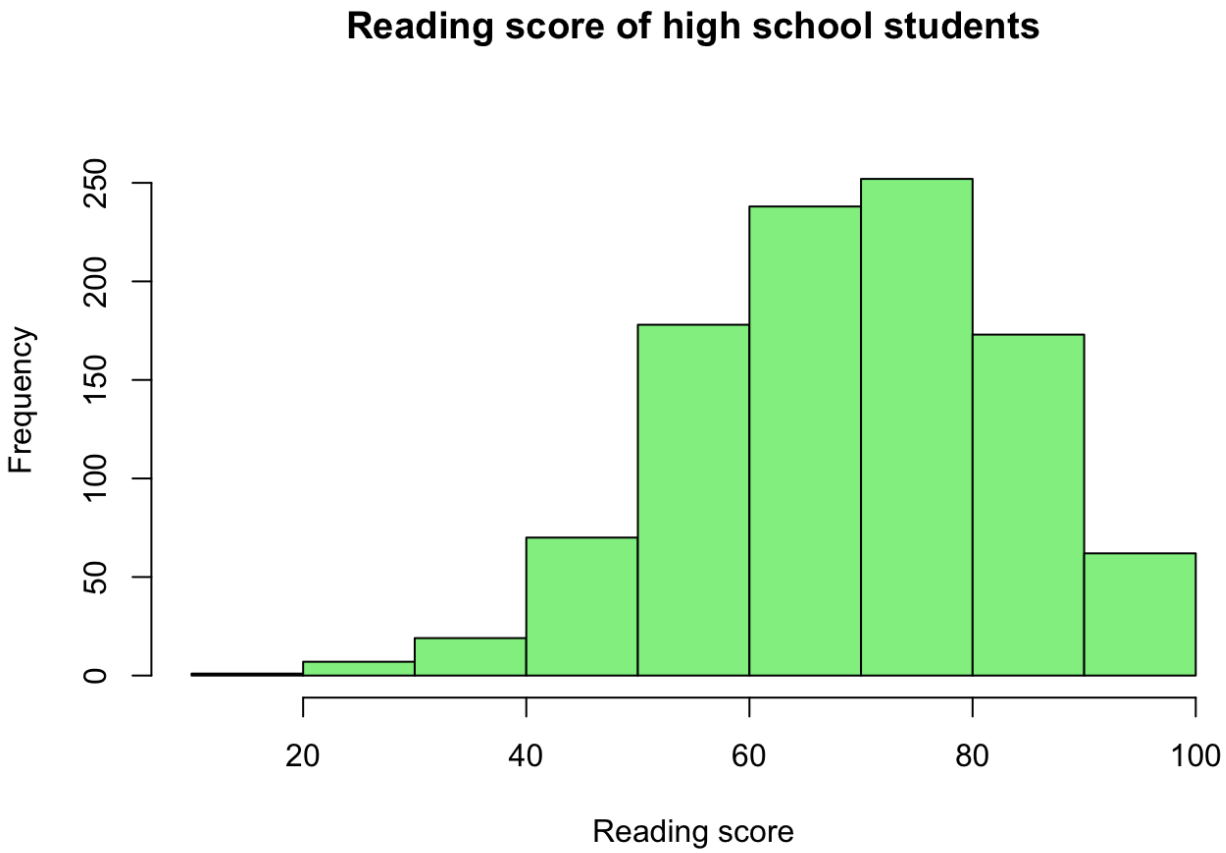
Normal Q-Q plot - Writing score



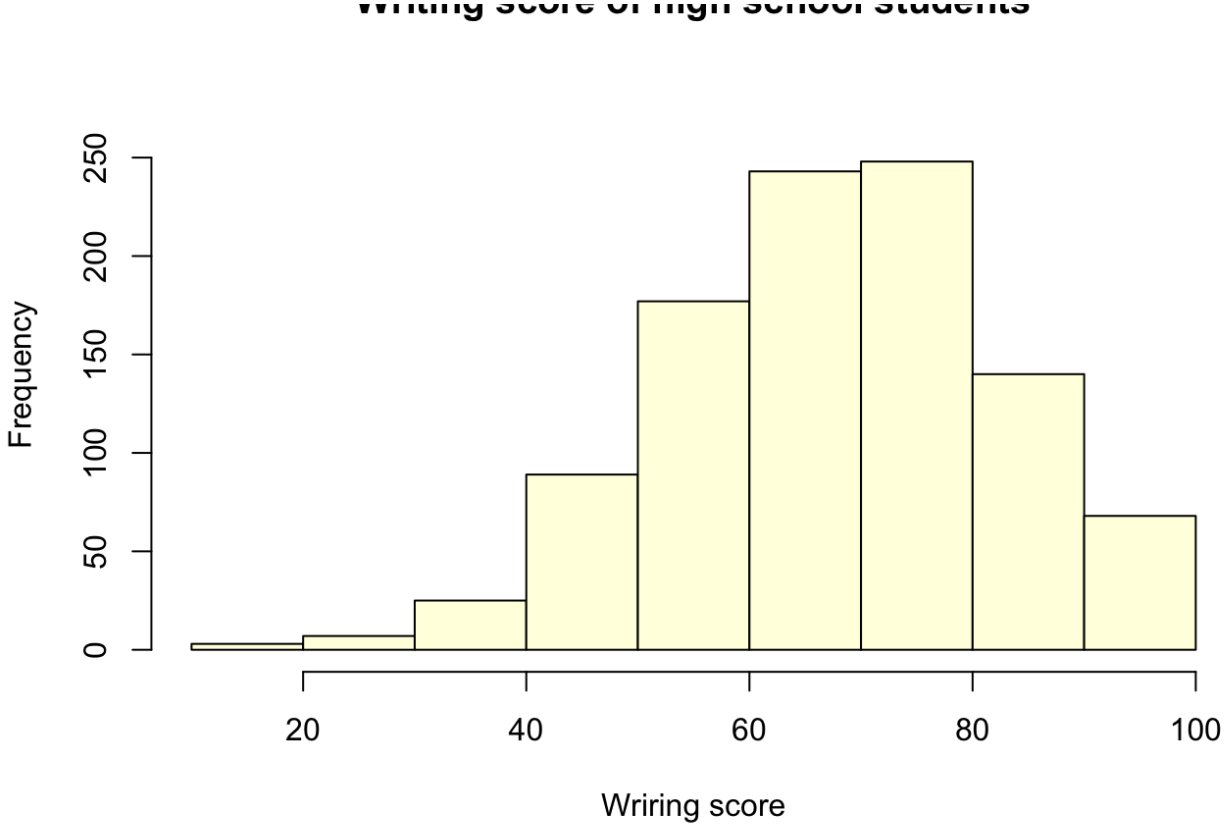
From the graph, it can be said that all the quantitative data falls along the y = x line with most points concentrated in the center of the line and fewer points towards the ends. The Q-Q plot shows us that the normality is probably a reasonable approximation for our dataset. It can also be observed that there are few outliers in the dataset but these are considerably small as compared to the size of the dataset. Let will try to find more about this through some other graphs.

A histogram is a graphical representation of frequency distribution. It can also be used to analyze whether our distribution is normal or not. Let us plot some histograms as well.



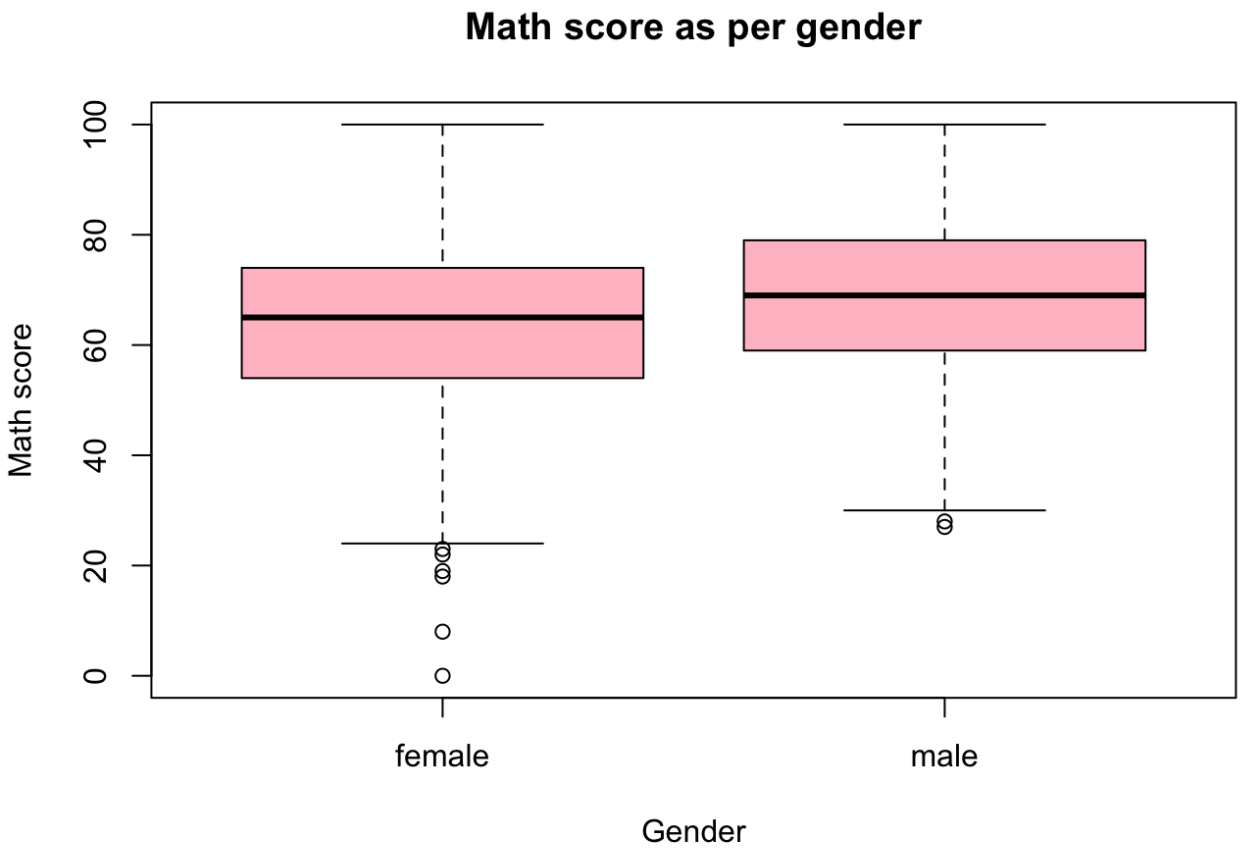


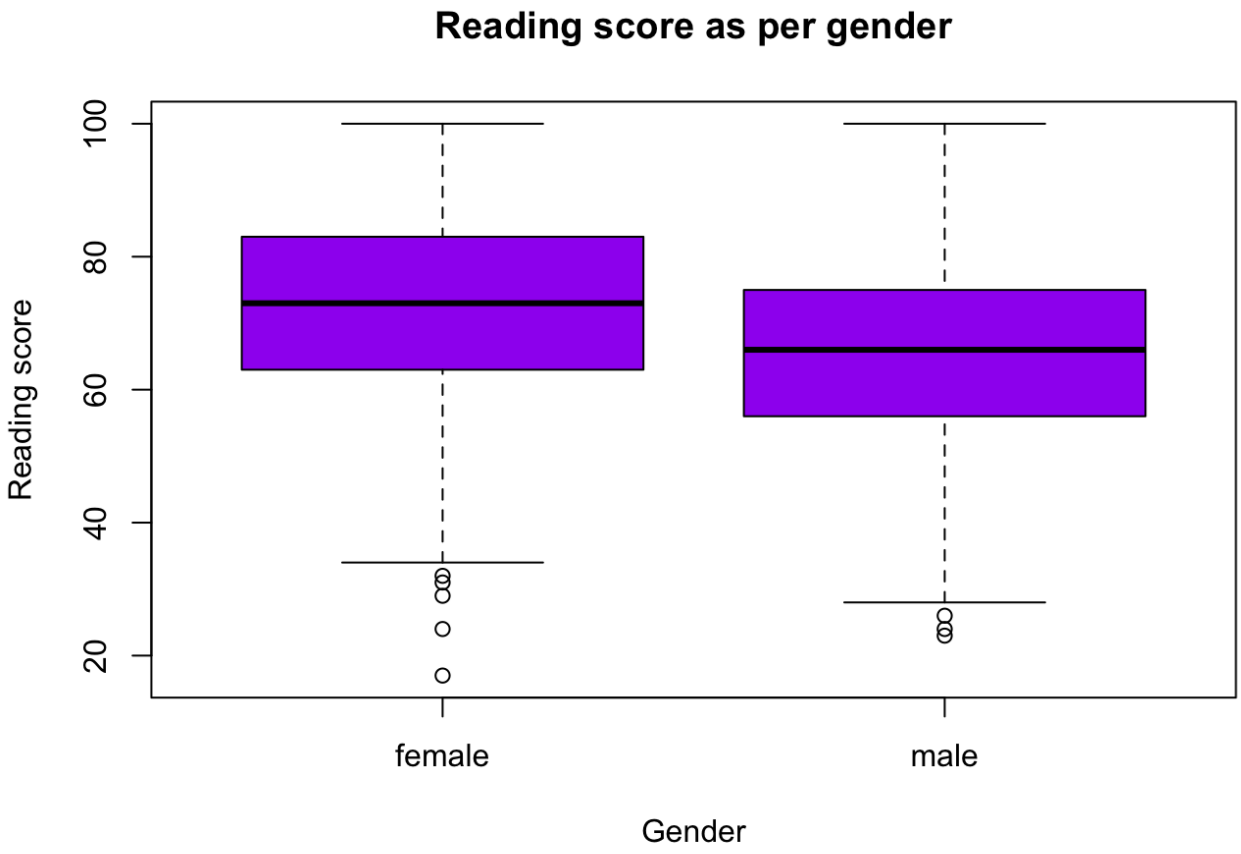
Writing score of high school students



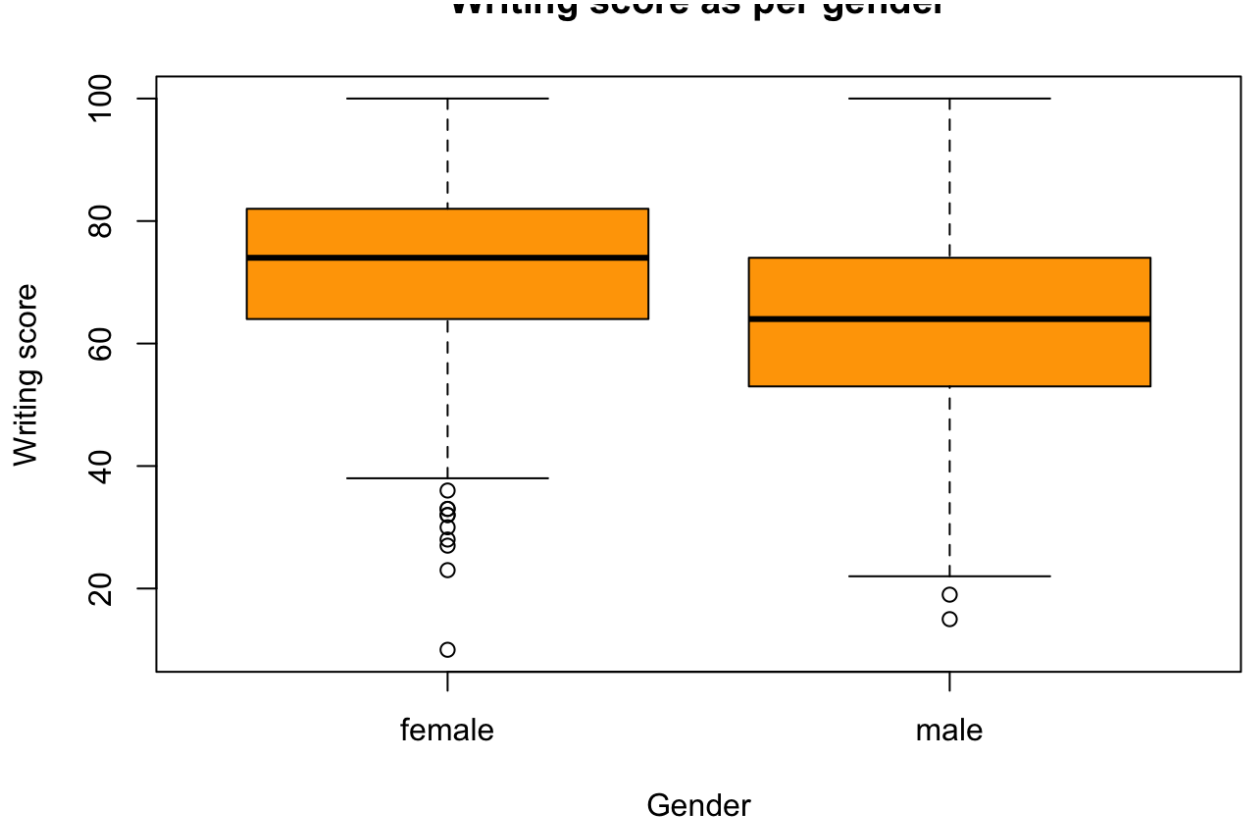
The approximate bell shape of our histogram depicts that the data is approximately normally distributed. The peak represents the highest values of the data.

Let us explore the categorical variables in our dataset. Let us try to find if there exists any relationship between the gender of the student and his/her score in various tests.





Writing score as per gender



It can be interpreted from the box plots that males on an average score higher than females in math. However, females on an average score higher than males in both reading and writing. It can also be observed that there are a few outliers present in our dataset and more of the outliers exist for females as compared to males. These boxplots also tell us the median marks for both males and females in each test. It can be said that there do exist a difference between the median marks of males and females in each test.

Let us plot a barchart to understand how the test preparation course and gender are related.

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: lattice
```

```
## Loading required package: ggformula
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggstance
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

```
## Loading required package: mosaicData
```

```
## Loading required package: Matrix
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected b
y this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.
```

```
##
## Attaching package: 'mosaic'
```

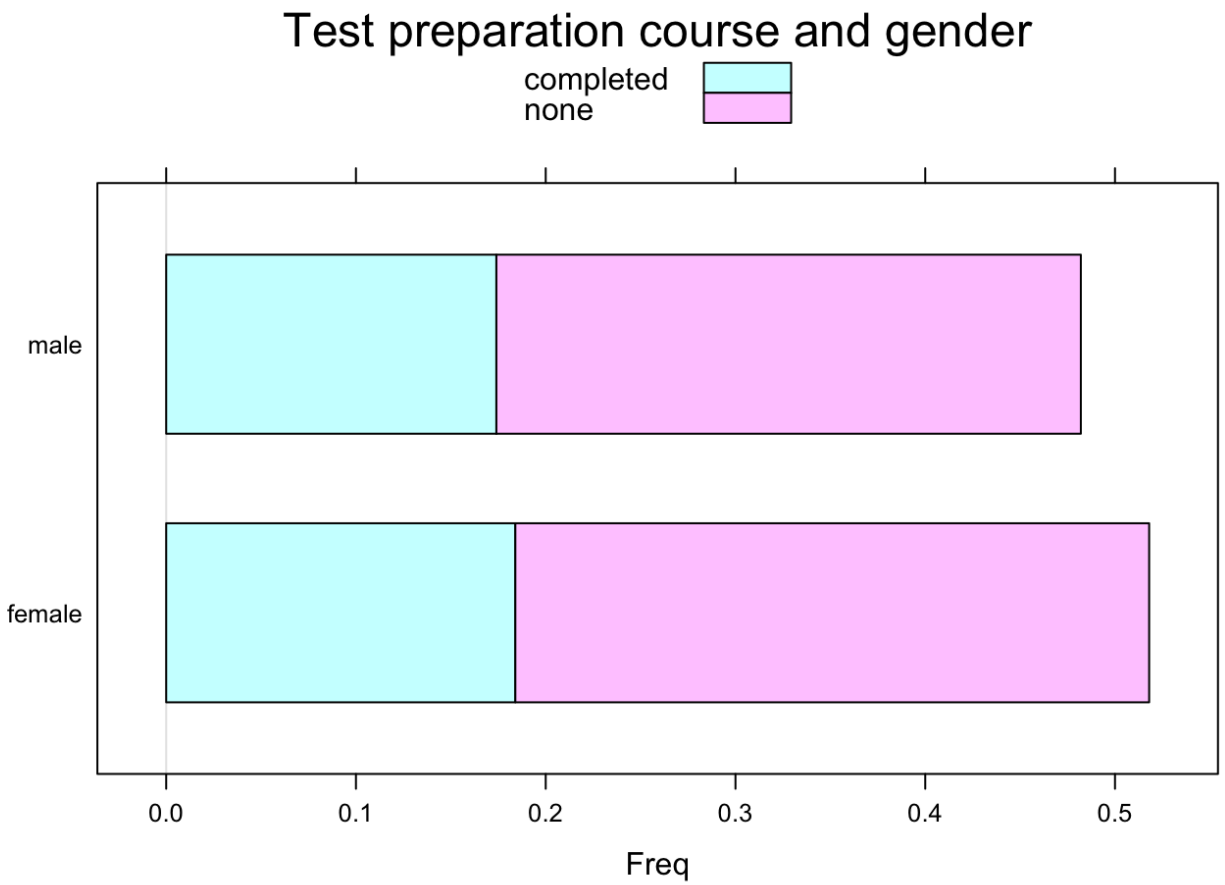
```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

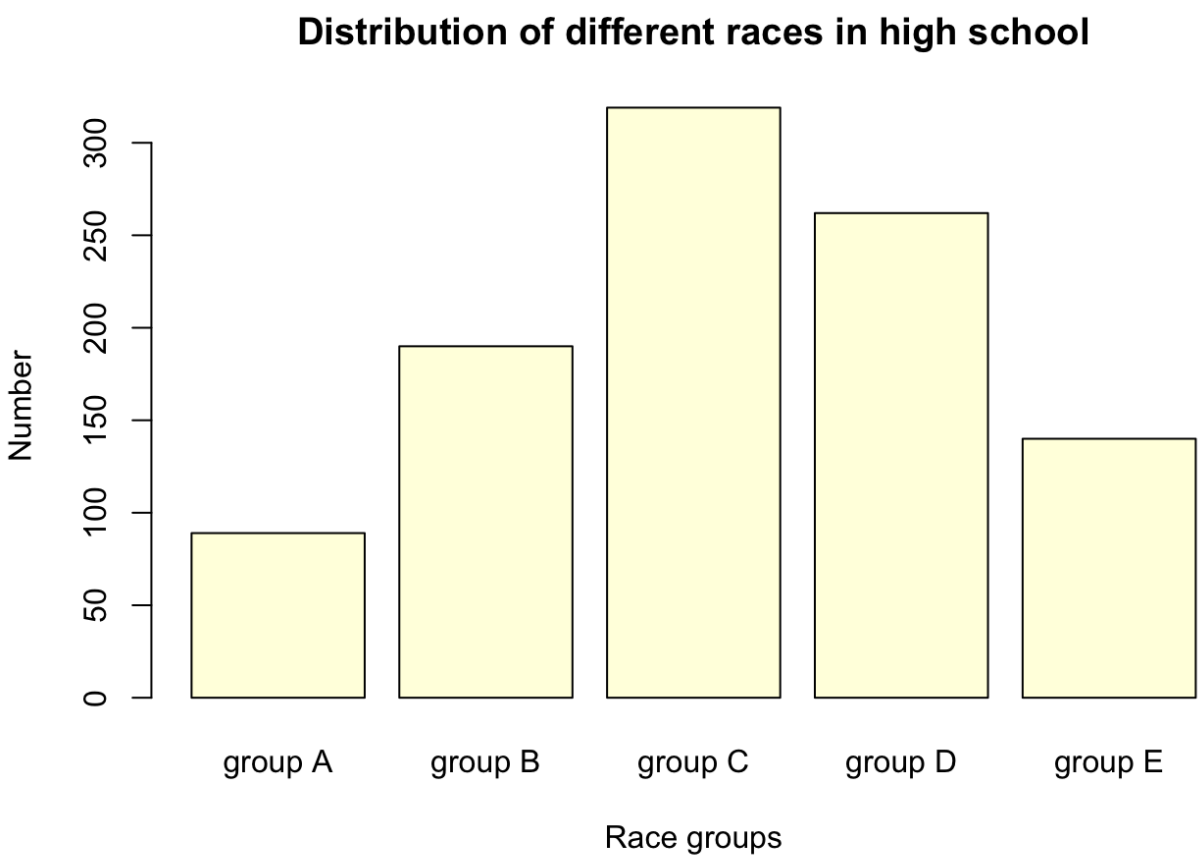
```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```



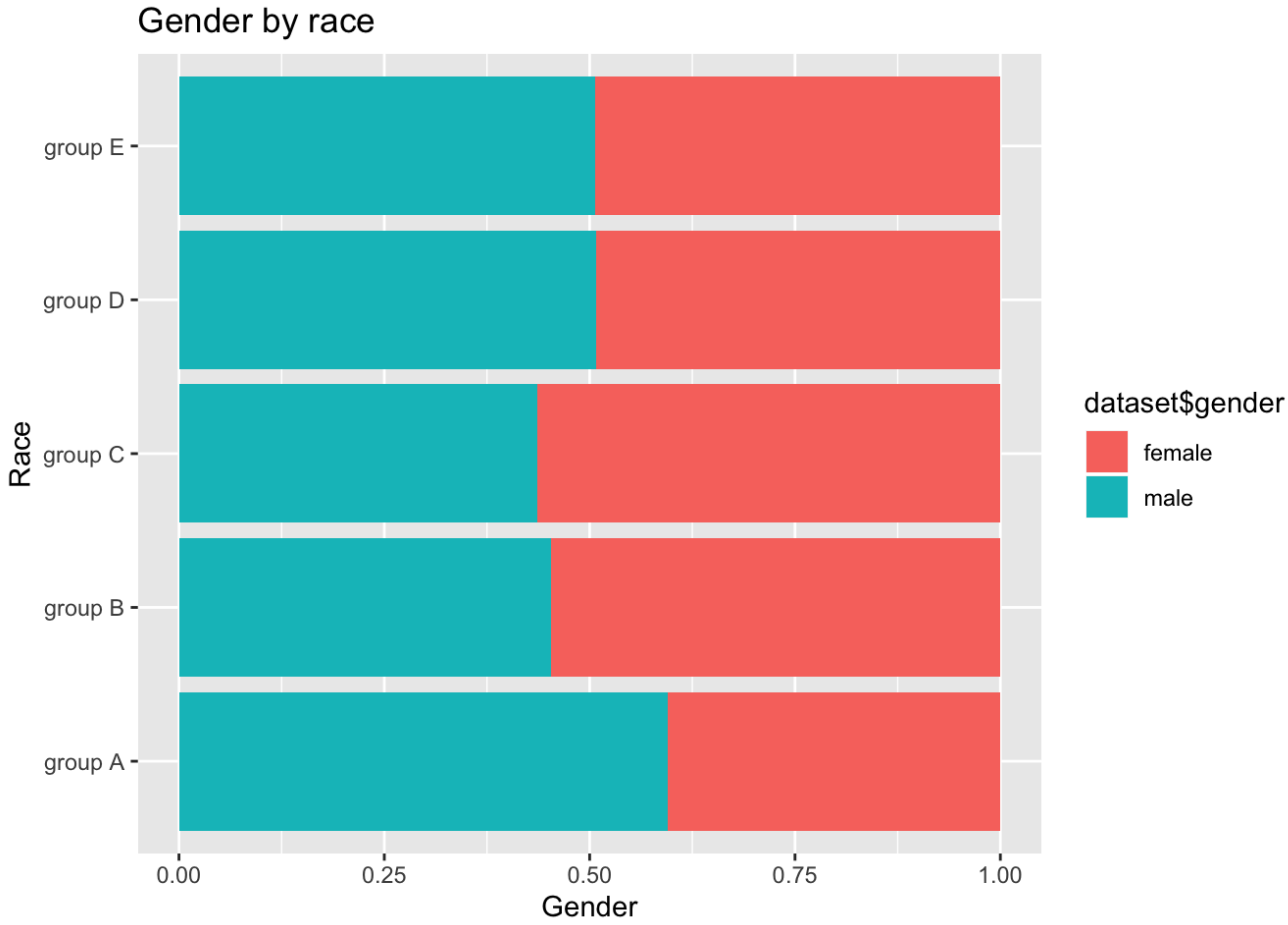
By plotting a simple stacked bar chart for our dataset, each segment has been placed after the previous one and the total value of the bar is all the segment values added together. It can be observed from the bar chart how the test preparation course is divided into completed and not completed category for both the genders and what relationship each category has with the total category.

Let us try to understand the distribution of different races in high school.



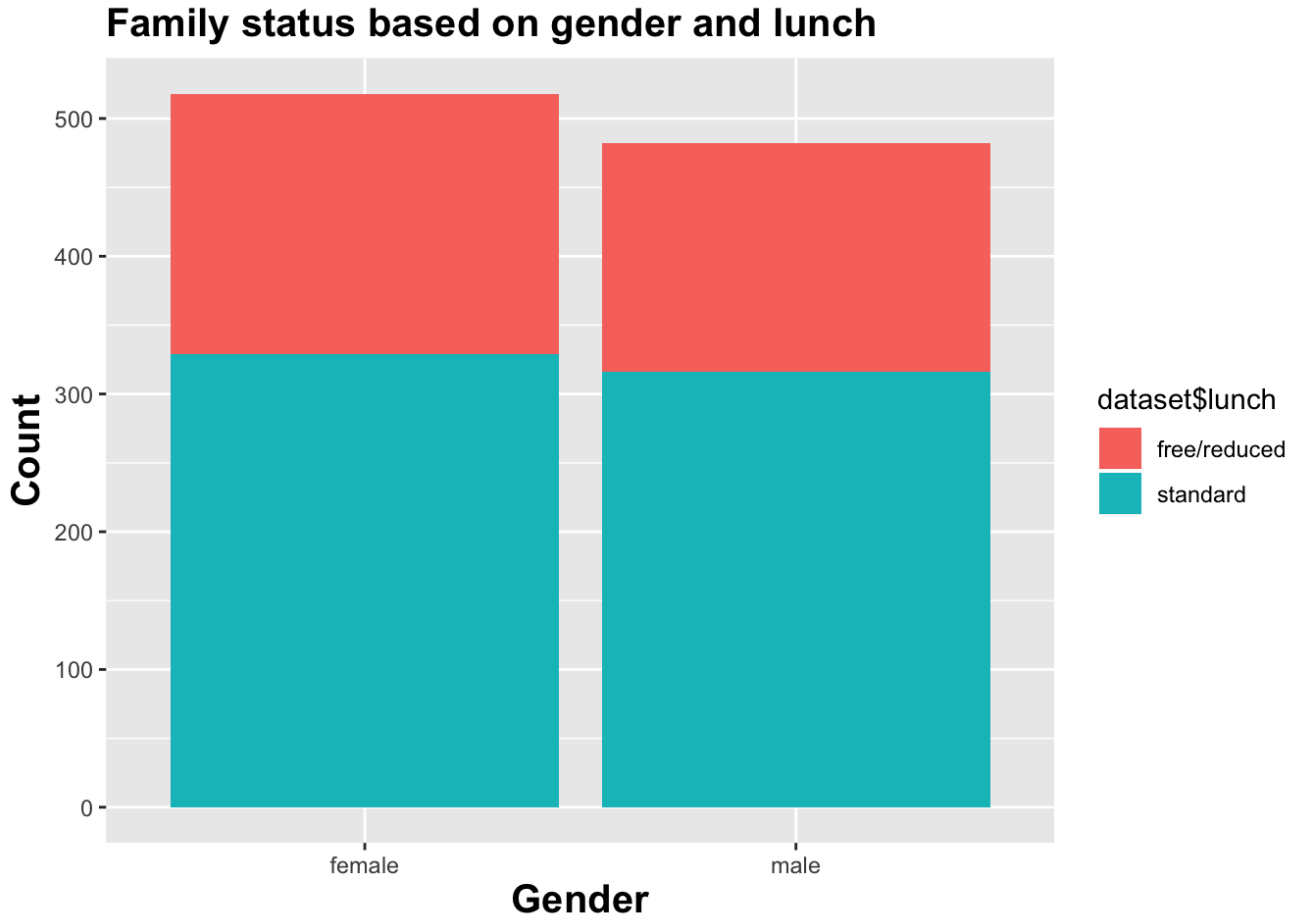
It can be observed from our plot that race ‘Group C’ is in majority in our dataset whereas race ‘Group A’ is in minority. As the actual names of these races are unknown, so we cannot claim which race is more popular among students in high schools in the United States.

Let us see the distribution of gender in the different races.



It is interesting to note from the graph that the majority of races have more females in high schools in the United States than males. It is a useful result in terms of gender equity in education.

Let us see if something interesting can be interpreted about the family status by finding a relationship between the gender and type of lunch students avail.



It can be observed from the ggplot that the majority of the students avail lunch at the standard cost irrespective of the gender. This can make us conclude that the majority of high school students in the United Status belong to a family who has a good status in terms of standard of living and income.

Part 3

Statistical Analysis

One sample t-test

Traditional statistical tool

The National Assessment of Educational Progress (NAEP) claims that the students in the United States are struggling readers. The studies conducted in this regard, claims that the average reading score of high school students in the United States is just 68 out of 100.

We are willing to analyse whether the claim about the reading score of high school students in the United States is true or not. As we do not know our population variance, we can use t-test. Let us first check the conditions to use t-test.

Conditions for use and check for condition:

1. The sample is representative of the population - The study says that the sample is of 1000 students belonging to different genders, race/ethnicity. The study even contains information about their test preparation status and their parental level of education. The sample seems to be representative of the population.

2. One quantitative variable of interest - We are taking the reading scores into account as our quantitative variable.
3. We want to make inference about the population mean using the sample mean - Yes
4. The population variance is unknown, so we will estimate using our sample data - Yes
5. We are assuming that the sample comes form a single population - Yes, all students are high school students in the United States.
6. We have plotted a Q-Q plot above (in Part 2) for the reading score to find out whether the population data is normally distributed or not - Yes, it seems normal.

Question of interest: What is the average reading score of high school students in the United States?

Parameter:

The population parameter we want to make an inference to is μ.

Hypothesis:

Null hypothesis: The true mean reading score of high school students in the United States is 68. $\mu_0 = 68$

$$H_0 : \mu = \mu_0$$

$$H_0 : \mu = 68$$

Alternate hypothesis: The true mean reading score of high school students in the United States is different from 68.

$$H_A : \mu \neq 68$$

Sample statistic:

The sample statistic is the sample mean reading score, $\bar{x}$.

Distribution of the test statistic:

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

p-value:

```
## [1] 0.01149516
```

Confidence interval:

Let us create a (1 - α) confidence interval for the value of the true population parameter.

```
## [1] -1.962341
```

Our t-critical value has come out to be close to the normal equivalent because our sample size is large and the t-distribution with 999 degrees of freedom is very close to an N(0,1) distribution.

$$P(\bar{x} - t_{(\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}}) = (1 - \alpha)$$

```
## [1] 68.26299
```

```
## [1] 70.07501
```

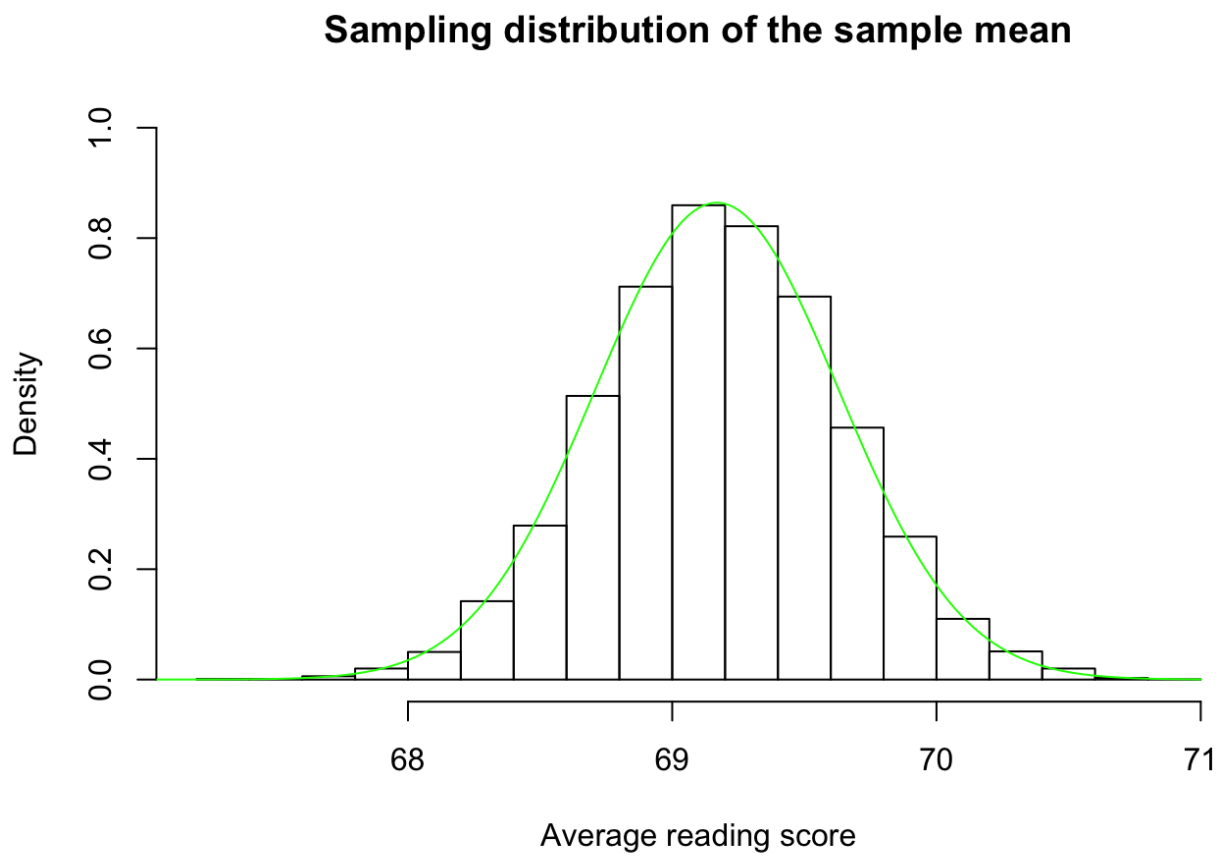
Sanity check:

```
##
##  One Sample t-test
##
## data:  dataset$reading.score
## t = 2.532, df = 999, p-value = 0.0115
## alternative hypothesis: true mean is not equal to 68
## 95 percent confidence interval:
##  68.26299 70.07501
## sample estimates:
## mean of x
##    69.169
```

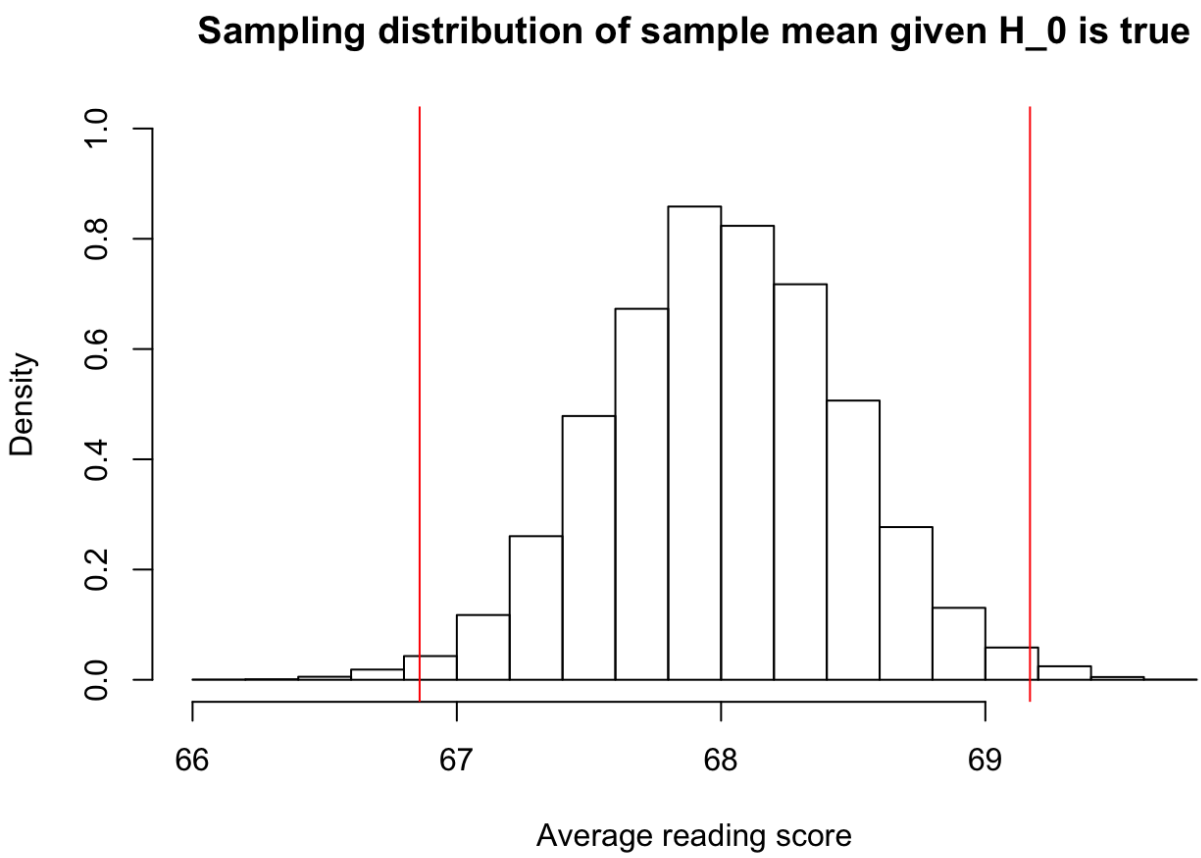
Interpretation:

There is strong evidence (p-value = 0.01149516) to suggest that the true mean reading score of high school students in the United States is different from 68. We reject the null hypothesis that the true mean reading score of high school students in the United States is 68 at the α = 0.05 level. With 95% confidence, we can say that the true mean reading score is between 68.26299 and 70.07501 which suggests that the true mean reading score is greater than 68.

Bootstrap Approach



We now have our sampling distribution and we need to find the p-value. The p-value is the probability that we observe a test statistic as or more unusual than the one we observed, given that the null hypothesis is true. If the null hypothesis is true then the true mean reading score of high school students in the United States is 68. We need to shift our distribution so that this is true.



Bootstrap p-value:

```
## [1] 0.0139
```

```
## [1] 0.01149516
```

As our sample size is large, our p-values from both the methods are fairly close.

Bootstrap confidence interval:

```
## [1] 68.24997 70.08803
```

```
##     2.5%    97.5%
## 68.28700 70.08205
```

```
## [1] 68.26299 70.07501
```

The interval using the empirical method, especially the quantile method, is wider which agrees with our p-value being a bit more conservative. Thus, we can say that the traditional t method is making our result more significiant and our confidence interval narrower as compared to the empirical method.

One sample test of proportion

Traditional statistical tool

We have a sample data where 518 out of 1000 students are female. So, according to our sample, the sample proportion is 51.8% or we can say that it is approximately 52%. Since we have a categorical variable named gender involved in our analysis, we will use one-sample test of proportion.

$$\hat{p} = 0.52$$

According to the findings of the US Department of Education, more than 50% of the students in high school are females. We are willing to analyze whether the claim made by the US Department of Education is true or not.

Let us first check the conditions to use one sample test of proportion.

Conditions for use and check for condition:

1. We have a categorical variable of interest with two categories - Gender: Male and Female
2. We are assuming that the sample comes form a single population - Yes, all students are high school students in the United States
3. Exact binomial test - No requirements
4. Normal approximation

$n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

$1000 \times 0.52 = 520$ and $1000 \times (1-0.52) = 480$

We can proceed with the one-sample proportion exact test and the requirements for the normal approximation are also met.

Question of interest: What is the proportion of females in high school in the United States?

Parameter:

The population parameter we want to make an inference to is the population proportion (p) of females in high schools.

Hypothesis:

Null hypothesis: The true proportion of females in high school is 50%. $p_0 = 0.5$

$$H_0 : p_F = p_0$$

$$H_0 : p_F = 0.50$$

Alternate hypothesis: The true proportion of females in high school is greater than 50%.

$$H_A : p_F > 0.50$$

Sample statistic:

The sample statistic is the sample proportion $\hat{p} = 0.52$.

Test statistic:

Fo exact test, there is no test statistic, so we will directly find the probability.

For normal approximation, we will use p_0 to find the test statistic.

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

```
## [1] 0.4
```

Distribution of the test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

p-value:

One-sided upper exact

```
##
##
##
## data:  52 out of 100
## number of successes = 52, number of trials = 100, p-value = 0.3822
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.4332319 1.0000000
## sample estimates:
## probability of success
##                   0.52
```

One sided upper normal approximation

```
## [1] 0.3445783
```

Confidence interval:

```
## [1] 0.4332319 1.0000000
## attr(,"conf.level")
## [1] 0.95
## attr(,"method")
## [1] "Score"
```

```
## [1] 0.4380656 1.0000000
```

Interpretation:

Using the exact binomial method for a one-sample test of proportion, there is no evidence (p-value = 0.3445783) to suggest that the true proportion of females in high school is greater than 50%. We fail to reject the null hypothesis that the true proportion of females in high school is equal to 50% at the α = 0.05 level. The true proportion of females in high school is between 0.4828245 and 1.0000000.

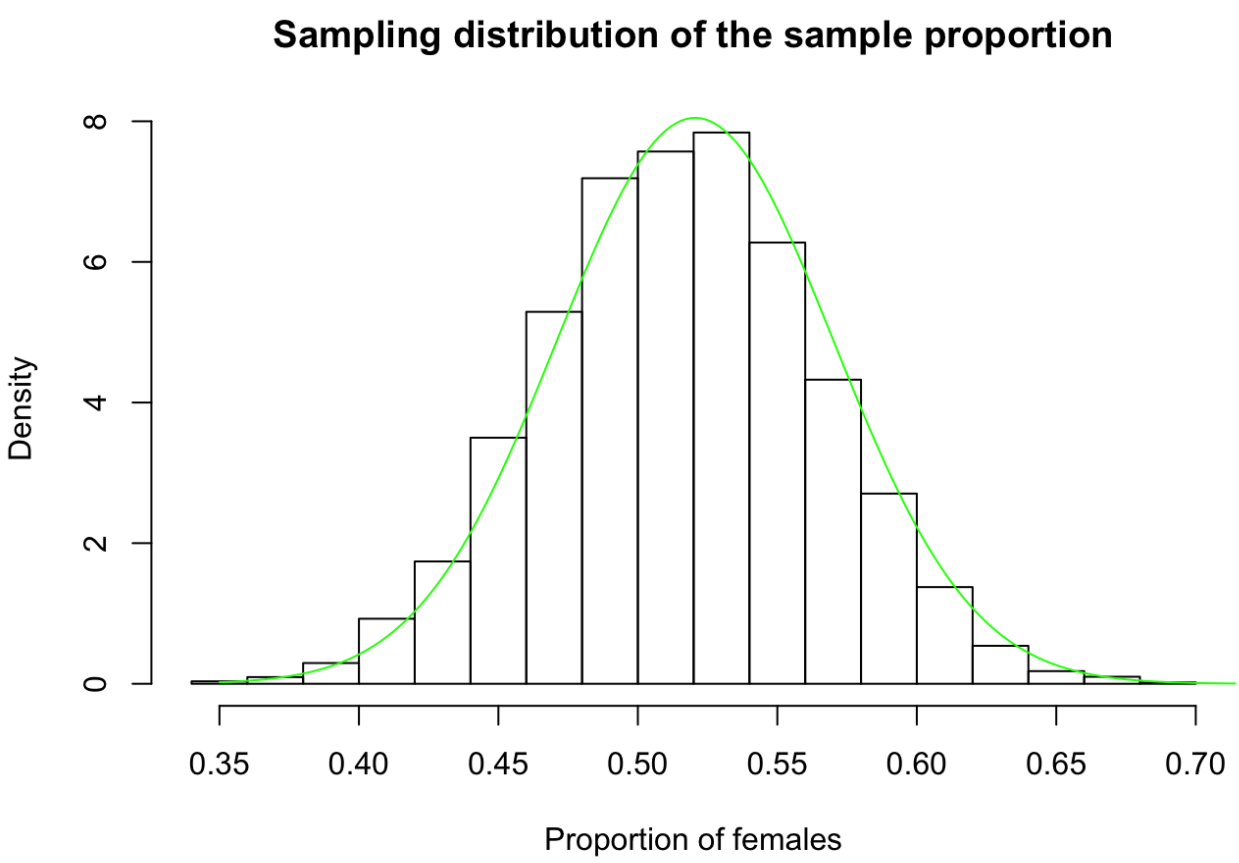
Bootstrap Approach

Let us create our data to describe our situation.

```
##   [1] female female female female female female female female female female
##  [11] female female female female female female female female female female
##  [21] female female female female female female female female female female
##  [31] female female female female female female female female female female
##  [41] female female female female female female female female female female
##  [51] female female male   male   male   male   male   male   male   male
##  [61] male   male   male   male   male   male   male   male   male   male
##  [71] male   male   male   male   male   male   male   male   male   male
##  [81] male   male   male   male   male   male   male   male   male   male
##  [91] male   male   male   male   male   male   male   male   male   male
## Levels: female male
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
## females
##  0  1
## 48 52
```



Bootstrap confidence interval:

Using this sampling distribution to find the 5th and 95th percentiles and then we will compare them to the other methods.

```
##   5% 100%
## 0.44 0.70
```

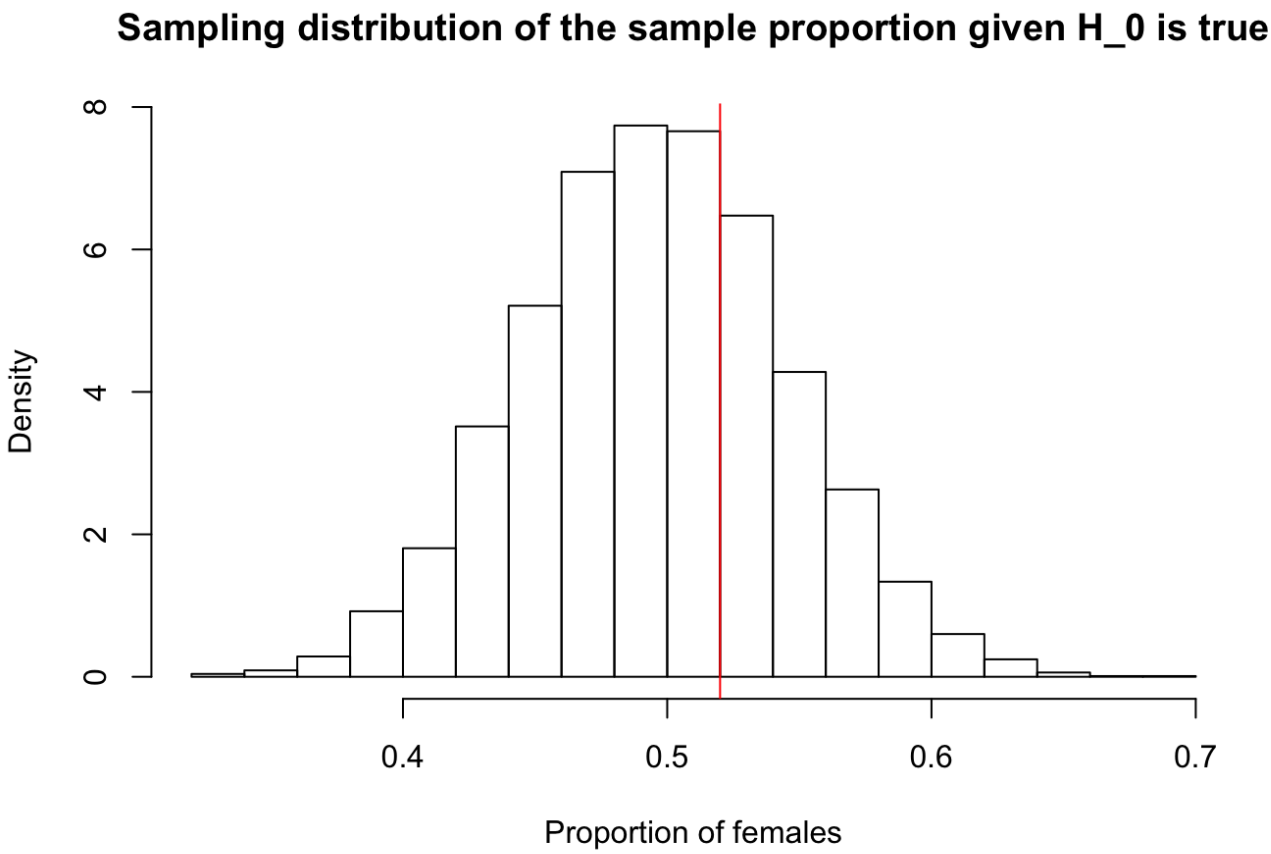
```
## [1] 0.4332319 1.0000000
## attr(,"conf.level")
## [1] 0.95
## attr(,"method")
## [1] "Score"
```

```
## [1] 0.4220784 1.0000000
```

We can observe that the one-sided confidence interval from the bootstrap stops at 0.70 because, in our empirical distribution, no value exceeds 0.70, however theoretically they could. Our empirical result is giving us a smaller confidence interval compared to the exact and normal approximations.

Bootstrap p-value:

We now have our sampling distribution and we need to find the p-value. In order to find the p-value, we need to create a sampling distribution under the assumption of the null hypothesis being true.




```
## [1] 0.3893
```

```
## [1] 0.3821767
```

```
## [1] 0.3445783
```

In our case, our bootstrap p-value is closer to the exact binomial p-value and it is more conservative than the normal approximation.

Two sample t-test for difference in means

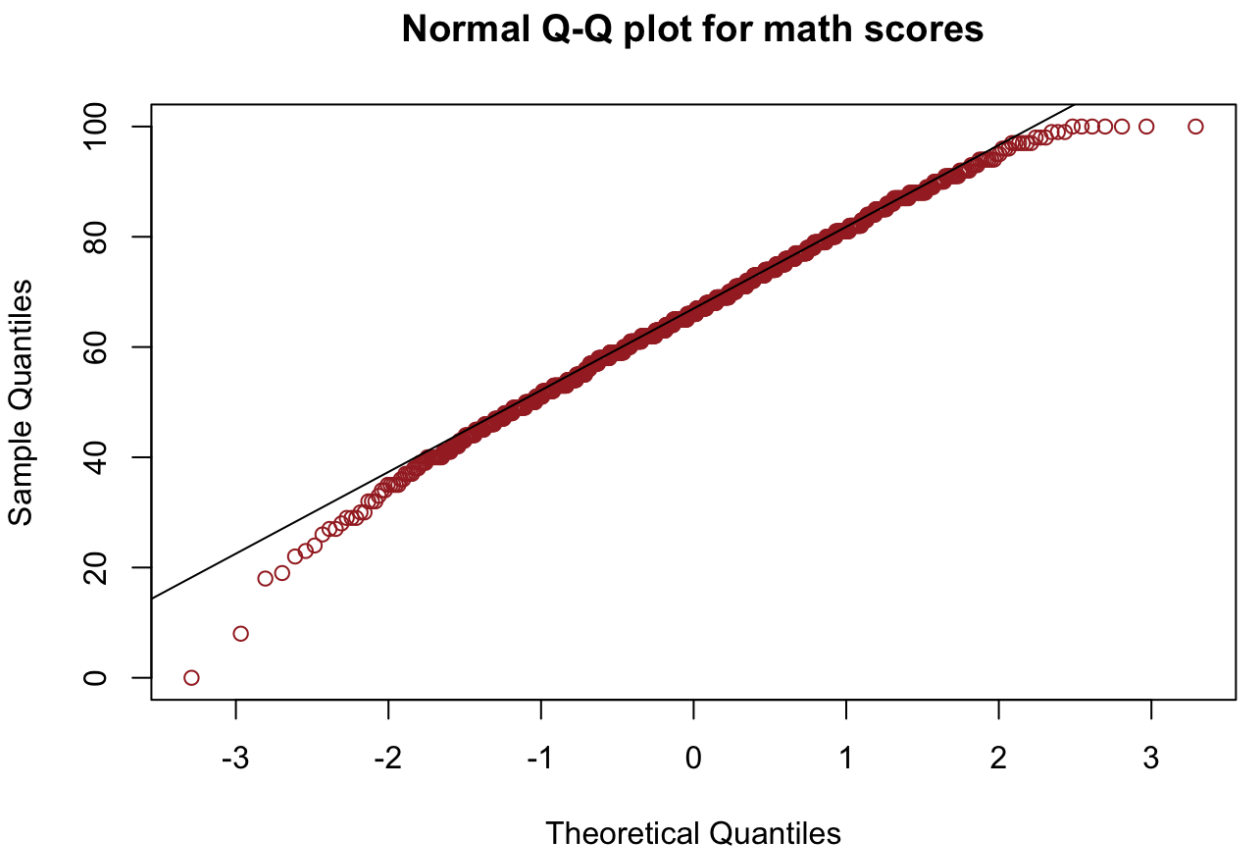
Traditional statistical tool

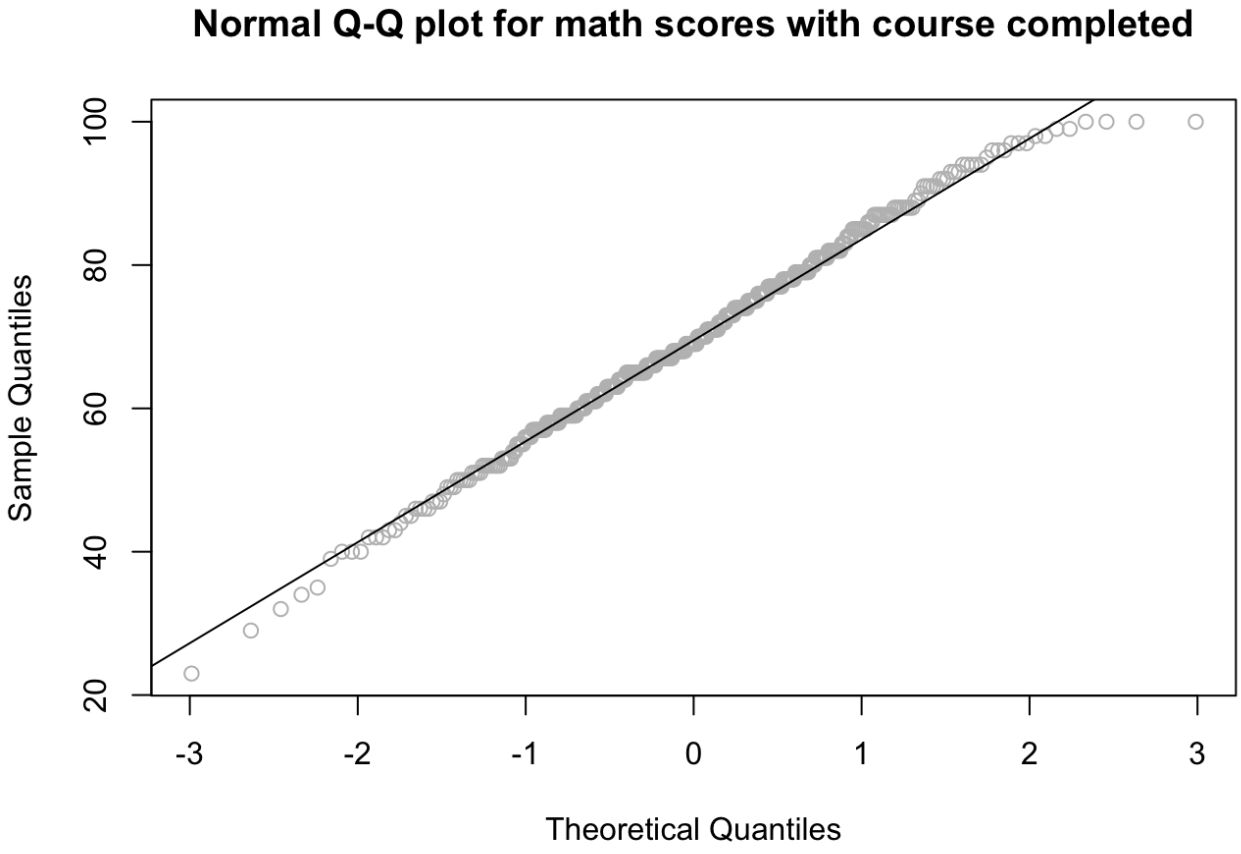
Some studies say that students who take the test preparation courses score higher than the ones who don't. While others disagree with this fact. Can the test preparation course have an effect on the math score for high school students? We are willing to find some insight into this through our study.

As we have two independent samples of interest, we can use the t-test. Let us first check the conditions to use this test.

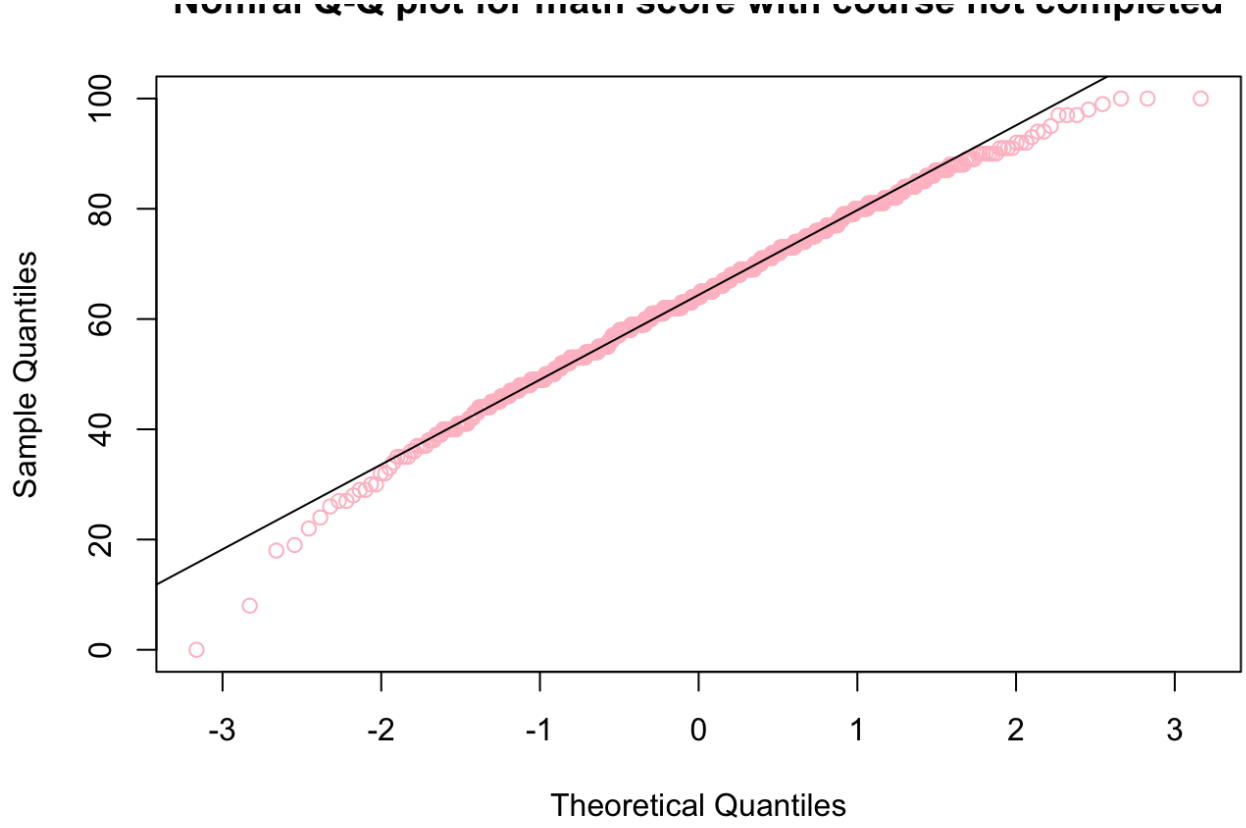
Conditions for use and check for condition:

1. The sample is representative of the population - The study says that the sample is of 1000 students belonging to different genders, race/ethnicity. The study even contains information about their test preparation status and their parental level of education. The sample seems to be representative of the population.
2. The question of interest has to do with the difference in the mean math score between two populations - Yes, students who have completed the test preparation course and have not completed the test preparation course and the difference in the mean math score for each population.
3. Two independent samples from two populations - Yes, considering students who have completed the test preparation course and who have not completed test preparation
4. We have plotted a Q-Q plot for the sample data to find out whether the population data is normally distributed or not - Yes, they seem normal





Nomral Q-Q plot for math score with course not completed



Question of interest:

Is there a difference in the average math scores between the students who complete the test preparation course and those who don't?

Parameter:

We are interested in the true population mean difference in math scores between those who complete a test preparation course and those who don't.

$\mu_c - \mu_n$

Hypothesis:

Null hypothesis: The true population mean math score of those who complete a test preparation course is the same as the true population mean math score of those who don't.

$H_0 : \mu_c - \mu_n = 0$ or $\mu_c = \mu_n$

Alternate hypothesis: The true population mean math score of those who complete the test preparation course is different from the true population mean math score of those who don't.

$H_A : \mu_c - \mu_n \neq 0$ or $\mu_c \neq \mu_n$

Sample statistic:

$$\bar{x}_c - \bar{x}_n$$

Test statistic:

$$t_{min(n_{c-1}, n_{n-1})} = \frac{(\bar{x}_c - \bar{x}_n) - (\mu_c - \mu_n)}{\sqrt{\frac{s_c^2}{n_c} + \frac{s_n^2}{n_n}}}$$

Distribution of test statistic:

$$t_{min(n_{c-1},n_{n-1})} = \frac{(\bar{x}_c - \bar{x}_n) - (\mu_c - \mu_n)}{\sqrt{\frac{s_c^2}{n_c} + \frac{s_n^2}{n_n}}} \sim t_{min(n_{c-1},n_{n-1})}$$

p-value:

```
## [1] 1.569151e-08
```

Confidence interval:

```
## [1] 3.708564
```

```
## [1] 7.526734
```

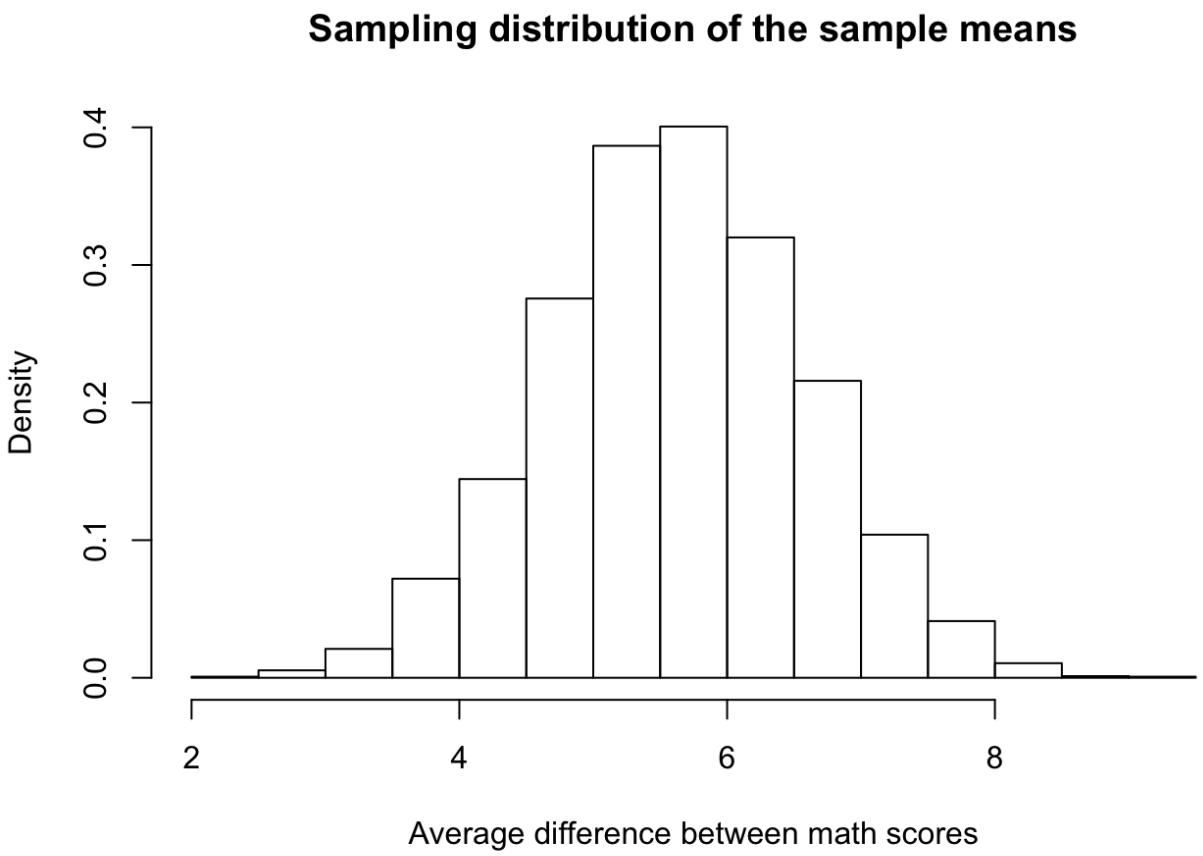
Sanity check:

```
##
##  Welch Two Sample t-test
##
## data:  (dataset$math.score[dataset$test.preparation.course == "completed"]) and (datase
t$math.score[dataset$test.preparation.course == "none"])
## t = 5.787, df = 770.08, p-value = 1.043e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.712041 7.523257
## sample estimates:
## mean of x mean of y
##  69.69553  64.07788
```

Interpretation:

There is no evidence (p-value = 1.569151e-08) to suggest that the true population mean math score is the same for the students who have completed the test preparation course and the ones who haven't. We reject the null hypothesis that there is no difference between the true population mean math score of students who completed the test preparation course and those who have not at the α = 0.05 level. With 95% confidence, the true difference in the mean math scores between those who have completed the test preparation and those who haven't is between 3.708564 and 7.526734. The null hypothesized difference between the mean math score is zero and zero is not in the 95% confidence interval which is consistent with the rejection of the null hypothesis. The values of the confidence interval suggest that on average those who complete the test preparation course score more than the ones who haven't completed the test preparation course.

Bootstrap Approach



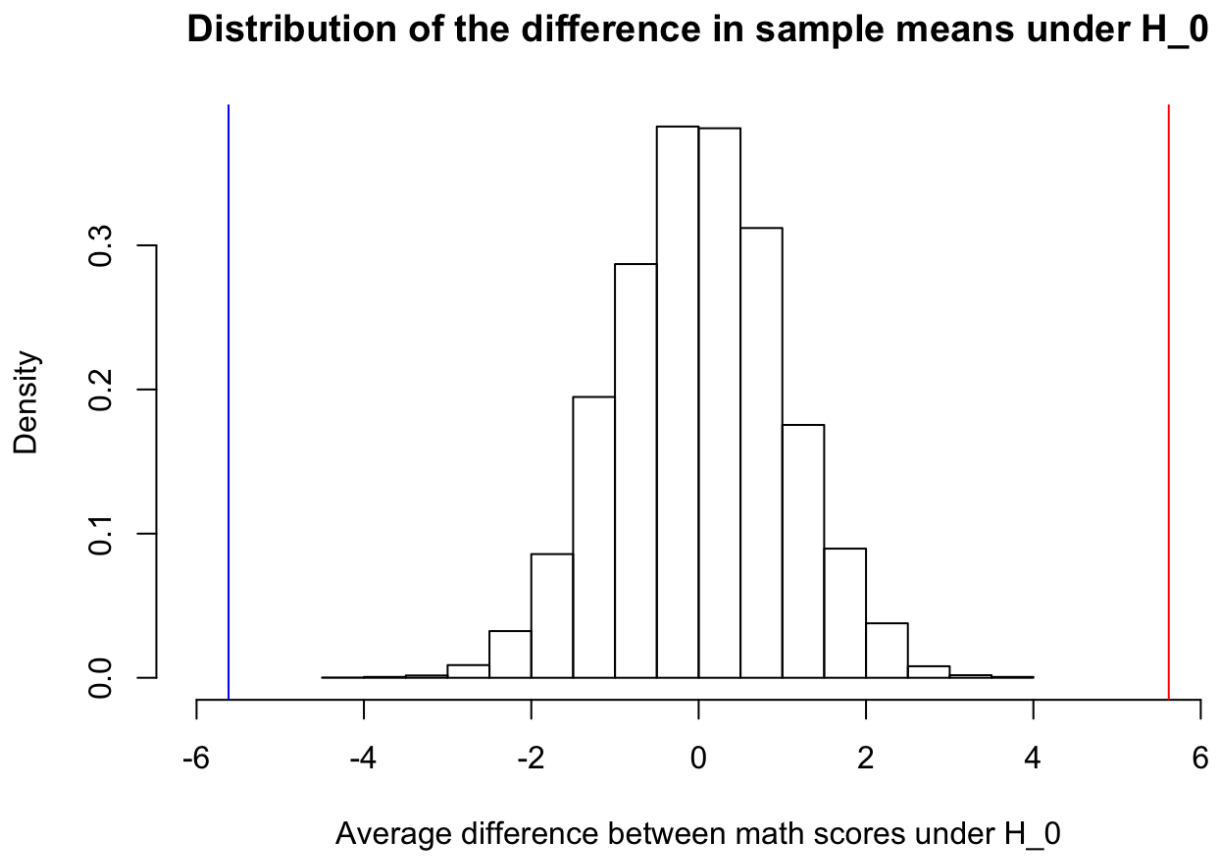
Bootstrap confidence interval:

```
##     2.5%    97.5%
## 3.695310 7.520599
```

```
## [1] 3.712041 7.523257
## attr(,"conf.level")
## [1] 0.95
```

Bootstrap p-value:

We need to shift our distribution so that the null hypothesis is true, that is, the mean difference between the math scores of the two groups is 0.




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

The interval using the traditional method is wider which agrees with our p-value being a bit more conservative. Thus, we can say that the empirical method is making our result more significiant and our confidence interval narrower as compared to the traditional method.

Two sample test for difference in proportions

Traditional statistical tool

Research shows that managing expenses in the United States while studying at a good school is a costly affair. We have a categorical variable named lunch in our dataset. Let us use a statistical method to understand whether the standard of living differs between males and females in high school based on the fact that they avail standard lunch at school.

We are willing to understand how gender and type of lunch availed are related. We can use two-sample test for difference in proportions in this case, but first, we need to check whether the conditions to use the test are met or not.

Conditions for use and check for condition:

1. The sample is representative of the population - The study says that the sample is of 1000 students belonging to different genders, race/ethnicity. The study even contains information about their test preparation status and their parental level of education. The sample seems to be representative of the population.
2. Categorical response variable with two categories - Yes
3. Two independent samples from two populations - Yes

4. np $\geq$ 10 and n(1 - p) $\geq$ 10 for both the populations - Yes

Question of interest:

Does the proportion of males who avail standard lunch at school differ from the proportion of females who avail standard lunch at school?

Parameter:

We are interested in the difference between the true population proportion of female students who avail standard lunch and the true population proportion of male students who avail standard lunch at school.

Hypothesis:

Null hypothesis: There is no difference between the true population proportion of female students who avail standard lunch and the true population proportion of male students who avail standard lunch at school.

$$H_0 : p_F - p_M = 0$$

Alternate hypothesis: There is a difference between the true population proportion of female students who avail standard lunch and the true population proportion of male students who avail standard lunch at school.

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

Sample statistic:

$$\hat{p}_F - \hat{p}_M$$

Test statistic:

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

Distribution of test statistic:

$$z = \frac{(\hat{p}_M - \hat{p}_F) - (p_M - p_F)}{\sqrt{\frac{\hat{p}_M(1-\hat{p}_M)}{n_M} + \frac{\hat{p}_F(1-\hat{p}_F)}{n_F}}} \sim N(0, 1)$$

p-value:

```
## [1] 0.4988486
```

Confidence interval:

```
## [1] -0.03884665
```

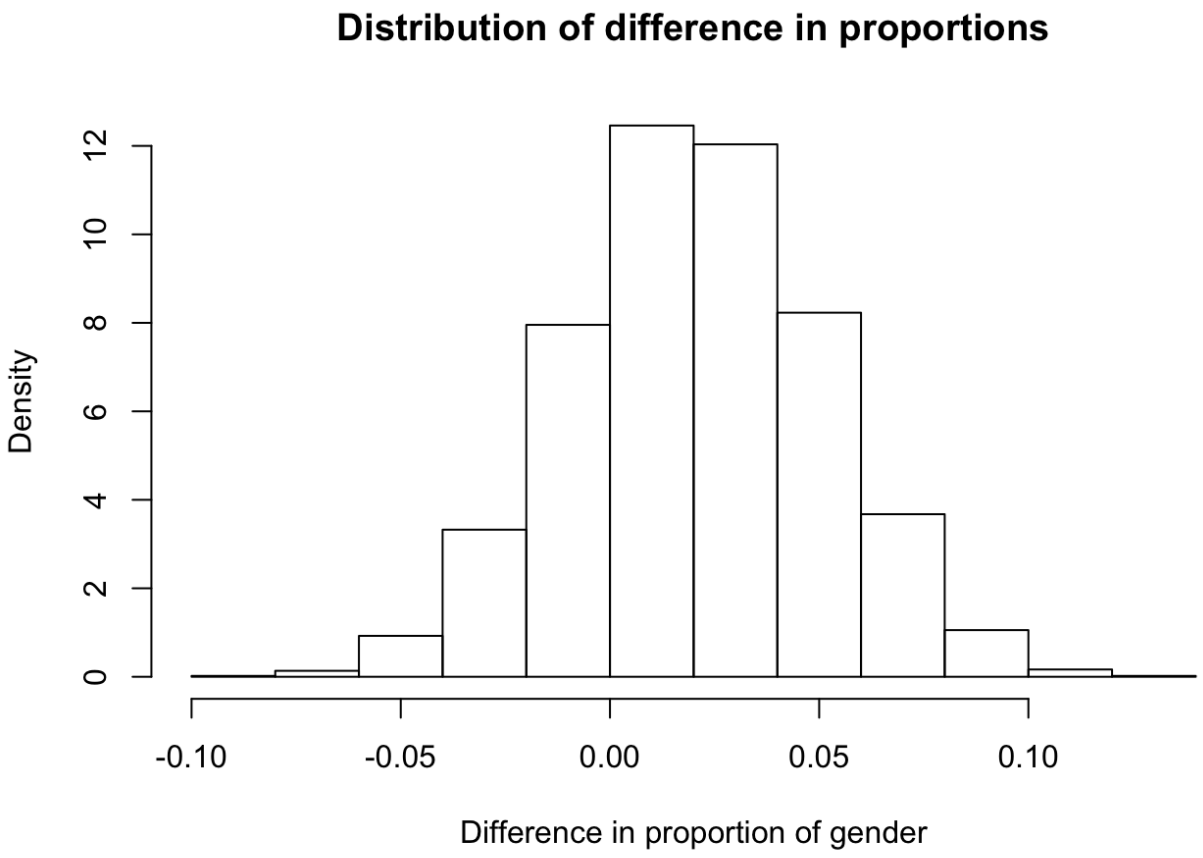
```
## [1] 0.0797797
```

Interpretation:

There is no evidence (p-value = 0.4988486) to suggest that there is a difference between the true population proportion of female students who avail standard lunch and the true population proportion of male students who avail standard lunch at school. We fail to reject the null hypothesis that the true population proportion of female students who avail standard lunch is the same as the true population proportion of male students who avail

standard lunch at school at the α = 0.05 level. With 95% confidence, we can say that the true population proportion difference is between 3.88% less standard lunch to 7.97% more standard lunch availed by the males than the females. The null hypothesized difference of 0 is in the confidence interval which agrees with our failure to reject the null hypothesis.

Bootstrap Approach



Bootstrap confidence interval:

```
##        2.5%       97.5%
## -0.03802748  0.07991297
```

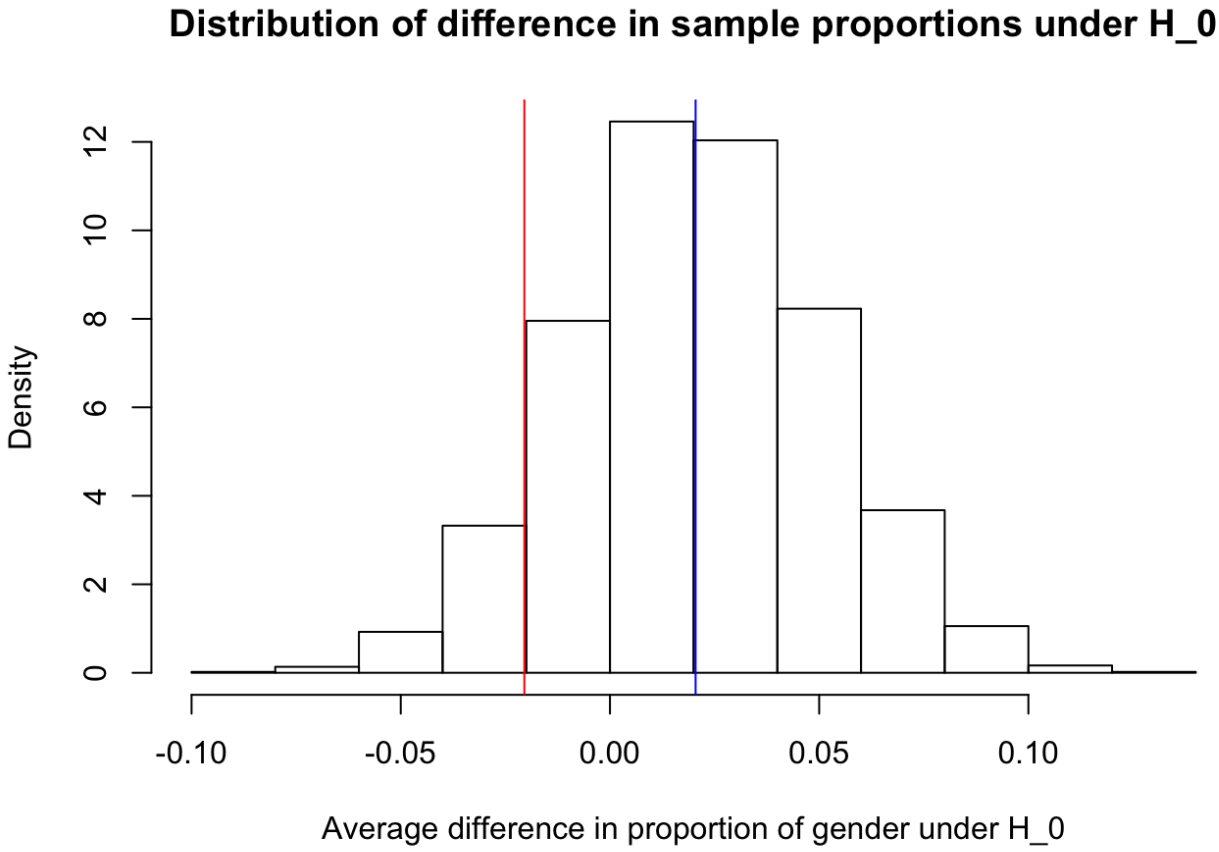
```
## [1] -0.03884665  0.07977970
```

In order to find the p-value, we need to shift our distribution or create a randomization distribution. Let us use the randomization approach to create a randomization distribution. We will stimulate many samples assuming the null hypothesis is true. Under the null hypothesis, there is does not exist any relationship between two variables. We will create many samples where the treatment group will be shuffled and compute the difference in groups for each of the samples. We will then create a histogram of the randomized statistics in order to approximate the null distribution.

```
## female   male
##    518    482
```

```
## [1] TRUE
```

```
## [1] TRUE
```




```
## [1] 0.5819
```

Using randomization also we observed that there is no evidence to suggest that there is a difference between the true population proportion of female students who avail standard lunch and the true population proportion of male students who avail standard lunch at school.

Chi-square goodness of fit test

Traditional statistical tool

We keep on reading news about the race or ethnicity of students in US schools. Some news claims that the white students are now in minority in the US schools. They posit that the enrollment of white students in US schools is decreasing each year. We are willing to find out the truth behind these claims through our dataset.

Our dataset has five race groups (A, B, C, D and E). Let us find out is any one race is more likely to be enrolled in US schools than the other? As we have races by the names Group A, Group B and so on, so we will be able to provide our interpretation like that and not with the actual race which exists in the world.

Our study has 5 categorical variables in the group race/ethnicity. So, it seems that we can use the chi-square goodness of fit test. Let us first check its conditions to use.

Conditions for use and check for condition:

1. Single categorical variable with more than 2 variables - Yes, Race/Ethnicity with five variables
2. The expected count of each is more than 5 - Yes

Let us first find out the proportion of each variable in the category race/ethnicity.

```
##
## group A group B group C group D group E
##      89     190     319     262     140
```

```
##
## group A group B group C group D group E
##   0.089   0.190   0.319   0.262   0.140
```

Question of interest:

Does the high school in the United States have students of all races in equal proportion?

Parameter:

We are interested in the true proportions: p_A, p_B, p_C, p_D, p_E.

Hypothesis:

Null hypothesis: The proportion of each race is the same and equal to 0.2.

$H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$

Alternate hypothesis: At least one of the proportion is not equal to 0.2.

H_A: Some $p_i \neq 0.2$

We have 1000 observations in our sample. If each of the race had the same frequency, then each race would have a count of (1000)(0.2) = 200. So, under the null hypothesis, the expected count np_i = 200.

Sample statistic:

We have five sample statistics: $\hat{p}_A, \hat{p}_B, \hat{p}_C, \hat{p}_D, \hat{p}_E$

Test statistic and distribution:

We need to compare our sample (observed) counts (O_i) with the expected count (E).

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

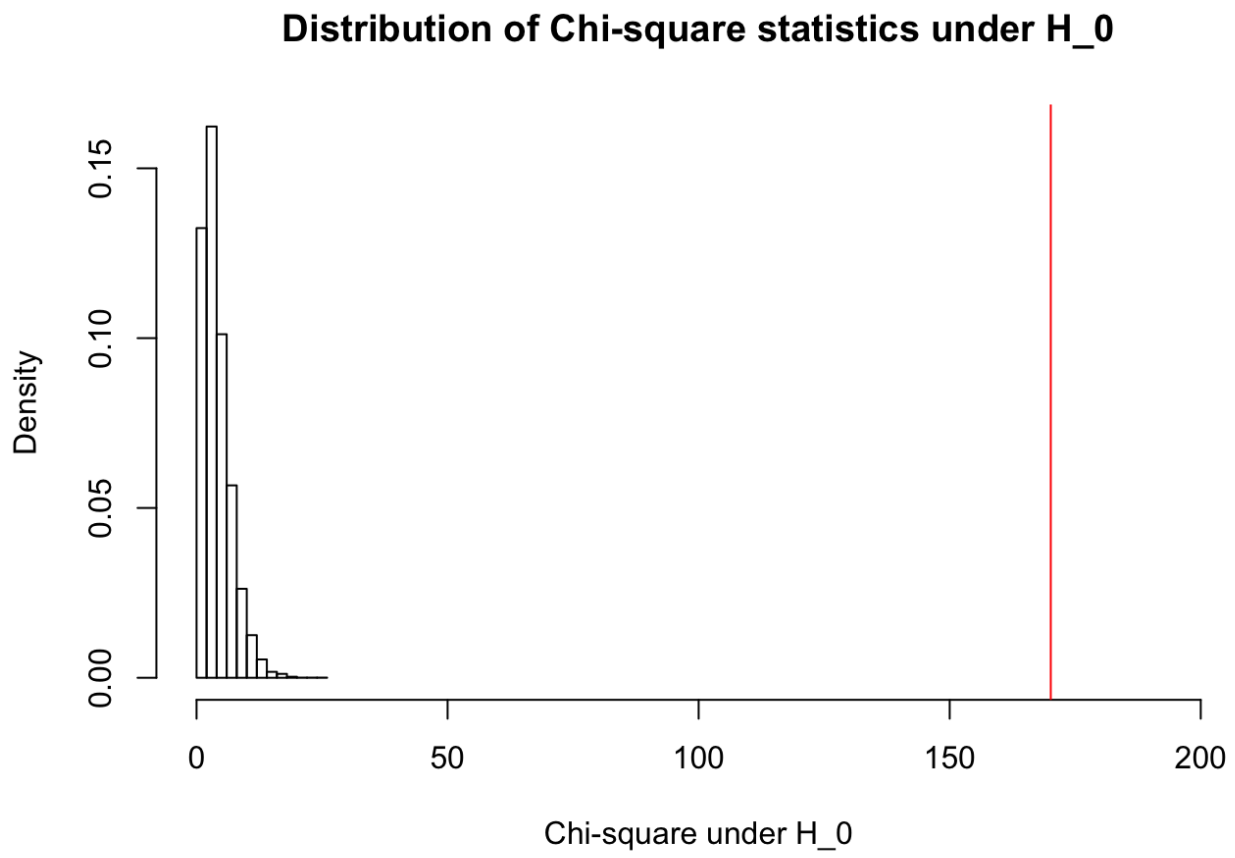
```
## [1] 170.13
```

p-value:

```
## [1] 9.807684e-36
```

Randomization Approach

```
## results_under_H_0
##   A   B   C   D   E
## 200 200 200 200 200
```



Randomized p-value:

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

Interpretation:

The data provides evidence that we observed a sample that is impossible under the null as our p-value is exactly equal to zero. So, the data provides strong evidence that the proportions of different races are very different than 0.2 at the α = 0.05 level.

Part 4

Discussion

In this study, we have performed a one-sample t-test, one sample test of proportion, two-sample t-test for difference in means, two-sample test for difference in proportions, chi-square goodness of fit test.

Using the one-sample t-test of mean, we tried to find out the average reading score of students and to check whether the claims made by NAEP regarding the average score of 68 out of 100 are true or not. The result provides strong evidence that the true mean reading score of students is more than 68.

Using the one-sample test of proportion, we worked to find out the proportion of females in high school in the United States. The resulting proportion can give us an insight into gender equity in high schools in the United States. Our result provides no evidence of the claim made by the US Department of Education that the proportion of females in high schools in the United States is more than 50%. We failed to reject our null hypothesis that the true proportion of females in high school is equal to 50%.

Using the two-sample test of difference in means, the question addressed in our study is trying to figure out whether there is a difference in the average math scores between the students who complete the test preparation course and those who don't. Our result provides no evidence to suggest both the population means are equal. Our result strongly claims that there exists a difference in the mean math scores between students who have taken the preparation course and those who haven't. It is quite obvious from our result that on average students who complete the test preparation course score higher than the students who don't complete the test preparation course.

Using the two-sample test of difference in proportions, we worked to find out the difference between the proportion of males who avail standard lunch at school and the proportion of females who avail standard lunch at school. The result obtained from our test states that there is no evidence to suggest that there is a difference between the true population proportion of female students who avail standard lunch and the true population proportion of male students who avail standard lunch at school. We can claim that an equal proportion of males and females avail the standard lunch at school.

Using the chi-square test, we have addressed the question regarding the different races in high schools in the United States. As studies always claim about equity in education in terms of race, we have tried to find out whether there exists an equal proportion of each race or not. The result provides strong evidence that the proportion of different races is not equal. The graphs clearly show that race 'Group C' is dominant in the high schools of the United States.

In spite of doing all these statistical tests to find interesting claims, there do exist limitations in our results due to the lack of information regarding the actual source of data. It seems like the data has been created in order to conduct a study on US high schools. As the dataset has just 1000 entries, it is difficult to say that the sampling has been done randomly and has been done at various schools.

The findings of our study can be used to draw conclusions or to take measures regarding the following points.

1. The gender distribution of students in US high schools. An effort can be put in to have gender equity in US schools by raising awareness.
2. The kind of background these students belong to by looking at their parental level of education. Measures can be taken to help the students whose parents have not studied much to come to school and study. We need to raise awareness regarding the importance of education in ones life.
3. The number of students who took the test preparation course in various subjects. If we can conduct a survey in schools and talk to teachers regarding the time they spend with weak students, we can even try to save the money of the parents who send their students to coachings apart from the school.
4. Racism has always been a big issue in the United States. Many studies provide evidence regarding the bully happening in schools with students not belonging to US. We can take measures to conduct counseling in the schools and let students know that each one of us is equal as a human being and bully is not acceptable at any ground.

In the current study, a lot of research has been done to find out the claims made by the US education department and to conduct different tests in order to see whether the claims hold true or not. However, there still exists many questions that can be solved in order bring useful insights from this dataset. We could have taken parental level of education into account to find out how does that affect students score in different tests or students of which race/ethnicity perform better in which test. Some advanced modeling techniques like the random forest, K-Nearest neighbor model, single decision tree and linear regression could have also helped to find out interesting insights from this dataset.

Part 5

Appendix

Dataset:

Dataset from Kaggle:

https://www.kaggle.com/spscientist/students-performance-in-exams (https://www.kaggle.com/spscientist/students-performance-in-exams)

Bibliography:

Gender equity in education in USA:

https://www2.ed.gov/about/offices/list/ocr/docs/gender-equity-in-education.pdf (https://www2.ed.gov/about/offices/list/ocr/docs/gender-equity-in-education.pdf)

Effect of test preparation course on scores:

https://nepc.colorado.edu/sites/default/files/Briggs_Theeffectofadmissionstestpreparation.pdf (https://nepc.colorado.edu/sites/default/files/Briggs_Theeffectofadmissionstestpreparation.pdf)

https://www.wsj.com/articles/SB124278685697537839 (https://www.wsj.com/articles/SB124278685697537839)

Effect of parental education on child's success:

https://degree.lamar.edu/articles/undergraduate/parents-education-level-and-childrens-success.aspx (https://degree.lamar.edu/articles/undergraduate/parents-education-level-and-childrens-success.aspx)

https://inclusiveschools.org/impact-of-parents-on-student-success/ (https://inclusiveschools.org/impact-of-parents-on-student-success/)

Race/Ethnicity classification in US schools:

https://nces.ed.gov/programs/coe/pdf/coe_cge.pdf (https://nces.ed.gov/programs/coe/pdf/coe_cge.pdf)

https://nces.ed.gov/programs/raceindicators/indicator_rbb.asp (https://nces.ed.gov/programs/raceindicators/indicator_rbb.asp)

https://www.publicschoolreview.com/blog/white-students-are-now-the-minority-in-u-s-public-schools (https://www.publicschoolreview.com/blog/white-students-are-now-the-minority-in-u-s-public-schools)

```
knitr::opts_chunk$set(echo = FALSE)
#Dataset
dataset <- read.csv(file = "StudentsPerformance.csv")
head(dataset)
summary(dataset)
#qqnorm graph
qqnorm(dataset$math.score, main = "Normal Q-Q plot - Math score", col = "Orange"); qqline
(dataset$math.score)
qqnorm(dataset$reading.score, main = "Normal Q-Q plot - Reading score", col = "Green"); qq
line(dataset$reading.score)
qqnorm(dataset$writing.score, main = "Normal Q-Q plot - Writing score", col = "Blue"); qql
ine(dataset$writing.score)
#histogram
hist(dataset$math.score, xlab = "Math score",
     main = "Math score of high school students", col = "Light Blue", ylim = c(0,280))
hist(dataset$reading.score, xlab = "Reading score",
     main = "Reading score of high school students", col = "Light Green", ylim = c(0,280))
hist(dataset$writing.score, xlab = "Wriring score",
     main = "Writing score of high school students", col = "Light Yellow", ylim = c(0,280
))
boxplot(dataset$math.score ~ dataset$gender, xlab = "Gender", ylab = "Math score", main =
"Math score as per gender", col = "Pink")
boxplot(dataset$reading.score ~ dataset$gender, xlab = "Gender", ylab = "Reading score", m
ain = "Reading score as per gender", col = "Purple")
boxplot(dataset$writing.score ~ dataset$gender, xlab = "Gender", ylab = "Writing score", m
ain = "Writing score as per gender", col = "Orange")
library(mosaic)
gender_course_table <- table(dataset$gender, dataset$test.preparation.course)
barchart(prop.table(gender_course_table), auto.key = list(title = "Test preparation course
and gender"))
plot(dataset$race.ethnicity, col = "Light Yellow", main = "Distribution of different races
in high school", ylab = "Number", xlab = "Race groups")
library(ggplot2)
a <- ggplot(dataset, aes(dataset$race.ethnicity, fill = dataset$gender))
a + geom_bar(position = "fill") + ggtitle("Gender by race") + labs(x = "Race", y = "Gende
r") + coord_flip()
b <- ggplot(dataset, aes(dataset$gender, fill = dataset$lunch))
b + stat_count() + ggtitle("Family status based on gender and lunch") + labs (x = "Gender"
, y = "Count") + theme(plot.title = element_text(face = "bold", size = 15, hjust = 0)) + t
heme(axis.title = element_text(face = "bold", size = 15))
#sample mean
x_bar <- mean(dataset$reading.score)
#null hypothesized population mean
mu_0 <- 68
#sample standard deviation
s <- sd(dataset$reading.score)
#sample size
n <- length(dataset$reading.score)
#t-test test statistic
t <- (x_bar - mu_0)/(s/sqrt(n))
#two-sided p-value
two_sided_p_value <- pt(q = t, df = n - 1, lower.tail = FALSE)*2
two_sided_p_value
qt(0.025, n - 1)
#lower bound
```

```
x_bar + (qt(0.025, n-1)*(s/sqrt(n)))
#upper bound
x_bar + (qt(0.975, n-1)*(s/sqrt(n)))
t.test(dataset$reading.score, alternative = "two.sided", mu = 68 )
set.seed(0)
#let us perform our simulations
num_sims1 <- 10000
#'result1' vector to store the results
result1 <- rep(NA, num_sims1)
#Loop to complete the simulations
for (i in 1:num_sims1) {
result1[i] <- mean(sample(x = dataset$reading.score, size = n, replace = TRUE))
}
#Let us plot our results
hist(result1, freq = FALSE, main = 'Sampling distribution of the sample mean', xlab = 'Ave
rage reading score', ylab = 'Density', ylim = c(0,1))
lines(x = seq(67, 71, 0.01), dnorm(seq(67, 71, 0.01), mean = x_bar, sd = s/sqrt(1000)), co
l = "Green")
set.seed(0)
#let us shift our sample so that our null hypothesis is true
reading_score_H0_true <- dataset$reading.score - mean(dataset$reading.score) + mu_0
num_sims1 <- 10000
#'result1_H0_true' vector to store the results
result1_H0_true <- rep(NA, num_sims1)
#Loop to complete the simulations
for (i in 1:num_sims1) {
result1_H0_true[i] <- mean(sample(x = reading_score_H0_true, size = n, replace = TRUE))
}
#Let us plot our results
hist(result1_H0_true, freq = FALSE, main = 'Sampling distribution of sample mean given H_0
is true', xlab = 'Average reading score', ylab = 'Density', ylim = c(0,1))
#line to show the values more extreme on the upper end
abline(v = x_bar, col = "red")
#line to show the values more extreme on the lower end
low_extreme_end <- mean(result1_H0_true) + (mean(result1_H0_true) - x_bar)
abline(v = low_extreme_end, col = "red")
set.seed(0)
#counts of values more extreme than the test statistic in our original sample, given null
 hypothesis is true
count_more_extreme_lowertail <- sum(result1_H0_true <= low_extreme_end)
count_more_extreme_uppertail <- sum(result1_H0_true >= x_bar)
bootstrap_pvalue <- (count_more_extreme_lowertail + count_more_extreme_uppertail)/num_sims
1
bootstrap_pvalue
#two-sided p-value from our traditional method
two_sided_p_value
#We need standard error to calculate our confidence interval. Here, the standard error is
 the standard deviation of result1 vector.
bootstrap_SE_x_bar <- sd(result1)
#confidence interval
c(x_bar - 2 * bootstrap_SE_x_bar, x_bar + 2 * bootstrap_SE_x_bar)
#5th and 95th quantiles to determine the bounds
c(quantile(result1, c(0.025, 0.975)))
#confidence interval from our traditional method
c(x_bar + (qt(0.025, n-1)*(s/sqrt(n))), x_bar + (qt(0.975, n-1)*(s/sqrt(n))))
```

```
z <- (0.52 - 0.50) / sqrt ((0.5 * (1 - 0.5))/100)
z
binom.test(x = 52, n = 100, p = 0.5, alternative = "greater")
pnorm(0.4, lower.tail = FALSE)
#exact binomial test
binom.test(x = 52, n = 100, p = 0.5, alternative = "greater")$conf.int
#normal approximation
c(0.52 - (1.64) * sqrt(((0.52) * (1 - 0.52))/100), 1)
set.seed(0)
females <- factor(rep(c("female", "male"), c(52, (100 - 52))))
females
set.seed(0)
#Let us recode female as 1 and male as 0.
females <- rep(c(1,0), c(52, (100 - 52)))
females
table(females)
set.seed(0)
#let us perform our simulations
num_sims2 <- 10000
#'result2' vector to store the results
result2 <- rep(NA, num_sims2)
#Loop to complete the simulations
for (i in 1:num_sims2) {
result2[i] <- mean(sample(x = females, size = 100, replace = TRUE))
}
#Let us plot our results
hist(result2, freq = FALSE, main = 'Sampling distribution of the sample proportion', xlab
 = 'Proportion of females', ylab = 'Density')
lines(x = seq(0.35, 0.75, 0.001), dnorm(seq(0.35, 0.75, 0.001), mean = mean(result2), sd =
sd(result2)), col = "Green")
#5th and 95th percentiles
set.seed(0)
c(quantile(result2, c(0.05, 1)))
#exact binomial test
binom.test(x = 52, n = 100, p = 0.5, alternative = "greater")$conf.int
#normal approximation
c(0.52 - (1.96) * sqrt(((0.52) * (1 - 0.52))/100), 1)
#Under the assumption that the null hypothesis is true, we have 50% females
set.seed(0)
#let us perform our simulations
females <- rep(c(1, 0), c(50, (100 - 50)))
num_sims2 <- 10000
#'result2' vector to store the results
result2 <- rep(NA, num_sims2)
#Loop to complete the simulations
for (i in 1:num_sims2) {
result2[i] <- mean(sample(x = females, size = 100, replace = TRUE))
}
#Let us plot our results
hist(result2, freq = FALSE, main = 'Sampling distribution of the sample proportion given H
_0 is true', xlab = 'Proportion of females', ylab = 'Density')
abline(v = 0.52, col = "Red")
set.seed(0)
#counts of values more extreme than the test statistic in our original sample, given null
 hypothesis is true
```

```
count_more_extreme_upper_tail <- sum(result2 >= 0.52)
bootstrap_pvalue <- count_more_extreme_upper_tail/num_sims2
bootstrap_pvalue
#Exact binomial p-value
binom.test(x = 52, n = 100, p = 0.5, alternative = "greater")$p.value
#Normal approximation p-value
pnorm(0.4, lower.tail = FALSE)
#Q-Q plot
qqnorm(dataset$math.score, col = "Brown", main = "Normal Q-Q plot for math scores"); qqlin
e(dataset$math.score)
qqnorm(dataset$math.score[dataset$test.preparation.course == "completed"], main = "Normal
 Q-Q plot for math scores with course completed", col = "Grey"); qqline(dataset$math.score
[dataset$test.preparation.course == "completed"])
qqnorm(dataset$math.score[dataset$test.preparation.course == "none"], main = "Nomral Q-Q p
lot for math score with course not completed", col = "Pink"); qqline(dataset$math.score[da
taset$test.preparation.course == "none"])
#sample means
x_bar_c <- mean(dataset$math.score[dataset$test.preparation.course == "completed"])
x_bar_n <- mean(dataset$math.score[dataset$test.preparation.course == "none"])
#null hypothesized population mean difference between two groups
mu_0 <- 0
#sample variances
s_c_v <- sd(dataset$math.score[dataset$test.preparation.course == "completed"]) ** 2
s_n_v <- sd(dataset$math.score[dataset$test.preparation.course == "none"]) ** 2
#sample size
n_c <- length(dataset$math.score[dataset$test.preparation.course == "completed"])
n_n <- length(dataset$math.score[dataset$test.preparation.course == "none"])
#t-test test statistic
t <- (x_bar_c - x_bar_n - mu_0)/sqrt((s_c_v/n_c) + (s_n_v/n_n))
#p-value
pval <- pt(q = t, df = (min(n_c, n_n) - 1), lower.tail = FALSE) * 2
pval
#lower bound
(x_bar_c - x_bar_n) + (qt(0.025, min(n_c, n_n) - 1) * sqrt((s_c_v/n_c) + (s_n_v/n_n)))
#upper bound
(x_bar_c - x_bar_n) + (qt(0.975, min(n_c, n_n) - 1) * sqrt((s_c_v/n_c) + (s_n_v/n_n)))
t.test((dataset$math.score[dataset$test.preparation.course == "completed"]), (dataset$mat
h.score[dataset$test.preparation.course == "none"]))
set.seed(0)
#Let us perform our simulations
num_sims3 <- 10000
#'result3' vector to store the results
result3 <- rep(NA, num_sims3)
#Loop to complete the simulations
for (i in 1:num_sims2) {
mean_completed <- mean(sample(x = dataset$math.score[dataset$test.preparation.course == "c
ompleted"], size = n_c, replace = TRUE))
mean_none <- mean(sample(x = dataset$math.score[dataset$test.preparation.course == "none"
], size = n_n, replace = TRUE))
result3[i] <- mean_completed - mean_none
}
#Let us plot our results
hist(result3, freq = FALSE, main = 'Sampling distribution of the sample means', xlab = 'Av
erage difference between math scores', ylab = 'Density')
c(quantile(result3, c(0.025, 0.975)))
```

```
#comparing this with our t method
t.test((dataset$math.score[dataset$test.preparation.course == "completed"]), (dataset$mat
h.score[dataset$test.preparation.course == "none"]))$conf.int
set.seed(0)
#Let us perform our simulations
num_sims3 <- 10000
#'result3_given_H0_true' vector to store the results
result3_given_H0_true <- rep(NA, num_sims3)
#Loop to complete the simulations
for (i in 1:num_sims2) {
shuffled_groups <- transform(dataset, test.preparation.course = sample(test.preparation.co
urse))
mean_completed <- mean(shuffled_groups$math.score[shuffled_groups$test.preparation.course
 == "completed"])
mean_none <- mean(shuffled_groups$math.score[shuffled_groups$test.preparation.course == "n
one"])
result3_given_H0_true[i] <- mean_completed - mean_none
}
#Let us plot our results
hist(result3_given_H0_true, freq = FALSE, main = 'Distribution of the difference in sample
means under H_0', xlab = 'Average difference between math scores under H_0', ylab = 'Densi
ty', xlim = c(-6,6))
diff_in_sample_means <- mean(dataset$math.score[dataset$test.preparation.course == "none"
]) - mean(dataset$math.score[dataset$test.preparation.course == "completed"])
abline(v = diff_in_sample_means, col = "Blue")
abline(v = abs(diff_in_sample_means), col = "Red")
#counts of values more extreme than the test statistic in our original sample, given H_0 i
s true
count_of_more_extreme_lower_tail <- sum(result3_given_H0_true <= diff_in_sample_means)
count_of_more_extreme_upper_tail <- sum(result3_given_H0_true >= abs(diff_in_sample_mean
s))
bootstrap_p_value <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tai
l)/num_sims3
bootstrap_p_value
#sample proportions
P_Male_Count = NROW(dataset$lunch[dataset$gender == 'male'])
P_Female_Count = NROW(dataset$lunch[dataset$gender == 'female'])
male_stand <- NROW(dataset$lunch[which(dataset$gender == 'male' & dataset$lunch == 'standa
rd')])
female_stand <- NROW(dataset$lunch[which(dataset$gender == 'female' & dataset$lunch == 'st
andard')])
P_Male_hat <- male_stand/P_Male_Count
P_Female_hat <- female_stand/P_Female_Count
#null hypothesized population proportion difference between the two groups
p_0 <- 0
#sample variance
sv_p_M <- (P_Male_hat * (1 - P_Male_hat))/P_Male_Count
sv_p_F <- (P_Female_hat * (1 - P_Female_hat))/P_Female_Count
#z-test test statistic
z <- (P_Male_hat - P_Female_hat - p_0)/sqrt(sv_p_M + sv_p_F)
#two sided p-value
two_sided_diff_prop_pval <- pnorm(q = z, lower.tail = FALSE)*2
two_sided_diff_prop_pval
#lower bound
(P_Male_hat - P_Female_hat) + (qnorm(0.025) * sqrt(sv_p_F + sv_p_M))
```

```
#upper bound
(P_Male_hat - P_Female_hat)  + (qnorm(0.975) * sqrt(sv_p_F + sv_p_M))
#Let us make the data
set.seed(0)
male_lunch <- rep(c(1,0), c(male_stand, P_Male_Count - male_stand))
female_lunch <- rep(c(1,0), c(female_stand, P_Female_Count - female_stand))
num_sims4 <- 10000
#'result4' vector to store the results
result4 <- rep(NA, num_sims4)
#Loop to complete simulations
for (i in 1:num_sims4) {
  prop_females <- mean(sample(male_lunch, size = P_Male_Count, replace = TRUE))
  prop_males <- mean(sample(female_lunch, size = P_Female_Count, replace = TRUE))
  result4[i] <- prop_females - prop_males
}
hist(result4, freq = FALSE, main = "Distribution of difference in proportions", xlab = "Di
fference in proportion of gender", ylab = "Density")
set.seed(0)
c(quantile(result4, c(0.025, 0.975)))
#Normal approximation
c((P_Male_hat - P_Female_hat) + (qnorm(0.025)*sqrt(sv_p_F + sv_p_M)), (P_Male_hat - P_Fema
le_hat) + qnorm(0.975) * sqrt(sv_p_F + sv_p_M))
#make the data
set.seed(0)
combined_df <- data.frame("lunch" = c(female_lunch, male_lunch), "gender" = rep(c("female"
, "male"), c(P_Female_Count, P_Male_Count)))
summary(combined_df$gender)
mean(combined_df$lunch[combined_df$gender == "female"]) == P_Female_hat
mean(combined_df$lunch[combined_df$gender == "male"]) == P_Male_hat
set.seed(0)
num_sims4 <- 10000
#make the data (recode as 0 and 1)
m <- rep(c(1,0), c(male_stand, P_Male_Count - male_stand))
f <- rep(c(1,0), c(female_stand, P_Female_Count - female_stand))
#'result4' vector to store the results
result4_H0_true <- rep(NA, num_sims4)
#Loop to complete simulations
for (i in 1:num_sims4) {
  #shuffled_group1 <- transform(combined_df, gender == sample(gender))
  prop_males <- mean(sample(m, size = P_Male_Count, replace = TRUE))
  prop_females <- mean(sample(f, size = P_Female_Count, replace = TRUE))
  result4_H0_true[i] <- prop_males - prop_females
}
hist(result4_H0_true, freq = FALSE, main = "Distribution of difference in sample proportio
ns under H_0", xlab = "Average difference in proportion of gender under H_0", ylab = "Dens
ity")
diff_in_sample_props <- P_Male_hat - P_Female_hat
abline(v = diff_in_sample_props, col = "Blue")
abline(v = -diff_in_sample_props, col = "Red")
#count of values more extreme than the test statistic given H_0 is true
#two-sided alternate hypothesis
count_of_more_extreme_lowertail <- sum(result4_H0_true <= -diff_in_sample_props)
count_of_more_extreme_uppertail <- sum(result4_H0_true >= diff_in_sample_props)
bootstrap_pval <- (count_of_more_extreme_lowertail + count_of_more_extreme_uppertail)/num_
sims4
```

```
bootstrap_pval
table(dataset$race.ethnicity)
prop.table(table(dataset$race.ethnicity))
sum(((table(dataset$race.ethnicity) - 200)^2)/200)
pchisq(q = 170.13, df = 5 - 1, lower.tail = FALSE)
results_under_H_0 <- rep(c("A", "B", "C", "D", "E"), 200)
table(results_under_H_0)
num_sims5 <- 10000
#'result4' vector to store my result
chisq_under_H0_true <- rep(NA, num_sims5)
#Loop to complete simulations
for (i in 1:num_sims5) {
  new_sample <- sample(results_under_H_0, 1000, replace = TRUE)
  chisq_under_H0_true[i] <- sum(((table(new_sample) - 200)^2)/200)
}
hist(chisq_under_H0_true, freq = FALSE, main = 'Distribution of Chi-square statistics unde
r H_0', xlab = 'Chi-square under H_0', ylab = 'Density', xlim = c(0,200))
abline(v = sum(((table(dataset$race.ethnicity) - 200)^2)/200), col = "Red")
sum(chisq_under_H0_true >= sum(((table(dataset$race.ethnicity) - 200)^2)/200))/num_sims5
```