STAT 423 - Analysis of Student Exam Performance

Bonita Lo - 30089974
Fabian Pradipto - 30067873
Steven La - 30022373
Terezia Pirhalova - 30089372

University of Calgary

April 12th, 2021

Table of Contents

Introduction

The purpose of our report will be to investigate a student performance data set to understand what might be affecting their test scores. We have obtained the data from kaggle.com. The purpose of this analysis is to gain more insight as to what helps make a difference with student performance and how different variables such as test preparation courses are linked with the test scores.

Our target population consists of highschool students in the United States who have completed the math, writing and reading tests. From these variables, we intend to estimate several parameters which will allow us to draw conclusions about student performance through hypothesis testing. The ideas we intend to investigate which will also be the focus of our investigation consists of whether the mean test scores in the three subject areas are equal or not across the 5 strata. This analysis will give us better insight to how different racial groups generally perform on math, reading and writing tests. We can explore the variance of scores within each stratum and see how each group performs on the three tests and then further look into how different racial group's overall test scores compare to each other. We can explore if they are statistically different or not.

We will also test the proportion of students with prior test preparation within each stratum. This will give us insight as to whether or not test preparation courses are linked with higher scores or not. Lastly, we will investigate the proportion of students with parents coming from an educational background in each stratum. We will compare the test scores of children with parents who have completed some highschool and finished highschool to parents who have completed any degree of university. We will investigate this categorical variable to understand whether there is any relationship between the entities that parent education is linked to how well a student performs on their math, writing and reading tests.

The proportion of students with parents coming from an educational background in each stratum (is parental education related to their child's test performance). The statistics computed from these three topics will allow us to conclude whether student test scores in each stratum are equal or not, whether the groups with higher test scores also have higher proportions of preparation and/or parents with an educational background.

The population parameters we will compute are population mean (μ) for topic 1 and population proportion *(P)* for topics 2 & 3. Since we are using stratified sampling, we will be comparing sub-populations by comparing each individual stratum for the three topics outlined above. This sub-population will be based on the race/ethnicity column from the dataset, meaning there will be 5 strata (group A, group B, group C, group D, group E).

Random Sampling Design and Implementation

Sampling Design:

For our project, we have decided to employ the stratified random sampling method, the data will be stratified according to their race. The reason why we choose stratified sampling is because the data for Race/Ethnicity is distributed as follows:

- Group A = 8.9% = 89 students
- Group B = 19.00% = 190 students
- Group C = 31.90% = 319 students
- Group D = 26.20% = 262 students
- Group E = 14.00% = 140 students
- Population Total = 1000

As we can see, Group A's representation is significantly lower compared to the other groups and we are concerned that students from this group would not be appropriately represented if simple random sampling is employed. We are also ensuring that our sample is proportional to the populations so that we can increase the accuracy of our results.

Sample Size Determination:

For our sample size determination, we have decided on estimating the mean math score of the students within 1.5 points using a 95% level of confidence. Fortunately there were no missing data in our chosen dataset and no additional wrangling is needed for both sample size determination and the actual sampling process. Using R, we have found that the number of samples that we need to take within each group are as follows.

- Group A = 25
- Group B = 52
- Group C = 87
- Group D = 72
- Group E = 38
- Total = Group A + Group B + Group C + Group D + Group E = 274

Implementation:

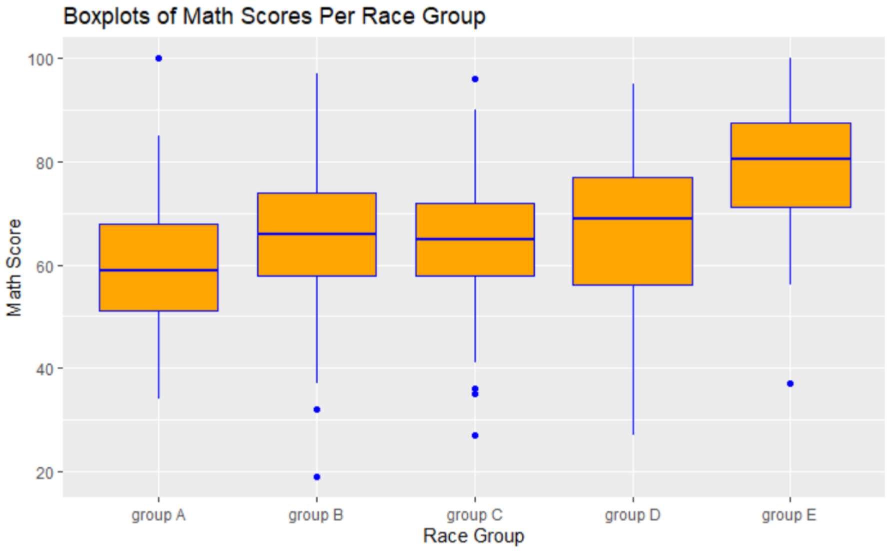
For our implementation,we simply need to separate the StudentsPerformance.csv using the race.ethnicity variable so that we have individual datasets for each group. We then randomly sample students within each group using na, nb, nc, nd, and ne that was calculated in the sample size determination (See Appendix A)

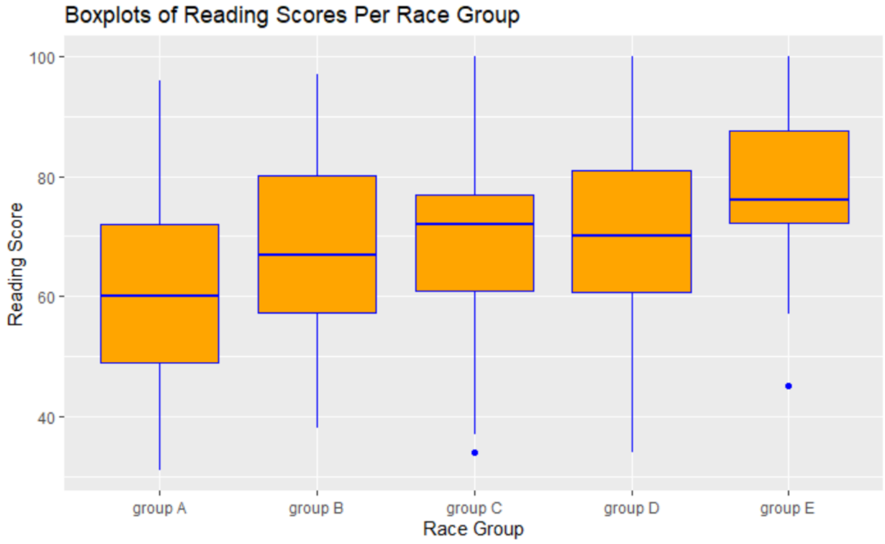
Analysis

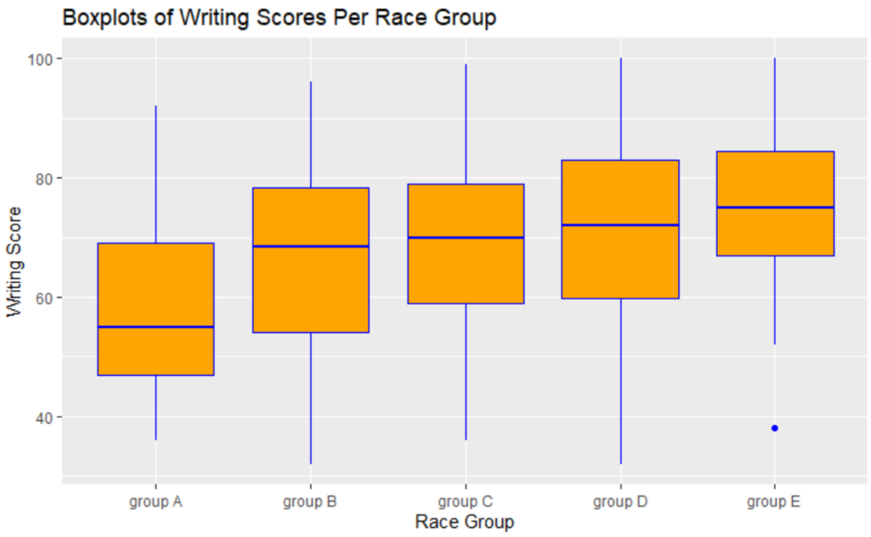
Mean test scores in the subject areas of math, reading, and writing

In this section, we want to estimate whether the mean test scores for math, reading, and writing are the same across the five different race groups. This can be done through hypothesis testing, where we will conduct Levene's test to determine whether the variances are equal across the 5 strata or not.

Our first step is to gain a greater visual understanding of whether the variances are equal across the five groups or not. This can be done by drawing boxplots. We have created three boxplots, in the order of the math, reading, and writing scores as seen below. Just from looking at the boxplots below, it appears that the length of the box plots are alike therefore the variances across the groups are in fact similar.







Our next step would be to conduct a Levene's test to statistically confirm whether the variances are in fact equal or not. Our hypothesis will be set to:

$$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2 = \sigma_E^2$$

$$H_A: \text{at least one } \sigma^2 \text{ is different}$$

After conducting the test on the three score groups separately, we computed the following: (See Appendix B)

A. Levene's test on the math scores:
   - P-value = 0.8859, Test Statistic = 0.28759

B. Levene's test on the reading scores:
   - P-value = 0.2984, Test Statistic = 1.23

C. Levene's test on the writing score:
   - P-value = 0.776, Test Statistic = 0.44503

Given that all the p-values from the three Levene's tests conducted were > 0.05, we therefore fail to reject the null hypothesis and we can conclude that the variances across all five groups are equal.

Now that we have concluded that the variances are equal, we can conduct an ANOVA test to determine whether the mean scores are equal across the five groups or not. We will again use hypothesis testing where:

$$H_0: \mu_A = \mu_B = \mu_C = \mu_D = \mu_E$$

$$H_A: \text{at least one } \mu \text{ is different}$$

From our ANOVA results, we got the following results: (See Appendix B)

A. Math scores:
   - p-value = 0.000000777, F-value = 9

B. Reading scores:
   - p-value = 0.0000292, F-value = 6.846

C. Writing scores:
   - p-value = 0.000317, F-value = 5.44

Given that all the p-values are < 0.05, we therefore reject the null hypothesis, and can conclude that the mean scores across the five strata groups are in fact not equal to each other. We know that the mean scores are not equal. Therefore, we can compute individual confidence intervals for each group, using an overall 95% level of confidence. This will allow us to gain additional insight into what the mean scores within each group is, and where the differences are.

We obtained the following results: (See Appendix B)

1. Math scores:
    a. Group A: $57.52 \leq \mu \leq 64.32$
    b. Group B: $61.44 \leq \mu \leq 68.44$
    c. Group C: $61.29 \leq \mu \leq 67.72$
    d. Group D: $63.20 \leq \mu \leq 70.25$
    e. Group E: $76.02 \leq \mu \leq 82.40$
2. Reading scores:
    a. Group A: $57.48 \leq \mu \leq 64.76$
    b. Group B: $65.02 \leq \mu \leq 71.90$
    c. Group C: $67.37 \leq \mu \leq 73.30$
    d. Group D: $66.55 \leq \mu \leq 73.37$
    e. Group E: $76.06 \leq \mu \leq 82.05$
3. Writing scores:
    a. Group A: $54.91 \leq \mu \leq 61.81$
    b. Group B: $63.00 \leq \mu \leq 71.08$
    c. Group C: $65.87 \leq \mu \leq 72.43$
    d. Group D: $66.59 \leq \mu \leq 74.02$
    e. Group E: $71.86 \leq \mu \leq 78.61$

With all the individual confidence intervals computed for each group, it confirms our results earlier from the ANOVA test that the mean scores are in fact unequal. Further insights from the computations above, we can clearly see that group E is the highest performing group across all three subject areas, group B, C, and D score within similar ranges, and group A consistently scores the lowest. This lets us perform further analysis, for example, does group E have a higher proportion of test preparation leading to higher scores?

To add further context, the overall 95% confidence interval across all five stratum was computed.

We got the following results (See Appendix B):

A. Math scores:
    - $65.46 \leq \mu_{\text{Math\_Score}} \leq 68.36$
B. Reading scores:
    - $68.88 \leq \mu_{\text{Reading\_Score}} \leq 71.68$
C. Writing scores:
    - $67.35 \leq \mu_{\text{Writing\_Score}} \leq 70.34$

With the results above, we can conclude that:

- With 95% confidence, the mean math score for all students was between 65.46% and 68.36%
- With 95% confidence, the mean reading score for all students was between 68.88% and 71.68%
- With 95% confidence, the mean writing score for all students was between 67.35% and 70.34%

As a result of our statistical analysis, we now understand that racial groups do not in fact score the same in math, reading, and writing. However, as a result of stratified sampling, we have a representative sample, and can estimate the overall averages for all students regardless of race.

Sample proportion of students with prior test preparation

In this section we will be exploring the sample proportion of students who have taken prior test preparation courses within each stratum.

Using R Studio , we have found that the result of our confidence interval is as follows: (See Appendix C)

$$0.2846149 \leq \text{test.preparation} \leq 0.3800424$$

From our test we can conclude that with 95% confidence that the sample proportion of students with prior test preparation within each stratum is somewhere between 0.2846149 and 0.3800424. Only about 28.46% to 38.00% of students chose to do test preparation courses before completing these exams.

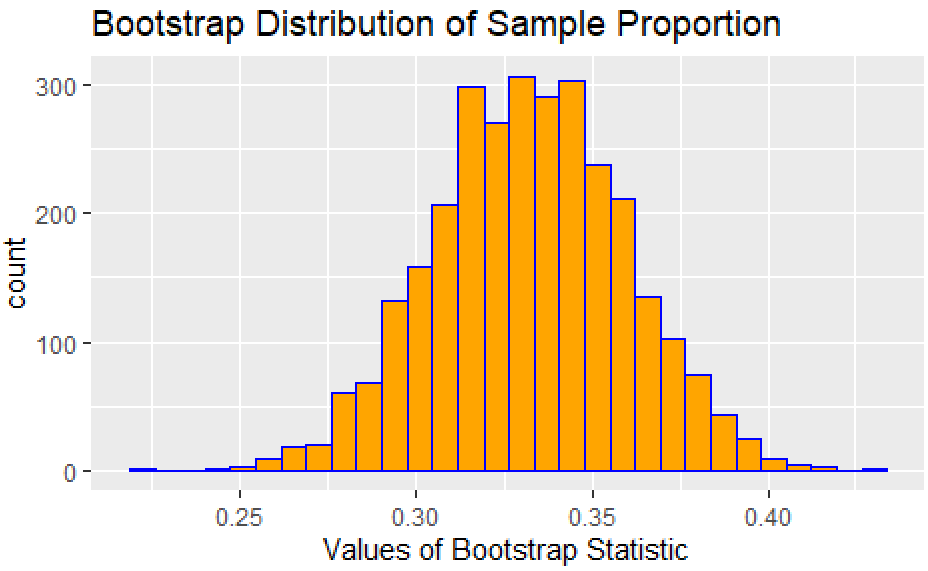
Furthermore, from the previous section, we concluded that group E had the highest scores, while group A consistently scored the lowest. We can further analyze this data, and determine the overall 95% confidence intervals for the percentage within each group that had prior test preparation.

From our results, we got the following intervals: (See Appendix C)

A. Group A: $0.2705 \leq p \leq 0.3941$
B. Group B: $0.2723 \leq p \leq 0.3923$
C. Group C: $0.2681 \leq p \leq 0.3966$
D. Group D: $0.2730 \leq p \leq 0.3917$
E. Group E: $0.2676 \leq p \leq 0.3971$

From the confidence intervals computed above, we can see that each group has almost the same proportion of test preparation. This leads us to the conclusion that perhaps prior test preparation is not a factor in how well a student performs after all.

We decided to additionally bootstrap our data regarding prior test preparation. Bootstrapping is an important resampling technique to estimate statistics on a population by sampling a dataset with replacement. For this bootstrap, we have used a sample size of 274 and have done 3000 repeats.



From the graph above, we can see that the bootstrap distribution of the sample proportion is approximately normally distributed. The overall proportion of students from the dataset that had prior preparation was 33.23%. As a result, a data vector containing 91 1s (yes) and 183 0s (no) was created. As a result, the 95% confidence interval computed wa:

$$0.2774 \leq p \leq 0.3867$$

Comparing our results to the confidence interval computed from stratified sampling, we can see that they are very similar, within a 0.01 range for the lower bound, and 0.006 for the upper bound.

As a result, if we did not have access to the dataset, bootstrapping would be a good alternative in performing our data analysis.

Analysis of parental level of education

In this portion, we are interested to investigate how parental level of education relates to the student's math, reading, and writing score. We wanted to investigate the proportion of students whose parents had some kind of college education and whether or not there is a difference in the scores between students with parents who have a different level of education. From our dataset we know that there are 6 different "parental.level.of.education" variable which includes:

- Associate's Degree
- Bachelor's Degree
- High School
- Master's Degree
- Some College
- Some High School

For the proportion analysis we will merge the 6 categories into binary variables which are divided as follows.

$n_{\text{College}} = n_{\text{Associate's degree}} + n_{\text{Bachelor's degree}} + n_{\text{Master's degree}} + n_{\text{Some college}}$

$n_{\text{High school graduates}} = n_{\text{High school}} + n_{\text{Some high school}}$

So that $\widehat{p}_{college} = \frac{n_{Associate} + n_{Bachelor's} + n_{Master's} + n_{Some\ college}}{n_{Student}}$

Proportion Estimate

Our first step will be to calculate the proportion estimate for each race group with parents that have some level of college education: (See Appendix B)

Group A = 44%
Group B = 63.462%
Group C = 60.920%
Group D = 66.667%
Group E = 78.947%

Hypothesis Test:

We will then test whether or not there is a difference between p.college and p.high_school with 95% confidence using the following hypotheses.
We are now interested to see the difference in proportion between students whose parents have some sort of college education compared to students whose parents didn't go to college.

We will first test whether or not there is a difference between p.college and p.high_school using the following hypotheses.

$$H_0: P_{College} = P_{High\ School}$$
$$H_A: P_{College} \neq P_{High\ School}$$

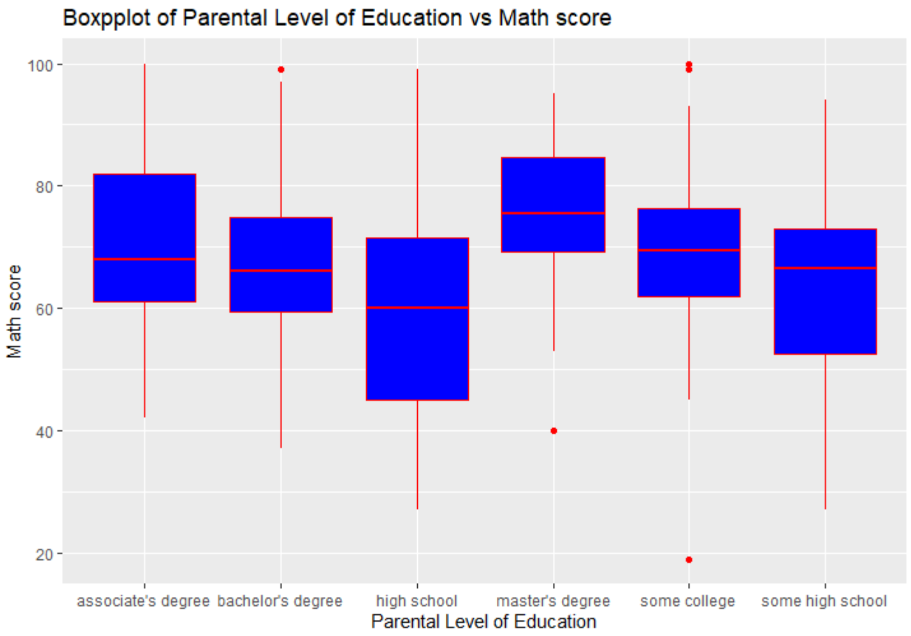
We know that the **Test statistics** = 105.005 and **p-value** = 0, since $0 < 0.05$ we reject our null hypothesis and conclude that the proportion between P_{College} and $P_{\text{High School}}$ differs (See Appendix D)

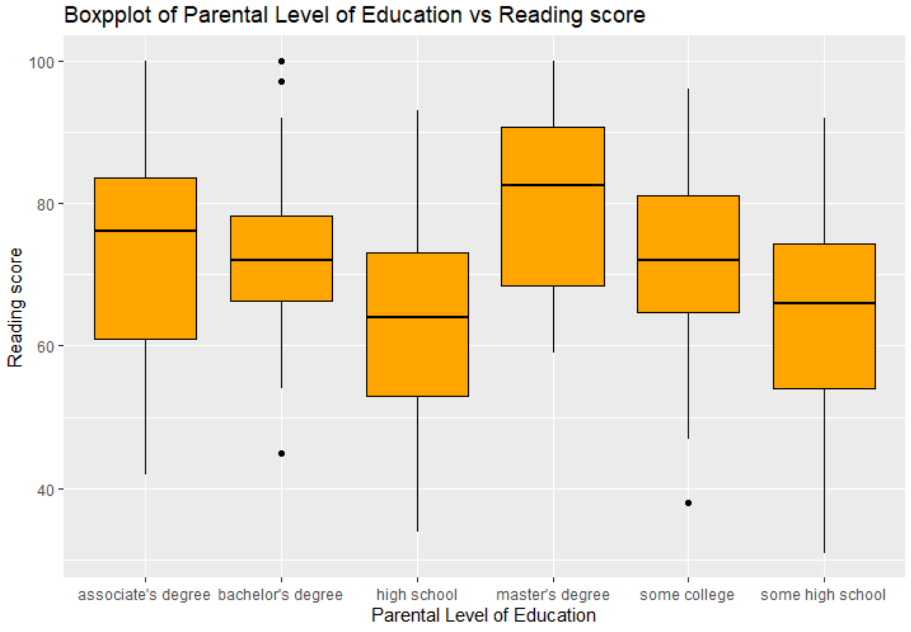
With the differing proportions, we are interested to estimate the P_{College} with 95% level of confidence which yields the following (See Appendix D):

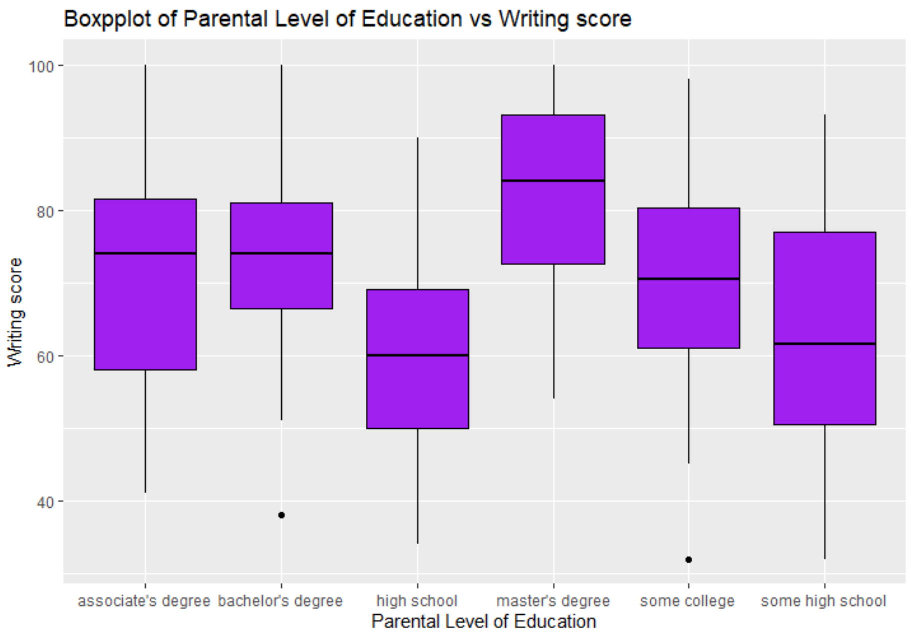
$$0.590 \leq p_{\text{College}} \leq 0.687$$

This means that we are 95% confident that the proportion of students whose parents have some sort of college education is between 0.590 and 0.687.

To further our investigation, we will next analyze how the parental.level.education relates to the student's math, reading, and writing score. We will be using the ANOVA test to see whether or not there is a difference in math.score, reading.score, and writing.score. Before testing our hypothesis we will need to plot the boxplot and run a Levene's test on each score.







From the graphs above, it appears that the length of the box plots are alike therefore the variances across the groups are in fact similar. Our next step would be to conduct a Levene's test to statistically confirm whether the variances are in fact equal or not. Our hypothesis will be set to:

Our Levene's test will use the following hypothesis for all 3 scores

$$H_0: \sigma_{\text{Associate's degree}}^2 = \sigma_{\text{Bachelor's degree}}^2 = \sigma_{\text{High school}}^2 = \sigma_{\text{Master's degree}}^2 = \sigma_{\text{Some college}}^2 = \sigma_{\text{Some high school}}^2$$

$$H_A: H_0 \text{ is false}$$

The result of out test is as follows (See Appendix D):

A. Levene's test on the math scores:
   - Test Statistic (F-value) = 1.7123, P-value = 0.1319
B. Levene's test on the reading scores:
   - Test Statistic (F-value) = 1.2962, P-value = 0.2658
C. Levene's test on the writing score:
   - Test Statistic (F-value) = 1.1089, P-value = 0.3559

Since all p-values > 0.05, we fail to reject the null hypothesis and conclude that the variance for the parental level of education are approximately equal for all 3 scores. We can also observe that students whose parents have a master's degree seems to be doing well when compared to other levels of education, we will now be testing our ANOVA with the following hypothesis

$$H_0: \mu_{\text{Associate's degree}} = \mu_{\text{Bachelor's degree}} = \mu_{\text{High school}} = \mu_{\text{Master's degree}} = \mu_{\text{Some College}} = \mu_{\text{Some high shool}}$$

$$H_A: \text{at least one } \mu \text{ is different}$$

Here are our results from the ANOVA test (See Appendix D):

A. Math scores:
   - F value = 5.827, p-value = 3.96 x 10^{-5}

B. Reading scores:
   - F value = 7.916, p-value = 5.67 x 10^{-7}

C. Writing scores:
   - F value = 11.18, p-value = 8.48 x 10^{-10}

Since all the p-values < 0.05, we can reject our null hypothesis and conclude that the math, reading, and writing score differs between the different parental levels of education. From this analysis, it is evident that the level of parental education among the students affects their exams scores.

Estimation of μ using ratio estimation and regression estimation.

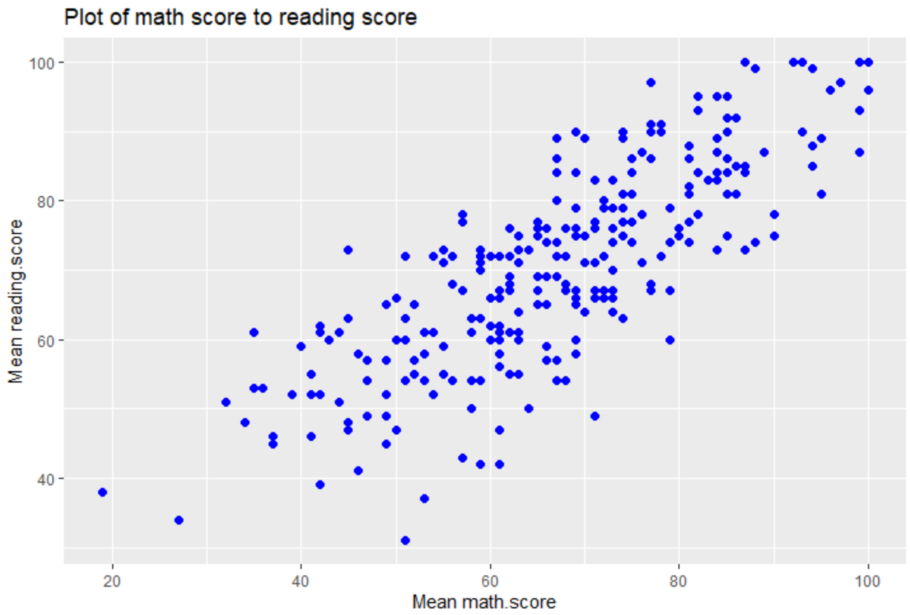
We are now going to explore several estimation techniques to find out which one is better for the data we used. We are going to use ratio and regression estimation to estimate $\mu_{math\ score}$.

Ratio Estimation

In ratio estimation we are estimating $\mu_{math\ score}$ with the help of a second variable so we can synchronize them both and create a more accurate estimate. The variables Y and X are defined as follows:

- Y_i: The main variable (number or categorical) on population element i
- X_i: The value of a subsidiary variable (numerical or categorical) on population element i

We will use the math score as Y and the reading score as X.



These variables also have coefficient correlation value r = 0.801, which indicates a fairly strong correlation.

The ratio estimate between the two variables is as follows (See Appendix E).

$\widehat{R} = \frac{Y}{X} = \frac{Mean\ of\ math.score}{Mean\ of\ reading.score} = 0.952$

We will also need the following values to calculate the confidence interval for both our ratio and mean of math score

- Variance of our ratio estimate = 84.228
- Population mean of our x variable = 69.169
- Sample size = 274
- Population size = 1000

Then we know that the 95% confidence interval of the estimate is:

$$0.939 \leq \widehat{R} \leq 0.965$$

This means we are 95% confident that the value of $\widehat{R}$is between 0.939 and 0.965.

We also know that our 95% confidence interval for the mean math score is

$64.930 \leq \mu_y \leq 66.782$

Hence we are 95% confident that the mean of math score is between 64.930 and 66.782.

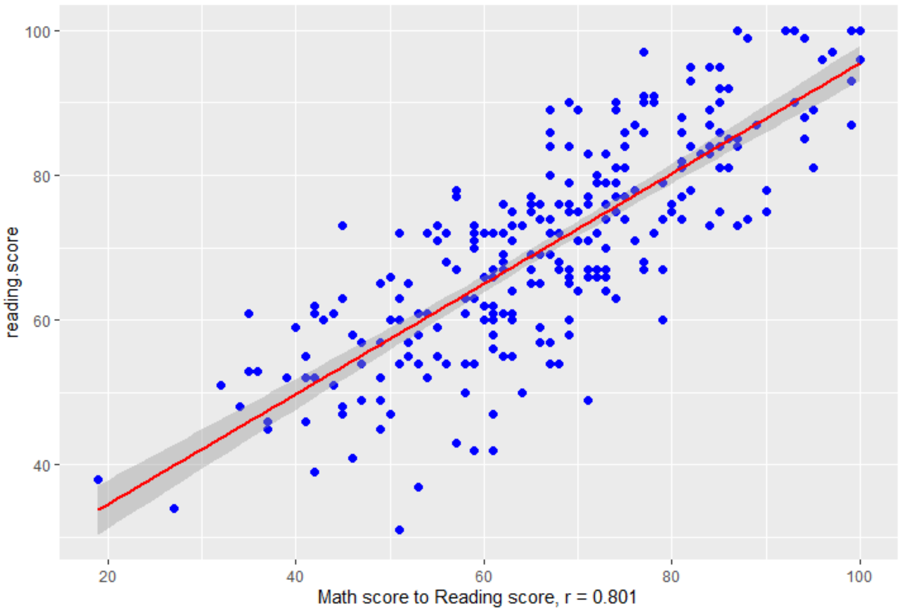
In addition to ratio estimation, we will also be using regression estimation to estimate math score as well and compare the two results

Regression Estimation

For regression estimation we will be using the same Y and X variable to estimate the mean math score so that

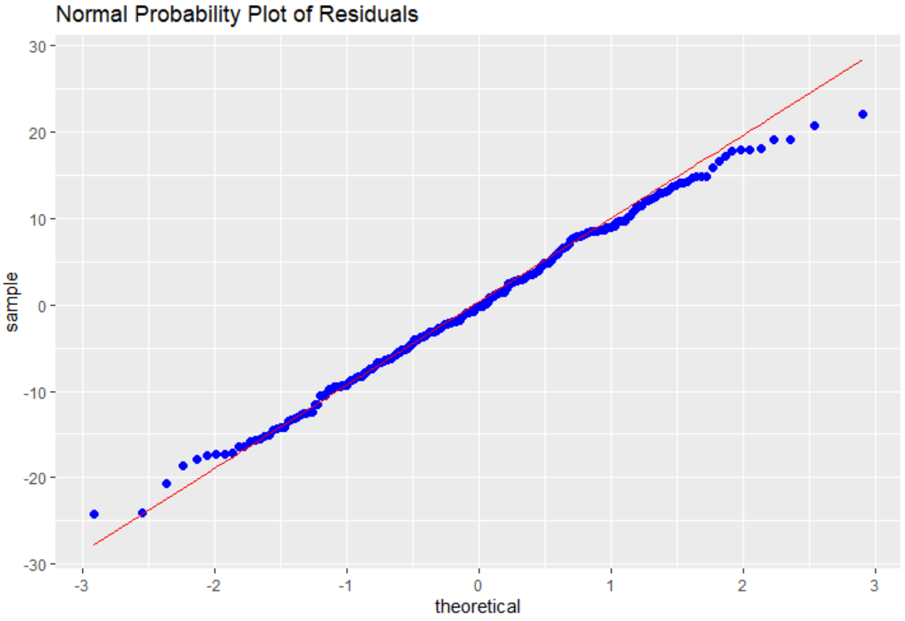
- Y_i: Math score
- X_i: Reading score

This is because we suspect that there is an intimate relationship between the two variables so that the student who did well in math may also do well in reading. The following are the scatterplot of math score to reading score with a “least-squares” regression line.

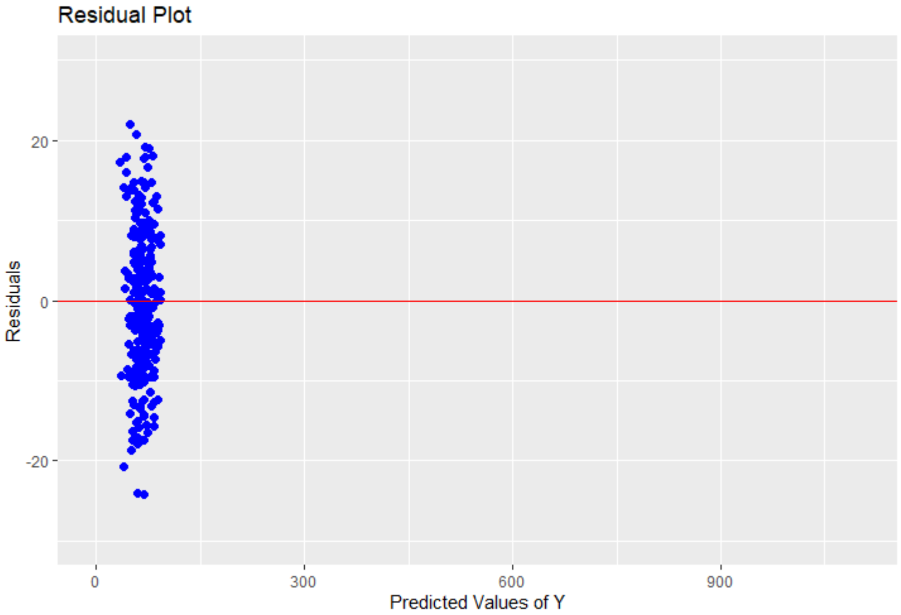


We can see that there is a strong correlation and a fairly linear relationship between the two variables which may indicate that regression estimation will provide an even more accurate 95% confidence interval for the mean math score.

Before we proceed on using regression estimation we will be checking for normality of residuals and homoscedasticity using the following plots.



We can see that the residuals follow the line and that it is approximately normal.



We can see that this graph does not have a rectangular shape which indicates that the data may not necessarily be homoscedastic. Despite this fact, we will proceed on using regression estimation for our mean math score.

To use the regression estimation we will need to create a model from the 2 variables which have the following result (See Appendix E).

$Math\ score\ =\ 7.686\ +\ 0.843(Reading\ Score)$

- Intercept/β_0 = 7.686
- Reading.score/β_1 (slope) = 0.843
- mse.math = $\frac{22320.28}{272}$= 82.060
- Margin of error (math) = 0.918

We also know that our mean math score using the regression estimate is $\mu_{Math}\ =\ 65.974$and our 95% confidence interval for the math score is

$65.056\ \leq\mu_y\leq\ 66.892$

We can see that the interval for the regression estimate is slightly tighter, to find out which is more efficient we can calculate the relative efficiency using the following formula.

<u>Relative Efficiency:</u>

The last step will be to calculate relative efficiency to compare which statistical approach is the most efficient.

$RE\ \ =\frac{MSE}{s_R^{\ 2}}=\frac{mse.math}{varrhat.student}\ =\ 0.974$

Since 0.974 < 1, we can conclude that regression estimation is more efficient than ratio estimation for this particular dataset and that if we wanted to estimate the mean of math score using ratio estimation we would only need 97.4% of the data points if we were to use regression estimation.

Conclusion

After conducting several tests and analysis on the student performance data set that was focused on gaining more insight on what factors may influence a high school students performance in writing exams of the three subjects of math, reading, and writing within the United States, we are able to compare the results to have a better understanding for the topics we decided to investigate.

The first topic we decided to investigate is whether the mean test scores in the three test subjects were equal or not across all five race groups. We can see that through hypothesis testing that we were able to conclude that mean test scores are in fact not equal and that there were race groups that clearly scored higher than other race groups consistently across the three subjects which can be seen in the detailed analysis above. In this case we can say that race does play a role in what the student scores on a test, meaning we can expect a student's test score to be in a range of somewhere depending on their racial group.

The second topic we decided to investigate is the proportion of students with prior test preparation and their test scores on the three subjects to see if test preparation may be linked to higher exam scores. We can see from the detailed analysis above that the overall proportion of students that had test preparation was 33.32%. We can then use this information to compare it with our results in the first topic above with the mean test scores across the five race groups. In our analysis we can see that students with prior test preparation does not affect a students performance on the tests.

The third topic we decided to investigate is the proportion of students with parents from an educational background (i.e whether they went to college or not at all) to see if it may be linked to higher test scores or not. Then we investigate the mean of test scores for the parental level of education of whether they went to college or not at all. We can see from the detailed analysis above that the proportion of students is likely between 59% and 69%. We then did hypothesis testing of mean scores and were able to conclude that at least one mean test score is different across the different parental educational levels. In this case we can say from our analysis that test scores across all three subjects differ depending on the parental level of education variable.

Lastly we tested both ratio and regression estimation to see which of the two would be more efficient for our data. From the above testing and detailed analysis we were able to conclude that regression estimation is more efficient.

Since we were able to successfully test this data and compare the results for a students performance on three test subjects of math, reading and writing based on several differentiating factors amongst the students, we can conclude that there are indeed several factors that may or may not affect a students performance on test scores of these three subjects, which we have analysed in depth for several topics.

R Studio Appendix

Appendix A - Random Sampling Design and Implementation

```r
student.df = StudentsPerformance
student.df
favstats(math.score~race.ethnicity ,data = student.df)

Nis = c(89, 190, 319, 262, 140)
ais = c(0.089, 0.19, 0.319, 0.262, 0.14)
sigmais = c(14.52, 15.47, 14.85, 13.77, 15.53)
moestudent = 1.5
studentpart1 = sum(((Nis^2)*(sigmais^2))/(ais))
studentpart2 = (sum(Nis)^2)*(moestudent/qnorm(0.975))^2
studentpart3 = sum(Nis*sigmais^2)
studentpart1

student.n = studentpart1/(studentpart2 + studentpart3)
student.n

student.na = ais[1]*student.n
student.na = ceiling(student.na)
student.nb = ais[2]*student.n
student.nb = ceiling(student.nb)
student.nc = ais[3]*student.n
student.nc = ceiling(student.nc)
student.nd = ais[4]*student.n
student.nd = ceiling(student.nd)
student.ne = ais[5]*student.n
student.ne = ceiling(student.ne)
cat("na = ", student.na, "\n")
cat("nb = ", student.nb, "\n")
cat("nc = ", student.nc, "\n")
cat("nd = ", student.nd, "\n")
cat("ne = ", student.ne, "\n")

cat("The overall sample size is", student.na + student.nb + student.nc +
student.nd + student.ne)
```

```
#Separates each groups into a data frame
groupA.df <- student.df[ which(student.df$race.ethnicity=='group A'), ]
groupB.df <- student.df[ which(student.df$race.ethnicity=='group B'), ]
groupC.df <- student.df[ which(student.df$race.ethnicity=='group C'), ]
groupD.df <- student.df[ which(student.df$race.ethnicity=='group D'), ]
groupE.df <- student.df[ which(student.df$race.ethnicity=='group E'), ]

#Randomly selects students according to the sample size determination
groupA_sample.df <- groupA.df[sample(nrow(groupA.df), 25), ]
groupB_sample.df <- groupB.df[sample(nrow(groupB.df), 52), ]
groupC_sample.df <- groupC.df[sample(nrow(groupC.df), 87), ]
groupD_sample.df <- groupD.df[sample(nrow(groupD.df), 72), ]
groupE_sample.df <- groupE.df[sample(nrow(groupE.df), 38), ]

groupA_sample.df

#Combines the randomly sampled data into a single data frame
studentsample.df <- rbind(groupA_sample.df, groupB_sample.df,
groupC_sample.df, groupD_sample.df, groupE_sample.df)
studentsample.df

favstats(math.score~race.ethnicity ,data = studentsample.df)
write.csv(studentsample.df,'studentsample.csv', row.names = TRUE)
```

Appendix B - Mean test scores in the three subject areas of math, reading, writing

```
studentsample = read.csv("C:/Users/bonit/Downloads/studentsample.csv")

# Boxplots for the three scores
ggplot(data = studentsample, aes(x = race.ethnicity, y = math.score)) +
geom_boxplot(col = "blue", fill = "orange") + coord_flip() + xlab("Race Group")
+ ylab("Math Score") + ggtitle("Boxplots of Math Scores Per Race Group")

ggplot(data = studentsample, aes(x = race.ethnicity, y = reading.score)) +
geom_boxplot(col = "blue", fill = "orange") + xlab("Race Group") + ylab("Reading
Score") + ggtitle("Boxplots of Reading Scores Per Race Group")

ggplot(data = studentsample, aes(x = race.ethnicity, y = writing.score)) +
geom_boxplot(col = "blue", fill = "orange") + xlab("Race Group") + ylab("Writing
Score") + ggtitle("Boxplots of Writing Scores Per Race Group")

attach(studentsample)
# Levene’s test
# math
levene.test(math.score, race.ethnicity, location=c("median"), correction.method
= "none")
```

```
# reading
levene.test(reading.score, race.ethnicity, location = c("median"),
correction.method = "none")
# writing
levene.test(writing.score, race.ethnicity, location = c("median"),
correction.method = "none")

# Anova
# Math
summary(aov(math.score~race.ethnicity, data=studentsample))
# Reading
summary(aov(reading.score~race.ethnicity, data=studentsample))
# Writing
summary(aov(writing.score~race.ethnicity, data=studentsample))

# Computation of stratified sampling confidence intervals
H = 5
ni = c(25, 52, 87, 72, 38) #A, B, C, D, E
Ni = c(89, 190, 319, 262, 140)
N = sum(Ni)
n = sum(ni)
yi.math = favstats(~math.score|race.ethnicity, data=studentsample)$mean
yi.reading = favstats(~reading.score|race.ethnicity, data=studentsample)$mean
yi.writing = favstats(~writing.score|race.ethnicity, data=studentsample)$mean
sd.math = favstats(~math.score|race.ethnicity, data=studentsample)$sd
sd.reading = favstats(~reading.score|race.ethnicity, data=studentsample)$sd
sd.writing = favstats(~writing.score|race.ethnicity, data=studentsample)$sd

# Math
ybar = sum(Ni*yi.math)/sum(Ni)
variance = (1/N^2)*sum(Ni^2*((Ni-ni)/Ni)*(sd.math^2/ni))
ybar - qt(0.975, n-H)*sqrt(variance)
ybar + qt(0.975, n-H)*sqrt(variance)

# Reading
ybar = sum(Ni*yi.reading)/sum(Ni)
variance = (1/N^2)*sum(Ni^2*((Ni-ni)/Ni)*(sd.reading^2/ni))
ybar - qt(0.975, n-H)*sqrt(variance)
ybar + qt(0.975, n-H)*sqrt(variance)

# Writing
ybar = sum(Ni*yi.writing)/sum(Ni)
variance = (1/N^2)*sum(Ni^2*((Ni-ni)/Ni)*(sd.writing^2/ni))
ybar - qt(0.975, n-H)*sqrt(variance)
ybar + qt(0.975, n-H)*sqrt(variance)

# individual confidence intervals

# overall level of confidence
level = 1 - (0.95)^5
level = 1 - level
```

```
Level
[1] 0.7737809
# family confidence level
alpha = 1 - (0.95)^(1/5)
alpha
[1] 0.01020622
# t-multiplier
t = qt(1 - alpha/2, 5-1)
t
[1] 4.577047

# math scores for Group A, B, C, D, E
yi.math - t*sqrt((sd.math^2/n)*((N-n)/N))
yi.math + t*sqrt((sd.math^2/n)*((N-n)/N))

# reading scores for Group A, B, C, D, E
yi.reading - t*sqrt((sd.reading^2/n)*((N-n)/N))
yi.reading + t*sqrt((sd.reading^2/n)*((N-n)/N))

# writing scores for Group A, B, C, D, E
yi.writing - t*sqrt((sd.writing^2/n)*((N-n)/N))
yi.writing + t*sqrt((sd.writing^2/n)*((N-n)/N))
```

Appendix C - The proportion of students with prior test preparation within each stratum

```
# stratified sampling confidence interval
ni = c(25, 52, 87, 72, 38) #A, B, C, D, E
Ni = c(89, 190, 319, 262, 140)
N = 1000
n = 274
pi = favstats(~test.preparation.course|race.ethnicity, data =
studentsample)$mean
sd = pi*(1-pi)
phat = sum(Ni*pi)/N
variance = sum((Ni/N)^2*((Ni-ni)/Ni)*(sd/(ni-1)))
phat - qnorm(0.975)*sqrt(variance)
phat + qnorm(0.975)*sqrt(variance)

# individual confidence intervals

pi = favstats(~test.preparation.course|race.ethnicity, data =
studentsample)$mean
sd = pi*(1-pi)
phat = 0.3323286
alpha = 0.01020622
```

```
phat - qnorm(1-alpha/2)*sqrt((sd/(n-1))*((N-n)/N))
phat + qnorm(1-alpha/2)*sqrt((sd/(n-1))*((N-n)/N))

# bootstrap code

ggplot(data = bootstrap, aes(x = sampprop)) + geom_histogram(col =
"blue", fill = "orange") + xlab("Values of Bootstrap Statistic") +
ggtitle("Bootstrap Distribution of Sample Proportion")

Phat = 0.3323286
n = 274
no_test = c(rep(0, n*(1-phat)))
test = c(rep(1, n*phat))
data = c(test, no_test)
Nresamples = 3000
sampprop = numeric(Nresamples)
for(i in 1:Nresamples)
{ yes = sum(sample(data, n, replace=TRUE))
  sampprop[i] = (yes)/n
}
bootstrap = data.frame(sampprop)
qdata(~sampprop, c(0.025, 0.975), bootstrap)
```

Appendix D - Parental level of education

<u>Proportion Estimate</u>

```
q3 = read.csv("~/UofC/STAT423/studentsample_renumbered.csv",
header=TRUE)
q3r = data.frame(q3$race.ethnicity, q3$parental.level.of.education)

q3rs=
aggregate(q3r$q3.parental.level.of.education~q3r$q3.race.ethnicity, FUN
= length)
gAdf = q3r[-c(26:274),]
gAdf

qAdff=
aggregate(gAdf$q3.race.ethnicity~gAdf$q3.parental.level.of.education,
FUN = length)
#group A proportion estimate
gA = 11/25

gBdf = q3r[-c(0:25, 78:274),]
gBdf
```

```
qBdff=
aggregate(gBdf$q3.race.ethnicity~gBdf$q3.parental.level.of.education,
FUN = length)
#group B proportion estimate
gB = 33/52

gCdf = q3r[-c(0:77, 165:274 ),]
qCdff=
aggregate(gCdf$q3.race.ethnicity~gCdf$q3.parental.level.of.education,
FUN = length)
#group C proportion estimate
gC = 53/87

gDdf = q3r[-c(0:164, 237:274),]
gDdf
qDdff=
aggregate(gDdf$q3.race.ethnicity~gDdf$q3.parental.level.of.education,
FUN = length)
#group D proportion estimate
gD = 48/72

gEdf = q3r[-c(0:236),]

qEdff=
aggregate(gEdf$q3.race.ethnicity~gEdf$q3.parental.level.of.education,
FUN = length)
#group E proportion estimate
gE = 30/38
```

Hypothesis Test

```
favstats(math.score~parental.level.of.education ,data = student.df)
phat.associate = 222/1000
phat.bachelor = 118/1000
phat.hs = 196/1000
phat.master = 59/1000
phat.some_college = 226/1000
phat.some_high_school = 179/1000
n.associate = 51
n.bachelor = 34
n.high_school = 51
n.master = 22
n.some_college = 68
n.some_high_school = 48
n.student = n.associate + n.bachelor + n.hs + n.master + n.some_college
+ n.some_hs
```

```
n.college = n.associate + n.bachelor + n.master + n.some_college
n.highschool = n.hs + n.some_hs

p.college = (222+118+59+226)/(1000)
p.highschool = (196+179)/(1000)

phat.prop =
((n.associate*phat.associate)+(n.bachelor*phat.bachelor)+(n.hs*phat.hs)+
(n.master*phat.master)+(n.some_college*phat.some_college)+(n.some_hs*pha
t.some_hs))/(51+34+51+22+68+48)
z.value =
(p.college-p.highschool-(0))/(phat.prop*(1-phat.prop)*((1/n.college)+(1/
n.highschool)))

(1-pnorm(z.value))*2
```

Confidence Interval

```
table(studentsample.df$parental.level.of.education)
table(student.df$parental.level.of.education)
phat.college =
(n.associate+n.bachelor+n.master+n.some_college)/n.student
p.college_size = 1000
s.college_size = 274
moe.college = qnorm(0.975)*sqrt(((p.college_size -
s.college_size)/p.college_size) *
(phat.college*(1-phat.college)/(274-1)))

lb.college = phat.college - moe.college
ub.college = phat.college + moe.college
```

Boxplot

```
#plots
ggplot(data=studentsample.df, aes(x = parental.level.of.education, y =
math.score)) + geom_boxplot(col='red', fill='blue') + xlab("Parental
Level of Education") + ylab("Math score") + ggtitle("Boxpplot of
Parental Level of Education vs Math score")
ggplot(data=studentsample.df, aes(x = parental.level.of.education, y =
reading.score)) + geom_boxplot(col='black', fill='orange') +
xlab("Parental Level of Education") + ylab("Reading score") +
ggtitle("Boxpplot of Parental Level of Education vs Reading score")
ggplot(data=studentsample.df, aes(x = parental.level.of.education, y =
writing.score)) + geom_boxplot(col='black', fill='purple') +
xlab("Parental Level of Education") + ylab("Writing score") +
ggtitle("Boxpplot of Parental Level of Education vs Writing score")
```

Levene’s test and ANOVA

```
#Levene's
leveneTest(math.score ~ parental.level.of.education, studentsample.df)
leveneTest(reading.score ~ parental.level.of.education,
studentsample.df)
leveneTest(writing.score ~ parental.level.of.education,
studentsample.df)

#aov
studentaov_math = summary(aov(math.score~parental.level.of.education,
data = studentsample.df))
studentaov_reading =
summary(aov(reading.score~parental.level.of.education,data =
studentsample.df))
studentaov_writing =
summary(aov(writing.score~parental.level.of.education,data =
studentsample.df))

studentaov_math
studentaov_reading
studentaov_writing
```

Appendix E - Estimation of μ using ratio estimation and regression estimation

```
#Ratio Estimation Estimating math score with reading score as an
estimate.
mean.math = mean(~math.score, data=studentsample.df) #y
mean.reading = mean(~reading.score, data=studentsample.df) #x
rhat.student = mean.math/mean.reading

#scatterplot
ggplot(data=studentsample.df, aes(x = math.score, y = reading.score)) +
geom_point(size=2, col='blue') + xlab("Mean math.score") + ylab("Mean
reading.score") + ggtitle("Plot of math score to reading score")

cor(studentsample.df$math.score, studentsample.df$reading.score)


dev.student = (studentsample.df$math.score  -
(rhat.student*studentsample.df$reading.score))
dev.student
varrhat.student = var(dev.student)
varrhat.student

popmean.x = mean(~reading.score, data=student.df)
popmean.x
mean.math
ssize.student = 274
psize.student = 1000
popmean.math

moe.rhat  = qnorm(0.975)*sqrt((1/(popmean.math^{2}))*((psize.student -
ssize.student)/(psize.student))*(varrhat.student/ssize.student))
moe.rhat

lb.studentr = rhat.student - moe.rhat
ub.studentr = rhat.student + moe.rhat

lb.studentr
ub.studentr

lb.meanstudent.rhat = popmean.x*lb.studentr
ub.meanstudent.rhat = popmean.x*ub.studentr
lb.meanstudent.rhat
ub.meanstudent.rhat

#Regression Estimation
```

```
n.student = 274
studentmodel = lm(math.score~reading.score, data=studentsample.df)
beta0.math = studentmodel$coefficients[[1]]
beta1.math = studentmodel$coefficients[[2]]
#scatterplot


aov(studentmodel)

mse.math = 22320.28/272
meanmath.regression = mean.math + (beta1.math*(popmean.x -
mean.reading))
meanmath.regression

moe.meanmath = qt(0.975, 274-2)*sqrt(((psize.student -
ssize.student)/(psize.student)) * (mse.math/ssize.student))
moe.meanmath

lb.meanmathregression = meanmath.regression - moe.meanmath
ub.meanmathregression = meanmath.regression + moe.meanmath

lb.meanmathregression
ub.meanmathregression

#Relative Efficiency
var(dev.student)/mse.math
```

References

Lohr, S. L. (2019). *Sampling Design and Analysis*. CRC Press.

Seshapanpu, Jakki. “Students Performance in Exams.” *Kaggle*, 9 Nov. 2018, www.kaggle.com/spscientist/students-performance-in-exams.

Stallard, Jim. “ Partially Completed Lecture Notes - February 1. ” D2L,1 Feb. 2021, d2l.ucalgary.ca.

Stallard, Jim. “ Partially Completed Lecture Notes - February 3. ” D2L, 3 Feb. 2021, d2l.ucalgary.ca.

Stallard, Jim. “ Partially Completed Lecture Notes - February 8. ” D2L, 8 Feb. 2021, D2l.ucalgary.ca.

Stallard, Jim. “ Partially Completed Lecture Notes - February 22. ” D2L, 8 Feb. 2021, D2l.ucalgary.ca.

Stallard, Jim. “ Partially Completed Lecture Notes - March 10. ” D2L, 8 Feb. 2021, D2l.ucalgary.ca.