Grau en Matemàtica Computacional i Analítica de Dades

Students Exam Scores

Anàlisi de Dades Complexes

Judit Yebra Valencia (1603614)

Índex

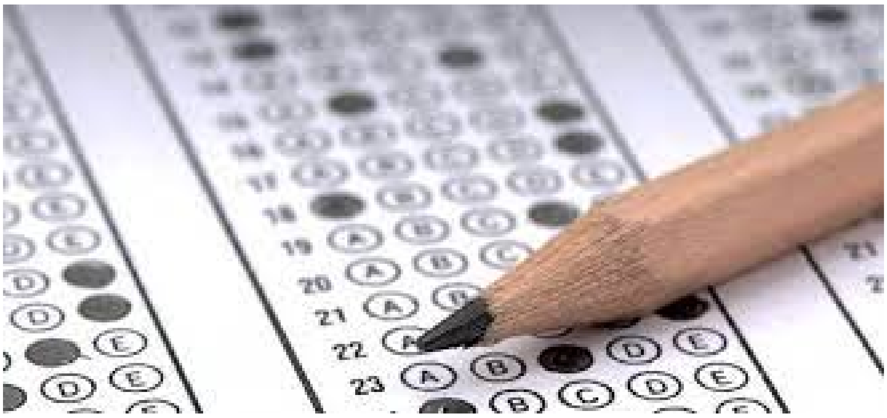
1 INTRODUCTION OF THE TOPIC

For thousands of years, schools have been a crucial component of human civilization, acting as institutions that impart knowledge and direct personal growth. From the Xia monarchy to the present day, the importance of schooling in society has remained steadfast. Over the time, the concept of teaching has evolved, becoming not only a place to acquire simple understanding but also a necessary and basic right for every individual.

However, the justice of the training method has become a subject of concern in recent times. The huge pressure on students to achieve great marks, do particular levels, and become recognized as the best has given rise to a disturbing reality: many students struggle with anxiety as they try to exceed themselves, frequently ignoring the difficulties they encounter outside the classroom. It is crucial to recognize that every person navigates special circumstances, and their struggles extend beyond the intellectual realm. In light of these considerations, a crucial question arises: Is the schooling system actually fair?

To explore this query, it is essential to delve into the world of data analysis. We can learn more about the elements that contribute to or hinder academic success by looking at various factors that affect school performance. We can look into the justice of the educational system from an experimental perspective by analyzing a dataset that includes various parameters, including socio-economic background, access to help at home, ethnic background, gender and more.

In conclusion, while schools have evolved to provide education and equal opportunities, the fairness of the educational system is a topic of debate. While efforts are being made to address disparities and accommodate diverse needs, there are still inherent challenges in ensuring fairness for every individual. This is expected to be proved through the analysis of a dataset which contains students' marks and some of their personal information.



2 DATA AND INFORMATION

2.1 Purpose of the project

In this project, some data about students, their marks and different factors of their lives will be analyzed, to check if it's true that some factors do make life at school easier.

2.2 Description of the data

The database that will be used for this project has been extracted from here. This dataset is conformed by over 30 thousand rows, where each one represents a different person, the information about them is divided in 9 columns, which are the following:

- **Col.1: ID:** The first column is basically an ID to differentiate each person.
- **Col.2: Gender:** The second column is the gender of the person only including males and females, it's written, so it must be changed to numbers.
- **Col.3: EthnicGroup:** The third column is the ethnic group to which the person belongs, it's also written, so it must also be changed to numbers. It has 5 possible groups.
- **Col.4: ParentEduc:** The fourth column is the educational background of the parents, it's also written, so it must be changed to numbers. It has 6 possible options which are: some high school, high school, some college, associate's degree, bachelor's degree and master's degree.
- **Col.5: LunchType:** The fifth column refers to the type of lunch the student has at school (it helps to understand the economic background of the person). It's also written, so it must be changed to numbers, there are only two possible outcomes, which are standard lunch or free/ reduced.
- **Col.6: TestPrep:** The sixth column refers to whether the student has the preparation course completed or not. It's also written, so it must be changed to numbers.
- **Col.7: MathScore:** The seventh column is the score 0-100 on the math test.
- **Col.8: ReadingScore:** The eight column is the score 0-100 on the reading test.
- **Col.9: WritingScore:** The ninth column is the score 0-100 on the writing test.

This dataset has been specifically selected without nulls (there was a bigger version with nulls), since the cleanse of the dataset is not the important part of the project. On the other hand, this dataset has a lot of columns where the information is written, so the first thing that has to be done it's the transformation into numbers so that the information is useful, this will be made through the creation of 'dummmy' variables, using the "One-Hot Encoding"method. It's not strictly necessary for what we're going to do, but it makes it clearer. The ID will also be removed since it's a numeration of the dataset, and it hasn't got any important information about the students.

The first column that will be changed is the Gender, whether it's a male or a female. Afterwards, the Ethnic Group will also be changed and classified from group A to E. The Parents' Education also has 6 possible options, being some high school, highschool, some college, associate's, bachelor's and master's. The lunch type it only has two options which are free/reduced and standard. Lastly, the TestPrep also has two options, whether it's completed or not.

Finally, the last three scores will be added up as a mean, since we want to analyze the marks in general and not in every topic.

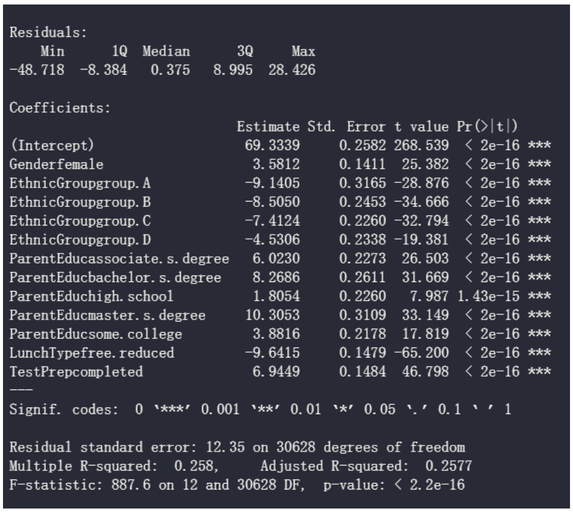
3 INITIAL APPROACH TO THE DATA

Once all the data is clean, the data analysis may start.

Firstly, as previously explained, there has been a cleanse. The index column has been removed, the marks have been added as a mean between the three and an onehotencoding has been done to the categorical variables.

3.1 Backward selection and the importance of the variables

Afterwards, a model for lineal regression has been made. This is useful to check which variables are actually important to estimate the mean score of the exams. Using the function AIC the result of the backward selection is that all the variables in the dataset are very influential, since all the pvalues are much lower than 0.05 and even 0.001 (that's why all of them have 3 stars next to them). Here it can be seen how the backward selection hasn't removed any of the variables:



```
Residuals:
    Min      1Q  Median      3Q     Max
-48.718  -8.384   0.375   8.995  28.426

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   69.3339     0.2582 268.539  < 2e-16 ***
Genderfemale                   3.5812     0.1411  25.382  < 2e-16 ***
EthnicGroupgroup.A            -9.1405     0.3165 -28.876  < 2e-16 ***
EthnicGroupgroup.B            -8.5050     0.2453 -34.666  < 2e-16 ***
EthnicGroupgroup.C            -7.4124     0.2260 -32.794  < 2e-16 ***
EthnicGroupgroup.D            -4.5306     0.2338 -19.381  < 2e-16 ***
ParentEducassociate.s.degree   6.0230     0.2273  26.503  < 2e-16 ***
ParentEducbachelor.s.degree    8.2686     0.2611  31.669  < 2e-16 ***
ParentEduchigh.school          1.8054     0.2260   7.987 1.43e-15 ***
ParentEducmaster.s.degree     10.3053     0.3109  33.149  < 2e-16 ***
ParentEducsome.college         3.8816     0.2178  17.819  < 2e-16 ***
LunchTypefree.reduced         -9.6415     0.1479 -65.200  < 2e-16 ***
TestPrepcompleted              6.9449     0.1484  46.798  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.35 on 30628 degrees of freedom
Multiple R-squared:  0.258,     Adjusted R-squared:  0.2577
F-statistic: 887.6 on 12 and 30628 DF,  p-value: < 2.2e-16
```

Figura 1: Results of the backward selection

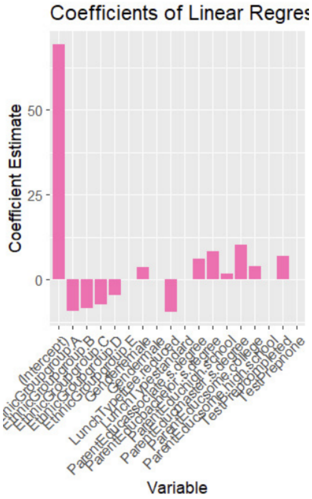


Figura 2: Correlation of the variables

In figure number 2, it can be seen the correlation between the variables (in the x axis) and the coefficient estimate, which is the mean score of the marks of the exams, (in the y axis).

In this graphic it is shown that the variables that make people who have them probably have higher marks than the mean are the people who have parents with master's degrees, bachelor's, and associate's degrees and also those who have completed the preparation for the tests. On the other hand, people who are from the Ethnic Groups A, B, C and D, and people who have the free or reduced lunch types have probably worse marks than the mean.

From this information it can be deduced that the ethnic group of the person has a very bad influence on their marks (however, we can't really analyze this information with a goal in mind since the ethnic groups are classified as letters and not as real ethnic groups nevertheless it can be really seen that being from the ethnic group E has a very nice influence). Another variable that is very low is the free meal, meaning that children who come from families with less money tend to do worse in school. On the other hand, the variables that are high are obviously the test preparation (of course someone will do better in a test if they're prepared for it), and it can also be seen that people whose parents have higher studies (master's, bachelor's and associate's) tend to do much better in school.

3.2 Residuals and plots

Afterwards, there has also been a search for residuals, since those could make the estimation a little bit wrong. However, after inspecting the variables and the correlation of the final mark in the dataset it has been decided it's not necessary to do a cleanse of residuals because there's not a mark that is weirdly surprising or that seems wrong. Here is a plot of the residuals:

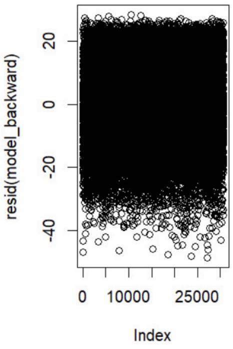


Figura 3: Plot of the residuals

In this plot it can be seen how there are some marks that don't behave as they should because at the end of the day people are not machines, so it's not completely predictable their behaviour. However, none of them has been considered very alarming, so the residuals have remained in the dataset to make it more realistic.

These are the plots of the variable MeanScore which is the mean of the three marks of the exams. The first plot is a boxplot, which shows exactly which is the mean and where is the highest density of marks inside a box. The second plot is a histogram, it shows which marks have the highest frequency and it can be really good to observe where the mean is.

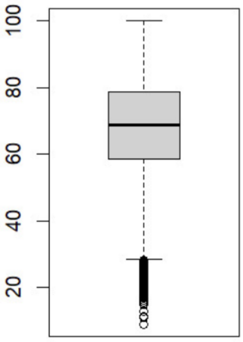


Figura 4: Boxplot of MeanScore

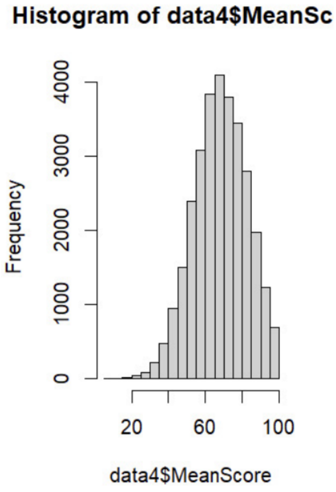


Figura 5: Histogram of MeanScore

3.3 Mean and median

3.3.1 Mean

The mean is the result of adding all the marks and dividing the result by the number of marks. This number can be useful to see approximately how the marks of the exam have been and to get a general idea of how the marks are.

In the following plots there's a boxplot of the mean where it can be seen that the mean is a little bit less than 68.3, next to it there's a histogram where it can also be seen that this mark has been obtained by approximately 250 students. Taking this information into account, the ICs will be judged as if having less than 68.3 was worse than the mean and having more is better, to understand the effects of the variables.

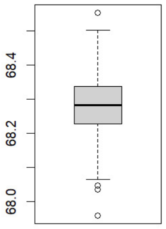


Figura 6: Boxplot of the mean

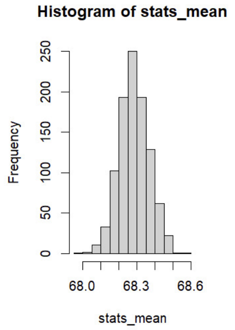


Figura 7: Histogram of the mean

3.3.2 Median

The median is the number (or numbers (two) if the total is odd), that stands exactly in the middle. This number is quite helpful to understand if the marks are very drastically opposed or almost all of them the same.

From the following plots it can be seen how the total number of marks is odd (that's why there are two medians) and that the median seems to be a little lower than the mean, which means that there are more very high marks than very low marks.

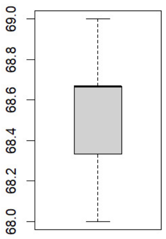


Figura 8: Boxplot of the median

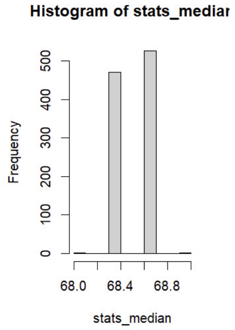


Figura 9: Histogram of the median

4 NON-PARAMETRIC BOOTSTRAP

4.1 What is it?

Non-parametric bootstrap is a statistical technique used to estimate the sampling distribution of a statistic **without making assumptions about the distribution**. It is particularly useful when the population distribution is unknown.

The non-parametric bootstrap works by **resampling** from the available data to create a large number of bootstrap samples. Each bootstrap sample is generated by randomly selecting observations from the original dataset with replacement, meaning that each observation has an equal chance of being selected multiple times or not at all. This process creates synthetic datasets that mimic the original data's characteristics.

The **statistic of interest**, such as a mean, median, or regression coefficient, is then computed for each bootstrap sample. This generates a distribution of bootstrap statistics, which provides an approximation of the sampling distribution of the statistic. From this distribution, confidence intervals and standard errors can be estimated, allowing for inference and hypothesis testing.

4.2 Goal of the research

As it has been previously noted, the parents' studies affect a lot on their childrens' marks. Now six different IC will be calculated: one for every kind of study group that there are in this dataset, which are the following: some high school, high school, some college, associate's degree, bachelor's degree and masters' degree.

To do this the rest of the variables have been maintained the same through all the research. The conditions imposed have been gender female, ethnic group A, lunch type free or reduced, and test preparation completed. The research has been done with a thousand reps and the IC has a 95% confidence.

From the table below it can be seen how there's more than ten points of difference in the mean between the lowest type of studies (some high school) and the highest (master's degree), and also how the highest number in the IC of some high school is the lowest number in the Ic of Bachelor's and Master's degree. From all of this information, we can conclude that the parents' studies really do have a huge influence in their children's marks. This can maybe be because they have help at home, or maybe because of the socio-economic status, which will be checked later.

	Mean	STD	IC
Some High School	62	2.19	(58,67)
High School	63	2	(59,67)
Some College	65	2.91	(59,71)
Associate's	69	2.23	(65,74)
Bachelor's	72	2.41	(67,77)
Master's	74	3.26	(67,80)

5 PARAMETRIC BOOTSTRAP

5.1 What is it?

Parametric bootstrap is a statistical technique used to estimate the sampling distribution of a statistic by **assuming a parametric model**. It is an extension of the non-parametric bootstrap that incorporates model assumptions to generate bootstrap samples.

In the parametric bootstrap, the first step is to fit a parametric model to the observed data. This model specifies the **assumed distribution** and its associated parameters. Common examples include the normal distribution, exponential distribution, or a regression model with specific functional forms. The parameter estimates from this fitted model.

The next step involves simulating new datasets, known as bootstrap samples, based on the fitted parametric model. This is done by randomly drawing observations from the fitted model. The number of observations in each bootstrap **sample** is **typically equal to the size of the original dataset**, and sampling is often done with replacement.

Once the bootstrap samples are generated, the **statistic of interest** is computed for each sample. This can include any statistic derived from the data, such as means, medians, regression coefficients, or hypothesis test statistics. By repeating the resampling process and computing the statistic for a large number of bootstrap samples, a sampling distribution of the statistic is obtained.

5.2 Goal of the research

Like it happened before, now the other variable which stood out a lot in the correlation will be analysed. This variable is the economic status. In this dataset, this is represented with the lunch type. Kids who have the free or reduced lunch type probably come from families who can't pay it.

Using the graphics of the non-parametric bootstrap, it can be seen that the best distribution to make the approach is the Normal. The following graphic is the result of the Some High School analysis. In it, it can be seen the IC and how the Normal Density and the Kernel Density are the ones that approach the density of the model:

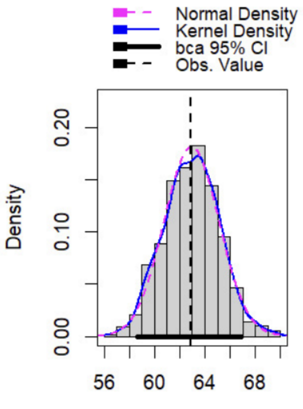


Figura 10: Histogram with densities

Therefore, the analysis of the type of lunch has been done following a Normal distribution. In the following table the mean, standard deviation and Ic of the two types of lunch can be compared, it has again been done with a thousand reps and the IC has a 95% confidence.

It's very remarkable to mention how the IC are much more concrete than the parents' studies. This is probably due to the economic situation of the household having a much major impact on the marks (it's a much more restrictive condition). This can also be judged since there's a very big difference in the mean which confirms the hypothesis. This test has been done taking into account all the other variables, and not setting just one value for all the other 4 variables. So comparing this results to the ones done by non-parametric bootstrap may be unfair. This may also be the reason why the ICs are much more bounded than the non-parametric ones.

	Mean	STD	IC
Free	62	13.7	(61.8,62.4)
Standard	71	13.5	(71.4,71.8)

6 CONCLUSIONS

First of all a background selection has been done, where it has been seen that all the variables have a big effect on the estimate of MeanScore. Starting from there after observing Figure 2 it has been observed that the most purposeful variables to observe were the parents' studies and the lunch type (economic situation). The Ethnic group would also have been interested but since it's just classified by letters it was considered that the information wouldn't be as useful. It has been then decided to apply non-parametric bootstrap for the first one and parametric bootstrap for the second one. In the second one, the parametric bootstrap has been applied using a Normal Distribution since it can be seen in Figure 10 that it's the one that matches the best the density histogram of the marks.

6.1 Parental studies

Upon analysing the marks of students in relation to their families' academic background, a convincing pattern emerges. The data clearly demonstrates a good relationship between parental training and the academic performance of their children. As parental education degree increases, there is a visible upward trend in the mean scores obtained by students. The mean score for students whose parents just have Some High School is 62. As we move up the educational ladder, we witness a steady increase. Kids whose parents have completed High School achieve a slightly higher mean tag of 63, while those with families who attended Some College improve to 65. The pattern becomes more obvious as we reach higher levels of education. Students whose parents have an Associate's Degree demonstrate a considerable rise in academic performance, with a mean mark of 69. Students whose parents have a Bachelor's degree continue to follow this pattern, earning an even higher mean of 72. Finally, students whose parents possess a Master's degree achieve the highest mean mark of 74.

The level of variation in the data is indicated by the standard deviations associated with each group's marks. As the level of parental studies increases, the standard deviation usually remains fairly stable, ranging from 2.19 to 3.26. This suggests that while the marks rise, the general consistency of performance remains similar across the various levels of maternal education.

In addition, the ICs usually become narrower as familial education level increases, indicating a higher level of trust in the estimated mean marks. This suggests that there is statistical significance to the observed variations.

The information study concludes by showing a strong correlation between parental education and students' academic performance. As parents' educational attainment increases, their kids tend to achieve higher marks. While personal factors and external influences even play a role in academic performance, the data suggests that parental education has a significant effect. It definitely underscores the importance of parental involvement and education options in fostering a conducive learning environment for students, ultimately contributing to their overall academic achievement.

6.2 Economic background

Upon analyzing the marks of students based on their socioeconomic status, categorized by their lunch type, an interesting comparison can be drawn between those two where they're classified as Free (indicating students from economically disadvantaged backgrounds) and Standard (representing students from more affluent backgrounds).

The information reveals a renowned discrepancy in the mean signs between the two groups. Students classified under Free lunch type have a mean mark of 62, while those categorized as Standard exhibit a considerably higher mean mark of 71. This significant 9 point variation suggests a clear disparity in academic performance based on socioeconomic factors.

The mean marks' standard deviations give information about the degree of variability within each type of lunch. Both groups exhibit relatively similar standard deviations, with values of 13.7 for the Free group and 13.5 for the Standard group. This suggests that there is a similar degree of mark variability across all economic groups.

The observed distinction between the two parties is further strengthened by analyzing the ICs. The IC for the Free ranges from 61.8 to 62.3, implying a fairly accurate estimate of the imply mark. Conversely, the IC for the Standard group extends from 71.4 to 71.8, highlighting a similar level of precision but with significantly higher mean mark estimates.

In summary, the evaluation showcases a large disparity in educational performance based on the economic backgrounds of students. Those from economically disadvantaged backgrounds tend to exhibit lower mean marks compared to their counterparts from more affluent backgrounds. This distinction emphasizes the impact of economic factors on academic outcomes, of course taking into account that there are other variables that influence this outcome.

7 BIBLIOGRAPHY

- Kaggle: Where the dataset has been extracted from.
- Boot: Used to work with the boot library.
- Parametric and non-parametric presentation from class.
- R-project: Extra information to understand the bootstrap.
- Examples: More examples of the bootstrap.

8 APPENDIX WITH CODE

```
#LIBRARY USED
library(tidyverse)
library(readxl)
library(boot)
library(ggplot2)

# INICIALIZATION OF THE DATASET
data = read.csv("C:/Users/Judit/Desktop/uab/segon/segon_semestre/adc/prac_final
    /Original_data_with_more_rows.csv")

#first we will remove the id
data2 <- data[,-1]

#now we will add the three marks as a final mark and delete the others
data3 <- mutate(data2, MeanScore= rowMeans(data2[6:8]))
data3 <- data3[,-c(6,7,8)]

#now we will do the One-Hot Encoding
dummy <- dummyVars("~.", data = data3)
data4 <- data.frame(predict(dummy, newdata = data3))
head(data4)

#BACKWARD SELECTION
# Fit a linear regression model
fit = lm(MeanScore ~ Genderfemale + Gendermale + EthnicGroupgroup.A +
    EthnicGroupgroup.B +
           EthnicGroupgroup.C + EthnicGroupgroup.D + EthnicGroupgroup.E +
    ParentEducassociate.s.degree + ParentEducbachelor.s.degree
         + ParentEduchigh.school +  ParentEducmaster.s.degree +  ParentEducsome
    .college + ParentEducsome.high.school + LunchTypefree.reduced
         + LunchTypestandard + TestPrepcompleted + TestPrepnone, data = data4)

# Perform stepwise selection using AIC criterion
model_backward <- stepAIC(fit, trace = TRUE, direction = "backward")
summary(model_backward)

#PLOTS
#plot of the residuals
plot(resid(model_backward))

#plot of the correlation
coef_data <- data.frame(
  coef_name = names(coef(fit)),
  coef_value = coef(fit)
)

ggplot(coef_data, aes(x = coef_name, y = coef_value)) +
  geom_bar(stat = "identity", fill = "hotpink") +
  xlab("Variable") +
  ylab("Coefficient Estimate") +
  ggtitle("Coefficients of Linear Regression Model") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#plots of MeanScore
boxplot(data4$MeanScore)
hist(data4$MeanScore)
```

```
#plots of the mean and the median
non_param_mean = function(x){
  x = mean(sample(x, size = length(x), replace = TRUE))
  return (x)
}
stats_mean = replicate(1000, non_param_mean(data4$’MeanScore’))
boxplot(stats_mean)
hist(stats_mean)

non_param_median = function(x){
  x = median(sample(x, size = length(x), replace = TRUE))
  return (x)
}
stats_median = replicate(1000, non_param_median(data4$’MeanScore’))
boxplot(stats_median)
hist(stats_median)

#NON-PARAMETRIC
#we will take into account the parent’s education (all the cases)
#SOME HIGHSCHOOL
cond1 <- data4$Genderfemale == 1
cond2 <- data4$EthnicGroupgroup.A == 1
cond3 <- data4$ParentEducsome.high.school == 1
cond4 <- data4$LunchTypefree.reduced == 1
cond5 <- data4$TestPrepcompleted == 1

#filter with the conditions
SHSMark <- data4 %>%
  filter(cond1, cond2, cond3, cond4, cond5)

#now we will do the mean with non parametric bootstrap and the IC 95%
mean_SHSMark <- boot(SHSMark, statistic = function(data, index) mean(data$
    MeanScore[index]), R = 1000)
print(mean_SHSMark)

SHS_IC <- boot.ci(mean_SHSMark, type = "bca", R=1000)
print(SHS_IC)
#we can see that an IC of 95% is (58, 67) approximately

#histogram to see the distribution
hist(mean_SHSMark)

#HIGHSCHOOL
cond1 <- data4$Genderfemale == 1
cond2 <- data4$EthnicGroupgroup.A == 1
cond3 <- data4$ParentEduchigh.school == 1
cond4 <- data4$LunchTypefree.reduced == 1
cond5 <- data4$TestPrepcompleted == 1

#filter with the conditions
HSMark <- data4 %>%
  filter(cond1, cond2, cond3, cond4, cond5)

#now we will do the mean with non parametric bootstrap and the IC 95%
mean_HSMark <- boot(HSMark, statistic = function(data, index) mean(data$
    MeanScore[index]), R = 1000)
print(mean_HSMark)

HS_IC <- boot.ci(mean_HSMark, type = "bca", R=1000)
print(HS_IC)
#we can see that an IC of 95% is (59, 67) approximately
```

```
#SOME COLLEGE
cond1 <- data4$Genderfemale == 1
cond2 <- data4$EthnicGroupgroup.A == 1
cond3 <- data4$ParentEducsome.college == 1
cond4 <- data4$LunchTypefree.reduced == 1
cond5 <- data4$TestPrepcompleted == 1

#filter with the conditions
SCMark <- data4 %>%
  filter(cond1, cond2, cond3, cond4, cond5)

#now we will do the mean with non parametric bootstrap and the IC 95%
mean_SCMark <- boot(SCMark, statistic = function(data, index) mean(data$
    MeanScore[index]), R = 1000)
print(mean_SCMark)

SC_IC <- boot.ci(mean_SCMark, type = "bca", R=1000)
print(SC_IC)
#we can see that an IC of 95% is (59, 71) approximately

#ASSOCIATE'S DEGREE
cond1 <- data4$Genderfemale == 1
cond2 <- data4$EthnicGroupgroup.A == 1
cond3 <- data4$ParentEducassociate.s.degree == 1
cond4 <- data4$LunchTypefree.reduced == 1
cond5 <- data4$TestPrepcompleted == 1

#filter with the conditions
ADMark <- data4 %>%
  filter(cond1, cond2, cond3, cond4, cond5)

#now we will do the mean with non parametric bootstrap and the IC 95%
mean_ADMark <- boot(ADMark, statistic = function(data, index) mean(data$
    MeanScore[index]), R = 1000)
print(mean_ADMark)

AD_IC <- boot.ci(mean_ADMark, type = "bca", R=1000)
print(AD_IC)
#we can see that an IC of 95% is (65, 74) approximately

#BACHELOR'S DEGREE
cond1 <- data4$Genderfemale == 1
cond2 <- data4$EthnicGroupgroup.A == 1
cond3 <- data4$ParentEducbachelor.s.degree == 1
cond4 <- data4$LunchTypefree.reduced == 1
cond5 <- data4$TestPrepcompleted == 1

#filter with the conditions
BDMark <- data4 %>%
  filter(cond1, cond2, cond3, cond4, cond5)

#now we will do the mean with non parametric bootstrap and the IC 95%
mean_BDMark <- boot(BDMark, statistic = function(data, index) mean(data$
    MeanScore[index]), R = 1000)
print(mean_BDMark)

BD_IC <- boot.ci(mean_BDMark, type = "bca", R=1000)
print(BD_IC)
#we can see that an IC of 95% is (67, 76) approximately
```

```
#MASTERS DEGREE
cond1 <- data4$Genderfemale == 1
cond2 <- data4$EthnicGroupgroup.A == 1
cond3 <- data4$ParentEducmaster.s.degree == 1
cond4 <- data4$LunchTypefree.reduced == 1
cond5 <- data4$TestPrepcompleted == 1

#filter with the conditions
MDMark <- data4 %>%
  filter(cond1, cond2, cond3, cond4, cond5)

#now we will do the mean with non parametric bootstrap and the IC 95%
mean_MDMark <- boot(MDMark, statistic = function(data, index) mean(data$
    MeanScore[index]), R = 1000)
print(mean_MDMark)

MD_IC <- boot.ci(mean_MDMark, type = "bca", R=1000)
print(MD_IC)
#we can see that an IC of 95% is (67, 80) approximately

#PARAMETRIC BOOTSTRAP- NORMAL DISTRIBUTION
#FREE / REDUCED LUNCH TYPE
# Estimate the parameters
mean_free <- mean(data4$MeanScore[data4$LunchTypefree.reduced == 1])
std_free <- sd(data4$MeanScore[data4$LunchTypefree.reduced == 1])

# Size of the sample
n_free <- sum(data4$LunchTypefree.reduced == 1)

# Parametric bootstrap
boot_free <- rnorm(1000, mean = mean_free, sd = std_free / sqrt(n_free))

# IC 95%
ic_free <- quantile(boot_free, c(0.025, 0.975))

print(mean_free)
print(std_free)
cat("95% Confidence Interval for FREE/REDUCED:", ic_free[1], "-", ic_free[2], "
    \n")
#we can see that an IC of 95% is (61.7, 62.4) approximately

#STANDARD LUNCH TYPE
mean_standard <- mean(data4$MeanScore[data4$LunchTypestandard == 1])
std_standard <- sd(data4$MeanScore[data4$LunchTypestandard == 1])

# Size of the sample
n_standard <- sum(data4$LunchTypestandard == 1)

# Parametric bootstrap
boot_standard <- rnorm(1000, mean = mean_standard, sd = std_standard / sqrt(n_
    standard))

# IC 95%
ic_standard <- quantile(boot_standard, c(0.025, 0.975))

print(mean_standard)
print(std_standard)
cat("95% Confidence Interval for FREE/REDUCED:", ic_standard[1], "-", ic_
    standard[2], "\n")
#we can see that an IC of 95% is (71.4, 71.8) approximately
```