IMPACT OF PARENTS' EDUCATION LEVEL ON STUDENT'S ALCOHOL CONSUMPTION

QUANTITATIVE DATA ANALYSIS, TSM

Outline

I. Introduction

Underage drinking has been one of the most significant public health issues which causes enormous health and safety risks. While drinking in high school has become a ritual that students see as a way of integration among their peers, as an experience for newfound freedom and independence of adulthood, excessive drinking habit phenomenon is increasingly worrying because it can cause an alcohol use disorder (AUD). This disorder can further result in long-lasting physical and emotional damage. In order to sort out the causes of student alcohol consumptions, educational and sociological experts are investigating different factors giving rise to students' alcohol consumption, including certain aspects of school life, the inconsistency of underage drinking laws, family history and relationships, the availability and exposure to alcohol for students at school events, social activities, and so on. Furthermore, we know that an individual's integration into society is strongly influenced by the socialisation and primary education he or she receives. The social environment in which the child grows up and the influence of parents on the internalisation of social norms are essential factors in the child's personal development.

We will focus in this study on family history and the possible impact on the frequency of alcohol consumption among students. This leads us to believe that the parents' level of education influences their child's drinking frequency. If this is the case, it would imply that the likelihood of each individual having an alcohol disorder would be greater for those from a lower social category. Indeed, we believe the lower the social class of a child is, the greater the social disadvantage for him or her, who will not benefit from the same chances and opportunities as a child from a higher social class. This issue was the motivation behind this study on the potential correlation between the parents' level of education and the frequency of alcohol consumption among students. Conducting this statistical study will enable us to establish whether or not the level of education of each parent has a significant impact on the frequency of the student's alcohol consumption, and, if possible, to determine what other factors may influence student's alcohol consumption level. The results we obtain from this study will eventually lead us to draw potential recommendations in order to understand better and help prevent Alcohol Use Disorders among students.

After introducing our research hypothesis, we will introduce our data selection along with the descriptions and limitations of the samples. Then, we will run statistical tests in order to test our hypothesis and interpret our results. Finally, we will cope with the implications of our study and possible recommendations we can draw based on the outcome of our research.

II. Research Hypothesis

Our reflection on the subject led us to elaborate the following research hypothesis :

The higher the level of education the parents have, the less the student consumes alcohol.

This statistical research aims to identify the eventual correlation, relationships and other causes of a parent's education level on the student's alcohol consumption level. Since correlation does not equal causality, the tests that we will be undertaking to support our research hypothesis will be set at a low significance level (0.05), with the confidence that a significance level of 5% can minimise the risks for the occurrence of Type 1 and Type 2 errors.

.

III. Data selection

a) Dataset: A general description

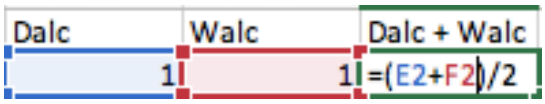
The materials used in this study includes compiled data from 2 secondary schools data sources. The dataset that has been used for this study includes compiled data from 2 secondary schools data sources. It includes the total number of 395 students from 2 high schools in Portugal, their genders, age, family size, parents' education level, parents' occupations, students' workday alcohol consumption and students' weekend alcohol consumption in the year 2008. The source of the data set is the University of Camerino and more precisely it was contributed by researchers Fabio Pagnotta and Hossain Mohammad Amran.

The variables in the dataset have been reduced to contain the needed variables required for our analysis. The columns that matter to us are Medu (Mother's education level), Fedu (Father's education level), Dalc (Student's weekday consumption of alcohol level) and Walc (Student's weekend consumption of alcohol level).

In this dataset, the parents' education levels are divided into 5 categories, (Zero education received, Primary education (until 4th grade), Primary education (from 5th to 9th grade), Secondary education (high school) and Higher education. Regarding the students' alcohol consumption levels, the type of data measurement that is used is categorical at an ordinary level. It is a ranking divided into 5 levels of consumption:

- 1 = Very low
- 2 = Low
- 3 = Medium
- 4 = High
- 5 = Very high

We added one column to our dataset, that is, the mean of Walc and Dalc in order to have the mean consumption level for each student, over a whole week. We have merged both columns by calculating the sum of the Dalc and the Walc columns and we divided the result by two, such as in the following :



Then, we had to adapt the initial ranking described above as some of the values end with .5. When averaged, the alcohol consumption values correspond to a total of 10 values, ranging from 1 to 5.5 (1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5). This variable 'Average alcohol consumption" now resembles a more continuous quantitative variable.

b) Data limitations

To begin with, the source is taken from a school data source in Portugal, but not from an official organization specialised in data gathering. The credibility of data remains to be proved and the data could be falsified.

Secondly, the data source was taken in the year of 2008. It has been more than a decade since it was updated, and the analysis derived from the data can be outdated as the data set does not take into consideration the change of recent years. The data was based on students from one school in Portugal, which restricts the study to a specific category of people (here, only students from one school), all from the same place. We know that when the sample is large and carried out on a random population (who do not necessarily live or work in the same area), it is more representative of the real population's distribution.

Regarding the education categories, the data are not specific enough. This is one of the common limitations of categorical variables. We are unsure of what each variable means precisely since we did not choose the variables and there is no further explanation about them. For instance, for the 'higher education' variable, it could mean Bachelor, Master or even PHD. If we overgeneralise it, the analysis drawn cannot be accurate. Same for the 'alcohol consumption' variable where the range is not clearly defined. How you precisely define a 'low' or 'high' level of alcohol consumption is unclear. After we averaged both columns of Weekend alcohol consumption and Weekday alcohol consumption, the dataset we are using to run our tests resulted in 10 values ranging from 1 to 5.5, which gives us a broader range of alcohol consumption level but remains hard to define.

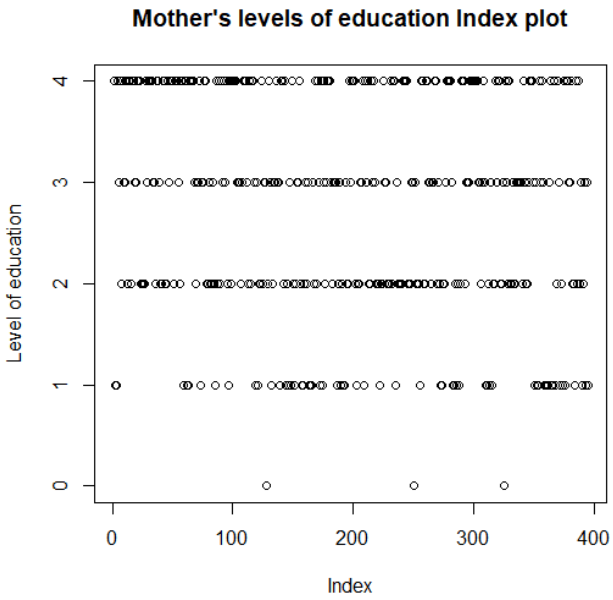
Besides, out of 395 students, only 22 students have a consumption ranked 4 and above, which means high to very high consumption. There's a high proportion of students in our sample who have a low drinking habit so it makes it harder to compare and find significant results related to our hypothesis as it is not a normal distribution. Thus, it hinders the possibility of several statistical tests, like regression tests.

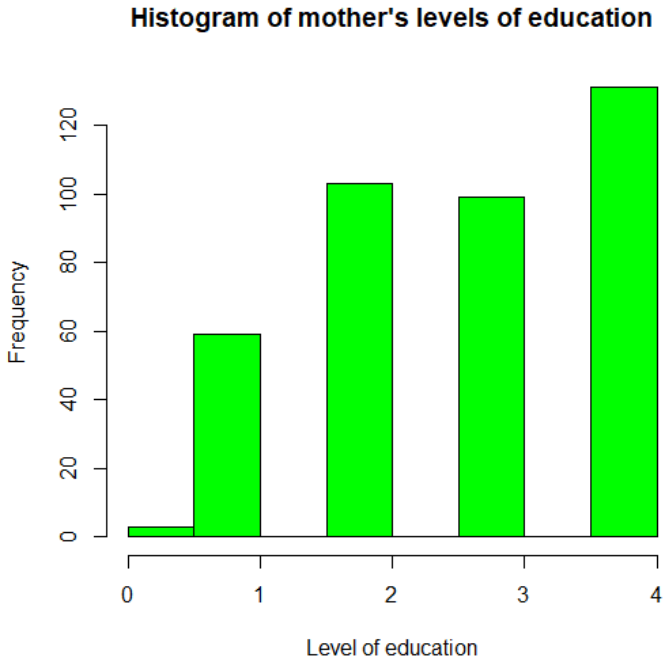
c) Sub-data descriptions

Mother's education level

First, it is important to notice that for the levels of the mother's and the father's education, we have a ranking from 0 to 4, which gives us the following categorical values:

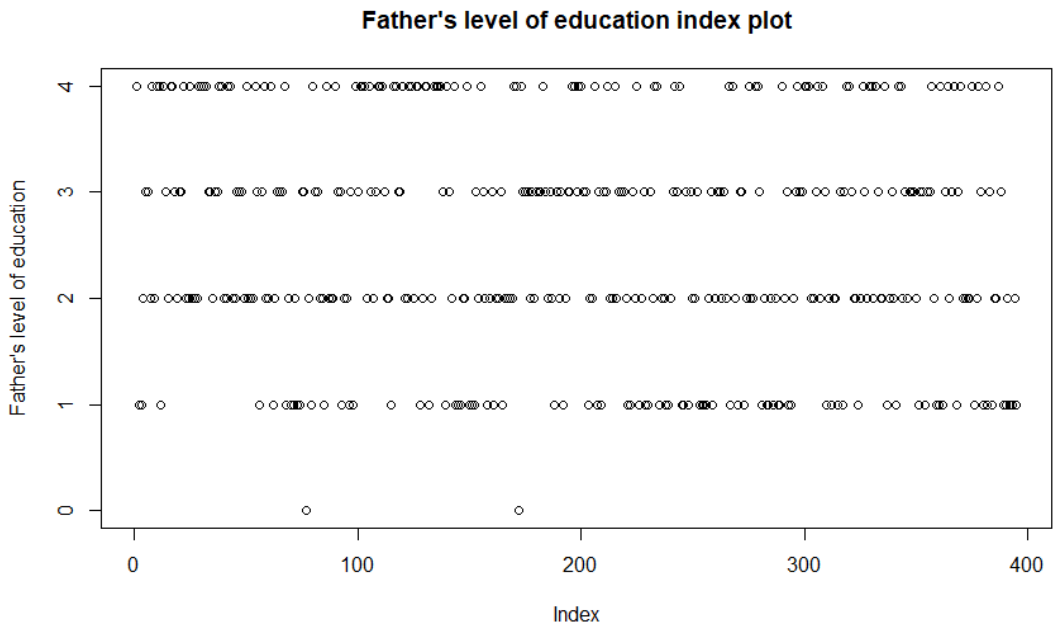
- 0 = none
- 1 = primary education (until 4th grade)
- 2 = primary education (5th to 9th grade)
- 3 = secondary education (high school)
- 4 = higher education



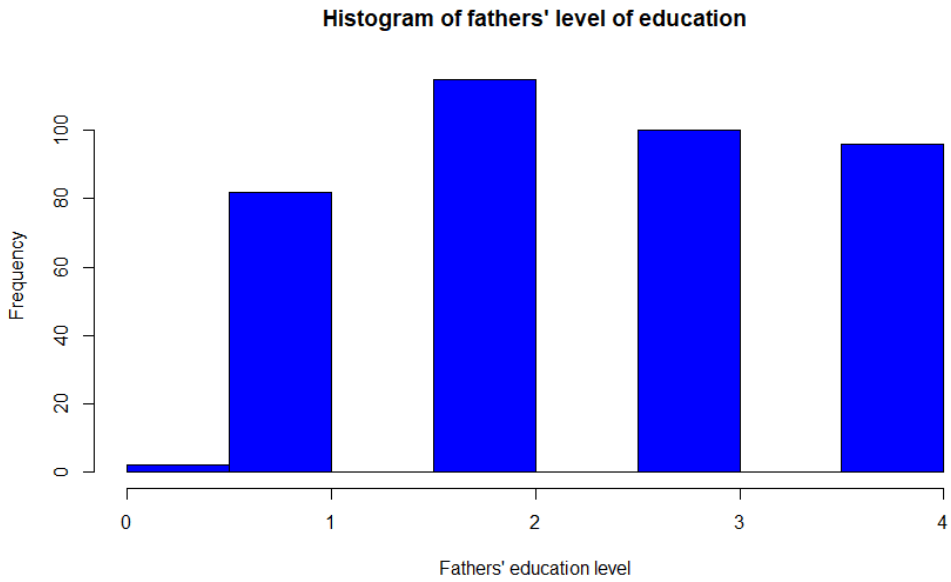


The index plot which can be seen above represents the distribution of categorical values regarding a mother's education level. As we can see from visual observations, the number of values increases as the level of education increases. These observations are verified by the histogram which also shows us that the number of values grows as the rank increases. The biggest amount of values recorded is in the rank 4, where there are 131 women with a higher education level. Ranks 2,3 and 4 are at least 40 values higher than the rank 1, which is the difference between the rank 1 and the rank 3. Nevertheless, the growing tendency is not proven completely as rank 2 is higher than rank 3 by the amount of values. Furthermore, the shape of the histogram shows us that the data is not normally distributed.

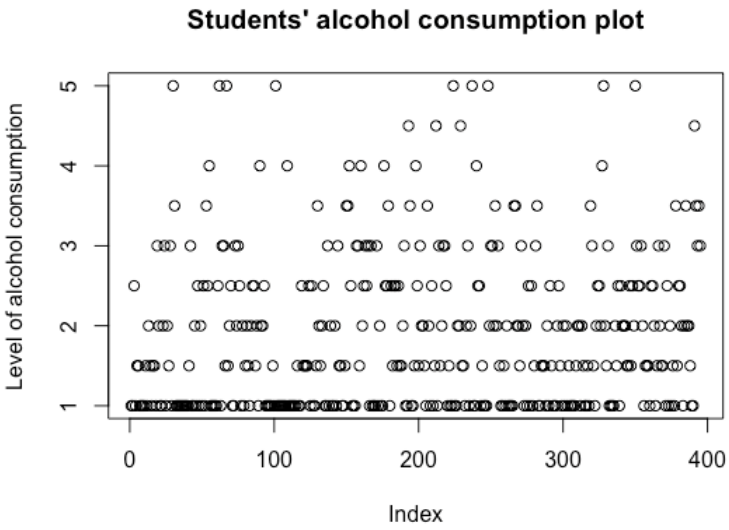
Father's education level



Regarding the father's education, the levels are almost similarly distributed between the five ranks. As the above index plot and histogram show, only two of the fathers are not educated, while ranks from 1 to 4 have a close distribution of numbers. The rank 2 has the highest recorded number of fathers (more than a 100), while ranks 1,3, and 4 have values ranging from 80 to 100. From the shape of the histogram, we can also conclude that the data is not normally distributed.

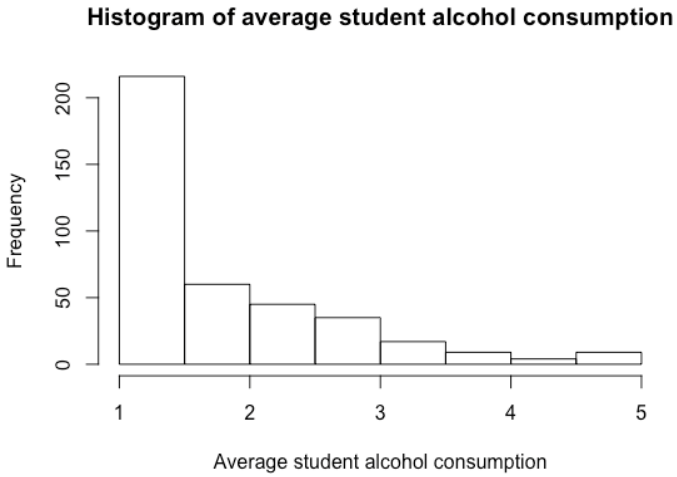


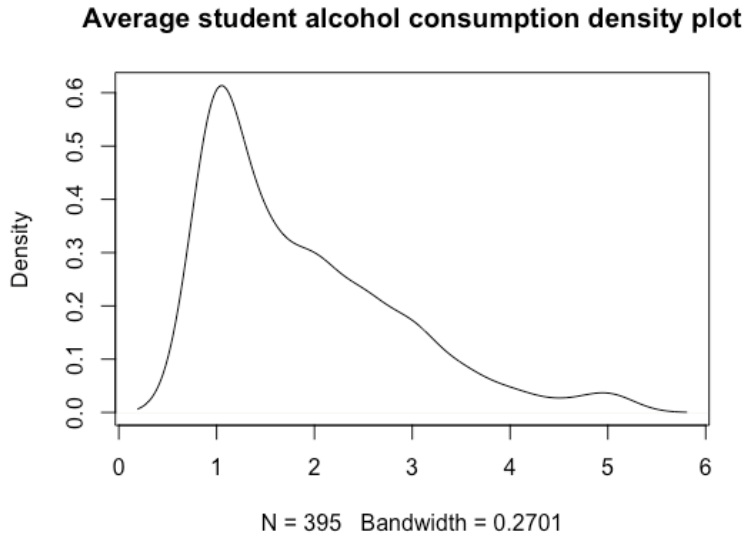
Student's alcohol consumption



The index plot of alcohol consumption among students differs from the two previous ones. Here, we have more values listed and these appear to be more like continuous variables. We observe that most students have a low alcohol consumption, ranging from 1 to about 3. As the alcohol consumption increases, the number of students decreases. We can deduct from this plot that the population does not seem to be normally distributed.

Furthermore, if we look at the histogram, most of the values are positioned at the left of the structure, meaning the plots are skewed to the extreme left which further implies that the data is not normally distributed.





Finally, we can see with the density plot that the distribution is positively skewed, meaning the median and the mode are to the left of the mean. We can also see from the density plot that the distribution is leptokurtic, as it has a high and thin peak.

From our descriptions, we can conclude that this data is not normally distributed. Nevertheless, according to the Central Limit Theorem, since our sample is large enough, we will still be able to apply the two-sample t-tests as part of the statistical analysis to test our hypothesis.

IV. Statistical tests

a) Analysis of Variance

We have to run first an analysis of variance (ANOVA) that will help us see if there is a difference of variance of student consumption, in each level of education. We chose ANOVA because it is a statistical test which estimates how much the variance of a quantitative dependent variable is influenced by one or more categorical independent variables. In our data sets, we have the dependent variable being the students' alcohol consumption level, and the independent variables, that is the fathers' and mothers' level of education. Then, it tests whether there is a difference in means of the groups at each level of the independent variable, by checking the variance in each individual group against the overall variance of the data. If one or more groups falls outside the range of variation which is predicted by the null hypothesis, then we will be able to say that our test is statistically significant. In our analysis of variance, the null hypothesis (H0) is no difference in means, and the alternate hypothesis (H1) is that the means are different from one another.

For our research, the one-way ANOVA is adopted. We are testing the effects of mothers' education level and fathers' education level on students' alcohol consumption.

ANOVA test for mothers' education level on student alcohol consumption

```
> VarResult_Medu<-aov(Alcohol~Medu,data=For_multiple)
> summary(VarResult_Medu)
             Df Sum Sq Mean Sq F value Pr(>F)
Medu          1    0.2  0.1823   0.185  0.667
Residuals   393  387.7  0.9865
```

ANOVA test for fathers' education level on student alcohol consumption

```
> VarResult_Fedu<-aov(Alcohol~Fedu,data=For_multiple)
> summary(VarResult_Fedu)
             Df Sum Sq Mean Sq F value Pr(>F)
Fedu          1    0.0  0.0197    0.02  0.888
Residuals   393  387.9  0.9869
```

Interpretation of the test

In the one-way ANOVA, we modelled the students' alcohol consumption as a function of the mothers' and fathers' education level respectively. First, we use aov() to run the model, then we use summary() to print the summary of the model.

The model summary first lists the independent variable being tested in the model (in this case we have only one, ‘Student Alcohol Consumption’) and the model residuals (‘Residuals’). All of the variation not explained by the independent variable is called residual variance. From the results, we can see that the F value for both tests is really low, which shows that the variation among sample means is smaller than the variation within groups (of education). This translates to very little difference against the null hypothesis. The variation in the dependent variable is not due to the independent variables. Furthermore, it is noted that the P-value of both cases is higher than the significance level 0.05, with 0.667 for mother’s education level and 0.888 for father’s education level. Therefore, we can conclude that there are no significant differences between each group and it appears that the education level of mother and father do not have a significant impact on students' alcohol consumption.

b) Two-sample t-tests

The analysis of variance showed us that there was no significant difference in means between each parents' level of education and the students' level of alcohol consumption. In order to further test our initial hypothesis that '***The higher the level of education the parents have, the less the student consumes alcohol'*,** we decided to conduct paired t-tests. The first one correlates the mean of alcohol consumption for students with low-educated mothers (education rank 3 and less), and the mean of alcohol consumption for students with highly-educated mothers (education rank 4). The second test was run similarly but on the fathers' education levels. These tests help in comparing the mean student alcohol consumption between well-educated parents and those with low education, to test our initial hypothesis. If for both the mother and the father, the consumption was higher when the parents were less educated, then our hypothesis would be validated. We can also check from the t-test whether the mean alcohol consumption for the different education levels falls in our "low consumption" (less than 3) rank or "high consumption" rank (3 and more).

T-test based on the mothers' education level

We will now perform the paired t-test by comparing the means of alcohol consumption among students who have mothers with high education or with low education.

Our Null hypothesis is that the true mean difference between the two groups is equal to 0, meaning there is no significant difference between both samples. Our alternative hypothesis is that the true mean difference is different to 0, meaning there is a significant difference between the means of both samples. The following is the result of the two-sample t-test of alcohol consumption based on the mothers' education level.

```
        Welch Two Sample t-test

data:  Ttestdatasample$Average.consumption.with.low.educated.mother and Ttestdatasample
$Average.consumption.with.high.educated.mother
t = 0.54707, df = 260.48, p-value = 0.5848
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1506976  0.2666472
sample estimates:
mean of x mean of y
 1.905303  1.847328
```

The mean alcohol consumption of students with low-educated mothers is 1.905, and that of students with highly-educated mothers is 1.847. Thus, the test indicates a slightly higher mean of alcohol consumption when the mothers are less educated. However, both means, regardless of the education levels, represent a rather "low alcohol consumption" in both samples, as it is less than 3. Also, the p-value is the significance level of the test. Here, the p-value is equal to 0.6131. It is greater than 0.05, so the average alcohol consumption of

students with low-educated mothers is not significantly different from the average alcohol consumption of students with highly educated mothers.

T-test based on the fathers' education level

We will now perform the paired t-test by comparing the means of alcohol consumption among students who have fathers with high education or with low education. Our Null hypothesis and alternative hypothesis remain the same as in the above T-test that we conducted with mothers' education level.

```
        Welch Two Sample t-test

data:  Ttestdatasample$Average.consumption.with.low.educated.father and Ttestdatasample
$Average.consumption.with.high.educated.father
t = -0.5067, df = 152.2, p-value = 0.6131
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2991083  0.1769999
sample estimates:
mean of x mean of y
 1.871237  1.932292
```

First, if we look at the two means, the mean of students with highly educated fathers is higher (1.932) than the one for low educated fathers (1.871). This indicates that the group of students with highly educated fathers drink on average more than students with low educated fathers. However, the p-value is equal to 0.6131. It is greater than 0.05, so the average alcohol consumption of students with low-educated fathers is not significantly different from the average alcohol consumption of students with highly educated fathers. This p-value indicates that we cannot verify that there is a real difference between the two samples, thus we cannot reject the null hypothesis, while our alternative hypothesis is rejected.

Through the use of the Welch t-test and the interpretation of our results, we find that students with low educated mothers consume on average more alcohol than those with highly educated mothers. On the other hand, among fathers with low education, students have on average a lower alcohol consumption than among fathers with high education. However, our tests show that there is no significant difference between the average consumption of students with low and high education parents. We therefore reject our initial hypothesis that the higher the level of education the parents have, the less the student consumes alcohol.

d) Conclusion

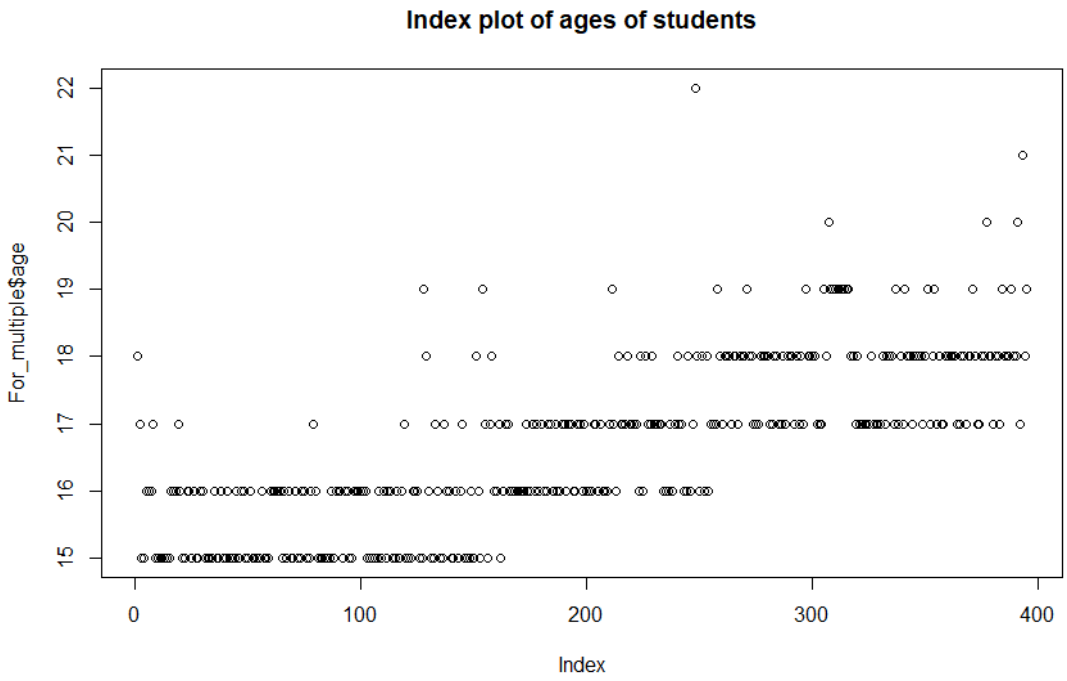
Relying on our results, we cannot establish any correlation between the level of education of parents and the level of alcohol consumption of students. This may be due to several factors. As we mentioned above, our database has several limitations which may influence the results we obtained on our tests. Indeed, the data from the sample is restricted to a defined group of people who were not randomly selected, which means that the data from the students' alcohol consumption may not reflect the real population. We know that the larger the sample and the more randomly selected it is, the more representative it will be of the real population. Even more importantly, two of the variables we are using are categorical. Although the student consumption variable is equivalent to a continuous variable with values ranging from 1 to 5.5, for the mother's and father's education levels, we only have 4 categories.

We assume that a more specific and diverse data sample about education levels and student consumption could have given us more significant results. Unfortunately, we cannot prove our assumptions as there is only one study that was carried out. However, we want to expand our study by making the same tests between students' alcohol consumption and the number of absences or their age, to determine whether the results are significantly different with richer data.

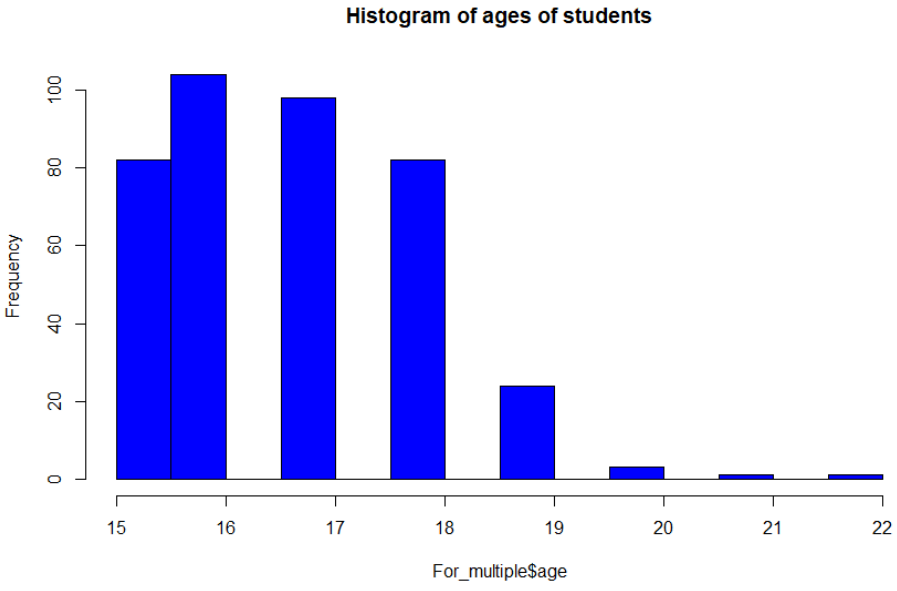
V. Going beyond our hypothesis

a) Additional data description

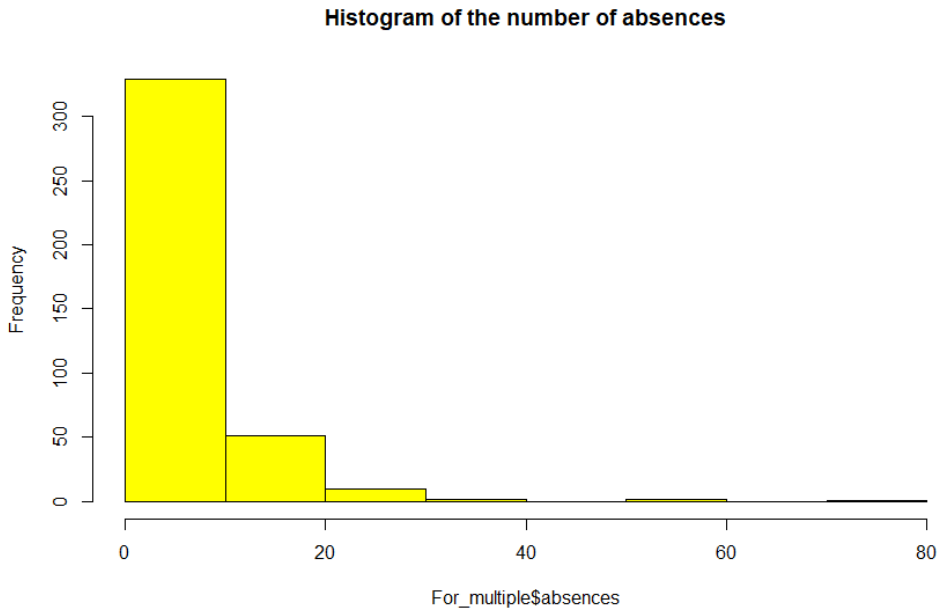
Age of students



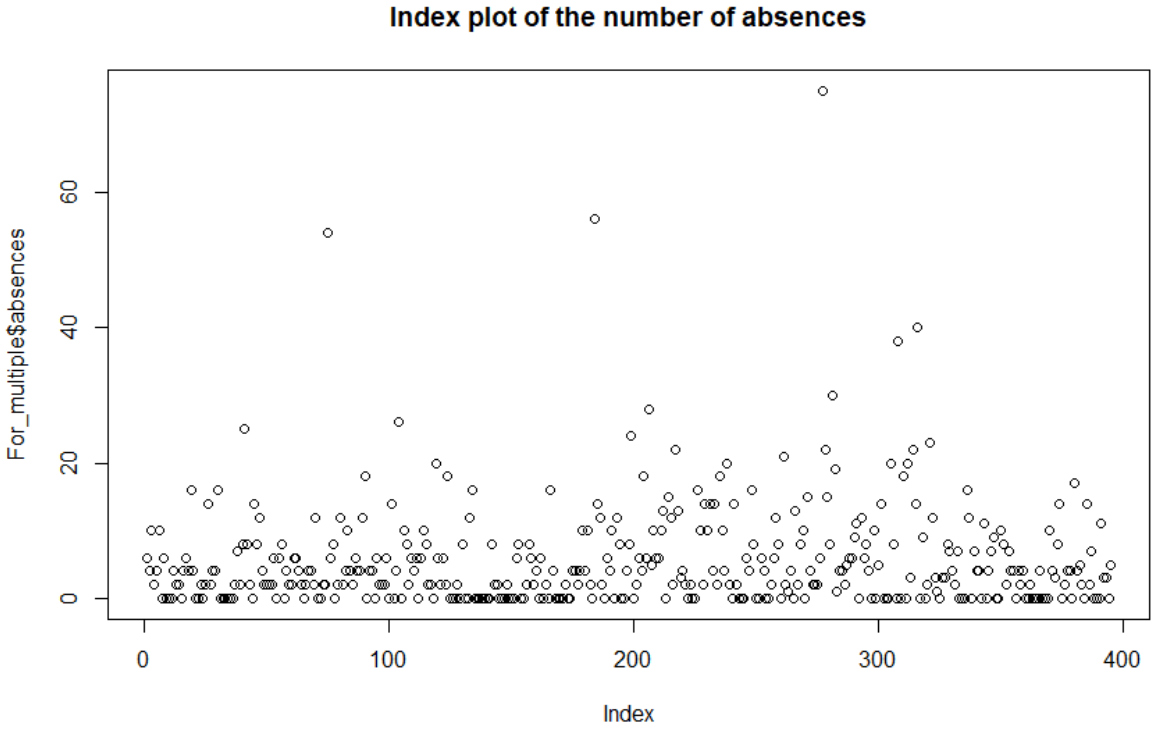
Given the variable "Age of students", we have 8 different values which is twice the number of what we had with parents' education levels. As we can see from the index plot and the histogram, values are mostly spread evenly. Most of the students are between 15 and 18 years old. In this case, from the shape of the histogram we can conclude that the data is not normally distributed.



Number of absences



The number of values in absences is also bigger than the number of categories we had for parents' education levels and the variable ranges from 0 up to 75 absences.



The picture that the histogram shows us is quite the same as it was with previous data. Most of the values are located within the smaller number of times students were absent. From the index plot, we can see that most of the values lie between 0 and 15 number of absences, while very few data are recorded above 15 absences. Both histogram and index plot show us that the data is not normally distributed.

b) Additional tests

We will perform paired two sample t-tests on the variables age and absences in order to examine the results, but we will not conduct any ANOVA test. Indeed, ANOVA tests require a continuous dependent variable (such as the average alcohol consumption), and categorical independent variables. In this case, age and absence aren't categorical, thus preventing us from using them as such.

T-test based on the number of absences

To check whether continuous variables such as absences can lead to more significant results, and can explain the lack of justification in our main categorical variables, we divided the data set in two. Absences of 14 or less are assumed as a low absence rate, with 359 students falling in this range. Absences of 15 and more are considered as a high absence rate, with 36 students falling in this range. The average alcohol consumption was calculated for each category and a paired two-sample t-test was run between them, as follows:

```
        Welch Two Sample t-test

data:  Absences$Avg.consumption.for.low.absences and Absences$Avg.consumption.for.high.absences
t = -1.8164, df = 40.379, p-value = 0.07673
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.74897913  0.03983336
sample estimates:
mean of x mean of y
 1.853760  2.208333
```

The results show that the average alcohol consumption for students attending more of the classes (with absences of 14 or less), is equal to 1.85, while the one of students who have missed their classes 15 or more times, is around 2.21. This indicates that students who attend most of their classes drink on average less than students with a higher number of absences. However, the p-value equals 0.07, which is just above 0.05, meaning there is no statistically significant difference between both groups.

T-test based on the age of the students

For this test, we decided to divide the sample in two: one column corresponds to the average student consumption for minors (less than 18 years old), and the other one corresponds to the average consumption of majors (18 years old and more). The following test will tell us if there is a significant difference in consumption for students according to their age.

```
        Welch Two Sample t-test

data:  AGE.R$Avg.consumption.minor and AGE.R$Avg.consumption.major
t = -0.95781, df = 193.67, p-value = 0.3394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3313992  0.1147389
sample estimates:
mean of x mean of y
 1.855634  1.963964
```

The results show that the average alcohol consumption for students who are minors equals 1.85, while the one of students who are major equals 1.96. This indicates that students who are 18 or more drink on average more than students who are less than 18 years old. The p-value equals around 0.34, which is way above 0.05, meaning that statistically, there is no significant difference between the means of both groups.

To conclude, the focus of our study is not on the influence of absences or age on alcohol consumption as we would need to do further research. We wished to determine whether using continuous data would provide us with more meaningful results. We can see that results on the new t-tests are more meaningful as they have a higher difference in means and lower p-values than in the tests we ran on parents' education level. Nonetheless, they remain not significant enough to reject the null hypothesis. The limitations of our data sample remain the main issue as it has influenced our test results and does not allow us to construct a study that is not partly skewed.

VI. Implications of the study

We conducted a study to show whether there was any influence of parents' level of education on the alcohol consumption of students. Our research hypothesis being, **the higher the level of education the parents have, the less the student consumes alcohol**. We started by describing our data selection and its limitations. Then, we conducted different statistical tests in order to test our hypothesis and interpret the results we obtained. To conclude, our research hypothesis cannot be justified and is rejected.

In fact, after the conduction of the ANOVA test and the paired T-tests, we found no significant evidence of a correlation between parents' education level and students' alcohol consumption. Therefore, based on the result shown, we cannot make further recommendations for students with alcohol dependency problems, nor for the health and education departments concerned by this issue.

Following the rejection of our hypothesis, and despite the limitations on data, we attempted to go further in our research by performing paired T-tests with two other external variables, the students' age and their absences from school to see if any more significant results could be found. Interestingly, the tests show differences in means although it remains not significant enough to show a potential correlation. Further tests would have to be conducted on this matter.

Finally, it is advised to use another dataset from a random sample and a normally distributed population to have more significant results. The categorical data of parents' education level would have to be better defined and distributed, since most of the values were in the highest education level, thus impacting the results. Furthermore, a more random and diverse sample taken from a larger population of students coming from different backgrounds would have led to more objective and representative results of the real population. If we wish to further our study, the above recommendations must be taken into account in order to obtain more meaningful and objective results, which would not be biased by the limitations of the data.