Lab 2

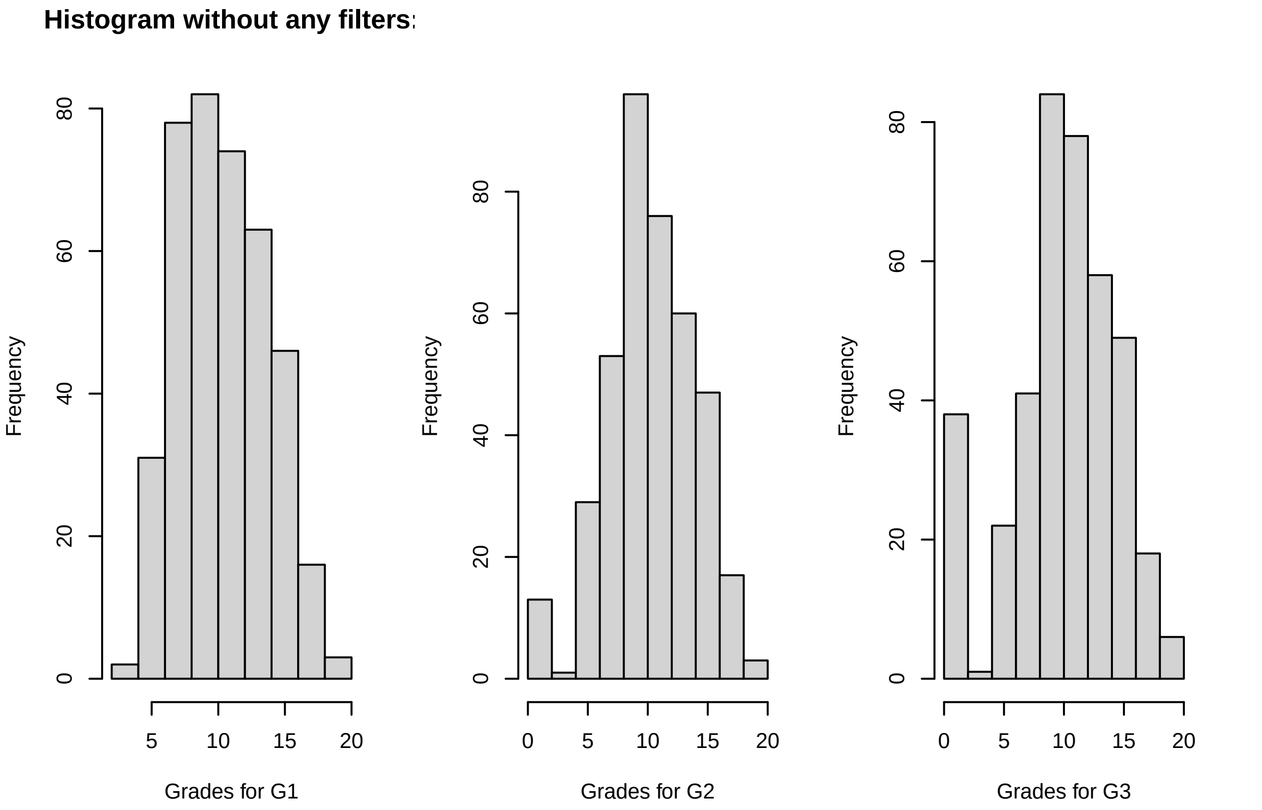
Bishal Thapa

3/4/2021

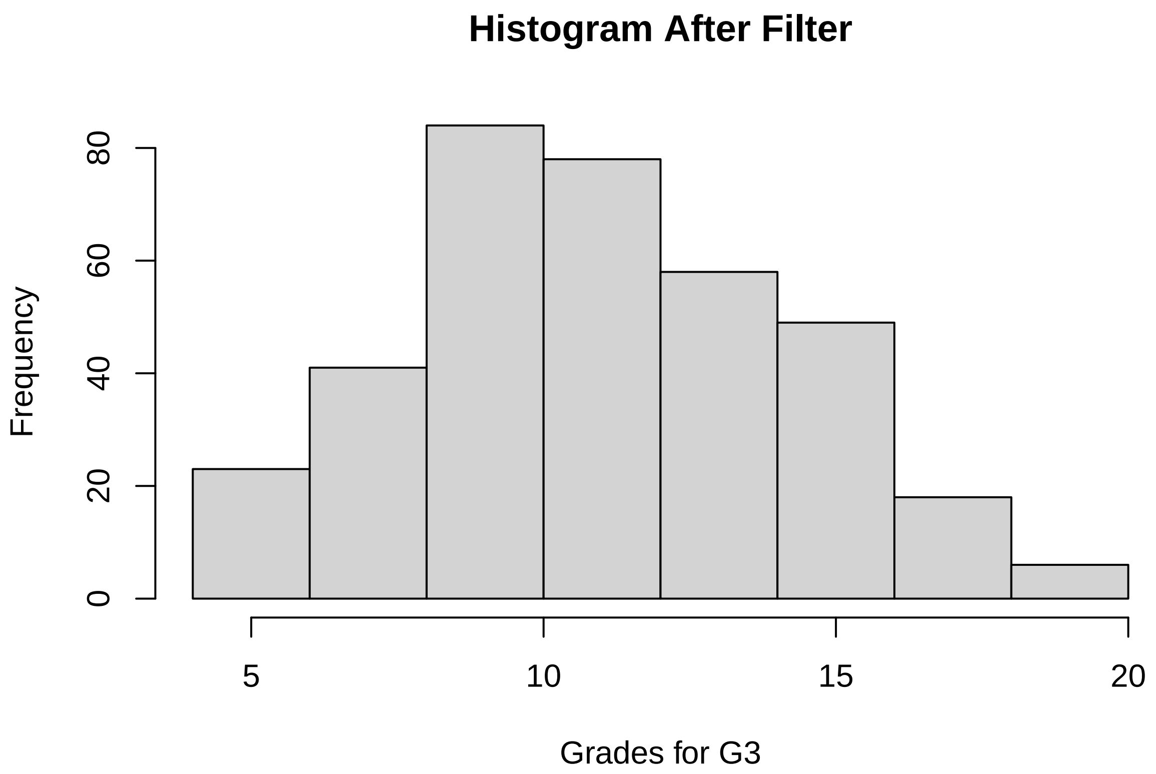
Introduction

The goal of this paper is to analyze a data set that recorded a Portuguese secondary education students' performance in mathematics (during the 2005-2006 school year).The main goal of this paper is to identify whether or not some of the other recorded information can help predict a student's grade. We will check if there is any correlation between students grade and drinking habits.We will also try to predict if there is any effects of paid extra math classes on student grades.For that we need to make sure the final grade reported is correct.So in order to do that we have removed any 0 grades posted as final G3 grades. This is because students are not likely to receive 0 as grade. This must have been an error or student left classes which resulted in 0 grade.The data set available is 'student-mat.csv'. In order to filter the student with 0 grade new data frame is extracted as 'students.completed'. We will use this data frame to perform our analysis.

Histogram of the grades G1-G3 without any filters is shown below:



We filtered the students with 0 grade so we could see the grades of only students that were taking the course.Histogram of grade G3 after filtering is shown below:



From the graph we can see Grade G3 seems to be normally distributed as the histogram takes the bell shape.

Hypothesis Testing for average grade

Now for the final grade G3 we are going to perform a hypothesis test to determine if the average final grade is significantly greater than 10 where 10 is the minimum passing grade.

For this our hypotheses is:

$$\begin{cases} H_0 : \mu = 10 \\ H_a : \mu \ > 10 \end{cases}$$

Now the mean (x) of final grade (G3)=11.52381

Also the S.D of final grade (G3)=3.22779

Number of students(n) = 357

Test Statistics:

$$Z = \frac{\bar{x} - \mu}{\frac{sd}{\sqrt{n}}}$$

$$Z = \frac{11.52381 - 10}{\frac{3.227797}{\sqrt{357}}}$$

So, Z=8.91986

Now, P value=P(X<-8.91986)

= pnorm(-8.91986)

= 2.33439e-19

Assume we are testing at significance level 5%.

$$Then, \alpha = 0.05$$

$$\therefore Since \quad \alpha > p - value \quad H_0 \quad is \quad rejected.$$

Hence, There is a strong evidence to support the claim that average final grade is more than 10.. As 10 is the passing grade, we can say that on average students score higher than the passing grade.

Confidence Interval for True Mean

It is also important to understand about the confidence interval for the true mean final score for Grades G3.

For confidence interval of 95% we need,

$$(\bar{X} - E, \bar{X} + E)$$

$$Since \quad our \quad \alpha = 0.05, \quad \frac{\alpha}{2} = 0.025$$

$$Now \quad E = Z_{\frac{\alpha}{2}} * (\frac{sd}{\sqrt{n}})$$

$$E = 1.96 * \frac{3.22779}{\sqrt{3}57}$$

$$\therefore E = 0.33483$$

$$(\bar{X} - E, \bar{X} + E) = (11.19, 11.86)$$

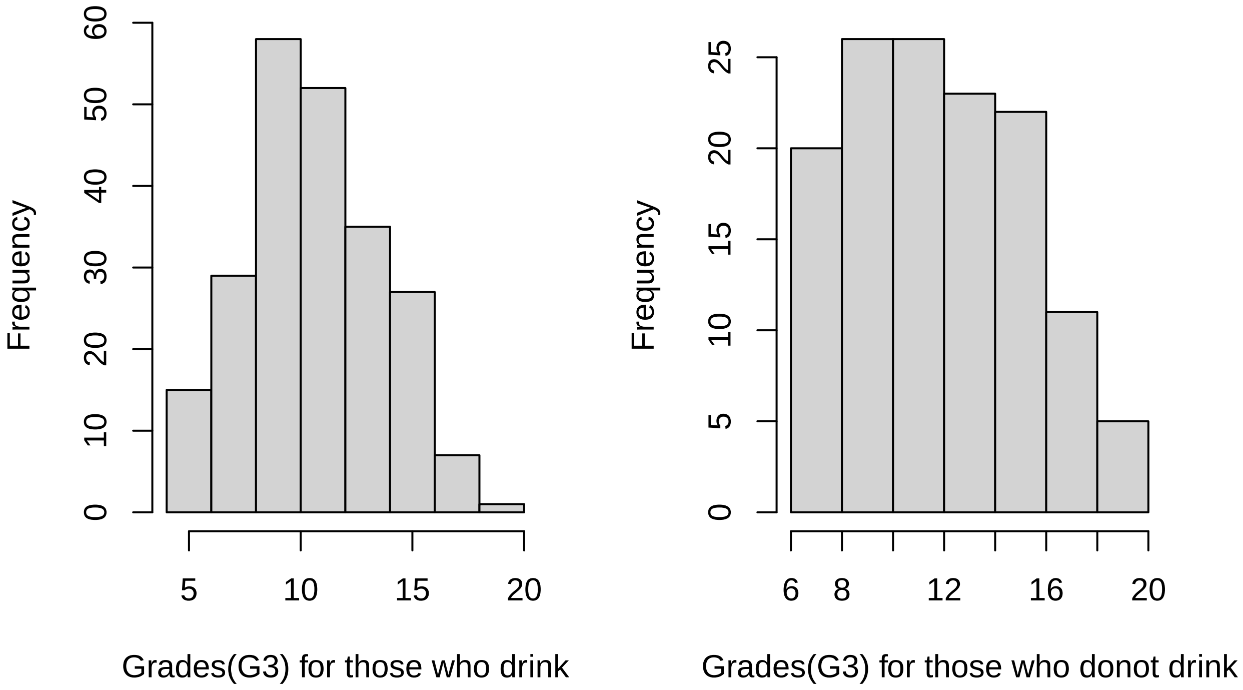
So we are 95% confident that the true mean of final grade(G3) lies between (11.19,11.86)

Hypothesis Testing for drinking vs non-drinking

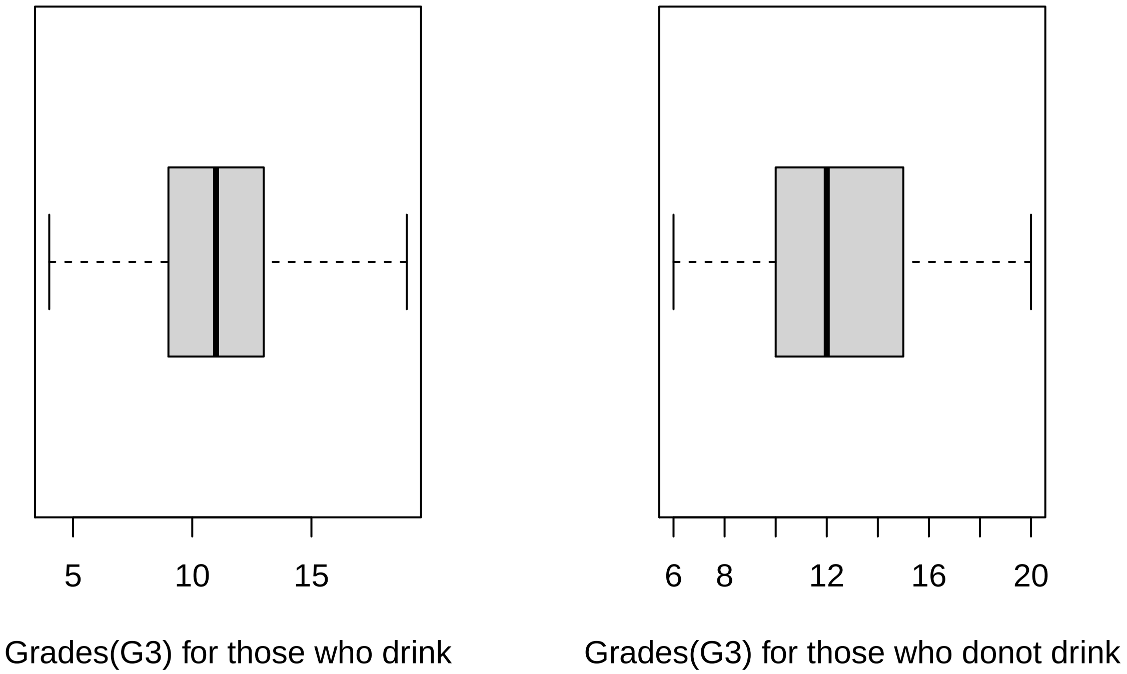
Furthermore we are interested in understanding how drinking habits influence the students grades. For that we need two groups of students: ones that drink and the other that does not drink. In order to get such students data frame from the data-set provided we are going to filter and store the students information based on their drinking habits. The dataframes will be categorized as Students.Alcohol and Students.Nonalcohol.

In the histograms below we can see the difference between students that drink and students that donot drink in terms of their grades.

Histogram for difference between Student groups:



Boxplots for difference between Student groups:



From the box-plot we can see the average grade of the students who do not drink is greater than the students that drink.We can see from box-plot that the grade for students who had alcohol is normally distributed while the grades of students who had no alcohol is right-skewed. The median grade of students who had no alcohol appears to be higher than for those who had alcohol. Also upper quartile for data on students who had no alcohol is higher than the upper quartile for the students who had alcohol. Also from the histogram we can verify there are more non-drinking students who score higher grades compared to the students that drink.The mean grade of drinking students group is 11.13 whereas the mean grade of non-drinking students is 12.19.Also, the number of students in drinking group was 224 whereas number of students in non-drinking group is 133.

With the above data we have collected, we can do a hypothesis test to determine if there is a significant difference in mean final grades between drinkers and non-drinkers.

For that our sample mean X of non-drinkers =12.19

Also, sample mean Y of drinkers = 11.13

Standard Deviation (S.D) of non-drinkers= 3.479704

Standard Deviation (S.D) of drinkers= 3.007648

For this our hypotheses is:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

Test Statistic:

$$Z = \frac{(\bar{X} - \bar{Y} - 0)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

$$Z = \frac{(12.19 - 11.13)}{\sqrt{\frac{3.4797^2}{133} + \frac{3.007^2}{224}}}$$

Therefore, Z= 2.924 Since this is a two-tail test, P value=2 * P(X<2.924)

```
= 2* pnorm(-2.924)

:. P=0.003455647
```

For significance level of 95%.

$$\alpha = 0.05$$

Since P-value < 0.05 we reject Null hypothesis.

Hence there is enough evidence to support the claim that there is a significant difference in mean final grades between drinkers and non-drinkers. This hypothesis supports that alcohol consumption may decrease student's grade.

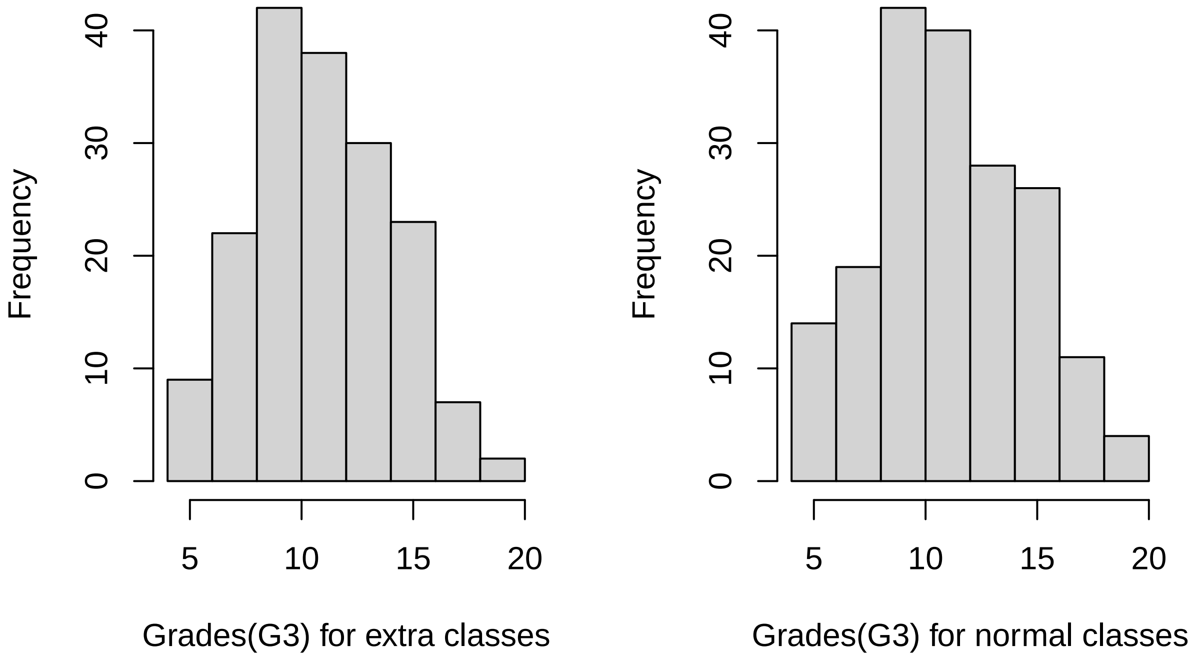
Hypothesis Testing for Extra classes vs Non-Extra classes

The average score of students who take extra classes is 11.427 whereas those who do not take classes is 11.614. The total number of students taking extra classes is 173 and who do not take extra classes is 184. So percentage of students taking the classes are 48.5%.

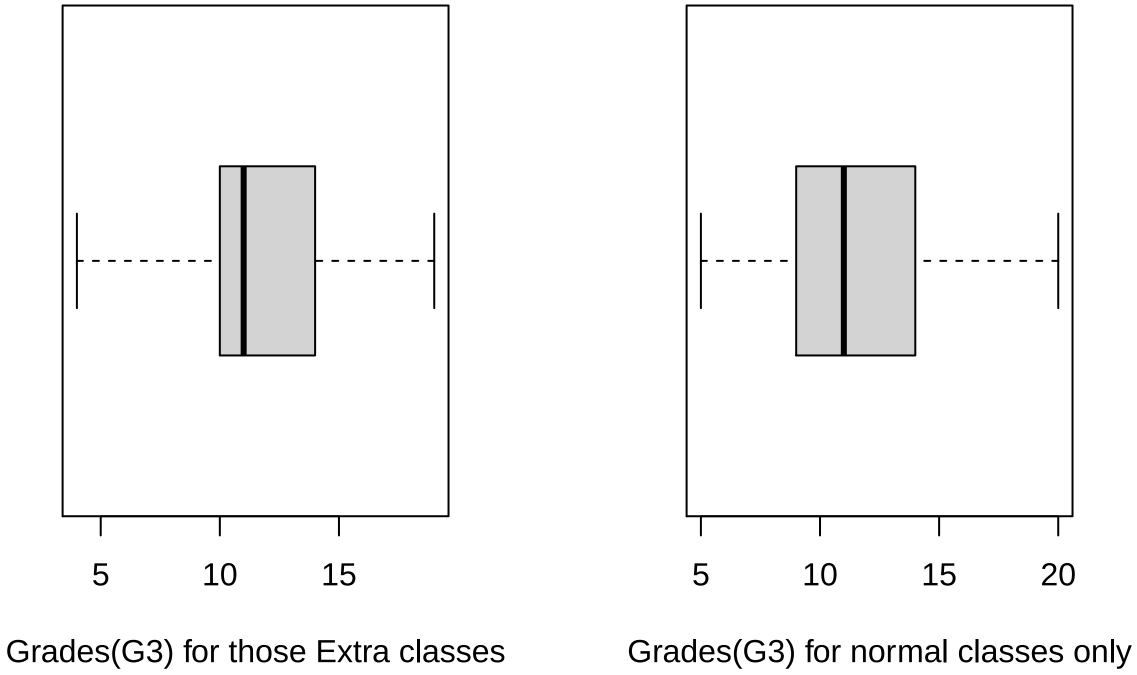
As the average grade for either group was not much different so there is no surprise with the result. Maybe the students who were weak in their math class took the extra class to improve to score.

Now we are going to analyze how the extra paid math classed affect the results. In order to do that we make 2 groups of students that take extra classes and that do not take extra classes.

Student that take extra classes versus the normal students.



Boxplots for difference between extra class vs normal group:



With the above data we have collected, we can do a hypothesis test to determine if there is a significant difference in mean final grades between students taking extra classes and normal students.The total number of students taking extra classes is 173 and who do not take extra classes is 184.

For that our sample mean X of extra class takers =11.427

Also, sample mean Y of other normal students = 11.614

Standard Deviation (S.D) of extra class takers= 3.038844

Standard Deviation (S.D) of normal students= 3.401706

For this our hypotheses is:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

Test Statistic:

$$Z = \frac{(\bar{X} - \bar{Y} - 0)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

$$Z = \frac{(11.427 - 11.614)}{\sqrt{\frac{3.038844^2}{173} + \frac{3.401706^2}{184}}}$$

Therefore, Z= -0.548 Since this is a two-tail test, P value=2 * P(X<-0.548)

```
= 2* pnorm(-0.548)

:. P=0.584
```

For significance level of 95%.

$$\alpha = 0.05$$

Since P-value $>$ 0.05 we do not reject Null hypothesis.

Hence there is not enough evidence to support the claim that there is a significant difference in mean final grades between extra class students and normal students. So students taking extra classes may not necessarily score higher grades than others.

Conclusion

To sum up, our hypothesis testing suggested that out of two factors, alcohol consumption factor could be useful in predicting the final grades. It looks like alcohol consumption was directly affecting students performance whereas extra classes did not have the same impact. We can assume that students are taking extra classes if they are below the passing grade so they can pass the class. We also learned that there is a strong evidence to support the claim that average final grade is more than 10.