SEMESTER 2

SESSION 2019/2020

PROBABILITY AND STATISTICAL DATA ANALYSIS

SECI 2143-02

PROJECT 2:

A STUDY ON THE INFLUENCE OF BACKGROUND AND ATTITUDES OF SECONDARY SCHOOL STUDENTS ON MATHEMATICS GRADES IN PORTUGAL

NAME:

HAM JING YI

MATRICS NUMBER:

A19EC0048

LECTURER'S NAME: DR. CHAN WENG HOWE

SUBMISSION DATE: 27 JUNE 2020

Table of Contents

INTRODUCTION

Today, basic education in Portugal is set to be compulsory to everyone. For me, I think that the education system is simple and good. However, according to The Portugal News, it showed that 50 % of the Portuguese of aged between 25 and 64 did not complete the secondary school (The Portugal News, 2020). I am curious about this. So, the dataset that I found is about the background, attitudes, and the grades of secondary students in Portugal. The sample size, n is 395. The dataset is collected by P. Cortez and A. Silva. It consists a lot of interesting social, gender and study information about students. The data is all about the students' mathematics courses in secondary school. The data is purposely used to predict students' final grades, or it has been used in exploratory data analysis (EDA). Based on the features and facts that I stated above, I decided to carry out a survey on the influences of background and attitudes on the Mathematics grades among secondary students in Portugal. This is to test that whether the background and attitudes of students really affect the students' grades in Mathematics subject. The background here is the jobs of students' parents. I also consider the address as the students' background too. The attitudes here are about how the students spend their time, their absences, their future study plans and more.

HYPOTHESIS TESTING

Test 1: 2 sample test to test the mean of Urban G1 and Rural G1.

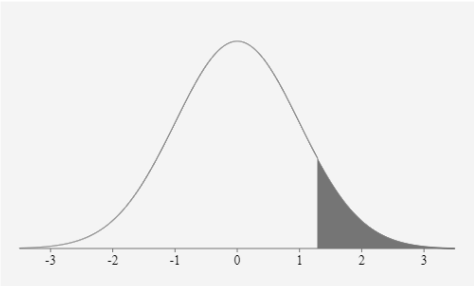
This is to test whether the mean of students' first period grades (G1) in urban address is higher than the mean of students' first period grades (G1) in rural address. Assume the population variances are different here and using 0.10 significance level. μ_1 is the population mean of students' first period grades (G1) in urban address while μ_2 is the population mean of students' first period grades (G1) in rural address.

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

$n_1 = 307$ $\bar{x}_1 = 11.0326$ $s_1 = 3.2858$

$n_2 = 88$ $\bar{x}_2 = 10.4773$ $s_2 = 3.4173$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$



Degree of freedom, $df = v = 136$

Critical value = $t_{0.10,136} = 1.2878$

Test Statistics, $T_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} = 1.3553$

It is a right tail test. Reject H_0 if $T_0 > 1.2878$

Decision:

Since $T_0 = 1.3553 > 1.2878$ We reject the null hypothesis, H_0 because the test statistics value lies within the critical region.

Conclusion:

There is sufficient evidence to conclude that the mean of students' first period grades (G1) in urban address is higher than the mean of students' first period grades (G1) in rural address at 0.10 significance level.

Test 2: 2 sample test to test the mean of Urban G2 and Rural G2.

This is to test whether the mean of students' second period grades (G2) in urban address is higher than the mean of students' second period grades (G2) in rural address. Assume the population variances are different here and using 0.05 significance level. μ_1 is the population mean of students' second period grades (G2) in urban address while μ_2 is the population mean of students' second period grades (G2) in rural address.

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

$n_1 = 307$ $\bar{x}_1 = 10.9674$ $s_1 = 3.6905$

$n_2 = 88$ $\bar{x}_2 = 9.8295$ $s_2 = 3.8929$

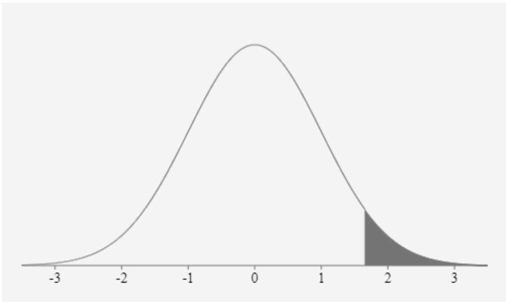
$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Degree of freedom, $df = v = 135$

Critical value = $t_{0.05,135} = 1.6562$

Test Statistics, $T_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} = 2.4451$

It is a right tail test. Reject H_0 if $T_0 > 1.6562$



Decision:

Since $T_0 = 2.4451 > 1.6562$ We reject the null hypothesis, H_0 because the test statistics value lies within the critical region.

Conclusion:

There is sufficient evidence to conclude that the mean of students' second period grades (G2) in urban address is higher than the mean of students' second period grades (G2) in rural address at 0.05 significance level.

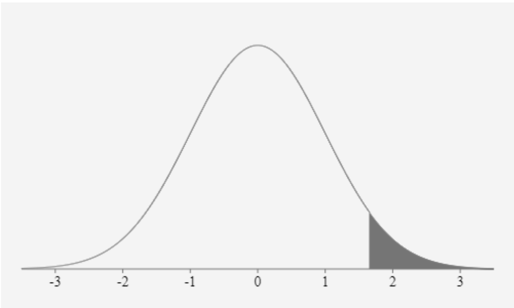
Test 3: 2 sample test to test the mean of Urban G3 and Rural G3.

This is to test whether the mean of students' final period grades (G3) in urban address is higher than the mean of students' final period grades (G3) in rural address. Assume the population variances is different here and using 0.05 significance level. μ_1 is the population mean of students' final period grades (G3) in urban address while μ_2 is the population mean of students' final period grades (G3) in rural address.

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

$n_1 = 307$	$\overline{x}_1 = 10.6743$	$s_1 = 4.5631$
$n_2 = 88$	$\overline{x}_2 = 9.5114$	$s_2 = 4.5561$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$



Degree of freedom, $df = v = 140$

Critical value = $t_{0.05,140} = 1.6558$

Test Statistics, $T_0 = \frac{\overline{x}_1 - \overline{x}_2 - \Delta_0}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} = 2.1101$

It is a right tail test. Reject H_0 if $T_0 > 1.6558$

Decision:

Since $T_0 = 2.1101 > 1.6558$ We reject the null hypothesis, H_0 because the test statistics value lies within the critical region.

Conclusion:

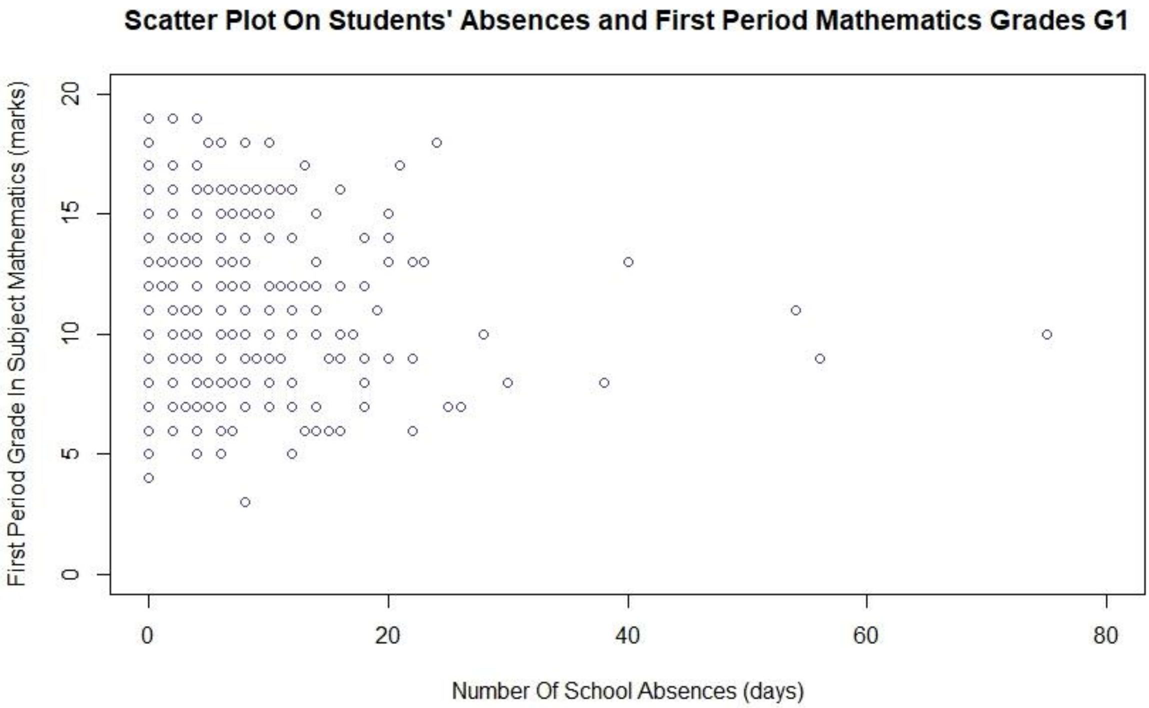
There is sufficient evidence to conclude that the mean of students' final period grades (G3) in urban address is higher than the mean of students' final period grades (G3) in rural address at 0.05 significance level.

Test 4: Pearson’s Product-Moment Correlation Coefficient to test absences and G1

This is to test the strength of relationship between number of school absences (absences) and first period grades (G1) of mathematics subject of secondary school students in Portugal. This test is carried out using absences as x and G1 as y.

x = number of school absences (absences)

y = first period grades of mathematics subject (G1)



r = -0.0310

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

From the value of r, it is a weak negative linear correlation.

In order to test whether there is linear correlation between these 2 variables. The hypothesis testing is carried out since the r is close to 0.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

α = 0.05, $n = 395$

degree of freedom, df = n - 2 = 393

Critical values = $-t_{0.025,393} = -1.966$ and $t_{0.025,393} = 1.966$

Test Statistics, $T_0 = \frac{r}{\sqrt{\frac{(1-r^2)}{n-2}}}$ = -0.6149

It is a two tails test. Reject H_0 if T_0 < -1.966 or T_0 > 1.966

Decision:

Since T_0 = -0.6149 is between -1.966 and 1.966 We fail to reject the null hypothesis, H_0 because the test statistics value does not lie within the critical region.

Conclusion:

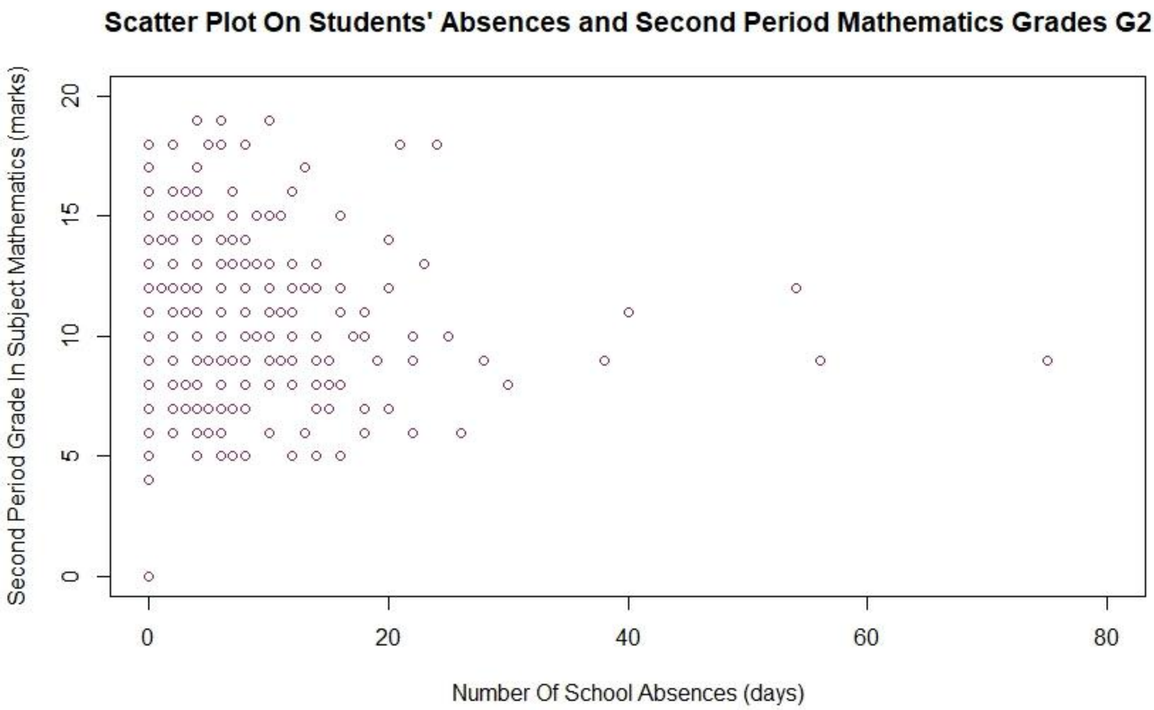
There is sufficient evidence to conclude that there is no linear correlation exist between number of school absences (absences) and first period grade in subject Mathematics (G1) at 0.05 significance level.

Test 5: Pearson's Product-Moment Correlation Coefficient to test absences and G2

This is to test the strength of relationship between number of school absences (absences) and second period grades (G2) of mathematics subject of secondary school students in Portugal. This test is carried out using absences as x and G2 as y.

x = number of school absences (absences)

y = second period grades of mathematics subject (G2)



$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

r = -0.0318

From the value of r, it is a weak negative linear correlation.

In order to test whether there is linear correlation between these 2 variables. The hypothesis testing is carried out since the r is close to 0.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

α = 0.05, $n = 395$

degree of freedom, df = n - 2 = 393

Critical values = $-t_{0.025,393} = -1.966$ and $t_{0.025,393} = 1.966$

Test Statistics, $T_0 = \frac{r}{\sqrt{\frac{(1-r^2)}{n-2}}}$ = -0.6303

It is a two tails test. Reject H_0 if T_0 < -1.966 or T_0 > 1.966

Decision:

Since T_0 = -0.6303 is between -1.966 and 1.966 We fail to reject the null hypothesis, H_0 because the test statistics value does not lie within the critical region.

Conclusion:

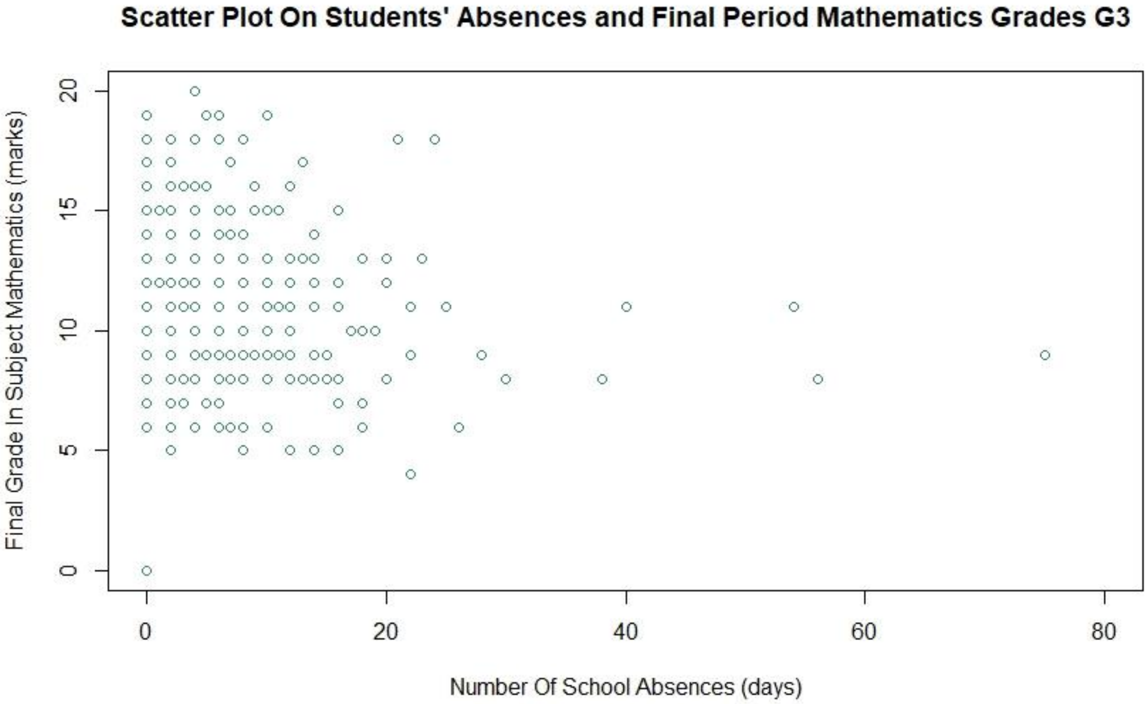
There is sufficient evidence to conclude that there is no linear correlation exist number of school absences (absences) and second period grades (G2) of mathematics subject at 0.05 significance level.

Test 6: Pearson's Product-Moment Correlation Coefficient to test absences and G3

This is to test the strength of relationship between number of school absences (absences) and final period grades (G3) of mathematics subject of secondary school students in Portugal. This test is carried out using absences as x and G3 as y.

x = number of school absences (absences)

y = final period grades of mathematics subject (G3)



$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

r = +0.0342

From the value of r, it is a weak positive linear correlation.

In order to test whether there is linear correlation between these 2 variables. The hypothesis testing is carried out since the r is close to 0.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

α = 0.05, $n = 395$

degree of freedom, df = n - 2 = 393

Critical values = $-t_{0.025,393} = -1.966$ and $t_{0.025,393} = 1.966$

Test Statistics, T_0 = $\frac{r}{\sqrt{\frac{(1-r^2)}{n-2}}}$ = 0.6793

It is a two tails test. Reject H_0 if T_0 < -1.966 or T_0 > 1.966

Decision:

Since T_0 = 0.6739 is between -1.966 and 1.966 We fail to reject the null hypothesis, H_0 because the test statistics value does not lie within the critical region.

Conclusion:

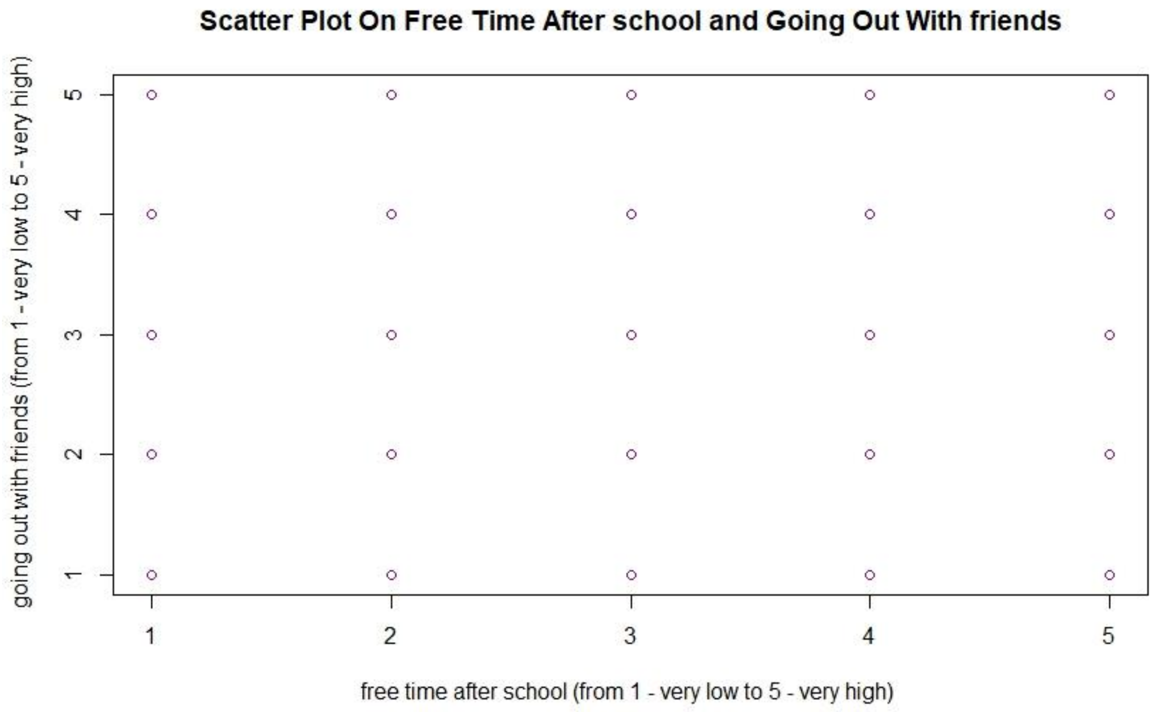
There is sufficient evidence to conclude that there is no linear correlation exist between number of school absences (absences) and final period grades (G3) of mathematics subjects at 0.05 significance level.

Test 7: Spearman's Rho Rank Correlation Coefficient on freetime and goout

This is to test the strength of relationship between going out with friends (goout) and free time after school (freetime) This test is carried out using freetime as x and goout as y.

x = free time after school (freetime) from 1 - very low to 5 - very high

y = going out with friends (goout) from 1 - very low to 5 - very high



$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

r_s = + 0.2852

In order to test whether there is linear correlation between these 2 variables. The hypothesis testing is carried out since the r is close to 0.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

α = 0.05, $n = 395$

degree of freedom, df = n - 2 = 393

Critical values = $-t_{0.025,393} = -1.966$ and $t_{0.025,393} = 1.966$

Test Statistics, $T_0 = \frac{r}{\sqrt{\frac{(1-r^2)}{n-2}}} = 5.8985$

It is a two tails test. Reject H_0 if $T_0 < -1.966$ or $T_0 > 1.966$

Decision:

Since $T_0 = 5.8985$ is greater than 1.966 We reject the null hypothesis, H_0 because the test statistics value lies within the critical region.

Conclusion:

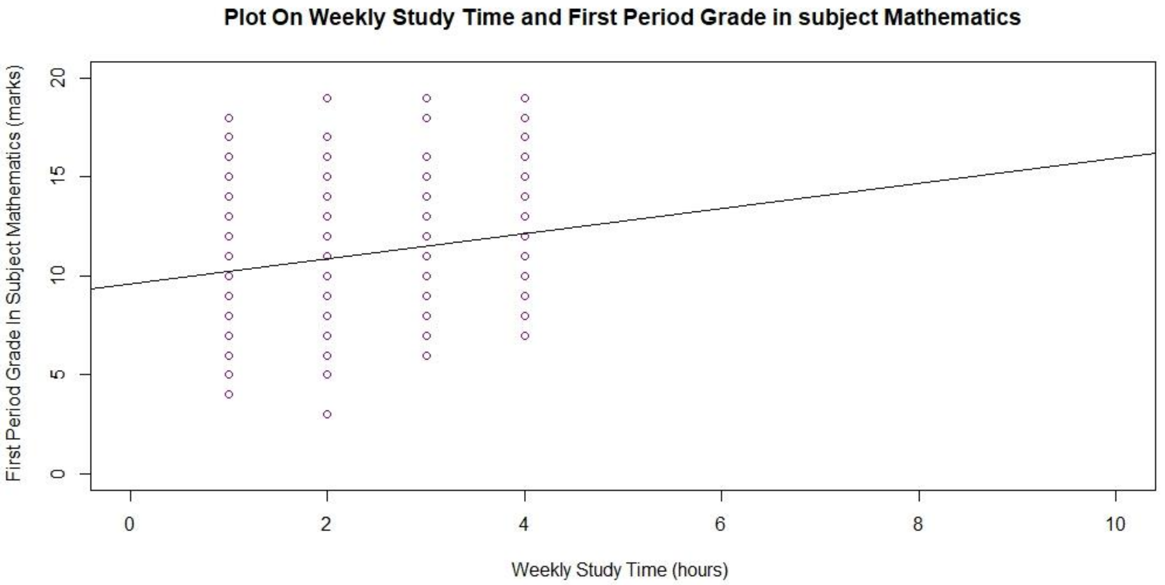
There is sufficient evidence to conclude that there is linear correlation exist between going out with friends (goout) and free time after school (freetime) at 0.05 significance level.

Test 8: Regression on studytime and G1

This is to test whether weekly study time(studytime) will affect the first period grade in Mathematics subject(G1).

x = weekly study time(studytime)

y = first period grade in Mathematics subject (G1)



Coefficient of determination, R^2 = + 0.0258

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

α = 0.05, $n = 395$

degree of freedom, df = n - 2 = 393

The regression line is $\hat{y}$ = 9.6159 + 0.6352x

Critical values = $-t_{0.025,393} = -1.966$ and $t_{0.025,393} = 1.966$

Test Statistics, $T_0 = \frac{b_1 - \beta_1}{Sb_1}$ = 3.226

It is a two tails test. Reject H_0 if T_0 < -1.966 or T_0 > 1.966

Decision:

Since T_0 = 3.226 is greater than 1.966 We reject the null hypothesis, H_0 because the test statistics value lies within the critical region.

Conclusion:

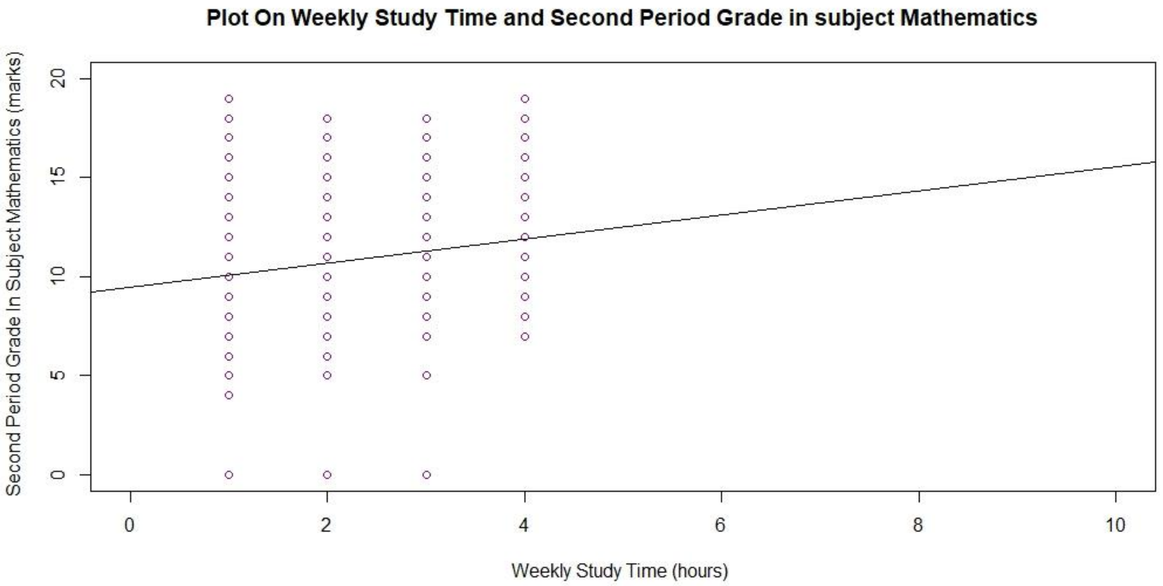
There is sufficient evidence to conclude that there is linear relationship exist between weekly study time(studytime) and first period grade in Mathematics subject(G1) at 0.05 significance level. The weekly study time(studytime) will affect the first period grade in Mathematics subject(G1).

Test 9: Regression on studytime and G2

This is to test whether weekly study time(studytime) will affect the second period grade in Mathematics subject (G2).

x = weekly study time(studytime)

y = second period grade in Mathematics subject(G2)



Coefficient of determination, R^2 = + 0.01846

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

α = 0.05, $n = 395$

degree of freedom, df = n - 2 = 393

Regression line , $\hat{y}$ = 9.474 + 0.609x

Critical values = $-t_{0.025,393} = -1.966$ and $t_{0.025,393} = 1.966$

Test Statistics, $T_0 = \frac{b_1 - \beta_1}{Sb_1}$ = 2.719

It is a two tails test. Reject H_0 if T_0 < -1.966 or T_0 > 1.966

Decision:

Since T_0 = 2.719 is greater than 1.966 We reject the null hypothesis, H_0 because the test statistics value lies within the critical region.

Conclusion:

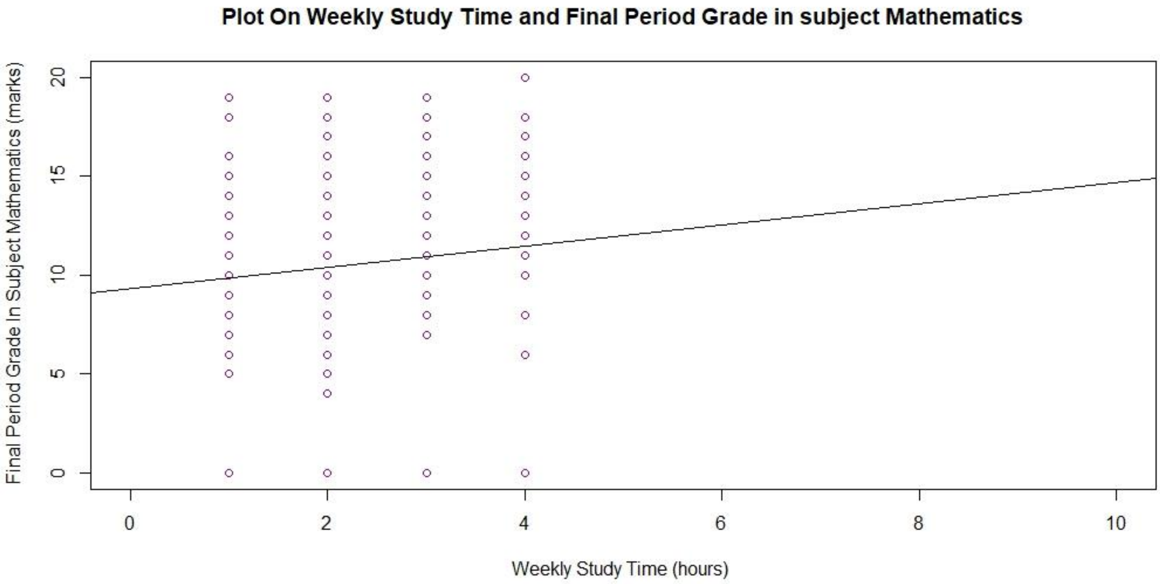
There is sufficient evidence to conclude that there is linear relationship exist between weekly study time (studytime) and second period grade in Mathematics subject(G2) at 0.05 significance level. The weekly study time (studytime) will affect the second period grade in Mathematics subject(G2).

Test 10: Regression on studytime and G3

This is to test whether weekly study time(studytime) will affect the final period grade in Mathematics subject(G3).

x = weekly study time(studytime)

y = final period grade in Mathematics subject(G3)



Coefficient of determination, R^2 = + 0.009569

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

α = 0.05, $n = 395$

degree of freedom, df = n - 2 = 393

Regression line, $\hat{y}$ = 9.328+ 0.534

Critical values = $-t_{0.025,393} = -1.966$ and $t_{0.025,393} = 1.966$

Test Statistics, $T_0 = \frac{b_1 - \beta_1}{Sb_1}$ = 1.949

It is a two tails test. Reject H_0 if T_0 < -1.966 or T_0 > 1.966

Decision:

Since -1.966 < T_0 = 1.949 < 1.966 We fail to reject the null hypothesis, H_0 because the test statistics value does not lie within the critical region.

Conclusion:

There is sufficient evidence to conclude that the linear relationship does not exist between studytime and G3 at 0.05 significance level. Weekly study time(studytime) will not affect the final period grade in Mathematics subject (G3).

Test 11: Goodness of fit test on Reason (Categories with Equal Probabilities)

Goodness of fit test is used to test an observed frequency distribution fits some claimed distribution. In this case, I assume it is categories with equal frequencies. This test is carried on variable reason to choose this school (reason) using 0.05 significance level. This is to test whether the observed frequencies in reason variable are same with the expected frequencies.

$H_0 : p_1 = p_2 = p_3 = p_4$
$H_1 :$ At least 1 of the 4 proportions is different from others.

E = $\frac{n}{k}$ = 98.75

Reason	Observed Frequency	Expected Frequency	Expected Probability
Course	145	98.75	0.25
Home	109	98.75	0.25
Reputation	105	98.75	0.25
Other	36	98.75	0.25

Test Statistics, $x_0^2 = \sum \frac{(o_i - E_i)^2}{E_i}$

= 62.9949

Number of different categories, k =4

Degree of freedom= k-1 =3

Critical value, ${x^2}_{0.05,3} = 7.8147$

It is always a right tail test. Reject H_0 if x_0^2 > 7.8147

Decision:

Since x_0^2 = 62.9949 >> 7.8147 We reject the null hypothesis, H_0 because the test statistics value lies within the critical region.

Conclusion:

There is sufficient evidence to conclude that there is at least 1 of the 4 proportions is different from others at 0.05 significance level.

Test 12: Chi Square Test of Independence on mothers' job and higher education

This test is carried on variable Mjob (mothers' jobs) and variable higher (wants to take higher education) using 0.05 significance level. This is to test there is a relationship exists between Mjob variable and higher variable.

H_0 : Variables are independent
H_1 : Variables are dependent

	no	yes
at_home	7	52
health	0	34
other	7	134
services	5	98
teacher	1	57

Test Statistics, $x_0^2 = \sum \frac{(o_{ij}-E_{ij})^2}{E_{ij}}$

$= 8.8482$

Critical value, $x^2{}_{0.05,4} = 9.488$

p-value = 0.06501

It is always a right tail test. Reject H_0 if $x_0^2 > 9.488$

Decision:

Since x_0^2 = 8.8482 < 9.488 and p-value = 0.0651 > α=0.05 We fail to reject the null hypothesis, H_0 because the test statistics value does not lie within the critical region.

Conclusion:

There is sufficient evidence to conclude that the variables mothers' jobs and variable higher (wants to take higher education) are independent at 0.05 significance level. There is no relationship between variables Mjob and higher.

Test 13: Chi Square Test of Independence on fathers' jobs and higher education

This test is carried on variable Fjob (fathers' jobs) and variable higher (wants to take higher education) using 0.05 significance level. This is to test there is a relationship exists between Fjob variable and higher variable.

H_0 : Variables are independent
H_1 : Variables are dependent

	no	yes
at_home	1	19
health	0	18
other	9	208
services	9	102
teacher	1	28

Test Statistics, $x_0^2 = \sum \frac{(o_{ij}-E_{ij})^2}{E_{ij}}$

= 3.637

Critical value, $x^2{}_{0.05,4} = 9.488$

p-value = 0.4574

It is always a right tail test. Reject H_0 if x_0^2 > 9.488

Decision:

Since x_0^2 = 3.637 < 9.488 and p-value = 0.4574 > α=0.05 We fail to reject the null hypothesis, H_0 because the test statistics value does not lie within the critical region.

Conclusion:

There is sufficient evidence to conclude that the variables fathers' jobs and variable higher (wants to take higher education) are independent at 0.05 significance level. There is no relationship between variables Fjob and higher.

DISCUSSION

Test 1 is test using 10% significance level means that there is still 0.10 probability to reject null hypothesis when it is true. The rest of the tests are test using 5% significance level means that there is still 0.05 probability to reject null hypothesis when it is true.

Firstly, from test 1, test 2 and test 3**.** I found out that the secondary school students with the urban address is the majority. By comparing 3 tests above, I found that mean of students' grades (G1, G2, G3) in urban address is higher than the mean of students' grades (G1, G2, G3) in rural address.

By comparing test 4, test 5, test 6. There is no relationship between variables

- **number of school absences (absences) and first period grade in subject Mathematics (G1)**
  The relationship between variables absences and G1 is extremely weak and the hypothesis testing shows that there is no linear correlation exist between these two variables When number of school absences increase, first period grade in subject Mathematics will not tend to either increase or decrease.

- **number of school absences (absences) and second period grade in subject Mathematics (G2)**
  the relationship between variables absences and G2 is extremely weak and the hypothesis testing shows that there is no linear correlation exist between these two variables. When number of school absences increase, second period grade in subject Mathematics will not tend to either increase or decrease.

- **number of school absences (absences) and final grade in subject Mathematics (G3)**
  The relationship between variables absences and G3 is extremely weak and the hypothesis testing shows that there is no linear correlation exist between these two variables. When number of school absences increase, final grade in subject Mathematics will also will not tend to either increase or decrease.

What I get by comparing these three tests is there is no relationship between number of school absences and all the period grade in the subject Mathematics. When number of school absences increase, all the grades includes first period, second and final grades in subject Mathematics will also will not tend to either increase or decrease.

When analysing the scatter plot in test 7 that applying the **Spearman's Rho Rank Correlation Coefficient.**There is relationship between variables going out with friends (goout) and free time after school (freetime) but it is weak. The direction of variables also cannot see clearly from the scatter plot above. However, we can obtain the direction through correlation analysis and the r_s is between -0.5 and 0.5 In order to provide a strong evidence, hypothesis testing is done and the result shows that there is relationship between these 2 variables. In other word, I can say that when the free time after school increase, going out with friends will also increase too.

From test 8, I found that weekly study time (studytime) will affect the first period grade in Mathematics subject(G1). However, there is only 2.58% of G1 in y is explained by studytime in x. To be more detailed, the regression line in this test is $\hat{y}$ = 9.6159 + 0.6352x means that students will still get about 9.6159 marks even though they did not study at all. At the same time, if there is an increase of 1 hour in weekly study time, they will get an increase in the first period grade by 0.6352 marks. From test 9, weekly study time (studytime) will affect the second period grade in Mathematics subject(G2). There is 1.846% of second period of grades (G2) in y is explained by weekly study time (studytime) in x through the Coefficient of determination, R-squared. The regression line $\hat{y}$ = 9.474 + 0.609x obtained shows that students will still get about 9.474 marks if they did not study at all. Then, if there is an increase of 1 hour in weekly study time, they will get an increase in the second period grade by 0.609 marks. From the test 10, there is no linear relationship exist between Weekly study time(studytime) and final period grade in Mathematics subject(G3) and the coefficient of determination is almost 0% provides a strong evidence of no existence of relationship between studytime and G3. The weekly study time(studytime) will not affect the final period grade in Mathematics subject(G3).

From test 11, there is at least 1 of the 4 proportions is different from others. This means that students will choose secondary school based on the course, home, reputation or other reason. One or more of them will have higher proportions compared with other. By comparing test 12 and test 13, I found that students' decision to take the higher education is not influenced by their mothers and fathers' jobs. We cannot estimate or predict students' decision to take higher education based on their parents' jobs.

CONCLUSION

The specification of target population is secondary school students who take Mathematics subject in Portugal. From the test 1, test 2 and test 3, students in urban area have higher grades in subject Mathematics compared to students in rural area. However, students' grades in subject Mathematics is not affected by their absences. We can say that although students' absences more, they will also get a good result. Or we can just say that students with full attendance maybe will get poor grades. In conclusion, when number of school absences increase, the three grades in subject Mathematics are not tend to either increase or decrease.

In words, the weekly study time of secondary students will affect their mathematics first and second periods of Mathematics grades. However, there are only weak relationships. There is evidence to show that 2.58% of first period and 1.846% of Mathematics grades can be explained by their weekly study time. Based on these 2, we can try to estimate students' grades by observing their weekly study time. However, the weekly study time will not cause an increase or decrease in final period grade in subject Mathematics. In the aspect of social, secondary school students nowadays will go out with friends when they have more free time after school and nowadays students will choose a secondary school based on some specific reason. Secondary school students who take subject mathematics in Portugal will not make decision to continue a higher level education based on their parents' jobs.

In a nutshell, I can conclude the background of students only have little influences on the mathematics grades. Living area will affect the grades but parents' jobs don't. Parents' jobs also do not affect the students' future study plan. Although students have high absence rate, their mathematics grades will not tend to increase or decrease. In conclusion, background and attitudes of students not really affect the secondary school students' grades in Mathematics subject in Portugal.

REFERENCES

(27 March, 2020). Retrieved from The Portugal News: https://www.theportugalnews.com/news/50-of-portuguese-between-25-and-64-did-not-complete-secondary-school/53517