



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

PROBABILITY AND STATISTICAL DATA ANALYSIS

SECI2143

SECTION: 02

LECTURER: DR NOR AZIZAH ALI

GROUP: 1

PROJECT 2: STUDENT PERFORMANCE DATA ANALYSIS

GROUP MEMBERS:

No	Name	Matric No
1	MUHAMMAD FARHAN BIN IBRAHIM	A21EC0072
2	MUHAMMAD ADAM FAHMI BIN MOHD TAUFIQ	A21EC0061
3	MUHAMMAD FIKRI BIN SHARUNAZIM	A21EC0075
4	MIKHAIL BIN YASSIN	A21EC0053

Table of contents

Introduction	3
Dataset	4
Statistical Analysis on Case Study	5
Execution of Tests Overall Execution – Compulsory Tests	6
Overall Execution – Optional Tests	12
Conclusion	14
References	15

Introduction

Inferential statistics are involved with producing inferences based on relationships found in the sample, to relationships in the population. Inferential statistics help us decide, for instance, whether the distinctions in the middle of groups that we see in our data are powerful and strong enough to provide support or claim for our hypothesis that group differences exist in common, in the whole population. In this Project 2 for Probability and Statistical Data Analysis course, we will implement all the necessary items we learned previously in making a conclusion. The conclusion that is to be made must adhere to several procedures such as producing a hypothesis statement, testing the hypothesis using appropriate methods and finalizing a decision. As a result, we gathered a set of data on student performance in Math and Portuguese in secondary school in order to conduct some analysis and draw conclusions based on our chosen dataset. This dataset was retrieved from a <https://www.kaggle.com/code/devansodariya/student-performance-analytics> website which was uploaded by Dev Ansodariya, who is a student at L.D. College of Engineering Ahmedabad, Gujarat, India. This data gathered a lot of interesting features like social, gender and study information about students. This includes the number of school absences, student's guardian, weekly study time, home-to-school travel time and so on.

In this era, we can see many reasons that can affect a student's performance in school. For instance, a student will perform if they spend a lot of time studying. Everyone has their own reasons and behaviours that affect their performance in school. That's why this data exists. The thing that we are interested in this question is whether we can decide and make suggestions for the situation of students who perform in their course subject. This question allows us to draw conclusions based on the analysis. It is in that way fundamentally different from descriptive statistics which merely summarize the data that has actually been measured. Can we find out the evidence of students performing more in their second period rather than a first period? Furthermore, in this data set, we expect to find out what factors make the students perform well in Math or Portuguese.

Dataset

In this project, we retrieved the data set from the Kaggle platform. The data set is about Student Performance. Through this data set, we are able to get the data on what has affected students' academic performance. The variables provided in this data set were their sex, age, health, guardians, students' grade, students' absence and the frequency of students usually going out with friends.

Around 363 students become respondents to the data set. They were students aged from 15 to 22 from school Gabriel Pereira and Mousinho da Silveira. The response from the respondents is very usable for us to conduct this project successfully. From the data between G1 and G2 we decide to create a hypothesis between the sample test. Other than that, the data between them also can create the correlation. As for the regression, we use the number of school absences that the student have and the rate number of times the student go out with friends. Lastly, we take the current health status and the student's guardian data for contingency using chi-square test of independence.

No.	Variables	Answer	Level of Measurement
1.	G1 (first-period grade)	0, 1, 2, ..., 20	Ordinal
2.	G2 (second-period grade)	0, 1, 2, ..., 20	Ordinal
3.	Study time	1 - (<2 hours) 2 - (2 to 5 hours) 3 - (5 to 10 hours) 4 - (>10 hours)	Ordinal
4.	Health (current health status)	1, 2, 3, 4, 5 (1-Very Bad to 5-Very Good)	Ordinal
5.	Guardian (student's guardian)	'Father' and 'Mother'	Nominal
6.	Absence (number of school absences)	Metric value	Ratio
7.	Go out (going out with friends)	1, 2, 3, 4, 5 (1-Very Low to 5-Very High)	Ordinal

Statistical Analysis on Case Study

Hypothesis statement

In my retrieved data for this Student Performance in Math or Portuguese for secondary school, the source does not include any specification on reliability between one factor to another. The data is free to consider any of the aspects that may suit our case study and statistical data to be analysed soon. Hence, We will be considering two-sample tests to compare the proportion for the first-period grade (G1) and second-period grade (G2). Based on the dataset, G1 which is the first-period grade has 363 samples. The same goes for G2, second-period grade. Both grades are numeric from 0 to 20. These grades are related to the course subject, Math or Portuguese. The definitions of proportion are as follows.

μ_1 = mean of first-period grade over a 363 data for G1.

μ_2 = mean of second-period grade over a 363 data for G2.

Next, the null and alternative hypotheses are to be defined. The mean of first-period grade over 363 data for G1 is equal to the mean second-period grade over 363 data for G2 will be the null hypothesis while the mean of first-period grade over 363 data for G1 is not equal to the mean second-period grade over 363 data for G2 included will be the alternative hypothesis.

The notations of the hypothesis are defined as follows.

H_0 = The mean of first-period grade over 363 data for G1 is equal to the mean second-period grade over 363 data for G2

H_1 = The mean of first-period grade over 363 data for G1 is not equal to the mean second-period grade over 363 data for G2

Therefore, these long definitions may be simplified into mathematical terms as follows to make us easier while doing the test.


$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Execution of Tests Overall Execution – Compulsory Tests

Hypothesis Dependant Two-Sample test:

We use a 0.05 significance level to test the claim that we made.

Data	
student_data	363 obs. of 7 variables 
Values	
alpha	0.05
g1	num [1:363] 5 5 7 15 6 15 12 6 16 14 ...
g2	num [1:363] 6 5 8 14 10 15 12 5 18 15 ...
n1	363
n2	363
s1	3.33185971109804
s2	3.76334259871914
t.alpha	1.96653881250995
tstats	1.3826

From the dataset process by using R we can make a conclusion regarding the value we get above.

Components	Values/Explanations
Test statistics	$t = 1.3826$
Critical Value	$cv = 1.9665$
Decision	Since, the test statistic value (1.3826) < critical value (1.966), thus we fail to reject H_0 at $\alpha = 0.05$.
Conclusion	There is not enough evidence to support that at a 0.05 level of significance, the mean of first-period grade is equal to the mean of second-period grade.

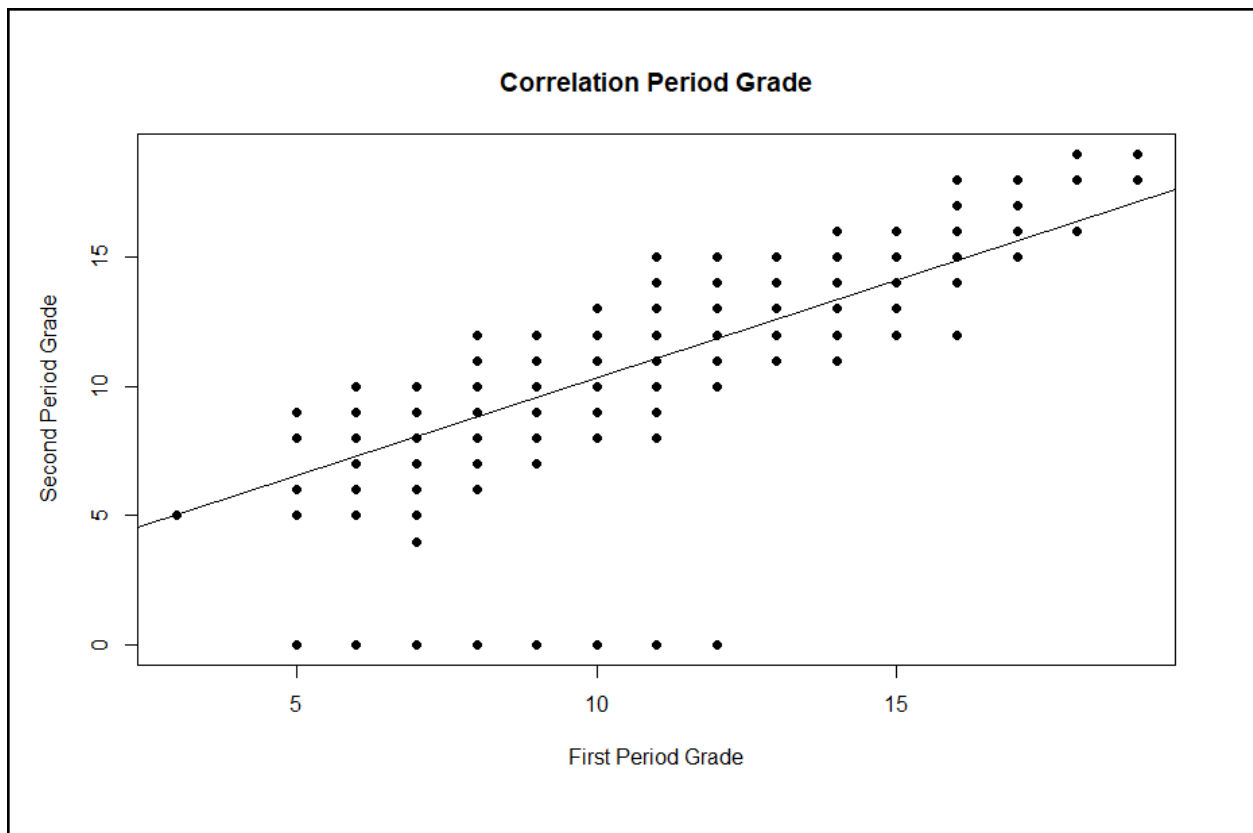
Correlation in First Period Grade and Second Period Grade

For this test, a random sample of 363 students were selected to check if there is a linear relationship between students' first-period grade (G1) and second-period grade (G2) at the 5% level of significance ($\alpha = 0.05$). Variables used in this test are first-period grade (G1) and second-period grade (G2).

Hypothesis statement:

$H_0: \rho = 0$. There is no linear correlation between first-period grade (G1) and second-period grade (G2).

$H_1: \rho \neq 0$. There exists a linear correlation between first-period grade (G1) and second-period grade (G2)



The graph above shows the relationship between first-period grade and second-period grade is high. So, the second-period grade affects the first-period grade. It is because students tend to study more to get better results for their tests or examinations, also students want to obtain better results from previous exams. The scatter plot and correlation analysis of the data indicates that there is a positive relationship between first-period grade and second-period grade.

```

Pearson's product-moment correlation

data: student_data$G1 and student_data$G2
t = 30.926, df = 361, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8211288 0.8779772
sample estimates:
      cor
0.8520459

```

From the data above, the r-value is 0.8520459 which indicates that there is a strong relationship between the variables tested. By looking at the result process by R above we can make the conclusion.

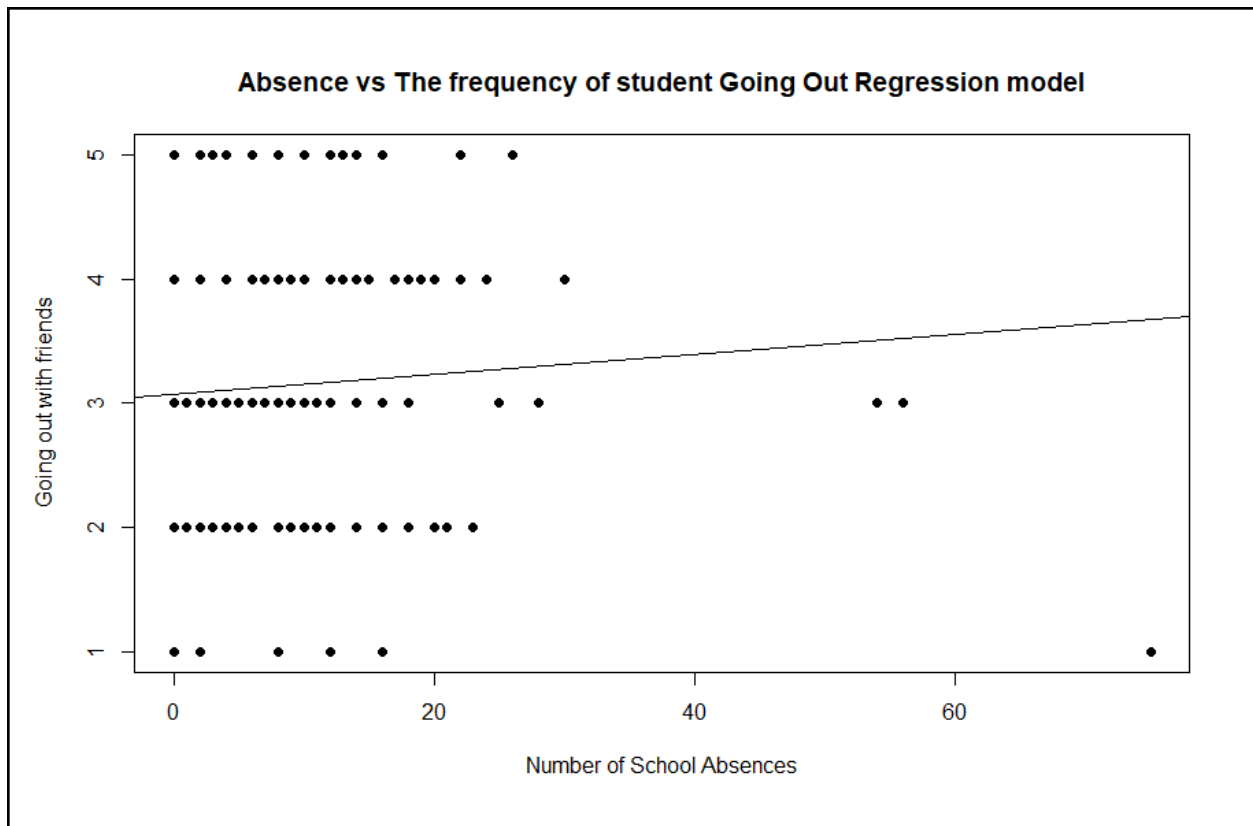
Components	Values/Explanation
Decision	$t = 30.926 > cv = 2.2e - 16$ Therefore, we reject H_0 .
Conclusion	There is sufficient evidence to support that at a 0.05 level of significance, there is a strong positive linear relationship between first-period grade and second-period grade.

Regression Test on number of school absences and going out with friends

For this test, the random sample of going out with students and the number of school absences among students in the data. We want to see if the number of school absences can predict the students going out with their friends at the 5% level of significance ($\alpha = 0.05$). The dependent variable (y) for this test is how often the students go out with their friends and the independent variable (x) for this test is the number of school absences. The objective of this test is to ensure the existence of a linear relationship between the dependent variable (y) and the independent variable (x).

$$H_0: \beta_1 = 0 \text{ (no linear relationship)}$$

$$H_1: \beta_1 \neq 0 \text{ (linear relationship exists)}$$



From the graph obtained, we can see that there is a positive linear relationship between the number of school absences by student and the rate number of times they go out with their friends.

```

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
  3.069418      0.008099

Call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-2.67684 -1.06942 -0.09372  0.89819  1.93058

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.069418   0.071676  42.824  <2e-16 ***
x            0.008099   0.007628   1.062   0.289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.12 on 361 degrees of freedom
Multiple R-squared:  0.003113, Adjusted R-squared:  0.0003514
F-statistic: 1.127 on 1 and 361 DF, p-value: 0.2891

b0 = 3.0694  b1 = 0.0081  sb1 = 0.007628  p-val = 0.000000000000000022
se = 1.12  t = 1.062

```

From the data process in R programming we found out the least square regression equation to be ($\hat{y} = 3.0694 + 0.0081x$). The value of intersection coefficient, β_0 is the estimated value of Y when the value is zero which is 3.0694. This will indicates that the value observed is the proportion of a number of school absences by student when there are no time they go out with their friends. The value of slope coefficient, β_1 measures the estimated change in the estimated change in the average value of Y when there is a change in X.

Here, β_1 is 0.0081 which tells us that the average of rate number of going out increase by 0.0081 on average, for each additional a number of school absence.

From the calculation in R above , the Regression Equation is

$$\hat{y} = 3.0694 + 0.0081x$$

Test Statistic:

$$t = \frac{b1 - B1}{sb1}$$

$$t = \frac{0.0081 - 0}{0.007628}$$

$$t = 1.062 \text{ (test statistic)}$$

Critical Value:

degree of freedom, $df = 363 - 2 = 361$

significance level, $\alpha = 0.05$

$t_{0.05, 361} = 1.960, -1.960$

Components	Values / Explanation
Decision	Since, the test statistic value (1.062) < critical value (1.960), thus we fail to reject H_0 at $\alpha = 0.05$.
Conclusion	There is not enough evidence to support that at 0.05 level of significance, the number of school absences affects going out with friends. There is no linear relationship between them.

Overall Execution – Optional Tests

Chi-Square Test of Independence in Two Way Contingency Test

We test the contingency of the guardian variables and the current health status of the students at the 5% level of significance ($\alpha = 0.05$) with a random sample of 363 students.

Health	Guardian	
	Father	Mother
1 (Very Bad)	5	41
2 (Bad)	15	24
3 (Moderate)	19	58
4 (Good)	14	49
5 (Very Good)	37	101

Hypothesis statement:

H_0 : The guardian variables and health do not have a relationship between them.

H_1 : The guardian variables and health do have a relationship between them.

Critical Value:

$$\alpha = 0.05$$

$$df = (5 - 1)(2 - 1) = 4$$

$$\text{Critical value} = 9.488$$

Expected counts:

Health	Guardian				Total
	Father		Mother		
	Observed	Expected	Observed	Expected	
1 (Very Bad)	5	11.40	41	34.60	46
2 (Bad)	15	9.67	24	29.33	39
3 (Moderate)	19	19.09	58	57.91	77
4 (Good)	14	15.62	49	47.38	63
5 (Very Good)	37	34.21	101	103.79	138
Total	90	90	273	273	363

The test statistic value:

Cell, ij	Observed Count, O_{ij}	Expected Count, E_{ij}	$[O_{ij} - E_{ij}]^2 / E_{ij}$
1,1	5	11.40	3.59
1,2	41	34.60	0.18
2,1	15	9.67	2.94
2,2	24	29.33	0.97
3,1	19	19.09	0.0004
3,2	58	57.91	0.0001
4,1	14	15.62	0.17
4,2	49	47.38	0.06
5,1	37	34.21	0.23
5,2	101	103.79	0.07
$\chi^2 =$			8.2105

The decision:

Components	Values/Explanations
Test statistics	$\chi^2 = 8.2105$
Critical Value	$\chi^2, k = 4, \alpha = 0.05 = 9.488$
Decision	Since, test statistic value (8.2105) < critical value (9.488), thus not to reject H_0 at $\alpha = 0.05$.
Conclusion	There is not enough evidence to support that at a 0.05 level of significance, the guardian variables and health do have a relationship between them.

By using the chi-square test of independence, it has shown that if a relationship exists between 2 qualitative variables which in this case the guardian and the current health status of student. This happens when one sample is drawn and does not show any casualty so that we can make assumptions. As we can see from the value we get above, the hypothesis we can make are the current health status and the student's guardian data do have a relationship between them. By using the two-way contingency table, we can conclude that the guardian will affect the well being of the student.

Conclusion

In conclusion, we found out there are a lot of factors that affect the performance of students in this dataset. Among the factors used are health, the number of students absent and the frequency of students usually going out with friends. For example, we can conclude that the health of students affects the students' grades. We found out that the good grades of the students basically came from the student themselves. If there is a will, there is a way.

The data that we retrieved and analyzed does not support strong enough evidence that we calculated. We noticed that after we have done our project, we have learned a lot of new things in order to study data and analysis such as using R programming to ease the process to calculate the mean, standard deviation etc of a big set of data. We found it a lot easier to use to do by practical ways to do analytical study. After analyzing the data that we are using, we became more observant of our surroundings and curious. We noticed that we can study every single thing that we want to study such as the population of people in a town and understand the real factors of issues through our analytical study. Thus, it's open wider our point of view of nature.

References

Devansodariya. (2022, May 26). *Student performance analytics*. Kaggle. Retrieved June 14, 2022, from <https://www.kaggle.com/code/devansodariya/student-performance-analytics>