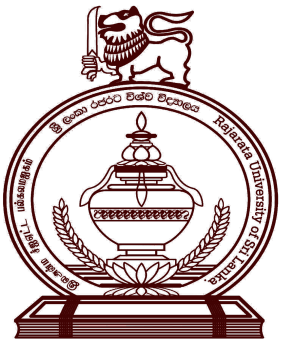


Probability and Statistics MAA 1302
Mid Semester Examination One - Report

Mohamed Infaz Rumy

ASP/18/19/021

Faculty of Applied Sciences

Rajarata University of Sri Lanka

Table of Contents

Introduction **1**
- Introduction about the Dataset 1
- Variables in the Dataset 1
- Objectives of the study 3

Methodology **4**
- Methodologies used for the analysis 4
- Softwares used to analyze data 4

Analysing Data **5**
- Relationship between categorical variables 5
- Relationship between quantitative variables 6

Conclusions **8**

Chapter 1: Introduction

1.1 Introduction about the Dataset

The Data were obtained in a survey of students math and Portuguese language courses in the secondary school and the data was collected during the 2005- 2006 school year from two public schools, from the Alentejo region of Portugal which are 'Gabriel Pereira' and 'Mousinho da Silveira' The Dataset that is used in the report is taken from the www.kaggle.com. Data was originally collected to use data mining to predict students' performance.

1.2 Variables in the Dataset

There are 33 variables and 395 observations available in the dataset and Table 1 shows the descriptions of the attribute. The selected dataset consists of categorical (qualitative) and numerical (quantitative) variables, therefore, statistical analysis was done on this report based on this single dataset.

Table 1: Description of variables.

Attribute	Description of attribute
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4 a)
Mjob	mother's job (nominal b)
Fedu	father's education (numeric: from 0 to 4 a)
Fjob	father's job (nominal b)
guardian	student's guardian (nominal: mother, father, or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 bad to 5 excellent)
reason	reason to choose this school (nominal: close to home, school

	reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first-period grade (numeric: from 0 to 20)
G2	second-period grade (numeric: from 0 to 20)
G3	the final grade (numeric: from 0 to 20)

1.3 OBJECTIVES OF THE STUDY

Out of the thirty-three variables mention in the dataset four variables have been selected for the further analysing. The analysing is done under two categories which are also the objectives of the study,

- To find out an association between categorical (qualitative) variables.
- To find out an association between numerical (quantitative) variables.

Selected variables for analysis data are,

- Numerical (Quantitative)
    - G1 - First-period grade (numeric: from 0 to 20)
    - G3 - The final grade (numeric: from 0 to 20)
- Categorical (Qualitative)
    - famsup - Family educational support (binary: yes or no)
    - higher - Wants to take higher education (binary: yes or no)

Chapter 2: Methodology

2.1 Methodologies used for the analysis

The selected four variables are analysed under two section which are,

- To analyse the association between the categorical variables data displayed in a two-way table and used the Chi-square statistic and made a decision.
- To analyse the association between the quantitative variables T Statistics has been used and made decisions based on it.

2.2 Softwares used to analyze data

The below table represents the Softwares and tools that have been used to analyze the dataset

Table 2: Software description

Software	Description Of the software
Google Sheets	To analysis the Dataset

Chapter 3: Analysing Data

3.1 Relationship between categorical variables

Step 1 Hypothesis

Null hypothesis H_0 : Student who is willing to get higher education doesn't have their family support

Alternative hypothesis H_a : Student who is willing to get higher education to have their family support

Step 2 The Chi-square Statistic

Table 3: Two-way table

Higher education	Family support	Family not support	Total
want to take higher education	234	141	375
do not want to take higher education	8	12	20
Total	242	153	395

Expected count = (Row total) x (Column total) / Total n for table

Table 4: Table of expected counts

Higher education	Family support	Family not support	Total
want to take higher education	229.7468354	145.2531646	375
do not want to take higher education	12.25316456	7.74683544	20
Total	242	153	395

Chi squared value = 4.014649165

Degrees of freedom: 1

Step 4 Making and Reporting a Decision

The P-Value is .045108. The result is significant at $p < .05$. Therefore we can avoid the null hypothesis which is that students who are willing to get higher education don't have their family support.

3.2 RELATIONSHIP BETWEEN QUANTITATIVE VARIABLES

Step 1 Hypothesis

Null hypothesis H_0: There is not a significant linear relationship(correlation) between First-period grade and the Final grade.

Alternative hypothesis H_a : There is a significant linear relationship between First-period grade and the Final grade.

Step 2 Significance level of $\alpha = 0.05$

Step 3 Critical value $t_{(\alpha/2,\ n-2)} = t_{(0.05/2,\ 395-2)} = 1.966$

Step 4 Test Statistic

Correlation between the First-period grade and the Final grade is 0.8014

0<r<1 this means positive correlation

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

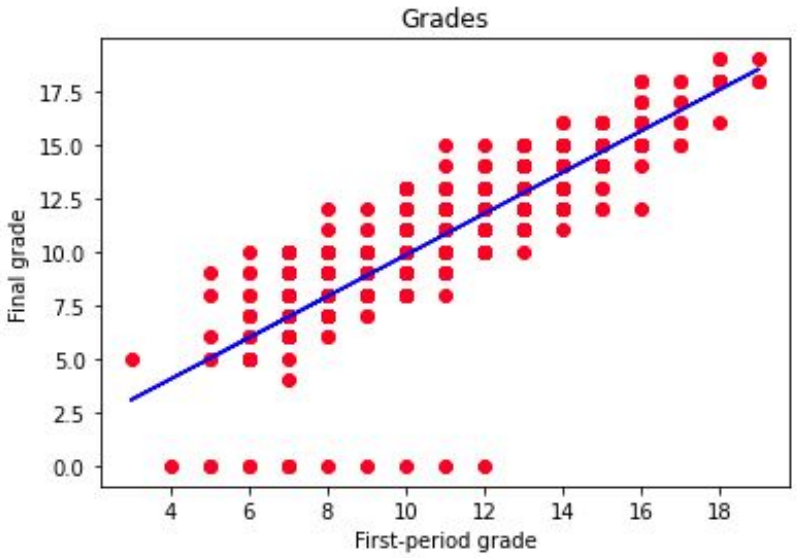
n= 395 , r = 0.8014 , r^2= 0.6422

t* = 44.402

Step 5 Draw the conclusion

44.402 > 1.966

Therefore we can avoid the null hypothesis which is that there is not a significant linear relationship (correlation) between First-period grade and the Final grade.



$Y = b_0 + b_1X$

X is the explanatory variable and Y is the dependent variable

b_0 is the intercept which is -1.6528 and b_1 is the slope of the line which is 1.1062

predicted Final Grade = 1.1062 (First period grade) - 1.6528

Descriptive statistics related to Final grade and First-period grade

	Mean	Std. Deviation	N
Final grade	10.41519	4.57564	395
First-period grade	10.908861	3.31499	395

Chapter 4: Conclusions

For analysing the dataset, four variables have been selected among the 33 variables in the dataset: two categorical variables and two quantitative variables.By analysing those two categorical variables and two quantitative variables, made two decisions.

References

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[Web Link]