Statistical Data Analysis of Student Goals

Mateusz Zaremba

November 4, 2019

Contents

1 Abstract

This should be a very brief explanation of your research paper (around 150 words). It normally includes information about the issue, why you are interested in that issue, your method/model, analysis results, discussions and conclusions.

This paper analyses the data gathered from surveying 625 undergraduate students. The authors of the survey tried to prove two hypothesis: 1) During students' junior years, they tend to primarily focus on getting good grades while during their senior years, the focus shifts towards a deep-understanding of the subject and 2) students' enjoyment and interest

tends to deteriorate as they progress through their studies. It is not obvious why this might be the case and if the student's sex or studied subject has any bearing. This is why the survey has 15 questions and probes 7 assessment categories. Each category consists of 3 to 1 questions and because the order of the questions is randomised, the student should not know the categories nor notice any patterns.

The data manipulation was done using R and tidywerse packages. A full analysis will be presented, including data: preparation, analysis, exploration and interpretation; calculation of confidence interval for a proportion, interpretation of the results using different kinds of graphs and an explanation of the methods used.

2 Introduction

This section should explain the topic, why it is important, and how you approach the issue

It is interesting how undergraduate students' goals change through-out their studies. They often experince various syndroms like: burnout, impostour, disheartening or even attempt a suicide. A Harvard graduate, Alex Chang, in his TEDx talk titled "The Unspoken Reality Behind the Harvard Gates" speaks about the pressure of gettings the best grades; how he was called for a jasmine tea to his tutor and asked if he couldn't give it his all, while he already was doing the best he could. He also recalls one tragic night when he and his roommates were woke up at 4am, to be informed that one of his friend has taken his own life.

Becasue this paper is going to be talking about student's course enjoyment, expectations and his or her focus on grades vs. understanding I would liketo give it another, less visible shade for there might be a lot more to say about a student who is at the bottom of the scale. It was assumed that a student, who might be at risk of developing mental health problems, would be someone who: is not enjoying the course, finds it not interesting but still primarly aims to perform better than others, and is lead by the fear of performing poorly. We will try to identify such students, calculate the confidence-interval-for-a-proportion of finding them, and test the hypotheses.

On top of this, we will explore the data and see if there are any other patterns we could draw conclusions from.

3 Data

Explain your dataset and how the was data is collected – e.g. your sampling strategy or information given by the project information. The data we will be analysing was originally sourced from *Elliot, A. J. and McGregor, H. A. (2001). A 2 x 2 achievement-goal framework. Journal of Personality and Social Psychology, 80, 3, 501-519.*

3.1 Initial Data

The data was converted from the original `.xlsx` format to `.csv` using *Microsoft Excel for Mac* and then it was loaded to *R* script using the *tidyverse* package - *readr*.

The initial number of students was also saved in a variable for later calcuations.

This is how the data looked like after loading it into the *R* script and before cleaning:

year	age	sex	subject	q1	q2	...	q12	interest	enjoy	mastgrad
3	19	1	1	7	2	...	5	7	7	1
3	20	2	1	7	2	...	2	6	6	4
3	21	1	1	1	1	...	1	7	7	1
3	NA	2	1	4	2	...	2	7	7	4

3.2 Cleaning the Data

3.2.1 Dropping

First, the *seq* column was dropped since it does not serve any purpose. Second, rows with empty cells were dropped because they could falsify the results.

3.2.2 Coding

The following coding informations was applied to the data:

Subject	Sex	Code
Management	Male	1
Law	Female	2
Tourism	-	3
General Economics	-	4
Accounting	-	5
Statistics	-	6

E.g., the code for *Male* was *1*, so the cells in the *sex* column containing *1* were replaced with a *Male* string; the code for *General Economics* was *4*, so the cells in the *subject* column containing *4* were replaced with a *General Economics* string. This is how the *sex* and *subject* columns looked like after applying the coding information:

sex	subject
Male	Management
Female	Management
Male	Management
Female	Management

3.2.3 Derandomization and Renaming

The students were presented with the questions in a randomized order. Becasue the random order was known, it was assumed that the collected data was put into the table in the random order. This would mean, that for proper data analysis the order would need to be derandomised, i.e., *question 6* from the *Performance avoidance* category is numbered *1* in the survey and in the data set. *question 12* from the *Mastery avoidance* category is numbered *2* in the survey and in the set, and so on. This also meant that in the process of derandomization previous lower case ***q*** letters in the column names, were changed to upper case ***Q***. E.g. of derandomization and renaming would be ***q1*** becoming ***Q6***.

Table 4 inlustrates the entire process:

Table 4: In the survey, questions from 1 to 12 were derandomized and renamed

Data set order	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12
Survey order	Q6	Q12	Q11	Q1	Q7	Q2	Q10	Q8	Q5	Q3	Q9	Q4

Questions 13, 14 and 15 - which were in fact, hypotheses testing questions and *Interest*, *Enjoyment*, *Importance focus* assessment-categories accordingly - kept the same order in the survey and in the data set, so they were only renamed for easier data manipulation:

Previous column name	New column name
interest	IG
enjoy	EJ
mastgrad	MG

3.2.4 Data collection

For a sake of data collection, I filled out the survey myself and added the results to the table.

3.3 Clean Data

This is how the data looked like after cleaning:

year	age	sex	subject	Q6	Q12	...	Q9	Q4	IR	EJ	MG
3	19	Male	Management	7	2	...	7	5	7	7	1
3	20	Female	Management	7	2	...	5	2	6	6	4
3	21	Male	Management	1	1	...	7	1	7	7	1
3	19	Female	Management	7	5	...	5	3	6	6	1

For the exact column order look at Table 4

4 Methodology

This section explains the statistical methods and/or your model. It is also a common practice to present the statistical model structure (i.e. equation) here as well.

The data was manipulated unsing `R` programming language and `tidyverse` packages:

- `ggplot2` - data visualisation
- `dplyr` - data manipulation
- `tibble` - data reimagining
- `readr` - data reading
- `tidyr` - data cleaning
- `purrr` - syntax simplification

Around 625 students were surveyed. They answered on a 7-level scale; 1 meaning the student felt the statement asked in the quesion is 'Not true of him/her' and 7 meaning the student felt it was 'Very true of him/her'. See the table below for a graphical explanation:

Not true of me	1	2	3	4	5	6	7	Very true of me

4.1 Interpretation

Apart from *Interest*, *Enjoyment* and *Importance focus* categories, all other consisted of 3 questions. For this reason, the assessment-categories' means were computed and saved for each individual student.

This resulted in 4 extra columns added to the original data set. An example of these can be seen below:

M1	M2	M3	M4
4.666667	4.333333	6.000000	3.333333
2.333333	2.333333	5.000000	2.000000
3.666667	1.333333	5.666667	1.333333
3.666667	3.666667	6.000000	5.333333
3.333333	3.333333	7.000000	4.000000

Table 10 presents the category, its label and interpretation (Based on the category's questions):

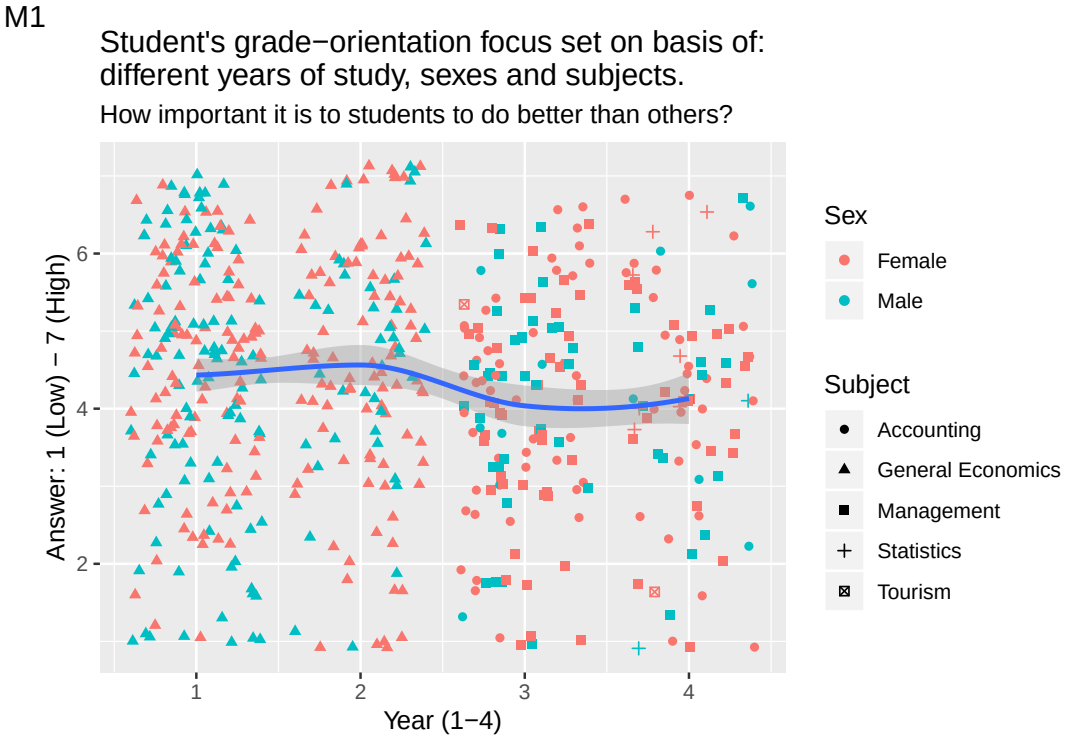
Table 9: Interpretation table

Category	Label	Interpretation
Performance approach	M1	Importance of doing better than others?
Performance avoidance	M2	Motivation based on the fear of performing poorly
Mastery approach	M3	Prevalence of mastery approach

Category	Label	Interpretation
Mastery avoidance	M4	Student's fear of not mastering the course
Interest	IR	Student expects the course to be interesting
Enjoyment	EJ	Student expects the course to be enjoyable
Importance focus	MG	Student's importance focus on understanding vs. grades

4.1.1 Data exploration

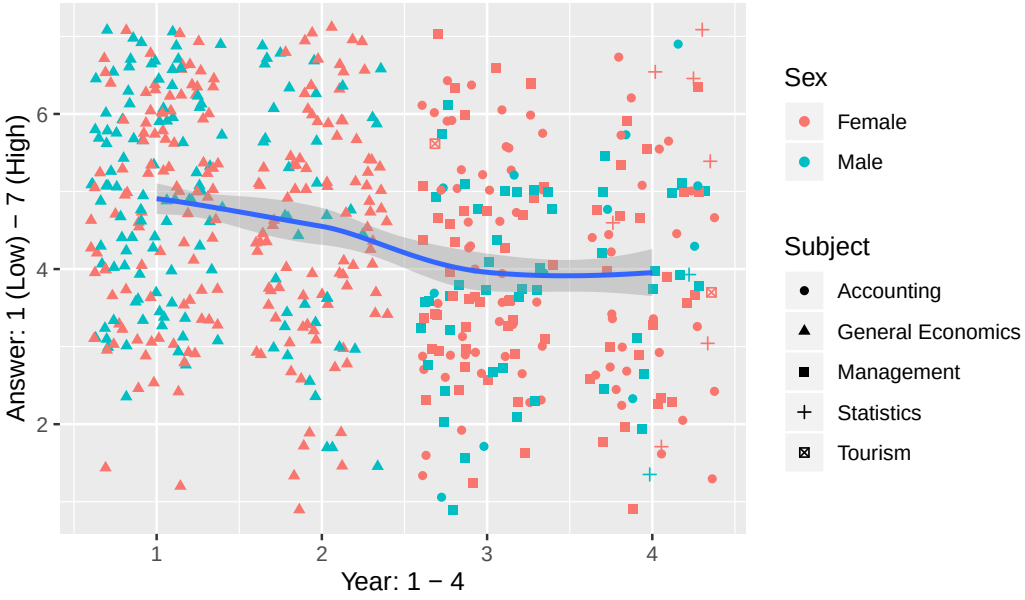
As a way of data exploration, the means from *M1*, *M2*, *M3* and *M4* assessment categories were plotted on basis of sex and subject



M2

Students' grade−orientation focus set on basis of: different years of study, sexes and subjects.

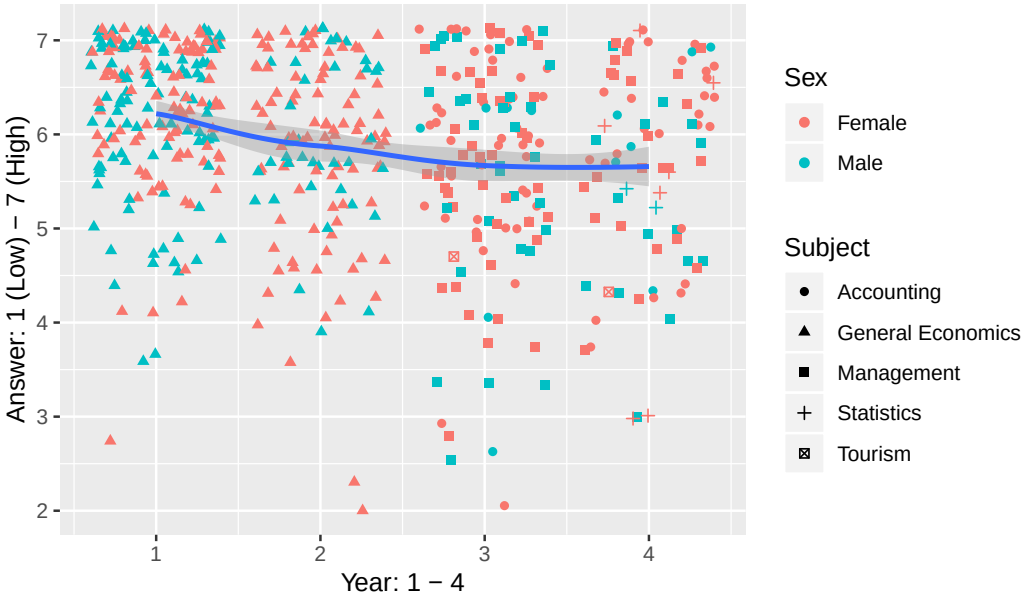
How motivated are students by fear of performing poorly?



M3

Students' focus on understanding set on basis of: different years of study, sexes and subjects.

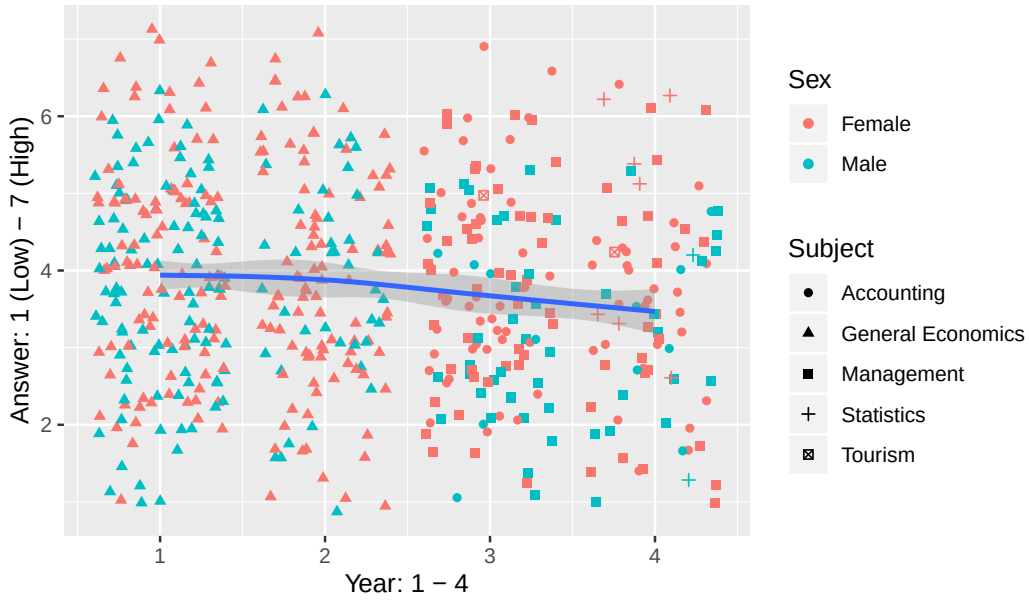
Prevalence of mastery approach.



M4

Students' focus on understanding set on basis of: different years of study, sexes and subjects.

Students' fear of not mastering the course.



4.2 Confidence Interval for a Proportion

4.2.1 Equations

$$\hat{p} = \frac{x}{n} = \frac{events}{trials}$$
$$Z = 1.96$$

$$\hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

4.2.2 Calculations

Calculating a 95% confidence interval for the student-at-risk population proportion.

$$\hat{p} = \frac{2}{625} = 0.0032$$

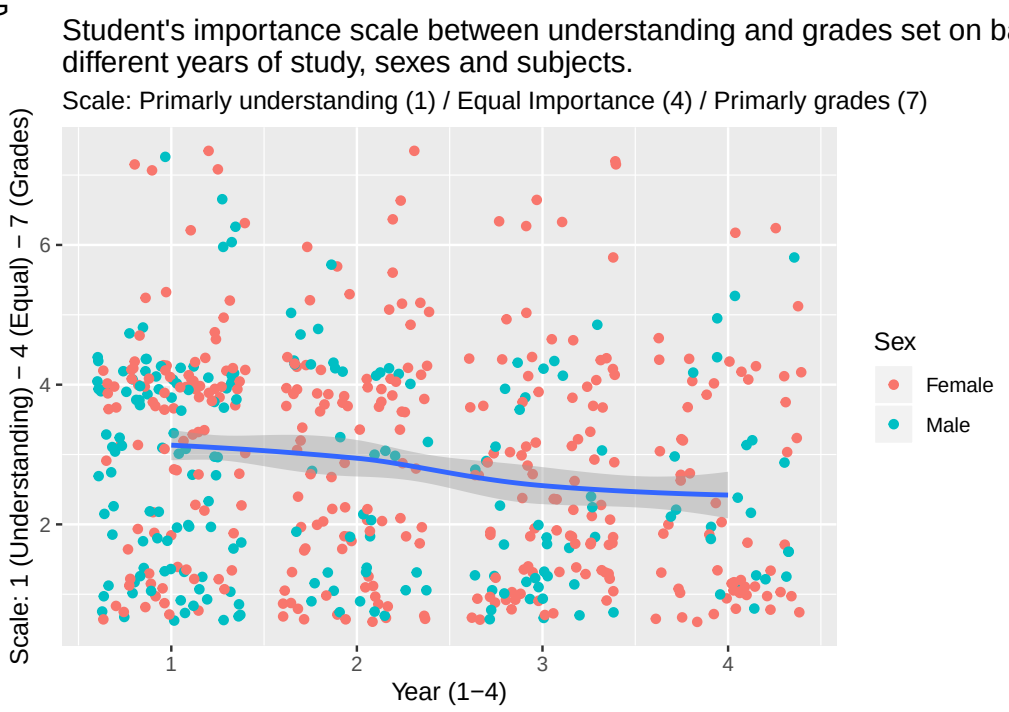
$$\hat{p} \pm Z = 0.0032 \pm 1.96\sqrt{\frac{0.0032(1-0.0032)}{625}} = 0.0032 \pm 0.0044$$

5 Results

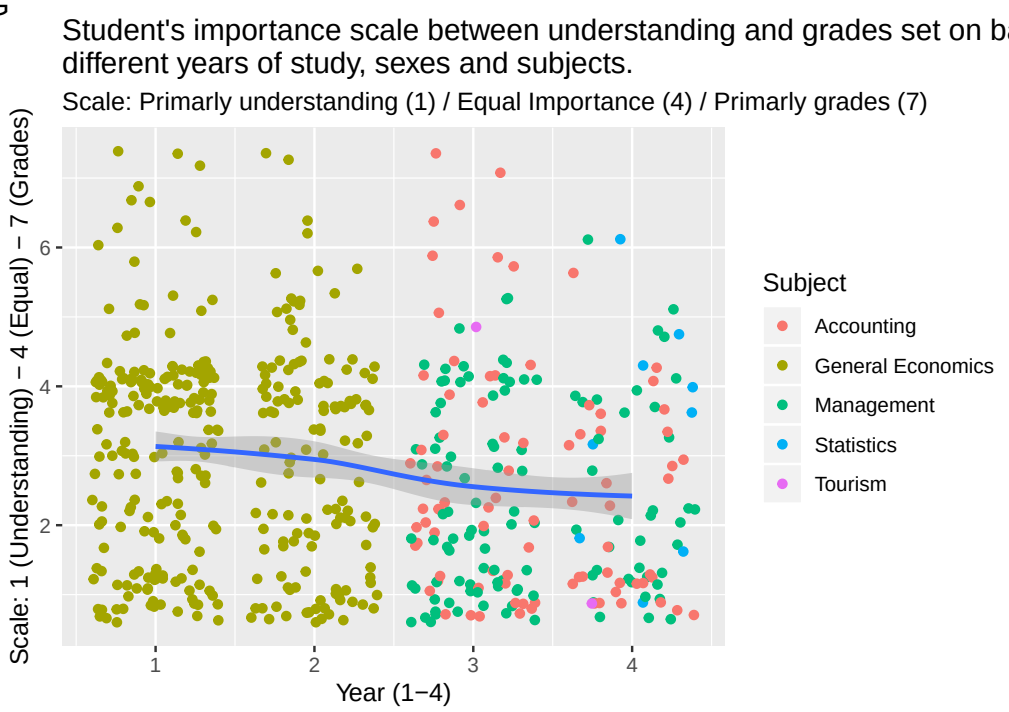
Present both your informal and formal analyses. We are going to test the hypothesis 1) and 2) on basis of sex and subject and we are going to find out if, and how many students are at risk of developing mental health issues.

5.1 Mastery Approach

MG



MG

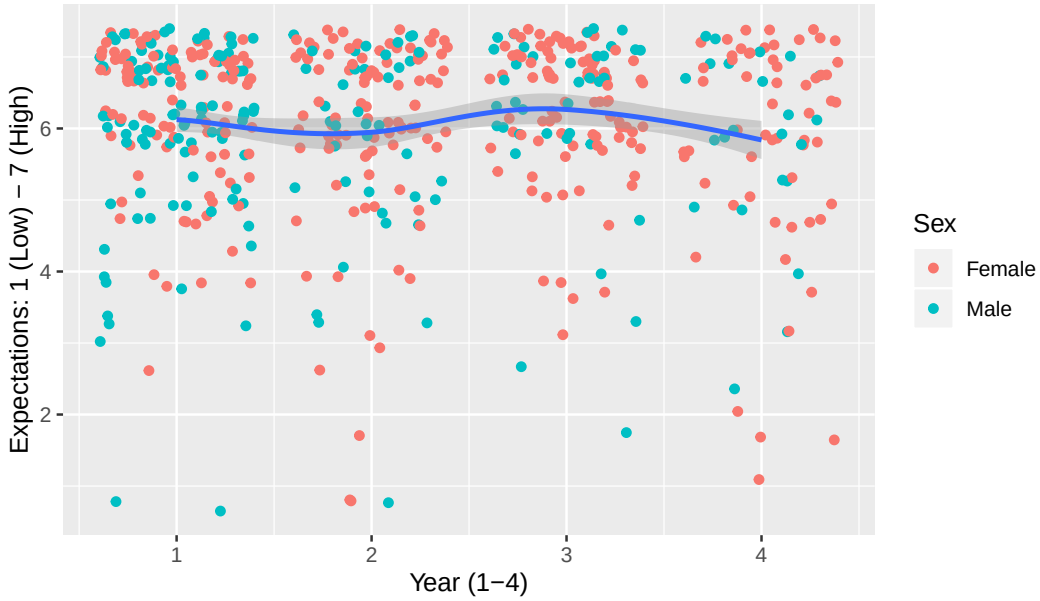


5.2 Interest

IR

Student's course interestedness expectations set on basis of: different years of study, sexes and subjects.

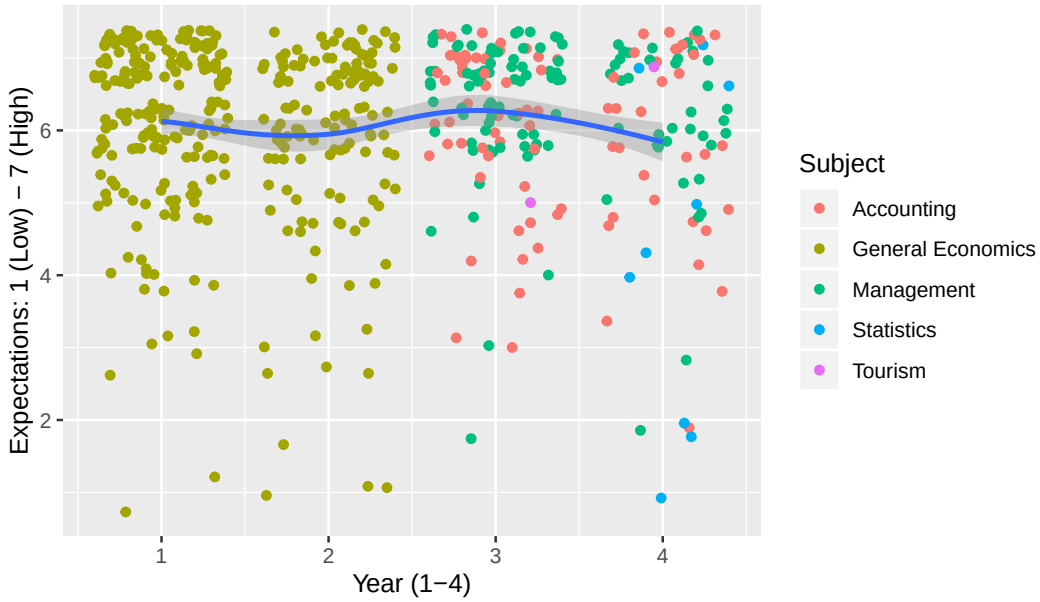
Student response: 'I expect my courses this semester to be very interesting'



IR

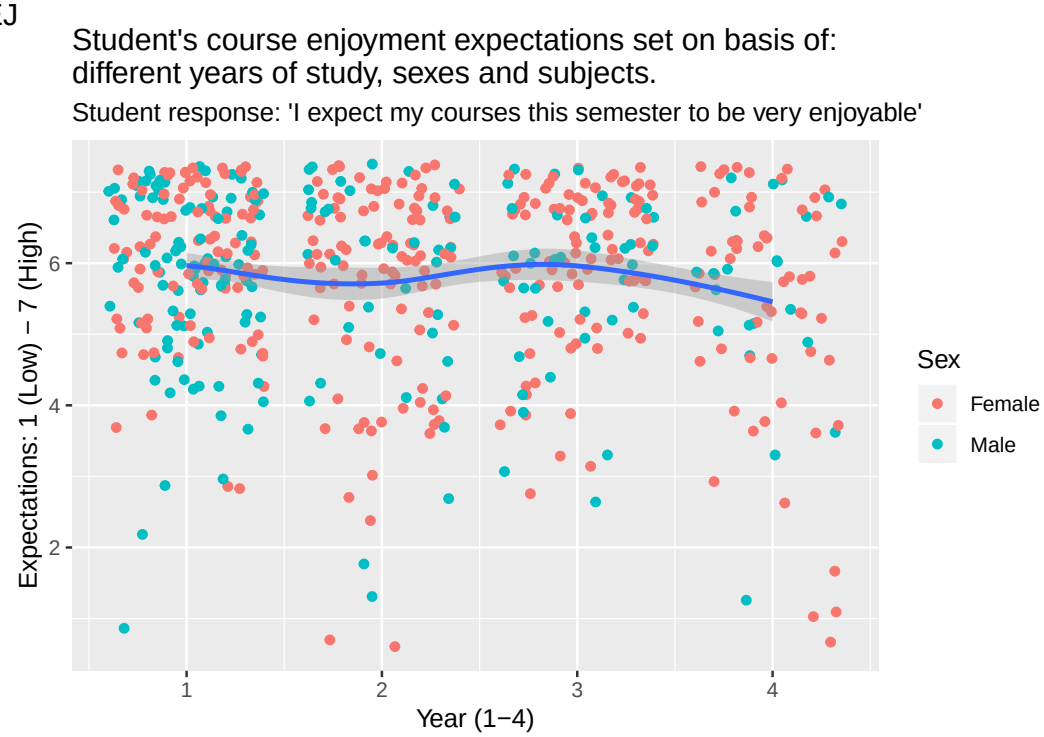
Student's course interestedness expectations set on basis of: different years of study, sexes and subjects.

Student response: 'I expect my courses this semester to be very interesting'

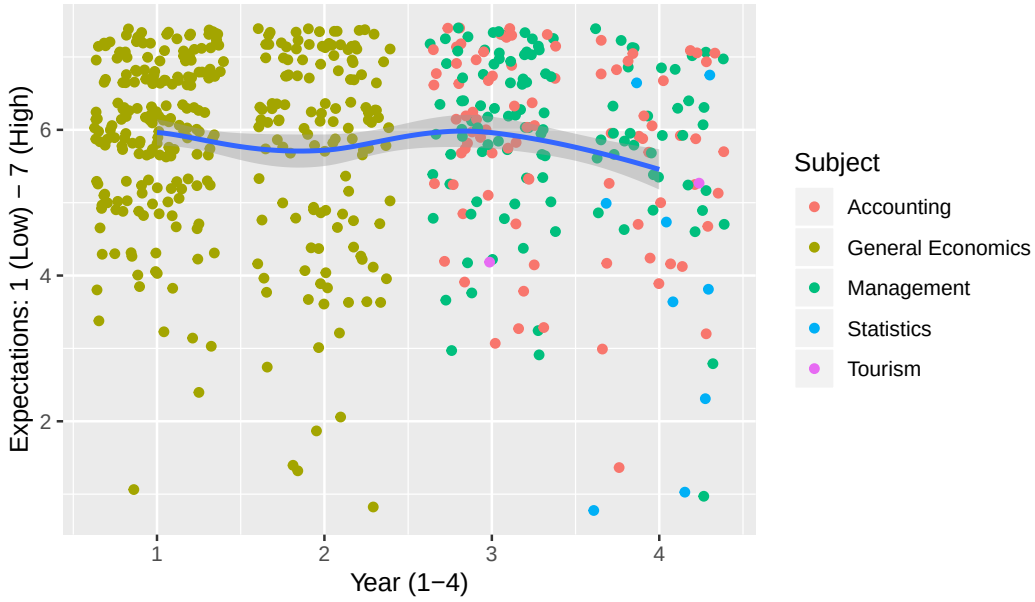


5.3 Enjoyment

EJ



EJ



5.4 Confidence Interval for a Proportion Interpretation

age	sex	subject	EJ	IR	MG	M1	M2
18	Female	General Economics	1	1	7	6	7
18	Female	General Economics	1	1	4	7	7

The upper confidence interval for a proportion is 0.0076279, and the lower is -0.0012279, which gives us a confidence interval, for proportion of finding a student-at-risk as: CI = (0.00762, -0.00122). This mean we could say with 95% confidence the percentage of the times we should expect to find a student at risk is between 0.7% and 0%.

6 Conclusion / Discussion

You need to conclude your project, discuss the results, discuss any reservations that you have about the study and list any future work.

The hypothese turned out to be true/false?

This might not seems like a lot but we have the actual data to back it up. We know that there are 2 students who might be at risk of developing mental health problems: they are both female, aged 18 years old and study general economics at their sophmore year. Wheter this data should be used directly to try to find these students is a different debate. What could be definietly done is try to announce this fact that there have been found students who might be at risk and that there is help available. Most universities offer counceling for students but the problem is to have already troubled and lonely students reach out for help. Maybe a university could develop their own programmes for sutdents which they can use to self-diagnose and if the system would detect that they might be experiencing mental health problems they could direct them towards a councelor at their univeristy. Such tests already exists but the reason why I think they should be university specific is for the fact that it makes the student feels that their university cares about their mental health and they do not need to feel ashamed to ask for help becasue it was their home institution who has made the first step.

Univerity should be the one reaching out, not the other way around.

so, what conclusions do you have bruh? Could analyse AGE