Statistical Data Analysis of Student Goals

Mateusz Zaremba

November 4, 2019

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 question 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

It is interesting how undergraduate students' goals change through-out their studies. They often experience various syndromes like burnout, impostor, 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 getting 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 woken up at 4am, to be informed that one of his friends has taken his own life.

Because this paper is going to be talking about student's course enjoyment, expectations and his or her focus on grades vs. understanding I would like to 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 primarily aims to perform better than others, and is led 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.

3 Data

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 Collection

The data was already collected but my own survey results were added to the data set.

3.2 Initialising

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 calculations.

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

3.3 Cleansing

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.

3.4 Coding

The following coding information 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

3.5 Assessment Categories

The survey's instructions provided 7 assessment categories and their appropriate questions. You can see them below with added labelling and interpretation.

Table 4: Interpretation table

Category	Label	Interpretation	Question
Performance Approach	M1	Importance of doing better than others	1, 2, 3
Performance Avoidance	M2	Motivation based on the fear of performing poorly	4, 5, 6
Mastery Approach	M3	Students' grade-orientation focus	7, 8, 9
Mastery Avoidance	M4	Students' fear of not mastering the course	10, 11, 12
Interest	IR	Student expects the course to be interesting	13
Enjoyment	EJ	Student expects the course to be enjoyable	14
Importance focus	MG	Students' importance focus on understanding vs. grades	15

3.6 De-randomization and Renaming

Random order (6, 12, 11, 1, 7, 2, 10, 8, 5, 3, 9, 4) was given to use in the survey's instructions. This meant, e.g., *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.

The column naming changed in the process: ***q1*** became ***Q6***, changing lower case ***q*** to upper case ***Q*** as well as the number of the question.

Table 4 illustrates the entire process and assigns questions to their appropriate categories:

Table 5: 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

3.7 Clean Data

This is how the data looked like after cleaning:

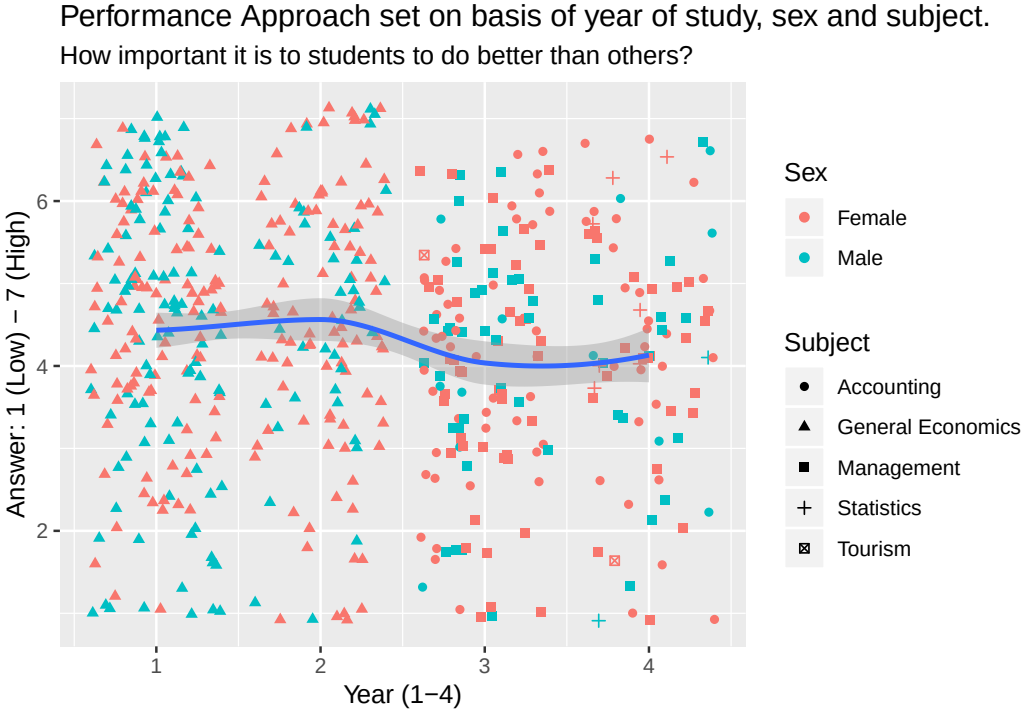
year	age	sex	subject	Q6	Q12	...	Q9	Q4	IR	EJ	MG
3	19	Male	Management	7	2	...	7	5	7	7	1
3	20	Female	Management	7	2	...	5	2	6	6	4
3	21	Male	Management	1	1	...	7	1	7	7	1
3	19	Female	Management	7	5	...	5	3	6	6	1

3.8 Visual Exploration

For all the students, the means from *M1*, *M2*, *M3* and *M4* assessment categories were shown on the graphs below; data smoothing was applied in attempt to capture any important patterns.

3.8.1 M1 - Performance Approach

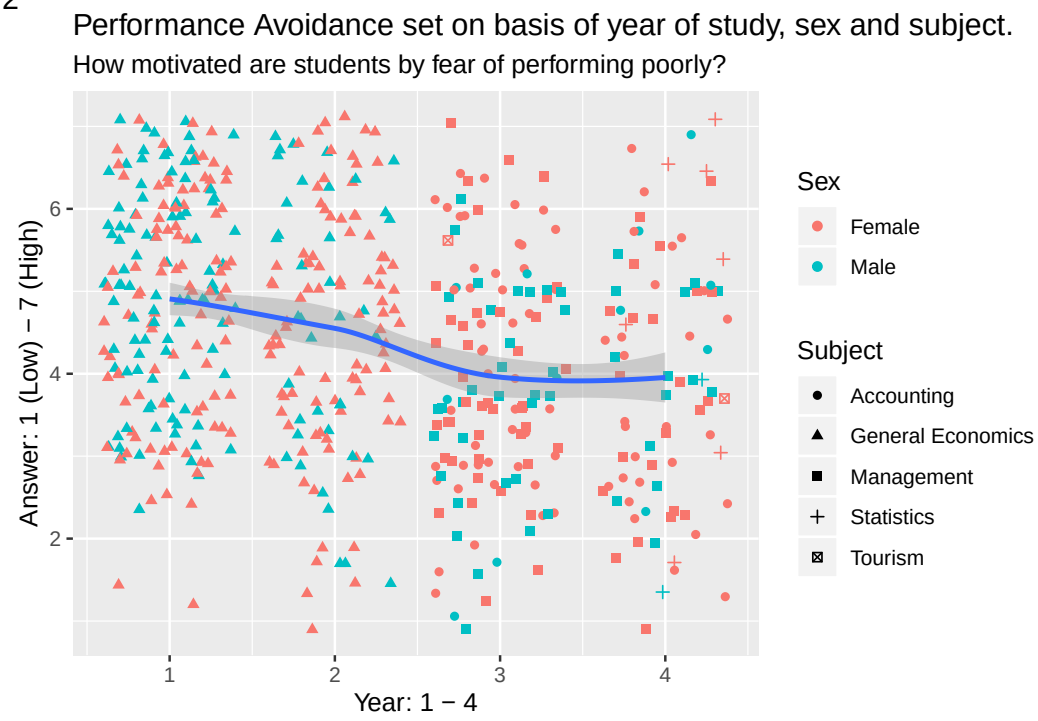
M1



It is visible that the students' performance approach almost does not change, and all the variance are within the confidence interval.

3.8.2 M2 - Performance Avoidance

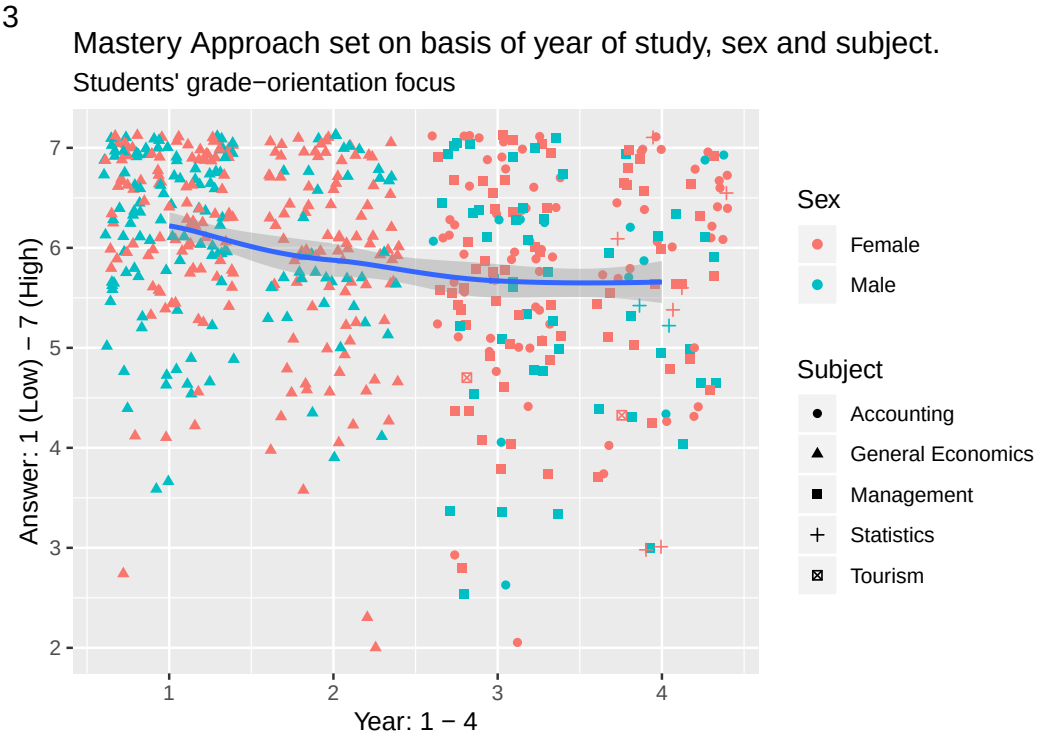
M2



Comparing year's 1 average results to year's 4, it is clearly visible that the students' performance avoidance drops by 1 scale point.

3.8.3 M3 - Mastery Approach

M3



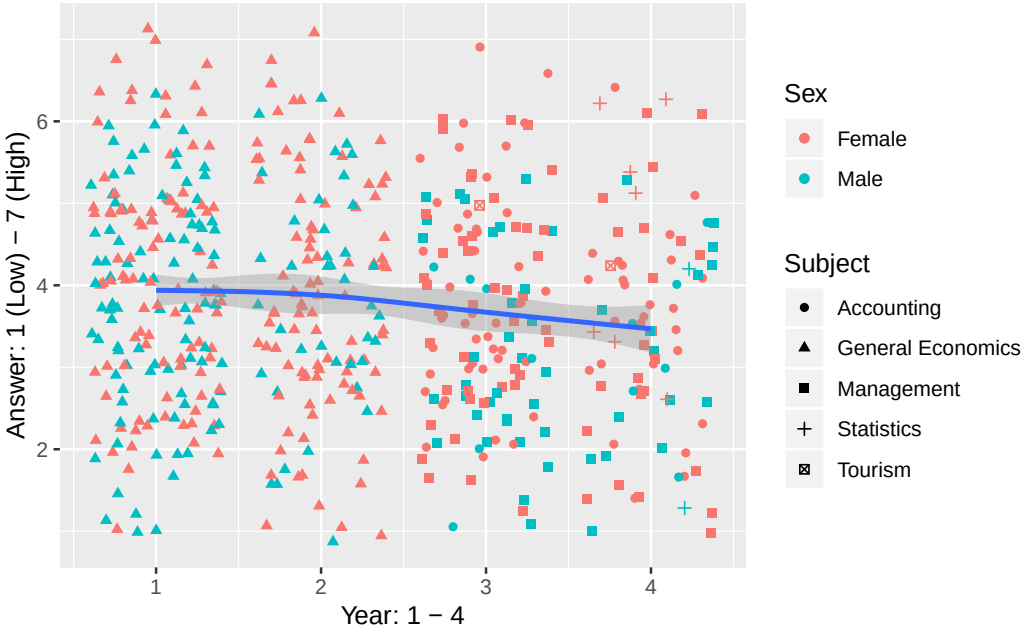
It is visible that the student's mastery approach drops by almost 1 scale point.

3.8.4 M4 - Mastery Avoidance

M4

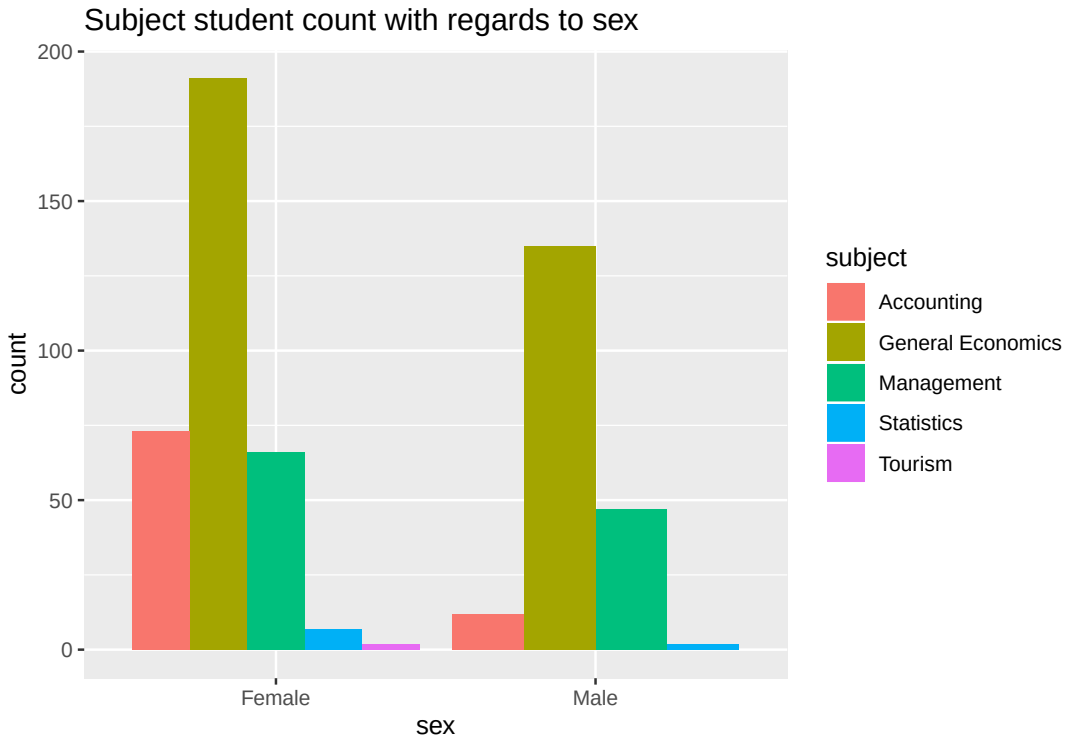
Mastery Avoidance set on basis of year of study, sex and subject.

Students' fear of not mastering the course.



There is a dropping tendency for *Mastery Avoidance* but it is not drastic.

3.8.5 Sex and Subject



4 Methodology

The data was manipulated using `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 question 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 Mean

Apart from *Interest*, *Enjoyment* and *Importance focus* categories, *Performance Approach*, *Performance Avoidance*, *Mastery Approach*, *Mastery Avoidance* categories consisted of 3 questions; for these 4 categories the means were computed and saved for each individual student.

This resulted in 4 extra columns added to the original data set. These results were using in data exploration and finding a student-at-risk. 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

4.2 Confidence Interval for a Proportion

To find students at risk with 95% confidence, we will calculate confidence interval for a proportion.

4.2.1 Equations

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

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

4.2.2 Calculations

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

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

$$Z = 1.96$$

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

5 Results

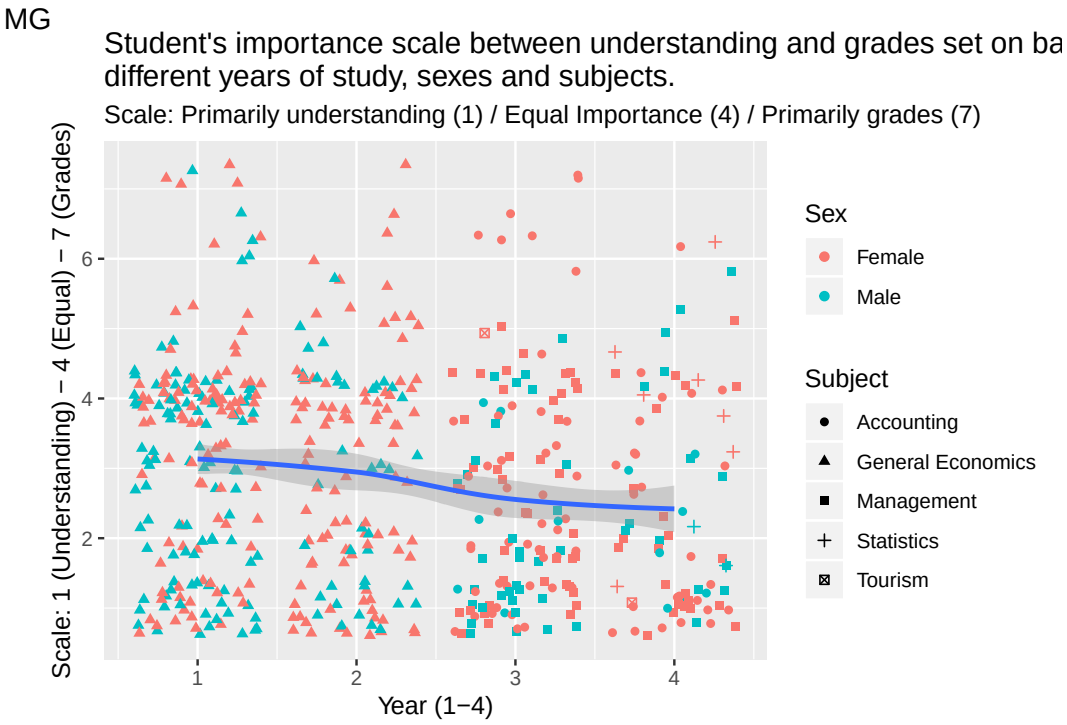
5.1 Hypothesis Testing

To test the hypotheses the following categories where picked:

- *Hypothesis 1 - Importance focus (MG)*
- *Hypothesis 2 - Enjoyment (EJ)* and *Interest (IR)*

The graphs below are a visual investigation of the hypotheses using their chosen categories. All the students are represented; student's sex was made distinguishable by a colour and student's subject was made visible by a shape.

5.2 Hypothesis 1



The hypothesis stated was:

> 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.

Visual investigation shows that students across all years, almost never primarily focus on grades. Most of the time they show equal interest in grades and understanding. Graph

smoothing does indeed show a slight shift towards deeper understanding at the more senior years but

Because the hypothesis wrongly assumes that students at their junior years primarily focus on getting good grades, and visual investigation clearly shows that the student's focus remains mostly equal, with a slight shift towards understanding at the senior year, the hypothesis is deemed wrong and no further investigation will be carried out.

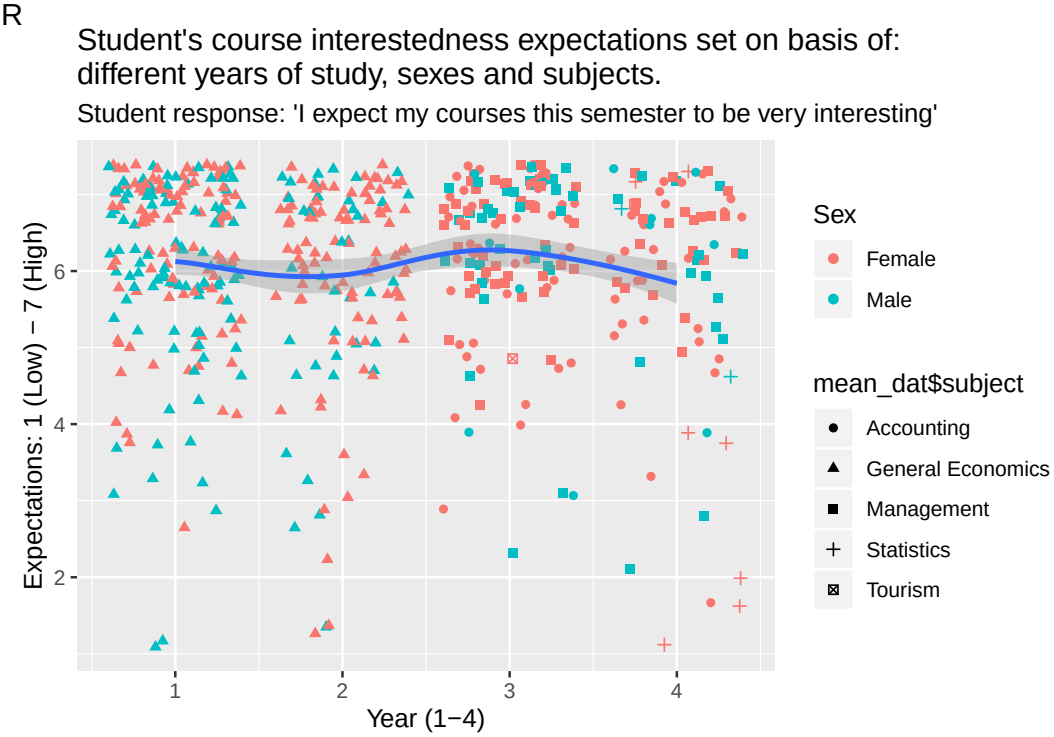
5.3 Hypothesis 2

Let us now remind ourselves the hypothesis 2:

> Students' enjoyment and interest tends to deteriorate as they progress through their studies.

5.3.1 Interest

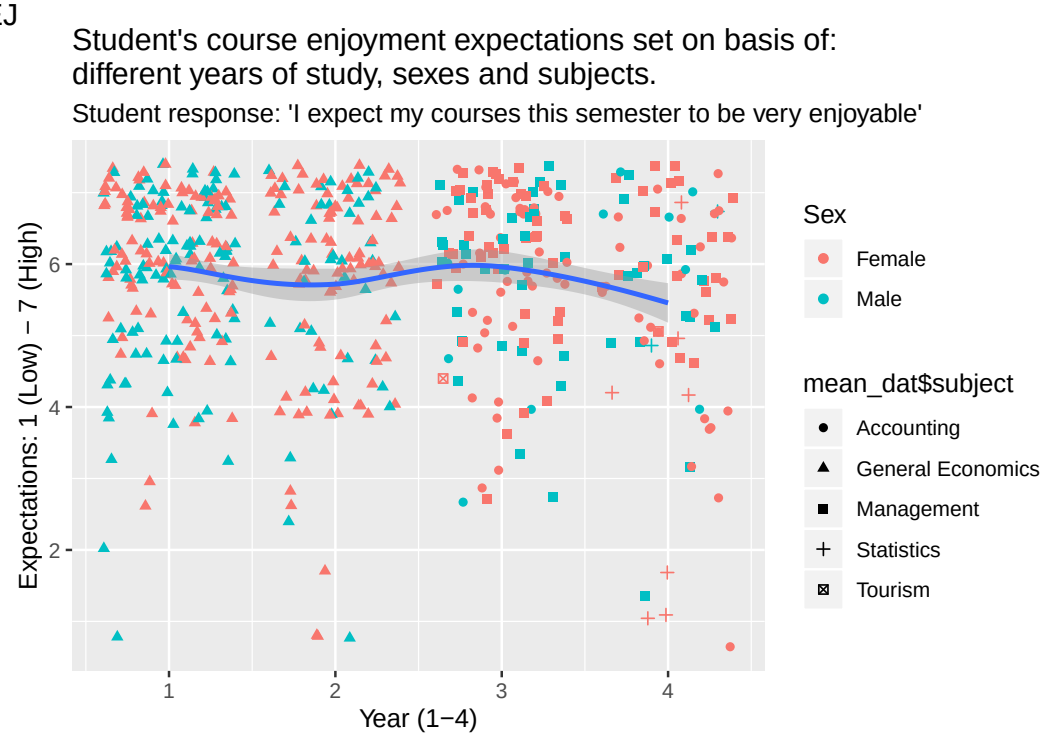
IR



Given the size of the population being quite small and graph smoothing not showing any visible patters, it is fair to say that the student's interest remains the same during their time at the university.

5.3.2 Enjoyment

EJ



Enjoyment seems to follow the same patter. Because its slight downwards trend is with the confidence interval it is deemed insignificant.

It is clearly visible that neither students' interest nor enjoyment deteriorates over the course of their studies. Visual investigation showed that it remains roughly the same.

5.4 Finding Students at Risk

To find out if there are any students at risk the arbitrary criteria was assumed:

- *Performance Approach* ≥ 6
- *Performance Avoidance* ≥ 6
- *Interest* ≤ 2
- *Enjoyment* ≤ 2
- *Importance focus* ≥ 4

Using this conservative criteria, 2 students were found (If we mitigated the scale only by 1 point, the number of students would go up to 7); they are not enjoying their course nor find it interesting, but their main focus remains on doing better than others and avoiding performing poorly with a main or neutral focus on grades. It was assumed that such students might be at risk of developing mental health problems.

Students at risk found using conservative criteria:

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

Students at risk found using mitigated criteria:

age	sex	subject	M1	M2	IR	EJ	MG
19	Male	Management	6.666667	5.000000	3	3	5
20	Female	Statistics	5.666667	6.666667	1	1	6
18	Female	General Economics	5.000000	5.666667	3	3	4
18	Male	General Economics	5.000000	6.000000	3	3	5
18	Female	General Economics	6.000000	7.000000	1	1	7
18	Female	General Economics	6.000000	5.000000	3	3	5
18	Female	General Economics	7.000000	7.000000	1	1	4

5.4.1 Confidence Interval for a Proportion of Students at Risk

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

Looking at data, we see there are in fact 2 students who might be at risk. They are both 18-year-old females studying *General Economic* at their sophomore years. Usually when we investigate the data, the extreme cases are dropped. But this time it would mean that we would be dismissing students who might be at real risk. Of course, it might be the case that they mixed up the scale or had a bad day, but this begs the question - should such cases be dismissed? Could universities address this data in any way? We might say that it is only a tiny fraction of the population and chances of finding such a student are less than 1%. So are the number of students taking their own lives at campuses.

I believe something could be done and such data should never be dismissed because we might be able to help these students. Maybe universities could develop their own programmes, that students could use to self-diagnose, whether they are at risk of developing mental health problems and if they are, it could direct them towards a university's counsellor as well as other places. Such tests already exist and can be found on the NHS website. The reason why I think they should be university specific is the fact that it makes the student feel like their university cares; it has made the first step so it must mean that it wants them to do well and is genuinely interested in their well-being. Being counselled should not be a gage of failure but a normal thing everyone in need should feel encouraged to use. I know that the universities are not being idle, but I wonder what results of a survey trying to find out if the students feel like the university empathizes with them, be like?

7 References

Chang, A. (2019). The Unspoken Reality Behind the Harvard Gates. [video] Available at: https://www.youtube.com/watch?v=WzP7oDCciGI [Accessed 9 Nov. 2019].

Mathsisfun.com. (2019). Confidence Intervals. [online] Available at: https://www.mathsisfun.com/data/confidence-interval.html [Accessed 9 Nov. 2019].

Medium. (2019). Cross Validation Explained: Evaluating estimator performance.. [online] Available at: shorturl.at/kuMV7 [Accessed 9 Nov. 2019].

Soltoff, B. (2019). Cross-validation methods. [Blog] Computing for the Social Sciences. Available at: https://cfss.uchicago.edu/notes/cross-validation/ [Accessed 9 Nov. 2019].

Statistics How To. (2019). Confidence Interval: How to Find a Confidence Interval: The Easy Way! - Statistics How To. [online] Available at: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/#CIProp [Accessed 9 Nov. 2019].

Statistics How To. (2019). Z-Score: Definition, Formula and Calculation - Statistics How To. [online] Available at: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/ [Accessed 9 Nov. 2019].

Statistics How To. (2019). Z-table (Right of Curve or Left) - Statistics How To. [online] Available at: https://www.statisticshowto.datasciencecentral.com/tables/z-table/ [Accessed 9 Nov. 2019].

Stulp, G. (2019). Different ways of calculating rowmeans on selected variables in a tidyverse framework - Gert Stulp. [online] Gertstulp.com. Available at: https://www.gertstulp.com/post/different-ways-of-calculating-rowmeans-on-selected-variables-in-a-tidyverse-framework/ [Accessed 9 Nov. 2019].

Wickham, H. and Grolemund, G. (2016). R for Data Science. Sebastopol, CA: O'Reilly Media.