Exploratory data analysis and statistical inference for students` results on Discrete Mathematics and Probability and Statistics at Faculty of computer science and engineering

Lenche Jovova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
lenche.jovova@students.finki.ukim.mk

***Abstract*— During the past decade Faculty of Computer Science and Engineering (FCSE) has become the most attractive faculty in North Macedonia. The interest for this faculty arises every year and this motivated me to investigate the student`s results in Probability and Statistics and Discrete Mathematics and to check for correlation with the results from the Final state exam. Hypothesis testing and exploratory data analysis were made, and the results are presented.**

Keywords*— *hypothesis testing, exploratory data analysis, statistical inference, FCSE

I. Introduction

Exploratory data analysis (EDA) is a powerful framework when it comes to having a big picture of which properties describe the process or the population of interest. It can never be the whole story, but nothing can serve as the foundation stone, as the first step [1]. When data analysis is done, EDA is often the first step taken after data is cleaned. On the other hand, to know how reasonable our model or our beliefs for the data are, hypothesis testing come into play. Our beliefs for the process or the population can be justified or not depending on the sample that describe the population. To draw conclusions for our process or population of interest, Statistical inference can be used.

In this paper, EDA and Statistical inference are used for drawing conclusions for the student`s results for Probability and Statistics and Discrete Mathematics. In Section II the used datasets are described, Section III includes the analysis of the results from Probability and Statistics and Section IV analyzes the correlation between the results from Discrete Mathematics and Probability and Statistics.

II. Dataset description

For the course Discrete Mathematics data is available from the last three academic years (2016/2017, 2017/2018, 2018/2019) and for Probability and Statistics data is available from the last four academic years (2015/2016, 2016/2017, 2017/2018, 2018/2019).

The datasets for Discrete Mathematics and Probability and Statistics include the following attributes:

- **Id number –** Identification number of the student, ordinal type
- **Name and Surname –** First name and Last name of the student, string type
- **Program-**the program the student follows, categorical variable
- **Professor –** the professor that teaches the course, string type
- **Total Points–** Total points the student has for the subject, continues data type
- **Grade –** Final grade of the student

Additionally, for Probability and Statistics despite the above attributes, the following are also used:

- **ExercisePart1 –** points gained on the first partial exam on exercises
- **ExercisePart2 –** points gained on the second partial exam on exercises
- **TheoryPart1 –** points gained on the first partial exam on theory
- **TheoryPart2 -** points gained on the second partial exam on theory

III. Analysis of the results from Probability and Statistics

The focus of the analysis for this subject is to compare the results for the students that follow the course regularly according to their year of enrollment in the faculty and the students that have previously followed the course but have not passed it. Also, the percentage that pass the subject from the regular students will be analyzed and the rank of the students according the results for Probability and Statistic will be compared with the rank of the students according their results from the Final state exam.

A. Regular vs. Non regular students

According to the data from the first and second partial exam for exercises, the regular and the students that once followed the subject seem to have different distributions.

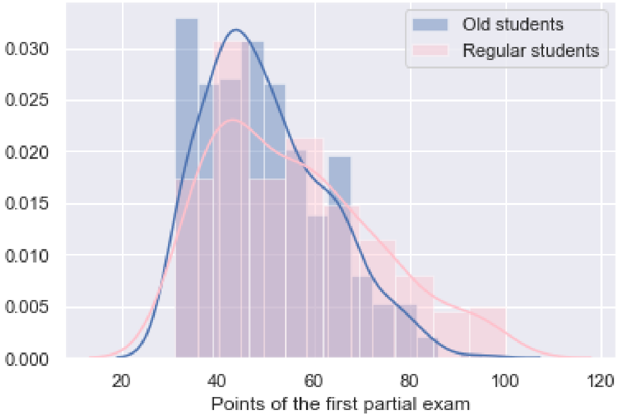

Fig. 1. Probabilty density function for the points of the first partial exam

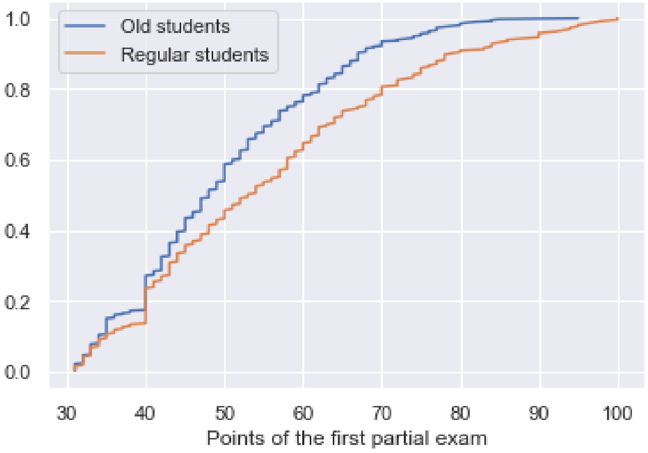

Fig. 2. Cumulative density function for the points of the first partial exam

From the graphs themselves is obvious that the regular students have slightly better performance on the exam. The distributions seem to have concession from normality, and we will perform Kolmogorov-Smirnov test, a non – parametric statistical test which tests the null hypothesis that the two groups come from the same population based on their cumulative density functions [2]. The test computes the distance between the cumulative density functions and based on that calculates Kolmogorov-Smirnov Statistic.

H_0: The two groups come from the same population

H_a: The two groups come from different populations

Tested groups: Old students that followed the course previously and failed and students that follow the subject regularly.

Statistical Test: Kolmogorov-Smirnov

Level of Significance: 0.05

When the test was performed on the two groups, a p-value of 0.0002 was obtained, which is below the level of significance of 0.05, hence I reject the null hypothesis in favor of the alternative.

The data from the second partial exam seem to slightly have some normality, as shown in the figure bellow:

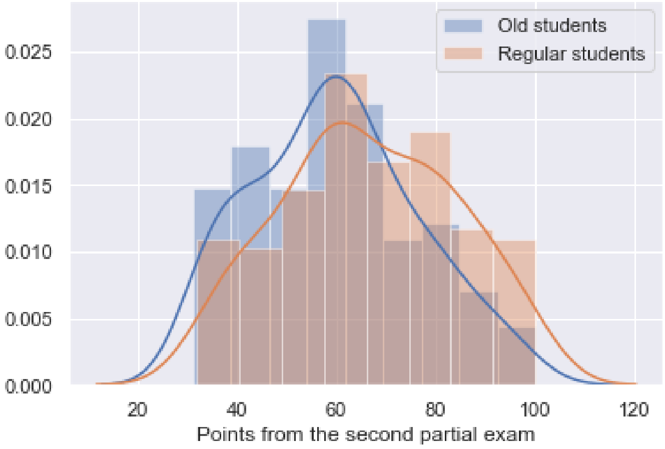

Fig. 3. Probabilty density function for the points of the second partial exam

For this data t-test for the mean of the two groups was performed. The two-sample t-test is parametric test for testing the means of two groups. The result of the test is a test statistic based on the difference of means and the pooled variance of the two groups [3].

H_0: The two groups have the same mean

H_a: The two groups have different means

Tested groups: Old students that followed the course previously and failed and students that follow the subject regularly.

Statistical Test: t-test for means

Level of Significance: 0.05

When the test was performed on the two groups, a p-value of 0.0001 was obtained, which is below the level of significance. The null hypothesis is rejected in favor of the alternative. Additionally, the test statistic has a value of – 3.82 which suggest that the regularly enrolled students for the course have higher mean than the students that previously failed the subject.

B. Analyze of the frequency of students that passed the subject throughout the years

For the available data I calculated contingency table in order to analyze the number of students that passed the exam successfully (either on partial exams or in any exam session) versus the number of students that did not passed any of the exams in the respective academic year. The analysis is done only for the students that regularly follow the course, i.e. students that follow the course for the first time. The table I bellow shows the number of the students that passed and did not passed the exams for Probability and Statistics.

TABLE I. CONTINGENCY TABLE FOR THE NUMBER OF STUDENTS THAT PASSED VERSUS THE NUMBER OF STUDENTS THAT FAILED

Academic Year	2015/16	2016/17	2017/18	2018/19	All
Passed					
No	114	122	156	175	567
Yes	69	54	72	82	277
All	183	176	228	257	844

Chi2 test for the contingency table will be performed. The null hypothesis that will be tested is the ratio of the passed versus failed students is not changing throughout the year [4].

H_0: The ratio of the passed/failed students is not changing throughout the years

H_a: The ratio of the passed/failed students is changing throughout the years

Tested groups: Students that regularly follow the course through four academic years

Statistical Test: Chi2 test for contingency table

Level of Significance: 0.05

When the test was performed, a p-value of 0.95 was obtained which suggests that we should accept the null hypothesis. Given the data from the last 4 years, the percentage of the students that have passed the exam has not changed.

C. Comparing the rank according the total points on Probability and Statistics and the rank from the total points on the Final state exam

The Final state exam is a national exam consisting of 4 subjects that the students must passed in order to make an application for Faculty enrollment. The students that want to enroll for FCSE must have chosen Mathematics in the Final state exam. The last three numbers (or the last two numbers for the program KNIA on FCSE) from the Id number that the student get depend on the Rank the student have according the points from the Final state exam, i.e. the student that had the highest points on the Final state exam gets Id number xxx001. We will compare if the first students of the Final state exam are the first students on the exams of Probability and Statistics as well. For this purpose, we will use Kendall Tau statistic for ordinal data [5]. The Kendall Tau statistic is based on the difference between the number of concordant and the number of discordant pairs in the two vectors. The Tau test is a non-parametric test for statistical dependence based on the Kendal Tau statistic.

The tested hypothesis and the groups, as well we the test and level of significance are the following:

H_0: The two groups are positively dependent

H_a: The two groups are independent

Tested groups: Rank if the students according to the Final state exam and results from the exams from Probability and Statistics from the last 4 years

Statistical Test: Tau

Level of Significance: 0.05

The test resulted with very low p values and more important, tau correlations very close to zero, for every year. These results not only lead to rejecting the null hypothesis, but also lead to the conclusion the results from the Final state exam maybe are not relevant.

IV. CORRELATION BETWEEN DISCRETE MATHEMATICS AND PROBABILITY AND STATISTICS

I wanted to check if there is any correlation between the results on Discrete Mathematics and the results on Probability and Statistics for the same students. On the scatter plot on Figure 4 are the shown the results from both subjects. There can be noticed a lot of outliers, students that have gained very high points in Discrete Mathematics, have very low points in Probability and Statistics.

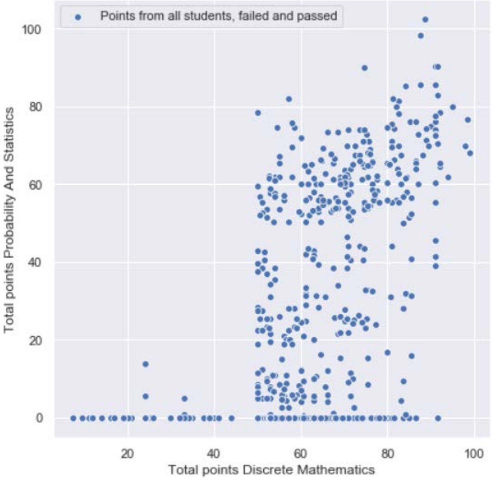

Fig. 4. Discrete Mathematics versus Probabililty and Statistics for all students

A moderate linear relationship is present and can be noticed especially in that part of the scatter plot where are the students that have gained more than 30 points in Probability and Statistics. On the scatter plot in Figure 5 is shown only that part of the plot.

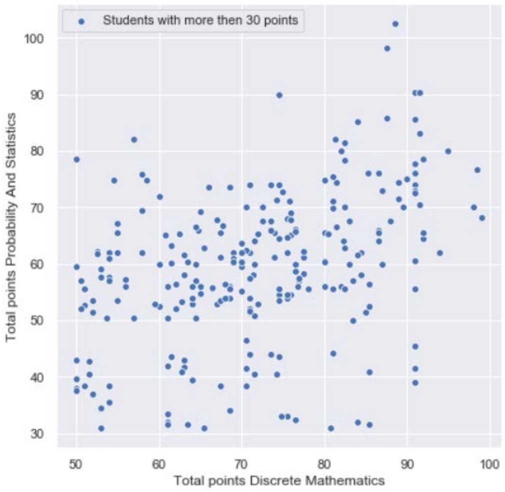

Fig. 5. Discrete Mathematics versus Probabililty and Statistics for students that have more than 30 points in Discrete Mathematics

A linear model will be made for the two variables Total Points in Discrete Mathematics and Total Points in Probability and Statistics for the students that obtained more than 30 points Probability and Statistics. The Pearson correlation coefficient for the two variables is 0.51 which confirms that there can be a moderate linear relationship. A linear regression model with ordinary least square method was performed [5]. The results from the model showed that the Total Points in Discrete Mathematics is significant variable and for every 1 point in Discrete Mathematics the total points for Probability and Statistics will increase for 0.43 The regression line is given in the figure 6.

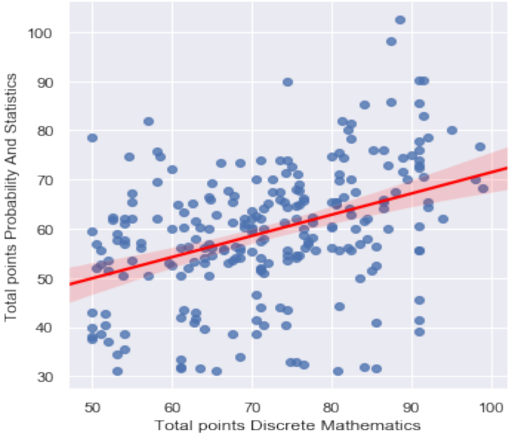

Fig. 6. Linear regression line for the model

The model leads us to the conclusion that based on the results in Discrete Mathematics, we could know what to expect in Probability and Statistics for the students that will pass the exam.

V. Conclusion

From the obtained results can be concluded that the students who follow Probability and Statistics for the first time have slightly better results than the students which have once followed the same subject. In the second group there are also high scores and further investigation need to be done to check whether the previous time they followed Probability and Statistics, they had focus on another subject, e.g. Algorithms and Data structures, which is in the same semester as Probability and Statistics and is also extensive course.

The ratio of the passed/failed students that is not changing through the years according to the results, suggests that there is a constant percentage of students that would pass Probability and statistics on their first following of the course.

The lack of correlation with the results from the Final state exam can be a sign that this condition for enrolling on the faculty is not the best choice, but to validate this, in future more courses will be included in the investigation.

VI. Future work

The goal was to check differences between regular and students that once followed the course, investigate relationship between the mentioned subjects and check for correlation with the Final state exam. My future work will include more detailed and extensive analysis as well as more subjects and more academic years. The obtained results can be then used to draw more precise conclusions that could help improving the faculty`s strategy.

Acknowledgment

Thanks to Marija Mihova, PhD for the provided anonymized datasets.

References

[1] John W. Tukey, “ Exploratory Data Analysis. Addison-Wesley”, 1977,pp.1-3

[2] Jean Dickinson Gibbons,Subhabrata Chakraborti, “Non parametric statistical inference”, Fourth Edition , Revised and Expandedn , pp.239-244

[3] Konstantin M.Zuev , “Statistical inference”, unpublished

[4] https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

[5] Peter D.Hoff, “A first course in Bayesian statistical methods”, Fourth Edition , Revised and Expandedn , pp.149-154