Computer-aided Simulations Lab - Lab G5 report

Stefania Budano - 314737

December 5, 2023

1 Introduction

In this lab we focused on making an improved version of the student career simulation previously developed in lab G4. We extended the model to include a more complex and realistic scenario.

We still maintain the basic structure as before, simulating the career of a single student and repeating the simulation until we reach a desired level of accuracy in the computation of the confidence interval for the averages of the graduation grade and the graduation time.

2 Main changes and improvements

All the main components of the previous simulator have remained the same. The main changes are the introduction of the possibility to reject a proposed exam grade based on a certain threshold, and the introduction of a preparedness parameter for each exam, which potentially increases the probability of passing it. This rejection threshold is also decreased every few sessions, in order to avoid the situation in which the student keeps rejecting the grades and could potentially never graduate.

Furthermore we have strived to make the simulation more realistic and detailed by not considering all exams as equal anymore, but each exam is characterized by its number of cfu, which corresponds to a different grade distribution and a different probability of passing.

We have also changed a bit the extraction of which exams to be tried in each session. Specifically, in the previous version at each session we extracted a number of exams to be tried from the total number of exams left in the curriculum. In this more realistic version, we consider the fact that a student can only take an exam if they have gained attendance in that course. This is translated in the code in the fact that we have split the list of exams to be taken in order to graduate into three semesters, and at the start of the simulation we append then to the list of exams we could try based on how many sessions have passed. For this purpose we have considered a model with 5 sessions per year (similarly to Politecnico, in which we have two winter sessions, two summer sessions, and one fall session), each with only one call for each exam, meaning that the first two sessions correspond to the end of the first semester, the 3rd (and 4th) correspond to the end of the second semester, and the 6th to the end of the third semester. For all sessions after the 6th we consider to have gained attendance of all subjects, and thus we can try any exams from those not yet passed.

3 Stochastic elements

All the stochastic elements present in the previous version of the simulator are still present.

The first random element is the number of exams tried in each session. This is still extracted from a discrete Poisson distribution with a certain average.

Second is the probability of passing or not an exam, which is still extracted from a uniform distribution and compared to a threshold.

A new stochastic element is given by the rejection threshold and the preparedness level for each exam. The rejection threshold is extracted from a custom distribution, and ranges from 18 (meaning that the student does not reject any passing grade), all the way up to 30. To each rejection threshold is corresponding a certain base increase factor, which represents a potential increase in the probability of passing exams, that is higher the higher the threshold is. Since no actual statistics on this could be found, the distribution from which we extract the rejection threshold has been custom made based on some reasonable assumptions: firstly there is a considerable percentage of students that do not reject any passing grade, and thus will have a threshold of 18, and secondly the fact that the most common thresholds are in the range 24-27, and higher ones are more rare. The base increase parameter has also been assigned to each threshold following a reasonable assumption, that is that students that have a higher threshold for accepting or rejecting a grade, will also tend to study more and be more prepared, and thus have a higher chance of passing exams.

The rejection threshold and the base increase parameter are extracted only once for each student. The preparedness level parameter, instead, is extracted every time we try an exam, and it is a parameter ranging from 1 to 5, to which again corresponds an increase parameter that potentially increases the probability of passing an exam. The value of the increase parameter has also been determined following an assumption that a student that is more prepared will have a higher probability of passing the exam.

For what concerns the grade extracted from the custom grade distribution, this has remained the same, extracted from a custom distribution based on real data, with the only difference that we use a different distribution for each different number of cfu of the exam. Specifically we have taken as example the grade distributions of three courses from the Data Science and Engineering course (specifically Machine Learning and Deep Learning for the 10 cfu distribution, Distributed Architectures for Big Data Processing and Analytics for the 8 cfu distribution, and Data Ethics and Data Protection for the 6 cfu distribution). The threshold with wich we compare the passing probability to determine if an exam is passed or not is also taken from these courses as the number of students who passed the exam divided by the number of students enrolled in the course. These data has been taken from the Politecnico website on the page for each course (We have considered to compute the distributions the data relative to all students, not only those enrolled in the course for the first time, in lighter blue in the graphs).

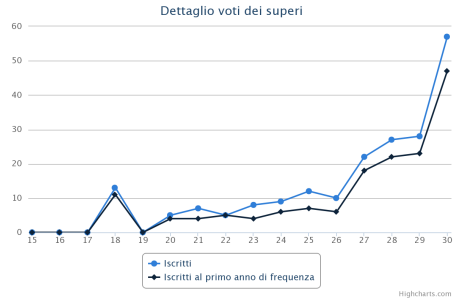

Figure 1: Grade distribution for the Machine Learning and Deep Learning course

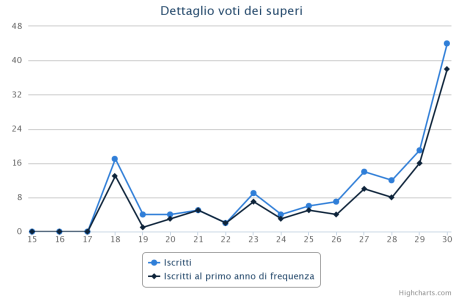

Figure 2: Grade distribution for the Distributed Architectures for Big Data Processing and Analytics course

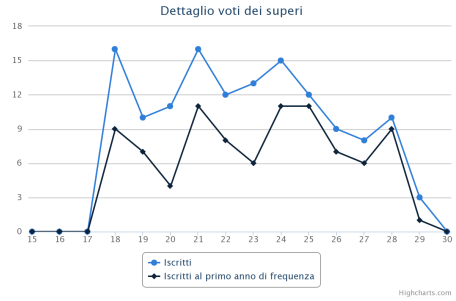

Figure 3: Grade distribution for the Data Ethics and Data Protection course

Finally the random elements introduced in the computation of the final graduation grade, given by the points for the thesis, presentation, and bonus points, are still extracted from uniform distributions, with the difference that we have considered different ranges depending on other factors, in order to simulate a more complex and realistic model. Specifically we have still maintained the points for the presentation as independent from other factors, while we have considered different ranges for the thesis points based on the average of the grades, based on the assumption that a student who has a higher average grade will also tend to produce a better thesis. Specifically if the average is from 18 to 22, we can only assign between 0 and 2 points out of the 4 total, if the average is between 22 and 26 we assign between 2 and 3 points, and if the average is higher than 26 we assign between 3 and 4 points.

Similarly we have assigned the bonus points based on how long the student took to graduate, and assign between 1 and 2 points if the student graduated "on time" in 2 years, between 0 and 1 point if they take up to 4 years, and no points if the take longer than that.

4 Input parameters

The input parameters of the simulation are:

- The total number of courses, equal to 13 (again taken from the Data Science and Engineering course study plan, considering two free ECTS courses in the second year)
- The average number of exams tried in each session, parameter of the Poisson distribution, equal to 3
- The number of sessions per year, equal to 5
- The accuracy and confidence level for the computation of the confidence intervals on the averages, both equal to 98%
- The maximum number of students, equal to the maximum number of times we potentially run the simulation, only introduce to avoid a loop in case we do not reach the required accuracy, equal to 10000
- The distributions for the grades and the probabilities of passing an exam, different for the different values of cfu an exam can have, as detailed in the previous section
- The lists of exams in each semester, where each exam is only characterized by its number of cfu
- The random seed ($np.random.seed(4)$)

5 Main assumptions

Unlike the previous version of the simulator, we are no longer considering all the exams as equal, with the same grade distribution and the same threshold for passing or not. But we are still assuming that all the exams with the same number of cfu have the same distribution, which is still a simplification. We are also assuming that each exam is somewhat independent from each other, when in reality some of them could be correlated, for example in terms of similar topics covered.

Also we are assuming that the preparedness for each exam in the same session is independent from the number of exams that we are trying, where in reality it could be that if we are trying more exams in the same session, the level of preparedness for each will be lower.

A further assumption that we are making is that the rejection threshold, with its base increase parameter, and the preparedness for each exam, only influence the probability of passing the exam, but not the actual grade, where in reality the grade could also be correlated to these other factors.

6 Output metrics

The output metrics we are computing in order to analyze the results, are the weighted average of the grades of the exams, the graduation time in years, the actual graduation grade, and also the average graduation grade for each length of graduation time, the average graduation grade for each value of the rejection threshold, and the average graduation time for each value

of the rejection threshold. We plot these computed metrics and observe the following results:

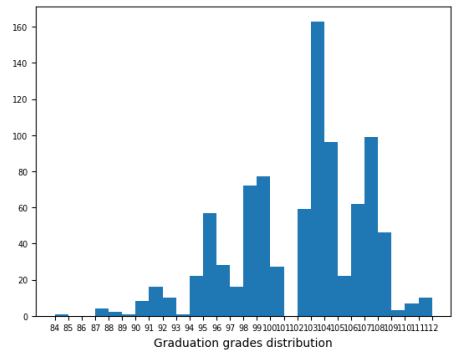

Figure 4: Distribution of the graduation grades

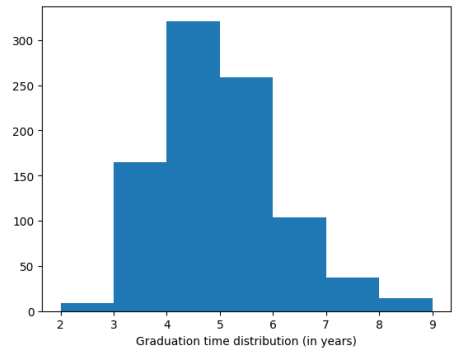

Figure 5: Distribution of the graduation times, in years

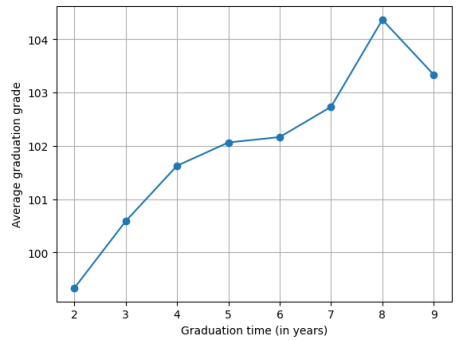

Figure 6: Plot of the average graduation grade for each value of the graduation time in years

7 Main data structures

Similarly to the previous iteration of this simulator, no particular data structures were needed to run the simulation and compute all the metrics, although the introduction of custom classes for both the student and the exams could be considered for possible extensions of the simulator, as will be described in a following section.

8 Observations

We can observe how the distributions of the graduation grades and times tend to follow a distribution similar to that of a normal or a beta. This would probably be even more evident if we made the simulation run for more times, as right now we reach the desired accuracy with a relatively small number of iterations (908).

More interesting could be to observe the impact of the rejection threshold on the average grade and the average graduation time. We can see how a higher threshold does indeed result in a higher average graduation grade, but it also increases the graduation time, as a student with a higher threshold will tend to reject more grades in order to achieve better results. In this sense it could be seen as a trade off between the grade and the time.

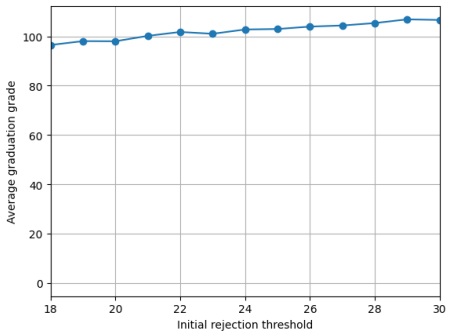

Figure 7: Plot of the average graduation grade for each value of the starting rejection threshold

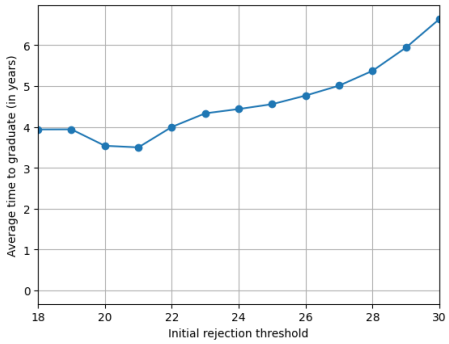

Figure 8: Plot of the average graduation time for each value of the starting rejection threshold

9 Potential improvements

This simulator could be further improved by introducing a correlation between the preparedness level and the grade of each exam, in addition to the correlation with the probability of passing. We could also make it even more realistic by introducing a different distribution and passing threshold for each separate exam, or by characterizing the exams also by their type (written, oral, project, theoretical/practical subject), and introducing a more complicated set of skills for the student, such as being better at written tests, or more practical subjects. If we were to introduce these improvements, it would be useful to also introduce separate classes for the Student and Exam items to keep track of all these parameters.