Lab4 - Simulating students careers

Martina Martini s306163
Politecnico di Torino

Assignment Develop a simulator to evaluate the graduation time and the final grade at a MSc course.

- Explain in details the random elements in the simulated system
- Explain all the input parameters
- Explain the main assumptions of the adopted simulation model
- Explain all all the output metrics
- Explain all the main data structures
- Explain some interesting correlations between output metrics and input parameters

For the simulation, use realistic values for all the parameters.

1 Random elements, input parameters and main assumptions

To simulate a student's academic career, some stochastic elements must be considered. Firstly, the choice of the courses for which taking the exam during a specific session is random, considering that the maximum number of exams to try in a single session is a fixed parameter. Secondly, the probability of getting a certain grade is given by a grade distribution, already provided as input. Moreover, I assumed the probability of passing or failing an exam follows the Bernoulli distribution, where each attempt is treated as an independent Bernoulli experiment with the same probability of success p and of failure $p - 1$. Similarly, given iid trials, the probability of accepting or rejecting a grade follows the Bernoulli distribution. Finally, some bonus points are added to the final grade based on the quality of the thesis (from 0 to 4), the goodness of the presentation (from 0 to 2) and other metrics like the speed of graduation or the number of lauds (from 0 to 2): all of these three variables are uniformly distributed. To conclude, a random seed is introduced for reproducibility purposes.

For what concerns the input parameters, I assume to set the following fixed variables: the total number of considered students is 100, the total number of courses for the MSc is 15, the number of sessions per year is 6 and the maximum number of exams that a student can take in one session is 4. Moreover, as disclosed in advance, the probability p of passing an exam is fixed at 0.7, while the one of accepting depends linearly on the grade that the student got. Finally, the initial grade distribution is already provided: 87/1849 for the grade 18, 62/1849 for 19, 74/1849 for 20, 55/1849 for 21, 99/1849 for 22, 94/1849 for 23, 117/1849 for 24 and 25, 136/1849 for 26,

160/1849 for 27, 215/1849 for 28, 160/1849 for 29 and 473/1849 for 30. In conclusion, a confidence level is set to 0.90 for the computation of the confidence intervals, but it can be tuned to find the best accuracies.

Regarding the assumption I made, some hypotheses are considered in the simulation. First of all, the time is expressed in terms of years: each year is interpreted as a complete cycle of 6 sessions. Also, each exam can be taken just once in a single session but the times that a student can try to pass an exam (in different sessions) are unbounded. Lastly, the time to write the thesis is not considered for the purposes of the computation of the graduation time.

2 Output metrics, data structures and further analyses

The output metrics (Fig 1) used to evaluate the results of the simulation consist of - in order of appearance -:

- The graduation grade for each student,
- The grade distribution of all the samples of students (the grades are rounded to the nearest integer for easier reading),
- The average grade for each student before adding the bonus points,
- The number of sessions taken by each student to graduate,
- The session distribution considering the entire sample of students,
- The maximum number of times that each student re-took to pass an exam,
- The most difficult exam to pass for each student (computed as the exam that the student must have re-taken the most before passing it).

For the plots of graduation grades, the average grades and the number of sessions, the confidence intervals are computed and displayed as well as the mean on the entire sample. Moreover, for each student, the average times that the student must have retaken the exam is computed as well as the number of times to pass the last exam before graduating. Finally, it is also displayed how many students found a certain course the most difficult counting how many of them re-took the exam more than one time.

The main data structures are very naive and include several lists, used for the final analyses, a few dictionaries and the class Student, whose attributes are the identification number, the number of remaining exams (they are initially set to the total number of courses), the grades he gets, the number of sessions and years he takes to graduate, the final grade and the list of the exams he does not pass each session.

Further analyses are added at the end of the simulation, such as some statistics about the graduation grades, the average grades, the average times to re-take an exam and the time taken to pass the last exam. As you can notice, different simulation running can bring pretty different results, even if the time to graduate is quite often around a year and a half and the mean graduation grade is usually around 100. Also, I tried to filter the data-frame overcited to evaluate the most talented students, who got the higher grades by re-taking on average the exams at most 1 time, trying at most 2 times the last exams and graduating in just 1 year.

In conclusion, another interesting measure is given by the correlation between the average graduation grade and the number of years taken to graduate: it can be seen that this mean remains stable when considering one or two years to graduate, but it increases when the years become more than two.

For completeness, the outputs are shown on the following page but for clearer evaluations, plots and results, please refer to the code provided.

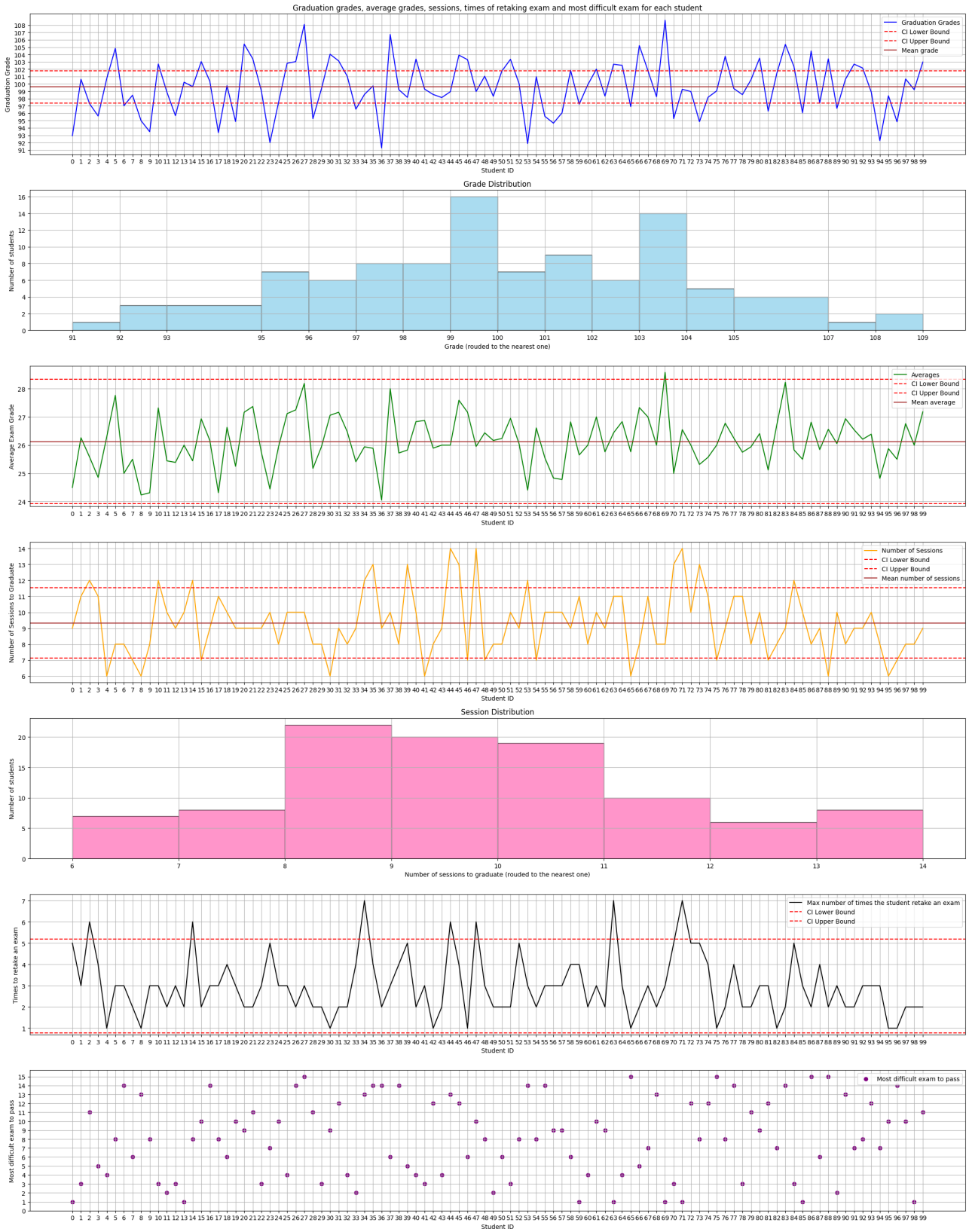

Figure 1: Main output metrics