Computational Statistics
Assignement 2 - The Bootstrap

Arthur Vieillevoye i6220340
Arthur Goffinet i6215214
Aaron Schapira i6211359

March 2023

Introduction

During the exercises, you will compute the non parametric bootstrap distribution of the Pearson's correlation coefficient between LSAT and GPA using B = 40000 bootstrap samples. In this assignment, work on these additional tasks.

Aditionally, here is the link to the project:

Note: The code for this assignment was realized using python on a Jupyter Notebook: Github

1 Question 1

Recompute using the complete enumeration bootstrap.

As we had to compute the enumeration bootstrap, we had to compute all the possible combination with repetition of length 15 of index from 1 to 15. Then, we would use those indexes to generate our samples. After some research we realised that it was possible to do it using the python library *itertools.product*. However, we felt like using a library was not the goal of the assignment. Additionally, if we were to implement the sequences with repetition and where order would matter, the algorithm would run during a very long time (between 4 and 6 hours).

We thus decided to implement a method to do it in a recursive way, which returned all combinations of number without repetition from size 1 to size 15 (where order does not matter). The size of the final array is $32,766$

Bootstrap Statistics	Correlation	Bias	Standard Deviation
	0.776374	-0.010570	0.177486
size of the Confidence interval	0.353952	0.987496	/

Bootstrap Statistics	Correlation	Bias	Standard Deviation
	0.776374	-0.010570	0.177483
size of the Confidence interval	0.353955	0.987496	/

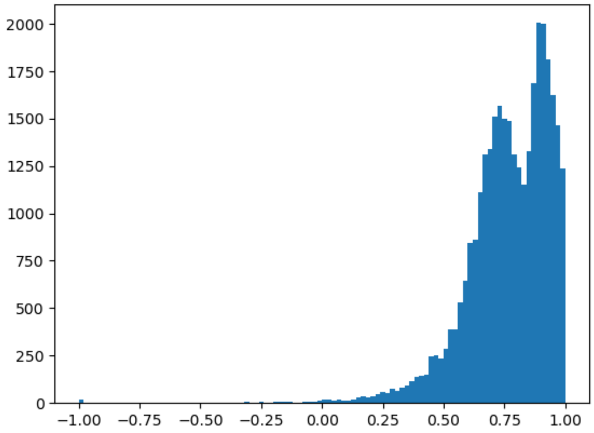

Figure 1: Histogram of the correlation of the complete enumeration bootstrap

We will investigate the time of this computation in section 3

2 Question 2

Use Gray codes for compositions to speedup computations.

We will first define what gray code is in order to work with it.

The gray codes technique should be faster because the algorithm creates the new sample by changing only one coordinates from the previously created sample. Thus, we have a list of binary values that changes from their neighbour by only one binary value.

The gray code gives the following results:

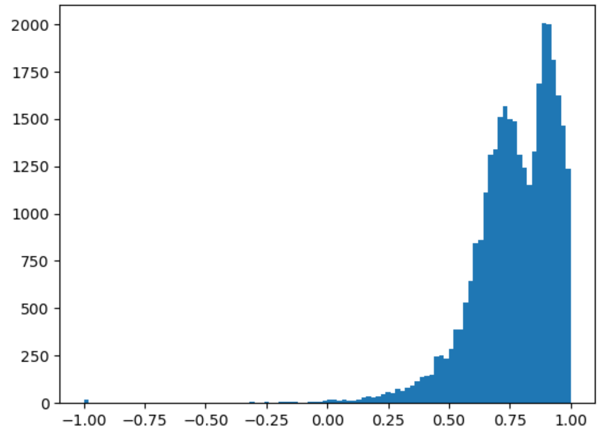

Figure 2: Histogram of the correlation of the Grey Code

We will investigate the time of this computation in section 3

3 Question 3

How much speedup can you get by using Gray codes? Show either experimentally or theoretically In this exercise, we used a similar set of data for the two techniques, then, we recorded the time taken by the two algorithms to compare them.

Running Time (s)	Question 1	Question 2
	106.1059 s	0.0452 s

As expected, the gray code is way faster. This can be explained from the definition of gray code and the generation of sample using it.

4 Question 4

Which observation(s) do you need to remove from the sample to make the Monte Carlo and complete enumeration bootstrap look more similar?

In order to get a similar distribution for the Monte Carlo as the complete enumeration one, we would need to give the Monte Carlo simulation a sample size of 15 with replacement when creating it. The following graph shows the result of such a sample. We can see that it looks then very similar to our original one, where we can see the same peaks around 0.8, with density starting to increase around a correlation of 0.4.

The difference between the Monte Carlo and the enumeration is explained by the fact that in the enumeration we use all the possible combinations where some of them have a low correlation. Due to this, the graph maps more of the exceptions (sub samples that have bad correlation). Where as for the Monte Carlo, we are creating a sample of 15 with repetitions. Due to this, the likelihood of having a bad sample is strongly reduced. This explains why Monte Carlo represents more the trend.

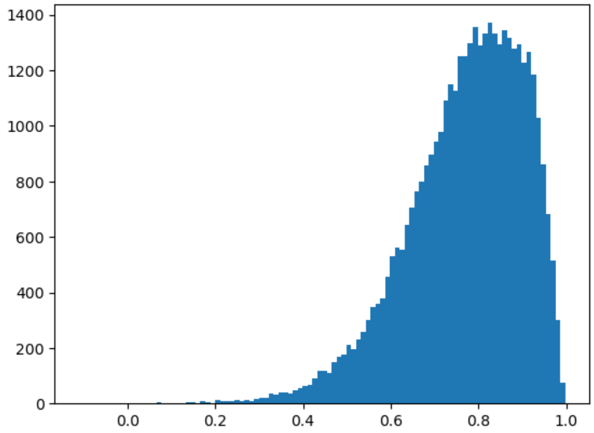

Figure 3: Histogram of the correlation of the Monte Carlo

On the other hand, if we plot the graph of the Monte Carlo simulation with sample size selected randomly from 2 to 15 with no replacement; we can see some changes. Those will be discussed in the next section.

5 Question 5

Explain why you obtain different results for Monte Carlo and complete enumeration bootstrap

Because with Monte Carlo, we use randomness, than as its name state, the complete enumeration bootstrap creates all the possible permutations of the data. Thus, the difference comes from the randomness of the sample sizes.

In the following graph, we can see a clear peak around 0.6, which can be explained by the randomness of the sample size. Indeed, there will always be some correlation coming from the original data, which results in the peak. Furthermore, for the same reasons, we can explain the small density for higher correlations, as we try to sample with random sizes it is expected that we often have bad results.

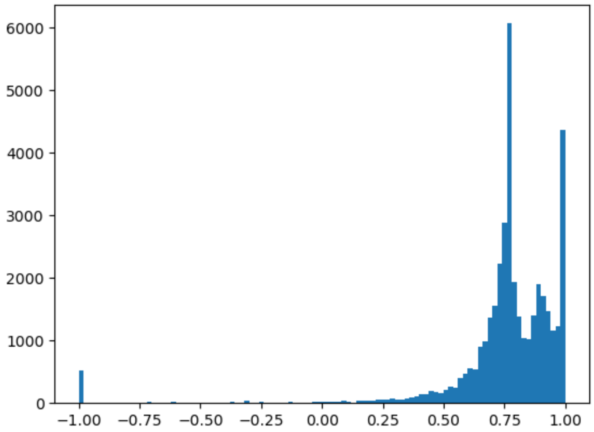

Figure 4: Histogram of the correlation of the Monte Carlo