Assignment 3 Research Track 2

Aurora Bottino s4814504

June 2nd,2023

1 Introduction

In the third assignment of Research Track 2, our task was to conduct a statistical analysis on the first assignment of Research Track 1, focusing on two distinct implementations: my own implementation and that of my colleague, Ludovica Danovaro (s4811864). I will indicate as "1" the quantities referred to Ludovica Danovaro's implementation and as "2" the quantities related to mine. The primary objective of the initial assignment in Research Track 1 was to develop a Python node capable of searching for and identifying a silver token within an environment, subsequently pairing it with a corresponding golden token. The arena itself consisted of twelve tokens arranged in concentric circles, with six gold tokens and six silver tokens. The golden tokens were positioned on a larger radius circle in comparison to the silver tokens.

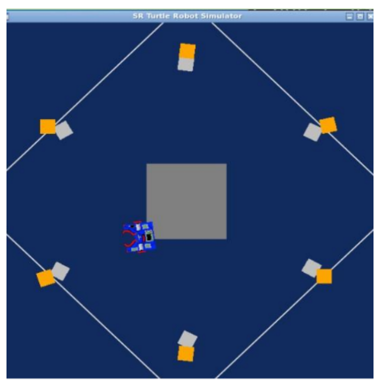

Figure 1: The arena

The experiments involved altering the appearance of the arena setup to investigate the performance of two different robotic controllers. By comparing the outcomes of my implementation with that of Ludovica Danovaro's, we can gain valuable insights into the efficiency of different approaches to solving the task. Through statistical analysis, we can identify any significant differences in performance between the two implementations and draw conclusions regarding the strengths and weaknesses of each approach.

2 General description of the experiments

The experiments conducted for the statistical analysis were focused on measuring the average time required to complete the silver-gold token pairing task. In order to ensure diverse scenarios and paths for the robot, the radius of the circle containing the golden tokens was deliberately varied during the simulations. To facilitate random token placement and introduce variability in the experiments, modifications were made to the "two-colours-assignment-arena" file. By

adjusting the random seed parameter, multiple iterations of the larger radius circle were generated, resulting in different positions for the golden tokens in each simulation. This approach allowed for randomized token configurations and added complexity to the task.
To ensure a robust analysis, 30 simulations were performed for each token assignment. For each simulation, a distinct random seed parameter was used, resulting in unique token placements and environmental setups. By conducting a sufficient number of simulations, we increase the reliability of our findings and reduce the impact of random fluctuations on the overall results.
The data collected from these simulations provides a comprehensive dataset for statistical analysis. By calculating the average time required for completion in each set of simulations with varying token assignments, we can compare the performance of the two implementations. Additionally, we can analyze the variability in the results and determine whether any observed differences are statistically significant.
The use of multiple simulations with different random seed parameters ensures that the statistical analysis captures the overall performance trends and accounts for the influence of random factors. This approach enhances the reliability and validity of the analysis, enabling us to draw meaningful conclusions about the relative efficiency of the robotic controllers in different token assignment scenarios.

3 Hypotesis

In order to initiate the analysis, we formulated a "null hypothesis" which assumes that there is no significant difference between the two implementations. This hypothesis suggests that both algorithms are equally effective in completing the task.
Contrarily, the "alternative hypothesis" challenges the null hypothesis by proposing that there is a specific difference between the two algorithms in terms of their performance. It implies that one algorithm may exhibit superiority or inferiority compared to the other. To start the analysis, we assume the null hypothesis to be true and the alternative hypothesis to be false. This means that we start by assuming there is no substantial difference between the two implementations. Our goal is to verify whether the means of the completion times for the task are the same for both cases, indicating equal efficiency.
To assess the validity of our assumption, we will conduct a T-Test. The T-Test is a statistical method that enables us to compare the means of two datasets and determine if they are significantly different from each other. By performing the T-Test on the completion times of the two implementations, we can infer whether there is a statistically significant difference between them.

4 Analysis

As previously mentioned, the experiments are based on the average time required to finish the task. In order to lead the analysis we used Microsoft Excel. In the last page you can find a table that shows the time needed to perform the task in my implementation and in Ludovica's implementation. Here you can find a plot that compares the times in the two cases.

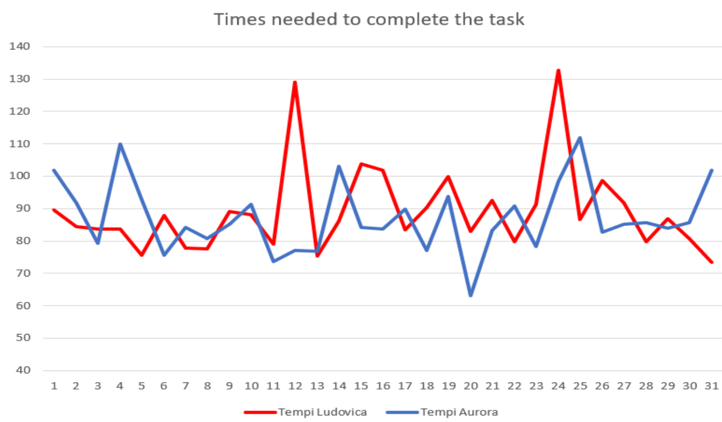

Figure 2: Comparison of the times needed to complete the task in the two cases

After taking the times, I computed the mean value μ, which is obtained by dividing the sum of observed values by the number of observations, n. The formula in the two cases (Ludovica: l, Aurora: a) is:

$$\mu_l = \frac{\sum_{n=1}^{N1} X_i}{N1} = 89,13309677 \qquad \mu_2 = \frac{\sum_{n=1}^{N2} X_i}{N2} = 86{,}68436667$$

Then I also computed the standard deviation. The standard deviation gives an idea of how close the entire set of data is to the average value. Data sets with a small standard deviation have tightly grouped, precise data. Data sets with large standard deviations have data spread out over a wide range of values. The formula in the two cases is:

$$\sigma_1 = \sqrt{\frac{\sum_{n=1}^{N1}(X_i-\mu_1)^2}{N1}} = 13,52927975 \qquad \sigma_2 = \sqrt{\frac{\sum_{n=1}^{N1}(X_i-\mu_2)^2}{N1}} = 10,84069019$$

where N1, N2 = 30 and X_i are the values of the times.
Based on the provided means, it seems that the average performance of the two implementations is relatively close. However, it is important to conduct a statistical analysis, such as the two-sample t-test, to determine if this observed difference is statistically significant or if it could have occurred by chance.
Therefore, it is important to perform the statistical analysis to obtain a more

accurate conclusion about the equality or difference between the two implementations. Simply comparing the means alone may not provide sufficient evidence to support a definitive conclusion.

5 T-Test

To go deep in the analysis, I decided to perform the T-test.T-Test, also known as Student's Test, is based on t-distribution and is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples. In other words, it checks if there are significant differences between the means of two groups. In my case the two groups are the time needed to perform the task in Ludovica's implementation and in mine. First of all I have computed the t-value = 0.72923473. The t-value has to be compared to the values available in a t-table A t-table shows also a degree of freedom (DoF) which is closely related to the sample size. In a t-test for com- paring the means of two groups, the degrees of freedom of the error correspond to the total number of observations minus 2 (DOF = N1 + N2 - 2), where N1 and N2 represent the sample sizes of the two groups. I have also calculated the pooled variance, which is:

$$\sigma^2_{pooled} = \frac{(N_1 - 1) \cdot s_1 + (N_2 - 1) \cdot s_2}{N_1 + N_2 - 2} = 150,2809871$$

The pooled variance leads to the pooled, estimated Standard Error of the sampling distribution of the difference of means. The formula is:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2_{pooled}}{N_1} + \frac{\sigma^2_{pooled}}{N_2}} = 3,165238138$$

We are interested in the differences, thus the t-statistics turns into:

$$t_{\bar{x}_1 - \bar{x}_2} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = 0,729234738$$

Let's consider a level of significance of 5% and a corresponding critical value of 2.042 (considering a t-distribution with 58 degrees of freedom), the calculated t-value of 0.729234738 is lower than the critical value. This indicates that there is not enough evidence to reject the null hypothesis.

6 Conclusions

The conclusions are the following: with a significance level of 5%, we cannot make the claim that one of the algorithms has a significantly different in time compared to the other. The results of the t-test suggest that, based on the available data, there is no significant statistical evidence to support the hypothesis of differences between the two algorithms in terms of time to complete the task.

RandomSeed	Time in Ludovica's implementation	Time in my implementation
10	89,631	101,834
11	84,444	91,701
12	83,590	79,259
13	83,689	109,813
14	75,586	92,693
15	87,828	75,652
16	77,921	84,191
17	77,512	80,714
18	89,136	85,232
19	87,985	91,318
20	79,109	73,759
21	128,936	77,177
22	75,256	76,769
23	86,182	102,957
24	103,678	84,212
25	101,921	83,711
26	83,328	89,778
27	90,432	77,159
28	99,967	93,801
29	82,934	63,197
30	92,517	83,189
31	79,692	90,757
32	91,187	78,221
33	132,641	98,462
34	86,519	111,796
35	98,713	82,681
36	91,768	85,228
37	79,845	85,677
38	86,893	83,829
39	80,812	85,764

Table 1: Table of times needed to complete the task