Research Track 2 - Assignment 3
Statistical Evaluation

Francesca Corrao

May 2023

1 Introduction

The goal of this assignment is to make a statistical evaluation. In particular the goal is to compare two different algorithm developed for the first Assignment of Research Track 1. The two algorithm taken into account are:

1. Robot 1: `https://github.com/Francesca-Corrao/Assignment1`
2. Robot 2: `https://github.com/Tabi43/First-Assignment-RT1`

In the following paragraphs the steps made in order to make a statistical evaluation are shown. In particular an Hypothesis is made and to prove it, whit a certain significance level, a test is design. In order to found the data to perform the test some experiment are done.

2 Hypothesis

The Hypothesis made is that:

> "the algorithm developed by me (1) has a higher rate of success than the one developed by my collegue (2) when the position and number of tokens in the map change"

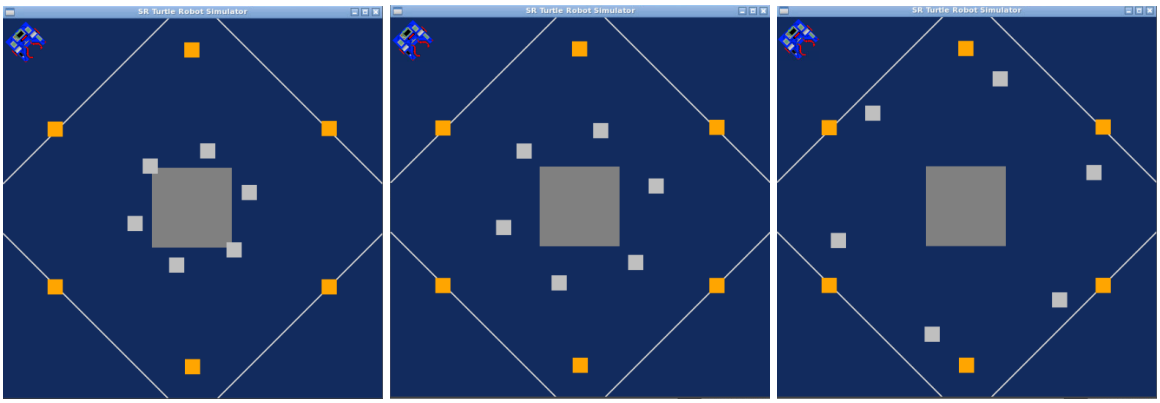
The *rate of success* is the number of silver-golden token paired over the number of token in the enviroment. A couple of silver-gold tokens is considered *paired* if during it's execution the robot takes the silver token and release it to the gold, once a couple is paired what happens to it doesn't matter,and for the sake of experiments will be considered paired in the end. This is lead to the following definition of null and alternativa Hypotesis:

- $H_0 : \mu_1 = \mu_2$ is the **null Hypotheis**;
- $H_1 : \mu_1 > \mu_2$ is the **alternative Hypothesis**.

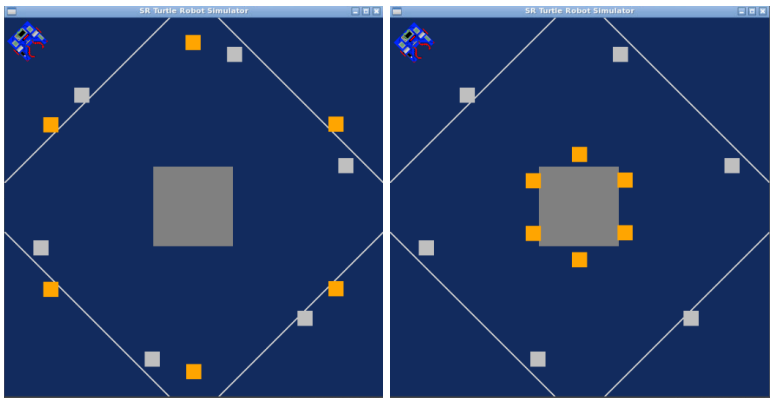
This means the test that should be used is a **one-tailed test** in particular a righ-tailed test since the goal is to prove that the rate of success of algorithm 1 is greater than the rate of success of algorithm 2.
To prove H_1 the null Hypothesis H_0 must be rejected with a **significant level** of 5%. An alternative could be to use a 1% level of significance but since the values of rate of success go between [0,1] and it's not expect to vary much it's better to choose 5%.
The distribution we aim to use is a t-distribution an in order to apply the central limit theorem, so that the shape of the sampling distribution is more similar that the one of a t-distribution we need a number of sampling greater than 30. This must be keep in mind when design the number of experiments to make.
Lastly a paired T-test is choose to test the hypothesis since we are comparing two different approaches applied to the same scenario.

3 Experiments

In the experiments both algorithms are executed whit different disposition of token in the enviroment and different number of token, and for each execution the rate of success is compute as defined above: number of token paired over total number of token. Let's see what are the different configuration taken into account.
To take into account how the position of token effects the performance of the algorithms 5 different maps have been chosen. In the first four the radius of the inner circle, in which the silver token are placed, is increased until both token are almost on the same circle. In the last one the position of silver token and gold token is swapped, so the gold token are on the inner circle and the silver token are on the outer circle. The different position of token in the environment taken into account for the experiment can be seen in Figure 1.



(a) original configuration (b) inner circle +0.3 (c) inner circle +1



(d) inner and outer circle the same (e) gold and silver swapped

Figure 1: Position of token in the enviroment

To see how the rate of success change with respect to the number of token on the map, the 5 different enviroment designed aboved are run whit 3 different number of token in the map: the original number and the case in which 2 token are added and 2 token are remove. In the end the experiment is perform with 6 token, 4 token and 8 token. To have a more robust measurement of the rate of success for each configuration (position-number of token) the algorithms are run 3 times. In the end the over all states evaluated are 45 (5x3x3) which is greater that 30 and therefor the central limit theorem allow us to say that the sample distribution has the shape of a t-distribution. The results obtained are shown in the table 1.

Position	n_token	Robot 1			Robot 2			Difference		
		rate_success			rate_success			$x_1 - x_2$		
original map		1	1	1	1	1	1	0	0	0
inner 1.2		1	1	1	1	1	1	0	0	0
inner 2	6	1	1	1	0	0	0	1	1	1
inner = outer		0,5	1	0.5	0	0	0	0.5	1	0.5
swap silver-gold		0,167	0,667	0.5	0,333	1	0.5	-0.167	-0.333	0
original map		1	1	1	1	1	1	0	0	0
inner 1.2		1	1	1	1	1	1	0	0	0
inner 2	4	1	1	1	1	1	1	0	0	0
inner = outer		1	1	1	1	1	1	0	0	0
swap silver-gold		0,5	0,5	0,75	1	0,75	1	-0.5	-0.25	-0.5
original map		0,75	0,75	0,75	0	0,625	0	0.75	0.125	0.75
inner 1.2		0,75	0,75	0,75	0,875	1	1	-0.125	-0.25	-0.25
inner 2	8	0,25	0,25	0,25	1	1	0,875	-0.75	-0.75	-0.625
inner = outer		0,625	0,75	0,75	0,125	0,125	0,125	0.5	0.625	0.625
swap silver-gold		0,125	0,125	0,125	0,125	0,625	0,625	0	-0.5	-0.5

Table 1: Experiment Result

4 Paired T-test

To test the Hypothesis a Paired T-test is performed with the data obtained from the experiments (Table 1). Since the goal is to prove $H_1 : \mu_1 > \mu_2$, so to reject the null Hypothesis $H_0 : \mu_1 = \mu_2$ with a significant level of 5%.
To prove this the following values are computed:

- the difference $d = x_1 - x_2$ between the two observation of each pair, last column of Table 1;
- the **mean** of the difference it $\bar{d} = 0,070$
- the **standard deviation** of the difference $s_d = 0,450$
- the **standard error** of the difference $SE(d) = \frac{s_d}{\sqrt{N}} = 0,067$
- the **t-value** $t = \frac{\bar{d}}{SE(d)} = 1,045$

Once the t-value it's found it's compered with the value on the t-table in this case (44 DOF and 5% level of significance) 1,680. Since the t-value computed is less then the one on the t-table the null Hypothesis H_0 can't be rejected with a confidence of 95%. In particular the p-value associated to the t-value it is 0.15 which means that the null hypothesis can be rejected whit a confidence of 85% which means a 15% level of significant which is 3 times the one requested.

5 Conclusion

In the end whit the experiments done so far the Hypothesis isn't proven. To make sure that the null Hypothesis can't be rejected the power of the test in detecting false null hypothesis could be increase by increasing the sample size and there for make more experiments. This is because as the sample size increases the mean and the standard errors will stabilize to the true parameter and there for the value of the t-test will be more correct. The experiment could be extended:

- by designing other map in which the position of the token is changed, by taking into account another configuration the sample size is 54 (6x3x3);
- by taking more measurement for each map, in fact by adding another measurement for each environment configuration with the different number of token sample size became 60 (5x4x3);
- by considering other number of token in the map, for example take into account also when one token is added or removed from the map (5 token and 7 token) this will bring the sample size to 75 (5x3x5).

***Probabilmente questa è da rimuovere.
To see if it's worthy to take other sample or the result obtained with the paired t-test is the same: another measurement for each map is assumed whit and optimistic approach. In the optimistic approach all the new measurement will advantage Robot 1, this means that for each set of data taken (3 for each map) the highest rate of success is taken for Robot 1 and the lowest for Robot 2. With this optimistic assumption and a new ideal data set of 60 samples the result found is that the t-value will be $t = 1.73$ that is greater that the one in the t-table and the corresponding p-value is $p = 0.044$ so the null Hypothesis could be discard with a level of significance of 5%. this scenario is very difficult to happen because even if only a few of the experiment benefit Robot 2 then the result found is similar to the one obtained before and so again the null Hyphotesis can't be discard with a confidence of 95%.