Statistical analysis

Abstract

The purpose of this assignment is to perform a statistical analysis between two algorithms, which accomplish a specific task, on the base of initial hypothesis, formulated by me. The assumption could be, for instance, the average of execution time or the number of failure and successes, when the number of tokens are changed; in any case it is important to point that any hypothesis must satisfy five different characteristics:

- It is provable or disprovable;
- It is simple and clear;
- It is testable;
- It is relevant to the problem;
- It is specific;

Once I have done this, hypothesis should be tested, by executing several times both algorithms and collecting data, for achieving some conclusions, which consist on accept or reject the initial hypothesis.

Introduction and hypothesis

My situation is based on the comparison between two searching algorithms, developed by me and my colleague Veronica Gavagna, in an environment with silver and gold tokens. For the sake of simplicity, I will refer my algorithm to as **Algorithm-B**, and the one of my colleague as **Algorithm-G**. Both end when any silver token is paired with only one gold by a mobile robot. An interesting analysis could be tested the number of successes and failures of them, by changing randomly the number of the tokens, starting from 2 up to 10 and setting, as constant values, the inner radius to 1 and external to 2. Both algorithm are tested under the same initial condition.

At the beginning I have formulated a **null hypothesis**, which is the one that a researcher wants to reject, based on the assumption of no difference between the two, and an **alternative hypothesis,** which is the one that a researcher wants to prove, based on the assumption of some differences in performance. Hence these are my assumptions:

- **Null hypothesis:** Algorithm-B and G have no difference in number of failures and successes;
- **Alternative hypothesis:** Algorithm-B and G have some differences in the performance of the two algorithms, in particular in the number of successes and failures;

Obviously my hypothesis are consistent with the 5 points that I have mentioned in the **abstract** section. The last thing that I want to point out is the following: I started from the null hypothesis and then, after testing and discussing the obtained results, I verified whether it is proved or disproved; if I am in the former case, the **alternative hypothesis** is rejected and the **null** is

accepted; in the latter one, the **null hypothesis** is rejected and the **alternative** is accepted. During the experiment, I use a level of significance of 5%.

Test and collecting data

It is the time to test my hypothesis, from the starting point of **null hypothesis**; as I have said before, I had to execute several times both algorithms and collecting data in a table. I have decided to execute them 80 times, by changing randomly the number of tokens from 2 to 10; then I have assigned **0 regarding failures** and **1 regarding successes**. For the sake of understanding better the results, I have made the decision to analyze the first 64 repetitions and, later, the last 16. The table below refers to the first 64 execution times, and the number of tokens has a range from 2 to 6.

Test	Number of tokens	Results of Algorithm-G	Results of Algorithm-B
1	2	1	1
2	2	1	1
3	2	1	1
4	2	1	1
5	2	1	1
6	2	1	1
7	2	1	1
8	2	1	1
9	2	1	1
10	2	1	1
11	2	1	1
12	2	1	1
13	2	1	1
14	2	1	1
15	2	1	1
16	2	1	1
17	3	1	1
18	3	1	1
19	3	1	1
20	3	1	1
21	3	1	1
22	3	1	1
23	3	1	1
24	3	1	1
25	3	1	1
26	3	1	1
27	3	1	1
28	4	1	1
29	4	1	1
30	4	1	1

31	4	1	1
32	4	1	1
33	4	1	1
34	4	1	1
35	4	1	1
36	4	1	1
37	5	1	1
38	5	1	1
39	5	1	1
40	5	1	1
41	5	1	1
42	5	1	1
43	5	1	1
44	5	1	1
45	5	1	0
46	5	1	1
47	5	1	1
48	5	1	1
49	5	1	1
50	5	1	1
51	5	1	1
52	6	1	1
53	6	1	1
54	6	1	1
55	6	1	0
56	6	1	1
57	6	1	1
58	6	1	1
59	6	1	1
60	6	1	1
61	6	1	1
62	6	1	0
63	6	1	1
64	6	1	0

Table of first 64 executions of both algorithms.

	Total successes	**Total failures**
Algorithm-G	64	0
Algorithm-B	60	4
Total	124	4

Table of total successes and failures of first 64 executions

For coming to some conclusion, I need to use a statistical test, which is the **chi square test;** it is a non-parametric test, based on the computation of ξ^2 (xi square) value, that expresses the level of difference between observed and expected data, based on an initial assumption; finally this value need to be compare with respect to a critical value on a reference table. Degree of freedom is another important thing to compute; it is equal to $(n_r - 1)(n_c - 1)=(2-1)(2-1)=1$, where $\mathbf{n_r}$ is the number of rows and $\mathbf{n_c}$ the number of columns. For computing the xi square value, I need to know E_{ij}, that is the expected frequency. Since I consider the **null hypothesis**, this number is equal to 62, regarding the successes, and 2 regarding failures ; in other words this means that both algorithms are expected to, respectively, succeed and failure: 124/2=62 and 4/2=2 runs out of 64. After doing that, the xi square value is computed as follow:

$\xi^2=((64-62)^2/62+(60-62)^2/62+(0-2)^2/2+(4-2)^2/2)=4.129032258$

The comparison of this result with the critical value, and its meaning, is discussing in the conclusion section.

Now I can add to the previous table the last 16 executions, and these are the results:

Test	Number of tokens	Results of Algorithm-G	Results of Algorithm-B
65	7	1	0
66	7	1	0
67	7	1	0
68	7	1	0
69	8	0	0
70	8	1	0
71	9	1	0
72	9	1	0
73	9	1	0
74	9	1	0
75	9	1	0
76	10	1	0
77	10	1	0
78	10	1	0
79	10	1	0
80	10	1	0

Table of remaining 16 executions; consider it as the continue of the table before

	Total successes	Total failures
Algorithm-G	79	1
Algorithm-B	60	20
Total	139	21

Table of full 80 executions

I compute the same previous quantities: E_{ij} (successes)=139/2=69.5, E_{ij} (failures)=21/2=10.5,

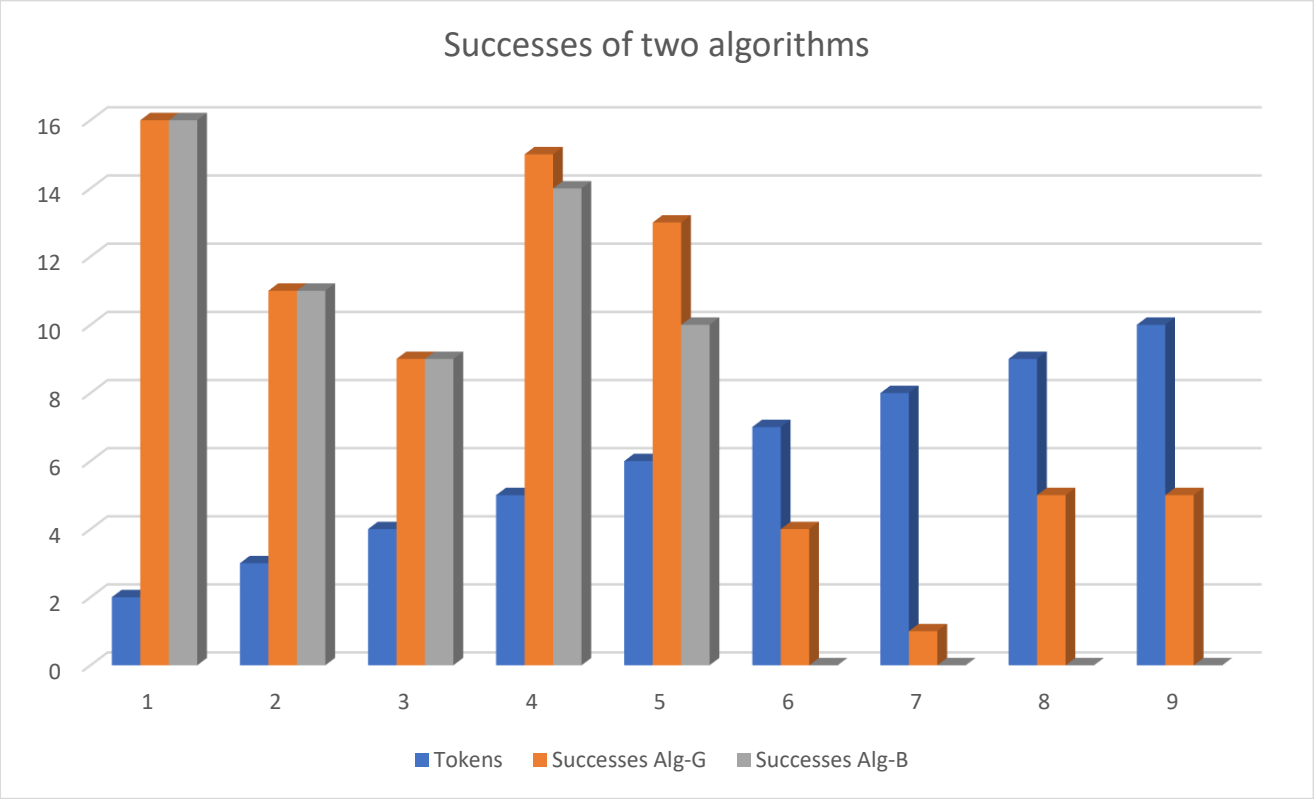
$\xi^2=((79-69.5)^2/69.5+(60-69.5)^2/69.5+(1-10.5)^2/10.5+(20-10.5)^2/10.5)=19.78759849$

The conclusion and the comparison between the previous result will be discussed soon.

For concluding this section and for the sake of clarity, I want to report the results of total successes and failures per number of tokens, through tables and bar charts.

Tokens	Successes Alg-G	Successes Alg-B
2	16	16
3	11	11
4	9	9
5	15	14
6	13	10
7	4	0
8	1	0
9	5	0
10	5	0

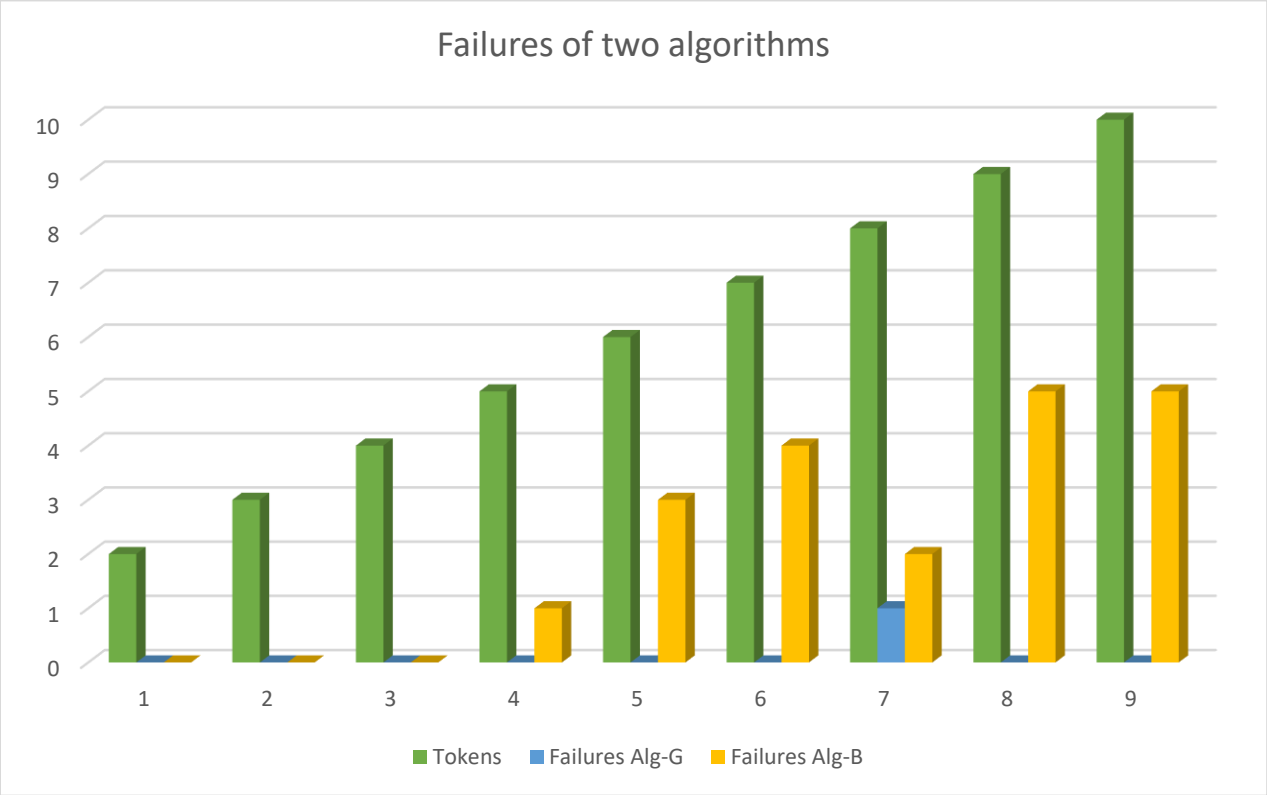
Table of successes of two algorithms



Bar chart of successes of two algorithms

Tokens	Failures Alg-G	Failures Alg-B
2	0	0
3	0	0
4	0	0
5	0	1
6	0	3
7	0	4
8	1	2
9	0	5
10	0	5

Table of failures of two algorithms



Bar chart of failures of two algorithms

Conclusion

For the sake of simplicity, I sum up the two xi square values that I have computed before:

- In the first 64 executions, the xi square value is 4.129032258; I call this as the first situation;
- In the last 16 executions, the xi square value is 19.78759849; I recall this as the second situation;

How can I interpret these results? I need to compare them with respect to a critical value, according to the table below; I am interested only in the first row of the table, because the degree of freedom is 1. Remember that if my statistic value is bigger than the red one, I can reject the **null hypothesis**; otherwise I cannot reject it. In the first case I could say that there is enough evidence to say there are some differences in the performance; in the second one I could say that there is not enough evidence to say that some differences exist.

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Reference table of chi square values

Hence the conclusion of the first situation is the following: using a level of significance of 5%, I can see that our statistic test is bigger than the critical value, so I can reject the **null hypothesis** and I can say that there are some significant differences in the performance of the two

algorithms; in addition I can affirm that the probability of committing an error, by rejecting the null hypothesis, is greater than 0.05. Regarding the second case, the evidence is clearer: using the same level of significance, I can see that my xi square value is not only bigger than the critical one, but also than the all after it; this means that, in this case, a level of significance is not useful and, according to the table above, I cannot say anything about the probability of committing and error if I reject the **null hypothesis**; this is consistent to the fact that, from the 65^{th} to the 80^{th} execution, the algorithm-B never ends. Also in this situation, it is possible to say with a much higher accuracy that I can reject the **null hypothesis**, and the **alternative** one can be accepted.

So in both situations I reached what a research want to prove: reject the **null hypothesis**, and accept the **alternative one**.