Metodologias Experimentais em Informática - UC

Exam Schedules - Hypothesis testing

Gui Costa, 2021186342, guibscosta@hotmail.com
Tomás Ventura, 2018279147, tventura@dei.uc.pt
Gustavo Gama, 2020180035, gustavo.p.gama@gmail.com

1. Introduction

In this paper we're going to perform hypothesis testing based on our previous reports (EDA and hypothesis). We're first going to recap the hypothesis used and also gonna showcase the data used with some confidence intervals, once that preliminary step is done we're gonna perform the hypothesis testing.

1.1. Hypothesis

As stated in our previous report of the second milestone, the hypothesis we will be testing and analysing in this document are the following:

First:
● H0 - Probability p has no effect on the execution time for both algorithms for a fixed N.
● H1 - Probability p has a significant difference between the algorithms for a fixed N

Second:
● H0 - For a lower N value (number of exams <= 20) and for a fixed p value (probability), there's no significant difference between both algorithms (code 1 and code 2) in terms of execution time.
● H1 - Code 1 is faster than code 2 for a lower N value.

Third:
● H0 - For a higher N value (number of exams >= 20) and for a fixed p value (probability), there's no significant difference between both algorithms (code 1 and code 2) in terms of execution time.
● H1 - Code 2 is faster than code 1 for a higher N value.
Some of those hypotheses may have been slightly changed.

1.2. Confidence Intervals (95% confidence)

The following tables show a few results obtained for the confidence intervals (95%) for the data used (an example of it) during our experiments.

1.2.1. Variation of the number of exams N.

The following table shows a few results obtained for the confidence intervals for our data for a fixed P=0.3 and for a few variations of N. The averages are the dependent variable execution time.

Code	N	Average Time (s)	Confidence Interval
Code 1	20	0.00043	(7.32e-05 ; 0.0008)
Code 2	20	0.00023	(0 ; 0.00047)
Code 1	30	0.60	(0 ; 1.560)
Code 2	30	0.064	(0.018 ; 0.11)
Code 1	40	9.723	(5.01 ; 14.43)
Code 2	40	19.61	(11.41 ; 27.80)

1.2.2. Variation of the probability of overlap P.

The following table shows a few results obtained for the confidence intervals for our data for a fixed N=0.3 and for a few variations of P. The averages are the dependent variable execution time.

Code	P	Average Time (s)	Confidence Interval
Code 1	0.15	0.007	(0.0027; 0.011)
Code 2	0.15	0.0178	(0.003; 0.032)
Code 1	0.25	0.041	(0.01 ; 0.072)
Code 2	0.25	0.049	(0.02 ; 0.076)
Code 1	0.3	0.612	(0; 1.560)
Code 2	0.3	0.064	(0.017 ; 0.11)

2. Hypothesis testing 1

<u>Hypothesis:</u> Probability p has no effect on the execution time for both algorithms for a fixed N.

In order to perform hypothesis testing, we've decided to use Two-way ANOVA with a p-value and a level of significance of 0.05. To perform this test we've decided to set a fixed value of exams N=35. We've arranged our data by cleaning up extreme values (those that reach 60 seconds, our max cap for execution time) in order to obtain a balanced design for our ANOVA.

So, for our model we have 2 factors, the probability P and the algorithm "Code". We have a 4x2 design with 30 different measures for each pair (Code,P):

- With 4 levels for our P values in the range [0.15 ; 0.35].
- With 2 levels for the algorithm : code 1 and 2.

Now that the configuration is set, we're going to present the results for the experiment with the two-way ANOVA; an overview of the data can be seen on the boxplots (Fig.1-2).

Box Plots

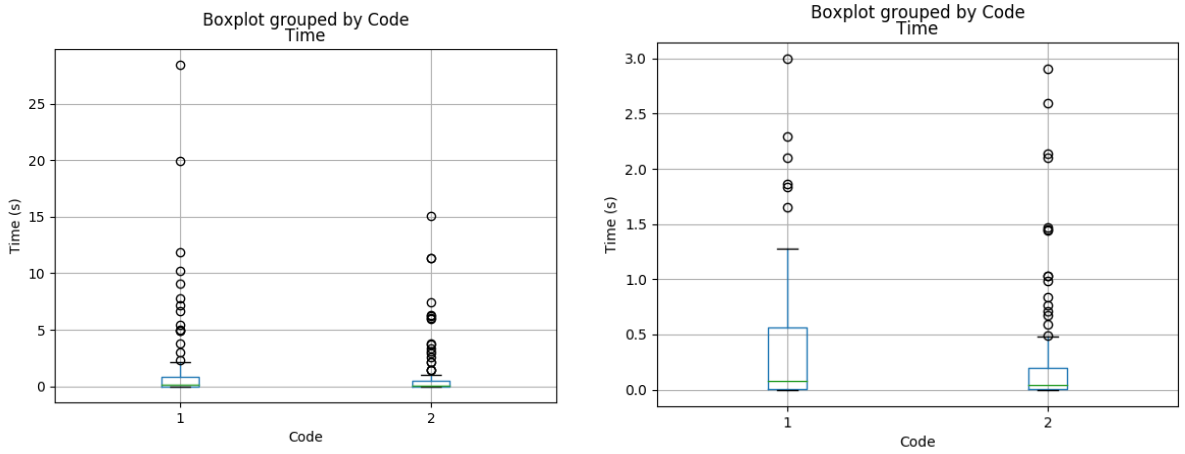


Fig.1 (Left): Box plot Time for each code - Hyp.3 (whole data)
Fig.2 (Right): Box plot for each code - Hyp.3 (zoomed in lower ranges of time)

Statistical test results:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
factor(Code)	1	11.2	11.25	1.289	0.2574	
factor(P)	3	206.4	68.81	7.887	4.95e-05	***
factor(Code):factor(P)	3	78.6	26.19	3.001	0.0313	*
Residuals	231	2015.4	8.72			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Fig.1: Hyp.3 - Two-way ANOVA results

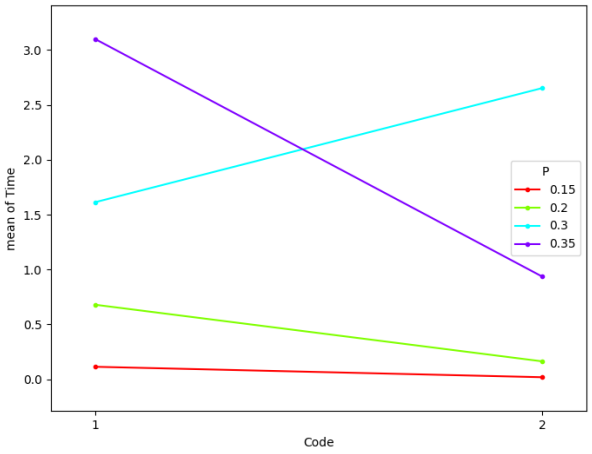


Fig.4: Interaction plot for the P values and codes - Hyp.3

Based on our results for this scenario/parameters, we can see the following (Fig.3):

- The p-value for the independent variable P is inferior to 0.05, i.e it's significant. The probability P significantly affects the execution time (the dependent variable).
- The p-value for the algorithm (code) is higher than 0.05, i.e it's not statistically significant. It suggests that the code doesn't impact the dependent variable (execution time).
- Concerning the interaction between the variable P and the code, the p-value is also significant, this suggests that the relation between the variables affects the results obtained for the dependent variable for this configuration. It can also be verified in figure 4.

After analyzing these results, it's important to check the ANOVA assumptions before accepting or rejecting the null hypothesis. For that, we're going to use visual tools and different tests (code execution traces) to check the assumptions.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   7  4.4406 0.0001185 ***
      231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 5: Levene Test - Hyp.3.

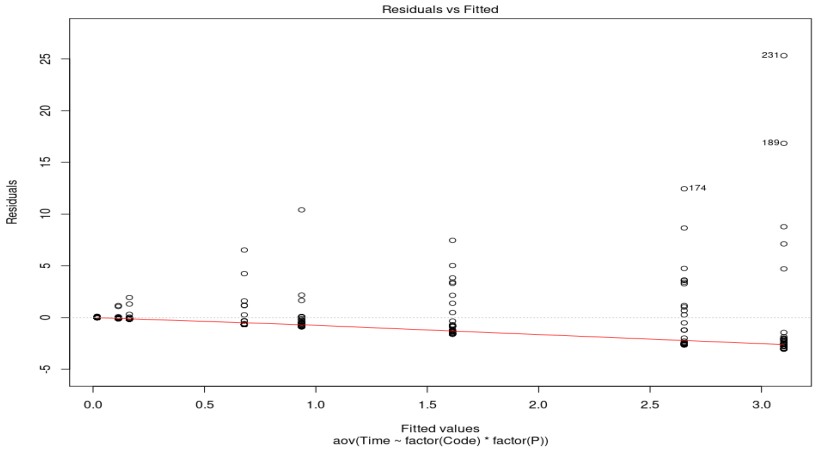


Fig.6: Residuals vs fitted plot, check homoscedasticity (constant variance).

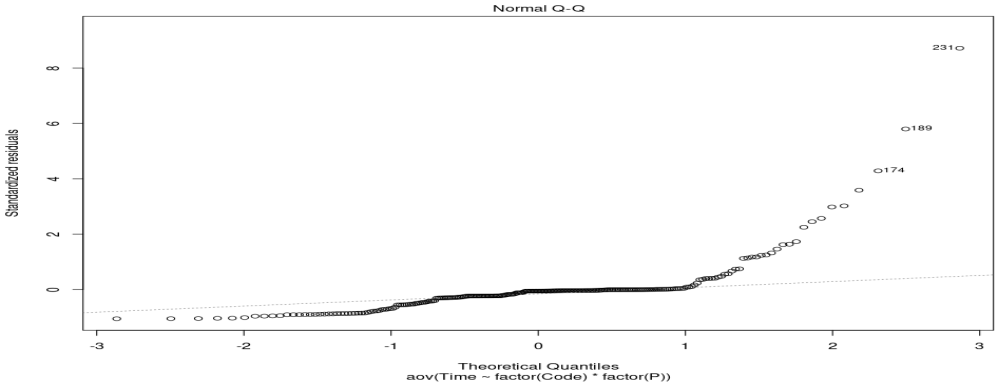


Fig.7: Normal QQ-plot.

Based on our results (Fig.5-8), we can see that the assumptions for the ANOVA are violated, the dataset fails on both tests which might be related to the nature of the data:

- *First assumption test violation (homogeneity of variances):* we have a significant p-value for our Levene Test (Fig.5), even though the "residuals vs fitted plot" (Fig.6) we can see that the spread of the residuals isn't really the same across the x-axis.
- *Second assumption test violation (the normality of the residuals)*: the Shapiro-Wilk Test (p-value <2.2e-16) showed that the p-value is significant, this suggests that the residuals are not drawn from a normal distribution. Also, in the fig.7, the QQ-plot clearly suggests a non-normality.

Anyways, the ANOVA is known for its robustness in case of broken assumptions, so to be sure of those results we've decided to perform a few non parametric tests before answering the hypothesis.

The non parametric tests chosen are an alternative for the Two-way ANOVA with interactions, they will allow to check if our results are similar with the test done previously on the first Two-way ANOVA and see their validity, here are the non parametric tests:

- Randomization tests which compute F statistics for each sample taken.
- ANOVA on ranks.

Randomization test (non parametric)

For this test, the results from table 1 allow us to obtain the same conclusions as the Two-way ANOVA done previously. Having that in mind, we can then proceed on the first Two-Way ANOVA and perform the post-hoc tests (Tukey).

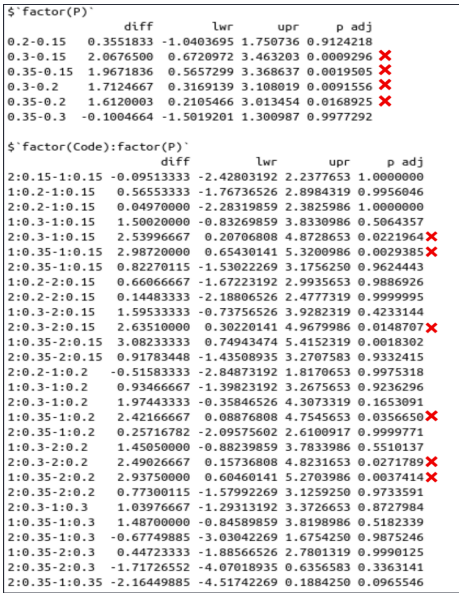
p-value P (probability P)	p-value C (code)	p-value CP (interaction)
0 (significant)	0.256 (non significant)	0.0258 (significant)

Table 1: Randomization test - Hyp.3

ANOVA on Ranks (non parametric)

We've also performed the ANOVA on ranks to give more confidence on our first Two-Way ANOVA and compare the results, but this is a very limited approach, we've obtained a few results that violate the homoscedasticity on the ANOVA on ranks and in general, rank based approaches are non robust when homoscedasticity is violated [1] (many Type I errors).
So we decided to stick with the Randomization test for our conclusions.

To finish, we can apply post-hoc tests to see the differences; here's the results of our Tukey test on the first Two-way Anova:



```
$`factor(P)`
               diff        lwr      upr     p adj
0.2-0.15   0.3551833 -1.0403695 1.750736 0.9124218
0.3-0.15   2.0676500  0.6720972 3.463203 0.0009296 ✗
0.35-0.15  1.9671836  0.5657299 3.368637 0.0019505 ✗
0.3-0.2    1.7124667  0.3169139 3.108019 0.0091556 ✗
0.35-0.2   1.6120003  0.2105466 3.013454 0.0168925 ✗
0.35-0.3  -0.1004664 -1.5019201 1.300987 0.9977292

$`factor(Code):factor(P)`
                        diff         lwr       upr     p adj
2:0.15-1:0.15 -0.09513333 -2.42803192 2.2377653 1.0000000
1:0.2-1:0.15   0.56553333 -1.76736526 2.8984319 0.9956046
2:0.2-1:0.15   0.04970000 -2.28319859 2.3825986 1.0000000
1:0.3-1:0.15   1.50020000 -0.83269859 3.8330986 0.5064357
2:0.3-1:0.15   2.53996667  0.20706808 4.8728653 0.0221964✗
1:0.35-1:0.15  2.98720000  0.65430141 5.3200986 0.0029385✗
2:0.35-1:0.15  0.82270115 -1.53022269 3.1756250 0.9624443
1:0.2-2:0.15   0.66066667 -1.67223192 2.9935653 0.9886926
2:0.2-2:0.15   0.14483333 -2.18806526 2.4777319 0.9999995
1:0.3-2:0.15   1.59533333 -0.73756526 3.9282319 0.4233144
2:0.3-2:0.15   2.63510000  0.30220141 4.9679986 0.0148707✗
1:0.35-2:0.15  3.08233333  0.74943474 5.4152319 0.0018302
2:0.35-2:0.15  0.91783448 -1.43508935 3.2707583 0.9332415
2:0.2-1:0.2   -0.51583333 -2.84873192 1.8170653 0.9975318
1:0.3-1:0.2    0.93466667 -1.39823192 3.2675653 0.9236296
2:0.3-1:0.2    1.97443333 -0.35846526 4.3073319 0.1653091
1:0.35-1:0.2   2.42166667  0.08876808 4.7545653 0.0356650✗
2:0.35-1:0.2   0.25716782 -2.09575602 2.6100917 0.9999771
1:0.3-2:0.2    1.45050000 -0.88239859 3.7833986 0.5510137
2:0.3-2:0.2    2.49026667  0.15736808 4.8231653 0.0271789✗
1:0.35-2:0.2   2.93750000  0.60460141 5.2703986 0.0037414✗
2:0.35-2:0.2   0.77300115 -1.57992269 3.1259250 0.9733591
2:0.3-1:0.3    1.03976667 -1.29313192 3.3726653 0.8727984
1:0.35-1:0.3   1.48700000 -0.84589859 3.8198986 0.5182339
2:0.35-1:0.3  -0.67749885 -3.03042269 1.6754250 0.9875246
1:0.35-2:0.3   0.44723333 -1.88566526 2.7801319 0.9990125
2:0.35-2:0.3  -1.71726552 -4.07018935 0.6356583 0.3363141
2:0.35-1:0.35 -2.16449885 -4.51742269 0.1884250 0.0965546
```

Fig.8: Tukey results for the initial Two-way ANOVA

We can see that there's significant differences (Fig.8) between different pairs of parameters P (overlaps probabilities) and also in the interaction Code and P.

In this setup with a more narrow range of overlap probability P, we can indeed notice an interaction between the factor P and the factor Code; this can be visually seen in the Fig.4, for P=0.35 and P=0.3. In this case of significant interaction but when the main effect isn't significant (the main effect Code), we could imagine that the effect of the Probability P has an opposite effect on the dependent variable based on the Code [2].

It's important to note one thing, this interaction only happened when we set a narrow range of P values. If we decided to spread the P values on a larger range, we couldn't obtain this interaction. For example, for the exact same setup as the previous ANOVA but with P in the range [0.15;0.6] (more spread), we actually obtain only a significant value for the factor P. The rest are non-significant (same statistical tests, i.e for both Two-Way ANOVA and on ranks). This is due to the fact that the probability overlap P increases the time execution very fast like we've seen during the milestone 1 on the EDA so if we have a wide range of P values levels we can't really have an interaction.

3. Hypothesis testing 2 and 3

3.1. Testing

Hypothesis 2: For a N <= 20 and a fixed P value, there's no significant difference between both algorithms (code 1 and code 2) in terms of execution time.

Hypothesis 3: For a N >= 20 and a fixed P value, there's no significant difference between both algorithms (code 1 and code 2) in terms of execution time.

For this hypothesis testing we've decided to set N=20 and the probability P=0.35. Based on those parameters and based on our results obtained for the previous hypothesis by using Two-Way ANOVA and the complications/violations of assumptions related to the data on that approach, i.e the assumptions of parametric statistics are not met and the

distribution is not normal, we've decided to use a simple non parametric test: *Mann-Whitney Test.*

This test is a non parametric test based on ranks that allow us to check if two samples come from the same distribution. This test also assumes independence of the data, the observations from both groups have to be independent, which corresponds to our case (different runs/seeds each time).

```
        Wilcoxon rank sum test with continuity correction

data:  Time by Code
W = 558, p-value = 0.06326
alternative hypothesis: true location shift is not equal to 0
```

Fig.10: Wilcoxon test (Mann-Withney) - Hyp.3

According to the statistical test on figure 10, we can see that at 0.05 significance level, the p-value is superior to 0.05 for the Hyp.3, this shows that the underlying hypothesis H0 for the Mann-Whitney test ("the two populations are equal") is not rejected, so for our hypothesis that also means that there are no real differences between our algorithms, in execution times.

For Hyp.2, this time N (Number of exams) value being for a lower bound we obtain analogous results, both algorithms execute extremely fast so we end up with similar populations and the statistical test doesn't allow it to make a difference. We obtain a p-value of 0.82. The main difference between the Hyp.2 and Hyp.3's test is that the data in Hyp.3 has more spread, and is not limited to the bottom range of execution time (0s).

3.2. Checking interaction between N and P for a fixed code.

Based on the fact that there seems to be no significant difference between the codes (1-2). We decided to check the interaction between the 2 independent variables P (probability) and N (number of exams) for a fixed Code (like code 1), in order to try to draw conclusions between those 2 independent variables, for that we performed similar steps as the part 1 of this paper (Two-way Anova), this time we're going to present the results straight away considering the same steps.

```
                     Df Sum Sq Mean Sq F value  Pr(>F)
factor(P)             2   1104     552   3.649 0.02736 *
factor(N)             2  13237    6618  43.732 < 2e-16 ***
factor(P):factor(N)   4   2266     567   3.744 0.00556 **
Residuals           261  39500     151
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig.11: Two-way ANOVA for factors P and N; the code 1 was used.

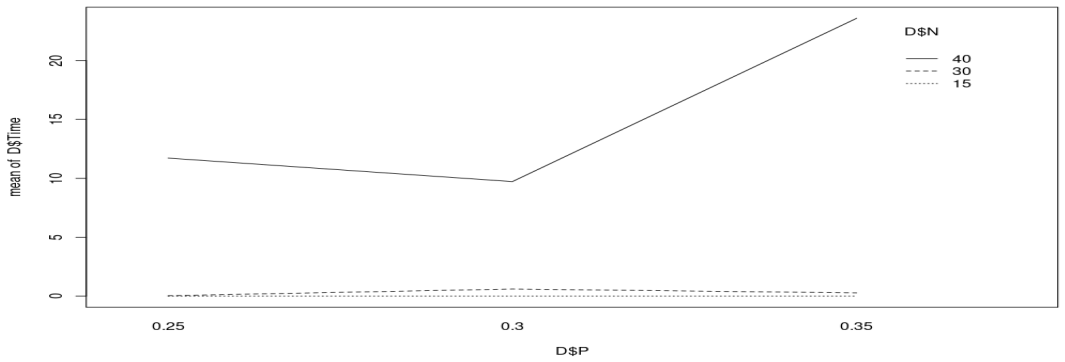


Fig.12: Interaction plot, factor P and N ; with code 1.

Once again, the assumptions for the ANOVA are violated in this case (based on our new Levene tests and Shapiro-Wilk tests, they are all with p-values<0.05), so we use the alternative for the Two-way ANOVA : Randomized tests. The results obtained on that test provide similar conclusions as the ANOVA results on Fig.11, i.e both factors are significant and affect the execution time as expected, but also the interaction between the 2 factors is also relevant (p-values are all inferior to 0.05). That interaction can be seen on the interaction plot presented on Fig.12 , we got unparallel lines for the different levels of the factors, specially N=40.

Based on those results we can conclude with statistically evidence that when we go for the higher ranges of our parameters (the independent variables), the execution time spikes highly. We can see on that interaction plot that for a difference level of 5% (0.3 to 0.35 for P) , we actually doubled our execution time when our N=40, leading to the exponential behavior of our data / problem of exams scheduling. Which confirms some of our results from the EDA part of the project.

4. Conclusion

Given all the tests we performed and the evidence we presented, we can conclude the following regarding the hypothesis we formed: **Hypothesis 1** - the null Hypothesis is rejected, meaning the probability P does in fact play a part regarding the final execution time of the code; **Hypothesis 2 and 3** - the null Hypothesis is not rejected. This means that based on our statistical experiments, we can't conclude that there are significant statistical differences between the codes. We couldn't really tell which one is better for each experiment with different parameters (higher/lower probabilities or higher/lower number of exams).

5.A few references

[1] https://en.wikipedia.org/wiki/ANOVA_on_ranks
[2] https://www.theanalysisfactor.com/interactions-main-effects-not-significant