Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Mestrado em Engenharia Informática
Metodologias Experimentais em Informática – 2021/2022
1º Ano – 1º Semestre

Experimental Methods in Computer Science

Milestone 3 - Hypothesis Testing

Coimbra, 22 de dezembro de 2021

2018298731	João Pedro Pacheco Silva	joaopedro@student.dei.uc.pt
2021209082	João Miguel Mendes Gonçalves	uc2021209082@student.uc.pt
2021205154	Filipe Ribeiro Saudade e Silva	filipe.silva@student.uc.pt

1. INTRODUÇÃO

Propõe-se neste trabalho realizar um estudo sobre a performance de dois algoritmos de *backtracking* aleatório, com o objetivo de adquirir competências na área de design experimental. O trabalho foi dividido em três metas, nas quais fizemos a exploração e regressão linear dos dados, a formulação e pré-registo de hipóteses e por fim o teste das mesmas.

Sendo este relatório referente à terceira e última meta, iremos então descrever o processo realizado para testar as duas hipóteses registadas na meta anterior sendo estas, a Hipótese 1: "O *code1* é mais rápido que o *code2* para números de exame maiores e probabilidades altas", e a Hipótese 2: "O *code2* é mais rápido que o *code1* para números de exame menores e probabilidades baixas".

2. DETALHES DE IMPLEMENTAÇÃO

2.1. Hardware usado

Na realização das experiências foram utilizados dois computadores, sendo que num deles os códigos foram executados numa máquina virtual em *Ubuntu*.

O computador onde foram obtidos os dados para o teste da Hipótese 1 tem as seguintes especificações:

- CPU: Intel(R) Core (TM) i7- 8750H CPU @ 2.20GHz 2.21 GHz
- GPU: NVIDIA GeForce GTX 1080ti
- RAM: 16GB
- Disco: SSD 256GB
- OS: Windows 10 Home

Para a Hipótese 2, foi utilizada uma máquina virtual (*VMWare Workstation* 16) com as seguintes especificações:

- CPU: 2 Processor Cores of Ryzen 7-5800H
- RAM: 2GB
- Disco: 20GB de 516GB (SSD)
- OS: Pop OS 21.10 – Ubuntu Distro

Foi utilizada a máquina virtual em *Ubuntu* para a Hipótese 2, visto que na primeira meta deste trabalho encontrámos muitos dos tempos a 0 quando a duração da execução tinha um valor na casa dos milissegundos. Isto ocorreu porque o *Windows* arredonda o tempo de processamento de cada execução, sendo então a solução realizar as experiências num ambiente *Linux*/*Ubuntu*.

Para a Hipótese 1, como as experiências eram mais demoradas, decidimos utilizar uma máquina em *Windows* visto que tinha mais poder de processamento que a máquina virtual.

2.2. Ferramentas usadas

Com a finalidade de automatizar as experiências, foram implementados *scripts* em *Bash* e *Python* que executavam ambos os códigos e guardavam tanto *output* como os parâmetros usados em documentos de texto (.txt).

Os resultados obtidos foram posteriormente importados o *RStudio,* onde efetuámos os testes às duas hipóteses.

3. SETUP EXPERIMENTAL

As hipóteses formuladas na meta anterior afirmam que o tempo necessário para encontrar uma solução é influenciado pela combinação de 3 fatores, a probabilidade, o número de exames e o algoritmo usado, pelo que teremos de aplicar o teste *Three-way* ANOVA ou uma das suas alternativas não paramétricas para poder ver se existe alguma interação entre estes.

De forma a poder aplicar esses testes definimos um conjunto de níveis para cada um dos fatores. Para o teste da Hipótese 1, definimos 2 níveis para o algoritmo usado (*code1* e *code2),* 3 níveis para a probabilidade (20%, 30% e 40%) e 3 níveis para o número de exames (30, 35 e 40). Para o teste da Hipótese 2 definimos, definimos 2 níveis para o algoritmo usado (*code1* e *code2),* 3 níveis para a probabilidade (5%, 15% e 25%), e 3 níveis para o número de exames (correspondentes aos valores 10, 20 e 30).

Para cada combinação de níveis destes 3 fatores, foram repetidas 30 experiências com *inputs* e *seeds* diferentes de forma a ter amostras independentes entre si. Os resultados dessas experiências encontram-se na pasta “Data”, sendo o ficheiro “Data1.txt” referente aos dados para a Hipótese 1 e “Data2.txt” referente aos dados para a Hipótese 2.

4. TESTE DE HIPOTESES

4.1. Formalização das hipóteses

Passando as hipóteses da meta anterior para o formato de hipótese nula (H_0) e hipótese alternativa (H_1) ficámos com o seguinte:

- Não existe diferenças nas médias devido ao fator código usado (*Code*)
  H^C_0: $\mu^C_1 = \mu^C_2$
  H^C_1: $\mu^C_1 \neq \mu^C_2$
- Não há interação entre os fatores código usado e número de exames
  H^{CE}_0: Para todo i e j, $\gamma^{CE}_{ij} = 0$
  H^{CE}_1: Existe pelo menos um i e j tal que $\gamma^{CE}_{ij} \neq 0$
- Não há interação entre os fatores código usado e probabilidade
  H^{CP}_0: Para todo i e j, $\gamma^{CP}_{ij} = 0$
  H^{CP}_1: Existe pelo menos um i e j tal que $\gamma^{CP}_{ij} \neq 0$
- Não há interação entre os fatores código usado, número de exames e probabilidade
  H^{CEP}_0: Para todo i e j, $\gamma^{CEP}_{ij} = 0$
  H^{CEP}_1: Existe pelo menos um i e j tal que $\gamma^{CEP}_{ij} \neq 0$

4.2. Pressupostos

O teste ANOVA assume que os dados têm uma distribuição normal e que a variância entre os grupos é constante. Assim sendo para poder aplicar o teste é necessário ver se os dados recolhidos cumprem esses pressupostos.

Para conferir a homogeneidade da variância recorremos ao *residuals versus fits plot* visível na Figura 1, no caso dos dados para o teste da primeira hipótese, e na Figura 2, no caso dos dados para o teste da segunda hipótese. Nesses gráficos vemos que a dispersão dos resíduos vai variando ao longo do eixo das abcissas, pelo que podemos assumir que a variância entre os grupos não é constante.

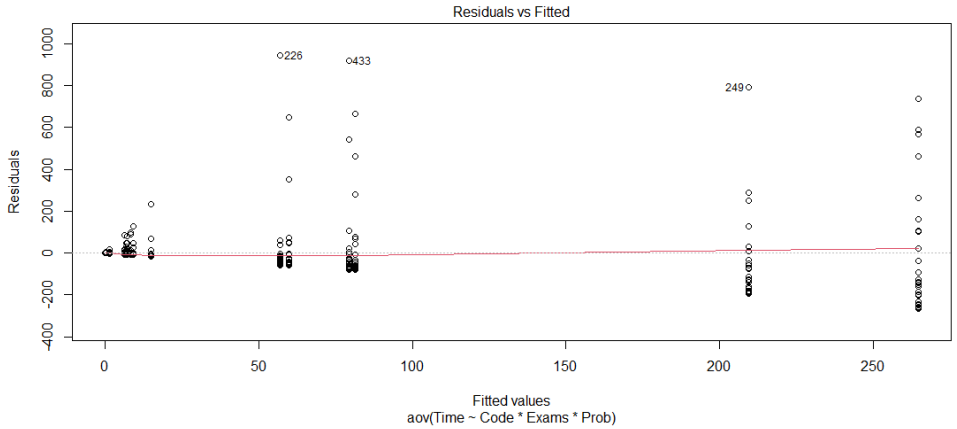

Figura 1 - Residuals-vs-fitted plot para a Primeira Hipótese

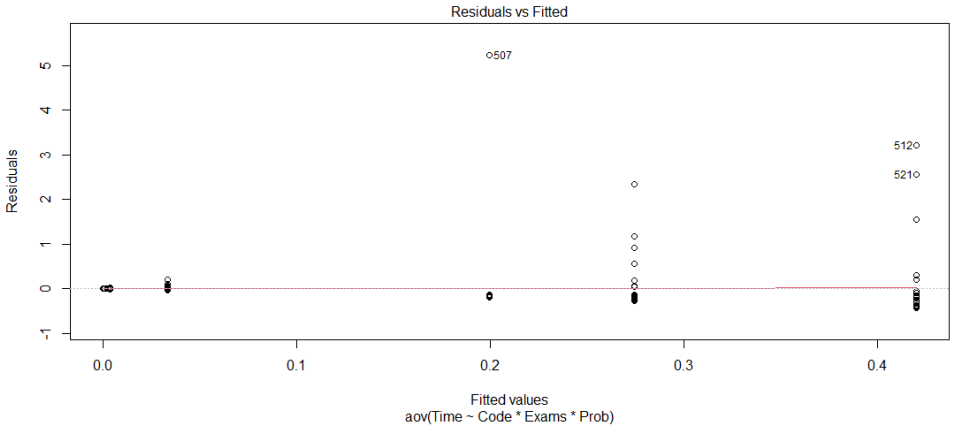

Figura 2 - Residuals-vs-fitted plot para a Segunda Hipótese

Para conferir a normalidade dos resíduos recorremos ao *normal Q-Q plot* visível na Figura 3, no caso dos dados para o teste da primeira hipótese, e na Figura 4, no caso dos dados para o teste da segunda hipótese. Nesses gráficos vemos que os pontos não seguem a reta de referência, pelo que podemos assumir que os dados não seguem uma distribuição normal.

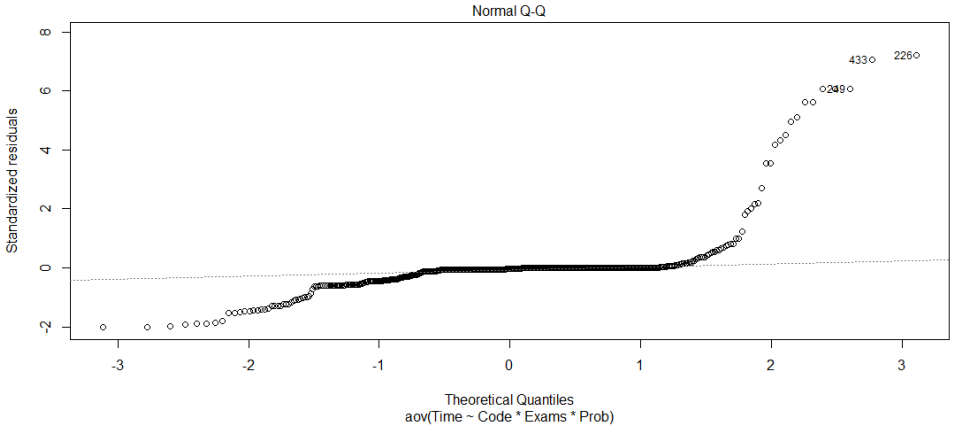

Figura 3 – Normal Q-Q plot para a Primeira Hipótese

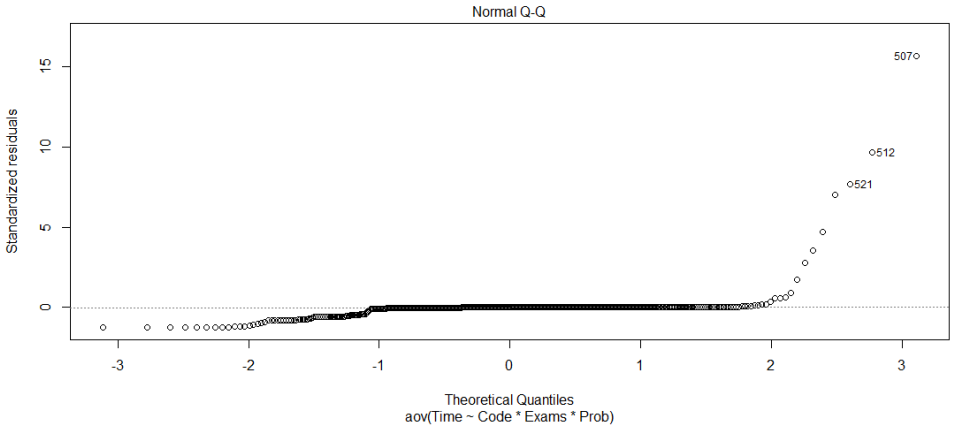

Figura 4 – Normal Q-Q plot para a Segunda Hipótese

4.3. *Randomization Test*

Uma vez que os pressupostos não foram cumpridos, tivemos que recorrer a uma alternativa não paramétrica para testar as hipóteses. A alternativa que escolhemos foi o *randomization test with unrestricted permutations* disponibilizado nas aulas teóricas, mas com as devidas adaptações para poder ser aplicado ao nosso trabalho. Os valores de *p-value* obtidos encontram se na Tabela 1.

	p-value	
	Teste da 1º hipótese	Teste da 2º hipótese
Code	0.2876	0.2556
Exams	0	0
Prob	0	0.0018
Code:Exams	0.5626	0.2622
Code:Prob	0.7104	0.7252
Exams:Prob	0	0
Code:Exams:Prob	0.9756	0.8572

Tabela 1 – Resultados do randomization test

Considerando um nível de significância de 0.05, podemos ver que as diferenças entre níveis do fator código utilizado (*Code*) não são suficientemente altas para rejeitar a H^{C}_{0} e que as interações desse fator com os restantes não são significantes o suficiente para rejeitar as H^{CE}_{0}, H^{CP}_{0} e H^{CEP}_{0}.

4.4. Análise *Post-hoc*

Apesar do *teste* não rejeitar nenhuma das hipóteses H^{C}_{0}, H^{CE}_{0}, H^{CP}_{0} e H^{CEP}_{0} decidimos realizar uma análise *Post-hoc*, onde comparamos as amostras dos dois códigos para cada combinação dos fatores probabilidade e número de exames. O intuito ao fazermos esta análise era ficar com uma ideia de como os valores de *p-value* variavam ao longo das diferentes comparações e com isso ver se o motivo das médias não se diferenciarem o bastante ser o facto dos valores escolhidos para os níveis da probabilidade e número de exames não serem suficientemente altos ou suficientemente baixos.

Para fazer as comparações usamos a versão *unpaired two-sample* do *randomization test* (também disponibilizada nas aulas teóricas), mas com uma pequena alteração para as comparações serem one-tailed em vez de two-tailed, ou seja, vermos se a média de uma amostra é maior/menor que outra em vez de vermos se estas são iguais ou diferentes.

Os resultados dessas comparações encontram-se na Tabela 2, no da Hipótese 1, e na Tabela 3, no caso da Hipótese 2.

Nas comparações da Tabela 2 a hipótese testada estava no seguinte formato:

H_0: $\mu_{code1} - \mu_{code2} \geq 0$

H_1: $\mu_{code1} - \mu_{code2} < 0$

		Exams		
		30	35	40
Prob	20	0.1768	0.0392	0.2574
	30	0.4722	0.6124	0.3108
	40	0.4156	0.3452	0.2482

Tabela 2 – Post-hoc para os dados da primeira hipótese

Nas comparações da Tabela 2 a hipótese testada estava no seguinte formato:

H_0: $\mu_{code1} - \mu_{code2} \leq 0$

H_1: $\mu_{code1} - \mu_{code2} > 0$

		Exams		
		10	20	30
Prob	5	*0*	0.502	*0.1484*
	15	*0*	0.5902	0.5476
	25	0.2768	0.4786	0.7714

Tabela 3 – Post-hoc para os dados da segunda hipótese

Olhando para os resultados obtidos podemos ver que para certas combinações de níveis ficamos mais próximos de rejeitar a H_0, mas em geral os valores dos p-values ao longo das diferentes comparações não seguem um padrão específico que nos permita generalizar o comportamento dos dois códigos.

Um aspeto que achamos estranho depois de fazer análise foi o facto de termos duas comparações onde a H_0 é rejeitada apesar do teste anterior ter indicado que não havia diferenças nas médias devido ao fator código usado.

5. CONCLUSÃO

Com esta última meta do estudo sobre a performance de dois algoritmos de *backtracking* aleatório, testámos duas hipóteses pré-registadas na meta anterior, sendo estas: Hipótese 1 - "O *code1* é mais rápido que o *code2* para números de exame maiores e probabilidades altas", e a Hipótese 2 - "O *code2* é mais rápido que o *code1* para números de exame menores e probabilidades baixas".

Foi decidido utilizar o *randomization test* visto que os dados não cumpriam os pressupostos do ANOVA e este revelou que não havia diferenças nas médias devido ao fator código usado.

Por fim, realizámos a análise *Post-hoc* para ver mais detalhadamente os valores de p-value nas comparações dos dois algoritmos em cada combinação de níveis e observámos que os valores obtidos dessas análises não seguiam nenhum padrão que nos permitisse deduzir algo para valores de probabilidades e exames maiores ou menores.

Em síntese, os resultados que obtivemos levam-nos a concluir que não há uma diferença significativa na performance dos dois algoritmos. No entanto achamos que as nossas conclusões poderiam ter sido diferentes caso tivéssemos realizado uma análise exploratória mais vasta, o que nos daria a possibilidade de registar outro tipo de hipóteses mais plausíveis.

6. REFERÊNCIAS

M. Anderson and C.Ter Braak, Permutation tests for multi-factorial analysis of variance, Journal of Statistical Computation and Simulation 2003, Vol. 73(2), pp. 85-113