PCC170 - Projeto e Análise de Experimentos Computacionais

Marco Antonio M. Carvalho

Departamento de Computação
Instituto de Ciências Exatas e Biológicas
Universidade Federal de Ouro Preto

Conteúdo

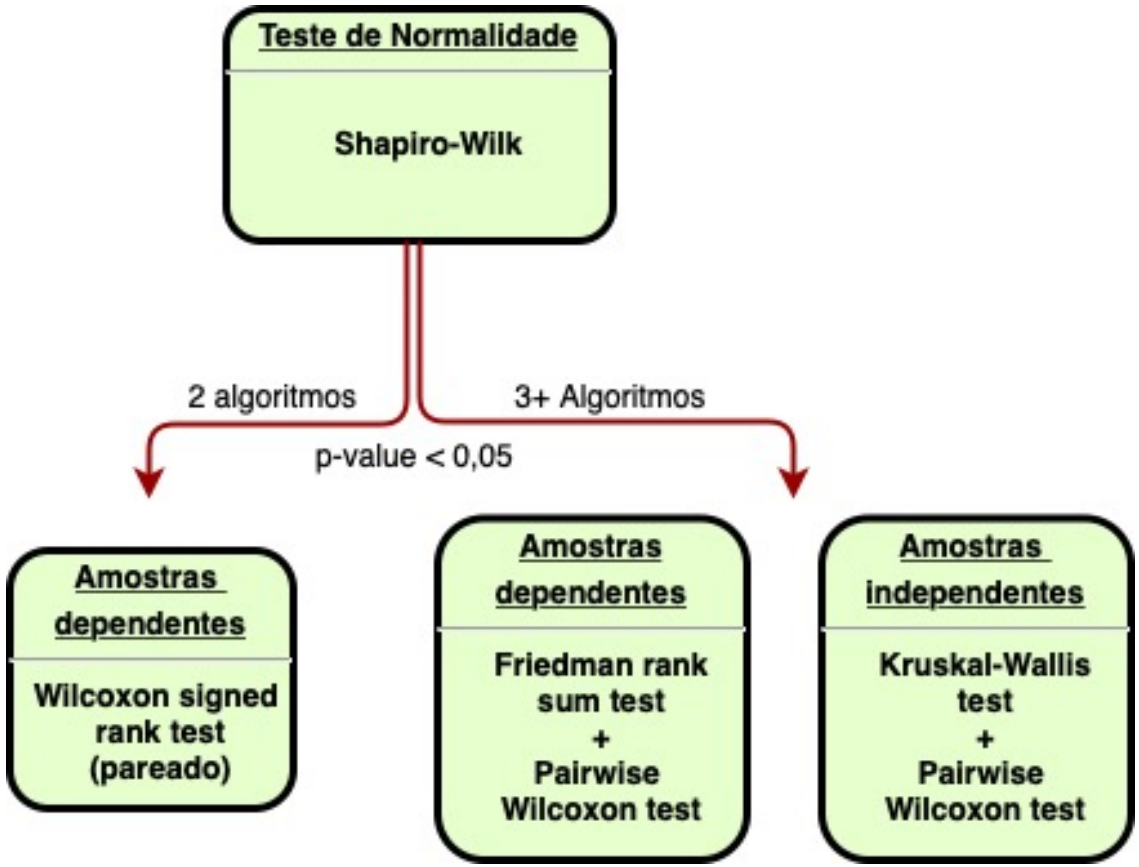
Fonte

Este material é parcialmente baseado no conteúdo de

- Alboukadel Kassambara. *Statistical tools for high-throughput analysis*. 2022. Disponível em `https://bityli.com/ixKGg`
- Chi Yau. *R tutorial: An R introduction to statistics*. 2022. Disponível em `https://bityli.com/qSzEd`
- Zach Bobbitt. *Statology: How to Perform the Friedman Test in R*. 2020. Disponível em `https://bityli.com/wwkFo`

Licença

Testes não paramétricos



Fluxograma de escolha de testes não paramétricos.

Introdução

Um método estatístico é chamado não paramétrico se não faz suposições sobre a distribuição da população ou o tamanho da amostra.

Em geral, as conclusões tiradas de métodos não paramétricos não são tão poderosas quanto os paramétricos.

No entanto, como os métodos não paramétricos fazem menos suposições, eles são mais flexíveis, mais robustos e aplicáveis a dados não quantitativos.

Wilcoxon signed rank test

O *Wilcoxon Signed Rank Test* é um método para comparação de duas amostras, no sentido de verificar se existem diferenças significativas entre os seus valores.

Por ser um método não paramétrico, as distribuições consideradas não devem ser normais e, no caso de amostras pareadas, estas são consideradas **dependentes**.

A hipótese nula é a de que não há diferença entre as médias das duas populações consideradas.

Como executar o teste

No *R*, crie uma série de dados para cada algoritmo (*Método A* e *Método B*) e atribua os valores de cada um.

O parâmetro *less* testa se a média do primeiro método é menor do que a do segundo.

Em outras palavras, está sendo testado se os dados são significativamente diferentes e se os dados do *Método A* são melhores do que os dados do *Método B*, supondo um problema de minimização.

Testes não paramétricos

```
> MetodoA <- c(214, 159, 13, 356, 789, 123)
> MetodoB <- c(159, 135, 123, 543, 12, 345)
> wilcox.test(MetodoA, MetodoB, paired=TRUE, alternative="
    less")
```

Testes não paramétricos

```
Wilcoxon signed rank test

data:  MetodoA and MetodoB
V = 0, p-value = 0.001953
alternative hypothesis: true location shift is
not equal to 0
```

Análise

São retornados um valor V e o *p-value*.

Se o *p-value* for menor do que um valor crítico dado pelo nível de significância α (normalmente 0,05), então o pressuposto de inexistência de diferença significativa é rejeitado no nível de significância α.

Ou seja, há evidência de que os dados testados diferem entre si.

O valor de V corresponde à soma dos *ranks* positivos associadas às diferenças com sinal positivo, ou seja, valores de V baixos indicam que o primeiro algoritmo é melhor.

Embora seja reportado, o valor de V não é interpretado diretamente.

Testes não paramétricos

[redacted] Consequently, we apply the non-parametric Wilcoxon signed-rank test [34] in order to investigate whether there is a significant difference between the solution values generated by the ALNS and those generated by the BRKGA. The test indicates that there is statistically significant evidence of a difference in the methods' results ($V = 0$ and p-value $= 0.001602$, for a significance level of 0.05). The value of V also states that the ALNS results are equal or smaller than those of the BRKGA. Therefore, it proves that the proposed ALNS is a competitive alternative method for the GMLP solution.

Fonte: Santos, Vinicius Gandra Martins, and Marco Antonio Moreira de Carvalho. *Adaptive large neighborhood search applied to the design of electronic circuits*. Applied Soft Computing 73 (2018): 14-23.

Testes não paramétricos

[redacted] Subsequently, given that the populations are not normally distributed, the non-parametric Wilcoxon signed-rank test (Rey & Neuhäuser, 2011) was applied to investigate whether or not there is a significant difference between the solution values generated by ALNS and VFS. For the Grid set, the test indicates ($V = 0$ and p-value equal to $1.923 \cdot 10^{-5}$) that there are statistical differences between the two methods. The value $V = 0$ also indicates that all the ALNS solution values are less than or equal to those generated by VFS. For the HB instance set, the test indicates ($V = 28.5$, p-value equal to $5.714 \cdot 10^{-5}$) that ALNS is significantly different from VFS. However, three VFS solution values are better than those generated by ALNS, and 24 ALNS solution values are better than those generated by VFS. Therefore, it may be concluded that ALNS performed better than VFS on the HB set.

Fonte: Santos, Vinícius Gandra Martins, and Marco Antonio Moreira de Carvalho. *Tailored heuristics in adaptive large neighborhood search applied to the cutwidth minimization problem*. European Journal of Operational Research 289.3 (2021): 1056-1066.

Friedman rank sum test

O *Friedman rank sum test* é um teste de comparações múltiplas de populações não consideradas normalmente distribuídas.

No *Friedman rank sum test*, considera-se que as amostras são **dependentes**, como os valores de solução de diferentes métodos para as mesmas instâncias.

A hipótese nula é de que não há diferença entre as populações.

Preparando os dados

No *R*, crie uma série de dados *data* em formato de matriz com três colunas: **instancia**, **metodo** e **resultado**.

Os dados da coluna **resultado** devem ser agrupados por instância.

No exemplo a seguir, há cinco métodos (de *A* a *E*) e os resultados de cada um deles para para sete instâncias.

Testes não paramétricos

```
1  > data <- data.frame(instancia = rep(1:7, each=5),
2                       metodo = rep(c("A", "B", "C", "D", "E"),
                             times=7),
3                       resultado = c(33.08, 38.69, 38.64,
                            33.66, 39.20, 60.83, 59.57, 60.33,
                            56.57, 60.35, 66.47, 62.65, 64.31,
                            59.48, 67.55, 69.35, 64.23, 66.23,
                            66.37, 70.93, 70.59, 65.61, 71.54,
                            73.35, 79.60, 72.90, 66.17, 71.91,
                            68.08, 84.33, 64.37, 60.23, 63.67,
                            60.22, 70.15))
4
5  > data
6
7  > boxplot(data$resultado~data$metodo)
```

Testes não paramétricos

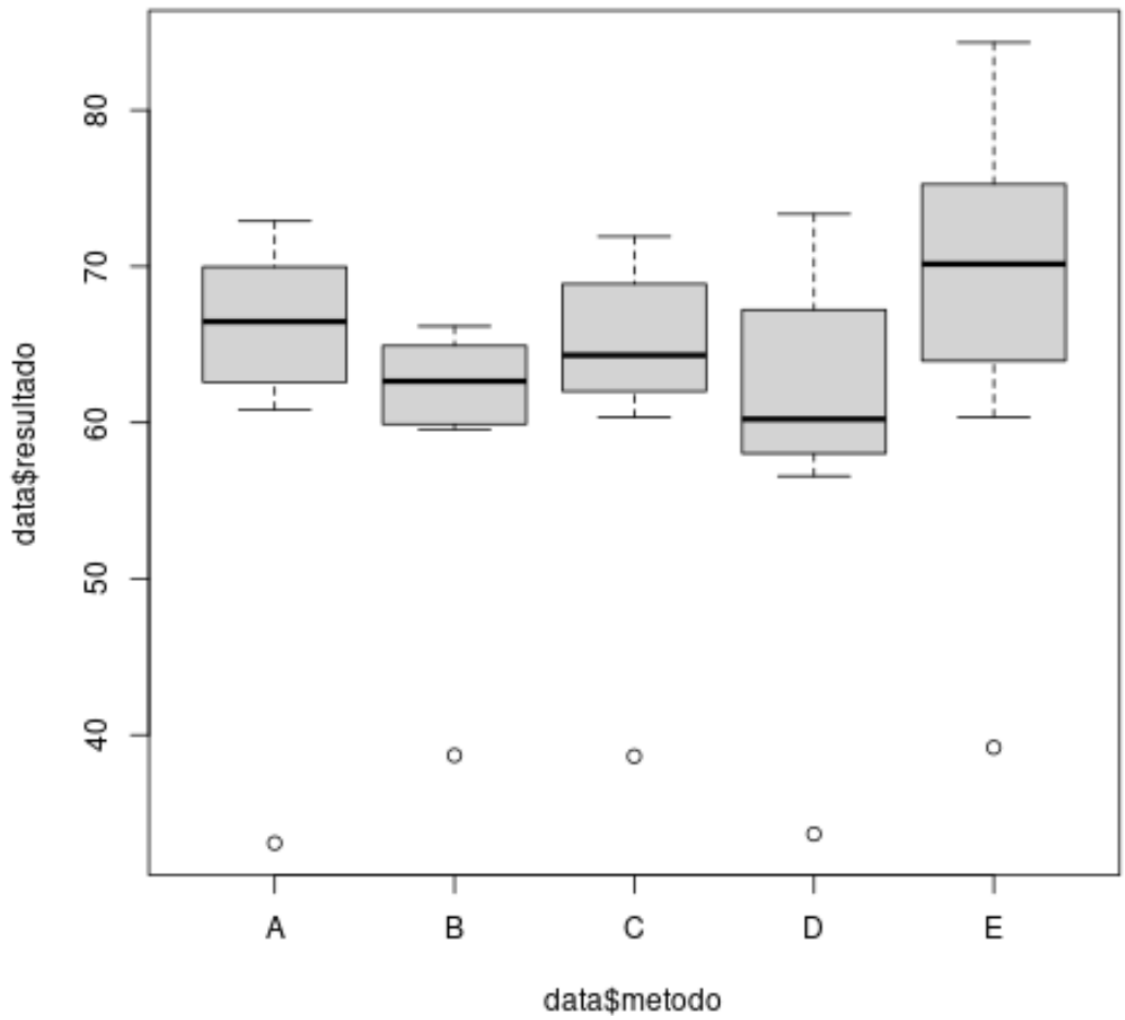
1	1	A	33.08
2	1	B	38.69
3	1	C	38.64
4	1	D	33.66
5	1	E	39.20
6	2	A	60.83
7	2	B	59.57
8	2	C	60.33
9	2	D	56.57
10	2	E	60.35
11	3	A	66.47
12	3	B	62.65
13	3	C	64.31
14	3	D	59.48
15	3	E	67.55

16	4	A	69.35
17	4	B	64.23
18	4	C	66.23
19	4	D	66.37
20	4	E	70.93
21	5	A	70.59
22	5	B	65.61
23	5	C	71.54
24	5	D	73.35
25	5	E	79.60
26	6	A	72.90
27	6	B	66.17
28	6	C	71.91
29	6	D	68.08
30	6	E	84.33

Testes não paramétricos

31	7	A	64.37
32	7	B	60.23
33	7	C	63.67
34	7	D	60.22
35	7	E	70.15

Testes não paramétricos



Como executar o teste

O *Friedman rank sum test* é exemplificado a seguir, utilizando os dados preparados.

```
1 > friedman.test(y=data$resultado, groups=data$metodo, blocks
    =data$instancia)
```

Análise

Como resultado é reportado o *p-value*, entre outros valores.

Se o *p-value* for menor do que o nível de significância α (normalmente 0,05) é possível rejeitar a hipótese nula de que os resultados são iguais para todos os métodos.

Caso contrário, não é possível afirmar que as populações são não idênticas.

Caso haja diferença significativa, realizamos um *pairwise Wilcoxon test* para descobrir onde ela reside.

Testes não paramétricos

```
        Friedman rank sum test

data:  data$resultado, data$metodo and data$instancia
Friedman chi-squared = 16.686, df = 4, p-value = 0.002224
```

Pairwise Wilcoxon test

A partir da saída do *Friedman rank sum test*, sabemos que há uma diferença significativa entre os grupos, mas não sabemos quais pares de grupos são diferentes.

É possível usar o *pairwise Wilcoxon test* para calcular comparações de pares entre níveis de grupo com correções para vários testes.

Como executar o teste

O *Friedman rank sum test* seguido do *pairwise Wilcoxon test* é exemplificado a seguir.

```
1 > friedman.test(y=data$resultado, groups=data$metodo,
      blocks=data$instancia)
2 > pairwise.wilcox.test(data$resultado, data$metodo, p.adj =
      "bonf")
```

Análise

Como resultado é reportada uma matriz triangular dos *p-values* entre todos os grupos comparados.

A análise é realizada para cada combinação linha $\times$ coluna.

Um *p-value* menor do que o nível de significância α (normalmente 0,05) nos permite concluir que há diferença significativa entre as populações.

Testes não paramétricos

```
  Pairwise comparisons using Wilcoxon rank sum exact test

data:  data$resultado and data$metodo

  A    B    C    D
B 1.00 -    -    -
C 1.00 1.00 -    -
D 1.00 1.00 1.00 -
E 1.00 0.03 1.00 1.00

P value adjustment method: bonferroni
```

Testes não paramétricos

Como reportar

The Friedman rank sum test was applied to evaluate if there is significant difference among the methods. The p-value of X.0e- X indicates that there is such a difference. The pairwise Wilcoxon test was applied to analyze where this difference lies. According to the test, method B significantly differs from method E. There is not a significant difference among the other combinations of methods.

Kruskal-Wallis test

O *Kruskal-Wallis test* é um teste de comparações múltiplas de populações não consideradas normalmente distribuídas.

O *Kruskal-Wallis test* possui o mesmo objetivo do *Friedman rank sum test*, porém, a diferença é que no *Kruskal-Wallis test* as amostras são **independentes**.

Este teste deve ser utilizado quando os dados **não são** pareados por instância, como indicadores de desempenho de diferentes componentes de um mesmo algoritmo, e.g., operadores de busca local.

A hipótese nula é de que as populações possuem distribuições idênticas.

Como executar o teste

No *R*, crie uma série de dados *data* em formato de matriz com três colunas: **instancia**, **metodo** e **resultado**.

Os dados da coluna **resultado** podem ser agrupados por instância.

No exemplo a seguir, novamente há cinco métodos (de *A* a *E*) e os resultados para sete instâncias, porém, os valores diferem do exemplo anterior.

Testes não paramétricos

```
1 > kruskal.test(data$resultado~data$metodo)
```

Análise

Como resultado é reportado o *p-value*, entre outros valores.

Se o *p-value* for menor do que o nível de significância α (normalmente 0,05) é possível concluir que há diferença significativa entre as populações.

Caso contrário, não é possível afirmar que as populações são não idênticas.

Caso haja diferença significativa, realizamos um *pairwise Wilcoxon test* para descobrir onde ela reside.

Testes não paramétricos

```
Kruskal-Wallis rank sum test

data:  resultados by metodos
Kruskal-Wallis chi-squared = 4.8136, df = 4,
p-value = 0.307
```

Como executar o teste

Usando os mesmos dados preparados anteriormente, o *Kruskal-Wallis test* seguido do *pairwise Wilcoxon test* é exemplificado a seguir.

Note que este é apenas a ilustração do uso dos dois testes. Como o exemplo anterior gerou *p-value* $\geq 0{,}05$, não seria necessário rodar o segundo teste.

```
1 > kruskal.test(resultados ~ metodos, data=comparacao)
2 > pairwise.wilcox.test(resultados, metodos, p.adjust="
      bonferroni")
```

Análise

Como resultado é reportada uma matriz triangular dos *p-values* entre todos os grupos comparados.

A análise é realizada para cada combinação linha $\times$ coluna.

Um *p-value* menor do que o nível de significância α (normalmente 0,05) nos permite concluir que há diferença significativa entre as populações.

Testes não paramétricos

```
  Pairwise comparisons using Wilcoxon rank sum exact test

data:  resultados and metodos

  A     B     C     D
B 0.005 -     -     -
C 0.011 0.054 -     -
D 0.150 0.173 0.002 -
E 0.100 0.730 0.001 0.950

P value adjustment method: bonferroni
```

Testes não paramétricos

The scores are well-distributed among the destroy neighborhoods. To further compare the neighborhoods we apply the Kruskal–Wallis test [35], whose result does not indicate any stochastic dominance (p-value = 0.307). Thus, despite the fact that the RShaw heuristic achieves the best scores in almost all sets of instances, the differences among the scores of the heuristics are not statistically significant.

There is a greater divergence of scores among the repair neighborhoods. The Kruskal–Wallis test indicates a stochastic dominance on the scores (p-value = 0.0001537). Consequently, we apply a pairwise t-test [36] in order to analyze where this dominance occurs. The test indicates that there is a significant difference between ICL and all the other neighborhoods, and also a difference between IRand and IBest (p-value = 0.0041). The test also indicates that there is no significant difference between IBest and IReg (p-value = 1.0) and between IRand and IReg (p-value = 0.0656).

Fonte: Santos, Vinicius Gandra Martins, and Marco Antonio Moreira de Carvalho. *Adaptive large neighborhood search applied to the design of electronic circuits*. Applied Soft Computing 73 (2018): 14-23.

Leitura recomendada

- Chi Yau. *Wilcoxon Signed-Rank Test*. 2022. Disponível em `https://bityli.com/MzgbV`.
- Chi Yau. *Kruskal-Wallis Test*. 2022. Disponível em `https://bityli.com/Wwuax`.
- Rey, D., & Neuhäuser, M. (2011). *Wilcoxon-signed-rank test*. In International encyclopedia of statistical science (pp. 1658-1659). Springer.
- W.H. Kruskal, W.A. Wallis. *Use of ranks in one-criterion variance analysis*, J. Amer. Statist. Assoc. 47 (260) (1952) 583-621.

Dúvidas?