Métodos não paramétricos de tipo ANOVA

Métodos não paramétricos de tipo ANOVA

- São uma alternativa quando há violações graves dos pressupostos dos modelos ANOVA clássicos:
  -normalidade dos erros;
  -homogeneidade de variâncias.

- **Mantém-se a necessidade de observações independentes.**

- Quando os pressupostos dos modelos ANOVA clássicos são válidos, os métodos não paramétricos têm menor potência.

Teste de Kruskal-Wallis

[Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. **47**, 583–621 and errata, ibid. **48**, 907–911 (1952).]

Variante não paramétrica de uma ANOVA a 1 factor (delineamento totalmente casualizado), com k níveis

- Nota: para o caso $k = 2$, o teste não paramétrico de Wilcoxon-Mann-Whitney compara hipóteses análogas às do teste de Kruskal-Wallis.

Notação

- 1 factor com k níveis

- Cada observação da variável resposta é identificada por 2 índices, Y_{ij}, onde:
  i indica o nível i do factor ($i = 1, \dots, k$),
  j indica a repetição correspondente à observação no nível i do factor ($j = 1, \dots, n_i$)

- Os valores observados são substituídos pelas suas ordens sobre todas as $n = \sum_{i=1}^{k} n_i$ observações.

- As ordens serão números naturais $1, 2, \dots, n$; o valor observado mais baixo é substituído por 1, o maior por n.
  A solução usual para o teste de Kruskal-Wallis com valores repetidos é atribuir a cada grupo de observações repetidas uma ordem comum dada pela média das ordens que caberiam ao conjunto dessas observações.

- Para os números naturais de 1 a n:

  soma, $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$

  média, $\frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}$

  variância, $\frac{1}{n-1}\sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2 = \frac{n(n+1)}{12}$

Teste de Kruskal-Wallis

Hipóteses

Seja $F_i\,(y)$ a função distribuição cumulativa das observações correspondentes ao nível i do factor.

- Hipótese nula:

$$H_0: F_i\,(y) = F(y) \quad (\forall y \in \mathbb{R}; \forall i = 1, \dots, k)$$

(Os k níveis do factor possuem idêntica distribuição; ou, de forma informal, os valores observados em cada nível são globalmente semelhantes, ou, as médias das ordens em cada nível não diferem)

- Hipótese alternativa:

$$H_1: \exists i, j\ :\ F_i\,(y) < F_j\,(y) \qquad (\forall y \in \mathbb{R})$$

(Pelo menos uma das distribuições dá valores mais elevados do que os das outras; ou, de forma informal, há pelo menos um nível do factor com valores globalmente maiores (ou menores) do que outros)

O teste diz-se não paramétrico porque não é importante conhecer a natureza exacta das distribuições $F_i\,(y)$

Teste de Kruskal-Wallis (cont.)

Admitamos que não há valores repetidos: os r_{ij} tomam os valores naturais de 1 a $n = \sum_{i=1}^{k} n_i$

Estatística do Teste de Kruskal-Wallis

$$H = \frac{\sum_{i=1}^{k} n_i (\bar{R}_i - \bar{R}_{..})^2}{\sigma_R^2} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

- k, número de níveis do factor
- $n = \sum_{i=1}^{k} n_i$
- n_i , número de observações no nível i do factor.
- $R_i = \sum_{j=1}^{n_i} r_{ij}$, a soma das ordens das n_i observações do nível i
- $\bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}$, média das ordens das n_i observações do nível i
- $\bar{R}_{..} = \frac{n+1}{2}$, média de todas as ordens
- $\sigma_R^2 = \frac{n(n+1)}{12}$, variância das ordens r_{ij}

Quanto maior $\sum_{i=1}^{k} n_i (\bar{R}_i - \bar{R}_{..})^2$, mais duvidosa será H_0

Nota: quando há valores repetidos, a estatística do teste é adaptada. Por exemplo, encontra-se explicada em Hollander, M., Wolfe, D.A., Chicken, E., 2014. Nonparametric Statistical Methods, third edition. John Wiley & Sons, New York.

Teste de Kruskal-Wallis (cont.)

Distribuição assintótica da estatística H

É possível determinar a distribuição exacta de H, admitindo que qualquer possível ordenação das n ordens é igualmente provável. No entanto, a distribuição exacta da estatística não é fácil de calcular. Conover (1980) apresenta um extrato da tabela da distribuição exacta da estatística de Kruskal-Wallis quando k=3.

Mas, para $n_i s$ grandes, e mesmo para $n_i s$ modestos ($n_i \geq 5$), prova-se que, sob H_0:

$$H \sim \chi^2_{k-1}$$

Região crítica: unilateral direita

Para um nível de significância α, rejeita-se H_0 se $H_{calc} > \chi^2_{\alpha(k-1)}$

No R,

\> kruskal.test (…)

Exercício 1: Num estudo sobre clones da variedade de videira Grenache, estudaram-se 3 clones relativamente ao rendimento (kg/planta) num ensaio com um delineamento totalmente casualizado, com 6 repetições de cada clone. Após o ajustamento de uma análise de variância clássica a um factor de efeitos fixos, a análise dos resíduos evidenciou a violação dos pressupostos da homogeneidade de variâncias e normalidade.

Um analista sugere a realização de um teste de Kruskal-Wallis. Os dados da experiência encontram-se seguidamente descritos.

a) Enuncie as hipóteses a testar no contexto deste estudo.
b) Calcule a estatística do teste de Kruskal-Wallis para a amostra observada.
c) Para um nível de significância (α) de 0.05, qual a conclusão que tira sobre o estudo efectuado?

clone	rendimento	**ordem**
GR1	12.725	**16**
GR1	16.550	**18**
GR1	14.432	**17**
GR1	6.538	**8**
GR1	10.775	**14**
GR1	11.808	**15**
GR2	9.025	**13**
GR2	7.575	**11**
GR2	7.317	**10**
GR2	7.188	**9**
GR2	8.717	**12**
GR2	4.888	**7**
GR3	2.317	**4**
GR3	2.088	**3**
GR3	3.875	**6**
GR3	2.045	**2**
GR3	2.475	**5**
GR3	1.588	**1**

Exercício 2: Num estudo sobre variedades tradicionais de alho, estudaram-se 4 variedades relativamente a características do bolbo. Concretamente, fez-se um ensaio com um delineamento totalmente casualizado com 5 repetições de cada variedade e em cada parcela avaliou-se o número de dentes presentes em 10 bolbos. Os valores obtidos em cada parcela, bem como o resultado obtido, no R, com o comando **kruskal.test(dentes~variedade, data=alhobd),** são seguidamente apresentados:

variedade	Nº dentes em 10 bolbos	ordem
V1	59	1
V1	65	5
V1	72	8
V1	81	11
V1	89	16
V2	61	2
V2	62	3
V2	73	9
V2	82	12
V2	90	17
V3	75	10
V3	88	15
V3	94	18
V3	96	19
V3	98	20
V4	63	4
V4	66	6
V4	67	7
V4	83	13
V4	87	14

```
> kruskal.test(dentes~variedade, data=alhobd)

        Kruskal-Wallis rank sum test

data:  dentes by variedade
Kruskal-Wallis chi-squared = 6.6571, df = 3, p-value = 0.08367
```

a) Enuncie as hipóteses a testar no contexto deste estudo.

b) Verifique que o valor da estatística do teste de Kruskal-Wallis para a amostra observada é 6.6571.

c) O que conclui sobre o estudo efectuado?

Teste de Friedman

[Friedman , M. (1937). The use of ranks to avoid the assumption of normality in the analysis of variance. *J. Amer. Statist. Assoc.* 32 : 675 – 701.]

- **Variante não paramétrica de uma ANOVA a 2 factores factorial , sem repetições nas células (sem interação).**

- **Aplica-se quando se sabe que um segundo factor (designado bloco) afecta a variável resposta e se deseja controlar os seus efeitos (por exemplo, terrenos, provadores).**

- Cada observação da variável resposta é identificada por 2 índices, Y_{ij}, onde:
  i indica o nível i do primeiro factor ($i = 1, \dots, a$),
  j indica o nível j do segundo factor (bloco) ($j = 1, \dots, b$)

Teste de Friedman

Hipóteses

- Hipótese nula:

H_0: Dado um qualquer bloco j, as distribuições dos valores de y são sempre iguais:

$$F_{ij}\,(y) = F_j\,(y) \quad (\forall y \in \mathbb{R}; \forall i = 1, \dots, a)$$

(de forma informal, os valores observados em cada nível do primeiro factor são globalmente semelhantes, tendo em conta eventuais diferenças em diferentes blocos; ou, a média das ordens em cada nível do primeiro factor não difere, tendo em conta eventuais diferenças em diferentes blocos)

- Hipótese alternativa:

H_1: Há níveis do primeiro factor com valores globalmente maiores (ou menores) do que outros, tendo em conta eventuais diferenças em diferentes blocos;

Teste de Friedman (cont.)

Admitamos que não há valores repetidos.

Estatística do Teste de Friedman

$$S = \frac{\sum_{i=1}^{a} b(\bar{R}_i - \bar{R}_{..})^2}{\sigma_B^2} = \frac{12}{ab(a+1)} \sum_{i=1}^{a} R_i^2 - 3b(a+1)$$

- a, número de níveis do primeiro factor
- b, número de níveis do segundo factor (bloco)
- $R_i = \sum_{j=1}^{b} r_{ij}$, a soma das ordens no nível i do primeiro factor
- $\bar{R}_{..} = \frac{a+1}{2}$, a ordem media global
- $\bar{R}_i$, a ordem média no nível i do primeiro factor
- $\sigma_B^2 = \frac{a(a+1)}{12}$, a variância média das ordens em cada bloco

Nota: quando há valores repetidos no bloco, atribui-se a cada grupo de observações repetidas uma ordem comum dada pela média das ordens que caberiam ao conjunto dessas observações e a estatística do teste é adaptada. Por exemplo, encontra-se explicada em Hollander, M., Wolfe, D.A., Chicken, E., 2014. Nonparametric Statistical Methods, third edition. John Wiley & Sons, New York.

Teste de Friedman (cont.)

Distribuição assintótica da estatística de Friedman

É possível determinar a distribuição exacta de S, mas não é fácil de obter. Por isso, é usual recorrer à seguinte distribuição assintótica:

$$S \sim \chi^2_{a-1}$$

Região crítica: unilateral direita

Para um nível de significância α, rejeita-se H_0 se $S_{calc} > \chi^2_{\alpha(a-1)}$

No R,

\> friedman.test (…)

Nota: Conover (1999) sugere uma estatística alternativa, $S_2 = \frac{(b-1)S}{b(a-1)S} \sim F_{a-1,(b-1)(a-1)}$. Contudo, os *softwares* aplicam correntemente a estatística S.

Exercício 3: Num estudo sobre clones da variedade de videira Aragonez, estudaram-se 5 clones relativamente ao número de cachos por planta num ensaio com um delineamento em blocos completos casualizados, com 6 blocos. Após o ajustamento de uma análise de variância clássica a dois factores de efeitos fixos sem interação, a análise dos resíduos evidenciou a violação do pressuposto da normalidade dos erros aleatórios. Como alternativa, é sugerida a realização de um teste de Friedman. Os dados da experiência, as respetivas ordenações do clone por bloco, bem como a soma das ordenações do clone por bloco ($R_i, para\ i = 1, \dots 5$), encontram-se seguidamente descritos.

a) Enuncie as hipóteses a testar no contexto deste estudo.

b) Calcule a estatística do teste de Friedman para a amostra observada.

c) Para um nível de significância (α) de 0.05, qual a conclusão que tira sobre o estudo efetuado?

	Número de cachos por planta					Ordem				
	Clone					Clone				
Bloco	RZ1	RZ2	RZ3	RZ4	RZ5	RZ1	RZ2	RZ3	RZ4	RZ5
B1	25	35	44	50	21	2	3	4	5	1
B2	14	46	49	58	6	2	3	4	5	1
B3	30	15	72	74	39	2	1	4	5	3
B4	48	40	63	78	17	3	2	4	5	1
B5	62	48	14	52	7	5	3	2	4	1
B6	44	35	34	62	26	4	3	2	5	1
						R_1=18	R_2=15	R_3=20	R_4=29	R_5=8

Exercício 3 no R

```
> rzcachos<-read.table("rzcachos.txt", header=T)

> friedman.test(y=rzcachos$Ncachos, groups=rzcachos$Clone, blocks=rzcachos$Bloco)
```

ou

```
➢friedman.test(Ncachos ~ Clone | Bloco, data = rzcachos)

Friedman rank sum test data:

Ncachos and Clone and Bloco

Friedman chi-squared = 15.6, df = 4, p-value = 0.003606
```

rzcachos.txt

Clone	Bloco	Ncachos
RZ1	B1	25
RZ1	B2	14
RZ1	B3	30
RZ1	B4	48
RZ1	B5	62
RZ1	B6	44
RZ2	B1	35
RZ2	B2	46
RZ2	B3	15
RZ2	B4	40
RZ2	B5	48
RZ2	B6	35
RZ3	B1	44
RZ3	B2	49
RZ3	B3	72
RZ3	B4	63
RZ3	B5	14
RZ3	B6	34
RZ4	B1	50
RZ4	B2	58
RZ4	B3	74
RZ4	B4	78
RZ4	B5	52
RZ4	B6	62
RZ5	B1	21
RZ5	B2	6
RZ5	B3	39
RZ5	B4	17
RZ5	B5	7
RZ5	B6	26

Exercício 4: Num ensaio de clones de Antão Vaz instalado de acordo com um delineamento em blocos completos casualizados, foram avaliados 7 clones de Antão Vaz em 3 blocos quanto à tolerância ao míldio, medindo-se por unidade experimental a proporção média de cachos afectados por planta. Os dados da experiência e as respetivas ordenações do clone por bloco, apresentam-se seguidamente.

	Proporção dos cachos afectados							Ordem						
	Clone							Clone						
	AN1	AN2	AN3	AN4	AN5	AN6	AN7	AN1	AN2	AN3	AN4	AN5	AN6	AN7
B1	0.375	0.917	0.125	0.250	0.583	0.750	0.333	4	7	1	2	5	6	3
B2	0.375	0.750	0.167	0.250	0.583	0.417	0.500	3	7	1	2	6	4	5
B3	0.500	0.417	0.167	0.125	0.250	0.750	0.333	6	5	2	1	3	7	4

Admitindo válida a distribuição assintótica da estatística de Friedman, o que conclui sobre o estudo efectuado ao nível $\alpha = 0.05$? Descreva em pormenor o teste realizado.

Exercício 4 no R

```
> mildio<-read.table("mildio.txt", header=T)

> friedman.test(INF ~ Clone | Bloco, data = mildio)

        Friedman rank sum test

data:  INF and Clone and Bloco
Friedman chi-squared = 13.714, df = 6, p-value = 0.033
```

Exercício 5: Num estudo sobre vinhos da casta Antão Vaz, foram avaliados 4 diferentes vinhos por 7 provadores, que fizeram a apreciação global do vinho numa escala de 1 a 4. Os valores obtidos para a apreciação global dos 4 vinhos pelos 7 provadores, bem como o resultado obtido no R, são seguidamente apresentados:

	Pontuação à apreciação global Vinho				Ordenação			
Provador	Vinho1	Vinho2	Vinho3	Vinho4	Vinho1	Vinho2	Vinho3	Vinho4
Provador 1	1	3	3	2	1	3.5	3.5	2
Provador 2	1	2	3	2	1	2.5	4	2.5
Provador 3	1	2	2	1	1.5	3.5	3.5	1.5
Provador 4	1	3	3	2	1	3.5	3.5	2
Provador 5	1	3	1	2	1.5	4	1.5	3
Provador 6	2	4	4	4	1	3	3	3
Provador 7	3	3	2	2	3.5	3.5	1.5	1.5

a) Enuncie as hipóteses a testar no contexto deste estudo.

b) Para um nível de significância (α) de 0.05, o que conclui sobre o estudo efetuado? E para um nível de significância (α) de 0.01?

```
> provaan<-read.table("provaan.txt", header=T)
> friedman.test(y=provaan$Score, groups=provaan$Vinho, blocks=provaan$Provador)

        Friedman rank sum test
data:  provaan$Score, provaan$Vinho and provaan$Provador

Friedman chi-squared = 10.138, df = 3, p-value = 0.01743
```