Taxas de erros de tipos I e II de alternativas não paramétricas à ANOVA com dois fatores: comparação entre delineamentos equilibrados e desequilibrados com dados discretos

Anabela Afonso
CIMA/IIFA e DMat/ECT, Universidade de Évora, Évora, Portugal,
aafonso@uevora.pt

Dulce G. Pereira
CIMA/IIFA e DMat/ECT, Universidade de Évora, Évora, Portugal,
dgsp@uevora.pt

Palavras–chave: Empates; Estatística de Wald; Testes de permutação; Transformação em ordens

Resumo: A existência de observações empatadas pode influenciar o desempenho dos procedimentos não paramétricos alternativos à ANOVA com dois fatores. Para comparamos o desempenho destas alternativas consideramos delineamentos equilibrados e desequilibrados, e que os dados são provenientes de distribuições discretas. Os nossos resultados mostram que o desempenho dos testes é afectado pelo tipo de delineamento (equilibrado ou desequilibrado), pelos efeitos presentes no modelo e pelo tamanho dos efeitos, e pela dimensão da amostra.

1 Introdução

Na análise de conjuntos de dados reais são várias as situações em que não se pode recorrer à ANOVA paramétrica devido a sérias violações dos seus pressupostos ou porque os dados são de tipo ordinal [2, 3].

Por isso, desde a segunda metade do século passado, foram propostos vários procedimentos não paramétricos alternativos à ANOVA com dois fatores.

Na literatura é possível encontrar vários trabalhos que estudam o desempenho destes métodos, onde são consideradas distribuições contínuas, com diferentes graus de assimetria e a presença de valores atípicos, e/ou com variâncias heterogéneas, tanto para delineamentos equilibrados como desequilibrados (e.g. [4, 5, 6]). Afonso e Pereira [2, 3] efetuaram estudos de simulação para analisar a probabilidade de erro de tipo I e a potência de alguns desses procedimentos alternativos considerando dados provenientes de distribuições discretas, o que propicia a existência de empates nas ordens das observações. Nestas análises foram considerados diferentes graus de dispersão e assimetria das distribuições, e apenas delineamentos equilibrados. De acordo com os vários estudos, nenhum dos procedimentos se destacou dos restantes por ter tido o melhor desempenho em todos os contextos.

Neste trabalho, estendemos a análise realizada nos trabalhos [2, 3] aos delineamentos desequilibrados para os procedimentos: transformação em ordens (RT), transformação normal inversa (INT), transformação em ordens alinhadas (ART), transformação ART combinada com INT (ART+INT), estatística L de Puri & Sen (L de PS), teste de van der Waerden, estatística de tipo Wald (WTS), estatística de tipo ANOVA (ATS) e teste de permutação de tipo Wald ($WTPS$).

2 Simulação

Para se comparar os diferentes procedimentos não paramétricos com a ANOVA paramétrica apresentamos um estudo de simulação. Neste estudo adota-se um modelo de efeitos fixos com interação:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

onde μ é a média global, α_i o efeito do nível i do fator A, $i = 1,...,L$, β_j o efeito do nível j do fator B, $j = 1,...,C$, γ_{ij} é o efeito da interação do nível i do fator A com o nível j do fator B e ϵ_{ijk} é o erro aleatório, $k = 1,...,n_{ij}$.
Os efeitos principais dos fatores A, B e da interação $A \times B$ foram modelados considerando:

$$\alpha_i = \begin{cases} c, & i = 1 \\ -c, & i = 2 \\ 0, & \text{caso contrário} \end{cases} ; \quad \beta_j = \begin{cases} c, & i = 1 \\ -c, & i = 2 \\ 0, & \text{caso contrário} \end{cases} ; \text{ e}$$

$$\gamma_{ij} = \begin{cases} c, & i = j \text{ e } i,j = 1,2 \\ -c, & i \neq j \text{ e } i,j = 1,2 \\ 0, & \text{caso contrário} \end{cases},$$

com $c = 0{,}25\sigma, 0{,}5\sigma$ e 1σ, onde σ representa o desvio-padrão da população amostrada. Os efeitos considerados satisfizem as restrições $\sum_{i=1}^{L} \alpha_i = \sum_{j=1}^{C} \beta_j = \sum_{i=1}^{L} \gamma_{ij} = \sum_{j=1}^{C} \gamma_{ij} = 0$.
Foram considerados delineamentos 3×3, equilibrados e desequilibrados, com amostras de dimensão global $N = \sum_i \sum_j n_{ij} = 27, 45, 90$, sendo as dimensões por célula apresentadas na (Tabela 1).
Foram consideradas distribuições discretas com diferentes graus de dispersão e assimetria: (i) Binomial assimétrica positiva: $B(K; 0{,}2)$ com $K = 25, 50, 100$; (ii) Binomial simétrica: $B(K; 0{,}5)$ com $K = 10, 20, 40$; (iii) Binomial Negativa: $BN(K; 0{,}4)$ com $K = 2, 4, 8$; (iv) Poisson: $P(\lambda)$ com $\lambda = 5, 10, 20$; e (v) Uniforme: $\{0,...,K\}$ com $K = 10, 20, 40$.
As taxas de erro de tipo I e II dos vários testes foram avaliadas considerando seis modelos distintos:

1. inexistência de efeitos principais e inexistência de interação (modelo nulo);
2. um efeito principal e inexistência de interação;
3. dois efeitos principais e inexistência de interação (modelo de efeitos principais);

Tabela 1: Dimensão global da amostra (N) e dimensões consideradas para as amostras (n_{ij}), por tipo de delineamento e grau de desequilíbrio

		Equilibrado	Desequilibrado	
			<deseq.	>deseq.
$N = 27$	$n_{11} = n_{12} = n_{13}$	3	2	
	$n_{21} = n_{22} = n_{23}$	3	3	
	$n_{31} = n_{32} = n_{33}$	3	4	
$N = 45$	$n_{11} = n_{12} = n_{13}$	5	5	2
	$n_{21} = n_{22} = n_{23}$	5	6	5
	$n_{31} = n_{32} = n_{33}$	5	4	8
$N = 45$	$n_{11} = n_{12} = n_{13}$	10	4	2
	$n_{21} = n_{22} = n_{23}$	10	10	9
	$n_{31} = n_{32} = n_{33}$	10	16	19

4. sem efeitos principais e existência de interação (modelo de interação);
5. um efeito principal e existência de interação;
6. dois efeitos principais e existência de interação (modelo completo).

Em cada cenário distribucional foram consideradas 1000 réplicas de Monte Carlo e, para cada um dos procedimentos, registou-se a distribuição empírica dos *valores p*, a *taxa de erro de tipo I empírica* e a *taxa de erro de tipo II empírica*, quando $\alpha = 1\%, 5\%$ e 10%.
Na análise do desempenho dos testes no controlo da probabilidade do erro de tipo I, foi usado o critério liberal de Bradley [1]. Segundo este critério, um teste de hipóteses é dito liberal se, tomadas k amostras de tamanho igual da mesma população, a taxa de rejeição da hipótese nula pelo teste, realizado em cada uma das k amostras, é maior do que $1{,}5\alpha$. Um teste é conservador quando a taxa de rejeição da hipótese nula pelo teste é menor do que $0{,}5\alpha$.

Os testes *ART*, *ATS* e *WTS* foram aplicados quer às observações originais (y) quer às respetivas ordens (ry). Para distinguir entre estas duas situações, na apresentação dos resultados utilizaram-se os posfixos y e ry, respetivamente. Na aplicação do teste INT o *score* normal foi definido como $\Phi^{-1}\left(\frac{r_i}{N+1}\right)$, onde Φ^{-1} denota o quantil da distribuição $N(0,1)$, r as ordens das observações y e N o número total de observações.
Para a aplicação dos procedimentos foram usados os pacotes ARTool, rankFD e GFD do programa R Project [7], e funções disponíveis em http://www.uni-koeln.de/∼luepsen/R/.

3 Resultados

Na impossibilidade de apresentar os resultados para todos os níveis de significância considerados, nesta seção apresentam-se apenas os resultados quando $\alpha = 0.05$. A eficácia dos testes foi avaliada com base na taxa de erro de tipo I empírica e na potência empírica.

3.1 Delineamentos equilibrados *vs.* desequilibrados

Nos resultados apresentados nas Tabelas 2 e 3 distingue-se o desempenho dos testes pelo tipo de desequilíbrio do delineamento.
A maioria dos testes que não verificam o critério de Bradley são classificados com liberais (valores a cinzento na Tabela 2). Com o aumento do desequilíbrio do delineamento aumenta o número de testes que violam o critério de robustez, e há um aumento do afastamento da taxa de erro de tipo I empírica destes testes relativamente ao nível de significância nominal. O teste *WTS* mostrou ser demasiado liberal, principalmente quando se testa a interação. O comportamento dos testes L e vdW não é consistente, pois são testes conservadores quando se testa a presença de interação e são liberais quando se testa a presença dos efeitos principais. De um modo geral, a taxa de erro de tipo I empírica dos métodos robustos tende a estar mais próxima

Tabela 2: Taxa de erro de tipo I empírica, com $\alpha = 0{,}05$, por desequilíbrio do delineamento: equilibrado (Equil.), menor desequilíbrio (<deseq.) e maior desequilíbrio (>deseq.). Valores a cinzento representam as situações em que o teste é conservativo e a cinzento e itálico quando o teste é liberal.

Efeito testado	A			B			AB		
Procedimento	Equil.	<deseq.	>deseq.	Equil.	<deseq.	>deseq.	Equil.	<deseq.	>deseq.
ANOVA	0,050	0,049	0,051	0,046	0,068	*0,090*	0,050	0,050	0,051
ART+INT	0,051	0,052	0,060	0,046	0,050	0,026	0,049	0,049	0,046
ART.ry	0,053	0,051	0,050	0,048	0,049	0,022	0,055	0,050	0,041
ART.y	0,052	0,054	0,062	0,049	0,052	0,026	0,053	0,055	0,052
ATS	0,043	0,044	0,060	0,040	0,044	0,058	0,038	0,043	0,068
ATS.y	0,041	0,038	0,041	0,037	0,039	0,036	0,035	0,036	0,041
INT	0,050	0,049	0,050	0,046	0,069	*0,092*	0,047	0,046	0,045
L	*0,272*	*0,178*	*0,160*	*0,181*	*0,104*	0,060	0,020	0,024	0,024
vdW	*0,281*	*0,187*	*0,179*	*0,190*	*0,116*	*0,076*	0,019	0,023	0,027
RT	0,050	0,048	0,047	0,047	0,068	*0,086*	0,052	0,047	0,040
WTPS	0,049	0,046	0,039	0,046	0,049	0,039	0,049	0,046	0,031
WTS.ry	*0,078*	*0,092*	*0,129*	0,074	*0,098*	*0,164*	*0,142*	*0,177*	*0,264*
WTS.y	*0,076*	*0,083*	*0,102*	0,072	*0,091*	*0,127*	*0,130*	*0,152*	*0,195*

Tabela 3: Potência empírica, com $\alpha = 0{,}05$ por desequilíbrio do delineamento: equilibrado (Equil.), menor desequilíbrio (<deseq.) e maior desequilíbrio (>deseq.). Valores a negrito representam os testes mais potentes e os valores a cinzento os testes que não são robustos (a itálico os liberais e a não itálico os conservativos).

Efeito testado	A			B			AB		
Procedimento	Equil.	<deseq.	>deseq.	Equil.	<deseq.	>deseq.	Equil.	<deseq.	>deseq.
ANOVA	**0,650**	**0,612**	**0,593**	0,652	0,624	*0,677*	**0,475**	**0,434**	**0,421**
ART+INT	0,642	0,606	**0,594**	0,635	0,576	**0,530**	0,452	0,410	0,395
ART.ry	0,635	0,592	0,567	0,630	0,563	*0,508*	0,442	0,391	0,360
ART.y	0,648	0,614	0,599	0,650	0,590	0,536	0,476	0,431	0,414
ATS	0,622	0,563	0,418	0,615	0,561	0,423	0,407	0,351	0,250
ATS.y	0,627	0,552	0,360	0,627	0,559	0,378	0,432	0,360	0,198
INT	**0,653**	**0,613**	**0,598**	0,647	**0,618**	*0,682*	0,461	0,418	0,409
L	*0,444*	*0,396*	*0,375*	*0,393*	*0,256*	0,131	*0,334*	*0,292*	*0,281*
vdW	*0,456*	*0,417*	*0,409*	*0,409*	*0,288*	*0,176*	*0,348*	*0,314*	0,317
RT	0,641	0,595	0,572	0,636	0,605	*0,666*	0,445	0,396	0,369
WTPS	0,638	0,581	0,497	0,638	0,588	0,499	0,441	0,382	0,236
WTS.ry	*0,684*	*0,657*	*0,644*	**0,679**	*0,665*	*0,685*	*0,570*	*0,565*	*0,597*
WTS.y	*0,687*	*0,652*	*0,631*	**0,687**	*0,665*	*0,659*	*0,577*	*0,559*	*0,559*

do nível de significância considerado nos delineamentos equilibrados do que nos desequilibrados.
Na Tabela 3 destacam-se a negritos os testes que revelaram ser os mais potentes nos diferentes desequilíbrios considerados para os delineamentos. A título informativo, apresenta-se a cinzento e itálico a potência empírica dos testes cuja probabilidade do erro de primeira espécie violou o critério de robustez adotado. A potência empírica dos procedimentos robustos é menor nos delineamentos desequilibrados, sendo esta mais reduzida nos delineamentos com maior desequilíbrio (>deseq.).

3.2 Tamanho da amostra

As Figuras 1 e 2 ilustram a eficácia dos testes quando se consideram diferentes tamanhos da amostra global, tendo em consideração o desequilíbrio do delineamento. De notar que nestas figuras não existem

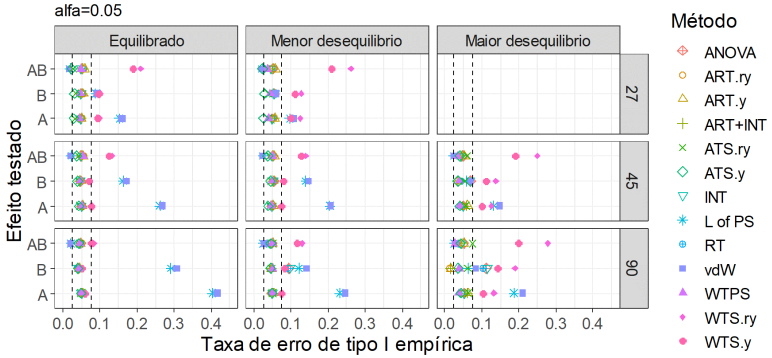

Figura 1: Taxa de erro de tipo I empírica por dimensão da amostra, efeito testado e desequilíbrio do delineamento, com $\alpha = 0{,}05$. As linhas verticais a tracejado indicam os limites de robustez do critério de Bradley.

reprentações quando $N = 27$ e o delineamento é mais desequilibrado uma vez que neste estudo não foi gerado este caso.

Na Figura 1, os testes que se encontram entre as duas linhas verticais a tracejado são considerados robustos, os que se estão à esquerda da primeira linha vertical são classificados como conservativos e os que se estão representados à direita da segunda linha vertical são liberais.

A maioria dos testes que não verificam o critério de Bradley são liberais (Figura 1). À medida que aumenta o tamanho global da amostra observa-se um aumento do número de testes que violam o critério de robustez e um aumento da taxa de erro de tipo I empírica, principalmente quando os delineamentos são mais desequilibrados. Estes comportamentos também são observados para a maior parte dos procedimentos à medida que aumenta o desequilíbrio do delineamento.

Com o aumento da dimensão da amostra observa-se um aumento da potência empírica dos testes (Figura 2). Nos testes robustos observa-se ainda que com o acentuar do desequilibrio dos delineamentos há

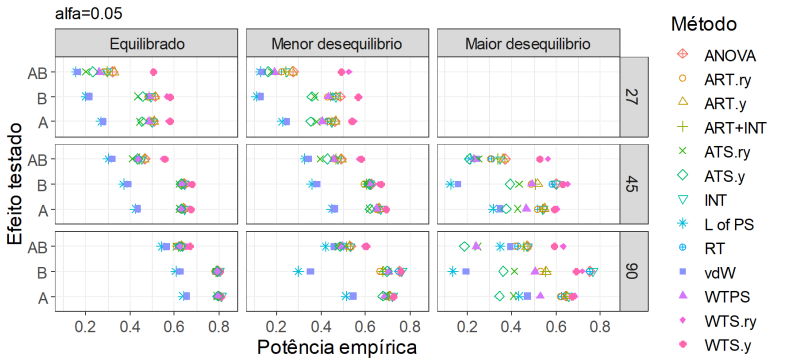

Figura 2: Potência empírica por dimensão da amostra, efeito testado e desequilíbrio do delineamento, com $\alpha = 0{,}05$.

uma redução na potência empírica.

3.3 Tamanho dos efeitos

Apenas se observa um aumento da taxa de erro de tipo I empírica nos testes liberais (L, vdW, WTS) com o aumento do tamanho do efeito dos níveis dos fatores e da interação, que depende da ponderação atribuída ao desvio-padrão da população amostrada (Figura 3). Em todos os outros procedimentos a taxa de erro de tipo I empírica mantém- se inalterada seja qual for o tamanho do efeito dos níveis dos fatores e da interação.

Quanto maior o tamanho dos efeitos maior é a capacidade dos testes para para rejeitar H_0 quando esta hipótese é falsa (Figura 4).

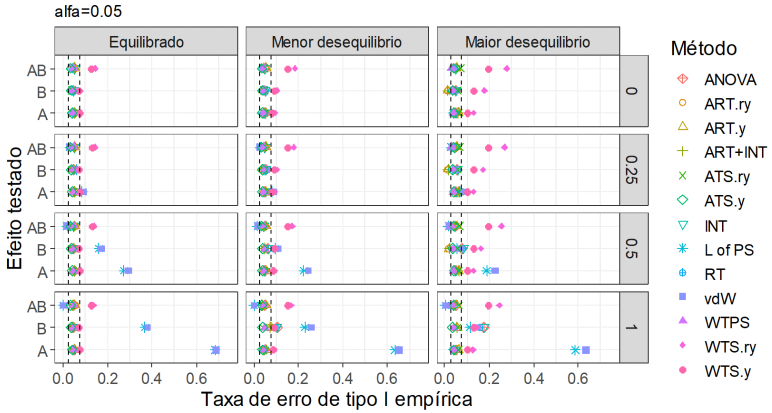

Figura 3: Taxa de erro de tipo I empírica por tamanho do efeito $c = 0{,}25\sigma, 0{,}5\sigma, 1\sigma$, efeito testado e desequilíbrio do delineamento, com $\alpha = 0{,}05$. As linhas verticais a tracejado indicam os limites de robustez do critério de Bradley.

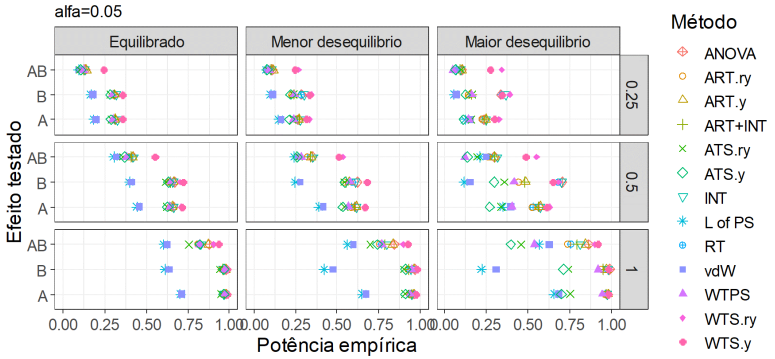

Figura 4: Potência empírica por tamanho do efeito $c = 0{,}25\sigma, 0{,}5\sigma$, 1σ, efeito testado e desequilíbrio do delineamento, com $\alpha = 0{,}05$.

3.4 Modelo

Uma vez que a taxa de erro de tipo I empírica corresponde à proporção de réplicas que rejeitaram H_0 quando H_0 (o efeito é nulo) é verdadeira, na Figura 5 só se representam as taxas associadas aos efeitos nulos considerados em cada modelo. Do mesmo modo, como a potência empírica corresponde à proporção de réplicas que rejeitaram H_0 quando H_0 falsa na Figura 6 só se representam as potências associadas aos efeitos não nulos do respetivo modelo.

No teste à interação, quer a taxa de erro de tipo I quer a potência empírica não é afectada pela presença ou não de efeitos principais (Figura 5). No teste aos efeitos principais na presença de interação observa-se um aumento na taxa de erro de tipo I empírica, que se agrava com o desequilibrio do delineamento.

Com o aumento desequilíbrio do delineamento parece haver uma diminuição da capacidade dos procedimentos decidirem corretamente (Figura 5 e Figura 6).

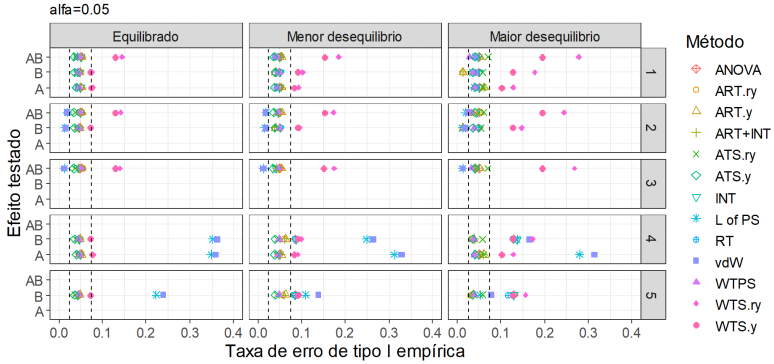

Figura 5: Taxa de erro de tipo I empírica por modelo, efeito testado e desequilíbrio do delineamento, com $\alpha = 0{,}05$. As linhas verticais a tracejado indicam os limites de robustez do critério de Bradley.

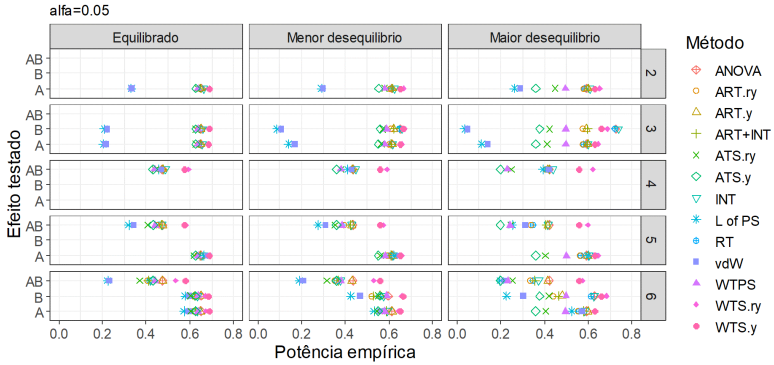

Figura 6: Potência empírica por modelo, efeito testado e desequilíbrio do delineamento, com $\alpha = 0{,}05$

4 Conclusão

Os nossos resultados mostram que de um modo geral os testes tendem a diminuir a capacidade de decidir corretamente quanto mais desequilibrado for o delineamento. Além disso, o desempenho dos testes é afectado pela dimensão da amostra, efeitos principais presentes no modelo e pelo tamanho dos efeitos.

O teste *WTS* mostrou ser demasiado liberal. O comportamento dos testes L de PS e de van der Waerden não é consistente, pois tanto são testes conservadores como liberais. Quando se testam os efeitos principais, a taxa de erro de tipo I empírica dos testes ART e $ART{+}INT$ é maior nos delineamentos desequilibrados do que nos equilibrados, e todos os procedimentos robustos apresentam potências menores nos delineamentos desequilibrados.

No teste à interação, a taxa de erro de tipo I empírica dos procedimentos é menor nos delineamentos desequilibrados, mas a taxa de erro de tipo II é menor nos delineamentos equilibrados.

A ANOVA paramétrica tem um comportamento estável, exceto nos delineamentos com maior desequilíbrio e o tamanho dos efeitos principais e da interação é elevado. Este teste mostrou ser robusto e muitas vezes mais potente do que os procedimentos alternativos considerados neste trabalho.

De futuro pretende-se estender este estudo à situação em que as variâncias são heterogéneas.

Agradecimentos

Este trabalho é financiado por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projeto "UID/MAT/04674/2019 (CIMA)".

Referências

[1] Bradley, J. V. (1978). Robustness? *British Journal of Mathematics and Statistical Psychology*, 31, 144–151.

[2] Afonso, A., Pereira, D. G. (2019). Comparação entre métodos não paramétricos para a análise de variância com dois fatores: um estudo de simulação. In *Classificação e Análise de Dados – Métodos e Aplicações III* (Eds. Bacelar-Nicolau, H., Sousa, F., Marcelo, C., Ferreira, A. S., Infante, P., Figueiredo, A.). Instituto Nacional de Estatística, 147–158.

[3] Pereira, D. G., Afonso, A. (2020) Taxas de erros de tipos I e II de procedimentos não paramétricos alternativos à ANOVA com dois fatores para dados discretos. *Atas do XXIII Congresso da SPE.* Sociedade Portuguesa de Estatística (Eds. Salgueiro, M. F., Vicente, P., Calapez, T., Marques, C., Silva, M. E.), 75-88.

[4] Hahn, S., Konietschke, F., Salmaso, L. (2014). A Comparison of efficient permutation tests for unbalanced ANOVA in two by two designs and their behavior under heteroscedasticity. In *Topics in Statistical Simulation: Research Papers from the 7th International Workshop on Statistical Simulation* (Eds. Melas, V.B., Mignani, S., Monari, P., Salmaso, L.). Springer, New York, USA, 257–269.

[5] Luepsen, H. (2017). The aligned rank transform and discrete variables - a warning. *Communications in Statistics – Simulation and Computation*, 46, 6923–6936.

[6] Pauly M., Brunner E., Konietschke F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society*, Series B, 77, 461–473.

[7] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.