Taxas de erros de tipos I e II de procedimentos não paramétricos alternativos à ANOVA com dois fatores para dados discretos

Dulce G. Pereira
CIMA/IIFA e Departamento de Matemática/ECT, Universidade de Évora, *dgsp@uevora.pt*

Anabela Afonso
CIMA/IIFA e Departamento de Matemática/ECT, Universidade de Évora, *aafonso@uevora.pt*

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

Resumo: Usualmente, nas alternativas à ANOVA paramétrica com dois fatores as observações são substituídas pelas suas ordens. No estudo do desempenho destas alternativas apenas têm sido consideradas distribuições contínuas. Contudo, quando os dados provêm de distribuições discretas propiciam a existência de muitos empates. Neste trabalho, recorrendo a um estudo por simulação, estudamos as taxas de erro de tipo I e II de várias dessas alternativas. Foram considerados delineamentos equilibrados com 2 fatores e dados provenientes de distribuições discretas. Os testes *WTS* e *USP* foram os que mostraram ser liberais. Os testes L de Puri & Sen e de *van der Waerden* não mostraram ter um bom desempenho.

1 Introdução

A análise de variância (ANOVA) com dois fatores, A e B, pretende testar se todos os níveis do fator A originam a mesma variância média

na variável resposta (quantitativa), isto é, se possuem um efeito médio igual (analogamente para o fator B), bem como determinar se existe interação entre os dois fatores.
A obtenção de conclusões através da ANOVA deve ser precedida da verificação de algumas condições, sob pena de poder conduzir a inferências erradas. Deve assegurar-se que a sua aplicação só tem lugar quando as observações são independentes e os dados provêm de populações normalmente distribuídas com variância comum.
Em muitas situações não podemos utilizar a ANOVA paramétrica porque os dados não são contínuos e por vezes são ordinais. Nestes casos, podemos recorrer a alternativas não-paramétricas. Muitas destas alternativas podem ser aplicadas quer a dados contínuos quer a dados discretos pois consistem em aplicar transformações aos dados originais (ordens, *scores* normais das ordens, ...) [4]. No entanto, existem poucos estudos sobre o desempenho destas alternativas considerando distribuições discretas. Mansouri *et al.* [6] comparam o desempenho do teste *aligned rank transform* com distribuições do erro contínuas e discretas (Binomial) e não encontraram grandes diferenças. Nos delineamentos 2×2, Kaptein *et al.* [3] mostraram que, no caso de escalas de Likert, a potência da *ANOVA type statistic* é superior à do teste F da ANOVA.
Neste trabalho, analisamos as probabilidades de erros de tipo I e II das técnicas não paramétricas, considerando um estudo de simulação quando os dados são provenientes de distribuições discretas e delineamentos equilibrados 2×5, 3×3 e 3×4.

2 Alternativas não paramétricas

Nos últimos anos foram propostas várias alternativas bastante distintas à ANOVA paramétrica com dois fatores. As técnicas *rank transform* (RT) e *inverse normal transformation* (INT) consistem na substituição das observações pelas suas ordens, ou pelos *scores* normais das ordens, respectivamente, e a posterior aplicação da ANOVA paramétrica usual. A técnica *aligned rank transform* (ART), bem

como a combinação deste método com o *INT* (*ART+INT*), subtrai todos os efeitos que não sejam de primeiro interesse antes de se realizar a ANOVA. A estatística *L* de Puri e Sen (*L de PS*), o teste de van der Waerden (*vdW*), e as *Wald type statistic* (*WTS*) e *ANOVA type statistic* (*ATS*) propõem as suas próprias estatísticas, em alguns casos à custa dos modelos lineares. As alternativas *Wald type statistic permutation* (*WTPS*), *constrainded synchronized permutations* (*CSP*) e *unconstrainded synchronized permutations* (*USP*) baseiam-se na permutação das observações. A descrição destas técnicas pode ser consultada, por ex., em Hahn *et al.* [2] e Luepsen [4].

Cada uma destas técnicas tem as suas vantagens e desvantagens, não existindo uma que seja melhor do que outra em todos os contextos (Tabela 1). Algumas são muito fáceis de implementar, outras apresentam problemas ao nível do erro de tipo I, da potência e lidam mal com a heterogeneidade de variâncias. Há alternativas que apresentam os mesmos problemas que a versão paramétrica quando a distribuição não é normal, nem todas são adequados para testar a interação e algumas têm problemas quando as amostras são pequenas. Os métodos que utilizam permutações nem sempre verificam o pressuposto de permutabilidade das observações, ou seja, a probabilidade dos dados observados ser invariante relativamente às permutações aleatórias dos índices.

Na Tabela 1 apresenta-se um resumo das principais caraterísticas de cada uma destas alternativas encontradas na literatura (e.g. [2, 4, 5, 7, 8]). A avaliação do desempenho destas técnicas, face a diferentes graus de assimetria, presença de *outliers* e heterogeneidade de variâncias, foi realizada com base em distribuições contínuas e considerando delineamentos equilibrados e/ou desequilibrados.

3 Simulação

No estudo de simulação foi considerado um modelo de efeitos fixos e com interação,

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

Tabela 1: Principais vantagens e desvantagens das alternativas à ANOVA paramétrica. (+ bom desempenho, − mau desempenho, ± o desempenho depende de algumas características, n.a. não aplicável)

Método	Fácil	Erro Tipo I	Potência	Dist. não normal	Heterogeneidade	Interação	Software	Permutabilidade	Amostras pequenas
RT	+	−	−	−	−	−	+	n.a.	−
INT	+	−	±				+	n.a.	−
ART	−	±	±	+	−	+	+	n.a.	+
ART+INT	−	±	±	+	−	−	+	n.a.	±
L de PS	+	±	±	+		+	+	n.a.	+
vdW	−	+	+	+	+	+	+	n.a.	±
WTS	−	±		+	+		+	n.a.	−
ATS	−	+	−	+	+	+	+	n.a.	+
WTPS	−	+	+	+	+	+	+	±	+
CPS	−	−	+				+	−	+
UPS	−	−	+				+	−	+

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, $k = 1,...,n$.

Os efeitos principais do fator A foram modelados considerando $\alpha_1 = c$, $\alpha_2 = -c$ e $\alpha_i = 0$ se $i \neq 1, 2$, com $c = 0{,}25\sigma, 0{,}5\sigma$ e 1σ onde σ representa o desvio-padrão da população amostrada. Os efeitos principais do fator B foram modelados considerando $\beta_1 = c$, $\beta_2 = -c$

e $\beta_i = 0$ se $i \neq 1, 2$. As interações $A \times B$ foram criadas definindo $\gamma_{11} = \gamma_{22} = c$, $\gamma_{12} = \gamma_{21} = -c$ e $\gamma_{ij} = 0$ nos restantes casos.

As taxas de erro de tipo I e II dos vários testes foram avaliadas considerando dois cenários distintos: (1) um efeito principal e inexistência de interação, ou seja, $c \neq 0$ para o efeito A e $c = 0$ para os outros efeitos; (2) um efeito principal e existência de interação, i.e., $c \neq 0$ para os efeitos A e $A \times B$ e $c = 0$ para o efeito B.

Foram considerados delineamentos equilibrados ($n_{ij} = 3, 5, 10$) com 2 fatores, A e B, com igual e desigual número de níveis ($2 \times 5, 3 \times 3$, 3×4) e dados provenientes de distribuições discretas, com diferentes parâmetros de modo a obter vários graus de dispersão e assimetria: (i) Binomial assimétrica positiva: $B(N; 0{,}2)$ com $N = 25, 50, 100$; (ii) Binomial simétrica: $B(N; 0{,}5)$ com $N = 10, 20, 40$; (iii) Binomial Negativa: $BN(N; 0{,}4)$ com $N = 2, 4, 8$; (iv) Poisson: $P(\lambda)$ com $\lambda = 5, 10, 20$; e (v) Uniforme: $\{0,...,N\}$ com $N = 10, 20, 40$.

Para cada cenário distribucional foram realizadas $M = 1000$ replicações tendo-se registado, para cada um dos testes descritos na secção anterior, a distribuição empírica dos *valores p*, a proporção de réplicas que rejeitaram H_0 quando H_0 verdadeiro (*taxa de erro de tipo I empírica*) e a proporção de réplicas que não rejeitaram H_0 quando H_1 verdadeiro (*taxa de erro de tipo II empírica*), ao nível de significância definido, $\alpha = 1\%, 5\%$ e 10%. 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 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 pode ser considerado robusto se a sua taxa de erro de tipo I empírica estiver no intervalo $[0{,}5\alpha; 1{,}5\alpha]$. O teste é considerado conservador se a taxa empírica estiver abaixo do limite inferior e é considerado liberal se estiver acima do limite superior.

Foram usados os pacotes ARTool, rankFD e GFD do programa R Project [9], e funções disponíveis em http://www.uni-koeln.de/~luepsen/R/ e http://static.gest.unipd.it/~salmaso/web/.

4 Resultados

Dado não ser possível mostrar todos os resultados, nas Figuras 1 e 2 ilustra-se o comportamento genérico da distribuição empírica dos *valores p* dos vários testes.

Quando o efeito em estudo não está presente, os testes de permutação *CSP* e *USP* são os que apresentam uma distribuição com maior dispersão (Figuras 1 e 2). Os testes *USP* e *WTS* distinguem-se de todos os restantes por na sua distribuição predominarem valores mais elevados. Os testes *L de PS* e *van der Waerden* destacam-se ora por apresentarem uma frequência maior de *valores p* menores ora pelo comportamento inverso (no teste ao efeito quando a interação está presente).

Quando o efeito em estudo está presente, a distribuição tende a ser assimétrica e a possuir vários valores atípicos (Figuras 1 e 2). A predominância de *valores p* mais elevados é maior nos testes *USP* e *WTS*, seguindo-se os testes *INT*, *ANOVA* e *ART*. Pelo contrário, uma maior frequência de *valores p* mais baixos é registada nos testes *L de PS* e *van der Waerden* (excepto no teste ao efeito isolado quando a interação está presente) e posteriormente nos testes *CSP* e *ATS*.

As taxas de erro de tipo I e II empíricas reportadas nas Tabelas 2 a 4 correspondem à média das taxas de erro produzidas por cada um dos testes em todos os cenários considerados. De acordo com o critério de Bradley [1], os testes *WTS* e *USP* são demasiado liberais. Os testes de *L de PS* e *van der Waerden* apresentam um comportamento instável; são muito conservadores na ausência de interação, mas na presença de interação a sua taxa de erro de tipo I ultrapassa o valor de α. O teste *CSP* é o que apresenta a maior taxa de erro de tipo II empírica. Os restantes testes apresentam taxas de erro empíricas semelhantes, embora o teste *ATS* seja o que menos vezes apresentou taxas de erro de tipo I empíricas superiores ao valor α nominal e o teste *ART* o que mais vezes ultrapassou o valor α.

Na análise que se segue excluir-se-ão os testes *USP*, *WTS*, *L de PS* e *van der Waerden* por violarem o critério de robustez [1].

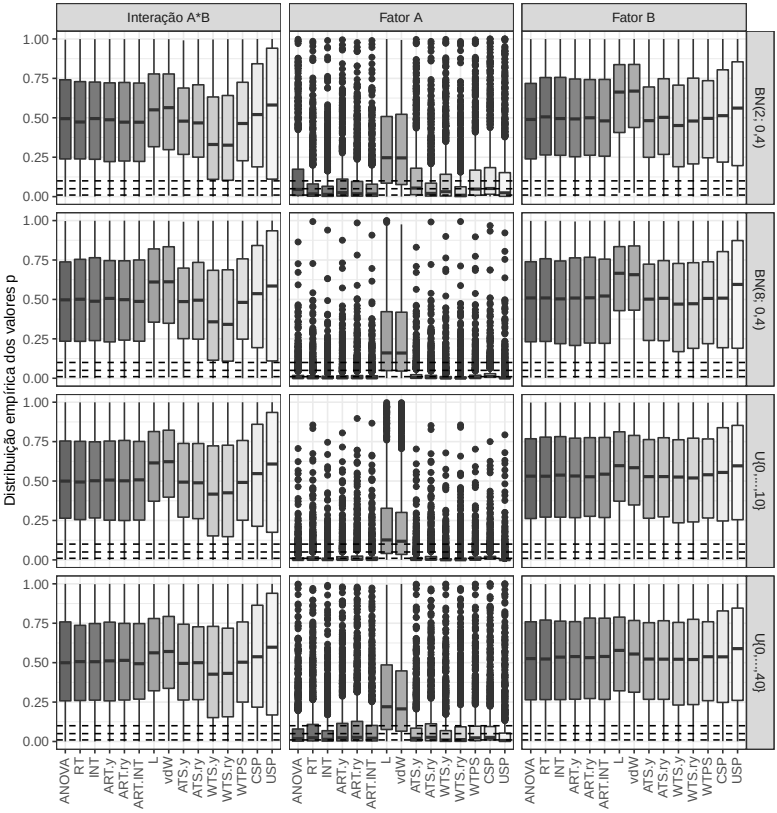

Figura 1: Distribuição empírica dos *valores p*, quando $n = 5$, $c = 0{,}5\sigma$ para o efeito A, $c = 0$ para os efeitos B e $A \times B$, e $\epsilon_{ijk} \frown BN(N;0{,}4)$ com $N = 2, 8$ e $\epsilon_{ijk} \frown U\{0,...,N\}$ com $N = 10, 40$, no delineamento 3×3. (as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

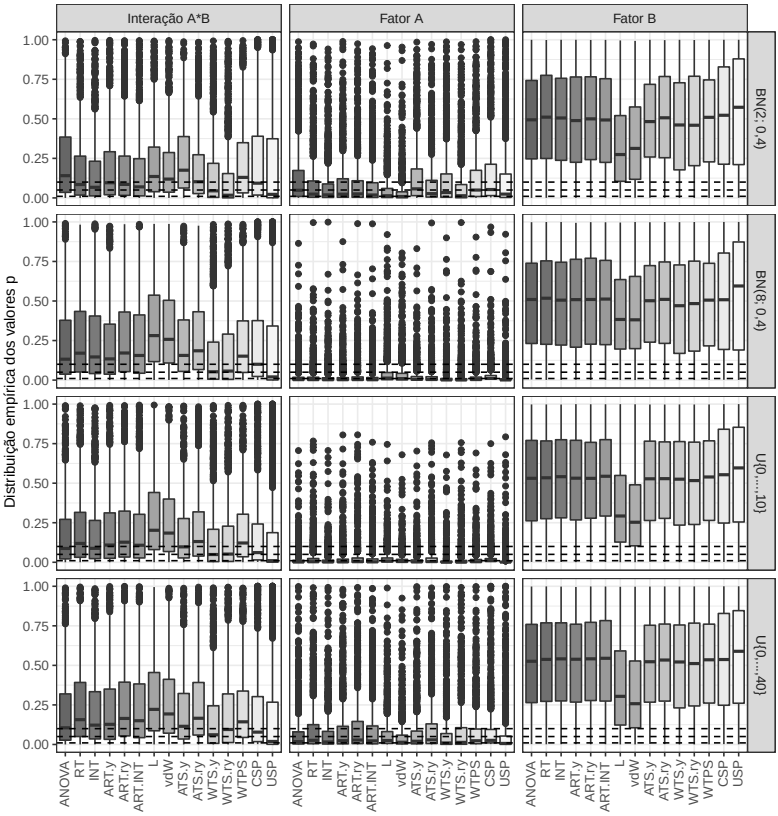

Figura 2: Distribuição empírica dos *valores p*, quando $n = 5$, $c = 0{,}5\sigma$ para os efeitos A e $A \times B$, $c = 0$ para o efeito B, e $\epsilon_{ijk} \frown BN(N; 0{,}4)$ com $N = 2, 8$ e $\epsilon_{ijk} \frown U\{0,...,N\}$ com $N = 10, 40$, no delineamento 3×3. (as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

Tabela 2: Média das taxas de erro de tipo I empíricas dos métodos no teste à presença de efeito principal B, para diferentes níveis de significância (α). Os valores de erro de tipo I que mais se afastam do nível de significância definido estão a itálico.

Interação	inexistente			existente		
α	0,01	0,05	0,10	0,01	0,05	0,10
ANOVA	0,010	0,049	0,098	0,010	0,049	0,099
RT	0,011	0,050	0,098	0,010	0,048	0,096
INT	0,010	0,049	0,099	0,009	0,045	0,092
ART.y	0,011	0,051	0,101	0,011	0,051	0,102
ART.ry	0,011	0,051	0,100	0,011	0,050	0,098
ART+INT	0,011	0,050	0,100	0,010	0,046	0,092
L de PS	*0,001*	*0,013*	*0,036*	*0,057*	*0,166*	*0,260*
vdW	*0,001*	*0,014*	*0,037*	*0,067*	*0,181*	*0,277*
ATS.y	0,006	0,037	0,083	0,006	0,037	0,083
ATS.ry	0,007	0,040	0,087	0,006	0,038	0,083
WTS.y	*0,036*	*0,096*	*0,155*	*0,037*	*0,097*	*0,155*
WTS.ry	*0,040*	*0,100*	*0,157*	*0,038*	*0,095*	*0,151*
WTPS	0,010	0,048	0,097	0,010	0,049	0,097
CSP	0,012	0,047	0,110	0,012	0,047	0,109
USP	*0,046*	*0,104*	*0,155*	*0,046*	*0,105*	*0,156*

Análise da distribuição: No teste ao efeito principal, a taxa de erro de tipo II empírica dos vários testes parece ser superior quando a distribuição dos erros é assimétrica e, além disso, a distribuição dos *valores p* apresenta maior dispersão. Quando a distribuição é simétrica, o tipo de achatamento influencia a dispersão dos *valores p* sendo mais elevada na distribuição platicúrtica.
A taxa de erro de tipo I empírica dos vários testes é semelhante qualquer que seja a distribuição considerada.

Análise do efeito do tamanho da amostra: Com amostras de reduzida dimensão ($n = 3$), o teste *ATS* revelou ser conservador

Tabela 3: Média das taxas de erro de tipo II empíricas dos métodos no teste à presença de efeito principal A, para diferentes níveis de significância (α).

Interação	inexistente			existente		
α	0,01	0,05	0,10	0,01	0,05	0,10
ANOVA	0,395	0,276	0,219	0,395	0,277	0,219
RT	0,394	0,278	0,220	0,403	0,283	0,224
INT	0,381	0,264	0,208	0,386	0,268	0,211
ART.y	0,393	0,279	0,223	0,392	0,279	0,222
ART.ry	0,398	0,284	0,227	0,405	0,288	0,230
ART+INT	0,390	0,276	0,219	0,393	0,278	0,221
ATS.y	0,424	0,293	0,230	0,424	0,294	0,230
ATS.ry	0,415	0,288	0,228	0,428	0,295	0,232
WTPS	0,409	0,284	0,224	0,407	0,284	0,224
CSP	0,605	0,485	0,337	0,605	0,485	0,337

quando $\alpha = 1\%$ tal como o teste *CSP* para $\alpha = 1\%$ e 5%.
A taxa de erro de tipo II empírica de todos os testes diminui com o aumento da dimensão da amostra por célula e há um aumento na dispersão dos *valores p*. Além disso, os testes tendem a apresentar uma taxa de erro de tipo II empírica semelhante.

Análise dos efeitos considerados: A média das taxas de erro de tipo I empíricas dos vários testes não se altera com a intensidade do efeito considerado. Quando o efeito não está presente, a percentagem de vezes que os testes *ART*, *INT*, *RT* e *ART+INT* decidem corretamente aumenta com a intensidade do efeito.
A taxa de erro de tipo II empírica de todos os testes diminui com o aumento da intensidade do efeito e os testes tendem a apresentar um comportamento semelhante. Quando se considera um efeito com intensidade $0{,}25\sigma$, de um modo geral, todos os testes não detectam a existência de interação. Contudo, à medida que se aumenta a intensidade do efeito os *valores p* reduzem, bem como a dispersão

Tabela 4: Média das taxas de erro de tipo I e de tipo II empíricas dos métodos no teste à existência de interacção AB, na presença de um efeito principal significativo, para diferentes níveis de significância (α). Os valores de erro de tipo I que mais se afastam do nível de significância definido estão a itálico.

Erro	tipo I			tipo II		
α	0,01	0,05	0,10	0,01	0,05	0,10
ANOVA	0,010	0,049	0,099	0,676	0,535	0,451
RT	0,011	0,050	0,100	0,697	0,556	0,471
INT	0,010	0,049	0,099	0,672	0,533	0,450
ART.y	0,012	0,053	0,104	0,677	0,537	0,453
ART	0,012	0,053	0,102	0,697	0,557	0,473
ART.INT	0,011	0,051	0,101	0,679	0,541	0,458
L de PS	*0,002*	*0,017*	*0,041*	0,849	0,710	0,609
vdW	*0,002*	*0,016*	*0,041*	0,829	0,686	0,583
ATS.y	0,005	0,033	0,077	0,723	0,579	0,487
ATS.ry	0,006	0,036	0,082	0,742	0,597	0,502
WTS.y	*0,065*	*0,141*	*0,206*	0,546	0,424	0,353
WTS.ry	*0,079*	*0,155*	*0,220*	0,528	0,415	0,350
WTPS	0,010	0,049	0,098	0,718	0,573	0,483
CSP	0,016	0,058	0,124	0,715	0,603	0,477
USP	*0,093*	*0,163*	*0,213*	0,458	0,377	0,332

da sua distribuição empírica, verificando-se que quando o efeito é 1σ e $n = 10$ os testes já decidem corretamente.

Análise dos empates: O número de empates depende tanto do número de observações por célula (n) como dos parâmetros das distribuições. Dado que já foi feita a análise do desempenho dos testes à dimensão da amostra, em que quanto maior a dimensão da amostra maior o número de empates presentes nos dados, focar-se-á apenas o efeito da alteração dos parâmetros das distribuições.
Na ausência de interação, o número de empates não altera a taxa de

erro de tipo I empírica.
A taxa de erro de tipo II empírica do teste à presença do efeito principal tende a não ser afectada pela existência, ou não, de interação.
O desempenho dos testes *ATS*, *WTPS* e *CPS*, na deteção da presença de interação, não mostrou ser sensível à alteração dos parâmetros das várias distribuições consideradas, a *ANOVA* paramétrica por vezes aprensentou ligeiras alterações no seu desempenho não se identificando qualquer padrão, e os restantes testes mostraram ser instáveis.

Análise dos efeitos dos delineamentos: De um modo geral, não se registam diferenças entre delineamentos no comportamento dos testes.
Quando não existe interação, os testes tendem a apresentar uma taxa de erro de tipo II empírica mais baixa no delineamento 2×5 e mais elevada no delineamento 3×3. Quando a interação está presente, a taxa de erro de tipo II empírica dos testes não é afectada pelo número de níveis dos fatores.

5 Conclusão

Com base nos resultados obtidos neste estudo de simulação, verificou-se que a distribuição empírica dos *valores p* das várias alternativas à ANOVA, bem como da ANOVA paramétrica, é afectada por vários fatores como sejam, a dimensão da amostra, a intensidade do efeito, a distribuição dos erros, o número de empates e pelo número de níveis dos fatores.
Os testes de permutação *CSP* e *USP* e o teste *WTS* são os que apresentam a maior dispersão na distribuição empírica das taxas de erro de Tipo I. Além disso, estes testes mostraram ser liberais. O comportamento dos testes *L de PS* e de *van der Waerden* não é consistente, tanto são testes conservadores como liberais. Entre os restantes testes, o teste ATS é o que apresenta menores taxas de erro

de tipo I empíricas, mas por sua vez é dos que apresenta maior taxa de erro de tipo II empírica.
A taxa de erro de tipo II empírica de todos os testes diminui com o aumento da dimensão da amostra por célula, com o aumento a intensidade do efeito e os testes tendem a apresentar um comportamento semelhante.
Na presença de interação, todos os testes apresentaram um fraco desempenho no teste à interação, apresentando taxas de erro de tipo II elevadas, mas estas tendem a diminuir com o aumento da dimensão da amostra por célula e com a intensidade dos efeitos.
Os testes *INT*, *ANOVA* e *ART* apresentam taxas de erro de tipo I e II empíricas semelhantes. Comportam-se melhor no estudo do efeito principal presente do que no estudo da interação, do que as restantes alternativas não paramétricas.
Com base nos resultados obtidos, não é aconselhável a utilização dos testes *USP*, *WTS*, *L de PS* e *van der Waerden*. Também é desaconselhada a utilização das restantes alternativas para testar a interação na presença de um efeito principal, especialmente quando a intensidade do efeito é pequena e a dimensão da amostra reduzida.
De futuro pretendemos estender este estudo considerando delineamentos desequilibrados, variâncias heterogéneas e a existência de células omissas.

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] 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 Melas V., Mignani S., Monari P., Salmaso L. (eds.): *Topics in Statistical Simulation. Springer Proceedings in Mathematics & Statistics*, 114, 257–269.

[3] Kaptein, M., Nass, C., Markopoulos, P. (2010). Powerful and consistent analysis of Likert-type rating scales. *Proceedings of CHI 2010*, 2391–2394

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

[5] Mansouri, H., Chang, G.-H. (1995). A comparative study of some rank tests for interaction. *Computational Statistics & Data Analysis*, 19, 85–96.

[6] Mansouri, H., Paige, R., Surles, J. G. (2004). Aligned rank transform techniques for analysis of variance and multiple comparisons. *Communications in Statistics - Theory and Methods*, 33, 2217–2232.

[7] 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.

[8] Toothaker, L. E., Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational Statistics*, 19, 237–273.

[9] 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/.