CLASSIFICAÇÃO E ANÁLISE DE DADOS

Métodos e Aplicações III - CLADMAp III

Editores

Helena Bacelar-Nicolau

Fernanda Sousa

Carlos Marcelo

Ana Sousa Ferreira

Paulo Infante

Adelaide Figueiredo

CLASSIFICAÇÃO E ANÁLISE DE DADOS

MÉTODOS E APLICAÇÕES III - CLADMAp III

Editores
Helena Bacelar-Nicolau
Fernanda Sousa
Carlos Marcelo
Ana Sousa Ferreira
Paulo Infante
Adelaide Figueiredo

Associação Portuguesa de
Classificação e Análise de Dados

Título
Classificação e Análise de Dados – Métodos e Aplicações III

Editores
Helena Bacelar-Nicolau (Universidade de Lisboa)
Fernanda Sousa (Universidade do Porto)
Carlos Marcelo (Instituto Nacional de Estatística)
Ana Sousa Ferreira (Universidade de Lisboa)
Paulo Infante (Universidade de Évora)
Adelaide Figueiredo (Universidade do Porto)

Impressão
Instituto Nacional de Estatística
Av. António José de Almeida
1000-043 LISBOA

1.ª Edição
Lisboa, Abril de 2019
ISSN 2183-8801
Depósito legal 454535/19
Tiragem: 200 exemplares

Comparação entre métodos não paramétricos para a análise de variância com dois fatores: um estudo de simulação

Anabela Afonso[1] · Dulce G. Pereira[2]

Resumo Nos últimos anos têm sido propostas várias alternativas não paramétricas à Análise de Variância (ANOVA) com dois fatores. Neste trabalho, realizamos um estudo de simulação para analisar a probabilidade de erro de Tipo I e a potência de alguns desses testes alternativos, quando os dados são provenientes de distribuições discretas e os delineamentos 3×3 equilibrados. Dois testes apresentaram taxas de erro de Tipo I empíricas superiores ao nível de significância nominal e dois testes exibiram fraca potência no teste à interação, sendo por isso desaconselhada a sua utilização.

Palavras-chave: Estatística de Wald, Testes de Permutação, Transformação em Ordens.

1 Introdução

A Análise de Variância (ANOVA) foi introduzida por Fisher, com aplicações iniciais no domínio da Agronomia e Biologia, mas são várias as aplicações na área da Epidemiologia e das Ciências Sociais, por exemplo. Num delineamento fatorial pretende-se estudar a influência de um ou mais fatores numa determinada variável resposta. Neste tipo de delineamento está implícito que as amostras são aleatórias e independentes. Na ANOVA para além de se assumir que a variável resposta é do tipo contínuo, também se pressupõe que as populações têm distribuição normal e as variâncias são homogéneas.

Quando se trabalha com dados reais encontramos vários problemas: i) a natureza dos dados nem sempre é contínua; por exemplo, é muito usual nas áreas da Biologia e da Ecologia obtermos dados de contagens e na área das Ciências Sociais predominam dados ordinais; ii) a normalidade das distribuições é muitas vezes

[1] CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, *aafonso@uevora.pt*
[2] CIMA/IIFA e Departamento de Matemática/ECT, da Universidade de Évora, *dgsp@uevora.pt*

violada devido quer à forma da distribuição quer à presença de valores atípicos; iii) vulgarmente as amostras são pequenas, o que associado a dados não normais não possibilita a aplicação do Teorema do Limite Central, o que põe em causa não só o pressuposto da normalidade como também a validade da distribuição F. Nestas situações a ANOVA paramétrica não é a técnica mais adequada o que levou ao desenvolvimento de alternativas.

Nos últimos anos têm sido propostas alternativas não paramétricas à ANOVA com dois fatores, sendo umas mais simples do que outras, não existindo uma alternativa que seja melhor que as restantes em todos os contextos. Na literatura o desempenho destas alternativas foi analisado considerando distribuições contínuas, preocupando-se com a assimetria e a presença de valores atípicos e/ou com variâncias heterogéneas, tanto para delineamentos equilibrados como desequilibrados. No entanto, não há estudos onde se utilizem distribuições discretas, que usualmente potenciam o surgimento de empates.

O principal objetivo deste trabalho é colmatar a lacuna existente ao nível das distribuições discretas, de forma a, se possível, generalizar as vantagens e desvantagens de algumas dessas alternativas: transformação em ordens, transformação normal inversa, transformação em ordens alinhadas, estatística L do teste de Puri e Sen, teste de van der Waerden, estatística de tipo Wald, estatística de tipo ANOVA, teste de permutação de tipo Wald e testes de permutação sincronizada.

2 ANOVA não paramétrica

As alternativas à ANOVA com dois fatores podem ser divididas em três grandes grupos cujas principais características são: 1) criam uma nova variável a partir das observações e realizam a ANOVA paramétrica; 2) propõem as suas próprias estatísticas de teste; 3) baseiam-se nas permutações das observações.

O teste de *transformação em ordens* (RT – *rank transform*), tal como o nome indica, consiste em transformar os dados originais (Y) em ordens (R) e posteriormente aplicar a ANOVA paramétrica à nova variável R (Conover e Iman, 1976). Este teste é muito simples, mas não deve ser utilizado para testar a interação na presença de efeitos principais significativos e vice-versa (Higgins e Tatshtoush, 1994, Beasley e Zumbo, 2009), pois pode apresentar um elevado erro de Tipo I e falta de potência (Sawilowsky, 2000).

A *transformação normal inversa* (INT – *inverse normal transformation*) consiste em transformar os dados originais (Y) em ordens (R), calcular os scores normais das ordens R, obtendo-se uma nova variável dependente Z e aplicar a ANOVA a essa nova variável dependente Z (van der Waerden, 1952). A salientar

que existem várias versões para calcular os scores normais (Beasley *et al.*, 2009). Mansouri e Chang (1995) mostraram que com erros normais e delineamentos equilibrados este teste é muito conservativo. Além disso, pode haver um erro de Tipo I muito elevado, se existirem outros efeitos principais significativos.

A *transformação em ordens alinhadas* (ART – *aligned rank transform*) aplicada a delineamentos fatoriais foi proposta por Higgins e Tashtoush (1994) para dados quantitativos, tendo mais tarde sido propostas alternativas que permitem a sua aplicação a dados ordinais (e.g. Peterson, 2002). Este método consiste em subtrair os efeitos de que não são de primeiro interesse antes de realizar a ANOVA, ou seja, aos resíduos adiciona-se o efeito de interesse (linha, coluna ou interação) e substituem-se os valores pelas ordens aos quais se aplica a ANOVA. Este teste é mais robusto do que o RT (Mansouri e Chang (1995) e, além disso, quando existem *outliers* ou a distribuição tem caudas pesadas este procedimento apresenta menores erros de Tipo I do que o teste F (Higgins e Tashtoush, 1994). Contudo, o erro de Tipo I aumenta com o número de observações por célula (Luepsen, 2017) e é sensível à heterocedasticidade (Leys e Schumann, 2010).

Para lidar com o problema da inflação do erro de Tipo I, Mansouri e Chang (1995) propuseram a *combinação das transformações* ART com INT (ART+INT).

Puri e Sen (1985) propuseram uma generalização do teste H de Kruskal-Wallis que consiste numa adaptação da *estatística L* do teste das ordens com distribuição qui-quadrado (teste L de PS). Na presença de efeitos nulos, este teste controla de forma adequada o erro de Tipo I tanto no teste aos efeitos principais como à interação. Mas, tem falta de potência para testar um efeito quando existirem outros efeitos não nulos no modelo (Toothaker e Newman, 1994).

O teste de van der Waerden foi generalizado para delineamentos fatoriais por Mansouri e Chang (1995). Este teste combina a transformação INT com a estatística L de PS. De um modo geral, o teste de van der Waerden apresenta um bom comportamento do erro de Tipo I e da potência. No entanto, tal como a estatística L de PS, sofre de alguma falta de potência quando existem poucas observações por célula e estamos na presença de outros efeitos não nulos.

Akritas *et al.* (1997) propuseram uma nova estatística de ordens baseada na *estatística de tipo Wald* (WTS – Wald type statistic). Apesar desta estatística ser assimptoticamente exata mesmo quando é violado o pressuposto de normalidade, para amostras de dimensão pequena ou média este teste tende a dar resultados muito liberais. Quando o número de níveis dos fatores é elevado não é recomendada a inferência baseada na distribuição assimptótica (Brunner et al., 1997).

Para contornar o liberalismo extremo do teste WTS, Brunner *et al.* (1997) propuseram o *teste de tipo ANOVA* (ATS – ANOVA type statistic) que consiste em calcular uma estatística de teste de tipo ANOVA às ordens das observações originais. Os graus de liberdade da estatística F são corrigidos usando a

aproximação de Box. O teste ATS é mais potente que o teste WTS (Shah e Madden, 2004), mas é muito conservativo quando a distribuição dos erros é enviesada (Pauly *et al.*, 2015).

Os testes de permutação têm sido propostos como uma alternativa quando as amostras são pequenas e as distribuições dos erros não satisfazem os pressupostos. De um modo geral, estes testes assentam na condição de permutabilidade que é satisfeita quando a probabilidade dos dados observados é invariante relativamente às permutações aleatórias dos índices.

Pauly *et al.* (2015) propuseram o *teste de permutação de tipo Wald* (WTPS – Wald type permutation test) que é mais potente que o teste ATS. No teste WTPS os dados originais são permutados entre si, sem restrições, e é calculada uma estatística de tipo Wald com base nas observações permutadas. Este teste mostrou ter um bom controlo da taxa de erro de Tipo I mesmo com distribuições enviesadas com erros homocedásticos.

Os testes de *permutação sincronizada* (CSP – Constrained Synchronized Permutations, e USP – Unconstrained Synchronized Permutations) foram introduzidos por Pesarini (2001) e Salmaso (2003). Posteriormente foram generalizados para delineamentos equilibrados por Basso *et al.* (2007) e mais recentemente para alguns tipos de delineamentos desequilibrados por Hahn e Salmaso (2017). Ao contrário do teste WTPS, os testes CSP e USP impõem restrições à forma como os dados são permutados entre os níveis dos fatores e usam uma estatística de teste que não é studentizada (*i.e.*, não é de tipo Wald). Estes procedimentos mostraram ser potentes e ter uma boa aderência ao valor nominal α (Hahn e Salmaso, 2017).

3 Simulação

No estudo de simulação levado a cabo considerou-se um modelo de ANOVA com dois fatores da forma:

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

onde μ é a média global, α_i o efeito do nível i do fator A, $i = 1, \dots, L$, β_j o efeito do nível j do fator B, $j = 1, \dots, C$, γ_{ij} é o efeito da interação entre o nível i do fator A e o nível j do fator B, e ε_{ijk} é o erro aleatório, $k = 1, \dots, n$. Em particular, estudámos dois cenários distintos:

- Ausência de interação e de efeitos principais (modelo nulo);
- Dois efeitos principais e existência de interação (modelo completo)

Foram considerados delineamentos equilibrados com amostras de dimensão $n = 3, 5, 10$, $L = C = 3$, e uma diversidade de cenários distribucionais com diferentes graus de dispersão e vários tipos assimetria:

- Binomial:
  - assimétrica positiva: $B(N; 0{,}2)$ com $N = 25, 50, 100$;
  - simétrica: $B(N; 0{,}5)$ com $N = 10, 20, 40$;
- Binomial Negativa: $BN(N; 0{,}4)$ com $N = 2, 4, 8$;
- Poisson: $P(\lambda)$ com $\lambda = 5, 10, 20$;
- Uniforme: $U\{0, \dots, N\}$ com $N = 10, 20, 40$.

Considerámos várias intensidades para os efeitos:

$$\alpha_i = \begin{cases} c, & i = 1 \\ -c, & i = 2 \\ 0, & c.c. \end{cases} \qquad \beta_j = \begin{cases} c, & i = 1 \\ -c, & i = 2 \\ 0, & c.c. \end{cases} \qquad \gamma_{ij} = \begin{cases} c, & i = j \text{ e } i, j = 1{,}2 \\ -c, & i \neq j \text{ e } i, j = 1{,}2 \\ 0, & c.c. \end{cases}$$

com $c = 0{,}25\sigma; 0{,}5\sigma$ e 1σ e σ o desvio-padrão da população amostrada.

Para cada cenário distribucional foram realizadas $M = 1000$ replicações, tendo-se registado:

- A distribuição empírica dos $valores\ p$,
- O número total de réplicas que ultrapassaram o nível de significância definido, $\alpha = 1\%, 5\%$ e 10%.

para cada um dos testes descritos na secção 2, i.e., RT, INT, ART, ART+INT, L de PS, van der Waerden (vdW), WTS, ATS, WTPS, CSP e USP, e ainda para o teste F da ANOVA. Os testes ART, WTS e ATS foram aplicados tanto aos dados originais como às ordens das observações (ART.r, WTR.r, ATS.r).

Para todos os testes foi usado o programa R project (R Core Team, 2016). Foram usadas as livrarias *ARTool*, *rankFD* e *GFD*, e as funções disponíveis em http://www.uni-koeln.de/~luepsen/R/ e http://static.gest.unipd.it/~salmaso/web/.

4 Resultados

Na análise do erro de Tipo I, verificou-se que a distribuição empírica dos $valores\ p$ é semelhante para todas as distribuições consideradas. Na Figura 1 apresenta-se, como exemplo, o caso particular em que os dados foram simulados com base na distribuição Poisson com média 5.

O teste WTS, aplicado quer às observações originais quer às ordens das observações, tende a apresentar menores taxas de erro de Tipo I que os restantes testes (Figura 1). Nos testes CPS e USP observou-se uma maior dispersão nas taxas de erro obtidas. Além disso, em mais de metade das vezes a taxa de erro ultrapassa o nível de significância definido.

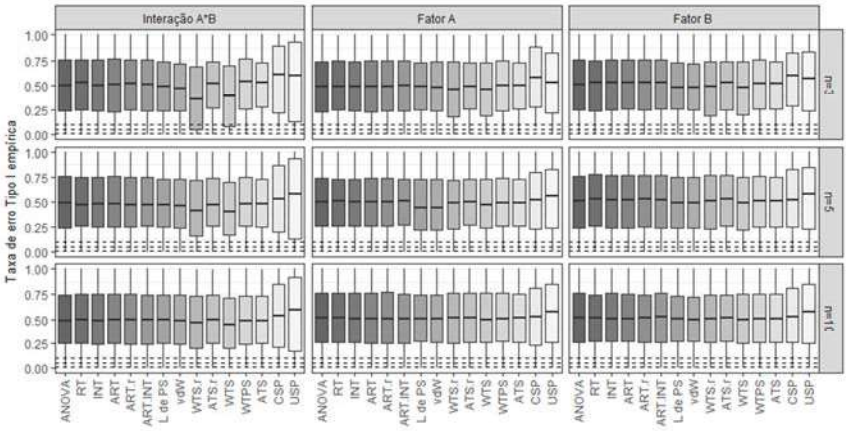

Figura 1 – Distribuição empírica das taxas de erro de Tipo I, quando $c = 0$ e $\varepsilon_{ijk} \sim P(5)$.
(as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

Na análise à interação os testes RT, INT, ART e ART+INT tendem a ter desempenhos piores quando se consideram distribuições simétricas (Binomial e Uniforme), ultrapassando na maior parte das vezes o valor de alfa nominal. Os testes ART e ART+INT são sensíveis à assimetria, reagindo mal à assimetria negativa.

Na Figura 2 é possível observar a proporção de vezes que não foi cometido o erro de Tipo I tendo em conta o nível de significância definido e a dimensão da amostra em cada célula, e considerando todas as distribuições.

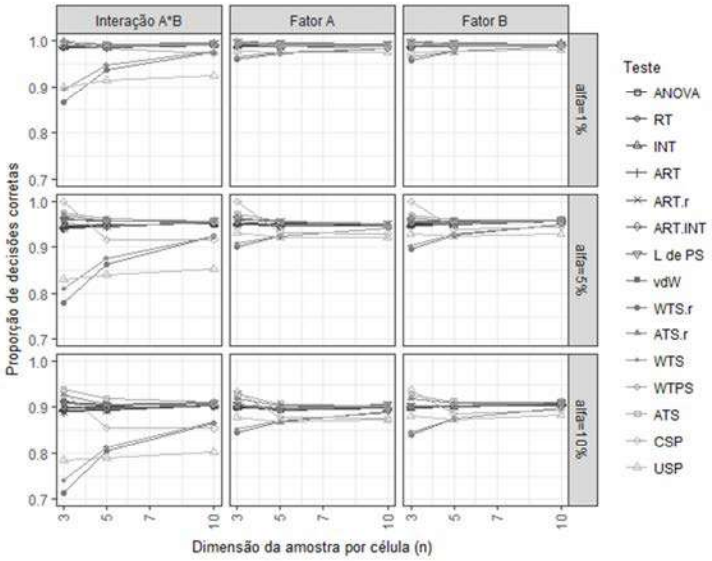

Figura 2 – Proporção de cenários em que não foi cometido o erro de Tipo I, por nível de significância nominal (α) e dimensão da amostra em cada célula.

De um modo geral, à medida que aumenta a dimensão da amostra (*n*), a taxa de erro de Tipo I de quase todos os testes aproxima-se do α nominal (Figura 2). Os testes USP e WTS são os que apresentaram os piores desempenhos. A taxa de erro de Tipo I empírica destes dois testes foi sempre superior à do nível de significância α definido, tanto no teste à interação quer nos testes aos efeitos principais.

Os resultados variam ligeiramente com o valor de α e a dimensão da amostra (Figura 2). O teste ATS é o mais estável em termos de bom comportamento quando $\alpha = 5\%$ e 10%. Quando $\alpha = 1\%$ os testes apresentam taxas de erro de Tipo I empíricas muito similares e próximas do α nominal.

Os testes RT, INT, ART e ART + INT tendem: i) a não ultrapassar o α nominal com o aumento do *n* no teste à interação; ii) a ter desempenhos piores quando se consideram distribuições simétricas (Binomial e Uniforme), ultrapassando na maior parte das vezes o α nominal, no teste à interação; iii) a ser sensíveis à assimetria, reagindo mal à assimetria negativa, nos testes aos efeitos principais.

Na análise da potência de teste, o tipo de assimetria e achatamento tem influência na distribuição empírica dos *valores p* dos testes aos efeitos principais (Figura 3). Estes testes parecem ser mais potentes quando se consideram distribuições assimétricas, e menos potentes quando a distribuição é simétrica

mesocúrtica. Contudo, quando se aumenta o valor do efeito *c* então os *valores p* diminuem e a sua distribuição aproxima-se da observada quando se considera que os erros têm distribuição Poisson.

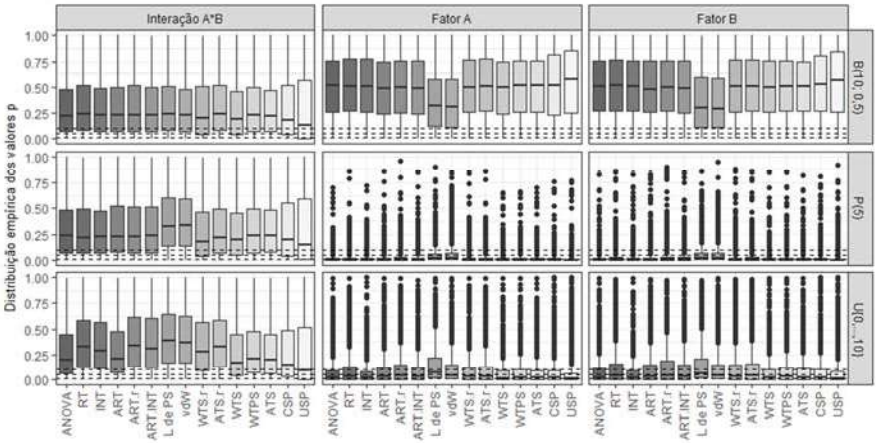

Figura 3 – Distribuição empírica dos *valores p* quando $n = 10$, $c = 0{,}25\sigma$ e $\varepsilon_{ijk} \sim B(10; 0{,}5)$ (cima), $\varepsilon_{ijk} \sim P(5)$ (meio) e $\varepsilon_{ijk} \sim U\{0, \ldots, 10\}$ (baixo).
(as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

A potência de teste aumenta com o aumento do *n*, sendo mais evidente nos testes aos efeitos principais do que no teste à interação (Figura 4). Os testes L de PS e van der Waerden são os que apresentam os piores desempenhos.

De um modo geral, à medida que aumenta a intensidade *c* do efeito observou-se um aumento da potência de teste (Figura 5). Todos os testes são mais potentes na avaliação dos efeitos principais do que no teste à interação. Os testes L de PS e van der Waerden são os menos potentes no teste à interação.

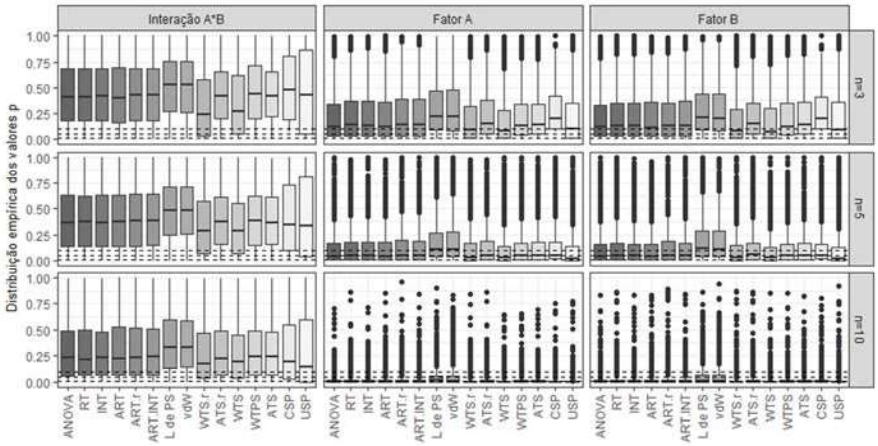

Figura 4 – Distribuição empírica dos *valores p* quando $n = 3, 5, 10$, $c = 0{,}25\sigma$ e $\varepsilon_{ijk} \sim P(5)$.
(as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

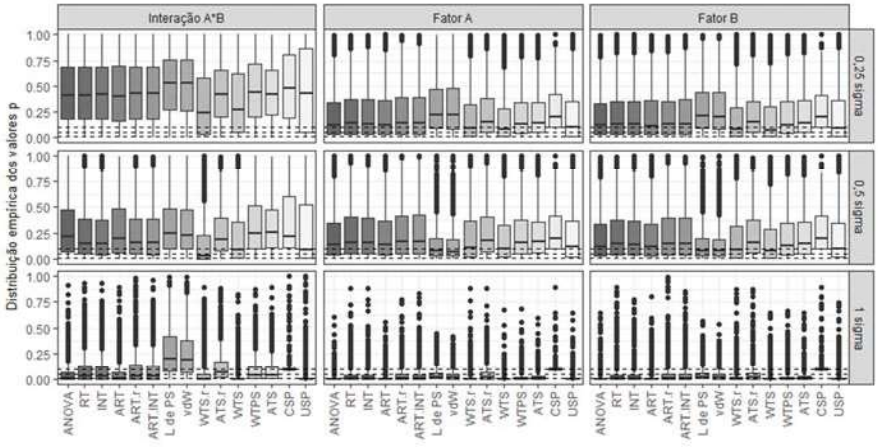

Figura 5 – Distribuição empírica dos *valores p* quando $n = 3$, $c = 0{,}25\sigma, 0{,}75\sigma, 1\sigma$ e $\varepsilon_{ijk} \sim P(5)$.
(as linhas horizontais tracejadas representam os níveis de significância de 1%, 5% e 10%)

5 Conclusão

Na análise da probabilidade de erro de Tipo I e da potência às alternativas não paramétricas à ANOVA com dois fatores 3×3 equilibrada, foram considerados vários cenários distribucionais que incluíram distribuições simétricas e assimétricas, bem como diferentes graus de dispersão e diferentes dimensões de amostra, para avaliar se estas características tinham influência no desempenho dos testes.

O tipo de distribuição e o número de observações em cada célula, de um modo geral, não afeta o erro de Tipo I, mas tem influência na potência dos testes tal como a intensidade dos efeitos. A potência de teste é menor quando os erros têm uma distribuição simétrica e aumenta com a dimensão das amostras.

Todos os testes são mais potentes na avaliação dos efeitos principais do que no teste à interação, sendo os testes L de PS e van der Waerden os menos potentes no teste à interação. Com o aumento do tamanho do efeito, estes dois testes aumentam substancialmente a sua potência no teste aos efeitos principais, alcançando em algumas situações desempenhos melhores que os outros testes. No entanto, estes testes são os que apresentam taxas de erro de Tipo I empírica mais próximas do α definido, bem como o teste ATS.

Desaconselha-se o uso dos testes USP, WTS e CSP porque apresentam taxas de erro de Tipo I empíricas mais afastadas do nível de significância nominal. Além disso, os testes CPS e USP são os que apresentam a maior dispersão nas distribuição dos *valores p*. No entanto, os testes WTS e USP mostraram ser os mais potentes no teste à interação.

A ANOVA paramétrica tem um comportamento estável e são poucos os métodos que a superam. No caso do erro de Tipo I, a ANOVA apenas tem um desempenho mais fraco que os testes ATS, L de PS, van der Waerden e WTPS, e um desempenho sempre melhor que os testes WTS e USP. No entanto, a ANOVA é mais potente que os testes L de PS e van der Waerden e menos potente que os testes USP e WTS. De um modo geral, a ANOVA comporta-se melhor no estudo dos efeitos principais do que no estudo da interação, quando comparada com as alternativas não paramétricas.

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

AKRITAS, M. G., ARNOLD, S. F. & BRUNNER, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs, *Journal of the American Statistical Association*, 92, 258-265.

BASSO, D., CHIARANDINI, M. & SALMASO, L. (2007) Synchronized permutation tests in replicated I x J designs, *Journal of Statistical Planning and Inference*, 137, 2564-2578.

BEASLEY, T. M. & ZUMBO, B. D. (2009). Aligned rank tests for interactions in split-plot designs: distributional assumptions and stochastic heterogeneity, *Journal of Modern Applied Statistical Methods*, 8, 16-50.

BEASLEY, T. M., ERICKSON, S. & ALLISON, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited?, *Behavior genetics*, 39, 580–595.

BRUNNER, E., DETTE, H. & MUNK, A. (1997). Box-type approximations in nonparametric factorial designs, *Journal of the American Statistical Association*, 92, 1494-1502.

CONOVER, W. J., & IMAN, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs, *Communication in Statistics – Theory and Methods*, 5, 1349-1368.

HAHN, S. & SALMASO, L. (2017). A comparison of different synchronized permutation approaches to testing effects in two-level two-factor unbalanced ANOVA designs, *Statistical Papers*, 58, 123-146.

HIGGINS, J. J. & TASHTOUSH, S. (1994). An aligned rank transform test for interaction, *Nonlinear World*, 1, 201-211.

LEYS, C. & SCHUMANN, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test, *Journal of Experimental Social Psychology*, 46, 684-688.

LUEPSEN, H. (2017). The aligned rank transform and discrete variables: a warning, *Communications in Statistics - Simulation and Computation*, 46, 6923-6936.

MANSOURI, H. & CHANG, G.-H. (1995). A comparative study of some rank tests for interaction, *Computational Statistics & Data Analysis*, 19, 85-96.

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.

PESARIN, F. (2001). *Multivariate permutation tests with applications in biostatistics,* Wiley & Sons, Chichester.

PETERSON, K. (2002). Six modifications of the aligned rank transform test for interaction, *Journal of Modem Applied Statistical Methods*, 1(1), 100-109.

PURI, M. L. & SEN, P. K. (1985). *Nonparametric methods in General Linear Models*, Wiley, New York.

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/.

SALMASO, L. (2003). Synchronized permutation tests in 2k factorial designs, *Communication in Statistics*, 32, 1419-1437.

SAWILOWSKY, S. S. (2000). Review of the rank transform in designed experiments, *Perceptual and motor skills*, 90, 489-497.

SHAH, D. A. & MADDEN, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments, *Phytopathology*, 94,33-43.

TOOTHAKER, L. E. & NEWMAN, D. (1994). Nonparametric competitors to the two-way ANOVA, *Journal of Educational Statistics*, 19, 237-273.

VAN DER WAERDEN, B. L. (1952). Order tests for the two-sample problem and their power, in *Indagationes Mathematicae (Proceedings)*, 55, 453-458.