Ciência Rural, Santa Maria, v.53:7, e20220146, 2023
http://doi.org/10.1590/0103-8478cr20220146
ISSNe 1678-4596
CROP PRODUCTION

# Familywise type I error of ANOVA and ANOVA on ranks in factorial experiments

**André Mundstock Xavier de Carvalho**[1,2*] **Matheus Romano de Souza**[2] **Tadeu Bernardes Marques**[2] **Davi Leite de Souza**[1] **Emanuel Fernando Maia de Souza**[3]

[1]Programa de Pós-graduação em Agronomia, Produção Vegetal, Universidade Federal de Viçosa (UFV), 38810-000, Rio Paranaíba, MG, Brasil. E-mail: andre.carvalho@ufv.br. *Corresponding author.
[2]Universidade Federal de Viçosa (UFV), Rio Paranaíba, MG, Brasil.
[3]Programa de Pós-graduação em Desenvolvimento Regional e Meio Ambiente, Programa de Pós-graduação em Ciências Ambientais, Universidade Federal de Rondônia (UNIR), Rolim de Moura, RO, Brasil.

**ABSTRACT**: This research evaluated the importance of a preliminary general analysis of variance (ANOVA) in the interpretation of data from factorial experiments under total nullity. For this, we evaluated the familywise type I error rate (accumulated FWER) of the F test for the unfolding of factorial ANOVA and factorial ANOVA on ranks, which were compared with the FWER for the global effect of treatments. In addition, we evaluated the FWER of the Tukey's test under total nullity for factorial experiments in the presence or absence of preliminary ANOVA protection (omnibus F test). The study was conducted by simulating data from 2,000 experiments, which were separated into four representative agricultural research scenarios. For both the parametric factorial ANOVA and the non-parametric factorial ANOVA, the FWER significantly exceeded the nominal level of 5%, even under total nullity. While the tests that control the total FWER in the factorials are not being used, the factorial ANOVA should not be performed without the preliminary ANOVA F test showing a significant effect. This, of course, does not apply to tests that are not multiple comparisons, such as Bonferroni, Dunn-Sidak and others, which do not need ANOVA protection. The same recommendation applies to the factorial ANOVA on ranks.
**Key words**: omnibus F-test, aligned rank transformation, multiplicity in multiway ANOVA.

## Erro tipo I familiar da ANOVA e da ANOVA on ranks em experimentos fatoriais

**RESUMO**: Este trabalho teve como objetivo avaliar a importância da análise de variância (ANOVA) geral preliminar na interpretação dos dados de experimentos fatoriais sob nulidade total. Para isso, avaliou-se as taxas de erro tipo I familiar (FWER acumulada) do teste F para os componentes do desdobramento da ANOVA fatorial e da ANOVA fatorial *on ranks*, em comparação à FWER para o efeito global de tratamentos. Além disso, avaliou-se a FWER do teste de Tukey sob nulidade total para experimentos fatoriais na presença ou ausência da proteção da ANOVA preliminar (teste F global). O estudo foi conduzido a partir da simulação de dados de 2000 experimentos, separados em quatro cenários representativos da pesquisa agrícola. Tanto para a ANOVA fatorial paramétrica, quanto para a ANOVA fatorial não-paramétrica, as FWER ultrapassaram significativamente o nível nominal de 5%, mesmo sob nulidade total. Enquanto os testes que controlam a FWER total nos fatoriais não estiverem sendo utilizados, a ANOVA fatorial não deve ser realizada sem que o teste F da ANOVA preliminar acuse um efeito significativo. O mesmo, evidentemente, não se aplica aos testes que não são de comparações múltiplas, como Bonferroni, Dunn-Sidak e outros, que não precisam proteção da ANOVA. A mesma recomendação se aplica à ANOVA fatorial *on ranks*.
**Palavras-chave**: teste F global, transformação de postos alinhada, multiplicidade na ANOVA fatorial.

In agricultural experimentation, factorial designs stand out from unstructured experiments for at least two major advantages. First, they allow a formal estimation of the interaction effect, ensuring safer generalizations. Second, factorials allow the reduction of the total number of two-by-two comparisons to be performed by multiple comparison tests. This allows a higher level of sensitivity of the multiple comparison tests (CARVALHO et al., 2023).

A non-parametric counterpart for the factorial ANOVA is the on-rank with aligned-rank transformation (ART) for the interaction, which allows a valid estimation of the significance of the interaction (DURNER, 2019). The approach via rank transformation is a valid technique of great utility, since it allows the unification and simplification of a whole set of non-parametric procedures (CONOVER, 2012; MONTGOMERY, 2017).

The analysis of variance (ANOVA) of a factorial experiment is performed by unfolding the sum of squares of the treatments in parts due to the main effects of each factor (A and B) and

Received 03.16.22 Approved 08.03.22 Returned by the author 09.23.22
CR-2022-0146.R1
Editors: Leandro Souza da Silva Marcos Toebe

the interaction (AxB). In the ANOVA of factorial experiments, it is common to proceed directly to the unfolding of the global effect of treatments, only calculating the F for the interaction between the factors and the isolated effects of the factors. However, a few statisticians recommend that this unfolding, even when orthogonal, should not be performed if the general ANOVA does not point to a significant effect of the treatments as a whole (BANZATO & KRONKA, 2006; CRAMER et al., 2016). This recommendation is either unknown or largely ignored in mainstream experimental statistics textbooks and the analysis packages of leading applications, which can lead to high familywise Type I error rates (FWER), since later means tests can indicate significant differences when a protection criterion is not used, such as the F test (in factorial experiments). Thus, this study evaluated the importance of the preliminary general ANOVA in the analysis of factorial experiments and the empirical accumulated FWER of the F test for components of the unfolding of the factorial ANOVA and factorial ANOVA on ranks. An additional aim was to evaluate the empirical accumulated FWER of the Tukey test under total nullity for factorial experiments in the presence or absence of the preliminary general ANOVA.

The study was conducted from the simulation of data from 2,000 experiments that were separated into two initial groups: a group of experiments with a 2 × 5 factorial structure and five repetitions, and another group of experiments with a 5 × 5 factorial structure and three repetitions. Each group was subdivided into two groups: those with higher coefficient of variation (CV) values (between 25% and 40%) and those with lower CV values (between 1% and 10%), totaling four scenarios with 500 experiments each. This sample number was defined considering the power of the one-sided Binomial test (for a proportion) in relation to the magnitude of the expected FWER. The data were simulated by considering a completely randomized model for a factorial structure ($y_{ijk} = m + a_i + b_j + a_ib_j + e_{ijk}$), where: $m$ is the overall average of the observations in each simulated experiment; $a_i$ is the average effect of factor a, level $i$; $b_j$ is the average effect of factor b, level $j$; $a_ib_j$ is the average effect of the interaction between factors a and b; and $e_{ijk}$ is the error estimate. Only the error component ($e_{ijk} \sim NID(0,\sigma)$) was considered as a random variable.

Data were simulated in Apache Open Office - Calc 4.1.7. We used the function "=NORMINV(RAND();mean;standard deviation)" for generating error values with a normal distribution, a procedure similar to that used by SOUSA et al. (2012). First, random values between 1–10 or 25–40 were generated, which were duly converted to feed the "standard deviation" parameter of the previously described function.

All data were previously submitted to the Jarque-Bera and Bartlett tests to verify the conditions of normality and homoscedasticity, respectively. When a simulated experiment did not meet one of these assumptions, it was discarded and replaced by another. The 2,000 simulated experiments were then individually submitted to an ANOVA following the factorial structure of the treatment decomposition into factors A, B, and the interaction. The preliminary F test (omnibus F test) was also performed for the global effect of treatments only, with a nominal α error of 5% always considered as a critical value. The means were then compared by the Tukey test to verify that the false positives indicated by the F test were also indicated by this standard test. If at least one significant difference was identified by the test, the result was counted as a false positive (accumulated FWER or EWER). The same procedure was also performed for the data submitted to an ANOVA on ranks (CONOVER, 2012), with the ART used to estimate the interaction (DURNER, 2019). Analyzes were performed using BioEstat 5.0, Microsoft Excel® and SPEED Stat 2.5 (CARVALHO et al., 2020).

The analysis of the simulated data showed that the empirical FWER for each of the isolated components of the factorial was close to the 5% limit (Table 1). However, according to the most commonly used interpretation of an ANOVA, it was sufficient for only one of the factorial components (A, B, or AxB) to be significant for it to consider that there was at least one mean that differed from the others. In this case, the FWER oscillated between 10.4% and 14.0% for the different scenarios (Table 1). However, when a preliminary ANOVA was applied, the FWER ranged between 2.6% and 4.2% (Table 1). That is, the usual interpretation of a factorial ANOVA, which disregards the previous verification of the significance of the F test for treatments, led to inflated FWERs even under total nullity. Although, this inflation was expected and is well known (since FWER = $1-(1-\alpha)^k$), it is important to demonstrate it empirically, as it is generally assumed that this problem occurs only under partial nullity (FRANE, 2021). These results; therefore, corroborate the recommendations of FLETCHER et al. (1989) and CRAMER et al. (2016). However, believing in the global F implies in the problem that the F for treatments reduces its

Table 1 - Percentage of experiments in which factorial ANOVA or preliminary ANOVA or Tukey's test indicated the existence of significant effects ($P \leq 0.05$) (empirical accumulated FWER or EWER) in the 500 experiments of each of the four scenarios evaluated.

| Source of Variation | -----2x4 factorial---- | | -----5x5 factorial---- | |
|---|---|---|---|---|
| | low CV | high CV | low CV | high CV |
| | ----------%---------- | | | |
| Treatments (preliminary ANOVA or omnibus F test) | 3.0 | 2.6 | 4.2 | 4.2 |
| Factor A | 3.8 | 3.4 | 6.2 | 3.4 |
| Factor B | 2.8 | 4.2 | 3.2 | 5.8 |
| Interaction AxB | 4.8 | 3.0 | 6.0 | 3.6 |
| Total (A, B and/or interaction) | 11.0* | 10.6* | 14.0* | 12.0* |
| Total (A, B and/or unfolding of the interaction (A/B's and B/A's) when the interaction was significant) | 11.0* | 10.4* | 13.8* | 11.8* |
| At least one false positive by the Tukey test (with factorial ANOVA protection) | 11.0* | 10.4* | 13.4* | 10.4* |
| At least one false positive by the Tukey test (with omnibus F test protection) | 3.0 | 2.6 | 4.2 | 4.2 |

Values followed by "*" indicate FWER statistically higher than 5% by the one-sided Binomial test (n=500; $P < 0.05$). Experiments under total nullity.

power as the number of treatments increases (LAZIC, 2018), which is especially relevant for factorials.

The analysis of the simulated data after the rank transformation (the ANOVA on ranks with ART to estimate the interaction) also showed that the FWER for each of the isolated components of the factorial were close to the nominal value of 5% (Table 2). Likewise, according to the usual interpretation for the factorial ANOVA on ranks, it was sufficient that only one of the factorial components (A, B, or AxB) was significant for it to be considered that there was at least one mean that differed from the others. In this case, the FWER fluctuated between 9.4 and 12.8% for the different scenarios considered (Table 2).

However, when the preliminary on ranks ANOVA was applied (to test the overall significance for the F of the treatments), the FWER ranged between 2.4% and 5.2% (Table 2). That is, like the parametric factorial ANOVA, the on ranks factorial ANOVA also led to inflated FWERs if we did not previously consider the significance for the omnibus F test.

Furthermore, disregarding the preliminary ANOVA had an impact on the FWER of the Tukey test. If no protection criteria were applied to the Tukey test, its FWER fluctuated between 21.4% and 38.4% (because in factorials the Tukey test only controls the FWER in each subfamily of comparisons). If we considered that the test of means should only be applied when one of the factors (A, B, or interaction) was significant, these error rates ranged between 10.4% and 13.4% (Table 1). Even if we only considered the F tests for the ramifications of the interaction (B's within $A_i$ and A's within $B_j$), these error rates did not approach acceptable levels (Table 1). The FWER of the Tukey's test was ≤ 5% when Fisher's protection criterion was considered using the significance of F for the global effect of treatments (preliminary ANOVA) (Table 1). These results; therefore, showed that the conclusions obtained by RODRIGUES (2015) do not apply to a factorial ANOVA under total nullity. As long as the tests of means are not adapted to control for the total FWER in the factorials, the general ANOVA continues to be useful to some extent.

As with the parametric factorial ANOVA, if no protection criteria were applied to the Tukey on ranks test, the FWERs greatly exceeded the nominal level of 5% (data not shown). Similarly, this problem can be avoided by the simple inclusion of omnibus F test (Table 2). The current discussion on the validity of non-parametric factorial ANOVA procedures (LUEPSEN, 2018; HARRAR et al., 2019) also requires this fact to be considered before suggesting the non-application of the ART for some situations.

Therefore, under total nullity in both the factorial ANOVA and the non-parametric factorial ANOVA, the control of the FWER will be guaranteed by the preliminary ANOVA (the significance of the F for the treatments) and not by the significance of the factorial components (A, B, and interaction). This could to be included in the routines of several analysis software to reduce the frequency of erroneous

Table 2 - Percentage of experiments in which the factorial ANOVA on ranks or the preliminary ANOVA on ranks or the Tukey on ranks test indicated the existence of significant effects ($P \leq 0.05$) (empirical accumulated FWER or EWER) in the 500 experiments of each of the four scenarios evaluated.

| Source of Variation (ANOVA on ranks) | -----2x4 factorial----- | | -----2x5 factorial----- | |
|---|---|---|---|---|
| | low CV | high CV | low CV | high CV |
| | ----------%---------- | | | |
| Treatments (preliminary ANOVA or omnibus F test) | 3.0 | 2.4 | 5.2 | 5.2 |
| Factor A | 3.6 | 3.4 | 5.6 | 3.8 |
| Factor B | 3.0 | 3.6 | 3.0 | 4.8 |
| Interaction AxB (*ART*[1]) | 5.2 | 2.6 | 4.8 | 4.0 |
| Total (A, B and/or interaction) | 10.8* | 9.8* | 12.8* | 11.6* |
| Total (A, B and/or interaction unfolding (A/B's and B/A's), when the interaction was significant) | 10.0* | 9.4* | 12.0* | 11.4* |
| At least one false positive by the Tukey on ranks test (with factorial ANOVA protection) | 10.4* | 9.8* | 11.2* | 10.4* |
| At least one false positive by Tukey on ranks test (with omnibus F test protection) | 3.0 | 2.4 | 5.2 | 5.2 |

[1]ART: aligned rank transformation to estimate the interaction. Values followed by "*" indicate FWER statistically higher than 5% by the one-sided Binomial test (n=500; $P < 0.05$). Experiments under total nullity.

statistical conclusions, such as in Minitab, Assistat, Sisvar, R (packages Easyanova, FrF2, Agricolae, among others).

Furthermore, the full unfolding of the interaction by the F test does not replace the preliminary ANOVA to ensure the control of the FWER for multiple comparisons, even under total nullity. The use of the Fisher protection criterion in the preliminary ANOVA possibly eliminates further concerns with the full unfolding of the interaction in the ANOVA.

Finally, the data allowed us to deduce that if the Tukey test needs an ANOVA to control the FWER in factorials under total nullity, uncontrolled FWER will always occur under partial nullity. This means the insertion of corrected or adapted versions of this and other multiple comparison tests is urgently required for FWER correction in the factorial ANOVAs of commonly used software and analysis packages in the agricultural sciences.

## ACKNOWLEDGEMENTS

We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil (CNPq); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil - Finance code 001; Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and Éder Matsuo for the suggestions.

## DECLARATION OF CONFLICT OF INTEREST

We have no conflict of interest to declare.

## AUTHORS' CONTRIBUTIONS

The authors contributed equally to the manuscript.

## REFERENCES

BANZATO, D. A.; KRONKA, S. N. **Experimentação Agrícola**. Jaboticabal: Funep, 2006.

CARVALHO, A. M. X.; et al. A brief review of the classic methods of experimental statistics. **Acta Scientiarum – Agronomy**, v.45: *in press*, 2023.

CARVALHO, A. M. X.; et al. SPEED Stat: a free, intuitive, and minimalist spreadsheet program for statistical analyses of experiments. **Crop Breeding and Applied Biotechnology**, v.20(3): e327420312, 2020. Available from: <https://doi.org/10.1590/1984-70332020v20n3s46>. Accessed: Jan. 18, 2021.

CONOVER, W. J. The rank transformation - an easy and intuitive way to connect many nonparametric methods to their parametric counterparts for seamless teaching introductory statistics courses. **WIREs Computational Statistics**, v.4, p.432-438, 2012. Available from: <https://doi.org/10.1002/wics.1216>. Accessed: Jan. 18, 2021.

CRAMER, A. O. J. et al. Hidden multiplicity in exploratory multiway ANOVA: prevalence and remedies. **Psychonomic Bulletin & Review**, v.23, p. 640–647, 2016. Available from: <https://doi.org/10.3758/s13423-015-0913-5>. Accessed: Jan. 18, 2021.

DURNER, E. Effective Analysis of Interactive Effects with Non-Normal Data Using the Aligned Rank Transform, ARTool and SAS® University Edition. **Horticulturae**, v.5, p.57-70, 2019. Available from: <doi:10.3390/horticulturae5030057>. Accessed: Jan. 18, 2021.

FLETCHER, H. J. et al. Controlling Multiple F Test Errors with an Overall F Test. **The Journal of Applied Behavioral Science**, v.25, p.101-108, 1989. Available from: <https://doi.org/10.1177/0021886389251008>. Accessed: Jan. 18, 2021.

FRANE, A.V. Experiment-Wise Type I Error Control: A Focus on 2 × 2 Designs. **Advances in Methods and Practices in Psychological Science**, v.4, p.1-20, 2021. Available from: <https://doi.org/10.1177/2515245920985137>. Accessed: Jan. 18, 2021.

HARRAR, S.W. et al. A comparison of recent nonparametric methods for testing effects in two-by-two factorial designs. **Journal of Applied Statistics**, v.46, p.1649-1670, 2019. Accessed: Jan. 18, 2021.

LAZIC, S. E. Four simple ways to increase power without increasing the sample size, **Laboratory Animals**, v.52, p.621-629, 2018. Available from: <https://doi.org/10.1177/0023677218767478>. Accessed: Jan. 18, 2021.

LUEPSEN, H. Comparison of nonparametric analysis of variance methods: A vote for van der Waerden. **Communications in Statistics - Simulation and Computation**, v.47, p.2547-2576, 2018. Available from: <https://doi.org/10.1080/03610918.2017.1353613>. Accessed: Jan. 18, 2021.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 9th Ed. Danvers: Wiley, 2017.

RODRIGUES, J. **Um estudo sobre testes de comparação de médias e sua aplicação condicional a um teste F global significativo na análise de variância**. 2015. 164p. Tese – Doutorado em Estatística e Experimentação Agronômica, USP.

SOUSA, C. A.; et al. Avaliação de testes estatísticos de comparações múltiplas de médias. **Ceres**, v.59, p.350-354, 2012. Available from: <https://doi.org/10.1590/S0034-737X2012000300008>. Accessed: Jan. 18, 2021.