Acta Scientiarum

http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v45i1.57742

Type I error in multiple comparison tests in analysis of variance

Josiane Rodrigues[1*], Sônia Maria De Stefano Piedade[2], Idemauro Antonio Rodrigues de Lara[2] and Francisco Humberto Henrique[3]

[1]Departamento de Tecnologia Agroindustrial e Socioeconomia Rural, Centro de Ciências Agrárias, Universidade Federal de São Carlos, Rodovia Anhanguera, km 174, 13600-970, Araras, São Paulo, Brazil. [2]Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, São Paulo, Brazil. [3]HNRQ AgTec, Rio Claro, São Paulo, Brazil. *Author for correspondence. E-mail: josirodrigues@ufscar.br

ABSTRACT. In a hypothesis test, a researcher initially fixes a type I error rate, that is, the probability of rejecting the null hypothesis given that it is true. In the case of means tests, it is important to present a type I error that is equal to the nominal pre-fixed level, such that this error remains unchanged across various scenarios, including the number of treatments, number of repetitions, and coefficient of variation. The purpose of this study is to analyse and compare the following multiple comparison tests for the control of both conditional and unconditional type I error rates, depending on a significant F-test in the analysis of variance: Tukey, Duncan, Fisher's least significant difference, Student–Newman–Keuls (SNK), and Scheffé. As an application, we present a motivation study and develop a simulation study using the Monte Carlo method for a total of 64 scenarios. In each simulated scenario, we estimate the comparison-wise and experiment-wise error rates, conditional and unconditional on a significant result of the overall F-test of analysis of variance for each of the five multiple comparison tests evaluated. The results indicate that the application of the means tests based only on the significance of the F-test should be considered when determining the error rates, as this can change them. In addition, we find that Fisher's test controls for the comparison-wise error rate, the Tukey and SNK tests control for the experiment-wise error rate, and the Duncan and Fisher tests control for the conditional experiment-wise error rate. Scheffé's test does not control for any of the error rates considered.

Keywords: comparison-wise error rate; experiment-wise error rate; means tests; Monte Carlo simulation.

Received on February 9, 2021.
Accepted on June 17, 2021.

Introduction

In agricultural experiments, a common problem arises while conducting a comparison among treatments of interest to determine the existence of a difference between them. The most common solution to this problem lies in the application of the analysis of variance (ANOVA) (Girardi, Cargnelutti Filho, & Storck, 2009).

The overall F-test in ANOVA checks the hypothesis of equality for the population means of the treatments. If the F-test is significant, then a comparison of means test is performed to investigate the possible differences between the pairs of specific means or a linear combination of them (Saville, 2014).

One of the dilemmas involved in the means tests is their conditional application to a significant F-test. According to Cardellino and Siewerdt (1992), this is a controversial question and should be investigated further. Rodrigues, Piedade, and Lara (2016), for example, noticed divergent results between the overall F-test and the means tests evaluated in their simulation study. In this motivational study, we present the agronomic experiment of Henrique and Laca-Buendía (2010), which compares five cultivars and a new genotype of cotton, and show divergent results between the overall F-test and certain means tests commonly used in agricultural research. Nevertheless, many authors recommend applying means tests only on a significant result of the F-test in ANOVA. Therefore, many questions remain to be answered in this field of study.

However, it is not possible to dissociate this study from the errors that can occur in a hypothesis test. This is because when the hypothesis for an average contrast is analysed, the test, whether or not applied based only on a significant result of the overall F-test, exhibits the probabilities of type I and type II errors, where the type I error rate can be of the comparison-wise or experiment-wise type (Ramos & Vieira, 2014). The comparison-wise error rate is the long-run proportion of the number of erroneous inferences observed divided by the total number of comparisons

made; the experiment-wise error rate is the long-run proportion of the number of experiments conducted with at least one erroneous inference divided by the total number of experiments (Boardman & Moffitt, 1971).

Several studies have been conducted to evaluate the means tests in relation to the type I error rate control and to propose modifications in the tests that are aimed at controlling this rate (Biase & Ferreira, 2011; Souza, Lira Junior, & Ferreira, 2012; Gonçalves, Ramos, & Avelar, 2015). However, the analysis of these concepts associated with the significance of the F-test requires an additional investigation.

In this context, the aim of this study is to analyse and compare the Tukey, Duncan, Fisher's least significant difference (Fisher's LSD), Student–Newman–Keuls (SNK), and Scheffé tests, which are used for conducting pair-wise comparisons between the means, with respect to the control of type I error rates, which are conditional and unconditional to a significant result of the overall ANOVA F-test.

Material and methods

Motivational study

We present an experiment conducted by Henrique and Laca-Buendía (2010), in which the aim was to compare five cultivars and a new genotype of cotton (*Gossypium hirsutum* L. r. *latiFolium* Hutch). The experiment was conducted in Uberaba, Minas Gerais State, Brazil, located at longitude 47°57'22" WGR, latitude 19°44'6.82" S, and an altitude of 775 m.

The experiment was carried out in a randomised block design with six treatments and four replicates. The plots consisted of four lines of 5 m with a 0.7 m spacing between the lines. The two central lines were considered to have a useful area of 3.5 m^2; the other two lines, one on each side, were the borders, with each line being an equivalent of a total of 10 plants. In agricultural experiments, it is quite common to use a number of repetitions equal to four, since the plots cover larger areas and more than one 'individual' is used per plot.

In the experiment, the following varieties were compared: Delta Opal (Delta and Pine), Delta Penta (Delta and Pine), BRS-Cedro (EMBRAPA), IAC-25 (IAC), EPAMIG Precoce I, and the progeny IAC-06/191 (IAC). The height of the first productive branch (average distance from the soil to the first branch in which there were bolls, in centimetres) and final stand (total number of plants at harvest time) were among the evaluated traits.

In the context of the variable height of the first productive branch, the overall ANOVA F-test is found to be significant at the 5% level, but the Scheffé test shows no difference between the means of the treatments at the same level of significance (Table 1).

Table 1. Multiple comparisons applied to the data of the height of the first productive branch, in centimetres, for five cultivars and a new genotype of cotton.

ANOVA F-test = 4,664*						
Treatment	Mean	Tukey	Duncan	Fisher's LSD	SNK	Scheffé
IAC-06/191	28.45	a	a	a	a	a
BRS-Cedro	25.55	ab	a	a	ab	a
IAC-25	22.25	ab	ab	ab	ab	a
Delta Opal	21.85	ab	ab	ab	ab	a
Delta Penta	18.15	b	b	b	b	a
EPAMIG Precoce I	17.40	b	b	b	b	a

*Significant at the 5% level; means followed by the same letter do not differ from each other at a significance level of 5%. The assumptions of normality and homoscedasticity were verified using the Shapiro–Wilk test (p-value = 0.3399) and Bartlett test (p-value = 0.2617), respectively.

For the variable final stand, the ANOVA F-test is not significant, but Duncan's and Fisher's LSD tests show differences between some of the treatment means (Table 2).

Table 2. Multiple comparisons applied to the data of the final stand of five cultivars and a new genotype of cotton.

ANOVA F-test = 2,648^{ns}						
Treatment	Mean	Tukey	Duncan	Fisher's LSD	SNK	Scheffé
Delta Penta	74.50	a	a	a	a	a
Delta Opal	73.75	a	a	a	a	a
BRS-Cedro	70.00	a	ab	a	a	a
IAC-25	66.75	a	ab	ab	a	a
IAC-06/191	57.50	a	ab	ab	a	a
EPAMIG Precoce I	51.50	a	b	b	a	a

^{ns}Not significant at the 5% level; means followed by the same letter do not differ from each other at a significance level of 5%. The assumptions of normality and homoscedasticity were verified using the Shapiro–Wilk test (p-value = 0.1575) and Bartlett test (p-value = 0.6062), respectively.

These results serve as the basis for establishing scenarios to study the type I error rates presented here, since the means tests can be applied irrespective of whether the F-test is significant or not. A completely randomised design was used to facilitate the simulation process, although the results can be applied to any other type of experimental design.

Study of simulation

A total of 128,000 experiments were simulated using the Monte Carlo method with 2,000 for each scenario for a total of 64 cases consisting of a combination of the following factors: 3, 5, 7, or 9 treatments; 3, 4, 10, or 20 replicates; and the coefficient of variation (CV) of 1, 5, 10, or 20%, without considering the treatment effect. The experiments were simulated using a completely randomised design

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, \ldots, a, \qquad j = 1, \ldots, r,$$

where y_{ij} represents the simulated value of the response obtained with the i-th treatment in its j-th repetition, μ is the mean that is arbitrarily set at 100, τ_i is the fixed effect of the i-th treatment (which is considered to be null), and ε_{ij} is the random error generated independently with a normal distribution having zero mean and a standard deviation (σ) varying according to the desired CV.

In all simulated scenarios, we ensured that the analyses were conducted based on the same randomly used seed, such that possible differences did not occur due to a random error of the simulation process, but rather due to the differences between the tests. Moreover, the nominal significance level adopted in all the cases was 5%. To simulate the experimental data for performing statistical analyses, an algorithm was developed using the R software (R Core Team, 2020).

Consequently, for each simulated scenario, the type I error rates were estimated by the comparisons and experiments conducted for each of the five tests (Tukey, Duncan, Fisher's LSD, SNK, and Scheffé). Furthermore, these error rates were considered in two different aspects: (i) the multiple comparison procedure was applied regardless of the overall F-test result, and (ii) it was applied only when the F-test was significant.

The comparison-wise error rate (α_c) is defined as the ratio of the number of erroneous inferences, such that $\mu_i \neq \mu_{i'}$ when $\mu_i = \mu_{i'}$, to the total number of comparisons performed. Thus, by taking the demonstrative scenario of a = 3, r = 3, and CV = 1%, the unconditional comparison-wise error rate can be estimated from the ratio of the total number of erroneous inferences to the total number of inferences (in this case, 2,000 experiments × 3 contrasts per experiment = 6,000 contrasts).

In the case where the means tests are applied only if the overall ANOVA F-test is significant, the comparison-wise error rate, that is conditional (α_1), can be estimated empirically based on the total number of type I errors observed in the experiments that presented a significant F-test and the total number of inferences made. Taking the demonstrative scenario of a = 3, r = 3, and CV = 1%, while considering that out of the 2,000 experiments simulated for this scenario, only 100 showed a significant result for the overall F-test, this error rate can be estimated by the ratio of the total number of erroneous inferences made in 100 experiments to the total number of inferences (6,000 contrasts).

A second way to estimate the conditional comparison-wise error rate (α_2) is to take the total number of erroneous inferences within the experiments with a significant result for the overall F-test and divide it by the total number of inferences made within these experiments. Considering the same scenario where a = 3, r = 3, and CV = 1% and the number of 100 experiments with a significant overall F-test, this rate can be estimated by the ratio of the total number of erroneous inferences made in the 100 experiments to the total number of inferences made within these experiments (in this case, 100 experiments × 3 contrasts per experiment = 300 contrasts).

Accordingly, the experiment-wise error rate (α_e) is defined as the ratio of the number of experiments with at least one erroneous inference ($\mu_i \neq \mu_{i'}$ when $\mu_i = \mu_{i'}$) divided by the total number of experiments. Thus, considering the scenario where a = 3, r = 3, and CV = 1%, the unconditional experiment-wise error rate can be estimated by the ratio between the number of trials with at least one erroneous inference among the three tested to the total number of experiments (in this case, 2,000).

The conditional experiment-wise error rate (α_3) can be estimated by taking the total number of experiments that presented a significant overall F-test as well as at least one comparison resulting in a type I error, which is divided by the total number of experiments. Considering that for the scenario where a = 3, r = 3, and CV = 1%, out of the 2,000 experiments simulated, only 100 presented a significant result for the overall

F-test, this error rate can be estimated by the ratio between the number of experiments (out of the 100 experiments) that presented at least one comparison resulting in a type I error and the total number of experiments (2,000).

Moreover, the same rate can be estimated based on the experiments that presented a significant result in the overall F-test. This rate (α_4) can then be calculated by dividing the total number of trials that presented a significant F-test and at least one comparison resulting in a type I error by the total number of trials out of the 2,000 trials with a significant F-test. For the scenario with a = 3, r = 3, and CV = 1%, considering that out of the 2,000 experiments simulated, only 100 presented a significant result for the overall F-test, this rate can be estimated by dividing the number of experiments (out of the 100) that presented at least one comparison resulting in a type I error and a total of 2,000 experiments with a significant overall F-test (in this case, 100).

To verify whether each of the rates differed from the established nominal significance level (α = 5%), a lower limit of 0.038 and an upper limit of 0.063 were used, which were calculated from a 95% confidence interval (CI) for the proportion $\hat{p}$ = 0.05, expressed as

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{2{,}000}},$$

Where $z_{\alpha/2}$ is the quantile value of the standard normal distribution at the α level of significance. Thus, the rates within this interval were not considered to be different from the established nominal level.

Results and discussion

In the 64 scenarios formed by the combination of the number of treatments (a = 3, 5, 7, and 9), the number of repetitions (r = 3, 4, 10, and 20), and the CV (CV = 1%, 5%, 10%, and 20%), the error rates, do not present differences with a high magnitude in the variation of the number of repetitions and the CV (Tables 3 to 6). The same results were noted by Girardi et al. (2009), who presented the comparison-wise and experiment-wise error rates for 80 scenarios formed by a variation in the number of treatments, number of repetitions, and CV of the experiments.

In all the simulated scenarios, the comparison-wise error rates are lower than the experiment-wise error rates (Tables 3 to 6) for the five multiple comparison tests evaluated, which is an expected result according to Girardi et al. (2009); this is because the equality between them would be obtained only if the totality of the contrasts was significant in all the experiments with at least a significant contrast.

Regarding the comparison-wise error rates α_c, the behaviour of the Tukey, SNK, and Scheffé tests are similar, since the presented estimates for this rate lie below the lower limit of the 95% CI calculated (0.0375) in all simulated scenarios (Tables 3 to 6). In general, an increase in the number of treatments causes a decrease in the error rate for these tests.

Further, the Duncan test presents the estimates for α_c below the lower limit of the 95% CI in most scenarios. However, for CV = 1%, the test shows that this rate can be controlled in most scenarios where a = 3 and a = 5; while for CV = 5%, 10%, and 20%, this rate can be controlled in scenarios with a = 3. Similar to the Tukey, SNK, and Scheffé tests, an increase in the number of treatments causes a significant decrease in the error rate in the Duncan test.

For Fisher's LSD test, it is observed that the estimates obtained for the comparison-wise error rates α_c always remain within the 95% CI calculated (Tables 3 to 6). In this case, the variation in the number of treatments does not cause significant changes to the results.

Thus, with respect to α_c, Fisher's LSD test proves to be the only one that can control for this error rate regardless of the number of treatments, repetitions, and the CV of the experiments; therefore, it is the most robust test in this case. The Duncan test, in turn, lies in an intermediate situation; it controls for this rate only in the scenarios in which the number of treatments is small, and is conservative in other cases. The Tukey, SNK, and Scheffé tests exhibit the worst performance; they are conservative in all simulated scenarios. In this case, the Scheffé test is the most conservative, followed by the Tukey and SNK tests.

In the context of the experiment-wise error rates α_e, the Tukey and SNK tests exhibit a similar behaviour by always presenting the estimates for this rate within the 95% CI (Tables 3 to 6). In both the cases, a variation in the number of treatments does not significantly influence the results.

Table 3. Type I error rates for multiple comparison tests according to the number of treatments (a) and the number of repetitions of treatments (r) with coefficient of variation = 1%.

Unconditional rates on a significant ANOVA F-test											
a	r	Tukey		Duncan		Fisher's LSD		SNK		Scheffé	
		α_c	α_e	α_c	α_e	α_c	α_e	α_c	α_e	α_c	α_e
3	3	0.0177•	0.0395	0.0430	0.0950••	0.0500	0.1120••	0.0257•	0.0500	0.0175•	0.0390
3	4	0.0193•	0.0465	0.0443	0.0980••	0.0488	0.1115••	0.0273•	0.0570	0.0133•	0.0335•
3	10	0.0192•	0.0490	0.0408	0.0965••	0.0482	0.1165••	0.0262•	0.0575	0.0153•	0.0395
3	20	0.0170•	0.0445	0.0362•	0.0850••	0.0425	0.1030••	0.0238•	0.0510	0.0128•	0.0325•
5	3	0.0079•	0.0495	0.0390	0.1800••	0.0513	0.2420••	0.0123•	0.0620	0.0033•	0.0225•
5	4	0.0074•	0.0540	0.0389	0.1990••	0.0533	0.2610••	0.0104•	0.0545	0.0033•	0.0260•
5	10	0.0071•	0.0480	0.0333•	0.1815••	0.0496	0.2625••	0.0103•	0.0515	0.0029•	0.0230•
5	20	0.0068•	0.0540	0.0316	0.1820••	0.0495	0.2780••	0.0094•	0.0585	0.0025•	0.0215•
7	3	0.0047•	0.0500	0.0370•	0.2770••	0.0551	0.3940••	0.0080•	0.0580	0.0014•	0.0175•
7	4	0.0032•	0.0455	0.0295•	0.2560••	0.0490	0.3930••	0.0049•	0.0505	0.0008•	0.0115•
7	10	0.0034•	0.0490	0.0265•	0.2605••	0.0479	0.4130••	0.0046•	0.0500	0.0004•	0.0080•
7	20	0.0030•	0.0455	0.0255•	0.2550••	0.0479	0.4290••	0.0041•	0.0460	0.0004•	0.0055•
9	3	0.0028•	0.0460	0.0288•	0.3125••	0.0494	0.4720••	0.0044•	0.0510	0.0003•	0.0080•
9	4	0.0027•	0.0530	0.0266•	0.3255••	0.0502	0.5025••	0.0046•	0.0545	0.0002•	0.0045•
9	10	0.0017•	0.0450	0.0231•	0.3225••	0.0578	0.5335••	0.0022•	0.0460	0.0001•	0.0025•
9	20	0.0019•	0.0475	0.0233•	0.3300••	0.0494	0.5470••	0.0025•	0.0480	0.0000•	0.0015•
Conditional rates on a significant ANOVA F-test											
		α_1	α_3	α_1	α_3	α_1	α_3	α_1	α_3	α_1	α_3
3	3	0.0175•	0.0390	0.0258•	0.0445	0.0258•	0.0445	0.0241•	0.0445	0.0175•	0.0390
3	4	0.0165•	0.0380	0.0218•	0.0395	0.0218•	0.0395	0.0218•	0.0395	0.0133•	0.0335•
3	10	0.0181•	0.0460	0.0253•	0.0515	0.0253•	0.0515	0.0238•	0.0510	0.0153•	0.0395
3	20	0.0161•	0.0420	0.0230•	0.0470	0.0230•	0.0470	0.0220•	0.0460	0.0128•	0.0325•
5	3	0.0073•	0.0445	0.0198•	0.0560	0.0210•	0.0560	0.0112•	0.0515	0.0032•	0.0225•
5	4	0.0062•	0.0425	0.0159•	0.0530	0.0188•	0.0530	0.0090•	0.0490	0.0033•	0.0260•
5	10	0.0062•	0.0390	0.0146•	0.0470	0.0162•	0.0470	0.0091•	0.0415	0.0029•	0.0230•
5	20	0.0055•	0.0420	0.0142•	0.0535	0.0166•	0.0535	0.0079•	0.0445	0.0024•	0.0215•
7	3	0.0045•	0.0480	0.0197•	0.0515	0.0230•	0.0515	0.0070•	0.0525	0.0014•	0.0175•
7	4	0.0026•	0.0330•	0.0097•	0.0415	0.0117•	0.0415	0.0037•	0.0345•	0.0007•	0.0115•
7	10	0.0026•	0.0340•	0.0092•	0.0430	0.0115•	0.0430	0.0038•	0.0350•	0.0003•	0.0080•
7	20	0.0023•	0.0315•	0.0085•	0.0430	0.0111•	0.0430	0.0032•	0.0315•	0.0003•	0.0055•
9	3	0.0024•	0.0350•	0.0112•	0.0500	0.0137•	0.0500	0.0038•	0.0385	0.0003•	0.0080•
9	4	0.0022•	0.0375	0.0099•	0.0510	0.0134•	0.0510	0.0036•	0.0385	0.0002•	0.0045•
9	10	0.0012•	0.0295•	0.0075•	0.0460	0.0113•	0.0460	0.0015•	0.0295•	0.0000•	0.0025•
9	20	0.0015•	0.0335•	0.0080•	0.0475	0.0110•	0.0475	0.0020•	0.0335•	0.0000•	0.0015•
		α_2	α_4	α_2	α_4	α_2	α_4	α_2	α_4	α_2	α_4
3	3	0.4487••	0.8764••	0.5805••	1.0000••	0.5805••	1.0000••	0.5753••	1.0000••	0.4487••	0.8764••
3	4	0.4342••	0.9620••	0.5527••	1.0000••	0.5527••	1.0000••	0.5527••	1.0000••	0.3980••	0.8481••
3	10	0.3949••	0.8932••	0.4919••	1.0000••	0.4919••	1.0000••	0.4863••	0.9902••	0.3881••	0.7669••
3	20	0.3849••	0.8936••	0.4893••	1.0000••	0.4893••	1.0000••	0.4888••	0.9787••	0.3948••	0.6914••
5	3	0.1651••	0.7946••	0.3544••	1.0000••	0.3750••	1.0000••	0.2472••	0.9196••	0.1444••	0.4017••
5	4	0.1470••	0.8018••	0.3009••	1.0000••	0.3556••	1.0000••	0.1989••	0.9245••	0.1269••	0.4905••
5	10	0.1589••	0.8297••	0.3117••	1.0000••	0.3446••	1.0000••	0.2287••	0.8829••	0.1260••	0.4893••
5	20	0.1321••	0.7850••	0.2663••	1.0000••	0.3102••	1.0000••	0.1827••	0.8317••	0.1139••	0.4018••
7	3	0.0952••	0.5889••	0.2424••	1.0000••	0.2822••	1.0000••	0.1400••	0.6441••	0.0816••	0.2147••
7	4	0.0793••	0.7951••	0.2352••	1.0000••	0.2839••	1.0000••	0.1125••	0.8313••	0.0662••	0.2771••
7	10	0.0777••	0.7906••	0.2159••	1.0000••	0.2685••	1.0000••	0.1088••	0.8139••	0.0476	0.1860••
7	20	0.0755••	0.7325••	0.1993••	1.0000••	0.2591••	1.0000••	0.1035••	0.7325••	0.0649••	0.1279••
9	3	0.0702••	0.7000••	0.2244••	1.0000••	0.2750••	1.0000••	0.1039••	0.7700••	0.0399	0.1600••
9	4	0.0600	0.7352••	0.1944••	1.0000••	0.2630••	1.0000••	0.0950••	0.7549••	0.0524	0.0882••
9	10	0.0428	0.6413••	0.1630••	1.0000••	0.2469••	1.0000••	0.0541	0.6413••	0.0388	0.0543
9	20	0.0456	0.7052••	0.1687••	1.0000••	0.2318••	1.0000••	0.0613	0.7052••	0.0277	0.0315

•Type I error rates below the lower limit of the 95% confidence interval (CI) (0.0375) for the empirical proportion of this rate. ••Type I error rates above the upper limit of the 95% CI (0.0625) for the empirical proportion of this rate.

Table 4. Type I error rates of multiple comparison tests according to the number of treatments (a) and the number of repetitions of treatments (r) with coefficient of variation = 5%.

Unconditional rates on a significant ANOVA F-test											
a	r	Tukey		Duncan		Fisher's LSD		SNK		Scheffé	
		α_c	α_e	α_c	α_e	α_c	α_e	α_c	α_e	α_c	α_e
3	3	0.0201•	0.0470	0.0433	0.0925••	0.0471	0.1040••	0.0288•	0.0540	0.0163•	0.0385
3	4	0.0215•	0.0510	0.0471	0.1025••	0.0516	0.1160••	0.0298•	0.0595	0.0160•	0.0365•
3	10	0.0201•	0.0495	0.0398	0.0930••	0.0471	0.1150••	0.0255•	0.0530	0.0165•	0.0410
3	20	0.0173•	0.0450	0.0383	0.0885••	0.0455	0.1090••	0.0246•	0.0545	0.0130•	0.0345•
5	3	0.0075•	0.0470	0.0364•	0.1720••	0.0483	0.2405••	0.0123•	0.0515	0.0034•	0.0255•
5	4	0.0064•	0.0445	0.0372•	0.1970••	0.0512	0.2645••	0.0101•	0.0495	0.0029•	0.0220•
5	10	0.0074•	0.0525	0.0340•	0.1840••	0.0502	0.2725••	0.0101•	0.0565	0.0030•	0.0225•
5	20	0.0059•	0.0480	0.0328•	0.1965••	0.0495	0.2815••	0.0081•	0.0530	0.0021•	0.0195•
7	3	0.0048•	0.0520	0.0354•	0.2765••	0.0522	0.3850••	0.0077•	0.0550	0.0013•	0.0140•
7	4	0.0035•	0.0450	0.0282•	0.2535••	0.0478	0.3910••	0.0051•	0.0470	0.0008•	0.0115•
7	10	0.0031•	0.0470	0.0251•	0.2410••	0.0449	0.4030••	0.0044•	0.0485	0.0005•	0.0100•
7	20	0.0025•	0.0405	0.0251•	0.2510••	0.0462	0.4160••	0.0038•	0.0420	0.0002•	0.0055•
9	3	0.0023•	0.0415	0.0285•	0.3315••	0.0622	0.4880••	0.0033•	0.0430	0.0003•	0.0075•
9	4	0.0040•	0.0445	0.0203•	0.3225••	0.0496	0.5085••	0.0058•	0.0460	0.0007•	0.0080•
9	10	0.0020•	0.0490	0.0228•	0.3160••	0.0475	0.5330••	0.0026•	0.0505	0.0001•	0.0035•
9	20	0.0035•	0.0520	0.0287•	0.3260••	0.0492	0.5535••	0.0030•	0.0530	0.0000•	0.0025•
Conditional rates on a significant ANOVA F-test											
		α_1	α_3	α_1	α_3	α_1	α_3	α_1	α_3	α_1	α_3
3	3	0.0193•	0.0445	0.0276•	0.0490	0.0276•	0.0490	0.0266•	0.0480	0.0163•	0.0385
3	4	0.0193•	0.0445	0.0276•	0.0485	0.0276•	0.0485	0.0256•	0.0485	0.0160•	0.0365•
3	10	0.0188•	0.0455	0.0243•	0.0480	0.0243•	0.0480	0.0238•	0.0475	0.0165•	0.0410
3	20	0.0158•	0.0405	0.0231•	0.0460	0.0231•	0.0460	0.0215•	0.0460	0.0130•	0.0345•
5	3	0.0068•	0.0405	0.0177•	0.0510	0.0188•	0.0510	0.0113•	0.0430	0.0034•	0.0255•
5	4	0.0056•	0.0365•	0.0158•	0.0465	0.0170•	0.0465	0.0092•	0.0400	0.0029•	0.0220•
5	10	0.0065•	0.0435	0.0153•	0.0515	0.0167•	0.0515	0.0088•	0.0460	0.0030•	0.0225•
5	20	0.0047•	0.0365•	0.0130•	0.0475	0.0144•	0.0475	0.0069•	0.0395	0.0021•	0.0195•
7	3	0.0043•	0.0410	0.0137•	0.0480	0.0154•	0.0480	0.0065•	0.0410	0.0013•	0.0140•
7	4	0.0030•	0.0345•	0.0099•	0.0425	0.0121•	0.0425	0.0042•	0.0355•	0.0008•	0.0115•
7	10	0.0025•	0.0355•	0.0095•	0.0440	0.0116•	0.0440	0.0037•	0.0365•	0.0005•	0.0100•
7	20	0.0020•	0.0285•	0.0087•	0.0410	0.0105•	0.0410	0.0029•	0.0285•	0.0002•	0.0055•
9	3	0.0021•	0.0345•	0.0103•	0.0445	0.0143•	0.0445	0.0030•	0.0355•	0.0003•	0.0075•
9	4	0.0034•	0.0350•	0.0135•	0.0510	0.0133•	0.0510	0.0049•	0.0365•	0.0007•	0.0080•
9	10	0.0016•	0.0330•	0.0080•	0.0480	0.0110•	0.0480	0.0020•	0.0330•	0.0001•	0.0035•
9	20	0.0026•	0.0360•	0.0116•	0.0465	0.0106•	0.0465	0.0024•	0.0370•	0.0000•	0.0025•
		α_2	α_4	α_2	α_4	α_2	α_4	α_2	α_4	α_2	α_4
3	3	0.4344••	0.9081••	0.5646••	1.0000••	0.5646••	1.0000••	0.5673••	0.9795••	0.4242••	0.7857••
3	4	0.4344••	0.9175••	0.5704••	1.0000••	0.5704••	1.0000••	0.5641••	1.0000••	0.4383••	0.7525••
3	10	0.4139••	0.9479••	0.5069••	1.0000••	0.5069••	1.0000••	0.5070••	0.9895••	0.4024••	0.8541••
3	20	0.3909••	0.8804••	0.5036••	1.0000••	0.5036••	1.0000••	0.4942••	1.0000••	0.3768••	0.7500••
5	3	0.1691••	0.7941••	0.3470••	1.0000••	0.3696••	1.0000••	0.2658••	0.8431••	0.1333••	0.500••
5	4	0.1534••	0.7849••	0.3397••	1.0000••	0.3655••	1.0000••	0.2358••	0.8602••	0.1318••	0.4731••
5	10	0.1505••	0.8446••	0.2980••	1.0000••	0.3252••	1.0000••	0.1966••	0.8932••	0.1333••	0.4368••
5	20	0.1301••	0.7684••	0.2736••	1.0000••	0.3042••	1.0000••	0.1828••	0.8315••	0.1102••	0.4105••
7	3	0.1062••	0.8541••	0.2867••	1.0000••	0.3214••	1.0000••	0.1602••	0.8541••	0.0969••	0.2916••
7	4	0.0869••	0.8117••	0.2341••	1.0000••	0.2868••	1.0000••	0.1228••	0.8352••	0.0724••	0.2705••
7	10	0.0711••	0.8068••	0.2170••	1.0000••	0.2635••	1.0000••	0.1005••	0.8295••	0.0524	0.2273••
7	20	0.0710••	0.6951••	0.2125••	1.0000••	0.2566••	1.0000••	0.1035••	0.6951••	0.0519	0.1341••
9	3	0.0719••	0.7752••	0.2315••	1.0000••	0.3224••	1.0000••	0.0873••	0.7977••	0.0407	0.1685••
9	4	0.0996••	0.6862••	0.2660••	1.0000••	0.2622••	1.0000••	0.1396••	0.7156••	0.0954••	0.1568••
9	10	0.0792••	0.6875••	0.1684••	1.0000••	0.2309••	1.0000••	0.0696••	0.6875••	0.0317	0.0729••
9	20	0.0744••	0.7741••	0.2505••	1.0000••	0.2290••	1.0000••	0.0669••	0.7956••	0.0277	0.0537

•Type I error rates below the lower limit of the 95% confidence interval (CI) (0.0375) for the empirical proportion of this rate. ••Type I error rates above the upper limit of the 95% CI (0.0625) for the empirical proportion of this rate.

Table 5. Type I error rates of multiple comparison tests according to the number of treatments (a) and the number of repetitions of treatments (r) with coefficient of variation = 10%.

Unconditional rates on a significant ANOVA F-test											
a	r	Tukey		Duncan		Fisher's LSD		SNK		Scheffé	
		α_c	α_e	α_c	α_e	α_c	α_e	α_c	α_e	α_c	α_e
3	3	0.0211•	0.0495	0.0435	0.0940••	0.0482	0.1070••	0.0291•	0.0540	0.0163•	0.0395
3	4	0.0203•	0.0475	0.0451	0.0980••	0.0505	0.1140••	0.0276•	0.0575	0.0158•	0.0375
3	10	0.0205•	0.0500	0.0416	0.0955••	0.0495	0.1190••	0.0258•	0.0570	0.0163•	0.0400
3	20	0.0178•	0.0465	0.0380	0.0865••	0.0455	0.1080••	0.0253•	0.0575	0.0126•	0.0330•
5	3	0.0075•	0.0450	0.0359•	0.1735••	0.0495	0.2460••	0.0111•	0.0515	0.0034•	0.0230•
5	4	0.0064•	0.0455	0.0365•	0.1915••	0.0506	0.2645••	0.0098•	0.0515	0.0029•	0.0235•
5	10	0.0078•	0.0560	0.0337•	0.1845••	0.0505	0.2785••	0.0107•	0.0585	0.0030•	0.0220•
5	20	0.0058•	0.0490	0.0323•	0.1970••	0.0496	0.2865••	0.0083•	0.0555	0.0022•	0.0190•
7	3	0.0050•	0.0540	0.0271•	0.2750••	0.0493	0.3830••	0.0079•	0.0565	0.0015•	0.0140•
7	4	0.0033•	0.0450	0.0280•	0.2495••	0.0485	0.3895••	0.0050•	0.0455	0.0007•	0.0105•
7	10	0.0029•	0.0475	0.0252•	0.2380••	0.0471	0.4015••	0.0043•	0.0500	0.0005•	0.0095•
7	20	0.0026•	0.0405	0.0249•	0.2535••	0.0483	0.4190••	0.0040•	0.0430	0.0003•	0.0060•
9	3	0.0022•	0.0435	0.0283•	0.3250••	0.0493	0.4900••	0.0033•	0.0435	0.0002•	0.0080•
9	4	0.0025•	0.0475	0.0264•	0.3205••	0.0495	0.5060••	0.0035•	0.0485	0.0003•	0.0080•
9	10	0.0032•	0.0520	0.0227•	0.3115••	0.0474	0.5330••	0.0038•	0.0535	0.0001•	0.0040•
9	20	0.0020•	0.0505	0.0295•	0.3230••	0.0485	0.5560••	0.0028•	0.0525	0.0000•	0.0025•
Conditional rates on a significant ANOVA F-test											
		α_1	α_3	α_1	α_3	α_1	α_3	α_1	α_3	α_1	α_3
3	3	0.0198•	0.0455	0.0271•	0.0470	0.0271•	0.0470	0.0268•	0.0470	0.0163•	0.0395
3	4	0.0186•	0.0425	0.0286•	0.0510	0.0286•	0.0510	0.0250•	0.0500	0.0158•	0.0375
3	10	0.0183•	0.0435	0.0250•	0.0480	0.0250•	0.0480	0.0230•	0.0480	0.0163•	0.0400
3	20	0.0170•	0.0440	0.0238•	0.0490	0.0238•	0.0490	0.0218•	0.0490	0.0126•	0.0330•
5	3	0.0070•	0.0400	0.0177•	0.0510	0.0191•	0.0510	0.0104•	0.0440	0.0034•	0.0230•
5	4	0.0056•	0.0370•	0.0154•	0.0475	0.0169•	0.0475	0.0086•	0.0400	0.0029•	0.0235•
5	10	0.0067•	0.0450	0.0151•	0.0510	0.0168•	0.0510	0.0092•	0.0460	0.0030•	0.0220•
5	20	0.0047•	0.0375	0.0128•	0.0485	0.0146•	0.0485	0.0069•	0.0410	0.0022•	0.0190•
7	3	0.0044•	0.0415	0.0173•	0.0515	0.0191•	0.0515	0.0066•	0.0425	0.0015•	0.0140•
7	4	0.0027•	0.0330•	0.0105•	0.0455	0.0130•	0.0455	0.0041•	0.0330•	0.0007•	0.0105•
7	10	0.0023•	0.0350•	0.0097•	0.0430	0.0118•	0.0430	0.0034•	0.0365•	0.0005•	0.0095•
7	20	0.0021•	0.0300•	0.0086•	0.0425	0.0111•	0.0425	0.0032•	0.0310•	0.0003•	0.0060•
9	3	0.0020•	0.0360•	0.0103•	0.0460	0.0131•	0.0460	0.0030•	0.0360•	0.0002•	0.0080•
9	4	0.0022•	0.0380	0.0106•	0.0530	0.0136•	0.0530	0.0031•	0.0390	0.0003•	0.0080•
9	10	0.0023•	0.0350•	0.0079•	0.0470	0.0109•	0.0470	0.0027•	0.0355•	0.0001•	0.0040•
9	20	0.0016•	0.0370•	0.0075•	0.0490	0.0106•	0.0490	0.0023•	0.0385	0.0000•	0.0025•
		α_2	α_4	α_2	α_4	α_2	α_4	α_2	α_4	α_2	α_4
3	3	0.4358••	0.9680••	0.5780••	1.0000••	0.5780••	1.0000••	0.5770••	1.0000••	0.4135••	0.8404••
3	4	0.4392••	0.8333••	0.5620••	1.0000••	0.5620••	1.0000••	0.5555••	0.9803••	0.4222••	0.7352••
3	10	0.4214••	0.9062••	0.5208••	1.0000••	0.5208••	1.0000••	0.5111••	1.0000••	0.4083••	0.8333••
3	20	0.3863••	0.8979••	0.4863••	1.0000••	0.4863••	1.0000••	0.4746••	1.0000••	0.3838••	0.6734••
5	3	0.1762••	0.7843••	0.3470••	1.0000••	0.3745••	1.0000••	0.2476••	0.8627••	0.1500••	0.4509••
5	4	0.1513••	0.7789••	0.3242••	1.0000••	0.3557••	1.0000••	0.2205••	0.8421••	0.1255••	0.4947••
5	10	0.1500••	0.8823••	0.2970••	1.0000••	0.3294••	1.0000••	0.2044••	0.9019••	0.1363••	0.4313••
5	20	0.1253••	0.7731••	0.2639••	1.0000••	0.3010••	1.0000••	0.1792••	0.8453••	0.1157••	0.3917••
7	3	0.1061••	0.8058••	0.3365••	1.0000••	0.3726••	1.0000••	0.1592••	0.8252••	0.1071••	0.2718••
7	4	0.0836••	0.7252••	0.2328••	1.0000••	0.2878••	1.0000••	0.1248••	0.7252••	0.0725••	0.2307••
7	10	0.0673••	0.8139••	0.2275••	1.0000••	0.2763••	1.0000••	0.0979••	0.8488••	0.0526	0.2209••
7	20	0.0706••	0.7058••	0.2044••	1.0000••	0.2621••	1.0000••	0.1059••	0.7294••	0.0515	0.1411••
9	3	0.0555	0.7826••	0.2255••	1.0000••	0.2865••	1.0000••	0.0852••	0.7826••	0.0364	0.1739••
9	4	0.0592	0.7169••	0.2017••	1.0000••	0.2570••	1.0000••	0.0826••	0.7358••	0.0399	0.1509••
9	10	0.0662••	0.7446••	0.1681••	1.0000••	0.2319••	1.0000••	0.0777••	0.7553••	0.0312	0.0851••
9	20	0.0446	0.7551••	0.1544••	1.0000••	0.2168••	1.0000••	0.0629••	0.7857••	0.0277	0.0510

•Type I error rates below the lower limit of the 95% confidence interval (CI) (0.0375) for the empirical proportion of this rate. ••Type I error rates above the upper limit of the 95% CI (0.0625) for the empirical proportion of this rate.

Table 6. Type I error rates of multiple comparison tests according to the number of treatments (a) and the number of repetitions of treatments (r) with coefficient of variation = 20%.

Unconditional rates on a significant ANOVA F-test											
a	r	Tukey		Duncan		Fisher's LSD		SNK		Scheffé	
		α_c	α_e	α_c	α_e	α_c	α_e	α_c	α_e	α_c	α_e
3	3	0.0205•	0.0485	0.0448	0.0950••	0.0490	0,1075••	0.0296•	0.0545	0.0161•	0.0390
3	4	0.0213•	0.0490	0.0456	0.1000••	0.0506	0.1140••	0.0278•	0.0575	0.0168•	0.0395
3	10	0.0203•	0.0490	0.0406	0.0945••	0.0480	0.1155••	0.0255•	0.0550	0.0163•	0.0400
3	20	0.0166•	0.0440	0.0380	0.0870••	0.0448	0.1065••	0.0245•	0.0555	0.0130•	0.0345•
5	3	0.0075•	0.0455	0.0353•	0.1675••	0.0493	0.2455••	0.0113•	0.0520	0.0034•	0.0235•
5	4	0.0064•	0.0450	0.0366•	0.1945••	0.0501	0.2660••	0.0097•	0.0505	0.0028•	0.0225•
5	10	0.0077•	0.0560	0.0337•	0.1810••	0.0503	0.2765••	0.0110•	0.0590	0.0031•	0.0235•
5	20	0.0058•	0.0480	0.0325•	0.1980••	0.0501	0.2855••	0.0081•	0.0525	0.0020•	0.0185•
7	3	0.0050•	0.0530	0.0352•	0.2750••	0.0522	0.3855••	0.0077•	0.0560	0.0015•	0.0155•
7	4	0.0036•	0.0455	0.0287•	0.2575••	0.0486	0.3915••	0.0049•	0.0465	0.0008•	0.0115•
7	10	0.0029•	0.0470	0.0248•	0.2400••	0.0452	0.4035••	0.0042•	0.0485	0.0005•	0.0095•
7	20	0.0025•	0.0395	0.0250•	0.2515••	0.0483	0.4185••	0.0037•	0.0415	0.0003•	0.0060•
9	3	0.0025•	0.0425	0.0280•	0.3240••	0.0491	0.4890••	0.0042•	0.0435	0.0002•	0.0080•
9	4	0.0025•	0.0485	0.0264•	0.3205••	0.0492	0.5130••	0.0035•	0.0510	0.0003•	0.0080•
9	10	0.0020•	0.0495	0.0225•	0.3110••	0.0474	0.5320••	0.0026•	0.0495	0.0001•	0.0040•
9	20	0.0021•	0.0535	0.0229•	0.3215••	0.0485	0.5575••	0.0029•	0.0550	0.0000•	0.0025•
Conditional rates on a significant ANOVA F-test											
		α_1	α_3	α_1	α_3	α_1	α_3	α_1	α_3	α_1	α_3
3	3	0.0193•	0.0450	0.0278•	0.0475	0.0278•	0.0475	0.0275•	0.0475	0.0161•	0.0390
3	4	0.0201•	0.0455	0.0283•	0.0505	0.0283•	0.0505	0.0256•	0.0505	0.0168•	0.0395
3	10	0.0183•	0.0430	0.0240•	0.0465	0.0240•	0.0465	0.0228•	0.0460	0.0163•	0.0400
3	20	0.0160•	0.0420	0.0235•	0.0480	0.0235•	0.0480	0.0211•	0.0480	0.0130•	0.0345•
5	3	0.0069•	0.0395	0.0177•	0.0510	0.0189•	0.0510	0.0105•	0.0440	0.0034•	0.0235•
5	4	0.0056•	0.0370•	0.0155•	0.0480	0.0170•	0.0480	0.0087•	0.0400	0.0028•	0.0225•
5	10	0.0066•	0.0450	0.0154•	0.0515	0.0167•	0.0515	0.0094•	0.0465	0.0031•	0.0235•
5	20	0.0048•	0.0380	0.0126•	0.0490	0.0149•	0.0490	0.0068•	0.0400	0.0020•	0.0185•
7	3	0.0043•	0.0400	0.0139•	0.0485	0.0154•	0.0485	0.0066•	0.0415	0.0015•	0.0155•
7	4	0.0030•	0.0340•	0.0106•	0.0445	0.0128•	0.0445	0.0041•	0.0345•	0.0008•	0.0115•
7	10	0.0024•	0.0360•	0.0097•	0.0450	0.0118•	0.0450	0.0035•	0.0370•	0.0005•	0.0095•
7	20	0.0020•	0.0285•	0.0084•	0.0410	0.0108•	0.0410	0.0029•	0.0290•	0.0003•	0.0060•
9	3	0.0022•	0.0340•	0.0098•	0.0445	0.0125•	0.0445	0.0177•	0.0345•	0.0002•	0.0080•
9	4	0.0022•	0.0380	0.0104•	0.0520	0.0133•	0.0520	0.0031•	0.0405	0.0003•	0.0080•
9	10	0.0015•	0.0335•	0.0077•	0.0470	0.0111•	0.0470	0.0019•	0.0335•	0.0001•	0.0040•
9	20	0.0016•	0.0365•	0.0077•	0.0490	0.0105•	0.0490	0.0022•	0.0375	0.0000•	0.0025•
		α_2	α_4	α_2	α_4	α_2	α_4	α_2	α_4	α_2	α_4
3	3	0.4296••	0.9473••	0.5859••	1.0000••	0.5859••	1.0000••	0.5851••	1.0000••	0.4145••	0.8210••
3	4	0.4432••	0.9009••	0.5610••	1.0000••	0.5610••	1.0000••	0.5519••	1.0000••	0.4261••	0.7821••
3	10	0.4263••	0.9247••	0.5161••	1.0000••	0.5161••	1.0000••	0.5131••	0.9892••	0.4083••	0.8602••
3	20	0.3809••	0.8750••	0.4895••	1.0000••	0.4895••	1.0000••	0.4756••	1.0000••	0.3768••	0.7187••
5	3	0.1759••	0.7745••	0.3470••	1.0000••	0.3705••	1.0000*•	0.2542••	0.8627••	0.1468••	0.4607••
5	4	0.1513••	0.7708••	0.3239••	1.0000••	0.3552••	1.0000••	0.2272••	0.8333••	0.1244••	0.4687••
5	10	0.1477••	0.8737••	0.3000••	1.0000••	0.3252••	1.0000••	0.2065••	0.9029••	0.1319••	0.4563••
5	20	0.1263••	0.7755••	0.2581••	1.0000••	0.3051••	1.0000••	0.1779••	0.8163••	0.1108••	0.3775••
7	3	0.1095••	0.8247••	0.2866••	1.0000••	0.3186••	1.0000••	0.1634••	0.8556••	0.0998••	0.3195••
7	4	0.0903••	0.7640••	0.2402••	1.0000••	0.2894••	1.0000••	0.1218••	0.7752••	0.0724••	0.2584••
7	10	0.0681••	0.8000••	0.2169••	1.0000••	0.2634••	1.0000••	0.0971••	0.8222••	0.0526	0.2111••
7	20	0.0718••	0.6951••	0.2049••	1.0000••	0.2642••	1.0000••	0.1035••	0.7073••	0.0515	0.1463••
9	3	0.0665••	0.7640••	0.2222••	1.0000••	0.2815••	1.0000••	0.3982••	0.7752••	0.0364	0.1797••
9	4	0.0584	0.7307••	0.2003••	1.0000••	0.2572••	1.0000••	0.0836••	0.7788••	0.0399	0.1538••
9	10	0.0472	0.7127••	0.1643••	1.0000••	0.2367••	1.0000••	0.0588	0.7127••	0.0312	0.0851••
9	20	0.0449	0.7448••	0.1584••	1.0000••	0.2157••	1.0000••	0.0624	0.7653••	0.0277	0.0510

•Type I error rates below the lower limit of the 95% confidence interval (CI) (0.0375) for the empirical proportion of this rate. ••Type I error rates above the upper limit of the 95% CI (0.0625) for the empirical proportion of this rate.

Further, Duncan and Fisher's LSD tests show that the values for α_e are above the upper limit of the 95% CI (0.0625), and the increase in the number of treatments of the scenarios cause a significant increase in this error rate (Tables 3 to 6).

The Scheffé test shows that the estimates for α_e mostly lie below the lower limit of 0.0375, except for two scenarios where a = 3 when CV = 1% and CV = 5% (Tables 3 and 4), and except for three scenarios with a = 3 when CV = 10% and CV = 20% (Tables 5 and 6), where the test controlled this rate. An increase in the number of treatments for the test causes a significant decrease in the error rate.

In general, in the context of α_e, Tukey and SNK are the only tests that control for this rate and are therefore considered to be robust in all experimental conditions, regardless of the number of treatments, the number of repetitions, and CV. The Scheffé test, in turn, is in an intermediate situation, since, to some extent, it controls the experiment-wise error rate only in the scenarios with three treatments, while it is conservative in other cases. The Duncan and Fisher's LSD tests depict the worst performance, showing that they are liberal in all simulated scenarios, with Fisher's LSD test being the most liberal among them.

According to Girardi et al. (2009), the equality of the comparison-wise error rate and experiment-wise error rate would be ideal for a multiple comparison test according to the level of significance established. However, according to Perecin and Barbosa (1988), a test that controls for the comparison-wise error rate can become very liberal when applied to the entire experiment, while a test that controls for the experiment-wise error rate can become conservative in a comparison.

Indeed, the Tukey and SNK tests prove to be conservative in terms of controlling for the comparison-wise error rate, while controlling for the experiment-wise error rate equal to the nominal significance level. A similar behaviour exhibited by the tests was observed by Boardman and Moffitt (1971), Bernhardson (1975), and Girardi et al. (2009). Fisher's LSD test, which controls for the comparison-wise error rate, proves to be liberal in terms of controlling for the experiment-wise error rate, as observed by Boardman and Moffitt (1971), Bernhardson (1975), Perecin and Barbosa (1988), and Girardi et al. (2009).

Regarding conditional error rates, the comparison-wise rates α_c are always equal to the conditional rates α_1 for the Scheffé test, while there is a tendency for the rates α_c to be slightly higher than α_1 for each of the other tests considered (Tables 3 to 6). Meanwhile, the conditional rates α_2 are always higher than the rates α_c for the five tests. The differences between these last two error rates are large when the number of treatments is small and decrease as the number of treatments increase.

Therefore, none of the means tests controls for the conditional comparison-wise error rates α_1 or α_2. For α_1, all the tests present estimates below the lower limit of the 95% CI calculated, being conservative regarding the control of this rate. For α_2, all the tests, in general, show a liberal behaviour with the estimates lying above the upper limit of the CI, except in some of the scenarios with a = 9 for the Tukey and SNK tests, and some of scenarios with a = 7 and all the scenarios with a = 9 for the Scheffé test, in which the tests control this error rate.

The unconditional experiment-wise error rates α_e are always equal to the conditional rates α_3 for the Scheffé test, whereas for each of the other tests, the rates α_e are slightly higher than the rates α_3 (Tables 3 to 6). Further, the rates α_e are always lower than the conditional rates α_4, and although the differences between them are considerable, they decrease as the number of treatments increase. For the Scheffé test, however, the differences between these rates for scenarios with a high number of treatments are not large.

Regarding the control of the conditional experiment-wise error rates, we observed that the Duncan and Fisher's LSD tests control for the error rate α_3, regardless of the number of treatments. The Tukey and SNK tests, in turn, control for this rate in all the scenarios with a = 3 and a = 3; a = 5, respectively. However, for the other scenarios, when they do not control this rate, both tests are conservative. The Scheffé test, however, proves to be conservative in most scenarios, controlling for this rate only in certain scenarios with a = 3. For α_4, all the tests show a liberal behaviour, except for the Scheffé test in certain scenarios with a = 9, where this rate is controlled.

Figures 1 and 2 summarise the tendency of the means tests, explaining every rate for the observed tests as the number of treatments increase. Only variation in the number of treatments is considered here, as this is the only factor that led to significant alteration of the results. Therefore, we consider r = 10, CV = 10%, and α = 5%. The lines between the dots represent what occurred to the rates with an increase in the number of treatments. For simplicity, only the case with r =10 and CV = 10% is reported; the results for the other simulated scenarios are similar.

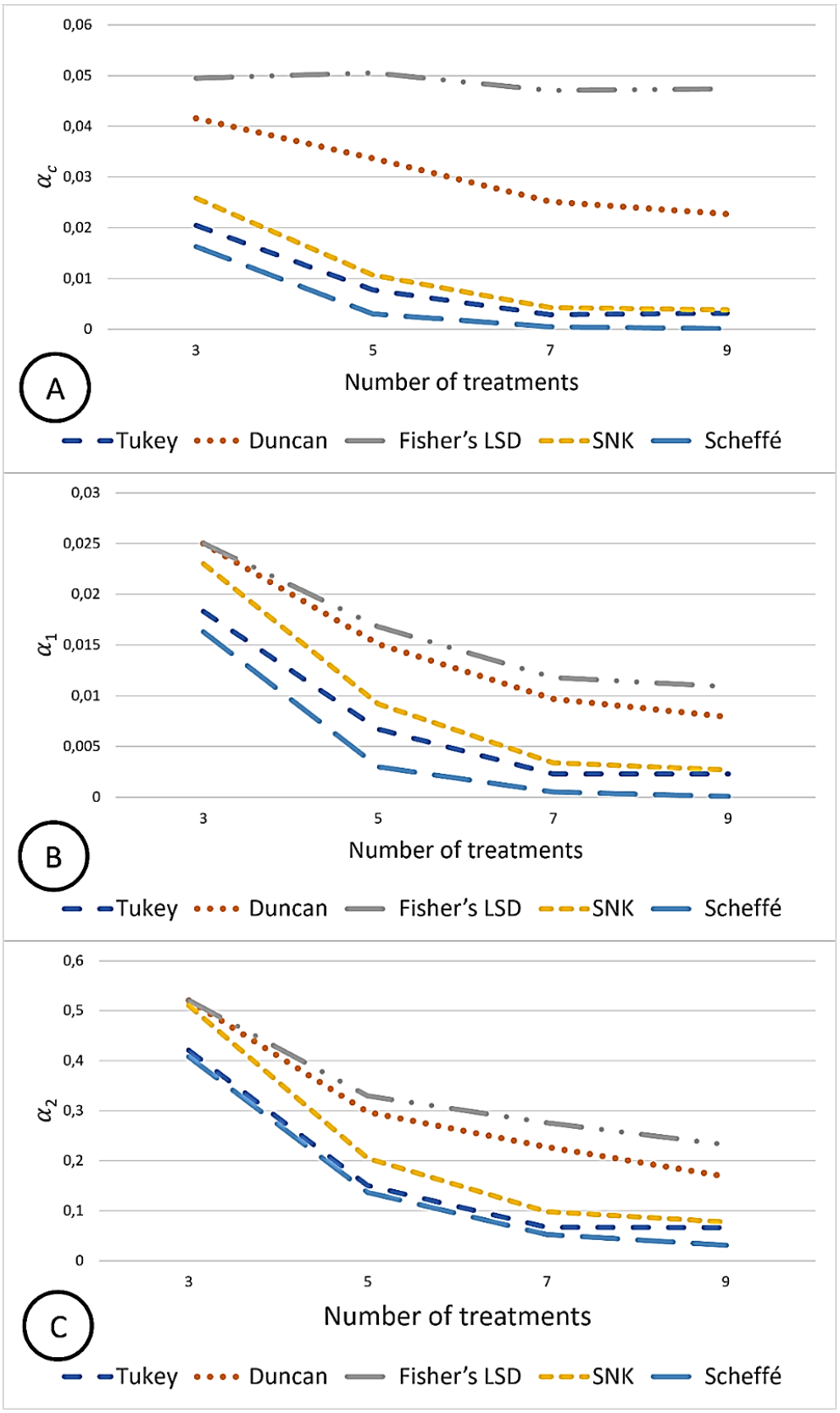

Figure 1. (A) Unconditional comparison-wise error rates α_c; (B) conditional comparison-wise error rates α_1; and (C) conditional comparison-wise error rates α_2 for the various multiple comparison tests with 10 replications, a nominal significance level of 5%, and coefficient of variation = 10%, according to the variation in the number of treatments.

Based on the above results, it can be inferred that the combined use of the overall ANOVA F-test and the multiple comparison tests can change the type I error rates of the means tests. Duncan and Fisher's LSD tests do not control for the experiment-wise error rate α_e, while an increase in the number of treatments leads to

an increase in this rate. However, considering the conditional experiment-wise error rate α_3, we find that the tests control for this rate, regardless of the number of treatments because the nominal significance level for the ANOVA F-test determines the upper limit for the α_3 error rates (Bernhardson, 1975).

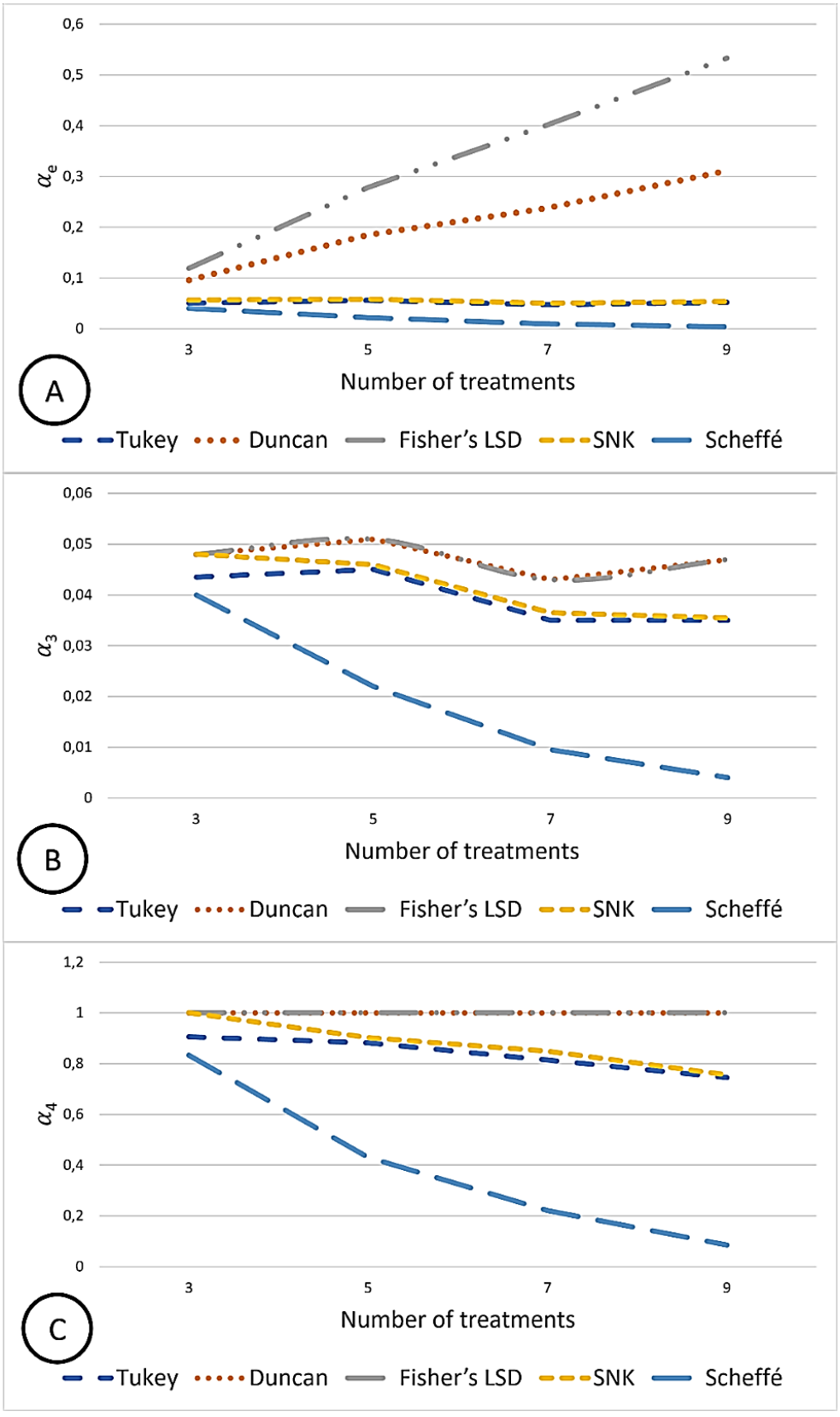

Figure 2. (A) Unconditional experiment-wise error rates α_e; (B) conditional experiment-wise error rates α_3; (C) conditional experiment-wise error rates α_4 for the various multiple comparison tests with 10 replications, nominal significance level of 5%, and coefficient of variation = 10%, according to the variation in the number of treatments.

In general, Fisher's LSD test controls for the comparison-wise error rate α_c, whereas the Tukey and SNK tests control for the experiment-wise error rate α_e. The Duncan and Fisher's LSD tests control for the conditional experiment-wise error rate α_3. The Scheffé test does not control for any of the error rates considered, possibly because the test can be used for all possible contrasts and not only for the pair-wise contrasts of means (Boardman & Moffitt, 1971).

Conclusion

Since each multiple comparison test controls for a different error rate, the choice of the test must depend on what error rates are intended to be controlled. If the decision is to control for the comparison-wise error rate α_c, then Fisher's LSD test is the most suitable. If the decision is to control for the experiment-wise error rate α_e, then the Tukey and SNK tests are recommended. If the intention is to control for the conditional experiment-wise error rate α_3, then the Duncan and Fisher's LSD tests can be used. Type I error rates, in general, did not show significant changes with a variation in the number of repetitions or the CV, but exhibited changes with a variation in the number of treatments of the trials. When choosing the most applicable test, in addition to the type I error rate, the power function should be considered. It is always desirable for tests with good performance to maintain the coverage of the type I error rate, which should simultaneously have great power. Research on the power function is extremely exhaustive and therefore should be developed by future studies.

Acknowledgements

We would like to thank *Coordenação de Aperfeiçoamento de Pessoal de Nível Superio*r (CAPES) for their financial support.

References

Bernhardson, C. S. (1975). Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, *31*(1), 229-232. DOI: https://doi.org/10.2307/2529724

Biase, N. G., & Ferreira, D. F. (2011). Testes de igualdade e de comparações múltiplas para várias proporções binomiais independentes. *Revista Brasileira de Biometria*, *29*(4), 549-570.

Boardman, T. J., & Moffitt, D. R. (1971). Graphical Monte Carlo Type I error rates for multiple comparison procedures. *Biometrics*, *27*(3), 738-744. DOI: https://doi.org/10.2307/2528613

Cardellino, R. A., & Siewerdt, F. (1992). Use and misuse of statistical tests for comparison of means. *Revista da Sociedade Brasileira de Zootecnia*, *21*(6), 985-995.

Girardi, L. H., Cargnelutti Filho, A., & Storck, L. (2009). Erro tipo I e poder de cinco testes de comparação múltipla de médias. *Revista Brasileira de Biometria*, *27*(1), 23-36.

Gonçalves, B. O., Ramos, P. S., & Avelar, F. G. (2015). Test Student-Newman-Keuls bootstrap: proposal, evaluation, and application productivity data of soursop. *Revista Brasileira de Biometria*, *33*(4), 445-470.

Henrique, F. H., & Laca-Buendía, J. P. (2010). Comportamento morfológico e agronômico de genótipos de algodoeiro no município de Uberaba - MG. *FAZU em Revista*, *7*, 32-36.

Perecin, D., & Barbosa, J. C. (1988). Uma avaliação de seis procedimentos para comparações múltiplas. *Revista de Matemática e Estatística*, *6*, 95-103.

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, AT: R Foundation for Statistical Computing. Retrieved on Jan. 9, 2021 from URL https://www.R-project.org/

Ramos, P. S., & Vieira, M. T. (2014). Bootstrap multiple comparison procedure based on the F distribution. *Revista Brasileira de Biometria*, *31*(4), 529-546.

Rodrigues, J., Piedade, S. M. S., & Lara, I. A. R. (2016). Aplicação condicional de testes de comparação de médias a um resultado significativo do teste F global na análise de variância. *Revista Brasileira de Biometria*, *34*(1), 1-22.

Saville, D. J. (2014). Multiple comparison procedures - Cutting the Gordian knot. *Agronomy Journal*, *107*(2), 730-735. DOI: https://doi.org/10.2134/agronj2012.0394

Souza, C. A., Lira Junior, M. A., & Ferreira, R. L. C. (2012). Avaliação de testes estatísticos de comparações múltiplas de médias. *Revista Ceres*, *59*(3), 350-354. DOI: https://doi.org/10.1590/S0034-737X2012000300008