Monte Carlo evaluation of the ANOVA's F and Kruskal-Wallis tests under binomial distribution

Eric B. Ferreira[1†], *Marcela C. Rocha*[2], *Diego B. Mequelino*[3]

[1] *Adjunct Lecturer III, Exact Sciences Institute, Federal University of Alfenas, Brazil, CEP: 37130-000.*
[2] *Lecturer D1-1, Federal Institute of Education, Science and Technology of Southern Minas Gerais, Brazil.*
[3] *Mathematics student, Federal University of Alfenas, Brazil.*

Resumo: *Para verificar a igualdade de mais de dois níveis de um fator de interesse em experimentos conduzidos em delineamento inteiramente casualizado costuma-se utilizar o teste F da análise de variância, que é considerado o teste mais poderoso para tal finalidade. No entanto, a validade de seus resultados depende da verificação das seguintes pressuposições: aditividade dos efeitos admitidos no modelo, independência, homocedasticidade e normalidade dos erros. O teste não-paramétrico de Kruskal-Wallis possui pressuposições mais moderadas e, portanto, é uma alternativa quando as pressuposições exigidas pelo teste F não se verificam. Entretanto, quanto mais fortes as pressuposições de um teste, melhor será seu desempenho e, se forem satisfeitas as hipóteses fundamentais da análise de variância, o teste F será a melhor opção. Neste trabalho; violou-se normalidade dos erros, ao simular variáveis-resposta binomiais, objetivando-se comparar os desempenhos dos testes F e Kruskal-Wallis quando uma das pressuposições da análise de variância não é satisfeita. Por meio da simulação Monte Carlo, foram simulados* 3.150.000 *experimentos para avaliar a taxa de erro tipo I e poder dos testes. Na maioria das situações, o poder do teste F foi superior ao do teste de Kruskal-Wallis e, ainda assim, o teste F controlou a taxas de erro tipo I.*

Palavras-chave: Poder; taxa de erro tipo I; simulação Monte Carlo; DIC.

Abstract: *To verify the equality of more than two levels of a factor under interest in experiments conducted under a completely randomized design (CRD) it is common to use the F ANOVA test, which is considered the most powerful test for this purpose. However, the reliability of such results depends on the following assumptions: additivity of effects, independence, homoscedasticity and normality of the errors. The nonparametric Kruskal-Wallis test requires more moderate assumptions and therefore it is an alternative when the assumptions required by the F test are not met. However, the stronger the assumptions of a test, the better its performance. When the fundamental assumptions are met the F test is the best option. In this work, the normality of the errors is violated. Binomial response variables are simulated in order to compare the performances of the F and Kruskal-Wallis tests when one of the analysis of variance assumptions is not satisfied. Through Monte Carlo simulation, were simulated* $3,150,000$ *experiments to evaluate the type I error rate and power rate of the tests. In most situations, the power of the F test was superior to the Kruskal-Wallis and yet, the F test controlled the Type I error rates.*

Keywords: Power; type I error rate; Monte Carlo simulation; Completely Randomized Design.

[†] *Corresponding author:* eric.ferreira@unifal-mg.edu.br.

Introduction

In experiment designs the comparison of means - in general - is made by comparing the effects of treatments, through multiple comparison tests. Before that, however, is usually done a test to detect the existence of differences between treatments, in which the null hypothesis of equality of means is tested against the alternative hypothesis that there is at least one average different from the others.

The analysis of variance (ANOVA) was the first method for the analysis of experimental data, developed by Ronald Fisher from the 1920s. In this method the comparison of means is performed using the F test, which is considered the most powerful parametric test for this purpose (SIEGEL; CASTELLAN, 2006).

A parametric test commonly brings strong assumptions such as those about the distribution of the data. Thus, the use of the ANOVA's F test depends on the verification of four assumptions to be valid. Such assumptions, called *fundamental assumptions of analysis of variance*, are: additive effects model, independence, homoscedasticity and normality of errors. Theoretically, if at least one of these assumptions is not met, the analysis of variance has no validity as a statistical analysis technique and becomes a simple mathematical treatment of data collected (LIMA; ABREU, 2000).

An alternative to circumvent the violation of assumptions required by the variance analysis is the use of nonparametric statistics for data analysis. The nonparametric corresponding of the ANOVA's F test, in experiments conducted in a completely randomized design (CRD), is the Kruskal-Wallis test, which is based on the observations rank and whose assumptions are: (i) independence among observations; (ii) observations from the same population within a treatment and that the treatments have roughly the same distribution and; (iii) the variables are continuous (KRUSKAL; WALLIS, 1952).

It is remarkable that the assumptions of the Kruskal-Wallis are milder than those of the F test, but it is important to note that the less extensive are the assumptions for performing a test, the more general will be its conclusions and the lower its efficiency. Thus, if the underlying assumptions of the analysis of variance are met, the F test will present better performance than the nonparametric Kruskal-Wallis (CAMPOS, 1983).

Vieira (2006) argues that non-normality of errors affects the efficiency in the estimation of treatment effects and results in loss of power and, furthermore, there is increased error in the level of significance of the test. However, that author asserts that small violations of this assumption does not affect substantially the result of analysis of variance.

This statement is reinforced by Feir and Toothaker (1974) who compared, via Monte Carlo simulation, the power and type I error rates of F and Kruskal-Wallis tests, especially in situations where the assumptions of the parametric procedure were not satisfied. For this comparison, the authors simulated various situations resulting from combinations of total sample sizes ($N = 28$ or $N = 68$), balanced and unbalanced experiments, equal and not equal variances, normal and exponential (non-normal) data. According to those authors, the Kruskal-Wallis test proved to be competitive considering the type I error rates, but the same did not happen with the power. Therefore, the authors concluded that the F test performed best in most cases, even when the normality and/or homocedascity were not met.

Reis and Ribeiro (2007) compared the performance of the F, Kruskal-Wallis and Friedman tests to data under normality or not, in experiments conducted in completely randomized designs (CRD) and a randomized block designs (RBD), respectively. For this comparison, authors simulated 1000 samples with a fixed number of treatments ($I = 5$) and 5, 10 and 25 replicates per sample, under normal, lognormal and binomial distribution, and estimated the type I error rate and power of tests. According to the authors, the F test for both the CRD and DBC presented empirical power greater than the nonparametric tests and still controlled the type I error rates in all simulated situations. Thus, the authors concluded that there is no need to replace the F test for their nonparametric competitors, even when the assumption of normality

is not satisfied.

Thus, this study aimed to compare, via Monte Carlo simulation, the performance of ANOVA's F and Kruskal-Wallis tests and recommend which one should be used when the response variable is binomial, therefore, do not satisfy the assumption of normality.

Methodology

We used a Monte Carlo simulation to assess the type I error rates and power of ANOVA's F and Kruskal-Wallis tests.

To simulate the experimental data as well as for the estimation of type I error and power, an algorithm was developed in R language (R CORE TEAM, 2012).

The experimental data were simulated from a binomial distribution with parameters n and p_i^*, where $n \in \mathbb{N}$ and $p_i^* \in [0,1]$, ie $Y_{ij} \sim Bin(n, p_i^*)$ with $i = 1, 2, \ldots, I$ and $j = 12, \ldots, J$, where I is the number of treatments, J is the number of replications and n is given by:

$$n = \frac{1}{CV^2} \tag{1}$$

where CV is the coefficient of variation.

The equation above was obtained from the coefficient of variation (CV), settling the average probability of success of the experiment in $\bar{p} = 0.5$ as follows:

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{n\bar{p}(1-\bar{p})}}{n\bar{p}} = \frac{\sqrt{n(0.5)(0.5)}}{n(0.5)} = \frac{\sqrt{n}}{n} \quad \Rightarrow \quad n = \frac{1}{CV^2}$$

Note that $n \in \mathbb{N}$ and $n > 1$ since n is the number of trials of the simulated binomial distribution. Thus, for the value of n, it was chosen the nearest natural number, where the value calculated in equation (1) were not natural numbers.

For the data simulation under complete H_0, the parameter $p_i^* = 0.5$ and was fixed and, under complete H_1, p_i^* was obtained, so that the treatments were centred at 0.5, from the equation:

$$p_i^* = \begin{cases} p_i + \left[0.5 - p_{\left(\frac{I+1}{2}\right)}\right], \text{if I is odd} \\ p_i + \left[0.5 - \dfrac{p_{\left(\frac{I}{2}\right)} + p_{\left(\frac{I}{2}+1\right)}}{2}\right], \text{if I is even.} \end{cases} \tag{2}$$

where p_i is given by:

$$p_i = \frac{i}{I + K} \tag{3}$$

where K is a penalty factor, whose function is to generate values of p_i close to each other.

Equation (3) was obtained to keep the treatment equally spaced in the range $[0, 1]$, ie the interval between each adjacent treatment was a fixed length p_i.

The factor K acts as pseudo-treatments that are added to I, thus increasing the number of treatments that should be equally spaced between 0 and 1. Thus, the real I treatments are confined to a subinterval contained in $[0, 1]$. Therefore, the higher $K \in \mathbb{N}$, the lower the subinterval, ie closer are the average of treatments.

For example, taking $K = 1$, $CV = 20\%$ and total of treatments $I = 3$, the values generated for p_i^* are: $p_1^* = 0.25$, $p_2^* = 0.5$ and $p_3^* = 0.75$. Thus, treatments T_1, T_2 and T_3 follows the binomial distributions $Bin(25, .25)$,$Bin(25, .50)$ and $Bin(25, 0.75)$, respectively, and are illustrated on Figure 1.

Keeping the same CV and number of treatments, and setting $K = 10$, the values generated for p_i^* are: $p_1^* = 0.42$, $p_2^* = 0.5$ and $p_3^* = 0.58$. Thus, treatments T_1, T_2 and T_3 follow the

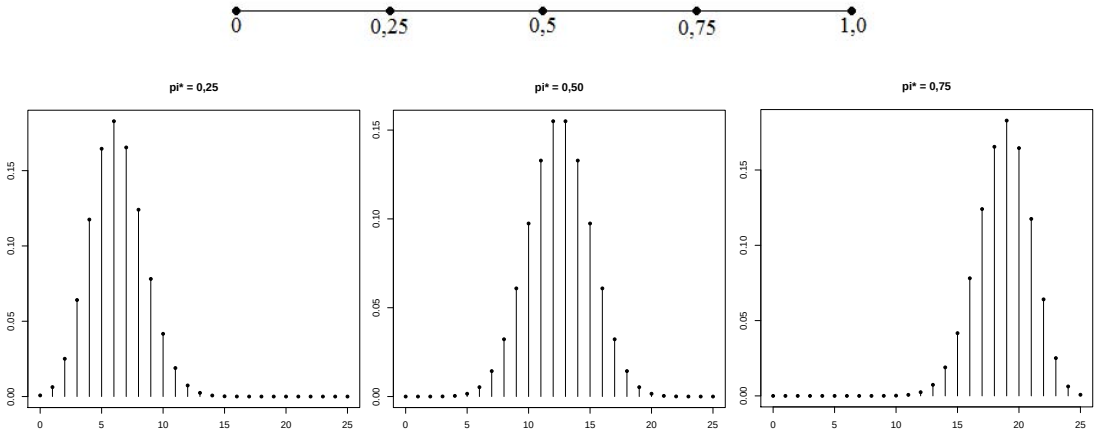


Figure 1: Values generated for p_i^* and their respective distributions, setting $CV = 20\%$ and $K = 1$.

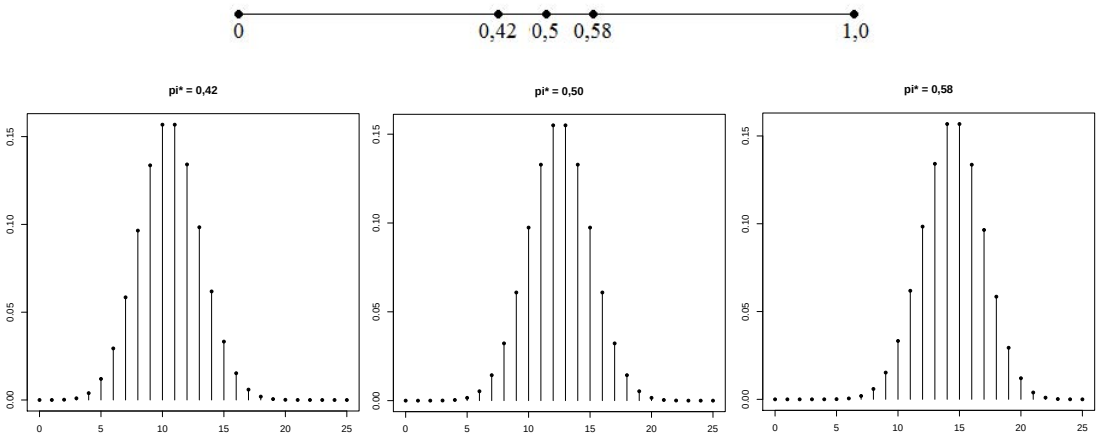


Figure 2: Values generated for p_i^* and their respective distributions, setting $CV = 20\%$ and $K = 10$.

binomial distributions $Bin(25, .42)$, $Bin(25, .50)$ and $Bin(25, 0.58)$, respectively, illustrated on Figure 2.

It is noteworthy that, for the simulation under complete H_0, the value of p_i^* was fixed and thus the penalty factor was only used in the simulation under H_1. The values adopted for the penalty factor were $K = 1, 10, 50, 100$.

We considered groups of experiments (I and II) and, for each scenario, we simulated 3000 experiments. The nominal level of significance was set to 5%.

In experiment I, we evaluated the experimentwise type I error rates of the F and Kruskal-Wallis tests and, for this, were simulated 630,000 experiments (210 sets $\times$ 3,000 experiments per scenario) without treatment effect (complete H_0). The 210 scenarios were the result of combinations between the number of treatments ($I = 3, 5, 10, 15, 20, 25, 30$), the number of replications ($J = 3, 4, 5, 10, 15, 20$) and coefficients of variation ($CV = 1\%, 5\%, 10\%, 15\%, 20\%$). The empirical type I error rate was computed by the ratio of the total number of wrong inferences (under H_0), and the total number of experiments (3000).

To verify the presence of differences between the type I error rate and the nominal level of significance (5%) we used the 99% exact confidence interval for proportions, given by:

$$IC_{1-\alpha}: \left[P_I = \frac{1}{1 + \frac{(n-y+1)F_{\alpha/2;\nu_1=2(n-y+1),\nu_2=2y}}{y}}; P_S = \frac{1}{1 + \frac{(n-y)}{(y+1)F_{\alpha/2;\nu_1=2(y+1),\nu_2=2(n-y)}}} \right] \quad (4)$$

where $F_{\alpha/2}$ is the superior quantile of F distribution with ν_1 and ν_2 degrees of freedom. If $y = 0$ then $P_I = 0$ and P_S is obtained in (4), or if $y = n$ then $P_S = 1$ and P_I is obtained in (4) (FERREIRA, 2005).

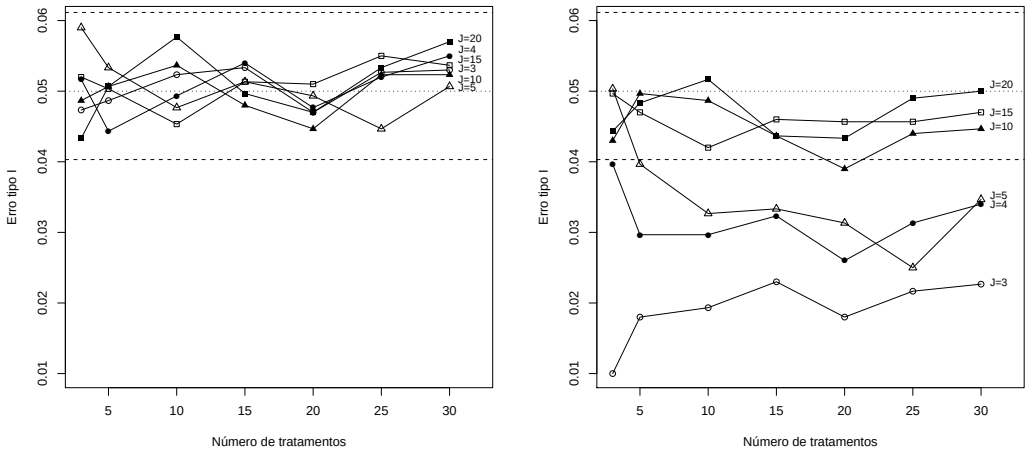
In group II, we computed the power of the tests. Overall, $2,520,000$ experiments were simulated (840 scenarios $\times$ 3000 experiments per scenario) with different treatment effects, so that the means of the treatments were equally spaced and $\tau_1 < \tau_2 < \ldots < \tau_I$. The 840 scenarios were obtained from the combinations of number of treatments ($I = 3, 5, 10, 15, 20, 25, 30$), the number of replications ($J = 3, 4, 5, 10, 15, 20$), the coefficients of variation ($CV = 1\%, 5\%, 10\%, 15\%, 20\%$) and the values of the penalty factor ($K = 1, 10, 50, 100$).

The power of the tests was estimated by $1 - \hat{\beta}$, where $\hat{\beta}$ is given by the ratio between the total number of wrong inferences (under H_1) and the total number of experiments (3000)[1].

Results and discussion

Type I error rate

On Figure 3 and Figure 4 we show the type I error rates of the F and Kruskal-Wallis tests, under the null, considering the nominal level of significance set at 5%.



(a) (b)

Figure 3: Type I error rate based on the number of treatments, number of replications (J) and $CV = 1\%$, for F (a) and Kruskal-Wallis (b) tests.

To check for differences between the nominal level of significance ($\alpha = 5\%$) and type I error rate, 99% exact confidence intervals for proportions were calculated. The confidence interval is

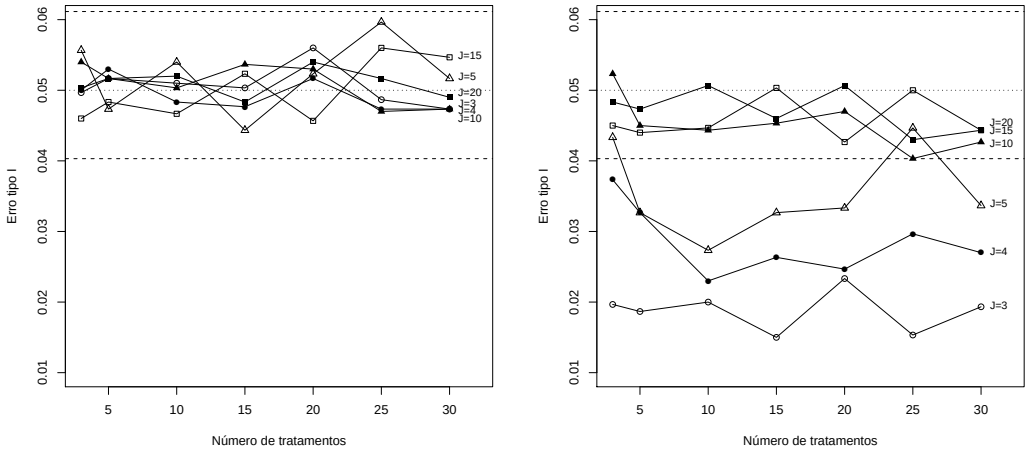
[1]Note that, although the simulation has been made under complete H_1, for computing the power, decisions were found to be correct when the test has been detected at least one pair of different means. Thus, power was calculated by an empirical number of times the test correctly rejected H_0 divided by the total number of experiments (3000).

$[0.04031, 0.06115]$. Thus, all figures show three lines: a dotted, referring to the significance level, and two dashed, referring to the extremes of the confidence interval.

As in Figures 3 and 4, in all cases studied, both tests have control of the type I error rate. However, in most cases where the number of replicates was less than or equal to 5, the type I error rates encountered for Kruskal-Wallis test were below the nominal level of significance, ie, the Kruskal-Wallis test was more conservative.

It was observed that for both tests, the number of treatments did not affect this rate as well as the coefficient of variation. About the coefficient of variation, similar result was obtained by Reis and Ribeiro (2007), who simulated binomial data with a fixed number of treatments ($I = 5$) and reported no changes in the type I error rates when increased the coefficient of variation.

Considering ($CV = 5\%, 10\%, 15\%$), the results were quite similar to ($CV = 1\%$) shown above and ($CV = 20\%$) shown in the figure below, therefore these results were not presented here.



(a) (b)

Figure 4: Type I error rate based on the number of treatments, number of replications (J) and $CV = 20\%$, for F (a) and Kruskal-Wallis (b) tests.

Evaluating the type I error rate according to the number of replications, it was found that increasing the number of repetitions caused an increase of the error rate of the Kruskal-Wallis test for all the coefficients of variation, leading it to stay closest to the nominal level of significance ($\alpha = 5\%$). The number of replicates had no effect on type I error rates of the ANOVA's F test. This result is corroborated by Reis and Ribeiro (2007), who had type I error rates closer to the nominal level of significance when the number of simulated repetitions was high.

Power rates

In this section we show the percentages of correct decisions for the ANOVA's F test and Kruskal-Wallis test, depending on the number of treatments, for the nominal significance value $\alpha = 5\%$, under complete H_1 and considering the various penalty factors (K) and coefficients of variation (CV).

The power of ANOVA's F test was less than the power of the nonparametric test in all situations, even with the breakdown of the assumption of normality of errors. Feir and Toothaker (1974), which simulated data from an exponential distribution, and Reis and Ribeiro (2007) that

simulated data lognormal and binomial distribution, obtained similar results when simulating non-normal data.

Evaluating the power in function of the CV, it was found that its increase cause a decrease in power for both F Kruskal-Wallis tests. A similar result was obtained by Reis and Ribeiro (2007), who stated that this decrease in power due to the fact that the increase in CV causes a departure from the normal distribution by reducing the sample size (n) of the binomial distribution.

Also, it was possible to observe that in all scenes, the empirical power increased with the number of treatments and repetitions. Reis and Ribeiro (2007) also observed the growth of power when increased the number of replications.

Below are the detailed results for each penalty factor, as well as graphical representations of some of these results.

Penalty factor $K = 1$

Both tests had a maximum power when they were considered the coefficients of variation 1%, 5% and 10%, despite the number of treatments and replications.

Considering the other coefficients of variation, the power of the F test was found to be equal to or higher than the nonparametric test because the F test maintained maximum power in most situations, except in the case where $CV = 20\%$, $I = 3$ and $J = 3$ (whose power was 99.63%), while Kruskal-Wallis showed larger falls, considering the same number of treatments and replicates (98.76% and 92.40% for $CV = 15\%$ and $CV = 20\%$, respectively).

Penalty factor $K = 10$

Both the ANOVA's F test and Kruskal-Wallis test reached maximum power in most situations where $CV = 1\%$ and $CV = 5\%$. However the parametric test showed maximum power in all situations, while its nonparametric competitor presented a fall when $CV = 5\%$, $I = 3$ and $J = 3$, obtaining power of 96%.

Considering $CV = 10\%$, 3 treatments and 3 replicates, the power of F and Kruskal-Wallis tests were (72.63%) and (42.23%), respectively. Taking into account the same number of treatments, both tests had a power greater than 80% for 4 or more repetitions. Both tests have reached the maximum power, regardless of the number of repetitions, when the number of treatment was less than 10.

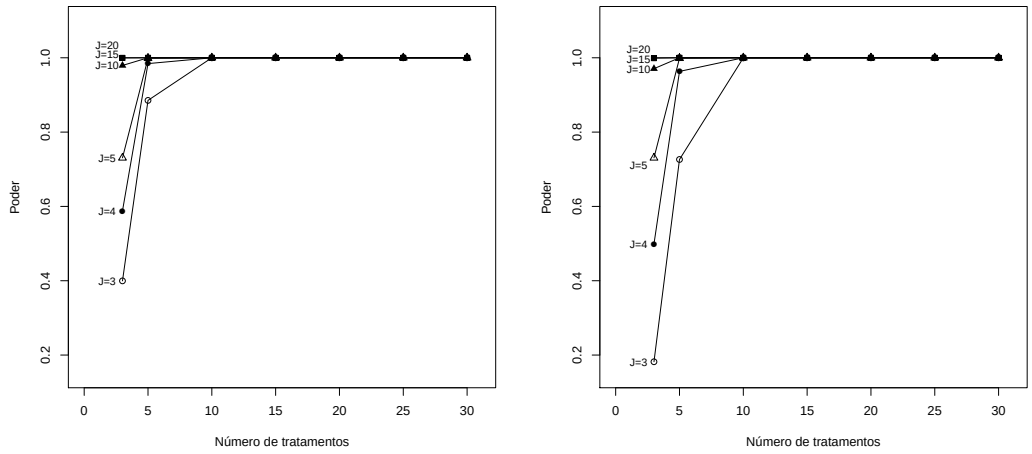
When considering $CV = 15\%$, shown on Figure 5, one can observe that, for the minimum number of treatments ($I = 3$), the power of the F test ranged from 39.96% to 100%, while the Kruskal-Wallis's power ranged between 18.20% and 100%. Note also that the empirical power curves of the F test are closer together than the Kruskal-Wallis test. Both tests showed maximum power when the number of treatments are equal to or greater than 10.

It is shown on Figure 6 the power of F and Kruskal-Wallis tests, where $CV = 20\%$. The growth rate of the power curves of the F test are closer together than the Kruskal-Wallis test. Taking into account three treatments and three replicates, F test achieved power between 24.96% and 99.20% while the Kruskal-Wallis ranged over 10.5% to 98.9%. Both tests had a maximum power when the number of treatments was less than 15, although they come close to that power, when considered 10 treatments (99.90% and 99.40%, respectively).

Penalty factor $K = 50$

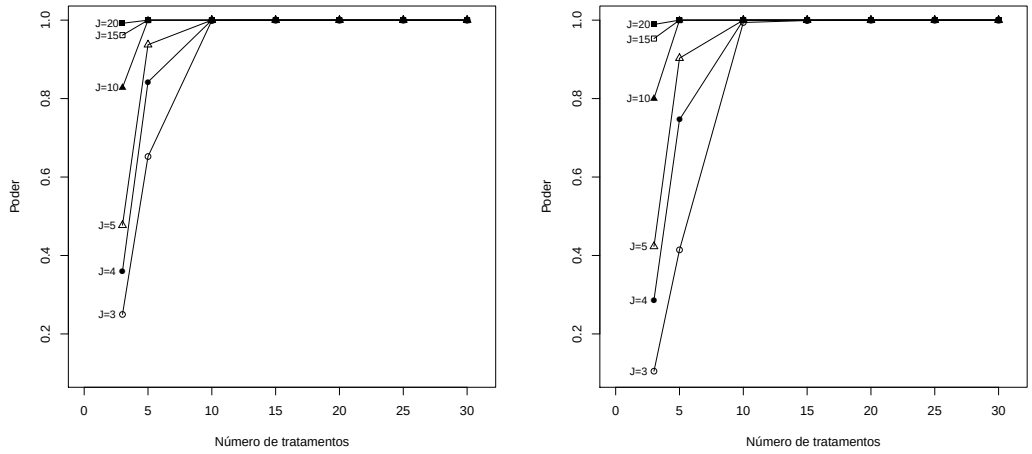
Regarding $CV = 1\%$, F test showed maximum power no matter how many treatments and replications, while the Kruskal-Wallis test showed a drop in power for the minimum number of treatments and replicates (98.80%).

On Figure 7 are shown the power of F and Kruskal Wallis tests in situations where $CV = 5\%$. Note that the empirical power curves of the F test are closer together than the Kruskal-Wallis test. Considering the minimum number of treatments ($I = 3$), the power of F test



(a) (b)

Figure 5: Power rates based on the number of treatments and replications, considering $K = 10$ and $CV = 15\%$, for F (a) and Kruskal-Wallis (b) tests.

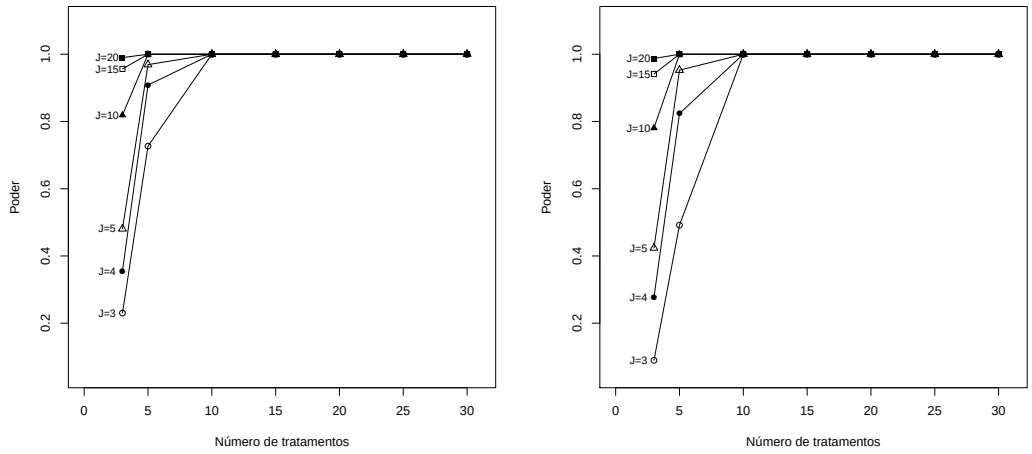


(a) (b)

Figure 6: Power rates based on the number of treatments and replications, considering $K = 10$ and $CV = 20\%$, for F (a) and Kruskal-Wallis (b) tests.

ranged from 23.03% and 99% and, in the same situation, the power of the Kruskal-Wallis ranged between 8.97% and 98.73%. Besides, both tests reached the maximum power when the number of treatment was less than 10.

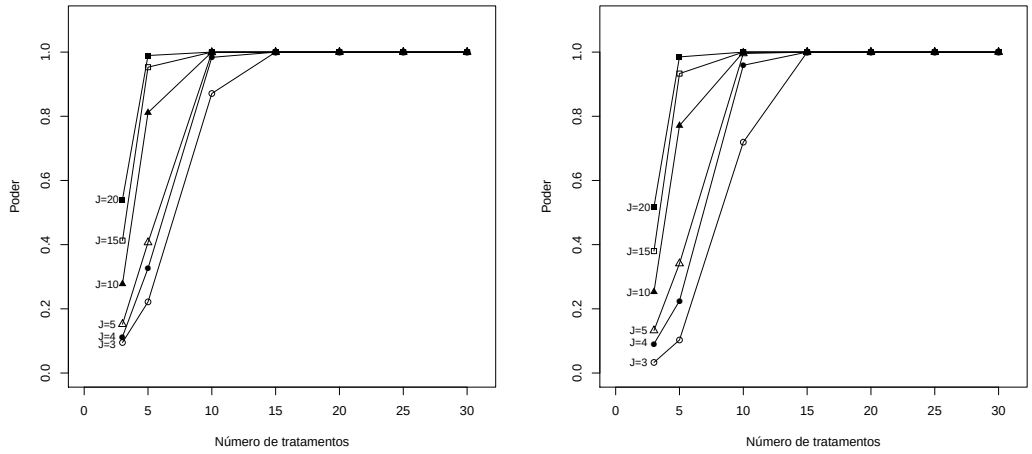
Regarding the situation where $CV = 10\%$, shown on Figure 8, it is observed that the growth rates of empirical power curves are more similar to each other for the F test than for Kruskal-Wallis test. When the number of treatments and replicates is 3, the power of the F test was 9.5%, while the power of the Kruskal-Wallis test was 3.3%, and considering that number of treatments, the power shown by the tests - to the maximum number of repetitions ($J = 20$) - was 54.03% and



(a) (b)

Figure 7: Power rates based on the number of treatments and replications, considering $K = 50$ and $CV = 5\%$, for F (a) and Kruskal-Wallis (b) tests.

51.73%, respectively. The power of the F test was above 80% for all treatments and replications, when the number of treatments was equal to more than 10 and, in this situation, the Kruskal-Wallis test was below this value (71.93%) when the number of repetitions was minimal ($J = 3$). Both tests showed the maximum power from 15 treatments on.

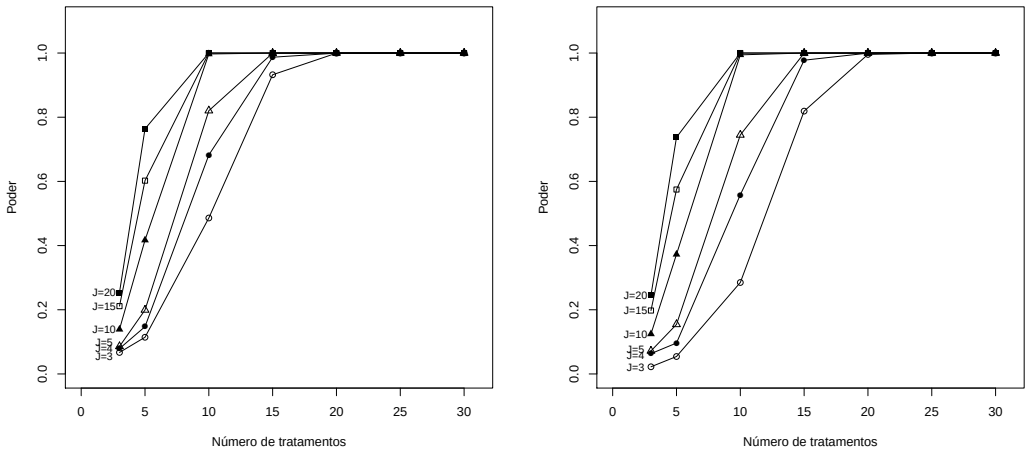


(a) (b)

Figure 8: Power rates based on the number of treatments and replications, considering $K = 50$ and $CV = 10\%$, for F (a) and Kruskal-Wallis (b) tests.

Considering $CV = 15\%$, shown on Figure 9, it is observed that the power varied between 6.7% and 25.36% and between 2.23% and 24.53% for the F and Kruskal-Wallis tests, respectively, for $I = 3$. Also, the curves of F test's empirical power are closer to each other, when compared to the Kruskal-Wallis test. When the number of treatments was less than 15, both tests had a power above 80%, and, with three replicates, F test showed the power 93.23% while the Kruskal-Wallis power presented 81.9%. Both tests presented maximum power when the number

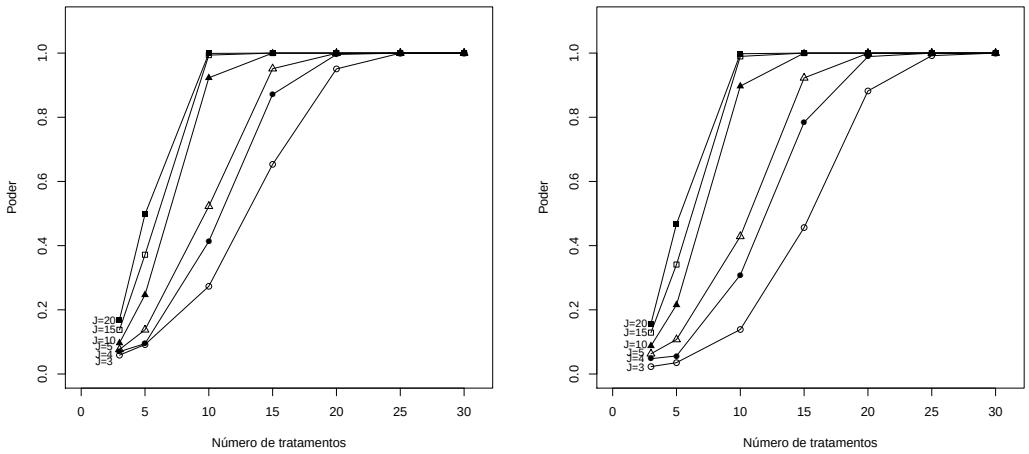
of treatments was equal to or greater than 25, although for 20 treatments the minimum power presented by both tests was close to 100%, (99.57% for the Kruskal-Wallis and 99.97% for the F test).



(a) (b)

Figure 9: Power rates based on the number of treatments and replications, considering $K = 50$ and $CV = 15\%$, for F (a) and Kruskal-Wallis (b) tests.

On Figure 10, we show the power of F test and Kruskal-Wallis tests setting $CV = 20\%$. It may be noted that the empirical power curves are closer together for the F test than for the Kruskal-Wallis test. Considering three treatments, the power of the F test ranged from 5.87% to 16.93% and the Kruskal-Wallis test ranged from 2.23% to 15.7%. Both had power over 80% for the number of treatments equal to or greater than 20 and showed maximum power only when the number of treatments was equal to 30, although they got close to 100% when the number of treatments was equal to 25 (99.97% and 99.20%, respectively).



(a) (b)

Figure 10: Power rates based on the number of treatments and replications, considering $K = 50$ and $CV = 5\%$, for F (a) and Kruskal-Wallis (b) tests.

Penalty factor $K = 100$

On Figure from 11 to 15 we present the power of F test and Kruskal-Wallis tests for $K = 100$. Considering this penalty factor, the F test had power greater than or equal to the Kruskal-Wallis test for all coefficients of variation.

Regarding $CV = 1\%$, shown on Figure 11, both tests presented the maximum power when the number of treatments is equal to or greater than 5. When considering three treatments, the power of the Kruskal-Wallis test was lower than 60% for a small number of repetitions ($J = 3$) and, in this situation, the F test showed power above 90%, growing as the number of repetitions increased.

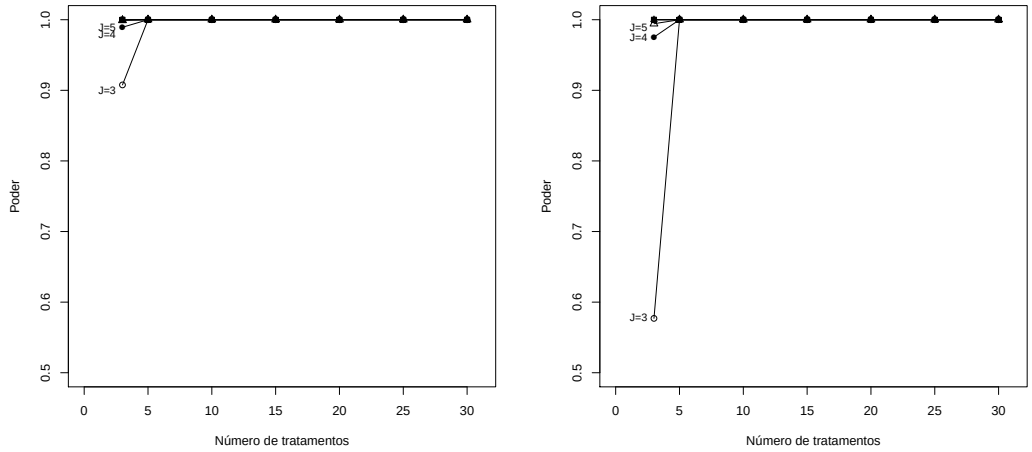
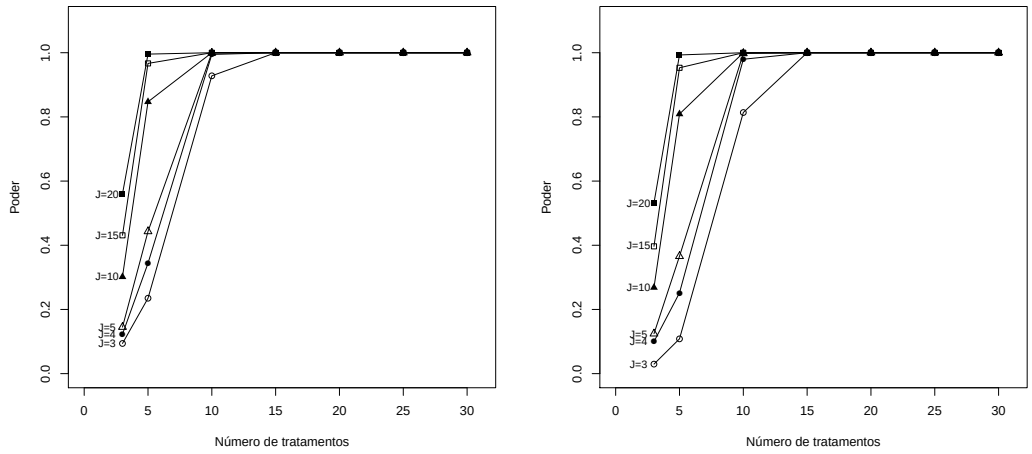


Figure 11: Power rates based on the number of treatments and replications, considering $K = 100$ and $CV = 1\%$, for F (a) and Kruskal-Wallis (b) tests.

In situations where $CV = 5\%$, shown on Figure 12, the two tests had low power when three treatments were considered, both staying below 60%, independent of the number of replications. In this case, the power of the Kruskal-Wallis test is close to zero for a small number of repetitions ($J = 3$), while the power of F test is close to 10%. As the number of treatments increase, increases the power of both tests, reaching the maximum power when the number of treatments is equal to or greater than 15.

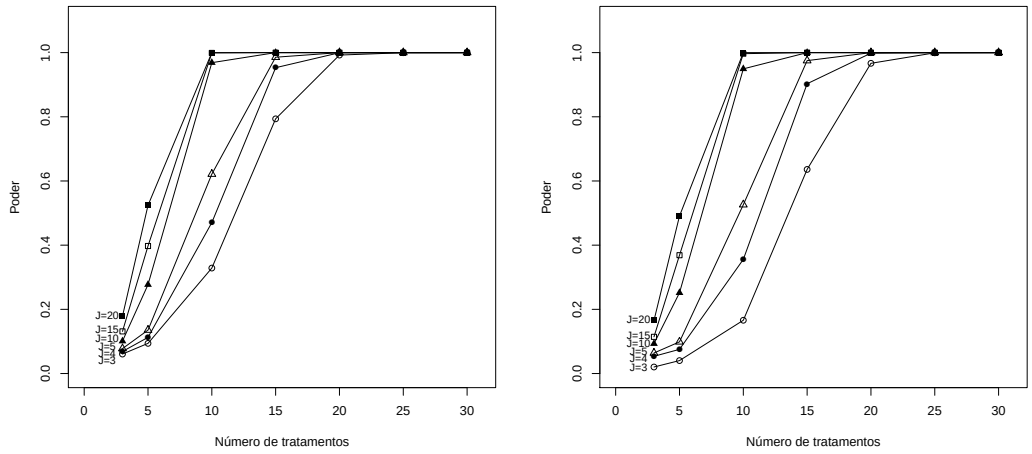
On Figure 13 we show the power of F and Kruskal-Wallis tests, considering $CV = 10\%$. Regarding three treatments, both tests presented power under 20% (17.87% and 16.53%, respectively), regardless of the number of repetitions. It should be noted also that the growth rates of the curves of the F test are more alike than the Kruskal-Wallis test. In the same situation, both the F test and the Kruskal-Wallis test had power greater than 80%, for all numbers of replications, when the number of treatments was equal to of greater than 20, although the F test has shown a power near that value (79.37%) when the number of treatments was equal to 15. Both tests showed maximum power with, at least, 25 treatments.

It is shown on Figure 14, the power of F and Kruskal-Wallis tests in situations where $CV = 15\%$ and in these situations, the empirical power curves remained closer to each other for both F and Kruskal-Wallis tests. Regarding three treatments, the power of the Kruskal-Wallis test ranged from 2% to 9.9%, in the same situation the F test ranged from 5.27% to 10.27%. Both tests exceeded 80% when the number of treatment was equal to or greater than 25. Tests have not reached the maximum power for all numbers of replicates in any number of treatments, although the power of both has been close to the maximum power in situations where the



(a) (b)

Figure 12: Power rates based on the number of treatments and replications, considering $K = 100$ and $CV = 5\%$, for F (a) and Kruskal-Wallis (b) tests.

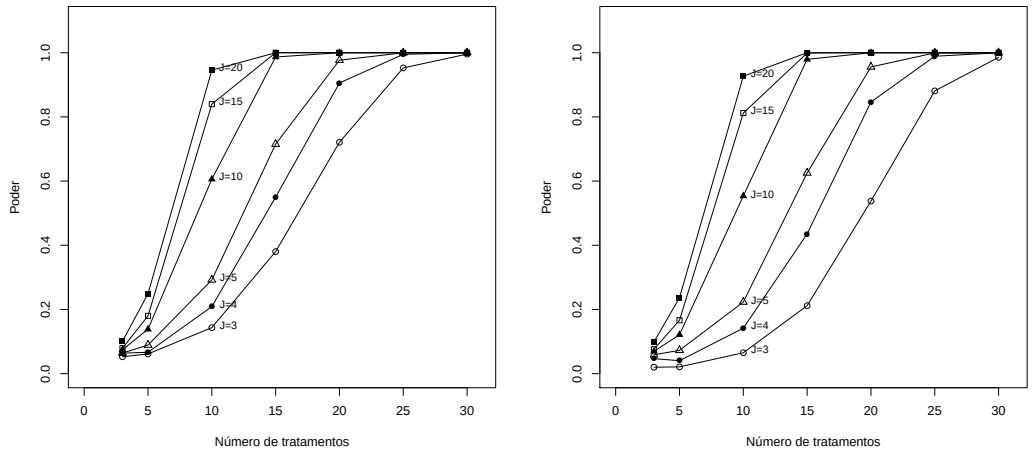


(a) (b)

Figure 13: Power rates based on the number of treatments and replications, considering $K = 100$ and $CV = 10\%$, for F (a) and Kruskal-Wallis (b) tests.

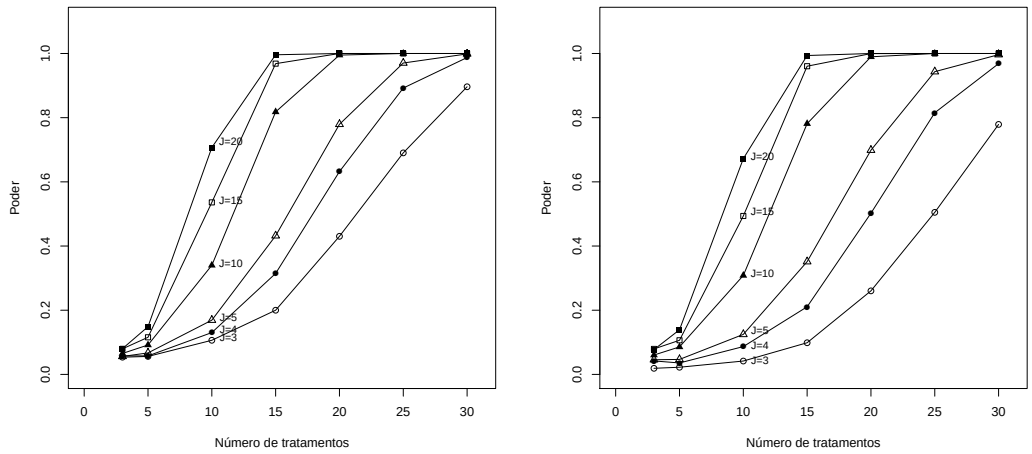
number of treatments was equal to 30 (more than 98.57% and 99.57% for Kruskal-Wallis and F, respectively).

On Figure 15 are displayed the power of F and Kruskal-Wallis tests when $CV = 20\%$. Note that the empirical power curves of the F test are closer together when compared with the Kruskal-Wallis test. Regarding three treatments, the power of the Kruskal-Wallis test ranged between 1.9% and 7.73%, according to the number of replications, while the power of the F test in the same situation, varied between 5.37% and 8.03%. The power of F test reached 80%, despite the number of repetitions, in situations where the number of treatments was maximum ($I = 30$) and in the same situation, the Kruskal-Wallis test did not reach that power for the minimum number of repetitions ($J = 3$), when he presented a power of 77.90%.



(a) (b)

Figure 14: Power rates based on the number of treatments and replications, considering $K = 100$ and $CV = 15\%$, for F (a) and Kruskal-Wallis (b) tests.



(a) (b)

Figure 15: Power rates based on the number of treatments and replications, considering $K = 100$ and $CV = 20\%$, for F (a) and Kruskal-Wallis (b) tests.

Conclusions

The main conclusion of this study was that the assumption of normality of residuals imposed by F test analysis of variance is not strong.

For data with binomial errors in an analysis of variance model of fixed effects in completely randomized design, the F test behaved in general, equally or better than its immediate nonparametric competitor, the Kruskal-Wallis test, controlling the nominal level of significance and presenting high rates of power.

Even when the samples sizes are small (few treatments and/or repetitions) the Kruskal-Wallis test was not better than F test.

This finding suggests that non-normal residuals in small samples should not be a factor that prevents the use of analysis of variance.

Acknowledgements

Acknowledgements to FAPEMIG and CNPq for the scholarships and to the organizer committee of the I Semana da Matemática da Unifal-MG.

References

CAMPOS, H. de. *Estatística experimental não-paramétrica.* 4ª ed. Piracicaba: FEALQ, 1983. 349p.

FEIR, B.; TOOTHAKER, L. The ANOVA F-test versus the Kruskal-Wallis test: a robustness study. *In: Annual Meeting of the American Educational Research Association*, Chicago, 1974. URL
http://eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchVal
Chicago, 1974.

KRUSKAL, W. H; WALLIS, W. A. Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, Washington, v. 47, p. 583-621, 1952.

LIMA, P. C.; ABREU, A. R. de. *Estatística experimental: ensaios balanceados.* Lavras: UFLA, 2000. 99 p.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0, URL
http://www.R-project.org/.

REIS, G. M.; RIBEIRO, J. I. Jr. Comparação de testes paramétricos e não paramétricos aplicados em delineamentos experimentais. In: III Simpósio Acadêmico de Engenharia de Produção, Viçosa, 2007.*Anais do III Simpósio Acadêmico de Engenharia de Produção*, URL
http:www.http://www.saepro.ufv.br/Image/artigos/SA03.pdf, Viçosa, 2007.

SIEGEL, S.; CASTELLAN, N. J. Jr. *Estatística não-paramétrica para ciências do comportamento.* 2 ed. Trad. S. I. C. Carmona. Porto Alegre: Artmed, 2006. 448p.

VIEIRA, S. *Análise de variância (ANOVA).* São Paulo: Atlas, 2006. 204p.