UNIVERSIDADE FEDERAL DE VIÇOSA

GLEYNNER GHIOTTO LIMA DE MENEZES

DESEMPENHO DE TESTES DE HOMOGENEIDADE DE VARIÂNCIAS EM DIFERENTES CENÁRIOS SIMULADOS

VIÇOSA - MINAS GERAIS
2021

GLEYNNER GHIOTTO LIMA DE MENEZES

DESEMPENHO DE TESTES DE HOMOGENEIDADE DE VARIÂNCIAS EM DIFERENTES CENÁRIOS SIMULADOS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Nerilson Terra Santos

Coorientador: Eduardo Campana Barbosa

VIÇOSA - MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

M543d
2021

Menezes, Gleynner Ghiotto Lima de, 1989-
Desempenho de testes de homogeneidade de variâncias em diferentes cenários simulados / Gleynner Ghiotto Lima de Menezes. – Viçosa, MG, 2021.
1 dissertação eletrônica (84 f.): il. (algumas color.).

Orientador: Nerilson Terra Santos.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 81-84.
DOI: https://doi.org/10.47328/ufvbbt.2021.069
Modo de acesso: World Wide Web.

1. Análise de variância. 2. Heterocedasticidade.
3. Estatística robusta . I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.538

Bibliotecário(a) responsável: Alice Regina Pinto CRB6 2523

GLEYNNER GHIOTTO LIMA DE MENEZES

DESEMPENHO DE TESTES DE HOMOGENEIDADE DE VARIÂNCIAS EM DIFERENTES CENÁRIOS SIMULADOS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 30 de junho de 2021.

Assentimento:

Gleynner Ghiotto Lima de Menezes
Autor

Nerilson Terra Santos
Orientador

AGRADECIMENTOS

Agradeço a Deus por estar sempre ao meu lado durante esta trajetória, por sua imensa bondade e por me guiar em todos os momentos.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade concedida para realização deste curso.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, pela concessão da bolsa de estudos.

Ao D. Sc. Nerilson Terra Santos, pela orientação, empenho, confiança, paciência, amizade, oportunidade e pelos ensinamentos transmitidos.

Ao meu coorientador D. Sc. Eduardo Campana Barbosa, pelas contribuições e sugestões valiosas.

Aos pesquisadores D. Sc.s: Luiz Alexandre Peternelli, José Ivo Ribeiro Júnior, Rodrigo Luiz Pereira Lara e Calisto Manuel Máquina, por aceitarem participar da minha banca de defesa.

Aos meus pais, Duclerk e Vilma, pela confiança, incentivos e orações, isso me deu forças para entregar o melhor de mim. Obrigado por seu amor incondicional!

Aos meus irmãos, Gleiph e Gleybber, pelo companheirismo, conselhos e risadas.

Aos meus amigos do PPESTBIO e LAPEA, em especial à Marianna e Sabrina, pelos momentos de descontração, companheirismo, pela troca de experiência e incentivo.

Além de, todos os professores e funcionários do Programa de Pós-Graduação em Estatística Aplicada e Biometria.

Muito obrigado a todos que de alguma forma contribuíram para realização desta etapa.

RESUMO

MENEZES, Gleynner Ghiotto Lima, M.Sc., Universidade Federal de Viçosa, junho de 2021. **Desempenho de testes de homogeneidade de variâncias em diferentes cenários simulados.** Orientador: Nerilson Terra Santos. Coorientador: Eduardo Campana Barbosa.

A confiabilidade nos resultados obtidos a partir dos testes de hipóteses estão sujeitos ao atendimento de pressuposições, o qual, quando pelo menos uma delas não é satisfeita, seu desempenho ou nível de confiança pode estar comprometido, levando a conclusões errôneas. Deste modo, existem diversos testes na literatura que foram propostos a fim de verificar a suposição de homogeneidade de variâncias em análises estatísticas, sendo esta tomada por diversos autores como o fator de maior influência sobre a sensibilidade dos resultados. No entanto, não existe um consenso sobre o melhor cenário de aplicação para cada um deles. Neste trabalho, pretende-se comparar os testes de homogeneidade de variâncias paramétricos de Bartlett, Levene, Brown- Forsythe, Cochran e Hartley, e os testes não paramétricos de Fligner-Killeen, Conover e Mood, através de um estudo de simulação utilizando o *software R*, onde, serão realizadas comparações segundo um Delineamento Inteiramente Casualisado sobre os seguintes aspectos de avaliação: proporção de heterogeneidade, proporção de desbalanceamento e diferentes distribuições de probabilidades. A hipótese de homocedasticidade foi adotada para analisar a taxa empírica do erro tipo I ($\hat{\alpha}$) e, a de heterocedasticidade, para analisar a taxa empírica do poder do teste ($\hat{\pi}$). Diante disso, foi observado que, sob distribuição normal, o teste paramétrico de Bartlett obtém o melhor controle da taxa empírica do erro tipo I e obtém alto poder nos cenários balanceados e desbalanceados. Quando os conjuntos de dados são provenientes de populações não normais, o teste paramétrico de Brown-Forsythe foi o mais indicado. Dentre os testes não paramétricos, o teste de Mood foi o mais indicado para atuar sobre as três distribuições de probabilidades avaliadas.

Palavras-chave: Heterocedasticidade. Robustez. Poder.

ABSTRACT

MENEZES, Gleynner Ghiotto Lima, M.Sc., Universidade Federal de Viçosa, June, 2020. **Performance of homogeneity tests of variances in different simulated scenarios.** Adviser: Neriolson Terra Santos. Co-adviser: Eduardo Campana Barbosa.

The reliability of the results obtained from the hypothesis tests are subject to the fulfillment of assumptions, or which, when at least one of them is not satisfied, its performance or level of confidence may be compromised, leading to erroneous granting. Thus, there are several tests in the literature that have been proposed in order to verify the assumption of homogeneity of variances in statistical analyses, which is taken by several authors as the major influencing factor on the sensitivity of the results. However, there is no consensus on the best application scenario for each of them. In this work, we intend to compare the Bartlett, Levene, Brown-Forsythe, Cochran and Hartley parametric variance homogeneity tests, and the Fligner-Killeen, Conover and Mood nonparametric tests, through a study of simulation of the use of R software, where comparisons will be performed according to an Completely Randomized Design on the following evaluation aspects: heterogeneity proportion, imbalance proportion and different probability distributions. The homoscedasticity hypothesis was adopted to analyze an empirical type I error rate ($\hat{\alpha}$) and the heteroscedasticity hypothesis to analyze an empirical test power rate ($\hat{\pi}$). Therefore, it was observed that, under normal distribution, Bartlett's parametric test obtained the best control of the empirical type I error rate and obtained high power in balanced and unbalanced scenarios. When the data sets come from non-normal populations, the Brown-Forsythe parametric test was the most indicated. Among the non-parametric tests, the Mood test was the most indicated to act on the three probability distributions evaluated.

Keywords: Heteroscedasticity. Robustness. Power.

LISTA DE SÍMBOLOS

α	Nível de significância $(\alpha = 0{,}05)$
$\hat{\alpha}$	Taxa empírica do erro tipo I
$\hat{\pi}$	Taxa empírica do poder do teste
σ_i^2	Variância referente a i-ésima população
H_0	Hipótese nula
H_A	Hipótese alternativa
k	Número de repetições da população
N	Tamanho total da amostra $(\sum_{i=1}^{t} r_i = N)$
r_i	Tamanho da amostra da população i
$\bar{r}_i$	Número de observações médio da i-ésima população
R_k	*Rank* da k-ésima observação ordenada
S_i^2	Estimador da variância para a população i
s_i^2	Estimativa da variância para a i-ésima população
s_{min}^2	Menor estimativa da variância populacional $= min(s_1^2,\ s_2^2, \dots, s_t^2)$
s_{max}^2	Maior estimativa da variância populacional $= max(s_1^2,\ s_2^2, \dots, s_t^2)$
t	Número de populações (ou tratamentos)
Y_{ik}	Valor observado da k-ésima observação da i-ésima população
$\bar{Y}_{i.}$	Média da i-ésima população $\left(\bar{Y}_{i.} = \sum_{k=1}^{r_i} \frac{Y_{ik}}{r_i}\right)$
$\tilde{Y}_{i.}$	Mediana da amostra referente a i-ésima população
Z_{ik}	Valor transformado de Y_{ik}
$\bar{Z}_{i.}$	Média dos valores transformados da i-ésima população $\left(\bar{Z}_{i.} = \sum_{k=1}^{r_i} \frac{Z_{ik}}{r_i}\right)$
$\bar{Z}_{..}$	Média geral dos Z_{ik} valores transformados $\left(\bar{Z}_{..} = \sum_{k=1}^{r} \frac{r_i \bar{Z}_{i.}}{N}\right)$

SUMÁRIO

1. Introdução

Segundo Martin e Games (1977), os testes de homogeneidade de variâncias são frequentemente associados à premissa de homogeneidade de variâncias (HOV), necessária para apoiar a precisão de certos testes de médias.

Uma outra razão para que se deva testar a igualdade de um conjunto de variâncias populacionais ocorre quando existe o interesse a priori na variabilidade. Teorias relacionadas a educação e a psicologia lançam hipóteses sobre a igualdade de variâncias, como, por exemplo, na área de medição da capacidade cognitiva, de acordo com estudos de Cattell (1971) e Gagné (1970). Além disso, na área da genética quantitativa, o teste de homogeneidade de componentes de variância é utilizado na comparação de parâmetros fenotípicos e genéticos entre populações (LYNCH; WALSH, 1998). Além disso, na área da medicina, os testes de homogeneidade de variâncias são utilizados para avaliar a variabilidade dos níveis de metilação do DNA, que é considerada um regulador da expressão do gene humano, tendo diferentes níveis médios e sua maior variabilidade pode estar associada à cânceres e outras doenças complexas (LI *et al.*, 2015).

Os testes de homogeneidade de variâncias têm também utilidade na avaliação da pressuposição de homogeneidade das variâncias das populações para as quais se deseja testar a igualdade de suas médias ou efeitos dos seus tratamentos (NOGUEIRA; PEREIRA, 2013), como exemplo, o teste F da ANOVA, a qual é uma das metodologias estatísticas mais utilizadas na análise de dados oriundos de pesquisa (JAN; SHIEH, 2014). Além da homogeneidade de variâncias, a ANOVA também pressupõe que os erros aleatórios associados ao modelo da análise estatística são independentes e identicamente distribuídos, segundo um modelo de probabilidade normal. Além disso, é amplamente conhecido que o teste F da ANOVA é tendencioso quando pelo menos uma de suas pressuposições são violadas (CHOI, 2005; COCHRAN, 1947; CRIBBIE *et al.*, 2012; HOEKSTRA; KIERS; JOHNSON, 2012).

A confiabilidade nos resultados obtidos a partir da aplicação dos testes de hipóteses está ligada ao atendimento de suas pressuposições. Nem sempre todas as pressuposições são satisfeitas, situação em que Micceri (1989) e Golinski e Cribbie (2009) se referem como algo corriqueiro. Neste caso, o nível de significância e a

sensibilidade das estatísticas de teste podem ser comprometidos, sendo, a sensibilidade mais afetada pela ausência de homogeneidade de variância das populações (BAYDILI; SIĞIRLI, 2017; MILLIKEN; JOHNSON, 1992). Riboldi *et al.* (2014) afirmam que, sob heterogeneidade de variâncias, o método dos mínimos quadrados ordinários não produz os melhores estimadores e o teste F, as comparações múltiplas, os contrastes ortogonais ou a estimação dos componentes de variância poderão ser fortemente afetados. Porém, desvios moderados da suposição de homogeneidade de variâncias podem ser aceitáveis e não afetar seriamente os resultados da ANOVA, fazendo com que os pesquisadores se preocupem apenas com maiores desvios dessa suposição (GLASS; PECKHAM; SANDERS, 1972; MIRTAGIOĞLU *et al.*, 2017).

Como a pressuposição de homogeneidade de variâncias demonstrar ter grande importância para os testes paramétricos, determinar o desempenho dos testes comumente utilizados na prática para verificar a suposição de homogeneidade de variâncias sob diferentes condições experimentais é essencial (BARTLETT, 1937; BROWN; FORSYTHE, 1974; CONOVER, 1999; VORAPONGSATHORN; TAEJAROENKUL; VIWATWONGKASEM, 2004; HATCHAVANICH, 2014; WANG *et al.*, 2017; WLUDYKA; NELSON, 1997).

Visto o interesse de algumas áreas sobre o estudo da variabilidade e sua importância na verificação da suposição de homocedasticidade, testar a HOV é uma etapa importante antes da realização da análise de variância. Desta forma, o presente trabalho tem o objetivo de avaliar o desempenho de alguns testes de homogeneidade de variâncias paramétricos e não paramétricos, usando cenários simulados, para identificar aqueles testes que fornecem os melhores desempenho em cada aspecto avaliado nesse estudo.

2. Objetivos

2.1. Objetivo Geral

Comparar alguns testes de homogeneidade de variâncias paramétricos e não paramétricos, sob diferentes cenários, em relação a taxa empírica do erro tipo I ($\hat{\alpha}$) e o poder empírico do teste ($\hat{\pi}$).

2.2. Objetivos Específicos

Especificamente, os cenários simulados visam identificar o teste mais adequado considerando a distribuição de probabilidades dos dados, a proporção de heterogeneidade das variâncias de populações e a proporção de desbalanceamento amostral.

3. Revisão Bibliográfica

Segundo Casella e Berger (2011), uma hipótese é tida como uma suposição a respeito de um parâmetro de uma ou mais populações. O teste de hipóteses tem o objetivo de decidir, com base em uma amostra da população, qual de suas hipóteses complementares é a verdadeira, a hipótese nula (H_0) ou a hipótese alternativa (H_A). Além disso, é o conjunto de características contidas nos dados que dirão se uma hipótese será rejeitada ou não (RICARDI, 2011). Em um teste de homogeneidade de variâncias, para $t > 2$ populações, a hipótese de nulidade e a hipótese alternativa são definidas como:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 \\ H_A : pelo\ menos\ uma\ das\ variânicas\ se\ difere\ das\ demais \end{cases}$$

em que σ_i^2 é a variância da i-ésima população, para $i = 1, 2, \ldots, t$.

Após formuladas as hipóteses estatísticas, o experimentador deverá escolher, dentre os testes existentes na literatura, àquele que mais se adequa ao conjunto de dados disponível, considerando as pressuposições e as características do teste, tal como a sua robustez e o seu poder. Por exemplo, um teste de homogeneidade é robusto a normalidade quando, em populações com distribuições distorcidas (ou seja, distribuições diferentes da normal), o teste não indica diferenças significativas de variâncias quando de fato essas diferenças não existem. Ou seja, quando não se rejeita H_0, dado que H_0 é verdadeira e os dados populacionais não são normalmente distribuídos. De modo mais geral, um teste é tido como robusto se não for sensível às suposições iniciais as quais está sujeito (ASSIS; SOUSA; DIAS, 2019). Por poder, entendemos como a capacidade do teste em detectar variações desiguais quando as variações são de fato desiguais, ou, de outra forma, é a probabilidade de rejeitar a hipótese nula quando ela é realmente falsa (CASELLA; BERGER, 2011).

Testar a homogeneidade de variâncias em populações normais é comum em muitas análises estatísticas (DYER; KEATING, 1980), porém, nem sempre a distribuição normal dos erros ocorrem. Alguns testes de HOV têm como pressuposição a normalidade dos dados (como os testes de Bartlett, Levene, Brown-Forsythe, Cochran, Hartley, entre outros testes paramétricos) e outros que são robustos à falta dela, como o teste paramétrico de Brown-Forsythe e outros testes não paramétricos (BAYDILI; SIĞIRLI, 2017; CONOVER; JOHNSON; JOHNSON, 1981; LI

et al., 2015; MIRTAGIOĞLU *et al.*, 2017; RAMSEY; RAMSEY, 2007; RIBOLDI *et al.*, 2014; SHARMA; KIBRIA, 2013).

Na literatura, diversos estudos utilizaram a simulação de dados para avaliar os testes de HOV e quais características dos dados podem afetar o seu desempenho, no entanto, o valor obtido para a taxa de erro tipo I segue algumas classificações. Segundo Biase e Ferreira (2011), um teste é considerado conservador quando a taxa empírica de erro tipo I é inferior ao nível nominal de significância (α); considerado liberal, se a taxa de erro tipo I observada é superior ao nível α; e, exato, quando a taxa empírica do erro tipo I é igual ao nível de significância adotado.

Riboldi *et al.* (2014) fizeram uso de simulação para comparar os testes paramétricos de Bartlett, Brow-Forsythe, O'Brien, Levene (resíduos absolutos e resíduos quadrado); e os testes não paramétricos de Siegel-Tukey, Ansari-Bradley, Klotz e Mood. Os autores concluíram que o teste de Bartlett, sob normalidade, é um teste exato e com alto poder, influenciado pelo desbalanceamento dos dados. De acordo com o mesmo estudo, o teste de Levene foi considerado robusto à normalidade. Já os testes de Brow-Forsythe e O'Brien, foram conservadores e tiveram baixo poder. Os testes não paramétricos se assemelharam quanto ao poder.

Vorapongsathorn, Taejaroenkul e Viwatwongkasem (2004) utilizaram o método de Monte Carlo para simular dados (alterando a distribuição de probabilidades, variâncias e tamanhos amostrais) e avaliaram o desempenho dos testes de homogeneidade de variâncias de Bartlett, Levene e Cochran. Os autores concluíram que o teste de Bartlett foi sensível a violação da suposição de normalidade, enquanto os demais se comportaram de forma robusta. O teste de Levene mostrou bom desempenho (ou seja, quando $\hat{\alpha}$ e $\hat{\pi}$ atendem aos níveis estabelecidos de α e $1-\beta$, respectivamente) para amostras pequenas e balanceadas. Ademais, existem diversos trabalhos que utilizaram a simulação de dados para avaliar o desempenho dos testes de hipóteses e suas pressuposições (LEE; KATZ; RESTORI, 2010; MIRTAGIOĞLU *et al.*, 2017) e, para avaliar a suposição de normalidade (CANTELMO; FERREIRA, 2007).

Tendo em vista os trabalhos apresentados acima, dentre outros existentes na literatura, foi proposto nesta dissertação a construção de uma obra em que se pudesse ser abrangido uma quantidade maior de tamanhos de amostras, testadas sob

diferentes distribuições de probabilidades e proporções de heterogeneidade de variâncias e, posteriormente, analisado o desempenho de cada teste com relação a taxa empírica de erro do tipo I e a taxa empírica de poder do teste. Para padronizar a notação de equações e de procedimentos dos testes de HOV, são apresentados na "Lista de Símbolos" a notação geral e os seus respectivos significados, os quais serão utilizados em todas as seções ao se referir aos testes de HOV. Para isso, suponha que os dados (isto é, os valores observados da variável resposta) tenham sidos coletados a partir de t amostras independentes, com cada amostra contendo r_i observações.

A seguir são apresentados os testes de homogeneidade de variâncias envolvidos neste estudo de simulação e suas respectivas estatísticas de teste. Esses testes de HOV podem ser divididos em testes paramétricos e não paramétricos.

3.1. Testes paramétricos

Os testes paramétricos são aqueles que utilizam parâmetros específicos da distribuição para testar suas hipóteses, tais como, média, desvio-padrão, variância ou proporção. No entanto, segundo Siegel e Castellan Jr. (2006), a validação dos resultados obtidos para os testes paramétricos depende da verificação de suas pressuposições, por exemplo, a normalidade dos dados, sendo essa uma pressuposição básica para aplicação da maioria dos testes paramétricos.

3.1.1. Teste de Bartlett

O teste de homogeneidade de variâncias de Bartlett nos permite avaliar se t populações possuem variâncias idênticas, contudo, este teste presume que a variável em estudo seja normalmente distribuída (CONOVER; JOHNSON; JOHNSON, 1981). O cálculo para este teste envolve uma estatística cuja distribuição amostral é aproximada pela distribuição de qui-quadrado com $t-1$ graus de liberdade, quando as t amostras aleatórias são de populações normais independentes (MONTGOMERY, 1997). A estatística de Bartlett para testar a hipótese de homogeneidade de variâncias é dada por

$$\chi^2 = \frac{(N-t)log\left[\frac{\sum_{i=1}^{t}(r_i-1)S_i^2}{(N-t)}\right] - \sum_{i=1}^{t}(r_i-1)\log(S_i^2)}{1+\frac{\left(\sum_{i=1}^{t}\frac{1}{r_i-1}\right)-\frac{1}{N-t}}{3(t-1)}} \qquad (1)$$

O resultado obtido pela equação (1) é comparado com o valor de qui-quadrado crítico (ou tabelado) dado por $\chi^2_{(t-1;1-\alpha)}$. O teste de Bartlett é derivado do teste da razão de verossimilhança e, portanto, uma de suas principais desvantagens é que este teste é extremamente sensível à desvios de normalidade e, como, na prática, heterogeneidade e não normalidade geralmente ocorrem simultaneamente, dificilmente saberemos devido a qual dos dois fatores (ou sob a influência de ambos) a hipótese nula pode ser rejeitada (ARSHAM; LOVRIC, 2011).

3.1.2. Teste de Levene

O teste de Levene, associado a uma estatística W, é utilizado para testar se t populações são homocedásticas, podendo fazer uso tanto dos valores absolutos dos resíduos quanto dos valores dos resíduos ao quadrado, transformando este em um teste de variâncias relativamente robusto à suposição de normalidade (WANG *et al.*, 2017). A estatística de teste é dada pela equação abaixo

$$W = \frac{(N-t)\sum_{i=1}^{t} r_i\,(\bar{Z}_{i.} - \bar{Z}_{..})^2}{(t-1)\sum_{i=1}^{t}\sum_{k=1}^{r_i}(Z_{ik} - \bar{Z}_{i.})^2} \qquad (2)$$

O valor transformado (Z_{ik}) pode ser obtido das seguintes formas:

- Desvios absolutos: $Z_{ik} = |Y_{ik} - \bar{Y}_{i.}|$; e

- Desvios quadrados: $Z_{ik} = (Y_{ik} - \bar{Y}_{i.})^2$

Em que

$\bar{Y}_{i.}$: valor médio da amostra referente a i-ésima população, ou seja, $\bar{Y}_{i.} = \sum_{k=1}^{r_i} \frac{Y_{ik}}{r_i}$.

Após obtido o valor da estatística de teste, dado pela equação (2), seu valor deve ser comparado com o valor crítico de F (ou seja, o valor tabelado de F), definido como $F_{tab} = F_{(t-1,\ N-t;\,1-\alpha)}$, em que α é o nível de significância (BAYDILI; SIĞIRLI, 2017). A hipótese nula de igualdade de variâncias é rejeitada se $W \geq F_{tab}$.

3.1.3. Teste de Brown-Forsythe

Brown e Forsythe (1974) propuseram a utilização da mediana da população ao invés da média populacional no cálculo dos valores residuais absolutos de Levene, esperando-se, assim, que o teste seja mais robusto quando a distribuição da

população apresentar desvios da normalidade (WANG *et al.*, 2017). O valor transformado (Z_{ik}) é obtido conforme mostrado abaixo

$$Z_{ik} = \left|Y_{ik} - \tilde{Y}_{i.}\right|$$

onde, $\tilde{Y}_{i.}$ é a mediana da amostra referente a i-ésima população. O teste de Brown-Forsythe segue a mesma estatística de teste de Levene, dada pela equação (2).

Assim, como no teste de Levene, a estatística W deve ser comparada com o valor de F_{tab}. Quando o valor de W for igual ou superior ao valor tabelado de F, rejeita-se a hipótese de homogeneidade de variâncias. Caso contrário, as variâncias populacionais serão estatisticamente iguais.

3.1.4. Teste de Cochran

O teste de Cochran é dado pela razão entre a maior variância e a soma de todas as variações amostrais populacionais (COCHRAN, 1941). Para aplicar o teste assume-se que o experimento seja balanceado, ou seja, $r_1 = r_2 = \cdots = r$. Para um projeto desbalanceado, ou seja, que possuem amostras de tamanho desiguais, pode-se usar tanto a média aritmética quanto a média harmônica de r_i no lugar de r para calcular o número de graus de liberdade (WANG *et al.*, 2017). Ele testa a homogeneidade de variâncias e, quando o valor obtido para a estatística C exceder o valor crítico, a hipótese nula é rejeitada. A estatística do teste é dada por

$$C = \frac{s^2_{i\,max}}{\sum_{i=1}^{t} s_i^2} = \frac{maior\ variância}{soma\ de\ todas\ as\ variâncias} \qquad (3)$$

O valor crítico do teste de Cochran (ou, valor tabelado) é obtido por meio do número de populações, t, e o número de graus de liberdade, $r - 1$ (COCHRAN, 1941). Sendo que cada uma das t populações tem $r - 1$ graus de liberdade, quando o experimento é balanceado, ou, a média aritmética (ou harmônica) do número de observações é utilizada para determinar o número de graus de liberdade em experimentos desbalanceados.

3.1.5. Teste de Hartley

Hartley (1950) desenvolveu o teste de Hartley (ou teste F-max) para comparar três ou mais variâncias populacionais a partir da razão entre a maior e a menor variância amostral. Esse teste requer que as amostras aleatórias sejam

independentes e com mesmo número de observações entre os tratamentos (ou número de repetições aproximados), obtidas de populações normalmente distribuídas. No entanto, o teste F-max também pode ser aplicado quando houver pequena diferença entre os tratamentos e, para isso, o maior tamanho de amostra deve ser usado para o cálculo do grau de liberdade (BHANDARY; DAI, 2008). A estatística de teste de Hartley, F_{max}, segue a equação (4), dada por

$$F_{max} = \frac{s_{max}^2}{s_{min}^2} \qquad (4)$$

Para concluir, é necessário comparar o valor de F_{max} com o valor crítico da tabela de Hartley, que é obtido em função do número de populações, t, e o número de graus de liberdade, $r - 1$. Em experimentos desbalanceados o grau de liberdade é obtido por: $\max\{r_1, r_2, \ldots, r_t\}$ -1 (BHANDARY; DAI, 2008).

3.2. Testes não paramétricos

Segundo Pontes e Corrente (2001), os testes não paramétricos utilizam para o cálculo de sua estatística, os postos atribuídos com base na ordenação (*ranks*) dos dados e não em seu valor real, isto significa, que para uma amostra aleatória, o *rank* corresponde ao número natural que indica a posição de determinado valor em uma lista ordenada. Desse modo, tais testes são livres de distribuição de probabilidades dos dados (GORBUNOVA; LEMESHKO, 2011), ou seja, não exigem que os dados sejam normalmente distribuídos. Porém, Odiase e Ogbonmwan (2008) afirmam que a maior dificuldade no uso dos testes não paramétricos é a indisponibilidade de valores críticos exatos.

3.2.1. Teste de Conover

Este teste proposto por Conover (1999) se baseia nos *ranks* dos dados amostrais. Como esse teste não pressupõe normalidade, ele é recomendado, principalmente, quando a suposição de normalidade não é satisfeita. No entanto, o teste exige que os dados sejam obtidos a partir de uma amostra de cada uma das t populações (MIRTAGIOĞLU et al., 2017). A estatística do teste de Conover é obtida por

$$T_2 = \frac{1}{D^2}\sum_{i=1}^{t}\frac{S_i^2}{r} - N\bar{S}^2 \qquad (5)$$

Em que

$S_i = \sum_{i=1}^{t} R_{ik}^2$

$R_{ik} = Rank(Z_{ik})$

$Z_{ik} = |Y_{ik} - \bar{Y}_{i.}|$

$\bar{S}_i = \frac{1}{N}\sum_{i=1}^{t} S_i$

$D^2 = \frac{1}{N-1}\left(\sum_{i=1}^{t}\sum_{k=1}^{r_i} R_{ik}^4 - N\bar{S}^2\right)$

A estatística do teste tem distribuição de qui-quadrado com $t-1$ graus de liberdade. A hipótese nula, de que as variâncias populacionais são estatisticamente iguais, é rejeitada se T_2 exceder ao quartil $(1-\alpha)$ da distribuição de qui-quadrado com $t-1$ graus de liberdade (CONOVER, 1999).

3.2.2. Teste de Fligner-Killeen

Este teste foi originalmente proposto por Fligner e Killeen (1976) para testar se as variâncias de t populações são homogêneas. Em sua proposta, os autores sugeriram classificar conjuntamente os valores absolutos $|Y_{ik}|$ e atribuir pontuações (scores) crescentes, tais como, $a_{N,i} = i$, $a_{N,i} = i^2$ ou $a_{N,i} = \Phi^{-1}((1 + i/(N+1))/2)$ com base nos *ranks*. Conover, Johnson e Johnson (1981) propuseram modificar o teste original de Fligner e Killeen e classificar $\left|Y_{ik} - \tilde{Y}_{i.}\right|$ ao invés de apenas $|Y_{ik}|$. Esta versão modificada é conhecida como teste de Fligner-Killeen com centralização na mediana. Após obtidos os scores $a_{N,i}'s$ atribuídos à $\left|Y_{ik} - \tilde{Y}_{i.}\right|$, o teste de Fligner-Killeen pode ser calculado utilizando a equação abaixo

$$\chi^2 = \sum_{i=1}^{t} r_i(\bar{A}_i - \bar{a})^2/V^2 \qquad (6)$$

onde $\bar{A}_i$ é a pontuação média para a i-ésima população e $\bar{a}$ é a pontuação média geral. Adicionalmente, tem-se que

$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_{N,i}$; e

$V^2 = \frac{1}{N-1}\sum_{i=1}^{N}(a_{N,i} - \bar{a})^2$.

A estatística de teste χ^2 tem distribuição assintótica de qui-quadrado com $t-1$ graus de liberdade. A hipótese de homogeneidade de variâncias é rejeitada se χ^2 exceder ao valor tabelado da distribuição de qui-quadrado, com $t-1$ graus de liberdade para o quartil $(1-\alpha)$.

3.2.3. Teste de Mood

Este teste, desenvolvido por A. M. Mood, em 1954, é baseado na soma dos desvios quadrados das classificações de uma amostra da classificação média das amostras combinadas (FAHOOME, 2002). Esse teste assume que as amostras são independentes e identicamente distribuídas sob a hipótese nula. Caso que, em particular, este requer médias iguais, ou uma transformação conhecida para alcançar médias iguais. Ao invés de utilizar R_k como sendo o posto de Y_{ik} quando as médias são iguais, o desvio $(Y_{ik} - \overline{Y_i})$ é utilizado para a classificação dos postos (CONOVER; JOHNSON; JOHNSON, 1981). A estatística para este teste é

$$M = \sum_{k=1}^{N} \left(R_k - \frac{N+1}{2} \right)^2 \qquad (7)$$

onde R_k é o rank da k-ésima observação ordenada e N é o número total de observações do conjunto de dados.

4. Materiais e Métodos

Neste trabalho foram avaliados os testes de homogeneidade de variâncias paramétricos (seção 3.1.) e não paramétricos (seção 3.2.) considerando vários aspectos que podem influenciar em sua habilidade de testar a homogeneidade de variâncias em diferentes populações. Essas populações podem ser entendidas como os tratamentos que se deseja comparar os seus efeitos em um experimento. Os aspectos avaliados nesta pesquisa foram: homogeneidade, desbalanceamento (incluindo diferentes tamanhos de amostras, ou seja, repetições em um experimento) e distribuição de probabilidade dos conjuntos de valores populacionais. Com esta finalidade, foram gerados conjuntos de dados referentes a diferentes cenários, que visam retratar aqueles aspectos anteriormente mencionados.

Os cenários simulados seguem um Delineamento Inteiramente Casualizado (DIC) de um fator em estudo com t níveis (tratamentos) cujo modelo estatístico é dado por

$$y_{ik} = \mu + \tau_i + e_{ik}$$

Em que

- y_{ik} é o valor observado para a k-ésima repetição da i-ésima população, tal que $i = 1, 2, \dots, t$ e $k = 1, 2, \dots, r_i$;
- μ é a média geral;
- τ_i é o efeito do i-ésimo tratamento, tal que: $\tau_i = \mu_i - \mu$ e μ_i é a média do i-ésimo tratamento;
- e_{ik} é o efeito do erro aleatório, tal que, $e_{ik} = y_{ik} - \mu_i$;

Os testes de homogeneidade de variâncias apresentados na seção anterior são aplicados aos diferentes cenários, criados para avaliá-los, segundo os parâmetros estabelecidos para a simulação. Esses cenários foram planejados de modo que pudessem ser avaliadas a robustez dos testes a violação de suas pressuposições (normalidade, balanceamento e homogeneidade de variâncias). Essa avaliação foi mensurada a partir das taxas empíricas do erro tipo I e do poder do teste. Para obter essas taxas, parte dos cenários foram gerados sob a condição que a hipótese de nulidade é verdadeira, ou seja, $H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$. A outra parte dos cenários

foram gerados sob a condição que H_0 é falsa, ou seja, pelo menos uma das variâncias se difere estatisticamente das demais.

Nos cenários simulados sob a condição de H_0 verdadeira foi calculada a taxa empírica do erro tipo I por

$$\hat{\alpha} = \frac{\#RH_o|H_0\ verdadeira}{\#H_0\ verdadeira} \quad (8)$$

Em que

$\hat{\alpha}$: taxa empírica do erro tipo I;

$\#RH_o|H_0\ verdadeira$: número de conjunto de dados dos cenários para os quais H_0 foi rejeitada entre todos os conjuntos de dados dos cenários simulados sob a condição que H_0 é verdadeira;

$\#H_0\ verdadeira$: número total de conjuntos de dados dos cenários simulados sob a condição que H_0 é verdadeira. Para o presente trabalho, $\#H_0\ verdadeira$ foi igual a 1.000.

Para classificar a robustez dos testes quanto a violação de suas pressuposições, a taxa empírica do erro tipo I foi comparada com o nível de significância de 5%, ou seja, $\alpha = 0{,}05$. Contudo, para compensar possíveis variações devido a erros de amostragem, foi estabelecida uma margem de variação para a taxa empírica do erro tipo I em cada uma das classes de robustez dos testes, utilizando uma adaptação do Critério Liberal de Bradley (1978), tal como utilizados por Ramsey (1994), Wang *et al.* (2017) e Kim e Cribbie (2018). O método adaptado para o presente trabalho é apresentado a seguir

1- Se $\hat{\alpha} \leq \alpha - 0{,}25\alpha$ (ou seja, $\hat{\alpha} \leq 0{,}0375$) o teste foi classificado como não robusto e conservador. Nesse caso, como a taxa empírica do erro tipo I é inferior ao nível de significância pré-estabelecido $\alpha = 0{,}05$, em condições de violação da pressuposição, o teste tende a rejeitar H_0 numa porcentagem menor do que o esperado. Esse resultado indica que o teste não foi capaz de controlar o erro tipo I quando a pressuposição não é satisfeita.
2- Se $\alpha - 0{,}25\alpha < \hat{\alpha} < \alpha + 0{,}25\alpha$ (ou seja, $0{,}0375 < \hat{\alpha} < 0{,}0625$) o teste é dito ser exato e robusto à violação da pressuposição. Esse resultado indica que

o teste é capaz de controlar a taxa de erro tipo I quando a pressuposição não é satisfeita.

3- Se $\hat{\alpha} \geq \alpha + 0{,}25\alpha$ (ou seja, para $\hat{\alpha} \geq 0{,}0625$) o teste é dito ser não robusto e liberal. Nesse caso, como a taxa empírica de erro tipo I é superior ao nível pré-estabelecido $\alpha = 0{,}05$, em condições de violação da pressuposição, o teste tende a rejeitar H_0 numa porcentagem maior do que o esperado. Esse resultado também indica que o teste não foi capaz de controlar o erro tipo I quando a pressuposição não é satisfeita.

A escolha por limites menos conservadores se justifica por estabelecer um nível aceitavelmente próximo à taxa nominal, uma vez que o uso de um critério conservador, de $\alpha \pm 0{,}1\alpha$, produziriam intervalos de classificação rigorosos (BRADLEY, 1978).

A mensuração do poder empírico do teste foi feita por meio da computação do número de vezes em que a hipótese H_0 foi rejeitada, quando heterogeneidade de variâncias foram intencionalmente adicionadas às amostras, tornando H_0 realmente falsa. Com isso, nos cenários onde se partiu do pressuposto de que H_0 é tida como falsa, foi calculado a proporção de rejeição da hipótese nula falsa, dado o total de simulações realizadas com esta configuração. Assim, nos cenários simulados sob a condição H_0 falsa foi calculada a taxa empírica do poder do teste por

$$\hat{\pi} = \frac{\#RH_o|H_0\ falsa}{\#H_0\ falsa} \qquad (9)$$

Em que

$\hat{\pi}$: taxa empírica do poder do teste;

$\#RH_o|H_0\ falsa$: número de conjunto de dados dos cenários para os quais H_0 foi rejeitada dentre todos os conjuntos de dados dos cenários simulados sob a condição que H_0 é falsa;

$\#H_0\ falsa$: número total de conjuntos de dados dos cenários simulados sob a condição que H_0 é falsa. No presente trabalho, fez-se $\#H_0\ falsa$ igual a 1.000.

Nesse trabalho, um teste foi classificado como poderoso se a taxa empírica do poder do teste foi igual ou superior a 0,80, ou seja, se $\hat{\pi} \geq 0{,}80$ (COCHRAN, 1947).

A avaliação destas taxas empíricas (erro tipo I e poder) foram obtidas para cada um dos testes (paramétricos e não paramétricos) em cada um dos cenários simulados. A idealização desses cenários tomou como base estudos anteriores que também utilizaram de simulação. Contudo, conforme mencionado anteriormente, os cenários simulados foram idealizados para avaliar os testes de HOV sob diferentes situações. Para facilitar o entendimento de cada um dos cenários simulados é descrito a seguir a notação utilizada na representação dos cenários:

1. Número de populações: todos os cenários simulados consistiram em cinco populações, ou seja, $t = 5$;
2. Proporção de heterogeneidade: Determinou, a princípio, se as populações simuladas seriam heterogêneas ou não. As proporções de heterogeneidade $(\sigma_1^2 : \sigma_2^2 : \sigma_3^2 : \sigma_4^2 : \sigma_5^2)$ avaliadas foram (a_j): 1:1:1:1:1 (a_1); 1:1:1:1:2 (a_2); 1:1:1:1:4 (a_3); 1:1:1:2:2 (a_4); 1:1:1:4:4 (a_5); e, 1:2:3:4:5 (a_6). Essas proporções têm o objetivo de observar o comportamento de cada teste quando H_0 é tomada com verdadeira (configuração a_1) e quando H_0 é tomada como falsa, proporções: a_2, a_3, a_4, a_5 e a_6. O aumento gradativo da proporcionalidade entre as variâncias populacionais nos cenários a_2, a_3, a_4, a_5 e a_6 têm o objetivo de avaliar o desempenho de cada um dos testes quanto a taxa de rejeição da hipótese nula falsa à medida que um maior grau de heterogeneidade é acrescido entre as populações.
3. Proporção de desbalanceamento: Em cada cenário foi definido se as populações simuladas seriam balanceadas ou desbalanceadas. Entenda-se por balanceada como sendo o conjunto de populações em que para todas elas foram extraídas o mesmo tamanho amostral, ou seja, todas as amostras possuem o mesmo número de repetições, isto é, $r_i = k$ para todo $i = 1, \cdots, t$. Por outro lado, conjunto desbalanceado é aquele em que existe pelo menos um $r_i \neq r_m$ para $i = 1, \cdots, m, \cdots, t$. O número de observações $(r_1 : r_2 : r_3 : r_4 : r_5)$ para o conjunto de populações avaliadas foram (b_n): $4:4:4:4:4$ (b_1); $6:6:6:6:6$ (b_2); $8:8:8:8:8$ (b_3); $10:10:10:10:10$ (b_4); $15:15:15:15:15$ (b_5); $25:25:25:25:25$ (b_6); $45:45:45:45:45$ (b_7);

$90:90:90:90:90$ (b_8); $140:140:140:140:140$ (b_9); $200:200:200:200:200$ (b_{10}); $4:6:8:10:15$ (b_{11}); $15:25:45:90:140$ (b_{12}); $10:45:90:140:200$ (b_{13}). Estas configurações têm o objetivo de avaliar o desempenho de cada teste de HOV contemplando uma vasta variedade de tamanhos de amostras, tanto nos cenários balanceados quanto nos desbalanceados. O fato de avaliar número de repetições por tratamento maior do que o usual se deve ao objetivo de buscar o número de repetições ideal, caso o número de repetições usual não forneça resultados satisfatórios do ponto de vista estatístico.

4. Distribuição de probabilidades: para cada cenário foi definido se o conjunto de dados simulado apresentava distribuição de probabilidades Normal (n), Qui-quadrado (q) ou Beta (*b*). A Normal tem distribuição simétrica (com média igual a 10) e as distribuições Qui-quadrado e Beta tiveram seus parâmetros definidos para que tenham distribuições assimétrica positiva (com graus de liberdade igual a 2 e parâmetro de não centralidade igual a zero) e negativa (com parâmetro alfa igual a 8 e beta igual a um), respectivamente. As distribuições adotadas para as respectivas populações foram (f_w): $n:n:n:n:n$ (f_1); $q:q:q:q:q$ (f_2); $b:b:b:b:b$ (f_3).

Dado os critérios de avaliação descritos acima, o *software R*, versão 4.0.3, foi utilizado para a condução de todo o processo de simulação. O algoritmo utilizado na geração dos cenários para a avaliação da taxa empírica do erro tipo I ($\hat{\alpha}$) foi:

1. Definir a proporção de heterogeneidade a_j para $j = 1$;
2. Definir a proporção de desbalanceamento b_n para $n = 1, \cdots, 13$;
3. Definir a distribuição de probabilidade f_w para $w = 1, 2, 3$;
4. Definir o cenário c_{jnw}. Gerar v-ésimo conjunto deste cenário identificando-o como c_{jnwv}.
5. Repetir os passos de 1 a 4 para $v = 1, \cdots, 1.000$.

O algoritmo utilizado na geração dos cenários para avaliar a taxa empírica do poder do teste ($\hat{\pi}$) foi:

1. Definir a proporção de heterogeneidade a_j para $j = 2, \cdots, 6$;
2. Definir a proporção de desbalanceamento b_n para $n = 1, \cdots, 13$;
3. Definir a distribuição de probabilidade f_w para $w = 1$;

4. Definir o cenário c_{jnw}. Gerar v-ésimo conjunto deste cenário identificando-o como c_{jnwv}.
5. Repetir os passos de 1 a 4 para $v = 1, \cdots, 1.000$;

5. Resultados

Neste trabalho, a comparação da performance dos testes de HOV foi realizada com base na taxa empírica do erro tipo I $(\hat{\alpha})$ e na taxa empírica do poder $(\hat{\pi})$ dos testes paramétricos e não paramétricos, taxas essas obtidas a partir de 104 cenários de base de dados simuladas. Para facilitar a avaliação da performance, a taxa empírica de erro $\hat{\alpha}$ obtida para cada teste em cada um dos cenários simulados sob hipótese de nulidade verdadeira, foi comparada com o nível de significância previamente estabelecido pela adaptação do Critério Liberal de Bradley (1978) resultando em três classes: conservador $(\hat{\alpha} \leq 0{,}0375)$, exato $(0{,}0375 < \hat{\alpha} < 0{,}0625)$ e liberal $(\hat{\alpha} \geq 0{,}0625)$. A taxa empírica do poder do teste, obtida para cada um dos cenários simulados sob a hipótese de nulidade falsa, foi classificado segundo o critério proposto por Cochran (1947) o qual estabelece que um teste é poderoso se $\hat{\pi} \geq 0{,}80$ e não poderoso se $\hat{\pi} < 0{,}80$.

5.1. Taxa empírica do erro tipo I

A taxa empírica do erro tipo I foi quantificada para os cenários simulados sob a proporção de heterogeneidade 1:1:1:1:1 (a_1), de acordo com as b_n proporções de desbalanceamento e as distribuições de probabilidades, f_w. Isto é, para a obtenção da taxa empírica $\hat{\alpha}$, o estudo de simulação foi realizado considerando a hipótese H_0 verdadeira, sob diferentes níveis de desbalanceamento e para as distribuições normal, qui-quadrado e beta.

5.1.1. Distribuição normal

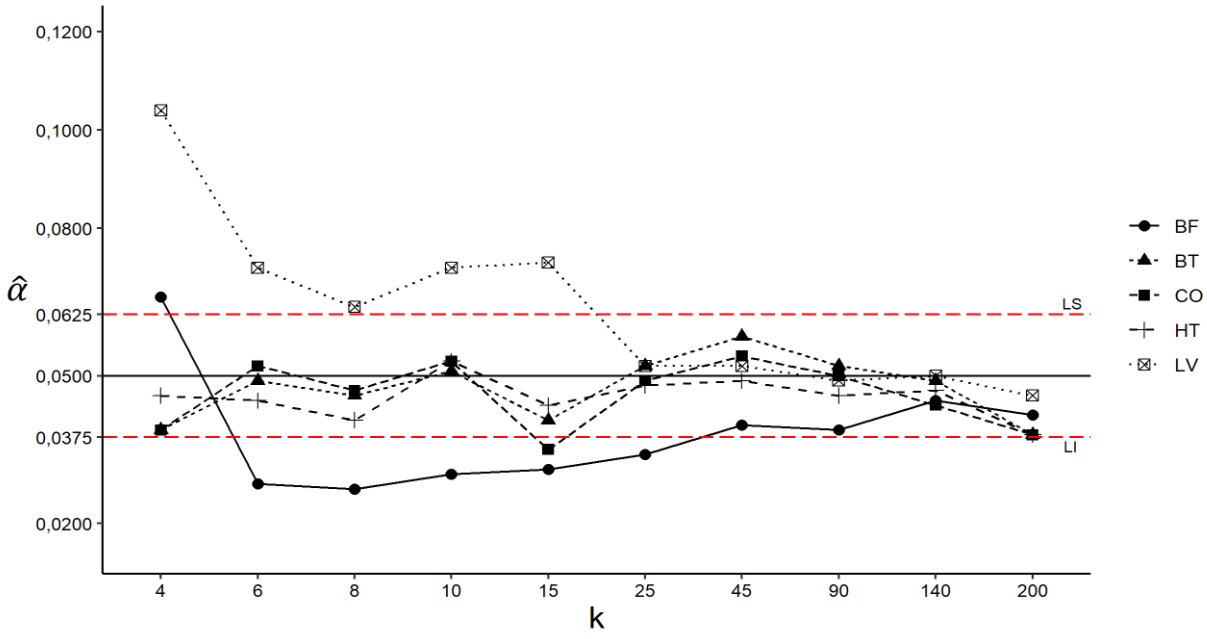
5.1.1.1. Cenários balanceados

Neste primeiro momento são abordados os resultados referentes aos cenários em que os dados amostrais são normalmente distribuídos e a hipótese de homogeneidade de variâncias é verdadeira, alterando apenas o número de repetições por tratamento durante as simulações. Deste modo, as simulações não violam as pressuposições dos testes.

A taxa empírica do erro tipo I $(\hat{\alpha})$ em função do número de repetições por tratamento (k) para cada um dos testes paramétricos é apresentada na Figura 1. Podemos observar que para $k = 4$, os testes de Brown-Forsythe e de Leneve foram classificados como liberais, ao passo que os testes de Bartlett, de Cochran e de Hartley foram categorizados como testes exatos. Quando amostras de tamanhos $k =$

$6, 8$ e 10 são empregadas, o teste de Levene tende a permanecer com classificação liberal, o teste de Brown-Forsythe passa a ter um comportamento conservador e, os testes de Bartlett, Cochran e Hartley, foram classificados como exatos. Para $k = 15$ apenas os testes de Bartlett e de Hartley são considerados exatos, ao passo que os testes de Cochran e de Brown-Forsythe conservou H_0 verdadeira numa proporção menor do que o esperado e o teste de Levene foi liberal. Por outro lado, quando $k \geq 25$, os testes paramétricos foram categorizados como exatos e, à medida que o número de repetições por tratamento aumenta, em geral, os testes tendem a apresentar uma diminuição da taxa empírica do erro tipo I. Exceção à essa generalidade é o teste de Brown-Forsythe que é classificado como conservador para $k = 25$ e exato quando $k \geq 45$, com a taxa empírica do erro tipo I aproximando-se de $\alpha = 0{,}05$ à medida que se aumenta o valor de k. Contudo, para $k = 200$ todos os testes são classificados como exatos.

Figura 1 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob distribuição normal, em função do número de repetições (k) por tratamento para os testes paramétricos. Onde: BF, é o teste de Brown-Fortsyhe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, é o teste de Levene; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).

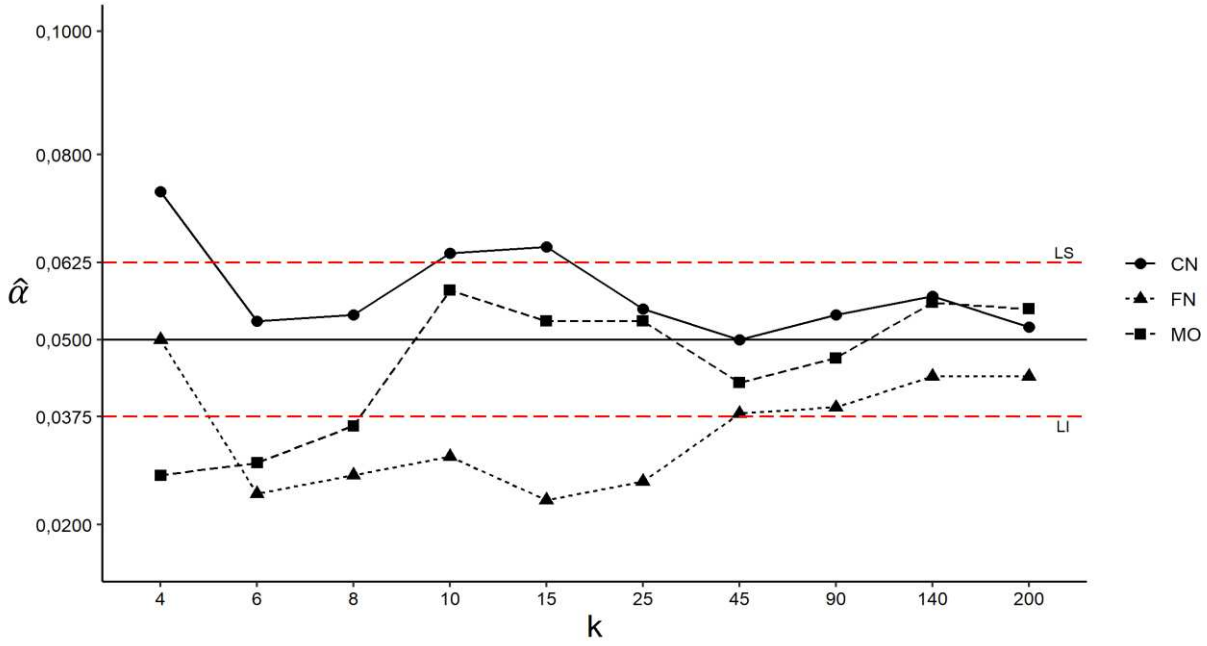


Considerando os testes que utilizam transformações dos valores originais observados sobre a análise de variância de um fator, como é o caso dos testes de Levene e de Brown-Forsythe, observa-se que, de modo geral, o teste de Brown-Forsythe tende a obter menores taxas empíricas de erro do tipo I. Por outro lado, ao se comparar o teste de Cochran com o teste de Hartley, que apresentam estatísticas

de teste semelhantes, observou-se que ambos tiveram comportamentos parecidos em relação a taxa empírica $\hat{\alpha}$, exceto, quando o cenário avaliado continha 15 repetições por tratamento, no qual o teste de Cochran foi conservador e o teste de Hartley exato.

Na Figura 2 é apresentada a taxa empírica do erro tipo I em função do número de repetições por tratamento para cada um dos testes não paramétricos. Nessa imagem é possível observar que cada teste teve comportamento distinto para $k \leq 25$, assim, cada teste teve a taxa empírica oscilando de modo diferente. No entanto, quando $k = 4$, os testes não paramétricos apresentaram diferentes classificações: Conover, foi liberal; Mood, conservador; e Fligner-Killeen, exato (com $\hat{\alpha} = 0{,}05$). Quando $k = 6$ e 8, tanto o teste de Fligner-Killeen quanto o teste de Mood foram conservadores, ao passo que o teste de Conover foi exato. Ao utilizar 10 e 15 repetições por tratamento os testes novamente atuaram de maneira distinta quanto as classificações da taxa empírica do erro tipo I, onde: o teste de Conover foi liberal; Fligner-Killeen, foi conservador; e, o teste de Mood passou a ser um teste exato. Ao considerar o cenário simulado com 25 repetições por tratamento, os testes de Conover e de Mood foram classificados como exatos, enquanto, o teste de Fligner-Killeen continuou sendo conservador. Além do mais, quando $k \geq 45$ observações por tratamento os três testes controlam a taxa empírica do erro tipo I e foram classificados como exatos.

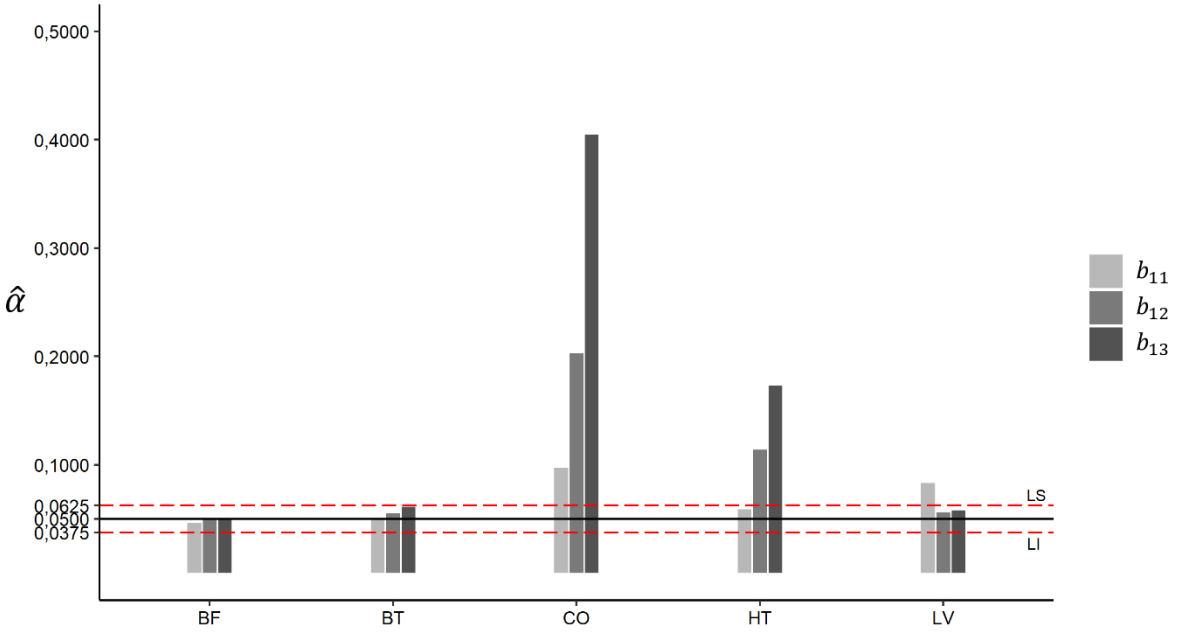
Figura 2 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob distribuição normal, em função do número de repetições (k) por tratamento para os testes não paramétricos. Onde: CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



5.1.1.2. Cenários desbalanceados

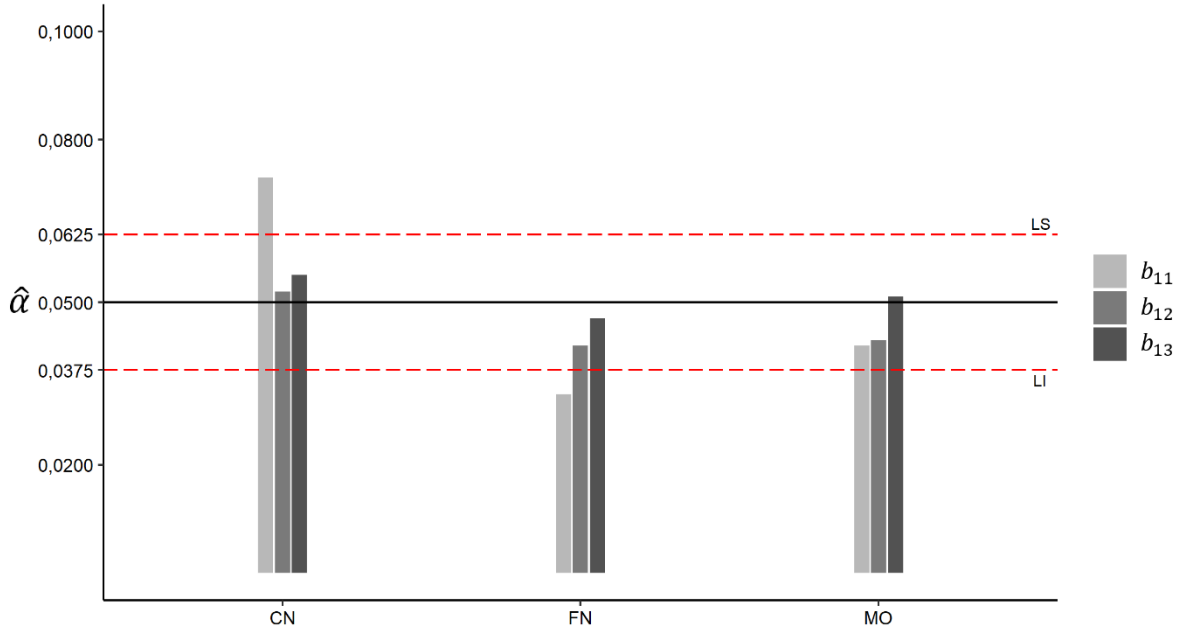
Na Figura 3 são apresentadas as taxas empíricas do erro tipo I referentes aos testes paramétricos, para os cenários simulados supondo distribuição normal e as proporções de desbalanceamento: $4:6:8:10:15$ (b_{11}), $15:25:45:90:140$ (b_{12}) e $10:45:90:140:200$ (b_{13}). Por meio desses resultados, nota-se que mesmo em cenários desbalanceados os testes de Bartlett e de Brown-Forsythe foram classificados como exatos em todas as proporções de desbalanceamento. O teste de Levene teve classificação liberal quando avaliado sob a proporção b_{11} e, nas proporções b_{12} e b_{13}, controlou a taxa empírica de erro $\hat{\alpha}$ e foi um teste exato. Quando as proporções de desbalanceamento foram analisadas sob a vista do teste de Cochran e do teste de Hartley, ambos apresentaram um aumento da taxa empírica do erro tipo I à medida em que o desbalanceamento entre tratamento foi maior. Contudo, o teste de Cochran foi liberal nos cenários b_{11}, b_{12} e b_{13}, enquanto o teste de Hartley foi exato na proporção de desbalanceamento b_{11}. Esse aumento da taxa empírica pode ser devido ao teste de Cochran utilizar a média aritmética (ou, média harmônica) do número de repetições por tratamento para o cálculo do número de graus de liberdade e, o teste de Hartley, utilizar a amostra populacional que contém o maior número de repetições.

Figura 3 - Taxa empírica do erro tipo I ($\hat{\alpha}$) , sob distribuição normal, para os testes paramétricos nos cenários $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$. Onde: BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, é o teste de Levene; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



Dentre os testes não paramétricos apresentados na Figura 4, observa-se diferentes comportamentos para a taxa empírica $\hat{\alpha}$ quando a proporção de desbalanceamento b_{11} é levada em consideração. Nesse cenário, o teste de Conover foi classificado como liberal e o teste de Fligner-killeen como conservador, ao passo que o teste de Mood foi considerado um teste exato. No entanto, quando a proporção de desbalanceamento b_{12} e b_{13} foram analisadas, os três testes não paramétricos alcançaram exatidão segundo o critério utilizado para a classificação.

Figura 4 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob distribuição normal, para os testes não-paramétricos nos cenários $b_{11}(4{:}\,6{:}\,8{:}\,10{:}\,15\,)$, $b_{12}\,(15{:}\,25{:}\,45{:}\,90{:}\,140)$ e $b_{13}\,(10{:}\,45{:}\,90{:}\,140{:}\,200)$. Onde: CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



5.1.2. Distribuição de qui-quadrado

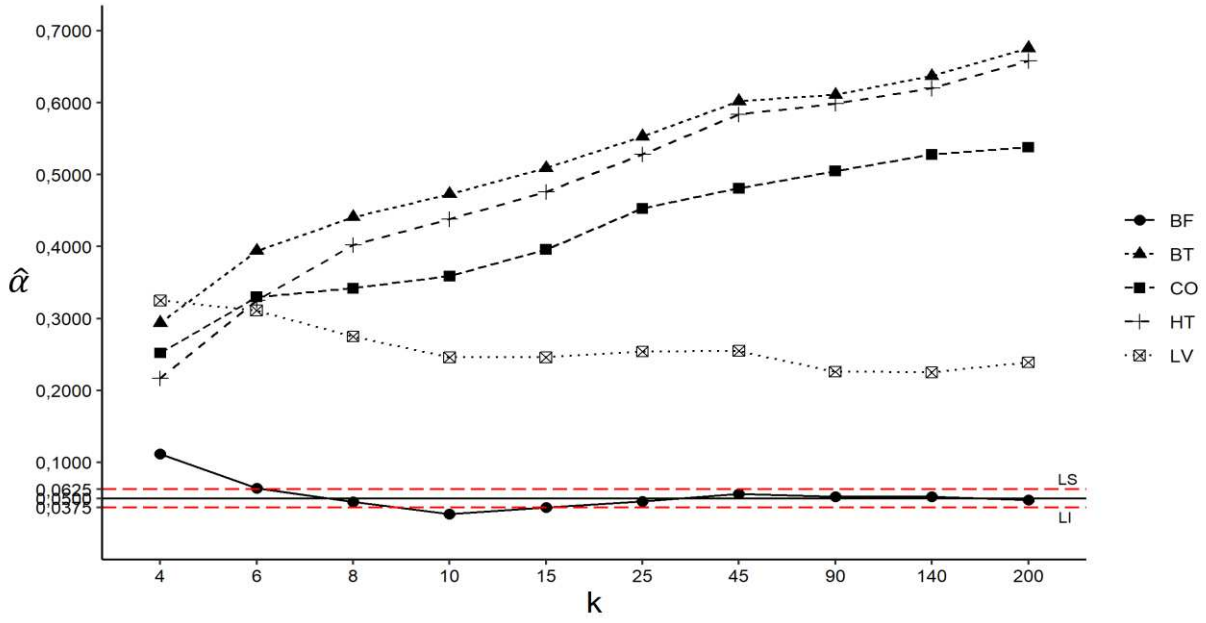
5.1.2.1. Cenários balanceados

Nesta seção são apresentados os resultados referentes aos cenários simulados sob a distribuição de qui-quadrado quando a hipótese nula de igualdade de variâncias é tomada como verdadeira. Desta forma, os cenários aqui gerados não satisfazem a pressuposição de normalidade.

Daremos início abordando os resultados relativos aos testes paramétricos nos cenários gerados sob a condição de que os tratamentos têm o mesmo número de repetições. A taxa empírica para o erro do tipo I ($\hat{\alpha}$) em função do tamanho da amostra (k) é ilustrada na Figura 5. Com base no estudo de simulação pode-se observar que, quando $k = 4$ e 6, todos os testes paramétricos foram classificados como liberais. No entanto, ao se utilizar amostras de tamanhos maiores, o teste de Brown-Forsythe foi classificado como exato, exceto para $k = 10$ e 15, em que o teste foi conservador. Os demais testes paramétricos foram rotulados como liberais, em todos os k tamanhos amostrais, demonstrando tendência a obter taxas empíricas de erro tipo I mais elevadas e tornando-se ainda maiores à medida que o número de repetições por tratamento aumentava, como é o caso dos testes de Bartlett, de Cochran e de Hartley.

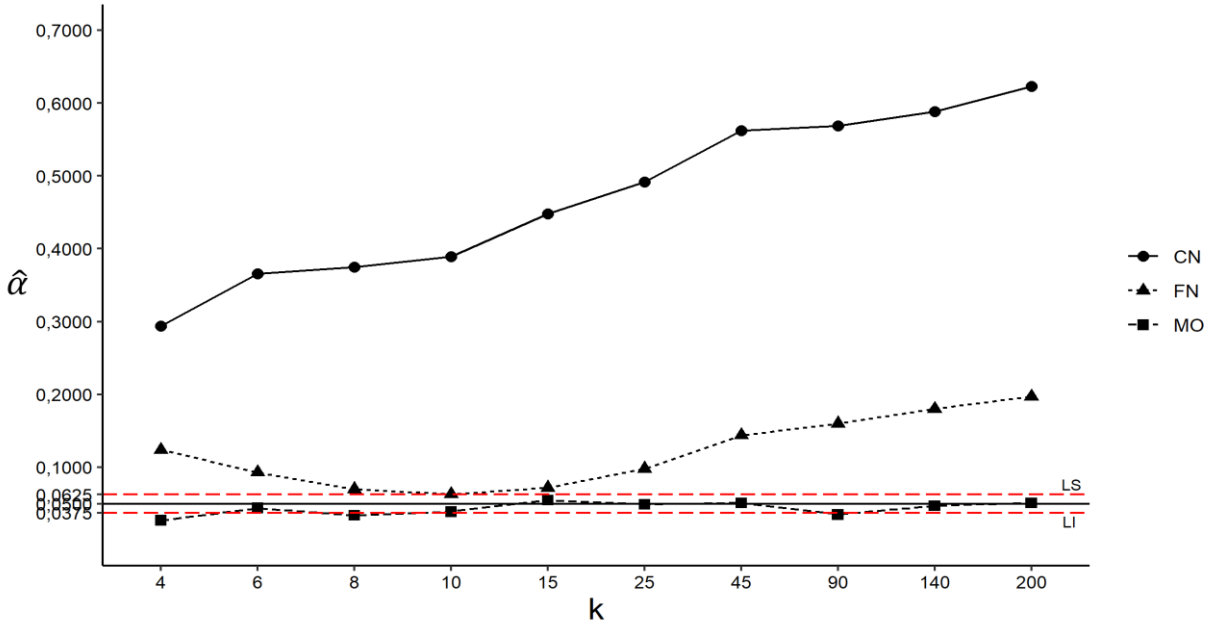
Dentre eles, o teste de Bartlett mostrou uma tendência para maior sensibilidade a ausência de normalidade. Essa característica é devido ao teste de Bartlett ser baseado em uma modificação da estatística do teste da razão de verossimilhança que é derivada sob o pressuposto de normalidade (CONOVER; JOHNSON; JOHNSON, 1981). O teste de Levene também atuou de modo liberal em todos os tamanhos de amostras avaliados, porém, menos sensível à desvios de normalidade quando comparados aos testes de Bartlett, Hartley e Cochran.

Figura 5 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob distribuição de qui-quadrado, em função do número de repetições (k) por tratamento nos testes paramétricos, onde: BF, é o teste de Brown-Fortsyhe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, é o teste de Levene; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



A taxa empírica, $\hat{\alpha}$, em função do número de repetições por tratamento (k) para cada um dos testes não paramétricos é ilustrada na Figura 6. Seguindo a adaptação do critério liberal de Bradley, nota-se que o teste de Mood foi exato na maior parte dos tamanhos de amostras quando a pressuposição de normalidade não foi satisfeita. Exceção a esse comportamento foi observado para $k = 4, 8$ e 90 repetições e o teste foi considerado como conservador diante desses cenários. Os testes de Conover e Fligner-Killeen foram considerados liberais para todos os k número de repetições e apresentaram tendência de crescimento da taxa empírica do erro tipo I à medida que o número de observações por tratamento se tornou maior. Ainda, o teste de Conover se mostrou mais sensível à desvios da normalidade e obteve taxa empírica variando entre $0{,}294$, para $k = 4$, até $0{,}623$ para $k = 200$.

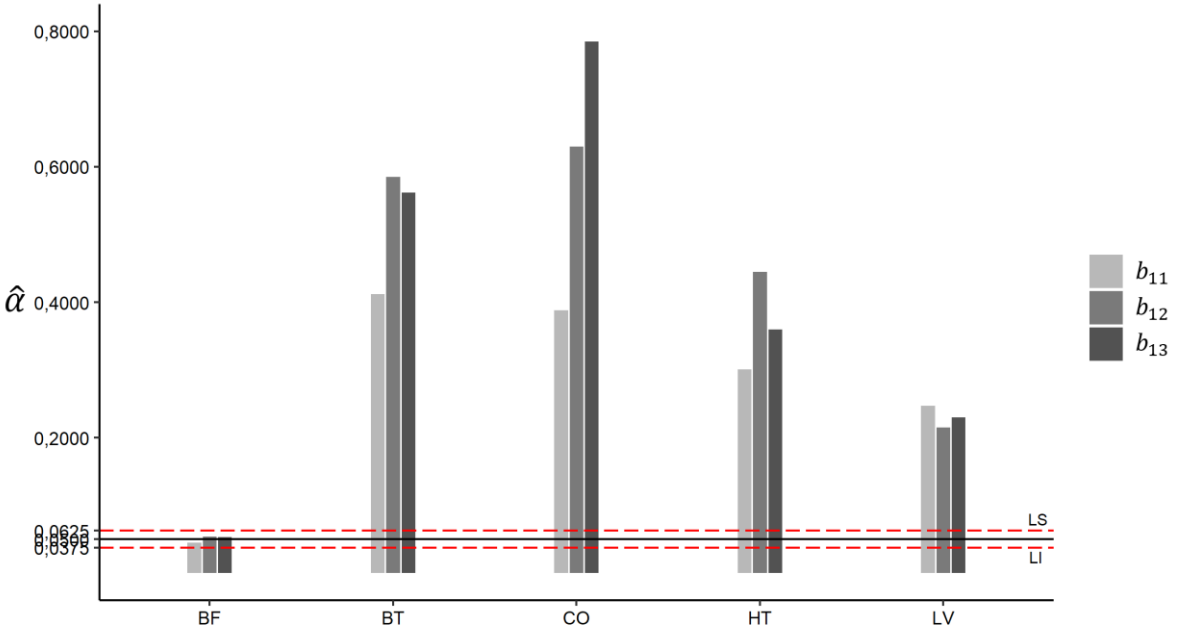
Figura 6 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob distribuição de qui-quadrado, em função do número de repetições (k) por tratamento nos testes não paramétricos, onde: CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; LS, é o limite superior adaptado de Bradley; e, LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



5.1.2.2. Cenários desbalanceados

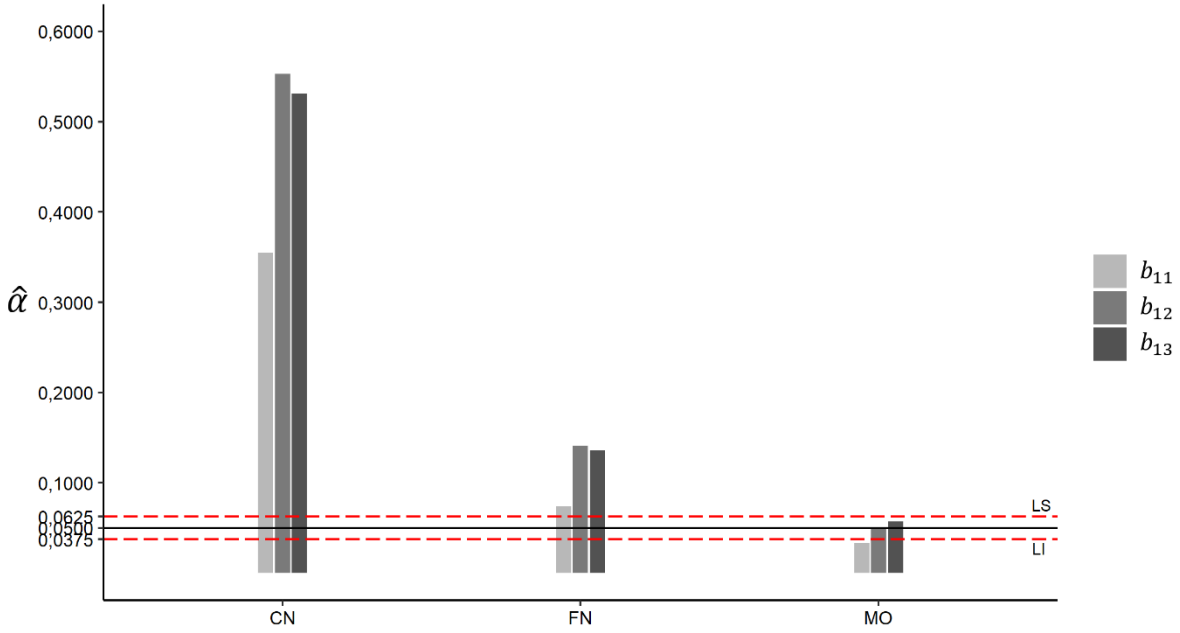
Os resultados para a taxa empírica do erro tipo I ($\hat{\alpha}$) obtidos por meio do estudo de simulação, sob a distribuição de qui-quadrado, para cada um dos testes paramétricos em cenários desbalanceados é exibida na Figura 7. Com base nesses resultados, pode ser observado que o uso da mediana na transformação dos valores observados tornou o teste de Brown-Forsythe robusto a violação da pressuposição de normalidade e, portanto, capaz de controlar a taxa empírica $\hat{\alpha}$ mesmo quando o número de repetições por tratamento era diferente. Esse fato pode ser observado para o teste Brown-Forsythe quando as proporções de desbalanceamento $4:6:8:10:15$ (b_{11}), $15:25:45:90:140$ (b_{12}) e $10:45:90:140:200$ (b_{13}) são analisadas. Quando avaliado os testes que utilizam, respectivamente, a média do número de repetições por tratamento e o maior número de observações entre as populações para a obtenção do número de graus de liberdade, como é o caso do teste de Cochran e de Hartley, foi notado que ambos os testes foram liberais e apresentaram tendência de aumento das taxas empíricas do erro $\hat{\alpha}$ para as proporções de desbalanceamento mais discrepantes (b_{12} e b_{13}). Altas taxas empíricas de erro do tipo I também foram observadas para o teste de Bartlett, assim, sendo considerado liberal em todos os cenários desbalanceados.

Figura 7 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob a distribuição de qui-quadrado, para os testes paramétricos nos cenários $b_{11}(4:6:8:10:15)$, $b_{12}(15:25:45:90:140)$ e $b_{13}(10:45:90:140:200)$. Onde: BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, é o teste de Levene; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



Com base nos resultados obtidos para os testes não paramétricos, apresentados na Figura 8, podemos observar que quando a avaliação foi realizada sob a proporção de desbalanceamento b_{11} nenhum dos testes foi exato. Nesse cenário, os testes de Conover e de Fligner-killeen foram liberais, enquanto, o teste de Mood foi conservador. Quando se aumentou a discrepância entre o número de repetições por tratamento, como para as proporções b_{12} e b_{13}, o teste de Mood foi robusto à não normalidade e ao desbalanceamento das amostras, sendo classificado como exato.

Figura 8 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob a distribuição de qui-quadrado, para os testes não paramétricos nos cenários $b_{11}(4:6:8:10:15)$, $b_{12}(15:25:45:90:140)$ e $b_{13}(10:45:90:140:200)$. Onde: CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



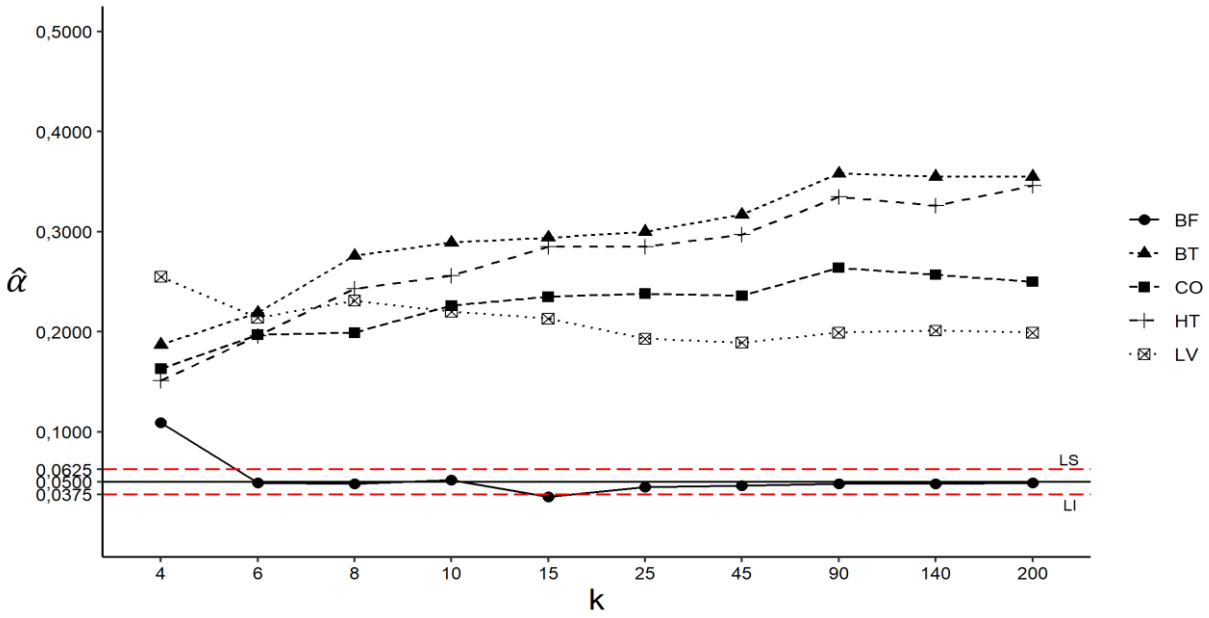
5.1.3. Distribuição beta

5.1.3.1. Cenários balanceados

No presente tópico são apresentados os resultados referentes aos cenários simulados supondo distribuição beta e homogeneidade de variâncias. Deste modo, os cenários aqui avaliados não seguem a distribuição normal que é requerida para alguns testes.

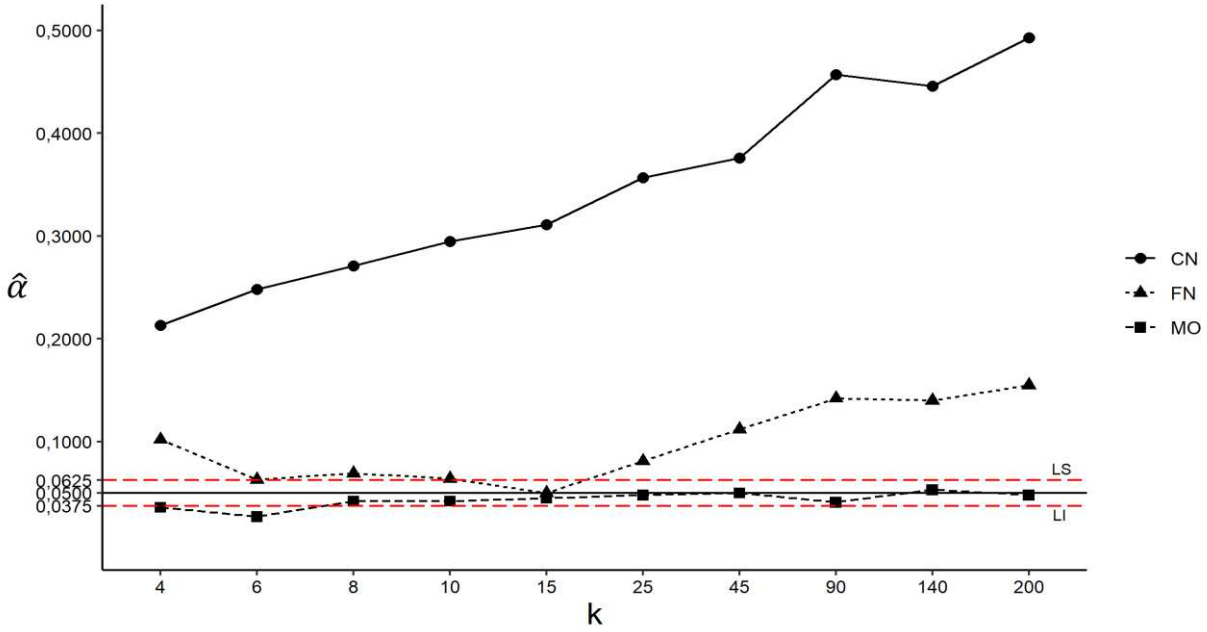
Nestes cenários, os testes de homogeneidade de variâncias tenderam a apresentar taxas empíricas de erro do tipo I relativamente menores e com curvas menos inclinadas quando comparados aos resultados obtidos para cenários simulados sob a distribuição de qui-quadrado. Esse fato pode ser visto na Figura 9 para os testes paramétricos. Quando $k = 4$, todos os cinco testes paramétricos foram considerados liberais. No entanto, para $k \geq 6$, o teste de Brown-Forsythe tende a atuar de forma exata, exceto, quando 15 repetições por tratamento foram usadas e, nesse caso, o teste foi conservador. Os demais testes foram liberais em todos os k tamanhos de amostras avaliados e, o teste de Bartlett, tendeu a apresentar as maiores taxas empíricas para o erro tipo I.

Figura 9 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob a distribuição beta, em função do número de repetições (k) por tratamento para os testes paramétricos, onde: BF, é o teste de Brown-Fortsyhe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, é o teste de Levene; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



Dentre os testes não paramétricos apresentados na Figura 10, é possível observar que quando $k = 4$ e 6 nenhum teste foi exato, de modo, que, os testes de Conover e Fligner-Killeen foram classificados como liberais e, o teste de Mood, foi conservador. No entanto, para $k \geq 8$, o teste de Mood foi exato e, então, robusto a não normalidade. Todavia, nestes cenários, o teste de Fligner-Killeen alcançou taxa empírica $\hat{\alpha} = 0{,}05$ quando $k = 15$ e atuou de forma liberal nos demais números de repetições por tratamento. Contudo, o teste Conover foi liberal para todos os números de observações simulados, com tendência de crescimento da taxa empírica $\hat{\alpha}$ para crescente número de repetições por tratamento.

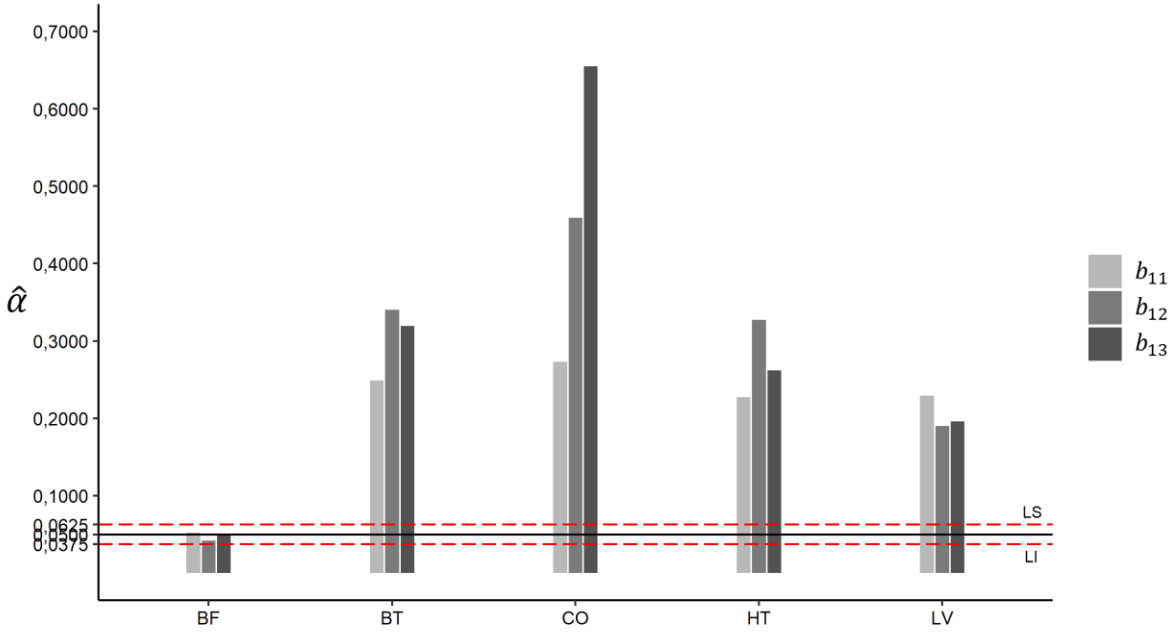
Figura 10 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob a distribuição beta, em função do número de repetições (k) por tratamento para os testes não paramétricos, onde: CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



5.1.3.2 Cenários desbalanceados

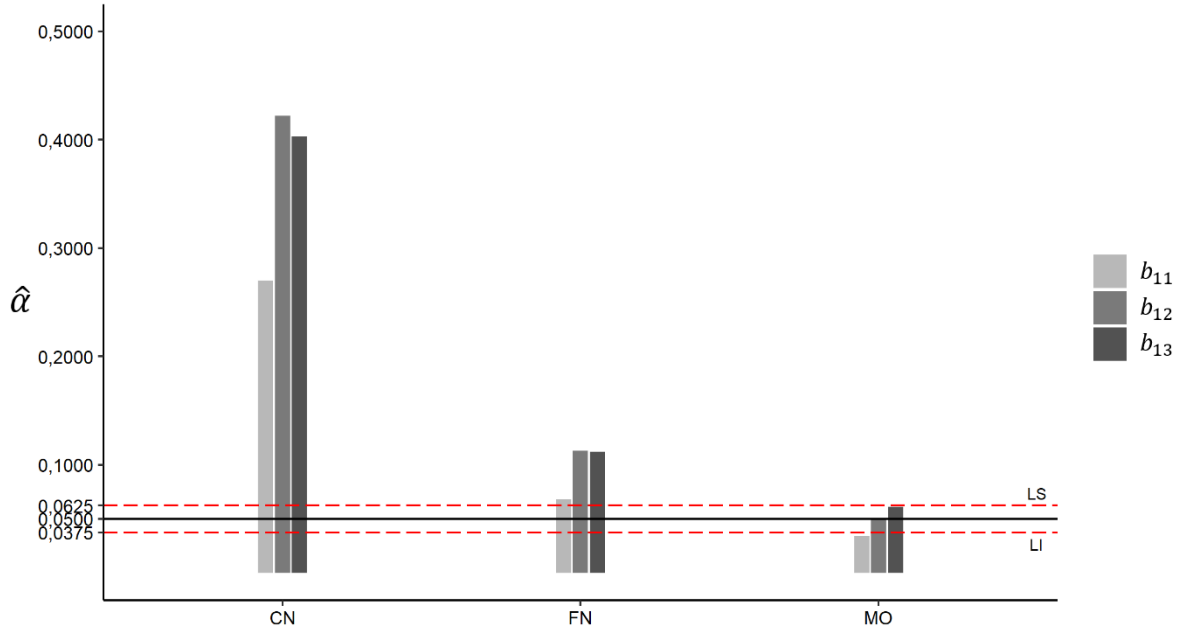
São apresentados na Figura 11 as taxas empíricas do erro tipo I para os cenários desbalanceados gerados sob a distribuição beta. Tal como observado para os cenários simulados sob a distribuição de qui-quadrado, o teste paramétrico de Brown-Forsythe foi o único considerado exato nos níveis de desbalanceamentos: $4:6:8:10:15$ (b_{11}), $15:25:45:90:140$ (b_{12}) e $10:45:90:140:200$ (b_{13}). Os testes de Bartlett, Cochran, Hartley e Levene não conseguiram controlar o erro tipo I dentro dos limites de classificação e foram liberais. O teste de Cochran mais uma vez apresentou tendência de crescimento da taxa empírica $\hat{\alpha}$ à medida que a proporção de desbalanceamento se tornava maior. Padrão similar de tendência pode ser observado para os testes de Bartlett e de Hartley. Dentre os testes liberais, o teste de Levene, que como o teste Brown-Forsythe faz uso da transformação dos valores observados sobre o teste F da ANOVA, alcançou a segunda menor taxa de erro $\hat{\alpha}$ nos níveis de desbalanceamento avaliados.

Figura 11 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob a distribuição beta, para os testes paramétricos nos cenários $b_{11}(4:6:8:10:15)$, $b_{12}(15:25:45:90:140)$ e $b_{13}(10:45:90:140:200)$. Onde: BF, o teste de Brown-Forsythe; BT, éo teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



De acordo com os resultados para os testes não paramétricos apresentados na Figura 12, observa-se características semelhantes aos resultados gerados para os cenários simulados sob as mesmas configurações para a distribuição de qui-quadrado. Pode ser observado que os testes de Conover e de Fligner-Killeen foram liberais ao serem avaliados nos três níveis de desbalanceamentos incluídos no estudo de simulação, isto é, para as proporções b_{11}, b_{12} e b_{13}. Nota-se também, que os testes Conover, Fligner-Killen e Mood sofrem aumento da taxa empírica de erro do tipo I quando os desbalanceamentos b_{12} e b_{13} são analisados. No entanto, o teste de Mood foi exato nessas duas proporções e, classificado como conversador, ao ser avaliado na proporção b_{11}.

Figura 12 - Taxa empírica do erro tipo I ($\hat{\alpha}$), sob a distribuição beta, para os testes não paramétricos nos cenários $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$. Onde: CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; LS e LI representam, respectivamente, os limites superior e inferior adaptados de Bradley (1978).



5.2. Taxa empírica do poder do teste

A taxa empírica do poder do teste ($\hat{\pi}$) foi obtida para os diferentes cenários a partir da frequência de rejeição da hipótese nula de homogeneidade de variâncias quando essa hipótese é realmente falsa na geração dos cenários simulados, isto é, heterogeneidade foram intencionalmente adicionadas aos conjuntos de dados. Um teste foi classificado como poderoso quando a taxa empírica do poder do teste foi igual ou superior a 0,80, ou seja, quando $\hat{\pi} \geq 0{,}80$.

Para melhor visualização e compreensão, iremos dividir os resultados em seções de acordo com as proporções de heterogeneidade e de desbalanceamento, conforme é disposto a seguir.

5.2.1. Proporção de heterogeneidade a_2

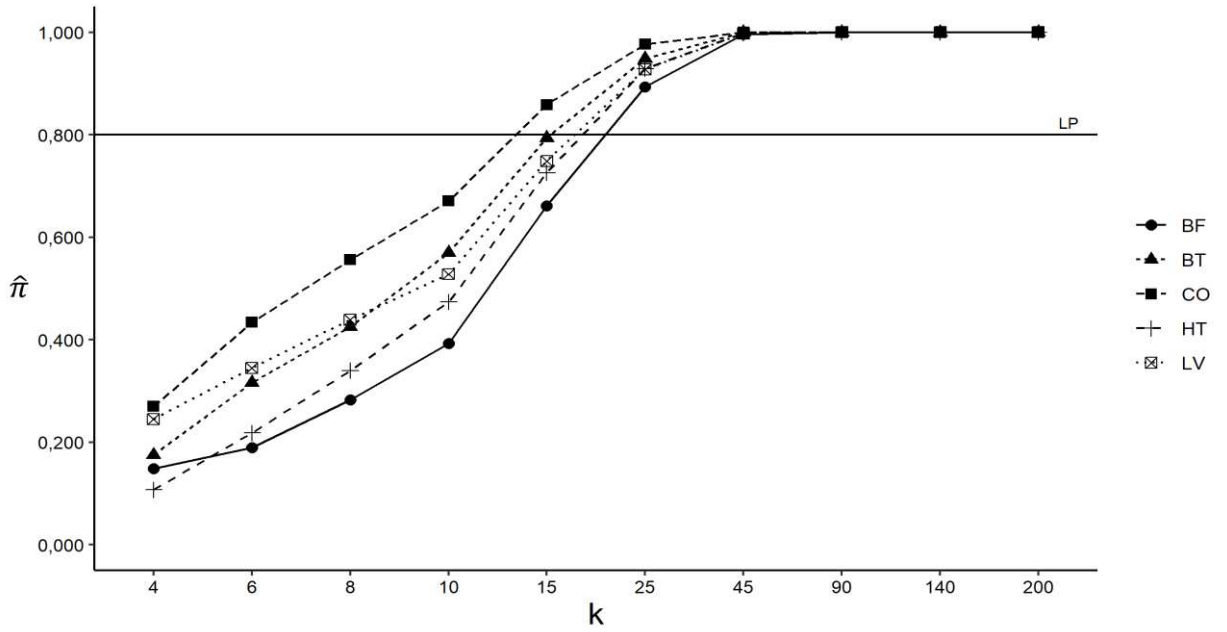
5.2.1.1. Cenários balanceados

São apresentados nesta seção os resultados relativos aos cenários cujos conjuntos de dados simulados são balanceados, normalmente distribuídos e a hipótese de homogeneidade de variâncias é tomada como falsa (ou seja, populações

são geradas sob a condição H_0 falsa), isto é, os conjuntos de dados foram simulados sob a proporção de heterogeneidade 1:1:1:1:2 (a_2).

A taxa empírica do poder, $\hat{\pi}$, em função do número de repetições por tratamento (k) para cada um dos testes paramétricos é apresentado na Figura 13. Podemos observar um crescente aumento da taxa empírica de poder do teste à medida que o número de repetições por tratamento foram se tornando maiores. Contudo, nenhum teste foi poderoso nos cenários cujo número de repetições por tratamento era igual ou inferior a 10 repetições. No entanto, ao se utilizar $k = 15$ repetições o teste de Cochran foi poderoso e, o teste de Bartlett, ficou próximo ao limiar da taxa empírica do poder do teste, com $\hat{\pi} = 0{,}794$. Quando $k = 25$, todos os testes paramétricos foram classificados como poderosos e, alcançaram poder máximo, com $k = 90$ repetições.

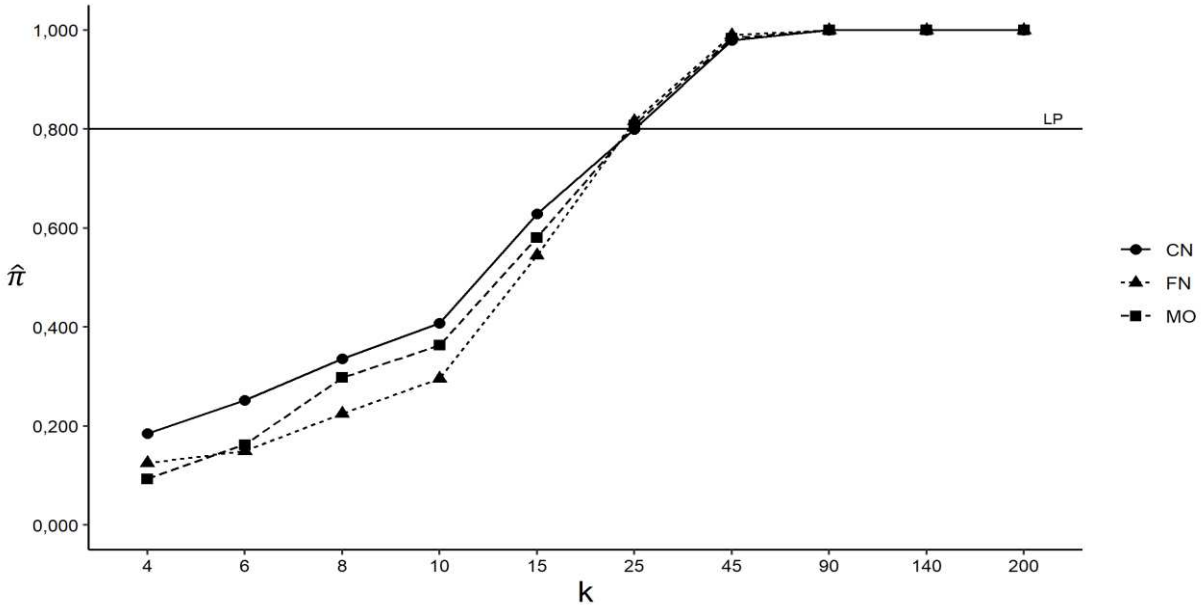
Figura 13 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal e proporção de heterogeneidade 1: 1: 1: 1:2 (a_2). Onde: BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene;e, LP é o limiar da taxa empírica do poder do teste.



Ao se observar as taxas empíricas de poder dos testes não paramétricos apresentados (Figura 14) percebe-se tendência semelhante àquela apresentada pelos testes paramétricos, isto é, os testes obtêm baixo poder quando um menor número de repetições é utilizado e essa taxa tende a crescer à medida que o número de repetições por tratamento é maior. Quando usadas amostras de tamanho $k \leq 15$ repetições, nenhum teste foi considerado poderoso de acordo com o critério de classificação do poder. No entanto, quando $k = 25$, o teste de Fligner-Killeen e o teste

de Mood são qualificados como poderosos, enquanto, o teste de Conover obtém taxa empírica do poder próximo ao limiar, com $\hat{\pi} = 0{,}799$. Utilizando 45 repetições por tratamento todos os testes não paramétricos são poderosos e com $k = 90$ observações todos eles possuem poder máximo.

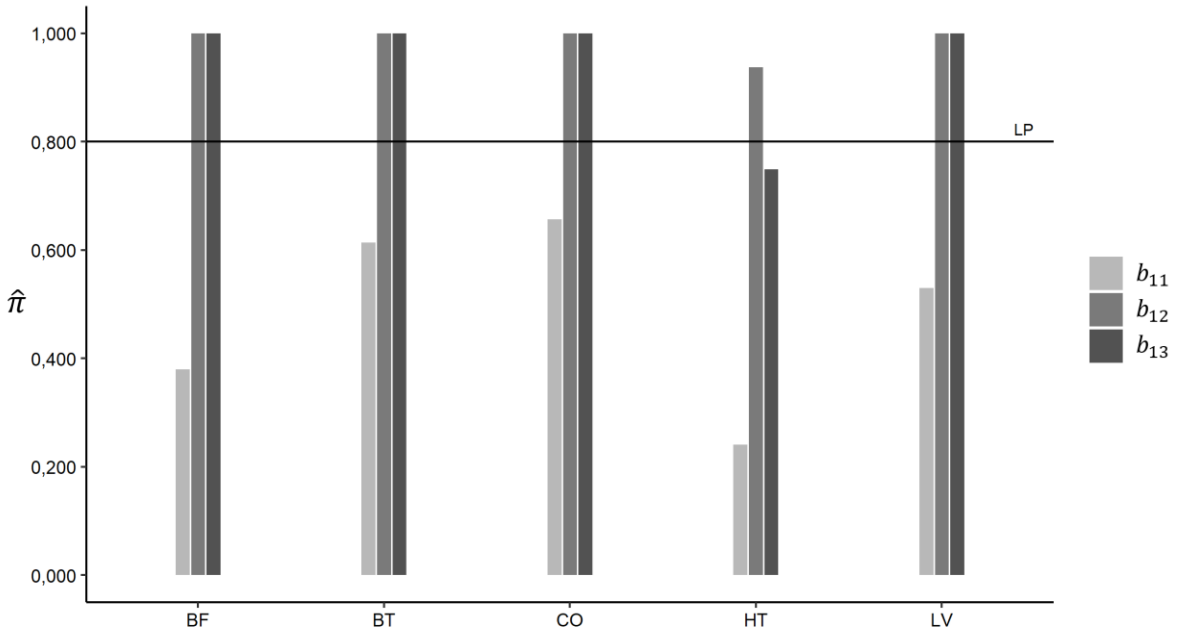
Figura 14 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não-paramétricos nos cenários simulados sob distribuição normal e proporção de heterogeneidade 1: 1: 1: 1: 2 (a_2). Onde: CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e, LP é o limiar da taxa empírica do poder do teste.



5.2.1.2. Cenários desbalanceados

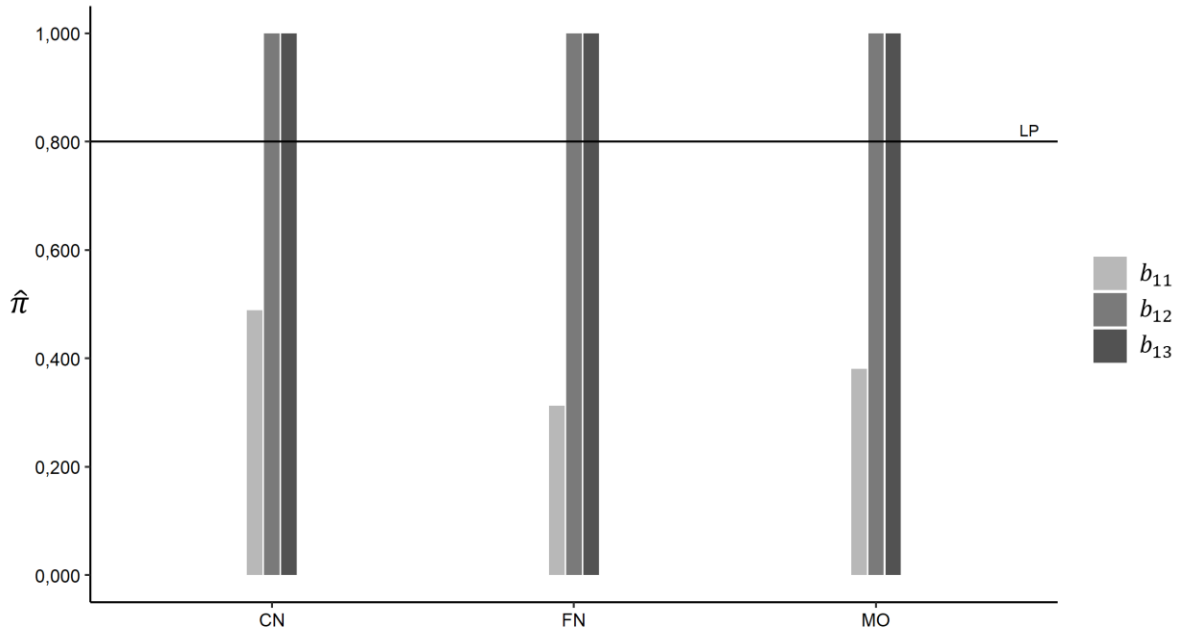
As taxas empíricas do poder para os testes paramétricos referentes aos cenários simulados sob as proporções de desbalanceamento $4:6:8:10:15$ (b_{11}), $15:25:45:90:140$ (b_{12}) e $10:45:90:140:200$ (b_{13}), para os conjuntos de dados normalmente distribuídos, de acordo com a proporção de heterogeneidade a_2, são apresentadas na Figura 15. Para a proporção de desbalanceamento b_{11} nenhum dos testes foi classificado como poderoso. No entanto, ao se considerar o desbalanceamento b_{12} todos os testes foram poderosos, ao passo que na proporção de desbalanceamento b_{13} apenas o teste de Hartley não foi poderoso.

Figura 15 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 1: 2 (a_2), nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde: BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e, LP é o limiar da taxa empírica do poder do teste.



Quando avaliado os testes não paramétricos (Figura 16) sob a proporção de desbalanceamento b_{11} foi verificado que nenhum destes foi poderoso. Esse resultado pode ser devido ao fato de o poder ser afetado pelo tamanho da amostra. Vale ressaltar que para $k \leq 15$ repetições por tratamentos (Figura 14) nenhum teste não paramétrico foi classificado como poderoso. Contudo, nas proporções de desbalanceamento b_{12} e b_{13} os três testes foram poderosos.

Figura 16 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 1: 2 (a_2), nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e, LP é o limiar da taxa empírica do poder do teste.



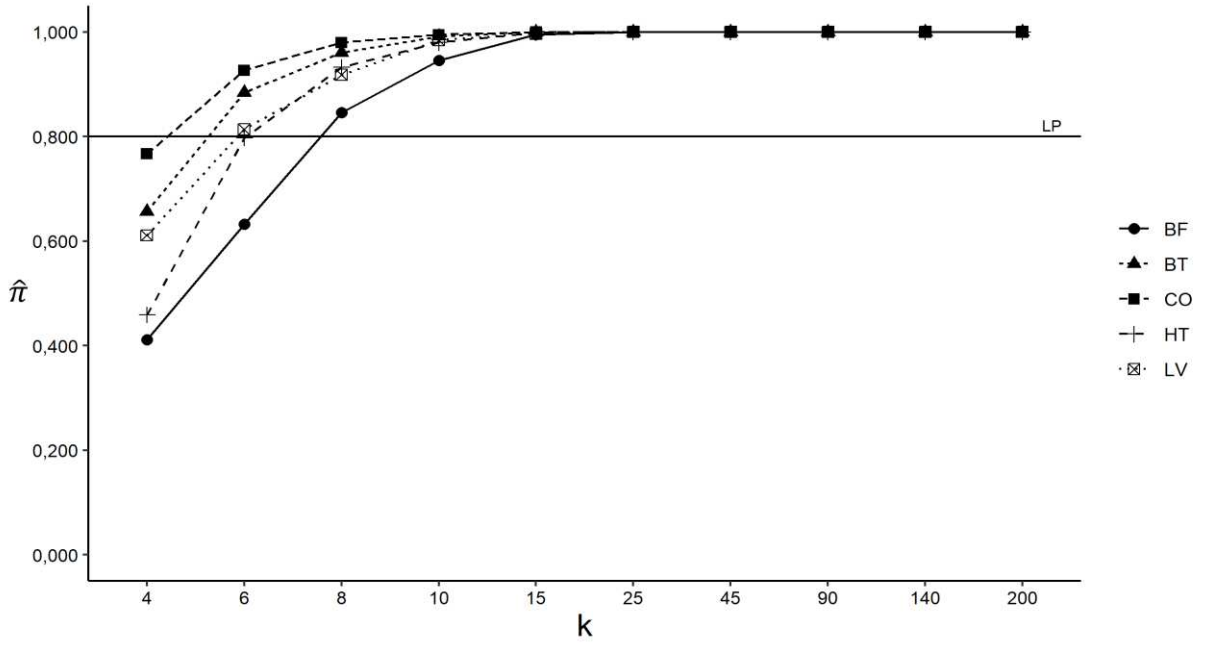
5.2.2. Proporção de heterogeneidade a_3

5.2.2.1. Cenários balanceados

Nesta seção são apresentados os resultados para o estudo de simulação realizado sob distribuição normal e proporção de heterogeneidade de variâncias a_3 $(1:1:1:1:4)$.

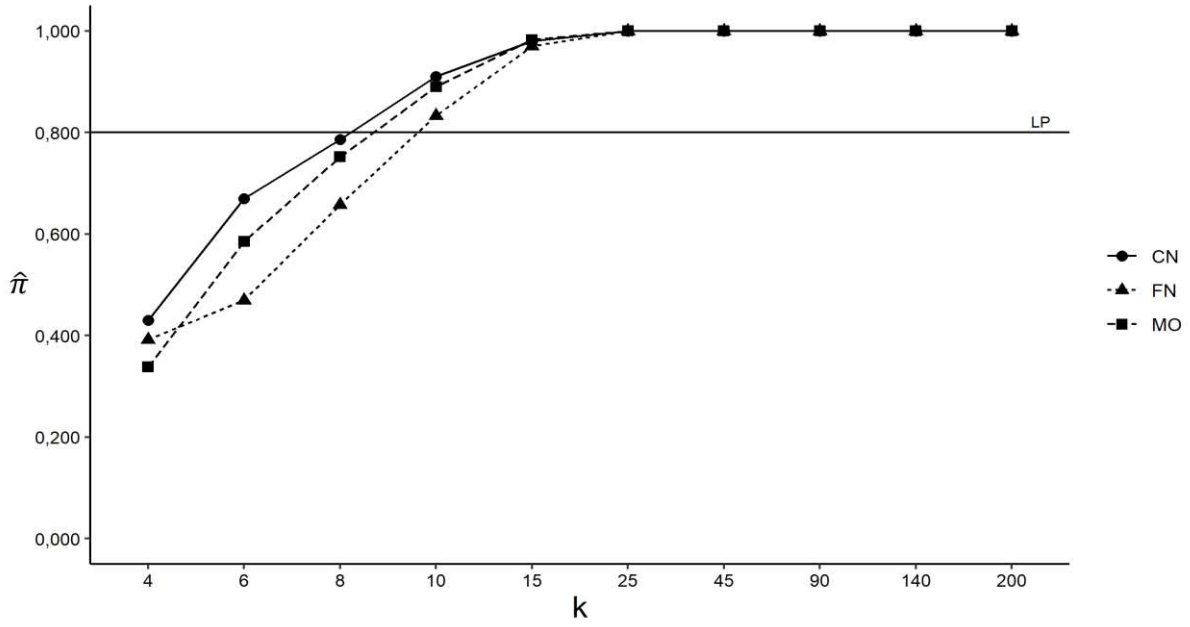
Pode ser observado na Figura 17 que, ao se utilizar a proporção de heterogeneidade de variâncias a_3, um teste tem o poder de detecção aumentado mesmo quando um menor número de repetições é utilizado e, os testes de Bartlett, de Levene e de Cochran são classificados como poderosos ao utilizar $k = 6$ repetições por tratamento, enquanto, o teste de Hartley alcança taxa empírica próxima ao limiar de poder ($\hat{\pi}_{HT} = 0{,}798$). Além do mais, com $k \geq 8$ repetições por tratamento todos os testes paramétricos avaliados foram classificados como poderosos e alcançam nível máximo de poder quando $k \geq 25$.

Figura 17 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 1: 4 (a_3), nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde: BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e, LP é o limiar da taxa empírica do poder do teste.



Ao observar os resultados obtidos para os testes não paramétricos (Figura 18) percebe-se que os testes ultrapassam o limiar da taxa empírica do poder quando $8 < k \leq 10$ observações e são classificados como poderosos. Vale ressaltar que o nível máximo para a taxa empírica do poder do teste é alcançado apenas com $k = 25$ repetições por tratamento e, nessas condições, a probabilidade de se cometer o erro do tipo II é reduzida. Contudo, mesmo com este aumento na proporção de heterogeneidade os testes não paramétricos tiveram dificuldade em obter alto poder para baixas repetições (por exemplo, quando $k \leq 8$).

Figura 18 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 1: 4 (a_3), nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e LP é o limiar da taxa empírica do poder do teste.

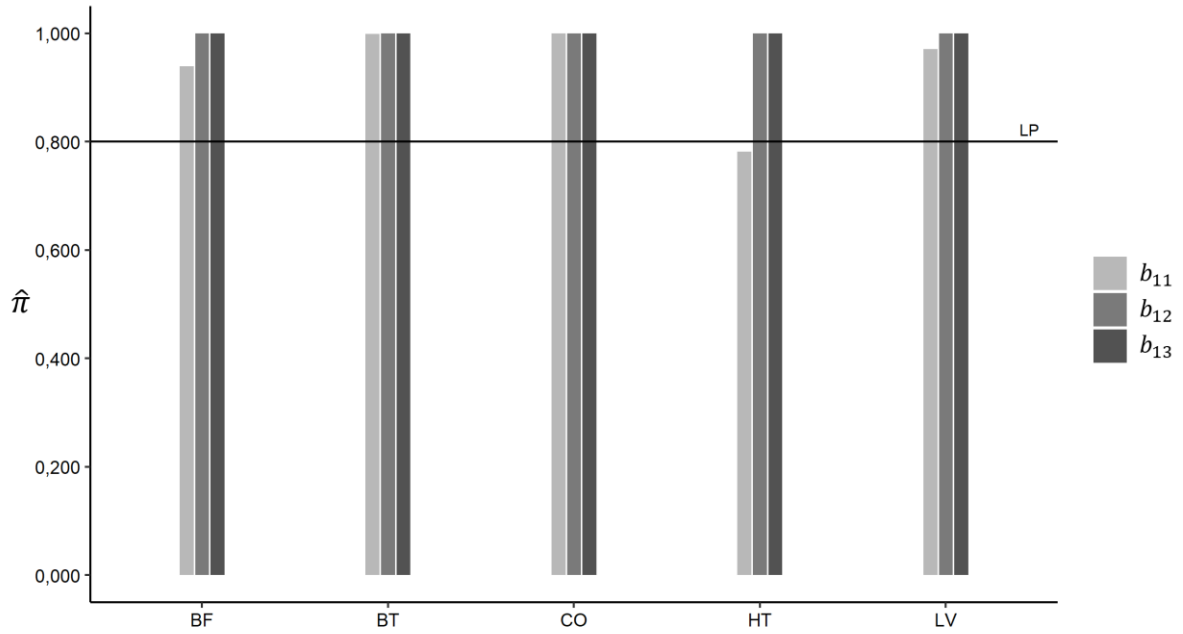


5.2.2.2. Cenários desbalanceados

Nessa seção são apresentadas as taxas empíricas do poder dos testes para cenários simulados sob distribuição normal e proporção de heterogeneidade a_3 nas proporções de desbalanceamento $4:6:8:10:15$ (b_{11}), $15:25:45:90:140$ (b_{12}) e $10:45:90:140:200$ (b_{13}).

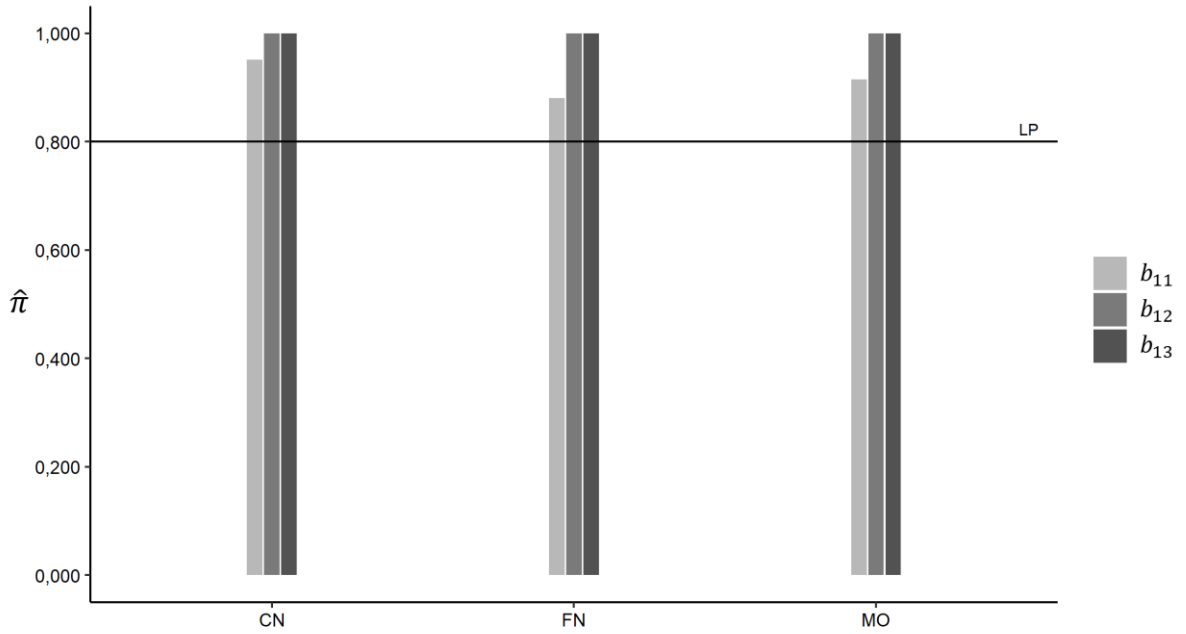
Na Figura 19 são apresentados os resultados referentes aos testes paramétricos. Por meio dela podemos observar que, os testes foram classificados como poderosos em todas as proporções de desbalanceamento, exceto, o teste de Hartley na proporção de desbalanceamento b_{11}. No mais, para as proporções de desbalanceamento b_{12} e b_{13} os cinco testes paramétricos são classificados como poderosos.

Figura 19 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 1: 4 (a_3), nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e LP é o limiar da taxa empírica do poder do teste.



Na Figura 20 são apresentadas as taxas empíricas do poder para os testes não paramétricos. Pode ser observado que para essa proporção de heterogeneidade de variâncias todos os testes foram classificados como poderosos em todas as proporções de desbalanceamentos, até mesmo na proporção de desbalanceamento b_{11}, que apresenta menor número de repetições por tratamento. Quando o estudo envolve os desbalanceamentos b_{12} e b_{13}, os três testes obtêm poder máximo ($\hat{\pi} = 1{,}000$) e têm probabilidade reduzida de cometer o erro do tipo II.

Figura 20 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 1: 4 (a_3), nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e LP é o limiar da taxa empírica do poder do teste.

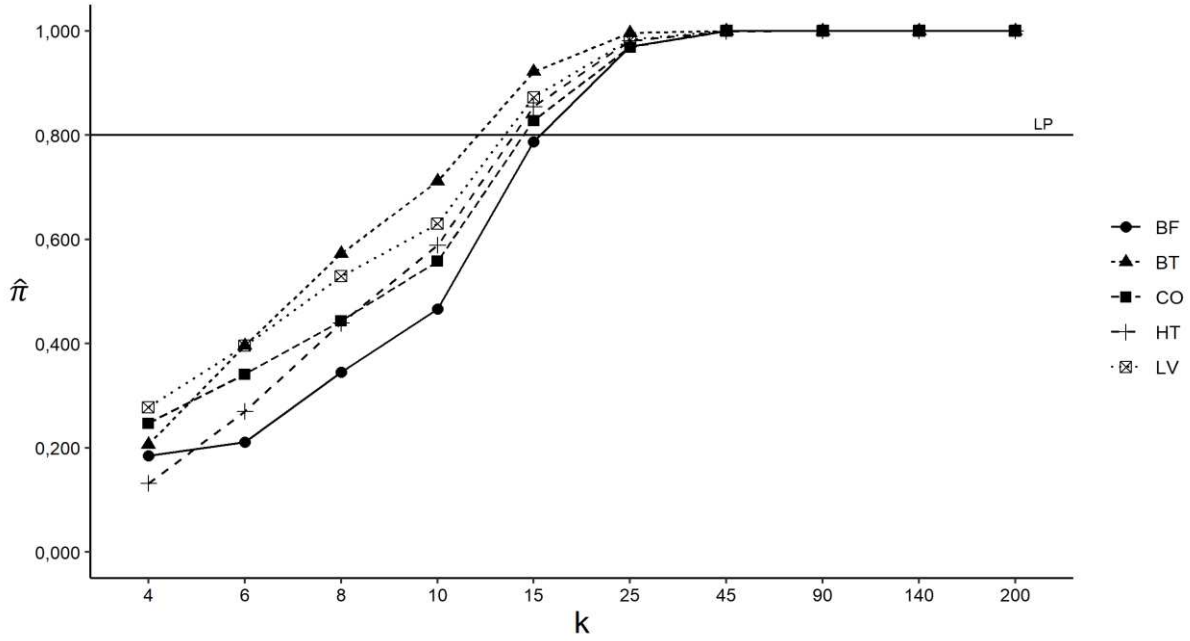


5.2.3. Proporção de heterogeneidade a_4

5.2.3.1. Cenários balanceados

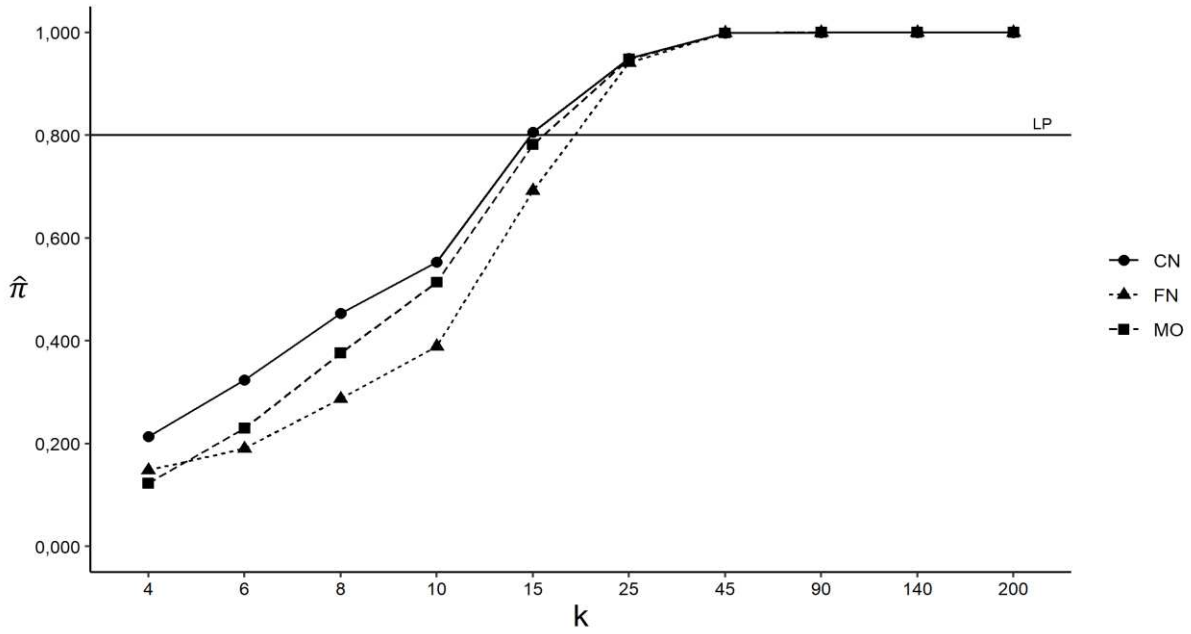
Neste tópico são apresentados os resultados referentes aos cenários simulados sob distribuição normal e proporção de heterogeneidade 1:1:1:2:2 (a_4) para cenários balanceados. Inicialmente, são abordados os resultados para os testes paramétricos apresentados na Figura 21. De acordo com esses resultados, observa-se crescente aumento da taxa empírica do poder do teste conforme o aumento do número de repetições por tratamento. Diante disso, é visto que para até 10 repetições por tratamento nenhum teste foi classificado como poderoso, contudo, com $k = 15$ repetições apenas o teste de Brown-Forsythe não alcança taxa empírica de poder $\hat{\pi} \geq 0{,}80$, os demais testes são tidos como poderosos. No entanto, quando se utilizam amostras com $k \geq 25$ observações por tratamento, os cinco testes paramétricos foram classificados como poderosos.

Figura 21 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 2: 2 (a_4) em cenários balanceados onde: BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e LP é o limiar da taxa empírica do poder do teste.



Quando os testes não paramétricos são analisados, como é apresentado na Figura 22, também foi observado um crescente aumento da taxa empírica de poder conforme o aumento do número de repetições por tratamento. Para $k \leq 10$, nenhum teste não paramétrico foi poderoso. Conover foi o teste a ser classificado como poderoso com o menor número de repetições, com $k = 15$. De acordo com a tendência observada, nota-se que os testes de Fligner-Killeen e Mood são classificados como poderosos quando $15 < k \leq 25$. Os três testes não paramétricos atingem poder máximo quando 90 repetições por tratamento são utilizadas.

Figura 22 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 2: 2 (a_4) em cenários balanceados onde: CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e, LP é o limiar da taxa empírica do poder do teste.

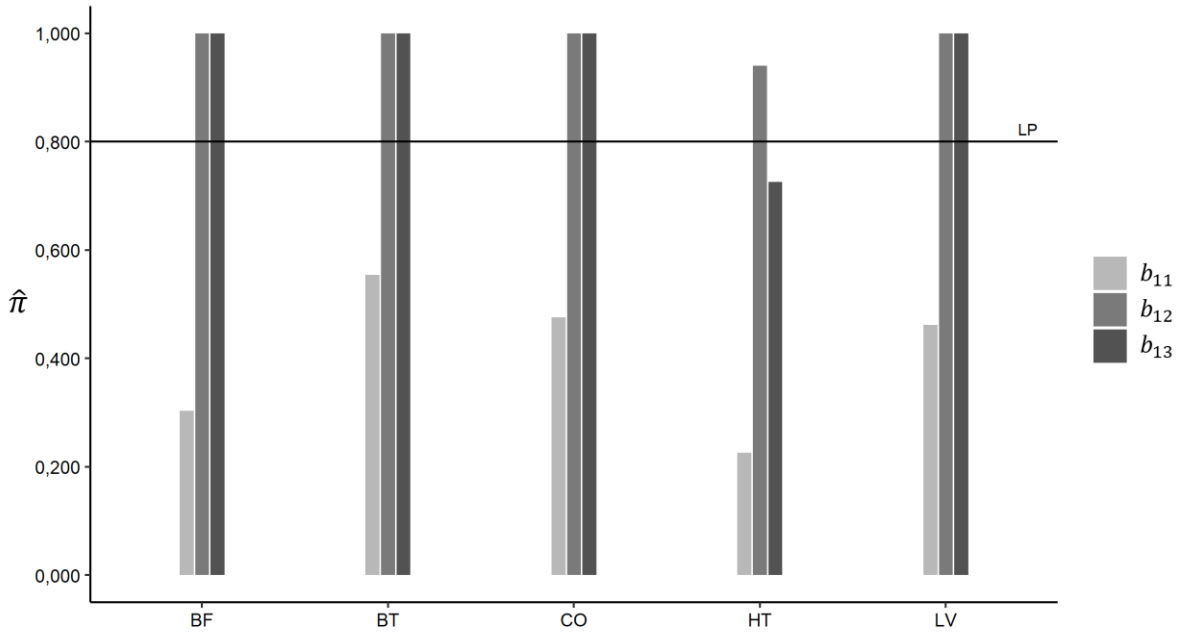


5.2.3.2. Cenários desbalanceados

Neste tópico são analisados os resultados oriundos dos cenários simulados com diferentes níveis de desbalanceamento ($4:6:8:10:15$, b_{11}; $15:25:45:90:140$, b_{12}; e $10:45:90:140:200$, b_{13}) para a proporção de heterogeneidade a_4.

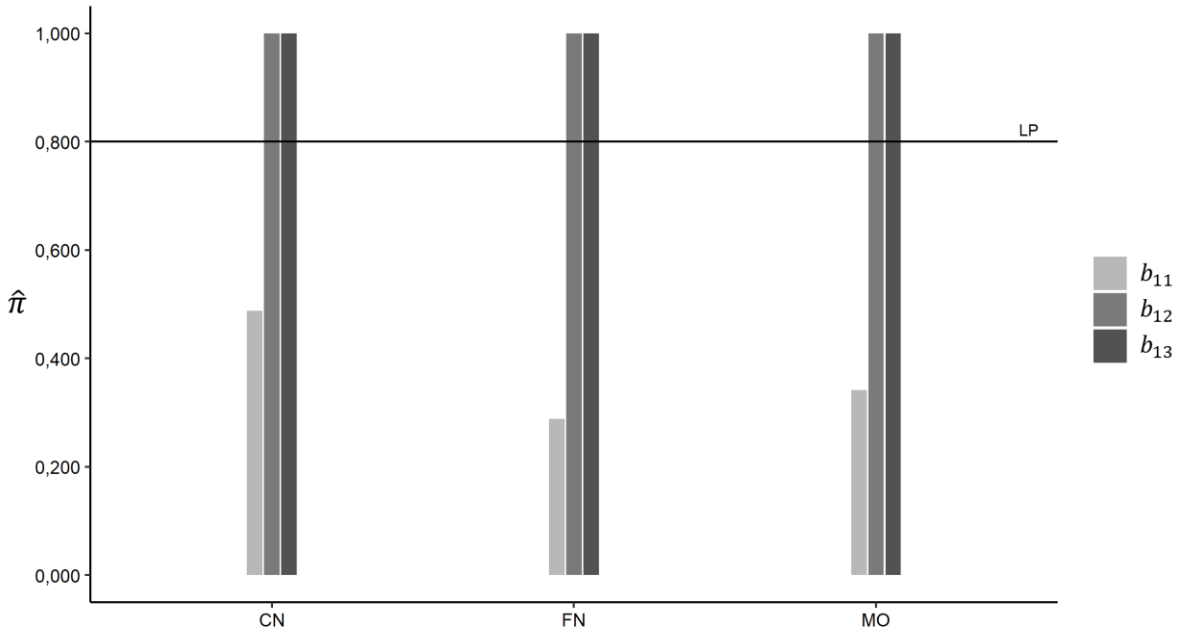
Na Figura 23 são apresentadas as taxas empíricas de poder para os testes paramétricos. Por meio desses resultados observa-se comportamento semelhante quando comparados aos resultados obtidos para os mesmos cenários simulados sob a condição de heterogeneidade a_2. Isto é, adicionar um segundo tratamento com variância populacional igual a dois não afetou de forma relevante os resultados obtidos e, nenhum dos testes foi poderoso para a proporção de desbalanceamento b_{11}. Além disso, quando a proporção de desbalanceamento b_{12} é avaliada todos os testes paramétricos foram poderosos e, para b_{13}, apenas o teste de Hartley não ultrapassou o limiar da taxa empírica de poder.

Figura 23 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 2: 2 (a_4) nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e LP é o limiar da taxa empírica do poder do teste.



Resultados semelhantes aos observados para a proporção de heterogeneidade a_2 também são observados para os testes não paramétricos quando a proporção a_4 é avaliada, conforme pode ser visto na Figura 24. A partir da imagem, nota-se que nenhum dos testes foi classificado como poderoso ao ser avaliado sob a proporção de desbalanceamento b_{11}. No entanto, ao adotar níveis de desbalanceamento onde tamanhos maiores de amostras são avaliados, como ocorre em b_{12} e b_{13}, os testes são poderosos e alcançam nível de poder máximo nesses cenários.

Figura 24 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 2: 2 (a_4) nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e LP é o limiar da taxa empírica do poder do teste.

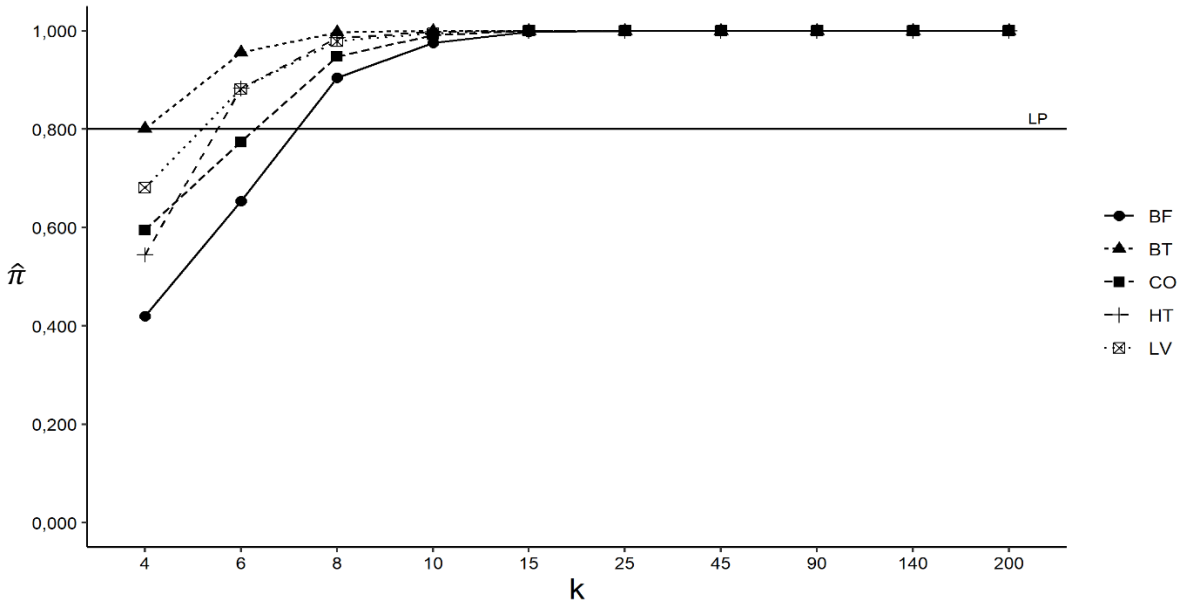


5.2.4. Proporção de heterogeneidade a_5

5.2.4.1. Cenários balanceados

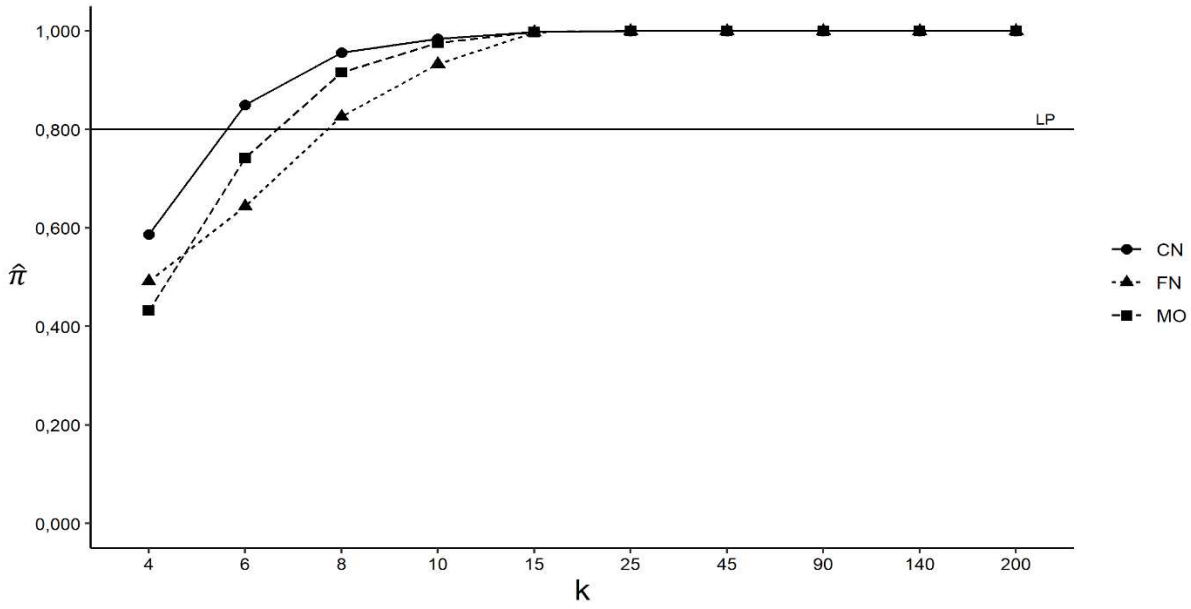
São apresentados nesse tópico os resultados referentes aos cenários simulados sob distribuição normal e proporção de heterogeneidade de variâncias dada por: 1:1:1:4:4 (a_5). Inicialmente, retrataremos os testes paramétricos expostos na Figura 25. Por meio desses resultados observa-se que o teste de Bartlett foi poderoso mesmo quando pequeno número de repetições é usado, ou seja, para $k = 4$. Ao aumentar o valor de k para 6 observações por tratamento, apenas os testes de Brown-Forsythe e de Cochran não foram poderosos. Contudo, para $k \geq 8$, todos os testes ultrapassam o limiar da taxa empírica de poder e foram poderosos, isto é, esses testes têm menor probabilidade de falhar ao fazer uma inferência sobre a hipótese de homogeneidade de variâncias. Vale ressaltar que o teste de Bartlett alcançou poder empírico máximo com 10 repetições por tratamento, ao passo que os testes Cochran, Hartley e Levene alcançam poder empírico igual a um com $k = 15$ repetições.

Figura 25 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 4: 4 (a_5) e cenários balanceados onde: BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e LP é o limiar da taxa empírica do poder do teste.



Quando os testes não paramétricos são analisados, é possível observar através da Figura 26, que adicionar a dupla de variâncias de valor quatro, enquanto as demais permanecem iguais a um, proporcionou um aumento no poder de detecção do teste e fez com que o teste de Conover fosse classificado como poderoso com $k = 6$ repetições. Além disso, esse teste atingiu o limiar da taxa empírica do poder do teste com o menor número de repetições por tratamento. Ademais, ao utilizar 8 repetições em amostras balanceadas, os três testes não paramétricos foram poderosos, ao passo que, o poder máximo para esses testes foi atingido quando $k = 15$ repetições.

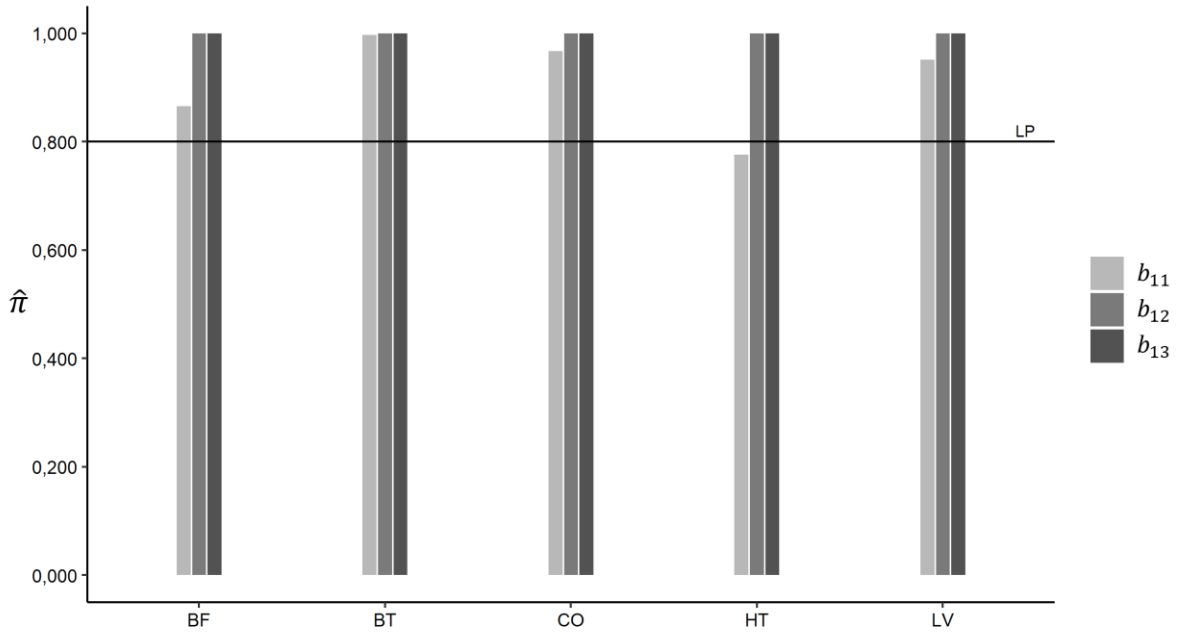
Figura 26 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 4: 4 (a_5) em cenários balanceados onde CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e LP é o limiar da taxa empírica do poder do teste.



5.2.4.2. Cenários desbalanceados

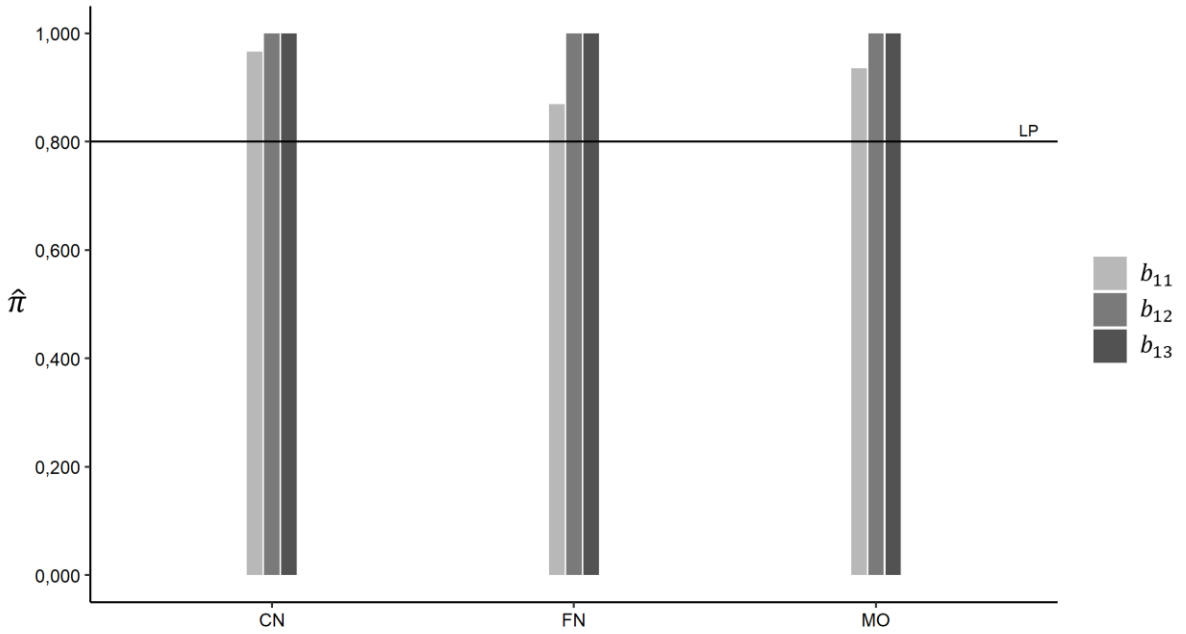
Ao avaliar os testes paramétricos sob a proporção de heterogeneidade de variâncias 1:1:1:4:4 (a_5) nas proporções de desbalanceamento $4:6:8:10:15$ (b_{11}), $15:25:45:90:140$ (b_{12}) e $10:45:90:140:200$ (b_{13}), apresentados na Figura 27, observou-se que todos os testes foram poderosos ao inferir sob esta proporção de heterogeneidade. Exceção a esse resultado foi o teste de Hartley que não ultrapassou o limiar da taxa empírica do poder quando avaliado na proporção de desbalanceamento b_{11}, que possui menor número repetições por tratamento. Contudo, para as proporções de desbalanceamento b_{12} e b_{13}, foi observado que os testes paramétricos tendem a possuir alto poder de detecção quando as amostras envolvem maior número de repetições e todos os testes alcançam o máximo da taxa empírica de poder.

Figura 27 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 4: 4 (a_5) nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e LP é o limiar da taxa empírica do poder do teste.



Ao retratar os resultados obtidos para os testes não paramétricos, sob as mesmas condições de simulação previamente comentadas para os testes paramétricos, é observado na Figura 28 alto poder de detecção para os testes de Conover, Fligner-Killeen e Mood ao atuar nas proporções de desbalanceamentos b_{11}, b_{12} e b_{13}. Contudo, ao utilizar proporções de desbalanceamento que contém maior número de repetições por tratamento, como em b_{12} e b_{13}, os três testes alcançaram nível de poder máximo.

Figura 28 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 1: 1: 4: 4 (a_5) nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e LP é o limiar da taxa empírica do poder do teste.

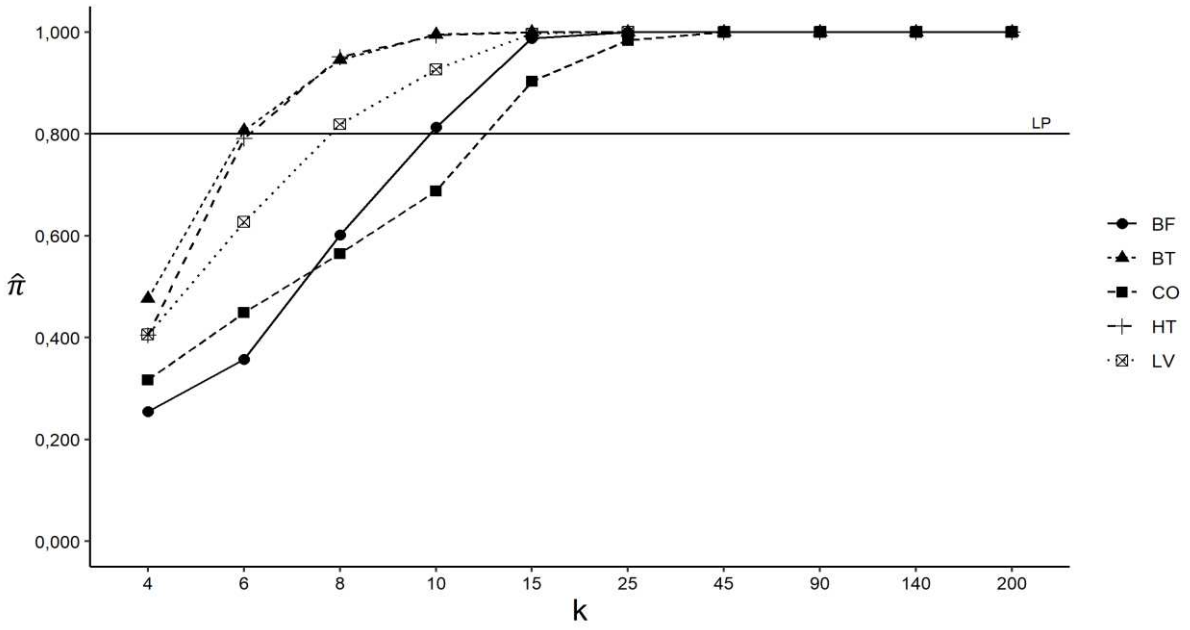


5.2.5. Proporção de heterogeneidade a_6

5.2.5.1. Cenários balanceados

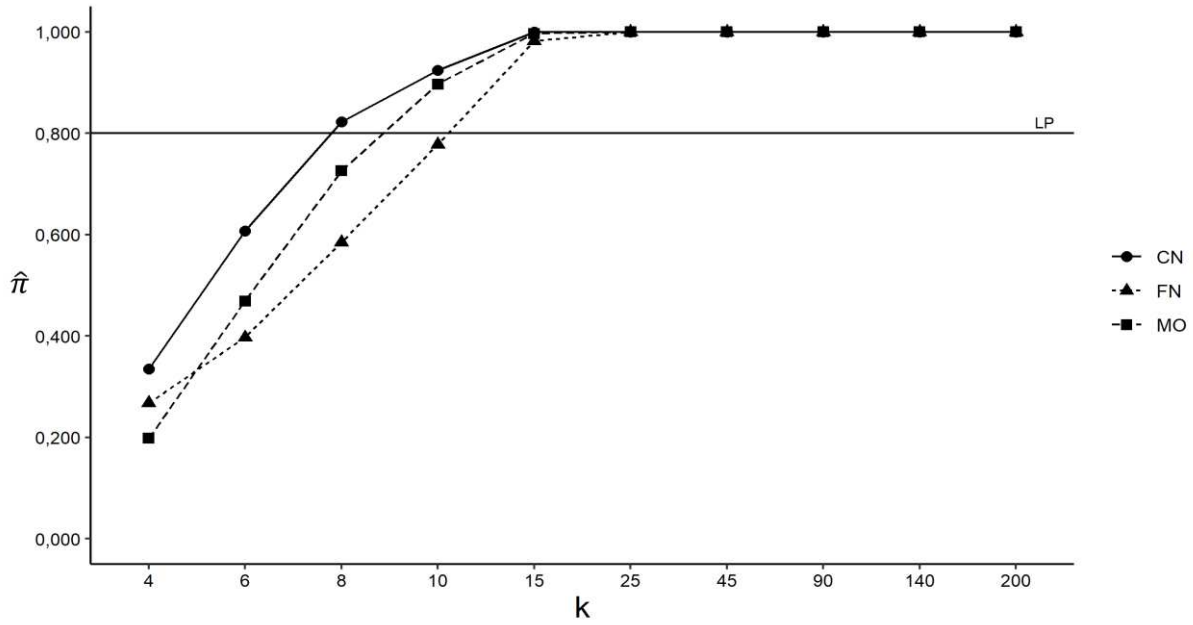
Nesta seção são apresentados os resultados para o estudo de simulação executado sob a proporção de heterogeneidade 1:2:3:4:5 (a_6). Na Figura 29, são apresentadas as taxas empíricas do poder para os testes paramétricos. Nela é observado um maior distanciamento entre as curvas e diferentes grau de inclinação. Foi constatado neste cenário que o teste de Bartlett foi classificado como poderoso mesmo lidando com pequenas amostras, com $k = 6$. Contudo, comportamento similar ao teste de Bartlett é observado para o teste de Hartley e, de acordo com a tendência observada, ele ultrapassa o limiar de poder com $6 < k \leq 8$ repetições. Ao se utilizar $k = 8$ repetições, apenas os testes de Brown-Forsythe e Cochran não foram classificados como poderosos. No entanto, tanto o teste de Brown-Forsythe quanto o teste de Cochran possuem curvas de ganho de poder menos inclinadas e necessitam de um maior número de repetições por tratamento para alcançar o limiar de poder e, assim, o teste de Brown-Forsythe foi poderoso ao utilizar 10 repetições e, o teste de Cochran, foi poderoso apenas quando $k = 15$.

Figura 29 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal proporção de heterogeneidade 1: 2: 3: 4: 5 (a_6) no cenário balanceado onde BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e LP é o limiar da taxa empírica do poder do teste.



Ao analisar os testes não paramétricos retratados na Figura 30, observou-se que eles precisaram de um maior número de repetições por tratamento para serem classificados como poderosos ao comparar com os resultados obtidos para os mesmos cenários (Figura 26) simulados sob a proporção de heterogeneidade a_5. O primeiro teste a ser classificado como poderoso foi o teste de Conover com 8 repetições em cada população. No entanto, ao verificar a tendência gerada pela curva da taxa empírica do poder do teste, notou-se que, o poder aumenta à medida que se aumenta o número de repetições entre os tratamentos, e o teste de Mood ultrapassou o limiar de poder com $8 < k \leq 10$ repetições. Ainda observando a tendência destacada pelos resultados exibidos na figura abaixo, o teste de Fligner-Killeen atinge o limiar com $10 < k \leq 15$ observações. Ademais, o teste de Conover foi o primeiro a atingir nível máximo de poder, com 15 repetições entre tratamentos e, os testes Fligner-Killeen e Mood, alcançam taxa empírica máxima de poder com $k = 25$ observações.

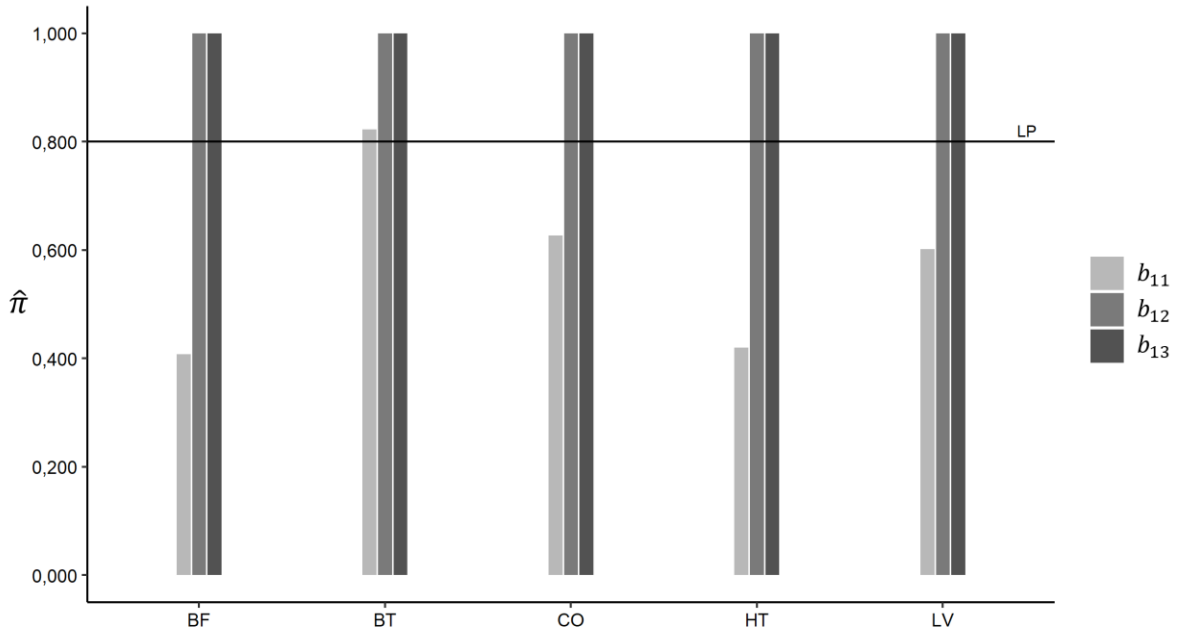
Figura 30 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 2: 3: 4: 5 (a_6) e balanceado onde CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e LP é o limiar da taxa empírica do poder do teste.



5.2.5.2. Cenários desbalanceados

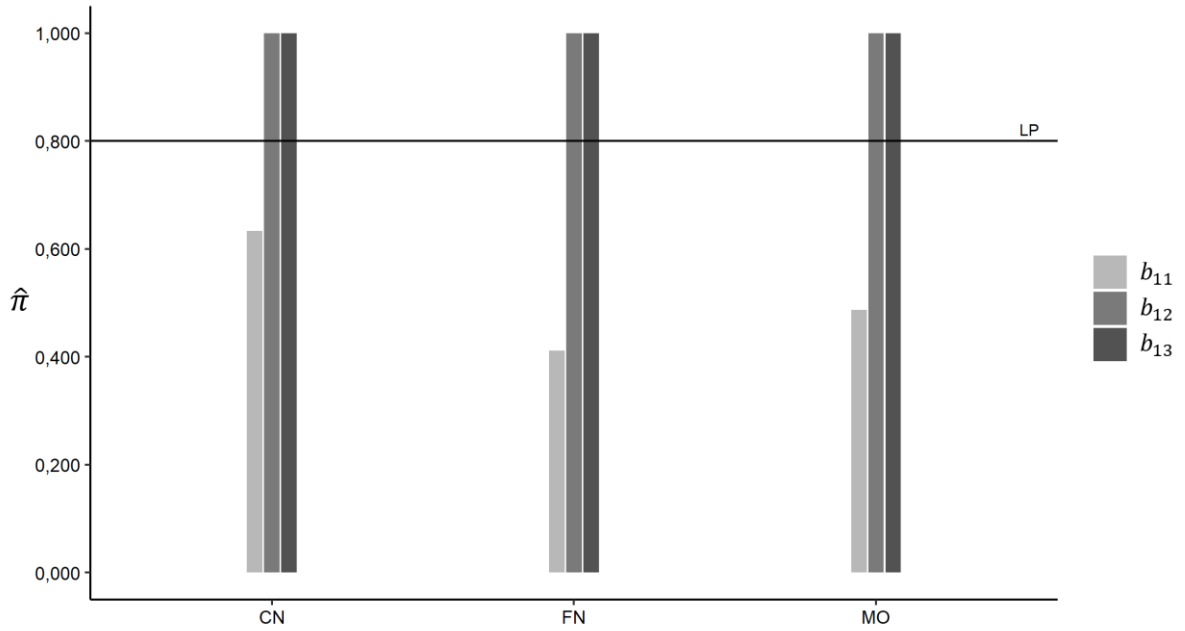
Serão abordados agora os resultados gerados pelo estudo de simulação para os cenários simulados de acordo com as proporções de desbalanceamento $4{:}\,6{:}\,8{:}\,10{:}\,15$ (b_{11}), $15{:}\,25{:}\,45{:}\,90{:}\,140$ (b_{12}) e $10{:}\,45{:}\,90{:}\,140{:}\,200$ (b_{13}). Na Figura 31 são apresentadas as taxas empíricas do poder para os testes paramétricos. Através dela é possível observar que o teste de Bartlett foi o único a ser classificado como poderoso nos três níveis de desbalanceamento avaliados, mesmo quando o desbalanceamento envolviam tratamentos com $r_i \leq 15$, para $i = 1, \ldots, 5$, ou seja, a proporção de desbalanceamento b_{11}. No entanto, ao avaliar as populações compostas por um maior número de observações, como em b_{12} e b_{13}, todos os testes paramétricos foram poderosos, com $\hat{\pi} = 1{,}000$.

Figura 31 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 2: 3: 4: 5 (a_6) nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde BF, o teste de Brown-Forsythe; BT, é o teste de Bartlett; CO, é o teste de Cochran; HT, é o teste de Hartley; LV, o teste de Levene; e LP é o limiar da taxa empírica do poder do teste.



Ao analisar os testes não paramétricos da Figura 32, pode ser observado que nenhum deles foi poderoso para o cenário desbalanceado b_{11}. Contudo, ao utilizar proporções de desbalanceamentos com número de repetições mais elevados, como em b_{12} e b_{13}, os testes não paramétricos de Conover, de Fligner-Killeen e de Mood alcançaram poder máximo e tem probabilidade reduzida de cometer o erro do tipo II.

Figura 32 - Taxa empírica do poder do teste ($\hat{\pi}$) para os testes não paramétricos nos cenários simulados sob distribuição normal, proporção de heterogeneidade 1: 2: 3: 4: 5 (a_6) nas proporções de desbalanceamento $b_{11}(4:6:8:10:15)$, b_{12} $(15:25:45:90:140)$ e b_{13} $(10:45:90:140:200)$ onde CN, é o teste de Conover; FN, é o teste de Fligner-Killeen; MO, é o teste de Mood; e LP é o limiar da taxa empírica do poder do teste.



6. Discussão

6.1. Discussão por teste de homogeneidade de variâncias

Nesta seção os mesmos resultados apresentados anteriormente serão agrupados por teste e discutidos seus resultados quanto as taxas empíricas do erro tipo I ($\hat{\alpha}$) e poder do teste ($\hat{\pi}$) para cada configuração de simulação avaliada nesse estudo. O objetivo dessa análise é identificar em quais cenários cada um dos testes apresentaram melhor performance, ou seja, cenários em que cada teste foi classificado como exato e poderoso. Tal identificação, a princípio, permitiria a um analista de dados escolher o teste mais apropriado de acordo com a base de dados que ele tem disponível.

Com essa finalidade, inicialmente são retratadas as taxas empíricas $\hat{\alpha}$ em cada distribuição de probabilidade, f_w, com $w = 1,2,3$, para cada proporção de desbalanceamento, b_n, onde: $4{:}\,4{:}\,4{:}\,4{:}\,4$ (b_1); $6{:}\,6{:}\,6{:}\,6{:}\,6$ (b_2); $8{:}\,8{:}\,8{:}\,8{:}\,8$ (b_3); $10{:}\,10{:}\,10{:}\,10{:}\,10$ (b_4); $15{:}\,15{:}\,15{:}\,15{:}\,15$ (b_5); $25{:}\,25{:}\,25{:}\,25{:}\,25$ (b_6); $45{:}\,45{:}\,45{:}\,45{:}\,45$ (b_7); $90{:}\,90{:}\,90{:}\,90{:}\,90$ (b_8); $140{:}\,140{:}\,140{:}\,140{:}\,140$ (b_9); $200{:}\,200{:}\,200{:}\,200{:}\,200$ (b_{10}); $4{:}\,6{:}\,8{:}\,10{:}\,15$ (b_{11}); $15{:}\,25{:}\,45{:}\,90{:}\,140$ (b_{12}); $10{:}\,45{:}\,90{:}\,140{:}\,200$ (b_{13}). Posteriormente, sob a distribuição normal, os testes foram classificados de acordo com a taxa empírica do poder do teste, $\hat{\pi}$, em cada uma das proporções de heterogeneidade, a_j, com $j = 2, \ldots, 6$, para cada uma das b_n proporções de desbalanceamento, para $n = 1, \ldots, 13$.

Iniciaremos retratando os resultados referentes ao teste de Bartlett, apresentados na Tabela 1. Pode ser observado que quando este teste é executado sob populações normalmente distribuídas, ele tende a controlar a taxa empírica de erro do tipo I ($\hat{\alpha}$) dentro dos limites adaptados para o Critério Liberal de Bradley (1978), mesmo em pequenas amostras, e foi considerado um teste exato tanto nos cenários balanceados quanto nos cenários desbalanceados, concordando, assim, com os resultados obtidos por Bhandary e Dai (2008). Por outro lado, o teste de Bartlett teve classificação liberal quando simulado em populações cuja distribuição não era normal (qui-quadrado e beta) em todos os b_n níveis de desbalanceamento avaliados. Resultados semelhantes a estes foram encontrados por Sharma e Kibria (2013) e Vorapongsathorn, Taejaroenkul e Viwatwongkasem (2004) quando, em seus estudos, utilizaram populações não normais no processo de simulação para avaliar os cenários balanceados e desbalanceados e o teste de Bartlett foi classificado como liberal. Devido a tais fatos, este foi considerado como não robusto à violação da

pressuposição de normalidade. Esse resultado pode ser devido ao teste de Bartlett ser derivado do teste da razão de verossimilhança e, por essa causa, ser mais sensível à desvios de normalidade (CONOVER; JOHNSON; JOHNSON, 1981).

Com relação a taxa empírica do poder do teste (Tabela 1), percebe-se que o teste de Bartlett tende a aumentar o poder de detecção à medida que aumenta o número de repetições por tratamento e/ou a proporção de heterogeneidade entre populações. Resultados análogos quanto ao crescente aumento de poder foram encontrados por Vorapongsathorn, Taejaroenkul e Viwatwongkasem (2004) e Baydili e Siğirli (2017). Desta forma, nota-se que quando existe um menor grau de heterogeneidade de variâncias entre populações (como na proporção a_2 e a_4) o teste de Bartlett tem poder comparável aos demais teste paramétricos, sendo superado apenas pelo teste de Cochran em a_2. Porém, ao adicionar maior grau de heterogeneidade, como em a_5, o teste de Bartlett foi poderoso com o menor número de repetições por tratamento, com $k = 4$. Riboldi *et al.* (2014) concluem em seu trabalho que, sob normalidade, o teste de Bartlett infere o resultado com alto poder e foi capaz de identificar eficientemente os casos com variâncias heterogêneas.

Tabela 1 - Resumo dos resultados do estudo de simulação para a taxa empírica de erro do tipo I e taxa empírica do poder para o teste paramétrico de Bartlett. As classificações são: C, para um teste conservador; E, para um teste exato; L, para um teste liberal; S, para um teste poderoso; e, N, para um teste não-poderoso. As proporções de desbalanceamento são: b_1, $4:4:4:4:4$; b_2, $6:6:6:6:6$; b_3, $8:8:8:8:8$; b_4, $10:10:10:10:10$; b_5, $15:15:15:15:15$; b_6, $25:25:25:25:25$; b_7, $45:45:45:45:45$; b_8, $90:90:90:90:90$; b_9, $140:140:140:140:140$; b_{10}, $200:200:200:200:200$; b_{11}, $4:6:8:10:15$; b_{12}, $15:25:45:90:140$; e b_{13}, $10:45:90:140:200$.

Classificação de acordo com a taxa empírica do erro tipo I												
Distribuição normal e variâncias homogêneas												
Proporção de desbalanceamento:												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
E	E	E	E	E	E	E	E	E	E	E	E	E
Distribuição de qui-quadrado e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L
Distribuição beta e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L

Classificação de acordo taxa empírica do poder do teste													
Distribuição normal e variâncias heterogêneas													
	Proporção de desbalanceamento:												
Proporção de heterogeneidade:	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
1:1:1:1:2 (a_2)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:1:4 (a_3)	N	N	S	S	S	S	S	S	S	S	S	S	S
1:1:1:2:2 (a_4)	N	N	N	N	S	S	S	S	S	S	N	S	S
1:1:1:4:4 (a_5)	S	S	S	S	S	S	S	S	S	S	S	S	S
1:2:3:4:5 (a_6)	N	S	S	S	S	S	S	S	S	S	S	S	S

Na Tabela 2 são apresentados os resultados obtidos através do estudo de simulação para o teste de Levene. Primeiramente, tratando-se da avaliação da taxa empírica do erro do tipo I ($\hat{\alpha}$) para os conjuntos de dados simulados sob distribuição normal, percebe-se que este obteve resultados comparáveis aos obtidos por Bhandary e Dai (2008) e Wang *et al.* (2017), não conseguindo controlar a taxa empírica $\hat{\alpha}$ e teve classificação liberal ao atuar em amostras com até 15 repetições por tratamento, no entanto, alcançou exatidão para amostras balanceadas de tamanho $k \geq 25$ repetições (para b_n, com $n = 6, \dots, 10$). Além disso, em cenários desbalanceados e sob a mesma distribuição, o teste de Levene foi liberal no cenário composto pelos menores números de repetições entre os tratamentos (isto é, $r_i \leq 15$, para $i = 1, \dots, 5$), como acontece em b_{11}, e foi classificado como exato ao atuar em cenários cujo número de repetições foram maiores, como ocorre para as

configurações de desbalanceamento b_{12} e b_{13}. De acordo com estudos de Veitch e Roscoe (1974), o teste de Levene foi liberal ao atuar em populações balanceadas com $k = 10$; exato quando $k = 20$ e 30, obtendo resultados semelhantes aos do presente trabalho; e, para amostra desbalanceadas o mesmo atuou de forma liberal, o que não ocorreu neste trabalho, conforme se vê em b_{12} e b_{13}. Analisando os cenários simulados sob a distribuição de qui-quadrado e beta, o teste de Levene não controlou a taxa empírica do erro tipo I e foi classificado como liberal em todas as b_n proporções de desbalanceamento, para $n = 1, \ldots, 13$. Baydili e Siğirli (2017) e Conover, Johnson e Johnson (1981) constataram que o teste de Levene resultou em taxas do erro tipo I liberais quando o experimento foi realizado sob distribuições assimétricas, conforme evidenciado neste estudo.

Quando a avaliação é feita levando-se em consideração a taxa empírica do poder do teste ($\hat{\pi}$), é observado na Tabela 2 que quando proporções de heterogeneidade menos discrepantes (a_2 e a_4) são analisadas, o teste necessita de um maior número de repetições por tratamento para ser considerado poderoso, ao passo que, ao adicionar um maior grau de heterogeneidade (proporções a_3, a_5 e a_6) o teste alcança $\hat{\pi} \geq 0{,}80$ para um número de repetições relativamente menor. Vale ressaltar que para a proporção de heterogeneidade a_3 o teste de Levene foi classificado como poderoso com o menor número de repetições por tratamento. Riboldi *et al.* (2014) encontraram em seu estudo que o teste de Levene possui alto poder e, dizem ainda, que este se assemelha ao teste de Bartlett quando se trata do poder do teste.

Tabela 2 - Resumo dos resultados do estudo de simulação para a taxa empírica de erro do tipo I e taxa empírica do poder para o teste paramétrico de Levene. As classificações são: C, para um teste conservador; E, para um teste exato; L, para um teste liberal; S, para um teste poderoso; e, N, para um teste não-poderoso. As proporções de desbalanceamento são: b_1, $4:4:4:4:4$; b_2, $6:6:6:6:6$; b_3, $8:8:8:8:8$; b_4, $10:10:10:10:10$; b_5, $15:15:15:15:15$; b_6, $25:25:25:25:25$; b_7, $45:45:45:45:45$; b_8, $90:90:90:90:90$; b_9, $140:140:140:140:140$; b_{10}, $200:200:200:200:200$; b_{11}, $4:6:8:10:15$; b_{12}, $15:25:45:90:140$; e b_{13}, $10:45:90:140:200$.

Classificação de acordo com a taxa empírica do erro tipo I												
Distribuição normal e variâncias homogêneas												
Proporção de desbalanceamento:												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	E	E	E	E	E	L	E	E
Distribuição de qui-quadrado e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L
Distribuição beta e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L

Classificação de acordo taxa empírica do poder do teste													
Distribuição normal e variâncias heterogêneas													
	Proporção de desbalanceamento:												
Proporção de heterogeneidade:	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
1:1:1:1:2 (a_2)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:1:4 (a_3)	N	S	S	S	S	S	S	S	S	S	S	S	S
1:1:1:2:2 (a_4)	N	N	N	N	S	S	S	S	S	S	N	S	S
1:1:1:4:4 (a_5)	N	S	S	S	S	S	S	S	S	S	S	S	S
1:2:3:4:5 (a_6)	N	N	S	S	S	S	S	S	S	S	N	S	S

Quando avaliado o teste de Brown-Forsyhte, cujos resultados são apresentados na Tabela 3, observa-se que, de modo geral, substituir a média pela mediana na transformação dos valores observados sob a estatística de teste de Levene tornou o teste de Brown-Forsyhte robusto a desvios da normalidade. Em seu estudo, Conover, Johnson e Johnson (1981) afirmam que todos os testes escolhidos como robustos usaram a mediana ao invés da média. Resultados similares foram encontrados por Ramsey e Ramsey (2007), Sharma e Kibria (2013), Li *et al.* (2015) e Baydili e Siğirli (2017) ao constatarem que o teste de Brown-Forsythe apresenta robustez a uma ampla gama de distribuições de probabilidades e tamanhos de amostras. Analisando inicialmente os cenários simulados sob distribuição normal, percebe-se que o teste de Brown-Forsyhte iniciou com desempenho liberal, para $k = 4$; conservou H_0 verdadeira numa proporção menor do que o esperado para $\alpha = 0{,}05$ quando o número de

repetições por tratamento variou entre $6 \leq k \leq 25$; e, foi exato para as proporções de desbalanceamento b_7 até b_{13}. No entanto, ao ser avaliado quanto ao controle da taxa de erro do tipo I em populações não normais, tem-se que sob as distribuições beta e qui-quadrado, os resultados empíricos mostraram que este teste tende a controlar a taxa de erro $\hat{\alpha}$ e foi considerado robusto nas proporções de desbalanceamento b_6 até b_{13}, isto é, para $k \geq 15$. Resultados parecidos foram apresentados por Algina, Blair e Coombs (1995) e Parra-Frutos (2013), onde encontraram que o teste de Brown-Forsythe teve bom desempenho (foi exato) em tamanhos maiores de amostra, porém, foi conservador em tamanhos de amostras menores (ou seja, para $k = 5$ repetições) nas diferentes distribuições. Shoemaker (2003) observou que, em distribuições simétricas e assimétricas, o teste de Brown-Forsythe controlou a taxa de erro tipo I no valor nominal para grandes tamanhos de amostras e foi conservador quando os tamanhos de amostras eram menores. Segundo Baydili e Siğirli (2017), quando a distribuição de qui-quadrado foi avaliada, este foi o teste a apresentar melhor resultado quanto a taxa de erro tipo I.

Contudo, ao avaliar a taxa empírica do poder ($\hat{\pi}$) para o teste de Brown-Forsythe (Tabela 3), os resultados indicaram que este perde em poder quando comparado com o teste de Levene (ver Tabela 2) e necessita de um maior número de repetições por tratamento para alcançar o limiar da taxa empírica do poder. Este fato pode ser verificado para os seguintes resultados simulados sob a condição H_0 falsa: para a proporção a_4, o teste de Brown-Forsythe ultrapassou o limiar da taxa empírica do poder apenas com 25 repetições, enquanto, que para o teste de Levene bastaram 15; nas proporções de heterogeneidade a_3 e a_5, foi poderoso com $k = 8$ repetições, ao passo que, para Levene, foi poderoso com 6 repetições; e em a_6, o teste foi poderoso a partir de $k = 10$ observações, contra $k = 8$ para o teste de Levene. Apesar do teste de Brown-Forsythe ser considerado poderoso, nosso estudo concorda com os estudos de Riboldi *et al.* (2014) em que constatou-se que este é superado pelo teste de Bartlett. Além disso, em nosso estudo, comparando o poder do teste, o teste de Brown-Forsythe foi superado também pelo teste de Levene. Parra-Frutos (2009) relatou em seu estudo que o teste de Brown-Forsythe (teste da mediana de Levene) teve bons resultados para o poder apenas quando a diferença entre as taxas de variâncias era alta.

Tabela 3 - Resumo dos resultados do estudo de simulação para a taxa empírica de erro do tipo I e taxa empírica do poder para o teste paramétrico de Brown-Forsythe. As classificações são: C, para um teste conservador; E, para um teste exato; L, para um teste liberal; S, para um teste poderoso; e, N, para um teste não-poderoso. As proporções de desbalanceamento são: b_1, $4:4:4:4:4$; b_2, $6:6:6:6:6$; b_3, $8:8:8:8:8$; b_4, $10:10:10:10:10$; b_5, $15:15:15:15:15$; b_6, $25:25:25:25:25$; b_7, $45:45:45:45:45$; b_8, $90:90:90:90:90$; b_9, $140:140:140:140:140$; b_{10}, $200:200:200:200:200$; b_{11}, $4:6:8:10:15$; b_{12}, $15:25:45:90:140$; e b_{13}, $10:45:90:140:200$.

Classificação de acordo com a taxa empírica do erro tipo I												
Distribuição normal e variâncias homogêneas												
Proporção de desbalanceamento:												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	C	C	C	C	C	E	E	E	E	E	E	E
Distribuição de qui-quadrado e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	E	C	C	E	E	E	E	E	E	E	E
Distribuição beta e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	E	E	E	C	E	E	E	E	E	E	E	E

Classificação de acordo taxa empírica do poder do teste													
Distribuição normal e variâncias heterogêneas													
	Proporção de desbalanceamento:												
Proporção de heterogeneidade:	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
1:1:1:1:2 (a_2)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:1:4 (a_3)	N	N	S	S	S	S	S	S	S	S	S	S	S
1:1:1:2:2 (a_4)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:4:4 (a_5)	N	N	S	S	S	S	S	S	S	S	S	S	S
1:2:3:4:5 (a_6)	N	N	N	S	S	S	S	S	S	S	N	S	S

Na Tabela 4 são apresentados os resultados do estudo de simulação de dados realizado para o teste paramétrico de Cochran. Por meio desses resultados é possível observar que para amostras balanceadas e sob distribuição normal, o teste de Cochran tende a controlar a taxa empírica do erro tipo I ($\hat{\alpha}$) e atuar como um teste exato. Resultados semelhantes aos de nosso estudo são encontrados por Wang *et al.* (2017), Bhandary e Dai (2008), Parra-Frutos (2013) e Lee, Katz e Restori (2010) ao concluírem sobre o controle do erro tipo I para o teste de Cochran quando a condição de normalidade é satisfeita. Exceção ao padrão de tendência observada é constatada quando $k = 15$ e o teste foi classificado como conservador (Tabela 4). No entanto, ao lidar com amostras desbalanceadas, o teste de Cochran conservou H_0 verdadeira numa quantidade de vezes maior do que o esperado (para $\alpha = 0{,}05$) e teve $\hat{\alpha}$ liberal para as proporções de desbalanceamento b_{11}, b_{12} e b_{13}. Resultados liberais em

experimentos desbalanceados também foram encontrados por Veitch e Roscoe (1974) quando utilizaram a média aritmética para a obtenção do grau de liberdade. Porém, em experimentos balanceados, o autor obteve que este foi um teste conservador, em que a taxa de erro tipo I decresce à medida que o número de repetições por tratamento aumenta. Esse comportamento liberal em cenários desbalanceados pode ser devido ao teste usar a média aritmética ou a média harmônica de r_i no lugar de r para calcular o número de graus de liberdade. Contudo, quando populações não normais são avaliadas, como ocorre para a distribuição beta e qui-quadrado, o teste de Cochran foi classificado como liberal em todas as proporções de desbalanceamento, tanto nos cenários balanceados quanto nos desbalanceados. Veitch e Roscoe (1974) e Baydili e Siğirli (2017) concluíram que o teste de Cochran foi um dos testes a apresentar piores resultados quanto ao controle do erro tipo I sob desvios da normalidade.

Padrão similar aos demais testes anteriormente comentados é verificado ao analisar a taxa empírica do poder do teste de Cochran (Tabela 4), dado que o poder possui relação direta com o tamanho da amostra e com a proporção de heterogeneidade entre os tratamentos. Ao analisar os cenários balanceados nas proporções de heterogeneidade a_2, a_4 e a_6 o teste de Cochran foi classificado como poderoso ao simular amostras com $k = 15$ repetições por tratamentos. Ao analisar as heterogeneidades a_3 e a_5, o teste de Cochran ultrapassou o limiar da taxa empírica de poder com $k = 8$ repetições. Nos cenários compostos por tratamentos com número de repetições diferentes, para $r_i \leq 15$ (b_{11}), o teste de Cochran não foi poderoso quando o estudo foi realizado sob as proporções de heterogeneidade a_2, a_4 e a_6. Este fato também pôde ser observado no estudo de Baydili e Siğirli (2017), onde os autores afirmaram que quando os tratamento possuíam diferentes número de repetições, para tamanhos pequenos e médios, o teste de Cochran obteve baixo poder para esses cenários. Entretanto, ao promover um aumento do número de repetições entre os tratamentos desbalanceados, tais como em b_{12} e b_{13}, fez com que o teste de Cochran fosse poderoso em todas as proporções de heterogeneidade simuladas.

Tabela 4 - Resumo dos resultados do estudo de simulação para a taxa empírica de erro do tipo I e taxa empírica do poder para o teste paramétrico de Cochran. As classificações são: C, para um teste conservador; E, para um teste exato; L, para um teste liberal; S, para um teste poderoso; e, N, para um teste não-poderoso. As proporções de desbalanceamento são: b_1, $4:4:4:4:4$; b_2, $6:6:6:6:6$; b_3, $8:8:8:8:8$; b_4, $10:10:10:10:10$; b_5, $15:15:15:15:15$; b_6, $25:25:25:25:25$; b_7, $45:45:45:45:45$; b_8, $90:90:90:90:90$; b_9, $140:140:140:140:140$; b_{10}, $200:200:200:200:200$; b_{11}, $4:6:8:10:15$; b_{12}, $15:25:45:90:140$; e b_{13}, $10:45:90:140:200$.

Classificação de acordo com a taxa empírica do erro tipo I												
Distribuição normal e variâncias homogêneas												
Proporção de desbalanceamento:												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
E	E	E	E	C	E	E	E	E	E	L	L	L
Distribuição de qui-quadrado e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L
Distribuição beta e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L

Classificação de acordo taxa empírica do poder do teste													
Distribuição normal e variâncias heterogêneas													
	Proporção de desbalanceamento:												
Proporção de heterogeneidade:	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
1:1:1:1:2 (a_2)	N	N	N	N	S	S	S	S	S	S	N	S	S
1:1:1:1:4 (a_3)	N	N	S	S	S	S	S	S	S	S	S	S	S
1:1:1:2:2 (a_4)	N	N	N	N	S	S	S	S	S	S	N	S	S
1:1:1:4:4 (a_5)	N	N	S	S	S	S	S	S	S	S	S	S	S
1:2:3:4:5 (a_6)	N	N	N	N	S	S	S	S	S	S	N	S	S

Quando verificado os resultados obtidos para o teste de Hartley, apresentados na Tabela 5, foi observado que, quando as amostras são balanceadas e normalmente distribuídas ele tende a controlar a taxa empírica do erro tipo I ($\hat{\alpha}$), atuando de forma exata mesmo quando amostras pequenas são simuladas, ou seja, para $k \geq 4$ repetições por tratamento. Um ponto importante a se considerar, assim como para o teste de Cochran, é que o teste de Hartley também sofre interferência quando amostras desbalanceadas são utilizadas e ele tende a conservar a hipótese nula como verdadeira em uma frequência maior do que a esperada para o teste ($\alpha = 0{,}05$), como pode ser observado para as proporções de desbalanceamento b_{12} e b_{13}. Desta forma, quanto maior a diferença entre o número de observações por tratamento, maior a probabilidade do teste de Hartley em inferir incorretamente sobre H_0. Resultados semelhantes foram constatados por Bhandary e Dai (2008) ao verificar que, sob

normalidade e balanceamento das amostras, o número de repetições por tratamento não teria impacto sobre o controle da taxa empírica $\hat{\alpha}$. No entanto, os autores identificaram inflação da taxa de erro $\hat{\alpha}$ à medida que a diferença entre o número de repetições por tratamento era maior. De acordo com estudo feito por Veitch e Roscoe (1974), o teste de Hartley foi conservador ao atuar em populações balanceadas com $k = 10$, com diminuição da taxa do erro tipo I à medida que o número de repetições eram maiores e, para experimentos desbalanceados, o mesmo atuou de forma liberal. Ainda se tratando da taxa empírica de erro do tipo I, tem-se que o teste de Hartley não é robusto ao atuar em populações cuja distribuição não é normal, tanto para tratamentos balanceados quanto para aqueles desbalanceados.

Analisando ainda o teste de Hartley (Tabela 5), foi observado forte relação entre o poder do teste e o número de repetições por tratamento. Por meio da taxa empírica do poder do teste ($\hat{\pi}$) observou-se que, a partir de $k = 25$ repetições em amostras balanceadas o teste de Hartley foi considerado poderoso em todos as proporções de heterogeneidade simuladas. No entanto, à medida que é adicionado um aumento à proporção de heterogeneidade, o teste fica mais poderoso em detectar diferenças significativas de variâncias quando ela existe e alcança o limiar da taxa empírica do poder com menor número de repetições por tratamento, como pode ser visto para a_3, a_5 e a_6. Todavia, quando o estudo envolve amostras desbalanceadas, o teste de Hartley obteve baixo poder quando o desbalanceamento b_{11} foi considerado em todos as proporções de heterogeneidade geradas sob a condição H_0 falsa. Contudo, teve tendência a ser poderoso quando as proporções de desbalanceamento com maiores observações foram avaliadas, b_{12} e b_{13}. Wang *et al.* (2017) encontraram que o teste de Hartley foi mais poderoso que os demais testes avaliados, o que vale a dizer, em nosso estudo, que o teste de Hartley obteve melhores taxa empíricas de poder quando comparado aos testes de Cochran e Brown-Forsythe.

Tabela 5 - Resumo dos resultados do estudo de simulação para a taxa empírica de erro do tipo I e taxa empírica do poder para o teste paramétrico de Hartley. As classificações são: C, para um teste conservador; E, para um teste exato; L, para um teste liberal; S, para um teste poderoso; e, N, para um teste não-poderoso. As proporções de desbalanceamento são: b_1, $4:4:4:4:4$; b_2, $6:6:6:6:6$; b_3, $8:8:8:8:8$; b_4, $10:10:10:10:10$; b_5, $15:15:15:15:15$; b_6, $25:25:25:25:25$; b_7, $45:45:45:45:45$; b_8, $90:90:90:90:90$; b_9, $140:140:140:140:140$; b_{10}, $200:200:200:200:200$; b_{11}, $4:6:8:10:15$; b_{12}, $15:25:45:90:140$; e b_{13}, $10:45:90:140:200$.

Classificação de acordo com a taxa empírica do erro tipo I												
Distribuição normal e variâncias homogêneas												
Proporção de desbalanceamento:												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
E	E	E	E	E	E	E	E	E	E	E	L	L
Distribuição de qui-quadrado e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L
Distribuição beta e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L
Classificação de acordo taxa empírica do poder do teste												
Distribuição normal e variâncias heterogêneas												
	Proporção de desbalanceamento:											
Proporção de heterogeneidade:	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
1:1:1:1:2 (a_2)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:1:4 (a_3)	N	N	S	S	S	S	S	S	S	S	N	S	S
1:1:1:2:2 (a_4)	N	N	N	N	S	S	S	S	S	S	N	S	N
1:1:1:4:4 (a_5)	N	S	S	S	S	S	S	S	S	S	N	S	S
1:2:3:4:5 (a_6)	N	N	S	S	S	S	S	S	S	S	N	S	S

A partir deste ponto iremos descrever sobre os resultados obtidos pelo estudo de simulação para os testes não paramétricos e, inicialmente, na Tabela 6, são apresentados os resultados para as taxas empíricas de erro do tipo I ($\hat{\alpha}$) e poder ($\hat{\pi}$) para o teste de Fligner-Killeen. Começaremos retratando os resultados obtidos para a taxa empírica do erro tipo I, $\hat{\alpha}$, em cenários gerados sob distribuição normal e homogeneidade de variâncias. Foi observado que, quando os números de repetições são menores ou iguais a 25, o teste tende a se comportar de modo conservador, exceto, para $k = 4$ em que o teste foi exato. Contudo, ao se aumentar o número de repetições por tratamento, o teste de Fligner-Killeen alcança exatidão e controla a taxa empírica de erro $\hat{\alpha}$ para os cenários simulados sob as proporções de desbalanceamento b_7, b_8, b_9 e b_{10} (ou seja, para $k \geq 45$). No entanto, ao atuar em cenários com amostras desbalanceadas, sob normalidade, o teste de Fligner-Killeen

foi conservador quando avaliou o desbalanceamento b_{11} e, exato, para b_{12} e b_{13}. Além do mais, quando a distribuição beta e/ou qui-quadrado foram simuladas, de modo geral, o teste de Fligner-Killeen demonstrou tendência a se comportar de forma liberal. Exceção a esse comportamento é registrado quando a proporção de desbalanceamento b_5 foi executada para a distribuição beta e o teste foi classificado como exato. Tais resultados são consistentes com o estudo de Conover, Johnson e Johnson (1981), em que, sob normalidade, constataram que este teste atendeu aos critérios por eles estabelecidos e atuou de forma exata nos cenários balanceados e desbalanceados. Com base ainda nesse estudo, quando o teste de Fligner-Killeen é submetido a distribuições assimétricas, ele tem taxa de erro tipo I inflada e atua de modo liberal tanto sob balanceamento quanto sob desbalanceamento das amostras.

Se tratando da taxa empírica do poder do teste ($\hat{\pi}$), percebe-se (na Tabela 6) forte relação entre o aumento do poder com o aumento do número de repetições. Assim, como nos demais testes avaliados anteriormente, é notado que quando o experimento é balanceado e o número de repetições é igual a 25, todos os testes atingem níveis satisfatórios de poder, ou seja, $\hat{\pi} \geq 0{,}80$. No entanto, percebe-se que o teste de Fligner-Killeen obteve melhores resultados relacionados ao poder quando maior grau de heterogeneidade entre variâncias estava envolvidas, ou seja, nas proporções de heterogeneidade 1:1:1:1:4 e 1:1:1:4:4. Conover, Johnson, Johnson (1981) afirmam que o teste de Fligner-Killeen tem poder relativamente alto e está entre os melhores para serem usados, com base em sua robustez e poder.

Tabela 6 - Resumo dos resultados do estudo de simulação para a taxa empírica de erro do tipo I e taxa empírica do poder para o teste não paramétrico de Fligner-Killeen. As classificações são: C, para um teste conservador; E, para um teste exato; L, para um teste liberal; S, para um teste poderoso; e, N, para um teste não-poderoso. As proporções de desbalanceamento são: b_1, 4:4:4:4:4; b_2, 6:6:6:6:6; b_3, 8:8:8:8:8; b_4, 10:10:10:10:10; b_5, 15:15:15:15:15; b_6, 25:25:25:25:25; b_7, 45:45:45:45:45; b_8, 90:90:90:90:90; b_9, 140:140:140:140:140; b_{10}, 200:200:200:200:200; b_{11}, 4:6:8:10:15; b_{12}, 15:25:45:90:140; e b_{13}, 10:45:90:140:200.

Classificação de acordo com a taxa empírica do erro tipo I												
Distribuição normal e variâncias homogêneas												
Proporção de desbalanceamento:												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
E	C	C	C	C	C	E	E	E	E	C	E	E
Distribuição de qui-quadrado e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L
Distribuição beta e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	E	L	L	L	L	L	L	L	L

Classificação de acordo taxa empírica do poder do teste													
Distribuição normal e variâncias heterogêneas													
	Proporção de desbalanceamento:												
Proporção de heterogeneidade:	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
1:1:1:1:2 (a_2)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:1:4 (a_3)	N	N	N	S	S	S	S	S	S	S	S	S	S
1:1:1:2:2 (a_4)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:4:4 (a_5)	N	N	S	S	S	S	S	S	S	S	S	S	S
1:2:3:4:5 (a_6)	N	N	N	N	S	S	S	S	S	S	N	S	S

Retratando os resultados para o teste não paramétrico de Conover, apresentados na Tabela 7, observou-se que quando os dados são normalmente distribuídos e assumem amostras balanceadas com $k \leq 15$ repetições, há uma certa irregularidade entre as classificações do teste e, para esses cenários, ocorre uma alternância entre o comportamento exato e liberal. Contudo, analisando ainda os conjuntos de dados balanceados, o teste de Conover passa a ter controle sob a taxa empírica do erro tipo I ($\hat{\alpha}$) quando as proporções de desbalanceamento b_7, b_8, b_9 e b_{10} são avaliadas, ou seja, $0{,}0375 < \hat{\alpha} < 0{,}0625$ para $25 \leq k \leq 200$. Resultados aproximados foram encontrados por Mirtagioğlu *et al.* (2017) ao constatarem que o teste manteve o erro tipo I no nível alfa com amostras de tamanho iguais a 30, no entanto, em nosso estudo, o controle do erro tipo I (exatidão) foi verificado para $k \geq 25$. Do mesmo modo que o teste de Conover teve tendência de ser liberal quando

número de repetições menores ou iguais a 15 foram adotadas, no cenário desbalanceado b_{11} ele atuou de forma liberal. Todavia, quando amostras com maior número de repetições foram usadas, como em b_{12} e b_{13}, o teste alcançou exatidão. No entanto, o teste de Conover foi classificado como liberal ao ser avaliado em populações cuja distribuição não é normal. Esse fato pode ser observado para todas as b_n proporções de desbalanceamento nas distribuições beta e qui-quadrado.

Considerando a taxa empírica do poder do teste ($\hat{\pi}$), foi observado (Tabela 7) que quando a proporção de heterogeneidade a_2 foi executada no estudo de simulação, esse foi o teste a alcançar o limiar da taxa empírica do poder com o maior número de observações, ou seja, o teste foi considerado como poderoso apenas com $k = 45$ observações por tratamento. Contudo, apesar de possuir padrão de comportamento semelhante aos demais testes, esse foi o teste não paramétrico a alcançar o limiar da taxa empírica de poder do teste com o menor número de repetições por tratamento para as proporções de heterogeneidade a_5 e a_6, isto é, com $k = 6$ e $k = 8$ repetições, respectivamente. Para os cenários desbalanceados, o teste de Conover não foi poderoso ao atuar sob as proporções a_2, a_4 e a_6, no desbalanceamento b_{11}, e para as configurações b_{12} e b_{13} obteve taxas empíricas $\hat{\pi} \geq 0{,}80$.

Tabela 7 - Resumo dos resultados do estudo de simulação para a taxa empírica de erro do tipo I e taxa empírica do poder para o teste não paramétrico de Conover. As classificações são: C, para um teste conservador; E, para um teste exato; L, para um teste liberal; S, para um teste poderoso; e, N, para um teste não-poderoso. As proporções de desbalanceamento são: b_1, $4:4:4:4:4$; b_2, $6:6:6:6:6$; b_3, $8:8:8:8:8$; b_4, $10:10:10:10:10$; b_5, $15:15:15:15:15$; b_6, $25:25:25:25:25$; b_7, $45:45:45:45:45$; b_8, $90:90:90:90:90$; b_9, $140:140:140:140:140$; b_{10}, $200:200:200:200:200$; b_{11}, $4:6:8:10:15$; b_{12}, $15:25:45:90:140$; e b_{13}, $10:45:90:140:200$.

Classificação de acordo com a taxa empírica do erro tipo I												
Distribuição normal e variâncias homogêneas												
Proporção de desbalanceamento:												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	E	E	L	L	E	E	E	E	E	L	E	E
Distribuição de qui-quadrado e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L
Distribuição beta e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
L	L	L	L	L	L	L	L	L	L	L	L	L
Classificação de acordo taxa empírica do poder do teste												
Distribuição normal e variâncias heterogêneas												
	Proporção de desbalanceamento:											
Proporção de heterogeneidade:	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
1:1:1:1:2 (a_2)	N	N	N	N	N	N	S	S	S	S	N	S	S
1:1:1:1:4 (a_3)	N	N	N	S	S	S	S	S	S	S	S	S	S
1:1:1:2:2 (a_4)	N	N	N	N	S	S	S	S	S	S	N	S	S
1:1:1:4:4 (a_5)	N	S	S	S	S	S	S	S	S	S	S	S	S
1:2:3:4:5 (a_6)	N	N	S	S	S	S	S	S	S	S	N	S	S

Considerando os resultados para o teste de Mood, apresentados na Tabela 8, percebe-se o melhor controle sobre a taxa empírica de erro do tipo I ($\hat{\alpha}$) quando comparado aos demais testes não paramétricos. No entanto, quando os dados possuem distribuição normal e o tamanho da amostra é igual ou inferior a 8 repetições (para experimentos balanceados), o teste de Mood foi conservador. Contudo, considerando as demais proporções de desbalanceamento (de b_4 até b_{13}) o teste de Mood controlou a taxa empírica $\hat{\alpha}$ dentro dos limites adaptados para o critério liberal de Bradley (1978) e foi considerado como exato nessas configurações. Esses resultados concordam com aqueles obtidos pelo estudo de simulação de Riboldi *et al.* (2014), onde, encontraram que, sob normalidade e $k = 10$, o teste de Mood foi exato. Analisando a distribuição de qui-quadrado, foi verificado que o teste de Mood foi robusto a essa distribuição em maior parte das proporções de desbalanceamento

avaliadas, exceto nos casos b_1, b_3, b_8 e b_{11} em que o teste foi conservador. Quando a distribuição beta é investigada, o teste assume resultados semelhantes aos obtidos para a distribuição de qui-quadrado e controlou a taxa empírica de erro $\hat{\alpha}$ em maior parte dos cenários, exceto para pequenas amostras (b_1 e b_2). Ainda, o teste foi classificado como conservador nos cenários simulados sob as proporções de desbalanceamento b_1, b_2 e b_{11}.

No momento em que a taxa empírica do poder do teste ($\hat{\pi}$) é considerada, foi observado (Tabela 8) comportamento padrão, semelhante àquele assumido pelos demais testes. Pôde-se constatar para menores níveis de heterogeneidades, como em a_2 e a_4, que o teste de Mood foi poderoso apenas com $k = 25$ repetições por tratamento e, quando há um aumento da heterogeneidade entre os tratamentos, como em a_5, o teste de Mood ultrapassa o limiar de poder com $k = 8$ repetições. Gorbunova e Lemeshko (2011) encontraram em seu estudo de simulação que o teste de Mood é o mais poderoso quando comparado aos testes não paramétricos de Siegel-Tukey, Ansari-Bradley, Capon e Klotz, no entanto, em nosso estudo, o teste de Mood perdeu em poder para o teste de Conover quando maior grau de heterogeneidade foi acrescido as amostras.

Tabela 8 - Resumo dos resultados do estudo de simulação para a taxa empírica de erro do tipo I e taxa empírica do poder para o teste não paramétrico de Mood. As classificações são: C, para um teste conservador; E, para um teste exato; L, para um teste liberal; S, para um teste poderoso; e, N, para um teste não-poderoso. As proporções de desbalanceamento são: b_1, $4:4:4:4:4$; b_2, $6:6:6:6:6$; b_3, $8:8:8:8:8$; b_4, $10:10:10:10:10$; b_5, $15:15:15:15:15$; b_6, $25:25:25:25:25$; b_7, $45:45:45:45:45$; b_8, $90:90:90:90:90$; b_9, $140:140:140:140:140$; b_{10}, $200:200:200:200:200$; b_{11}, $4:6:8:10:15$; b_{12}, $15:25:45:90:140$; e b_{13}, $10:45:90:140:200$.

Classificação de acordo com a taxa empírica do erro tipo I												
Distribuição normal e variâncias homogêneas												
Proporção de desbalanceamento:												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
C	C	C	E	E	E	E	E	E	E	E	E	E
Distribuição de qui-quadrado e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
C	E	C	E	E	E	E	C	E	E	C	E	E
Distribuição beta e variâncias homogêneas												
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
C	C	E	E	E	E	E	E	E	E	C	E	E

Classificação de acordo taxa empírica do poder do teste													
Distribuição normal e variâncias heterogêneas													
	Proporção de desbalanceamento:												
Proporção de heterogeneidade:	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
1:1:1:1:2 (a_2)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:1:4 (a_3)	N	N	N	S	S	S	S	S	S	S	S	S	S
1:1:1:2:2 (a_4)	N	N	N	N	N	S	S	S	S	S	N	S	S
1:1:1:4:4 (a_5)	N	N	S	S	S	S	S	S	S	S	S	S	S
1:2:3:4:5 (a_6)	N	N	N	S	S	S	S	S	S	S	N	S	S

6.2. Discussão geral

Neste tópico serão destacados os testes de homogeneidade de variâncias (HOV) que obtiveram os melhores resultados quanto aos aspectos avaliados nesse estudo de simulação: proporção de heterogeneidade, desbalanceamento e distribuição de probabilidades. Em cada um dos cenários, criados com o intuito de avaliar os testes de HOV (satisfazendo ou não suas pressuposições), foram destacados aqueles que obtiveram os melhores resultados quanto a taxa empírica do erro tipo I ($\hat{\alpha}$) e poder do teste ($\hat{\pi}$).

6.2.1. Taxa empírica do erro tipo I

Considerando os resultados referentes às taxas empíricas do erro tipo I para os testes paramétricos e não paramétricos, dispostos na seção anterior (item 6.2.1. **Taxa empírica do erro tipo I**), serão destacados aqui aqueles que tiveram os melhores

desempenho (foram exatos) em cada um dos aspectos avaliados nesse estudo quando a hipótese H_0 foi dada como verdadeira.

Quando o estudo envolveu os cenários simulados sob distribuição normal, verificou-se que o teste paramétrico de Bartlett (Tabela 1) teve o melhor desempenho e foi exato em todos os níveis de desbalanceamento investigados, isto é, ele controlou a taxa empírica de erro tipo I nos cenários balanceados e desbalanceados. Resultados semelhantes são verificados no estudo de Hatchavanich (2014) ao constatar que o teste de Batlett não foi afetado pelo tamanho da amostra quando os dados são normalmente distribuídos. Ainda sob normalidade, foi notado que os testes de Cochran (Tabela 4) e de Hartley (Tabela 5) também tiveram resultados satisfatórios quanto ao controle do erro tipo I, porém, estes testes seriam indicados apenas quando os tratamentos têm o mesmo número de repetições. Os autores Milliken e Johnson (1992) recomendam, com base nos estudos de Conover, Johnson e Johnson (1981), que quando os dados são aproximadamente normais é aconselhável utilizar o teste de Bartlett ou de Hartley, no entanto, quando houver desbalanceamento o teste de Bartlett deve ser utilizado. Dentre os testes não paramétricos, àqueles que mais se adequaram a este tipo de configuração foram os testes de Conover e Mood, sendo válido ressaltar que o teste de Mood teve melhor desempenho tanto nos cenários balanceados quanto nos desbalanceados.

Analisando os cenários simulados sob a distribuição de qui-quadrado, observou que o teste de Brown-Forsythe (Tabela 3) obteve o melhor desempenho dentre os testes paramétricos e foi robusto a violação da pressuposição de normalidade nos experimentos balanceados e desbalanceados. Dentre os testes não paramétricos, o teste de Mood (Tabela 8) foi o que teve melhor performance quanto ao controle da taxa empírica $\hat{\alpha}$ e foi robusto também ao desbalanceamento dos dados.

Resultados semelhantes são encontrados ao analisar a taxa empírica do erro tipo I ($\hat{\alpha}$) para os conjuntos de dados gerados sob a distribuição beta. Foi verificado que substituir a média pela mediana na estatística de teste de Levene tornou o teste paramétrico de Brown-Forsythe robusto a desvios da pressuposição de normalidade, conforme mencionado por Conover, Johnson e Johnson (1981), e robusto também ao desbalanceamento das amostras. De acordo com Baydili e Siğirli (2017), o teste de Brown-Forsythe obteve taxas de erro tipo I próximas do nível de significância para

todas as distribuições avaliadas em seu estudo, especialmente em grandes tamanhos de amostra . Dentre os não paramétricos, o teste de Mood obteve melhor performance sob balanceamento e desbalanceamento.

6.2.2. Taxa empírica do poder do teste

Aqui serão destacados aqueles testes que, além de ter bom desempenho quanto ao controle da taxa empírica do erro tipo I (referidos em 6.2.1.), também foram avaliados como poderosos nos cenários em que os conjuntos de dados estavam sob distribuição normal para as a_j proporções de heterogeneidade, com $j = 2, \dots, 6$.

Dentre os testes paramétricos indicados para avaliar a homogeneidade de variâncias, o teste de Bartlett recebe maior destaque por fornecer (em média) resultados mais precisos e com alto poder, seguido pelos testes de Cochran e de Hartley, que possuem poder equivalente entre si. Apesar de o teste de Brown-Forsythe ser considerado poderoso, este necessita de um maior tamanho de amostra para alcançar o limiar da taxa empírica de poder quando comparado aos demais testes paramétricos.

Ao observar a taxa empírica do poder para os testes não paramétricos, foi observado que o teste de Mood infere a respeito de H_0 com maior poder quando comparado com o teste de Conover, necessitando, assim, de um menor número de repetições por tratamento avaliar a homocedasticidade com maior poder de detecção.

Comportamento semelhante entre os testes paramétricos e não paramétricos foi observado quando ambos tiveram aumento da taxa empírica do poder do teste ($\hat{\pi}$) à medida que o número de repetições por tratamento e/ou a proporção de heterogeneidade entre as populações foram maiores. Contudo, foi constatado também, que os testes não paramétricos tendem a precisar de um maior número de repetições para serem considerados como poderosos quando comparados aos testes paramétricos.

7. Conclusão

Avaliar os testes de homogeneidade de variâncias por meio de simulação de dados foi importante pois permitiu verificar o comportamento de cada um deles sob os diferentes aspectos avaliados nesse estudo, isto é, sob os diferentes níveis de heterogeneidade, desbalanceamento e distribuição de probabilidades do conjunto de dados simulados. Diante disso, foi factível de investigar os testes de HOV que apresentaram melhor desempenho quanto a cada um desses aspectos, comparando-os com relação as taxas empíricas do erro tipo I ($\hat{\alpha}$) e as taxas empíricas do poder do teste ($\hat{\pi}$).

Os resultados obtidos nos permitiram destacar, dentre os testes paramétricos, o desempenho do teste de Bartlett, que sob normalidade foi um teste exato, preciso (isto é, os valores da taxa empírica $\hat{\alpha}$ tiveram baixa variação em torno valor nominal) e com alto poder, não sendo afetado pelo desbalanceamento dos dados. O teste de Levene teve poder próximo ao obtido pelo teste de Bartlett, porém, sob normalidade, foi exato apenas quando um maior número de repetições por tratamento foi utilizado e, liberal, para as distribuições de qui-quadrado e beta. O teste de Brown-Fosythe, que é baseado em uma modificação do teste de Levene, foi robusto a violação da suposição de normalidade e robusto também ao desbalanceamento dos dados, porém, foi o teste paramétrico menos poderoso. Os testes de Cochran e de Hartley, sob distribuição normal, tiveram comportamento semelhantes e foram classificados como exatos e poderosos quando os tratamentos possuíram o mesmo número de repetições.

Os testes não paramétricos de Conover, Fligner-Killeen e de Mood se assemelham quanto a taxa empírica do poder, porém, em média, o teste de Mood tende a ser poderoso com o menor número de repetições. Quando avaliados sob normalidade, os testes de Conover e de Mood tendem a serem exatos, porém, apenas o teste de Mood foi robusto a violação dessa suposição (sob balanceamento e desbalanceamentos das amostras).

Em síntese, quando os dados possuem distribuição normal recomenda-se utilizar o teste paramétrico de Bartlett. Caso os dados apresentem desvios da normalidade, sugere-se utilizar o teste de Brown-Forsythe, devido a sua robustez, ou o teste não paramétrico de Mood. Esses testes podem ser aplicados tanto em

amostras com o mesmo número de repetições quanto naquelas com número de repetições diferentes.

Outro ponto importante é que, na ausência de normalidade, tanto os testes paramétricos de Bartlett, Levene, Cochran e Hartley quanto os testes não paramétricos de Fligner-Killeen e de Conover foram liberais em todos os níveis de desbalanceamento e tiveram aumento da taxa empírica do erro tipo I à medida que o número de repetições por tratamento foi maior.

Como tais determinações são baseadas em conjunto de dados gerados por simulação, algumas generalizações requerem cautela. Por exemplo, o cenário a ser avaliado pode conter características diferentes dos aspectos aqui simulados, como, avaliar uma quantidade de tratamentos diferente ou outro tipo de distribuição de probabilidades dos dados. Além disso, divergências podem ser encontradas se as condições verdadeiras forem muito diferentes daquelas aqui investigadas.

8. Referências

ALGINA, J.; BLAIR, R. C.; COOMBS, W. T. A Maximum Test for Scale: Type I Error Rates and Power. **Journal of Educational and Behavioral Statistics**, v. 20, n. 1, p. 27–39, 1995.

ARSHAM, H.; LOVRIC, M. **Bartlett's Test**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

ASSIS, J. P.; SOUSA, R. P.; DIAS, S. C. T. **Glossário de Estatística**. Mossoró: EdUFERSA, 2019.

BARTLETT, M. S. Properties of Sufficiency and Statistical Tests. **Journal of the Royal Statistical Society**, v. 160, p. 268–282, 1937.

BAYDILI, K. N.; SIĞIRLI, D. Comparison of Variance Homogenity Tests for Different Distributions. **Turkiye Klinikleri Journal of Biostatistics**, v. 9, n. 3, p. 197–212, 2017.

BHANDARY, M.; DAI, H. An Alternative Test for the Equality of Variances for Several Populations When the Underlying Distributions are Normal. **Communications in Statistics: Simulation and Computation**, v. 38, n. 1, p. 109–117, 2008.

BIASE, N. G.; FERREIRA, D. F. Testes de igualdade e de comparaçõoes múltiplas para várias proporções binomiais independentes. **Revista Brasileira de Biometria**, v. 29, n. 4, p. 549–570, 2011.

BRADLEY, J. V. Robustness? **British Journal of Mathematical and Statistical Psychology**, v. 31, p. 144–152, 1978.

BROWN, M. B.; FORSYTHE, A. B. Robust Tests for the Equality of Variances. **Journal of the American Statistical Association**, v. 69, n. 346, p. 364–367, 1974.

CANTELMO, N. F.; FERREIRA, D. F. DESEMPENHO DE TESTES DE NORMALIDADE MULTIVARIADOS AVALIADO POR SIMULAÇÃO MONTE CARLO. **Ciênc. agrotec**, v. 31, n. 6, p. 1630–1636, 2007.

CASELLA, G.; BERGER, R. L. **Inferência Estatística**. 1. ed. São Paulo: Cengage Learning, 2011.

CATTELL, R. B. **Abilities: their structure, growth, and action**. Boston: Houghton Mifflin, 1971.

CHOI, P. T. Statistics for the reader: What to ask before believing the results. **Canadian Journal of Anesthesia**, v. 52, n. SUPPL, p. 1–5, 2005.

COCHRAN, W. G. The Distribution Of The Largest Of A Set Of Estimated Variances As A Fraction Of Their Total. **Annals of Eugenics**, v. 11, n. 1, p. 47–52, 1941.

COCHRAN, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. **Biometics**, v. 3, n. 1, p. 22–38, 1947.

CONOVER, W. J. **Practical Nonparametric Statistics**. 3. ed. New York: John Wiley and Sons, 1999.

CONOVER, W. J.; JOHNSON, M. E.; JOHNSON, M. M. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. **Technometrics**, v. 23, n. 4, p. 351–361, 1981.

CRIBBIE, R. A. et al. Effect of non-normality on test statistics for one-way independent groups designs. **British Journal of Mathematical and Statistical Psychology**, v. 65, n. 1, p. 56–73, 2012.

DYER, D. D.; KEATING, J. P. On the Determination of Critical Values for Bartlett's Test. **Journal of the American Statistical Association**, v. 75, n. 370, p. 313–319, 1980.

FAHOOME, G. Twenty nonparametric statistics and their large sample approximations. **Journal of Modern Applied Statistical Methods**, v. 1, n. 2, p. 248–268, 2002.

FLIGNER, M. A.; KILLEEN, T. J. Distribution-Free Two-Sample Tests for Scale. **Journal of the American Statistical Association**, v. 71, n. 353, p. 210–213, 1976.

GAGNÉ, R. M. **The conditions of learnin**. 2. ed. Canadá: Holt, Rinehart and Winston, 1970.

GLASS, G. V.; PECKHAM, P. D.; SANDERS, J. R. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. **Review of Educational Research**, v. 42, n. 3, p. 237–288, 1972.

GOLINSKI, C.; CRIBBIE, R. A. The expanding role of quantitative methodologists in advancing psychology. **Canadian Psychology**, v. 50, n. 2, p. 83–90, 2009.

GORBUNOVA, A. A.; LEMESHKO, B. Y. Application of Variance Homogeneity Tests Under Violation of Normality Assumption. 2011.

HARTLEY, H. O. The Maximum F-Ratio as a Short-Cut Test for Heterogeneity of Variance. **Biometrika**, v. 37, n. 3, p. 308–312, 1950.

HATCHAVANICH, D. a Comparison of Type I Error and Power of Bartlett's Test , Levene's Test and O'Brien's Test for Homogeneity of Variance Tests. **Southeast-Asian J. of Sciences**, v. 3, n. 2, p. 181–194, 2014.

HOEKSTRA, R.; KIERS, H. A. L.; JOHNSON, A. Are assumptions of well-known statistical techniques checked, and why (not)? **Frontiers in Psychology**, v. 3, p. 1–9, 2012.

JAN, S.-L.; SHIEH, G. Determining sample size for precise control analysis with heterogeneous variances. **Journal of Educational and Behavioral Statistics**, v. 39, n. 2, p. 91–116, 2014.

KIM, Y. J.; CRIBBIE, R. A. ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. **British Journal of Mathematical and Statistical Psychology**, v. 71, n. 1, p. 1–12, 2018.

LEE, H. B.; KATZ, G. S.; RESTORI, A. F. A Monte Carlo Study of Seven Homogeneity of Variance Tests. **Journal of Mathematics and Statistics**, v. 6, n. 3, p. 359–366, 2010.

LI, X. et al. A Comparative Study of Tests for Homogeneity of Variances with Application to DNA Methylation Data. **PLoS ONE**, v. 10, n. 12, p. 1–12, 2015.

LYNCH, M.; WALSH, B. **Genetics and Analysis of Quantitative Traits**. Inc. Sunderland, MA: Sinauer Associates, 1998.

MARTIN, C. G.; GAMES, A. P. Tests for Homogeneity of Variance: Nonnormality and Unequal Samples. **Journal of Educational Statistics**, v. 2, n. 3, p. 187–206, 1977.

MICCERI, T. The Unicorn, The Normal Curve, and Other Improbable Creatures. **Psychological Bulletin**, v. 105, n. 1, p. 156–166, 1989.

MILLIKEN, G. A.; JOHNSON, D. E. **Analysis of Messy Data Volume 1: Designed Experiments**. 1. ed. Londres: Chapman and Hall, 1992.

MIRTAGIOĞLU, H. et al. A Monte Carlo Simulation Study for Comparing Performances of Some Homogeneity of Variances Tests. **Journal of Applied Quantitative Methods**, v. 12, n. 3, 2017.

MONTGOMERY, D. C. **Design and analysis of experiments**. 4. ed. New York: John Wiley and Sons, 1997.

MOOD, A. M. On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests. **The Annals of Mathematical Statistics**, v. 25, n. 3, p. 514–522, 1954.

NOGUEIRA, D. A.; PEREIRA, G. M. Desempenho de testes para homogeneidade de variâncias em delineamentos inteiramente casualizados. **Sigmae**, v. 2, n. 1, p. 7–22, 2013.

ODIASE, J. I.; OGBONMWAN, S. M. Critical values for the mood test of equality of dispersion. **Missouri Journal of Mathematical Sciences**, v. 20, n. 1, p. 1–13, 2008.

PARRA-FRUTOS, I. The behaviour of the modified Levene's test when data are not normally distributed. **Computational Statistics**, v. 24, n. 4, p. 671–693, 2009.

PARRA-FRUTOS, I. Testing homogeneity of variances with unequal sample sizes. **Computational Statistics**, v. 28, n. 3, p. 1269–1297, 2013.

PONTES, A. C. F.; CORRENTE, J. E. COMPARAÇÕES MÚLTIPLAS NÃO-PARAMÉTRICAS PARA O DELINEAMENTO COM UM FATOR DE CLASSIFICAÇÃO SIMPLES. **Rev. Mat. Estat**, v. 19, p. 179–197, 2001.

RAMSEY, P. H. Testing Variances in Psychological and Educational Research. **Journal of Educational Statistics**, v. 19, n. 1, p. 23–42, 1994.

RAMSEY, P. H.; RAMSEY, P. P. Testing variability in the two-sample case. **Communications in Statistics: Simulation and Computation**, v. 36, n. 2, p. 233–248, 2007.

RIBOLDI, J. et al. PRECISÃO E PODER DE TESTES DE HOMOCEDASTICIDADE PARAMÉTRICOS E NÃO-PARAMÉTRICOS AVALIADOS POR SIMULAÇÃO. **Revista Brasileira de Biometria**, v. 32, n. 3, p. 334–344, 2014.

RICARDI, F. Q. Testing the hypothesis. **Medwave**, v. 11, n. 7, 2011.

SHARMA, D.; KIBRIA, B. M. G. On some test statistics for testing homogeneity of variances: a comparative study. **Journal of Statistical Computation and Simulation**, v. 83, n. 10, p. 1944–1963, 2013.

SHOEMAKER, L. H. Fixing the F test for equal variances. **American Statistician**, v. 57, n. 2, p. 105–114, 2003.

SIEGEL, S.; CASTELLAN JR., N. J. **Estatística não-Paramétrica Para Ciências do**

Comportamento. 2. ed. Porto Alegre: Artmed, 2006.

VEITCH, W. R.; ROSCOE, J. T. Homogeneity of variance: an empirical comparison of 4 statistical tests. **Journal of Experimental Education**, v. 43, n. 2, p. 73–78, 1974.

VORAPONGSATHORN, T.; TAEJAROENKUL, S.; VIWATWONGKASEM, C. A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions. **Songklanakarin J. Sci. Technol.**, v. 26, n. 4, p. 537–547, 2004.

WANG, Y. et al. Comparing the Performance of Approaches for Testing the Homogeneity of Variance Assumption in One-Factor ANOVA Models. **SAGE Journals**, v. 77, n. 2, p. 305–329, 2017.

WLUDYKA, P. S.; NELSON, P. R. An Analysis-of-Means-Type Test for Variances From Normal Populations. **Technometrics**, v. 39, n. 3, p. 274–285, 1997.