UNIVERSIDADE FEDERAL DE ALFENAS
UNIFAL-MG

HELEN MARIA PEDROSA DE OLIVEIRA

COMPARAÇÃO DE TESTES PARA A IGUALDADE DE MÉDIAS SOB HETEROCEDASTICIDADE: SIMULAÇÃO E APLICAÇÕES

ALFENAS/MG
2016

HELEN MARIA PEDROSA DE OLIVEIRA

COMPARAÇÃO DE TESTES PARA A IGUALDADE DE MÉDIAS SOB HETEROCEDASTICIDADE: SIMULAÇÃO E APLICAÇÕES

Dissertação apresentada à Universidade Federal de Alfenas, como parte dos requisitos para obtenção do título de Mestre em Estatística Aplicada e Biometria.
Área de concentração: estatística aplicada e biometria.
Orientador: Dr. Eric Batista Ferreira

ALFENAS/MG
2016

Dados Internacionais de Catalogação-na-Publicação (CIP)
Biblioteca Central da Universidade Federal de Alfenas

Oliveira, Helen Maria Pedrosa de.
Comparação de testes para a igualdade de média sob heterocedasticidade: simulação e aplicações / Helen Maria Pedrosa de Oliveira. -- Alfenas/MG, 2016.
75 f.

Orientador: Eric Batista Ferreira.
Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Alfenas, 2016.
Bibliografia.

1. Estatística. 2. Estatístico - Testes. 3. Estatistica matematica. I. Ferreira, Eric Batista. II. Título.

CDD-519

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Alfenas / UNIFAL-MG
Programa de Pós-graduação em Estatística Aplicada e Biometria
Rua Gabriel Monteiro da Silva, 700. Alfenas - MG CEP 37130-000
Fone: (35) 3299-1392 (Secretaria) (35) 3299-1121(Coordenação)
https://www.unifal-mg.edu.br/ppgeab/

HELEN MARIA PEDROSA DE OLIVEIRA

"COMPARAÇÃO DE TESTES PARA A IGUALDADE DE MÉDIAS SOB HETEROCEDASTICIDADE: SIMULAÇÃO E APLICAÇÕES"

A Banca Examinadora, abaixo assinada, aprova a Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Alfenas. Linha de Pesquisa: Modelagem Estatística e Estatística Computacional.

Aprovado em: 29 de fevereiro de 2016.

Prof. Dr. Eric Batista Ferreira
Instituição: UNIFAL-MG
Assinatura: Eric Batista Ferreira

Prof. Dr. Marcelo Silva de Oliveira
Instituição: UFLA
Assinatura:

Prof. Dr. Quintiliano Siqueira Schroden Nomelini
Instituição: UFU
Assinatura:

Dedico à minha mãe, Joana,
que fez dos meus sonhos os seus.

AGRADECIMENTOS

Agradecer é sempre um gesto de reconhecimento e gratidão à pessoas que diretamente ou indiretamente contribuíram para que este trabalho tivesse início, meio e fim. Por isso, agradeço:

À Deus, acima de todas as coisas, por ter me dado inspiração para concluir este trabalho.

À minha família, em especial, à minha mãe Joana, por ser meu maior exemplo de honestidade e força. Te amo.

À minha irmã e sobrinho, Hirlana e Emanuel, que sempre me impulsionaram a continuar caminhando.

Ao meu namorado, Francislei, pela generosidade, calma e paciência durante a minha ausência.

A todos os amigos que souberam me aconselhar nos momentos de angústia e aflição.

Ao Prof. Dr. Eric Batista Ferreira, não só pela competência e sabedoria na orientação, mas pelo carinho e amizade que demonstrou em todos os momentos de nossa jornada, sempre confiando em meu potencial.

Aos membros da banca examinadora, Prof. Dr. Marcelo Silva de Oliveira, Prof. Dr. Quintiliano Siqueira Schroden Nomelini, Prof. Dr. Denismar Alves Nogueira, Profª. Drª. Roberta Bessa Veloso Silva , pelo incentivo e colaboração.

Aos docentes e discentes, Dr. Flávio Bertin Gandara, Daniel Paula Perez Braga e Celeide Pereira, que gentilmente nos cederam os dados reais para as aplicações, enriquecendo o trabalho.

À Universidade Federal de Alfenas e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, principalmente a seus docentes e servidores.

Agradeço à CAPES, pela auxílio financeiro.

Enfim, agradeço a todos que contribuíram para que eu pudesse chegar até aqui. Contei com pessoas que fizeram dos meus sonhos os seus, e dos meus objetivos sua própria luta. Nada na vida faria sentido sem tê-los comigo.

RESUMO

Em experimentação, geralmente a comparação de várias médias é feita por meio de testes para detectar a existência de diferenças entre os tratamentos. Um dos testes mais utilizados para a comparação de médias é o teste F, no contexto da Análise da Variância. Entretanto, a credibilidade desse teste está diretamente ligada ao cumprimento de quatro pressuposições, que são: aditividade dos termos do modelo, os erros devem seguir uma distribuição normal, serem independentes e possuírem variâncias homogêneas. De acordo com alguns pesquisadores, a pressuposição que mais afeta o desempenho do teste F é a quebra da homogeneidade. Contudo, na literatura existem diversos testes alternativos ao F, quando se quebra alguma das pressuposições. O objetivo deste trabalho foi a comparação de oito testes para a igualdade de médias sob heterocedasticidade. A avaliação dos testes foi feita por meio de simulação Monte Carlo analisando as taxas de erro tipo I e poder, ao longo de cenários resultantes da combinação de diferentes números de tratamentos (5, 10, 15 e 20), repetições (3 e 20), graus de heterogeneidade (1, 2, 4, 8, 16, 32, 64, 128, 256) e erros padrões da diferença entre as médias (0,5, 1; 2; 4 e 8). De maneira geral, os testes se mostraram pouco sensíveis ao aumento da heterogeneidade da variância, o que não aconteceu com o teste de Welch. Nas condições avaliadas, os testes de melhor desempenho foram Kruskal-Wallis e o F no contexto da ANAVA, seguidos do *bootstrap* paramétrico de Krishnamoorthy, Lu e Mathew. Já os testes de pior desempenho foram o *bootstrap* não-paramétrico de Reddy, Kumar e Ramu e o *bootstrap* não-paramétrico de Zhow e Wong.

Palavras-chave: Estatística Computacional. Delineamento Experimental. Inferência.

ABSTRACT

In experimentation, usually several comparison means is made by testing to detect the existence of differences among treatments. One of the most widely used tests for comparison of averages is the F test in the context of analysis of variance. However, the credibility of this test is directly linked to compliance with four assumptions which are: additivity of the terms of the model, errors should follow a normal distribution, be independent and possess homogeneous variances. According to some researchers, the assumption that most affects test performance F is breaking the homogeneity. However, in the literature there are several alternative tests to F, when you break any of the assumptions. The objective of this study was the comparison of eight tests for equality of means under heteroskedasticity. The evaluation of the tests was made by Monte Carlo simulation analyzing the error rates of type I and power over scenarios resulting from the combination of different numbers of treatments (5, 10, 15 and 20) repeats (3 and 20) , degree of heterogeneity (1, 2, 4, 8, 16, 32, 64, 128, 256) and standard error of the difference between the mean (0.5, 1, 2, 4 and 8). In general, the tests proved insensitive to increased heterogeneity of variance, which did not occur with Welch test. The evaluated conditions, improved performance tests were Kruskal-Wallis and F in the context of ANOVA, followed by parametric bootstrap Krishnamoorthy, Lu and Mathew. Already the worst performance tests were non-parametric bootstrap Reddy, Kumar and Ramu and the non-parametric bootstrap Zhow and Wong.

Key words: Computational Statistics. Experimental Design. Inference.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em experimentação, geralmente a comparação de várias médias é feita por meio de um teste para detectar a existência de pelo menos um tratamento diferente dos demais, no qual a hipótese nula de igualdade das médias é testada contra a hipótese alternativa de que haja, pelo menos, uma média que difere das outras. Se a hipótese nula for rejeitada, então as médias dos tratamentos são categorizadas ordinalmente por meio de um teste de comparações múltiplas.

A análise de variância (ANAVA), desenvolvida por Ronald Fisher a partir da década de 1920, é uma técnica estatística que possibilita averiguar, por meio de estatísticas de somas de quadrados, se as médias dos tratamentos são estatisticamente iguais, ou não.

A comparação de médias na análise de variância é feita por meio do teste F, que é um teste paramétrico amplamente utilizado. Entretanto, um teste paramétrico é caracterizado por possuir suposições fortes, em especial sobre a distribuição de origem dos dados, que é considerada conhecida. Em particular, a utilização do teste F depende da verificação de quatro pressuposições: aditividade dos efeitos admitidos no modelo; e independência, homocedasticidade e normalidade dos erros. Entretanto, alguns testes funcionam de maneira satisfatória mesmo fora das condições ideais, isto é, mesmo não atendendo à todas as pressuposições e os testes que possuem essa característica são chamados de robustos. No universo acadêmico, o teste F divide opiniões, alguns autores o definem como robusto, entretanto outras pesquisas defendem que a credibilidade do teste F está diretamente ligada ao cumprimento das pressuposições, o que nem sempre acontece.

Nos casos em que essas pressuposições não são atendidas, a literatura apresenta como alternativa testes que foram construídos para falta a alguma pressuposição e também métodos que não exijam nenhuma condição para sua realização, como por exemplo, os testes de James e Welch, que não requerem homogeneidade na variância do erros e o método de reamostragem *bootstrap* que não possui pressuposição.

Existem diversos estudos sobre a quebra de alguma pressuposição, que apresentam testes para a igualdade de médias sob condições adversas. Entretanto, muitos desses estudos se mostram frágeis, com relação ao número de cenários e testes comparados. Desta forma, é relevante comparar o desempenho de vários testes sob mesmas condições, avaliando o desempenho dos mesmos sob diferentes números de tratamentos, repetições, erros padrões de diferença entre as médias, por exemplo. Assim, o objetivo deste trabalho é avaliar o desempenho, em termos

de erro tipo I e poder, dos oito testes mais utilizados na literatura para comparação de médias, sob heterocedascidade. Os testes selecionados para este estudo foram:

1. O teste F na Anava (1920);
2. O teste de Welch (1951);
3. O teste de James (1954);
4. O teste *bootstrap* paramétrico de Krishnamoorthy, Lu e Mathew (2007);
5. O teste *bootstrap* não-paramétrico com abordagem gráfica de Reddy, Kumar e Ramu (2010);
6. A correção do teste *bootstrap* não-paramétrico com abordagem gráfica de Reddy, Kumar e Ramu (2016);
7. O teste *bootstrap* não-paramétrico de de Zhou e Wong (2011);
8. O teste de Kruskal-Wallis (1952).

Com relação à estes testes os objetivos específicos do trabalho são:

• Avaliar a taxa de erro tipo I e o poder em diferentes cenários, resultantes da combinação de número de tratamentos, repetições, grau de heterogeneidade e número de erros padrões de diferença entre as médias.

• Eleger o(s) teste(s) com melhor desempenho, para situações de heterocedasticidade, que seja exato em termos de erro tipo I e possua máximo poder.

• Ilustrar o comportamento dos testes aplicado em situações reais.

O presente trabalho está dividido da seguinte maneira, a seção 2 tem como função apresentar um panorama geral sobre testes de significância, e em especial, testes para a comparação de médias, juntamente com as definições de erro tipo I e o poder, que são conceitos fundamentais para a avaliação do desempenho de um teste. Além disso, o capítulo ainda apresenta pesquisas no campo da Estatística Experimental, que abordam testes para a comparação de médias, com propostas alternativas à Anava, sob a quebra das pressuposições, juntamente com a descrição de todos os testes estudados neste trabalho.

Na seção 3 são relatadas as etapas cumpridas na pesquisa, com o tipo de simulação, realização da quebra da pressuposição de homogeneidade da variância dos erros, número de

tratamentos e repetições testados, *softwares* utilizados e descrição das aplicações utilizadas para ilustrar o desempenho de todos os testes.

A seção 4 é composta pela apresentação dos resultados da pesquisa, descrevendo o erro tipo I e o poder de todos os testes de acordo com cada cenário, juntamente com as recomendações de utilização dos mesmos sob a quebra da homogeneidade da variância dos erros. Além disso, também é feita a apresentação dos resultados dos testes em duas aplicações com massa de dados reais.

Por fim, a seção 5 discute as conclusões referente às contribuições da utilização de testes para a comparação de médias no contexto experimental, apresentando os melhores testes, principalmente sob heterocedasticidade.

2 REFERENCIAL TEÓRICO

Esta seção tem como principal objetivo descrever as etapas para realizar um teste Estatístico e sua a avaliação, apresentar um panorama geral sobre testes para a comparação de médias e suas pressuposições, caso hajam, e para finalizar descrever detalhadamente os oito testes alvo de estudo neste trabalho.

2.1 TESTES ESTATÍSTICOS

Segundo Oliveira et al. (2009, p. 300) "um teste estatístico verifica a validade, ou não, de hipóteses sobre a população, mediante critérios estatísticos".

Um teste estatístico deve ser construído e avaliado segundo dois critérios de desempenho, critérios este que servem para dimensionar os teste, e também classificá-los. Esses critérios são: riscos (ou probabilidade) de decisões erradas e custo da tomada da decisão.

De acordo com estes critérios, existem três tipos diferentes de testes estatísticos: testes de significância, testes mais poderosos e testes sequenciais.

Entretanto, em muitos casos, os testes de significância e os testes mais poderosos são indistintamente chamados de teste de hipóteses. Para compreender a diferença entre os três testes, antes é preciso ter conhecimento de alguns conceitos.

2.1.1 Estrutura dos testes

Geralmente, os teste têm a seguinte estrutura:

- Hipótese: existe uma hipótese principal sob julgamento, chamada de hipótese nula, representada por H_0. Se rejeita, então uma outra hipótese candidata é considerada como verdadeira, representada por H_1 e chamada de hipótese alternativa.
- Estatística de teste: é a estatística a ser aplicada na amostra para extrair a informação que esta contém sobre o parâmetro de interesse.

- Regra de Decisão: procedimento pelo qual opta-se por rejeitar ou não rejeitar a hipótese nula.

- Conclusão: é a decisão que se tomou depois da execução de um teste.

2.1.2 Erro tipo I e poder de um teste

Ao se realizar um teste estatístico é importante observar que as inferências realizadas estão sujeitas a erros, o erro tipo I e o erro tipo II, que estão relacionados ao desempenho dos testes.

Se a hipótese nula for rejeitada quando ela é de fato falsa ou aceita quando é verdadeira, nenhum erro de julgamento está sendo cometido. Caso contrário, comete-se um de dois possíveis erros. Esses erros são chamados erro tipo I e erro tipo II. O erro de se rejeitar a hipótese nula, quando esta é verdadeira, é o erro tipo I; e o erro de se aceitar a hipótese nula, quando esta é falsa, é o erro tipo II, como mostra a Tabela 1.

Tabela 1 – Representação tabular das decisões possíveis em um teste.

	A verdade na população	
Decisão tomada	H_0 verdadeira	H_0 falsa
Não rejeita-se H_0	Decisão correta	Erro tipo II
Rejeita-se H_0	Erro tipo I	Decisão correta

Fonte: Da autora.

Define-se α como a probabilidade de se cometer o erro tipo I , que também pode ser chamado de nível de significância do teste ou tamanho do teste. Já a probabilidade de se aceitar H_0, quando ela é verdadeira, que consiste em uma decisão correta, corresponde ao valor $1 - \alpha$.

Define-se β como a probabilidade de se cometer o erro tipo II. Agora quando se rejeita H_0 e ela é de fato falsa, isso também consiste em uma decisão correta com probabilidade $1 - \beta$ de acontecer, este valor recebe o nome de poder do teste.

Um teste é considerado exato se tem tamanho real igual ao tamanho nominal α. Testes que tenham um tamanho real menor que o nominal são considerados conservadores e aqueles com taxas de erros tipo I maiores que o nível nominal são liberais.

Diante dessas definições de erros e suas probabilidades é possível classificar os testes estatísticos, já mencionado anteriormente. Oliveira et al. (2009) os classifica em função de suas

características de desempenho em relação à probabilidade de se cometer o erro tipo I e II. Essa classificação é dada na Tabela 2.

Tabela 2 – Classificação dos testes estatísticos.

	Possibilidade de controlar	
Tipos de testes	Erro tipo I	Erro tipo II
Testes de significância	α fixado pelo pesquisador, normalmente pequeno (5% ou 1%).	β não é fixado e não tem garantia de ser mínimo.
Testes mais poderosos	α fixado pelo pesquisador, normalmente pequeno (5% ou 1%).	β não é fixado, mas tem garantia de ser mínimo. Porém, β mínimo não garante que β seja pequeno.
Testes sequenciais	α fixado pelo pesquisador, normalmente pequeno (5% ou 1%).	β fixado pelo pesquisador, normalmente pequeno (5% ou 1%).

Fonte: Adaptado de Oliveira et al. (2009).

Neste trabalho, os testes abordados se caracterizam como testes de significância. Vale ressaltar, que nos testes de significância somente a hipótese nula precisa ser definida, sendo que a hipótese alternativa H_1 resulta da negação de H_0.

2.2 SIMULAÇÃO MONTE CARLO

O nome Monte Carlo está relacionado com a cidade de mesmo nome, no Principado de Mônaco, onde há vários cassinos. De acordo com Bussab e Morettin (2001) este nome é originário, principalmente, em razão dos jogos de azar, decorrentes da roleta, que é um mecanismo simples para gerar números aleatórios.

A simulação é usada para servir como uma primeira avaliação de um sistema para gerar novas estratégias de análise e regras de decisões antes de se correr o risco de experimentá-las no sistema real. Além disso, Santos (2001) afirma que todo processo simulado que envolve um componente aleatório de qualquer distribuição é considerado como pertencente ao método Monte Carlo .

A simulação Monte Carlo, segundo Dachs (1988), é utilizada com os recursos e as técnicas computacionais em que amostras são geradas de acordo com determinadas distribuições teóricas conhecidas, visando estudar o comportamento de diferentes técnicas estatísticas que poderiam ser empregadas num dado problema.

2.3 REVISÃO SOBRE TESTES PARA A COMPARAÇÃO DE MÉDIAS NA EXPERIMENTAÇÃO

A detecção da existência de diferenças entre tratamentos é feita por meio de testes de comparação de médias, que verifica quanto da variação observada entre as médias é devida exclusivamente aos efeitos dos tratamentos.

Comparar efeitos de tratamentos é alvo de estudo de vários autores, além de ser altamente aplicável em várias áreas do conhecimento. Segue uma pequena amostra de como isso tem sido realizado: Lima (2008) que buscava verificar o efeito da descafeinação do café sobre a atividade antioxidante e prevenção de lesão hepática em ratos; Santos, Sousa e Silva (2011) que avaliaram o bolo de puba da mandioca elaborado com açúcar e rapadura por meio da análise sensorial; Fernandes et al. (2009) que investigou a composição em ácidos graxos e qualidade da carne de tourinhos Nelore e Canchim alimentados com dietas à base de cana-de-açúcar e dois níveis de concentrado; Janebro et al. (2008) que testaram o efeito da farinha da casca do maracujá-amarelo nos níveis glicêmicos e lipídicos de pacientes diabéticos tipo 2.

Em grande parte das pesquisas, a comparação de médias é feita por meio do teste F, no contexto da análise de variância, embora existam inúmeros testes com a mesma função. O teste F é um teste paramétrico, que é aquele que pressupõe uma distribuição de probabilidade dos dados. Entretanto, alguns testes paramétricos são conhecidos por possuírem pressuposições fortes. No caso do teste F as pressuposições são: aditividade dos termos do modelo, a independência dos erros, que também devem seguir uma distribuição normal e devem possuir variâncias homogêneas.

Alguns testes funcionam de maneira satisfatória mesmo fora das condições ideais, isto é, mesmo não atendendo à todas as pressuposições. Os testes que possuem essa característica são chamados de robustos. No universo acadêmico, o teste F divide opiniões, alguns autores o definem como robusto, entretanto outras pesquisas defendem que a credibilidade do teste F está diretamente ligada ao cumprimento das pressuposições, o que nem sempre acontece.

Nos casos em que essas pressuposições não são atendidas, alguns autores ressaltam a ocorrência de problemas. De acordo com Lima e Abreu (2000), a não aditividade dos efeitos implica na falta de homogeneidade dos erros e, além disso, há perda de precisão do experimento, porque o erro experimental é acrescido do componente de não aditividade.

Já Vieira (2006) afirma que a dependência dos erros resulta na perda de validade dos

testes de significância, pois compromete inferências acerca das médias e provoca aumento do nível de significância.

Ferreira, Rocha e Mequelino (2012) afirmam que a não normalidade dos erros afeta a eficiência na estimação dos efeitos de tratamentos e implica em perda do poder dos testes e, além disso, há aumento do erro no nível de significância dos testes. Já Cochran (1947) afirma que é possível que a não normalidade seja acompanhada de menor eficiência na estimação dos efeitos e haja perda correspondente de poder no teste F.

Adicionalmente, vários pesquisadores relatam que a quebra da homogeneidade da variância dos erros pode afetar o desempenho do teste F. Cochran (1947) revela que, se for efetuada a análise de variância quando a variância dos erros não forem homogêneas, será quase certo que ocorrerá perda de eficiência na estimação dos efeitos de tratamentos, e haverá, também, uma perda de sensibilidade nos testes de significância: quanto maiores forem as diferenças na variância, maiores serão estas perdas. Scheffée (1959) afirma que o teste F é robusto para a não-normalidade, contudo este não apresenta o mesmo desempenho sob heterocedasticidade associada à distribuição normal, mostrando-se liberal neste caso. Zhang (2014) complementa destacando que, nesta situação, mesmo para grandes amostras, o teste F continua aceitando H_0 mesmo quando esta hipótese não é verdadeira. Almeida, Elian e Nobre (2008) afirmam que o teste F no contexto da análise de variância com um fator para comparar médias de populações normais independentes apresenta desvios no que tange ao tamanho do teste quando os erros possuem variâncias populacionais diferentes. O teste baseado na estatística F é sensível à falta de homogeneidade de variâncias, pois, sob heterocedasticidade, o tamanho real do teste não coincide com o nível de significância fixado pelo pesquisador. Desta forma, o resultado do teste F fica bastante comprometido quando as variâncias dos erros são heterocedásticas.

Contudo, na literatura existem diversas alternativas ao teste F no contexto experimental, sob a quebra das pressuposições, como por exemplo: as transformações de dados, os teste não-paramétricos, as técnicas de reamostragem, entre outros.

Testes não paramétricos são aqueles livres da distribuição de probabilidade dos dados estudados. Entretanto, estes são caraterizados, por vezes, como menos poderosos que os testes paramétricos.

Já as transformações de dados são feitas na tentativa de tornar a variância dos erros homogêneas. Contudo, vale ressaltar que o principal problema das transformações de dados é a interpretabilidade dos resultados, já que após a aplicação de alguma transformação a interpre-

tação dos resultados também deve seguir o mesmo padrão, o que nem sempre faz sentido no contexto da pesquisa, e acada sendo negligenciado por muitos pesquisadores.

Um método da reamostragem é a técnica que retira repetidas reamostras dentro da mesma amostra, sendo que a amostra original representa a população da qual foi extraída e as reamostras tem o objetivo de amplificar a informação da amostra original. Estas reamostras podem ser extraídas com ou sem reposição. Um dos principais métodos de reamostragem é o método *bootstrap*, que utiliza a reamostragem com reposição.

O método de reamostragem *bootstrap*, que foi introduzido por Efron (1979), sendo uma técnica estatística de reamostragem utilizada em diversos contextos, fundamentando-se na ideia de que, na impossibilidade de coletar infinitas amostras da população, assumi-se a amostra única e dela retiram-se reamostras. Os métodos de *bootstrap* mais utilizados são o não-paramétrico e o paramétrico.

Bastos (2014) define de forma bem clara estes dois métodos. O método *bootstrap* não-paramétrico é utilizado quando a distribuição de probabilidade da variável aleatória é desconhecida, como mostra a Figura 1, em que m é o tamanho da amostra, B é o número de reamostras e $Y_{1i} = X_{j1}, Y_{2i} = X_{j2}, \ldots, Y_{mi} = X_{jm}$, em que i é a i-ésima reamostra de $j_1, j_2, \ldots, j_m \in \{1, \ldots, m\}$.

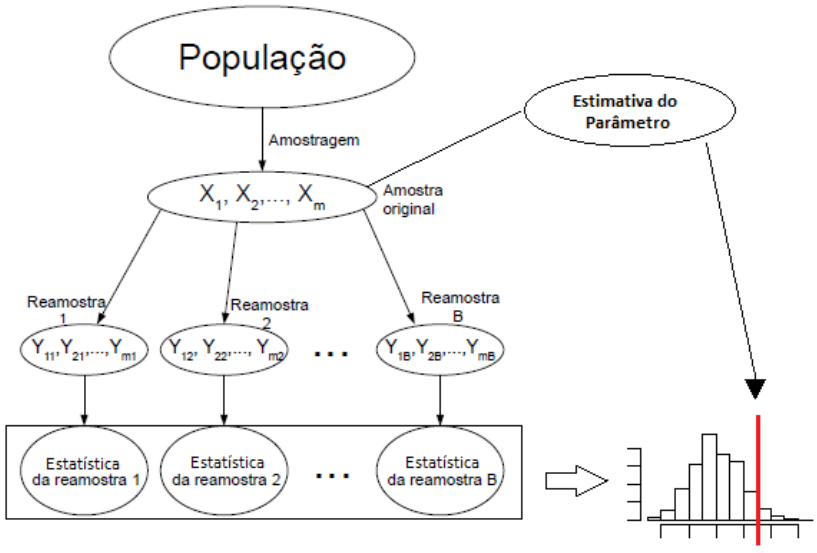

Figura 1 – Funcionamento do *bootstrap* não-paramétrico.

Fonte: Adaptado de Bastos (2014, p. 30).

O método *bootstrap* não-paramétrico consiste na reamostragem com reposição da amostra original, formando pseudoamostra de mesmo tamanho, na qual é estimado o parâmetro de

interesse. Esse processo é repetido um número grande e finito de vezes, tendo, assim, o mesmo número de estimativas. Essa série de estimativas representa uma amostra da distribuição do estimador permitindo realizar inferência sobre o parâmetro de interesse.

Já o método *bootstrap* paramétrico é utilizado quando a distribuição da variável aleatória é conhecida, com parâmetros desconhecidos, consistindo, assim, na realização de sorteios aleatórios, utilizando a estimativa do parâmetro desconhecido, obtida a partir da amostra aleatória disponível, gerando dados da distribuição de interesse, como apresentado na Figura 2.

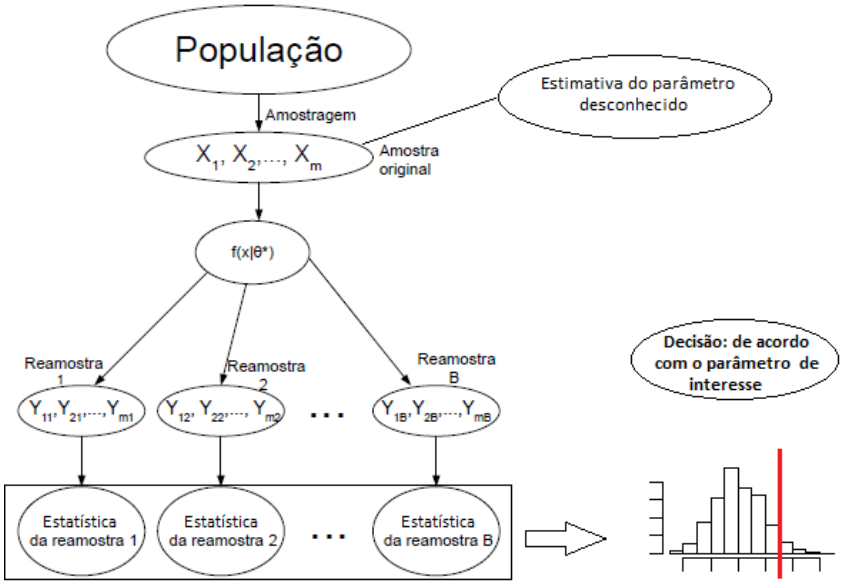

Figura 2 – Funcionamento do *bootstrap* paramétrico.

Fonte: Adaptado de Bastos (2013, p. 28).
Nota: θ^* é a estimativa de θ.

Muitos pesquisadores têm apostado na avaliação do teste F, no contexto da ANAVA, sob quebra de alguma das pressuposições, outros vêm trazendo novos testes com alternatina a ANAVA sob quebra de pressuposições. Zimmermann (1987) estudou o efeito da heterogeneidade de variância dos erros e da distribuição de probabilidade dos dados sobre o erro tipo I e o poder do teste F, no delineamento em blocos casualizados e quadrado latino. Os cenários foram gerados a partir de: distribuições probabilísticas (normal, uniforme, logística, Laplace, Weibull, exponencial e Cauchy), número de tratamentos (3,5,7,9 e 11), número de repetições (3, 4 e 7) e variância dos tratamentos (homogênea, moderadamente heterogênea - razão de $\frac{4}{1}$ entre a maior e a menor; e extremamente heterogênea - razão de $\frac{16}{1}$). Os resultados do trabalho mostraram que a não-normalidade não influenciou no desempenho do teste tanto em termos de erro tipo I, quanto em termos de poder. Entretanto, o estudo revelou que a heterogeneidade da variância

exerceu influência no poder do teste, principalmente quando associada à distribuição normal.

Krishnamoorthy, Lu e Mathew (2006) fizeram um estudo em termos de erro tipo I e poder dos testes de Welch, James e um teste *bootstrap* paramétrico proposto pelos autores (KLM) para a comparação de médias quando as variâncias do erros são desconhecidas e arbitrárias. Os teste foram avaliados por meio de 100000 simulações Monte Carlo divididos nos seguintes cenários resultantes da combinação entre cinco número de tratamentos (2, 3, 6, 10 e 20), quatro número de repetições (2, 3, 5 e 8) e as variâncias dos erros, cujo grau de heterogeneidade entre a maior e a menor é de até 100, sendo distribuídos de forma arbitrária. Já para o poder, o número de tratamentos se restringiu em apenas 3 e 10. Os estudos deles revelaram que o *bootstrap* paramétrico é o melhor entre os três testes com relação à taxa de erro tipo I. O teste *bootstrap* se comporta de forma muito satisfatória, mesmo para pequenas amostras, enquanto o teste de Welch não apresenta bom desempenho quando os tamanhos das amostras são pequenos e/ou o número de médias a serem comparadas é grande. Já o teste de James possui melhor desempenho do que de Welch porém não é tão bom quanto àquele proposto pelos autores.

Cribbie et al. (2012), também conduziu um estudo de erro tipo I e poder dos testes de Welch, James e KLM. Os cenários estudados são resultados de: números de tratamentos (3 e 20), tamanhos das amostras para experimentos balanceados e desbalanceados (variando de 19 à 40), variâncias populacionais (homogêneas, moderadamente heterogêneas - razão entre elas de 4 para 1, e extremamente heterogênea - razão entre elas de 9 para 1), médias populacionais (que variam de 0 à 0,8) e distribuição normal simétrica e assimétrica (assimetria = 1,75, curtose = 8,90 e assimetria = 6,18, curtose = 113,94). De acordo com os autores, o resultados indicaram que, embora as taxas de erro do tipo I de todos os testes foram satisfatórios sob normalidade ou assimetria moderada, entretanto os testes se mostraram liberais quando as distribuições foram extremamente assimétricas. De maneira geral, pode-se afirmar que o teste KLM, controlou a taxa de erro tipo I satisfatoriamente, independentemente do tamanho da amostra, dos valores das variâncias dos erros ou o número de médias a serem comparadas. Já com relação ao poder, os teste não apresentaram um resultado eficaz, pois estes variaram de 10% à 75% sendo que quanto mais assimétricos, menor o poder dos testes.

A pesquisa feita por Zhang (2014) propõe o estudo em termos de erro tipo I do teste KLM e do teste F para a comparação de médias sob heterocedasticidade da variância dos erros, com experimentos desbalanceados. O estudo foi feito utilizando cenários resultantes da combinação de: níveis de significância (0,01; 0,05 e 0,1), número de tratamentos (3 e 10), número

de repetições (3 e 7) e variância do erros ($\sigma^2; 0 < \sigma_i^2 \leq 1$). De acordo com o autor, quando as variâncias da população são heterogêneas, o teste F no contexto da Anava não rejeita a hipótese nula, mesmo para grandes amostras, ou seja, o teste F apresentou um comportamento conservador. O estudo revelou também que o teste KLM se mostrou competitivo ao teste F, apresentando um comportamento satisfatório em termos de erro tipo I para todos os cenários simulados.

Feir e Toothaker (1974) compararam, via simulação Monte Carlo, o poder e as taxas de erro tipo I dos testes F na análise de variância e Kruskal-Wallis. Para tal comparação, os autores simularam diversas situações resultantes das combinações entre tamanhos de amostra (28 e 68), variâncias iguais e diferentes (razão entre a maior e a menor de até 4 para 1), normalidade e não-normalidade dos dados (dados exponenciais). Segundo esses autores, o teste de Kruskal-Wallis mostrou-se competitivo com o teste F, considerando-se as taxas de erro tipo I, mas o mesmo não aconteceu com o poder do teste. Diante disso, os autores concluíram que o teste F teve melhor desempenho que o teste de Kruskal-Wallis, mesmo quando a normalidade e/ou homogeneidade das variâncias não foram satisfeitas.

Reis e Ribeiro (2007) compararam o desempenho do teste F da análise de variância e do teste de Kruskal-Wallis, para dados sob normalidade ou não, em experimentos conduzidos em delineamento inteiramente casualizado (DIC) e delineamento em blocos casualizados (DBC). Para o estudo de simulação foram adotados: número de tratamentos (5), número de repetições (5, 10 e 25) e distribuições (normal, lognormal e binomial). Segundo os autores, o teste F, tanto para o DIC quanto para o DBC, apresentou poder empírico maior que o poder do teste não paramétrico e ainda controlou as taxas de erro tipo I, em todas as situações simuladas. Dessa forma, os autores concluíram que não há necessidade de substituir o teste F da análise de variância pelos seus respectivos competidores não-paramétricos, mesmo quando a pressuposição de normalidade não é satisfeita.

Já no trabalho realizado por Ferreira, Mequelino e Rocha (2012) o objetivo foi comparar o desempenho do teste F o do teste de Kruskal-Wallis sob a quebra da pressuposição de normalidade dos erros. Para o estudo de simulação foram considerados dois grupos (I e II). No grupo I, foram avaliadas as taxas de erro tipo I por meio dos cenários resultantes das combinações entre o número de tratamentos (3, 5, 10, 15, 20, 25 e 30), o número de repetições (3, 4, 5, 10, 15 e 20) e os coeficientes de variação ($1\%, 5\%, 10\%, 15\%$ e 20%). Já no grupo II, foi avaliado o poder dos testes de acordo com: número de tratamentos (3, 5, 10, 15, 20, 25 e 30), o número de repetições (3, 4, 5, 10, 15 e 20), os coeficientes de variação ($1\%, 5\%, 10\%, 15\%$ e 20%) e

os valores do fator de penalidade (1; 10; 50; 100), que funciona como pseudo-tratamentos, aumentando assim o número de tratamentos que devem estar igualmente espaçados entre 0 e 1. Os autores observaram que ambos os testes apresentaram controle da taxa de erro tipo I, embora, em muitos casos, o teste de Kruskal-Wallis mostrou-se mais conservador. Agora, o teste F se mostrou igualmente poderoso ou superior ao teste de Kruskal-Wallis em todas as situações, mesmo com a quebra da pressuposição de normalidade dos erros.

Reddy, Kuman e Ramu (2010) apresentaram um procedimento gráfico usando o método *bootstrap* não-paramétrico como uma alternativa à ANAVA para testar a hipótese de igualdade de várias médias. Segundo os autores, este procedimento não só testa a significância entre as médias, mas também identifica a fonte da heterogeneidade das mesmas.

Já Zhou e Wong (2011) trabalham com análise fatorial de dados de experimentos com microarranjos. Para isso, os autores trouxeram uma proposta de *bootstrap* não-paramétrico. Na pesquisa, os mesmos discutem o alto poder do método quando o experimento envolve múltiplos fatores, entretanto não apresentam um estudo de simulação que comprovem este resultado.

Desta forma, nota-se a grandeza e a importância desse estudo para a estatística experimental, sendo que a busca por testes robustos tem sido e continuará sendo alvo de muitas pesquisas, já que atender as pressuposições de um teste é uma tarefa que nem sempre é passível de ser feita.

2.4 TESTES DE SIGNIFICÂNCIA

Neste trabalho foi feita a comparação de oito testes de significância, descritos na sequência, a saber:

1. O teste F na Anava (1920);
2. O teste de Welch (1951);
3. O teste de James (1954);
4. O teste *bootstrap* paramétrico de Krishnamoorthy, Lu e Mathew (2007);
5. O teste *bootstrap* não-paramétrico com abordagem gráfica de Reddy, Kumar e Ramu (2010);

6. A correção do teste *bootstrap* não-paramétrico com abordagem gráfica de Reddy, Kumar e Ramu(2016);

7. O teste *bootstrap* não-paramétrico de de Zhou e Wong (2011);

8. O teste de Kruskal-Wallis (1952).

2.4.1 Análise de variância (ANAVA)

Para compreender o teste F no contexto da Anava e suas pressuposições, primeiramente é preciso conhecer a distribuição F. De acordo com Salsburg (2009), a distribuição F, apesar de receber este nome, não foi proposta por Fisher, mas por Snedecor, que o homenageou.

A variável aleatória F consiste na razão entre duas variáveis com distribuição qui-quadrado central dividido pelos seus respectivos graus de liberdade, ou seja,

$$F = \frac{U/m}{V/n} \tag{2.1}$$

onde U e V são variáveis independentes com distribuição qui-quadrado com m e n graus de liberdade, respectivamente. Por sua vez, uma variável aleatória qui-quadrado central é formada pela soma de quadrados de variáveis normais padrão independentes, ou seja,

$$U = \sum_{i=1}^{I} e_i^2 \tag{2.2}$$

onde e_i são variáveis normais padrão independentes.

- Teste F na ANAVA e suas pressuposições

No Delineamento Inteiramente Casualizado (DIC) para experimentos balanceados as observações são expressas pelo seguinte modelo linear:

$$y_{ij} = \mu + \tau_i + e_{ij} \tag{2.3}$$

em que y_{ij} é o valor da parcela que recebeu o tratamento i na repetição j, com $i = 1, \ldots, I$ e $j = 1, \ldots, J$; μ é uma constante comum a todas as parcelas (geralmente, o valor da média

geral do experimento); τ_i é o efeito do i-ésimo tratamento, e_{ij} é o erro aleatório associado à observação y_{ij}.

De acordo com o modelo (2.3), os erros e_{ij} podem ser escritos da seguinte maneira:

$$e_{ij} = y_{ij} - \mu - \tau_i$$

De posse de y_{ij} e estimando μ e τ_i, é possível predizer esses erros. Essas predições recebem o nome de resíduos.

O modelo expresso pela equação (2.3) pode ser fixo ou aleatório. Um modelo é dito aleatório quando seu único efeito fixo é a média (μ), ou seja, exceto a média, todos os efeitos admitidos no modelo seguem uma distribuição de probabilidade. Se o modelo contiver apenas efeitos fixos, com exceção do erro, é denominado modelo fixo. Quando o modelo contém efeitos aleatórios e fixos, este é chamado de modelo misto.

Nesse trabalho foi adotado o modelo fixo. Observe que neste tipo de modelo, se os dados de um experimento seguem alguma determinada distribuição então os erros obrigatoriamente também seguem a mesma distribuição, já que a média e os efeitos de tratamentos são constantes; o que não é necessariamente verdade no modelo misto e no modelo aleatório. Esta é uma informação relevante quanto se diz respeito ao cumprimento das pressuposições.

A Anava é amplamente utilizada, porém a credibilidade de seus resultados depende da verificação de quatro pressuposições:

1. Os erros devem seguir uma distribuição normal;

2. Os erros devem ser independentes;

3. Os termos do modelo devem ser aditivos;

4. A variância dos erros deve ser homogênea.

De acordo com o modelo adotado e as condições de existência de uma distribuição F, descritas nas equações (2.1) e (2.2), podem-se fazer duas afirmações sobre os erros experimentais $e_{i,j}$: seguem assumidamente uma distribuição normal e são independentes, já que os dados do experimento são normais e independentes.

A terceira pressuposição é a aditividade dos termos do modelo (2.3). A não aditividade desses termos implica na falta de homogeneidade dos mesmos. Além disso, de acordo

com Lima e Abreu (2000), há perda de precisão, porque o erro experimental é acrescido do componente de não aditividade.

A seguir foi apresentado o processo para se obter a análise de variância adotando o delineamento inteiramente casualizado juntamente com a descrição da quarta pressuposição do teste F na análise de variância.

Considere um experimento com I tratamentos e J repetições por tratamento. Denomina-se Y_{ij} a variável resposta em estudo e y_{ij} os dados observados, em que i é referente ao tratamento e j à repetição. Vale ressaltar que ao longo de todo o trabalho, esta notação adotada para a Anava foi padronizada para os demais testes.

Os dados provenientes do experimento podem ser dispostos em uma tabela, como ilustrado na Tabela 3.

Tabela 3 – Disposição dos dados de um experimento em formato de tabela.

	Repetições				
Tratamentos	1	2	...	J	Totais dos tratamentos
1	y_{11}	y_{12}	$\cdots$	y_{1J}	T_1
2	y_{21}	y_{22}	$\cdots$	y_{2J}	T_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	y_{I1}	y_{I2}	$\cdots$	y_{IJ}	T_I

Fonte: Da autora.

Para a realização da análise de variância é preciso obter os graus de liberdade, as somas de quadrado, os quadrados médios e a estatística de teste. Para isto vamos utilizar os dados da Tabela 3.

1. **Graus de Liberdade**

   - Graus de liberdade de tratamentos

   $$gl_{trat} = I - 1$$

   - Graus de liberdade total

   $$gl_{total} = IJ - 1$$

   - Graus de liberdade de resíduos

   $$gl_{res} = (IJ - 1) - (I - 1) = IJ - I = I(J - 1)$$

2. **Soma de Quadrados**

   - Soma de quadrados de tratamentos

$$SQ_{trat} = \frac{\sum_{i=1}^{I} T_i^2}{J} - \frac{\left(\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}\right)^2}{IJ}$$

   - Soma de quadrados total

$$SQ_{total} = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}^2 - \frac{\left(\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}\right)^2}{IJ}$$

   - Soma de quadrados de resíduos

$$SQ_{res} = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}^2 - \frac{\sum_{i=1}^{I} T_i^2}{J}$$

3. **Quadrados Médios**

   - Quadrados médios de tratamentos

$$QM_{trat} = \frac{SQ_{trat}}{gl_{trat}}$$

   - Quadrados médios de resíduos

$$QM_{res} = \frac{SQ_{res}}{gl_{res}}$$

4. **Estatística de Teste**

$$F_c = \frac{QM_{trat}}{QM_{res}} \tag{2.4}$$

Observe que, no contexto experimental, o denominador da estatística de teste, ou seja, o QM_{res}, que é o erro experimental é formado pela composição dos erros de todas as parcelas, e por isso, deve-se admitir que eles vêm de populações com variância comum e é isso que possibilita escrevê-los como um único erro experimental, caracterizando a última pressuposição.

De acordo com a Figura 3, a estatística de teste F_c é uma razão de duas qui-quadrado centrais, sobre seus graus respectivos graus de liberdade. Desta forma, sob H_0, F_c segue uma distribuição F central com $I-1$ e $I(J-1)$ graus de liberdade. Contudo, de acordo com Montgomery (2000), sob H_1, temos $N(\mu_1,1)^2 + N(\mu_2,1)^2 + \ldots + N(\mu_I,1)^2$, o que resulta em uma variável qui-quadrado não central, e neste caso, a estatística de teste F_c segue uma distribuição F não central.

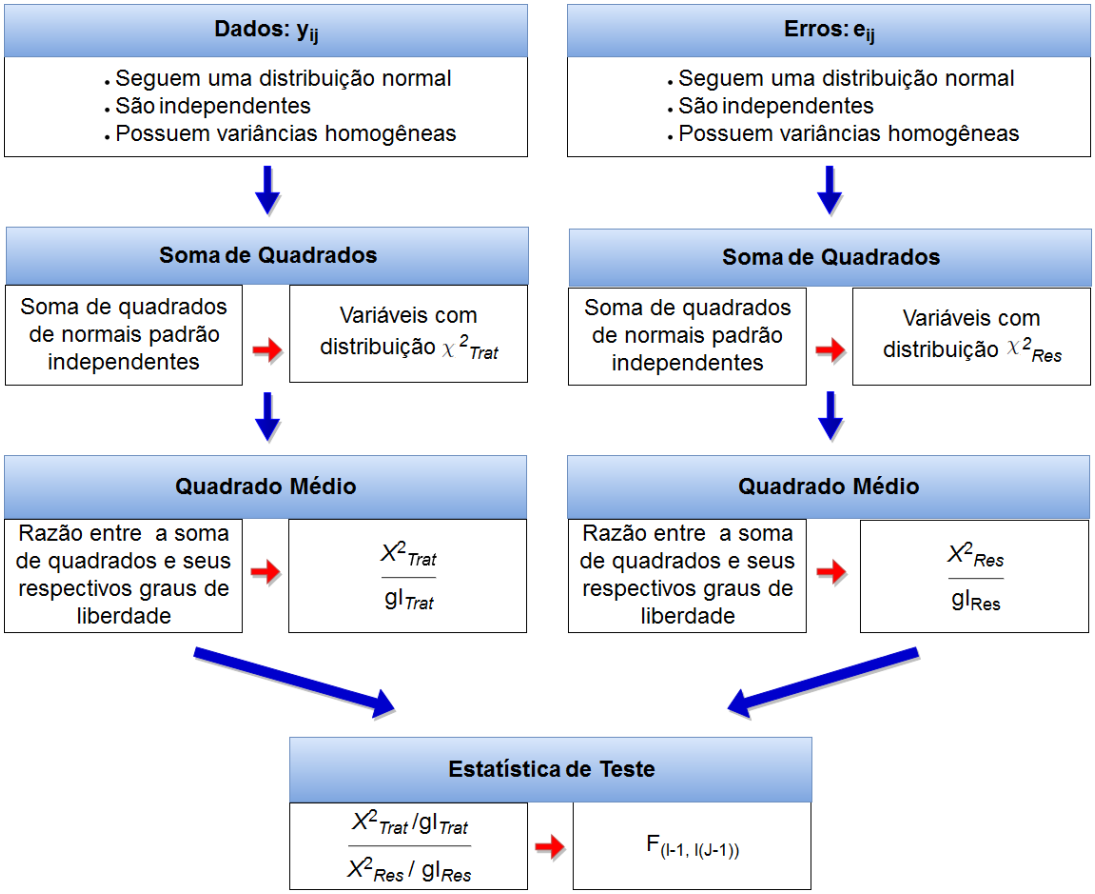

Figura 3 – Processo da análise de variância.
Fonte: Da autora.

Com as equações descritas, é possível obter a Tabela 4, com a sistematização da análise de variância.

Tabela 4 – Sistematização da análise de variância.

Fontes de variação	gl	SQ	QM	Estatística de Teste
Tratamentos	gl_{trat}	SQ_{trat}	QM_{trat}	F_c
Resíduos	gl_{res}	SQ_{res}	QM_{res}	
Total	gl_{total}	SQ_{total}		

Fonte: Da autora.

É importante ressaltar que

$$E[QM_{trat}] = \sigma^2_{erro} + \sigma^2_{trat}, \text{ onde } \sigma^2_{trat} = \frac{J\sum_{i=1}^{I}\tau_i^2}{I-1} \tag{2.5}$$

$$E[QM_{res}] = \sigma^2_{erro} \tag{2.6}$$

Observe que contexto experimental, $\sigma^2_{erro}+\sigma^2_{trat} \geq \sigma^2_{erro}$, pois a menor variação possível é a variação ao acaso gerada por variáveis não-controladas, ou seja, observe na equação (2.4), que neste âmbito o teste F é unilateral, o que não acontece necessariamente fora deste contexto.

$$\begin{cases} H_0 : \dfrac{\sigma^2_{erro} + \sigma^2_{trat}}{\sigma^2_{erro}} = 1 \iff \sigma^2_{trat} = 0 \\ H_1 : \dfrac{\sigma^2_{erro} + \sigma^2_{trat}}{\sigma^2_{erro}} > 1 \iff \sigma^2_{trat} > 0 \end{cases} \tag{2.7}$$

É importante ressaltar que, de acordo com Vieira (2006), a ANAVA, embora exija o cálculo de variâncias, na realidade compara médias de tratamentos. Observe que se $\sigma^2_{trat} = 0$ com $I \neq 1$ e $J \neq 0$ então $\frac{J\sum_{i=1}^{I}\tau_i^2}{I-1} = 0$ se e, somente se, $\tau_1 = \ldots = \tau_I = 0$, e assim, $\mu_1 = \mu_2 = \ldots = \mu_I$. Portanto o par de hipóteses formalizado no teste F da análise de variância é dado por:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases} \tag{2.8}$$

De acordo com o modelo adotado (2.3), $\mu_i = \mu + \tau_i$. Assim, as equações (2.7) e (2.8) também pode ser escritas da seguinte forma, levando em consideração que μ é uma constante comum à todas as observações:

$$\begin{cases} H_0 : \tau_1 = \ldots = \tau_I = 0 \\ H_1 : \tau_i \neq \tau_k \text{ para algum } i \neq k \end{cases} \tag{2.9}$$

Sendo F_c obtido na análise de variância e $F_{(\alpha,\nu_1,\nu_2)}$ um determinado quantil, em que

α é o nível de significância adotado pelo pesquisador e ν_1, ν_2 são os graus de liberdade do teste F, neste caso, $(I-1)$ e $I(J-1)$, respectivamente, a regra de decisão desse teste é: se $F_c \leq F_{(\alpha,\nu_1,\nu_2)}$ não há indícios para rejeitar H_0, isto é, as médias dos tratamentos devem ser consideradas estatisticamente iguais; caso contrário, conclui-se que pelo menos uma das médias difere das demais.

2.4.2 O teste de Welch (W)

O teste de Welch foi desenvolvido por um pesquisador de mesmo nome em 1951. Este é um teste paramétrico, pois assume que os dados seguem uma distribuição normal independente. Para este teste, o par de hipóteses é dado por:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases}$$

.

Este teste tem o objetivo de ponderar o teste F na presença de heterocedasticidade da variância dos erros.

Sejam W_i, $\bar{Y}^*$ e Ω dados por

$$W_i = \frac{J}{S_i^2}, \text{ com } i = 1, \ldots, I \tag{2.10}$$

em que S_i^2 é a variância amostral do tratamento I;

$$\bar{Y^*} = \frac{\sum_{i=1}^{I} W_i \bar{Y_i}}{\sum_{i=1}^{I} W_i}, \tag{2.11}$$

em que $\bar{Y_i}$ é a média de cada tratamento.

$$\Omega = \frac{\sum_{i=1}^{I}\left(1 - \frac{W_i}{\sum_{i=1}^{I} W_i}\right)^2}{J-1} \tag{2.12}$$

A estatística de teste proposta por Welch (1951), sob H_0 é dada por:

$$W = \frac{\frac{\sum_{i=1}^{I} W_i\left(\bar{Y}_i - \bar{Y}^*\right)^2}{(I-1)}}{1 + \frac{2(I-2)\Omega}{(I^2-1)}} \sim F_{(I-1,f)} \tag{2.13}$$

em que $f = \frac{I^2-1}{3\Omega}$. Observe que, neste caso, o autor utiliza o ajuste de Satterthwaite-Welch para determinar os graus de liberdade do teste, como pode ser visto em Satterthwaite (1946).

A regra de decisão do teste consiste em comparar o W obtido no teste de Welch com um determinado quantil $F_{(\alpha,\nu_1,\nu_2)}$ em que α é o nível de significância adotado pelo pesquisador e ν_1, ν_2 são os graus de liberdade do teste F, neste caso, segundo Welch (1951), é dada por $(I-1)$ e $\frac{I^2-1}{3\Omega}$, respectivamente. Se $W \leq F_{(\alpha,\nu_1,\nu_2)}$ não há indícios para rejeitar H_0 e conclui-se que as médias dos tratamentos são estatisticamente iguais, caso contrário, deve-se rejeitar H_0 e conclui-se que pelo menos uma das médias dos tratamentos difere das demais.

2.4.3 O teste de James (JA)

O teste de James (1954) é um teste paramétrico, pois assume que as variáveis aleatórias seguem uma distribuição normal independente e o par de hipóteses é dado pela sentença:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases}$$

.

O teste de James é definido por partes de acordo com o tamanho da amostra e neste tra-

balho abordou-se o caso em que o tamanho da amostra é grande. Nessas condições, a estatística de teste é dada por:

$$JA = \frac{\sum_{i=1}^{I} W_i(\bar{Y}_i - \bar{Y})^2}{I-1} \sim \chi^2_{I-1} \tag{2.14}$$

em que $\bar{Y}_i$ é a média de cada tratamento, $\bar{Y}$ é a média geral do experimento e W_i já foi definida na equação (2.10).

Assim, a regra de decisão entre rejeitar ou não a hipótese nula deve ser tomada comparando JA obtido na realização do teste com um determinado $\chi^2_{(\nu_1,\alpha)}$, em que α é o nível de significância e ν_1 são os graus de liberdade do teste, neste caso, segundo James (1954), $I-1$. Se $JA \leq \chi^2_{(\nu_1,\alpha)}$ não há indícios para rejeitar H_0 e conclui-se que as médias dos tratamentos são estatisticamente iguais, caso contrário, deve-se rejeitar H_0 e conclui-se que pelo menos uma das médias dos tratamentos difere das demais.

2.4.4 O teste *bootstrap* paramétrico de Krishnamoorthy, Lu e Mathew (KLM)

Este teste foi proposto por Krishnamoorthy, Lu e Mathew em 2007 para testar a igualdade de várias médias normais com variâncias desconhecidas, cujo par de hipóteses é dado por:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases}$$

.

O método *bootstrap* paramétrico é utilizado quando a distribuição da variável aleatória é conhecida, com parâmetros desconhecidos. Neste caso, o teste foi utilizado quando a distribuição da variável aleatória é normal, com média (μ) e variâncias desconhecidas. As amostras foram geradas a partir de modelos paramétricos com os parâmetros substituídos por suas estimativas e, sem perda de generalidade, adotou-se a média (μ) igual a zero.

Considerando I como o número de tratamentos e J o número de repetições, a estatística

de teste (T_{NB}) é dada por:

$$T_{NB}(Z_i, \chi^2_{J-1}; S_i^2) = \sum_{i=1}^{I} \frac{Z_i^2(J-1)}{\chi^2_{J-1}} - \frac{\left[\sum_{i=1}^{I}\left(\frac{\sqrt{J}Z_i(J-1)}{S_i\chi^2_{J-1}}\right)\right]^2}{\sum_{i=1}^{I}\left(\frac{J(J-1)}{S_i^2\chi^2_{J-1}}\right)} \tag{2.15}$$

Para um determinado $(s_i^2, \ldots, s_I^2)$ realização de $(S_i^2, \ldots, S_I^2)$ e nível α, o teste *bootstrap* paramétrico rejeita H_0 quando

$$P(T_{NB}(Z_i, \chi^2_{J-1}; s_i^2) > T_{N0}) < \alpha \tag{2.16}$$

em que T_{N0} é um valor observado de T_N, ou seja, um determinado quantil da distribuição obtido quando se aplica o conjunto de dados em (2.17).

$$T_N\left(\bar{Y_1}, \ldots, \bar{Y_I}; {S_1}^2, \ldots, {S_I}^2\right) = \sum_{i=1}^{I} \frac{J}{{S_i}^2}{\bar{Y_i}}^2 - \frac{\left[\sum_{i=1}^{I} J\frac{\bar{Y_i}}{{S_i}^2}\right]^2}{\sum_{i=1}^{I}\frac{J}{{S_i}^2}} \tag{2.17}$$

Se fixado $s_1, \ldots, s_I$ a estatística de teste não depende de nenhum parâmetro desconhecido, e por isso pode ser estimada utilizando a simulação Monte Carlo de acordo com o Algoritmo.

Algoritmo

Para um dado J, $(\bar{y_1}, \ldots, \bar{y_I})$ e $(s_1^2, \ldots, s_I^2)$:

Calcular T_N em (2.17) e chame de T_{N0}

Seja B número de reamostragens, então para $b = 1, \ldots, B$:

Gerar $Z_i \sim N(0,1)$ e $\chi^2_{J-1}, i = 1, \ldots, I$

Calcular $T_{NB}(Z_i, \chi^2_{J-1}; s_i^2)$ usando (2.15)

Se $T_{NB}(Z_i, \chi^2_{J-1}; s_i^2) > T_{N0}$, fixar $Q_b = 1$

$\left(\frac{1}{B}\right)\sum_{j=1}^{B} Q_b$ é uma estimativa do valor-p.

A regra de decisão desse teste é baseada na seguinte inequação: se o valor-p < α (nível

de significância) deve-se rejeitar H_0, isto é, pode-se concluir que pelo menos uma das médias dos tratamentos difere das demais; caso contrário, conclui-se que as médias são estatisticamente iguais.

2.4.5 O teste *bootstrap* não-paramétrico de Reddy, Kumar e Ramu (RKR)

O teste *bootstrap* não-paramétrico RKR foi proposto por Reddy, Kumar e Ramu em 2010. Neste teste, o par de hipóteses é dado pela expressão:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases}$$

.

A inferência sobre H_0 se dá por meio de um procedimento gráfico denominado cartas de controle, que é um tipo de gráfico utilizado para o acompanhamento de um processo. Este gráfico determina estatisticamente uma faixa denominada limites de controle que é delimitada pela linha superior (limite superior de controle) e uma linha inferior (limite inferior de controle), cujo objetivo é verificar, por meio do gráfico, se o processo está sob controle, que neste caso, quer dizer se as médias são consideradas iguais.

O procedimento para testar H_0 que pode ser obtido de acordo com os seguintes passos:

Reuna todos os dados do experimento em uma única amostra, denotada como a amostra conjunta $\{Z_c, c = 1, \ldots, IJ\}$.

1. Extraia as B amostras de *bootstrap* de tamanho IJ, com reposição da amostra conjunta Z_c. A b-ésima amostra de *bootstrap* de tamanho IJ é dada por

$$\{Y^*_{bi},\ i = 1, \ldots, IJ \text{ e } b = 1, 2, \ldots, B\} \tag{2.18}$$

2. Calcule a média da b-ésima amostra de *bootstrap*

$$\bar{Y_b} \ = \ \frac{1}{IJ} \sum_{i=1}^{IJ} Y^*_{bi},\ b = 1, 2, \ldots, B \tag{2.19}$$

3. Obtenha a distribuição de amostragem da média usando as B estimativas de *bootstrap* e calcule a média e o erro padrão da média

$$\bar{Y^*} = \frac{1}{B}\sum_{b=1}^{B}\bar{Y}_b \tag{2.20}$$

$$S^*_{IJ} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\bar{Y}_b - \bar{Y^*})^2} \tag{2.21}$$

4. A linha de decisão inferior (LDI) e a linha de decisão superior (LDS) para a comparação de cada um dos $\bar{y}_i$ são dados por

$$\begin{aligned} LDI &= \bar{Y^*} - Z_{\frac{\alpha}{2}} S^*_{IJ} \\ LDS &= \bar{Y^*} + Z_{\frac{\alpha}{2}} S^*_{IJ} \end{aligned} \tag{2.22}$$

em que $Z_{\frac{\alpha}{2}}$ é o valor crítico ao nível de significância α e a LDI e a LDS são, respectivamente, o limite superior de controle e o limite inferior de controle.

5. A Estatística de teste intervalar é dada por:

$$E_T = \left[\bar{Y^*} \pm Z_{\frac{\alpha}{2}} S^*_{IJ}\right] \quad \text{isto é,} \quad E_T = [LDI; LDS]$$

6. Faça o gráfico de $\bar{y}_i$ contra as linhas de decisão.

Nesse teste a regra de decisão é feita por meio de uma abordagem gráfica da seguinte forma: se qualquer um dos pontos $\bar{y}_i$ traçados encontra-se fora das respectivas linhas de decisão (fora do intervalo de confiança), H_0 é rejeitada ao nível α e conclui-se que pelo menos uma das médias dos tratamentos difere estatisticamente das demais; caso contrário, deve-se concluir que as médias são estatisticamente iguais.

2.4.6 Correção do teste *bootstrap* não-paramétrico de Reddy, Kumar e Ramu (CRKR)

No teste original de RKR as I médias amostrais computadas são comparadas com limites de controle (que nada mais são do que um intervalo de confiança) da média geral do experimento, ou seja, a distribuição nula de *bootstrap* é construída com médias vindas de IJ observações reamostradas. Porém, rejeita-se H_0 toda vez que pelo menos uma das médias amostrais $\bar{X}_i^{(J)}$ cai fora do intervalo construído com médias gerais $\bar{X}_i^{(IJ)}$.

Ora, como a variância da distribuição de $\bar{X}_i^{(IJ)}$ é muito menor que a variância da distribuição de $\bar{X}_i^{(J)}$, a probabilidade de uma média amostral cair fora dos limites de controle é muito grande.

A ideia dessa correção é tornar essa comparação mais coerente. Propõe-se aqui a distribuição nula de *bootstrap* seja construída por médias de J repetições de *bootstrap* da amostra original. Assim, após construídos os limites de controle, pode-se adotar o mesmo procedimento e compará-los com as médias amostrais.

Além disso, os limites de controle do teste proposto por Reddy, Kumar e Ramu (2010), ao adotar o nível de significância α, em tese, teriam $100(1-\alpha)\%$ de confiança. Porém, ao situar os pontos que representam as médias dos tratamentos sobre os limites de controle, implicitamente está-se admitindo a ocorrência de I eventos independentes, cuja probabilidade pode ser entendida como o produto das probabilidades de cada deles.

Assim, quando os autores afirmam que, se alguma das médias dos tratamentos cair fora dos limites de controle, H_0 deve ser rejeitada ao nível α de significância, eles estão enganados. Na verdade, supondo que os tratamentos são independentes, a probabilidade de alguma média cair fora dos limites de controle é o complementar da probabilidade de todas caírem dentro, ou seja

$$\begin{aligned} P[\text{Algum } \bar{Y}_i < LDI \text{ ou } \bar{Y}_i > LDS] &= 1 - P[LDI < \bar{Y}_i < LDS, \forall i] \\ &= 1 - (1-\alpha)(1-\alpha)\ldots(1-\alpha) \\ &= 1 - (1-\alpha)^I \\ &= \alpha^* \end{aligned}$$

Portanto, num teste desse tipo, α e α^* devem ser levados em consideração.

Nesta proposta, o par de hipóteses também é dado pela expressão:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases}$$

.

E para se decidir entre rejeitar, ou não, H_0 basta seguir o Algoritmo.

Algoritmo

Seja a amostra conjunta $\{Z_c, c = 1, \ldots, IJ\}$ e $B^* = \dfrac{B}{I}$ o número de reamostragem que foi feito no teste.

1. Extraia B^* amostras de *bootstrap* de tamanho IJ, com reposição da amostra conjunta Z_c. A b-ésima amostra de *bootstrap* é dada por

$$\{Y^*_{bij},\ i = 1, \ldots, I, j = 1, \ldots, J, \text{ e } b = 1, 2, \ldots, B^*\} \tag{2.23}$$

2. Particione a amostra de *bootstrap* em I subconjuntos de tamanho J e calcule a média de cada subconjunto.

$$\bar{Y}_{bi} = \frac{1}{J}\sum_{i=1}^{J} Y^*_{bij}, i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; \forall b \tag{2.24}$$

3. Obtenha a distribuição de amostragem da média usando as B estimativas de *bootstrap* e calcule a média e o erro padrão da média.

$$\bar{Y^*} = \frac{1}{B^* I}\sum_{b=1}^{B^*}\sum_{i=1}^{I} \bar{Y}_{bi} \tag{2.25}$$

$$S^*_J = \sqrt{\frac{1}{B^* I}\sum_{b=1}^{B^*}\sum_{i=1}^{I}(\bar{Y}_{bi} - \bar{Y}^*)^2} \tag{2.26}$$

Observe que a média tanto do teste original RKR como da correção CRKR são iguais, contudo o erro padrão da média difere nos dois testes, já que no RKR o S^*_{IJ} é resultante

de IJ observações e no CRKR o S_J^* considera-se somente J observações.

4. A linha de decisão inferior (LDI) e a linha de decisão superior (LDS) para a comparação de cada um dos $\bar{y}_i$ são dados por

$$\begin{aligned} LDI &= \bar{Y}^* - Z_{\frac{\alpha^*}{2}} S_J^* \\ LDS &= \bar{Y}^* + Z_{\frac{\alpha^*}{2}} S_J^* \end{aligned} \tag{2.27}$$

   em que $Z_{\frac{\alpha^*}{2}}$ é o valor crítico ao nível de significância α^* .

5. A Estatística de teste intervalar é dada por:

$$E_T = \left[\bar{Y}^* \pm Z_{\frac{\alpha^*}{2}} S_{IJ}^* \right] \qquad \text{isto é,} \qquad E_T = [LDI; LDS]$$

6. Faça o gráfico de $\bar{y}_i$ contra as linhas de decisão.

A regra de decisão desse teste é tomada observando a seguinte sentença: se qualquer um dos pontos $\bar{y}_i$ traçados encontra-se fora das respectivas linhas de decisão (fora do intervalo de confiança), H_0 é rejeitada ao nível α^* e conclui-se que pelo menos uma das médias difere das demais; caso contrário, não há indícios para rejeitar H_0 e deve-se concluir que as médias são estatisticamente iguais.

2.4.7 O teste *bootstrap* não-paramétrico de Zhou e Wong (ZW)

Zhou e Wong (2011) desenvolveram um método *bootstrap* não-paramétrico como alternativa à análise de variância, na presença da quebra das pressuposições. Neste teste, o par de hipóteses também é dado pela expressão:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases}$$

.

A distribuição nula F_{ZW} é simulada da seguinte maneira:

1. Seja $e_{ij} = y_{ij} - \mu - \tau_i$. Amostre $e_{ij}(i = 1, \ldots, I; j = 1, \ldots, J)$ com reposição e denote por e^*_{ij}.

2. Recomponha os dados fazendo $y^*_{ij} = \widehat{\mu} + \widehat{\tau_i} + e^*_{ij}$, em que $\widehat{\mu}, \widehat{\tau_i}$ são as estimativas de mínimos quadrados do modelo $y_{ij} = \mu + \tau_i + e_{ij}$.

3. Calcule F^*_{ZW} usando os dados y^*_{ij} por meio da mesma Estatística de teste utilizada no processo da Anava.

4. Repita os passos 1, 2 e 3 B-vezes para obter $F^{(1)*}_{ZW}, F^{(2)*}_{ZW}, \ldots, F^{(B)*}_{ZW}$.

5. Por fim, as B estimativas de F^*_{ZW} irão compor uma distribuição empírica e será comparada com a F_{ZW} amostral, que também pode ser obtido por meio da Anava.

A regra de decisão é feita da seguinte forma: se o número de estimativas *bootstrap* maiores que F_{ZW} superar α (nível de significância adotado pelo pesquisador) então deve-se rejeitar H_0 e concluir que pelo menos uma das médias difere das demais; caso contrário, as médias são consideradas estatisticamente iguais.

2.4.8 O teste de Kruskal-Wallis (KW)

De acordo com Campos (1983) o teste de Kruskal-Wallis (1952), cujas condições exigidas são bem gerais, sobretudo no que diz respeito à distribuição da população da qual a amostra foi retirada, ao contrário do teste F da análise de variância, é considerado uma extensão do teste de Wilcoxon, que compara duas amostras independentes. O teste de Kruskal-Wallis é realizado para comparar I amostras $(I > 2)$, cujas pressuposições são: observações independentes; dentro de um mesmo tratamento, todas as observações devem ser provenientes da mesma população, os tratamentos devem ter aproximadamente a mesma distribuição e as variáveis devem ser contínuas.

Este teste é baseado em postos e tem por objetivo decidir se as I amostras foram retiradas de populações diferentes, a partir da premissa de que se as amostras são provenientes da mesma população ou de populações iguais, os postos médios devem ser iguais (ou muito próximos) e

devem diferir se as amostras forem retiradas de populações diferentes, são o que afirmam Siegel e Castellan (2006).

A hipótese nula é de que todos os tratamentos foram retirados de uma mesma população ou de populações com a mesma mediana, e a hipótese alternativa é de que há pelo menos um tratamento com medianas diferente dos demais. Para o teste, os dados devem ser organizados em uma tabela, na qual cada tratamento deve representar uma linha e os dados devem ser ordenados para que as respectivas medianas possam ser encontradas, como ilustrado na Tabela 5, em que y_{ij} são os dados observados no i-ésimo tratamento e na j-ésima repetição.

Tabela 5 – Organização dos dados do experimento em formato de tabela.

	Repetições			
Tratamentos	1	2	...	J
1	y_{11}	y_{12}	$\cdots$	y_{1J}
2	y_{21}	y_{22}	$\cdots$	y_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	y_{I1}	y_{I2}	$\cdots$	y_{IJ}

Fonte: Da autora.

Os IJ dados da Tabela 5 devem ser substituídos pelos valores de seus postos, ou seja, os dados observados são classificados de acordo com sua posição relativa - em ordem crescente - e, a partir disso, são atribuídos postos correspondentes a sua classificação. De acordo com Campos (1983), nos casos em que há empates de valores, é atribuída a média dos valores de classificação que os postos receberiam se não houvesse empate.

Para a estatística de teste serão utilizados os valores dos postos e não os dados reais do experimento. Considerando-se apenas os postos, é possível encontrar a mediana de cada amostra. Desta forma, pode-se fixar o nível de significância do teste (α) e formalizar o par de hipóteses:

$$\begin{cases} H_0 : \theta_1 = \theta_2 = \ldots = \theta_I \\ H_1 : \text{pelo menos uma das medianas se difere das demais} \end{cases} \tag{2.28}$$

em que θ_i é a mediana de cada tratamento. Entretanto, como a distribuição de probabilidade dos dados é simétrica a média coincide com a mediana e, desta forma, pode-se adotar o seguinte par de hipóteses:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases}$$

.

Segundo Campos (1983), o cálculo da estatística de teste KW deve levar em conta a ocorrência de empates. Em experimentos nos quais há empates entre duas ou mais observações, é necessário fazer a correção do efeito destes. Desta forma, o valor de KW deve ser dividido por C, que é dado por:

$$C = 1 - \frac{\left[\sum_{i=1}^{g}(t_i^3 - t_i)\right]}{N^3 - N}$$

em que:

g é o número de agrupamentos de postos diferentes empatados;

t_i é o número de postos empatados no i-ésimo agrupamento;

N é o número total de observações, ou seja, $N = IJ$.

Assim, nos casos em que não ocorrem empates, o valor da estatística de teste pode ser obtido por:

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^{I} J(\bar{R}_i - \bar{R})^2$$

$$ou$$

$$KW = \left[\frac{12}{N(N+1)} \sum_{i=1}^{I} J\bar{R}_i^2\right] - 3(N+1)$$

em que:

I é o número de tratamentos;

J é o número de repetições;

N é o número total de observações;

R_i é a soma dos postos do i-ésimo tratamento;

$\bar{R}_i$ é a média dos postos do i-ésimo tratamento;

$\bar{R} = \frac{N+1}{2}$ é a média dos postos na amostra combinada.

E, considerando empates entre as observações, KW pode ser obtido da seguinte maneira:

$$KW = \frac{\left[\frac{12}{N(N+1)} \sum_{i=1}^{I} J\bar{R}_i^2\right] - 3(N+1)}{C}$$

Nos casos em que há três tratamentos e cinco observações por tratamento (totalizando, no máximo, quinze observações no experimento), há tabela própria para os valores críticos de KW, caso contrário, à medida que o número de repetições por tratamento cresce, a distribuição do valor calculado para a estatística de teste KW tende à distribuição χ^2 com $(I-1)$ graus de liberdade.

A regra de decisão desse teste é feita por meio da comparação do KW obtido no teste com um determinado $\chi^2_{(\alpha,\nu_1)}$ em que α é o nível de significância adotado pelo pesquisador e ν_1 são os graus de liberdade do teste χ^2, neste caso, $(I-1)$. Se $KW \leq \chi^2_{(\alpha,\nu_1)}$ não há indícios para rejeitar H_0 e pode-se concluir que as médias dos tratamentos são estatisticamente iguais; caso contrário, deve-se rejeitar H_0 e concluir que pelo menos uma das médias difere das demais.

3 MATERIAL E MÉTODOS

Nesta seção foram apresentados o material e os métodos utilizados para conduzir a pesquisa. Primeiramente foi descrito o estudo de simulação, composto pelo delineamento adotado, a obtenção dos dados que resultaram nos cenários simulados e os softwares utilizados neste estudo. Posteriormente, foram apresentadas duas aplicações usadas para ilustrar os testes em situações reais.

3.1 ESTUDO DE SIMULAÇÃO

Todos os testes comparados consideraram o modelo de Delineamento Inteiramente Casualizado (DIC) balanceado. Os dados oriundos desse tipo de delineamento são expressados pelo seguinte modelo linear:

$$y_{ij} = \mu + \tau_i + e_{ij} \tag{3.1}$$

em que y_{ij} é o valor da parcela que recebeu o tratamento i na repetição j; μ é uma constante comum a todas as parcelas (geralmente, o valor da média geral do experimento); τ_i é o efeito fixo do i-ésimo tratamento, com a restrição $\sum_i \tau_i = 0$; e_{ij} é o erro associado a y_{ij}, sendo que $e_{ij} \sim N(0; \sigma_i^2)$.

Como as variâncias residuais dos tratamentos (σ_i^2) podem ser diferentes, elas definem o grau de heterocedasticidade (δ), que foi imposto segundo a razão entre a maior e a menor variância dentre os tratamentos:

$$\delta = \frac{\max(\sigma_i^2)}{\min(\sigma_i^2)}, \text{ com } \delta \in \{1, 2, 4, 8, 16, 32,64,128,256\}\,.$$

Os oito testes mencionados foram avaliados via simulação Monte Carlo, computando-se suas taxas de erro tipo I e poder. Com essa avaliação, foi realizada uma comparação entre eles, verificando qual apresentou melhor desempenho na igualdade de médias sob heterocedasticidade crescente.

Para avaliação da taxa de erro tipo I, foram gerados dados segundo o modelo (3.1) sob a hipótese nula de igualdade entre as médias. Ao longo de $MC = 1000$ repetições do processo de Monte Carlo, foi contabilizada a proporção de rejeições equivocadas de H_0 para cada teste.

Para a avaliação do erro tipo I foram considerados 72 cenários, resultantes da combinação dos nove graus de heterocedasticidade, quatro números de tratamentos (5, 10, 15 e 20) e dois números de repetições (3 e 20). Para verificar a existência de diferença entre a taxa de erro tipo I praticada pelo teste e o nível nominal de significância estabelecido (5%) foi utilizado o intervalo de confiança exato para proporção, com 99% de confiança, que é dado por:

$$IC(p)_{1-\alpha}: \left[\frac{1}{1+\dfrac{(n-y+1)F_{(\frac{\alpha}{2};2(n-y+1),2y)}}{y}}; \frac{1}{1+\dfrac{(n-y)}{(y+1)F_{(\frac{\alpha}{2};2(y+1),2(n-y))}}} \right] \tag{3.2}$$

em que n é o número de ensaios, y o número de sucessos, α o nível de significância e $F_{(\frac{\alpha}{2},,,)}$ é o quantil superior da distribuição F com seus respectivos graus de liberdade. Neste caso particular, n é o número de simulações Monte Carlo e y é o números de sucessos que se deseja obter ao longo de toda simulação. Assim, adotou-se $n = 1000, y = 50$ e $\alpha = 0.01$. Logo,

$$IC(p)_{1-\alpha} = [0{,}0339; 0{,}0705]$$

Para a avaliação dos testes, toda proporção $\hat{p} \in IC(p)_{1-\alpha}$, foi considerada estatisticamente igual à 5%.

Para a avaliação do poder, foi imposta a hipótese alternativa completa, que é definida da seguinte forma: $H_1 : \tau_1 \neq \tau_2 \neq \ldots \neq \tau_I$. A distância imposta entre o menor efeito de tratamento e o maior foi igual a ϕ erros padrões, sendo $\phi \in \{0{,}5; 1; 2; 4; 8\}$. Os demais efeitos de tratamento (intermediários) foram distribuídos de maneira equidistante.

Para o poder foram gerados 360 cenários, resultantes da combinação dos nove graus de heterocedasticidade, quatro números de tratamentos (5, 10, 15 e 20), dois números de repetições (3 e 20) e dos cinco valores de ϕ.

É importante ressaltar, que para os testes *bootstrap*, foram realizadas B = 2000 reamostragens por teste.

Todas as rotinas foram programadas e executadas no software R (R CORE TEAM, 2015) e a simulação foi realizada no computador com *Microsoft Windows* 7 professional 64 bits, com a seguinte configuração: processador Intel Core I5 - 650 (3.20 GHZ, 4 MB L3 CACHE); memória 4GB DDR3 1333 MHZ e disco rígido de 500 GB e no servidor torre, com dois processadores instalados, 32GB de memória RAM. Memória cache de 15 MB, capacidade de processamento

de 12 threads simultânea. Controladora de memória integrada de três canais, compatível com DDR3 de 1333 MHz e sistema operacional Red Hat Enterprise Linux Versão 6 de 64 bit.

3.2 APLICAÇÃO 1 - ANÁLISE SENSORIAL DE QUEIJO MINAS PADRÃO

Os dados utilizados para essa ilustração são provenientes do experimento de Storti, Ferreira e Pereira (2014). O experimento foi conduzido no Laboratório de Laticínios do Departamento de Ciência dos alimentos (DCA/UFLA) e o queijo foi fabricado no Laticínio Verde Campo, em Lavras, MG. Seiscentos litros de leite pasteurizado com 3,5% de gordura, acidez entre 15 e 16 $^{\circ}D$ (Dornic), foram utilizados na fabricação dos queijos. Os três tratamentos continham as mesmas proporções dos ingredientes, sendo diferenciada na adição do prebiótico inulina: sem inulina (testemunha), adição de 2% de inulina, adição de 4% de inulina. Amostras de queijo foram coletadas nos tempos de 0, 15, 30, 45 dias de maturação e submetidas às análises sensoriais.

Para as análises sensoriais foram selecionados provadores, os quais receberam um pré-treinamento onde foi observado a disponibilidade de tempo, atenção, aptidão e responsabilidades. Dentre esses, foi recrutado um grupo de 20 pessoas para receber o treinamento. Por fim, 8 provadores que apresentaram notas para todos os tratamentos (dados balanceados) foram utilizados nessa ilustração.

Os queijos elaborados com 0 %, 2 % e 4 % de inulina foram submetidos à análise sensorial aos 0, 15, 30 e 45 dias de maturação utilizando a escala não estruturada de 9 pontos, que possibilita averiguar a intensidade do sabor, textura, aparência e cor característicos. Entretanto, neste estudo, foi feito um recorte do experimento original, utilizando-se os dados do tratamento 0% de inulina ao longo do tempo. Neste recorte foram considerados 8 provadores, que por serem treinados, foram modelados como repetições e não blocos.

Primeiramente, foi realizada a análise de variância seguida do teste de Shapiro-Wilk, para avaliar a normalidade dos erros, e do teste de Bartlett, para avaliar a homogeneidade da variância dos erros, ambos à 5% de significância. Para a avaliação da homogeneidade, escolheu-se o teste de Bartlett pois, de acordo com o estudo feito por Nogueira e Pereira (2013), este é o melhor teste para tal avaliação em experimentos conduzidos em DIC. Posteriormente, foram aplicados os oito testes em estudo (à 5% de significância) nas variáveis selecionadas, com o

objetivo de verificar quais concordam e quais discordam na decisão de rejeitar, ou não, a hipótese nula. Os testes que apresentaram melhor desempenho no estudo de simulação foram considerados mais acurados também no exemplo prático.

3.3 APLICAÇÃO 2 - EFEITO DO ESPAÇAMENTO NO DESENVOLVIMENTO DE MUDAS

Este experimento foi realizado na Escola Superior de Agricultura Luiz de Queiroz, no departamento de Ciências Biológicas e os dados foram gentilmente cedidos pelo docente Dr. Flávio Bertin Gandara e pelo discente Daniel Palma Perez Braga.

O objetivo do trabalho foi analisar o crescimento das mudas do grupo zoocóricas, cujas espécies são: Acacia polyphylla, Cariniana estrelensis, Guazuma ulmifolia e Luehea grandiflora Croton urucurana em diferentes espaçamentos, com intuito de identificar o melhor desenvolvimento delas e preenchimento de copa.

Neste estudo, os tratamentos foram os espaçamentos (50, 100 e 150 cm) de plantio entre as mudas. Como as espécies fazem parte do mesmo grupo biológico, estas foram consideradas homogêneas (repetições), totalizando trinta e oito repetições em cada espaçamento. A coleta de dados foi realizada com medições da altura da planta e da área e do volume da copa.

Para a análise dos dados, primeiramente foi realizada a análise de variância seguida do teste de Shapiro-Wilk e do teste de Bartlett, ambos à 5% de significância, verificando e selecionando qual das variáveis apresenta, respectivamente, normalidade e heterogeneidade na variância dos erros e determinando qual é o grau dessa heterogeneidade. Posteriormente, foram aplicados os oito testes em estudo (à 5% de significância) nas variáveis selecionadas, com o objetivo de verificar quais concordam e quais discordam na decisão de rejeitar, ou não, a hipótese nula. Os testes que apresentaram melhor desempenho no estudo de simulação serão considerados mais acurados também nesta aplicação.

4 RESULTADOS E DISCUSSÃO

Nesta seção foram descritos os resultados referentes ao desempenho de todos os testes em termos de erro tipo I e poder ao longo dos nove graus de heterogeneidade adotados.

4.1 ERRO TIPO I

Primeiramente, foi construído um gráfico do erro tipo I de todos os testes em relação ao grau de heterogeneidade da variância dos erros, como apresentado na Figura 4. Este gráfico revelou o comportamento dos testes sob graus crescentes de heterogeneidade.

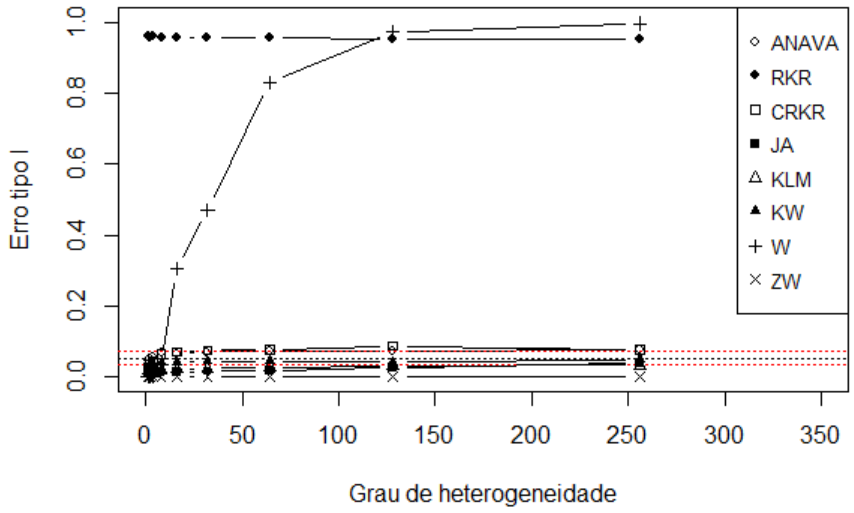

Figura 4 – Taxa do erro tipo I dos testes em relação ao grau de heterogeneidade.
Fonte: Da autora.

A Figura 4 e a Tabela 6 mostraram que o teste KW foi rigorosamente exato, praticando o nível nominal de significância adotado para todos os graus de heterogeneidade, ou seja, pode-se dizer que este teste não é sensíveis à falta de homogeneidade. Um resultado semelhante foi obtido por Feir e Toothaker (1974), que estudaram o comportamento do teste KW, sob falta de normalidade (dados exponenciais) e heterogeneidade (razão entre a maior e a menor de até 4 para 1), mostrando que este é competitivo com o teste F, considerando-se as taxas de erro tipo I.

Tabela 6 – Taxa de erro tipo I de todos os testes ao longo dos graus de heterogeneidade.

	Grau de Heterogeneidade (δ)								
Testes	1	2	4	8	16	32	64	128	256
ANAVA	0,047*	0,052*	0,056*	0,067*	0,063*	0,069*	0,073	0,074	0,074
RKR	0,960	0,957	0,961	0,957	0,954	0,958	0,956	0,952	0,954
CRKR	0,025	0,035*	0,042*	0,065*	0,070*	0,075	0,076	0,087	0,077
JA	0,011	0,012	0,010	0,012	0,011	0,015	0,018	0,026	0,039*
KLM	0,018	0,019	0,019	0,019	0,021	0,023	0,026	0,031	0,032
KW	0,031*	0,033*	0,038*	0,040*	0,041*	0,044*	0,043*	0,041*	0,048*
W	0,000	0,000	0,000	0,034*	0,303	0,471	0,829	0,973	0,996
ZW	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Fonte: Da autora.

Nota: * valores que se encontram dentro do IC exato para proporção, com 99% de confiança.

Em seguida, destaca-se o teste F, que se mostrou exato com delta variando de 1 à 32 e apresentou comportamento ligeiramente liberal para os demais graus de heterogeneidade. Montgomery (2000) concorda com esse resultado quando afirma que, mesmo sob falta de homogeneidade na variância dos erros, o desempenho do teste F é pouco afetado para o modelo misto com experimentos balanceados, como é o caso deste trabalho.

O teste KLM também se mostrou competitivo em termos de erro tipo I, apresentando um comportamento ligeiramente conservador para todo grau de heterogeneidade. Um resultado semelhante foi detectado no estudo de Zhang (2014), que revelou um comportamento satisfatório do teste KLM em termos de erro tipo I, sob quebra da homogeneidade das variâncias, para todos os cenários simulados, resultante da combinação de número de tratamentos (3 e 10), número de repetições (3 e 7) e variância do erros ($\sigma^2; 0 < \sigma_i^2 \leq 1$).

O teste de James se mostrou conservador para todos os cenário, exceto quando o grau de heterogeneidade é igual a 9, pois neste caso, o teste é exato. Já o teste de Welch revelou-se o mais sensível a quebra da pressuposição, revelando-se conservador sob homogeneidade e baixa heterogeneidade da variância dos erros e praticando taxa de erro tipo I proporcionalmente crescente ao grau de heterocedasticidade.

Observe que o teste RKR praticou mais de 90% de erro tipo I. Já a correção desse teste é pouco sensível à quebra da pressuposição de homogeneidade da variância dos erros, se mostrando um ligeiramente liberal à medida que a heterogeneidade cresce. Vale ressaltar que a correção CRKR apresentou um comportamento tão bom quanto ao teste F em relação ao erro tipo I.

Já o teste ZW, mesmo no contexto de extrema heterogeneidade, não rejeitou a hipótese

nula em nenhum dos casos.

Observe que, de maneira geral, tem-se indícios que os testes foram pouco sensíveis à falta de homogeneidade de variâncias, exceto para o teste de Welch.

Também foi realizada a análise do erro tipo I de todos os testes separadamente.

A Figura 5 e a Tabela 7 mostram o desempenho do teste de Kruskal-Wallis.

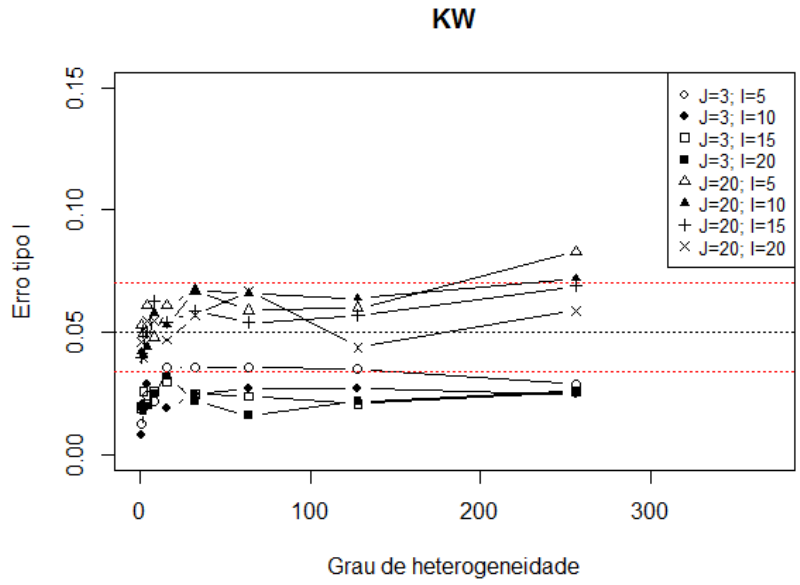

Figura 5 – Taxa do erro tipo I do teste KW ao longo do grau de heterogeneidade.
Fonte: Da autora.

Tabela 7 – Erro tipo I do teste KW com relação aos graus de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,013	0,019	0,024	0,022	0,036*	0,036*	0,036*	0,035*	0,029
	20	0,053*	0,050*	0,061*	0,048*	0,061*	0,067*	0,059*	0,060*	0,083
10	3	0,008	0,021	0,029	0,025	0,019	0,025	0,027	0,027	0,025
	20	0,042*	0,041*	0,044*	0,058*	0,053*	0,067*	0,066*	0,064*	0,072
15	3	0,019	0,026	0,021	0,026	0,030	0,025	0,024	0,021	0,026
	20	0,040*	0,050*	0,050*	0,063*	0,054*	0,059*	0,054*	0,057*	0,069*
20	3	0,020	0,018	0,020	0,025	0,032	0,022	0,016	0,022	0,026
	20	0,046*	0,040*	0,055*	0,055*	0,047*	0,057*	0,067*	0,044*	0,059*

Fonte: Da autora.
Nota: * valores que se encontram dentro do IC exato para proporção, com 99% de confiança.

Este teste é exato quando o número de repetições é pequeno com o tamanho da amostra igual a cinco e tende a ser conservador se o tamanho da amostra cresce. Já para o número de repetições igual a vinte, o teste tem desempenho exato àquele fixado inicialmente, de 5%, para

todos os tamanhos de amostra testados, exceto para a amostra de tamanho cinco, quando grau de heterogeneidade é maior que 128, tornando-se um pouco liberal. Desta forma, há de se destacar que o deste não é afetado pelos graus de heterogeneidade. Este resultado pode ser comprovado também na pesquisa de Ferreira, Mequelino e Rocha (2012) que simularam sob os cenários resultantes das combinações entre o número de tratamentos (3, 5, 10, 15, 20, 25 e 30), o número de repetições (3, 4, 5, 10, 15 e 20) e os coeficientes de variação ($1\%, 5\%, 10\%, 15\%$ e 20%). Quebrando normalidade, os autores comprovaram que o teste KW controlou da taxa de erro tipo I, tendendo a ser conservador em alguns casos.

A Figura 6 juntamente com a Tabela 8 mostraram o comportamento do teste F.

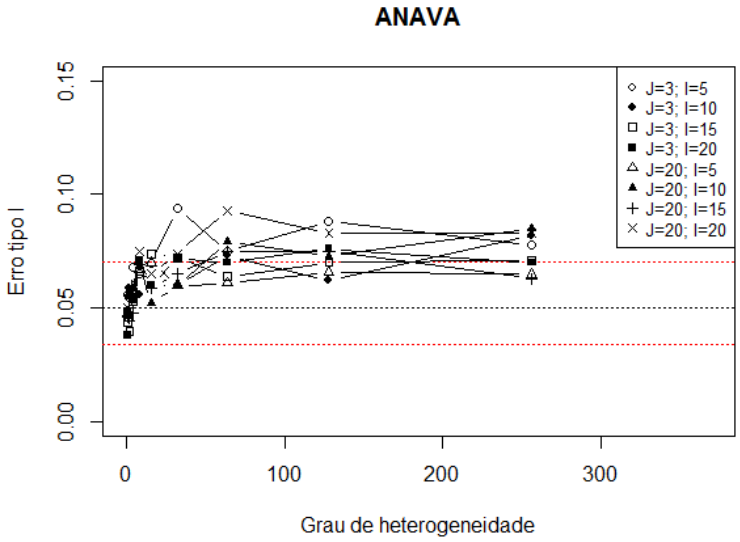

Figura 6 – Taxa do erro tipo I do teste F no contexto de ANAVA.
Fonte: Da autora.

Tabela 8 – Erro tipo I do teste F com relação aos graus de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,056*	0,055*	0,068*	0,065*	0,070*	0,094	0,075	0,088	0,078
	20	0,048*	0,046*	0,059*	0,067*	0,070*	0,060*	0,061*	0,066*	0,065*
10	3	0,049*	0,059*	0,059*	0,056*	0,060*	0,060*	0,073	0,062*	0,082
	20	0,048*	0,058*	0,053*	0,069*	0,052*	0,060*	0,079	0,073	0,085
15	3	0,044*	0,040*	0,053*	0,066*	0,074	0,072	0,064*	0,070*	0,071
	20	0,046*	0,056*	0,048*	0,069*	0,059*	0,065*	0,075	0,075	0,063*
20	3	0,038*	0,056*	0,055*	0,071	0,060*	0,072	0,070*	0,076	0,070*
	20	0,050*	0,049*	0,055*	0,075	0,065*	0,074	0,093	0,083	0,083

Fonte: Da autora.
Nota: * valores que se encontram dentro do IC exato para proporção, com 99% de confiança.

O gráfico revela que para pouca repetição, o teste F tende a ser liberal à medida que aumenta o grau de heterogeneidade, entretanto à medida em que se aumenta o tamanho da amostra, o teste apresenta comportamento exato. Agora, para o número de repetições igual a vinte, o teste F apresenta comportamento contrário ao anterior, sendo exato para as amostras de menor tamanho e liberal para grandes amostras de acordo com o crescimento da heterogeneidade.

O teste KLM, apresentado na Figura 7 e na Tabela 9 , revelam o comportamento conservador do teste em termos de erro tipo I. Entretanto, para os cenários simulados com pouca repetição e grau de heterogeneidade crescente, à medida que cresce o tamanho da amostra, o teste que era ligeiramente conservador, torna-se exato. Já para os casos que que há grande número de repetições, o teste apresenta comportamento exato e tende a ser conservador quando cresce o tamanho da amostra, mesmo para o mais alto grau de heterogeneidade, o que comprova que este teste é pouco influenciado pela quebra da pressuposição de homogeneidade da variância dos erros.

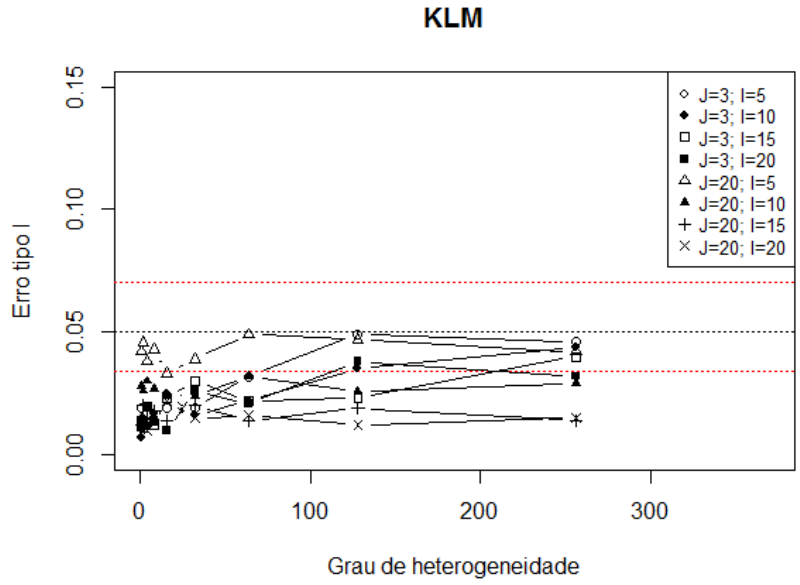

Figura 7 – Taxa do erro tipo I do teste *bootstrap* paramétrico KLM.
Fonte: Da autora.

Resultado semelhante foi obtido por Krishnamoorthy, Lu e Mathew (2006) com cenários resultantes da combinação entre cinco número de tratamentos (2, 3, 6, 10 e 20), quatro número de repetições (2, 3, 5 e 8) e as variâncias dos erros, cujo grau de heterogeneidade entre a maior e a menor é de até 100, sendo distribuídos de forma arbitrária. Eles também comprovaram o bom desempenho do teste KLM, mesmo para pequenas amostras. Assim como Cribbie et al.

(2012), que afirmou que o teste KLM, controlou a taxa de erro tipo I satisfatoriamente, independentemente do tamanho da amostra (variando de 19 à 40), dos valores das variâncias dos erros (homogêneas, moderadamente heterogêneas - razão entre elas de 4 para 1 - e extremamente heterogênea - razão entre elas de 9 para 1) ou o número de médias a serem comparadas (3 e 20).

Tabela 9 – Taxa do erro tipo I do teste KLM.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,019	0,014	0,013	0,014	0,019	0,019	0,032*	0,049*	0,046*
	20	0,042*	0,046*	0,038*	0,043*	0,033	0,039*	0,049*	0,047*	0,042*
10	3	0,007	0,015	0,012	0,017	0,025	0,016	0,022	0,035*	0,044*
	20	0,028	0,026	0,030	0,027	0,024	0,024	0,032	0,026	0,029
15	3	0,014	0,011	0,020	0,012	0,023	0,030	0,022	0,023	0,040*
	20	0,012	0,020	0,015	0,018	0,014	0,020	0,014	0,019	0,014
20	3	0,011	0,012	0,019	0,014	0,010	0,026	0,021	0,038*	0,032
	20	0,017	0,014	0,010	0,013	0,022	0,015	0,016	0,012	0,015

Fonte: Da autora.
Nota: * valores que se encontram dentro do IC exato para proporção, com 99% de confiança.

A Figura 8 e a Tabela 10 revelaram o comportamento do teste de Welch. Este teste é conservador sob a pressuposição de homogeneidade da variância dos erros e vai ser tornando cada vez mais liberal quando o grau de heterogeneidade aumenta.

Tabela 10 – Erro tipo I do teste de Welch.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,000	0,000	0,000	0,031	0,296	0,716	0,919	0,987	0,997
	20	0,000	0,000	0,000	0,000	0,000	0,029	0,422	0,814	0,973
10	3	0,000	0,000	0,000	0,057*	0,559	0,952	1,000	1,000	1,000
	20	0,000	0,000	0,000	0,000	0,000	0,032	0,646	0,988	1,000
15	3	0,000	0,000	0,001	0,093	0,746	0,989	1,000	1,000	1,000
	20	0,000	0,000	0,000	0,000	0,000	0,034*	0,777	0,997	1,000
20	3	0,000	0,000	0,001	0,095	0,827	0,996	1,000	1,000	1,000
	20	0,000	0,000	0,000	0,000	0,000	0,024	0,868	0,999	1,000

Fonte: Da autora.
Nota: * valores que se encontram dentro do IC exato para proporção, com 99% de confiança.

Vale destacar que de todos os teste abordados neste trabalho, o teste de Welch é o que apresentou mais sensibilidade aos graus crescentes de heterogeneidade de variância, sendo uma característica indesejável em um teste. Já a pesquisa feita por Krishnamoorthy, Lu e Mathew (2006) com cenários resultantes da combinação entre cinco número de tratamentos (2, 3, 6, 10

e 20) e quatro número de repetições (2, 3, 5 e 8), comprovou que o teste de Welch não apresenta bom desempenho quando os tamanhos das amostras são pequenos e/ou o número de médias a serem comparadas é grande.

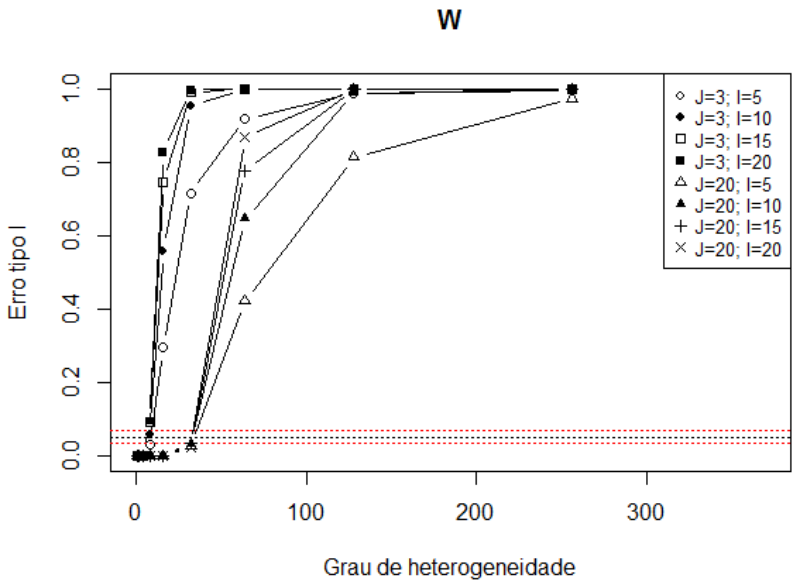

Figura 8 – Taxa do erro tipo I praticada pelo teste de Welch.
Fonte: Da autora.

O comportamento do teste de James pode ser observado na Figura 9 e na Tabela 11.

Tabela 11 – Erro tipo I do teste de James.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,046*	0,038*	0,034*	0,041*	0,046*	0,074	0,102	0,130	0,208
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10	3	0,017	0,032	0,014	0,022	0,030	0,024	0,027	0,044*	0,053*
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
15	3	0,016	0,012	0,014	0,022	0,012	0,014	0,011	0,016	0,042*
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20	3	0,011	0,014	0,020	0,018	0,007	0,012	0,008	0,020	0,011
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Fonte: Da autora.
Nota: * valores que se encontram dentro do IC exato para proporção, com 99% de confiança.

Note que o teste JA é altamente conservador, sendo que este praticou menos de 5% de erro tipo I para todos os cenários, exceto quando o número de tratamento é 5 e o número de repetições é 3, pois neste caso, ele apresenta um comportamento exato para homogeneidade e baixa heterogeneidade da variância dos erros e vai se tornando liberal à medida que o grau de

heterogeneidade cresce.

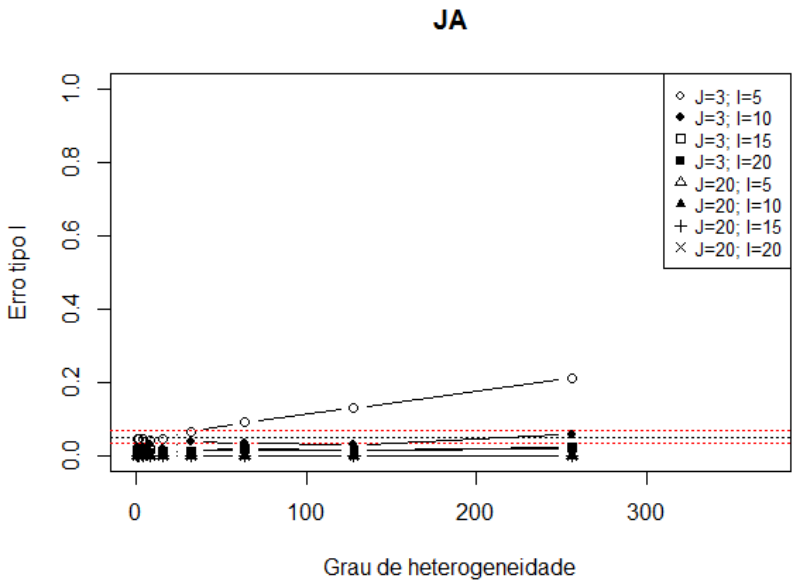

Figura 9 – Taxa do erro tipo I praticada pelo teste de James.
Fonte: Da autora.

A Figura 10 e a Tabela 12 mostram que o teste RKR é extremamente liberal, praticando mais de 75% de taxa de erro tipo I para todos os cenários simulados, independentemente do grau de heterogeneidade.

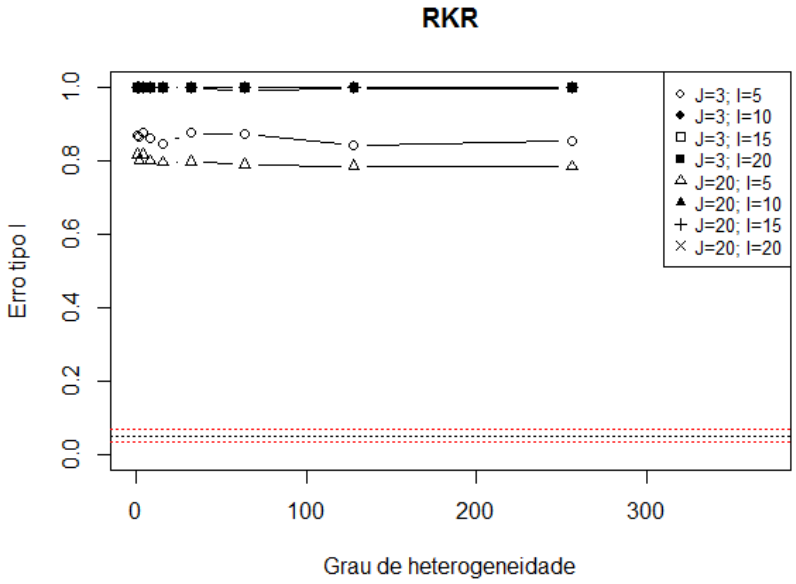

Figura 10 – Taxa do erro tipo I do teste *bootstrap* não-paramétrico RKR.
Fonte: Da autora.

Observe que a quebra da pressuposição não afetou o desempenho do teste RKR sendo

Tabela 12 – Taxa do erro tipo I do teste RKR.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,868	0,865	0,875	0,861	0,846	0,874	0,872	0,840	0,854
	20	0,815	0,800	0,817	0,799	0,795	0,796	0,788	0,784	0,783
10	3	0,999	1,000	1,000	0,999	1,000	0,999	1,000	0,999	1,000
	20	0,998	0,997	0,999	0,998	0,998	0,998	0,991	0,996	0,995
15	3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	20	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	20	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Fonte: Da autora.
Nota: * valores que se encontram dentro do IC exato para proporção, com 99% de confiança.

esta uma característica importante do mesmo, que apresenta o comportamento semelhante desde a variâncias homogêneas até o extremo grau de heterogeneidade, entretanto o teste não apresentou comportamento satisfatório.

Já o teste CRKR se mostrou conservador para o cenário com pouca repetição e tamanho de amostra igual a 5, sendo exato sob erros homogêneos e tornando-se liberal à medida em que cresce o tamanho das amostras e o grau de heterogeneidade, como pode ser visto na Figura 11 e na Tabela 13. Agora, para o número de repetições igual a vinte com tamanho da amostra igual a cinco, o teste apresenta comportamento conservador sob homogeneidade, e tende a ser exato assim que os graus de heterogeneidade crescem. Entretanto, para os demais tamanhos de amostra, o teste se mostra liberal na quebra da pressuposição. Vale ressaltar o quanto a correção proposta neste trabalho melhorou o desempenho do teste em termos de erro tipo I, basta comparar a Figura 10 com a Figura 11.

Tabela 13 – Taxa do erro tipo I do teste CRKR.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,004	0,012	0,012	0,017	0,018	0,025	0,031	0,025	0,024
	20	0,019	0,030	0,023	0,036*	0,036*	0,058*	0,043*	0,04*1	0,045*
10	3	0,021	0,029	0,040*	0,056*	0,066*	0,050*	0,059*	0,076	0,068*
	20	0,029	0,034*	0,054*	0,069*	0,065*	0,070*	0,079	0,088	0,076
15	3	0,026	0,031	0,048*	0,072	0,071	0,085	0,089	0,101	0,080
	20	0,025	0,048*	0,055*	0,096	0,092	0,103	0,096	0,117	0,097
20	3	0,036*	0,046*	0,053*	0,078	0,106	0,106	0,098	0,118	0,114
	20	0,043*	0,050*	0,057*	0,100	0,108	0,103	0,116	0,133	0,119

Fonte: Da autora.
Nota: * valores que se encontram dentro do IC exato para proporção, com 99% de confiança.

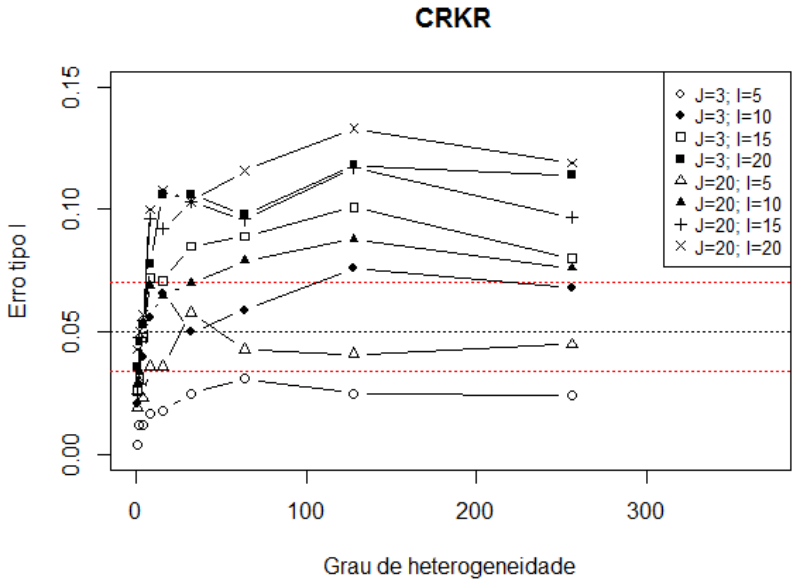

Figura 11 – Taxa do erro tipo I do teste CRKR.
Fonte: Da autora.

Como apresentado na Figura 12, o teste não rejeitou a hipótese nula, quando esta é verdadeira em nenhum dos cenários adotados, mesmo sob alto grau de heterogeneidade, ou seja, o teste não pratica taxa de erro tipo I e também não é influenciado pela heterogeneidade da variância dos erros. Desta forma, para todo cenário simulado o teste praticou 0% de erro tipo I, sendo desnecessária a apresentação da tabela.

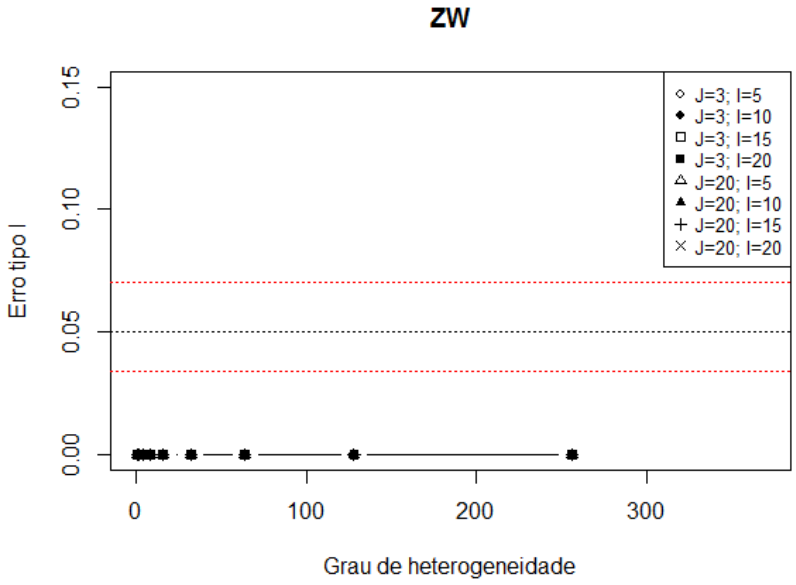

Figura 12 – Taxa do erro tipo I do teste *bootstrap* não-paramétrico de Zhow e Wong.
Fonte: Da autora.

4.2 PODER

Primeiramente foi feito um estudo do comportamento dos testes em relação ao poder, fixando $\phi = 1$, que neste caso é o número de erros padrões da diferença entre a maior e a menor média adotada no estudo. Esta escolha foi feita levando em consideração a representatividade dos cenários em que $\phi = 1$, sendo que os demais apresentam comportamento semelhante à este.

A Figura 13 e a Tabela 14 mostraram o desempenho geral dos testes em relação ao poder.

Tabela 14 – Poder de todos os testes ao longo dos graus crescentes de heterogeneidade.

	Grau de Heterogeneidade (δ)								
Testes	1	2	4	8	16	32	64	128	256
ANAVA	0,835	0,710	0,685	0,694	0,716	0,725	0,720	0,729	0,733
RKR	0,999	0,998	0,996	0,997	0,998	0,998	0,998	0,998	0,998
CRKR	0,313	0,241	0,217	0,235	0,253	0,265	0,265	0,267	0,274
JA	0,045	0,050	0,060	0,093	0,147	0,212	0,293	0,371	0,443
KLM	0,438	0,320	0,315	0,381	0,482	0,574	0,667	0,735	0,774
KW	0,739	0,613	0,606	0,642	0,687	0,706	0,714	0,727	0,731
W	0,000	0,000	0,010	0,301	0,474	0,557	0,926	0,993	0,999
ZW	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Fonte: Da autora.

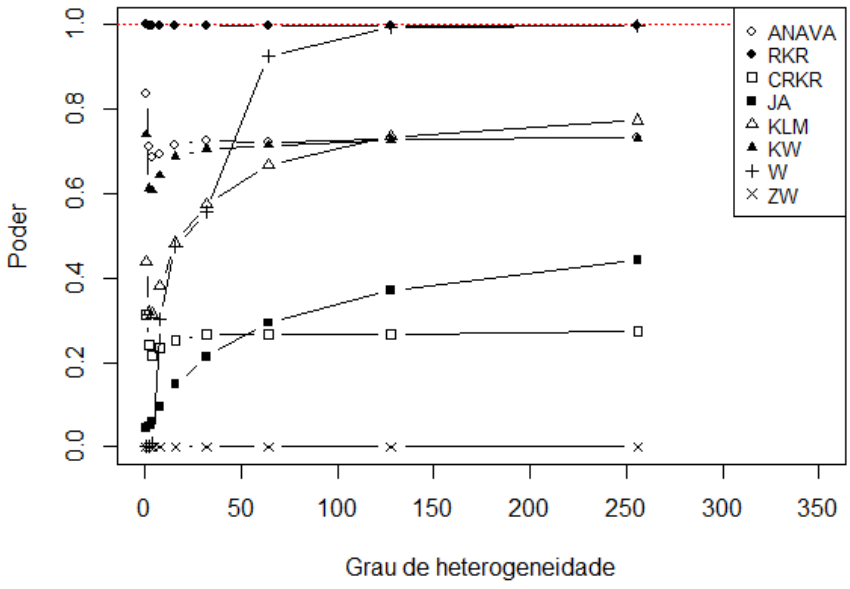

Figura 13 – Poder de todos os testes em relação ao grau de heterogeneidade.
Fonte: Da autora.

Observe que todos os testes estabilizaram o comportamento a partir do grau de heterogeneidade igual a 64, exceto o teste de Welch que novamente se mostrou mais sensível aos graus de heterogeneidade.

A Figura 13 e a Tabela 14 revelaram também que tanto sob a condição de homogeneidade da variância dos erros quanto sob heterogeneidade, os testes KW e F apresentaram excelente desempenho em relação ao poder e, assim como na taxa de erro tipo I, obtiveram melhor desempenho. De maneira semelhante, mas quebrando a normalidade, Ferreira, Mequelino e Rocha (2012) revelaram que o teste F se mostrou igualmente poderoso ou superior ao teste de Kruskal-Wallis, para simulação dos cenários resultantes: número de tratamentos (3, 5, 10, 15, 20, 25 e 30), o número de repetições (3, 4, 5, 10, 15 e 20), os coeficientes de variação ($1\%, 5\%, 10\%, 15\%$ e 20%) e os valores do fator de penalidade (1, 10, 50 e 100), que funciona como pseudo-tratamentos, aumentando assim o número de tratamentos que devem estar igualmente espaçados entre 0 e 1. Já o estudo de Zimmermann (1987) revelou que a heterogeneidade da variância (moderadamente heterogênea - razão de 4 para 1 entre a maior e a menor; e extremamente heterogênea - razão de 16 para 1) exerceu influência no poder do teste F, principalmente quando associada à distribuição normal, o que não foi comprovado neste estudo. Feir e Toothaker (1974) mostraram que o teste KW não foi competitivo em relação ao F quando avaliado o poder, para falta de normalidade (dados exponenciais) e heterogeneidade (razão entre a maior e a menor de até 4 para 1). Já neste trabalho, o teste de KW teve um desempenho tão bom quanto o teste F.

A Figura 13 e a Tabela 14 também apresentaram um altíssimo poder do teste RKR. Entretanto, é importante salientar que o teste também foi caracterizado pela alta taxa de erro tipo I, como mostrou a Figura 10. O teste de Welch também apresenta o mesmo problema, observe que o poder do teste aumentou à medida em que o grau de heterogeneidade cresceu. Contudo, a Figura 8 revelou que a taxa de erro tipo I do teste apresenta o mesmo comportamento com os graus crescentes de heterogeneidade. Já o teste de James não apresentou, em média, um poder satisfatório, variando de 4% à 50% dependendo do grau de heterogeneidade. Cribbie et al. (2012) encontrou um resultado semelhante, sob falta de normalidade (assimetria = 1,75, curtose = 8,90 e assimetria = 6,18, curtose = 113,94), os testes de JA e W não apresentaram um resultado eficaz, pois estes variaram de 10% à 75% sendo que quanto mais assimétricos, menor o poder dos testes.

Já os testes de ZW e CRKR tiveram comportamento não satisfatório com relação ao

poder, ambos ficando abaixo de 40%.

Agora, foi feito um estudo do poder de todos os testes separadamente.

A Figura 14 e a Tabela 15 revelam o desempenho do teste de Kruskal-Wallis, que se mostrou sensível aos graus crescentes de heterogeneidade.

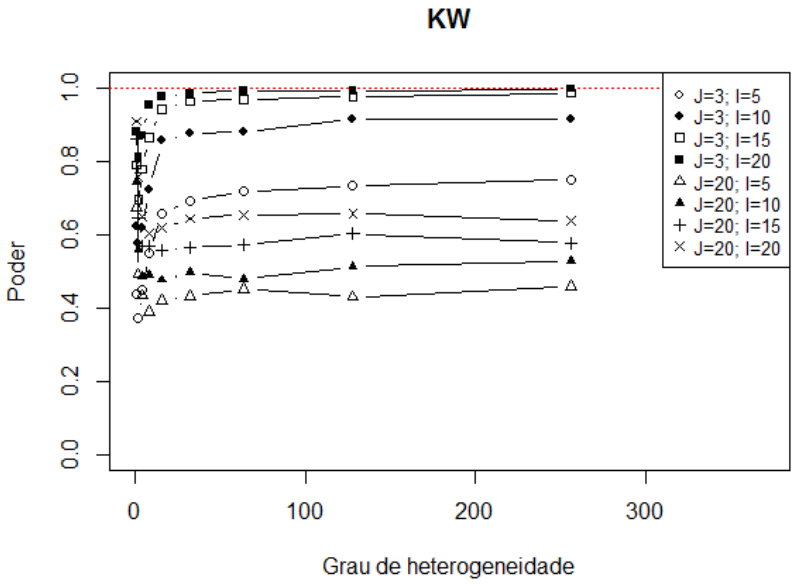

Figura 14 – Poder do teste de Kruskal-Wallis ao longo dos graus de heterogeneidade.
Fonte: Da autora.

Tabela 15 – Poder do teste de Kruskal-Wallis ao longo dos graus de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,440	0,373	0,448	0,548	0,658	0,692	0,718	0,733	0,75
	20	0,674	0,491	0,434	0,391	0,421	0,433	0,452	0,430	0,46
10	3	0,622	0,578	0,618	0,722	0,855	0,874	0,881	0,915	0,914
	20	0,742	0,558	0,487	0,489	0,476	0,497	0,479	0,513	0,528
15	3	0,789	0,694	0,778	0,865	0,941	0,962	0,969	0,975	0,985
	20	0,859	0,646	0,569	0,570	0,557	0,566	0,572	0,602	0,578
20	3	0,881	0,810	0,87	0,952	0,976	0,984	0,992	0,990	0,997
	20	0,908	0,757	0,65	0,604	0,618	0,644	0,654	0,658	0,638

Fonte: Da autora.

Para os cenários com pouca repetição, o teste apresenta o menor poder sob homogeneidade e posteriormente tem um crescimento à medida em que a heterogeneidade vai crescendo, até estabilizar. Ainda nestes casos, o poder é proporcional ao tamanho da amostra. Porém, nos cenários envolvendo vinte repetições, é sob a pressuposição que o poder atinge seu máximo caindo posteriormente antes de estabilizar. Observe que quanto maior o tamanho da amostra,

maior o poder do teste. De maneira geral, neste teste, menos repetição implica em melhor desempenho.

A Figura 15 e a Tabela 16 mostraram que, para o número de repetições igual a três, o poder do teste F é proporcional ao tamanho da amostra. Já para muita repetição, à medida em que os erros deixam de ser homogêneos os testes vão perdendo poder e posteriormente estabilizam, sendo que neste âmbito, os cenários com maior tamanho de amostra apresentam melhor desempenho.

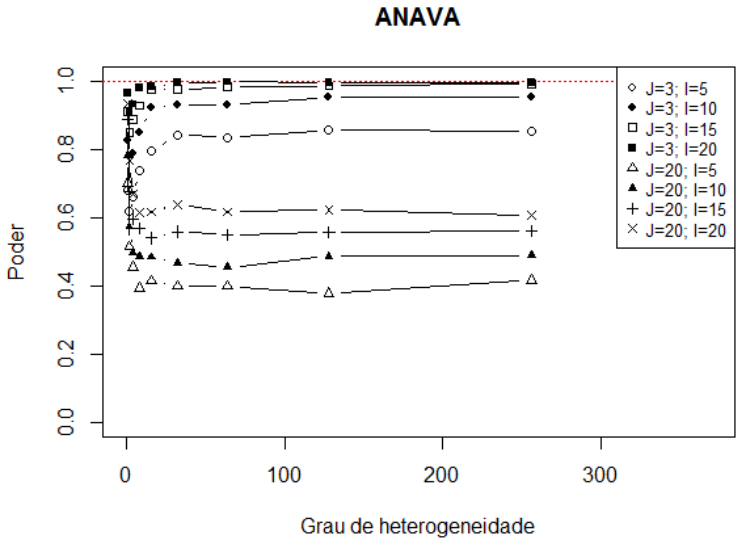

Figura 15 – Poder do teste F ao longo dos graus de heterogeneidade.
Fonte: Da autora.

É importante ressaltar, que embora a homogeneidade seja uma pressuposição do teste F no contexto experimental, a falta dela não afetou bruscamente o comportamento do teste em termos de poder.

Tabela 16 – Poder do teste F ao longo dos graus de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,680	0,619	0,660	0,738	0,797	0,841	0,835	0,856	0,853
	20	0,701	0,514	0,456	0,393	0,415	0,399	0,399	0,377	0,416
10	3	0,825	0,781	0,788	0,850	0,920	0,929	0,930	0,951	0,952
	20	0,782	0,572	0,495	0,486	0,484	0,467	0,454	0,487	0,490
15	3	0,911	0,849	0,888	0,928	0,974	0,976	0,982	0,985	0,992
	20	0,886	0,675	0,597	0,569	0,540	0,559	0,550	0,557	0,561
20	3	0,965	0,908	0,931	0,980	0,983	0,994	0,997	0,996	0,996
	20	0,933	0,766	0,668	0,614	0,617	0,639	0,617	0,623	0,608

Fonte: Da autora.

O poder do teste KLM é apresentado na Figura 16 e na Tabela 17.

Para J=3, sob homogeneidade o teste KLM não apresenta tão bom desempenho quanto na quebra de pressuposição. Neste caso, poder é diretamente proporcional ao grau de heterogeneidade.

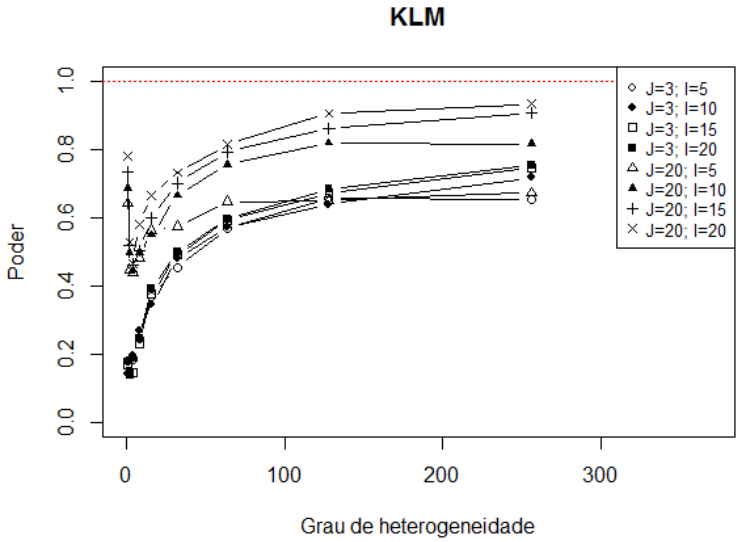

Figura 16 – Poder do teste *bootstrap* paramétrico de Krishnamorthy, Lu, Mathew.
Fonte: Da autora.

Tabela 17 – Poder do teste KLM ao longo dos graus crescentes de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,180	0,142	0,184	0,247	0,371	0,455	0,568	0,656	0,654
	20	0,641	0,447	0,440	0,482	0,562	0,573	0,646	0,650	0,672
10	3	0,142	0,138	0,195	0,268	0,345	0,481	0,572	0,639	0,717
	20	0,686	0,496	0,445	0,497	0,549	0,664	0,755	0,818	0,815
15	3	0,169	0,142	0,146	0,230	0,377	0,492	0,593	0,671	0,745
	20	0,732	0,519	0,463	0,502	0,600	0,701	0,793	0,862	0,905
20	3	0,179	0,149	0,187	0,246	0,392	0,499	0,597	0,684	0,753
	20	0,779	0,528	0,466	0,579	0,666	0,732	0,815	0,905	0,932

Fonte: Da autora.

Já para J=20, o teste perde poder com pouco heterogeneidade, entretanto ela volta a crescer à medida que a heterogeneidade cresce. Este teste também se mostrou sensível à heterogeneidade crescente da variância dos erros para a avaliação do poder.

O comportamento do teste de Welch pode ser visto na Figura 17 e na Tabela 18. Observe que sob homogeneidade todos os cenários mostram que o teste não tem poder, ou seja, ele não consegue detectar diferença entre as médias quando estas são de fato diferentes.

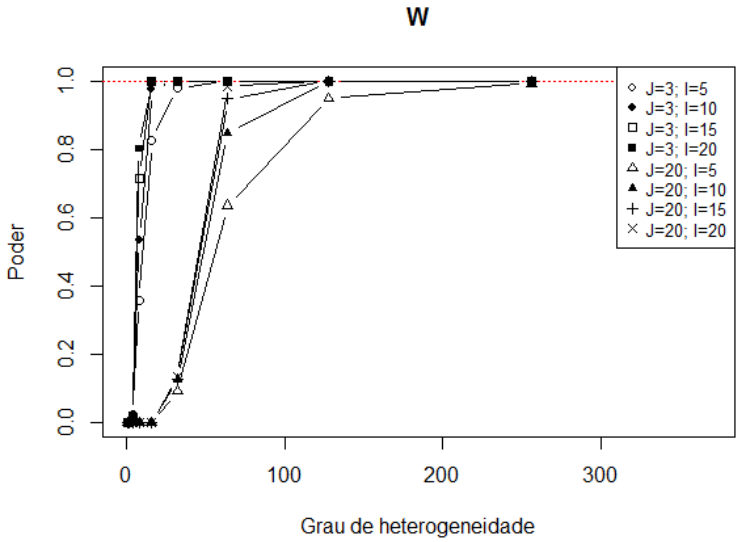

Figura 17 – Poder do teste de Welch em relação aos graus de heterogeneidade.
Fonte: Da autora.

Tabela 18 – Poder do teste de Welch para os graus crescentes de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,000	0,000	0,025	0,358	0,824	0,981	0,997	1,000	1,000
	20	0,000	0,000	0,000	0,000	0,000	0,092	0,635	0,948	0,993
10	3	0,000	0,000	0,018	0,534	0,977	1,000	1,000	1,000	1,000
	20	0,000	0,000	0,000	0,000	0,000	0,126	0,847	1,000	1,000
15	3	0,000	0,000	0,016	0,716	0,998	1,000	1,000	1,000	1,000
	20	0,000	0,000	0,000	0,000	0,000	0,128	0,948	1,000	1,000
20	3	0,000	0,000	0,021	0,801	0,999	1,000	1,000	1,000	1,000
	20	0,000	0,000	0,000	0,000	0,000	0,134	0,984	1,000	1,000

Fonte: Da autora.

Já quando os graus de heterogeneidade crescem, o poder do teste obtém o mesmo comportamento, crescendo com maior velocidade à medida que o tamanho da amostra cresce, atingindo o ápice. Entretanto, a taxa de erro tipo I do teste apresenta o mesmo comportamento, mostrando um desempenho indesejável do teste.

A Figura 18 e a Tabela 19 revelam que para pouca repetição o teste de James apresenta melhor desempenho, atingindo maior poder quanto menor o número de tratamentos, o que não era de se esperar, já que este teste foi proposto para experimentos com grandes amostras. Vale ressaltar também que a pressuposição de homogeneidade implica em menor poder do teste, quando comparado com os demais graus de heterogeneidade. Já para o caso em que o número de repetições é igual a vinte, o teste comete quase 100% de erro tipo II para todos os cenários,

ou seja, toma-se a decisão errônea de aceitar H_0, mesmo que esta seja falsa.

Tabela 19 – Poder do teste de James ao longo dos graus crescentes de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,196	0,214	0,255	0,384	0,603	0,789	0,952	0,994	1,000
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,008	0,070
10	3	0,089	0,083	0,118	0,183	0,291	0,434	0,636	0,849	0,968
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
15	3	0,047	0,063	0,066	0,112	0,177	0,284	0,453	0,642	0,832
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20	3	0,034	0,042	0,047	0,071	0,112	0,194	0,310	0,479	0,679
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Fonte: Da autora.

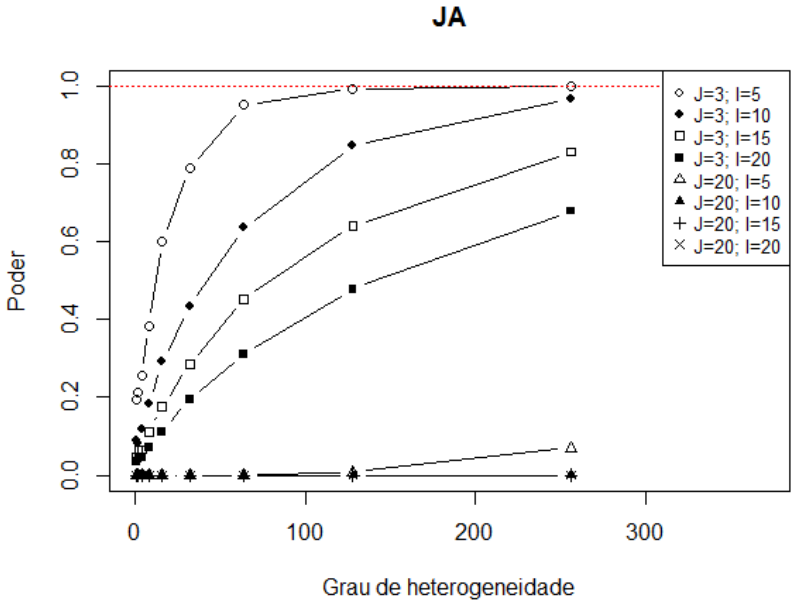

Figura 18 – Poder do teste de James em relação aos graus de heterogeneidade.
Fonte: Da autora.

O desempenho, em termos de poder, do teste RKR, pode ser comprovado na Figura 19 e na Tabela 20. Observe que em todos os cenários e para todos os graus de heterogeneidade o teste se mostrou altamente poderoso. Logo, o teste não é sensível à quebra da pressuposição de homogeneidade. Contudo, o teste também se caracterizou por alta taxa de erro tipo I, revelando que este não tem bom desempenho.

Tabela 20 – Poder do teste RKR ao longo dos graus crescentes de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,998	1,000	0,999	0,999	1,000	1,000	1,000	1,000	1,000
	20	1,000	0,986	0,973	0,977	0,985	0,987	0,987	0,987	0,989
10	3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	20	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
15	3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	20	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	20	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Fonte: Da autora.

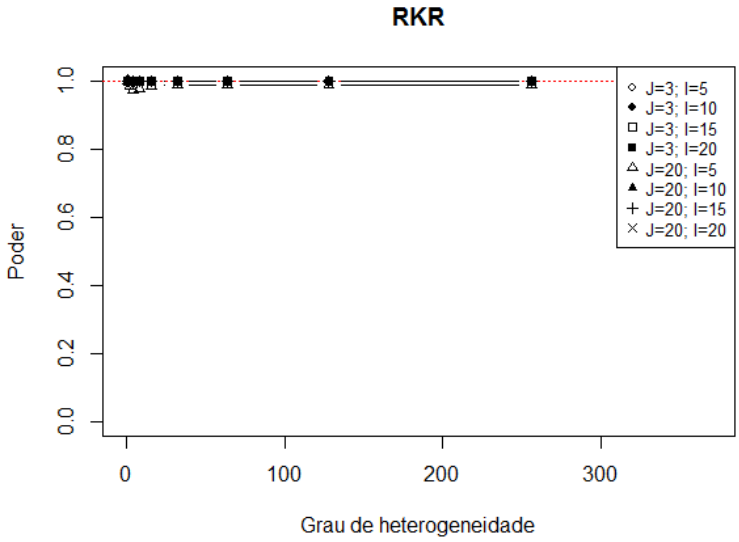

Figura 19 – Poder do teste *bootstrap* não paramétrico de Rddy, Kumar e Ramu.
Fonte: Da autora.

Já a correção do teste anterior é apresentada na Figura 20 e na Tabela 21.

Tabela 21 – Poder do teste CRKR ao longo dos graus crescentes de heterogeneidade.

Trat	Rep	Grau de Heterogeneidade (δ)								
		1	2	4	8	16	32	64	128	256
5	3	0,096	0,110	0,102	0,143	0,142	0,159	0,169	0,155	0,155
	20	0,417	0,264	0,210	0,202	0,213	0,203	0,196	0,209	0,206
10	3	0,176	0,175	0,179	0,184	0,200	0,242	0,229	0,228	0,242
	20	0,457	0,286	0,229	0,246	0,263	0,277	0,281	0,264	0,262
15	3	0,191	0,178	0,228	0,225	0,24	0,284	0,270	0,272	0,280
	20	0,471	0,335	0,283	0,290	0,31	0,330	0,328	0,320	0,324
20	3	0,193	0,187	0,220	0,257	0,298	0,312	0,299	0,340	0,325
	20	0,510	0,395	0,285	0,333	0,362	0,318	0,352	0,355	0,399

Fonte: Da autora.

Observe que este teste apresentou um desempenho insatisfatório, sendo que em todas as combinações de cenários o poder variou entre 10% e 50% ao longo de todos os graus de heterogeneidade. Veja também que mesmo pouco poderoso, o teste é pouco sensível aos graus de heterogeneidade.

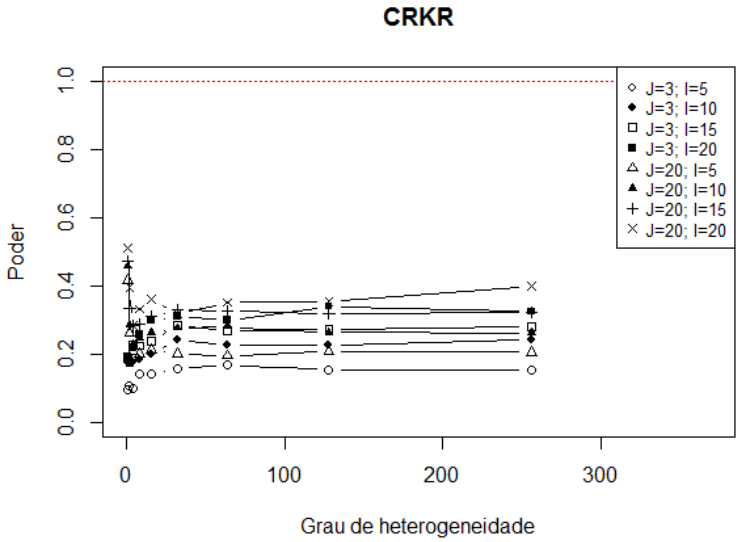

Figura 20 – Poder do teste CRKR.
Fonte: Da autora.

Já o teste ZW nunca rejeita a hipótese nula, mesmo que esta não seja, de fato, verdadeira, como mostra a Figura 21.

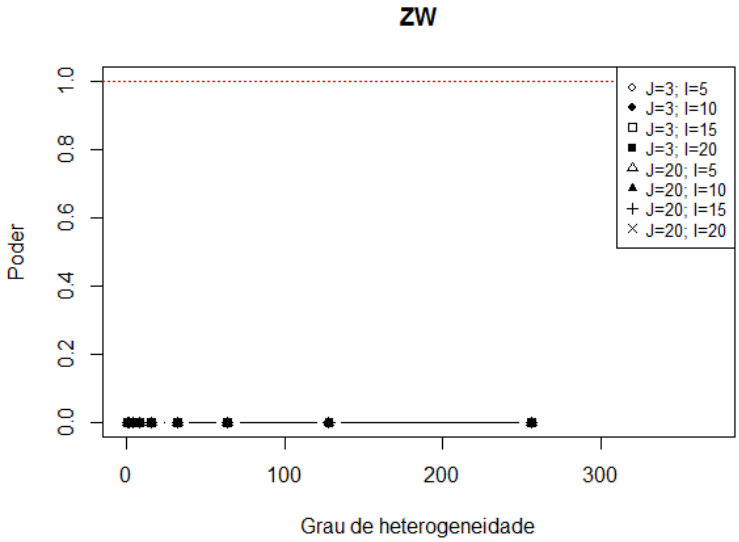

Figura 21 – Poder do teste *bootstrap* não paramétrico de Zhow e Wong.
Fonte: Da autora.

Observe que, assim como no erro tipo I, o teste não foi afetado pelos graus de heteroge-

neidade, já que não possui poder. Desta forma, para todo cenário simulado o teste praticou 0% de poder, sendo desnecessária a apresentação da tabela.

Agora, foi feita uma análise do comportamento dos testes em termos de poder levando em consideração o desempenho deles em cada grau de heterogeneidade, como mostra a Figura 22.

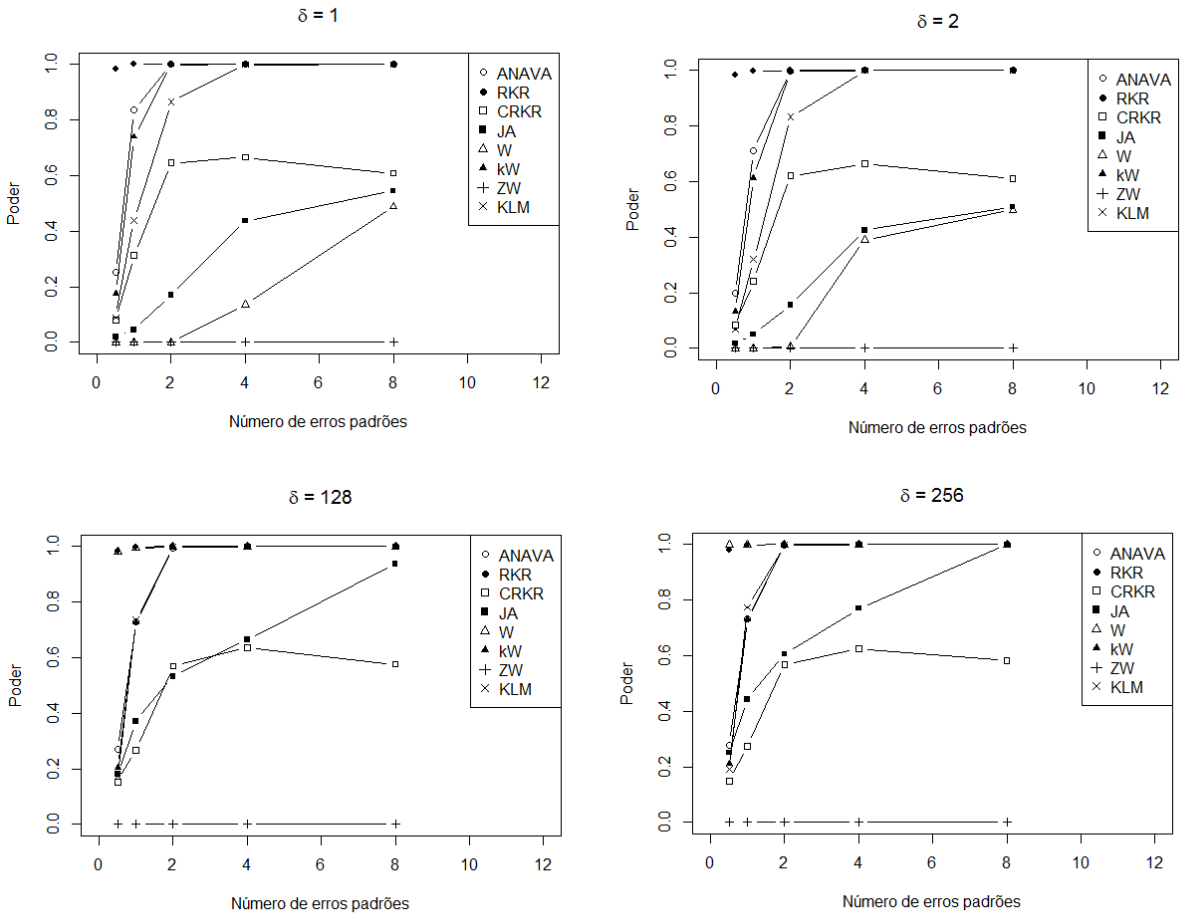

Figura 22 – Poder dos testes para $\delta \in \{1, 2, 128, 256\}$.
Fonte: Da autora.

A Figura 22 revela que o teste com melhor desempenho em termos de poder é o teste F na ANAVA, KW e KLM, que com 2 e 4 erros padrões de diferença entre a maior e a menor média já conseguiram atingir 100% de poder. É importante lembrar que, embora o teste RKR apresente alto poder, ele também apresenta alta taxa de erro tipo I, o que não é um comportamento desejável. Já o teste de James, só conseguiu detectar a diferença entre os tratamentos, quando estes eram, de fato, bem distintos. Agora os testes CRKR e ZW não conseguiram atingir um bom desempenho em termos de poder, mesmo quando os tratamentos apresentam grandes diferenças entre si. Observe também que o teste de Welch é altamente influenciado pelo grau

de heterogeneidade, sendo que para os casos de homogeneidade e baixa heterogeneidade o teste apresenta baixo poder, já para os casos de grande heterogeneidade, ele atinge o máximo de poder.

Entretanto, é importante ressaltar que os testes ficam mais sensíveis a diferença entre os tratamentos, atingindo o máximo de poder com maior rapidez, à medida em que cresce o grau de heterogeneidade.

4.3 APLICAÇÃO 1 - ANÁLISE SENSORIAL DE QUEIJO MINAS PADRÃO

Primeiramente, dentre as variáveis em estudo, somente a cor apresentou heterogeneidade da variância dos erros, de acordo com o teste de Bartlett à 5% de significância. Nesta etapa, ainda foi calculado o grau de heterogeneidade, sendo que a variável cor apresentou $\delta = 10{,}250$. Note, na Figura 23, como a variabilidade das notas, segundo a escala não estruturada de 9 pontos, aumentou ao longo do tempo de maturação do queijo minas padrão. Este é um comportamento esperado, já que ao longo do tempo o queijo tende a apresentar um aspecto mais amarelado, resultante da desidratação e também do acúmulo de gordura em sua superfície, o que não acontece no início da maturação, sendo a variabilidade mais homogênea. Já o teste de Shapiro-Wilk, à 5% de significância, revelou que os erros podem ser considerados normais.

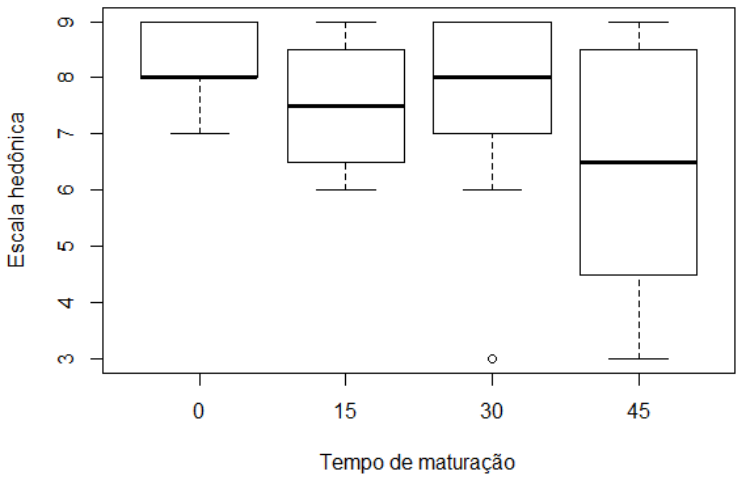

Figura 23 – *Boxplot* da variância dos erros ao longo do tempo de maturação.
Fonte: Da autora.

Posteriormente foram aplicados os testes estudados neste trabalho para os resultados da

variável cor. Os valores-p de todos os testes, à 5% de significância, podem ser vistos na Tabela 22.

Tabela 22 – Valores-p dos oito testes em estudo para a aplicação do queijo minas padrão.

Testes	JA	W	ZW	KW	ANAVA	KLM	CRKR	RKR
Valores-p	0,954	0,921	0,510	0,197	0,192	0,168	0^*	1^*

Fonte: Da autora.

Nota: * os teste RKR e CRKR não possuem valores-p, pois são baseados em cartas de controle, neste caso, 1 significa que deve-se rejeitar H_0 e 0 significa que não se deve rejeitar H_0.

Observe que os testes JA, W, ZW, KW, ANAVA, KLM e CRKR concordam em não rejeitar a hipótese nula, ou seja, em média o queijo apresenta estatisticamente a mesma coloração ao longo do tempo de maturação, quando não há inserção de inulina, observe na Tabela 23 a apresentação das médias.

Tabela 23 – Médias em cada tempo de maturação.

Tempo	Médias
0	8,250
15	7,500
30	7,500
45	6,375

Fonte: Da autora.

Entretanto, o teste RKR indica rejeitar H_0, ou seja, pelo menos uma das médias de cor do queijo minas padrão ao longo do tempo difere das demais.

Veja que, neste experimento, tem-se $\delta = 10{,}250$, $I = 4$ e $J = 8$. Comparando essas informações com o estudo de simulação, pode-se determinar a taxa de erro tipo I (α) e taxa de erro tipo II (β) para esta aplicação, levando em consideração os cenários simulados que mais se aproximam dos dados reais, ou seja, $\delta = 8$, $I = 5$ e $J = 3$, como mostra a Tabela 24.

Tabela 24 – Erro tipo I e erro tipo II para a aplicação do queijo minas padrão.

Testes	JA	W	ZW	KW	ANAVA	KLM	CRKR	RKR
α	0,041*	0,031	0,000	0,022	0,065*	0,014	0,017	0,861
β	0,616	0,642	1,000	0,452	0,262	0,753	0,857	0,001

Fonte: Da autora.

Nota: * valores que se encontram dentro do IC exato para a proporção, com 99% de confiança.

Observe na Tabela 24, que o teste com melhor desempenho neste caso é o teste F no contexto da ANAVA, pois é exato em termos de erro tipo I e pratica cerca de $26{,}2\%$ de taxa

de erro tipo II, isto é, adotando a recomendação desse teste em não rejeitar H_0 existe $26{,}2\%$ de chance de tomar uma decisão errada aceitando a hipótese nula, quando esta é falsa. Vale ressaltar que o teste RKR não obteve bom desempenho em termos de erro tipo I nos cenários testados, desta forma adotar sua decisão pode levar o pesquisador a um resultado errôneo.

4.4 APLICAÇÃO 2 - EFEITO DO ESPAÇAMENTO NO DESENVOLVIMENTO DE MUDAS

Dentre as variáveis em estudo (altura da planta, área e volume da copa), de acordo com o teste de Bartlett à 5% de significância, as variáveis área e volume apresentaram heterogeneidade na variância dos erros, cujo grau de heterogeneidade $\delta = 6{,}309$ para a área e $\delta = 19{,}467$ para o volume.

A Figura 24 revela que a variabilidade tanto da área da copa quanto do volume aumentam à medida que o espaçamento também cresce. Também pode-se perceber a presença de *outlier* no conjunto de dados.

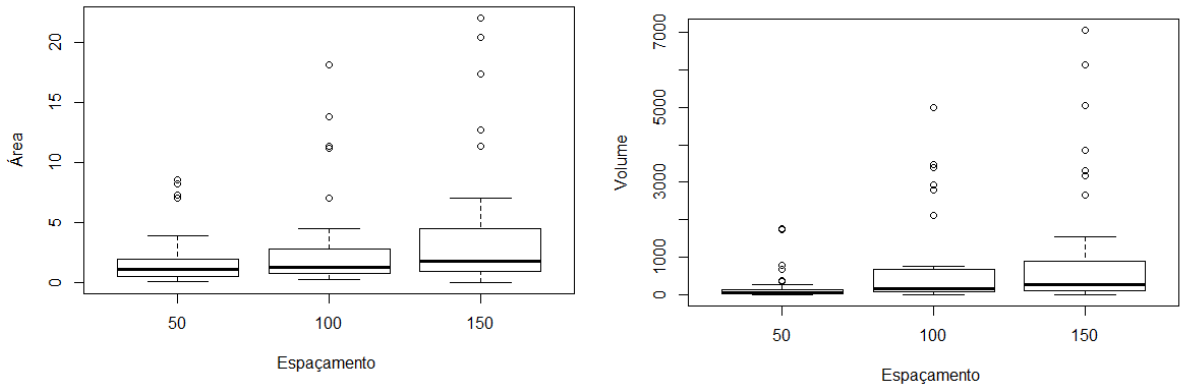

Figura 24 – *Boxplot* da variância dos erros da área e do volume.
Fonte: Da autora.

O teste de Shapiro-Wilk, à 5% de significância, revelou que os erros não podem ser considerados normais, tanto para a variável área, quanto para a variável volume.

A Tabela 25 mostra os valores-p para a variável área de todos os testes estudados neste trabalho à 5% de significância.

Para tomar uma decisão com relação aos testes, comparou-se os dados obtidos no expe-

Tabela 25 – Valores-p dos oito testes em estudo para a variável área.

Testes	KW	KLM	RKR	CRKR	ANAVA	ZW	W	JA
Valores-p	$3{,}67 \times 10^{-05}$	0,029	1^*	0^*	0,056	0,521	0,672	0,916

Fonte: Da autora.

Nota: * os teste RKR e CRKR não possuem valores-p, pois são baseados em cartas de controle, neste caso, 1 significa que deve-se rejeitar H_0 e 0 significa que não se deve rejeitar H_0.

rimento ($I = 3$, $J = 38$ e $\delta = 6{,}309$) com o estudo de simulação proposto neste trabalho. A Tabela 26 mostra a taxa de erro tipo I (α) e taxa de erro tipo II (β) para esta aplicação, levando em consideração os cenários simulados que mais se aproximam dos dados reais, ou seja, $\delta = 8$, $I = 5$ e $J = 20$.

Tabela 26 – Taxa de erro tipo I e erro tipo II para a variável área.

Testes	KW	KLM	RKR	CRKR	ANAVA	ZW	W	JA
α	0,048*	0,043*	0,799	0,036*	0,067*	0,000	0,000	0,000
β	0,609	0,518	0,023	0,798	0,607	1,000	1,000	1,000

Fonte: Da autora.

Nota: * valores que se encontram dentro do IC exato para a proporção, com 99% de confiança.

Observe que os testes KW, KLM e RKR sugerem rejeitar H_0 de igualdade das médias, ou seja, de acordo com estes testes pelo menos uma das médias difere das demais. Já os testes CRKR, ANAVA, ZW, W e JA, indicam que, para a variável área, a média em todos os espaçamentos são estatisticamente iguais. A Tabela 27 revela as médias do experimento para a variável área.

Tabela 27 – Médias da área da copa em cada espaçamento de plantio.

Espaçamento	Médias
50	1,815
100	3,173
150	4,197

Fonte: Da autora.

Observe que de acordo com os resultados do estudo de simulação, recomenda-se rejeitar a hipótese nula, pois os testes de KW e KLM apresentam as menores taxas de erro tipo II, levando em consideração a exatidão da taxa de erro tipo I, ou seja, em média eles cometem 5% de chance de fazer uma recomendação errada (rejeitar H_0 verdadeira). Também vale destacar que, embora na aplicação o teste F indique a não rejeição de H_0 (valor-p=0,056), o estudo de

simulação mostra que este resultado é estatisticamente igual a 0,05, ou seja, assim como KW e KLM, a ANAVA sugere que as médias são diferentes.

Agora, na Tabela 28 encontram-se os valores-p dos testes simulados nesta pesquisa para a variável volume.

Tabela 28 – Valores-p dos oito testes em estudo para a variável volume.

Testes	W	KW	ANAVA	KLM	RKR	CRKR	ZW	JA
Valores-p	$2{,}28 \times 10^{-256}$	$6{,}04 \times 10^{-06}$	0,015	0,020	1^*	0^*	0,503	0,730

Fonte: Da autora.
Nota: * os teste RKR e CRKR não possuem valores-p, pois são baseados em cartas de controle, neste caso, 1 significa que deve-se rejeitar H_0 e 0 significa que não se deve rejeitar H_0.

Observe que os testes W, KW, ANAVA, KLM e RKR concordam em rejeitar a hipótese nula, ou seja, considerando a variável volume pelo menos uma das médias dos espaçamentos difere das demais. Entretanto, os testes CRKR, ZW e JA indicam a não rejeição de H_0, ou seja, as médias de todos os espaçamentos podem ser consideradas iguais, observe na Tabela 29 as médias em cada um dos espeçamentos.

Tabela 29 – Médias do volume da copa em cada espaçamento de plantio.

Espaçamento	Médias
50	212,908
100	699,430
150	1065,037

Fonte: Da autora.

Para tomar uma decisão com relação aos testes, comparou-se os dados obtidos no experimento ($I = 3$, $J = 38$ e $\delta = 10{,}250$) com o estudo de simulação proposto neste trabalho, a Tabela 30 mostra a taxa de erro tipo I (α) e taxa de erro tipo II (β) para esta aplicação, levando em consideração os cenários simulados que mais se aproximam dos dados reais, ou seja, $\delta = 8$, $I = 5$ e $J = 20$.

Tabela 30 – Taxa de erro tipo I e erro tipo II para a variável volume.

Testes	W	KW	ANAVA	KLM	RKR	CRKR	ZW	JA
α	0,000	0,048*	0,067*	0,043*	0,799	0,036*	0,000	0,000
β	1,000	0,609	0,607	0,518	0,023	0,798	1,000	1,000

Fonte: Da autora.
Nota: * valores que se encontram dentro do IC exato para a proporção, com 99% de confiança.

De acordo com os resultados do estudo de simulação, recomenda-se rejeitar a hipótese nula, pois os testes de KW, ANAVA e KLM apresentam as menores taxas de erro tipo II, levando

em consideração a exatidão da taxa de erro tipo I, ou seja, em média eles cometem 5% de chance de fazer uma recomendação errada (rejeitar H_0 verdadeira). Também vale destacar que os testes CRKR, ZW e JA não tiveram tão bom desempenho em termos de erro tipo I e poder nos cenários próximos ao desta aplicação.

5 CONCLUSÃO

O estudo feito neste trabalho revelou quais são os testes com melhor desempenho para a comparação de médias na quebra da homogeneidade da variância dos erros. Tanto tem termos de poder, quanto em termos de erro tipo I, o teste de Kruskal-Wallis mostrou maior eficiência, seguido do teste F, no contexto da análise de variância. Estes testes podem ser recomendados, mesmo nos cenários com alta heterocedasticidade, além de se mostrarem pouco sensíveis ao número de tratamentos, repetições ou diferença entre as médias. Dentre eles, o teste F é o único teste paramétrico, e desta forma, podemos concluir que ele é um teste robusto para falta de homogeneidade da variância dos erros, com grau de heterogeneidade variando de 1 à 256.

O teste *bootstrap* paramétrico de Krishnamoorthy, Lu e Mathew se mostrou competitivo ao teste KW e ao teste F, pois mesmo sendo conservador em termos de erro tipo I, apresentou alto poder para os cenários com alto grau de heterogeneidade, podendo ser recomendado nestes casos.

De maneira geral, os testes se mostraram pouco sensíveis à heterogeneidade da variância dos erros, o que era esperado, já que eles foram construídos como uma alternativa ao teste F na quebra dessa pressuposição. Contudo, o teste de Welch não teve o mesmo comportamento, se mostrando bastante sensível à falta da homogeneidade.

A correção proposta no trabalho, apresentou melhor desempenho em termos de erro tipo I que o teste original proposto por Reddy, Kumar e Ramu. Entretanto, seu poder não foi tão eficiente, mas é importante lembrar que a taxa de erro tipo I do teste original é bem mais alta do que aquela fixada no estudo. Assim, pode-se considerar que a nova proposta melhorou o desempenho do teste.

REFERÊNCIAS

ALMEIDA, A.; ELIAN, S.; NOBRE, J. Modificações e alternativas aos testes de Levene e de Brown e Forsythe para igualdade de variâncias e médias. **Revista Colombiana de Estatística**. (Bogotá), v. 31, n. 2, p. 241-260, 2008.

BASTOS, R. L. **Proposição de testes bootstrap para o índice de qualidade sensorial**. 2013. 125f. Dissertação de Mestrado, Universidade Federal de Lavras, Lavras, 2014.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. 6 ed. São Paulo: Atual, 2009, 526 p.

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

CRIBBIE, R. A. et al. Effects on 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.

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

DACHS, J. N. W. **Estatística computacional: uma introdução em Turbo Pascal**. Rio de Janeiro, 1988, 236 p.

EFRON, B.*Bootstrap* methods: another look at the jack- knife. In: ANNALS OF STATISTIC, n. 1, 1979. Hayward. **Anais...** Hayward, 1979. p. 1-26.

FEIR, B.; TOOTHAKER, L. The ANOVA F-test versus the Kruskal-Wallis test: a robustness study. In: ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, Chicago. **Anais ...**. Chicago, 1974. p. 1-37.

FERNANDES, A. R. M. Composição em ácidos graxos e qualidade da carne de tourinhos Nelore e Canchim alimentados com dietas à base de cana-de-açúcar e dois níveis de concentrado. **Revista Brasileira de Zootecnia**, (Lavras), v. 38, n. 12, p. 328-337, 2009.

FERREIRA, E. B.; ROCHA, M. C.; MEQUELINO, D. B. Monte Carlo evaluation of the ANOVA's F and Kruskal-Wallis tests under binomial distribution. **Sigmae**, v. 1, n. 1, p. 126 - 139, 2012.

JAMES, G. S. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. **Biometrika** v. 41, 1954.

JANEBRO, D. I. et al. Efeito da farinha da casca do maracujá-amarelo (Passiflora edulis f. flavicarpa Deg.) nos níveis glicêmicos e lipídicos de pacientes diabéticos tipo 2. **Revista Brasileira de Farmacognosia Brazilian Journal of Pharmacognosy**, v. 18, p.724-732, 2008.

KRISHNAMOORTHY, K.; LU, F.; MATHEW, T. A parametric bootstrap approach for ANOVA with unequal variances: fixed and randow models. **Computational Statistics & Data Analysis**, v. 51, n. 12, p. 5731 - 5742, 2007.

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

LIMA, A. R. **Efeito da descafeinação do café sobre a atividade antioxidante e prevenção da lesão hepática em ratos**. 2008. 90 f. Dissertação de Mestrado em Ciência dos Alimentos, Universidade Federal de Lavras, Lavras, 2008.

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

MONTGOMERY, D. C. **Design and analysis the experiments.** 5. ed. New York: John Wiley, 2000, 684 p.

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

OLIVEIRA, M. S. et al. **Introdução à Estatística**. Lavras: Editora UFLA, 2009, p. 334.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015.

REDDY, M. K.; KUMAR, B. N.; RAMU, Y. Bootstrap method for testing of equality of several means. **Inter Stat**, 2010.

REIS, G. M.; RIBEIRO, J. I. Jr. Comparação de testes paramétricos e não paramétricos aplicados em delineamentos experimentais. Viçosa. In: ANAIS DO III SIMPÓSIO ACADÊMICO DE ENGENHARIA DE PRODUÇÃO. **Anais ...** Viçosa, 2007. p. 1-13.

SALSBURG, D. **Uma senhora toma chá... como a estatística revolucionou a ciência do século XX**. Rio de Janeiro: Jorge Zahar, 2009.

SANTOS, A. C. dos. **Definição do tamanho amostral usando simulação Monte Carlo para os testes de normalidade univariado e multivariado baseados em assimetria e curtose**. 2001. 71f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2001

SANTOS, D. N.; SOUSA, M. S. B.; SILVA, R. A. Análise sensorial do bolo de puba da mandioca -manuê- elaborado com açúcar cristal e rapadura. **Revista Brasileira de Produtos Agroindustriais**, (Campina Grande), v. 13, n. 3, p. 229-234, 2011.

SATTERTHWAITE, F.E. An Approximate Distribution of Estimates of Variance Components. **Biometrics Bulletin**, v. 2, n. 6, p. 110-114, 1946.

SCHEFFÉE, H. **The analysis of variance**. New Yor: Wiley, 1959. 478 p.

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

STORTI, L. B.; FERREIRA, E. B.; PEREIRA, C. A importância dos experimentos em faixas na sensometria: o caso do queijo minas Padrão com inulina. **Sigmae**, (Alfenas), v. 3, n. 2, p. 25-33, 2014.

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

WELCH, B. L., On the comparison of several mean values: an alternative approach. **Biometrics** v. 38, p. 330-336, 1951.

ZIMMERMANN, F. J. P. Efeito de Heterogeneidade de variância e distribuição de probabilidade dos dados sobre o poder e tamanho do teste F. **Revista Agropecuária Brasileira** , v. 22, n. 11, p. 1209-1213, 1987.

ZHANG, G. A parametric bootstrap approach for one-way ANOVA under unequal variances with unbalanced data. **Communications in Statistics - Simulation and Computation**, p. 37-41, 2014.

ZHOU, B.; WONG, W. H. A bootstrap-based non-parametric ANOVA method with applications to fatorial microarray data. **Statistica Sinica**, v. 21, n. 22, p. 485 - 514, 2011.