COMPARAÇÃO DE TESTES PARA IGUALDADES DE MÉDIAS SOB HETEROCEDASTICIDADE: SIMULAÇÃO E APLICAÇÃO

Helen Maria Pedrosa de OLIVEIRA [1]
Eric Batista FERREIRA[1]

- RESUMO: Geralmente a comparação de várias médias é feita por meio de testes para detectar a existência de diferenças entre tratamentos. Um dos testes mais utilizados neste caso é o teste F, no contexto da Análise da Variância. Entretanto, sua credibilidade está ligada ao cumprimento de quatro pressuposições: aditividade dos termos do modelo, os erros devem seguir uma distribuição normal, serem independentes e possuírem variâncias homogêneas. Contudo, na literatura existem testes alternativos ao F, quando se quebra alguma das pressuposições. O objetivo deste trabalho foi a comparação de sete testes para a igualdade de médias sob heterocedasticidade. A avaliação dos testes foi feita analisando as taxas de erro tipo I e poder, ao longo de cenários resultantes da combinação de números de tratamentos, repetições, graus de heterogeneidade e erros padrões da diferença entre as médias. De maneira geral, eles se mostraram pouco sensíveis ao aumento da heterogeneidade, o que não aconteceu com o teste de Welch. Nas condições avaliadas, os testes de melhor desempenho foram Kruskal-Wallis e o F, 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 de Zhow e Wong.
- PALAVRAS-CHAVE: Estatística Computacional; delineamento experimental; inferência.

1 Introdução

Comparar o efeito que alguns tratamentos exercem em determinado experimento é alvo do estudo de vários pesquisadores. Entretanto, em grande parte das pesquisas essa comparação é feita por meio do teste F, no contexto da análise de variância (ANAVA), embora existam inúmeros testes com a mesma função.

[1]Universidade Federal de Alfenas - UNIFAL-MG, Instituto de Ciências Exatas, CEP: 37130-000, Alfenas, MG, Brasil. E-mail: *helen.pedrosa@yahoo.com.br*; *ericbferreira@gmail.com*

De acordo com Vieira (2006), a análise de variância é uma técnica estatística que possibilita averiguar, por meio da comparação de variâncias, se as médias dos tratamentos são estatisticamente iguais.

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 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. Caso pelo menos uma dessas pressuposições não seja satisfeita, a análise de variância fica comprometida e, por sua vez, a validade do teste F pode ser contestada.

Entretanto, no meio acadêmico há divergências sobre o desempenho do teste F. Alguns autores o denominam como um teste robusto, ou seja, possui desempenho satisfatório mesmo não atendendo à todas as pressuposições, como revela o estudo feito por Feir e Toothaker (1974) para a falta de normalidade. Por outro lado, há quem defenda a vertente de que sua credibilidade está diretamente ligada ao cumprimento das pressuposições, o que nem sempre acontece no âmbito das aplicações.

Cochran (1947), por exemplo, revela que efetuar a análise de variância quando a verdadeira variância dos erros são heterogêneas, acarretará na perda de eficiência na estimação dos efeitos de tratamentos, e haverá, também, uma perda de sensibilidade nos testes de significância, sendo que quanto maiores forem as diferenças na variância, maiores serão estas perdas. Scheffée (1959) revela 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. De acordo com esses autores, o resultado do teste F fica bastante comprometido quando as variâncias dos erros são heterogêneas.

Nos casos em que as pressuposições do teste F não são atendidas, a literatura apresenta como alternativa testes que foram construídos quando ocorre a quebra de 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.

O método de reamostragem *bootstrap*, introduzido por Efron(1979), é 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 $j_1, j_2, \ldots, j_m \in \{1, \ldots, m\}$.

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

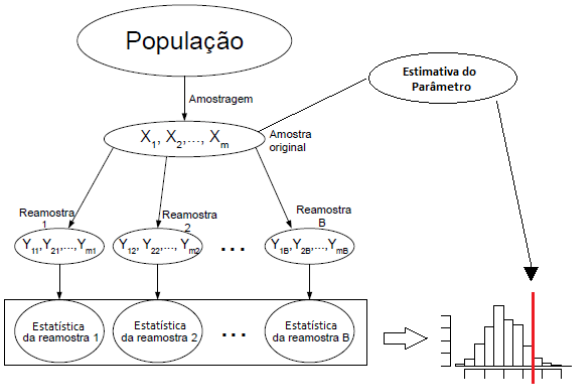

Figura 1 - Funcionamento do *bootstrap* não-paramétrico. Adaptado de Bastos (2014, p. 30).

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.

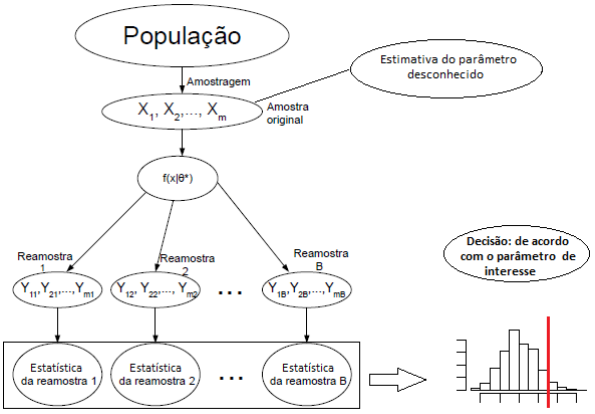

Figura 2 - Funcionamento do *bootstrap* paramétrico. Nota: θ^* é a estimativa de θ. Adaptado de Bastos (2013, p. 28).

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.

Existem diversos estudos que avaliam o desempenho de testes para a igualdade de médias sob condições adversas, como por exemplo, Zimmermann (1987) estudou o desempenho do teste F sob falta de homogeneidade na variância dos erros, Krishnamoorthy, Lu, Mathew (2007) avaliaram o teste de Welch, James e um *bootstrap* paramétrico proposto pelos autores para variâncias desconhecidas e arbitrárias, já Ferreira, Rocha e Mequelino (2012) que estudaram o comportamento do teste F e do teste de Kruskal-Wallis sob falta de normalidade. Entretanto, muitos desses estudos se mostram frágeis, devido ao baixo número de testes de médias comparados e ao baixo número de cenários avaliados. Desta forma, o objetivo do presente trabalho é avaliar o desempenho de testes para a comparação de médias sob a condição de heterocedasticidade, em cenários formados pelas combinações entre o número de tratamentos, repetições, graus de heterocedasticidade e erros padrões da diferença entre as médias.

2 Testes

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:

$$y_{ij} = \mu + \tau_i + e_{ij} \tag{1}$$

em que y_{ij} é o valor da parcela que recebeu o *i-ésimo* tratamento em sua *j-ésima* repetição, 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 fixo do *i-ésimo* tratamento, com a restrição $\sum_i \tau_i = 0$; e_{ij} é o erro associado a y_{ij}, sendo que $e_{ij} \sim N(0; \sigma_i^2)$.

Nos testes de comparações de médias, o par de hipóteses é dado por:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \\ H_1 : \text{pelo menos uma das médias se difere das demais} \end{cases} \tag{2}$$

O par de hipóteses também pode ser escrito em termos dos efeitos dos tratamentos:

$$\begin{cases} H_0 : \tau_1 = \ldots = \tau_I = 0 \\ H_1 : \tau_I \neq \tau_K \text{ para algum } I \neq K \end{cases} \tag{3}$$

De acordo com o modelo adotado em (1) e levando em consideração que μ é uma constante comum à todas as observações, pode-se substituir μ_i por $\mu + \tau_i$, como foi feito nas equações (3).

2.1 Teste F no contexto de análise de variância (ANAVA)

De acordo com Salsburg (2009), a *distribuição F*, apesar de receber este nome, não foi proposta por Fisher, mas por Snedecor, que o homenageou. O teste F se baseia na razão entre duas qui-quadrado sobre seus respectivos graus de liberdade, como mostra a Figura 3.

A estatística desse teste é dada por:

$$F_c = \frac{\text{Quadrado médio de tratamentos}}{\text{Quadrado médio do resíduo}} \sim F_{(I-1,I(J-1))}$$

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.

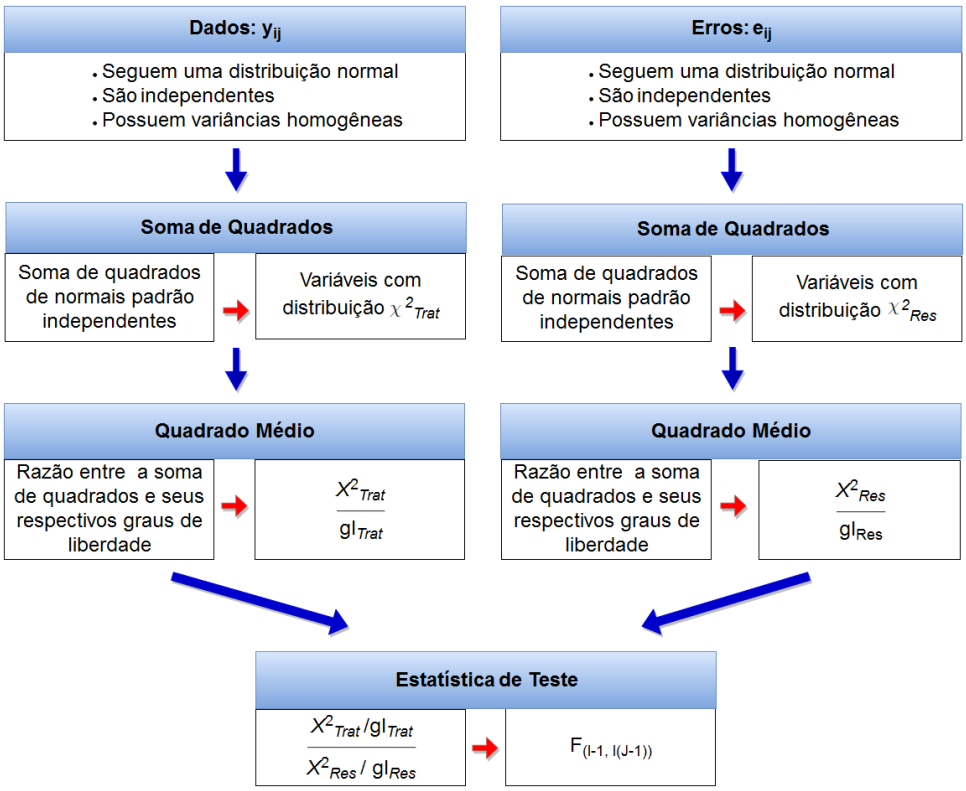

Figura 3 - Processo da análise de variância.

Vale destacar que, no meio acadêmico, este teste é amplamente utilizado e é considerado, por muitos, como o melhor teste para comparação de médias.

2.2 O teste de Welch (W)

O teste de Welch foi desenvolvido por um pesquisador de mesmo nome em 1951 (WELCH, 1951). Este é um teste paramétrico, pois assume que os dados seguem uma distribuição normal independente. Este teste tem o objetivo de ponderar o teste F na presença de heterogeneidade da variância dos erros.

Sejam W_i, $\bar{Y}^*$ e Ω dados por:

$$W_i = \frac{J}{S_i^2}; \qquad \bar{Y^*} = \frac{\sum_{i=1}^{I} W_i \bar{Y_i}}{\sum_{i=1}^{I} W_i}; \qquad \Omega = \frac{\sum_{i=1}^{I} \left(1 - \frac{W_i}{\sum_{i=1}^{I} W_i}\right)^2}{J-1}$$

em que S_i^2 e $\bar{Y_i}$ representam a variância e a média amostral do *i-ésimo* tratamento, respectivamente.

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

$$W = \frac{\frac{\sum_{i=1}^{I} W_i \left(\bar{Y_i} - \bar{Y}^*\right)^2}{(I-1)}}{1 + \frac{2(I-2)\Omega}{(I^2-1)}} \sim F_{\left(I-1, \frac{I^2-1}{3\Omega}\right)} \tag{4}$$

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).

2.3 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. O teste de James é definido por partes, de acordo com o tamanho da amostra e, neste trabalho, abordou-se o caso em que o tamanho da amostra é grande. Nessas condições, a estatística de teste é dada por:

$$JA = \frac{\sum_{i=1}^{I} W_i (\bar{Y_i} - \bar{Y})^2}{I-1} \sim \chi^2_{I-1} \tag{5}$$

em que $\bar{Y_i}$ é a média do *i-ésimo* tratamento, $\bar{Y}$ é a média geral do experimento e W_i é como já definido para o teste de Welch.

2.4 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, e consiste no método de reamostragem *bootstrap* paramétrico, que é utilizado quando a distribuição da variável aleatória é conhecida, com parâmetros desconhecidos (KRISHNAMOORTHY; LU; MATHEW, 2007).

A estatística de teste (T_{NB}) é dada por:

$$T_{NB}(Z_i, \chi^2_{J-1}; S_i^2) = \sum_{i=1}^{I} \frac{Z_i^2(J-1)}{\chi^2_{J-1}} - \frac{\left[\sum\limits_{i=1}^{I}\left(\frac{\sqrt{J}Z_i(J-1)}{S_i\chi^2_{J-1}}\right)\right]^2}{\sum\limits_{i=1}^{I}\left(\frac{J(J-1)}{S_i^2\chi^2_{J-1}}\right)} \tag{6}$$

Para um determinado $(s_1^2, \ldots, s_I^2)$ realização de $(S_1^2, \ldots, S_I^2)$ e nível α, o teste *bootstrap* paramétrico rejeita H_0 quando

$$P(T_{NB}(Z_i, \chi^2_{J-1}; s_i^2) > T_{N0}) < \alpha \tag{7}$$

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 (8).

$$T_N\left(\bar{Y}_1, \ldots, \bar{Y}_I; {S_1}^2, \ldots, {S_I}^2\right) = \sum_{i=1}^{I} \frac{J}{{S_i}^2}\bar{Y}_i^2 - \frac{\left[\sum\limits_{i=1}^{I} J\frac{\bar{Y}_i}{{S_i}^2}\right]^2}{\sum\limits_{i=1}^{I}\frac{J}{{S_i}^2}} \tag{8}$$

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

1. Para um dado $J, (\bar{y_1}, \ldots, \bar{y_I})$ e $(s_1^2, \ldots, s_I^2)$:
2. Calcular T_N em (8) e chame de T_{N0}
3. Seja B o número de reamostragens, então para $b = 1, \ldots, B$:
4. Gerar $Z_i \sim N(0,1)$ e $\chi^2_{J-1}, i = 1, \ldots, I$
5. Calcular $T_{NB}(Z_i, \chi^2_{J-1}; s_i^2)$ usando (6)

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

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

2.5 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 (REDDY; KUMAR; RAMU, 2010). 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:

Reúna 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

$$\{Y^*_{bt},\ t = 1, \ldots, IJ \text{ e } \ b = 1, 2, \ldots, B\}$$

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

$$\bar{Y}_b \quad = \quad \frac{1}{IJ} \sum_{t=1}^{IJ} Y^*_{bt},\ b = 1, 2, \ldots, B$$

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

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

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 dadas por

$$LDI = \bar{Y}^* - Z_{\frac{\alpha}{2}} S^*_{IJ} \qquad LDS = \bar{Y}^* + Z_{\frac{\alpha}{2}} S^*_{IJ}$$

em que $Z_{\frac{\alpha}{2}}$ é o valor crítico da normal padrão 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:

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

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

Zhou e Wong (2011) desenvolveram um método *bootstrap* não-paramétrico para a comparação de médias como alternativa à análise de variância. A estatísitca de teste 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.

2.7 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.

A estatística do teste de Kruskal-Wallis pode ser obtida por:

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

em que 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.

3 Material e métodos

Neste artigo, todos os testes comparados consideram o modelo descrito em (1), com média geral μ arbitrada zero, sem perda de generalidade, τ_i efeito fixo do *i-ésimo* tratamento, com a restrição $\sum_i \tau_i = 0$ e os erros $e_{ij} \sim N(0; \sigma_i^2)$. Como adotou-se a quebra da pressuposição de homogeneidade da variância dos erros, as variâncias residuais dos tratamentos (σ_i^2) podem ser diferentes. Logo, elas definem o grau de heterocedasticidade (δ), que foi imposto segundo a razão entre a maior e a menor variância dentre os tratamentos, assim como outros trabalhos presentes na literatura, como o artigo de Nogueira e Pereira (2013):

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

Os sete testes mencionados foram avaliados via simulação Monte Carlo, computando-se suas taxas de erro tipo I e poder.

Para avaliação da taxa de erro tipo I foram gerados dados segundo 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:

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

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 a 5%.

Para o estudo 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).

3.1 Aplicação - 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.

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, juntamente com o resultado da aplicação de queijo minas padrão.

4.1 Erro tipo I

Primeiramente, o erro tipo I de todos os testes foi avaliado em um gráfico como apresentado na Figura 4 juntamente com as informações descritas na Tabela 1.

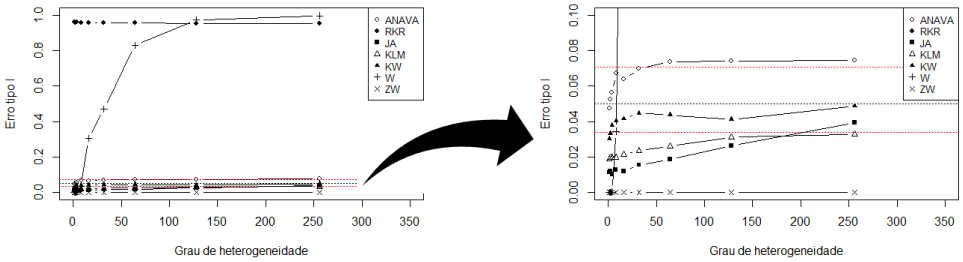

Figura 4 - Taxa do erro tipo I dos testes em relação ao grau de heterocedasticidade, seguido da ampliação do gráfico evidenciando a região do $IC(p)_{1-\alpha}$.

Observe 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 sensível 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 heterocedasticidade (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.

Em seguida, destaca-se o teste F, que se mostrou exato com $1 \leq \delta \leq 32$ 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.

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 ($\sigma^2; 0 < \sigma_i^2 \leq 1$).

Tabela 1 - 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
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

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

O teste de James se mostrou conservador para todos os cenários, exceto quando o grau de heterogeneidade é igual a 256 pois, neste caso, o teste é exato. Já o teste de Welch revelou-se o mais sensível à 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.

O teste RKR praticou mais de 90% de erro tipo I, ou seja, rejeitou a hipótese nula verdadeira pelo menos 90% das vezes. Já o teste ZW, mesmo no contexto de extrema heterogeneidade, não rejeitou a hipótese nula em nenhum dos casos.

Assim, de maneira geral, tem-se indícios que os testes foram pouco sensíveis à falta de homogeneidade de variâncias, exceto para o teste de Welch, que foi proporcionalmente afetado à medida em que o grau de heterogeneidade aumentou.

Também foi realizada a análise do erro tipo I de todos os testes separadamente, revelando com mais detalhe o comportamento dos mesmos em cada cenário simulado.

A Figura 5 mostra o desempenho do teste de Kruskal-Wallis. 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 e grau de heterogeneidade maior que 128. Desta forma, há de se destacar que o teste 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.

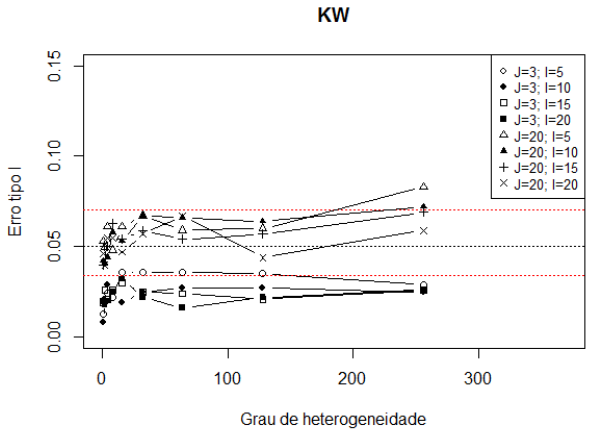

Figura 5 - Taxa do erro tipo I do teste KW ao longo do grau de heterogeneidade.

A Figura 6 mostra o comportamento do teste F, no contexto de ANAVA.

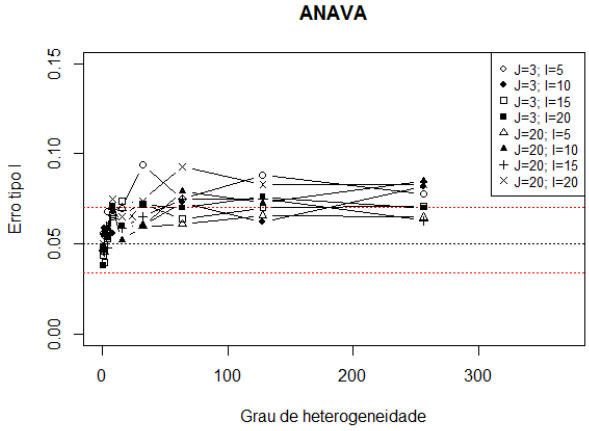

Figura 6 - Taxa do erro tipo I do teste F no contexto de ANAVA.

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.

O teste KLM, apresentado na Figura 7 revela o comportamento conservador do teste em termos de erro tipo I. Entretanto, para os cenários simulados com pouca repetição, 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 aumenta 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.

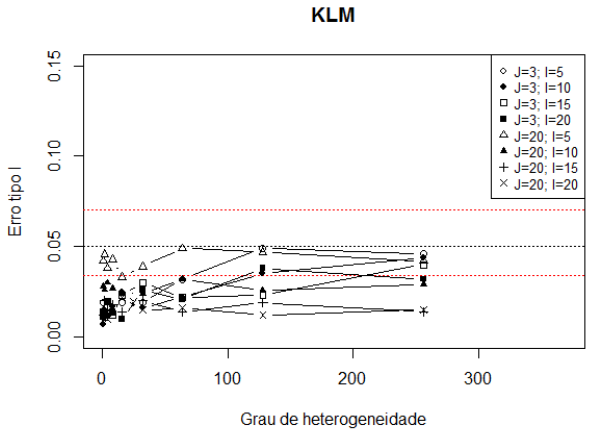

Figura 7 - Taxa do erro tipo I do teste *bootstrap* paramétrico KLM.

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).

A Figura 8 revela o comportamento do teste de Welch, que 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. Este comportamento se mantém em todos os cenários simulados.

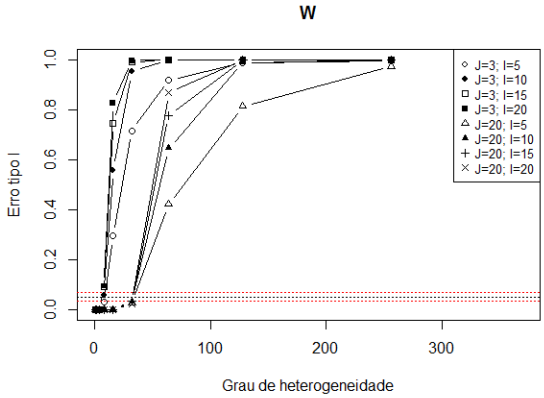

Figura 8 - Taxa do erro tipo I praticada pelo teste de Welch.

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.

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

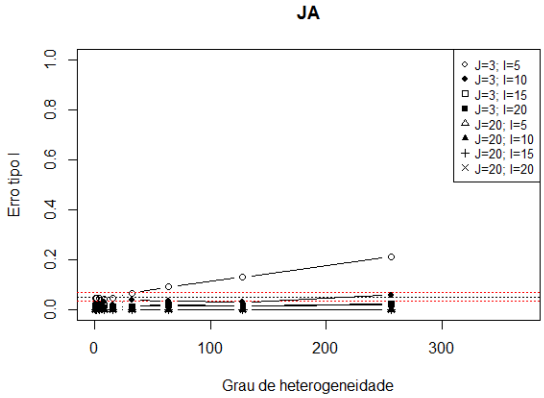

Figura 9 - Taxa do erro tipo I praticada pelo teste de James.

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.

A Figura 10 mostra que o teste RKR é extremamente liberal, praticando mais de 75% de taxa de erro tipo I para todos os cenários simulados.

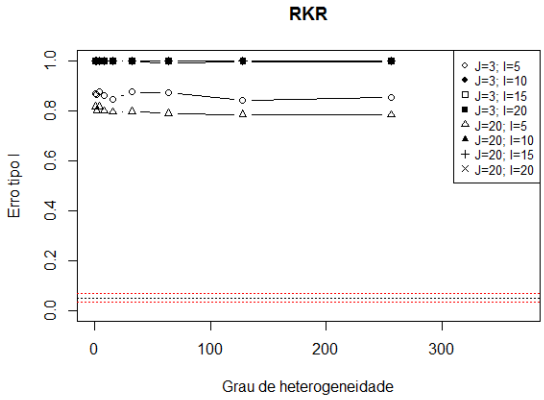

Figura 10 - Taxa do erro tipo I do teste *bootstrap* não-paramétrico RKR.

Já o teste ZW não rejeitou a hipótese nula, quando esta é verdadeira em nenhum dos cenários adotados, ou seja, praticou 0% de erro tipo I, dispensando a apresentação de seu gráfico.

4.2 Poder

Agora, em relação ao poder, foi feito um estudo 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 ϕ é igual a 1, sendo que os demais apresentam comportamento semelhante a este.

A Figura 11 e a Tabela 2 mostram o desempenho geral dos testes em relação ao poder. Todos os testes estabilizaram o comportamento a partir do grau de heterogeneidade igual a 64, no entanto, o teste de Welch se mostrou mais sensível aos graus de heterogeneidade.

A Figura 11 e a Tabela 2 revelam 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 pressuposição da normalidade, Ferreira, Rocha e Mequelino (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

Tabela 2 - 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
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

que devem estar igualmente espaçados entre 0 e 1.

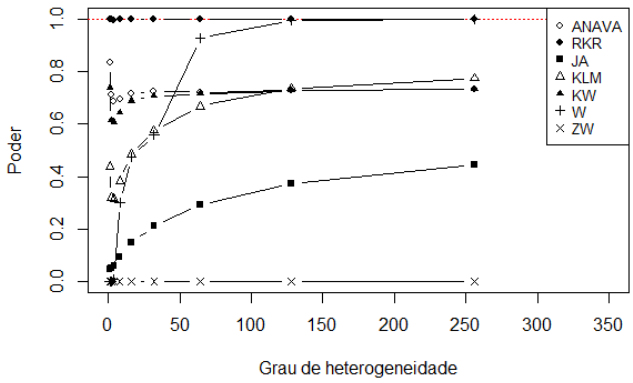

Figura 11 - Poder de todos os testes em relação ao grau de heterogeneidade.

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 11 e a Tabela 2 também mostram 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 4. O teste KLM se mostrou sensível à heterogeneidade crescente da variância dos erros para a avaliação do poder, já que quanto maior a heterogeneidade mais poderoso se revela o teste, apresentando o poder igual ou superior ao teste F e KW a partir de $\delta = 128$. Já o teste de Welch, por sua vez, também apresentou aumento de poder à medida que o grau de heterogeneidade cresceu. Contudo, a Figura 4 revelou que a taxa de erro tipo I do teste apresenta o mesmo comportamento com os graus crescentes de heterogeneidade.

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, concluindo que os testes de JA e W não apresentaram um resultado eficaz para o poder.

Já o teste de ZW não teve um comportamento satisfatório com relação ao poder, ficando abaixo de 40%.

Também foi feito um estudo do poder de todos os testes separadamente, dando maior ênfase na influência do número de tratamentos e repetições no comportamento dos testes.

A Figura 12 revela o desempenho do teste de Kruskal-Wallis, que se mostrou pouco sensível aos graus crescentes de heterogeneidade.

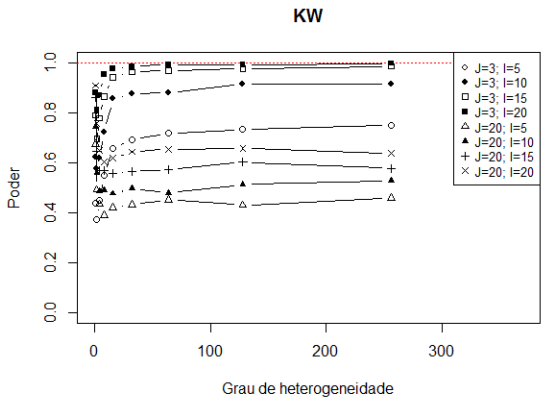

Figura 12 - Poder do teste de Kruskal-Wallis ao longo dos graus de heterogeneidade.

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 aumentando, 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 13 mostra 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. 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.

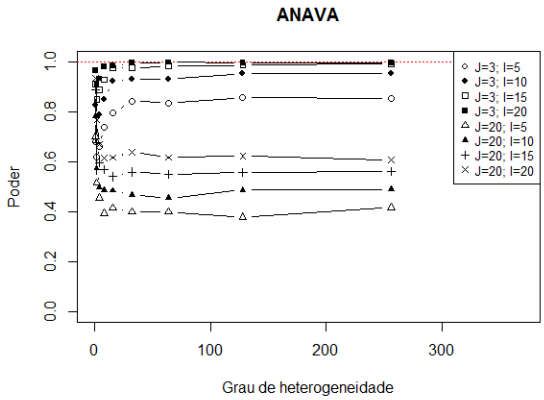

Figura 13 - Poder do teste F ao longo dos graus de heterogeneidade.

O poder do teste KLM é apresentado na Figura 14. Para J=3, sob homogeneidade o teste *bootstrap* paramétrico de Krishnamorthy, Lu, Mathew não apresenta tão bom desempenho quanto na quebra de pressuposição. Neste caso, poder é diretamente proporcional ao grau de heterogeneidade. Já para J=20, o teste perde poder com pouca heterogeneidade, entretanto ele volta a crescer à medida que a heterogeneidade cresce.

O comportamento do teste de Welch pode ser visto na Figura 15. 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. Já quando os graus de heterogeneidade crescem, o poder do teste de Welch 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 16 revela 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

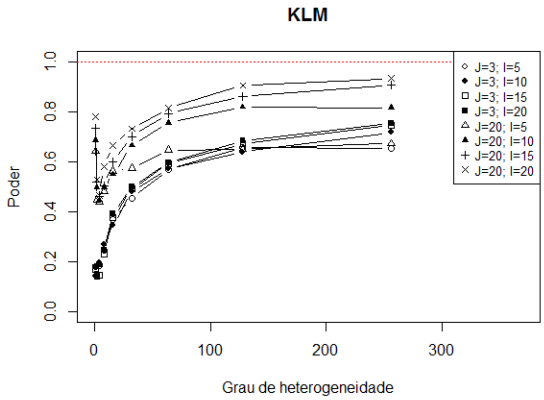

Figura 14 - Poder do teste *bootstrap* paramétrico de Krishnamorthy, Lu, Mathew.

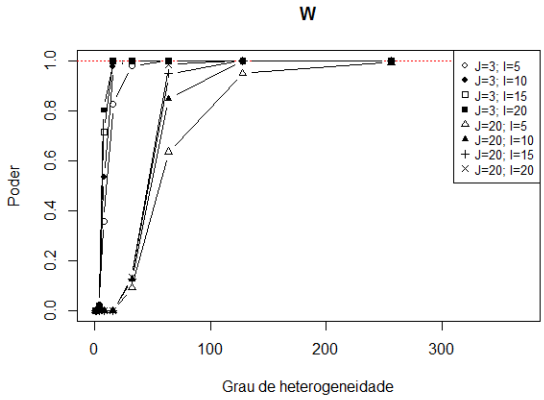

Figura 15 - Poder do teste de Welch em relação aos graus de heterogeneidade.

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.

O desempenho, em termos de poder, do teste RKR, pode ser comprovado na Figura 17. Observe que em todos os cenários 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.

Já o teste ZW nunca rejeita a hipótese nula, mesmo que esta não seja, de fato, verdadeira. Assim como no erro tipo I, o teste não foi afetado pelos graus de heterogeneidade, já que não possui poder. Desta forma, para todos os cenários

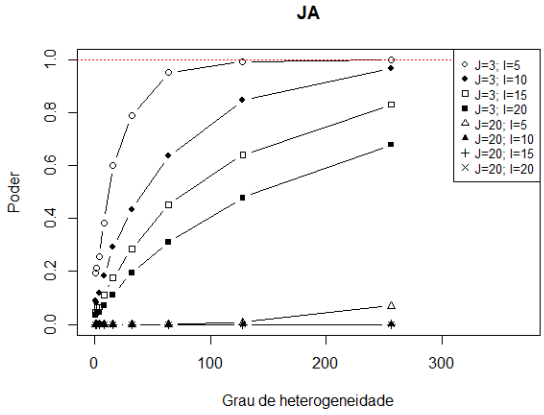

Figura 16 - Poder do teste de James em relação aos graus de heterogeneidade.

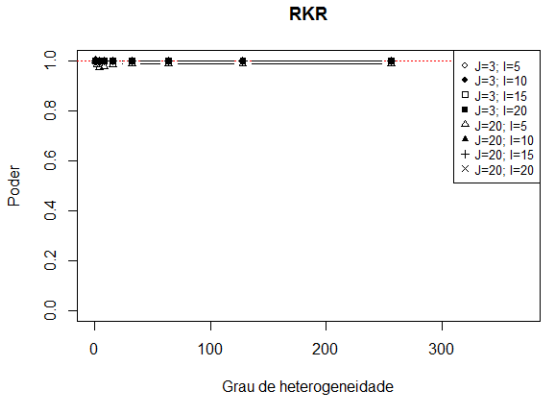

Figura 17 - Poder do teste *bootstrap* não paramétrico de Rddy, Kumar e Ramu.

simulados o teste praticou 0% de poder.

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 18, para $\delta \in \{1, 2, 128, 256\}$.

A Figura 18 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,

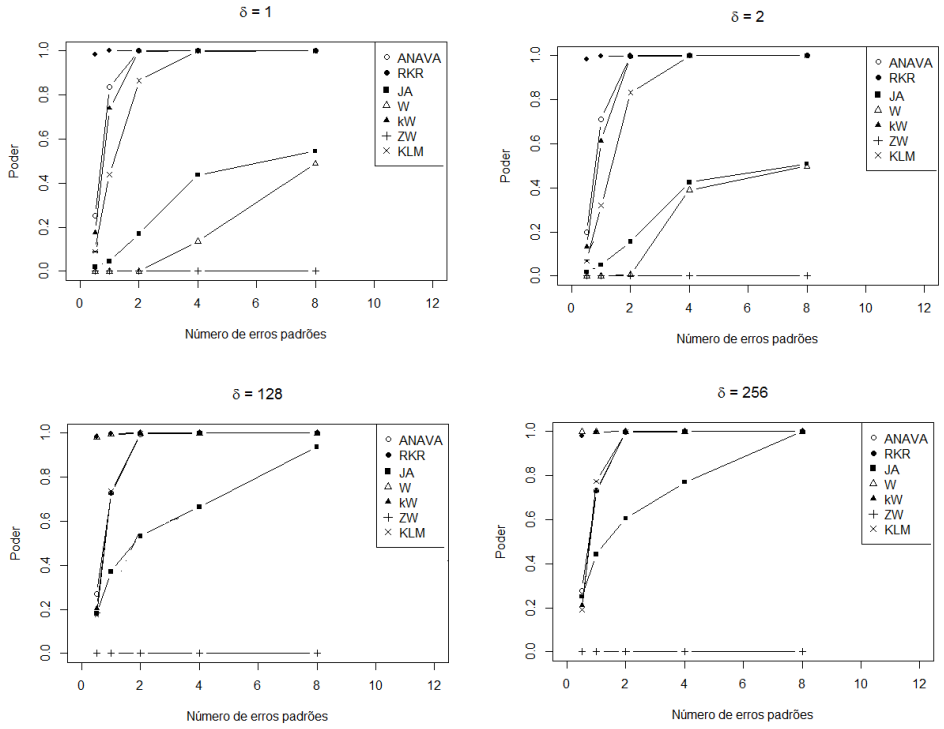

Figura 18 - Poder dos testes para $\delta \in \{1, 2, 128, 256\}$.

bem distintos. Agora o teste ZW não conseguiu 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 Resultado da 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 $\delta = 10,250$.

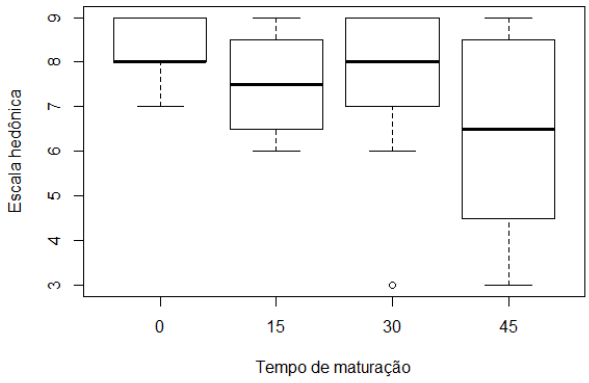

Figura 19 - *Boxplot* da variância dos erros ao longo do tempo de maturação.

É possível notar, na Figura 19, de que maneira a variabilidade das notas, segundo a escala não estruturada de 9 pontos, aumentou ao longo do tempo de maturação do queijo minas padrão. Este comportamento era 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.

Posteriormente foram aplicados os testes estudados neste trabalho para os resultados da variável cor. Na Tabela 3 estão apresentadas as médias em cada tempo de maturação.

Tabela 3 - Médias em cada tempo de maturação

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

Os valores-p de todos os testes, à 5% de significância, podem ser vistas na Tabela 4.

Observe que os testes JA, W, ZW, KW, ANAVA e KLM 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 3 a apresentação das médias.

Tabela 4 - Valores-p dos oito testes em estudo para a aplicação do queijo minas padrão

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

Nota: * o teste RKR não possue valor-p, pois é baseado em cartas de controle, neste caso, 1 significa que deve-se rejeitar H_0.

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 5.

Tabela 5 - Erro tipo I e erro tipo II para a aplicação do queijo minas padrão

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

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

Na Tabela 5 é possível notar 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.

Conclusões

Os resultados deste estudo revelaram quais são os testes com melhor desempenho para a comparação de médias na quebra da homogeneidade da variância dos erros. Tanto em 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.

O teste *bootstrap* não-paramétrico de Zhow e Wong apresentou o pior comportamento deste estudo, já que o mesmo nunca rejeita a hipótese nula, mesmo que ela seja falsa, ou seja, não consegue detectar diferença entre as médias.

Agradecimentos

O presente artigo foi realizado com apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil.

OLIVEIRA, H. M. P.; FERREIRA, E. B. Test Comparison for means equality under heteroscedasticity: simulation and application. *Rev. Bras. Biom.,* Lavras, v.35, n.1, p.132-159, 2017.

- *ABSTRACT: Generally, comparison of several means is made by testing to detect the existence of differences between treatments. One of the most widely used tests in this case is the F test, within the Analysis of Variance. However, your credibility is linked to the fulfillment of four assumptions: additivity of the terms of the model, errors should follow a normal distribution, be independent and possess homogeneous variances. However, in the literature there are alternative tests to F, when you break any of those assumptions. The aim of this study was to provide the comparison of seven tests for equality of means under heteroscedasticity. The evaluation of the tests was done by analyzing rates of type I error and power over scenarios resulting from the combination of treatments numbers repetitions degree of heterogeneity and standard error difference between means. Overall, the tests showed little sensitiveness to the increasing of heterogeneity, which it was not observed with the Welch test. Based on the evaluated conditions, the tests that presented the best performance were Kruskal-Wallis and F, followed by the parametric bootstrap Krishnamoorthy, Lu and Mathew. The tests with the worst performances were the bootstrap nonparametric Reddy, Kumar and Ramu and the Zhow and Wong.*

- *KEYWORDS: Computational statistics; experimental design; inference.*

Referências

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

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

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.

EFRON, B. Bootstrap methods: another look at the jack-knife. In: *Annals of Statistic*, Hayward. *Anais...* Hayward, n, 1, p. 1-26, 1979.

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, p.1-37, 1974.

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.

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.

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

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.

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.

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

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.

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.

Recebido em 26.04.2016.

Aprovado após revisão em 13.07.2016.