Desempenho de testes para homogeneidade de variâncias em delineamentos inteiramente casualizados

Denismar A. Nogueira[1†], *Giselle M. Pereira*[2]

[1] *Professor Adjunto III, Instituto de Ciências Exatas, Universidade Federal de Alfenas (Unifal-MG).*
[2] *Graduanda em Ciências Biológicas, Universidade Federal de Alfenas (Unifal-MG).*

Resumo: *Por volta de 1920, Fisher propôs a análise de variância, que visa a decomposição da variação total em fontes de variação conhecidas. Para validade dos resultados da análise de variância, esta depende que algumas condições pressupostas sejam atendidas. Uma das razões de se ignorar a checagem das pressuposições é a dificuldade de encontrar testes adequados para tal finalidade. A hipótese de homogeneidade de variâncias é o pressuposto mais importante da análise de variâncias. A violação de qualquer outra suposição pode resultar em heterogeneidade do erro experimental, e isso reforça ainda mais a necessidade de seu estudo. Com isso, os objetivos desta pesquisa foram implementar e estudar o desempenho no controle do erro tipo I e poder de 15 testes para homogenenidade de variâncias, utilizando simulação de Monte Carlo, em variadas configurações de tratamentos e repetições. Em situações de normalidade e delineamento inteiramente casualizado as propostas baseadas na verossimilhança apresentaram os melhores resultados seguidas da proposta bayesiana apresentada por Samiuddin. As variações do teste de Levene tiveram resultados modestos em situações de poucas repetições o que também ocorreu com as de Cochran.*

Palavras-chave: Heterocedasticidade, Erro tipo I, Poder.

Abstract: *By 1920, Fisher proposed the analysis of variance, which aims to decompose the total variation in sources of variation known. For validity of the results of the analysis of variance, this depends on some conditions are met presupposed. One reason to ignore checking of assumptions is the difficulty of finding adequate tests for such purpose. The assumption of homogeneity of variances is the most important assumption of the analysis of variance. Violation of any other assumption may result in heterogeneity of experimental error, and this further reinforces the need for their study. Thus, the objectives of this research were to study the performance and implement the control of type I error and power of 15 tests for homogenenidade variances using Monte Carlo simulation, in varied settings of treatments and replicates. In normal and randomized design proposals based on the likelihood showed the best results followed the proposal presented by Bayesian Samiuddin. Variations Levene's test had modest results in situations of low reps which also happened to Cochran.*

Keywords: Heteroscedasticity, Type I error, power.

[†] *Autor correspondente:* denismar@unifal-mg.edu.br.

Introdução

Comparações de variâncias sempre ocorrem em várias áreas da estatística, com o intuito de minimizar, controlar e acompanhar a variabilidade, o que é de fundamental importância na produção, no melhoramento genético, no controle de qualidade etc.

Quando se está interessado na realização de inferências há sempre a necessidade de verificar algumas pressuposições dos métodos utilizados. Para ser valida uma análise de variância, esta também depende que algumas condições pressupostas sejam atendidas. A normalidade dos erros é uma delas. Muitos autores consideram que a de maior importância seja a de homogeneidade de variâncias dos erros.

Essas pressuposições muitas vezes não são checadas e, desta forma, podem comprometer a validade dos resultados dos testes e das estimações realizadas. Uma das razões de se ignorar a checagem das pressuposições para validade da análise é a dificuldade de se encontrarem recursos computacionais. A maior parte dos softwares estatísticos não avaliam estas pressuposições ou até não possuem rotinas para isso. Na literatura, os testes existentes para se verificar a hipótese de homogeneidade de variâncias são específicos para certos modelos, o que dificulta a sua aplicação.

A importância do teste de homogeneidade de variâncias em muitas áreas da Experimentação é baseada na premissa de que muitos testes de hipóteses sobre médias ou efeitos de tratamentos são realizados pressupondo que as variâncias das populações amostradas sejam iguais. A violação dessa hipótese pode afetar o desempenho do método e comprometer os resultados de diferentes formas, segundo Johnson e Wichern (1998). Vários modelos estão disponíveis na literatura para este tipo de estudo e é sabido que a heterocedasticidade dos resíduos é um fator que pode afetar a inferência, podendo impactar diretamente nas conclusões (FERREIRA et al., 2006). A presença de heterogeneidade de variâncias pode também ter um sério efeito na validade do teste F, especialmente quando os tamanhos des amostras são desbalanceados (O'BRIEN, 1978; KEYES; LEVY, 1997).

Segundo Gomez e Gomez (1984), a heterogeneidade de variância pode ocorrer de duas maneiras. Em uma delas, a variância ocorre sem nenhuma relação com a média e na outra existe uma relação entre estas. Em ciências biológicas é comum a presença de correlação positiva entre média e variância. Grupos com grandes médias tendem a apresentar grandes variâncias e grupos de pequenas médias apresentam pequenas variâncias.

Um dos objetivos desse estudo foi testar o desempenho no controle da taxa de erro tipo I e o poder dos testes utilizados por pesquisadores diante de situações em que a pressuposição de homogeneidade de variância não é atendida. Um teste é classificado como rigoroso, quando a taxa de erro tipo I cometida por ele é menor que o nível nominal de significância e considerado como liberal se este for maior. O erro do tipo I é o erro cometido ao rejeitar H_0 quando, na realidade, é verdadeira, ou seja, dizer que há diferença entre as variâncias, sendo que essa diferença não existe. A probabilidade de cometer este erro é designada por α. O segundo é o erro tipo II que se comete ao aceitar H_0 quando, na realidade, é falsa, ou seja, afirmar que as variâncias são todas iguais quando na verdade existe uma diferença entre elas. A probabilidade de cometer este erro do tipo II é designada por β. Em um teste de hipóteses é obviamente desejável que se reduza ao mínimo as probabilidades α e β (STEEK; TORRIE, 1980).

O Poder de um teste tem por objetivo conhecer o quanto este controla o erro do tipo II, ou qual a probabilidade de rejeitar a hipótese nula se realmente for falsa. Na prática, é importante que se tenham testes com níveis de significância próximos do nível de significância nominal e que o poder seja elevado, mesmo em situações de amostras pequenas. O poder de um teste de hipóteses é dado pelo complementar de β, e é afetado diretamente pelo tamanho da amostra, pelo nível de significância adotado e pelo verdadeiro valor do parâmetro a ser testado.

Pretendeu-se com esse trabalho, por meio de simulações de Monte Carlo, avaliar o desempenho em relação ao erro tipo I e poder, de 15 testes para homogeneidade de variâncias em 16 combinações diferentes para o número de tratamentos e repetições em três diferentes níveis de significância. Foram abordadas apenas situações envolvendo normalidade e balanceamento.

Metodologia

Para exemplificar uma hipótese de homogeneidade de variâncias, sejam t amostras de tamanho n cada, provindas de suas respectivas populações de tratamentos. Para tal considerou-se o seguinte modelo:

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

em que y_{ij} a j-ésima observação da i-ésima população de tratamento, para i = 1, ..., t e j = 1, ..., n. Ainda no modelo μ pode ser a média geral, τ_i o efeito fixo do i-ésimo tratamento e ε_{ij} o erro experimental associado a cada observação, assumindo distribuição normal com média 0 e variância σ_i^2.

Como definido, deseja-se verificar a hipótese $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_t^2 = \sigma^2$.

Para testar a hipótese de homogeneidade (H_0) inicia-se com a proposta apresentada por Bartlett, que é considerada por muitos como o melhor teste para comparação de variâncias que se baseia na razão de verossimilhanças dada por:

$$\Lambda = \frac{(2\pi)^{-\frac{2}{n}}(\sigma^2)^{-\frac{2}{n}}exp^{-\frac{n}{2}}}{(2\pi)^{-\frac{2}{n}}\prod_{i=1}^{t}(\sigma_i^2)^{-\frac{2}{n_i}}exp^{-\frac{n}{2}}} = \frac{(\sigma^2)^{-\frac{2}{n}}}{\prod_{i=1}^{t}(\sigma_i^2)^{-\frac{2}{n_i}}} = \frac{\prod_{i=1}^{t}(\sigma_i^2)^{\frac{2}{n_i}}}{(\sigma^2)^{\frac{2}{n}}}. \tag{2}$$

Sob H_0, $-2Ln(\Lambda)$ tem distribuição assintótica de qui-quadrado com $\nu = t - 1$ graus de liberdade. Sob H_1 têm-se t médias e t variâncias e, sob H_0, t médias e 1 variância comum a todas. Assim, $B_0 = nLn(\widehat{\sigma}^2) - \sum_{i=1}^{t}(n_i Ln(\widehat{\sigma}^2))$ tem distribuição assintótica de qui-quadrado com $\nu = t - 1$ graus de liberdade, sob H_0. Bartlett (1937) propôs uma correção e mudanças para melhorar a aproximação e, desta forma, a estatística de Bartlett (1937) para o teste da hipótese é:

$$B_1 = \frac{(n-t)Ln(S_p^2) - \sum_{i=1}^{t}(n_i - 1)Ln(S_i^2)}{1 + \frac{1}{3(t-1)}\left[\sum_{i=1}^{t}\left(\frac{1}{n_i - 1}\right) - \frac{1}{n-t}\right]}, \tag{3}$$

sendo S_i^2 o estimador da variância amostral e $S_p^2 = \sum_{i=1}^{n}\nu_i S_i^2 \Big/ (n-t)$ o estimador não-viesado da variância comum.

Em 1989, Boos e Brownie propuseram uma modificação no teste de Bartlett, que agora considera o estimador do coeficiente de curtose (β_2). Este fato se deve a tentativa de minimizar a influência da ausência de normalidade das populações amostrais que afetam, assim, o controle das taxas de erro tipo I e o poder (Ferreira, 2005). O teste aproxima de $(\beta_2 - 1)\chi^2_{(t-1)} \Big/ 2$ sob H_0, em que $\beta_2 = E(Y-\mu)^4 \Big/ \sigma^4$ é o coeficiente de curtose da população amostral. Assim, a estatística do teste é dada:

$$B_2 = \frac{2}{\left[\dfrac{n\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2}{[\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2]^2}\right]-1}\left[(n-t)Ln(S_p^2)-\sum_{i=1}^{t}(n_i-1)Ln(S_i^2)\right], \quad (4)$$

que segue assintoticamente a distribuição de qui-quadrado com $\nu = t-1$ graus de liberdade, sob H_0. O $\overline{y}_{i.}$ é a média amostral do i-ésimo tratamento.

Outra proposta avaliada, seguindo a mesma linha de raciocínio, também apresentada por Boos e Brownie (1989) é considerar uma outra opção de correção usando a curtose e a estatística do teste é dada por:

$$B_3 = \frac{2B_1}{\left[\dfrac{n\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2}{[\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2]^2}\right]-1}, \quad (5)$$

que segue também assintoticamente a distribuição de qui-quadrado com $\nu = t-1$ graus de liberdade, sob H_0. Para ambos os casos com correção de curtose se o estimador do coeficiente de curtose for menor que 1 a estatística é considerada nula e o valor-p será 1.

Dixon e Massey um pouco antes, em 1969, propuseram nova versão para o teste de Bartlett que se considerou, neste estudo, como a versão 4 do teste de Bartlett. A proposta é diferente das propostas anteriores, e a estatística $B_4 = LM/(b-M)$ segue uma distribuição F com $\nu_1 = t-1$ e $\nu_2 = 3(t-1)/A^2$ graus de liberdade, em que

$$M = (n-t)Ln(S_p^2) - \sum_{i=1}^{t}(n_i-1)Ln(S_i^2);$$

$$A = \frac{1}{3(t-1)}\left[\sum_{i=1}^{t}\left(\frac{1}{n_i-1}\right)-\frac{1}{n-t}\right];$$

$$L = 3(t-1)/A^2;$$

$$b = L/(1-A+2/L).$$

Outra proposta muito utilizada na literatura foi apresentada por Levene (1960) para a comparação de variâncias e baseia-se em uma transformação nos dados originais com a realização de uma análise de variância com um fator onde o teste F permite avaliar a existência de efeitos entre os tratamentos. Na verdade estes efeitos são as variâncias, testando-se assim, a presença de homogeneidade de variâncias. A transformação nada mais é do que a obtenção dos resíduos. A ideia acabou criando uma família de testes com modificações na transformação. Para tanto, seja $\overline{y}_i$ a média amostral da i-ésima população de tratamentos e seja $z_{ij} = |y_{ij} - \overline{y}_{i.}|$ uma transformação realizada nos valores originais, a estatística do teste é dada por:

$$L_5 = \frac{\sum_{i=1}^{t} n_i \left(\overline{z}_{i.} - \overline{z}_{..}\right)^2 \Big/ (t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{n_i} \left(\overline{z}_{i.} - \overline{z}_{..}\right)^2 \Big/ (n-t)}, \tag{6}$$

sendo $\overline{z}_{i.}$ a média de cada tratamento da variável transformada e $\overline{z}_{..}$ a média geral da variável transformada. Sob H_0, a estatística L_5 segue uma distribuição F, com $\nu 1 = t - 1$ e $\nu 2 = n - t$ graus de liberdade.

Uma segunda opção utiliza uma outra transformação definida por $z_{ij} = \left(\overline{y}_{ij} - \overline{y}_{i.}\right)^2$. A estatística permanece a mesma e será denotada por L_6.

Em 1974, uma sugestão foi apresentada por Brown e Forsythe para melhoria do teste de Levene, que considera a utilização do estimador da mediana no lugar da média. Na literatura, esta versão é conhecida como teste de Brown-Forsythe. Para tanto, seja $\widetilde{y}_{i\cdot} = md_i$ a mediana amostral da i-ésima população de tratamentos e seja $z_{ij} = |y_{ij} - \widetilde{y}_{i.}|$ uma transformação realizada nos valores originais. A estatística permanece a mesma utilizada no teste de Levene (1960) e trataremos esta de L_7.

Segundo Rubin (1983) e Mehrotra (1997) citados por Argaç (2002) há uma falha no teste original de Brown e Forsythe (1974), a falha é especialmente na consideração dos graus de liberdade da aproximação Ã distribuição F. A estatística Brown-Forsythe e Levene, utilizam $\nu 1 = t - 1$ onde, segundo estes autores, deveriam utilizar uma aproximação proposta por Box (1954), ficando então os graus de liberdade do numerador da seguinte forma:

$$\nu 1 = \frac{\left[\sum_{i=1}^{t}\left(1 - \frac{n_i}{n}\right) S_i^2\right]^2}{\sum_{i=1}^{t} S_i^4 + \left[\sum_{i=1}^{t} \frac{n_i S_i^2}{n}\right]^2 - 2\sum_{i=1}^{t} \frac{n_i S_i^4}{n}}. \tag{7}$$

O $\nu 2$ e a estatística, que trataremos de L_8, são os mesmos do teste de Brown e Forsythe (1974). Sob H_0, a estatística L_8 segue uma distribuição F, com $\nu 1$ dado na equação (7) e $\nu 2 = n - t$ graus de liberdade.

Samiuddin (1976) propôs um teste para avaliar a hipótese de igualdade das variâncias usando a análise bayesiana. Samiuddin considerou uma distribuição a priori não informativa para μ_i e σ_i^2. A verossimilhança assumida no teste é proporcional a:

$$\prod_{i=1}^{t}\left(\frac{1}{\sigma_i}\right) exp\left[-\frac{1}{2}\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2 \Big/ \sigma_i^2\right].$$

A posteriori conjunta é definida como o produto das duas distribuições (Priori e Verossimilhança) e a integração da posteriori em relação aos μ_i's nos permite obter a distribuição marginal de σ_i^2's. O autor utiliza uma transformação de Wilson-Hilferty (Wilson e Hilferty, 1931) para aproximar uma qui-quadrado pela normal. O autor também mostra que $\phi_i = (1/\sigma_i^2)^{\frac{1}{2}}$ segue aproximadamente uma distribuição normal com média m_i e variância a_i^2 sendo, portanto:

$$mi = \left(\frac{n_i - 1}{\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2}\right)^{\frac{1}{2}} \left[1 - \frac{2}{9(n_i - 1)}\right]$$

$$a_i^2 = \frac{2}{\left[9\left(\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2\right)^{\frac{2}{3}}(n_i - 1)^{\frac{1}{3}}\right]}.$$

Quando a hipótese de homogeneidade de variâncias é verdadeira então a estatística do teste bayesiano é dado por:

$$S_9 = \sum_{i=1}^{t}(m_i - m)^2/a_i^2,$$

sendo $m = \left(\sum_{i=1}^{t} m_i/a_i^2\right) \Big/ \left(\sum_{i=1}^{t} 1/a_i^2\right)$. Sob H_0, a estatística do teste bayesiano tem distribuição assintótica de qui-quadrado com $\nu = t - 1$ graus de liberdade.

Segundo O'Neill e Mathews (2000), os softwares estatísticos que oferecem as variadas formas do teste de Levene ignoram o fato dos delineamentos desbalanceados, e de que as variáveis analisadas podem ser não-normais e, por isso, a estatística do teste F utilizada pelo teste de Levene pode não seguir uma distribuição F. A proposta se baseia na estimação utilizando a análise de variância pelo método de mínimos quadrados ponderados (MQP). A estatística analisa a variável $z_{ij} = |y_{ij} - \overline{y}_{i.}|$ e a estatística do teste com base na análise de variância por mínimos quadrados ponderados é dada por:

$$OM_{10} = \frac{N - t}{t - 1} \frac{\sum_{i=1}^{t} w_{0i}(\overline{z}_{i.} - \overline{\overline{z}})^2}{\sum_{i=1}^{t} w_{1i} \sum_{j=1}^{n}(z_{ij} - \overline{z}_{i.})^2},$$

sendo $N = \sum_{i=1}^{t} n_i$ e os pesos dados por

$$w_{0i} = n_i \left[1 + \frac{2}{\pi}\left(\sqrt{n_i(n_i - 2)} + sen^{-1}\frac{1}{n_i - 1} - n_i\right)\right]^{-1}$$

e

$$w_{1i} = \left[1 + \frac{2}{\pi(n_i - 1)}\left(\sqrt{n_i(n_i - 2)} + sen^{-1}\frac{1}{n_i - 1}\right)\right]^{-1},$$

em que $\overline{z}_{i.}$ é a i-ésima média de z_{ij}, e $\overline{\overline{z}}$ a média ponderada de $\overline{z}_{i.}$ usando w_{0i}. Para a situação de delineamentos balanceados ($n_i = n$ para todo i) a estatística é simplesmente um múltiplo da estatística F da análise de variância por mínimos quadrados ordinários (MQO)

$$F_{QMP} = m \times F_{QMO},$$

sendo

$$m = \frac{b - c}{b + (n - 1)c},$$

em que

$$b = 1 - \frac{2}{\pi}$$

e

$$c = \frac{2}{\pi}\left(\frac{1}{n-1}\right)\left(\sqrt{n(n-2)} + sin^{-1}\frac{1}{n-1} - (n-1)\right).$$

Este multiplicador (m) converte a estatística por quadrados mínimos ordinários em uma estatística por quadrados mínimos ponderados. Segundo os autores, esta estatística é importante especialmente para amostras pequenas, pois o valor de m tende a 1 quando n tende ao ∞.

Além das propostas frequentistas e bayesiana apresentadas, uma outra vertente são os procedimentos computacionais como Jackknife. Segundo Manly (1997), procedimentos Jackknife são aqueles em que, a partir de uma amostra, é feito o descarte de uma observação e em seguida é aplicado o estimador nos valores restantes. O descarte é de uma observação por vez para todas as observações amostrais assim, o número de estimadores é o mesmo do número de observações. Essas observações no final sofrem um determinado tipo de alteração e passam a se chamar pseudovalores. Layard (1973), propôs mudanças no procedimento do Jackknife de Miller (1968) para testar a hipótese de igualdade de variâncias. O teste é baseado no procedimento de Levene (1960), porém considerando pseudovalores. Seja o pseudovalor $U_{ij} = n_i Ln(S_i^2) - (n_i - 1)Ln(S_{i(j)}^2)$, em que $S_{i(j)}^2$ é o estimador da variância da i-ésima população de tratamento, após a eliminação da j-ésima observação. Desta forma, o teste de Layard é dada por:

$$L_{11} = \frac{\sum_{i=1}^{t} n_i \left(\overline{U}_{i.} - \overline{U}_{..}\right)^2 \Big/ (t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{n_i} \left(\overline{U}_{i.} - \overline{U}_{..}\right)^2 \Big/ (n-t)}, \tag{8}$$

sendo $\overline{U}_{i.}$ a média de cada tratamento do pseudovalor e $\overline{U}_{..}$ a média geral do pseudovalor. Sob H_0, a estatística L_{11} segue uma distribuição F, com $\nu 1 = t - 1$ e $\nu 2 = n - t$ graus de liberdade.

O'Brien em 1978 define uma nova proposta do teste de Layard usando pseudovalores que não sofreram transformação logarítmica sendo eles $V_{ij} = n_i S_i^2 - (n_i - 1)S_{i(j)}^2$. A estatística do teste (O_{12}) é a mesma do teste de Layard somente com a substituição dos pseudovalores.

Em Argaç (2002) é citado que James em 1951 sugeriu uma aproximação Ã distribuição de qui-quadrado para o teste clássico proposto por Cochran em 1937. A estatística do teste é dada por:

$$C_{13} = \sum_{i=1}^{t} \frac{n_i}{S_i^2}\left[\overline{y}_{i.} - \sum_{j=1}^{t} h_j \overline{y}_{.j}\right]^2,$$

sendo $h_j = \left(n_j/S_j^2\right) \Big/ \sum_{k=1}^{t} n_k/S_k^2$. Sob H_0 a estatística de Cochran tem distribuição assintótica de qui-quadrado com $\nu = t-1$ graus de liberdade. Nesta mesma linha, Welch, também em 1951, propôs uma modificação ao teste clássico de Cochran, onde a estatística segue uma distribuição F.

A proposta de Welch apresenta a seguinte estatística:

$$W_{14} = \frac{\sum_{i=1}^{t} \frac{n_i}{S_i^2}\left[\overline{y}_{i.} - \sum_{j=1}^{t} h_j \overline{y}_{.j}\right]^2}{(t-1) + 2(t-2)(t+1)^{-1}\sum_{i=1}^{t}(n_i-1)^{-1}(1-h_i)^2}.$$

Sob H_0 a estatística W_{14} segue uma distribuição F, com $\nu 1 = t-1$ e $\nu 2$ graus de liberdade dado por:

$$\nu 2 = \frac{t^2 - 1}{3\sum_{i=1}^{t}(n_i - 1)^{-1}(1 - h_i)^2}.$$

Hartung et al. (2002) desenvolveram uma modificação no teste de Welch para correção da liberalidade deste teste na presença de amostras pequenas e aumento no número de tratamentos. A estatística do teste é dada por:

$$WM_{15} = \frac{\sum_{i=1}^{t} \frac{n_i}{(\varphi_i S_i^2)} \left[\overline{y}_{i.} - \sum_{j=1}^{t} h_j^* \overline{y}_{.j}\right]^2}{(t-1) + 2(t-2)(t+1)^{-1}\sum_{i=1}^{t}(n_i - 1)^{-1}(1 - h_i^*)^2},$$

sendo $h_i^* = h_j^* = \left(n_j/(\varphi_j S_j^2)\right) \Big/ \sum_{k=1}^{t} n_k/(\varphi_k S_k^2)$; $\varphi_i = (n_i + \delta_1)/(n_i + \delta_2)$ e δ_1 e δ_2 números reais escolhidos cada qual para satisfazer $1 \leq \varphi_i \leq c_i$, $c_i = (n_i - 1)/(n_i - 3)$. Sob H_0 a estatística WM_{15} segue uma distribuição F, com $\nu = t - 1$ e $\nu 2^*$ graus de liberdade dado por:

$$\nu 2^* = \frac{t^2 - 1}{3\sum_{i=1}^{t}(n_i - 1)^{-1}(1 - h_i^*)^2}.$$

Simulações

Considerando um delineamento inteiramente casualizado (DIC) foram comparados por simulação de Monte Carlo, testes (B_1, B_2, B_3, B_4, L_5, L_6, L_7, L_8, S_9, OM_{10}, L_{11}, O_{12}, C_{13}, W_{14} e WM_{15}) para homogeneidade de variâncias.

Para isso, foram realizadas simulações nas hipóteses nulas (H_0) completas e hipóteses alternativas (H_1) para o estudo de desempenho das taxas de erro tipo I e poder. As simulações dos resíduos foram realizadas assumindo distribuição normal com média zero e variância de acordo com a situação. Para o estudo do erro tipo I assumiram-se variâncias constantes entre os tratamentos.

No caso do estudo do poder, sob hipótese alternativa verdadeira, as variâncias foram consideradas diferentes entre tratamentos. As diferenças foram simuladas de acordo com uma razão (δ) entre a maior e a menor variância.

Todas simulações foram realizadas no software R (R CORE TEAM, 2012), de forma a se ter um coeficiente de variação experimental de 15% para todos os casos. O número de iterações foi 10.000 e os números de tratamentos (t níveis) foram 3, 5, 10 e 20 e o mesmo ocorreu para as repetições (n níveis).

Para o teste WM_{15}, o valor de φ_i foi considerado a média entre 1 e c_i. Os valores de δ foram 4, 8, 16 e 32. Os valores-p observados foram confrontados com três níveis nominais de significância (0,01; 0,05 e 0,10).

Resultados e Discussão

Na Tabela 1 são apresentadas as taxas de erro tipo I em função do número de tratamentos, repetições e nível de significância de 0,05 para os testes de homogeneidade. As taxas observadas foram confrontadas com os intervalos de 99% de confiança para proporções (Leemis e Trivedi, 1996). De acordo com os resultados foi verificado que o teste de Bartlet, considerando o nível nominal de 5%, foi o único a controlar o erro tipo I ao nível de significância para todas as configurações adotadas. A versão proposta por Boos e Brownie (B_2) apresentou controle da taxa de erro tipo I com o aumento do número de repetições, mas se manteve liberal para poucas repetições. A terceira versão do teste de Bratlett (B_3) controlou o erro tipo I para situções com até 5 tratamentos e acima desse número, passou a ser conservador. A proposta bayesiana apresentada por Samiuddin (S_9) teve um comportamento sempre rigoroso, controlando a taxa de erro Ã medida que se aumentou o número de repetições. O mesmo ocorreu com a versão Jacknife apresentada por O'Neil e Mathews (OM_{10}). Os demais testes não controlaram a taxa de erro tipo I, como apresentado na Tabela 1. Resultados similares foram observados para os níveis de significância 0,01 e 0,10.

Para uma representação gráfica de todos os testes, percebeu-se um comportamento semelhante entre os desempenhos dos testes quando considerou-se 3 e 5 tratamentos. Devido a isso, optou-se por fixar o número de tratamentos como 3 e representar, na Figura 1, o desempenho destes com o aumento dos tamanhos amostrais.

Para δ=1 tem-se a representação da taxa de erro tipo I. Alguns testes tiveram resultados de destaque. O teste B_4 de Dixon e Massey apresentou maior poder em relação aos demais testes mas não controlou o nível de significância nominal para a taxa de erro tipo I, sendo de comportamento liberal, com taxas elevadas, para grande número de tratamentos e amostras pequenas.

Outro teste de elevado poder foi a proposta B_2 mas que também não controlou a taxa de erro tipo I para pequenas amostras, sendo esta obtida somente a partir de 10 repetições. De comportamento similar ao teste B_2, a proposta original de Levene (L_5) apresentou curva do poder inferior ao B_2, mas superior aos demais.

O teste original de Bartlett (B_1) teve sua curva de poder inferior aos supracitados mas foi a única proposta a controlar o erro tipo I em todas as situações, como apresentado na Tabela 1.

Ainda considerando amostras pequenas, a versão B_3 apresentou um poder próximo Ã proposta original de Bartlett e o teste bayesiano de Samiuddin (S_9), mantendo um rigor no controle da taxa de erro tipo I e, como esperado, menor poder nestas.

Os testes L_7, L_8, OM_{10}, L_{11} e O_{12} tiveram desempenhos muito inferiores, valores de taxa de erro tipo I e Poder não passaram de 5% para amostras de tamanho 3. Para o teste de Cochran (C_{13}) verificou-se uma ausência de controle da taxa de erro tipo I para pequenas amostras, com diminuição desta taxa a medida que as amostras aumentam.

As versões W_{14} e WM_{15} dependem diretamente da proposta C_{13}, por isso W_{14} também apresentou um comportamento assintótico para o controle da taxa de erro tipo I a medida que as amostras aumentavam.

A versão modificada WM_{15}, controlou a taxa de erro tipo I em todos os casos, mas apresentou comportamento rigoroso em amostras pequenas.

A ausência de controle da taxa de erro tipo I ou o rigor nesta influenciou o comportamento do poder destes testes baseados na proposta de Cochran.

Tabela 1: Taxas de erro dos testes para hogeneidade de variâncias em função do número de tratamentos (t) e do número de repetições (n) para o nível nominal de significância $\alpha = 0,05$

t	n	B_1	B_2	B_3	B_4	L_5	L_6	L_7	L_8	S_9	OM_{10}	L_{11}	O_{12}	C_{13}	W_{14}	WM_{15}
3	3	0,046*	0,101	0,054*	0,141	0,069	$0,000^{\dagger}$	$0,000^{\dagger}$	$0,000^{\dagger}$	$0,016^{\dagger}$	$0,000^{\dagger}$	$0,006^{\dagger}$	$0,000^{\dagger}$	0,197	$0,034^{\dagger}$	$0,003^{\dagger}$
5	3	0,046*	0,105	0,052*	0,174	0,122	$0,009^{\dagger}$	$0,000^{\dagger}$	$0,000^{\dagger}$	$0,011^{\dagger}$	$0,000^{\dagger}$	$0,006^{\dagger}$	$0,000^{\dagger}$	0,322	0,063	$0,005^{\dagger}$
10	3	0,045*	0,125	0,049*	0,227	0,254	0,164	$0,000^{\dagger}$	$0,000^{\dagger}$	$0,008^{\dagger}$	$0,001^{\dagger}$	$0,002^{\dagger}$	$0,000^{\dagger}$	0,566	0,139	$0,010^{\dagger}$
20	3	0,047*	0,142	$0,039^{\dagger}$	0,321	0,486	0,565	$0,000^{\dagger}$	$0,000^{\dagger}$	$0,004^{\dagger}$	$0,001^{\dagger}$	$0,001^{\dagger}$	$0,000^{\dagger}$	0,820	0,287	$0,022^{\dagger}$
3	5	0,048*	0,083	0,061	0,090	0,086	0,062	$0,004^{\dagger}$	$0,003^{\dagger}$	$0,039^{\dagger}$	$0,010^{\dagger}$	$0,043^{\dagger}$	$0,000^{\dagger}$	0,125	0,045*	$0,019^{\dagger}$
5	5	0,052*	0,075	0,050*	0,101	0,098	0,077	$0,003^{\dagger}$	$0,001^{\dagger}$	$0,034^{\dagger}$	$0,009^{\dagger}$	$0,035^{\dagger}$	$0,012^{\dagger}$	0,181	0,053*	$0,018^{\dagger}$
10	5	0,047*	0,074	$0,044^{\dagger}$	0,119	0,135	0,113	$0,001^{\dagger}$	$0,000^{\dagger}$	$0,026^{\dagger}$	$0,011^{\dagger}$	$0,025^{\dagger}$	$0,011^{\dagger}$	0,312	0,084	$0,024^{\dagger}$
20	5	0,048*	0,075	$0,037^{\dagger}$	0,149	0,181	0,167	$0,000^{\dagger}$	$0,000^{\dagger}$	$0,021^{\dagger}$	$0,011^{\dagger}$	$0,014^{\dagger}$	$0,007^{\dagger}$	0,489	0,130	$0,031^{\dagger}$
3	10	0,050*	0,058	0,051*	0,068	0,063	0,049*	$0,032^{\dagger}$	$0,027^{\dagger}$	0,048*	$0,010^{\dagger}$	0,052*	$0,021^{\dagger}$	0,080	0,047*	$0,031^{\dagger}$
5	10	0,048*	0,057	0,047*	0,070	0,072	0,058	$0,029^{\dagger}$	$0,021^{\dagger}$	$0,044^{\dagger}$	$0,009^{\dagger}$	0,048*	$0,025^{\dagger}$	0,108	0,050*	$0,031^{\dagger}$
10	10	0,048*	0,050*	$0,039^{\dagger}$	0,078	0,078	0,072	$0,024^{\dagger}$	$0,016^{\dagger}$	$0,040^{\dagger}$	$0,011^{\dagger}$	$0,040^{\dagger}$	$0,024^{\dagger}$	0,152	0,060	$0,031^{\dagger}$
20	10	0,047*	0,051*	$0,037^{\dagger}$	0,081	0,091	0,093	$0,016^{\dagger}$	$0,009^{\dagger}$	$0,037^{\dagger}$	$0,009^{\dagger}$	$0,030^{\dagger}$	$0,017^{\dagger}$	0,217	0,068	$0,033^{\dagger}$
3	20	0,049*	0,053*	0,051*	0,058	0,057	0,050*	$0,038^{\dagger}$	$0,035^{\dagger}$	0,048*	$0,010^{\dagger}$	0,050*	$0,035^{\dagger}$	0,066	0,057	$0,043^{\dagger}$
5	20	0,051*	0,052*	0,048*	0,060	0,061	0,056	$0,034^{\dagger}$	$0,029^{\dagger}$	0,049*	$0,008^{\dagger}$	0,049*	$0,037^{\dagger}$	0,073	0,049*	$0,039^{\dagger}$
10	20	0,049*	0,047*	$0,043^{\dagger}$	0,061	0,063	0,057	$0,032^{\dagger}$	$0,024^{\dagger}$	0,045*	$0,009^{\dagger}$	$0,040^{\dagger}$	$0,034^{\dagger}$	0,091	0,050*	$0,038^{\dagger}$
20	20	0,047*	0,048*	$0,042^{\dagger}$	0,063	0,066	0,067	$0,024^{\dagger}$	$0,018^{\dagger}$	$0,042^{\dagger}$	$0,011^{\dagger}$	$0,039^{\dagger}$	$0,031^{\dagger}$	0,117	0,055*	$0,037^{\dagger}$

*Não diferente do nivel nominal, baseado no IC exato para proporções com 99% de confiança.

† diferente do nivel nominal, estando abaixo do limite inferior do IC exato para proporções com 99% de confiança.

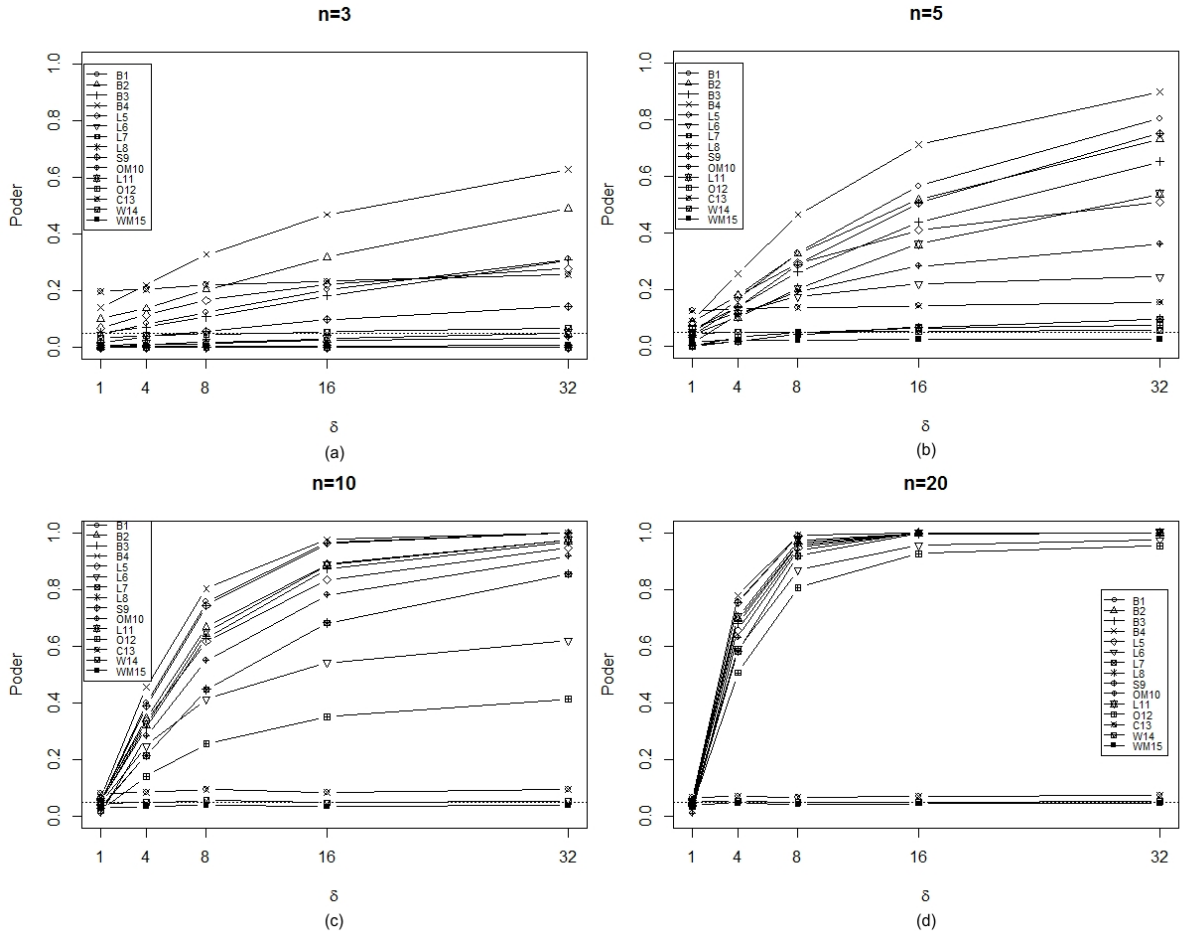

Figura 1: Representação gráfica do desempenho de todos os testes considerando t=3, tamanho da amostra (n) e razão entre variâncias populacionais (δ).

De modo geral, para o número de tratamentos igual a 3, os testes de maior destaque foram B_1, B_2, B_4 e S_9. As propostas tendem a ter um comportamento mais parecido, entre si, com o aumento no número de repetições, mas observa-se que com 5 repetições, valor muito usado na prática, as curvas de poder se distanciaram (Figura 2). As propostas L_6 e O_{12} tiveram os piores desempenhos de todos os testes.

As versões que se baseiam na proposta de Bartlett (B_1, B_2, B_3 e B_4) não foram exploradas para as propostas que foram desenvolvidas pois suas varições, segundos os autores, vem para contornar a dependencia Ã distribuição normal.

No estudo, as simulações foram baseadas na normalidade e, por isso, todas tiveram desempenho inferior Ã proposta original, com resalva para o teste B_4 que, por não controlar o erro tipo I (liberal), teve uma curva de maior poder.

Uma supresa foi o desempenho apresentado pela proposta de Samiuddin (S_9) baseada em um método bayesiano. Mostrou-se um teste com uma curva de poder próxima dos melhores desempenhos, apesar de ser conservador em amostas com 3 e 5 repetições. A proposta de O'Neill e Mathews (OM_{10}) apresentou uma curva de poder abaixo dos demais para pequenas amostras e um rigor no controle da taxa de erro tipo I. Merece destaque pelo fato de ser conservador em todas as configurações e, em situações de amostras grandes, apresentou um elevado poder. Com amostras de tamanho 5 ou mais o teste controlou também o erro tipo I para 1% e 10% de significância.

Todas as versões que foram propostas para a melhoria de algum teste clássico, tiveram resultados piores nas situações estudadas. Apenas as versões B_4, que apresentou uma curva de

poder melhor que o teste original de Bartlett, com resalva de ser um teste liberal em todas as configurações estudadas, e a versão *jackknife* ($L11$) do teste original de Levene que apresentou uma curva de poder um pouco melhor em situações com amostra de tamanhos 10 e 20.

O aumento do número de tratamentos surtiu um efeito de um leve aumento no poder dos testes, sendo mais significativo quando se estudaram 20 tratamentos. O teste B_4 apresentou um aumento na taxa de erro tipo I e, por isso, credita-se um aumento também do poder. Os testes S_9 e as versões jackknife (L_{11}, O_{12}) passaram a ser mais rigorosos no controle da taxa de erro tipo I com um aumento no poder, como os demais. Esses resultados são mais evidentes para amostras maiores que 3 repetições.

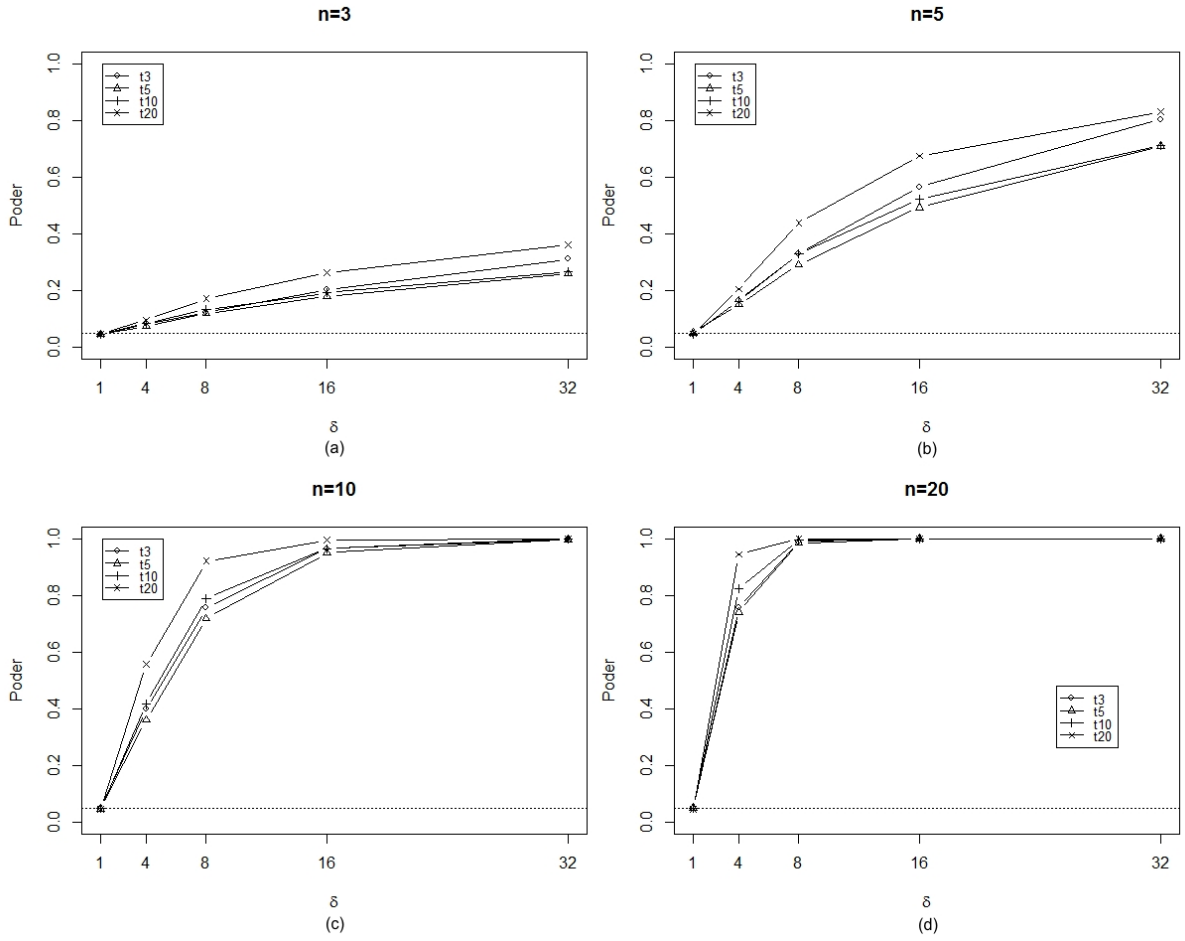

Figura 2: Representação gráfica do desempenho do teste original de Bartlett de acordo com o número de tratametos (t), tamanho da amostra (n) e razão entre variâncias populacionais (δ).

Na Figura 2 são representados graficamente os resultados referente ao desempenho quanto a taxa de erro tipo I e poder do teste de Bartlet (B_1) para todas as configurações estudadas. É possível observar o controle do erro tipo I e o aumento do poder do teste a medida que o número de repetições aumenta. O aumento da razão entre as variâncias (*delta*), como era esperado, possibilitou ao teste maior poder, ou seja, maior capacidade de diagnosticar diferenças. Pode-se verificar que, para um $\delta = 8$ com 10 repetições, o poder já se encontra próximo de 80%. Considerando um número de repetições qualquer observa-se também um aumento do poder com o aumento do número de tratamentos.

Os testes baseados na proposta de Levene (1960) (L_5, L_6, L_7, L_8, L_{11}, O_{12}) apresentaram desempenho frustrante, com destaque apenas para a versão original L_5 e a versão jackknife de Layard (1973) (L_{11}). Na Figura 3 são apresentados os desempenhos para o número de

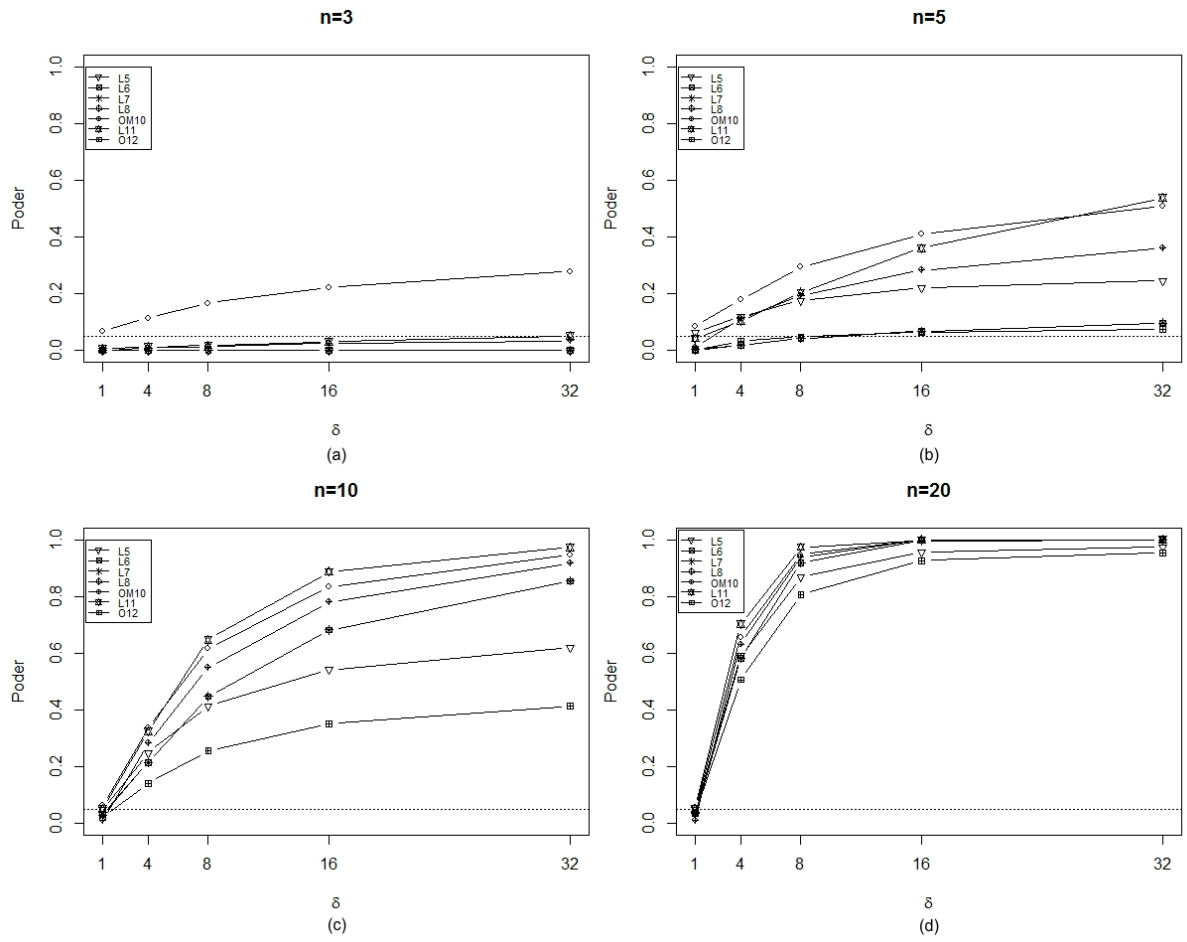

Figura 3: Representação gráfica do desempenho dos testes que são baseados na proposta de Levene considerando t=3, tamanho da amostra (n) e razão entre variâncias populacionais (δ).

tratamentos igual a 3. Nenhum dos seis testes controlou o erro tipo I nas configurações estudadas, exceto algumas poucas combinações. Apresentaram baixo poder para amostras pequenas, mesmo com a máxima diferença estudada entre as variâncias (Figura 3). A versão *jackknife* apresentada por Layard, com tamanhos de amostras 5, apresentou um resultado bem próximo do teste original de Levene e, com amostras maiores do que 10, apresentou um poder superior em todas as configurações estudadas. Outro teste que merece destaque é a proposta apresentada por O'Neill e Mathews (OM_{10}) que tem como finalidade corrigir o teste de Levene para situações de desbalanceamento mas, como não foram estudados estes casos, esta se mostrou de desempenho pior do que a versão original de Levene. As propostas de Brown-Forsythe (L_7) e Brown-Forsythe modificado (L_8) apresentaram igual comportamento e tiveram desempenho pior do que as versões supracitadas. As versões com transformações definidas na mediana são recomendadas quando a distribuição é não-normal, com o intuito de aumentar a robustez do teste. As versões L_6 e O_{12} tiveram os piores desempenhos nestas situações. Com um tamanho de amostra de 20 os resultados foram semelhantes. Para outros números de tratamentos, as versões L_5 e L_6 tiveram um poder maior que as demais, mas apresentaram um comportamento liberal, com erros tipo I muito elevados, da ordem de 40% para 20 tratamentos. As demais versões (L_7, L_8, L_{11} e O_{12}) se mostraram rigorosas no controle do erro tipo I, e para amostras de tamanho 3, o poder não ultrapassou 5%, como anteriormente comentado. Os desempenhos só melhoram quando o tamanho da amostra é, pelo menos, 10.

Conclusões

Em amostras pequenas os testes apresentaram desempenhos variados. Como destaque, o teste de Bartlett (B_1) manteve o controle do erro tipo I em todas as configurações e níveis de significância estudados, com curva de poder entre os melhores desempenhos. A proposta original de Levene (L_5) não controlou a taxa de erro tipo I para pequenas amostras, apresentando aumento desta Ã medida que o número de tratamentos aumentava e, consequentemente, maior poder. As versões S_9, OM_{10} e L_{11} apresentaram comportamento rigoroso e poder baixo para pequenas amostras. Com amostras maiores do que 10 repetições os testes apresentaram comportamentos parecidos para a curva de poder, sendo estes elevados para δ acima de 8.

Agradecimento

A Fapemig pelo apoio financeiro.

Referências

ARGAÇ, D. Testing for homogeneity in a general one-way classication with fixed effects: power simulations and comparative study. *Computational Statistics and Data Analysis*, 2002.

BARTLETT, M. S. Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society - Serie A*, v,60, p.268-282, 1937.

BOOS, D. D.; BROWNIE, C. Bootstrap methods for testing homogeneity of variances *Technometrics*, v.31, n.1, p.69-82, 1989.

BOX, G. E. P.; ANDERSEN, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumptions. *Journal of the Royal Statistical Society, Series B*, v.17, n.1, p.1-26, 1955.

BOX, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Statist.*, v.25, p.290-302, 1954.

BROWN, M. B.; FORSYTHE, A. B. Robust tests for equality of variances. *Journal of the American Statistical Association*, v.69, n.346, p.364-367, 1974.

BROWN, M. B.; FORSYTHE, A. B. The use of weighted contrasts in analysis of models with heterogeneity of variance. Proceedings of the Business and Economics Statistics Section *American Statistical Association*, p.347-352, 1983.

DIXON, W. J.; MASSEY, F. J. Introduction to statistical analysis. *McGraw-Hill Book*, New York, n.3, p.308-309 , 1969.

FERREIRA, D. F. *Estatística básica*, Editora UFLA, Lavras - MG, p.664, 2005.

FERREIRA, D. F.; DEMÉTRIO, C. G. B.; MANLY, B. F. J.; MACHADO, A. DE A.; VENCOVSKY, R. Statistical models in agriculture: biometrical methods for evaluating phenotypic stability in plant breeding. *Cerne*, v.12, n.4, p.373-388, 2006.

GOMEZ, K. A.; GOMEZ, A. A. Statistical procedures for agricultural research. *John Wiley*, n.2, p.680, 1984.

HARTUNG, J.; ARGAÇ, D.; MAKAMBI, K. H., Small sample properties of tests on homogeneity in one-way ANOVA and meta-analysis. *Statist. Papers* , n.43, p.197-235, 2002.

JAMES, G. S. The comparison of several groups of observations when the ratios of population variances are unknown. *Biometrika*, n.38, p.324-329, 1951.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis.* Prentice Hall: New Jersey, p.816, 1998.

KEYES, T. K.; LEVY, M. S. Analysis of levene's test under design imbalance. *Journal of Educational and Behavioral Statistics*, v.22, p.227-236, 1997.

LAYARD, M. N. J. Robust large-sample tests for homogeneity of variances. *Journal of the American Statistical Association*, v.68, n.341, p.195-198, 1973.

LEEMIS, L.; TRIVEDI, K. S. A comparison of approximate interval estimators of the Bernoulli parameter. *The American Statistician*, v.50, p.63-68. 1996.

LEVENE, H. Robust tests for equality of variances. *Contribution to Probability and Statistics. Stanford, CA: Stanford University Press*, p.278-292, 1960.

MANLY, B. F. J. Randomization, bootstrap and Monte Carlo methods in biology. *University of Otago*, New Zealand, p. 356, 1997.

MEHROTRA, D. V. Improving the Brown Forsythe solution to the generalized Behrens Fisher problem. *Comm. Statist. Simulation Comput.*, v.26, p.1139 1145, 1997.

MILLER, R. G., Jr. Jackknifing variances. *Annals of Mathematical Statistics*, v.39, n.2, p.567-582, 1968.

O'BRIEN, R. G. A robust technique for testing heterogeneity of variance effects in factorial design. *Psychometrika*, v.43, n.3, p.327-342, 1978.

O'NEILL, M. E.; MATHEWS, K. L. A weighted least squares approach to levene's test of homogeneity of variance. *Australian e New Zealand Journal Statistical*, v.42, n.1, p.81-100, 2000.

O'NEILL, M. E.; MATHEWS, K. L. Levene tests of homogeneity of variance for general block and treatment designs. *Biometrics*, v.51, p.216-224, 2002.

R DEVELOPMENT CORE TEAM. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. 2012. ISBN 3-900051-07-0, Disponível: http://www.R-project.org/.

RUBIN A. S. *Experimentação em genética.* Editora UFLA, Lavras - MG, n.2, p.303, 2005.

SAMIUDDIN, M. Bayesian test of homogeneity of variance. *Journal of the American Statistical Association*, v.71, n.354, p.515-517, 1976.

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

WILSON, E. B.; HILFERTY, M. M. The distribution of chi-square. *Proceedingdof the National Academy of Science*, v.17, p.684-688, 1931.