PRECISÃO E PODER DE TESTES DE HOMOCEDASTICIDADE PARAMÉTRICOS E NÃO-PARAMÉTRICOS AVALIADOS POR SIMULAÇÃO

João RIBOLDI[1]
Márcia Helena BARBIAN[1]
Ana Beatriz da Silva KOLOWSKI[2]
Lisiane Priscila Roldão SELAU[1]
Vanessa BielefeldtLeottiTORMAN[1]

- RESUMO: No presente trabalho compararam-se os testes de homogeneidade de variâncias paramétricos de Bartlett, Brow-Forsythe, O'Brien, Levene com as opções absoluto e quadrado; e os testes não-paramétricos de Siegel-Tukey, Ansari-Bradley, Klotz e Mood, utilizando-se simulação de dados. Para tanto foram simulados no SAS 10.000 experimentos para oito diferentes cenários de análise de variância em classificação simples e erros com distribuição normal, uniforme ou exponencial. Foram realizadas comparações sob hipóteses de homocedasticidade(para avaliar taxa de erro tipo I) e heterocedasticidade (para avaliar poder). Adicionalmente, consideraram-se efeitos nulos e não-nulos de tratamentos e dados balanceados ou não. Os resultados obtidos permitem destacar, dentre os testes paramétricos investigados, a performance do teste de Bartlett, que sob normalidade, é um teste exato, preciso e com alto poder, influenciado pelo desbalanceamento dos dados. O teste de Levene (absoluto) se assemelha ao de Bartlett quanto ao poder, mas é impreciso e liberal, sob qualquer distribuição, enquanto que o teste de Levene(quadrado), mesmo que seja impreciso, é exato para distribuição subjacente normal e levemente liberal para as demais, com poder inferior aos testes de Bartlett e de Levene(absoluto). O teste de Brow-Forsythe é conservador, impreciso com poder de moderado a alto. O teste de O'Brien foi o de pior performance dentre os testes paramétricos, sendo conservador, impreciso e com poder instável. Os testes não-paramétricos de Siegel-Tukey, Ansari-Bradley, Klotz e Mood se assemelham em performance quanto ao poder, são imprecisos e, com exceção do teste de Klotz, são exatos na ausência de efeito de tratamentos. Recomenda-se utilizar o teste de Bartlett se dados aproximadamente normais, caso contrário, dada a sua maior robustez, o teste de Levene preferentemente com as duas opções, quadrado para exatidão e absoluto para poder.

- PALAVRAS-CHAVE: Testes de homocedasticidade, simulação, precisão, poder.

1 Introdução

1 Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Caixa Postal 15080, CEP: 91509-900, Porto Alegre, RS, Brasil, E-mail: *joao.riboldi@ufrgs.br*, *mhbarbian@gmail.com, lisianeselau@gmail.com*, *vanessa.leotti@ufrgs.br*

2 StatSoft South America, Analista Estatístico, Porto Alegre, E-mail: *ana.kolowski@statsoft.com.br*

As pressuposições de aditividade, variância constante (homocedasticidade), normalidade e erros não correlacionados (independência) em modelos lineares têm vários objetivos, mas essencialmente visam facilitar a interpretação dos resultados, tornar as técnicas estatísticas mais simples e possibilitar testes de hipóteses.O não atendimento destas pressuposições pode afetar mais ou menos gravemente as conclusões tomadas com base nos modelos estabelecidos e nas técnicas de análise a eles associados (JOHNSON; WICHERN, 1998).

A validade exata dessas pressuposições é meramente teórica e, na prática, o que se deseja é a validade aproximada, uma vez que os procedimentos obtidos através dos modelos lineares são razoavelmente robustos e pouco se perde se a validade das pressuposições é apenas aproximada.Os efeitos dos desvios das pressuposições são variados e a gravidade depende sempre da situação em particular.

A variância constante ou homogeneidade de variâncias (homocedasticidade) é, na maioria das vezes, um requisito necessário para a análise de variância (ANOVA). Sob heterogeneidade de variância, o método dos mínimos quadrados ordinários não produz os melhores estimadores e o teste F, as comparações múltiplas, os contrastes ortogonais, ou a estimação dos componentes de variância poderão ser fortemente afetados. Isto é, dependendo do grau de severidade da heterogeneidade das variâncias, a análise de variância pode ter sua significância estatística comprometida (MILLIKEN; JOHNSON, 1992).

O procedimento clássico de testar a hipótese de igualdade de variâncias dentro de cada tratamento, isto é, contra a hipótese alternativa de que pelo menos duas variâncias de tratamentos diferem, é o teste da razão de verossimilhança sob a suposição de normalidade. A distribuição da estatística no teste da razão de verossimilhança para igualdade de variâncias em populações normais depende da curtose da distribuição (BOX, 1953). Assim, o teste é sensível a desvios de normalidade. O teste de Bartlett (1937) é uma modificação do teste da razão de verossimilhança com objetivo de melhorar a aproximação à distribuição qui-quadrado.

Outro procedimento é transformar os valores originais da variável dependente, derivando a variável de dispersão e então executar a análise de variância nesta variável. O nível de significância para o teste de homogeneidade de variâncias é o p-valor do teste F na análise de variância da variável de dispersão. Os testes de homogeneidade de variâncias de Levene, Brow-Forsythe e O'Brien, usam este procedimento. O teste de Levene (1960) usa os desvios em relação à média dos grupos (tratamentos), podendo-se tomar o valor absoluto dos desvios (Levene absoluto) ou os quadrados dos desvios (Levene quadrado). O teste de Brow-Forsythe (1974) usa o absoluto dos desvios em relação à mediana, enquanto que o teste de O'Brien (1979) usa uma função dos desvios em relação a média dos tratamentos.

Existem também os testes não-paramétricos, como os de Siegel-Tukey (1960), Ansari-Bradley (1960), Klotz (1962) e Mood (1954), que são baseados em postos. Eles são adequados e eficazes para detectar diferenças de dispersão entre populações que são idênticas em todos os outros aspectos, incluindo posição.

Assim, o usuário de testes de igualdade de variâncias é contemplado com um elenco grande e confuso de opções. Tem-se tornado padrão o uso do teste de Bartlett, que tem maior precisão para a probabilidade do erro tipo I e alto poder quando a distribuição subjacente dos dados é normal, mas pode ser muito inexato se a distribuição é ligeiramente não-normal (CONNOVER et al., 1981).

Na literatura tem sido feito referência à sensibilidade do teste de Bartlett a desvios da normalidade, enquanto que os testes de Box e Levene mostram-se mais robustos e, portanto menos sensíveis a falta de normalidade. Gartside (1972) estudou oito testes e concluiu que só o procedimento de Box foi robusto, mas apresentou baixo poder, além de depender do processo de dividir cada amostra em subamostras menores de forma aleatória. Layard (1973) chegou a conclusões similares para o teste de Box. O procedimento de jackknife parece ser o melhor dos seis procedimentos investigados por Hall (1972) em um estudo intensivo de simulação, enquanto Keselman et al.(1979) concluíram que o procedimento jackknife tinha erros instáveis (erro tipo I) quando as amostras têm tamanhos desiguais. Eles concluíram do estudo com dez testes que "os atuais testes para heterogeneidade de variâncias são ou sensíveis para a não normalidade ou, se robustos, falham em poder. Portanto esses testes não podem ser recomendados para o propósito de testar a validade da suposição de homogeneidade de variâncias". Os quatro testes estudados por Levy (1978) foram todos grosseiramente afetados pela violação da suposição subjacente de normalidade.

Olejink e Algina (1987), nos resultados de suas simulações, mostraram que, enquanto todos os testes baseados na ANOVA da variável de dispersão são razoavelmente robustos à distribuição subjacente, o teste de Brown-Forsythe parece ter melhor poder para detectar a diferença de variâncias e mais preciso para a probabilidade do erro tipo I. Conover et al. (1981) conduziram, por simulação, um estudo comparativo entre testes de homogeneidade de variâncias e concluíram que os testes de Brown-Forsythe e de Klotz são superiores em robustez e poder.

Milliken e Johson (1992), com base no trabalho de Conover et al. (1981), recomendam adotar o seguinte procedimento: (i) se os dados são aproximadamente normais deve-se usar os testes de Bartlett ou de Hartley, e o teste de Bartlett quando houver desbalanceamento; (ii) para amostras grandes deve-se usar o teste de Box, que é muito poderoso nestes casos, mas com sensível perda de poder no caso de amostras pequenas; (iii) noutros casos utilizar o teste de Levene, que se mostra tão eficiente quanto os testes de Bartlett e Hartley para dados normalmente distribuídos e mais eficiente que eles para dados não normalmente distribuídos.

Dada a importância da verificação da suposição de homocedasticidade para a análise de variância, o presente trabalho tem como objetivo comparar cinco testes de homogeneidade de variâncias paramétricos e quatro não-paramétricos que podem ser utilizados nesta tarefa.

Este artigo está organizado em cinco seções. Na segunda seção é detalhada a metodologia utilizada para comparar os nove testes por meio de simulações. Na terceira seção são apresentados os principais resultados das comparações quanto a exatidão, precisão e poder dos testes. A quarta seção trata da discussão dos achados neste artigo.Por fim, na última seção são apresentadas as considerações finais do estudo.

2 Metodologia

No presente trabalho compararam-se os testes de homogeneidade de variâncias paramétricos de Bartlett, Brow-Forsythe, O'Brien, Levene absoluto e quadrado; e os não-paramétricos de Siegel-Tukey, Ansari-Bradley, Klotz (qui-quadrado) e Mood, que se encontram implementados no software SAS, utilizando-se simulação de dados. Foram

considerados oito diferentes cenários em classificação simples, ou seja, numa estrutura de delineamento completamente casualizado com um fator de tratamentos. Os diferentes cenários foram pensados para permitir comparações sob hipóteses de homogeneidade de variâncias (para avaliar taxa de erro tipo I) e heterogeneidade de variâncias (para avaliar poder).

O uso dos testes não-paramétricos Siegel-Tukey, Ansari-Bradley, Klotz e Mood somente é recomendado nos casos em que os tratamentos possuem a mesma mediana, testando, portanto, a hipótese de que populações que são idênticas em forma e locação são também idênticas em dispersão, isto é, se possuem variabilidade não distinta. Por isso em alguns momentos considerou-se efeitos nulos e não-nulos de tratamentos. Adicionalmente, alguns cenários possuíam balanceamento e outros não, para verificar a influência desse fator no desempenho dos testes. O tamanho da amostra variou de 29 a 48 observações.

A descrição de cada caso avaliado é a seguinte:

A. Três tratamentos, com efeitos não-nulos de tratamentos e mesma variância, com 10 repetições;

B. Três tratamentos, com efeitos nulos de tratamentos e mesma variância, com 10 repetições;

C. Três tratamentos, com efeitos nulos de tratamentos e diferentes variâncias, com 10 repetições;

D. Três tratamentos, com efeitos não-nulos de tratamentos e diferentes variâncias, com 10 repetições;

E. Doze tratamentos, com efeitos nulos de tratamentos e diferentes variâncias, com quatro repetições;

F. Doze tratamentos, com efeitos não-nulos de tratamentos e diferentes variâncias, com quatro repetições;

G. Quatro tratamentos, com efeitos nulos de tratamentos, diferentes variâncias e diferentes repetições por tratamento(7, 6, 8 e 8 repetições);

H. Quatro tratamentos, com efeitos não-nulos de tratamentos, diferentes variâncias e diferentes repetições por tratamento(7, 6, 8 e 8 repetições).

Mais detalhes sobre os parâmetros considerados em cada caso estão apresentados na Tabela 1. Para cada caso, foram consideradas três possibilidades de distribuição para o erro aleatório do modelo: Normal, Uniforme ou Exponencial. No caso da distribuição Normal, a média foi sempre fixada em zero, e as variâncias foram as apresentadas na Tabela 1. Para a distribuição Uniforme, considerou-se sempre o limite inferior igual a zero, e os limites superiores foram os da Tabela 1. As médias consideradas para a distribuição Exponencial também estão apresentadas na Tabela 1.

Para cada caso, foram simulados no SAS,versão 9.2, 10.000 experimentos com as diferentes configurações estabelecidas. Os testes foram realizados através dos procedimentos *ANOVA* e *NPAR1WAY* do SAS.

Para cada experimento simulado, obtiveram-se os níveis mínimos de significância (nms) de cada um dos testes de homogeneidade de variâncias. Segundo Mood *et al.*

(1974), sob hipótese de homogeneidade de variâncias, caso as exigências do teste sejam satisfeitas, a distribuição dos nms será Uniforme no intervalo [0,1).

Tabela 1 - Detalhes dos casos estudados

Caso	Número do Tratamento	Número de Repetições	Média Geral	Variância do Tratamento	Efeito do Tratamento	Variância da Distribuição Normal	Limite superior da Distribuição Uniforme	Média da Distribuição Exponencial
A	1	10	40	100	11	100	34,6410	0,1
	2	10	40	100	-6	100	34,6410	0,1
	3	10	40	100	-5	100	34,6410	0,1
B	1	10	40	100	0	100	34,6410	0,1
	2	10	40	100	0	100	34,6410	0,1
	3	10	40	100	0	100	34,6410	0,1
C	1	10	50	16	0	16	13,8564	0,25
	2	10	50	256	0	256	55,4256	0,0625
	3	10	50	4096	0	4096	221,7025	0,0156
D	1	10	140	16	-120	16	13,8564	0,2500
	2	10	140	256	-60	256	55,4256	0,0625
	3	10	140	4096	180	4096	221,7025	0,0156
E	1	4	0,325	0,0048	0	0,0048	0,2406	14,3963
	2	4	0,325	0,0259	0	0,0259	0,5571	6,2177
	3	4	0,325	0,0246	0	0,0246	0,5429	6,3812
	4	4	0,325	0,0127	0	0,0127	0,3909	8,8619
	5	4	0,325	0,0057	0	0,0057	0,2608	13,2842
	6	4	0,325	0,1131	0	0,1131	1,1650	2,9735
	7	4	0,325	0,0032	0	0,0032	0,1970	17,5863
	8	4	0,325	0,0734	0	0,0734	0,9387	3,6904
	9	4	0,325	0,0005	0	0,0005	0,0748	46,2910
	10	4	0,325	0,0022	0	0,0022	0,1612	21,4834
	11	4	0,325	0,0002	0	0,0002	0,0447	77,4597
	12	4	0,325	0,0007	0	0,0007	0,0917	37,7965
F	1	4	0,325	0,0048	0,0875	0,0048	0,2406	14,3963
	2	4	0,325	0,0259	0,5550	0,0259	0,5571	6,2177
	3	4	0,325	0,0246	0,2425	0,0246	0,5429	6,3812
	4	4	0,325	0,0127	0,2850	0,0127	0,3909	8,8619
	5	4	0,325	0,0057	-0,0050	0,0057	0,2608	13,2842
	6	4	0,325	0,1131	0,4900	0,1131	1,1650	2,9735
	7	4	0,325	0,0032	0,0500	0,0032	0,1970	17,5863
	8	4	0,325	0,0734	0,3425	0,0734	0,9387	3,6904
	9	4	0,325	0,0005	0,1150	0,0005	0,0748	46,2910
	10	4	0,325	0,0022	0,0100	0,0022	0,1612	21,4834
	11	4	0,325	0,0002	-0,0900	0,0002	0,0447	77,4597
	12	4	0,325	0,0007	0,0000	0,0007	0,0917	37,7965
G	1	7	9,6897	16,2860	0	16,2860	13,9796	0,2478
	2	6	9,6897	1,8670	0	1,8670	4,7333	0,7319
	3	8	9,6897	9,6960	0	9,6960	10,7865	0,3212
	4	8	9,6897	2,7860	0	2,7860	5,7819	0,5991
H	1	7	9,6897	16,2860	-5,1182	16,2860	13,9796	0,2478
	2	6	9,6897	1,8670	1,9770	1,8670	4,7333	0,7319
	3	8	9,6897	9,6960	-1,0647	9,6960	10,7865	0,3212
	4	8	9,6897	2,7860	4,0608	2,7860	5,7819	0,5991

Então para os casos A e B (casos onde a hipótese de homogeneidade de variâncias estava atendida), testou-se a aderência dos nms à distribuição Uniforme [0,1), utilizando-se o teste de Kolmogorov-Smirnov. Também para estes casos, as taxas de erro tipo I foram calculadas para cada teste como a proporção de vezes que a hipótese de nulidade foi rejeitada, ou seja, calculando-se a proporção de experimentos simulados em que o nms era menor ou igual a 0,05. Testes binomiais exatos considerando um nível de significância nominal de 1%, para testar a hipótese: H_0: α=5% versus H_1: $\alpha \neq$5%, para os diferentes testes, foram procedidos.Se a hipótese de nulidade (H_0) não for rejeitada, quando a taxa de erro tipo I observada não for significativamente diferente do nível nominal ($p > 0,01$), tem-se a situação ideal e, nesses casos, o teste será considerado exato. Quando a hipótese de nulidade (H_0) for rejeitada, se a taxa de erro tipo I observada for considerada significativamente ($p < 0,01$) inferior ao nível nominal, o teste será considerado conservador. Caso contrário, se a taxa de erro tipo I observada for considerada significativamente ($p < 0,01$) superior ao nível nominal, o teste será considerado liberal.

Nos casos C a H (onde a hipótese de homogeneidade de variâncias não estava atendida), estimou-se o poder dos testes, calculando-se a proporção de experimentos simulados em que o nms era menor ou igual a 0,05.

3 Resultados

As taxas de erro tipo I estimadas para os diferentes testes sob hipótese de homogeneidade de variâncias, ou seja, para os casos A e B, encontram-se na Tabela 2, juntamente com os resultados do teste binomial exato.

Tabela 2 - Taxas de erro tipo I estimadas dos testes (em percentual)

Testes	*Caso*[I] e *Distribuição dos erros*					
	A Normal	A Uniforme	A Exponencial	B Normal	B Uniforme	B Exponencial
Bartlett	4,84	0,75-	0,75-	4,84	0,75-	0,75-
Levene Quadrado	5,04	5,75+	5,75+	5,04	5,75+	5,75+
Levene Absoluto	5,93+	6,42+	6,42+	5,93+	6,42+	6,42+
Brow-Forsythe	3,14-	2,55-	2,55-	3,14-	2,55-	2,55-
O'Brien	3,48-	4,28-	4,28-	3,48-	4,28-	4,28-
Siegel-Tukey	5,6+	2,97-	2,97-	4,68	4,98	4,98
Ansari-Bradley	5,81+	3,08-	3,08-	4,65	4,99	4,99
Klotz	1,33-	0,47-	0,47-	3,82-	4,21-	4,21-
Mood	3,97-	2,09-	2,09-	4,72	4,82	4,82

I – Caso A: Três tratamentos, com efeitos não-nulos de tratamentos e mesma variância; **Caso B:**Idem Caso A mas com efeitos nulos de tratamentos.

+ Taxas de erro tipo I superiores ao valor nominal de 5% de significância ($p<0,01$)

- Taxas de erro tipo I inferiores ao valor nominal de 5% de significância ($p<0,01$)

Não houve diferença entre as taxas de erro tipo I dos testes paramétricos ao se comparar os casos A e B com a mesma distribuição dos erros. O teste de Bartlett teve taxa de erro tipo I fortemente desviada do valor nominal de 5%, sendo significativamente inferior, constituindo-se num teste conservador no caso de erros não-normais. Performance similar verifica-se para o teste de Levene (quadrado), com a diferença que para erros não-normais, a taxa de erro tipo I é significativamente superior ao valor nominal de 5%, constituindo-se, portanto, num teste liberal. Os testes de Levene (absoluto), Brow-Forsythe e O'Brien se desviaram significativamente do valor nominal mesmo no caso de erros normais, observando-se comportamentos liberal no teste de Levene (absoluto) e conservador nos testes de Brow-Forsythe e O'Brien. Os testes não-paramétricos foram exatos no caso B, com efeitos nulos de tratamentos, ou seja, nos casos de não diferença de posição, com exceção do teste de Klotz que teve performance conservadora independentemente da diferença de posição ou não.

Pelo teste de Kolmogorov-Smirnov de aderência a distribuição Uniforme[0,1) dos nms dos testes, o teste de Bartlett se mostrou preciso, com nms aderindo à distribuição Uniforme quando a distribuição dos erros é normal, e impreciso no caso das outras distribuições. Os demais testes, paramétricos e não-paramétricos, se mostraram imprecisos independentemente da distribuição dos erros.

O poder estimado para cada teste sob hipótese de heterogeneidade de variâncias, isto é, para os casos C a H, encontram-se na Tabela 3. Quanto ao poder, os testes de Bartlett e Levene(absoluto) apresentaram melhor performance, pouco afetados pela distribuição subjacente dos erros, mas com sensível diminuição no poder devido ao desbalanceamento dos dados nos casos G e H. Os testes de Levene (quadrado) e de Brow-Forsythe apresentam poder de moderado a alto, em geral. Já o teste de O'Brien é o de pior performance quanto ao poder.

Os testes não-paramétricos de Siegel-Tukey, Ansari-Bradley, Klotz e Mood se assemelharam em performance, apresentando poder que varia de alto a baixo na dependência da distribuição dos erros, também influenciado pelo desbalanceamento dos dados. É interessante observar que apesar de os testes de Siegel-Tukey, Ansari-Bradley, Klotz e Mood suporem que os grupos não diferem em posição, isto não pareceu ter afetado o poder dos testes nos casos D, F e H, onde os efeitos de tratamentos não são nulos.

4 Discussão

Destaca-se a performance do teste de Bartlett quando a distribuição subjacente é normal, sendo nos casos de variâncias homogêneas um teste exato, preciso, aderindo à distribuição uniforme; e com alto poder, identificando eficientemente os casos com variâncias heterogêneas, sendo afetado pelo desbalanceamento dos dados. Muito embora mantenha elevado poder, com outras distribuições subjacentes, é muito conservador e impreciso, perdendo robustez. Desta forma, concorda-se com a recomendação de Milliken e Johson (1992) de que o teste de Bartlett deve ser a primeira opção no caso de dados pelo menos aproximadamente normais.

Tabela 3 - Poder (em percentual) dos testes para cada caso

Distribuição	*Testes*	*Caso*[I]					
		C	D	E	F	G	H
Normal	Bartlett	100,00	100,00	100,00	100,00	65,22	65,22
	Levene	89,98	89,98	91,19	91,19	39,44	39,44
	Levene	99,82	99,82	99,55	99,55	54,71	54,71
	Brow-	99,00	99,00	75,76	75,76	28,35	28,35
	O'Brien	85,20	85,20	35,06	35,06	27,54	27,54
	Siegel-	98,45	100	93,09	99,85	37,93	58,42
	Ansari-	98,52	100	93,06	99,86	38,01	58,42
	Klotz	98,85	99,92	89,95	94,48	39,79	35,37
	Mood	99,11	100	94,68	99,79	42,84	54,19
Uniforme	Bartlett	100,00	100,00	100,00	100,00	61,12	61,12
	Levene	97,35	97,35	92,20	92,20	62,02	62,02
	Levene	99,94	99,94	99,74	99,74	69,48	69,48
	Brow-	99,65	99,65	78,53	78,53	35,05	35,05
	O'Brien	95,79	95,79	44,99	44,99	50,72	50,72
	Siegel-	67,82	100	57,11	100	15,38	30,25
	Ansari-	68,36	100	57,34	100	15,48	30,11
	Klotz	37,69	100	39,31	100	10,14	11,05
	Mood	59,73	100	54,85	100	14,19	25,45
Exponencial	Bartlett	99,98	99,98	99,92	99,92	72,50	72,50
	Levene	64,19	64,19	83,40	83,40	26,33	26,33
	Levene	99,01	99,01	99,75	99,75	55,26	55,26
	Brow-	85,12	85,12	55,46	55,46	17,21	17,21
	O'Brien	57,83	57,83	26,92	26,92	19,67	19,67
	Siegel-	24,92	100	12,67	100	5,5	81,75
	Ansari-	25,42	100	12,73	100	5,53	81,93
	Klotz	11,22	100	10,21	99,98	4,09	52,59
	Mood	22,63	100	13,96	100	5,31	75,76

I – Caso C: Três tratamentos, com efeitos nulos de tratamentos e diferentes variâncias; **Caso D**: Idem Caso C mas com efeitos não-nulos de tratamentos; **Caso E**: Doze tratamentos, com efeitos nulos de tratamentos e diferentes variâncias; **Caso F**: Idem Caso E mas com efeitos não-nulos de tratamentos; **Caso G**: Quatro tratamentos, com efeitos nulos de tratamentos, diferentes variâncias e diferentes repetições; **Caso H**: Idem Caso G mas com efeitos não-nulos de tratamentos

O teste de Levene (absoluto) se assemelha ao de Bartlett quanto ao poder, mas é impreciso e liberal, sob qualquer distribuição. O teste de Levene(quadrado), mesmo que seja impreciso, é exato para distribuição subjacente normal e levemente liberal para as

demais, com poder inferior aos testes de Bartlett e de Levene(absoluto). Diante disso, indicar-se-ia o uso do teste de Levene como alternativa ao teste de Bartlett, quando a distribuição dos erros se afastar consideravelmente da normal, dada a sua maior robustez, preferentemente com as duas opções, quadrado para exatidão e absoluto para poder.

Os resultados concordaram, em parte, com os achados de Olejink e Algina (1987) de que os testes baseados na ANOVA da variável de dispersão são razoavelmente robustos à distribuição subjacente. Por outro lado, os autores encontraram que o teste de Brown-Forsythe parece ter bom desempenho em relação ao erro tipo I, o que não se confirmou nesse estudo, uma vez que o teste se mostrou conservador. Olejink e Algina (1987) e Conover et al. (1981) também encontraram superioridade do teste de Brown-Forsythe em relação a poder, o que também não ocorreu nesse estudo, já que ele foi superado tanto pelo teste de Bartlett como pelo teste de Levene (absoluto).

Os testes não-paramétricos de Siegel-Tukey,Ansari-Bradley e Mood se assemelham em performance, sendo exatos na ausência de efeito de tratamentos, imprecisos, apresentando poder de moderado a alto e influenciado pela distribuição subjacente, pela presença de efeito de tratamentos e pelo desbalanceamento dos dados. O teste de Klotz, quanto ao poder, apresentou comportamento similar aos outros testes não-paramétricos, mas teve performance conservadora independentemente da presença de efeito de tratamentos. Os achados foram totalmente divergentes dos de Conover et al. (1981) que concluíram que o teste de Klotz era superior em poder, já que em muitos dos casos este teste apresentou o menor poder.

Conclusão

Através do estudo de simulação com oito cenários, verificou-se grande variação no desempenho dos testes. Nenhum dos testes investigados foi superior em todas as medidas de desempenho avaliadas. Assim, ao invés de recomendar apenas um teste, adotar-se-á procedimento similar à recomendação apresentada por Milliken e Johson (1992). Recomenda-se utilizar o teste de Bartlett se dados aproximadamente normais, caso contrário, o teste de Levene preferentemente com as duas opções, quadrado para exatidão e absoluto para poder.

RIBOLDI, J.; BARBIAN, M. H.; KOLOWSKI, A. B. S.; SELAU, L. P. R.; TORMAN, V. B. L. Accuracy and power of parametric and non-parametric homocedasticy tests assessed for simulation. *Rev. Bras. Biom.*, São Paulo, v.32, n.3, p.xxx-xxx, 2014.

▪ *ABSTRACT: In the present study were compared the parametric tests of homogeneity of variances Bartlett, Brown-Forsythe, O'Brien, Levene with the absolute and square options; and non-parametric tests Siegel-Tukey, Ansari-Bradley, Klotz, and Mood, using simulation data. For this were simulated 10,000 experiments at SAS for eight different scenarios of analysis of variance for simple classification and errors with normal, uniform or exponential distribution. Comparisons under the homoscedasticity assumptions were made (for robustness) and heteroscedasticity (for power). Additionally, were considered null and non-null treatments and balanced or unbalanced data purposes. The results indicate that among the parametric tests investigated, Bartlett test showed better performance with high power, influenced by the unbalanced data and accuracy when the error distribution is normal. Levene test (absolute) resembles Bartlett in power, but it is inaccurate and liberal in any distribution of errors. Levene*

test (square) is exact for normal underlying distribution and slightly liberal for the other, less power than the Bartlett and Levene (absolute) test. Brown-Forsythe test is inaccurate, imprecise with power from moderate to high. O'Brien test were the worst performance, among the parametric tests, being conservative, imprecise and unstable power. Non-parametric tests Siegel-Tukey, Ansari-Bradley, Klotz and Mood are similar in power, are inaccurate and, with the exception of the Klotz test are accurate in the absence of treatment effects. We recommend using the Bartlett test data is approximately normal, otherwise, given their greater robustness, the Levene test preferably with two options, square for accuracy and for absolute power.

▪ *KEYWORDS: Homoscedasticity tests; simulation; precision; power.*

Referências

ANSARI, A. R.; BRADLEY, R. A. Rank-sum tests for dispersions. *Annals of Mathematical Statistics*, v.31, p.1174-1189, 1960.

BARTLETT, M. S. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society, Series A*, v.160, p.268–282, 1937.

BOX, G.E.P. Non-normality and tests on variances. *Biometrika,* v.40, p.318-335, 1953.

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

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

GARTSIDE, P. S. A study of methods for comparing several variances. *Journal of the American Statistical Association*, v.67, p.342-346, 1972.

HALL, I. J. Some comparisons of tests for equality of variances. *Journal of Statistical Computing and Simulation*, v.1, p.183-194, 1972.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. New Jersey: Prentice Hall, 1998.816p.

KESELMAN, H. J.; GAMES, P. A.; CLINCH J. J. Tests for homogeneity of variance. *Journal of Statistical Computation and Simulation*, v.8, p.113-129, 1979.

KLOTZ, J. Nonparametric tests for scale. *Annals of Mathematical Statistics*, v.33, p.498-512. 1962.

LAYARD, M. W. J. Robust large-sample tests for homogeneicity of variance *Journal of the American Statistical Association,* v.68, p.195-198, 1973.

LEVENE, H. Robust Tests for the equality of variance. In: Olkin, I(Ed.) *Contributions to Probability and Statistics,* Palo Alto, California: Stanford University Press, 1960. p.278–292.

LEVY, K. J. An empirical study of the cube-root test for homogeneity of variances with respect o the effects of non-normality and power. *Journal of Statistical Computing and Simulation,* v.7, p.71-78, 1978.

MILLIKEN, G.A.; JOHNSON, D.E. *Analysis of messy data. volume I: designed experiments*. Londres: Chapman & Hall, 1992. 473p.

MOOD, A. M. On the asymptotic efficiency of certain nonparametric two-sample tests. *Annals of Mathematical Statistics,* v.25, p.514-522, 1954.

MOOD, A.M.; GRAYBILL, F.A.; BOES, D.C. *Introduction to the theory of statistics*. 3.ed.New York: Mcgraw-Hill, 1974. 564p.

O'BRIEN, R. G. A General ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, v.74, p.877–880, 1979.

OLEJINK, S. F.; ALGINA, J. Type I error rates and power estimates of selected parametric and non-parametric tests of scale. *Journal of Educational Statistics*, v.12, p.45-61, 1987.

SIEGEL, S.; TUKEY, J. W.A nonparametric sum of ranks procedure for relative spread in unpaired samples. *Journal of the American Statistical Association*, v.55, p.429-445, 1960.

Recebido em 16.06.2014

Aprovado após revisão em 30.09.2014