Revista Colombiana de Estadística
Diciembre 2008, volumen 31, no. 2, pp. 241 a 260

Modificações e alternativas aos testes de Levene e de Brown e Forsythe para igualdade de variâncias e médias

Modifications and Alternatives to the Tests of Levene and Brown & Forsythe for Equality of Variances and Means

Antônia de Almeida[1,a], Silvia Elian[1,b], Juvêncio Nobre[2,c]

[1]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brasil

[2]Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Fortaleza, Brasil

Resumo

Os testes usuais para comparar variâncias e médias, teste de Bartlett e teste F, supõem que as amostras sejam provenientes de populações com distribuições normais. Para o teste de igualdade de médias, a suposição de homogeneidade de variâncias também é necessária. Alguns problemas se destacam quando tais suposições básicas são violadas, como tamanho excessivo e baixo poder. Neste trabalho descrevemos inicialmente o teste de Levene para igualdade de variâncias, que é robusto à não normalidade, e o teste de Brown e Forsythe para igualdade de médias quando existe desigualdade de variâncias. Apresentamos várias modificações do teste de Levene e do teste de Brown e Forsythe, propostas por diferentes autores. Analisamos e aplicamos uma forma do teste modificado de Brown e Forsythe a um conjunto de dados reais. Este teste é uma alternativa robusta com relação a desvios de normalidade e homocedasticidade e também na presença de observações discrepantes. Na comparação de variâncias, destaca-se o teste de Levene com centralização na mediana.

***Palavras chave*:** teste de Levene, teste de Brown e Forsythe, médias aparadas, variâncias winsorizadas, *bootstrap*.

[a]Mestre em Estatística. E-mail: erilaniaalmeida@yahoo.com.br
[b]Professor doutor. E-mail: selian@ime.usp.br
[c]Professor adjunto I. E-mail: juvencio@ufc.br

Abstract

The usual tests to compare variances and means (e.g. Bartlett's test and F-test) assume that the sample comes from a normal distribution. In addition, the test for equality of means requires the assumption of homogeneity of variances. In some situation those assumptions are not satisfied, hence we may face problems like excessive size and low power. In this paper, we describe two tests, namely the Levene's test for equality of variances, which is robust under nonnormality; and the Brown and Forsythe's test for equality of means. We also present some modifications of the Levene's test and Brown and Forsythe's test, proposed by different authors. We analyzed and applied one modified form of Brown and Forsythe's test to a real data set. This test is a robust alternative under nonnormality, heteroscedasticity and also when the data set has influential observations. The equality of variance can be well tested by Levene's test with centering at the sample median.

***Key words*:** Levene's test, Brown and Forsythe's test, Trimmed Means, Winsorized variances, *bootstrap*.

1. Introdução

Os testes de Levene (1960) e de Brown & Forsythe (1974*a*) têm-se constituído como técnicas úteis para comparação de médias e variâncias quando as suposições básicas dos testes de igualdade de variâncias e de igualdade de médias não são satisfeitas. Estes testes foram sofrendo modificações ao longo do tempo, propostas por diversos autores.

O teste de Bartlett para homogeneidade de variâncias não é robusto para divergência de normalidade. Visando contornar esse problema, propõe-se o uso do teste de Levene para a comparação de variâncias de grupos de observações provenientes de distribuições contínuas e não necessariamente normais. O teste de Levene é robusto à não normalidade, apesar que algumas deficiências foram destacadas por alguns autores, que também apresentaram algumas alternativas mais eficientes.

O teste F obtido através da análise de variância com um fator para comparar médias de populações normais independentes apresenta desvios no que tange ao tamanho do teste quando os grupos possuem variâncias populacionais diferentes. Para esse problema foram propostas várias soluções, entre elas o teste de Brown e Forsythe. Vários autores apontam inadequações no teste de Brown e Forsythe e apresentam algumas modificações para o mesmo.

O objetivo principal do presente artigo é apresentar e discutir as modificações propostas aos testes de Levene e de Brown e Forsythe. O artigo está organizado da seguinte forma: na seção 2 é apresentado o teste de Levene, que testa igualdade de variâncias quando os dados são oriundos de distribuições contínuas, não necessariamente distribuições normais, e algumas de suas modificações. Entre essas modificações está a proposta por Brown & Forsythe (1974*a*), que considera as distâncias das observações com relação às medianas amostrais ao invés das médias amostrais. Nesta versão, o teste se torna mais robusto para amostras pequenas e pode ser encontrado, por exemplo, no pacote computacional MINITAB 14. A seção

3 se destina ao estudo do teste de igualdade de médias com amostras independentes de populações normais com variâncias desiguais (teste de Brown e Forsythe). Também nesta seção são apresentadas algumas modificações deste teste, propostas por diferentes autores. Na seção 4 apresentamos testes aleatorizados para igualdade de médias e de variâncias em problemas de bioequivalência. Uma aplicação do teste modificado de Brown e Forsythe a um conjunto de dados reais e algumas conclusões encontram-se nas seções 5 e 6. Existem ainda alternativas bayesianas aos testes descritos, que não serão abordadas aqui, mas que podem ser encontradas, por exemplo, em Pereira & Stern (2003).

Para a execução do teste de Brown e Forsythe modificado foi desenvolvido um programa em linguagem de programação R (R Development Core Team 2007). O programa calcula a estatística do teste de Brown e Forsythe modificado e o tamanho do teste, que é estimado via *bootstrap*.

2. Teste de Levene e modificações

Muitas técnicas estatísticas requerem a suposição de igualdade de variâncias das variáveis de interesse para as populações envolvidas. O teste padrão de homogeneidade de variâncias (teste de Bartlett) é uma ferramenta eficiente somente se as variáveis possuem distribuição aproximadamente normal. Quando a suposição de normalidade é violada, o tamanho do teste (taxa de rejeição da hipótese nula, quando ela é verdadeira) pode ser muito maior do que o nível de significância fixado. Um procedimento relativamente insensível a desvios da normalidade é o teste de Levene. Este teste é robusto, já que, na ausência de normalidade, seu tamanho real é próximo do nível de significância fixado para uma grande variedade de distribuições de probabilidade.

Levene (1960) propôs uma estatística para testar igualdade de variâncias para estudos balanceados; posteriormente foi generalizada para estudos desbalanceados. A estatística é obtida a partir de uma análise de variância com um único fator, já que os níveis são as populações; cada observação i substituída pelo desvio absoluto da variável em relação à média do seu respectivo grupo.

Suponha que sejam tomadas $k \geq 2$ amostras aleatórias independentes entre si, digamos, $X_{i1}, \ldots, X_{in_i}$, $i = 1, \ldots, k$. A amostra i representa uma coleção de n_i variáveis aleatórias independentes e identicamente distribuídas (iid) com distribuição G_i, com média μ_i e variância σ_i^2, para G_i , μ_i e σ_i^2 desconhecidos. A hipótese nula de igualdade de variâncias,

$$\mathcal{H}_0 : \sigma_1^2 = \cdots = \sigma_k^2, \qquad i = 1, \ldots, k \tag{1}$$

é testada contra a hipótese alternativa que nem todas as variâncias são iguais, i.e.,

$$\mathcal{H}_1 : \sigma_i^2 \neq \sigma_j^2, \qquad \text{para algum } i \neq j, \quad j = 1, \ldots, k \tag{2}$$

Denotamos os desvios absolutos das variáveis X_{ij} com relação às médias amostrais dos grupos $\overline{X}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$ por $Z_{ij} = \left| X_{ij} - \overline{X}_{i\cdot} \right|$, $j = 1, \ldots, n_i$,

$i = 1, \ldots, k$ e definimos a estatística

$$W_0 = \left(\frac{n-k}{k-1}\right) \frac{\sum_{i=1}^{k} n_i \left(\overline{Z}_{i.} - \overline{Z}_{..}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_{i.}\right)^2} \tag{3}$$

em que $\overline{Z}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} Z_{ij}$, $\overline{Z}_{..} = n^{-1} \sum_{i=1}^{k} n_i \overline{Z}_{i.}$ e $n = \sum_{i=1}^{k} n_i$. O teste de Levene consiste em rejeitar $\mathcal{H}_0$ se $W_0 > F_{(k-1,n-k),(1-\alpha)}$; $F_{(k-1,n-k),(1-\alpha)}$ representa o quantil de ordem $1-\alpha$ da distribuição $F_{(k-1,n-k)}$ e α é o nível de significância do teste.

Portanto, o teste é uma análise de variância com um fator na variável desvio absoluto Z_{ij}. O uso de Z_{ij} ao invés de Z_{ij}^2 faz com que o critério do teste se torne menos sensível à ausência de normalidade, por exemplo, para distribuições com caudas pesadas. Mesmo assim, como em geral as variáveis aleatórias não são normalmente distribuídas nem independentes, pois $\mathrm{Cor}(Z_{ij}, Z_{ij'}) = O(n_i^{-2})$, para $j \neq j'$, o qual implica que W_0 sob a hipótese nula (1) não possui distribuição F.

No entanto, para uma variedade de distribuições G_i, por exemplo, distribuições normais, distribuições simétricas com caudas pesadas tais como a exponencial dupla e a t de Student com quatro graus de liberdade, em níveis de significâncias usuais, $\alpha = 0.01, 0.05$ ou 0.10 e amostras para cada grupo de tamanho pelo menos igual a 10 (i.e., $n_i \geq 10$, $i = 1, \ldots, k$), o teste de Levene se mostra robusto. Brown & Forsythe (1974*a*), num estudo de simulação, verificaram que, neste caso, o quantil de ordem $1-\alpha$ da distribuição nula de W_0, estimado pelo método de Monte Carlo, é aproximadamente igual a $F_{(k-1,n-k),(1-\alpha)}$. Verificaram ainda que a falta de robustez devia-se à assimetria das distribuições e não à existência de correlação entre os desvios. Estes fatos levaram à construção de formas alternativas do teste de Levene.

Para distribuições assimétricas, como a distribuição quiquadrado com 4 graus de liberdade, e distribuições com caudas extremamente pesadas, como a Cauchy, p.e., Brown & Forsythe (1974*a*) observaram que o teste de Levene tende a ser *liberal*, i.e., o tamanho do teste é maior que o nível de significância fixado. Por esse motivo, uma modificação do método de Levene é proposta pelos autores. A alteração consiste em substituir o estimador clássico do parâmetro de localização, $\overline{X}_{i.}$, em (3), por estimadores mais robustos.

Substituindo a média $\overline{X}_{i.}$ pela mediana do grupo, M_i, em (3), i.e., utilizandose $Z_{ij}^{(m)} = |Z_{ij} - M_i|$ ao invés de Z_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$, definese

$$W_{50} = \left(\frac{n-k}{k-1}\right) \frac{\sum_{i=1}^{k} n_i \left(\overline{Z}_{i.}^{(m)} - \overline{Z}_{..}^{(m)}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Z_{ij}^{(m)} - \overline{Z}_{i.}^{(m)}\right)^2} \tag{4}$$

em que $\overline{Z}_{i.}^{(m)} = n_i^{-1} \sum_{j=1}^{n_i} Z_{ij}^{(m)}$ e $\overline{Z}_{..}^{(m)} = n^{-1} \sum_{i=1}^{k} n_i \overline{Z}_{i.}^{(m)}$ e $n = \sum_{i=1}^{k} n_i$.

Utilizou-se ainda a estatística W_{10}, definida a partir de W_0, substituindo a média $\overline{X}_{i.}$ por $\widetilde{X}_i$, em que $\widetilde{X}_i$ representa a média aparada a 10% do i-ésimo grupo.

Brown & Forsythe (1974*a*) realizaram um estudo de simulação cujos resultados indicaram que a igualdade de variâncias em distribuições de caudas pesadas pode ser melhor testada por uma estatística da forma W_{10} e, em distribuições assimétricas, por uma estatística similar a W_{50}. Portanto, quando se tem indícios de desvios de normalidade, a estimativa da média para cada grupo na estatística de Levene deve ser substituída por uma estimativa mais robusta do parâmetro de localização. A perda no poder observada quando é usada W_{10} ao invés de W_0 é pequena, relativa ao aumento da probabilidade de uma rejeição falsa da hipótese nula causada pela não normalidade.

Carrol & Schneider (1985) apresentam um interessante estudo sobre as estatísticas W_0 e W_{50}. Analisam inicialmente o motivo pelo qual os testes de Levene utilizando as estatísticas W_0 e W_{50} têm tamanhos próximos dos níveis corretos para distribuições simétricas. Posteriormente, investigam a razão pela qual o teste de Levene que utiliza a mediana (estatística W_{50}) tem tamanho aproximadamente igual ao nível de significância para distribuições assimétricas, enquanto o teste de Levene baseado em (3) não possui essa propriedade. Os autores consideraram estes baseados em $W_{ij} = \left|X_{ij} - \widehat{\theta}_i\right|$ e em $G(W_{ij})$, com $G(\cdot)$ representando uma função monótona com derivada $g(\cdot)$ e $\widehat{\theta}_i$ um estimador de θ_i, e analisaram a eficiência dos testes sob hipóteses nulas de variâncias iguais. O modelo considerado foi

$$X_{ij} = \theta_i + e_{ij}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, n_i$$

com θ_i representando um parâmetro de localização da distribuição de X_{ij}, p.e., a média, a mediana ou a média aparada a 10%. As variáveis aleatórias e_{ij} são consideradas independentes e identicamente distribuídas, já que a análise foi feita somente sob a hipótese de igualdade de variâncias (1). Como conseqüência, as conclusões obtidas serão relativas à significância e não ao poder dos testes. Além disso, admitese que o estimador $\widehat{\theta}_i$ é tal que, quando $n_i \to \infty$: $\sqrt{n_i}\big(\widehat{\theta}_i - \theta_i\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \xi^2\big)$, em que $\xi^2 > 0$, para $i = 1, \ldots, k$.

Foi considerada uma classe geral de testes baseados em $G(W_{ij})$, com estatística de teste dada por

$$F_n\big(\widehat{\theta}_i\big) = \frac{\mathrm{QME}\big(\widehat{\theta}_i\big)}{\mathrm{QMD}\big(\widehat{\theta}_i\big)} \tag{5}$$

em que $\mathrm{QME}\big(\widehat{\theta}_i\big) = \sum_{i=1}^{k} n_i\big(\overline{R}_{i\cdot} - \overline{R}_{\cdot\cdot}\big)^2/(k-1)$, $\mathrm{QMD}\big(\widehat{\theta}_i\big) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(R_{ij} - \overline{R}_{i\cdot}\big)^2/(n-k)$, $\overline{R}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} R_{ij}$ e $\overline{R}_{\cdot\cdot} = n^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij}$ e $F_n(\theta_i)$ é a estatística (5) calculada em função de $L_{ij} = G(Z_{ij})$, com $Z_{ij} = |X_{ij} - \theta_i|$.

A seguir, apresentamos um teorema que será útil no desenvolvimento do trabalho, que pode ser demonstrado usando os resultados de Bickel (2005) e de Carrol & Ruppert (1982) e é apresentado em Carrol & Schneider (1985).

Teorema 1. *Para $F_n\big(\widehat{\theta}_i\big)$ e $F_n(\theta_i)$ definidas anteriormente, temos*

$$F_n\big(\widehat{\theta}_i\big) - F_n(\theta_i) - Q_n \xrightarrow{\mathbb{P}} 0$$

em que $Q_n = \dfrac{\sum_{i=1}^{k} \gamma^2 H_i^2 - 2\gamma H_i \sqrt{n_i}\big(\overline{L}_{i.} - \overline{L}_{..}\big)}{\sigma^2(k-1)}$, $\gamma = \int_0^\infty g(u)\big\{f(u) - f(-u)\big\}\,du$, $H_i = \sqrt{n_i}\Big\{\big(\widehat{\theta}_i - \theta_i\big) - \sum_{i=1}^{k}\big(\frac{n_i}{n}\big)\big(\widehat{\theta}_i - \theta_i\big)\Big\}$, *com* f *representando a função densidade de probabilidade de* e_{ij}.

Em particular, se $\widehat{\theta}_i$ é um estimador de θ_i para $G\big(|X_{ij} - \theta_i|\big) = |X_{ij} - \theta_i|$, de modo que $G(u) = u$, segue que $g(u) = 1$ e $\gamma = \int_0^\infty f(u)\,du - \int_0^\infty f(-u)\,du = \mathbb{P}(e_{ij} > 0) - \mathbb{P}(e_{ij} < 0) = \mathbb{P}(X_{ij} - \theta_{ij} > 0) - \mathbb{P}(X_{ij} - \theta_{ij} < 0)$.

Podemos interpretar $F_n(\theta_i)$ como uma medida de variabilidade das quantidades $G\big(|X_{ij} - \theta_i|\big)$ para os diferentes valores de θ_i, $i = 1, \ldots, k$ e $F_n\big(\widehat{\theta}_i\big)$ como um preditor de $F_n(\theta_i)$. Situações nas quais $Q_n = 0$, implicando que $F_n\big(\widehat{\theta}_i\big) - F_n(\theta_i) = o_p(1)$, são indicativas de que a estatística do teste de Levene, $F_n(\widehat{\theta}_i)$, é eficiente na detecção da variabilidade em $G\big(|X_{ij} - \theta_i|\big)$ para $i = 1, \ldots, k$. Desta forma, o teste de Levene terá assintoticamente nível de significância correto somente quando $\gamma = 0$, i.e., para θ_i igual à mediana de X_{ij}. Isso explica a razão para a qual, em distribuições simétricas, a média ou mediana podem ser utilizadas sem alterações significativas no tamanho do teste. Já, para distribuições assimétricas, $\gamma \neq 0$, a menos que θ_i seja a mediana de X_{ij}.

Um segundo caso especial interessante ocorre quando $G(u) = u^2$, ou seja, quando o teste é baseado em $R_{ij} = \big(X_{ij} - \widehat{\theta}_i\big)^2$. Neste caso $g(u) = 2u$ e $\gamma = 2\mathbb{E}[X_{ij} - \theta_i]$. Assim, para grandes amostras de distribuições assimétricas, o teste teria tamanho igual ao nível de significância somente quando a centralização é na média, situação em que $\gamma = 0$. Se $\gamma \neq 0$, então pelo teorema 1, temos que para centralização na mediana, o tamanho do teste é diferente do nível de significância fixado (conclusão obtida por Conover et al. (1981), via simulação).

Observa-se então que o teste de Levene terá nível assintótico correto somente quando as estimativas utilizadas estiverem estimando a mediana de cada população. Dessa maneira, para distribuições simétricas, as formas centralizadas na média e mediana serão igualmente eficientes. Para distribuições assimétricas, temos que assintoticamente, só a centralização na mediana é viável e, se utilizada a centralização na média, o teste de Levene terá tamanho diferente do nível de significância fixado.

O poder do teste de Levene de homogeneidade de variâncias, que emprega os desvios calculados usando a mediana amostral, pode ser aumentado com as modificações propostas por Hines & O'Hara Hines (2000). Os autores propõem identificar e remover os chamados zeros estruturais, visto que o procedimento de Levene pode apresentar falhas; por exemplo, ignorar a falta de independência dos desvios envolvidos, não explorar o fato de que médias (ou medianas) e variâncias de algumas variáveis aleatórias não são funcionalmente independentes, como nas distribuições Poisson e Binomial. Os zeros estruturais correspondem a valores de Z_{ij} nulos, que sempre existem para n_i ímpar, pois nesse caso, a mediana coincide com um dos valores de X_{ij}. Segundo os autores, a presença desses zeros estruturais pode tornar o teste ineficiente para detectar desigualdade de variâncias. Hines & O'Hara Hines (2000) mostram que a retirada de zeros estruturais au-

menta o poder do teste para detectar essas desigualdades, principalmente quando n é pequeno. Sugerem ainda a complementação da análise através do posterior uso de contrastes quando o resultado do teste aponta para a desigualdade de variâncias. Outra modificação para o teste de Levene pode ser encontrada em O'Neill & Mathews (2000). Os autores propuseram uma forma alternativa do teste de Levene construída com base no procedimento de mínimos quadrados ponderados.

3. Teste de Brown e Forsythe e modificações

Consideremos X_{ij} a j-ésima observação da i-ésima amostra correspondente ao i-ésimo grupo, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$, com independência entre observações de grupos distintos, de forma que $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. A estatística F da análise de variância com um fator para testar a hipótese nula de igualdade de médias no nível de significância α contra a hipótese alternativa que nem todas as médias são as mesmas, i.e., testar

$$\mathcal{H}_0 : \mu_1 = \cdots = \mu_k \tag{6}$$

versus

$$\mathcal{H}_1 : \mu_t \neq \mu_g, \qquad \text{para algum } t \neq g, \quad t, g = 1, \ldots, k \tag{7}$$

é dada por

$$F = \frac{\sum_{i=1}^{k} n_i \big(\overline{X}_{i.} - \overline{X}_{..}\big)^2/(k-1)}{\sum_{i=1}^{k}(n_i - 1)S_i^2/(n-k)}$$

em que $S_i^2 = \sum_{j=1}^{n_i - 1} \big(X_{ij} - \overline{X}_{i.}\big)^2/(n_i - 1)$ e $\overline{X}_{..} = n^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$. Se as variâncias populacionais são iguais, então sob $\mathcal{H}_0$, estatística de teste possui distribuição $F(k-1, n-k)$.

O teste baseado na estatística F é sensível à falta de homogeneidade de variâncias, pois, sob heterocedasticidade, o tamanho real do teste não coincide com o nível de significância. O problema de comparar médias de distribuições normais independentes para três ou mais grupos quando há heterocedasticidade é conhecido na literatura como problema de Behrens-Fisher generalizado. Várias soluções foram propostas para esse problema, entre elas, o teste proposto por Brown & Forsythe (1974*b*), que leva seu nome e será descrito a seguir.

Considere agora que $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$, com X_{ij} e $X_{i'j'}$ independentes para quaisquer $\{i, j\} \neq \{i', j'\}$, e que o interesse é testar (6) *vs.* (7) no nível de significância α. A estatística de teste proposta por Brown e Forsythe é definida como

$$F^* = \frac{\sum_{i=1}^{k} n_i \big(\overline{X}_{i.} - \overline{X}_{..}\big)^2}{\sum_{i=1}^{k} \big(1 - \frac{n_i}{n}\big) S_i^2} \tag{8}$$

Os valores críticos de F^*, sob $\mathcal{H}_0$, são obtidos da distribuição $F(k-1, f)$, em que f, obtida pela aproximação de Satterthwaite (1941) (para detalhes, veja

Brown & Forsythe 1974*b*), é dado por

$$f^{-1} = \sum_{i=1}^{k} \frac{c_i^2}{n_i - 1} \qquad \text{e} \qquad c_i = \frac{\left(1 - \frac{n_i}{n}\right) S_i^2}{\sum_{i=1}^{k} \left(1 - \frac{n_i}{n}\right) S_i^2}$$

Através de estudos de simulação, utilizando diferentes valores para variâncias populacionais e diferentes tamanhos amostrais, Brown & Forsythe (1974*b*) compararam o tamanho e o poder do teste de igualdade de médias populacionais na situação de desigualdade de variâncias utilizando as estatísticas F, F^* e as propostas por Welch (1951) e James (1954).

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

$$W = \frac{\sum_{i=1}^{k} w_i \left(\overline{X}_{i.} - \widetilde{X}_{..}\right)^2 / (k-1)}{\left(1 + \frac{2(k-2)}{(k^2-1)} \sum_{i=1}^{k} \frac{\left(1 - \frac{w_i}{u}\right)^2}{n_i - 1}\right)}$$

em que $w_i = n_i / S_i^2$, $i = 1, \ldots, k$, $u = \sum_{i=1}^{k} w_i$ e $\widetilde{X}_{..} = \sum_{i=1}^{k} w_i \overline{X}_{i.}/u$. Sob $\mathcal{H}_0$ em (6), temos $W \xrightarrow{\mathcal{D}} F(k-1, f)$, em que

$$f^{-1} = \frac{3}{k^2 - 1} \sum_{i=1}^{k} \frac{\left(1 - \frac{w_i}{u}\right)^2}{n_i - 1}$$

Por outro lado, a estatística de teste proposta por James (1954) é

$$J = \sum_{i=1}^{k} \frac{w_i \left(\overline{X}_{i.} - \widetilde{X}_{..}\right)^2}{k - 1}$$

De acordo com James (1954), sob $\mathcal{H}_0$ em (6),

$$\mathbb{E}\left[J > a\left(1 + \frac{3a + (k+1)}{2(k^2-1)} \sum_{i=1}^{k} \frac{\left(1 - \frac{w_i}{u}\right)^2}{n_i - 1}\right)\right] = \alpha$$

com a representando o quantil de ordem $1 - \alpha$ da distribuição quiquadrado com $k - 1$ graus de liberdade (χ^2_{k-1}).

Brown & Forsythe (1974*b*) verificaram, através de um estudo de simulação, que o teste F apresenta acentuados desvios no tamanho quando as variâncias dos grupos são desiguais, e os outros três testes pequenas flutuações no tamanho. Para amostras de tamanho pequeno, verificaram que o teste baseado na estatística J apresenta uma característica liberal, i.e., o tamanho do teste é maior do que o nível de significância fixado. O teste baseado na estatística F^* variou em tamanho um pouco mais do que o teste baseado na estatística W e em situações com mais de 10 observações por grupo; a diferença entre os níveis de significância e os tamanhos foi pequena, para os dois testes em questão. Na situação de homoscedasticidade, os testes baseados nas estatísticas W e F^* apresentaram poder similar ao teste

F clássico. Observou-se ainda que para grupos com médias discrepantes com variâncias pequenas com relação ao tamanho da amostra do grupo, o teste baseado na estatística W é mais poderoso do que o teste baseado na estatística F^*. Tal fato pode ser explicado pela diferente ponderação de médias: W ponderava as médias usando $\left(n_i/S_i^2\right)$, enquanto que a estatística F^* utiliza n_i. Dessa forma, médias extremas com variâncias pequenas com relação ao tamanho da amostra do grupo tenderiam a inflacionar o valor de W, mais do que F^*, ocorrendo o inverso para médias discrepantes com variâncias grandes com relação ao tamanho da amostra do grupo. Quando médias discrepantes vem acompanhadas de variâncias grandes, o teste de Brown e Forsythe mostrou-se superior.

Mehrotra (1997) aponta uma inadequação na aproximação proposta por Brown & Forsythe (1974*b*) para a distribuição da estatística de teste F^* sob $\mathcal{H}_0$ em (6). Adicionalmente, Mehrotra (1997) mostra que sob $\mathcal{H}_0$, $F^* \approx F(f_1, f_2)$, com

$$f_1 = \frac{\left(\sum_{i=1}^k \sigma_i^2 - \frac{\sum_{i=1}^k n_i\sigma_i^2}{n}\right)^2}{\sum_{i=1}^k \sigma_i^4 + \left(\frac{\sum_{i=1}^k n_i\sigma_i^2}{n}\right)^2 - 2\frac{\sum_{i=1}^k n_i\sigma_i^4}{n}} \tag{9}$$

e

$$f_2 = \frac{\left[\sum_{i=1}^k \left(1-\frac{n_i}{n}\right)\sigma_i^2\right]^2}{\sum_{i=1}^k \frac{\left(1-\frac{n_i}{n}\right)^2\sigma_i^4}{n_i-1}} \tag{10}$$

e não como havia sido proposta por Brown & Forsythe (1974*b*), em que a distribuição era aproximadamente igual a $F(k-1, f_2)$. Na prática, os valores de σ_i^2, presentes em (9) e (10) devem ser substituídos por estimadores consistentes, p.e., S_i^2.

Keselman & Wilcox (1999) mostraram que, sob condições semelhantes envolvendo variâncias heterogêneas e também não normalidade, o teste modificado proposto por Brown e Forsythe com hipótese nula (6) e estatística de teste (10) apresenta um aumento na probabilidade do erro do tipo I em modelos não balanceados. Os autores propõem um procedimento que consiste em um teste para igualdade dos parâmetros de localização no qual a estatística de teste utiliza estimadores robustos dos parâmetros de localização e de dispersão, por exemplo, médias aparadas e variâncias "winsorizadas", ao invés dos estimadores usuais. Sugerem ainda que os valores críticos associados a um particular nível de significância sejam obtidos através do método de *bootstrap*.

A idéia é substituir a hipótese de igualdade de médias (6), por exemplo, pela de igualdade de médias aparadas

$$\mathcal{H}_0 : \mu_{a1} = \cdots = \mu_{ak} \tag{11}$$

Os autores consideraram que as n_i observações da população i, $X_{i1}, \ldots, X_{in_i}$ são independentes, com $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, em que $\sigma_i^2 \neq \sigma_{i'}^2$ para algum $i \neq i'$, $i, i' = 1, \ldots, k$.

Sejam $X_{(1)i} \leq X_{(2)i} \leq \cdots \leq X_{(n_i)i}$ as observações ordenadas do i-ésimo grupo e $J_i = [n_i\gamma]$, com γ representando a proporção de observações aparadas de cada cauda da distribuição. Dessa forma, o tamanho amostral efetivo do i-ésimo grupo é $h_i = n_i - 2J_i$. A i-ésima média amostral aparada é $\overline{X}_{ai.} = h_i^{-1} \sum_{j=J_i+1}^{n_i-J_i} X_{(j)i}$. Define-se a i-ésima média winsorizada por $\overline{X}_{wi.} = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$, em que

$$Y_{ij} = \begin{cases} X_{(J_i+1)i}, & \text{se } X_{ij} \leq X_{(J_i+1)i}; \\ X_{ij}, & \text{se } X_{(J_i+1)i} < X_{ij} < X_{(n_i-J_i)i}; \\ X_{(n_i-J_i)i}, & \text{se } X_{ij} \geq X_{(n_i-J_i)i}. \end{cases}$$

A variância amostral winsoriada do i-ésimo grupo é definida por $S^2_{wi} = (n_i - 1)^{-1} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{X}_{wi.}\right)^2$. De acordo com Wilcox (1996), uma estimativa da variância da média amostral aparada é dada por $\widetilde{S}^2_{wi} = (n_i - 1)S^2_{wi}/(h_i(h_i - 1))$. De posse desses estimadores robustos, podemos ter uma versão robusta da estatística de teste de Welch, em que são utilizadas as médias aparadas dos grupos ($\overline{X}_{wi.}$) ao invés das médias amostrais ($\overline{X}_{i.}$) e as variâncias amostrais winsorizadas (S^2_{wi}) ao invés das variâncias amostrais usuais (S^2_i) e $\sum_{i=1}^{n} h_i$ ao invés de n, de forma a obter:

$$F_a = \frac{\sum_{i=1}^{k} w_{ai}\left(\overline{X}_{ai.} - \widetilde{X}_{a..}\right)^2/(k-1)}{\left(1 + \frac{2(k-2)}{(k^2-1)} \sum_{i=1}^{k} \frac{\left(1-\frac{w_{ai}}{u_a}\right)^2}{h_i - 1}\right)}$$

com $w_{ai} = h_i/S^2_{wi}$, $i = 1, \ldots, k$, $u_a = \sum_{i=1}^{k} w_{ai}$ e $\widetilde{X}_{a..} = \sum_{i=1}^{k} w_{ai}\overline{X}_{ai.}/u_a$ e $f_a = (k^2 - 1)/c^*$, com $c^* = \left(3\sum_{i=1}^{k}(1 - w_{ai}/u_a)^2/(h_i - 1)\right)$. Definida a estatística de teste, o valor crítico é estimado através da obtenção de sua distribuição empírica via métodos de reamostragem, por exemplo o *bootstrap* (para mais detalhes sobre o método *bootstrap*, veja Davison & Hinkley 1997). O procedimento utilizado é sucintamente descrito a seguir.

Sejam $C_{ij} = X_{ij} - \overline{X}_{ai.}$, $i = 1, \ldots, k$ e $j = 1, \ldots, n_i$, os valores das variáveis originais centralizados pela média aparada. Para o i-ésimo grupo, determinase a amostra *bootstrap* selecionando aleatoriamente, com reposição, n_i observações dentre as C_{ij}, $j = 1, \ldots, n_i$, de forma a obter a amostra *bootstrap* $X^*_{i1}, \ldots, X^*_{in_i}$ para $i = 1, \ldots, k$. Denotando F^*_a o valor da estatística baseada na amostra *bootstrap*, repete-se o procedimento B vezes, obtendo $F^*_{a1}, \ldots, F^*_{aB}$. O nível de significância estimado p^* é a proporção de vezes que a estatística é maior do que a do teste baseado nos dados originais. Obtido p^*, se $p^* \leq \alpha$, rejeitamos a hipótese $\mathcal{H}_0$ em (11). Wilcox (1996) sugere a utilização de $B = 599$, de forma a obter um controle satisfatório na probabilidade do erro do tipo I.

4. Testes aleatorizados para igualdade de médias e de variâncias e aplicações em bioequivalência

4.1. Testes aleatorizados para igualdade de médias e variâncias

Manly (1995), Francis & Manly (2001, 2002) introduzem testes aleatorizados de igualdade de médias e variâncias que podem ser utilizados como alternativas aos testes de Levene e de Brown e Forsythe. Segundo os autores, a tarefa aparentemente simples de comparar médias e variâncias de duas ou mais populações é na verdade bastante difícil quando as amostras são provenientes de distribuições "muito distantes" da normal. De acordo com Manly & Francis (2002), há três problemas potenciais:

1. Se não existe desigualdade nas variâncias entre as populações, os testes para comparar as médias podem ter propriedades pobres (tamanho alto ou poder baixo) devido à não normalidade.

2. Se existem diferenças nas variâncias mas não nas médias, os testes de igualdade de médias podem ter tamanhos excessivos devido à sensibilidade à desigualdade de variâncias.

3. Testes para comparar variâncias mesmo quando não afetados por diferenças nas médias, podem ter propriedades pobres como conseqüência da não normalidade.

Um teste aleatorizado é um particular teste de permutação baseado em aleatorização. É realizado da seguinte maneira: a estatística de teste é calculada para cada um dos resultados das permutações dos dados. Essas permutações, incluída aquela representando os resultados obtidos inicialmente, constituem o conjunto referência para determinar o nível de significância. A proporção de permutações, no conjunto referência, com valores da estatística de teste maiores ou iguais (ou, para certas estatísticas de teste, menores ou iguais) ao valor obtido experimentalmente é o nível descritivo (valor p). Definir a significância com base na distribuição de estatísticas de teste geradas por permutar os dados é característica de todos os testes de permutação. Um teste de permutação é chamado teste aleatorizado quando a permutação dos dados é feita através de escolhas aleatórias. A hipótese nula para um teste aleatorizado é de que a distribuição de probabilidades da variável para cada unidade experimental é a mesma qualquer que seja a associação dos tratamentos. Assim, sob a hipótese nula, a escolha de unidades experimentais para os tratamentos teria sido feita de forma aleatória.

Manly & Francis (2002) apresentam, como proposta para o problema de comparar médias e variâncias, um teste aleatorizado conjunto, cuja hipótese nula é que as amostras comparadas provêm da mesma distribuição. Os autores propõem o uso da estatística da razão de verossimilhanças para testar se as amostras são provenientes de uma mesma distribuição normal. Se X_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$

correspondem às amostras aleatórias das populações, a estatística de teste é calculada como

$$T_0 = \sum_{i=1}^{k} n_i \ln\left(\frac{V_T}{V_i}\right)$$

em que $V_i = \sum_{j=1}^{n_i} n_i^{-1}\left(X_{ij} - \overline{X}_i\right)^2$ é o estimador de máxima verossimilhança da variância da i-ésima população e $V_T = n^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}\right)^2$ é o estimador de máxima verossimilhança da variância comum. A estatística T_0 é sensível tanto a diferenças entre médias como a diferenças entre variâncias. O fato das amostras serem provenientes de uma distribuição normal é irrelevante se a significância de um valor observado é obtida por aleatorização. Dessa forma T_0 é considerado significante ao nível α quando exceder o quantil de ordem $(1-\alpha)$ do conjunto consistindo do próprio T_0 e de $M-1$ valores de T_0 calculados após permutação aleatória dos dados observados às amostras. Se T_0 for calculado para um conjunto de dados e for encontrado um resultado não significante com o teste aleatorizado, é razoável concluir que não há qualquer evidência para diferenças entre as distribuições, em termos de médias e variâncias. Por outro lado, se T_0 for significante, então há evidências de diferenças. Para identificar se as diferenças são nas médias, variâncias ou ambas, é necessário algum teste adicional. Os autores constroem ainda testes separados para igualdade de médias e igualdade de variâncias, que podem ser encontrados em Manly (1995) e Francis & Manly (2001). Posteriormente, esta metodologia é aplicada a problemas de bioequivalência conforme descrita brevemente a seguir.

4.2. Aplicações em bioequivalência

Um problema comum na ciência ambiental é a comparação entre um local sob controle e um local "tratado" que pode estar danificado, para decidir se os dois são similares em termos da distribuição de alguma medida de saúde ambiental. Nessa área, o uso de um teste de significância padrão tem dois problemas. Primeiro, não é razoável supor que o local tratado e o sob controle terão exatamente a mesma média para a variável em estudo, mesmo na ausência de qualquer dano no local tratado. Neste caso, tomadas amostras grandes de cada local, haverá uma alta probabilidade de se detectar diferenças, independentemente da extensão na qual o local tratado esteja danificado. Além disso, quando o teste para uma diferença entre os dois locais não apresentar resultados significantes, isso não significará necessariamente que não exista uma diferença importante. Uma explicação alternativa seria que o tamanho da amostra não foi suficientemente grande para detectar a diferença. Essas considerações sugerem que a questão de interesse pode não ser se há uma diferença significante entre os locais, mas sim se a diferença é de importância prática. Uma forma de abordar esse tipo de problema é usando o conceito de bioequivalência.

Na área farmacêutica, uma nova droga pode ser bioequivalente a uma droga padrão se sua potência é, por exemplo, mais do que 80% da potência da droga padrão. Da mesma forma, um local tratado pode ser considerado bioequivalente a outro sob controle em termos da biomassa (peso total da matéria viva em uma área

determinada) da vegetação se a média da biomassa por unidade de área no local tratado, μ_t, é maior do que 80% da média no local sob controle, μ_c. Nesse caso, a bioequivalência pode ser examinada testando a hipótese nula $\mathcal{H}_0 : \mu_t \leq 0.8\mu_c$ contra a hipótese alternativa $\mathcal{H}_1 : \mu_t > 0.8\mu_c$. Um resultado significante fornece evidências de bioequivalência, já um resultado não significante sugere que o local tratado pode estar prejudicado. Nessa área, duas diferentes drogas ou formulações de uma mesma droga são chamadas bioequivalentes se elas são absorvidas pelo sangue e se tornam disponíveis no mesmo ritmo e concentração.

O teste de bioequivalência consiste em verificar se o medicamento genérico apresenta a mesma biodisponibilidade no organismo que o respectivo medicamento de referência. O medicamento de referência é aquele que passou por pesquisa clínica para comprovar sua eficácia e segurança antes do registro junto ao Ministério da Saúde, através da ANVISA (Agência Nacional de Vigilância Sanitária –Brasil). A biodisponibilidade relaciona-se à quantidade absorvida e à velocidade do processo de absorção do fármaco ou princípio ativo (substância existente na formulação do medicamento, responsável pelo seu efeito terapêutico). Quando dois medicamentos apresentam a mesma biodisponibilidade no organismo, sua eficácia clínica é considerada comparável. A bioequivalência, na grande maioria dos casos, assegura que o medicamento genérico é equivalente terapêutico do medicamento de referência, ou seja, que apresenta a mesma eficácia clínica e a mesma segurança em relação ao medicamento de referência.

Manly (2004) discute como realizar testes unilaterais para bioequivalência quando as fontes de variação não possuem distribuição normal com heterocedasticidade. Os exemplos apresentados pelo autor envolvem a medida dos níveis de arsênio em uma região de munição de um campo do exército e a vegetação que cobre um local minado recuperado. O local minado era tido como recuperado se sua produção fosse maior do que 90% da produção do local sobre controle. O procedimento proposto consiste inicialmente na aplicação do teste aleatorizado conjunto para diferença de médias e variâncias nos dois locais. Posteriormente, o autor desenvolve e aplica um teste unilateral para diferença de médias com estatísticas

$$t_1 = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \qquad \text{ou} \qquad t_2 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

com respectivo valor p, baseado no número de configurações dos dados aleatorizados, que são maiores ou iguais ao valor da estatística de teste observado. Uma posterior avaliação da confiabilidade do teste aleatorizado era feita através do procedimento *bootstrap*.

5. Aplicação

Nesta seção apresentamos uma aplicação do teste de Brown e Forsythe com as modificações propostas por Keselman & Wilcox (1999), em um conjunto de dados fornecido pelo Centro de Estatística Aplicada (CEA) IME-USP. Estes dados constituem uma parte do conjunto analisado em Elian & Santos (2003).

Segundo este trabalho, no Brasil, as micros e pequenas empresas, que representam 98% do total, exercem um papel significante na economia brasileira. Estudos realizados mostram que essas empresas são responsáveis por 35 milhões de empregados e 20% do produto interno bruto. Da mesma forma que 1.5 milhões de empresas do estado de São Paulo estão iniciando as suas atividades, em contrapartida, 1 milhão estão decretando falência. Uma das prováveis razões pela qual as empresas fracassam é a inexistência de um planejamento. O estudo realizado se baseou em questionários aplicados a 115 empreendedores de micro (0 a 9 empregados) e pequenas empresas (10 a 99 empregados). Diferentes variáveis foram obtidas através do questionário, porém foram considerada para análise as variáveis **variação percentual do faturamento bruto** (VALFAT), que representa a variação percentual do faturamento bruto no período de 2000 a 2002, considerando como base 1999; e a variável **treinamento**: se a empresa participou alguma vez do curso de treinamento oferecido pelo Serviço Brasileiro de Apoio às micros e pequenas Empresas (SEBRAE): Sim ou Não. O objetivo é verificar se as médias da variável VALFAT são iguais para as empresas que realizam e não realizam o treinamento. Os dados estão dispostos na tabela 1.

Inicialmente, foram calculadas algumas medidas descritivas para a variável VALFAT, apresentadas na tabela 2. Observamos que 50% dos menores valores de VALFAT com (sem) treinamento estão entre 0 e 138.0 (127.0) e que o máximo dessa variável é um valor muito maior do que o 3º quartil. Analisando a figura 1, concluímos que para os dois níveis da variável treinamento, a variável apresenta valores discrepantes, inclusive com evidências de assimetria.

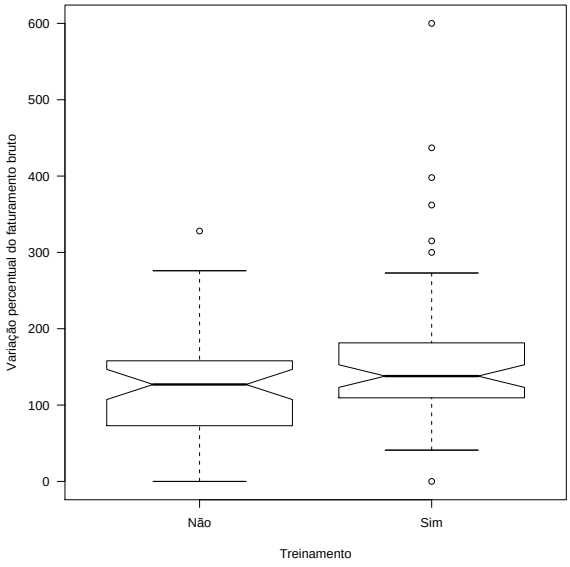

Figura 1: Boxplot para a variação percentual de faturamento bruto segundo o treinamento das empresas.

A análise dos histogramas na figura 2 sugere que os dados não são normalmente distribuídos, existindo inclusive indícios de assimetria. As respectivas tracejadas

Tabela 1: Variação percentual de faturamento bruto segundo o treinamento das empresas.

Empresa	Variação	Treinamento	Empresa	Variação	Treinamento
1	259	Sim	54	85	Sim
2	41	Sim	55	129	Sim
3	398	Sim	56	300	Sim
4	95	Sim	57	140	Sim
5	120	Sim	58	245	Sim
6	127	Sim	59	92	Sim
7	53	Sim	60	180	Não
8	172	Sim	61	138	Não
9	192	Sim	62	0	Não
10	110	Sim	63	166	Não
11	151	Sim	64	110	Não
12	176	Sim	65	276	Não
13	0	Sim	66	40	Não
14	437	Sim	67	133	Não
15	165	Sim	68	42	Não
16	127	Sim	69	60	Não
17	144	Sim	70	133	Não
18	600	Sim	71	0	Não
19	136	Sim	72	110	Não
20	130	Sim	73	89	Não
21	82	Sim	74	108	Não
22	84	Sim	75	201	Não
23	163	Sim	76	154	Não
24	111	Sim	77	133	Não
25	110	Sim	78	96	Não
26	187	Sim	79	151	Não
27	102	Sim	80	158	Não
28	139	Sim	81	68	Não
29	147	Sim	82	73	Não
30	120	Sim	83	72	Não
31	100	Sim	84	133	Não
32	216	Sim	85	100	Não
33	362	Sim	86	38	Não
34	191	Sim	87	257	Não
35	125	Sim	88	24	Não
36	127	Sim	89	129	Não
37	104	Sim	90	70	Não
38	120	Sim	91	85	Não
39	109	Sim	92	184	Não
40	273	Sim	93	203	Não
41	157	Sim	94	27	Não
42	125	Sim	95	125	Não
43	141	Sim	96	100	Não
44	198	Sim	97	137	Não
45	77	Sim	98	148	Não
46	138	Sim	99	225	Não
47	170	Sim	100	100	Não
48	145	Sim	101	180	Não
49	62	Sim	102	110	Não
50	315	Sim	103	328	Não
51	152	Sim	104	230	Não
52	206	Sim	105	130	Não
53	97	Sim			

Tabela 2: Medidas descritivas para a variação percentual de faturamento bruto segundo o treinamento das empresas.

VALFAT	Treinamento	
	Sim	Não
Média	162.4	125.1
Desvio padrão	100.4	70.6
Mediana	138.0	127.0
1º Quartil	109.0	72.8
3º Quartil	187.0	160.0
Mínimo	0.0	0.0
Máximo	600.0	328.0
Número de empresas	59.0	46.0

nos gráficos representam as densidades obtidas sob normalidade e sob uma suavização não paramétrica. Analisando a tabela 2, observa-se uma grande diferença entre os desvios padrão nos dois grupos, sugerindo desigualdade das variâncias populacionais.

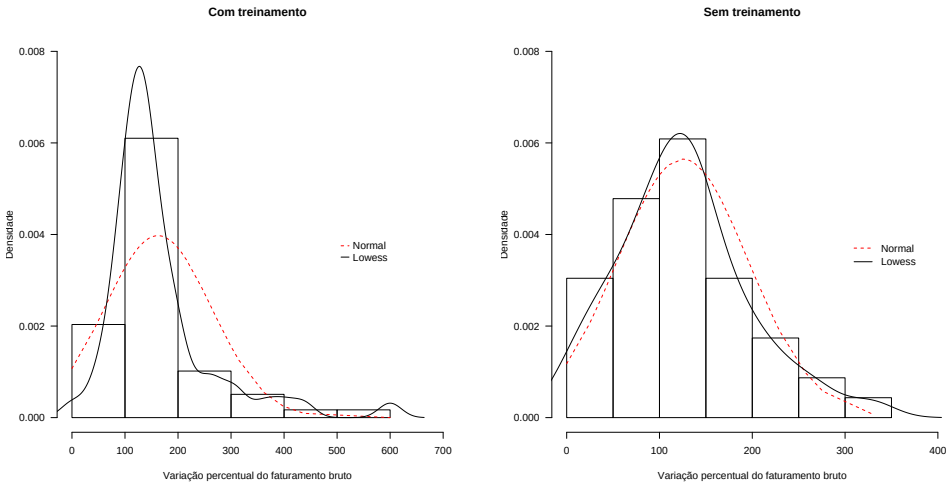

Figura 2: Histograma para a variação percentual de faturamento bruto segundo o treinamento das empresas.

Através da forma modificada do teste de Brown e Forsythe descrita na seção 3, desejamos testar a hipótese nula de igualdade de médias $\mathcal{H}_0 : \mu_1 = \mu_2$ versus $\mathcal{H}_0 : \mu_1 \neq \mu_2$, em que μ_1 (μ_2) representa a média da variação percentual do faturamento bruto das empresas que fazem (não) o curso de treinamento. Nosso objetivo é aplicar o teste de Brown e Forsythe modificado para testar as hipóteses descritas. Para este fim, desenvolvemos uma função em linguagem R (R Development Core Team 2007).

O teste foi realizado para aparos de 30, 20, 10, 5, 1 e 0%. A estatística de teste, F_a, foi calculada e o respectivo valor crítico foi estimado através da construção da distribuição empírica via *bootstrap*, conforme descrito na seção 3. Para aparos de 30, 20 e 10% usamos $B = 100, 250, 599$ e 1000. Para os demais aparos utilizamos $B = 599$, conforme sugerido por Wilcox (1996) e pelo fato de que não foram observadas alterações relacionadas à quantidade de reamostras nos aparos de 30, 20 e 10%. O valor de p^* foi obtido e a decisão tomada (se $p^* \leq \alpha$, rejeitamos a hipótese nula de igualdade de médias). Na tabela 3 exibimos os resultados obtidos e a correspondente decisão em um nível de significância de 5%.

Tabela 3: Resultados obtidos na realização do teste de Brown e Forsythe modificado no nível de significância de 5%.

% de Aparo	B	F_a	Valor p	Decisão
30	100	46.0994	0.200	Não rejeita $\mathcal{H}_0$
	250	46.0994	0.156	Não rejeita $\mathcal{H}_0$
	599	46.0994	0.145	Não rejeita $\mathcal{H}_0$
	1000	46.0994	0.165	Não rejeita $\mathcal{H}_0$
20	100	76.7002	0.100	Não rejeita $\mathcal{H}_0$
	250	76.7002	0.128	Não rejeita $\mathcal{H}_0$
	599	76.7002	0.132	Não rejeita $\mathcal{H}_0$
	1000	76.7002	0.130	Não rejeita $\mathcal{H}_0$
10	100	137.4555	0.070	Não rejeita $\mathcal{H}_0$
	250	137.4555	0.076	Não rejeita $\mathcal{H}_0$
	599	137.4555	0.075	Não rejeita $\mathcal{H}_0$
	1000	137.4555	0.071	Não rejeita $\mathcal{H}_0$
5	599	201.4128	0.038	Rejeita $\mathcal{H}_0$
1	599	264.4095	0.023	Rejeita $\mathcal{H}_0$
0	599	264.4095	0.025	Rejeita $\mathcal{H}_0$

Observa-se que quanto maior a porcentagem de aparo, maior é o valor p, nos levando à não rejeição (aceitação) da hipótese nula. Para aparos de 5 e 1%, rejeitamos a hipótese nula de igualdade de médias. Da mesma forma, rejeitamos a hipótese de igualdade de médias quando não há aparos, ou seja, a porcentagem de aparos, γ, é 0%. Era de se esperar esse comportamento, pois as medianas dos dois grupos são próximas, porém com médias muito distintas, devido provavelmente à presença de valores discrepantes, conforme observado na figura 1. Uma alta porcentagem de aparos teria o efeito de excluir tais pontos discrepantes, tornando a média aparada próxima da mediana e levando portanto à não rejeição da hipótese nula (concordando com a informação contida na figura 1, com relação a igualdade das medianas). Com o objetivo de comparação, foram realizados os testes com as estatísticas F^*, W (Welch) e t.[1] Os resultados estão resumidos na tabela 4.

O valor da estatística de Welch obtido, W, coincidiu numericamente com o apresentado para estatística F^*. Tal fato pode ter ocorrido devido ao uso de amostras de tamanhos grandes. Quando realizamos os testes baseados nas estatísticas F^*, W e t, a hipótese de igualdade de médias foi rejeitada no nível

[1] O teste t referese ao teste t usual para igualdade de médias para amostras independentes, sob suposição de normalidade, com variâncias desconhecidas e desiguais.

Tabela 4: Resultados obtidos via testes F^*, W (Welch) e t no nível de significância de 5%.

Estatística de teste	Valor p	Decisão
$F^* = 4.98$	0.02783	Rejeita $\mathcal{H}_0$
$W = 4.98$	0.02783	Rejeita $\mathcal{H}_0$
$t = 2.23$	0.02794	Rejeita $\mathcal{H}_0$

de significância de 5%, ou seja, conclui-se que a média populacional da variável variação percentual do faturamento bruto para empresas que fazem o curso de treinamento difere da correspondente média das empresas que não fazem o curso de treinamento. Em contrapartida, o teste de Brown e Forsythe modificado, que utiliza uma estatística robusta, rejeita a hipótese nula de igualdade de médias no mesmo nível de significância de 5% somente para aparos iguais ou inferiores a 5%. Já para aparos de 10, 20 e 30%, a decisão é pela não rejeição da hipótese nula.

Na particular situação analisada, observa-se que a maior média amostral ocorreu no grupo de maior variância amostral. De acordo com a literatura, o teste baseado na estatística F^* é superior ao teste baseado em W nesse caso. No entanto, ambos os testes levaram à mesma conclusão, com mesmo valor p. Tal fato pode ser conseqüência de dispormos de apenas dois grupos, ou ainda, conforme já mencionado, aos tamanhos de amostra, relativamente grandes. O teste proposto por Keselman & Wilcox (1999) pode ser considerado uma alternativa robusta, não somente com relação a desvios de normalidade e de homocedasticidade, mas também com relação à presença de observações discrepantes. Nesse sentido, sugerimos sua utilização na presença desses problemas, restrita no entanto a pequenas porcentagens de aparo.

De maneira informal, consideramos aceitável uma porcentagem de aparos de 5%, que não implicaria em perda excessiva de informação e fornecereia robustez com relação à presença de valores discrepantes.

6. Considerações finais

O objetivo principal do presente trabalho é apresentar um levantamento das principais modificações aos testes de Levene e de Brown e Forsythe propostas na literatura. As aplicações na área de bioequivalência foram introduzidas devido ao grande destaque que esse estudo tem recebido atualmente e também a seu extremo interesse prático. Na comparação de variâncias, sugerimos o uso do teste de Levene com centralização na mediana, já que é equivalente ao centrado na média para distribuições simétricas e se mostra mais eficiente para distribuições assimétricas. Destaca-se ainda sua popularidade e fácil obtenção nos pacotes computacionais. Quanto às modificações no teste de Brown e Forsythe, consideramos o teste de Keselman e Wilcox muito bem elaborado e interessante do ponto de vista teórico. Sua aplicabilidade no entanto é mais difícil e está restrita à existência de programas computacionais específicos. O programa utilizado na aplicação realizada na seção 5 pode ser encontrado em http://www.dema.ufc.br/∼juvencio.

Agradecimentos

Agradecemos ao Professor Dr. Bryan F.J. Manly que, em sua visita ao Brasil em 2004 e 2005, nos despertou para o problema, e aos dois árbitros pelas valiosas sugestões na melhoria do nosso trabalho. O terceiro autor gostaria de agradecer ao CNPq pelo suporte financeiro.

[Recibido: febrero de 2008 — Aceptado: septiembre de 2008]

Referências

Bickel, P. J. (2005), 'One-step Haber Estimates in the Linear Model', *Journal of the American Statistical Association* **70**, 428–434.

Brown, M. B. & Forsythe, A. B. (1974*a*), 'Robust Tests for the Equality of Variances', *Journal of the American Statistical Association* **69**, 364–367.

Brown, M. B. & Forsythe, A. B. (1974*b*), 'The Small Sample Behavior of Some Statistics which Test the Equality of Several Means', *Technometrics* **16**, 129–132.

Carrol, R. J. & Ruppert, D. (1982), 'Robust Estimation in Heteroscedastic Linear Models', *Annals of Statistics* **10**, 429–441.

Carrol, R. J. & Schneider, H. (1985), 'A Note on Levene's Test for Equality of Variances', *Statistics and Probability Letters* **3**, 191–194.

Conover, W. J., Johnson, M. E. & Johnson, M. M. (1981), 'A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data', *Technometrics* **23**, 351–361.

Davison, A. C. & Hinkley, D. V. (1997), *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge, United States.

Elian, S. N. & Santos, L. D. (2003), Relatório de Análise Estatística Sobre o Projeto: "Tipos Psicológicos Associados a Variáveis Estratégicas Em Empreendedores de Pequena e Micro Empresa", Technical report, São Paulo, IME-USP.

Francis, R. I. C. C. & Manly, B. F. J. (2001), 'Bootstrap Calibration to Improve the Reliability of Tests to Compare Means and Variances', *Environmetrics* **12**, 713–729.

Hines, W. G. S. & O'Hara Hines, R. J. (2000), 'Increased Power with Modified Forms of the Levene (Med) Test for Heterogeneity of Variance', *Biometrics* **56**, 451–454.

James, G. S. (1954), 'Tests of Linear Hypotheses in Univariate and Multivariate Analysis when the Ratios of the Population Variances are Unknown', *Biometrika* **41**, 19–43.

Keselman, H. J. & Wilcox, R. R. (1999), 'The Improved Brown and Forsythe Test for Mean Equality: Some Things Can't be Fixed', *Communications in Statistics-Simulation* **28**, 687–698.

Levene, H. (1960), Robust Test for Equality of Variances, *in* I. O. et al., ed., 'Contributions to Probability and Statistics: Essays in Honor of Harold Hotteling', Stanford University Press, California, United States, pp. 278–292.

Manly, B. F. J. (1995), 'Randomization Tests to Compare Means with Unequal Variation', *Sankhyã* **57**, 200–222.

Manly, B. F. J. (2004), 'One-sided Tests of Bioequivalence with Nonnormal Distributions and Unequal Variances', *Journal of Agricultural, Biological and Environmental Statistics* **9**, 270–283.

Manly, B. F. J. & Francis, R. I. C. C. (2002), 'Testing for Mean and Variance Differences with Samples from Distributions that May Be Non-Normal with Unequal Variances', *Journal of Statistical Computation and Simulation* **72**, 633–646.

Mehrotra, D. V. (1997), 'Improving the Brown-Forsythe Solution to the Generalized Behrens-Fisher Problem', *Communications in Statistics-Simulation and Computation* **26**, 1139–1145.

O'Neill, M. E. & Mathews, K. (2000), 'A Weighted Least Squares Approach to Levene's Test of Homogeneity of Variance', *Austral. & New Zealand J. Statist.* **42**, 81–100.

Pereira, C. A. B. & Stern, J. M. (2003), 'Evidence and Credibility: Full Bayesian Significance Test for Precise Hypothesis', *Entropy* **1**, 99–110.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Satterthwaite, F. E. (1941), 'Synthesis of Variance', *Psychometrika* **6**, 309–316.

Welch, B. L. (1951), 'On the Comparison of Several Mean Values: An Alternative Approach', *Biometrika* **38**, 330–336.

Wilcox, R. R. (1996), *Statistics for the Social Sciences*, Academic Press, New York, United States.