ALTERNATIVAS PARA O TESTE t COM VARIÂNCIAS HETEROGÊNEAS AVALIADAS POR MEIO DE SIMULAÇÃO

ROBERTA BESSA VELOSO SILVA[1]
DANIEL FURTADO FERREIRA[2]

RESUMO – Conduziu-se este trabalho com o objetivo de avaliar os riscos de se tomar decisões erradas (erro tipo I e erro tipo II), com o aumento da diferença entre as variâncias populacionais, por meio de simulação computacional, utilizando-se o teste t de Student com o número de graus de liberdade sendo aproximado pelas alternativas de Satterthwaite (1946), valor mínimo $\upsilon = min(\ n_1 - 1,\ n_2 - 1)$ e pelo método de bootstrap. Duas populações foram geradas, e a variância da população 1 foi igual a um ($\sigma_1^2 = 1$), e a da população 2 foi especificada em função da razão σ_2^2 / σ_1^2, a qual assume os valores 1, 2, 8 e 16. Usando essas abordagens diferentes para o teste t, avaliaram-se as taxas de erro tipo I e tipo II. Todos os critérios controlaram adequadamente a taxa de erro tipo I; o critério de t com alteração dos graus de liberdade foi mais rigoroso que os demais critérios quando as amostras apresentaram tamanhos diferentes. Essa aproximação foi a que apresentou maiores taxas de erro tipo II para as situações de maiores heterogeneidades de variância. O procedimento de bootstrap foi melhor com relação ao controle da taxa de erro tipo II para situações de tamanhos de amostras diferentes (n_1=5 e n_2=30, n_1=10 e n_2=30) e para razões de variâncias maiores que 1.

TERMOS PARA INDEXAÇÃO: Erros tipo I e tipo II, Monte Carlo e bootstrap.

ALTERNATIVES FOR EVALUATING t TEST WITH HETEROGENEOUS VARIANCES BY MONTE CARLO SIMULATION

ABSTRACT – This work aimed to measure the type I and II error rates with the increases of the difference among populational variances through computational simulation using Student ***t*** test with degrees of freedom proposed by Satterthwaite (1946), or degrees of freedom given by $\upsilon = min(\ n_1 - 1,\ n_2 - 1)$ and an alternative given by bootstrap method. Two populations were generated. The variance of the first population was considered equal to 1, and the variance of the population 2 was specified in function of the ratio σ_2^2 / σ_1^2, which assumes the values of 1, 2, 8 and 16. Using these three different approaches the type I and II error rates were evaluated. All the approaches controlled appropriately the type I error rates; the Student ***t*** with degrees of freedom given by $\upsilon = min(\ n_1 - 1,\ n_2 - 1)$ was more rigorous than the other approaches when the samples had different sizes. This approach presented larger type II error rates than the others to the situations of great variance heterogeneity. The bootstrap procedure better controlled the type II error rates to situations of different sample sizes (n_1=5 and n_2=30, n_1=10 and n_2=30) and of variances ratios larger than 1.

INDEX TERMS: Type I and type II error rates, Monte Carlo, bootstrap.

INTRODUÇÃO

O pesquisador depara-se, muitas vezes, com a necessidade de comparar duas médias populacionais, e cada população na experimentação é conhecida por tratamento. Tais estudos comparativos podem ser feitos por meio de duas formas básicas: (a) comparações pareadas, em que a amostra selecionada na população é avaliada antes e após a aplicação de um tratamento; (b) comparações independentes, em que as duas populações que se deseja comparar são amostradas de forma independente. O segundo caso é o mais freqüente nas mais diversas áreas de pesquisa.

1. Acadêmica do 9º módulo do curso de Administração da UNIVERSIDADE FEDERAL DE LAVRAS/UFLA, Caixa Postal 37 – 37200-000 – Lavras, MG, bolsista CNPq, roberta@ufla.br
2. Professor Adjunto do Departamento de Ciências Exatas da UFLA, bolsista do CNPq, danielff@ufla.Br

Para o caso das comparações independentes, existe uma pressuposição básica para garantir que o teste de t seja exato. Essa pressuposição refere-se à homocedasticia das variâncias populacionais, ou seja, $\sigma_1^2 = \sigma_2^2$ (Snedecor & Cochran, 1980). Independentemente dessa pressuposição, um estimador não-viesado da variância da diferença entre duas médias amostrais independentes $\left(\bar{X}_1 - \bar{X}_2\right)$ é:

$$S^2_{\bar{X}_1 - \bar{X}_2} = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

sendo S_1^2 e S_2^2 os estimadores das variâncias das populações 1 e 2, obtidos em amostras de tamanhos n_1 e n_2, respectivamente.

Se as variâncias são homogêneas, ou seja $\sigma_1^2 = \sigma_2^2 = \sigma^2$, um melhor estimador ($S_p^2$) da variância comum σ^2 é dada pela média ponderada dos estimadores S_1^2 e S_2^2, usando como peso os graus de liberdade, o qual é dado por:

$$S_p^2 = \frac{\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2}{n_1 + n_2 - 2}$$

Com esse estimador substituído na expressão de $S^2_{\bar{X}_1 - \bar{X}_2}$, o teste t é considerado exato e adequado para a hipótese de igualdade das médias das duas populações. O problema ocorre quando $\sigma_1^2 \neq \sigma_2^2$, e nesse caso o teste depende da razão entre as variâncias populacionais (σ_2^2 / σ_1^2), que é desconhecida. Ainda para o caso de heterogeneidade, porém sob normalidade, esse teste é apenas aproximado e é conhecido por problema de Behrens-Fisher (Moreno et al., 1999; Akahira, 2002). Para grandes ou pequenos valores da razão σ_2^2 / σ_1^2, o teste da hipótese H_0: $\mu_1 = \mu_2$ e os intervalos de confiança podem ser seriamente comprometidos (Borges & Ferreira, 1999). Os riscos de se cometer o erro tipo I, ou seja, de rejeitar uma hipótese verdadeira, e o do tipo II, de aceitar uma hipótese falsa, aumentam consideravelmente. Com o aumento desses erros, o pesquisador tem grande chance de tomar decisões erradas.

Nesse caso, a hipótese H_0: $\mu_1 = \mu_2$ pode ser avaliada, usando-se a estatística:

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

No entanto, essa estatística (*t*') não segue a distribuição exata de *t* de Student, sob a hipótese de igualdade das médias populacionais e com variâncias populacionais geterogêneas. Na avaliação dessa estatística, duas aproximações são bastante comuns na literatura. A primeira, derivada por Satterthwaite (1946), refere-se ao cálculo do número de graus de liberdade associados a ***t´,*** de tal forma que a distribuição ***t*** de Student possa ser usada. Nesse caso, o número de graus de liberdade (υ) é estimado por:

$$\upsilon = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2 \Bigg/ \left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1} \right]$$

(υ pode ser arredondado para o inteiro mais próximo, se for utilizar valores tabelados da distribuição ***t*** de Student).

A segunda aproximação é a de Cochran & Cox (1957), estatística que usa valor médio ponderado para a estatística ***t´,*** cujo estimador é:

$$t^* = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$$

em que ***t**** refere-se ao valor determinante da região de rejeição da hipótese de igualdade das médias populacionais; $w_1 = S_1^2 / n_1$ e $w_2 = S_2^2 / n_2$; t_1 e t_2 são os valores críticos (tabelados) da distribuição ***t*** (unilateral), com n_1 - 1 e n_2 - 1 graus de liberdade, respectivamente, Para o valor nominal de significância estipulado previamente.

Confrontando essas alternativas, Borges & Ferreira (1999) demonstraram por simulação que os dois tipos de aproximações não diferiram quanto ao poder do teste e nem quanto às taxas de erro tipo I.

Uma terceira alternativa é considerar os graus de liberdade (Triola, 1999) como sendo estimados por:

$$\upsilon = min\left(n_1 - 1,\ n_2 - 1 \right)$$

Essa alternativa é pouco usada e nenhum estudo é conhecido confrontando-a com as demais. Esse critério parece ser mais conservador por fornecer menores valores para o número de graus de liberdade do que o

critério de Satterthwaite (1946), mas nenhuma comprovação científica foi encontrada na literatura consultada.

Uma quarta abordagem pode ser realizada, para avaliar os riscos de se tomar decisões erradas (erros tipo I e II), mediante o uso de métodos de computação intensiva, Bootstrap (Manly, 1998). Nesse tipo de procedimento, as amostras aleatórias obtidas de cada população são reamostradas, com reposição, milhares de vezes. As duas amostras originais são combinadas, utilizando a hipótese $H_0 : \mu_1 - \mu_2 = 0$, em uma única amostra de tamanho n_1+n_2 e as reamostragens são feitas com reposição, formando duas novas amostras de bootstrap de tamanhos n_1 e n_2 das populações 1 e 2, respectivamente. A estatística ***t'*** foi calculada para cada uma das reamostragens realizadas. Essa estatística é representada por ***tb'***, ***t'*** de bootstrap. Com base nos milhares de valores obtidos e pelo confronto do valor de ***t'*** na amostra original com a distribuição de ***tb'*** empírica gerada pelas simulações, é possível testar a hipótese de interesse. Se o valor de ***t'*** é um valor atípico dentre os valores de ***tb',*** a hipótese deve ser rejeitada.

O teste de t original sem nenhuma correção nos graus de liberdade e a abordagem de bootstrap poderiam ser aplicados sem nenhum problema se as variâncias das populações forem iguais. Entretanto, nas várias aplicações esse fato pode não ocorrer e, conseqüentemente, os resultados podem ser comprometidos quando $\sigma_1^2 \neq \sigma_2^2$. Desse modo, os riscos de se cometer o erro tipo I de rejeitar uma hipótese verdadeira, e do tipo II de aceitar uma hipótese falsa, aumentam consideravelmente se for utilizado o teste t sem nenhum ajuste dos graus de liberdade, podendo levar o pesquisador a tomar decisões erradas.

O presente trabalho foi realizado tendo por objetivo avaliar os riscos de se tomar decisões erradas por meio dos erros tipo I e tipo II, considerando diferenças entre as duas variâncias populacionais, por simulação computacional, comparando-se três alternativas: a) aproximação dos graus de liberdade proposto por Satterthwaite (1946); b) os graus de liberdade dados por $\upsilon = min(\ n_1 - 1,\ n_2 - 1)$; e c) o método de Bootstrap.

METODOLOGIA

Para avaliar as probabilidades de se cometer o erro tipo I e o erro tipo II, duas populações foram simuladas computacionalmente. Para a geração dessas populações, fixou-se a variância da população 1 em $\sigma_1^2 = 1$, sem perda de generalidade. A variância da população 2 foi obtida em função da razão $\delta = \sigma_2^2 / \sigma_1^2$, a qual considerou os valores 1, 2, 8 e 16. Para cada uma dessas situações, as populações foram geradas com médias iguais e com médias diferentes. Para o caso de médias populacionais diferentes, foram consideradas as diferenças padronizadas entre as médias populacionais de k=1, 2, 3 e 4 erros-padrão.

Para o caso da avaliação do erro tipo I, as médias populacionais foram iguais e dadas por $\mu_1 = \mu_2 = 100$ (sem perda de generalidade). Para o caso de médias populacionais diferentes, os valores paramétricos foram fixados em função da diferença de erros padrão da diferença de médias por:

$$\mu_2 = \mu_1 + k\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};\ com\ \mu_1 = 100$$

Para cada uma das diferentes situações, foram retiradas 2.000 amostras independentes de cada população, sendo aplicado o teste ***t*** para a hipótese da igualdade entre as médias populacionais, H_0: $\mu_1=\mu_2$. Diferentes tamanhos de amostra foram considerados em ambas as populações, quais sejam, $n_1=n_2=5$, $n_1=5$ $n_2=30$, $n_1=10$ $n_2=30$, $n_1=n_2=30$. O teste ***t*** foi realizado mediante aproximações de Satterthwaite (1946) (tc) e $\upsilon = min(\ n_1 - 1,\ n_2 - 1)$ (tgl). O número de erros cometidos nas 2.000 repetições foi computado para os níveis de significância de 5% e de 1%.

Para a geração das amostras, foi utilizado o método de Monte Carlo, considerando-se o modelo estatístico para cada população:

$$y_{ij} = \mu_i + e_{ij}$$

em que y_{ij} é o valor do j-ésimo elemento amostral observado (j=1, 2, ..., n_i) para a i-ésima população (i=1, 2), μ_i é a média paramétrica, e e_{ij} é o valor do erro amostral não observável associado y_{ij}, com distribuição NID (0, σ_i^2).

Esse modelo foi adotado para cada população, com os parâmetros μ_i e σ_i^2, estipulados conforme a situação simulada. Para a obtenção do valor de $e_{ij,}$ foi usado um algoritmo em Pascal para inversão da função da distribuição normal, conforme procedimento relatado por Dachs (1988), baseado no teorema da probabilidade integral. Baseado nesse teorema, se U ~ U (0,1) (uniforme 0-1) e sendo F a função de distribuição normal, então:

$$z_{ij} = F^{-1}(U)$$

tem função de densidade de probabilidade N(0, 1). Esse valor foi, então, multiplicado por σ_i para obtenção dos valores de e_{ij} (Morgan, 1995).

Em cada uma das 2.000 simulações, também foi aplicado o método de bootstrap. As amostras de cada população foram agrupadas e a amostra composta, assim formada, foi reamostrada com reposição, gerando-se novas amostras aleatórias de tamanho n_1 e n_2 das populações 1 e 2, respectivamente. Em cada reamostragem, calculou-se a estatística:

$$tb' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Após terem sido realizadas 2.000 reamostragens de cada uma das 2.000 amostras de Monte Carlo, para cada configuração de tamanhos amostrais (n_1 e n_2), razão de variâncias (σ_2^2/σ_1^2) e valores das médias (μ_1 e μ_2), sendo obtidos 2.000 estimativas de tb', essas foram ordenadas e tomados os seus quantis $tb'_{(\alpha/2)}$ e $tb'_{(1-\alpha/2)}$. Esses quantis referem-se àqueles valores que deixam uma proporção de $\alpha/2$ e $1-\alpha/2$ estimativas de tb' menores do que eles. Se o valor de ***t'*** das amostras originais for menor do que $tb'_{(\alpha/2)}$ ou maior do que $tb'_{(1-\alpha/2)}$, o teste da hipótese H_0: $\mu_1=\mu_2$ é considerado como significativo para o valor nominal α da significância, fixado em 5% ou em 1%.

Para cada configuração dos parâmetros de simulação, foram realizadas 4.000.000 reamostragens e cálculos das estatísticas de bootstrap e 2.000 estatísticas do teste t com ajuste dos graus de liberdade pelas duas alternativas apresentadas. Um total de 240.120.000 análises foi realizado.

RESULTADOS E DISCUSSÃO

Na Figura 1 verifica-se que as taxas de erro tipo I foram devidamente controladas para o procedimento de t de bootstrap (tb), tanto para o nível nominal de 5% quanto de 1%, para amostras de (a) diferentes tamanhos ou (b) de mesmos tamanhos e grandes. Na situação (b), os três procedimentos não apresentaram nenhuma diferença nas taxas de erro tipo I, para os níveis nominais de 5% ou de 1%. A opção do t com graus de liberdade dados pelo $\upsilon = min(\, n_1 - 1,\ n_2 - 1)$, tgl, apresentou-se muito rigorosa, para 5% ou para 1%, para o caso de amostras com tamanhos diferentes, Figura 1 (a). Isso significa que o controle da taxa de erro tipo I foi realizada a contento, embora com taxas inferiores aos níveis nominais adotados. Para variâncias iguais, todas as opções controlaram as taxas de erro tipo I adequadamente. Nenhuma tendência de alteração nas taxas de erro tipo I foi percebida com o aumento da razão de variâncias de 2 para 16, a não ser o aumento do rigor do teste tgl. As situações em que as amostras são de ta-

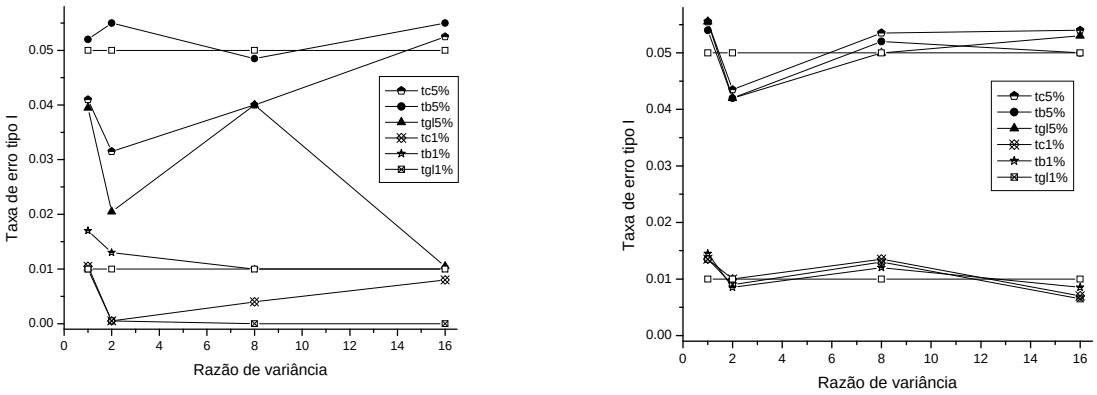


FIGURA 1 – Representação gráfica da taxa de erro tipo I para o teste t, em função de diferentes razões entre variâncias populacionais, para as alternativas de Satterthwaite (tc), valor mínimo (tgl) e bootstrap (tb), nos valores nominais de significância de 5% e 1% e para os tamanhos de amostras (a): n_1=5 e n_2=30 e (b) n_1=30 e n_2=30.

manhos diferentes é que, em geral, são problemáticas para o controle da taxa de erro tipo I. O teste tgl apresentou um aumento do rigor à medida que as variâncias tornavam-se mais heterogêneas, para essa situação. Esse resultado está de acordo com aqueles apontados por Zar (1999), para essa situação de tamanhos de amostras diferentes em populações com heterogeneidade de variâncias.

Na Figura 2 estão apresentados os resultados das taxas de erro tipo II para n_1=5 e n_2=30 e para (a) k=1 e (b) k=4. Verifica-se que tanto para o caso de diferenças de médias populacionais de 1 erro-padrão (k=1) ou de 4 erros padrão (k=4), o procedimento de t com a modificação nos graus de liberdade (tgl) foi o que apresentou maiores taxas de erro tipo II, considerando grandes razões entre as variâncias (δ). Para k=1, o tgl apresentou resultados de taxas de erro tipo II similar aos demais critérios, considerando δ=1 e níveis nominais de 5% ou de 1%. Nessa situação, de k=1, considerando agora $\delta \geq 2$, esse procedimento destacou-se dos demais no sentido de cada vez apresentar maiores taxas de erro tipo II relativas. O tb foi, nessa situação, de k=1, o melhor procedimento, apresentando as menores taxas de erro tipo II. Para o caso de δ=1, tb apresentou taxas de erro tipo II praticamente idênticas às demais, tanto para 5% quanto para 1%. Para a situação de k=4 (Figura 2b), o comportamento de todos os critérios foi bastante similar no caso de k=1, no entanto, as diferenças destacadas anteriormente foram evidenciadas. Existe uma tendência, em ambas as situações, de os procedimentos apresentarem um aumento das taxas de erro tipo II com o aumento de δ. O critério de bootstrap apresentou uma exceção desse comportamento, ou seja, à medida que
δ aumenta, o erro tipo II diminui. Esse último resultado ocorre em situações em que para a população de maior variância é sorteada a amostra de maior tamanho (Zar, 1999).

É interessante observar que no caso de k=4 e $\delta \leq 2$, os procedimentos tgl e tc apresentam praticamente as mesmas taxas de erro tipo II com valores inferiores aos do tb. A partir de δ>2 há uma inversão do comportamento dos procedimentos quanto às taxas de erro tipo II. Existiu uma tendência de queda para as taxas do tb, ao passo que as dos tc e tgl tenderam a aumentar. Esse é um fato bastante interessante que destaca que o tb deve ser um procedimento recomendado para situações de grande heterogeneidade e variâncias, em amostras de diferentes tamanhos e com o maior tamanho associado à população de maior variância.

Os resultados mencionados referentes às taxas de erro tipo II (Figura 2), no entanto, não se confirmam para as situações de tamanhos amostrais iguais. Assim, observando-se a Figura 3 (a e b), verifica-se que o procedimento de bootstrap foi quase sempre inferior ao dos outros critérios, tanto para pequenas razões de variâncias e pequenos tamanhos amostrais, quanto para grandes valores de ambos os critérios. Para a situação de amostras de tamanhos iguais a 30 e valor de significância de 5%, os três procedimentos foram idênticos, independentemente das razões de variâncias. Para 1% e amostras de tamanho 30, a magnitude do pior resultado do teste de bootstrap, comparado aos demais, foi menor do que a encontrada para pequenas amostras (Figura 3a). Os resultados do critério tc e tgl são similares quando as amostras possuem o mesmo tamanho.

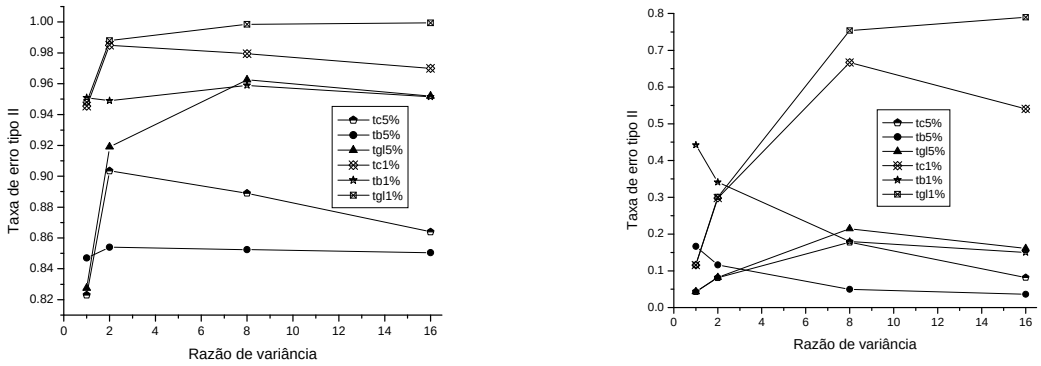
Na Figura 4 pode-se observar que, tanto para variâncias iguais (δ=1), quanto diferentes (δ=16), considerando pequenos tamanhos amostrais (n_1=n_2=5) e k=4, o tc e o tgl apresentaram taxas de erro tipo II iguais e menores do que o tb para os níveis de 5% e 1%.

Para a situação de variâncias iguais (Figura 4a), tanto para 1% quanto para 5%, os procedimentos tgl e tc apresentaram similares taxas de erro tipo II e sempre inferiores ao critério de bootstrap. As taxas de erro tipo II do critério de bootstrap tenderam a se aproximar das taxas dos outros dois critérios, tornando-se idênticas a elas quando as amostras eram maiores.

Para a situação de variâncias diferentes (Figura 4b), houve uma maior distinção dos três critérios para pequenas amostras e uma menor distinção para grandes amostras. Nesse caso, o critério de bootstrap apresentou, em geral, menores taxas de erro tipo II, do que os demais procedimentos estudados. Essa diferença foi mais acentuada para o valor nominal de 1% se comparada com as apresentadas para 5%. Dessa forma, pode se inferir que o procedimento de bootstrap foi mais adequado para situações de tamanhos de amostras diferentes (n_1=5 e n_2=30, n_1=10 e n_2=30) e para razões de variâncias maiores que 1.

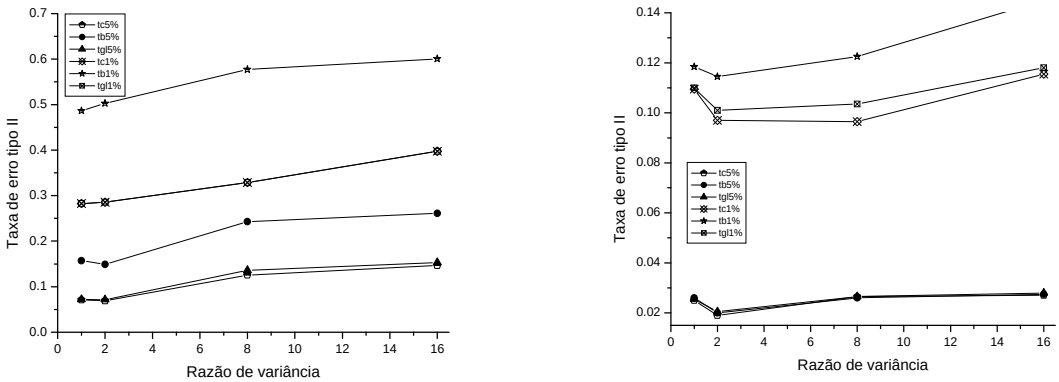
Os três testes avaliados mostraram diferenças no controle das taxas de erro tipo I e tipo II. A alternativa tgl mostrou-se conservadora nas situações adversas, ou seja, para grandes heterogeneidades de variâncias e amostras de diferentes tamanhos apresentando taxa de erro tipo I menor do que o valor nomi-

nal e maiores taxas de erro tipo II. O teste tb apresentou melhores resultados nas situações de amostras de diferentes tamanhos e maiores heterogeneidades de variâncias. Esse resultado mostra que o maior esforço computacional requerido (Manly, 1998) deve ser compensado pela melhor performance do teste nesse caso.



a) b)

FIGURA 2 – Representação gráfica da taxa de erro tipo II para o teste t, em função de diferentes razões entre variâncias populacionais, para as alternativas de Satterthwaite (tc), valor mínimo (tgl) e bootstrap (tb), nos valores nominais de significância de 5% e 1% e para os tamanhos de amostras n_1=5 e n_2=30 e para (a) k=1 e (b) k=4.



a) b)

FIGURA 3 – Representação gráfica da taxa de erro tipo II para o teste t, em função de diferentes razões entre variâncias populacionais para as alternativas de Satterthwaite (tc), valor mínimo (tgl) e bootstrap (tb), nos valores nominais de significância de 5% e 1% e para (a) k=4 e os tamanhos de amostras n_1=n_2=5 e (b) n_1=n_2=30.

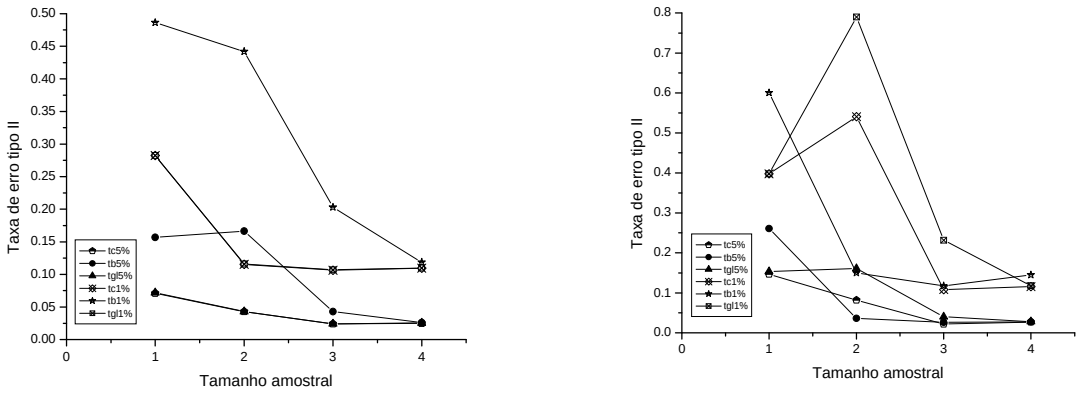


FIGURA 4 – Representação gráfica da taxa de erro tipo II para o teste t em função dos diferentes tamanhos amostrais 1: (n_1=n_2=5); 2: (n_1=5 n_2=30); 3: (n_1=10 n_2=30); 4: (n_1=n_2=30) para (a) δ=1 e (b) δ=16 com k=4 e para as alternativas de Satterthwaite (tc), valor mínimo (tgl) e bootstrap (tb) nos valores nominais de significância de 5% e 1%.

CONCLUSÕES

Todos os critérios controlam adequadamente a taxa de erro tipo I; o critério de t com alteração dos graus de liberdade (tgl) foi mais rigoroso que os demais critérios quando as amostras têm tamanhos diferentes.

O t com graus de liberdade dados por $\upsilon = min(\ n_1 - 1,\ n_2 - 1)$ (tgl) foi o procedimento que apresentou maiores taxas de erro tipo II para as situações de maiores heterogeneidades de variância.

O procedimento de bootstrap foi melhor com relação ao controle da taxa de erro tipo II para situações de tamanhos de amostras diferentes (n_1=5 e n_2=30, n_1=10 e n_2=30) e para razões de variâncias maiores que 1.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAHIRA, M. Confidence interval for the difference of means: application to the Behrens-Fisher type problem. **Statistical Papers**, v. 43, n. 2, p. 273-284, 2002.

BORGES, L. C; FERREIRA, D. F. Comparação de duas aproximações do teste t com variâncias heterogêneas através de simulação. **Ciência e Agrotecnologia**, Lavras, v. 23, n. 2, p. 390-403, abr./jun. 1999.

COCHRAN, W. G.; COX, G. M. **Experimental designs.** 2. ed. Singapure: John Wiley & Sons, 1957. 611 p.

DACHS, J. N. W. **Estatística computacional.** Uma introdução ao Turbo Pascal. Rio de Janeiro: Livros Técnicos e Científicos, 1988. 236 p.

MANLY, B. F. J. **Randomization, bootstrap and Monte Carlo methods in biology**. 2. ed. London: Chapman and Hall, 1998. 424 p.

MORENO, E.; BERTOLINO, F.; RACUGNO, W. Default bayesian analysis of the behrens-fisher problem. **Journal of Statistical Planning and Inference,** Amsterdam, n. 81, v. 2, p. 323-333, nov., 1999.

MORGAN, B. J. T. **Elements of simulation**. London: Chapman & Hall, 1995. 351 p.

SATTHERTHWAITE, F. E. An approximate distribution of estimates of variance components. **Biometric Bulletin**, London, v. 2, p. 110-114, 1946.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical methods.** 7. ed. Ames: The Iowa State University, 1980. 507 p.

TRIOLA, M. F. **Introdução a estatística:** livros técnicos e científicos. 7. ed. Rio de Janeiro: [s.n.], 1999. 410 p.

ZAR, J. H. **Biostatistical analysis.** 4. ed. Upper Saddle River: Prentice-Hall, 1999. 931 p.