# VIABILIDADE DO USO DA FUNÇÃO DISCRIMINANTE DE FISHER: COMPARAÇÃO COM A MANAVA

Katia Alves CAMPOS[1]
Augusto Ramalho MORAIS[2]
Crysttian Arantes PAIXÃO[3]

- RESUMO: Os métodos de análise multivariada permitem o estudo simultâneo quando são obtidas várias variáveis respostas por parcela. Uma opção para o tratamento de dados multivariados é a transformação, por meio da função discriminante linear de Fisher (FDF). Após redução do espaço *p*-dimensional ao unidimensional, aplica-se a análise de variância univariada (ANAVA). Este trabalho teve por objetivos avaliar a eficiência da transformação dos dados multivariados por meio da FDF e comparar a capacidade de detecção de diferenças entre tratamentos pela ANAVA desses dados com os resultados obtidos por meio da análise de variância multivariada (MANAVA). Foram realizadas simulações com o intuito de avaliar as taxas de aceitação das hipóteses de nulidade para tratamentos, em quatro níveis de correlações, igualdade de médias e de variâncias para tratamentos e desigualdade entre médias e variâncias. Aos valores desses cenários foram aplicadas a ANAVA, a MANAVA e a ANAVA da função discriminante linear de Fisher. Os resultados das simulações indicam que a FDF é uma alternativa adequada para avaliação de dados.
- PALAVRAS-CHAVE: Simulação; correlação; transformação de dados; redução de espaço; análise de variância.

## 1 Introdução

Os métodos multivariados surgem como uma necessidade da análise de dados em pesquisas que envolvem um conjunto de variáveis, como no caso de experimentos com mudas de café, pois permitem a avaliação dessas variáveis simultaneamente. Entre as metodologias multivariadas, podem ser citadas a análise de agrupamento, a técnica de componentes principais, a análise fatorial, a análise de correlação canônica, a análise discriminante e a análise de variância multivariada (BARROSO; ARTES, 2003; CHATFIELD; COLLINS, 1980; FERREIRA, 2018; HAIR *et al*., 2009; MANLY; ALBERTO, 2019; MINGOTI, 2007).

As técnicas multivariadas são aplicadas em diversas áreas do conhecimento, tais como educação, genética, zootecnia, psicologia, odontologia e economia. Porém, na área

[1] Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais - IFSULDEMINAS, CGE, Caixa Postal 49, CEP: 37750-000, Machado, MG, Brasil. E-mail: *katia.campos@ifsuldeminas.edu.br*
[2] Universidade Federal de Lavras - UFLA, Departamento de Estatística, Caixa Postal 3037, CEP: 37200-000, Lavras, MG, Brasil, E-mail: *armorais@ufla.br*
[3] Universidade Federal de Santa Catarina - UFSC, Departamento de Ciências Naturais e Sociais, Caixa Postal 101, CEP: 89520-000, Curitibanos, SC, Brasil, E-mail: *crysttian.arantes.paixao@ufsc.br*

agrícola são poucos os estudos que as utilizam. Os trabalhos com essa temática são encontrados em maior número sobretudo em zootecnia, ou ligados a testes para novas cultivares, ou apenas como técnica coadjuvante a análise univariada (FONSECA e SILVA, 1999; FONSECA *et al*., 2002; LEDO *et al*., 2003; TORRES FILHO *et al*., 2005; CARNEIRO *et al*., 2006; BEZERRA NETO *et al*., 2007).

A proposta inicial de Fisher (1936) para a análise de discriminante linear de Fisher (FDF) foi estabelecer um critério para separação de três populações de plantas por meio de medidas de suas folhas. A ideia era transformar as observações multivariadas, por meio de combinações lineares dessas variáveis, em observações univariadas de tal forma que as variáveis transformadas se apresentassem o mais separadas possível. Assim, a análise discriminante é uma técnica de análise multivariada criada para diferenciar ou discriminar populações e classificar ou alocar indivíduos em populações pré-definidas. Para a discriminação, estabelecem-se funções das variáveis observadas que sejam responsáveis ou possam explicar as diferenças entre populações. Para a alocação ou classificação, determinam-se as funções que além de separar as populações, estas sejam capazes de alocar ou classificar novos indivíduos em uma das populações (FERREIRA, 2018; HAIR *et al*., 2009; MANLY e ALBERTO, 2019; MINGOTI, 2007).

Uma possível aplicação da FDF, que foi abordada por Pimentel-Gomes (2009), é a sua utilização para a transformação de dados multivariados em uma nova variável, por meio da variável canônica principal, o que possibilita nova opção de análise de variância dos dados multivariados, além da análise de variância multivariada (MANAVA). Desse modo, por meio da aplicação da FDF às observações multivariadas, pode-se reduzir o espaço *p* dimensional a um espaço unidimensonal, a partir do qual torna-se possível realizar a análise de variância univariada (ANAVA).

Nesta abordagem em que se faz a transformação dos dados multivariados em univariados, calcula-se apenas a primeira função discriminante linear, associada ao maior autovalor e seu autovetor de $\mathbf{R}^{-1}\mathbf{H}$, em que **R** é a matriz das somas de quadrados e de produtos devidos aos efeitos dos fatores não controlados e **H** é a matriz formada pelas somas de quadrados e de produtos devidos aos efeitos dos tratamentos. Uma justificativa para a adoção apenas da primeira função discriminante linear se deve à porcentagem de informação que pode ser calculada por (1) e normalmente grande parte da informação se esgota com esta primeira função (HAIR *et al*., 2009; MANLY e ALBERTO, 2019; PIMENTEL GOMES, 2009).

$$PI = \frac{\lambda_l}{\sum_{i=1}^{s} \lambda_s} \qquad (1)$$

A ideia inicial apresentada por Fisher era a de encontrar uma combinação linear das variáveis originais $y_1, y_2, \ldots, yp$, sendo cada $y_l$ ($l = 1, 2, \ldots, p$) um vetor de dados da l-ésima variável, como: $Z = a_1y_1 + a_2y_2 + \ldots + a_1y_1$, em que $a_l$ são os coeficientes do autovetor associado ao maior autovalor $\lambda_1$ de $\mathbf{R}^{-1}\mathbf{H}$.

Sobre os coeficientes de ponderação $a_l$, Simeão e Padovani (2008) afirmam que podem ser interpretados como coeficientes de um modelo de regressão múltipla ou de análise fatorial e como tal, servem para identificar as variáveis que mais contribuem para distinguir os grupos dentro de uma mesma função.

A função $Z$ é uma combinação linear das variáveis originais construída de tal maneira que o teste F de tratamentos dessa análise de variância tenha o valor máximo. Após a transformação dos dados, segue-se a análise de variância normalmente, mas o valor da estatística do teste F, pode ser obtido diretamente dos graus de liberdade e do valor de $\lambda_1$, maior autovalor, por $F_{max}$ (2), em que a fração que multiplica o maior autovalor $\lambda_1$ é a razão entre os números de graus de liberdade associados aos resíduos e aos efeitos dos tratamentos.

$$F_{max} = \frac{(I\text{-}1)\,(J\text{-}1)}{I\text{-}1}\,\lambda_1 \qquad (2)$$

Simeão e Padovani (2008) revisaram os aspectos teóricos da análise de discriminate linear, com o levantamento histórico e posterior evolução; afirmaram que mesmo não existindo a homogeneidade entre as matrizes de covariâncias, a função discriminante linear é aplicada para se efetuar a discriminação.

A técnica de análise de discriminante linear de Fisher (FDF), tem sido utilizada como técnica adicional à análise de variância univariada, em experimentos com mudas de cafeeiro; isto é a variável transformada foi considerada como outra variável respostas. Santana *et al*. (2011) na avaliação do uso de adubação foliar no período de viveiro e de Silva *et al*. (2012) na comparação de composições de substratos e de manejos em três cultivares de *Coffea arabica* e Campos *et al*., (2016) utilizaram quatro critérios de seleção para testar os agrupamentos das características das mudas de café e compararam a interpretação dos resultados com os obtidos pela análise de variância univariada e também multivariada e concluíram que ambas as técnicas, multivariada e univariada das características transformadas, foram capazes de detectar as mesmas diferenças dos tratamentos, sendo a FDF mais simples.

Existem trabalhos que, além de estudar alguns fatores, fazem testes para decidir qual característica da muda é mais informativa, como ocorre em métodos de seleção de variáveis, por meio de técnicas diversas ou mesmo por facilidade e importância dessas características. Campos *et al*. (2016) estudaram as possíveis combinações de sete características usuais em experimentos com mudas de cafeeiro, altura, diâmetro, comprimento de raiz, fitomassa seca da parte aérea e da parte radicular, área foliar e número de folhas e concluíram que a análise de todas as sete características é a mais informativa e obtiveram resultados semelhantes na comparação da análise de variância dos dados transformados à análise de variância multivariada.

Facilitar a análise estatística de experimentos com várias variáveis resposta é o objeto de diversas pesquisas e o que se propõe aqui é o uso dos dados transformados por meio da função discriminante linear de Fisher, capaz de, por meio de uma combinação linear que representa a soma ponderada das variáveis independentes, transformar o espaço multidimensional em unidimensional, sem perda de informações (HAIR *et al*., 2009).

Diante da carência de técnicas experimentais específicas para avaliação de dados provenientes de experimentos nos quais são obtidas várias variáveis respostas, esse trabalho tem como objetivo validar a técnica da função discriminante linear de Fisher, como transformação que aplicada às variáveis originais para reduzir o espaço $p$-dimensional a um

espaço unidimensional. Para isso, foram realizadas por meio da simulação de dados de sete variáveis respostas com base em resultados obtidos de experimento com mudas de cafeeiro, conduzido no delineamento inteiramente casualizado e da comparação dos resultados obtidos por meio da análise de variância multivariada com aqueles obtidos com a análise dos dados transformados.

## 2 Material e métodos

Para a simulação dos dados e análises estatística foi considerado um experimento realizado no delineamento inteiramente casualizado, com cinco repetições e quatro tratamentos, cujos resultados foram adaptados de Campos *et al.*, (2016).

### 2.1 Material

Os dados analisados foram gerados a partir de seis situações experimentais ou cenários que retratam possíveis experimentos, com sete variáveis respostas ($V_1$, $V_2$, $V_3$, $V_4$, $V_5$, $V_6$ e $V_7$). Os valores iniciais das médias e das variâncias populacionais (Tabela 1) que serviram de base para a realização das simulações foram obtidos a partir de um experimento com mudas de cafeeiro (*Coffea arabica*) L., cultivar Catuaí Vermelho IAC- 44, cujo objetivo do estudo original foi testar três fontes de adubação orgânica: (A) esterco curtido de aves de postura (70 $l/m^3$), (B) esterco curtido de gado bovino leiteiro (300 $l/m^3$), (C) húmus de minhocas (200 $l/m^3$) e, (D) um tratamento testemunha (controle) que recebeu apenas a adubação química comum a todos os outros tratamentos.

Tabela 1 - Valores populacionais de média e de variância que serviram de base para a realização das simulações das sete variáveis respostas

| Parâmetro | Variável resposta | | | | | | |
|---|---|---|---|---|---|---|---|
| | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ | $V_6$ | $V_7$ |
| Média | 21,03 | 2,56 | 21,22 | 0,67 | 0,18 | 20,24 | 7,41 |
| Variância | 22,35 | 0,10 | 4,44 | 0,10 | 0,01 | 60,08 | 1,17 |

Os seis cenários propostos para a simulação dos dados experimentais são combinações entre as três hipóteses sobre as médias e as duas hipóteses sobre as variâncias dos erros, descritas na sequência.

#### 2.1.1 Hipóteses sobre a média

Foram realizadas três simulações sobre a média:

2.1.1.1 Simulação sob $H_0$ verdadeira: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$,

2.1.1.2 Simulação sob $H_0$ falsa: $H_1 : \begin{cases} \mu_1 = \mu \\ \mu_2 = \mu_3 = \mu_4 = \mu + 2\sigma \end{cases}$, diferença entre as médias obtida pela adição de dois desvios padrão.

2.1.1.3 Simulação sob $H_0$ falsa: $H_1: \begin{cases} \mu_1 = \mu \\ \mu_2 = \mu_3 = \mu_4 = \mu + 8\sigma \end{cases}$, diferença entre as médias obtida pela adição de oito desvios padrão.

Em que: $\mu_i$ é a média populacional de cada variável, no i-ésimo tratamento, com $i$ = 1, 2, 3 e 4, para cada variável resposta simulada, $\mu$ é a média populacional de cada variável resposta simulada, e $\sigma$ é a raiz quadrada da variância populacional de cada variável resposta simulada. Os valores de $\mu$ e $\sigma^2$ foram apresentados na Tabela 1.

### 2.1.2 Hipóteses sobre as variâncias

Para as variâncias, foram simulados dois cenários:

2.1.2.1 Simulação sob $H_0$ verdadeira: $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$, situação favorável

2.1.2.2 Simulação sob $H_0$ falsa**:** $H_1: \begin{cases} \sigma_1^2 = \sigma^2 \\ \sigma_2^2 = 4\,\sigma^2 \\ \sigma_3^2 = 16\,\sigma^2 \\ \sigma_4^2 = 64\,\sigma^2 \end{cases}$, situação desfavorável

em que: $\sigma_i^2$ é a variância populacional de cada variável, no i-ésimo tratamento, com i = 1, 2, 3 e 4, para cada variável resposta simulada, $\sigma^2$ é a variância populacional de cada variável resposta simulada, apresentada na Tabela 1.

Para facilitar a apresentação dos resultados das análises, cada cenário utilizado para a simulação foi denominado por um binômio, cujo o primeiro termo faz referência a hipótese sobre as médias dos tratamentos e o segundo termo sobre as variâncias dos erros, conforme apresentado na Tabela 2.

Tabela 2 - Descrição das combinações de hipóteses adotadas nos cenários propostos e respectivas denominações

| Hipótese sobre as médias ($\mu$) | Diferença entre as médias | Hipótese sobre as variâncias ($\sigma^2$) | Razão entre as variâncias | Denominação dos cenários simulados |
|---|---|---|---|---|
| Verdadeira | 0 $\sigma$ | verdadeira | 0 $\sigma^2$ | $H_0 - H_0$ |
| Verdadeira | 0 $\sigma$ | Falsa | $4^i$ $\sigma^2$ | $H_0 - H_1$ |
| Falsa | 2 $\sigma$ | verdadeira | 0 $\sigma^2$ | $H_1$ 2$s$ $- H_0$ |
| Falsa | 8 $\sigma$ | verdadeira | 0 $\sigma^2$ | $H_1$ 8$s$ $- H_0$ |
| Falsa | 2 $\sigma$ | Falsa | $4^i$ $\sigma^2$ | $H_1$ 2$s$ $- H_1$ |
| Falsa | 8 $\sigma$ | Falsa | $4^i$ $\sigma^2$ | $H_1$ 8$s$ $- H_1$ |

$i$ = 0 , 1 , 2, 3

### 2.1.3 Níveis de correlação

Para cada um dos cenários propostos, foram estabelecidos quatro níveis de correlação entre os erros das variáveis respostas para a simulação dos dados: correlação baixa ($\rho$ = 0, 2), correlação média ($\rho$ = 0, 5), correlação alta ($\rho$ = 0, 9) e sem considerar correlação ($\rho$ = 0, 0). Portanto, cada um dos seis cenários propostos para a simulação foram gerados 1.000

experimentos, em cada uma das 121 possibilidades de combinação do número de variáveis respostas correlacionadas (Tabela 3), em cada um dos níveis de correlação estabelecidos.

Tabela 3 - Número de variáveis de cada subconjunto, número de subconjuntos possíveis e constituição dos subconjuntos formados com agrupamentos variando de 2 a 7 variáveis que apresentaram correlações nas simulações para cada um dos quatro níveis adotados

| Número de variáveis Correlacionadas | Número de subconjuntos | Constituição dos subconjuntos de variáveis correlacionadas |
|---|---|---|
| 0 | 1 | - |
| 2 | 21 | $V_1$ e $V_2$; $V_1$ e $V_3$; ...; $V_6$ e $V_7$ |
| 3 | 35 | $V_1$, $V_2$ e $V_3$; $V_1$, $V_2$ e $V_4$; ...; $V_5$, $V_6$ e $V_7$ |
| 4 | 35 | $V_1$, $V_2$, $V_3$ e $V_4$; ...; $V_4$, $V_5$, $V_6$ e $V_7$ |
| 5 | 21 | $V_1$, $V_2$, $V_3$, $V_4$ e $V_5$; ...; $V_3$, $V_4$, $V_5$, $V_6$ e $V_7$ |
| 6 | 7 | $V_1$, $V_2$, $V_3$, $V_4$, $V_5$ e $V_6$; ...; $V_2$, $V_3$, $V_4$, $V_5$, $V_6$ e $V_7$ |
| 7 | 1 | $V_1$, $V_2$, $V_3$, $V_4$, $V_5$, $V_6$ e $V_7$ |
| Soma | 121 | |

## 2.2 Métodos

Para avaliação dos testes em cada variável resposta simulada, contabilizou-se o número de vezes em que a hipótese nula foi aceita, apresentando os resultados por meio de porcentagem (número de aceite de $H_0$/1000) x 100%.

A fim de avaliar a eficiência da transformação dos dados multivariados por meio da função discriminante linear de Fisher e comparar seus resultados aos testes da análise de variância multivariada, cada conjunto de dados simulados foram submetidos aos testes descritos na sequência:

### 2.2.1 Avaliação das pressuposições para análise de variância

Para cada uma das sete variáveis respostas ($V_1$, $V_2$, $V_3$, $V_4$, $V_5$, $V_6$ e $V_7$) simuladas, em cada um dos diferentes cenários propostos, na ausência de correlação, foi realizada a análise de variância (univariada) segundo Pimentel Gomes (2009), sendo avaliados os pressupostos de normalidade do erro, por meio do teste de Shapiro-Wilk (SHAPIRO e WILK, 1965), e homogeneidade de variâncias, por meio do teste de Bartlett (BARTLETT, 1937), além do teste F para tratamentos. A avaliação dentro dos critérios propostos considerou o cálculo dos percentuais de aceitação das hipóteses nulas, sendo que a contagem se deu sob nível de significância de 5%.

### 2.2.2 Obtenção da função discriminante de Fisher

As funções discriminantes lineares amostrais são determinadas após calcular os $s$ autovalores não nulos ($\lambda_s$) de $\mathbf{R}^{-1}\mathbf{H}$, sendo $s = \min(I - 1, p)$, em que $\mathbf{R}$ e $\mathbf{H}$, são, respectivamente, como na análise multivariada de variância, as matrizes de soma de quadrados e de produtos devidos aos efeitos dos fatores não controlados e dos tratamentos,

$I$ é o número de tratamentos e $p$ o número de características avaliadas; e nesse trabalho, $p$ = 7, $I$ = 4, resultando em $s$ = 3.

Assim, após a simulação, para cada um dos seis cenários propostos, foi obtida a função discriminante linear de Fisher (FDF). Para isso, foi escrita a combinação linear $Z = a_1y_1 + a_2y_2 + \ldots + a_ly_l$ que transforma o espaço $p$ dimensional em unidimensional. Para essa transformação foi necessário encontrar o autovetor $t$ que maximiza a razão soma de quadrados das distâncias entre as médias das populações e sua variância: $\frac{t'Ht}{t'Rt}$, na qual: $t$ é o autovetor associado ao maior autovalor ($\lambda_1$) de $\mathbf{R}^{-1}\mathbf{H}$.

Os dados transformados, por meio da aplicação das FDF, foram submetidos aos mesmos testes propostos para as variáveis respostas simuladas na ausência de correlação, de modo a avaliar as características de normalidade e homogeneidade de variância desta nova variável.

Para os cenários propostos sob heterocedasticidade ($H_0 - H_1$, $H_1$ 2$s$ $- H_1$ e $H_1$ 8$s$ $- H_1$), foi verificada a possibilidade do uso da transformação dos dados por meio da FDF, como alternativa de tratamento dos dados heterocedásticos.

De maneira a validar a FDF foram estimados o percentual de explicação dado por (PI = $\frac{\lambda_1}{\sum_{i=1}^{s} \lambda_s}$), indicado por Padovani e Aragon (2005).

### 2.2.3 Análise de variância multivariada

Os subconjuntos de variáveis respostas gerados também foram submetidos às análises de variância multivariadas (MANAVA), sendo os tratamentos avaliados por meio dos quatro testes aproximados de máximo autovalor $\theta$ de Roy; $\Lambda$ de Wilks; Hotelling e Lawley (U) e Pillai (V) de acordo com sugestões apresentadas por Ferreira (2018) e Haase e Ellis (1987).

### 2.2.4 Comparação FDF e MANAVA

De modo a comparar os resultados obtidos por cada uma das duas metodologias multivariadas, FDF e MANAVA, foram confrontados os valores p obtidos para as análises univariadas, realizadas com os dados transformados pelas respectivas FDF estimadas aos valores de p obtidos pelos quatro testes F aproximados para as análises de variância multivariada.

A comparação foi realizada considerando a contagem dos valores de p que não rejeitaram a hipótese nula da igualdade das médias dos tratamentos ($H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$) e posterior cálculo da porcentagem de aceitação considerando o nível de significância de 5%.

### 2.2.5 Efeito da correlação nos testes multivariados

De maneira a avaliar a influência da correlação existente entre as variáveis resposta simuladas nos resultados dos testes F, tanto da ANAVA realizada com os dados transformados por meio da FDF, quanto com os aproximados da MANAVA, em cada hipótese simulada, foram agrupados os percentuais de aceitação da hipótese nula dos testes realizados com o subconjunto formado com todas as variáveis simuladas ($V_1$, $V_2$, $V_3$, $V_4$, $V_5$,

$V_6$ e $V_7$) pelo nível de correlação estipulado para a simulação ($\rho$=0,0; $\rho$=0,2; $\rho$ = 0, 5 e $\rho$ = 0, 9) e o número de variáveis respostas simuladas com correlação (0, 2, 3, 4, 5, 6, e 7).

Estas separações foram estabelecidas de maneira a permitir a comparação do comportamento da aceitação por meio dos testes propostos com o aumento tanto no nível de correlação, quanto na quantidade de variáveis correlacionadas, em ambos os casos o primeiro valor representa a ausência da correlação.

### 2.2.6 Programa computacional

Todas as simulações realizadas e suas respectivas análises, tanto para os dados reais quanto para os dados simulados, foram feitas por meio de rotinas específicas elaboradas no programa estatístico R (R CORE TEAM, 2019).

## 3 Resultados e discussão

### 3.1 Avaliação das pressuposições para as simulações univariadas

Do total de simulações para cada um dos cenários testados estudou-se o comportamento dos dados que foram simulados sem considerar correlação. Nas Tabelas 4, 5 e 6 estão representados os percentuais de aceitação das hipóteses nulas para o efeito dos tratamentos ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$); para as variâncias ($H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$) e para normalidade dos erros $e_i \sim N\left(0, s^2\right)$ das análises univariadas das sete variáveis respostas.

Pode-se observar na Tabela 4, que as simulações das hipóteses sobre as médias sob $H_0$ verdadeira, que de modo geral, houve um bom controle do erro tipo I, pois todas as variáveis respostas simuladas apresentaram aceitação da igualdade dos efeitos dos tratamentos superior a 95% (teste F) para o cenário $H_0{-}H_0$. A análise dos pressupostos de normalidade dos erros e homocedasticidade de variância também aceitaram as hipóteses nulas, pois o teste de Shapiro-Wilk detectou que em mais de 96% houve o aceite da normalidade, o teste de Bartlett detectou que em mais de 99% dos casos houve aceite de homocedasticidade.

Quanto aos dados simulados sob $H_0{-}H_1$ os testes de Shapiro-Wilk, Bartlett e $F$, apresentaram percentagem de aceitação da igualdade dos efeitos dos tratamentos pouco inferior ao dos resultados simulados sob $H_0{-}H_0$; os percentuais de aceitação de homocedasticidade variou de 95,97% até 96,77% o que pode ser considerado alto, já que na simulação esperava-se detectar mais casos de heterogeneidade; no entanto são inferiores ao caso de homocedasticidade que detectou mais de 99% de aceite de $H_0$. No caso de normalidade dos erros, observa-se (Tabela 4) que sob o cenário $H_0{-}H_1$, a porcentagem de detecção de $H_0$ variou de 93,53% a 94,23%, reforçando a pressuposição de que a presença de certo nível de heterocedasticidade, demanda maiores cuidados com relação à normalidade dos erros, pois houve uma queda na aceitação da hipótese nula.

Tabela 4 - Percentagens de aceitação da normalidade dos erros (Shapiro-Wilk), da homogeneidade de variâncias (Bartlett) e do teste F para tratamentos, em 5%, sob a hipótese nula ($H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$), obtidas para os dados simulados com $\rho = 0$, e sob cenário homocedástico ($H_0-H_0$) e heterocedástica ($H_0-H_1$)

| Cenário | Variável | Shapiro-Wilk | Bartlett | Teste F (p<0,05) |
|---|---|---|---|---|
| $H_0-H_0$ | $V_1$ | 97,13 | 99,47 | 95,47 |
| | $V_2$ | 96,97 | 99,57 | 95,53 |
| | $V_3$ | 97,37 | 99,60 | 95,07 |
| | $V_4$ | 97,13 | 99,30 | 94,97 |
| | $V_5$ | 96,77 | 99,53 | 96,03 |
| | $V_6$ | 97,07 | 99,43 | 95,30 |
| | $V_7$ | 97,53 | 99,67 | 95,07 |
| $H_0-H_1$ | $V_1$ | 94,13 | 95,97 | 90,17 |
| | $V_2$ | 93,67 | 96,77 | 90,43 |
| | $V_3$ | 94,23 | 96,93 | 90,80 |
| | $V_4$ | 93,53 | 95,97 | 90,60 |
| | $V_5$ | 94,17 | 96,67 | 90,47 |
| | $V_6$ | 93,90 | 96,80 | 90,33 |
| | $V_7$ | 93,57 | 96,50 | 91,07 |

De modo geral, o teste Bartlett obteve elevados valores de aceite da homocedasticidade. Os cenários simulados sob a igualdade das variâncias dos tratamentos ($H_0-H_0$, $H_12s-H_0$ e $H_18s-H_0$) obtiveram o percentual mínimo 99,30% das amostras geradas consideradas homocedásticas, em duas variáveis respostas: $V_4$ (Tabela 4: cenário sob a hipótese nula para as médias e para as variâncias) e $V_1$ (Tabela 5: cenário sob a hipótese alternativa para as médias, com diferença estipulada em oito desvios padrão e hipótese nula para as variâncias). Entretanto, com o aumento na diferença entre as médias para oito desvios padrão, houve rejeição em 100% da hipótese nula, por meio do teste F em 5% (Tabela 5).

Do mesmo modo, para o teste F houve uma redução na porcentagem de aceitação de $H_0$. No nível de 1% observa-se uma porcentagem de aceite de $H_0$ de 96% (aproximadamente) e no nível de 5%, um percentual de 90%. E esta redução em relação ao cenário $H_0-H_0$ é devida à presença de variâncias heterogêneas.

Já para os dados simulados sob o cenário de $H_0$ falsa para as médias dos tratamentos, mas homocedástico ($H_1-H_0$), observa-se na Tabela 5, a porcentagem de aceite de normalidade dos erros não foi afetada com o aumento nas diferenças entre as médias de tratamentos; o mesmo ocorreu com aceite da homogeneidade de variâncias (teste de Bartlett). Para o teste F, a diferença de dois desvios padrão, no nível de significância 1%, apresenta valores de aceitação da igualdade de tratamentos mínima 47,77% para a variável resposta $V_2$ e máxima 51,20% ($V_4$). Este erro de detecção nas diferenças dos efeitos dos tratamentos, erro tipo II, pode ser explicado em função da variabilidade dos dados, indicando que diferença entre médias de dois desvios padrão não é facilmente detectada.

Tabela 5 - Percentagens de aceitação da normalidade dos erros (Shapiro-Wilk), da homogeneidade de variâncias (Bartlett) e do teste F para tratamentos, 5%, sob hipótese nula ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$), dos dados simulados sem a presença de correlação ($\rho = 0$) e sob hipóteses homocedásticas, com diferenças na médias dadas por dois ($H_1 2s - H_0$) e oito desvios padrão ($H_1 8s - H_0$)

| Cenários | Variável | Shapiro-Wilk | Bartlett | Teste F (p<0,05) |
|---|---|---|---|---|
| | $V_1$ | 96,93 | 99,53 | 19,20 |
| | $V_2$ | 96,70 | 99,37 | 19,13 |
| | $V_3$ | 97,10 | 99,43 | <0,0001 |
| $H_1 2s - H_0$ | $V_4$ | 97,33 | 99,73 | 21,67 |
| | $V_5$ | 96,70 | 99,50 | 19,77 |
| | $V_6$ | 96,20 | 99,43 | 20,50 |
| | $V_7$ | 96,70 | 99,60 | 20,20 |
| | $V_1$ | 97,13 | 99,30 | <0,0001 |
| | $V_2$ | 96,80 | 99,57 | <0,0001 |
| | $V_3$ | 97,17 | 99,60 | <0,0001 |
| $H_1 8s - H_0$ | $V_4$ | 96,47 | 99,43 | <0,0001 |
| | $V_5$ | 97,13 | 99,63 | <0,0001 |
| | $V_6$ | 97,40 | 99,53 | <0,0001 |
| | $V_7$ | 97,17 | 99,60 | <0,0001 |

Para os dados simulados sob a heterocedasticidade ($H_0 - H_1$, $H_1 2s - H_1$ e $H_1 8s - H_1$) com amplitude de até 64 vezes a variância populacional ($\sigma^2$), o percentual de aceitação da homogeneidade de variâncias ficou entre 97,37% para variável $V_2$ e 95,80% para a variável $V_3$, ambas obtidas para o cenário $H_1 8s - H_1$ (Tabela 6), resultado que reforça a necessidade de ao analisar dados univariadamente.

Para o cenário heterocedástico, Tabela 6, quando a diferença entre as médias foi estabelecida em dois desvios padrão, mesmo se considerado o nível de significância p<0,05, o percentual de aceitação da igualdade entre o efeito dos tratamentos foi superior a 86% dos casos. Vale o mesmo raciocínio feito para os dados simulados com médias diferentes e homocedásticos; e, mesmo tendo aumentado a diferença entre médias para oito desvios padrão, houve aceitação superior a 14%. Tal fato reforça a pressuposição de que a heterogeneidade da variância interfere na detecção da igualdade dos tratamentos, pois o cenário $H_0 - H_1$, também apresentou diminuição no percentual de aceitação da hipótese nula para as médias dos tratamentos, novamente indicando a necessidade dos testes dos pressupostos da análise de variância ao optar pelo estudo univariado.

Observa-se, também, que os percentuais de aceitação de normalidade e homogeneidade de variâncias foram menores do que aqueles observados sob $H_0$. Para o teste de Bartlett, talvez fosse esperado até um menor percentual de aceitação, já que a simulação é feita sob $H_1$; mas, o teste de Shapiro-Wilk, os dados são simulados sob normalidade, logo esperava-se uma aceitação maior. Com o aumento na diferença entre as médias (Tabela 6) de dois para oito desvios padrão, não houve diferenças significativas nas porcentagens de aceitação de normalidade e homocedasticidade. Isto sugere que a

simulação foi realizada dentro dos critérios propostos dando confiabilidade aos dados obtidos.

Tabela 6 - Percentagens de aceitação da normalidade dos erros (Shapiro-Wilk), da homogeneidade de variâncias (Bartlett) e do teste F para tratamentos, 5%, sob hipótese nula ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$), dos dados simulados com $\rho = 0$, e sob hipótese heterocedástica, com diferenças nas médias dadas por dois ($H_1 2s - H_1$) ou oito desvios padrão ($H_1 8s - H_1$)

| Cenários | Variável | Shapiro-Wilk | Bartlett | Teste F (p<0,05) |
|---|---|---|---|---|
| $H_1 2s - H_1$ | $V_1$ | 93,10 | 96,13 | 88,10 |
| | $V_2$ | 93,37 | 96,33 | 87,97 |
| | $V_3$ | 94,30 | 96,43 | 88,30 |
| | $V_4$ | 94,47 | 96,30 | 88,13 |
| | $V_5$ | 93,97 | 96,60 | 88,57 |
| | $V_6$ | 94,00 | 96,17 | 87,77 |
| | $V_7$ | 93,90 | 96,17 | 87,23 |
| $H_1 8s - H_1$ | $V_1$ | 94,33 | 96,67 | 27,30 |
| | $V_2$ | 93,43 | 97,37 | 28,37 |
| | $V_3$ | 92,90 | 95,80 | 29,33 |
| | $V_4$ | 93,40 | 96,43 | 28,37 |
| | $V_5$ | 93,93 | 97,07 | 27,37 |
| | $V_6$ | 94,43 | 96,57 | 27,83 |
| | $V_7$ | 93,80 | 96,63 | 26,87 |

### 3.2 Eficiência da FDF

Com o intuito de validar a viabilidade do uso da função discriminante linear de Fisher (FDF) na transformação de conjuntos de variáveis respostas em dados univariados para posterior análise de variância univariada (ANAVA), após a obtenção da FDF, estimaram-se, os valores médios da porcentagem de informação ou explicação, por meio da expressão $PI = \frac{\lambda_1}{\sum_{i=1}^{s} \lambda_s}$, em cada um dos seis cenários na Tabela 7.

Tabela 7 - Valores médios da porcentagem de informação (PI) dada pelo maior autovalor ($\lambda_1$), porcentagem da aceitação da normalidade dos erros, teste de Shapiro-Wilk, 5% (S-W) e da homogeneidade de variâncias, teste de Bartlett, 5% (B), para cada cenário proposto

| Teste | Cenário proposto | | | | | |
|---|---|---|---|---|---|---|
| | $H_0$- $H_0$ | $H_0$- $H_1$ | $H_1$2s- $H_0$ | $H_1$8s- $H_0$ | $H_1$2s- $H_1$ | $H_1$8s- $H_1$ |
| PI | 44,79 | 55,87 | 63,31 | 94,48 | 55,92 | 70,94 |
| S–W | 96,92 | 97,27 | 96,94 | 96,94 | 97,25 | 97,22 |
| B | 96,87 | 93,33 | 96,90 | 96,85 | 92,93 | 93,39 |

Os valores da porcentagem de explicação ficaram no intervalo [44,79%; 94,48%], indicando que o uso da FDF é bastante informativo. Ao considerar os cenários gerados, tem-se que $H_0{-}H_0$ apresentou os menores valores de informação quando comparados a todos os outros, sugerindo que a FDF deve apresentar restrições quando os efeitos dos tratamentos forem nulos, podendo até aumentar o erro tipo I.

Os cenários simulados sob heterocedasticidade, tanto sob a igualdade dos efeitos dos tratamentos quanto com diferenças entre suas médias de dois desvios padrão ($H_0{-}H_1$ e $H_1 2s{-}H_1$), têm percentuais médios de informação muito semelhantes e tal semelhança entre o percentual de informação destes dois cenários e seus valores médios estimados abaixo de 56% reforçam a necessidade de atenção, pois podem ser indícios de que a FDF é menos rigorosa e detecta diferenças quando estas não existem, sob variâncias heterogêneas. Os outros cenários simulados sob a hipótese alternativa para tratamentos apresentaram percentuais de informação superiores a 60%.

Há grande incremento no percentual de informação, com o aumento na diferença entre médias dos tratamentos.

Os cenários simulados sob $H_0$ falsa para os tratamentos, nos quais, também, foi mantida a homogeneidade de variâncias, apresentam os mais elevados percentuais de informação, sendo aumentados conforme se aumenta a diferença entre as médias, por exemplo, $H_1 2s{-}H_0$ (63,31%) e $H_1 8s{-}H_0$ (94,48%), sugerindo que a FDF, quando aplicada a dados homocedásticos, é capaz de detectar mais eficientemente as diferenças dos efeitos dos tratamentos, e esta capacidade aumenta com o aumento na diferença das médias. Entretanto estes percentuais caem, ao se considerar os dados heterocedásticos, os valores percentuais de explicação para dois desvios padrão caem para 55,92% e 70,94% para oito desvios padrão, o que pode indicar que mesmo com grandes diferenças entre os efeitos dos tratamentos quando existirem indicações de heterogeneidade das variâncias é necessária uma análise mais criteriosa dos dados antes de proceder à transformação dos dados pela FDF.

As observações sobre o percentual de informação sugerem que há a necessidade de outros testes, quando a ANAVA dos dados transformados por meio da FDF detectar diferença entre os efeitos dos tratamentos juntamente com valores pequenos do percentual

de informação (para os dados simulados sob a igualdade dos efeitos dos tratamentos e homocedásticos os valores de informação foram menores que 56%), pois pode ser indício de erro tipo I e não existir diferença significativa entre os efeitos dos tratamentos.

Quanto as pressuposições testadas, os valores de aceitação da hipótese nula para a normalidade dos erros (Tabela 7) mostram que todos os cenários apresentaram dados transformados por meio da FDF com valores superiores a 96% de aceitação da normalidade dos erros. Fato que viabiliza o uso da transformação dos dados por meio da FDF, pois os percentuais de aceitação da normalidade do erro, testados por meio do teste Shapiro-Wilk, por meio das respectivas FDF, foram semelhantes aos obtidos quando foi realizada a avaliação univariada na ausência de correlação, Tabela 4.

Ao comparar estes resultados com os apresentados nas Tabelas 4 e 5, verifica-se que os cenários sob homogeneidade de variâncias apresentam valores de aceitação da hipótese de normalidade dos erros dos dados transformados, em qualquer critério, no intervalo formado pelos menor e maior percentuais da avaliação das variáveis respostas geradas em cada cenário. Por exemplo, se considerar $H_1 2s - H_0$ (Tabela 5), os valores estão [96,20; 97,33]%, respectivamente para as variáveis $V_4$ e $V_6$, enquanto os dados transformados apresentaram 96,94% da aceitação da normalidade.

Os cenários gerados sob a heterocedasticidade, apresentaram elevação do percentual de aceitação da normalidade quando comparado ao estudo univariado sem a presença de correlação (Tabelas 4 e 6) com a aceitação da normalidade dos dados transformados, por exemplo, o menor valor individual do cenário $H_1 8s - H_1$ foi atribuído à variável $V_3$ (92,90%) e para os dados transformados o aceite ficou em 97,22%.

A aceitação da homocedasticidade (Tabela 7), tem uma amplitude maior que a encontrada para a normalidade dos erros, [92,93%; 96,90%]. E, como esperado, os menores valores de aceitação da hipótese nula da homogeneidade de variâncias foram atribuídos aos cenários simulados sob a violação dessa hipótese. O teste de Bartlett dos dados transformados pela FDF rejeitaram em maior número de vezes a homocedasticidade, quando comparado ao mesmo teste aplicado às variáveis geradas sem a presença de correlação e estudadas individualmente (Tabelas 4 a 6). Univariadamente os valores simulados sob homocedasticidade apresentaram aceite superior a 99% e, para os dados transformados, este valor cai para aproximadamente 96%. Os cenários sob a hipótese alternativa da homogeneidade das variâncias apresentam valores próximos a 93%, entretanto o estudo univariado apresentou aceite sempre superior a 95%.

Estes resultados sugerem, também, a necessidade de testes nas variáveis originais, antes da transformação dos dados, pois se existir violação da homogeneidade das variâncias, deve-se tomar providências para controlar a violação; o que corrobora com Hair et al. (2009), que relatam ser a FDF sensível às quebras de suposições.

Os cenários propostos sob a hipótese alternativa para efeito dos tratamentos com amplitudes diferentes de médias, estipuladas pela diferença do número de desvios padrão adicionados à média geral, não apresentam grandes alterações no número de vezes em que são aceitas as hipóteses nulas dos pressupostos.

Os elevados valores de aceitação das hipóteses nulas dos pressupostos da normalidade dos erros e da homogeneidade de variâncias dos dados transformados por meio da FDF, mesmo para os cenários simulados sob heterocedasticidade, são indicativos da eficiência da FDF ao lidar com dados que violam tais pressupostos.

### 3.3 Comparação FDF/MANAVA

De modo a verificar se a aplicação da função discriminante linear de Fisher (FDF) apresenta resultados compatíveis aos encontrados por meio da análise de variância multivariada (MANAVA), foram comparados os valores p da análise de variância (ANAVA) para os dados transformados aos valores-p obtidos em cada um dos quatro testes F aproximados da MANAVA, máximo autovalor $\theta$ de Roy, $\Lambda$ de Wilks, Hotelling e Lawley (U) e Pillai (V).

Para os cenários simulados sob a hipótese nula, $H_0{-}H_0$ e $H_0{-}H_1$, esperava-se que tanto o teste F, da ANAVA com os dados transformados, quanto os testes F aproximados, da MANAVA, aceitassem a igualdade entre os efeitos dos tratamentos com percentuais elevados. E de maneira inversa, as hipóteses geradas sob a hipótese alternativa, $H_1 2s{-}H_0$, $H_1 8s{-}H_0$, $H_1 2s{-}H_1$ e $H_1 8s{-}H_1$, apresentassem percentuais pequenos de aceite da igualdade dos tratamentos, visto que foram simuladas com no mínimo dois desvios padrão de diferença.

Em diversas situações dentre as simuladas, os testes F aproximados não detectam as diferenças entre os efeitos dos tratamentos de forma unânime ou esperada, entretanto pode ser observado que qualquer que seja o subconjunto de variáveis resposta estudado, para qualquer hipótese simulada, quando não há equivalência dos testes F aproximados, há um ranqueamento do poder desses testes, em ordem decrescente, Pillai (V), $\Lambda$ de Wilks, Hotelling-Lawley (U) e $\theta$ de Roy (Figura 1). Tal constatação contraria a afirmação de Chatfield e Collins (1980) de que os testes Pillai (V), $\Lambda$ de Wilks, Hotelling-Lawley (U) são assintoticamente equivalentes e diferem pouco em poder para pequenas amostras.

A aceitação da hipótese nula dada por meio do valor-p do teste $F$ da ANAVA, dos dados transformados pela FDF, é semelhante ao encontrado por $\theta$ de Roy, nos cenários $H_1 8s$-$H_0$ e $H_1 8s$-$H_1$, Figura 1. Uma possível explicação para o comportamento menos rigoroso das ANAVAS dos dados transformados por meio da FDF pode estar na forma como ela é construída, que determina o teste F de tratamentos da análise de variância com máximo valor possível (PIMENTEL GOMES, 2009).

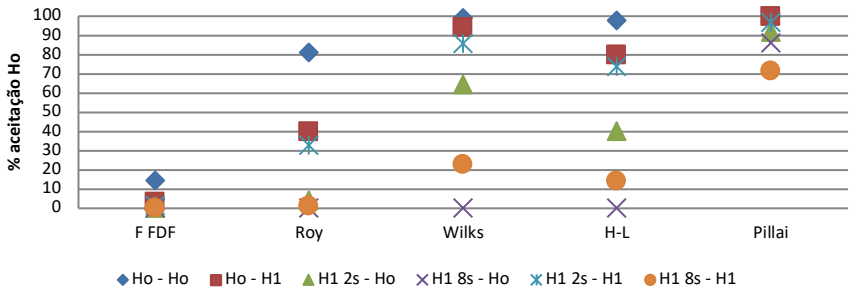


Figura 1 - Porcentagem de aceitação da hipótese nula para tratamentos (%), obtida em cada cenário ($H_0 - H_0$, $H_0 - H_1$, $H_1 2s - H_0$, $H_1 8s - H_0$, $H_1 2s - H_1$ e $H_1 8s - H_1$) para os testes F, da análise dos dados transformados pela FDF (F FDF), e em cada teste F aproximado da análise multivariada (Roy, Wilks, H-L e Pillai).

Na Tabela 8, estão os percentuais de aceitação da igualdade dos efeitos dos tratamentos para os cenários simulados $H_0 - H_0$ e $H_0 - H_1$. Em ambos os casos, os dados transformados por meio das FDF foram mais sensíveis, com menores valores do erro tipo I.

Tabela 8 - Porcentagem de aceitação da hipótese nula para os dois cenários simulados sob $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (com p<0,01), nos diferentes critérios de seleção de variáveis e testes F da análise da FDF e da análise multivariada

| Cenário | FDF | Teste da análise multivariada | | | |
|---|---|---|---|---|---|
| | Teste F | $\Theta$ Roy | $\Lambda$ de Wilks | Hotelling e Lawley (U) | Pillai (V) |
| $H_0-H_0$ | 14,26 | 81,03 | 98,97 | 97,61 | 99,53 |
| $H_0-H_1$ | 3,11 | 39,68 | 93,91 | 79,59 | 99,81 |

Os testes da MANAVA são, em geral, mais rigorosos que a ANAVA dos dados transformados por meio da FDF, quando observados os dados simulados sob igualdade de efeito de tratamentos, pois a porcentagem de aceitação de $H_0$ dos testes multivariados foram maiores do que pelo uso da FDF. Tomando apenas o teste $\theta$ de Roy, que apresenta o menor rigor dos quatro testes F aproximados, a aceitação da igualdade dos efeitos dos tratamentos chega a ser 5,7 vezes a aceitação da encontrada pelo teste F dos dados transformados. Esta tendência se mantém para o cenário sob heterogeneidade de variâncias, $H_0-H_1$, no qual apresenta diferenças maiores, a aceitação por meio do teste $\theta$ de Roy ficou mais de 12 vezes superior ao aceite estabelecido pelo teste F dos dados transformados. Tais diferenças sugerem que a FDF apresente maiores erros do tipo I.

Tanto os testes F aproximados quanto o teste F dos dados transformados por FDF, para o cenário $H_0-H_1$, apresentaram índices de aceitação da hipótese nula de tratamentos menores dos que os encontrados para o cenário $H_0-H_0$, sugerindo que a heterocedasticidade deva ser contornada antes da aplicação de tais testes. A mesma interpretação foi feita quando do estudo da eficiência da FDF. Vale ressaltar que o teste F aproximado (V) de Pillai, o mais rigoroso em todas as situações, não foi afetado pela presença de diferenças nas variâncias no cenário simulado $H_0-H_1$, apresentando percentuais de aceitação da hipótese nula maiores do que os encontrados para os cenários sem variância $H_0-H_0$.

Na Tabela 9, estão apresentadas as porcentagens de aceitação de $H_0$, nos cenários de variâncias homocedásticas e com diferença entre as médias dos tratamentos de dois e oito desvios padrão. A ANAVA dos dados transformados consegue detectar as diferenças de tratamentos desde o primeiro cenário, $H_12s-H_0$, sendo os percentuais de aceitação da hipótese nula para tratamentos menores que 1%.

Tabela 9 Porcentagem de aceitação da hipótese nula do teste F da análise da FDF e dos testes da análise de multivariada, para os cenários simulados sobre a hipótese nula falsa das médias dos tratamentos, com p<0,01

| Cenário | FDF | Teste da análise multivariada | | | |
|---|---|---|---|---|---|
| | Teste F | $\theta$ Roy | $\Lambda$ de Wilks | Hotelling e Lawley (U) | Pillai (V) |
| $H_1 2s - H_0$ | 0,01 | 4,15 | 64,40 | 40,11 | 91,57 |
| $H_1 8s - H_0$ | <0,0001 | <0,0001 | 0,10 | <0,0001 | 85,93 |
| $H_1 2s - H_1$ | 1,32 | 32,71 | 85,54 | 73,86 | 96,57 |
| $H_1 8s - H_1$ | <0,0001 | 1,24 | 22,64 | 14,13 | 71,30 |

Entretanto, a MANAVA quando comparada a ANAVA dos dados transformados por meio da FDF é mais rigorosa e aceita a hipótese nula, mesmo existindo diferença de dois desvios padrão, aumentando o erro do tipo II dos testes.

Com o aumento na diferença entre as médias dos tratamentos para oito desvios padrão $H_1 8s - H_0$, os testes já se apresentam compatíveis, persistindo ainda, apenas o teste F aproximado de Pillai (V) com altos índices de aceitação da igualdade dos efeitos dos tratamentos (maior erro do tipo II).

Esses resultados sugerem que o aumento da diferença dos efeitos dos tratamentos faz com que a MANAVA apresente menor erro do tipo II, exceto o F aproximado de Pillai que mesmo com diferenças de oito desvios padrão não as consegue detectar.

Os cenários gerados com médias diferentes e variâncias heterocedásticas, estão na Tabela 9. Novamente pode-se perceber que, quando se aumenta a diferença entre as médias, a aceitação da igualdade das médias dos tratamentos é menor do que os valores encontrados para os cenários simulados sem variâncias distintas.

Como aconteceu para os cenários de médias distintas e homocedásticos o teste F da ANAVA dos dados transformados por meio da FDF foi capaz de detectar as diferenças dos efeitos dos tratamentos de maneira mais eficaz.

A presença de heterogeneidade de variâncias parece atrapalhar a rejeição da hipótese nula, principalmente para a MANAVA, e assim como ocorrido para as hipóteses homocedásticas.

Os erros encontrados, tanto o tipo I quanto o tipo II, podem ser explicados se considerar a afirmação feita por diversos autores, dentre eles Manly e Alberto (2019), os quais ressaltam que a igualdade ou desigualdade de efeitos univariados pode não ser confirmada com a análise de variância multivariada.

### 3.4 Efeitos da correlação nos testes propostos

Foram estimados os percentuais de aceitação de $H_0$ para tratamentos, sob os diferentes cenários, em função dos níveis de correlação e do número de variáveis correlacionadas, para a avaliação do comportamento dos resultados obtidos pelo teste F oriundos dos dados transformados pela FDF e dos resultados dos testes da análise multivariada, considerando o nível nominal de significância de 1%.

O teste F da ANAVA dos dados transformados por meio da FDF é pouco afetado tanto pelo nível de correlação quanto pela quantidade de variáveis correlacionadas. Na Figura 2, observa-se que não houve efeito do aumento no nível de correlação entre as variáveis sobre a porcentagem de aceitação de $H_0$, em todas os cenários simulados.

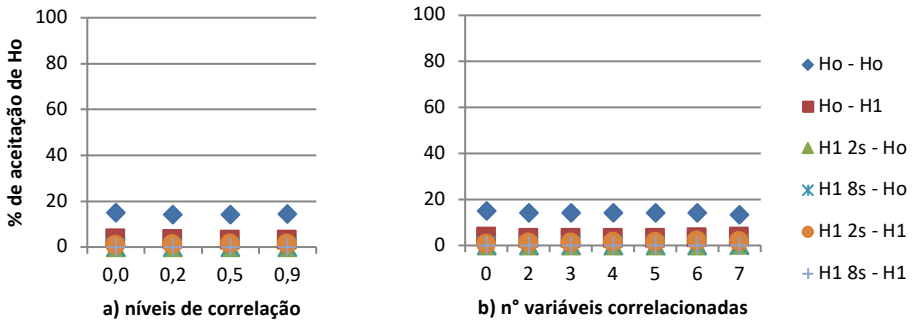


Figura 2 - Representação gráfica do comportamento da aceitação da hipótese nula de médias de tratamento pelo teste F dos dados transformados pelas FDF, nos a) níveis de correlação (0,0; 0,2; 0,5 e 0,9) e b) número de características correlacionadas (0, 2, 3, 4, 5, 6, 7).

Do mesmo modo, o número de variáveis correlacionadas não afetou a porcentagem de aceite de $H_0$ , em todo os cenários. Para o cenário $H_0 - H_0$, as taxas de aceitação do $H_0$ ficaram próximas a 15%, enquanto que para os demais cenários, as porcentagens de aceitação de $H_0$ ficaram menores que 5%.

A porcentagem de aceitação da hipótese $H_0$, obtida pelo critério $\theta$ de Roy (Figura 3), variou de acordo com o cenário simulado; o nível de correlação somente afetou o cenário $H_1 2s - H_0$, que na ausência de correlação tinha aproximadamente 0% de aceitação, passando para 7,5% no nível de alta correlação entre as variáveis; sugestionando que a correlação interfere na precisão do erro do tipo II. E, com o aumento do número de variáveis correlacionadas, houve o acréscimo na porcentagem de aceitação de $H_0$ nos cenários $H_1 2s - H_1$, $H_1 2s - H_0$ e $H_1 8s - H_1$.

Nos outros cenários, os valores da aceitação de $H_0$ permaneceram praticamente constantes com o aumento do número de variáveis correlacionadas.

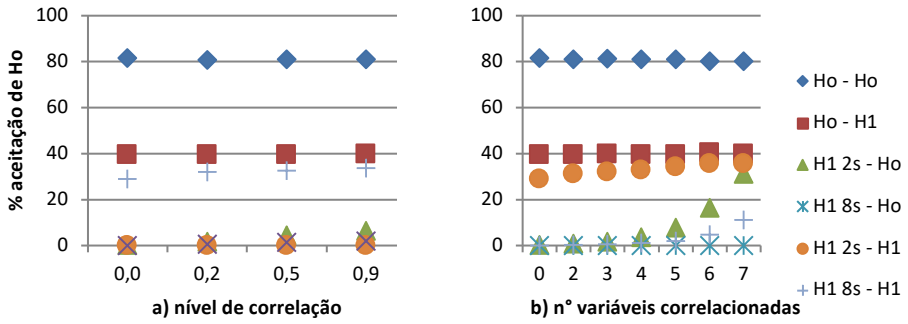


Figura 3 - Representação gráfica do comportamento da aceitação da hipótese nula de médias de tratamento pelo teste F aproximado máximo autovalor $\theta$ de Roy, nos níveis de correlação (0,0; 0,2; 0,5 e 0,9) e b) número de características correlacionadas (0, 2, 3, 4, 5, 6, 7).

Na Figura 4, estão representados os percentuais de aceite de $H_0$, para tratamentos, agrupados por efeitos da correlação, tanto por nível quanto pelo número de características geradas com correlação, para os critérios multivariados $\Lambda$ de Wilks e Hotelling-Lawley.

Nela, percebe-se que os maiores incrementos foram para o cenário homocedástico com menor diferença entre as médias ($H_1 2s - H_0$) e também que, quando se aumentam as diferenças para oito desvios padrão, o nível de correlação mantém sua influência, mas com variações menores, entretanto o número de variáveis continua levando a uma maior aceitação da hipótese nula, principalmente para $\Lambda$ de Wilks.

De modo geral, os testes $\Lambda$ de Wilks e Hotelling-Lawley são semelhantes com diferenças apenas na aceitação menor por parte do critério Hotelling-Lawley nos cenários simulados sob diferença entre as médias dos tratamentos. Portanto, Hotelling-Lawley apresenta menor erro do tipo II, quando comparado ao teste $\Lambda$ de Wilks, porém há, mesmo na ausência de correlação, altos índices de aceite de $H_0$ que são incrementados com o aumento no nível de correlação e no número de variáveis correlacionadas.

A tendência, quando existe, para esses três testes F aproximados ($\theta$ de Roy, $\Lambda$ de Wilks e Hotelling-Lawley) é de aumento na aceitação com o aumento tanto do nível de correlação, quanto da quantidade de variáveis correlacionadas. No primeiro caso o incremento é gradual, enquanto que o número de variáveis tem uma aceleração de crescimento a partir de quatro variáveis.

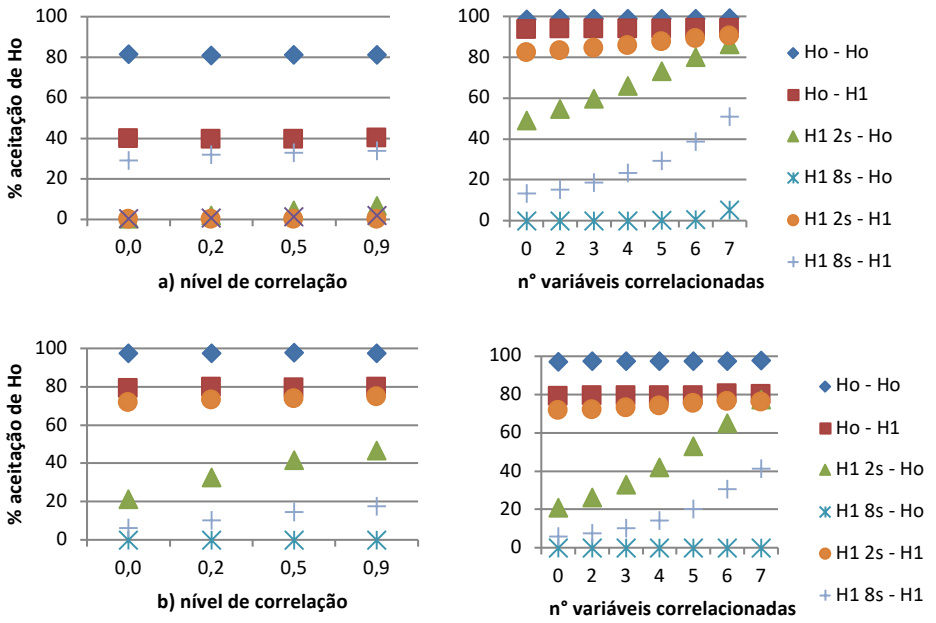


Figura 4 - Representação gráfica do comportamento da aceitação da hipótese nula de médias de tratamento nos níveis de correlação (0,0; 0,2; 0,5 e 0,9) e no número de características correlacionadas (0, 2, 3, 4, 5, 6, 7), obtidas pelos testes F aproximado a) Λ de Wilks e b) Hotelling-Lawley.

O critério Pillai (Figura 5) apresenta, para os cenários simulados sob a hipótese alternativa, uma tendência crescente de aceitação com o aumento do nível/número de variáveis com correlação. Dentre os quatro testes multivariados estudados, o critério de Pillai é o mais rigoroso em todos os cenários, e mesmo com diferenças de oito desvios padrão e variâncias homogêneas ($H_1 8 - H_0$) esse teste apresentou aceite de $H_0$, superior a 80%, mesmo na ausência de correlação.

De forma geral, os testes aproximados da MANAVA foram mais influenciados pela presença de correlação e do aumento no número de variáveis correlacionadas que o teste F da ANAVA dos dados transformados por meio da FDF quando na introdução de correlação, tanto em nível quanto em número de variáveis correlacionadas, sugerindo que os testes F aproximados apresentam aumento do erro tipo II, com a presença de correlação.

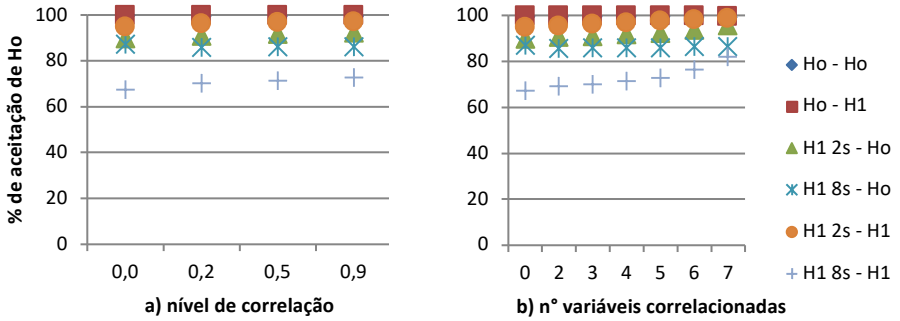


Figura 5 - Representação gráfica do comportamento da aceitação da hipótese nula de médias de tratamento pelo teste F aproximado Pillai, nos a) níveis de correlação (0,0; 0,2; 0,5 e 0,9) e b) número de características correlacionadas (0, 2, 3, 4,5, 6, 7).

### 3.5 Efeitos da correlação na validação da FDF

Os testes realizados para a validação do uso da FDF, porcentagem de informação atribuída às FDF, além da normalidade do erro (Shapiro-Wilk) e homocedasticidade (Bartlett) dos dados transformados por meio da FDF também foram estudadas sobre a interferência do nível de correlação e do número de variáveis geradas sob correlação.

O percentual médio de explicação dos cenários simulados $H_0{-}H_0$, $H_0{-}H_1$ e $H_1 2s{-}H_1$ que foram os com menor índice dos seis cenários simulados, permanece constante com o aumento do nível de correlação e do número de variáveis correlacionadas, não ultrapassando a 60%. Entretanto os outros cenários apresentam diminuição, aproximadamente 10%, da informação com o aumento do nível de correlação. Com o aumento do número de variáveis correlacionadas, a retração da explicação do maior autovalor chega a aproximadamente 20% para o cenário $H_1 2s{-}H_0$ (Figura 6).

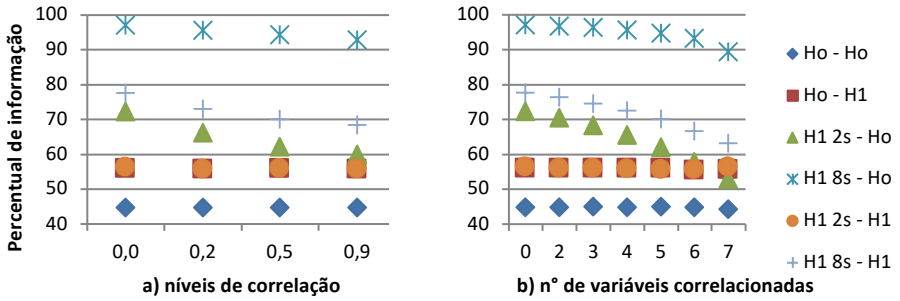


Figura 6 - Representação gráfica do comportamento do percentual de informação (PI) estimado para os dados transformados por meio da FDF, nos diferentes cenários e para a) os níveis de correlação (0,0; 0,2; 0,5 e 0,9) e b) o número de características correlacionadas (0, 2, 3, 4, 5, 6, 7).

Ao estudar a normalidade dos erros sob correlação, quer em níveis quer em número de variáveis correlacionada, verificou-se que o intervalo de aceitação da normalidade fica entre 96,5 e 97,5 % (Figura 7).

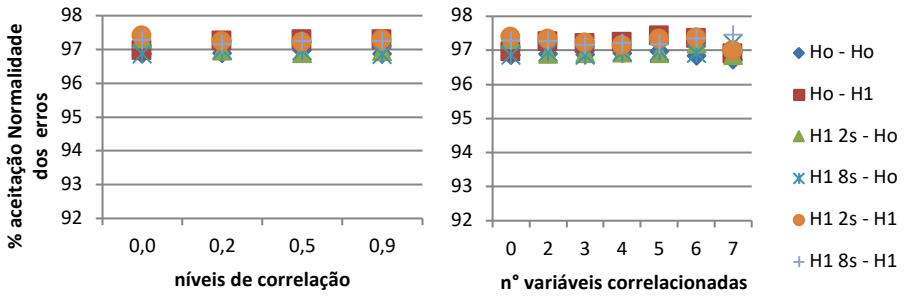


Figura 7 - Representação gráfica do comportamento da porcentagem de aceitação ($p<0,05$), da normalidade dos erros (teste Shapiro-Wilk), estimado para os dados transformados por meio da FDF nos níveis de correlação (0,0; 0,2; 0,5 e 0,9) e no número de características correlacionadas (0, 2, 3, 4, 5, 6, 7) para os diferentes cenários.

A aceitação da homogeneidade de variâncias dos dados transformados por FDF separou os cenários gerados homocedásticos dos heterocedásticos; conforme Figura 8. Com o aumento do nível de correlação e do número de variáveis há estabilização dos grupos, com tendência constante. E a homogeneidade de variâncias dos dados transformados pela FDF também separa os cenários gerados, sendo que os homocedásticos ficam constantes, próximos a 97% de aceitação da hipótese nula, e os heterocedásticos, entre 92 e 94%.

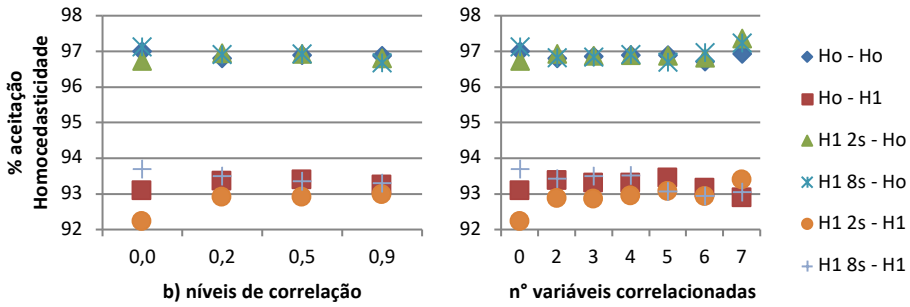


Figura 8 - Representação gráfica do comportamento da porcentagem de aceitação ($p<0,05$), da homocedasticidade de variâncias (teste Bartlett), estimados para os dados transformados por meio da FDF nos níveis de correlação (0,0; 0,2; 0,5 e 0,9) e no número de características correlacionadas (0, 2, 3, 4, 5, 6, 7) para os diferentes cenários.

De modo geral percebeu-se que, mesmo o teste F e os pressupostos da normalidade dos erros e homogeneidade de variâncias dos dados transformados por meio da FDF não terem sido influenciados pelo nível de correlação e pelo número de variáveis correlacionadas, a análise pela FDF apresentou queda no percentual de explicação com seus aumentos.

## Conclusões

A utilização da técnica de transformação de dados multivariados por meio da função discriminante de Fisher (FDF) em dados univariados mostrou-se como alternativa viável a ser aplicada na análise de experimentos.

A análise de variância dos dados transformados por meio da FDF detecta diferenças compatíveis com a análise de variância multivariada, destacando-se pela facilidade do processo de decisão.

Os elevados valores de aceitação das hipóteses nulas dos pressupostos da normalidade dos erros e da homogeneidade de variâncias dos dados transformados por meio da FDF, mesmo para os casos simulados sob heterocedasticidade, são indicativos da eficiência da FDF ao lidar com dados que violam tais pressupostos.

## Agradecimentos

Ao Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – IFSULDEMINAS e ao CNPq pelo apoio. Aos revisores e editores pelos comentários e sugestões.

CAMPOS, K. A.; MORAIS, A. R.; PAIXÃO, C. A. Feasibility of Fisher's discriminating function: comparison with MANAVA. *Rev. Bras. Biom*. Lavras, v.38, n.2, p.159-184, 2020.

- *ABSTRACT: The multivariate analysis methods allow the simultaneous study when several variable responses are obtained by plot. An option for the treatment of multivariate data is the transformation, using Fisher's linear discriminant function (FDF). After reduction of the p-dimensional to the unidimensional space, the univariate analysis of variance (ANOVA) is applied. The objectives of this paper were to evaluate the transformation efficiency of the multivariate data through the FDF and to compare the detection capacity of differences between treatments by the ANOVA of these data with the results obtained by means of the multivariate analysis of variance (MANOVA). Simulations were carried out to evaluate the acceptance rates of the null hypotheses for treatments, in four levels of correlations, equality of averages and variances for treatments and inequality between averages and variances. It was applied ANOVA, MANOVA and ANOVA of FDF to the values of these simulations. The results of the simulations indicate that FDF is a proper alternative for data evaluation.*

- *KEYWORDS: Simulation; correlation; data transformation; space reduction; analysis of variance.*

## Referências

BARROSO, L. P.; ARTES, R. Análise multivariada. In: Simpósio de Estatística Aplicada a Experimentação Agronômica, 10.; Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, 48., 2003, Lavras. *Anais*... Lavras: UFLA, 2003. 1 CD-ROM.

BARTLETT, M. S. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A*, v.160, n.2, p.268-282, 1937.

BEZERRA NETO, F. *et al*. Desempenho de sistemas consorciados de cenoura e alface avaliados através de métodos uni e multivariados. *Horticultura Brasileira*, v.25, n.4, p.514-520, 2007.

CAMPOS, K. A., PAIXÃO, C. A., MORAIS, A. R. Alternative for evaluation of coffee seedlings using Fisher's discriminant analysis. *Revista Ciência Agronômica*, v.47, p.299, 2016.

CARNEIRO, P. L. S. et al. Estudo de populações de ovinos Santa Inês utilizando técnicas de análise multivariada. *Revista Científica de Produção Animal*, v.8, n.1, p.40-50, 2006.

CHATFIELD, C.; COLLINS, A. J. *Introduction to multivariate analysis*. Gembloux: Presses Agroomiques, 1980. 362 p.

FERREIRA, D. F. *Estatística multivariada*. 3.ed. Lavras: Editora UFLA, 2018. 624 p.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v.7, n.2, p.179-188, 1936.

FONSECA, J. R.; SILVA, H. T. Identificação de duplicidades de acessos de feijão por meio de técnicas multivariadas. *Pesquisa Agropecuária Brasileira*, v.34, n.3, p.409-414, 1999.

FONSECA, R. et al. Avaliação de frangos de corte utilizando técnicas de análise multivariada: I., características de carcaça. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, v.54, n.5, p.525-529, 2002.

HAASE, R. F.; ELLIS, M. V. Multivariate analysis of variance. *Journal of Counseling Psychology*, Washington, v.34, n.4, p.404-413, 1987.

HAIR, J. F. et al. *Análise multivariada de dados*. 6.ed. Porto Alegre: Bookman, 2009. 688 p.

LEDO, C. A. S.; FERREIRA, D. F.; RAMALHO, M. A. P. Análise de variância multivariada para cruzamentos dialéticos. *Ciência e Agrotecnologia*, v.27, n.6, p.1214-1221, 2003.

MANLY, B. J. F.; ALBERTO, J.A.N. *Métodos estatísticos multivariados: uma introdução*. 4.ed. Porto Alegre: Bookman, 2019. 254 p.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada:* uma abordagem. Belo Horizonte: UFMG, 2007. 297p.

PADOVANI, C. R. P.; ARAGON, F. F. Programa computacional para método de discriminante de Fisher. *Revista Energia na Agricultura*, v.20, n.1, p.1-10, 2005.

PIMENTEL GOMES, F. *Curso de estatística experimental*. 15.ed. Piracicaba: FEALQ, 2009. 451 p.

R CORE TEAM. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2019. Disponível em: <http://www.R-project.org>. Acesso em: 12 abr. 2019.

SANTANA, S. L. A. *et al*. Fertilização foliar em mudas de cafeeiro com organominerais líquidos. *Tecnologia & Ciência Agropecuária*, v.5, n.3, p 9-13, 2011.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normatily: complete samples. *Biometrika*, v.52, n.3/4, p.591-611, 1965.

SILVA, E. A. et al. Desenvolvimento de mudas de cafeeiro *Coffea arabica* L sob diferentes composições de substratos. *Enciclopédia Biosfera*, v.8, n.14, p.337-346, 2012.

SIMEÃO, S. F. A. P.; PADOVANI, C. R. Utilização da função discriminante quadrática em ciências experimentais. *Revista Energia na Agricultura*, v.23, n.1, p.116-134, 2008.

TORRES FILHO, R. A. *et al*. Estudo da divergência genética entre linhas de suínos utilizando técnicas de analise multivariada. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, v.57, n.3, p.390-395, 2005.

Recebido em 03.06.2019

Aprovado após revisão em 21.11.2019

## ANEXO A: Script para FDF

```
## o exemplo citado é de um fatorial 2X5 com três blocos, para fins de modelo vou tratá-lo como um experimento com 10 tratamentos.

## introduzir os graus de liberdade dos tratamentos (h); blocos(s) e resíduos (r)

vh=9;vs=2;vr=18

## mostrar o arquivo de dados e as variáveis respostas

total<- read.table("dados.txt", h=T);attach(total);

s<-factor(subst);o<- factor(org);bl<- factor(bloco);a1<-org;a2<-org^2;

tr<-factor(trat)

## introduzir as matrizes, utilizou-se um arquivo para cada um: total (com trat, combinação dos fatores; subst, 2 níveis; org, 5 doses; blocos e sete variáveis respostas: ALT,DIAM,RAIZ,MSPA,MSR,AREA,NFOLHA), tratamentos (com os fatores em estudo, sem considerar o fatorial) e blocos (com a soma das variáveis por bloco).

t7=cbind(ALT,DIAM,RAIZ,MSPA,MSR,AREA,NFOLHA);detach(total)

tratamentos<- read.table("fatorial.txt",h=T); attach(tratamentos); h7=cbind(ALT,DIAM,RAIZ,MSPA,MSR,AREA,NFOLHA);detach(tratamentos)

blocos<- read.table("bloco.txt", h=T);attach(blocos); b7=cbind(ALT,DIAM,RAIZ,MSPA,MSR,AREA,NFOLHA);detach(blocos)

## função que calcula a matriz de soma de quadrados e de produtos

sqp=function(x){n=nrow(x);p=ncol(x);xb=x[1,];W=matrix(0,p,p);for(ii in 2:n){aux=x[ii,]-xb;W=W+(ii-1)*aux%*%t(aux)/ii;xb=xb+aux/ii};S=W/(n-1);list(vetmedia=xb,covariancia=S,SQP=W,COR=cor(x))}

## anexar as SQP e calcular matriz de resíduos ®

T<-sqp(t7)$SQP;S<-(sqp(b7)$SQP)/(vh+1);H<-(sqp(h7)$SQP)/(vs+1);R<-T-S-H

##calcular o máximo autovalor e autovetor

p=r=ncol(H);teta=eigen(solve(R)%*%H)

## % de explicação

l7<-max(Re(teta$values))/sum(Re(teta$values));l7

## encontrar a nova variável transformada pela FDF

Fisher7<-rbind(Re(teta$vectors[,1]));FDF7=t7%*%t(Fisher7);Fisher7;FDF7

## Proceder a ANAVA normalmente depois da transformação.

FDF.7<-aov(FDF7~tr+bl);

anova(FDF.7)

res.FDF.7<-resid(FDF.7);
```

```
W7=shapiro.test(res.FDF.7);W7
B7=bartlett.test(res.FDF.7,tr);B7
```