

**UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS RURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA**

**PRESSUPOSTO DA NORMALIDADE
MULTIVARIADA PARA O TESTE DE RAZÃO DE
VEROSSIMILHANÇA ENTRE DOIS GRUPOS DE
CARACTERES DE MAMONEIRA**

TESE DE DOUTORADO

Betania Brum

Santa Maria, RS, Brasil.

2012

**PRESSUPOSTO DA NORMALIDADE MULTIVARIADA PARA
O TESTE DE RAZÃO DE VEROSSIMILHANÇA ENTRE DOIS
GRUPOS DE CARACTERES DE MAMONEIRA**

Betania Brum

Tese apresentada ao Curso de Doutorado do Programa de Pós-Graduação
em Agronomia, Área de concentração em Produção Vegetal, da Universidade
Federal de Santa Maria (UFSM, RS), como requisito parcial
para obtenção do grau de
Doutor em Agronomia

Orientador: Prof. Sidinei José Lopes

Santa Maria, RS, Brasil.

2012

B893p Betania Brum

Pressuposto da normalidade multivariada para o teste de razão de verossimilhança entre dois grupos de caracteres de mamoneira / por Betania Brum. – 2012.

116 p. : il. ; 30 cm

Orientador: Sidinei José Lopes.

Tese (doutorado) – Universidade Federal de Santa Maria, Centro de Ciências Rurais, Programa de Pós-Graduação em Agronomia, RS, 2012

1. Mamoeira 2. Razão de verossimilhança 3. Poder do teste 4. Distribuição normal multivariada 5. Distribuição normal multivariada contaminada I. Lopes, Sidinei José II. Título.

CDU 634.651

Ficha catalográfica elaborada por Simone G. Maisonave – CRB 10/1733
Biblioteca Central da UFSM

© 2012

Todos os direitos autorais reservados a Betania Brum. A reprodução de partes ou do todo deste trabalho só poderá ser feita com autorização por escrito do autor.

Endereço: Rua General Câmara, nº 740, Bairro Centro, São Gabriel, RS, 97300-000
Fone (0xx)55 99352559; End. Eletr: bbufsm@gmail.com

**Universidade Federal de Santa Maria
Centro de Ciências Rurais
Programa de Pós-Graduação em Agronomia**

A Comissão Examinadora, abaixo assinada,
aprova a Tese de Doutorado

**PRESSUPOSTO DA NORMALIDADE MULTIVARIADA PARA O
TESTE DE RAZÃO DE VEROSSIMILHANÇA ENTRE DOIS GRUPOS
DE CARACTERES DE MAMONEIRA**

elaborada por
Betania Brum

como requisito parcial para obtenção do grau de
Doutor em Agronomia

COMISSÃO EXAMINADORA:

Sidinei José Lopes, Dr. (UFSM)
(Presidente/Orientador)

Lindolfo Storck, Dr. (UFSM)

Alberto Cargnelutti Filho, Dr. (UFSM)

Paulo Henrique de Oliveira, Dr. (UTFPR)

José Pedro Pereira Trindade, Dr. (EMBRAPA PECUÁRIA SUL)

Santa Maria, 29 de fevereiro de 2012.

*Aos meus amados
pais, Ildo e Terezinha*

Dedico este trabalho

AGRADECIMENTOS

À Deus, pelo maravilhoso dom da vida.

À Universidade Federal de Santa Maria e ao Programa de Pós-Graduação em Agronomia pela oportunidade de realização do curso de doutorado.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão da bolsa de doutorado.

À minha família, pelo sempre apoio e ajuda incondicional, principalmente aos meus pais, Ildo e Terezinha, exemplos de honestidade e perseverança, meus ídolos. Pai e mãe, obrigada pelo incentivo, amor e compreensão de vocês em mais uma etapa importante da minha vida.

Ao meu noivo André, agradeço o amor e a compreensão na realização deste trabalho, bem como, o apoio, o incentivo e o bom humor sempre presentes nos momentos em que tudo parecia impossível. André, obrigada pelas dicas “únicas” de Excel, que tornaram meu trabalho mais rápido e eficiente.

Ao meu orientador e grande amigo, Professor Sidinei, pelos preciosos ensinamentos, pelo carinho, apoio e amizade durante os nove anos de orientação. Durante a realização deste trabalho, agradeço tua disponibilidade, sabedoria, atenção, paciência e incentivo. Muito obrigada por tudo!

Aos meus co-orientadores, professores, Lindolfo Storck e Alberto Cargnelutti Filho, agradeço a disponibilidade, a atenção e as dicas valiosíssimas na concepção e andamento deste trabalho. Ao professor Alberto, agradeço ainda o auxílio na programação do R. Ao professor Storck, agradeço também o incentivo, apoio e a amizade durante os nove anos de convivência junto ao setor de Experimentação Agrícola. Professor, obrigada por compartilhar de forma tão clara e eficiente os teus conhecimentos.

Ao Professor Daniel Furtado Ferreira, pelas dicas importantíssimas que deram origem a este trabalho e pelo empréstimo da função `mvrnormCT`

do programa R. Professor, muito obrigada pela tua atenção e disponibilidade nas respostas das minhas dúvidas via email, as quais possibilitaram o andamento e realização deste trabalho.

Aos demais professores do Programa de Pós-Graduação em Agronomia, pela oportunidade de convívio e pelos ensinamentos.

Agradecimento especial, as minhas amigas Caroline Almeida Gulart e Tatiani Reis da Silveira pela parceria em todos os momentos.

Enfim, a todos que estiveram presentes direta ou indiretamente nesta fase de minha vida e que contribuíram para a realização deste trabalho:

Muito Obrigada!

*“Uma jornada de mil quilômetros
começa com um único passo”*

(Lao Tsé)

RESUMO

Tese de Doutorado
Programa de Pós-Graduação em Agronomia
Universidade Federal de Santa Maria

PRESSUPOSTO DA NORMALIDADE MULTIVARIADA PARA O TESTE DE RAZÃO DE VEROSSIMILHANÇA ENTRE DOIS GRUPOS DE CARACTERES DE MAMONEIRA

AUTORA: BETANIA BRUM

ORIENTADOR: Prof. Dr. SIDINEI JOSÉ LOPES

Local e Data da Defesa: Santa Maria, 29 de fevereiro de 2012.

O teste de razão de verossimilhança para a independência entre dois grupos de variáveis permite identificar se há ou não relação de dependência entre dois grupos de variáveis, ou seja, se as covariâncias entre os dois grupos são nulas. Esse teste pressupõe normalidade multivariada dos dados, o que limita sua aplicação, em muitos estudos da área agrônoma, em que se necessita utilizar, por exemplo, a análise de correlação canônica. O objetivo deste trabalho é avaliar o erro tipo I e o poder do teste de razão de verossimilhança (TRV) para independência entre dois grupos de variáveis em diversos cenários, constituídos pelas combinações de: 16 tamanhos de amostra; 40 combinações de número de variáveis dos dois grupos; e, nove graus de correlação entre as variáveis (para o poder); sob distribuição normal multivariada e distribuição normal multivariada contaminada, bem como, comparar, nos diferentes cenários, duas fórmulas para cálculo da estatística do teste. Dessa forma, foram avaliados o efeito de 640 e 5760 cenários sobre as taxas de erro tipo I e poder, respectivamente, em cada uma das distribuições de probabilidade e fórmulas. A avaliação do desempenho do TRV foi realizada por meio de simulação computacional pelo método Monte Carlo, utilizando-se 2000 simulações em cada um dos cenários gerados. Em situação de normalidade multivariada: quando o número de variáveis é grande ($p= 24$), o TRV para a independência entre dois grupos de variáveis, controla as taxas de erro tipo I e apresentou poder elevado, em tamanhos de amostra superiores a 100 e 500, com uso das fórmulas para pequenas e grandes amostras, respectivamente; e, para tamanhos amostrais pequenos ($n= 25, 30$ e 50), o teste apresenta bom desempenho, desde que, o número de variáveis não exceda a 12; e, a fórmula escolhida, seja para pequenas amostras. Sob distribuição normal multivariada contaminada, o TRV para a independência entre dois grupos de variáveis possui elevados valores de poder, mas não é robusto, pois apresenta elevadas taxas de erro tipo I, em qualquer cenário avaliado.

Palavras-chave: Erro tipo I. Poder do teste. Distribuição normal multivariada contaminada.

ABSTRACT

Master thesis
Pos-Graduation Program in
Federal University of Santa Maria

ASSUMPTION OF MULTIVARIATE NORMALITY FOR THE LIKELIHOOD RATIO TEST BETWEEN TWO GROUPS OF CHARACTERS OF CASTOR BEANS

AUTHOR: BETANIA BRUM

ADVISER: Prof. Dr. SIDINEI JOSÉ LOPES

Place and Date of the defense: Santa Maria, 29th February, 2012.

The likelihood ratio test for independence between two groups of variables allows us to identify whether there is a dependency relationship between two groups of variables, ie, if the covariance between the two groups are zero. This test assumes normality multivariate data, which limits its application, in many studies of agronomic area, times when you need use, for example, the canonical correlation analysis. The objective of this study is to evaluate the type I error and power of the likelihood ratio test (LRT) for independence between two groups of variables in different scenarios, consisting of combinations of: sample sizes 16, 40 number of combinations of two variables groups, and nine degrees of correlation between variables in matrices (for power); multivariate normal distribution under normal and contaminated, as well as compare the different scenarios, two formulas for calculating the test statistic. Thus, were evaluated the effect of 640 and 5760 scenarios on rates of type I error and power, respectively, in each one of probability distributions and formulas. The assessment of performance of LRT was performed through computational simulation by Monte Carlo method, using 2000 simulations in each of the scenarios generated. In multivariate normal situation: when the number of variables is large ($p = 24$), the LRT for independence between two groups of variables, controls the type I error rates and has high power in sample sizes greater than 100 and 500, with use of formulas for small and large samples, respectively; and, for sample sizes small ($n = 25, 30$ and 50), the test presents good performance, provided that, the number of variables does not exceed to 12; and, the formula chosen, whether for small samples. Under contaminated multivariate normal distribution, the LRT for independence between two groups of variables have high values of power, but is not robust, because it has high rates of type I error in any scenario evaluated.

Key words: Type I error rate. Power of the test. Contaminated multivariate normal.

LISTA DE TABELAS

- Tabela 1 - Variância e média de variáveis de planta adulta e de produção de grãos de mamoneira, em 50 plantas do híbrido Lyra. Santa Maria, 2011..... 36
- Tabela 2 – Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com três variáveis no grupo I e três e quatro variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações..... 53
- Tabela 3 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com três variáveis no grupo I, cinco e seis variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações..... 54
- Tabela 4 – Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com três variáveis no grupo I, sete e oito variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações..... 55
- Tabela 5 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com três variáveis no grupo I; nove e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras) em 2000 simulações..... 56
- Tabela 6 – Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com quatro variáveis no grupo I; e quatro e cinco variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações..... 57
- Tabela 7 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com quatro variáveis no grupo I; seis e sete variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações..... 58
- Tabela 8 – Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com quatro variáveis no grupo I; e oito e nove variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações..... 59
- Tabela 9 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com quatro e cinco variáveis no grupo I; e 10 e cinco variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações..... 60

Tabela 10 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com cinco variáveis no grupo I; seis e sete variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	61
Tabela 11 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com cinco variáveis no grupo I; oito e nove variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	62
Tabela 12 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com cinco e seis variáveis no grupo I; 10 e seis variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	63
Tabela 13 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com seis variáveis no grupo I; sete e oito variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	64
Tabela 14 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com seis variáveis no grupo I; nove e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	65
Tabela 15 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com sete variáveis no grupo I; sete e oito variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	66
Tabela 16 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com sete variáveis no grupo I; nove e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	67
Tabela 17 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com oito variáveis no grupo I; oito e nove variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	68
Tabela 18 – Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com oito variáveis no grupo I; 10 e nove variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	69

Tabela 19 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com nove e 10 variáveis no grupo I; 10 e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.....	70
Tabela 20 – Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com 11 e 12 variáveis no grupo I; 10 e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	71
Tabela 21 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com 13 e 14 variáveis no grupo I; 10 e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	72
Tabela 22 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com três variáveis no grupo I e três a cinco variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.....	79
Tabela 23 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 3 variáveis no grupo I e 6 a 8 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.....	80
Tabela 24 – Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 4 variáveis no grupo I e 4 a 7 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.....	81
Tabela 25 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 4 variáveis no grupo I e 8 a 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.....	82
Tabela 26 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 5 variáveis no grupo I e 5 a 7 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.....	83
Tabela 27 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 5 variáveis no grupo I e 8 a 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.....	84
Tabela 28 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 6 variáveis no	

grupo I e 6 a 8 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	85
Tabela 29 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 6 e 7 variáveis no grupo I e 7, 9 e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	86
Tabela 30 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 7 variáveis no grupo I e 8 a 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	87
Tabela 31 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 8 variáveis no grupo I e 8 a 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	88
Tabela 32 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 9 e 10 variáveis no grupo I e 9 e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	89
Tabela 33 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 11, 12 e 13 variáveis no grupo I e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	90
Tabela 34 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 14 variáveis no grupo I e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.	91

LISTA DE FIGURAS

- Figura 1 - Distribuição normal multivariada de seis variáveis (representando o comportamento de seis variáveis, duas a duas) para $n = 2000$ 30
- Figura 2 - Distribuição normal multivariada contaminada de seis variáveis (representando o comportamento de seis variáveis, duas a duas), para $n = 2000$, com grau de contaminação igual a 20%. 32
- Figura 3 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha = 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n = 25$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $p = 0,05$ 44
- Figura 4 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha = 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n = 30$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $p = 0,05$ 44
- Figura 5 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha = 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n = 50$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $p = 0,05$ 45
- Figura 6 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha = 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n = 100$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $p = 0,05$ 46
- Figura 7 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha = 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n = 200$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $p = 0,05$ 46
- Figura 8 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha = 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n = 300$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $p = 0,05$ 47
- Figura 9 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha = 5\%$), com diferentes números de variáveis e duas fórmulas,

para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n= 400$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $p = 0,05$ 47

Figura 10 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis (combinações de variáveis dos dois grupos) e duas fórmulas: para pequenas e grandes amostras, no tamanho de amostra ($n= 25$), considerando a distribuição normal multivariada contaminada, em 2000 simulações. 74

Figura 11 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis (combinações de variáveis dos dois grupos) e duas fórmulas: para pequenas e grandes amostras, no tamanho de amostra ($n= 30$), considerando a distribuição normal multivariada contaminada, em 2000 simulações. 74

Figura 12 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis (combinações de variáveis dos dois grupos) e duas fórmulas: para pequenas e grandes amostras, no tamanho de amostra ($n= 50$), considerando a distribuição normal multivariada contaminada, em 2000 simulações. 75

SUMÁRIO

1	INTRODUÇÃO.....	17
2	REVISÃO BIBLIOGRÁFICA	21
2.1	Testes de hipóteses: elementos e tipos de erro.....	21
2.2	Poder do teste	22
2.3	Testes de hipóteses baseados na razão de verossimilhança	22
2.4	Teste de razão de verossimilhança para a independência entre dois grupos de variáveis	26
2.5	Teste de razão de verossimilhança para a independência entre dois grupos de variáveis e a análise de correlação canônica	27
2.6	Distribuição normal multivariada e normal multivariada contaminada.....	29
2.7	O teste de razão de verossimilhança e a normalidade multivariada.....	32
3	MATERIAL E MÉTODOS	35
3.1	Banco de dados.....	35
3.2	Obtenção das amostras normais multivariadas.....	36
3.2.1	Erro tipo I.....	36
3.2.1.1	Matrizes de covariância e vetores de médias	36
3.2.1.2	TRV para a independência entre dois grupos de variáveis	38
3.2.1.3	Cenários e simulação Monte Carlo	38
3.2.2	Poder	39
3.2.2.1	Matrizes de covariância e vetores de médias	39
3.2.2.2	Cenários e simulação Monte Carlo	40
3.3	Obtenção das amostras normais multivariadas contaminadas.....	41
3.4	Obtenção das matrizes de covariância e simulação no programa R.....	41
4	RESULTADOS E DISCUSSÃO.....	43
4.1	Distribuição normal multivariada.....	43
4.1.1	Erro tipo I.....	43
4.1.2	Poder	50
4.1.3	Recomendações do TRV para a independência entre dois grupos de variáveis, sob distribuição normal multivariada.....	52
4.2	Distribuição normal multivariada contaminada	73
4.2.1	Erro tipo I.....	73
4.2.2	Poder	77
5	CONCLUSÕES.....	93
6	CONSIDERAÇÕES FINAIS.....	95
	REFERÊNCIAS.....	97

1 INTRODUÇÃO

O estudo de fenômenos biológicos, dos quais fazem parte as ciências agrárias é complexo, ou seja, influenciado por uma série de fatores, tais como: solo, variáveis meteorológicas, pragas, doenças, os quais, normalmente, apresentam relação entre si. Por isso, em algumas situações não é suficiente avaliar o comportamento isolado de determinada variável para representar um fenômeno, mas considerar as interações e o contexto em que está inserido. Para tal, há a necessidade de utilização de métodos de análise multivariada (MAM). Os MAM consistem em técnicas estatísticas nas quais se considera o estudo de múltiplas variáveis, medidas simultaneamente no mesmo objeto ou elemento amostral.

Nos últimos anos, com o advento da informática e aperfeiçoamento de aplicativos estatísticos com módulos de análise multivariada, esses métodos têm sido muito utilizados, nas mais diversas áreas do conhecimento, das quais se destacam: educação, geologia, química, física, ergonomia e biológicas (MINGOTI, 2007).

Apesar de tamanha utilidade, as análises multivariadas ainda são pouco exploradas, o que ocorre por falta de conhecimento adequado sobre o assunto e sobre sua potencialidade. Além disso, assim como na análise univariada, o pesquisador precisa levar em consideração a qualidade dos dados obtidos e a representatividade da população, bem como, o atendimento às pressuposições de cada método.

Embora, as análises multivariadas ainda sejam pouco exploradas nas ciências agrárias, um exemplo de ferramenta multivariada que tem se popularizado é a análise de correlação canônica, que é uma técnica estatística de interdependência para avaliar as associações lineares entre dois grupos de variáveis. As correlações canônicas referem-se às correlações entre variáveis canônicas, ou seja, entre combinações lineares de variáveis do grupo 1 e do grupo 2, utilizando-se como coeficientes dessas combinações, valores que maximizem a correlação entre os dois conjuntos de variáveis. Dessa forma, ao avaliar as relações entre dois grupos de variáveis X e Y tem-se: $X = [X_1, X_2, \dots, X_p]$, vetor de medidas de p caracteres que constituem o grupo 1; e, $Y = [Y_1, Y_2, \dots, Y_q]$: vetor de medidas de q caracteres que constituem o grupo 2 (CRUZ; REGAZZI, 1997). Portanto, essa análise mede o grau de associação que existe entre dois conjuntos de variáveis.

Antes de realizar esta análise é importante verificar se os vetores X e Y são independentes entre si, ou não correlacionados; pois, caso isso ocorra, a análise de correlação

canônica perde sua utilidade, por que a correlação entre todas as combinações lineares $a_k'X$ e $b_k'X$ será igual a zero (MINGOTI, 2007). Para testar esta situação, deve-se realizar uma análise da matriz de covariâncias ou de correlações cruzadas, para verificar se essas são próximas ou não da matriz nula (MINGOTI, 2007).

Esta condição de independência entre os dois grupos de variáveis, pode ser verificada pelo teste de razão de verossimilhança para a independência entre dois grupos de variáveis, ou seja, se as covariâncias entre os dois grupos são nulas (FERREIRA, 2008).

Esse teste, como parte das técnicas de inferência multivariada, pressupõe que as variáveis observadas apresentem distribuição normal multivariada, o que poderia limitar sua aplicação, em estudos da área agrônômica, nos quais se dispõe de dados que, muitas vezes, não seguem a distribuição normal multivariada e para os quais não existe transformação adequada.

Apesar da importância do teste de razão de verossimilhança da independência entre dois grupos de variáveis, as informações disponíveis na literatura são escassas e não foram encontrados trabalhos de avaliação da robustez deste teste, em caso de violação do pressuposto de normalidade multivariada.

Existem vários relatos sobre o fraco desempenho de outros TRV, quando se utilizam tamanhos de amostra pequenos ($n \leq 50$) e/ou quando estes são combinados ao aumento do número de variáveis (HAIR et al., 2009; SILVA et al., 2008; PINTO, 2009; RIBEIRO, 2010). Em relação ao grau de correlação entre as variáveis, alguns TRV para matrizes de covariâncias apresentam maior poder quando em presença de maior correlação (PINTO, 2009; ALISSON et al., 2009); enquanto, no teste de Bartlett multivariado, Silva et al. (2008) relatam ausência de efeito do grau de correlação entre as variáveis no erro tipo I e poder do teste.

Apesar da grande utilidade dos testes de razão de verossimilhança, há relatos, em que este possui desempenho ruim quando os dados não seguem a distribuição normal; bem como, quando tamanho de amostra é pequeno e/ou quando a relação tamanho de amostra e número de variáveis é pequena. Portanto, trabalhos que avaliem a qualidade desses testes, em tais situações, são importantes para que o pesquisador não incorra em erros ao aplicá-los.

O objetivo deste trabalho é avaliar o erro tipo I e o poder do teste de razão de verossimilhança (TRV) para independência entre dois grupos de variáveis em diversos cenários, constituídos pelas combinações de: 16 tamanhos de amostra; 40 combinações de número de variáveis dos dois grupos; e, nove graus de correlação entre as variáveis nas matrizes de covariância (para o poder), sob distribuição normal multivariada e normal

multivariada contaminada, bem como, comparar, nos diferentes cenários, duas fórmulas para cálculo da estatística do TRV.

2 REVISÃO BIBLIOGRÁFICA

2.1 Testes de hipóteses: elementos e tipos de erro

Teste de hipótese é uma ferramenta utilizada em inferência estatística e consiste em uma regra de decisão a respeito de um parâmetro populacional, estimado por uma amostra, que será avaliado por meio de uma hipótese, a qual pode ser aceita ou rejeitada (OLIVEIRA et al., 2009).

Em estatística, uma hipótese é uma afirmação sobre um parâmetro da população ou sobre a natureza da distribuição de probabilidade de uma variável aleatória (MOOD et al., 1974). Os testes estatísticos admitem dois tipos de hipóteses: a hipótese nula (H_0) e a hipótese alternativa (H_1). A hipótese nula é uma afirmação que diz que o parâmetro de interesse é tal qual especificado, enquanto a hipótese alternativa é declaração que difere da nula, ou seja, o parâmetro é maior, menor ou diferente do que o especificado em H_0 (OLIVEIRA et al., 2009).

As hipóteses H_1 e H_0 são mutuamente excludentes, aceitando-se uma, automaticamente, rejeita-se a outra. Por isso, é necessário tomar cuidado para minimizar os erros durante a realização de um teste de hipóteses (FONSECA; MARTINS, 2009).

Ao se realizar um teste de hipótese, podem-se cometer dois tipos de erros: o erro tipo I e o erro tipo II. O erro tipo I consiste em rejeitar a hipótese nula quando ela é verdadeira, ou seja, quando ela não deveria ser rejeitada. Portanto, é uma afirmativa de que os resultados são estatisticamente significativos, quando na realidade, eles não são. A probabilidade de se cometer o erro tipo I é representada por α (alfa) e é denominada nível de significância do teste (FONSECA; MARTINS, 2009). O outro erro que pode ser cometido em um teste de hipótese é o erro tipo II, que é o ato de se aceitar a hipótese nula quando ela não deveria ser aceita; e, a probabilidade de cometer o erro tipo II é conhecida por β .

Para que os testes de hipóteses apresentem resultados aceitáveis, deve-se minimizar as probabilidades de se cometer os erros tipo I e II, o que não é fácil, pois, para um tamanho de amostra fixo, a redução em um determinado tipo de erro, resultará em incremento no outro (SILVA et al., 2008). Portanto, a probabilidade de se incorrer em erro tipo I é inversamente proporcional à probabilidade de cometer o erro tipo II (BORGES; FERREIRA, 2003). Para reduzir, simultaneamente, os erros tipo I e II, pode-se aumentar o tamanho de amostra, o qual não depende do valor de significância adotado (CARMER, 1976).

2.2 Poder do teste

O poder de um teste de hipóteses é a probabilidade de se rejeitar a hipótese nula, quando, de fato, ela é falsa e deveria ser rejeitada, o que resulta em uma decisão correta. Quanto maior essa probabilidade, maior será o poder do teste. Essa probabilidade é conhecida como $1-\beta$ (HAIR et al., 2009).

Existem três fatores que influenciam o poder de um teste (HAIR et al., 2009): i) tamanho do efeito – é a magnitude do efeito que está sendo estudado, por exemplo: no teste de hipótese para independência entre grupos de variáveis, o tamanho do efeito é dado pelo grau de correlação entre as variáveis; quanto maior o grau de correlação, maior será o tamanho do efeito e mais fácil a detecção de efeitos significativos; ii) erro tipo I (α) - a redução no erro tipo I, ocasiona redução nos valores de poder, ou seja, quando o pesquisador diminui a probabilidade de detectar um efeito significativo incorretamente, a chance de um efeito correto também diminui; iii) Tamanho de amostra - amostras muito pequenas tornam difícil a identificação de efeitos significativos em testes de hipóteses, pois nesta situação o teste é pouco sensível. O aumento no tamanho de amostra resultará em aumento do poder do teste, considerando um nível alfa fixo. Por outro lado, amostras muito grandes podem aumentar muito o poder do teste, o que significa que até efeitos muito pequenos serão detectados, tornando praticamente todos os efeitos significativos.

O ideal em um teste de hipóteses é que rejeitasse com alta probabilidade (em 100% dos casos) a hipótese H_0 , quando ela é falsa (poder) e que não rejeitasse para nenhum dos casos a hipótese H_0 verdadeira (erro tipo I) (CANTELMO; FERREIRA, 2007). Porém, como na realidade não existe teste de hipóteses ideal, espera-se que um bom teste apresente taxas de erro tipo I iguais ao nível mínimo de significância (α) estabelecido pelo pesquisador; e, que, possua o maior poder possível, quando comparado a testes concorrentes (CANTELMO; FERREIRA, 2007).

2.3 Testes de hipóteses baseados na razão de verossimilhança

Uma maneira lógica para decidir se uma amostra é oriunda da população 1 ou da população 2 é optar pela população que tenha maior probabilidade de acontecer para aquele

determinado valor do parâmetro (MOOD et al., 1974). Segundo Veloso (2006), partindo dessa lógica, nasceu o método da razão de verossimilhança, desenvolvido por Neyman e Pearson em 1928, o qual é muito utilizado para obtenção de estatísticas de testes de hipóteses, tanto na área univariada, quanto na multivariada (BIASE, 2009).

O método da máxima verossimilhança possibilita a estimação de diferentes parâmetros de um modelo estatístico a partir de um conjunto de dados amostrados na população. Assumindo que determinada variável amostrada tem distribuição normal (modelo estatístico), com parâmetros (média e variância desconhecidas), pode-se estimar esses parâmetros pela máxima verossimilhança a partir da medição de uma amostra da população. O método ajusta valores para os parâmetros que maximizem a probabilidade dos dados amostrados (parâmetros que maximizem a função de verossimilhança) para o modelo proposto (PORTUGAL,1995).

Considerando y_1, y_2, \dots, y_n uma amostra aleatória independente e identicamente distribuída, com função de probabilidade ou função densidade de probabilidade dada por $f(y|\theta)$, na qual θ é o parâmetro da função; a função de verossimilhança de $y = (y_1, y_2, \dots, y_n)$ é dada por: $l(\theta; y) = f(y|\theta) = f(y_1, y_2, \dots, y_n|\theta)$. Toda a informação presente na amostra ou experimento é representada pela função de verossimilhança (PORTUGAL,1995; RESENDE, 2007).

Ao testar a hipótese H_0 , de que um parâmetro θ faça parte de algum subespaço \mathbb{R}^s , o qual é conhecido como conjunto nulo e expresso por $\Omega_0 \subset \mathbb{R}^s$. Esse subespaço, geralmente diz respeito às restrições impostas no espaço paramétrico; e nessa situação, a hipótese nula equivale ao espaço restrito e a hipótese alternativa ao espaço irrestrito (FERREIRA, 2008).

Considerando que a distribuição de y_j em \mathbb{R}^p dependa do parâmetro θ e as seguintes hipóteses: $H_0: \theta \in \Omega_0$ e $H_1: \theta \in \Omega$, em que H_0 pertence ao espaço restrito e H_1 ao espaço irrestrito, a estatística da razão de verossimilhança será (FERREIRA, 2008):

$$\Lambda = \frac{L\Omega_0(Y; \hat{\theta})}{L\Omega(Y; \hat{\theta})}$$

Em que: $L\Omega_0(Y; \hat{\theta})$ é o máximo da função de verossimilhança para o espaço restrito e $L\Omega(Y; \hat{\theta})$ é o máximo da função de verossimilhança para o espaço irrestrito.

Se a razão Λ for grande, provavelmente a hipótese H_0 não será rejeitada, caso contrário, a hipótese H_1 deverá ser escolhida (FERREIRA, 2008). Como é muito difícil obter a distribuição nula de Λ ; a estatística de teste passa a ser calculada por $-2\ln(\Lambda)$, que tem

distribuição assintoticamente qui-quadrado. Para maiores detalhes sobre os testes de razão de verossimilhança consultar Ferreira (2008).

Os testes de razão de verossimilhança são muito populares, apesar disso, existem vários exemplos de situações, em que estes apresentam um desempenho ruim (LEHMANN, 2006). Ao destacar as razões da popularidade do TRV, o autor também evidencia alguns problemas associados a cada uma delas:

- a) A probabilidade de que $L_x(\theta) = p_\theta(x)$ ou seja, a probabilidade de x , como uma função de θ , é considerado uma medida (relativa) de suporte, que a observação x fornece para o parâmetro θ , assim a razão de verossimilhança (Λ) fornece alternativas para melhor explicação da hipótese; o que, de acordo com o autor, parece ser persuasivo, mas não tem nenhuma fundamentação lógica muito forte.
- b) em muitos problemas padrão, o TRV corrobora com ensaios obtidos por outros princípios, normalmente, gerando testes satisfatórios, embora hajam contra exemplos (PERLAN; WU, 1999; MENÉNDEZ et al., 1992); no entanto, em muitos problemas padrão, nos quais, normalmente, há um conjunto completo de estatísticas, todos princípios levam a testes com pequenas diferenças ou iguais; e,
- c) em condições adequadas o TRV, possui boas propriedades assintóticas, o que pouco infere sobre pequenas amostras.

Destacando a falta de embasamento teórico e os muitos contra-exemplos do TRV, Lehmann (2006) sugere a geração de estudos do comportamento do TRV para pequenas amostras; e, destaca que essa sugestão já foi dada por Perlman e Wu (1999).

Existem diversos testes baseados na razão de verossimilhança que são utilizados para testar matrizes de covariância de populações normais. Esses testes são importantes na estatística multivariada, para que em algumas técnicas, tais como: componentes principais, análise de variância e correlação canônica, se obtenha maior eficácia (FERREIRA, 2008). O autor salienta que os testes de hipóteses para matrizes de covariância apresentam menor robustez, em relação aos testes sobre vetores de médias, por isso, desvios de normalidade comprometem muito o desempenho daqueles testes.

Existem diversos trabalhos avaliando a qualidade de testes de hipóteses para matrizes de covariância. Pinto (2009) faz recomendações em relação à qualidade e uso dos testes para verificar se uma covariância amostral (S) é igual a uma covariância teórica (Σ_0), ou seja: ($H_0: \Sigma_0 = S$); teste de razão de verossimilhança (exato e aproximado) (ALT, 1985); teste do determinante (ANDERSON, 2003); estudo sobre a distribuição exata do determinante da matriz de covariâncias amostral ($|S|$) (ANDERSON, 2003) e teste de Sullivan (SULLIVAN

et al., 2007); bem como, as seguintes modificações em testes pré-existentes: adaptação do teste de Sullivan; adaptação do teste de Hayter e Tsui (1994) para os autovalores de Σ ; adaptação do teste T^2 de Hotelling para os autovalores de Σ (ANDERSON, 2003); Teste dos autovalores usando o número de condição; e, transformação da informação contida na matriz de covariâncias para uma variável univariada (SILVA; MINGOTI, 2008). Dentre as várias propostas para trabalhos futuros, Pinto (2009) enfatiza a necessidade de estudar o efeito da não normalidade multivariada nos testes propostos.

Os testes para matrizes de covariâncias: da variância generalizada, de razão de verossimilhança, do *Stepdown* (SULLIVAN et al., 2007) e do teste VMAX (MACHADO et al. (2008) são construídos sob a hipótese de independência entre as observações. Ribeiro (2010) avaliou o comportamento destes testes, quando essa suposição é violada; e, propôs correções (ajuste de modelo de séries temporais multivariados) nestes testes, afim de que, detectem a autocorrelação das observações. Concluindo, o autor recomenda o uso do teste da razão de verossimilhança e o teste *Stepdown* proposto por Sullivan et al. (2007), pois estes apresentam um melhor desempenho.

Em análise de regressão múltipla, Carvalho (1998), comparando os valores de poder do teste de razão de máxima verossimilhança para os coeficientes de regressão com o teste t, observou que a estatística $-2 \ln(\lambda)$ do TRV apresenta maior probabilidade de não rejeitar H_0 , quando ela é falsa (poder), do que a estatística t. Por meio de estudos experimentais e de simulação, o autor concluiu que o teste da razão de máxima verossimilhança demonstrou grande poder, constituindo-se no melhor critério para testar a hipótese de nulidade sobre a igualdade de coeficientes de regressão.

Veloso (2006) utilizou simulação Monte Carlo para avaliar o desempenho do teste de identidade de modelos não lineares sob diferentes tamanhos amostrais, número de equações, robustez e diferentes modelos (Logístico, Gompertz e Langmuir), e obteve como resultados gerais, robustez do TRV em distribuições não normais, quando em tamanhos de amostra superior a 50 ($n= 50$ e 100). Essa robustez não foi verificada quando o tamanho de amostra era de 15 observações.

2.4 Teste de razão de verossimilhança para a independência entre dois grupos de variáveis

O TRV para a independência entre dois grupos de variáveis consiste em identificar se há ou não relação de dependência entre dois grupos de variáveis, ou seja, se as covariâncias entre os dois grupos são nulas ($\sum xy = 0$).

Quando os vetores X e Y são normais multivariados, a independência entre os dois grupos de variáveis pode ser avaliada pelo teste de razão de verossimilhança para a independência entre dois grupos de variáveis (JOHNSON; WICHERN, 2007).

Considerando o seguinte vetor aleatório normal:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \\ Y_1 \\ \vdots \\ Y_q \end{bmatrix} \text{ com média } (\mu) \text{ e covariância } (\Sigma) \text{ e covariância particionada dada por:}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \text{ são definidas as seguintes hipóteses (FERREIRA, 2008):}$$

$$H_0: \sum xy = 0 \text{ versus } H_1: \sum xy \neq 0,$$

A estatística qui-quadrado do TRV para a independência entre dois grupos de variáveis para grandes amostras é baseada no critério de Wilks: $\chi_c^2 = n \times \ln\left(\frac{|S_{XX}||S_{YY}|}{|S|}\right)$, com pxq graus de liberdade para grandes amostras, em que: n é o tamanho de amostra; \ln é o logaritmo natural; S_{XX} é a matriz de covariância do grupo de variáveis X; S_{YY} é a matriz de covariância do grupo de variáveis Y; e, S é a matriz de covariância dos grupos XY. Para pequenas amostras, essa estatística de teste foi corrigida por Bartlett, em 1937 (FERREIRA, 2008; JOHNSON; WICHERN, 2007 e MINGOTI, 2007): $\chi_c^2 = \left[n - 1 - \frac{1}{2}(p + q + 1)\right] \times \ln\left(\frac{|S_{XX}||S_{YY}|}{|S|}\right)$, com pxq graus de liberdade, na qual: p e q são os números de variáveis do grupo X e Y respectivamente.

O TRV compara a variância amostral generalizada sob H_0 : $\begin{vmatrix} S_{11} & 0 \\ 0 & S_{22} \end{vmatrix} = |S_{11}||S_{22}|$, com a variância irrestrita S .

Nesse teste, o valor da estatística calculada é comparada ao valor crítico tabelado da distribuição qui-quadrado, com $p \times q$ graus de liberdade e em nível α de significância escolhido para o teste; sempre que o valor calculado for maior do que o valor tabelado,

rejeita-se a hipótese nula e conclui-se que os dois grupos considerados no estudo não são independentes. Para maiores detalhes e exemplos de aplicação deste teste, consultar Ferreira (2008) e Mingoti (2007).

2.5 Teste de razão de verossimilhança para a independência entre dois grupos de variáveis e a análise de correlação canônica

A análise de correlação canônica é uma técnica estatística multivariada que possibilita avaliar as relações lineares existentes entre dois grupos ou conjuntos p e q , de variáveis X e Y . As correlações canônicas referem-se às correlações entre combinações lineares de variáveis, de modo que a correlação entre essas combinações seja máxima (CRUZ; REGAZZI, 1997).

A idéia é determinar as combinações lineares dos dois grupos que obtenham a máxima correlação. A seguir determina-se o segundo par de maior correlação e que seja não correlacionado com o primeiro. O procedimento segue até esgotarem-se as dimensões de ambos os grupos ou do menor deles (FERREIRA, 2008).

Para entender a relação do TRV para a independência entre dois grupos de variáveis, com a análise de correlação canônica, é necessário descrever brevemente a metodologia da análise de correlação canônica, que em sua totalidade, pode ser encontrada em Ferreira (2008), Mingoti (2007), Carvalho et al. (2004), Johnson e Wichern (2007) e Cruz e Regazzi (1997).

Considerando o seguinte vetor aleatório normal:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \\ Y_1 \\ \vdots \\ Y_q \end{bmatrix} \text{ com média } (\mu) \text{ e covariância } (\Sigma).$$

Assumindo-se $p \neq q$, a matriz de covariância é particionada e dada por:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

A covariância Σ_{12} representa as covariâncias entre pares de variáveis pertencentes

aos dois grupos (X e Y), o que significa que os p q elemento Σ_{12} medem a associação entre os dois grupos de variáveis. Caso os valores de p e q sejam grandes, a interpretação das covariâncias conjuntamente é complicada. Por isso, Hotelling (1935, 1936) propõe o estudo das relações entre os vetores X e Y, por meio de combinações lineares destes vetores (U e V), geradas de forma a estarem fortemente correlacionadas entre si, as quais chamou de correlações canônicas.

Sejam U_k e V_k as combinações lineares das variáveis X e Y, respectivamente: $U_k = a_k'X$ e $V_k = b_k'Y$, sendo que a_k e b_k são os vetores não nulos dessas combinações, escolhidos de forma que a correlação entre as variáveis U_k e V_k seja máxima e que as duas novas variáveis tenham variâncias iguais a 1.

Portanto, a correlação canônica entre as variáveis canônicas U_k e V_k é definida por:

$$\text{Corr}(U_k; V_k) = \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a} \sqrt{b' \Sigma_{22} b}}$$

Sendo:

$\text{Var}(U_k) = a' \Sigma_{11} a$, a variância das variáveis do grupo X;

$\text{Var}(V_k) = b' \Sigma_{22} b$, a variância das variáveis do grupo Y; e,

$\text{Cov}(U_k; V_k) = a' \Sigma_{12} b$, a covariância entre as variáveis do grupo X e Y.

Em cada estágio deste procedimento, é gerado um par canônico, ou seja, duas combinações lineares, uma referente às variáveis contidas no vetor X e outra referente às variáveis do vetor Y. O número de correlações canônicas é igual ao número de variáveis que constitui o menor grupo (p ou q) e sua magnitude sempre decresce com a ordem em que são estimados (CRUZ; REGAZZI, 1997).

O TRV para a independência entre dois grupos de variáveis é um procedimento muito útil e que deve ser realizado antes de se executar a análise de correlação canônica, pois é importante verificar se existe dependência ou relação entre os dois grupos de variáveis (MINGOTI, 2007; FERREIRA, 2008). Para que a análise de correlação canônica seja eficiente, os dois grupos de variáveis devem ser correlacionados, ou seja, a hipótese nula do TRV para a independência entre dois grupos de variáveis deve ser rejeitada. Caso os vetores X e Y dos dois grupos da correlação canônica forem independentes ou não correlacionados, a análise perde sua utilidade. Isso ocorre por que $\sum xy = 0$ e, conseqüentemente, as variáveis

canônicas $U_k = a_k'X$ e $V_k = b_k'Y$ terão covariância nula para todos os pares de vetores a_k e b_k (FERREIRA, 2008; MINGOTI, 2007).

2.6 Distribuição normal multivariada e normal multivariada contaminada

Para realizar o TRV para a independência entre dois grupos de variáveis é necessário que os vetores X e Y tenham distribuição normal multivariada (JOHNSON; WICHERN, 2007).

A distribuição normal multivariada é uma generalização da normal univariada quando se tem duas ou mais variáveis aleatórias conjuntamente em um estudo (MINGOTI, 2007; FERREIRA, 2008). Essa distribuição é muito importante na estatística multivariada, pois, a maioria das técnicas pressupõe normalidade dos dados para sua efetivação.

A função densidade de probabilidade da distribuição univariada, com média μ e variância σ^2 , é dada por: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}}$.

Um vetor aleatório com dimensão p , ou seja, $X=(X_1, X_2...X_p)$ tem distribuição normal p -variada e denota-se por $X \sim N(\mu, \Sigma_{p \times p})$, quando a função densidade for:

$$f(x_1, x_2, \dots, x_p) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{\left\{-\frac{1}{2}(x - \mu)' \Sigma^{-1}(x - \mu)\right\}}$$

para todo vetor $x \in \mathfrak{R}^p$, em que $x=(x_1, x_2...x_p)'$, $-\infty < x_i < \infty$; $i= 1, 2, \dots, p$; $\mu=(\mu_1, \mu_2, \dots, \mu_p) \in \mathfrak{R}^p$; $\Sigma_{p \times p}$ positiva definida:

$$\Sigma_{p \times p} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}$$

Em que:

$\sigma_{ii} = var(x_i)$, $i = 1, 2, \dots, p$ e $\sigma_{ij} = \sigma_{ji} Cov(X_i, X_j)$, para $i, j = 1, 2, \dots, p$, com $i \neq j$.

As propriedades da distribuição normal multivariada, bem como, maiores detalhes são encontrados em Ferreira (2008).

Um exemplo gráfico de dados gerados sob distribuição normal multivariada (representando o comportamento de seis variáveis, duas a duas) para tamanho de amostra (n) igual a 2000 observações é representada na Figura 1.

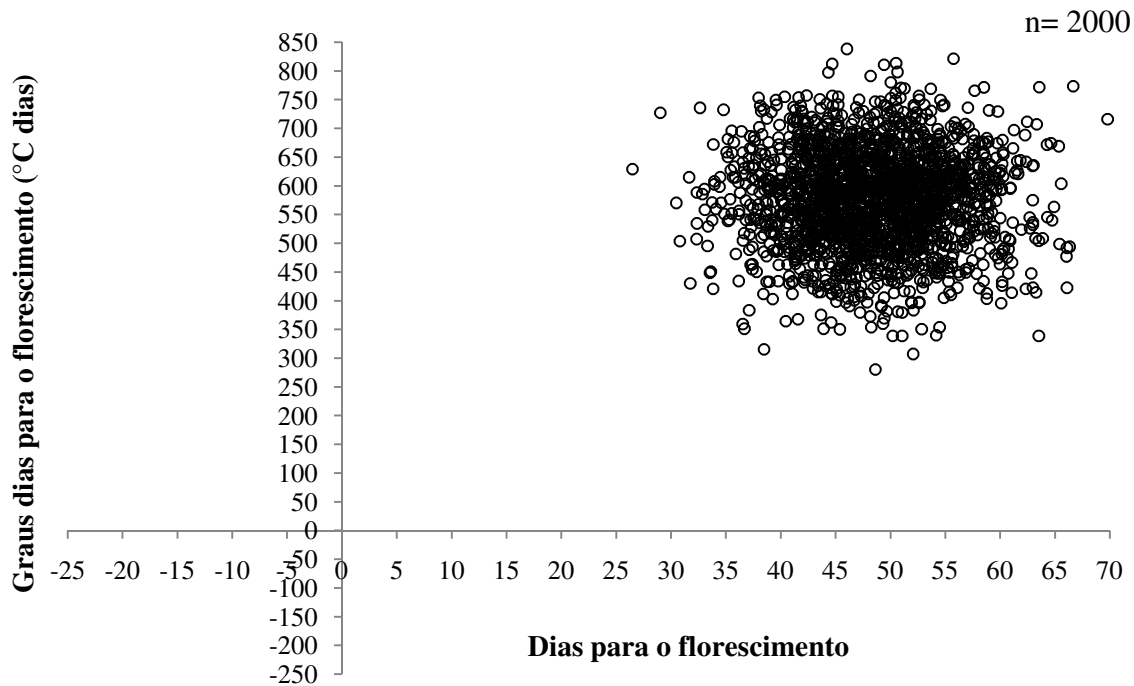


Figura 1 - Distribuição normal multivariada de seis variáveis (representando o comportamento de seis variáveis, duas a duas) para $n= 2000$.

Em estudos estatísticos é comum a ocorrência de erros, e estes podem ser classificados de duas formas: **erros aleatórios**, os quais se referem a pequenas flutuações aleatórias nos dados, causadas por fatores não controlados (SALSBURG, 2009; FARIAS, 2009). Este tipo de erro é previsto e considerado nos métodos estatísticos tradicionais, como independentes, normalmente distribuídos, com variância conhecida e média nula; e, **os erros grosseiros**, que se originam de eventos não aleatórios, e tem pouca ou nenhuma ligação com o valor medido (FARIAS, 2009). Estes invalidam os métodos estatísticos tradicionais e prejudicam a análise de dados.

Os erros grosseiros podem ser divididos em: desvios sistemáticos ou permanentes (*bias*) e valores espúrios (*outliers*) (CHEN; ROMAGNOLI, 1998). Os desvios sistemáticos são valores maiores ou menores das medidas em relação aos valores reais (exemplo calibração incorreta de um aparelho); e, os *outliers* são medidas obtidas a partir de algum comportamento anormal; valores atípicos; uma observação que apresenta um grande afastamento das demais pertencentes ao conjunto de dados (FARIAS, 2009). A expressão “erros grosseiros” designa pontos experimentais que parecem não se adequar a uma distribuição particular de probabilidade, definida pela maioria dos demais pontos experimentais (SCHWAAB; PINTO, 2007). Estes erros são conhecidos como contaminantes.

Portanto, em uma distribuição de probabilidade contaminada assume-se que a maioria das medições em um estudo seja oriunda da distribuição de probabilidade cujos parâmetros

queremos estimar (por exemplo, distribuição normal, com parâmetros: média e variância), dentre estes valores, existem algumas medições que foram contaminadas por alguns valores originários de outras distribuições (SALSBURG, 2009).

A densidade de probabilidade de uma variável aleatória X , com distribuição normal multivariada contaminada é dada por (FERREIRA, 2008):

$$fX(x) = \delta(2\pi)^{-p/2}|\Sigma_1|^{-1/2}exp\left\{-\frac{(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1)}{2}\right\} \\ + (1 - \delta)(2\pi)^{-p/2}|\Sigma_2|^{-1/2}exp\left\{-\frac{(x - \mu_2)^T \Sigma_2^{-1}(x - \mu_2)}{2}\right\}$$

Nos primeiros estudos estatísticos, os pesquisadores recorriam à prática de descartar observações, cujos valores parecessem falhos. No entanto, os teoremas matemáticos da estatística exigem que todas as observações de um estudo sejam consideradas igualmente válidas (SALSBURG, 2009). Conforme o autor, somente em estudos do final da década de 1960, desenvolvidos em Princeton, com simulação Monte Carlo de um grande número de distribuições normais contaminadas, foi “resolvido” o problema do que deve ser feito com medições aparentemente erradas (contaminadas).

Na distribuição normal, essa contaminação faz com que o conjunto de dados produza resultados bem diferentes do esperado na população. Além disso, quando os dados estão contaminados a média é uma medição muito pobre, que não representa a realidade populacional, pois até mesmo uma única medição errada (contaminante) a influenciará significativamente (SALSBURG, 2009).

Em experimentos agrícolas, ocorrem exemplos de distribuições normais contaminadas, tais como: i. supondo que um experimento em campo, com dada cultura agrícola, tenha algumas plantas atacadas por um animal; os resultados de determinadas medidas (por exemplo: altura de plantas) nas plantas atacadas será bem diferente dos resultados observados nas plantas sem ataque; ii. outro exemplo pode ocorrer, é quando, em determinada data, ao se coletar medidas de variáveis de plantas ou animais do experimento se utiliza um aparelho com má calibração (diferente da calibração utilizadas nas demais datas); iii. a coleta de variáveis de um experimento realizada por dois indivíduos de forma diferenciada (sem padronização) e/ou não respeitando a casualização resultará em valores discrepantes. Muitos outros exemplos poderiam ser citados, o que significa que é comum o pesquisador da área agrícola se deparar com uma situação onde há presença de dados contaminados.

Nessa situação, o pesquisador precisa recorrer a métodos de análise que diminuam a influência das observações contaminadoras. Porém, antes de recorrer a tais métodos é

importante que seja conhecida a resposta do método estatístico tradicional que se pretende utilizar aos diferentes graus de contaminação da distribuição normal.

Na Figura 2, tem-se um exemplo de dados gerados sob distribuição normal multivariada contaminada, com as mesmas variáveis e tamanho amostral utilizados na distribuição normal multivariada (Figura 1). Neste caso, a contaminação utilizada foi a contaminação de posição, ou seja, apenas as médias foram contaminadas, sendo o grau de contaminação igual a 20%. Os valores das matrizes de covariância são os mesmos utilizados para o caso da distribuição normal.

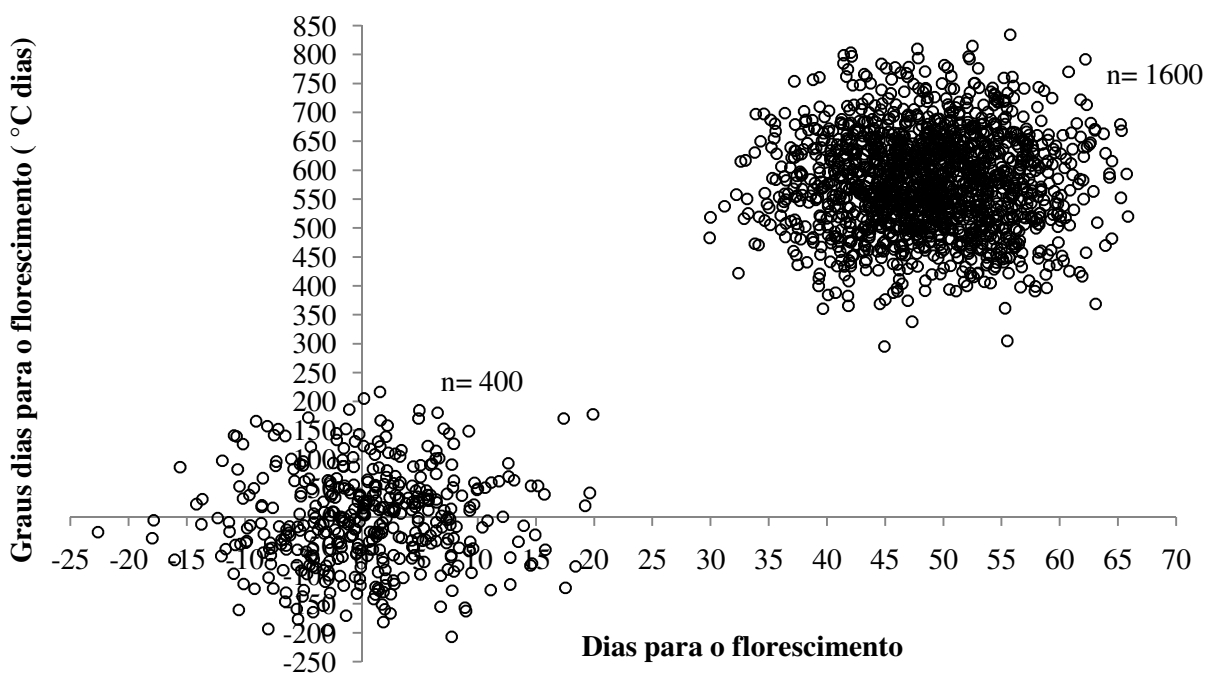


Figura 2 - Distribuição normal multivariada contaminada de seis variáveis (representando o comportamento de seis variáveis, duas a duas), para $n= 2000$, com grau de contaminação igual a 20%.

2.7 O teste de razão de verossimilhança e a normalidade multivariada

A normalidade multivariada é um pressuposto requerido nos vários testes de hipótese baseados na razão de verossimilhança. Johnson e Wichern (2007) afirmam que no TRV para a independência entre dois grupos de variáveis, ela é fundamental, pois a distribuição assintótica qui-quadrado para a estatística de teste da razão de verossimilhança é baseada no pressuposto de que os dados vêm do modelo paramétrico e que o parâmetro satisfaça a hipótese nula (KENT, 1982).

Alguns trabalhos mostram a importância da normalidade para o TRV: Kent (1982) avaliou o efeito de um modelo não paramétrico univariado, com distribuição desconhecida em diversas situações e concluiu que, na maioria dos casos, a estatística da razão de verossimilhança não segue uma distribuição assintótica qui-quadrado e o teste não é robusto.

Em trabalhos recentes, é possível encontrar relatos sobre a robustez de diversos testes de hipóteses baseados na razão de verossimilhança, quando o pressuposto de normalidade é violado (CURRAN et al., 1996; ALLISON et al., 1999; YANAGIHARA, 2005; SILVA et al., 2008); em alguns desses trabalhos são propostos, com sucesso, correções para as estatísticas de teste em condição de não-normalidade.

Estudando o efeito da não-normalidade sobre distribuições assintóticas das distribuições nulo e não nulo de critérios de razão de verossimilhança para quatro testes de hipóteses de matrizes de covariância (teste para verificar que a matriz de covariância é igual a uma dada matriz; teste de esfericidade; teste de independência e teste para igualdade de matrizes de covariância em k grupos), Yanagihara et al. (2005) verificaram que as taxas de erro tipo I, do teste com base no pressuposto de normalidade tendem a exceder o nível mínimo de significância assintoticamente. Por outro lado, o poder do teste é maior do que o poder do caso normal.

Os autores afirmam que se vetores são observados de forma independente e identicamente distribuídos de acordo com a distribuição normal multivariada, as distribuições nulas baseadas na razão de verossimilhança, para testar estruturas de covariância convergem para o centro da distribuição qui-quadrado, quando o tamanho da amostra n tende ao infinito.

A explicação para distorções nos TRV sob não normalidade, se deve ao fato de que, a distribuição normal é completamente descrita pelos dois primeiros momentos, a média e a variância. Se a distribuição é não-normal, então, distribuições assintóticas das distribuições nulo e não nulo das estatísticas de teste dependem do quarto momento (curtose) da distribuição da população (YANAGIHARA et al., 2005).

3 MATERIAL E MÉTODOS

3.1 Banco de dados

O banco de dados de variáveis de mamoneira (híbrido Lyra), do qual foram utilizadas as médias e variâncias para gerar as amostras aleatórias multivariadas foi obtido do experimento de Brum (2009).

O experimento com a cultura de mamoneira (híbridos Sara e Lyra) foi conduzido no ano agrícola 2007/2008, no Departamento de Fitotecnia da Universidade Federal de Santa Maria (UFSM), com coordenadas 29° 43' 23'' S e 53° 43' 15'' W e altitude de 95 m.

As sementes (previamente medidas quanto a largura, comprimento, espessura e peso) foram identificadas com números de um a 180 e semeadas na área experimental, no dia 19 de outubro de 2007, seguindo as recomendações para o cultivo da mamoneira no Rio Grande do Sul (WREGGE et al., 2007), juntamente com as demais sementes que compuseram as seis parcelas (três parcelas do híbrido Sara e três parcelas do híbrido Lyra) de 15m de comprimento x 16m de largura . Em campo, as plantas foram avaliadas desde a emergência até a colheita, originando as variáveis que foram divididas em grupos de: plântulas, plantas adultas e produção.

As médias e as variâncias, utilizadas para gerar as matrizes de covariância, e, assim, as amostras aleatórias multivariadas, para cálculo do erro tipo I e poder, do TRV para a independência entre dois grupos de variáveis, pertencem aos grupos de planta adulta e de produção do híbrido Lyra (Tabela 1).

3.2 Obtenção das amostras normais multivariadas

3.2.1 Erro tipo I

3.2.1.1 Matrizes de covariância e vetores de médias

Para a obtenção das amostras normais multivariadas, utilizadas nos cálculos das taxas de erro tipo I do TRV para a independência entre dois grupos de variáveis, sob a hipótese nula, foram utilizadas matrizes com covariâncias nulas e variância e médias reais. Essas foram obtidas de 14 variáveis de plantas adultas (Grupo I) e de 10 variáveis de produção de grãos do híbrido Lyra de mamoneira (Grupo II) (Tabela 1).

Tabela 1 - Variância e média de variáveis de planta adulta e de produção de grãos de mamoneira, em 50 plantas do híbrido Lyra. Santa Maria, 2011.

Variáveis (unidade de medida) - sigla	Variância	Média
Grupo I- Planta adulta		
1. Dias para o florescimento (dias) - diasflor	40,3282	48,7200
2. Graus-dia para o florescimento (°C dia) - gdflor	7134,1381	578,4340
3. Altura de planta no início do florescimento (cm) - hflor	16,4368	25,6100
4. Comprimento do caule com folhas verdadeiras (cm) - compcaule	12,6981	17,8900
5. Diâmetro do caule a 10 cm do solo (cm) - diamcaule	9,7405	19,7860
6. Número de folhas por planta - nfolhas	1,3486	9,2800
7. Área Foliar média – método comprimento e largura (cm ²) - afcl	14920,2997	439,5941
8. Área Foliar média – método comprimento da nervura principal (cm ²) - afcnp	13579,6812	351,6296
9. Altura de inserção do primeiro rácemo (cm) - h1rac	16,4368	25,6100
10. Dias para o fim do florescimento a partir da emergência (dias) - daefimflor	209,7322	98,6800
11. Dias para o fim do florescimento a partir da semeadura (dias) - dasfimflor	179,4883	109,0680
12. Altura do rácemo mais alto (cm) - hracalto	579,6017	133,8560
13. Graus-dia para o fim do florescimento (°C dia) - gdfimflor	38673,9307	1628,0140
14. Número de inflorescências - ninflo	72,4600	10,5340
Grupo II- Produção de grãos		
1. Comprimento médio dos racemos (cm) - comprac	89,7463	31,4256
2. Número de racemos por planta - nrac	54,7902	9,1600
3. Número médio de cápsulas por rácemo - ncapra	160,0415	40,4371
4. Massa média de cápsulas por racemo (g) - massacapra	938,8002	77,9063
5. Massa total de cápsula por planta (g) - massacapl	86702,2193	648,6912
6. Massa média de grãos por rácemo (g) - massagrac	820,5939	64,3425
7. Massa de grãos por planta (g) - massagraopl	43732,5893	462,7205
8. casca (g) - cascag	9858,8491	186,8053
9. Percentual de casca (%) - casca%	31,2163	28,3524
10. Massa de cem grãos (g) - massa100	17,8999	51,0854

As matrizes foram obtidas pela combinação de variáveis dos dois grupos (14 de planta adulta e 10 de produção de grãos), iniciando-se com a combinação 3+3 (3 variáveis no grupo I e 3 variáveis no grupo II) até a combinação 14+10, totalizando 40 combinações, conforme apêndice A. Dessa forma, a matriz de covariância para a combinação 3+3 (Σ_{XY}), por exemplo, para cálculo das taxas de erro tipo I (sob a hipótese nula) foi a seguinte:

$$\Sigma_{XY} = \begin{pmatrix} & \text{diasflor} & \text{gdflor} & \text{hflor} & \text{comprac} & \text{nrac} & \text{ncapra} \\ \text{diasflor} & \mathbf{40,3282} & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,0000 \\ \text{gdflor} & 0,0000 & \mathbf{7134,1381} & 0,0000 & 0,0000 & 0,0000 & 0,0000 \\ \text{hflor} & 0,0000 & 0,0000 & \mathbf{16,4368} & 0,0000 & 0,0000 & 0,0000 \\ \text{comprac} & 0,0000 & 0,0000 & 0,0000 & \mathbf{89,7463} & 0,0000 & 0,0000 \\ \text{nrac} & 0,0000 & 0,0000 & 0,0000 & 0,0000 & \mathbf{54,7902} & 0,0000 \\ \text{ncapra} & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,0000 & \mathbf{160,0415} \end{pmatrix}$$

Na qual, nas três primeiras linhas e colunas da matriz têm-se os valores das variâncias das três primeiras variáveis de planta adulta e nas três últimas linhas e colunas da matriz, os valores de variância das três primeiras variáveis de produção (Tabela 1). Assim, para o exemplo anterior, a matriz referente ao grupo 1 (Σ_{XX}) e ao grupo 2 (Σ_{YY}) foram, respectivamente:

$$\Sigma_{XX} = \begin{pmatrix} & \text{diasflor} & \text{gdflor} & \text{hflor} \\ \text{diasflor} & \mathbf{40,3282} & 0,0000 & 0,0000 \\ \text{gdflor} & 0,0000 & \mathbf{7134,1381} & 0,0000 \\ \text{hflor} & 0,0000 & 0,0000 & \mathbf{16,4368} \end{pmatrix}$$

e

$$\Sigma_{YY} = \begin{pmatrix} & \text{comprac} & \text{nrac} & \text{ncapra} \\ \text{comprac} & \mathbf{89,7463} & 0,0000 & 0,0000 \\ \text{nrac} & 0,0000 & \mathbf{54,7902} & 0,0000 \\ \text{ncapra} & 0,0000 & 0,0000 & \mathbf{160,0415} \end{pmatrix}$$

O vetor de médias dos grupos I (μ_X) e do grupo II (μ_Y) (é, respectivamente: $\mu_X = 48,7200; 578,4340; 25,6100$ e $\mu_Y = 31,4256; 9,1600; 40,4371$).

3.2.1.2 TRV para a independência entre dois grupos de variáveis

As hipóteses estabelecidas no TRV para a independência entre dois grupos de variáveis foram: $H_0: \sum xy = 0_{p+q}$ (os dois grupos de variáveis são independentes) e $H_1: \sum xy \neq 0_{p+q}$ (os dois grupos de variáveis não são independentes), sendo 0_{p+q} a matriz nula.

As estatísticas qui-quadrado do TRV para a independência entre dois grupos de variáveis (fórmulas) testadas foram: $\chi_c^2 = n \times \ln\left(\frac{|S_{XX}||S_{YY}|}{|S_{XY}|}\right)$, com pxq graus de liberdade para grandes amostras; e, com a correção de Bartlet (FERREIRA, 2008; JOHNSON; WICHERN, 2007 e MINGOTI, 2007): $\chi_c^2 = \left[n - 1 - \frac{1}{2}(p + q + 1)\right] \times \ln\left(\frac{|S_{XX}||S_{YY}|}{|S_{XY}|}\right)$, com pxq graus de liberdade, para pequenas amostras.

Nesse teste, o valor da estatística calculada é comparado ao valor crítico tabelado da distribuição qui-quadrado, com pxq graus de liberdade e em nível de 5% de significância; sempre que o valor calculado for maior do que o valor tabelado, rejeita-se a hipótese nula e conclui-se que os dois grupos considerados no estudo não são independentes.

3.2.1.3 Cenários e simulação Monte Carlo

Para avaliar a probabilidade de cometer o erro tipo I no TRV para a independência entre dois grupos de variáveis, foi realizada simulação computacional pelo método de Monte Carlo, considerando-se os seguintes cenários: 16 tamanhos de amostra - 25, 30, 50, 100, 200, 300, 400, 500, 600, 750, 1000, 1500, 2000, 3000, 4000 e 5000; 40 combinações de número de variáveis entre os dois grupos - iniciando em 3+3, 3+4, 3+5, 3+6, 3+7, 3+8, 3+9, 3+10, 4+4, 4+5, 4+6, 4+7, 4+8, 4+9, 4+10 até 14+10; e, um grau de correlação entre as variáveis das matrizes de covariância: 0; totalizando 640 cenários (16 x 40 x 1) na fórmula para pequenas amostras e 640 cenários na fórmula para grandes amostras.

Para cada cenário, foram gerados 2000 simulações com distribuição normal multivariada; o que resultou em 1280000 simulações (640 x 2000) por fórmula. Em cada simulação (2000) foram obtidos os valores de qui-quadrado calculado e o p-valor do TRV para independência entre dois grupos de variáveis. A taxa de erro tipo I foi calculada pela razão entre o número de vezes em que a hipótese nula foi rejeitada ($p\text{-valor} \leq 0,05$) e o número total de simulações (2000).

Para verificar se as taxas de erro tipo I diferiram do nível de significância adotado ($\alpha=5\%$) foi utilizado o limite inferior (3,475%) e o limite superior (6,255%) do intervalo de confiança (IC) exato de 99%, para uma proporção $\hat{p} = 0,05$, o qual foi calculado por:

$IC = \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, em que $n=2000$ simulações e $Z_{\alpha/2} = 2,575829304$ obtido da normal padrão com média zero e desvio padrão um. Portanto, somente taxas de erro fora deste intervalo (3,475% a 6,255%) foram consideradas como diferentes do valor nominal.

3.2.2 Poder

3.2.2.1 Matrizes de covariância e vetores de médias

Para obtenção das amostras normais multivariadas, utilizadas nos cálculos de poder do teste de razão de verossimilhança para a independência entre dois grupos de variáveis foram utilizadas as mesmas variâncias e vetor de médias utilizados no cálculo das taxas de erro tipo I (Tabela 1); e, covariâncias dadas por: $Cov_{XY} = s_p \times s_{p+1} \times r$, em que S_p é o desvio padrão da variável p e S_{p+1} , os desvios padrão das demais variáveis pertencentes a matriz; e, r o grau de correlação entre as variáveis de cada matriz ($r= 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8$ e $0,9$). Portanto, o total de matrizes geradas para cálculo de poder foi 360 (40 combinações de variáveis x 9 graus de correlação). Assim, covariância entre as variáveis *diasflor* e *gdflor*, considerando a matriz abaixo com $r= 0,1$, foi obtida da seguinte forma: $Cov(diasflor - gdflor) = s_{diasflor} \times s_{gdflor} \times 0,1$.

A matriz de covariância (Σ_{XY}) entre os dois grupos de variáveis, para a combinação 3+3, para cálculo do poder (matriz construída sob a hipótese alternativa), considerando o grau de correlação entre as variáveis igual a 0,1, por exemplo, foi a seguinte:

$$\Sigma_{XY} = \begin{pmatrix} & diasflor & gdflor & hflor & comprac & nrac & ncapra \\ diasflor & \mathbf{40,3282} & 53,6383 & 2,5746 & 6,0161 & 4,7006 & 8,0338 \\ gdflor & 53,6383 & \mathbf{7134,1381} & 34,2436 & 80,0164 & 62,5205 & 106,8531 \\ hflor & 2,5746 & 34,2436 & \mathbf{16,4368} & 3,8408 & 3,0010 & 5,1289 \\ comprac & 6,0161 & 80,0164 & 3,8408 & \mathbf{89,7463} & 7,0123 & 11,9846 \\ nrac & 4,7006 & 62,5205 & 3,0010 & 7,0123 & \mathbf{54,7902} & 9,3641 \\ ncapra & 8,0338 & 106,8531 & 5,1289 & 11,9846 & 9,3641 & \mathbf{160,0415} \end{pmatrix}$$

As matrizes referentes ao grupo 1 (Σ_{XX}) e grupo 2 (Σ_{YY}) são as seguintes:

$$\Sigma_{XX} = \begin{pmatrix} & \text{diasflor} & \text{gdflor} & \text{hflor} \\ \text{diasflor} & \mathbf{40,3282} & 53,6383 & 2,5746 \\ \text{gdflor} & 53,6383 & \mathbf{7134,1381} & 34,2436 \\ \text{hflor} & 2,5746 & 34,2436 & \mathbf{16,4368} \end{pmatrix}$$

e

$$\Sigma_{YY} = \begin{pmatrix} & \text{comprac} & \text{nrac} & \text{ncapra} \\ \text{comprac} & \mathbf{89,7463} & 7,0123 & 11,9846 \\ \text{nrac} & 7,0123 & \mathbf{54,7902} & 9,3641 \\ \text{ncapra} & 11,9846 & 9,3641 & \mathbf{160,0415} \end{pmatrix}$$

Os vetores de médias dos grupos 1 (μ_X) e 2 (μ_Y), foram os mesmos utilizados na geração das amostras aleatórias multivariadas no cálculo das taxas de erro tipo I.

Em relação à aplicação do teste para cálculo do poder, a única diferença em relação ao procedimento aplicado para cálculo das taxas de erro tipo I foi que o TRV para a independência entre dois grupos de variáveis foi construído sob a hipótese alternativa, ou seja, as matrizes de covariância apresentavam correlação entre as variáveis ($r= 0,1$ a $0,9$). Em outras palavras, as amostras aleatórias normais multivariadas foram geradas de forma a avaliar o poder do teste em rejeitar a hipótese H_0 , que por construção é falsa.

3.2.2.2 Cenários e simulação Monte Carlo

Na simulação Monte Carlo, para os cálculos de poder, foi considerado os seguintes cenários: 16 tamanhos de amostra - 25, 30, 50, 100, 200, 300, 400, 500, 600, 750, 1000, 1500, 2000, 3000, 4000 e 5000; 40 combinações de número de variáveis entre os dois grupos - iniciando em 3+3, 3+4, 3+5, 3+6, 3+7, 3+8, 3+9, 3+10, 4+4, 4+5, 4+6, 4+7, 4+8, 4+9, 4+10 até 14+10 e, nove graus de correlação entre as variáveis ($r= 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8$ e $0,9$), totalizando 5760 cenários ($16 \times 40 \times 9$) e 11520000 simulações (5760×2000). Portanto, foram estudados 11520000 simulações na fórmula para pequenas amostras e 11520000 simulações na fórmula para grandes amostras.

O poder foi calculado pela razão entre o número de vezes em que a hipótese nula foi rejeitada e o número total de simulações (2000), considerando o nível de significância de 5%.

3.3 Obtenção das amostras normais multivariadas contaminadas

As matrizes de variância e covariância utilizadas para a obtenção das amostras normais multivariadas contaminadas foram as mesmas utilizadas para gerar as amostras normais multivariadas.

No entanto, inseriu-se na programação do R um vetor de médias para contaminar as observações da população original (Tabela 1), ou seja, a contaminação efetuada foi a de posição, mantendo-se a mesma matriz de covariância. O grau de contaminação das médias foi de 20%, o que significa que a população contaminante continha 20% das observações da população de referência. Para fins de exemplo, o vetor de médias utilizado para contaminação da menor matriz (3+3; descrita no item 3.2.1.1 para erro tipo I e no item 3.2.2.1, para o poder) foi: 0; 0,1; 0,01; 0; 1; 10. Para as demais matrizes aumentou-se os valores, mantendo-se a mesma progressão observada nesses valores. Dessa forma, na matriz com três variáveis do grupo I e três variáveis do grupo II, por exemplo, o vetor de médias utilizado para contaminação foi: 0; 0,1; 0,01; 0,001; 0,0001; 0,00001; 0; 1; 10; 100; 1000; 10000. Este vetor de médias foi obtido por tentativa e erro, de forma que, em cada cenário, o vetor escolhido resultasse em 2000 simulações sem distribuição normal multivariada, quando avaliadas pelo *mvShapiro.Test*.

Os cenários e o procedimento para cálculo das taxas de erro tipo I e poder foram os mesmos utilizados na distribuição normal multivariada.

3.4 Obtenção das matrizes de covariância e simulação no programa R

Para obtenção das matrizes no cálculo de erro tipo I e poder do teste foi utilizado o aplicativo *Microsoft Excel 2007*. As simulações foram realizadas com o uso do método Monte Carlo; e, os cálculos das taxas de erro tipo I e poder foram obtidos por meio de programação desenvolvida no software R para Windows, versão 2.13.0 (R Development Core Team, 2010).

A obtenção das amostras normais multivariadas no R foi realizada com o uso da função `mvrnorm`, do pacote MASS. Para obtenção das amostras normais multivariadas contaminadas utilizou-se a função `mvrnormCT`, desenvolvida pelo Professor Daniel Furtado Ferreira.

Para confirmar a normalidade (distribuição normal multivariada) e a não-normalidade (distribuição normal multivariada contaminada) das amostras aleatórias multivariadas geradas, utilizou-se a função `mvShapiro.Test` ($\alpha= 5\%$), pertencente ao pacote `mvShapiroTest` do software R (VILLASENOR ALVA; ESTRADA, 2009). Dessa forma, cada uma das 2000 simulações/cenário foi avaliada quanto ao atendimento ou não do pressuposto de normalidade multivariada, garantindo que a condição estabelecida para a amostra (normalidade multivariada ou não-normalidade multivariada) fosse atendida.

4 RESULTADOS E DISCUSSÃO

4.1 Distribuição normal multivariada

A normalidade multivariada das amostras geradas no programa R foi confirmada pelo teste `mvShapiro.Test` (VILLASENOR ALVA; ESTRADA, 2009), em todas as 2000 simulações geradas por cenário. Além disso, o teste se mostrou confiável, pois manteve as taxas de erro tipo iguais ao nível mínimo de significância ($\alpha= 5\%$) adotado.

Após confirmação da normalidade multivariada das amostras geradas no programa R foi realizado o procedimento de simulação para avaliação das taxas de erro tipo I e poder do TRV para independência entre dois grupos de variáveis.

4.1.1 Erro tipo I

Os valores das taxas de erro tipo I do TRV para independência entre dois grupos de variáveis, sob distribuição normal multivariada, em tamanhos de amostra pequenos, $n= 25$ e 30 (Figura 3 e 4) aumentam expressivamente, com o acréscimo no número de variáveis.

Este comportamento ocorre com o uso de ambas as fórmulas de cálculo da estatística de teste qui-quadrado (para pequenas e grandes amostras). Porém, nos tamanhos de amostra (n)= 25 , 30 e 50 (Figuras 3, 4 e 5) as taxas de erro tipo I obtidas por meio da fórmula para grandes amostras são expressivamente maiores do que as calculadas com o uso da fórmula corrigida (pequenas amostras), principalmente quando o número de variáveis é de 14 a 22 .

Em todas as situações (Figura 3 a 9), observa-se que o efeito nas taxas de erro tipo I é devido, principalmente, ao número total de variáveis utilizada nos dois grupos de variáveis e não devido ao número de variáveis em cada grupo, pois diferentes combinações de número de variáveis em cada grupo (APÊNDICES A a D) produzem resultados semelhantes no erro tipo I (pontos repetidos para um mesmo p nas Figuras 3 a 9).

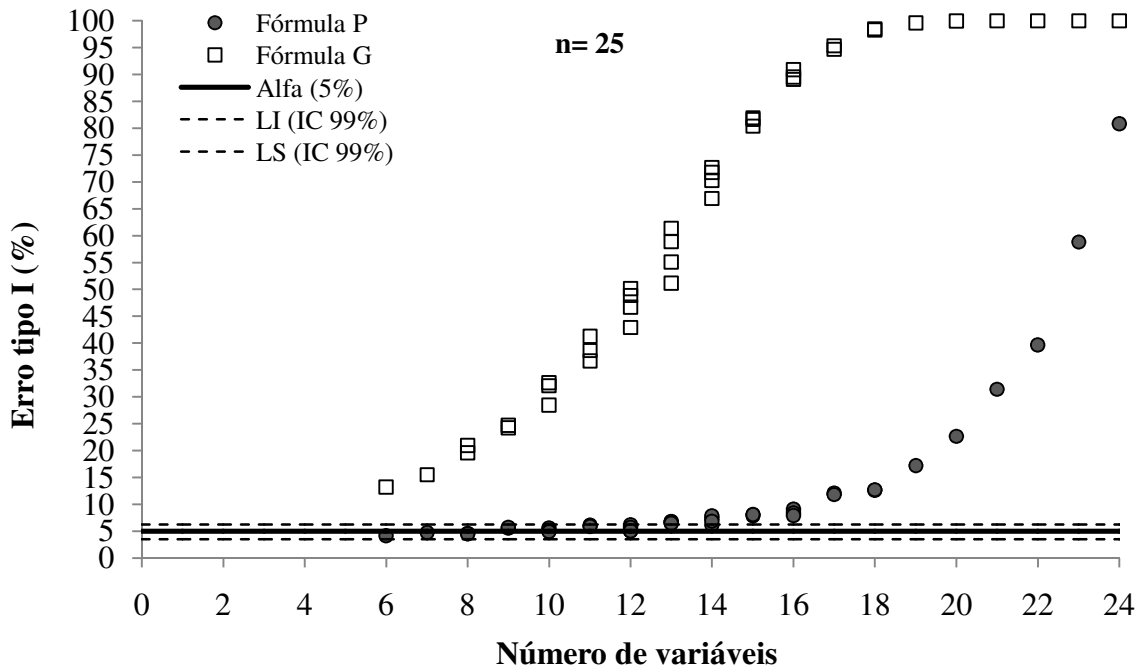


Figura 3 - Erro tipo I do TRV para a independência entre dois grupos de variáveis ($\alpha= 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n= 25$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $\hat{p} = 0,05$.

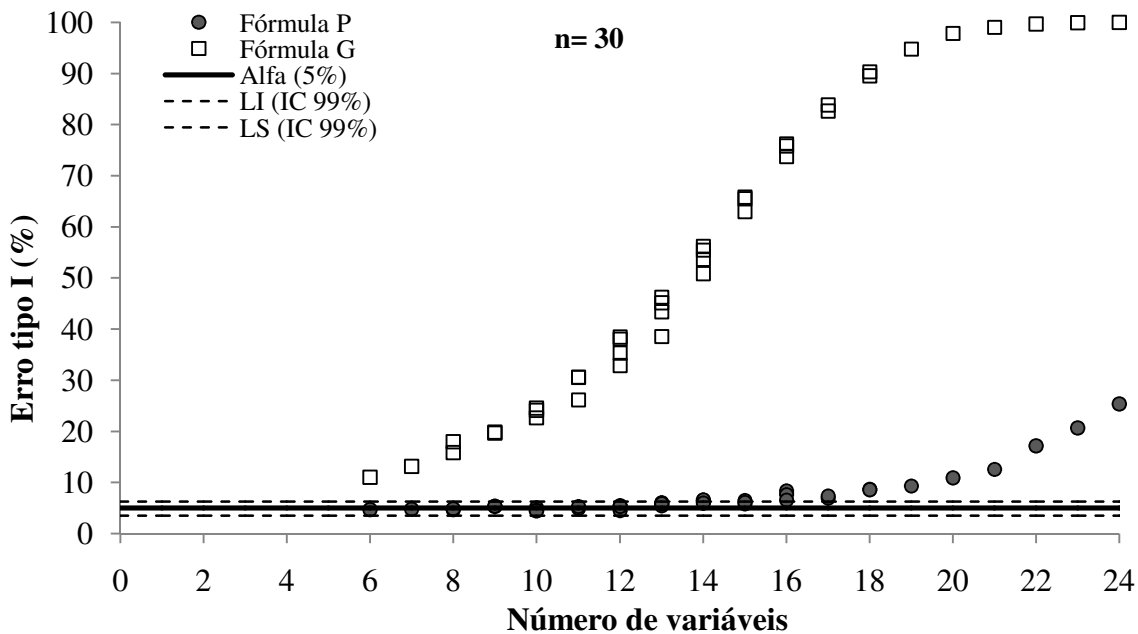


Figura 4 - Erro tipo I do TRV para a independência entre dois grupos de variáveis ($\alpha= 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n= 30$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $\hat{p} = 0,05$.

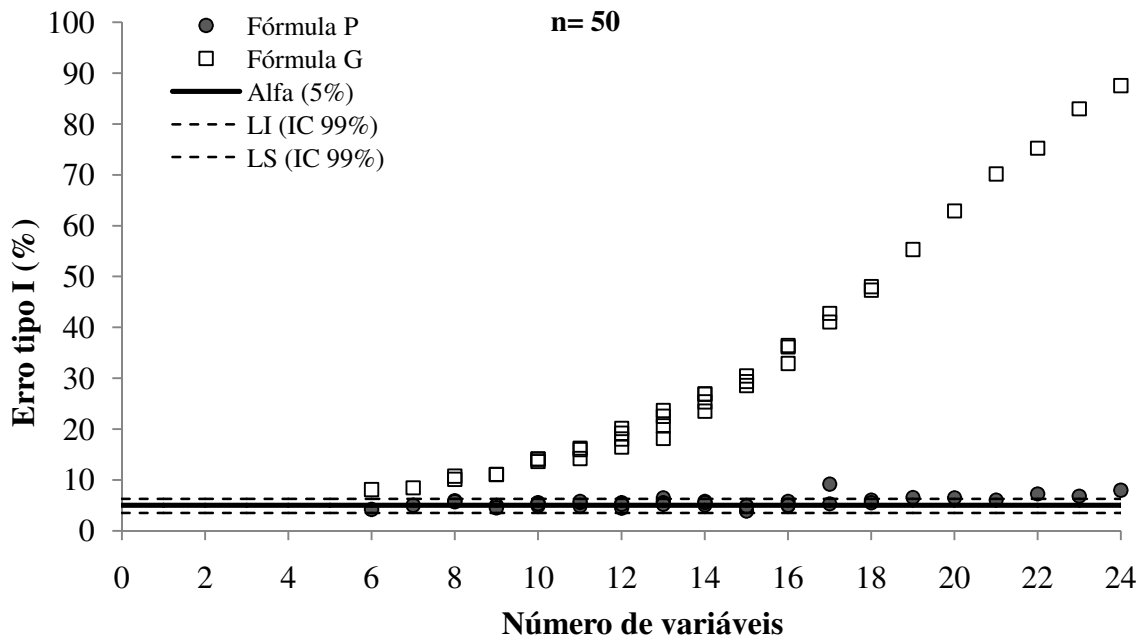


Figura 5 - Erro tipo I do TRV entre dois grupos de variáveis ($\alpha = 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n = 50$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $\hat{p} = 0,05$.

Em tamanhos de amostra de 100 a 400 observações, as diferenças entre as fórmulas são bem menores (Figuras 6, 7, 8 e 9); e, em tamanhos de amostra variando de 500 a 5000 observações (APÊNDICE B a D), as duas fórmulas comportam-se de maneira idêntica, mantendo, praticamente, todas as taxas de erro tipo I dentro do intervalo de 99% de confiança e não alterando estas ao aumentar o número de variáveis.

Utilizando-se a fórmula para pequenas amostras, o nível mínimo de significância de 5% do TRV da independência entre dois grupos de variáveis não é violado em pequenos tamanhos de amostra (25 e 30), quando o número de variáveis no estudo é de no máximo 12 (Figuras 3 e 4). Porém, nesse caso, ao utilizar 14 ou mais variáveis, as taxas de erro tipo I aumentam significativamente e o teste torna-se liberal (excede o limite superior de 99% do IC= 6,255%). Um teste de hipótese é classificado como liberal, quando as taxas de erro tipo I são significativamente superiores ao nível mínimo de significância estabelecido (α) (SILVA et al., 2008). A partir de $n=50$ (Figura 5), aumentos no número total de variáveis não causam efeito na taxa de erro tipo I e essa se mantém próximo a 5%, com o uso da fórmula para pequenas amostras. Neste caso, as taxas de erro tipo I ultrapassam o limite superior de 99% do IC apenas com $n= 22, 23$ e 24 variáveis.

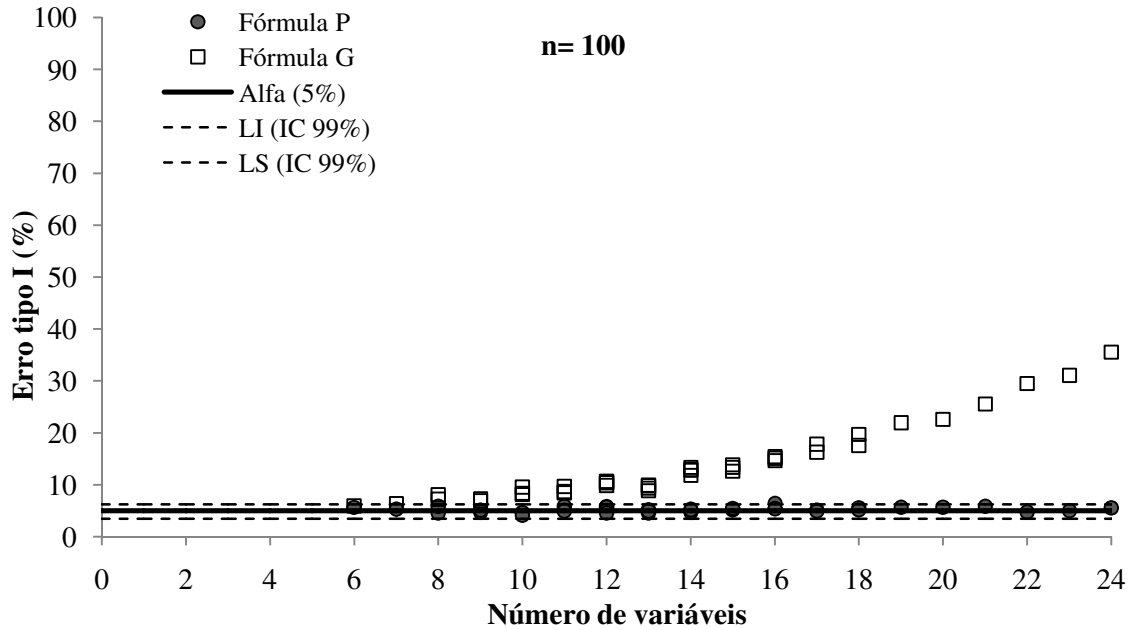


Figura 6 - Erro tipo I do TRV para a independência entre dois grupos de variáveis ($\alpha= 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n= 100$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $\hat{p} = 0,05$.

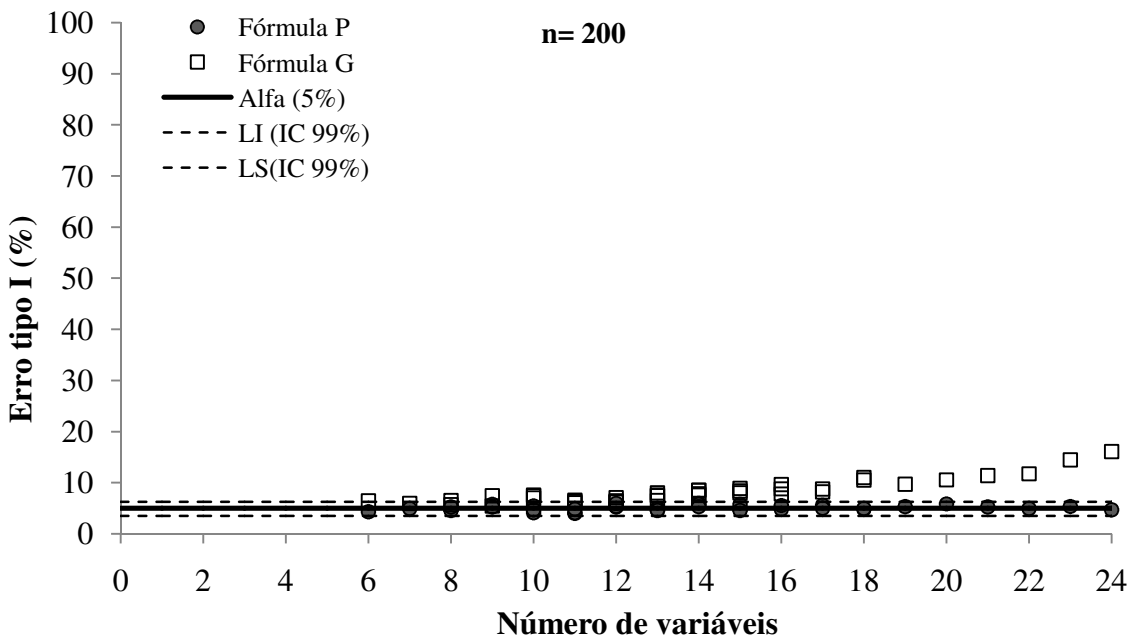


Figura 7 - Erro tipo I do TRV para a independência entre dois grupos de variáveis ($\alpha= 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n= 200$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $\hat{p} = 0,05$.

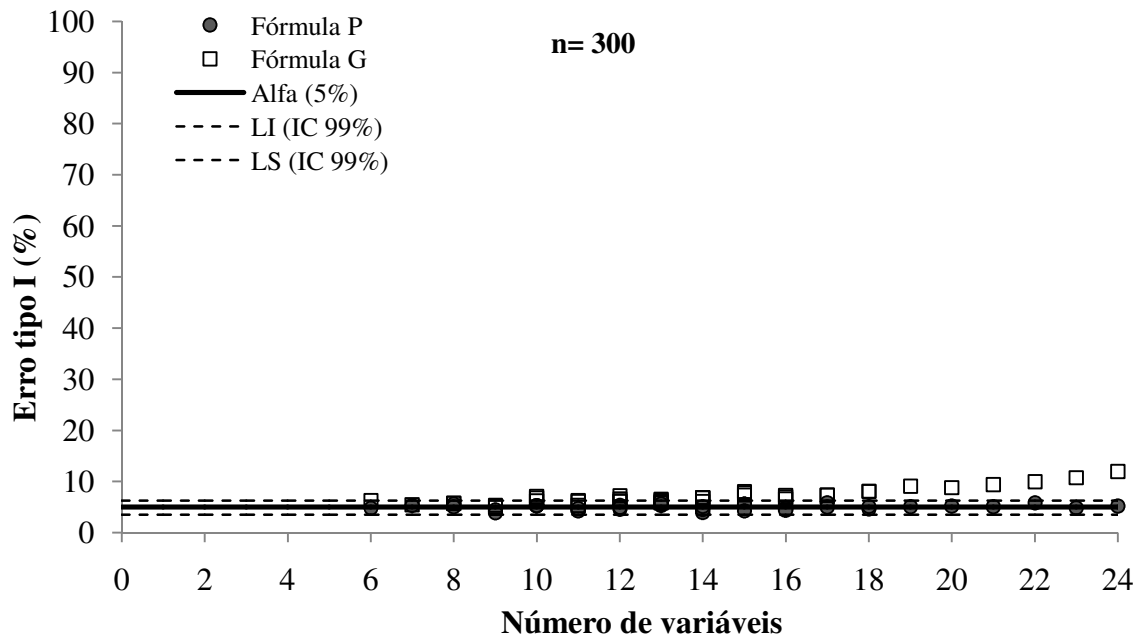


Figura 8 - Erro tipo I do TRV para a independência entre dois grupos de variáveis ($\alpha= 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n= 300$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $\hat{p} = 0,05$.

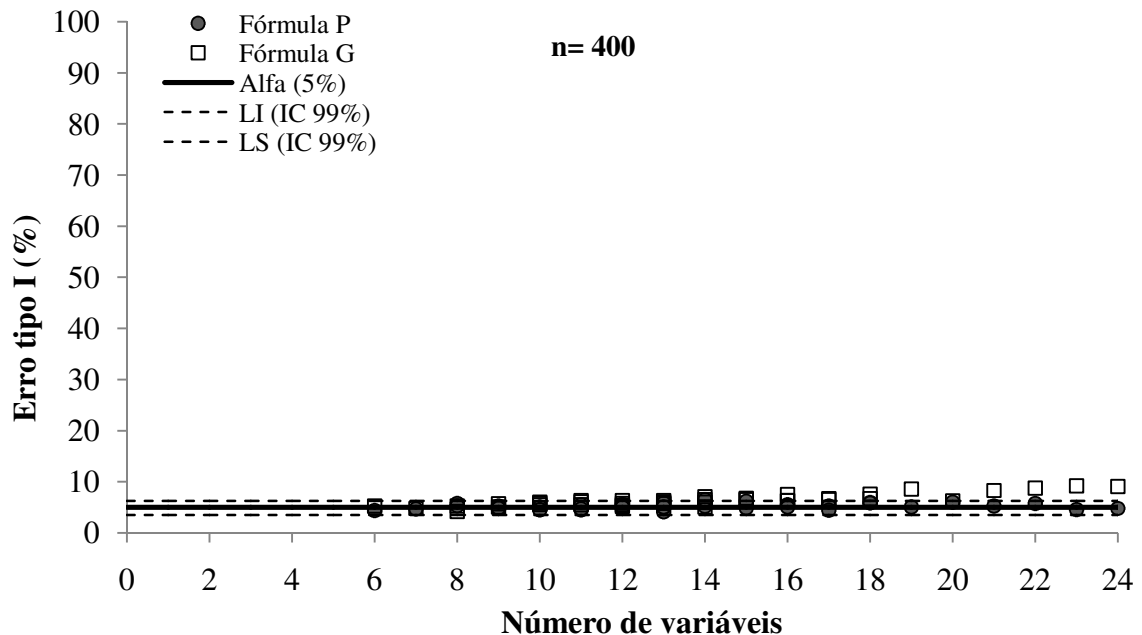


Figura 9 - Erro tipo I do TRV para a independência entre dois grupos de variáveis ($\alpha= 5\%$), com diferentes números de variáveis e duas fórmulas, para pequenas (Fórmula P) e para grandes amostras (Fórmula G), de cálculo da estatística de teste χ^2 , para o tamanho de amostra ($n= 400$), sob distribuição normal multivariada, em 2000 simulações, com o limite inferior (LI) e o limite superior (LS) do intervalo de confiança (IC) exato de 99%, para uma proporção $\hat{p} = 0,05$.

Na fórmula para grandes amostras, ao utilizar-se $n=25, 30$ e 50 (Figuras 3 a 5), mesmo com o menor número de variáveis ($p=6$), o erro tipo I é maior do que 5% , aumentando expressivamente com a elevação do número de variáveis ($p \geq 8$). Aumentando-se o tamanho de amostra para 100 observações e utilizando-se nove ou menos variáveis, o erro tipo I não é violado. A partir de $n=400$ (Figura 9), aumentos no número de variáveis, praticamente, não alteram o erro tipo I, mantendo-o próximo a 5% .

A explicação para as diferenças nas taxas de erro tipo I do TRV para a independência entre dois grupos de variáveis, encontradas entre as duas fórmulas de cálculo da estatística qui-quadrado, é dada por Bartlett (1938), ao afirmar que a fórmula corrigida (para pequenas amostras) $\chi_c^2 = \left[n - 1 - \frac{1}{2}(p + q + 1) \right] \times \ln \left(\frac{|S_{XX}||S_{YY}|}{|S|} \right)$ poderá ser usada, na prática, para p e q maiores que dois, mas devido a natureza do fator de correção e à origem da tabela de aproximação, ela começa a falhar com o aumento de $(p + q)/n$; o uso da razão $\Lambda = \left(\frac{|S_{XX}||S_{YY}|}{|S|} \right)$, parece improvável para valores maiores de p e q , exceto em extensas investigações (por exemplo, correlacional) para a qual n também seria grande e a aproximação, válida. Em outras palavras, a utilização da fórmula corrigida quando o tamanho amostral é pequeno e o número de variáveis dos dois grupos ($p+q$) é grande, ou vice-versa, torna essa estatística de teste pouco confiável.

A correção de Bartlett (fórmula para pequenas amostras) do TRV, para a independência entre dois grupos de variáveis, faz com que, em tamanhos de amostra pequenos ($n=25, 30$ e 50) sejam observados menores valores da estatística de teste qui-quadrado, em relação à estatística para grandes amostras. Isso significa que ao se utilizar a fórmula sem correção, ou seja, de grandes amostras, quando n é pequeno ocorra maior probabilidade de rejeição da hipótese H_0 , quando esta hipótese deveria ser aceita (erro tipo I).

Em tamanhos de amostras grandes, a distribuição da estatística de teste do TRV tem maior aproximação à distribuição qui-quadrado; e, a fórmula para pequenas amostras considera essa aproximação, dada pela correção de Bartlett (1938); ou seja, sob a suposição de que os vetores são observados de forma independente e identicamente distribuídos de acordo com a distribuição normal multivariada, as distribuições nulas baseadas na razão de verossimilhança, para testar estruturas de covariância convergem para o centro da distribuição qui-quadrado, quando o tamanho da amostra n tende ao infinito (YANAGIHARA, 2005).

Não foram encontrados trabalhos comparando as duas estatísticas qui-quadrado do TRV da independência entre dois grupos de variáveis, mas para outros testes para matrizes de covariâncias baseados na razão de verossimilhança com distribuição aproximada qui-

quadrado, são vários os relatos do efeito de: distribuições multivariadas, tamanho de amostra, grau de correlação, número de variáveis, tamanho de população, entre outros. Avaliando o comportamento de dois testes para igualdade de matrizes de variâncias de k populações (teste de Bartlett multivariado e a sua versão *bootstrap*) com o uso de simulação Monte Carlo, em populações normais e não-normais, em diferentes combinações entre os tamanhos amostrais (n), número de variáveis (p), correlação (ρ) e número de populações (k); Silva et al. (2008) relatam que o teste de Bartlett multivariado, controla o erro tipo I em quase todas as combinações, para $k=2$, porém torna-se muito liberal ao aumentar k para 5. Em ambas situações ($k=2$ e 5), as taxas de erro tipo I aumentam muito com aumento no número de variáveis em tamanhos de amostra variando de 5 a 50. Tais resultados corroboram aos observados neste trabalho.

Estudando vários testes para matrizes de covariância, em amostras com $p=2, 3$ e 5 variáveis, em diferentes estruturas das matrizes (modificação na variância; na correlação; e, em ambas); dentre eles, o teste de razão de verossimilhança, para testar a hipótese de que a matriz de covariância de um processo (Σ) é igual a uma matriz particular (Σ_0); Pinto (2009) verificou que o teste da razão de verossimilhança aproximado pela distribuição qui-quadrado não controla o erro tipo I, para amostras de tamanho $n=10$, nas situações bivariadas; e, para amostras de tamanho $n=10$ e 25, nas situações com $p=3$ e 5 variáveis, concluindo que este não é um bom teste para estudos que envolvem pequeno número de amostras. Esse resultado, assim como os observados nesse trabalho, evidencia que os testes de razão de verossimilhança para matrizes de covariância, aproximados pela distribuição qui-quadrado, não controlam o erro tipo I, em situações de pequenos tamanhos de amostra, associados ao aumento de número de variáveis.

Considerando apenas o erro tipo I, a recomendação é que, ao se trabalhar com tamanhos de amostras pequenos ($n \leq 30$), se utilize, no máximo, 12 variáveis; para $n=50$, deverá utilizar, no máximo, 18 variáveis. Em tamanhos de amostra variando de 100 a 300, poderá ser adotado qualquer número de variáveis dentro do intervalo estudado ($p=6$ a 24 variáveis). Quando se trabalha com tamanhos de amostra superior a 500 pode-se optar por qualquer número de variáveis dentro do intervalo avaliado, bem como, por qualquer fórmula. Portanto, em todas as situações, a fórmula para pequenas amostras deverá ser escolhida, pois apresenta desempenho superior ($n \leq 500$) ou igual ($n > 500$) à de grandes amostras.

4.1.2 Poder

Nas tabelas 2 a 21, são apresentados os valores de poder do TRV para a independência entre dois grupos de variáveis, em função dos diferentes cenários estabelecidos, nos quais variam os valores de: tamanho de amostra (n), número de variáveis (p), correlação (r) e fórmula para pequenas e grandes amostras, em condições de normalidade multivariada.

No geral, há grandes diferenças de valores de poder entre as duas fórmulas (para pequena e grandes amostras), quando em pequenas amostras e correlação fraca ($r \leq 0,3$). Essas diferenças são maiores quando se considera conjuntamente menor: n , p e r . Mantendo-se fixo o p e o n , e aumentando a correlação observa-se aumento nos valores de poder, em ambos os casos, como era de se esperar, já que ao aumentar a correlação aumenta a probabilidade de se rejeitar H_0 . Tal fato mostra que o TRV para a independência entre dois grupos de variáveis é eficiente em detectar pequenas mudanças ($r= 0,1$) no grau de correlação das matrizes de covariância dos dois grupos de variáveis, o que diminui a probabilidade de se cometer o erro tipo II.

Da mesma forma que o observado para o erro tipo I, os tamanhos de amostra influenciam o poder do teste. Em situação de tamanhos de amostra ($n= 25, 30$ e 50) e correlações menores ($r=0,1$ e $0,2$), os valores de poder são expressivamente menores (Tabelas 2 a 22), independentemente do número de variáveis; enquanto que, nessas situações, os valores de erro tipo I foram, geralmente, elevados (Figuras 3 a 5). Pinto (2009) observou que, para H_1 próximo de H_0 (caso de $r=0,1$; neste trabalho) vários testes de matrizes de covariância, inclusive o TRV, são pouco poderosos em pequenas amostras ($n=10$) e que o poder dos testes aumenta com o aumento no tamanho de amostra. Ribeiro (2010) concluiu que, geralmente, ao aumentar o tamanho de amostra o poder dos testes aumenta. Além disso, o autor lembra que o bom comportamento apresentado pelo TRV e pelo proposto por Sullivan et al. (2007), para dados autocorrelacionados, corroboram aos dados apresentados por Pinto (2009), que afirma que esses são os melhores testes para dados independentes.

Utilizando a fórmula para grandes amostras, ocorrem valores elevados de poder em tamanhos de amostra a partir de 100 observações (Tabelas 2 a 21), mesmo em correlações baixas ($r= 0,2$) e no menor número de variáveis (seis) (Tabela 2). Em tamanhos de amostra superiores a 200, e, considerando correlação mínima de 0,2; praticamente todos os cenários apresentam poder próximo a 100% (Tabelas 2 a 21). A partir de $n= 750$, todos os cenários (inclusive a menor correlação, $r= 0,1$) atingem valores de poder de 100%, tanto para fórmulas

pequenas, quanto para grandes (Tabelas 2 a 21). Essa situação mostra que, no TRV, a utilização de amostras muito grandes pode levar o pesquisador a decisões incorretas, uma vez que, ao dispor de uma matriz de covariância, cujo grau de correlação é baixo ($r= 0,1$), o teste rejeitará a hipótese H_0 , indicando que não há independência entre os dois grupos de variáveis (grupos são correlacionados), quando na verdade, essa decisão (rejeitar H_0) foi causada pelo elevado tamanho de amostra. Por isso, nas situações em que n é grande, deve-se interpretar os resultados significativos dos testes de hipóteses, considerando a significância prática, pois amostras grandes podem tornar os testes estatísticos muito sensíveis. Isso significa que até efeitos muito pequenos e sem importância prática serão estatisticamente significativos (HAIR et al., 2009).

Na avaliação da qualidade de ajuste de modelos de equações estruturais, Hair et al. (2009) afirmam que esse efeito se deve a sensibilidade da estatística qui-quadrado do TRV ao tamanho de amostra, especialmente quando o tamanho de amostra é maior do que 200 observações; quando n é menor ou igual a 100, o teste qui-quadrado mostrará ajuste aceitável (diferenças não significativas nas matrizes de covariância estimadas e observadas), mesmo quando as relações do modelo mostram-se não significativas. Os autores concluem que, em função disso, o emprego do teste qui-quadrado em equações estruturais somente será confiável no intervalo de tamanho de amostra de 100 a 200 observações.

Fixando-se o tamanho de amostra e a fórmula, observa-se que há forte influência do número de variáveis e do grau de correlação nos valores de poder. Com $p \geq 10$ variáveis e $r \geq 0,2$, já podem ser observados valores elevados de poder, quando $n \geq 100$, em ambas as fórmulas (Tabelas 2 a 21). Em correlações médias a elevadas, ($r= 0,6$ a $0,9$) (CARVALHO et al., 2004), os valores de poder são próximos a 100% mesmo nos menores tamanhos de amostra e no menor número de variáveis considerado (seis) (Tabelas 2 a 21).

Resultados semelhantes são reportados por Pinto (2009), que observou que os valores de poder do TRV (para uma covariância específica) com distribuição exata e do TRV com distribuição aproximada qui-quadrado, são semelhantes ao aumentar o tamanho de amostra para 50 e 100 observações. No entanto, a concordância ou proximidade entre os valores de poder das estatísticas dos testes, naquele caso, ocorrem em tamanhos de amostra inferiores ao observado nesse estudo ($n \geq 500$) do TRV para independência entre dois grupos de variáveis com utilização da fórmula para grandes amostras.

4.1.3 Recomendações do TRV para a independência entre dois grupos de variáveis, sob distribuição normal multivariada

Sob distribuição normal multivariada, para que o TRV para a independência entre dois grupos de variáveis seja robusto devem ser consideradas as seguintes situações: i. com utilização da fórmula para pequenas amostras, o erro tipo I é controlado até $p=12$ variáveis, mesmo em tamanhos de amostra pequenos ($n= 25$ e 30), mas nessa situação apresenta poder baixo; a partir de $n \geq 100$ e $r \geq 0,2$, os valores de poder são elevados. Quando n é pequeno e aumenta-se o número de variáveis, os valores de poder e do erro tipo I aumentam; ii. utilizando-se a fórmula para grandes amostras, o teste torna-se liberal (viola o erro tipo I), quando o tamanho de amostra é pequeno ($n \leq 50$), para qualquer número de variáveis. Em tamanhos de amostra pequenos, as taxas de erro tipo I aumentam expressivamente ao elevar os valores de p . Em $n= 100$ e 200 , ocorre redução do erro tipo I a um nível menor de 5% até $p= 9$ e 12 variáveis, respectivamente. Considerando maior número de variáveis, a redução do erro tipo I a níveis aceitáveis só ocorrerá em $n \geq 500$; e, iii. com o número máximo de variáveis ($p= 24$), pode-se afirmar que as fórmulas praticamente se igualam nos valores de taxas de erro tipo I a partir de $n \geq 500$, bem como, aumentos no p não provocam alterações significativas nas taxas de erro tipo I, que se mantêm próximas a 5% ; enquanto que, o poder aumenta nessa situação.

Tabela 8 – Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada, com quatro variáveis no grupo I; e oito e nove variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.

4+8									
	r=0,1	r=0,2	r=0,3	r=0,4	r=0,5	r=0,6	r=0,7	r=0,8	r=0,9
nFórmula para pequenas amostras.....								
25	10,85	22,45	41,10	61,60	78,85	90,10	98,10	99,80	100,00
30	10,40	28,45	53,40	76,55	91,50	97,90	99,70	100,00	100,00
50	19,05	59,40	90,50	97,95	99,95	100,00	100,00	100,00	100,00
100	48,15	97,55	100,00	100,00	100,00	100,00	100,00	100,00	100,00
200	89,20	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
300	99,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
400	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
500	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
600	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
750	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
...
5000	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
nFórmula para grandes amostras.....								
25	61,55	77,35	89,90	96,30	99,25	99,85	100,00	100,00	100,00
30	53,55	75,40	92,65	97,35	99,20	99,95	100,00	100,00	100,00
50	44,25	82,75	97,90	99,90	100,00	100,00	100,00	100,00	100,00
100	61,35	98,90	100,00	100,00	100,00	100,00	100,00	100,00	100,00
200	91,80	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
300	99,05	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
400	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
500	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
600	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
750	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
...
5000	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4+9									
	r=0,1	r=0,2	r=0,3	r=0,4	r=0,5	r=0,6	r=0,7	r=0,8	r=0,9
nFórmula para pequenas amostras.....								
25	12,45	23,25	39,90	56,20	76,40	91,35	96,50	99,75	100,00
30	12,00	28,15	51,00	73,95	88,80	97,80	99,65	100,00	100,00
50	21,10	59,40	89,00	98,95	100,00	100,00	100,00	100,00	100,00
100	48,80	97,20	99,95	100,00	100,00	100,00	100,00	100,00	100,00
200	88,75	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
300	99,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
400	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
500	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
600	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
750	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
...
5000	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
nFórmula para grandes amostras.....								
25	69,95	82,4	93,65	97,45	99,25	99,85	100	100	100
30	60,45	81,35	92,2	97,65	99,4	100	100	100	100
50	49,75	84,60	98,25	99,85	100,00	100	100	100	100
100	62,85	98,90	100,00	100,00	100,00	100	100	100	100
200	91,85	100,00	100,00	100,00	100,00	100	100	100	100
300	99,30	100,00	100,00	100,00	100,00	100	100	100	100
400	100,00	100,00	100,00	100,00	100,00	100	100	100	100
500	100,00	100,00	100,00	100,00	100,00	100	100	100	100
600	100,00	100,00	100,00	100,00	100,00	100	100	100	100
750	100,00	100,00	100,00	100,00	100,00	100	100	100	100
...
5000	100,00	100,00	100,00	100,00	100,00	100	100	100	100

4.2 Distribuição normal multivariada contaminada

A não-normalidade multivariada das amostras geradas no programa R foi confirmada pelo teste `mvShapiro.Test` (VILLASENOR ALVA; ESTRADA, 2009), em todas as 2000 simulações geradas por cenário. Além disso, o teste se mostrou confiável, pois manteve as taxas de erro tipo iguais ao nível mínimo de significância ($\alpha= 5\%$) adotado.

Após a confirmação da não-normalidade multivariada das amostras geradas no programa R, foi realizado o procedimento de simulação para avaliação das taxas de erro tipo I e poder do TRV para independência entre dois grupos de variáveis.

4.2.1 Erro tipo I

O comportamento da taxa de erro tipo I do TRV para a independência entre dois grupos de variáveis, quando os dados seguem a distribuição normal multivariada contaminada (Figuras 10 a 12 e APÊNDICES E a H) é muito diferente do observado quando do atendimento do pressuposto de normalidade multivariada.

Observa-se que não há diferenças nas taxas de erro tipo I entre as duas fórmulas de cálculo da estatística qui-quadrado (para pequenas e grandes amostras), em nenhum tamanho de amostra avaliado (25 a 5000) (Figuras 10 a 12 e APÊNDICES E a H), ou seja, a violação do pressuposto de normalidade multivariada afeta, de maneira similar, ambas estatísticas de teste (para pequenas e grandes amostras) do TRV para a independência entre dois grupos de variáveis.

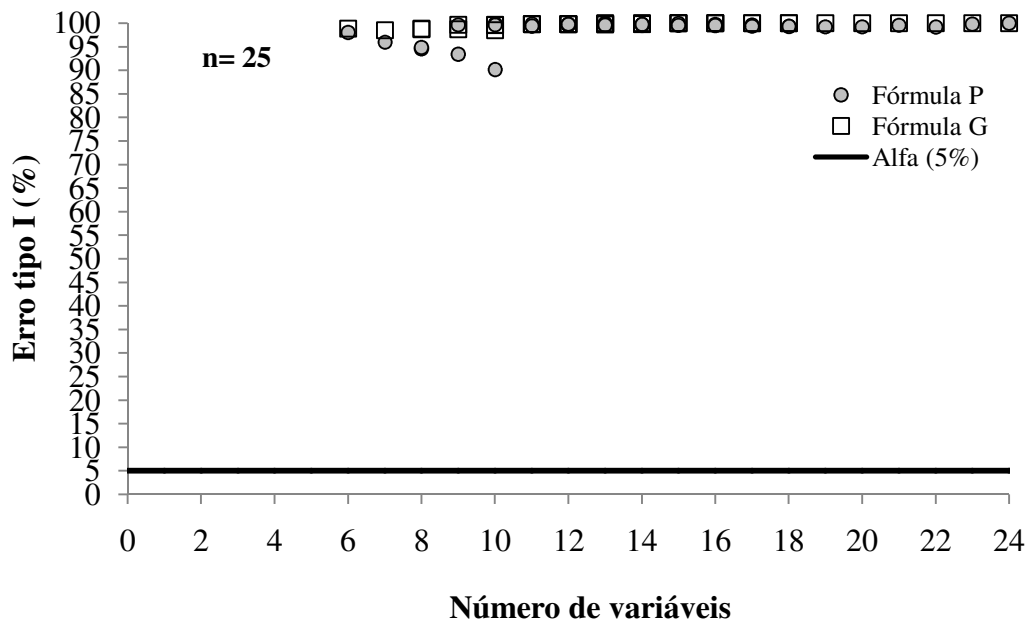


Figura 10 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis (combinações de variáveis dos dois grupos) e duas fórmulas: para pequenas e grandes amostras, no tamanho de amostra ($n= 25$), considerando a distribuição normal multivariada contaminada, em 2000 simulações.

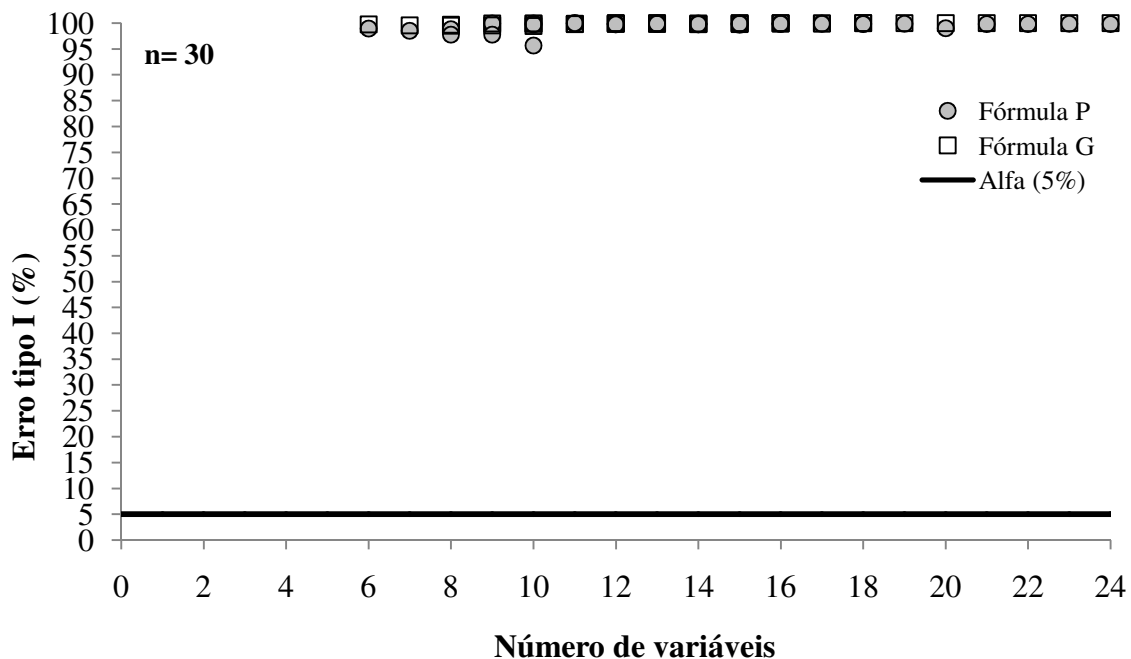


Figura 11 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis (combinações de variáveis dos dois grupos) e duas fórmulas: para pequenas e grandes amostras, no tamanho de amostra ($n= 30$), considerando a distribuição normal multivariada contaminada, em 2000 simulações.

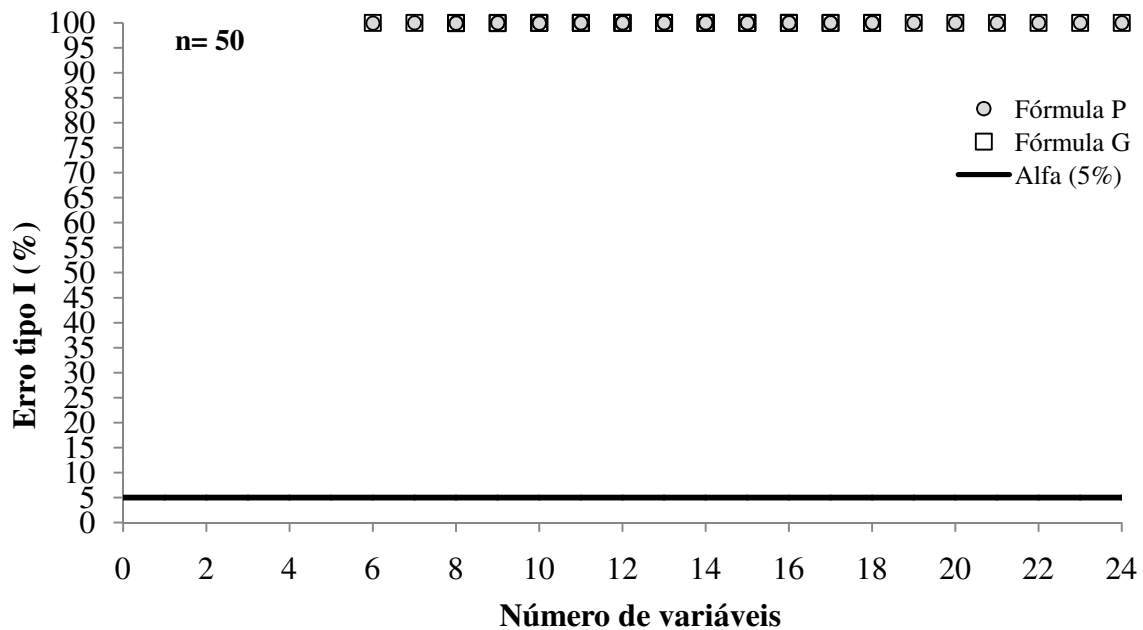


Figura 12 - Erro tipo I do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha = 5\%$), em diferentes números de variáveis (combinações de variáveis dos dois grupos) e duas fórmulas: para pequenas e grandes amostras, no tamanho de amostra ($n = 50$), considerando a distribuição normal multivariada contaminada, em 2000 simulações.

Nos casos em que a amostra multivariada foi gerada sob situação de normalidade multivariada, as diferenças nas taxas de erro tipo I entre as duas fórmulas foram muito acentuadas, considerando tamanhos de amostra menores ($n = 25, 30$ e 50) (Figuras 3 a 5). Provavelmente, o grande desvio de normalidade das amostras geradas faz com que a maior aproximação à distribuição qui-quadrado, dada pela correção de Bartlett, na fórmula para pequenas amostras, não tenha sido detectada.

Outra observação importante é o fato de que, em qualquer tamanho de amostra (25 a 5000) e número de variáveis estudadas (seis a 24), os valores das taxas de erro tipo I são sempre violados (Figuras 10 a 12 e APÊNDICES E a H), ou seja, superior ao nível mínimo de significância estabelecido (5%); sendo que a maioria é bem próxima a 100% . Quando houve normalidade multivariada, a violação do erro tipo I só ocorre em situações em que se usa a fórmula para pequenas amostras, com muitas variáveis ($p > 13$) e n pequeno ($n = 25$ e 30); enquanto que, ao se utilizar a fórmula para grandes amostras, o erro tipo I é violado em tamanhos de amostra pequenos ($n = 25, 30$ e 50), com qualquer número de variáveis ($p = 6$ a 24).

Em estudo de estimativa de componentes principais, utilizando o TRV, para testar se as matrizes de covariância têm autovalores diferentes e autovetores idênticos (H_0), ou se as

matrizes são arbitrárias (H_1), Boente et al. (2009) mostram um exemplo da influência de contaminação da distribuição normal, com $p=4$ variáveis e $n=100$. Neste estudo, a estatística do teste com distribuição normal foi de 1,7463 e o p -valor= 18,63%, levando a não rejeição da hipótese nula; ao inserir quatro outliers, no conjunto de dados 1, e três, no conjunto de dados 2, a estatística de teste foi 4,7880, correspondendo a um p -valor de 2,86% e à rejeição da hipótese nula. Isso mostra que, um pequeno número de outliers (contaminação) pode alterar significativamente os resultados das estatísticas de teste do TRV.

Em trabalho de Silva et al. (2008), de avaliação da robustez de dois testes para igualdade de matrizes de variâncias de k populações (teste de Bartlett multivariado e a sua versão *bootstrap*), em situação de não normalidade multivariada, os autores relatam as seguintes situações:

- a) Na distribuição uniforme, ambos os testes foram mais conservadores do que no caso da distribuição normal, pois a taxa de erro tipo I manteve-se muito abaixo de 5%; e,
- b) Na distribuição gama (com parâmetro $\lambda=1,5$ e $k=2$), o teste de Bartlett multivariado foi liberal em relação à taxa de erro tipo I, em todos os tamanhos de amostra avaliados (5 a 60), resultado que corrobora com os observados nesse trabalho. Além disso, as taxas de erro tipo I foram incrementadas ao aumentar-se o tamanho da população (k) de dois para cinco e 10. Porém, diferentemente deste trabalho, os autores observaram que quanto maior o n , maiores as taxas de erro tipo I. Conforme relatado pelos autores, esse resultado foi inesperado, levando a conclusão de que o teste de Bartlett multivariado para a igualdade de covariâncias de k populações é muito sensível a violações de normalidade e afetado por desvios de assimetria. Segundo os autores, resultados semelhantes já foram reportados por Zhang e Boss (1992). A conclusão de Silva et al. (2008) é de que o teste de *bootstrap* foi considerado superior ao assintótico (teste de Bartlett multivariado para a igualdade de covariâncias de k populações) e robusto, controlando o erro tipo I.

Outra situação de efeito de não normalidade em TRV para matrizes de covariâncias é descrito por Allison et al. (1999), que por meio de simulação, avaliaram a robustez do TRV em um estudo genético para características quantitativas de pares de irmãos. Para simular a não-normalidade, os autores utilizaram dois modelos: um modelo com distribuição marginal qui-quadrado com dois graus de liberdade para cada irmão (distribuição bem enviesada); e, outro modelo com a distribuição de Laplace-Padrão (EVANS et al., 1993), gerando uma cauda simétrica, mas leptocúrtica. Como resultados gerais, os autores relatam que alguns desvios de normalidade não produzem efeitos drásticos sobre as taxas de erro tipo I do TRV; enquanto que, outros, afetam significativamente as taxas de erro tipo I do teste, é o caso da

presença de acentuada curtose e assimetria, que resulta em elevadas taxas de erro tipo I, quando em presença de correlação entre irmãos. Estes resultados confirmam os reportados neste trabalho, de que as taxas de erro tipo I do TRV são muito elevadas quando as amostras apresentam grandes desvios de normalidade.

Veloso (2006), ao avaliar o desempenho do teste de identidade de modelos não lineares (cuja estatística de teste é dada pela razão de verossimilhanças), sob diferentes tamanhos amostrais, número de equações, robustez e diferentes modelos (Logístico, Gompertz e Langmuir) relata a ocorrência de taxas de erro tipo I elevadas, apenas em situação de tamanho de amostra pequeno ($n= 15$), e sob distribuição uniforme para o caso da aplicação do teste de identidade no modelo Logístico e Langmuir; porém, o teste torna-se conservativo ao se aumentar o tamanho de amostra para 50 e 100 e o número de equações a serem comparadas. Em relação ao poder, o autor observou que o teste mostrou valores de poder coerentes com o controle do erro tipo I para amostras maiores, $n = 50$ e $n = 100$; sendo que, para as amostras pequenas ($n = 15$), os valores de poder foram considerados adequados a partir de três equações. Estes resultados discordam dos observados nesse trabalho, no qual os valores das taxas de erro tipo I, sob distribuição não normal, são sempre elevados, independentemente do tamanho de amostra avaliado. Essa discordância pode ser devido ao fato de que, os desvios de normalidade naquele estudo, provavelmente foram menores do que os utilizados neste.

4.2.2 Poder

Nas tabelas 22 a 34, estão os valores de poder do TRV para a independência entre dois grupos de variáveis, em função dos diferentes cenários estabelecidos, nos quais variam os valores de: tamanho de amostra (n), número de variáveis (p), correlação (r) e fórmula para pequenas e grandes amostras, em condições de normalidade multivariada contaminada.

Os valores de poder do TRV para a independência entre dois grupos de variáveis, considerando a amostra multivariada proveniente da distribuição normal multivariada contaminada não apresentam grandes diferenças entre as duas fórmulas de cálculo da estatística de teste (para pequenas e grandes amostras). Resultado semelhante ocorreu com os valores das taxas de erro tipo I, do TRV para a independência entre dois grupos de variáveis, quando sob distribuição normal multivariada contaminada. Por outro lado, difere do ocorrido

no caso da distribuição normal multivariada, na qual, em situação de pequenos tamanhos de amostra ($n= 25$ e 30) e correlações fracas ($r=0,1$ e $0,2$), existem grandes diferenças entre as duas fórmulas.

Sob distribuição normal contaminada, os valores de poder do TRV para a independência entre dois grupos de variáveis, são sempre próximos a 100% e, praticamente, não se alteram com a modificação de tamanho de amostra, número de variáveis e grau de correlação. Lembrando que, nestas mesmas modificações, as taxas de erro tipo I foram sempre próximas a 100%, pode-se afirmar que estes valores elevados de poder não são reais. No caso de atendimento do pressuposto de normalidade, quando em tamanhos de amostra pequenos ($n= 25$ e 30) e na menor correlação ($r=0,1$), os valores de poder obtidos foram menores aos apresentados quando o pressuposto foi violado. Ainda, sob normalidade, o tamanho de amostra, o número de variáveis e o grau de correlação afetam fortemente o poder do teste, sendo que, no geral, aumentos nesses três fatores tendem a elevar o poder (Tabelas 2 a 21).

O fato do TRV da independência entre dois grupos de variáveis sob distribuição normal multivariada contaminada conduzir a elevadas taxas de erro tipo I e poder significa que a violação da normalidade multivariada, nesse teste, induz o pesquisador a rejeitar em torno de 100% das vezes a hipótese H_0 , de que dois grupos de variáveis são independentes, ou não correlacionados. Porém, a probabilidade de rejeitar H_0 , quando ela é falsa não foi afetada quando do não atendimento do referido pressuposto.

Situações em que se tem, simultaneamente, erro tipo I e poder elevados podem ocorrer e caracterizam testes liberais. Nestes casos, os valores de poder elevados não podem ser considerados reais (CANTELMO; FERREIRA, 2007). Ao avaliar o desempenho do teste de normalidade Shapiro-Wilk multivariado, os autores relatam tal situação, na qual o teste apresentou valores de erro tipo I sempre elevados, sendo próximos a 100%, ao aumentar o número de variáveis para um determinado tamanho de amostra fixo (n variou de 10 a 200), enquanto os valores de poder foram próximos a 100%, em qualquer situação.

Tabela 34 - Poder (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis sob normalidade multivariada contaminada, com 14 variáveis no grupo I e 10 variáveis no grupo II, em diferentes tamanhos de amostras (n) e graus de correlação (r) e fórmulas (para pequenas e grandes amostras), em 2000 simulações.

14+10									
n	r=0,1	r=0,2	r=0,3	r=0,4	r=0,5	r=0,6	r=0,7	r=0,8	r=0,9
.....Fórmula para pequenas amostras.....									
25	100,00	100,00	99,95	100,00	100,00	100,00	100,00	100,00	100,00
30	99,75	99,80	99,75	99,90	100,00	100,00	100,00	100,00	100,00
50	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
100	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
...
5000	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
.....Fórmula para grandes amostras.....									
14+10									
25	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
30	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
50	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
100	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
...
5000	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

Relatos sobre o efeito de não-normalidade em diversos testes de hipóteses para matrizes de covariância, baseados na razão de verossimilhança, são encontrados na literatura. Alisson et al. (1999) observaram que numa mistura de distribuições normais, aumentos no tamanho de amostra de 100 para 500 observações, conduziram há uma diminuição considerável nas taxas de erro tipo I do TRV; por outro lado, nas distribuições com elevados desvios de curtose, não há nenhuma melhora ao se aumentar o tamanho de amostra. Os autores justificaram o ocorrido devido a convergência lenta do TRV para limitar a distribuição qui-quadrado, quando as distribuições são muito leptocúrticas ou platicúrticas, ou em outras palavras, quando os dados são não normalmente distribuídos, a teoria da verossimilhança não garante que o TRV, se aproxima de uma distribuição qui-quadrado quando o tamanho da amostra aumenta. Curran et al. (1996) realizaram simulações de Monte Carlo para investigar o desempenho de três estatísticas de teste qui-quadrado (uma delas baseada na máxima verossimilhança) na análise fatorial confirmatória, sob variadas condições de tamanho da amostra, especificação do modelo e da distribuição multivariada. Em situação de normalidade multivariada e modelo correto, o teste baseado na máxima verossimilhança não mostrou nenhuma evidência de viés em todos os tamanhos de amostra (100 a 1000). Quanto maior a não-normalidade dos dados, maior a superestimação do teste. Para os modelos incorretos, a taxa de erro tipo I do teste foi inflacionada com o aumento de não-normalidade.

Muitos pesquisadores afirmam que quando há disponibilidade de tamanho de amostra maior do que 100, os testes de hipóteses apresentam elevada robustez quando o pressuposto de normalidade é violado, porém essa certeza só deve ser aceita para testes de hipóteses que envolvam médias (ALISSON et al., 1999), pois conforme Box (1953), os testes de hipótese para variâncias são menos robustos e sob não normalidade não apresentam a robustez tão útil apresentada pelos testes de comparação de médias. Ferreira (2008) salienta que desvios de normalidade afetam muito o desempenho dos testes sobre matrizes de covariâncias.

A explicação para os efeitos da não-normalidade nos TRV se deve ao fato de que a distribuição normal é completamente descrita pelos dois primeiros momentos, a média e a variância, sendo os momentos de terceira e quarta ordem (assimetria e curtose, respectivamente) iguais a zero. Se a distribuição é não normal, os valores dos momentos de terceira e quarta ordem são diferentes de zero. Nesse caso, o parâmetro de máxima verossimilhança estimado é consistente, mas não eficiente; e, o mínimo da função de máxima verossimilhança não é mais distribuído como uma grande amostra central do qui-quadrado (CURRAN et al.,1996). Aumentando-se os desvios de normalidade, a distribuição torna-se cada vez mais assimétrica, e é caracterizada por caudas mais grossas ou mais finas (em comparação com a curva normal), ou ambas; e então parâmetros adicionais são necessários para descrever esta distribuição mais complexa (CURRAN et al.,1996).

Na prática, isso significa que sob distribuição não normal, as estimativas obtidas pelo TRV serão viesadas, o que, em muitos destes testes, resulta em altas taxas de erro tipo I, levando o pesquisador a decidir pela rejeição da hipótese nula (em função da ausência de normalidade), quando ela não deveria ser rejeitada.

5 CONCLUSÕES

O TRV para a independência entre dois grupos de variáveis, sob distribuição normal multivariada, apresenta bom desempenho (taxas de erro tipo I reduzidas e poder elevado), quando o aumento no número de variáveis é acompanhado por acréscimos no tamanho amostral.

Sob distribuição normal multivariada, a fórmula para pequenas amostras do TRV para a independência entre dois grupos de variáveis deverá ser escolhida, pois apresenta desempenho superior ($n \leq 500$) ou igual ($n > 500$) à de grandes amostras.

Sob distribuição normal multivariada contaminada (20% de contaminação das observações, nas médias), o TRV para a independência entre dois grupos de variáveis não é robusto, pois apresenta elevadas taxas de erro tipo I, em qualquer combinação de n e p .

6 CONSIDERAÇÕES FINAIS

Em situação de normalidade multivariada: quando o número de variáveis é grande ($p=24$), o teste de razão de verossimilhança para a independência entre dois grupos de variáveis, controla as taxas de erro tipo I e possui poder elevado, quando se utiliza tamanhos de amostra superiores a 100 e 500, com uso das fórmulas para pequenas e grandes amostras, respectivamente; e, quando os tamanhos amostrais são pequenos ($n=25, 30$ e 50), o teste apresenta bom desempenho, desde que, o número de variáveis não exceda 12; e, a fórmula escolhida, seja para pequenas amostras.

Sob distribuição normal contaminada, especificamente, com 20% de contaminação das médias, o TRV para a independência entre dois grupos de variáveis não é robusto, pois, apesar de apresentar valores de poder próximo a 100%, estes não podem ser considerados reais, uma vez que, também possui taxas de erro tipo I sempre próximas a 100%, as quais não se alteram com as modificações nos cenários estudados.

Estudos futuros poderiam contemplar a avaliação da robustez do TRV da independência entre dois grupos de variáveis, em diferentes graus de contaminação da normal multivariada, bem como, sob diferentes tipos de distribuições não normais multivariadas.

REFERÊNCIAS

ALISSON, D. B. et al. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. **American Journal of Human Genetics**, Cambridge, v. 65. p.531–544, aug. 1999.

ALT, F. Multivariate Quality Control. In: KOTZ, S.; JOHNSON, N. L. (org.). **Encyclopedia of Statistical Sciences**. v.6. New York: John Wiley & Sons, 1985. p. 110–122.

ANDERSON, T. W. **An Introduction to Multivariate Statistical Analysis**, New York: John Wiley, 2003. 681p.

ANDREWS, D. F. et al. **Robust Estimates of Location: Survey and Advances**. Princeton: Princeton University Press, 1972. 373p.

BARTLETT, M.S. Further aspects of the theory of multiple regression. **Proceedings of the Cambridge Philosophical Society**, Cambridge, v. 34, n. 1. p. 33-40, jan. 1938.

BIASE, N. G. **Inferências sobre populações binomiais: testes frequentistas e bayesianos**. 2009. 152 f. Tese de Doutorado (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2009.

BOENTE, G.; PIRES, A. M.; RODRIGUES, I. M. Robust test for the common principal components model. **Journal of statistical planning and inference**, Indiana, v. 139, n. 4 p. 1332-1347, apr. 2009.

BORGES, L. C.; FERREIRA, D. F. Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normal e não-normais dos resíduos. **Revista Matemática e Estatística**, São Paulo, v. 21, n.1, p. 67-83, jan./abr. 2003.

BOX, G. Non-normality and tests on variances. **Biometrika**, London, v.40, p.318–335, dec.1953.

BRUM, B. **Relações entre variáveis de: sementes, plântulas, plantas, produção e óleo em mamoneira**. 2009. 120f. Dissertação (Mestrado em Agronomia) - Curso de Pós-graduação em Agronomia, Universidade Federal de Santa Maria, RS.

CANTELMO, N. F.; FERREIRA, D. F. Desempenho de testes de normalidade multivariados avaliados por simulação monte carlo. **Ciência e Agrotecnologia**, Lavras, v. 31, n. 6, p. 1630-1636, nov./dez. 2007.

CARMER, S. G. Optimal significance levels for application of the least significant difference in crop performance trials. **Crop Science**, Madison, v.16, n.1, p. 95-99, jan./fev. 1976.

CARVALHO, J. R. P. de. Comparação entre duas estatísticas para testar a hipótese de igualdade de coeficientes de regressão. **Pesquisa Agropecuária Brasileira**, Brasília, v.33, n.5, p.633-639, mai. 1998.

CARVALHO, F. I. F. et al. **Estimativas e implicações da correlação no melhoramento vegetal**. Pelotas : UFPel, 2004. 142p.

CHEN, J., ROMAGNOLI, J.A., A Strategy for Simultaneous Dynamic Data Reconciliation and Outlier Detection, **Computers and Chemical Engineering**, USA, v. 22, n. 4-5, p. 559-562, abr./mai. 1998.

COLENGHI, F. K. R. **Estudo comparativo de testes de hipótese multivariados para o vetor de médias via simulação de Monte Carlo**. 2008. 146f. Dissertação (Mestrado em Estatística) - Universidade Federal de Minas Gerais, Minas Gerais, 2008.

CRUZ, C.D., REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. 2. ed. Viçosa: UFV, 1997. 390 p.

CURRAN, P. J.; STEPHEN, G. W.; FINCH, J. F. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. **Psychological Methods**, Washington, v. 1, n.1, p.16-29, mar.1996.

DONOHU, D. L. Breakdown properties of multivariate location estimators. Ph.D. Qualifying Paper, Harvard University, 1982.

EVANS, M.; HATINGS, N.; PEACOCK, B. Laplace distribution. In: EVANS, M.; HATINGS, N.; PEACOCK, B. (eds) **Statistical distributions**. New York: John Wiley & Sons, p. 92-94, 1993.

FARIAS, A. C. **Avaliação de estratégias para reconciliação de dados e detecção de erros grosseiros**. 2009. 180f. Dissertação (Mestrado em Engenharia) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

FERREIRA, D. F. **Estatística Multivariada**. 1. ed. Lavras: Ed. da UFLA, 2008. v.1. 662 p.

FONSECA, J. S. de; MARTINS, G. de A. **Curso de estatística**. 6 ed. São Paulo: Atlas, 2009. 320p.

HAIR, J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009. 595p.

HAYTER, A. J.; TSUI, K.L. Identification and quantification in multivariate quality control problems, **Journal of Quality Technology**, Milwaukee, v.26, n.3, p.197-208, july, 1994.

HOTELLING, H. The most predictable criterion. **Journal of Educational Psychology**, Washington, v.26, n. 2, p. 139-142, feb. 1935.

HOTELLING, H. Relations between two sets of variates. **Biometrika**, London, v.28, n. 3-4, p. 321-377, 1936.

JOHNSON, R. A; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6.ed. New Jersey: Prentice Hall, 2007. 800p.

KENT, J. T. Robust Properties of Likelihood Ratio Test. **Biometrika**, London, v. 69, n. 1, apr. 1982.

LEHMANN, E. L. (2006) .On Likelihood Ratio Tests,.in **Optimality: The Second Erich L. Lehmann Symposium**, ed. by J. Rojo. Beachwood, OH: Institute of Mathematical Statistics, 1-8. Disponível em:< http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.lnms/1196283951>. Acesso em: 12 dez. 2011.

MACHADO, M. A. G., MAGALHAES, M. S., COSTA, A. F. B. Gráfico de controle de VMAX para o monitoramento da matriz de covariâncias. **Produção**, São Paulo, vol.18, no 2, p. 222-239, _____,2008.

MENÉNDEZ, J. A., RUEDA, C.; SALVADOR, B. Dominance of likelihood ratio tests under cone constraints. **Annals of statistics**, New York, v. 20, n.4, p.2087-2099, dec. 1992.

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

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

NEYMAN, J.; PEARSON, E. S. On the use and Interpretation of certain test criteria for purposes of statistical inference. **Biometrika**. London, v. 20A, n.1-2, p. 175-240 e p.263-294, dec. 1928.

OLIVEIRA, M. S. de. et al. **Introdução à estatística**. Lavras: Ed. da UFLA, 2009. 334p.

PERLMAN, M. D.; WU, L. The Emperor's new tests. **Statistical Science**, Cornell, v.14, n.4, p. 355–381, nov. 1999.

PINTO, L. P. **Estudo comparativo de testes de hipótese multivariados para matrizes de covariâncias via simulação de Monte Carlo e suas aplicações em Controle de Qualidade**. 2009. 109f. Dissertação (Mestrado em Estatística) - Universidade Federal de Minas Gerais, Minas Gerais, 2009.

PORTUGAL, M. S. Notas introdutórias sobre o princípio de máxima verossimilhança: estimação e testes de hipóteses. **DECON**, Porto Alegre, 1995. Disponível em: <<http://www.ufrgs.br/decon/publionline/textosdidaticos/Textodid04.pdf>>. Acesso em: 21 dez. 2011.

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2010. Disponível em: <http://www.Rproject.org>. Acesso em: 19 Abr. 2011.

RESENDE, M. D. V. de. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007. 561 p.

RIBEIRO, R. L. F. **Testes de hipótese multivariados para matrizes de covariâncias em processos autocorrelacionados com aplicações em controle de qualidade**. 2010. 138f. Dissertação (Mestrado em Estatística) - Universidade Federal de Minas Gerais, Minas Gerais, 2010.

SALSBURG, D. **Uma senhora toma chá...como a estatística revolucionou a ciência no século XX**. Rio de Janeiro: ZAHAR, 2009. 286p.

SCHWAAB, M.; PINTO, J. C., 2007, **Análises de Dados Experimentais I - Fundamentos de Estatística e Estimação de Parâmetros**. 1 ed. Rio de Janeiro: e-papers, 2007, 462p.

SILVA, R. A.; MINGOTI, S. **Estudos de testes estatísticos para matriz de covariâncias de processos multivariados**, Relatório técnico de iniciação científica para fapemig/ prpf, Universidade Federal de Minas Gerais, 2008.

SILVA, R. B. V.; FERREIRA, D. F. Comparação da robustez de três alternativas do teste de igualdade de duas médias populacionais sob não-normalidade por meio de simulação Monte Carlo. **Acta Scientiarum (UEM)**, Maringá, v. 24, n.1, p. 1771-1776, 2002.

SILVA, R. B. V.; FERREIRA, D. F.; NOGUEIRA, D. A. Robustness of asymptotic and bootstrap tests for multivariate homogeneity of covariance matrices. **Ciência e Agrotecnologia**, Lavras, v. 32, n.1, p. 157-161, jan./fev. 2008.

STAHEL, W. **Robust estimation: infinitesimal optimality and covariance matrix estimation**. Thesis, ETH, Zurich (in German), 1981.

SULLIVAN, J. H. et al. Step-down analysis for change in the covariance matrix and other parameters, **Journal of Quality Technology**, Milwaukee, v. 39, n.1, p.66-84, jan. 2007.

VELOSO, R. B. **Verificação por simulação monte Carlo da eficiência da aplicação do teste de razão de verossimilhança para identidade de modelos lineares**. 2006. 65 f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2006.

VILLASENOR ALVA , J. A.; ESTRADA, E. G. A Generalization of Shapiro–Wilk's Test for Multivariate Normality Preview. **Communications in Statistics - Theory and Methods**, Ontario, v. 38, n. 11, p. 1870-1883, may. 2009.

WREGGE, M. S. et al. **Zoneamento agroclimático para mamona no Rio Grande do Sul**. Pelotas: Embrapa Clima Temperado, 2007. 30p. (Documentos, 192).

YANAGIHARA , H.; TONDA, T.; MATSUMOTO, C. The effects of nonnormality on asymptotic distributions of some likelihood ratio criteria for testing covariance structures under normal assumption. **Journal of Multivariate Analysis**, USA, v.96, p. 237-264, oct. 2005.

ZHANG, J. I.; BOOS, D. D. Bootstrap critical values for testing homogeneity of covariance matrices. **Journal of the American Statistical Association**, Alexandria, v. 87, n. 418, p. 425-429, June 1992.

APÊNDICES

Apêndice A – Taxas de Erro tipo I (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis e duas fórmulas (para pequenas e grandes amostras) de cálculo da estatística de teste χ^2 , no tamanho de amostra (n= 25, 30, 50 e 100), sob distribuição normal multivariada, em 2000 simulações.

Nº de variáveis do G1 e G2 ¹	Total de variáveis	n= 25		n= 30		n= 50		n= 100	
		FP ²	FG ³	FP	FG	FP	FG	FP	FG
3x3	6	4,15	13,20	4,70	11,00	4,20	8,10	5,70	5,90
3x4	7	4,75	15,55	4,85	13,15	5,00	8,50	5,30	6,35
3x5	8	4,55	19,55	4,70	15,85	5,90	10,10	4,55	8,05
4x4	8	4,50	20,95	4,90	17,95	5,70	10,75	5,85	7,20
3x6	9	5,60	24,20	5,30	19,65	4,95	11,15	4,70	6,90
4x5	9	5,70	24,70	5,40	19,85	4,50	11,05	5,10	7,35
3x7	10	5,65	28,45	5,10	22,70	5,60	14,15	4,70	8,40
4x6	10	5,15	32,10	4,60	24,55	4,95	14,10	4,70	8,15
5x5	10	4,95	32,60	4,45	24,10	5,35	13,65	4,15	9,55
3x8	11	6,10	36,65	5,10	26,20	4,90	14,20	5,85	8,45
4x7	11	6,00	38,65	4,95	30,55	5,55	15,95	5,10	9,65
5x6	11	5,85	41,25	5,35	30,65	5,85	16,30	4,90	8,70
3x9	12	6,15	42,95	5,35	32,85	4,85	16,50	5,80	10,40
4x8	12	5,70	46,65	4,50	35,30	4,45	18,05	5,65	10,65
5x7	12	5,05	48,90	5,45	38,45	5,50	19,15	5,05	10,40
6x6	12	5,05	50,10	5,30	38,10	5,20	20,10	4,55	9,85
3x10	13	6,75	51,15	6,10	38,60	5,95	18,20	5,15	8,90
4x9	13	6,85	55,05	5,95	43,40	6,45	20,60	5,00	9,90
5x8	13	6,55	58,90	5,45	46,30	5,60	22,50	4,55	9,80
6x7	13	6,45	61,40	5,80	45,15	5,20	23,65	4,90	9,40
4x10	14	7,35	66,95	6,50	50,85	5,50	23,50	4,70	11,85
5x9	14	6,20	70,25	6,25	53,65	5,10	25,35	5,20	12,95
6x8	14	7,85	71,75	6,55	55,50	5,85	26,85	5,30	12,85
7x7	14	6,90	72,65	5,95	56,16	5,60	26,90	5,20	13,40
5x10	15	7,85	80,40	6,45	63,00	3,90	28,55	5,30	12,65
6x9	15	8,15	81,95	5,85	65,80	4,80	29,40	5,15	13,40
7x8	15	8,05	81,65	6,15	65,50	4,80	30,45	5,50	13,90
6x10	16	9,10	89,10	8,30	73,75	4,95	32,90	6,50	14,70
7x9	16	8,45	89,50	7,65	75,85	5,75	36,20	5,50	15,10
8x8	16	7,90	90,85	6,60	76,30	5,00	36,40	5,35	15,40
7x10	17	12,05	94,70	6,95	82,65	9,15	42,80	5,10	16,25
8x9	17	11,85	95,35	7,40	83,85	5,35	41,10	4,95	17,80
8x10	18	12,60	98,25	8,65	89,55	6,00	47,30	5,55	17,55
9x9	18	12,75	98,50	8,55	90,40	5,50	48,00	5,20	19,75
9x10	19	17,20	99,55	9,35	94,85	6,60	55,30	5,65	21,95
10x10	20	22,70	99,90	10,90	97,90	6,55	62,95	5,65	22,65
11x10	21	31,45	100,0	12,60	99,10	6,10	70,15	5,85	25,60
12x10	22	39,65	100,0	17,15	99,70	7,30	75,30	4,75	29,55
13x10	23	58,85	100,0	20,65	99,95	6,85	83,05	5,00	31,05
14x10	24	80,80	100,0	25,35	100,0	7,95	87,55	5,55	35,60

¹G1 e G2, são o grupo 1 e o grupo 2 de variáveis, do TRV para a independência entre dois grupos de variáveis, respectivamente. ² FP, fórmula para pequenas amostras. ³FG, fórmula para grandes amostras,

Apêndice B – Taxas de Erro tipo I (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis e duas fórmulas (para pequenas e grandes amostras) de cálculo da estatística de teste χ^2 , no tamanho de amostra (n= 200, 300, 400 e 500), sob distribuição normal multivariada, em 2000 simulações.

Nº de variáveis do G1 e G2 ¹	Total de variáveis	n= 200		n= 300		n= 400		n= 500	
		FP ²	FG ³	FP	FG	FP	FG	FP	FG
3x3	6	4,30	6,45	4,85	6,25	4,35	5,25	4,85	5,30
3x4	7	4,95	5,95	5,35	5,45	4,70	4,85	4,65	5,65
3x5	8	4,55	6,50	5,55	5,7	5,75	4,20	5,85	5,70
4x4	8	5,25	5,70	5,10	5,75	5,35	5,25	5,35	5,25
3x6	9	5,85	7,45	4,45	5,05	5,35	5,70	5,15	5,70
4x5	9	5,25	5,45	3,9	5,4	4,90	5,05	5,80	5,20
3x7	10	4,10	7,10	5,25	6,2	5,05	5,90	5,70	5,35
4x6	10	5,45	7,65	5,25	6,75	4,65	6,00	5,40	5,15
5x5	10	4,70	7,45	5,35	7,05	4,50	5,60	5,00	5,00
3x8	11	4,00	6,55	4,6	6,05	4,50	5,50	5,85	5,65
4x7	11	4,15	6,05	4,25	5,3	5,60	6,05	4,85	5,55
5x6	11	4,90	6,40	4,70	6,35	4,70	6,40	5,05	6,00
3x9	12	5,40	6,45	4,55	7,2	5,10	5,20	5,00	5,95
4x8	12	5,25	6,25	5,35	6,65	5,10	5,85	4,70	6,10
5x7	12	5,40	6,20	5,30	6,15	5,65	5,45	3,65	6,45
6x6	12	5,95	7,15	4,95	6,20	5,10	6,30	4,40	6,30
3x10	13	4,95	6,60	5,35	6,25	4,15	5,85	5,45	6,05
4x9	13	4,95	7,55	5,8	5,8	5,00	5,85	5,00	6,65
5x8	13	4,60	7,45	6,30	6,60	5,90	6,40	4,85	6,25
6x7	13	4,65	8,00	5,55	5,95	5,05	6,05	4,80	5,45
4x10	14	5,95	8,60	4,7	6,05	5,75	5,05	5,00	5,65
5x9	14	5,55	8,50	3,95	6,80	5,20	7,15	4,15	5,15
6x8	14	5,75	7,90	4,50	6,90	6,35	6,60	4,80	5,60
7x7	14	5,35	7,55	5,15	6,75	6,50	6,00	4,45	5,95
5x10	15	4,70	8,00	5,55	7,40	5,30	6,50	4,75	5,95
6x9	15	5,75	8,40	5,65	7,75	4,90	6,85	5,55	6,75
7x8	15	4,55	8,90	4,25	8,10	6,20	6,65	5,30	6,50
6x10	16	5,25	7,85	4,60	6,65	5,50	6,30	5,45	6,65
7x9	16	4,80	8,85	4,85	7,35	5,30	6,30	4,25	6,30
8x8	16	5,60	9,65	4,40	6,95	5,10	7,50	5,50	5,85
7x10	17	5,55	8,75	5,90	7,30	5,25	6,70	4,60	6,35
8x9	17	5,00	8,20	5,00	7,30	4,45	6,60	5,05	5,45
8x10	18	4,80	10,50	4,65	8,15	5,95	7,65	4,50	6,60
9x9	18	4,95	11,00	5,15	8,10	5,90	6,70	4,55	6,50
9x10	19	5,30	9,70	5,05	9,05	5,05	8,55	5,00	6,60
10x10	20	5,90	10,60	5,20	8,85	6,00	6,35	5,25	6,80
11x10	21	5,25	11,40	5,05	9,40	5,25	8,35	4,60	7,60
12x10	22	4,95	11,80	5,80	9,95	5,75	8,75	4,85	8,10
13x10	23	5,40	14,45	4,90	10,80	4,55	9,20	5,05	8,00
14x10	24	4,70	16,10	5,20	11,95	4,85	9,05	4,85	8,25

¹G1 e G2, são o grupo 1 e o grupo 2 de variáveis, do TRV para a independência entre dois grupos de variáveis, respectivamente. ²FP, fórmula para pequenas amostras. ³FG, fórmula para grandes amostras.

Apêndice C – Taxas de Erro tipo I (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis e duas fórmulas (para pequenas e grandes amostras) de cálculo da estatística de teste χ^2 , no tamanho de amostra (n= 600, 750, 1000 e 1500), sob distribuição normal multivariada, em 2000 simulações.

Nº de variáveis do G1 e G2 ¹	Total de variáveis	n= 600		n= 750		n= 1000		n= 1500	
		FP ²	FG ³	FP	FG	FP	FG	FP	FG
3x3	6	4,80	4,85	4,50	4,75	5,30	4,90	4,25	5,50
3x4	7	5,35	6,10	3,60	5,05	4,75	5,35	5,00	4,60
3x5	8	4,35	4,85	4,75	5,40	5,45	5,40	4,35	5,95
4x4	8	5,10	6,20	5,30	3,55	4,85	4,95	4,85	4,75
3x6	9	5,35	4,75	4,25	4,95	5,35	5,50	4,40	4,55
4x5	9	4,85	5,95	4,70	5,10	4,40	4,45	4,95	4,85
3x7	10	3,75	6,15	4,85	5,90	5,10	5,65	4,95	5,65
4x6	10	4,30	6,05	5,40	5,25	5,05	5,95	4,40	4,90
5x5	10	4,85	5,45	5,75	5,85	4,50	5,50	4,65	5,10
3x8	11	4,65	5,70	4,90	6,65	4,10	4,85	5,40	5,30
4x7	11	4,20	4,80	4,45	4,75	5,05	5,30	5,10	5,30
5x6	11	5,55	4,80	4,85	4,70	5,40	5,00	5,00	5,50
3x9	12	5,05	4,90	4,90	5,65	5,60	5,55	4,55	5,10
4x8	12	5,10	5,30	4,90	5,40	5,70	5,10	4,80	5,45
5x7	12	4,85	5,60	4,60	5,70	4,90	5,80	4,85	4,70
6x6	12	4,45	5,60	5,00	5,65	4,85	5,40	4,55	5,70
3x10	13	5,35	6,20	4,85	5,80	4,20	5,40	4,90	4,50
4x9	13	4,85	6,90	4,35	5,35	4,45	5,60	5,25	4,00
5x8	13	4,80	6,55	4,10	5,30	4,80	5,15	5,05	4,10
6x7	13	4,25	5,80	4,65	5,45	4,90	6,15	4,75	4,15
4x10	14	5,55	5,40	5,10	6,40	5,15	5,65	5,75	5,15
5x9	14	5,00	5,60	5,30	6,45	4,95	5,20	5,50	5,25
6x8	14	5,10	5,45	5,10	6,75	4,30	4,95	5,45	5,15
7x7	14	4,50	5,35	4,00	6,25	3,85	5,45	4,95	5,65
5x10	15	4,70	5,75	4,35	5,55	5,40	5,90	4,55	5,60
6x9	15	4,90	5,40	4,75	5,05	5,20	4,45	4,25	5,60
7x8	15	5,35	6,20	5,10	5,60	4,80	5,50	4,10	5,40
6x10	16	5,50	6,10	4,85	5,90	5,30	5,90	5,95	5,25
7x9	16	5,35	6,35	4,70	5,30	5,30	4,55	4,50	5,50
8x8	16	4,70	6,15	5,15	5,40	4,70	4,95	5,70	5,05
7x10	17	5,15	7,10	4,45	5,50	5,35	5,40	4,35	5,20
8x9	17	4,55	6,50	4,60	4,95	4,35	5,45	5,10	5,40
8x10	18	4,55	6,70	5,15	5,90	3,85	5,65	5,30	5,05
9x9	18	4,90	6,55	5,40	5,90	3,85	5,80	4,85	5,60
9x10	19	4,75	6,35	5,70	5,55	4,00	6,20	4,75	5,65
10x10	20	5,60	6,10	5,55	6,20	5,50	6,65	5,25	5,05
11x10	21	4,60	7,20	4,95	5,75	4,90	6,70	5,00	6,90
12x10	22	4,40	6,70	4,00	5,80	4,70	5,80	5,75	6,05
13x10	23	5,90	6,95	5,15	6,20	5,05	6,55	4,25	6,20
14x10	24	5,20	7,10	5,00	6,70	5,70	6,60	4,90	5,80

¹G1 e G2, são o grupo 1 e o grupo 2 de variáveis, do TRV para a independência entre dois grupos de variáveis, respectivamente. ²FP, fórmula para pequenas amostras, ³FG, fórmula para grandes amostras.

Apêndice D – Taxas de Erro tipo I (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis e duas fórmulas (para pequenas e grandes amostras) de cálculo da estatística de teste χ^2 , no tamanho de amostra ($n= 2000, 3000, 4000$ e 5000), sob distribuição normal multivariada, em 2000 simulações.

Nº de variáveis do G1 e G2 ¹	Total de variáveis	n= 2000		n= 3000		n= 4000		n= 5000	
		FP ²	FG ³	FP	FG	FP	FG	FP	FG
3x3	6	5,70	4,90	4,30	5,60	4,95	4,45	5,60	4,25
3x4	7	5,75	5,05	4,60	6,05	4,80	4,30	5,16	5,05
3x5	8	4,40	4,50	4,40	4,10	4,95	4,70	4,65	4,95
4x4	8	4,50	5,50	4,40	5,65	4,30	5,00	5,10	5,60
3x6	9	5,15	5,30	4,90	4,90	5,05	5,15	5,90	5,95
4x5	9	5,45	5,45	4,80	4,40	4,25	5,05	5,70	4,20
3x7	10	3,70	5,55	5,15	5,80	4,80	5,50	5,10	4,80
4x6	10	5,05	6,00	5,65	5,25	4,90	5,65	5,25	5,00
5x5	10	4,45	5,70	4,80	5,60	4,40	5,50	5,40	4,55
3x8	11	4,50	5,35	5,25	4,75	5,80	4,40	5,40	5,65
4x7	11	5,00	5,25	5,05	5,15	4,50	4,55	4,75	4,85
5x6	11	4,90	5,40	5,15	5,55	4,75	5,15	4,75	4,80
3x9	12	5,85	5,05	4,75	4,95	4,45	5,90	5,10	4,65
4x8	12	5,30	5,05	4,55	4,50	5,15	4,50	4,40	4,75
5x7	12	5,80	3,75	4,95	4,95	4,85	5,30	4,25	5,35
6x6	12	5,85	4,45	4,90	3,55	4,90	4,75	4,55	5,00
3x10	13	5,45	5,70	4,90	4,95	5,70	5,35	4,95	5,25
4x9	13	4,95	5,50	5,00	5,70	5,70	4,55	4,45	5,35
5x8	13	5,00	5,35	5,20	5,95	4,95	5,10	4,65	5,35
6x7	13	4,90	5,35	4,65	5,15	4,15	5,55	5,25	5,80
4x10	14	5,60	5,40	4,45	6,00	4,40	5,00	5,35	4,40
5x9	14	5,55	5,35	4,30	5,85	5,40	5,00	5,35	4,35
6x8	14	5,20	5,55	4,15	5,00	5,30	4,75	5,30	5,00
7x7	14	4,45	5,85	4,95	5,80	5,20	5,15	5,50	4,80
5x10	15	5,25	5,15	5,55	5,00	5,10	4,85	4,60	5,00
6x9	15	5,60	5,25	4,15	5,10	4,65	5,60	5,50	4,95
7x8	15	5,40	5,30	5,05	4,90	4,95	5,30	5,05	4,85
6x10	16	4,35	6,45	4,65	5,10	4,60	5,05	4,10	5,00
7x9	16	4,80	5,10	4,15	4,75	5,15	5,10	5,10	5,15
8x8	16	5,05	5,45	4,95	5,30	5,35	4,70	4,90	3,90
7x10	17	5,45	5,35	4,75	5,60	4,80	5,00	5,00	4,80
8x9	17	5,40	6,10	5,20	5,70	5,35	5,10	5,35	5,40
8x10	18	4,90	5,05	4,35	5,05	5,10	5,35	4,05	5,15
9x9	18	5,30	5,00	3,85	5,65	5,15	4,45	4,05	5,20
9x10	19	4,95	5,85	4,20	5,40	5,80	5,15	5,25	4,95
10x10	20	4,30	5,55	5,40	4,80	4,55	5,35	5,00	4,80
11x10	21	5,40	5,25	4,55	5,25	5,25	7,00	5,05	5,05
12x10	22	5,10	5,40	5,40	5,20	5,20	5,60	5,55	4,90
13x10	23	5,35	4,75	5,05	5,35	5,20	4,85	5,65	4,50
14x10	24	4,35	5,20	4,75	5,05	5,85	6,10	4,90	5,15

¹G1 e G2, são o grupo 1 e o grupo 2 de variáveis, do TRV para a independência entre dois grupos de variáveis, respectivamente. ²FP, fórmula para pequenas amostras. ³FG, fórmula para grandes amostras.

Apêndice E - Taxas de Erro tipo I (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis e duas fórmulas (para pequenas e grandes amostras) de cálculo da estatística de teste χ^2 , no tamanho de amostra (n= 25, 30, 50 e 100), sob distribuição normal multivariada contaminada, em 2000 simulações.

Nº de variáveis do GI e G2 ¹	Total de variáveis	n= 25		n= 30		n= 50		n= 100	
		FP ²	FG ³	FP	FG	FP	FG	FP	FG
3x3	6	98,00	98,85	98,95	99,75	99,95	100,00	100,00	100,00
3x4	7	96,00	98,45	98,50	99,60	100,00	100,00	100,00	100,00
3x5	8	94,55	98,80	98,8	99,70	100,00	100,00	100,00	100,00
4x4	8	94,80	98,70	97,75	99,50	99,95	99,95	100,00	100,00
3x6	9	99,65	99,65	99,95	99,95	100,00	100,00	100,00	100,00
4x5	9	93,40	98,70	97,75	99,55	100,00	99,95	100,00	100,00
3x7	10	99,85	99,65	99,95	99,95	100,00	100,00	100,00	100,00
4x6	10	99,65	99,6	99,75	99,90	100,00	100,00	100,00	100,00
5x5	10	90,10	98,45	95,70	99,45	99,95	100,00	100,00	100,00
3x8	11	99,40	99,75	100,00	99,85	100,00	100,00	100,00	100,00
4x7	11	99,75	99,75	99,90	99,85	100,00	100,00	100,00	100,00
5x6	11	99,45	99,75	100,00	99,90	100,00	100,00	100,00	100,00
3x9	12	99,70	99,80	99,90	99,90	100,00	100,00	100,00	100,00
4x8	12	99,60	99,80	99,85	100,00	100,00	100,00	100,00	100,00
5x7	12	99,70	99,75	99,90	99,95	100,00	100,00	100,00	100,00
6x6	12	99,80	99,75	99,75	99,90	100,00	100,00	100,00	100,00
3x10	13	99,55	99,85	99,90	99,95	100,00	100,00	100,00	100,00
4x9	13	99,90	99,70	100,0	99,90	100,00	100,00	100,00	100,00
5x8	13	99,70	99,85	99,85	100,00	100,00	100,00	100,00	100,00
6x7	13	99,55	99,95	99,90	99,90	100,00	100,00	100,00	100,00
4x10	14	99,70	99,85	99,95	99,95	100,00	100,00	100,00	100,00
5x9	14	99,70	99,90	99,80	99,85	100,00	100,00	100,00	100,00
6x8	14	99,50	99,95	99,75	99,85	100,00	100,00	100,00	100,00
7x7	14	99,70	99,75	99,80	99,85	100,00	100,00	100,00	100,00
5x10	15	99,55	99,95	99,95	99,85	100,00	100,00	100,00	100,00
6x9	15	99,90	100,00	99,70	99,95	100,00	100,00	100,00	100,00
7x8	15	99,60	100,00	99,90	100,00	100,00	100,00	100,00	100,00
6x10	16	99,70	100,00	99,90	99,95	100,00	100,00	100,00	100,00
7x9	16	99,50	100,00	99,90	99,95	100,00	100,00	100,00	100,00
8x8	16	99,50	100,00	99,90	100,00	100,00	100,00	100,00	100,00
7x10	17	99,65	99,95	99,95	99,95	100,00	100,00	100,00	100,00
8x9	17	99,40	100,00	99,90	99,95	100,00	100,00	100,00	100,00
8x10	18	99,30	100,00	99,90	100,00	100,00	100,00	100,00	100,00
9x9	18	99,30	100,00	99,85	100,00	100,00	100,00	100,00	100,00
9x10	19	99,25	100,00	99,90	100,00	100,00	100,00	100,00	100,00
10x10	20	99,25	100,00	99,90	100,00	100,00	100,00	100,00	100,00
11x10	21	99,60	100,00	99,85	100,00	100,00	100,00	100,00	100,00
12x10	22	99,15	100,00	99,85	100,00	100,00	100,00	100,00	100,00
13x10	23	99,75	100,00	99,90	100,00	100,00	100,00	100,00	100,00
14x10	24	99,95	100,00	99,85	100,00	100,00	100,00	100,00	100,00

¹GI e G2, são o grupo 1 e o grupo 2 de variáveis, do TRV para a independência entre dois grupos de variáveis, respectivamente. ²FP, fórmula para pequenas amostras. ³FG, fórmula para grandes amostras.

Apêndice F - Taxas de Erro tipo I (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis e duas fórmulas (para pequenas e grandes amostras) de cálculo da estatística de teste χ^2 , no tamanho de amostra (n= 200, 300, 400 e 500), sob distribuição normal multivariada contaminada, em 2000 simulações.

Nº de variáveis do GI e G2 ¹	Total de variáveis	n= 200		n= 300		n= 400		n= 500	
		FP ²	FG ³	FP	FG	FP	FG	FP	FG
3x3	6	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x4	7	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x5	8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x4	8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x6	9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x5	9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x7	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x6	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x5	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x8	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x7	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x6	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x9	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x8	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x7	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x6	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x10	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x9	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x8	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x7	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x10	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x9	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x8	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x7	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x10	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x9	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x8	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x10	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x9	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x8	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x10	17	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x9	17	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x10	18	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
9x9	18	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
9x10	19	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
10x10	20	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
11x10	21	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
12x10	22	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
13x10	23	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
14x10	24	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

¹GI e G2, são o grupo 1 e o grupo 2 de variáveis, do TRV para a independência entre dois grupos de variáveis, respectivamente. ²FP, fórmula para pequenas amostras. ³FG, fórmula para grandes amostras.

Apêndice G - Taxas de Erro tipo I (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis e duas fórmulas (para pequenas e grandes amostras) de cálculo da estatística de teste χ^2 , no tamanho de amostra (n= 600, 750, 1000 e 1500), sob distribuição normal multivariada contaminada, em 2000 simulações.

Nº de variáveis do GI e G2 ¹	Total de variáveis	n= 600		n= 750		n= 1000		n= 1500	
		FP ²	FG ³	FP	FG	FP	FG	FP	FG
3x3	6	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x4	7	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x5	8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x4	8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x6	9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x5	9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x7	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x6	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x5	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x8	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x7	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x6	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x9	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x8	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x7	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x6	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x10	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x9	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x8	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x7	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x10	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x9	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x8	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x7	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x10	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x9	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x8	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x10	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x9	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x8	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x10	17	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x9	17	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x10	18	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
9x9	18	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
9x10	19	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
10x10	20	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
11x10	21	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
12x10	22	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
13x10	23	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
14x10	24	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

¹GI e G2, são o grupo 1 e o grupo 2 de variáveis, do TRV para a independência entre dois grupos de variáveis, respectivamente. ²FP, fórmula para pequenas amostras. ³FG, fórmula para grandes amostras.

Apêndice H - Taxas de Erro tipo I (%) do teste de razão de verossimilhança para a independência entre dois grupos de variáveis ($\alpha= 5\%$), em diferentes números de variáveis e duas fórmulas (para pequenas e grandes amostras) de cálculo da estatística de teste χ^2 , no tamanho de amostra (n= 2000, 3000, 4000, 5000), sob distribuição normal multivariada contaminada, em 2000 simulações.

Nº de variáveis do GI e G2 ¹	Total de variáveis	n= 2000		n= 3000		n= 4000		n= 5000	
		FP ²	FG ³	FP	FG	FP	FG	FP	FG
3x3	6	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x4	7	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x5	8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x4	8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x6	9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x5	9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x7	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x6	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x5	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x8	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x7	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x6	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x9	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x8	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x7	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x6	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3x10	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x9	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x8	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x7	13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
4x10	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x9	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x8	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x7	14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5x10	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x9	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x8	15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6x10	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x9	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x8	16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7x10	17	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x9	17	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8x10	18	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
9x9	18	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
9x10	19	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
10x10	20	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
11x10	21	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
12x10	22	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
13x10	23	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
14x10	24	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

¹GI e G2, são o grupo 1 e o grupo 2 de variáveis, do TRV para a independência entre dois grupos de variáveis, respectivamente. ²FP, fórmula para pequenas amostras. ³FG, fórmula para grandes amostras.

Apêndice I – Rotinas de simulação no Programa R

Exemplo de rotina de simulação no Programa R para gerar as amostras normais multivariadas e para cálculo das taxas de erro tipo I

```

rm(list=ls())
simula <- 2000
saida <- as.data.frame(matrix(0,simula,7))
names(saida) <-
c("nobs","pvariaveis","pvalorShapiro","pvariaveisgrupo1","pvariaveisgrupo2","qui","pvalorqui")
for(j in 1:simula)
{
sigma <- matrix(c(40.3282,0.0000,0.0000,0.0000,0.0000,0.0000,
0.0000,7134.1381,0.0000,0.0000,0.0000,0.0000,
0.0000,0.0000,16.4368,0.0000,0.0000,0.0000,
0.0000,0.0000,0.0000,89.7463,0.0000,0.0000,
0.0000,0.0000,0.0000,0.0000,54.7902,0.0000,
0.0000,0.0000,0.0000,0.0000,0.0000,160.0415),6,6)
mu <- c(48.7200,578.4340,25.6100,31.4256,9.1600,40.4371)
n <-25
library(MASS)
x <- mvrnorm(n, mu, sigma)
mvw <- mvShapiro.Test(x)
s = var(x)
g1 = x[,1:3]
s1 = var(g1)
g2 = x[,4:6]
s2 = var(g2)
qui = -((dim(x)[1]-1)-((dim(x)[2]+1)/2))*(log(det(s)/(det(s1)*det(s2))))
pvalor = pchisq(qui, df=(dim(g1)[2]*dim(g2)[2]), ncp=0, lower.tail = FALSE, log.p = FALSE)
saida$nobs[j] <- n
saida$pvariaveis[j] <- dim(x)[2]
saida$pvalorShapiro[j] <- as.numeric(mvw[2])
saida$pvariaveisgrupo1[j] <- dim(g1)[2]
saida$pvariaveisgrupo2[j] <- dim(g2)[2]
saida$qui[j] <- qui
saida$pvalorqui[j] <- pvalor
}
write.table(saida,"C:\\R\\Betania2011\\errotipoI\\3X3\\peq25.csv", sep = ";", dec = ".", row.names = FALSE)
# Calculo do Erro Tipo I do teste qui quadrado para pequenas amostras com nível nominal de 5%
y = as.numeric(saida[,7])
ErroTipoI = length(y[y <= 0.05])/simula*100
ErroTipoI

```

Exemplo de rotina de simulação no Programa R para gerar as amostras normais multivariadas e para cálculo do poder (r= 0,1)

```

rm(list=ls())
simula <- 2000
saida <- as.data.frame(matrix(0,simula,7))
names(saida) <-
c("nobs","pvariaveis","pvalorShapiro","pvariaveisgrupo1","pvariaveisgrupo2","qui","pvalorqui")
for(j in 1:simula)
{
sigma <- matrix(c(40.3282,53.6383,2.5746,6.0161,4.7006,8.0338,
53.6383,7134.1381,34.2436,80.0164,62.5205,106.8531,
2.5746,34.2436,16.4368,3.8408,3.0010,5.1289,
6.0161,80.0164,3.8408,89.7463,7.0123,11.9846,
4.7006,62.5205,3.0010,7.0123,54.7902,9.3641,
8.0338,106.8531,5.1289,11.9846,9.3641,160.0415),6,6)

```

```

mu <-c(48.7200,578.4340,25.6100,31.4256,9.1600,40.4371)
n <-25
library(MASS)
x <- mvrnorm(n, mu, sigma)
mvw <- mvShapiro.Test(x)
s = var(x)
g1 = x[,1:3]
s1 = var(g1)
g2 = x[,4:6]
s2 = var(g2)
qui = -((dim(x)[1]-1)-((dim(x)[2]+1)/2))*(log(det(s)/(det(s1)*det(s2))))
pvalor = pchisq(qui, df=(dim(g1)[2]*dim(g2)[2]), ncp=0, lower.tail = FALSE, log.p = FALSE)
saida$nobs[j] <- n
saida$pvariaveis[j] <- dim(x)[2]
saida$pvalorShapiro[j] <- as.numeric(mvw[2])
saida$pvariaveisgrupo1[j] <- dim(g1)[2]
saida$pvariaveisgrupo2[j] <- dim(g2)[2]
saida$qui[j] <- qui
saida$pvalorqui[j] <- pvalor
}
write.table(saida,"C:\\R\\Betania2011\\Poder\\3x3-0.1peq25.csv", sep = ";", dec = ".", row.names = FALSE)
# Calculo do poder do teste qui quadrado para pequenas amostras com nível nominal de 5%
y = as.numeric (saida[,7])
Poder = length(y[y <= 0.05])/simula*100
Poder

```

Exemplo de rotina de simulação no Programa R para gerar as amostras normais multivariadas contaminadas e para cálculo das taxas de erro tipo I

```

rm(list=ls())
simula <- 2000
saida <- as.data.frame(matrix(0,simula,7))
names(saida) <-
c("nobs","pvariaveis","pvalorShapiro","pvariaveisgrupo1","pvariaveisgrupo2","qui","pvalorqui")
for(j in 1:simula)
{
mvrnormCT <- function(n, mu1, Sigma1, mu2, Sigma2,delta)
{
u <- runif(n)
if (u[1] <= delta) X <- mvrnorm(1, mu1, Sigma1) else
X <- mvrnorm(1, mu2, Sigma2)
for (i in 2:n)
{
if (u[i] <= delta) y <- mvrnorm(1, mu1, Sigma1) else
y <- mvrnorm(1, mu2, Sigma2)
X <- rbind(X, y)
}
return(X)
}
}
n<-25
Sigma1<-matrix (c(40.3282,0.0000,0.0000,0.0000,0.0000,0.0000,
0.0000,7134.1381,0.0000,0.0000,0.0000,0.0000,
0.0000,0.0000,16.4368,0.0000,0.0000,0.0000,
0.0000,0.0000,0.0000,89.7463,0.0000,0.0000,
0.0000,0.0000,0.0000,0.0000,54.7902,0.0000,
0.0000,0.0000,0.0000,0.0000,0.0000,160.0415),6,6)
mu1<-c(0,0.1,0.01,0,1,10)
Sigma2= Sigma1

```

```

mu2<- c(48.7200,578.4340,25.6100,31.4256,9.1600,40.4371)
delta<-0.2
library(MASS)
x<- mvrnormCT (n, mu1, Sigma1, mu2, Sigma2, delta)
mvw <- mvShapiro.Test(x)
s = var(x)
g1 = x[,1:3]
s1 = var(g1)
g2 = x[,4:6]
s2 = var(g2)
qui = -((dim(x)[1]-1)-((dim(x)[2]+1)/2))*(log(det(s)/(det(s1)*det(s2))))
pvalor = pchisq(qui, df=(dim(g1)[2]*dim(g2)[2]), ncp=0, lower.tail = FALSE, log.p = FALSE)
saida$nobs[j] <- n
saida$pvariaveis[j] <- dim(x)[2]
saida$pvalorShapiro[j] <- as.numeric(mvw[2])
saida$pvariaveisgrupo1[j] <- dim(g1)[2]
saida$pvariaveisgrupo2[j] <- dim(g2)[2]
saida$qui[j] <- qui
saida$pvalorqui[j] <- pvalor
}
write.table(saida, "C:\R\Betania2011\ResNCT-Erro\3x3peq25.csv", sep = ";", dec = ".", row.names =
FALSE)
# Calculo do Erro Tipo I do teste qui quadrado para pequenas amostras com nível nominal de 5%
y = as.numeric (saida[,7])
ErroTipoI = length(y[y <= 0.05])/simula*100
ErroTipoI

```

Exemplo de rotina de simulação no Programa R para gerar as amostras normais multivariadas contaminadas e para cálculo de poder ($r=0,1$)

```

rm(list=ls())
simula <- 2000
saida <- as.data.frame(matrix(0,simula,7))
names(saida) <-
c("nobs","pvariaveis","pvalorShapiro","pvariaveisgrupo1","pvariaveisgrupo2","qui","pvalorqui")
for(j in 1:simula)
{
mvrnormCT <- function(n, mu1, Sigma1, mu2, Sigma2,delta)
{
u <- runif(n)
if (u[1] <= delta) X <- mvrnorm(1, mu1, Sigma1) else
X <- mvrnorm(1, mu2, Sigma2)
for (i in 2:n)
{
if (u[i] <= delta) y <- mvrnorm(1, mu1, Sigma1) else
y <- mvrnorm(1, mu2, Sigma2)
X <- rbind(X, y)
}
return(X)
}
n<-25
Sigma1<-matrix (c(40.3282,53.6383,2.5746,6.0161,4.7006,8.0338,
53.6383,7134.1381,34.2436,80.0164,62.5205,106.8531,
2.5746,34.2436,16.4368,3.8408,3.0010,5.1289,
6.0161,80.0164,3.8408,89.7463,7.0123,11.9846,
4.7006,62.5205,3.0010,7.0123,54.7902,9.3641,
8.0338,106.8531,5.1289,11.9846,9.3641,160.0415),6,6)
mu1<-c(0,0.1,0.01,0.1,10)
Sigma2= Sigma1

```

```

mu2<- c(48.7200,578.4340,25.6100,31.4256,9.1600,40.4371)
delta<-0.2
library(MASS)
x<- mvrnormCT (n, mu1, Sigma1, mu2, Sigma2, delta)
library(mvShapiroTest)
mvw <- mvShapiro.Test(x)
s = var(x)
g1 = x[,1:3]
s1 = var(g1)
g2 = x[,4:6]
s2 = var(g2)
qui = -((dim(x)[1]-1)-((dim(x)[2]+1)/2))*(log(det(s)/(det(s1)*det(s2))))
pvalor = pchisq(qui, df=(dim(g1)[2]*dim(g2)[2]), ncp=0, lower.tail = FALSE, log.p = FALSE)
saida$noobs[j] <- n
saida$pvariaveis[j] <- dim(x)[2]
saida$pvalorShapiro[j] <- as.numeric(mvw[2])
saida$pvariaveisgrupo1[j] <- dim(g1)[2]
saida$pvariaveisgrupo2[j] <- dim(g2)[2]
saida$qui[j] <- qui
saida$pvalorqui[j] <- pvalor
}
write.table(saida,"C:\\R\\Betania2011\\ResNCT-Poder\\3x3-0.1peq25.csv ", sep = ";", dec = ".", row.names =
FALSE)

# Calculo do poder do teste qui quadrado para pequenas amostras com nível nominal de 5%
y = as.numeric (saida[,7])
Poder = length(y[y <= 0.05])/simula*100
Poder

```