



HENRIQUE JOSÉ DE PAULA ALVES

**PROPOSIÇÃO DE TESTE DE MÉDIAS ROBUSTOS E NÃO
ROBUSTOS SOB DISTRIBUIÇÕES NORMAIS
CONTAMINADAS E NÃO NORMAIS**

LAVRAS – MG

2020

HENRIQUE JOSÉ DE PAULA ALVES

**PROPOSIÇÃO DE TESTE DE MÉDIAS ROBUSTOS E NÃO ROBUSTOS SOB
DISTRIBUIÇÕES NORMAIS CONTAMINADAS E NÃO NORMAIS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística Multivariada, para a obtenção do título de Doutor.

Prof. DSc. Daniel Furtado Ferreira
Orientador

Prof. DSc. Renato Ribeiro de Lima
Coorientador

LAVRAS – MG

2020

**Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca
Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

Paula Alves, Henrique José de.

Proposição de teste de médias robustos e não robustos sob distribuições normais contaminadas e não normais / Henrique José de Paula Alves. – Lavras : UFLA, 2020.

112 p. : il.

Orientador: Prof. DSc. Daniel Furtado Ferreira.

Coorientador: Prof. DSc. Renato Ribeiro de Lima .

Tese(doutorado)–Universidade Federal de Lavras, 2020.

Bibliografia.

1. comedian. 2. T2. 3. alta dimensão. I. Ferreira, Daniel Furtado. II. Lima, Renato Ribeiro de. III. Título.

HENRIQUE JOSÉ DE PAULA ALVES

**PROPOSIÇÃO DE TESTE DE MÉDIAS ROBUSTOS E NÃO ROBUSTOS SOB
DISTRIBUIÇÕES NORMAIS CONTAMINADAS E NÃO NORMAIS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística Multivariada, para a obtenção do título de Doutor.

APROVADA em 18 de fevereiro de 2020.

Prof. DSc. Ben Dêivide de Oliveira Batista	UFSJ
Prof. DSc. Peter de Matos Campos	UFSJ
Prof. DSc. Augusto Ramalho de Moraes	UFLA
Prof. DSc. Devanil Jaques de Souza	UFLA

Prof. DSc. Daniel Furtado Ferreira
Orientador

Prof. DSc. Renato Ribeiro de Lima
Coorientador

**LAVRAS – MG
2020**

Agradeço a Deus pois sem ele eu não teria forças para essa longa jornada, agradeço a meus professores, meu orientador, meus pais, meus irmãos e aos meus colegas de curso que me ajudaram na conclusão da tese.

AGRADECIMENTOS

Vencer, antes de tudo, é uma conquista conjunta de professores, dos orientadores, de amigos, de familiares e do estudante que conclui um Doutorado. Assim, agradeço a muitas pessoas que ajudaram a fazer a diferença.

Ao meu orientador, Daniel Furtado Ferreira, pela ajuda, empenho e paciência nesses anos. Foram muitos os e-mails, idas e vindas até a sua sala. Ao colega Rafael Lemos Bastos, por ter me indicado ao professor Daniel, meu orientador.

Aos colegas de pós-graduação com os quais tivemos bons momentos durante a fase de estudos. Quantas horas dedicadas para adquirir um pouco mais de conhecimento, quantas conversas, risadas, viagens e festas que ajudaram a descontrair. Também pela ajuda e colaborações sempre em boa hora.

Aos amigos que, de alguma forma, colaboraram para que se pudesse desenvolver e concluir essa tese. Mesmo nas simples atitudes, como em palavras de incentivo e apoio, foram fundamentais para que eu pudesse chegar aqui.

Aos meus pais, “Seu Juca” (In memoriam) e “Dona Nem”, que lutaram por toda uma vida para poder dar a mim e a meus irmãos a oportunidade de alcançarmos o sucesso. Hoje, posso dizer que todo esse empenho foi de grande importância para o meu sucesso. A vocês, meu amor e gratidão eternos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Agradeço ao Laboratório de Computação Científica (LCC) do Departamento de Estatística da UFLA. Sem esse laboratório, os resultados científicos apresentados nesta tese não seriam possíveis.

Ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária pela oportunidade no curso de Doutorado, o qual possibilitou, através de seus professores e disciplinas, desenvolver esse trabalho através da sua estrutura.

A Deus, pela inspiração, pela força, pelos momentos concedidos para fazer com que essa vitória pudesse ser valorizada, eu agradeço eternamente.

“Há homens que lutam um dia e são bons. Há outros que lutam um ano e são melhores. Há os que lutam muitos anos e são muito bons. Porém, há os que lutam toda a vida. Esses são os imprescindíveis.”
Bertolt Brecht

RESUMO

Em uma análise estatística, o processo inferencial para a média de uma população inclui o chamado teste de hipóteses. Considerando o caso univariado e se tratando de uma população normalmente distribuída, quando não se conhece a variância populacional utiliza-se a distribuição t-student para tomar a decisão de rejeitar ou não a hipótese nula de que a média populacional é igual a uma dada constante. Para o caso multivariado, nas mesmas condições, utiliza-se o teste T^2 proposto por Harold Hotelling para testar as hipóteses sobre um vetor de médias populacional. Uma outra forma de realizar esses testes de hipóteses é por meio do teste da razão de verossimilhanças, denominado de LRT (likelihood ratio test). Na presença de outliers, ambos os testes T^2 e LRT não devem ser utilizados, pois seus estimadores que são o vetor de médias e a matriz de covariâncias amostrais são influenciados por valores discrepantes. Na literatura, existem muitos estudos sobre a construção de estimadores para o vetor de médias e para a matriz de covariâncias que são robustos à presença de outliers, possuindo alto ponto de ruptura e mantendo a propriedade de afim-equivariância. Porém, muitos destes estimadores estão condicionados à existência dos dois primeiros momentos. Um caso especial destes estimadores é o estimador robusto comedian que não está condicionado a existência dos dois primeiros momentos e, em síntese, considera medianas correlacionadas. Sendo assim, nesta tese, busca-se construir testes robustos alternativos ao T^2 de Hotelling tradicional e ao teste da razão de verossimilhanças, utilizando os estimadores comedian do vetor de médias e matriz de covariâncias. Estes testes foram gerados por meio de adaptações feitas tanto no teste T^2 quanto no teste LRT , sendo que alguns desses testes são assintóticos e outros são construídos através dos estimadores bootstrap paramétrico comedian para o vetor de médias e matriz de covariâncias. Os testes adaptados ao teste LRT possuem a vantagem de poderem ser utilizados para dados de alta dimensão. O desempenho destes testes foi avaliado e comparado ao teste T^2 tradicional considerando distribuições normais contaminadas e distribuições não normais, utilizando simulações Monte Carlo. Foram considerados o poder do teste e a taxa de erro tipo I como medidas avaliativas. Como resultado, verificou-se que o desempenho da versão bootstrap paramétrico com traço do teste da razão de verossimilhanças $TLRPBT$ (Trace Likelihood Ratio Bootstrap Parametric test) teve melhor desempenho entre todos os testes considerados. Um outro fato é que a versão assintótica do teste da razão de verossimilhanças com traço $ATLRT$ (Asymptotic Trace Likelihood Ratio Test) não deve ser descartado entre todos os testes propostos, pois o mesmo resistiu controlando o erro tipo I e poder em situações extremas de assimetria.

Palavras-chave: outliers. comedian. *bootstrap*. Monte Carlo. T^2 . LRT . alta dimensão.

ABSTRACT

In a statistical analysis, the inferential process for the mean of a population includes the so-called hypothesis test. Considering the univariate case and dealing with a normally distributed population, when the population variance is not known, the t-student distribution is used to make the decision to reject or not the null hypothesis that the population average is equal to a given constant. For the multivariate case, under the same conditions, the T^2 test proposed by Harold Hotelling is used to test the hypotheses about a population mean vector. Another way to perform these hypothesis tests is through the likelihood ratio test, called the *LRT* (likelihood ratio test). In the presence of outliers, both the T^2 and *LRT* tests should not be used, as their estimators, which are the vector of means and the sample covariance matrix, are influenced by outliers. In the literature, there are many studies on the construction of estimators for the vector of averages and for the covariance matrix that are robust to the presence of outliers, possessing a high breaking point and maintaining the affine-equivalence property. However, many of these estimators are conditioned to the existence of the first two moments. A special case of these estimators is the comedian robust estimator, which is not conditioned to the existence of the first two moments and, in summary, considers correlated medians. Therefore, in this thesis, we seek to build robust tests alternative to the traditional Hotelling T^2 and the likelihood ratio test, using the comedian estimators of the mean vector and covariance matrix. These tests were generated through adaptations made both in the T^2 test and in the *LRT* test, with some of these tests being asymptotic and others being constructed using the comedic parametric bootstrap estimators for the mean vector and covariance matrix. Tests adapted to the *LRT* test have the advantage that they can be used for high-dimensional data. The performance of these tests was evaluated and compared to the traditional T^2 test considering normal contaminated and non-normal distributions, using Monte Carlo simulations. The power of the test and the type I error rate were considered as evaluative measures. As a result, it was found that the performance of the parametric bootstrap version with trace of the *TLRPBT* likelihood ratio test (Trace Likelihood Ratio Bootstrap Parametric test) had better performance among all the considered tests. Another fact is that the asymptotic version of the *ATLRT* trace likelihood ratio test (Asymptotic Trace Likelihood Ratio Test) should not be discarded among all the proposed tests, as the same resisted controlled type I error and power in extreme situations asymmetry.

Keywords: outliers. comedian. *bootstrap*. Monte Carlo. T^2 . *LRT*. high dimension.

LISTA DE FIGURAS

Figura 2.1 – O algoritmo bootstrap.	29
Figura 4.1 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	57
Figura 4.2 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	58
Figura 4.3 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	59
Figura 4.4 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	60
Figura 4.5 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	61
Figura 4.6 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	62
Figura 4.7 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	63

Figura 4.8 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	64
Figura 4.9 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	65
Figura 4.10 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	66
Figura 4.11 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	87
Figura 4.12 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	88
Figura 4.13 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	89
Figura 4.14 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	90
Figura 4.15 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	91

Figura 4.16 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	92
Figura 4.17 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	93
Figura 4.18 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	94
Figura 4.19 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	95
Figura 4.20 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	96
Figura 4.21 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	98
Figura 4.22 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	99
Figura 4.23 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	100

Figura 4.24 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	101
Figura 4.25 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	102
Figura 4.26 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with t distribution with 1 d.f. and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	106
Figura 4.27 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with t distribution with 1 d.f. and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	107
Figura 4.28 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with multivariate exponential distribution, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.	108

LISTA DE TABELAS

Tabela 2.1 – Tipos de erros cometidos em um teste de hipóteses.	25
Tabela 2.2 – Variância assintótica (k) de T_j^1 para o modelo normal padrão.	37
Tabela 4.1 – Type I error rates of the three tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate Normal (N), Student t with 5 degrees of freedom (t_5), Student t with 30 degrees of freedom (t_{30}) and contaminated normal (CN) distributions.	56
Tabela 4.2 – Type I error rates of the three tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate Normal (N), Student t with 5 degrees of freedom (t_5), Student t with 30 degrees of freedom (t_{30}) and contaminated normal (CN) distributions.	56
Tabela 4.1 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate normal (N), multivariate Student t with 5 degrees of freedom (t_5) degrees of freedom and contaminated normal (CN) distributions.	85
Tabela 4.2 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate Normal (N), multivariate Student t with 5 degrees of freedom (t_5) and contaminated normal (CN) distributions.	86
Tabela 4.3 – Type I error rates of the three tests with $\alpha = 5\%$ and $p = 200$, considering the multivariate Normal (N), multivariate Student t with 5 degrees of freedom (t_5) and multivariate contaminated normal (CN) distributions.	97
Tabela 4.4 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate Student t -distribution with 1 degrees of freedom (t_1).	104
Tabela 4.5 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate Student t -distribution with 1 degrees of freedom (t_1).	105
Tabela 4.6 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate exponential distribution.	108
Tabela 4.7 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate exponential distribution.	108
Tabela 4.8 – Type I error rates ($\delta = 0$) and power ($\delta = 0.5$ and 10) of the three tests with $\alpha = 5\%$, $p = 200$ and $n = 200$, considering the multivariate exponential distribution, where the values in parenthesis express the margin of error of Monte Carlo simulations.	109

SUMÁRIO

	PRIMEIRA PARTE	14
1	INTRODUÇÃO	15
2	REFERENCIAL TEÓRICO	17
2.1	O conceito e classificação de formas quadráticas	18
2.2	Distâncias quadráticas	19
2.3	Os testes tradicionais T^2 de Hotelling e LRT (likelihood ratio test)	20
2.3.1	O teste T^2 de Hotelling	20
2.3.2	O teste LRT (likelihood ratio test)	21
2.4	O método de estimação robusta median absolute deviation (MAD)	22
2.4.1	A propriedade de afim-equivariância de locação e de escala	22
2.4.2	Ponto de ruptura de um estimador	23
2.5	O estimador robusto comedian e suas propriedades	23
2.6	Erros cometidos ao se testar uma hipótese	25
2.6.1	Erro tipo I e erro tipo II	25
2.7	A distribuição normal multivariada	26
2.8	A distribuição normal multivariada contaminada	27
2.8.1	Geração de variáveis aleatórias com distribuição normal multivariada contaminada	27
2.9	A distribuição t-student multivariada	28
2.10	O algoritmo <i>bootstrap</i>	28
2.10.1	O <i>bootstrap</i> paramétrico	30
2.11	O método Monte Carlo	30
2.12	O teste Binomial exato	31
2.13	Adaptações robustas e não robustas aos testes T^2 e LRT	32
2.14	Testes adaptados para dados de alta dimensão	38
3	CONSIDERAÇÕES FINAIS	40
	REFERÊNCIAS	41
	SEGUNDA PARTE - ARTIGOS	44
	ARTIGO 1 - Proposition of new alternative tests adapted to the traditional T^2 test.	44
1	INTRODUCTION	47

2	THE ROBUST ESTIMATOR COMEDIAN	49
3	METHODS	51
3.1	The parametric bootstrap T^2 test	52
3.2	The robust parametric bootstrap T^2 test	53
3.3	The exact binomial test	54
4	RESULTS AND DISCUSSION	55
4.1	Type I error rates	55
4.2	Power	57
5	CONCLUSION	68
	References	70
	ARTIGO 2 - Proposition of new robust tests to the multivariate normal mean vector for high-dimensional data.	71
1	INTRODUCTION	73
2	METHODS	76
2.1	The parametric bootstrap T^2 test	78
2.2	The robust parametric bootstrap T^2 test	79
2.3	The asymptotic LR trace test	80
2.4	The TLR parametric bootstrap test	81
2.5	The robust TLR parametric bootstrap test	81
2.6	The exact binomial test	82
3	RESULTS AND DISCUSSION	84
3.1	Type I error rates	84
3.2	Power	86
3.3	Special case of $p = 200$: type I error rates and power	97
3.4	Special cases	103
3.4.1	Multivariate t_1 distribution	103
3.4.2	Multivariate exponential distribution	105
4	CONCLUSION	110
	References	112

PRIMEIRA PARTE

1 INTRODUÇÃO

Em qualquer análise estatística de dados coletados em uma amostra, o objetivo final é realizar inferências de modo a extrapolar os resultados da amostra para a população. Isso equivale a realizar afirmativas a respeito dos parâmetros com base nos resultados observados nesta mesma amostra. Estas inferências podem ser realizadas tanto no caso univariado quanto no caso multivariado.

Uma das formas de realizar inferências se dá por meio de testes de hipóteses. Em particular, nesta tese tem-se o interesse de realizar testes de hipóteses para o vetor de médias populacionais considerando populações normais e não normais para o caso multivariado. Sob normalidade, o teste mais utilizado é o teste T^2 de Hotelling, que é utilizado quando número de observações n da amostra é maior que a dimensão dos dados p . Outro teste estatístico é o teste da razão de verossimilhanças, ao qual é designado a sigla LRT (Likelihood Ratio Test). Entretanto, ambos os testes T^2 e LRT apresentam problemas de utilização quando $p \geq n$ devido à singularidade da matriz de covariâncias.

Ainda, tem-se que esses testes possuem limitações quanto ao seu uso sob distribuições não normais. Mesmo sob normalidade, esses testes podem ser influenciados pela presença de observações discrepantes ou outliers, devido à não robustez dos estimadores da média e da matriz de covariâncias populacionais, uma vez que esses testes dependem desses estimadores. Então, torna-se necessário a obtenção de novas estatísticas de teste robustas para se aplicar testes de hipóteses concernentes ao vetor de médias populacional.

Na literatura existem métodos de estimação robustos a outliers, que geram estimadores robustos com alto ponto de ruptura e afim-equivariantes de locação e de escala e que têm potencial de serem utilizados tanto no teste T^2 de Hotelling como no teste LRT . As propriedades ponto de ruptura e afim-equivariância são muito importantes para gerar estimadores robustos a outliers, pois indicam, respectivamente, o quão resistente um estimador é na presença de outliers, além de preservarem a propriedade de robustez em qualquer combinação linear desse estimador.

Nesse sentido, um dos métodos de estimação robusta baseado em medianas correlacionadas, denominado comedian, foi sugerido por M. Falk, em 1997, gerando estimadores robustos afim-equivariantes de locação e de escala e com alto ponto de ruptura que podem ser utilizados em ambos os testes T^2 de Hotelling e LRT .

Visando solucionar os problemas de não normalidade e outliers que afetam o uso dos testes T^2 e LRT , o objetivo deste trabalho é propor novos testes alternativos para esses testes tradicionais, visando a utilização tanto sob normalidade quanto sob não-normalidade com ou sem contaminação, considerando ambas as distribuições assintótica e bootstrap paramétrico. Nas versões robustas desses novos testes substituíram-se os estimadores da média e da covariância na estatística do teste T^2 tradicional e LRT pelos respectivos estimadores robustos *comedian*.

Como objetivos específicos, o desempenho desses novos testes foi avaliado por meio da taxa de erro tipo I e o poder do teste. O desempenho dessas novas propostas foi comparado com o desempenho do teste T^2 de Hotelling tradicional em simulações Monte Carlo. Foram consideradas as distribuições normal, normal contaminada, t-student multivariada e diferentes tamanhos amostrais.

Esta tese está estruturada da seguinte maneira. Na primeira parte apresenta-se todo o referencial teórico e revisão bibliográfica necessários para o entendimento dessa tese. Apresenta-se também as considerações finais, mencionando possíveis trabalhos futuros. Na segunda parte a metodologia, os resultados e a discussão dos mesmos é apresentada no formato de dois artigos.

2 REFERENCIAL TEÓRICO

Quando um pesquisador deseja verificar se algum parâmetro da distribuição de uma variável de interesse é condizente com o valor por ele estipulado ou, quando se deseja saber se os parâmetros de duas populações independentes são semelhantes, os testes de hipóteses são uma ferramenta adequada muito utilizada. Sendo assim, estatísticas de teste são formuladas e seus valores são obtidos a partir dos dados amostrais para validar as hipóteses de interesse.

Um interesse muito comum é testar a hipótese de que a média populacional é igual a uma constante. Para o caso em que a variável de interesse segue uma distribuição normal univariada com média μ e variância σ^2 , ambos desconhecidos, uma estatística de teste muito utilizada é a estatística $t = (\bar{X} - \mu)/(S/\sqrt{n})$ que segue uma distribuição t-student com $n - 1$ graus de liberdade, sendo n o tamanho amostral, \bar{X} a média amostral e S o desvio-padrão amostral.

Porém, existem situações em que muitas variáveis são medidas na mesma população, como no caso multivariado e a hipótese de interesse envolve um vetor de parâmetros da distribuição de probabilidade conjunta dessas variáveis. Uma estratégia, então, é realizar um teste de hipóteses para cada parâmetro de interesse de cada uma das variáveis separadamente. No entanto, como estas variáveis são medidas quase sempre em uma mesma unidade experimental, é possível que exista correlação entre elas sendo então razoável a realização de um teste de hipóteses multivariado capaz de avaliar simultaneamente os parâmetros dessa distribuição conjunta.

Por levar em consideração a correlação entre as variáveis, os testes multivariados tendem a ser mais poderosos que os testes univariados quando estes são utilizados em conjunto para avaliar as hipóteses multivariadas. Suponha que o interesse seja testar

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \quad (2.1)$$

cujas médias são provenientes de uma população normal multivariada com variâncias desconhecidas, em que $\boldsymbol{\mu}$ é o vetor de médias de uma população e $\boldsymbol{\mu}_0$ é um vetor de constantes que supõe-se ser conhecido. Um dos testes de hipóteses multivariado mais utilizado é o teste T^2 proposto por Harold Hotelling, sendo este adequado para testar hipóteses sobre o vetor de médias populacional no caso de uma população normal multivariada. Tal teste é uma extensão do teste t-student univariado (FERREIRA, 2018).

Um outro teste de hipóteses bastante utilizado na literatura é o teste da razão de verossimilhanças *LRT* (likelihood ratio test), que considera em seu núcleo uma forma quadrática além da matriz de covariâncias amostral \mathbf{S} .

Sob distribuições não-normais e na presença de outliers, os testes T^2 de Hotelling e *LRT* não devem ser utilizados devido à quebra de pressuposição como normalidade dos dados e variâncias heterogêneas, devendo ser utilizado outra estatística de teste. Além disso, tais testes não podem ser obtidos quando o tamanho amostral n é menor do que ou igual a dimensão dos dados p , devido à singularidade da matriz de covariâncias amostral \mathbf{S} . Para resolver esses problemas pesquisadores da área têm desenvolvido adaptações alternativas para os testes T^2 de Hotelling e *LRT* cujas distribuições são proporcionais a uma distribuição F de Snedecor, porém não exatas.

Como citado anteriormente, o teste T^2 e o teste *LRT* englobam em seu núcleo o conceito de formas quadráticas e de distâncias. Em particular, ambos utilizam uma distância quadrática já bem difundida na literatura, conhecida como distância quadrática de Mahalanobis.

2.1 O conceito e classificação de formas quadráticas

Segundo Ferreira (2018), as formas quadráticas são funções que apresentam somente termos quadráticos ou duplos produtos de seus elementos. Uma forma quadrática, denotada por $Q(\mathbf{x})$, é dada pela seguinte expressão

$$Q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} = \langle \mathbf{x}, \mathbf{A} \mathbf{x} \rangle = \langle \mathbf{A} \mathbf{x}, \mathbf{x} \rangle \quad (2.2)$$

em que $\mathbf{A}_{p \times p}$ é uma matriz com elementos de entrada a_{ij} , com $i = 1, \dots, p$, $j = i, \dots, p$ e \mathbf{x} um vetor n -dimensional com $\mathbf{x} \neq \mathbf{0}$. A quantidade $Q(\mathbf{x})$ representa uma função do vetor \mathbf{x} sendo os elementos da matriz \mathbf{A} conhecidos. Considerando a matriz \mathbf{A} fixa, uma forma quadrática pode ser classificada conforme o valor numérico que a forma quadrática $Q(\mathbf{x})$, como em (2.2), assume (FERREIRA, 2018). Assim, $\forall \mathbf{x} \neq \mathbf{0}$:

- se a forma quadrática assume valores positivos, ou seja, se $Q(\mathbf{x}) > 0$, então a forma quadrática $Q(\mathbf{x})$ é positiva definida e a matriz \mathbf{A} é positiva definida;
- se a forma quadrática assume valores negativos, ou seja, se $Q(\mathbf{x}) < 0$, então a forma quadrática $Q(\mathbf{x})$ é negativa definida e a matriz \mathbf{A} é negativa definida;

- se a forma quadrática assume valores positivos incluindo o zero, ou seja, se $Q(\mathbf{x}) \geq 0$, então a forma quadrática $Q(\mathbf{x})$ é semipositiva definida e a matriz \mathbf{A} é semipositiva definida;
- se a forma quadrática assume valores negativos incluindo o zero, ou seja, se $Q(\mathbf{x}) \leq 0$, então a forma quadrática $Q(\mathbf{x})$ é seminegativa definida e a matriz \mathbf{A} é seminegativa definida.

2.2 Distâncias quadráticas

Uma forma quadrática muito utilizada em estatística multivariada é a distância quadrática, cuja expressão geral é dada por

$$d^2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\Psi}^2 = (\mathbf{x} - \mathbf{y})^{\top} \Psi (\mathbf{x} - \mathbf{y}), \quad (2.3)$$

em que \mathbf{x} , \mathbf{y} são vetores p -dimensionais e de mesma ordem, Ψ é uma matriz positiva definida, denominada de métrica e $\|\mathbf{a}\|$ se refere a norma de um vetor \mathbf{a} dada por $\|\mathbf{a}\| = \sqrt{\mathbf{a}^{\top} \mathbf{a}}$ (FERREIRA, 2018).

A distância $d^2(\mathbf{x}, \mathbf{y})$ tem as seguintes propriedades. Dado um terceiro vetor \mathbf{z} , tem-se:

- $d^2(\mathbf{x}, \mathbf{y}) \geq 0 \quad \forall \mathbf{x}, \mathbf{y}$;
- $d^2(\mathbf{x}, \mathbf{y}) = 0$ se, e somente se, $\mathbf{x} = \mathbf{y}$;
- $d^2(\mathbf{x}, \mathbf{y}) \leq d^2(\mathbf{x}, \mathbf{z}) + d^2(\mathbf{z}, \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z}$ não colineares.

De acordo com os valores assumidos pela métrica Ψ têm-se diferentes distâncias. Por exemplo, se $\Psi = \mathbf{I}$ tem-se a distância euclidiana quadrática que não leva em conta a existência de correlação entre as variáveis. Uma distância quadrática que merece destaque nessas notas é a distância quadrática de Mahalanobis que, conforme já mencionado, está presente nos testes tradicionais T^2 e LRT . Essa distância é obtida tomando $\Psi = \mathbf{S}^{-1}$ na expressão (2.3) em que \mathbf{S} é a matriz de covariâncias amostral e é representada por

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix}. \quad (2.4)$$

Assim, a distância quadrática de Mahalanobis é dada pela seguinte expressão

$$d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top \mathbf{S}^{-1}(\mathbf{x} - \mathbf{y}). \quad (2.5)$$

Segundo Ferreira (2018) nessa definição de distância as condições de escala são iguais e os efeitos da correlação são considerados, sendo assim, esta é muito rica em informações.

Pode-se, agora, apresentar as expressões relacionadas aos testes T^2 e LRT . Do mesmo modo, as distribuições de probabilidade aproximadas associadas a estas estatísticas de teste são apresentadas considerando a hipótese nula H_0 como definida na expressão (2.1).

2.3 Os testes tradicionais T^2 de Hotelling e LRT (likelihood ratio test)

O teste T^2 de Hotelling é baseado na distância quadrática de Mahalanobis, definida na expressão (2.5), entre o vetor de médias amostrais $\bar{\mathbf{X}}$ e o vetor de médias populacionais hipotético $\boldsymbol{\mu}_0$ assumindo-se a hipótese nula verdadeira, ponderada pela matriz de covariâncias amostrais \mathbf{S} . Assume-se também normalidade p -variada com parâmetros populacionais $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, ambos desconhecidos.

2.3.1 O teste T^2 de Hotelling

Segundo Ferreira (2018), a estatística T^2 é dada pela seguinte expressão

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \quad (2.6)$$

em que $\bar{\mathbf{X}} = \sum_{j=1}^n \mathbf{X}_j / n$ é o vetor de médias amostrais e $\mathbf{S} = \sum_{j=1}^n (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})^\top / (n - 1)$ é matriz de covariâncias amostrais.

A distribuição de $\bar{\mathbf{X}}$ é normal multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}/n$ e independente de $(n - 1)\mathbf{S}$, que possui distribuição Wishart denotada por $W_p(\nu = n - 1, \boldsymbol{\Sigma})$ (FERREIRA, 2018). Sendo assim

$$T^2 \sim \frac{\chi_p^2 \delta^2}{\chi_{\nu+1-p}^2}, \quad (2.7)$$

sendo que os termos $\chi_p^2 \delta^2$ e $\chi_{\nu+1-p}^2$ são independentes, sendo o termo $\delta^2 = n\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ o parâmetro de não-centralidade e χ^2 é o quantil de uma distribuição qui-quadrado com $\nu + 1 - p$

graus de liberdade. Trata-se de uma forma quadrática de variáveis normais e cujo núcleo é a inversa de uma distribuição Wishart dividida pelos respectivos graus de liberdade. Essa distribuição, sob H_0 , é proporcional a uma distribuição F não-central e a distribuição é dada por

$$T^2 \sim \frac{\nu p}{\nu + 1 - p} F(p, \nu + 1 - p, \delta^2), \quad (2.8)$$

em que $F(p, \nu + 1 - p, \delta^2)$ é o quantil de uma distribuição F com p e $\nu + 1 - p$ graus de liberdade e parâmetro de não centralidade δ^2 . Rejeita-se H_0 se $T^2 > \frac{\nu p}{\nu + 1 - p} F(p, \nu + 1 - p, \delta^2)$.

A função densidade de probabilidade da estatística T^2 é dada por

$$f_{T^2}(t^2; \delta^2) = \frac{\exp\{-\frac{1}{2}\delta^2\}}{\nu \Gamma(\frac{\nu+1-p}{2})} \sum_{j=0}^{\infty} \frac{\left(\frac{\delta^2}{2}\right)^j \left(\frac{t^2}{\nu}\right)^{p/2+j-1} \Gamma\left(\frac{\nu+1}{2} + j\right)}{j! \Gamma(p/2 + j) (1 + t^2/\nu)^{(\nu+1)/2+j}}, \quad t^2 \geq 0, \quad (2.9)$$

em que $\nu = n - 1$ nesse caso são os graus de liberdade e $\delta^2 = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$ é o parâmetro de não-centralidade. Vale ressaltar que o parâmetro ν não precisa ser inteiro, basta apenas ser real positivo (FERREIRA, 2018). O teste T^2 tem bom desempenho sob normalidade multivariada e homoscedasticidade (FERREIRA, 2018).

Considerando as hipóteses dadas em (2.1), um outro teste estatístico utilizado é o teste da razão de verossimilhanças (FERREIRA, 2018).

2.3.2 O teste LRT (likelihood ratio test)

Seja $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ com $\boldsymbol{\Sigma}$ desconhecida. A estatística de teste baseada na razão de verossimilhanças (Λ) é dada pela seguinte expressão

$$-2 \log(\Lambda) = n[\log |\mathbf{S} + \mathbf{H}| - \log |\mathbf{S}|], \quad (2.10)$$

em que $\mathbf{H} = (\bar{\mathbf{X}} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\top$ e \log é o logaritmo natural. Considere $\Omega \in \mathbb{R}^s$ o espaço paramétrico irrestrito e $\Omega_0 \subseteq \mathbb{R}^r$ o espaço paramétrico restrito, com $\Omega_0 \subset \Omega$. Sob certas condições de regularidade, a expressão (2.10) segue uma distribuição assintótica qui-quadrado com $s - r$ graus de liberdade (FERREIRA, 2018) sob a hipótese nula H_0 . Então, a região de rejeição de H_0 , R , como na expressão (2.1), é dada por

$$R = \{\mathbf{X} | \lambda = -2 \log(\Lambda(\mathbf{X})) > \chi_{\alpha, r-s}^2\}, \quad (2.11)$$

em que α é o nível de significância adotado e $\chi_{\alpha, r-s}^2$ é o $100\alpha\%$ quantil superior de uma distribuição qui-quadrado com $r - s$ graus de liberdade.

Seguindo os objetivos gerais apresentados nesta tese, novos testes de hipóteses são propostos, com as versões robustas e bootstrap paramétrico, sendo que estes foram avaliados pelo método Monte Carlo. Assim sendo, as próximas seções apresentam o estimador robusto para o vetor de médias e matriz de covariâncias comediano, o método bootstrap paramétrico e o método de avaliação Monte Carlo. Contudo, salienta-se que antes de se definir o estimador robusto comediano é preciso apresentar o método de estimação robusta para o caso univariado “Median Absolute Deviation” (*MAD*).

2.4 O método de estimação robusta median absolute deviation (*MAD*)

Falk (1997), em seu trabalho estudou o método *MAD*, que funciona como uma medida de escala para a mediana e é dado pela seguinte expressão

$$MAD(X) = med(|X - med(X)|) \quad (2.12)$$

que, segundo Falk (1997), o *MAD* funciona como uma única medida de estimativa ancilar de escala.

Criado por Falk (1997), o comediano é um método de estimação robusto a outliers, baseado em estimadores que envolvem medianas correlacionadas. Medianas correlacionadas são uma medida alternativa de dependência entre duas variáveis aleatórias. Este estimador apresenta as propriedades de afim-equivariância de locação e de escala e com alto ponto de ruptura.

Em se tratando dos estimadores robustos para o vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, assume-se que \mathbf{X} representa uma matriz dos dados que foram observados com \mathbf{X}_i^\top sendo descrito como a i -ésima linha de \mathbf{X} , $i = 1, 2, \dots, n$. Defina $\mathbf{A}_{p \times p}$ como uma matriz não-singular e $\mathbf{B} = \mathbf{1}_n \mathbf{b}^\top$, em que $\mathbf{1}_n$ é um vetor n -dimensional formado de uns e \mathbf{b} é um vetor de constantes p -dimensional.

2.4.1 A propriedade de afim-equivariância de locação e de escala

Segundo Hubert, Rousseeuw e Aelst (2008), define-se $\hat{\boldsymbol{\mu}}(\mathbf{X}) \in \mathbb{R}^p$ como um estimador de médias afim-equivariante de locação se, e somente se, para qualquer matriz $\mathbf{A}_{p \times p}$ não singular, a seguinte condição for satisfeita

$$\hat{\boldsymbol{\mu}}(\mathbf{X}_1\mathbf{A} + \mathbf{b}^\top, \dots, \mathbf{X}_n\mathbf{A} + \mathbf{b}^\top) = \hat{\boldsymbol{\mu}}(\mathbf{X}_1, \dots, \mathbf{X}_n)\mathbf{A} + \mathbf{b}^\top \quad \forall \mathbf{b} \in \mathbb{R}^p. \quad (2.13)$$

Diz-se que $\hat{\boldsymbol{\Sigma}}(\mathbf{X})$ é um estimador equivariante de escala se, e somente se, vale a seguinte condição

$$\hat{\boldsymbol{\Sigma}}(\mathbf{X}_1\mathbf{A} + \mathbf{b}^\top, \dots, \mathbf{X}_n\mathbf{A} + \mathbf{b}^\top) = \mathbf{A}^\top \hat{\boldsymbol{\Sigma}}(\mathbf{X}_1, \dots, \mathbf{X}_n)\mathbf{A}. \quad (2.14)$$

Em outras palavras um estimador é afim-equivariante de locação e de escala se o estimador de uma combinação linear for igual a combinação linear do estimador original.

Introduzido formalmente e de forma concisa por Hampel e Ronchetti (1986), o ponto de ruptura é uma medida muito utilizada em assuntos que envolvam estatística robusta. É um conceito assintótico mas que foi adaptado para amostras finitas por Donoho e Huber (1983).

2.4.2 Ponto de ruptura de um estimador

O ponto de ruptura de um estimador para amostras infinitas $\delta_n^*(T, \mathbf{Z})$, em que T é um estimador e \mathbf{Z} é uma vetor aleatório de interesse, é definido informalmente como a menor fração de contaminação da amostra por outliers que possa comprometer a qualidade ou veracidade das estimativas a serem obtidas. Em outras palavras, trata-se da maior proporção de contaminação que um estimador possa suportar e, ainda assim, fornecer informação confiável sobre o parâmetro considerado.

Os testes propostos foram avaliados de acordo com o controle da taxa de erro tipo I e o poder do teste. Desta forma, essas medidas são apresentadas. Porém, ressalta-se que essas medidas estão diretamente relacionadas com os dois tipos de erros que podem ser cometidos ao se testar uma hipótese.

2.5 O estimador robusto comedian e suas propriedades

Considere X e Y duas variáveis aleatórias, então o comedian entre X e Y é definido como

$$COM(X, Y) = med((X - med(X))(Y - med(Y))), \quad (2.15)$$

em que o termo *med* denota a mediana.

A expressão 2.15 generaliza o método MAD (median absolute deviation), que é obtido fazendo-se $X = Y$ na expressão (2.15). O estimador robusto $COM(X, Y)$ tem o maior ponto de

ruptura possível (FALK, 1997), isto é, a maior proporção de contaminação que um estimador pode suportar e ainda manter sua confiabilidade. Uma de suas vantagens é que o comedian sempre existe enquanto que $COV(X, Y)$ requer a existência dos dois primeiros momentos de X e Y . Este é simétrico, invariante para a média e covariância (FALK, 1997). Isso significa que $COM(X, aY + B) = aCOM(X, Y)$. Ainda, este estimador tem forte consistência e normalidade assintótica.

Sajesh e Srinivasan (2012) propuseram o uso da versão multivariada do estimador comedian. Defina $\mathbf{X}_{p \times p}$ como uma matriz de dados com linhas \mathbf{X}_i^\top , $i = 1, 2, \dots, n$ e colunas \mathbf{X}_j , $j = 1, 2, \dots, p$. O conceito de matriz mediana $COM(\mathbf{X})$ é definido como

$$COM(\mathbf{X})_{p \times p} = \begin{bmatrix} MAD^2(\mathbf{X}_1) & COM(\mathbf{X}_1, \mathbf{X}_2) & \dots & COM(\mathbf{X}_1, \mathbf{X}_p) \\ COM(\mathbf{X}_1, \mathbf{X}_2) & MAD^2(\mathbf{X}_2) & \dots & COM(\mathbf{X}_2, \mathbf{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ COM(\mathbf{X}_p, \mathbf{X}_1) & \dots & \dots & MAD^2(\mathbf{X}_p) \end{bmatrix}. \quad (2.16)$$

A matriz $COM(\mathbf{X})$ pode ser usada para substituir \mathbf{S} na expressão (2.6). Porém como alternativa robusta para a matriz de covariâncias, a matriz $COM(\mathbf{X})$ é, em geral, não positiva (semi-) definida (FALK, 1997). Este problema é frequente em estimação robusta da matriz de covariâncias.

Para lidar com este problema Maronna e Zamar (2002) adotaram os seguintes passos para obter estimativas robustas para o vetor de médias e matriz de covariâncias.

1. defina $\boldsymbol{\delta}(\mathbf{X}) = \mathbf{D}COM(\mathbf{X})\mathbf{D}^\top$ como a matriz de correlação mediana multivariada em que \mathbf{D} é uma matriz diagonal com elementos $1/MAD(\mathbf{X}_i)$, $i = 1, 2, \dots, p$;
2. calcule os autovalores λ_i e os autovetores \mathbf{e}_j de $\boldsymbol{\delta}(\mathbf{X})$, $j = 1, 2, \dots, p$ e chame de \mathbf{E} a matriz cujas colunas são os \mathbf{e}'_j s. Então $\boldsymbol{\delta}(\mathbf{X}) = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top$, em que $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$;
3. seja $\mathbf{Q} = \mathbf{D}\mathbf{E}$, em que \mathbf{D} é definida no passo 1 e $\mathbf{Z}_i = \mathbf{Q}^{-1}\mathbf{x}_i$, $i = 1, 2, \dots, p$;
4. as estimativas robustas para o vetor de médias $\mathbf{m}(\mathbf{X})$ e matriz de covariâncias $\mathbf{S}(\mathbf{X})$ são definidas como

$$\mathbf{S}(\mathbf{X}) = \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}^\top \quad \text{e} \quad \mathbf{m}(\mathbf{X}) = \mathbf{Q}\mathbf{l} \quad (2.17)$$

em que $\boldsymbol{\Gamma} = \text{diag}(MAD(\mathbf{Z}_1)^2, \dots, MAD(\mathbf{Z}_p)^2)$ e $\mathbf{l} = (\text{med}(\mathbf{Z}_1), \dots, \text{med}(\mathbf{Z}_p))^\top$.

As estimativas podem ser fornecidas utilizando de um processo iterativo substituindo δ por S e repetindo os passos 2, 3 e 4.

Estas estimativas são positiva definida e aproximadamente afim-equivariantes. Além disso, as estimativas obtidas pelo método comedian têm um alto ponto de ruptura. A eficiência do método aumenta com o aumento da dimensão dos dados (SAJESH; SRINIVASAN, 2012).

A propriedade de afim-equivariância é muito importante no âmbito de estudos aplicados e envolve transformação nos dados, permitindo que a escala da variável original possa ser alterada sem perda de coerência nos resultados (OLIVE, 2008).

2.6 Erros cometidos ao se testar uma hipótese

Em geral, quaisquer testes estatísticos utilizados para realizar inferências que envolvam testes de hipóteses são considerados apropriados se esses controlam a taxa de erro tipo I e tem alto poder. Em qualquer teste de hipóteses, existem dois tipos de erros que podem ser cometidos. A Tabela 2.1 ilustra esses dois erros.

Tabela 2.1 – Tipos de erros cometidos em um teste de hipóteses.

	Não rejeitar H_0	Rejeitar H_0
H_0 verdadeira	Decisão correta	Erro tipo I
H_0 falsa	Erro tipo II	Decisão correta

2.6.1 Erro tipo I e erro tipo II

O erro tipo I é cometido quando se rejeita a hipótese nula H_0 sendo esta verdadeira. Em geral, o erro tipo I ocorre com probabilidade igual ao nível nominal de significância α . Já o erro tipo II ocorre quando se aceita H_0 sendo que a mesma deveria ser rejeitada, sendo que esse erro ocorre com probabilidade β .

Em estudos de simulação, o controle do erro tipo I e o poder do teste é obtido por meio de taxas e proporções. A taxa de erro tipo I é definida como a proporção de vezes que um teste está sujeito a cometer o erro tipo I. O que se espera é que essa proporção seja baixa e controlada, podendo ser no máximo igual ao nível nominal de significância α .

O poder de um teste é definido como a proporção de vezes em que o teste não comete o erro tipo II. O que se espera é que essa proporção seja alta. Esse pode ser afetado por três situações:

- **Tamanho da amostra:** mantendo todos os outros parâmetros iguais, quanto maior o tamanho da amostra, maior o poder do teste;
- **Nível de Significância:** quanto maior o nível nominal de significância (α), maior o poder do teste. Um aumento do nível de significância leva à redução da região de aceitação. Como resultado, tem-se maior chance de rejeitar a hipótese nula. Isto significa que ocorre uma diminuição da chance de aceitar a hipótese nula quando ela é falsa, isto é, menor chance de cometer um erro do tipo II. Então, o poder do teste aumenta;
- **O verdadeiro valor do parâmetro a ser testado:** quanto maior a diferença entre o verdadeiro valor do parâmetro e o valor especificado pela hipótese nula, maior o poder do teste.

Um teste estatístico é considerado de bom desempenho se tem alto poder e controla o erro tipo I.

Novos testes propostos foram avaliados em cenários considerando algumas distribuições multivariadas como as distribuições normal, normal contaminada, t-student e Cauchy. Neste sentido, as respectivas funções densidade de probabilidade são agora apresentadas.

2.7 A distribuição normal multivariada

Considere um vetor $\mathbf{X} = [X_1, \dots, X_p]^\top$ no qual cada X_i , $i = 1, \dots, p$ é uma variável aleatória normalmente distribuída com média μ_i e variância σ_i^2 . A distribuição conjunta desses componentes gera a distribuição normal multivariada. Então, a distribuição de \mathbf{X} é denotada por $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ em que $\boldsymbol{\mu}$ é um vetor de médias p-dimensional, sendo $\mu_i = E[X_i]$ e a matriz de covariâncias $\boldsymbol{\Sigma}$ é de dimensão $p \times p$, sendo o i, j -ésimo elemento, a $cov(X_i, X_j)$, $i, j = 1, \dots, p$. A função densidade de probabilidade normal multivariada, segundo Johnson e Wichern (2007), será dada pela seguinte expressão

$$f(\mathbf{X}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})}, \quad (2.18)$$

em que p é a dimensão dos dados, $\boldsymbol{\Sigma}$ é a matriz de covariâncias populacional, \mathbf{X} é vetor que contém os dados amostrados e $\boldsymbol{\mu}$ é o vetor de médias populacional. A expressão (2.18) tem suporte em \mathbb{R}^p .

2.8 A distribuição normal multivariada contaminada

Outra distribuição importante é a distribuição normal multivariada contaminada. Considere um vetor aleatório $\mathbf{X} = [X_1, \dots, X_p]^\top \in \mathbb{R}^p$ com distribuição normal multivariada contaminada. A função densidade de probabilidade de um vetor aleatório normal contaminado é dada pela seguinte expressão

$$f(\mathbf{X}) = \delta(2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_1|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{X}-\boldsymbol{\mu}_1)} + (1-\delta)(2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_2|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{X}-\boldsymbol{\mu}_2)}, \quad (2.19)$$

em que δ é a probabilidade que o processo tem de ser realizado por uma normal k variada ($N_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$), $(1-\delta)$ é a probabilidade que o processo tem de ser realizado por uma normal k variada ($N_k(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$), $\boldsymbol{\Sigma}_i$ é uma matriz positiva definida, $\boldsymbol{\mu}_i \in \mathbb{R}^p$ é o vetor de médias, $i = 1, 2$ e $0 \leq \delta \leq 1$.

2.8.1 Geração de variáveis aleatórias com distribuição normal multivariada contaminada

Segundo Johnson e Wichern (2007) a geração de realizações de variáveis aleatórias da distribuição normal contaminada a partir da função densidade dada pela expressão (2.19) pode ser realizada seguindo os seguintes passos:

- I - Gerar um valor u de uma distribuição uniforme contínua com valores entre 0 e 1. Se $u \leq \delta$ avance para o passo II. Caso contrário execute o passo III;
- II - gerar $\mathbf{X} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$;
- III - gerar $\mathbf{X} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$;
- IV - repetir o passo I até que n observações sejam geradas.

Existe um problema ao usar a expressão (2.19) que não se situa na geração de uma realização da variável aleatória, mas sim na seleção dos parâmetros (JOHNSON; WICHERN, 2007). O número de parâmetros da expressão (2.19) é $m^2 + 3m + 1$ que correspondem a $2m$ médias, $2m$ variâncias, $m^2 - m$ correlações e a probabilidade δ de mistura.

Em geral, tem-se

$$f(\mathbf{x}) = \delta f_1(\mathbf{x}) + (1-\delta) f_2(\mathbf{x}). \quad (2.20)$$

A expressão (2.20) é denominada equação de mistura de distribuições e serve para qualquer distribuição de probabilidades.

2.9 A distribuição t-student multivariada

Uma outra distribuição muito importante é a t-student multivariada. Dado um vetor aleatório $\mathbf{X} = [X_1, \dots, X_p]^t \in \mathbb{R}^p$ com função densidade de probabilidade representada pela seguinte expressão

$$f(\mathbf{x}) = \frac{|\boldsymbol{\Sigma}|^{-1/2} \Gamma[(v+p)/2]}{\Gamma(1/2)^p \Gamma(1/2) v^{p/2}} \left[1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{v} \right]^{-(v+p)/2}. \quad (2.21)$$

Diz-se que \mathbf{X} tem função densidade de probabilidade t-student multivariada com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, com v graus de liberdade, cuja notação é dada por $\mathbf{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, v)$. Quando $v \rightarrow \infty$ a expressão (2.21) se aproxima da distribuição normal multivariada (LANGE; LITTLE; TAYLOR, 1989). Um caso específico é a distribuição de Cauchy, a qual é obtida fazendo-se $v = 1$ na expressão (2.21).

Foi apresentado na seção (2.3) que ambas estatísticas de teste T^2 e LRT têm distribuição assintótica qui quadrado sendo, portanto, não exata. Com o intuito de melhorar essa aproximação e, conseqüentemente, obter ganho de poder, neste trabalho foram realizados procedimentos de reamostragem bootstrap paramétrico. Foram, então, construídos novos testes que assintoticamente seguem uma distribuição bootstrap paramétrico. Antes de elucidarmos passo a passo o algoritmo para a construção desses novos testes, serão apresentados os algoritmos bootstrap e bootstrap paramétrico.

2.10 O algoritmo *bootstrap*

O termo *bootstrap*, em Estatística, se refere a fazer inferências sobre a distribuição amostral de uma estatística obtida por reamostragem da amostra original com reposição. Na medida que a distribuição de amostragem “imita” a distribuição original, as inferências são precisas. A precisão melhora à medida que o tamanho da amostra original aumenta, bastando aplicar o teorema central do limite (EFRON; TIBSHIRANI, 1994).

Segundo Efron e Tibshirani (1994), o método *bootstrap* consiste em um número de técnicas que faz parte de uma grande cobertura de técnicas não-paramétricas que são comumente

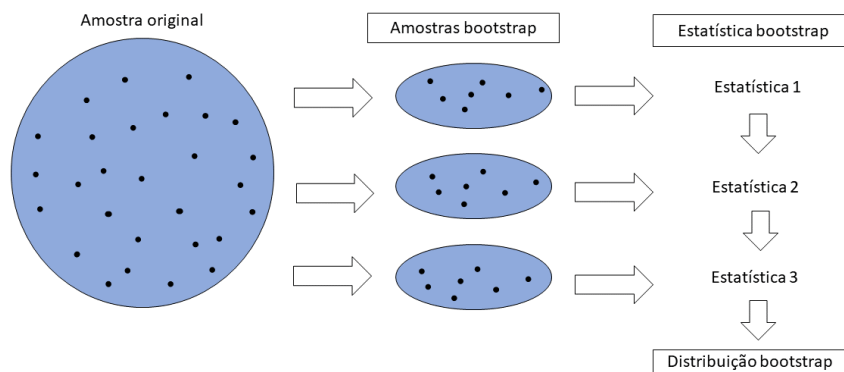
chamadas de métodos de reamostragem. Algumas dessas técnicas são mais antigas do que a técnica *bootstrap*. Cita-se o método de permutação (FISHER, 1935) e o método jackknife, iniciado com Quenouille (1949).

O *bootstrap* foi introduzido em 1979 por Bradley Efron como um método baseado no cálculo para estimação do erro padrão de uma estimativa. Ele tem a vantagem de ser completamente automático. A estimativa *bootstrap* do erro padrão não necessita de cálculos teóricos.

Problemas de inferência estatística frequentemente envolvem algum aspecto de estimação de uma distribuição de probabilidade F com base em uma amostra aleatória. A função de distribuição empírica \hat{F} é uma única estimativa de toda a distribuição F (EFRON; TIBSHIRANI, 1994). Uma maneira óbvia de estimar algum aspecto de interesse de F , como sua média, mediana ou coeficiente de correlação, é usar o aspecto correspondente \hat{F} . Este é o princípio do “plug-in”. O método *bootstrap* é uma aplicação direta do princípio do “plug-in”.

A Figura 2.1 apresenta um esquema do algoritmo *bootstrap*, no qual amostras com reposição de mesmo tamanho são geradas a partir da amostra original. De cada amostra gerada obtém-se uma estatística *bootstrap* e, posteriormente, obtém-se a distribuição de *bootstrap* dessa estatística.

Figura 2.1 – O algoritmo bootstrap.



Fonte: Do próprio autor.

A obtenção dessas amostras *bootstrap* podem ser feitas de forma empírica ou de acordo com algum modelo de probabilidade. Nesse caso, o método *bootstrap* pode ser classificado em *bootstrap* não-paramétrico e *bootstrap* paramétrico. O primeiro faz uso de reamostragem com reposição de uma distribuição empírica \hat{F} . O segundo associa um modelo paramétrico a

função de distribuição F . Este método apresenta duas diferentes vantagens sobre os métodos tradicionais:

1. quando é utilizado um modelo não-paramétrico para a função de distribuição F o analista não precisa fornecer hipóteses paramétricas sob a forma da população subjacente;
2. quando for utilizado um modelo paramétrico para a função de distribuição F , este fornece respostas mais precisas que fórmulas e o mesmo pode fornecer respostas que não estão presentes na literatura.

Nesta tese optou-se por utilizar o método bootstrap paramétrico o qual descrevemos sucintamente a seguir.

2.10.1 O *bootstrap* paramétrico

Suponha que tenha sido coletado uma amostra aleatória $\mathbf{X} = (x_1, x_2, \dots, x_n)$ de uma função de distribuição $f(x|\theta)$ conhecida. Suponha ainda que existe o interesse em estimar um parâmetro θ dessa distribuição f com base na amostra aleatória \mathbf{X} , com θ podendo ser estimado por $\hat{\theta}$. Este é seu estimador de máxima verossimilhança, sendo θ não necessariamente um escalar, podendo ser também um vetor de parâmetros. Então, as amostras *bootstrap* paramétrico são obtidas a partir de $f(\mathbf{X}|\hat{\theta})$ não sendo reamostradas dos dados mas, sim, sendo amostras aleatórias reais geradas utilizando-se o princípio do “plug-in”.

Para o *bootstrap* paramétrico, as amostras x_i^* são geradas a partir de uma função densidade (caso contínuo) ou função de probabilidade (caso discreto) $f(\mathbf{X}|\hat{\theta})$, que supõe-se conhecida (EFRON; TIBSHIRANI, 1994).

Uma vez apresentados os conceitos que visam melhorar a aproximação distribucional dos novos testes propostos, com o intuito de obter ganho em poder, torna-se necessário apresentar o método que foi utilizado para avaliar o erro tipo I e poder destes mesmos testes. Trata-se do método Monte Carlo, que envolve simulação estatística.

2.11 O método Monte Carlo

Trata-se de um mecanismo computacionalmente intensivo originado na década de 1940. Este método pode ser descrito como um estudo de simulação estatística que utiliza sequências de números aleatórios para desenvolver simulações (METROPOLIS; ULAM, 1949).

Segundo Metropolis e Ulam (1949) o método Monte Carlo se torna muito útil na solução de integrais que não possuem solução analítica. Em estatística, esse método permite realizar inferências pelas estimativas obtidas pelo mesmo e quando as suposições de alguns modelos são violadas, além de ser usado na avaliação do desempenho de testes estatísticos. Para este último caso, o método Monte Carlo é utilizado para validar testes quanto ao controle da taxa de erro tipo I e poder.

Mesmo sendo de grande utilidade, ressalta-se que seu uso só é justificado se o procedimento for adequado em substituir um sistema real. Segundo Metropolis e Ulam (1949), estudos de simulação tentam reproduzir um problema real por meio de um ambiente controlado. Por meio deste método pode ser determinado uma solução para um problema real, utilizando uma sequência de números aleatórios para esta representação.

As taxas de erro tipo I e o poder do teste serão avaliados via procedimento Monte Carlo e estes não estão livres de erro, como no caso o erro Monte Carlo. Então, utiliza-se o teste Binomial exato para tomar a decisão se os testes estatísticos são considerados exato, liberal ou conservativo.

2.12 O teste Binomial exato

Seja z o número de hipóteses rejeitadas em N simulações Monte Carlo para um nível nominal de significância α . Segundo Oliveira e Ferreira (2010), utilizando a relação existente entre as distribuições F de Snedecor e Binomial, com probabilidade de sucesso $p = \alpha$, a estatística de teste é dada pela seguinte expressão

$$F = \left(\frac{z+1}{N-z} \right) \left(\frac{1-\alpha}{\alpha} \right) \quad (2.22)$$

que, sob a hipótese nula, segue uma distribuição F com $V_1 = 2(N-z)$ e $V_2 = 2(z+1)$ graus de liberdade. Se $F \leq F_{0,005}$ ou $F \geq F_{0,995}$, a hipótese nula é rejeitada com 1% de significância.

Se a hipótese nula for rejeitada e a taxa de erro tipo I for considerada significativamente inferior ao nível nominal adotado, o teste pode ser considerado conservativo; se a hipótese nula for rejeitada e a taxa de erro tipo I for considerada significativamente superior ao nível nominal adotado, o teste pode ser considerado liberal; e a hipótese nula não for rejeitada o teste pode ser considerado exato. Em geral, adota-se o nível nominal de significância $\alpha = 5\%$.

Na próxima seção apresenta-se uma revisão bibliográfica de propostas de adaptações robustas e não robustas dos testes T^2 tradicional e LRT .

2.13 Adaptações robustas e não robustas aos testes T^2 e LRT

Na literatura existem algumas propostas de adaptações do teste T^2 de Hotelling tradicional com o intuito de utilizá-lo sob distribuições não normais ou sob distribuições normais com e sem contaminação. Nesta seção, algumas dessas propostas são citadas e detalhadas para facilitar o entendimento do leitor. Tais propostas podem ser encontradas em Dempster (1958), Dempster (1960), Tiku (1982), Bai e Saranadasa (1996), Mudholkar e Srivastava (2000), entre outros.

Uma das formas de obter uma adaptação do teste T^2 de Hotelling é substituir seus estimadores por estimadores robustos a outliers. Segundo Mudholkar e Srivastava (2000), a não-robustez da estatística T^2 de Hotelling pode manifestar substancial perda de poder do teste. O desenvolvimento de estimadores robustos pode ser classificado em três categorias: M-estimadores, que são baseados na abordagem do tipo máxima verossimilhança; os L-estimadores, que são baseados em uma combinação linear de estatísticas de ordem; e os R-estimadores, que são baseados em “ranks”. No caso multivariado, os M-estimadores são mais apropriados e aqueles que são baseados na distância quadrática robusta de Mahalanobis serão considerados neste trabalho.

No contexto dos M-estimadores robustos baseados na distância quadrática robusta de Mahalanobis alguns trabalhos foram desenvolvidos com o intuito de contornar os problemas apresentados pela estatística T^2 de Hotelling dos quais podem-se citar, por exemplo, (TIKU, 1982; MUDHOLKAR; SRIVASTAVA, 2000; SRIVASTAVA; MUDHOLKAR, 2001; WILLEMS et al., 2002; WILLIAMS et al., 2006; ALFARO; ORTEGA, 2009), em que o último trabalho citado se refere à proposição de novas cartas de controle de qualidade baseadas na estatística T^2 de Hotelling robusta à presença de outliers.

Tiku (1982) propôs e avaliou, por meio do poder do teste e taxa de erro tipo I, um teste adaptado ao T^2 de Hotelling, robusto na presença de outliers. Este teste foi construído para testar a hipótese sobre o vetor de médias no caso de uma e duas populações, baseado nos M-estimadores robustos. A construção desse novo teste, para o caso em que se tem normalidade multivariada, foi por meio do método médias “trimmed” (aparadas) considerando os quantis da distribuição normal padrão, sendo estes ainda ponderados por seus respectivos desvios. O

percentual de aparamento foi de 10% do tamanho da amostra n . Nesse sentido, Tiku (1982) obteve os estimadores para o vetor de médias e matriz de covariância robustos a presença de outliers, em que não foi considerado o problema de não singularidade da matriz de covariâncias amostral \mathbf{S} para o caso onde se tem mais variáveis p do que observações n . Tiku (1982) derivou a distribuição assintótica dessa novo teste, que segue uma distribuição F central, como é o caso do teste T^2 tradicional. Como resultado ele concluiu que o desempenho desse novo teste, tanto em poder como taxa de erro tipo I, foi melhor em diversos cenários simulados para n e p . Ele também realizou um estudo comparativo considerando as distribuições normal, qui-quadrado, normal contaminada e t-student multivariada com três variáveis ($p = 3$), verificando que em todas as situações, o novo teste teve melhor desempenho.

Mudholkar e Srivastava (2000) construíram um teste alternativo ao teste T^2 de Hotelling tradicional para o vetor de médias de populações multivariadas utilizando alternativas robustas “stepwise”. A ideia foi utilizar médias “trimmed” obtidas passo-a-passo e substituir esse novo estimador do vetor de médias obtido na estatística T^2 de Hotelling tradicional. Em outras palavras, a nova estatística de teste, F_i , foi obtida considerando um ordenamento das variáveis baseado na sua importância “a priori” de acordo com uma escolha flexível de um nível de significância α_i , $i = 1, 2, \dots, p$, considerando p testes independentes, com um percentual de aparamento de 20%, 30% e 33%, respectivamente. Então, o i -ésimo teste considera um nível de significância α_i que era dado por $(1 - \alpha) = \prod_{i=1}^p (1 - \alpha_i)$. Caso algum desses i testes fosse significativo a hipótese nula é rejeitada. Para obter os p-valores p_i dos p testes independentes F_i utilizou-se combinações dos métodos de Tippett, Fisher, Liptak e Mudholkar (MUDHOLKAR; SRIVASTAVA, 2000). Para avaliar o desempenho desse novo teste, Mudholkar e Srivastava (2000) realizaram simulações Monte Carlo considerando os tamanhos amostrais 10 (5) 50 para várias populações bivariadas e trivariadas, seguindo as seguintes distribuições, para $p = 2, 3$, para o caso de uma e duas populações: $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $0,8N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0,2N_p(\boldsymbol{\mu}, 9\boldsymbol{\Sigma})$, $0,8N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0,2N_p(\boldsymbol{\mu}, 16\boldsymbol{\Sigma})$, t-student bivariada e trivariada com três graus de liberdade. Os resultados mostraram melhor desempenho do teste proposto tanto para o controle da taxa de erro tipo I quanto para o poder do teste, quando comparado ao T^2 tradicional.

Seguindo o mesmo objetivo, Srivastava e Mudholkar (2001) construíram um novo teste considerando a hipótese de não normalidade multivariada. Então, eles consideraram o aparamento de amostras multivariadas “trimming” por coordenadas e usaram o vetor de médias “trimmed” para construir e desenvolver o novo teste \tilde{T}^2 com uma média “trimmed” robusta. A

teoria assintótica para os casos univariados e multivariados e simulações Monte Carlo foram usados para construir uma aproximação para a distribuição nula de \tilde{T}^2 . O poder do novo teste foi avaliado e a conclusão foi que o novo teste superou o T^2 tradicional em todos os cenários para n e p . Foram considerados populações normais, normais contaminadas com 20% de contaminação, populações não-normais contaminadas (caudas longas), populações de caudas curtas e populações assimétricas.

Uma das propostas deste trabalho é utilizar o estimador robusto a outliers comedian proposto por Falk (1997) para produzir uma adaptação do teste T^2 de Hotelling e avaliar a taxa de erro tipo I e poder deste novo teste em alguns cenários para o número de observações n e a dimensão dos dados p e compará-lo ao teste T^2 tradicional sob normalidade com contaminação e outras distribuições. Nesse sentido Sajesh e Srinivasan (2012) e Palma e Gallo (2016) avaliaram a “resistência” desse estimador a outliers, mas não obtiveram uma nova estatística de teste para o vetor de médias populacional adaptado ao teste T^2 de Hotelling.

A estatística T^2 de Hotelling é uma importante ferramenta para realizar inferência sobre a média de uma população normal multivariada (WILLEMS et al., 2002). Mas quando desejar-se realizar testes de hipóteses e construir intervalos de confiança para a média, a estatística de teste T^2 definida na expressão (2.6) pode ser afetada por valores discrepantes ou outliers. Então uma alternativa robusta à presença de outliers deve ser construída utilizando algum estimador robusto. Willems et al. (2002) utilizaram o estimador altamente robusto “Minimum Covariance Determinant” (MCD) construído por Rousseeuw e Driessen (1999) para produzir uma estatística de teste altamente robusta a presença de outliers. Os autores substituíram os estimadores média e matriz de covariâncias utilizados na expressão (2.6) pelos respectivos estimadores robustos MCD. Uma alternativa rápida para calcular o MCD é utilizar o FAST-MCD (ROUSSEEUW; DRIESSEN, 1999). Os autores verificaram que os resultados obtidos foram satisfatórios em todos os cenários avaliados.

Willems et al. (2002) construíram uma estatística de teste usando o MCD reponderado que apresentou alta eficiência, e, ainda, derivaram a distribuição desta nova estatística verificando ser ela um múltiplo da distribuição F . A precisão dessa nova distribuição foi avaliada por meio de simulações Monte Carlo, bem como o poder e a robustez do novo teste para o vetor de médias baseado nessa nova estatística.

Seja $\mathbf{X}_n = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ uma amostra de uma distribuição normal p -variada $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ conforme definido na expressão (2.18) com $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ desconhecidos. Para avaliar a hipótese H_0 :

$\boldsymbol{\mu} = \boldsymbol{\mu}_0$ utiliza-se a estatística de teste T^2 clássica. A hipótese H_0 é rejeitada com um nível de significância α se

$$T^2 \geq \frac{(n-1)p}{n-p} F(p, n-p, 1-\alpha) \quad (2.23)$$

em que T^2 é dado na Equação 2.6 e $F(p, n-p, 1-\alpha)$ é o quantil $1-\alpha$ dado pela distribuição F de Snedecor com p e $n-p$ graus de liberdade. Outras aplicações da estatística T^2 incluem a construção de intervalos de confiança e teste de hipótese para duas populações normais.

Dados n pontos em \mathbb{R}^p , o método MCD é determinado pelo subconjunto de tamanho $h = \lfloor \gamma n \rfloor$, com $0,5 \leq \gamma \leq 1$ cuja matriz de covariâncias tem o menor determinante, em que $\lfloor \cdot \rfloor$ representa o menor inteiro. A estimativa da média MCD, T , é definida como a média desse subconjunto e a estimativa de covariância MCD é definida como um múltiplo C da matriz de covariância desse subconjunto. O fator de multiplicação consiste em um fator c_γ e um fator de correção finito-amostrai. O fator de correção dado em Croux e Haesbroeck (1999) que, segundo Willems et al. (2002) fornece o estimador de Fisher-consistente MCD de covariâncias de um modelo normal e é igual a $c_\delta = \gamma / F_{\chi_{p+2}^2}(q_\gamma)$ em que $q_\gamma = \chi_{p,\gamma}^2$. Para utilizar o MCD reponderado, Willems et al. (2002) adotaram a seguinte estratégia. Observações com distância menor que $d_i(T, C)$ recebem peso 1, enquanto as outras recebem peso 0 (WILLEMS et al., 2002). O fator $c_\delta = (1-\delta) / F_{\chi_{p+2}^2}(q_\delta)$ em que $q_\delta = \chi_{p,\delta}^2$ fornece o MCD reponderado consistente de um modelo normal. O termo δ é a fração de observações que obtiveram peso 0 e têm uma distância robusta maior que $2q$ (outliers). A constante $d_{n,p}$ é um fator de correção finito-amostrai dado em Pison, Aelst e Willems (2001).

Segundo Willems et al. (2002), o estimador MCD reponderado é afim-equivariante para a média e matriz de covariâncias, preservando as propriedades ponto de ruptura e afim-equivariância. Willems et al. (2002) afirmam ter preferido usar $\delta = 0,75$ porque este valor garante uma melhor eficiência e um ponto de ruptura de 25% que é mais realístico.

Contudo, estes testes de hipóteses e intervalos de confiança são baseados nos estimadores da média e covariância clássicos e são afetados por outliers. Nesse sentido, Willems et al. (2002) propuseram utilizar os estimadores para a média e matriz de covariâncias MCD reponderado como alternativa para a média e matriz de covariância amostral do teste T^2 tradicional. Então a nova estatística de teste é dada por

$$T_R^2 = n(T^1 - \boldsymbol{\mu})^\top (C^1)^{-1} (T^1 - \boldsymbol{\mu}), \quad (2.24)$$

em que $T^1 = \sum_{i=1}^n w_i \mathbf{x}_i / \sum_{i=1}^n w_i$ e $C^1 = c_\delta d_{n,p} \sum_{i=1}^n w_i (\mathbf{x}_i - T^1) / \sum_{i=1}^n w_i$ em que T^1 e C^1 são a média e covariância reponderadas, respectivamente. Os pesos w_i são baseados nas distâncias robustas das observações, que são dadas por $d_i(T, C) = \sqrt{(x_i - T)^\top (C)^{-1} (x_i - T)}$.

A distribuição finito-amostal de T_R^2 é desconhecida mas tem propriedades similares aos estimadores clássicos (WILLEMS et al., 2002). Então, para uma amostra de tamanho n de uma $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ os estimadores MCD reponderados para a média e matriz de covariâncias satisfazem as seguintes propriedades

1. Deve existir algum k tal que $T^1 \sim N_p(\boldsymbol{\mu}, k \frac{1}{n} \boldsymbol{\Sigma})$;
2. Deve existir m e c tal que $mc^{-1} C^{-1} \sim W_p(\boldsymbol{\Sigma}, m)$ e $E[C^1] = c \boldsymbol{\Sigma}$;
3. T^1 e C^1 são independentes.

Estas propriedades permitem obter uma aproximação para a distribuição de T_R^2 , que segue uma distribuição F aproximada de forma análoga ao T^2 clássico.

Willems et al. (2002) obtiveram a seguinte aproximação

$$T_R^2 \approx d F_{p,q}. \quad (2.25)$$

O fator multiplicação d e q , os graus de liberdade do numerador e do denominador da distribuição F, serão obtidos conhecendo a média e a variância da distribuição dada na expressão 2.25 (WILLEMS et al., 2002). Para determinar d e q , basta considerar uma amostra de uma $N_p(\mathbf{0}, \mathbf{I}_p)$. Por definição da distribuição F segue que

$$E[T_R^2] = d \frac{q}{q-2} \quad (2.26)$$

e

$$Var[T_R^2] = d^2 \frac{2q^2(p+q-2)}{p(q-4)(q-2)^2}. \quad (2.27)$$

Então, d e q são dados por

$$d = E[T_R^2] \frac{q-2}{q} \quad \text{e} \quad q = \left(\frac{Var[T_R^2] p}{2E[T_R^2]} - 1 \right)^{-1} (p+2) + 4. \quad (2.28)$$

Segundo Willems et al. (2002), uma vez que a média e a variância da distribuição de T_R^2 não pode ser obtida analiticamente, elas deverão ser aproximadas por simulação, que resultará

numa aproximação precisa para a verdadeira distribuição de T_R^2 . No entanto, Willems et al. (2002) usaram a distribuição assintótica de T_R^2 para concluir que

$$T_R^2 \sim k\chi_p^2, n \rightarrow \infty \quad (2.29)$$

e

$$E[T_R^2] \xrightarrow{n \rightarrow \infty} kp \quad \text{Var}[T_R^2] \xrightarrow{n \rightarrow \infty} K^2 2p \quad (2.30)$$

em que K é dado pela Tabela 2.2 e é dependente do valor de p .

Tabela 2.2 – Variância assintótica (k) de T_j^1 para o modelo normal padrão.

p	1	2	3	5	10	20	50
k	1.258	1.145	1.112	1.085	1.063	1.050	1.040

Para obter os valores das constantes d e q , Willems et al. (2002) realizaram simulações Monte Carlo considerando vários tamanhos amostrais $n = 10, 20, 30, 50, 70, 100, 140, 200$ e de dimensões $p = 5, 10, 15$. Foram gerados $m = 3000$ amostras $X_j, j = 1, \dots, m$ de uma distribuição normal padrão. Para cada X_j foram calculados $T_{(j)}^1$ e $C_{(j)}^1$, os parâmetros média e covariância do MCD reponderado e o correspondente valor $T^{2(j)R}$. Segundo Willems et al. (2002), a média e a variância $T^{2(j)R}$ são dados por

$$m(T^{2(j)R}) = \frac{1}{m} \sum_{j=1}^m T^{2(j)R} \quad \text{e} \quad S^2(T^{2(j)R}) = \frac{1}{m-1} \sum_{j=1}^m (T^{2(j)R} - m(T^{2(j)R}))^2. \quad (2.31)$$

Os resultados mostraram que

- Se $p = 1$ ou $p = 2$ então $E[T_R^2]$ é aproximado por $f(n) = k + \alpha/(n^\beta)$ e $f_2(n) = 2k + \alpha_2/(n_2^\beta)$, respectivamente, sendo que os valores α e β foram determinados pelos autores;
- Se $p > 2$, então é preciso resolver o seguinte sistema para obter os valores de $\hat{\alpha}_\beta$ e $\hat{\beta}_p$:

$$\ln(\hat{\alpha}_\beta) - \hat{\beta}_p \ln(7p^2) = \ln\left(\frac{\gamma_7}{p^{\delta_7}}\right) \quad \text{e} \quad \ln(\hat{\alpha}_\beta) - \hat{\beta}_p \ln(10p^2) = \ln\left(\frac{\gamma_{10}}{p^{\delta_{10}}}\right);$$

- Então, o valor de $E[T_R^2]$ é aproximado por $\hat{f}_p(n) = kp + \frac{\hat{\alpha}_p}{n^{\hat{\beta}_p}}$.

Segundo Willems et al. (2002), $\text{Var}[T_R^2]$, iniciando pelo ajuste de $S^2(T_R^2)$ é derivada para algum p fixado por $h_p(n) = k^2 2p + \varepsilon_p/n^{\zeta_p}$. Se $p \neq 3$ então $\text{Var}[T_R^2] \approx h_p(qp^2) - k^2 2p$.

A distribuição aproximada de T_R^2 pode ser obtida por meio das estimativas obtidas de $E[T_R^2]$ e $Var[T_R^2]$. Utilizando de simulações, Willems et al. (2002) obtiveram os valores de d e q dados pela expressão (2.28) e verificaram a precisão dessa aproximação para diversos tamanhos de n e p com $m = 3000$ dados $Z_j, j=1, \dots, 3000$, com $Z_j \sim N_p(\mathbf{0}, \mathbf{I}_p)$. Para cada Z_j foi calculado a estatística T_R^2 , e verificaram que para as 3000 estatísticas T_R^2 foi aproximada a distribuição dada pela expressão (2.25). Então, Willems et al. (2002) concluíram que a distribuição aproximada T_R^2 obtida por simulação, é uma boa aproximação usando os valores de d e q obtidos pela expressão (2.28).

Willems et al. (2002) tinham como objetivo construir inferências para a estatística T_R^2 , em que o foco era testar a hipótese para a média populacional considerando uma amostra. Então, eles compararam a distribuição aproximada de T_R^2 com outras distribuições aproximadas que são menos precisas. Os autores comparam a distribuição aproximada de T_R^2 com um múltiplo da distribuição qui-quadrado ($k\chi_p^2$). A conclusão foi que para tamanhos amostrais n pequeno T_R^2 teve maior precisão. Quando se aumenta o tamanho da amostras e este converge para infinito o desempenho de ambas é praticamente igual. Também compararam a distribuição de T_R^2 com a distribuição do T^2 clássico e verificaram que a distribuição aproximada do T_R^2 foi mais precisa.

Quanto ao poder do teste, Willems et al. (2002) concluíram que o teste T_R^2 foi muito mais poderoso quando comparado a outros testes. Porém, para tamanhos amostrais pequenos, houve eventual perda de poder no teste T_R^2 , sendo esta perda aceitável. Quanto ao desempenho de T_R^2 na presença de outliers, Willems et al. (2002) geraram uma distribuição normal multivariada com 10% de contaminação. Mesmo sob essas condições, os autores concluíram que o desempenho do teste T_R^2 esteve perto do nível nominal de 5% considerado, apresentando uma melhoria comparado ao teste T^2 clássico.

Na próxima seção apresenta-se uma discussão em torno de algumas propostas de testes alternativos para dados de alta dimensão ($n < p$).

2.14 Testes adaptados para dados de alta dimensão

Baseado nos trabalhos de Dempster (1958), Dempster (1960), Bai e Saranadasa (1996b), Srivastava e Du (2008) propuseram um teste alternativo ao teste T^2 para dados de alta dimensão. Os autores consideraram, em suas hipóteses, um vetor de médias de variáveis aleatórias normais independentes e identicamente distribuídas. O teste proposto é invariante sob transformações de escala em cada componente do vetor aleatório. Os autores apresentaram resultados teóricos

e simulados e observaram que o desempenho dos testes propostos foi satisfatório. O interesse dessa proposta foi motivado por dados de DNA. A metodologia proposta só é válida para dados normais de alta dimensão.

Em seu novo trabalho, Srivastava (2009), com algumas alterações na metodologia proposta em Srivastava e Du (2008), produziu um teste alternativo ao teste T^2 para dados não normais de alta dimensão. Porém, para esse novo teste, os vetores aleatórios que compunham sua matriz de dados simulados possuíam, cada um, os quatro primeiros momentos finitos. Os autores encontraram as distribuições nula e não-nula do novo teste proposto.

Embora não seja o objetivo destas notas, não se pode deixar de mencionar os trabalhos de Chen, Qin et al. (2010), Lee et al. (2012), Bai e Saranadasa (1996b), em que os autores propuseram um teste alternativo ao teste T^2 de Hotelling tradicional, considerando duas populações multivariadas e dados de alta dimensão.

Wang, Peng e Qi (2013) produziram um teste de médias para dados de alta dimensão que envolve duas populações multivariadas, sendo este alternativo ao teste T^2 tradicional. Neste trabalho, os autores propuseram um teste de verossimilhanças Jackknife empírico em que sua distribuição nula foi aproximada por uma qui-quadrado independente da dimensão dos dados. Um estudo simulado mostrou que o novo teste proposto teve alto poder e controle do erro tipo I em todas as dimensões avaliadas.

Em trabalhos anteriormente já citados, como em Srivastava e Du (2008), Srivastava (2009), Chen, Qin et al. (2010), Lee et al. (2012), Srivastava, Katayama e Kano (2013), Wang, Peng e Qi (2013), os autores também optaram ora pelo traço da matriz de covariâncias \mathbf{S} , ora pelo traço da matriz de correlação \mathbf{R} , como em Ledoit e Wolf (2002). Em ambas as situações, o problema de não utilização do teste T^2 para dados de alta dimensão foi contornado.

Baseado nessas propostas, nesta tese foram propostos novos testes alternativos ao teste LRT , considerando o traço da matriz \mathbf{S} ao invés do determinante desta mesma matriz. Desta forma, será considerado a variância total dos dados ao invés da variância generalizada. Na próxima seção serão apresentados, em forma de dois artigos, a metodologia, os resultados e a discussão em relação ao desempenho dos novos testes propostos.

3 CONSIDERAÇÕES FINAIS

Nesta tese, objetivando realizar testes de hipóteses que envolvam o vetor de médias de uma população multivariada, foram propostos novos testes adaptados ao teste T^2 de Hotelling tradicional e ao teste da razão de verossimilhanças LRT . Essas metodologias propostas tem a vantagem de serem robustas a outliers e as versões *bootstrap* paramétrico melhoram a aproximação assintótica distribucional desses novos testes. Além disso, os novos testes adaptados ao teste LRT propostos possuem a vantagem de serem válidos para dados de alta dimensão. Os resultados apresentados nos dois artigos indicam que a versão *bootstrap* paramétrico do teste LRT teve melhor desempenho em todos os cenários avaliados. Esse teste controlou as taxas de erro tipo I e poder. A metodologia proposta possibilita estudos futuros como, por exemplo, uma adaptação da versão *bootstrap* paramétrico do teste LRT para testar as hipóteses de diferença significativa sobre os vetores de médias de duas populações multivariadas, com variâncias desconhecidas e heterogêneas. Esse problema é conhecido como problema de aproximação de Behrens-Fisher.

REFERÊNCIAS

- ALFARO, J. L.; ORTEGA, J. F. A comparison of robust alternatives to hotelling's T^2 control chart. **Journal of Applied Statistics**, Taylor & Francis, v. 36, n. 12, p. 1385–1396, 2009.
- BAI, Z.; SARANADASA, H. Effect of high dimension: by an example of a two sample problem. **Statistica Sinica**, JSTOR, v. 6, p. 311–329, 1996.
- BAI, Z.; SARANADASA, H. Effect of high dimension: by an example of a two sample problem. **Statistica Sinica**, JSTOR, p. 311–329, 1996.
- CHEN, S. X.; QIN, Y.-L. et al. A two-sample test for high-dimensional data with applications to gene-set testing. **The Annals of Statistics**, Institute of Mathematical Statistics, p. 808–835, 2010.
- CROUX, C.; HAESBROECK, G. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. **Journal of Multivariate Analysis**, Elsevier, v. 71, n. 2, p. 161–190, 1999.
- DEMPSTER, A. P. A high dimensional two sample significance test. **The Annals of Mathematical Statistics**, JSTOR, p. 995–1010, 1958.
- DEMPSTER, A. P. A significance test for the separation of two highly multivariate small samples. **Biometrics**, JSTOR, v. 16, n. 1, p. 41–50, 1960.
- DONG, K. et al. Shrinkage-based diagonal hotelling's tests for high-dimensional small sample size data. **Journal of Multivariate Analysis**, Elsevier, v. 143, p. 127–142, 2016.
- DONOHU, D. L.; HUBER, P. J. The notion of breakdown point. **A festschrift for Erich L. Lehmann**, v. 157184, 1983.
- EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. University of Toronto: CRC press, 1994.
- FALK, M. On mad and comedians. **Annals of the Institute of Statistical Mathematics**, Springer, v. 49, n. 4, p. 615–644, 1997.
- FERREIRA, D. F. **Estatística multivariada**. 3. ed. Universidade Federal de Lavras: Editora UFLA, 2018. 624 p.
- FISHER, R. A. The design of experiments. 1935. **Oliver and Boyd, Edinburgh**, 1935.
- HAMPEL, F. R.; RONCHETTI, E. R. Robust statistics: The approach based on influence functions. **PJ and Stahel, WA**, 1986.
- HUBERT, M.; ROUSSEEUW, P. J.; AELST, S. V. High-breakdown robust multivariate methods. **Statistical Science**, JSTOR, p. 92–119, 2008.
- JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate correspondence analysis**. USA: Prentice-Hall, Upper Saddle River, NJ, 2007.
- LANGE, K. L.; LITTLE, R. J. A.; TAYLOR, J. M. G. Robust statistical modeling using the t distribution. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 84, n. 408, p. 881–896, 1989.

- LEDOIT, O.; WOLF, M. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 30, n. 4, p. 1081–1102, 2002.
- LEE, S. H. et al. Order test for high-dimensional two-sample means. **Journal of Statistical Planning and Inference**, Elsevier, v. 142, n. 9, p. 2719–2725, 2012.
- MARONNA, R. A.; ZAMAR, R. H. Robust estimates of location and dispersion for high-dimensional datasets. **Technometrics**, Taylor & Francis, v. 44, n. 4, p. 307–317, 2002.
- METROPOLIS, N.; ULAM, S. The monte carlo method. **Journal of the American statistical association**, Taylor & Francis, v. 44, n. 247, p. 335–341, 1949.
- MUDHOLKAR, G. S.; SRIVASTAVA, D. K. A class of robust stepwise alternatives to hotelling's T^2 tests. **Journal of Applied Statistics**, Taylor & Francis, v. 27, n. 5, p. 599–619, 2000.
- OLIVE, D. J. Applied robust statistics. **Preprint M-02-006**, 2008.
- OLIVEIRA, I. R. Cardoso de; FERREIRA, D. F. Multivariate extension of chi-squared univariate normality test. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 80, n. 5, p. 513–526, 2010.
- PALMA, M. A. D.; GALLO, M. A co-median approach to detect compositional outliers. **Journal of Applied Statistics**, Taylor & Francis, v. 43, n. 13, p. 2348–2362, 2016.
- PAN, G. M.; ZHOU, W. Central limit theorem for hotelling's T^2 statistic under large dimension. **The Annals of Applied Probability**, JSTOR, p. 1860–1910, 2011.
- PISON, G.; AELST, S. V.; WILLEMS, G. Small sample corrections for lts and mcd. 2001.
- QUENOUILLE, M. H. Approximate tests of correlation in time-series. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 11, n. 1, p. 68–84, 1949.
- ROUSSEEUW, P. J.; DRIESSEN, K. V. A fast algorithm for the minimum covariance determinant estimator. **Technometrics**, Taylor & Francis Group, v. 41, n. 3, p. 212–223, 1999.
- SAJESH, T. A.; SRINIVASAN, M. R. Outlier detection for high dimensional data using the comedian approach. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 82, n. 5, p. 745–757, 2012.
- SRIVASTAVA, D. K.; MUDHOLKAR, G. S. Trimmed T^2 : A robust analog of hotelling's T^2 . **Journal of Statistical Planning and Inference**, Elsevier, v. 97, n. 2, p. 343–358, 2001.
- SRIVASTAVA, M. S. A test for the mean vector with fewer observations than the dimension under non-normality. **Journal of Multivariate Analysis**, Elsevier, v. 100, n. 3, p. 518–532, 2009.
- SRIVASTAVA, M. S.; DU, M. A test for the mean vector with fewer observations than the dimension. **Journal of Multivariate Analysis**, Elsevier, p. 386–402, 2008.
- SRIVASTAVA, M. S.; KATAYAMA, S.; KANO, Y. A two sample test in high dimensional data. **Journal of Multivariate Analysis**, Elsevier, v. 114, p. 349–358, 2013.

TIKU, M. L. Robust statistics for testing equality of means or variances. **Communications in Statistics-Theory and Methods**, Taylor and Francis, v. 11, n. 22, p. 2543–2558, 1982.

WANG, R.; PENG, L.; QI, Y. Jackknife empirical likelihood test for equality of two high dimensional means. **Statistica Sinica**, JSTOR, p. 667–690, 2013.

WILLEMS, G. et al. A robust hotelling test. **Metrika**, Springer, p. 125–138, 2002.

WILLIAMS, J. D. et al. Distribution of hotelling's T^2 statistic based on the successive differences estimator. **Journal of Quality Technology**, American Society for Quality, v. 38, n. 3, p. 217, 2006.

SEGUNDA PARTE

ARTIGO 1 - PROPOSITION OF NEW ALTERNATIVE TESTS ADAPTED TO THE T^2 TEST.

**Publicado no journal Communications in Statistics - Simulation and Computation - Doi:
<https://doi.org/10.1080/03610918.2019.1693596>**

ABSTRACT

New alternative tests of the T^2 Hotelling's test for testing hypotheses on the mean vector of a normal p -variate population were proposed. These test were based on comedian robust estimator of the covariance matrix using an asymptotic T^2 distribution and a parametric bootstrap distribution to the null distribution of the statistical tests. The performance of these new tests was evaluated under normal and non-normal distributions through Monte Carlo simulations. The contaminated normal populations were also considered to evaluate the effects of outliers in the tests performance. The type I error rates and power were computed in all Monte Carlo simulations by using the R software. The parametric bootstrap test based on the T^2 test statistic had equivalent performance of the T^2 original test. This test was recommend because it is easy to implement and computationally fast.

Keywords: power. type I error. parametric bootstrap. robust. comedian.

1 INTRODUCTION

Too often we are faced with the problem of ascertaining whether the unknown mean vector $\boldsymbol{\mu}$ of a p -variate normal population is equal to a fixed vector $\boldsymbol{\mu}_0$. Consider a random sample of size n of p -variate vectors, $\mathbf{X}_j = [X_{j1}, \dots, X_{jp}]^\top$ from this population, where $j = 1, 2, \dots, n$ and p refers to the number of variates in each random vector. Generally, the p variates components of each random vectors are correlated and their covariance matrix is the positive definite matrix $\mathbf{\Sigma}$ of order $p \times p$. Therefore the interest is to test the following hypotheses

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \quad (1.1)$$

considering the underlying distribution is a multivariate normal distribution.

One of the most commonly statistical tests for this purpose is the Hotelling's T^2 , where the test statistic is given by

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \quad (1.2)$$

and $\bar{\mathbf{X}} = \sum_{j=1}^n \mathbf{X}_j / n$ is the sample mean vector, $\mathbf{S} = \sum_{j=1}^n (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})^\top / (n - 1)$ is the sample covariance matrix and n is the sample size. Under H_0 , normality and homoscedastic covariance assumptions the T^2 statistic follows a $(n - 1)pF_{p, n-p} / (n - p)$ distribution, where $F_{p, n-p}$ is the F distribution with p and $n - p$ degrees of freedom.

Under asymmetric non normal distributions or in the presence of outliers the use of the T^2 test statistic (1.2) is not recommended. First, because the main assumption for this test is the multivariate normality and second, since even under normality, the test statistics considers in its expression the sample mean vector $\bar{\mathbf{X}}$ and the covariance matrix \mathbf{S} , which are heavily influenced by outliers as in the univariate case (WILLEMS et al., 2002). Besides those two limitations, the test statistic T^2 can not be computed when $p \geq n$ because of singularity of \mathbf{S} (BAI; SARANADASA, 1996). Bai e Saranadasa (1996) noted that the power of the traditional T^2 test can be adversely affected when $p < n$, if the sample covariance matrix \mathbf{S} is quase singular. This was also shown in Pan e Zhou (2011).

It is a challenge testing the hypotheses in (1.1) under asymmetric non normal distributions or in the presence of outlier due to its influence on the estimators of the mean vector and the covariance matrix used in the test statistic (1.2). Some results of robust tests can be found in Tiku (1982), Mudholkar e Srivastava (2000), Willems et al. (2002).

In the current paper new statistical tests are proposed adapting the T^2 using comedian affine-equivariant robust estimators of the mean vector and the covariance matrix (SAJESH; SRINIVASAN, 2012). The basic idea is to replace the estimators $\bar{\mathbf{X}}$ and \mathbf{S} for their robust comedian estimators and improve the tests accuracy by mean of the parametric bootstrap distribution under the null hypothesis. These new tests have potentially some advantages over the most recent adapted tests proposed above as, for example, they can show a better behaviour under non-normality and in the presence of outliers. Also, they are computationally simple to implement and fast to execute.

The performance of the new proposed tests was evaluated by Monte Carlo simulations computing the type I error rates and power. In section 2 we introduce the proposed new tests. In section 3 we present the results of the evaluation of the type I error rates and power by Monte Carlo simulations. To evaluated the tests control of the type I error rates, we use the exact binomial test presented by Oliveira e Ferreira (2010). In section 4 we present our conclusions.

2 THE ROBUST ESTIMATOR COMEDIAN

Consider X and Y two random variables. Falk (1997) defined the robust estimator comedian by the following expression

$$COM(X, Y) = med[(X - med(X))(Y - med(Y))], \quad (2.1)$$

in which the term “med” denotes the median. The comedian is an advantageous estimation method, since it always exists while $Cov(X, Y)$ requires the existence of the first two moments of X and Y . Besides that, it is symmetric, invariant to location and scale transformation. This means that $COM(X, aY + B) = aCOM(X, Y)$, where $a \in \mathbb{R}$ and B is a vector $n \times 1$ of constants. This estimator has also strong asymptotic consistency and normality (FALK, 1997). The comedian generalizes the MAD (median absolute deviation) method, also introduced by (FALK, 1997), which can be obtained by making $X = Y$. The MAD will be useful for building the p -variant version of comedian.

Sajesh e Srinivasan (2012) provide the use of the p -variate version of the comedian estimator. Set $\mathbf{X}_{n \times p}$ as a data matrix with vector columns \mathbf{X}_j , $j = 1, 2, \dots, p$. A comedian matrix $COM(\mathbf{X}) = \{MAD(\mathbf{X}_i, \mathbf{X}_j), \text{ if } i = j; COM(\mathbf{X}_i, \mathbf{X}_j), \text{ if } i \neq j\}$ is a symmetric $p \times p$ matrix.

The comedian matrix has an issue of being non-positively defined, a requirement of a covariance matrix. To circumvent this problem, Maronna e Zamar (2002) adopted the following steps to obtain robust estimates for the mean vector and covariance matrix (PALMA; SRINIVASAN; SAJESH, 2013):

1. let $\boldsymbol{\delta}(\mathbf{X}) = \mathbf{DCOM}(\mathbf{X})\mathbf{D}^\top$ be the median correlation matrix, where \mathbf{D} is a diagonal matrix with elements $1/MAD(\mathbf{X}_i)$, $i = 1, 2, \dots, p$;
2. calculate the eigenvalues λ_i and eigenvectors \mathbf{e}_j of the $\boldsymbol{\delta}(\mathbf{X})$, $j = 1, 2, \dots, p$ and let \mathbf{E} be a matrix whose columns are the \mathbf{e}'_j 's. Hence, $\boldsymbol{\delta}(\mathbf{X}) = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top$, where $\boldsymbol{\Lambda} = diag(\lambda_1, \dots, \lambda_p)$;
3. let $\mathbf{Q} = \mathbf{DE}$, where \mathbf{D} is defined in step 1 and $\mathbf{Z}_i = \mathbf{Q}^{-1}\mathbf{x}_i$, $i = 1, 2, \dots, n$;
4. the robust estimators for the mean vector $\mathbf{m}(\mathbf{X})$ and covariance matrix $\mathbf{S}(\mathbf{X})$ are

$$\mathbf{S}(\mathbf{X}) = \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}^\top \quad \text{and} \quad \mathbf{m}(\mathbf{X}) = \mathbf{Q}\mathbf{l} \quad (2.2)$$

where $\boldsymbol{\Gamma} = diag(MAD(\mathbf{Z}_1)^2, \dots, MAD(\mathbf{Z}_p)^2)$ and $\mathbf{l} = (med(\mathbf{Z}_p))^\top$, $p = 1, \dots, p$.

Estimates can be provided through an iterative process by replacing δ by \mathcal{S} and repeating the steps 2, 3 and 4.

Both estimator are affine-equivariant estimator of location and scale, respectively, and the comedian covariance matrix is positive definite estimator as required (SAJESH; SRINIVASAN, 2012). In addition, the estimates obtained by the COMEDIAN method have a high breaking point. The efficiency of the method increases with increasing data size (SAJESH; SRINIVASAN, 2012).

3 METHODS

In this section we describe the new propositions based on modifications of the T^2 statistic test and the Monte Carlo simulations for evaluating their performance. We consider a sample of size n from a normal p -variate distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The null hypothesis H_0 is given in (1.1). In all cases the traditional T^2 test, where the test statistic is given in (1.2), is applied. Two cases were considered to evaluate the performance of the tests. First, the type I error rates were considered, simulating sample sizes from populations with mean equal to the null hypothesis mean vector value $\boldsymbol{\mu}_0$. Second, samples were generate under H_1 , i. e., with $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. In both cases we consider simulating samples from normal and non-normal distributions. Initially, we consider the multivariate Student t -distribution with 5 and 30 degrees of freedom as well as a contaminated multivariate normal population for the non-normal cases. Some combinations of the number of variables p and the sample size n were considered.

The covariance matrix $\boldsymbol{\Sigma}$ has the compound symmetry structure given by

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} = \sigma^2[(1 - \rho)\mathbf{I} + \rho\mathbf{J}], \quad (3.1)$$

where \mathbf{J} is a square matrix ($p \times p$) with all entries equal to 1 and \mathbf{I} is a identity matrix of order $p \times p$. Without loss of generality $\sigma^2 = 1$ and $\rho = 0.9$. We justify the chosen degrees of freedom for the multivariate Student t -distribution since it shows a more heavy tails than normal and violates the main assumptions that are essential for the use of the T^2 statistic. The t_{30} multivariate distribution are more close to the normal distribution than the multivariate t_5 -distribution. The contaminated multivariate normal distribution is given by the following expression: $wN_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) + (1 - w)N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$. We set $w = 0.9$, $\boldsymbol{\Sigma}_1$ using (2.1) and get the covariance matrix $\boldsymbol{\Sigma}_2$ using the following restriction: $|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1| = \Delta$, thus $\boldsymbol{\Sigma}_2 = \Delta^{1/p}\boldsymbol{\Sigma}_1$ and set $\Delta = 2$. We consider the sample sizes of $n = 10, 50, 70, 100$ and 200 and significance level $\alpha = 5\%$. We also considered the number of variables $p = 2$ and 5 . We used 2,000 Monte Carlo simulations to evaluate estimate the empirical type I error and power of each test. In the parametric bootstrap versions, from the original sample, we consider 2,000 resamples from a $N(\mathbf{0}, \mathbf{S}^\bullet)$ distribution, where the mean is considered $\boldsymbol{\mu} = \mathbf{0}$ to impose H_0 to generate the null distribution of the statistic and \mathbf{S}^\bullet is the original sample covariance matrix computed using the traditional or the robust estimator.

Without loss of generality, the value of the vector $\boldsymbol{\mu}_0$ under the null hypothesis in (1.1) was considered the null p -dimensional vector $\mathbf{0}$ and the true population mean vector was also $\boldsymbol{\mu} = \mathbf{0}$. Under the alternative hypothesis H_1 in (1.1), the true population mean vector $\boldsymbol{\mu}$ was chosen considering a fixed generalized Mahalanobis distance $\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ between $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_0$, that is given by

$$\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0), \quad (3.2)$$

where $\boldsymbol{\mu}_0 = \mathbf{0}$ in this case, without loss of generality. The true mean vector is computed by trial and error and the final value was used as parameter in each of the above distribution under H_1 . The chosen values of δ were 0, 0.5, 1, 1.5, 3, 5, 10.

Therefore, we proposed two new tests based on the traditional T^2 test including the parametric bootstrap as well as the robust versions using the mean and covariance matrix comedian estimators (FALK, 1997; MARONNA; ZAMAR, 2002; SAJESH; SRINIVASAN, 2012). This performance was check only by Monte Carlo simulations. Below we describe each of the new proposed tests for the hypothesis $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$.

3.1 The parametric bootstrap T^2 test

We construct the parametric bootstrap T^2 test, called T_{PB}^2 ($T2PB$), where PB stands for parametric bootstrap, adopting the following steps:

1. from the original sample, the parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ were estimated, respectively, by \mathbf{S}^* and $\bar{\mathbf{X}}^*$, where \mathbf{S}^* and $\bar{\mathbf{X}}^*$ are the original sample covariance matrix and vector mean, respectively. The test statistic is computed by

$$T^{*2} = n(\bar{\mathbf{X}}^* - \boldsymbol{\mu}_0)^\top \mathbf{S}^{*-1}(\bar{\mathbf{X}}^* - \boldsymbol{\mu}_0). \quad (3.3)$$

2. using the covariance estimates \mathbf{S}^* , a random sample of size n is generated from a p -variate normal distribution imposing H_0 , that is, by setting $\boldsymbol{\mu} = \boldsymbol{\mu}_0$. Also $\boldsymbol{\Sigma} = \mathbf{S}^*$. Therefore a sample of size n is generated from a $N(\boldsymbol{\mu}_0, \mathbf{S}^*)$ distribution;
3. in each parametric bootstrap sample the sample mean $\bar{\mathbf{X}}_{PB}$ and the sample covariance matrix \mathbf{S}_{PB} were estimated;

4. in this step the test statistic is computed by

$$T_{PB}^2 = n(\bar{\mathbf{X}}_{PB} - \boldsymbol{\mu}_0)^\top \mathbf{S}_{PB}^{-1}(\bar{\mathbf{X}}_{PB} - \boldsymbol{\mu}_0); \quad (3.4)$$

5. the steps (2) to (4) are repeated B times and a set of size $B + 1$ was constructed with the test statistic values computed in (2.4) and the original value computed in (2.3). The null distribution of the parametric bootstrap test was obtained using this set. Therefore, if the i th member of this set was represented by T_i^2 , $i = 1, 2, \dots, B + 1$, then the p -value was computed by

$$p\text{-value} = \frac{\sum_{i=1}^{B+1} I(T_i^2 \geq T^{*2})}{B + 1}, \quad (3.5)$$

where $I(T_i^2 \geq T^{*2})$ is the indicator function.

6. The null hypothesis in (1.1) should be rejected at the significance level α if the p -value $\leq \alpha$.

3.2 The robust parametric bootstrap T^2 test

The robust parametric bootstrap T^2 test, called T_{RPB}^2 ($T2RPB$), in which RPB stands for robust parametric bootstrap, are performed by adopting the same steps of the previous test. However some of them are modified as described in the following sequence. In the step 1, the estimators \mathbf{S}^* and $\bar{\mathbf{X}}^*$ are replaced by comedian estimators \mathbf{X}_R and \mathbf{S}_R and the original test statistic is computed by

$$T^{*2} = n(\bar{\mathbf{X}}_R - \boldsymbol{\mu}_0)^\top \mathbf{S}_R^{-1}(\bar{\mathbf{X}}_R - \boldsymbol{\mu}_0). \quad (3.6)$$

In the step 2, the sample of size n is generated from a $N(\boldsymbol{\mu}_0, \mathbf{S}_R)$ distribution, where again, the null hypothesis is imposed by considering the multivariate normal mean equal to $\boldsymbol{\mu}_0$, the null value of the population mean. In step 3, the mean and the sample covariance in each parametric bootstrap sample were denoted respectively by $\bar{\mathbf{X}}_{RPB}$ and \mathbf{S}_{RPB} . In the step 4, the test statistic is computed by

$$T_{RPB}^2 = n(\bar{\mathbf{X}}_{RPB} - \boldsymbol{\mu}_0)^\top \mathbf{S}_{RPB}^{-1}(\bar{\mathbf{X}}_{RPB} - \boldsymbol{\mu}_0). \quad (3.7)$$

The steps 5 and 6 are identically as described in the previous test.

3.3 The exact binomial test

The test type I error rates were evaluated by Monte Carlo simulations. According to Oliveira e Ferreira (2010) these estimates are not error-free. Therefore, an exact binomial test is used to make the decision whether each of the modified or the original T^2 tests is considered exact, liberal or conservative. In this sense, considering a nominal level of significance of 1%, the following null hypothesis is tested

$$H_0 : \alpha = 5\% \quad \text{versus} \quad H_1 : \alpha \neq 5\%. \quad (3.8)$$

The statistic of the exact binomial test is given by

$$F = \left(\frac{z+1}{N-z} \right) \left(\frac{1-\alpha}{\alpha} \right), \quad (3.9)$$

where z is the number of rejection of the null hypothesis by one of the tests in the nominal significance level of α and N is the number of Monte Carlo simulations performed. If the null hypothesis is rejected and the type I error is considered significantly ($P \leq 0.01$) lower than the nominal level of 5%, the test can be considered conservative; if the null hypothesis is rejected and the type I error is considered significantly ($P \leq 0.01$) higher than the nominal level of 5%, the test can be considered liberal; and if the null hypothesis is not rejected, ($P > 0.01$), the test can be considered exact. In the next section we present our main results.

4 RESULTS AND DISCUSSION

Our results are presented and discussed in two steps. First the type I error rates are presented and second, the power of the tests. In order to evaluate the performance of the proposed new tests (see section 2), the results were compared with those of the classical T^2 test. The test performance was evaluated under multivariate normal, Student t with 5 and 30 degrees of freedom and contaminated normal distributions.

4.1 Type I error rates

The type I error rates of the three studied tests in Monte Carlo simulations were presented in Table 4.1 for $p = 2$ and $\alpha = 0.05$. The exact binomial test was applied to classify the tests in exact, liberal, or conservative. We noted that most of the tests were exact (size equal the nominal level of significance α) in all scenarios. Also an isolated results for T^2 with $n = 50$ under multivariate normality indicated it as a conservative test. The expected results for a test is that its test size be equal of the nominal significance level α , what is equivalent to say it is an exact test. However, when it is considered conservative, its use is acceptable. A important fact to be noticed is that the traditional T^2 test controls adequately the type I error rates when normality is violated showing it is a robust test. The same results was observed for all parametric bootstrap tests.

In order to observe if there is any different type I error rates control pattern among the tests we increased the number of variables and maintained the same distributions and sample sizes of Table 4.1. The results are shown in Table 4.2 for the type I error rates with $p = 5$. We noticed that the scenario did not change. The T^2 test maintained the same behavior even when multivariate normality assumptions were violated. The T^2PB showed two conservative type I error results for the multivariate Student t -distribution. None of the tests presented liberal results for type I error rates.

Some of the significantly results for the null hypothesis of exact test observed in the previous cases were attributed to the type I error rates, due to the binomial test multiplicity. Therefore the other property of a test, its power, should be considered to choose the best one. We hope that at least one of our proposed tests performs better than the traditional T^2 . Thus, in the next section the Monte Carlo results for the power of these tests were shown.

Tabela 4.1 – Type I error rates of the three tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate Normal (N), Student t with 5 degrees of freedom (t_5), Student t with 30 degrees of freedom (t_{30}) and contaminated normal (CN) distributions.

Test	Model	Sample sizes n				
		10	50	70	100	200
T2	N	0.0525	0.0380 ⁻	0.0505	0.0485	0.0460
	t_5	0.0450	0.0495	0.0420	0.0485	0.0510
	t_{30}	0.0515	0.0515	0.0530	0.0445	0.0495
	CN	0.0520	0.0450	0.0420	0.0490	0.0570
T2PB	N	0.0510	0.0475	0.0455	0.0485	0.0475
	t_5	0.0510	0.0430	0.0405	0.0485	0.0560
	t_{30}	0.0500	0.0530	0.0540	0.0430	0.0515
	CN	0.0535	0.0435	0.0440	0.0430	0.0515
T2RPB	N	0.0430	0.0480	0.0525	0.0450	0.0535
	t_5	0.0430	0.0390	0.0520	0.0500	0.0485
	t_{30}	0.0555	0.0475	0.0525	0.0450	0.0455
	CN	0.0530	0.0455	0.0525	0.0490	0.0615

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5% (section 3.3).

Tabela 4.2 – Type I error rates of the three tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate Normal (N), Student t with 5 degrees of freedom (t_5), Student t with 30 degrees of freedom (t_{30}) and contaminated normal (CN) distributions.

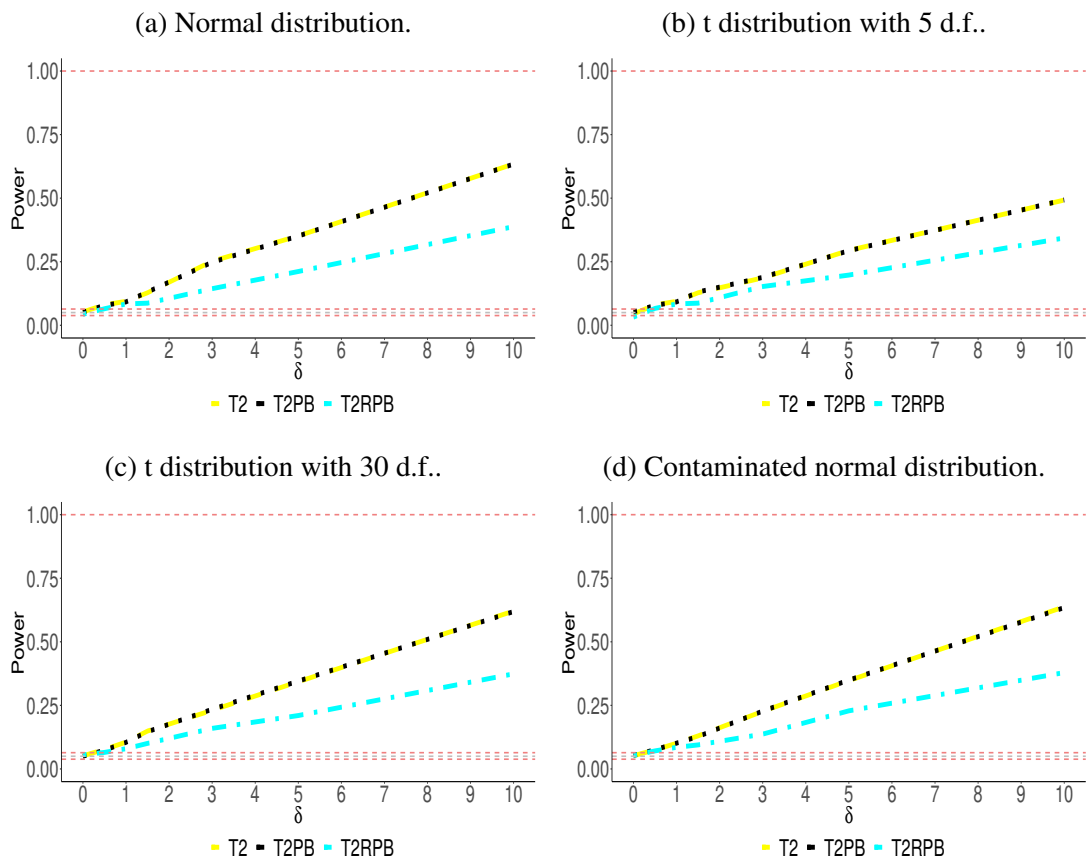
Test	Model	Sample sizes n				
		10	50	70	100	200
T2	N	0.0525	0.0545	0.0430	0.0500	0.0565
	t_5	0.0420	0.0545	0.0435	0.0440	0.0445
	t_{30}	0.0550	0.0430	0.0550	0.0530	0.0470
	CN	0.0475	0.0535	0.0465	0.0475	0.0520
T2PB	N	0.0520	0.0545	0.0430	0.0500	0.0560
	t_5	0.0545	0.0505	0.0440	0.0440	0.0465
	t_{30}	0.0545	0.0410	0.0555	0.0525	0.0480
	CN	0.0475	0.0540	0.0555	0.0470	0.0515
T2RPB	N	0.0505	0.0595	0.0465	0.0460	0.0605
	t_5	0.0315 ⁻	0.0385	0.0430	0.0405	0.0365 ⁻
	t_{30}	0.0450	0.0510	0.0530	0.0460	0.0605
	CN	0.0430	0.0525	0.0460	0.0475	0.0500

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5% (see section 3.3).

4.2 Power

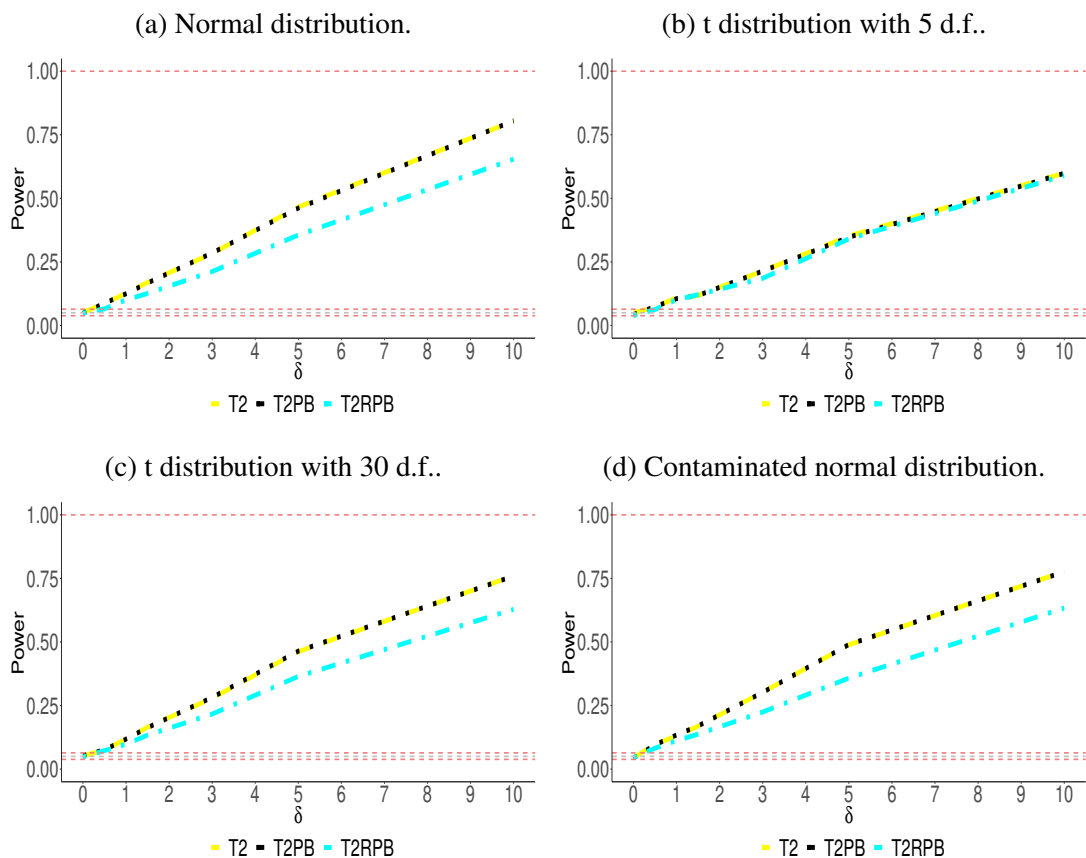
The performance in power of the new tests was compared with T^2 power in several circumstances related with several different distributions, number of variables p and sample sizes. In all cases the power curves were plotted in function of the Mahalanobis distance between the true population vector mean $\boldsymbol{\mu}$ and the hypothetical vector mean $\boldsymbol{\mu}_0$. In the previous section we saw that almost all of the proposed tests are exact. The $T2$ and $T2PB$ shows better power performance over all other tests, as can be seen in Figure 4.1. Even under normality and without any assumptions violation, the $T2PB$ performed equal to the traditional T^2 test (Figure 4.1(a)). In the case of Figure 4.1(d) a multivariate contaminated normal distribution with 10% outliers was considered. In this case, the power of the $T2PB$ is the same of the $T2$ test and for a fixed value of δ , the power values are almost equally of the multivariate normal case and higher than of those in the others distributions. The performance of the robust bootstrap test is always lower, with lower power than the $T2$ and $T2PB$ tests power.

Figura 4.1 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



To verify if the patterns observed in the previous circumstances shown in Figure 4.1 will use the same scenarios with larger sample sizes where the same number of variables were considered. The sample sizes of $n = 50, 70, 100,$ and 200 were simulated. In the circumstances of Figure 4.2, Figure 4.3, Figure 4.4, and Figure 4.5 the previous behavior pattern of the tests changed only for the multivariate Student t -distribution with 5 degrees of freedom (Figure 4.2(b), 4.3(b), 4.4(b) and 4.5(b)). Since this distribution has a heavy tail and all the tests studied are based on the multivariate normality assumption, the $T2RPB$ achieved at least the same performance of the $T2$ and $T2PB$ tests. In general, considering the numbers of observations n and the number of variables $p = 2$, the $T2$ and $T2PB$ are the most prominent tests. The next results refer to cases where sample sizes and distributions are the same, but the number of variables p is increased from 2 to 5.

Figura 4.2 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



The $T2$, $T2PB$ and $T2RPB$ tests showed the same power performance for sample size of $n = 10$ under normality and when the normality assumption was violated (Figure 4.6). The Figures 4.7, 4.8, 4.9 and 4.10 show the same power behavior for the proposed new tests and

Figura 4.3 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

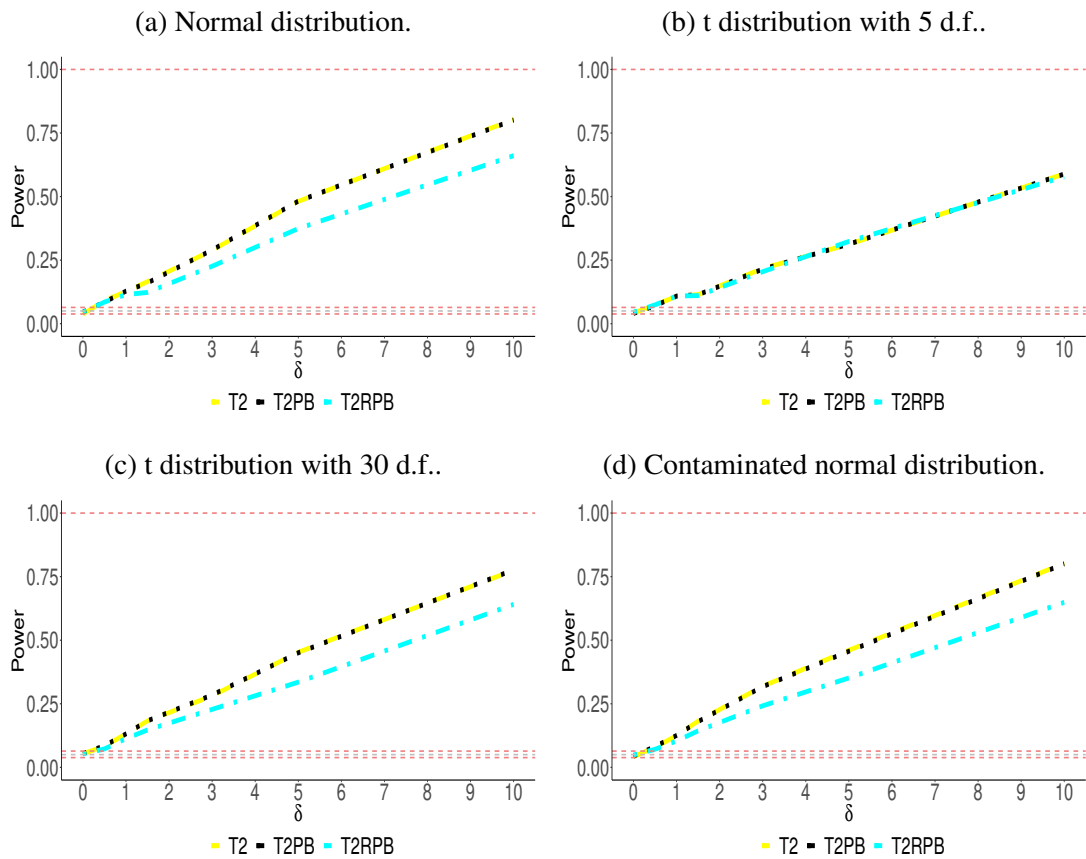
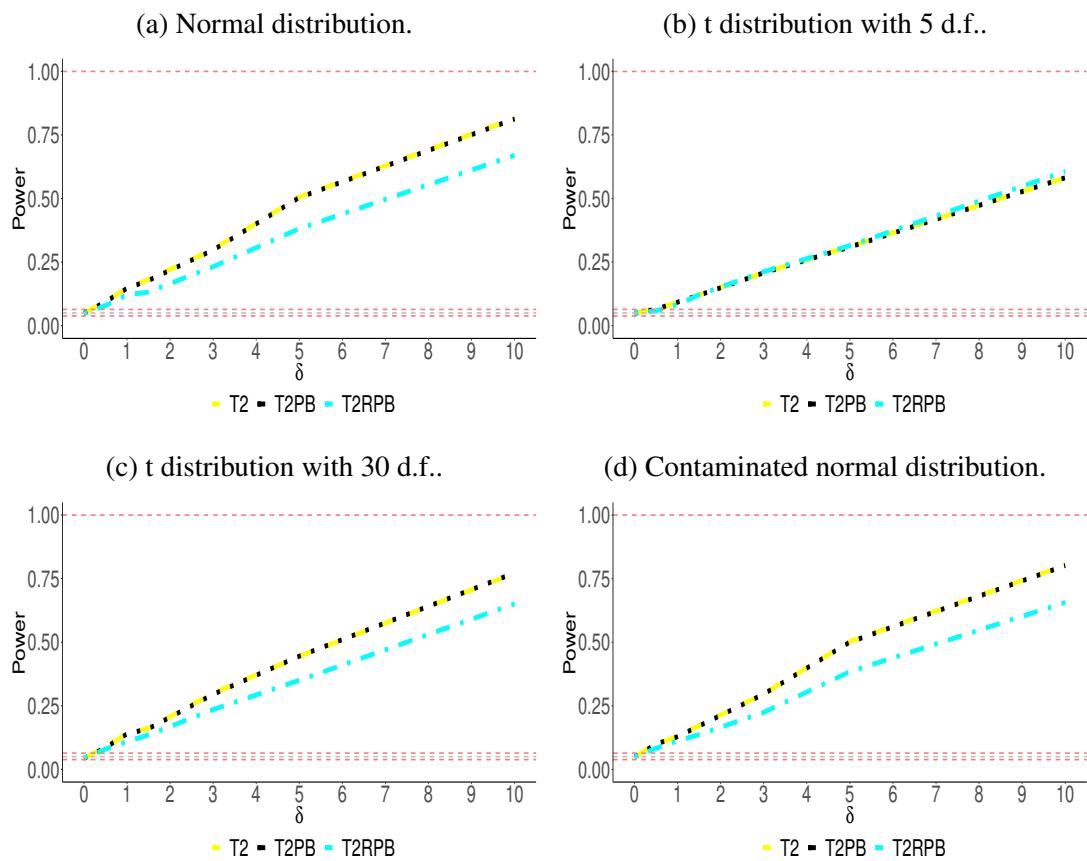
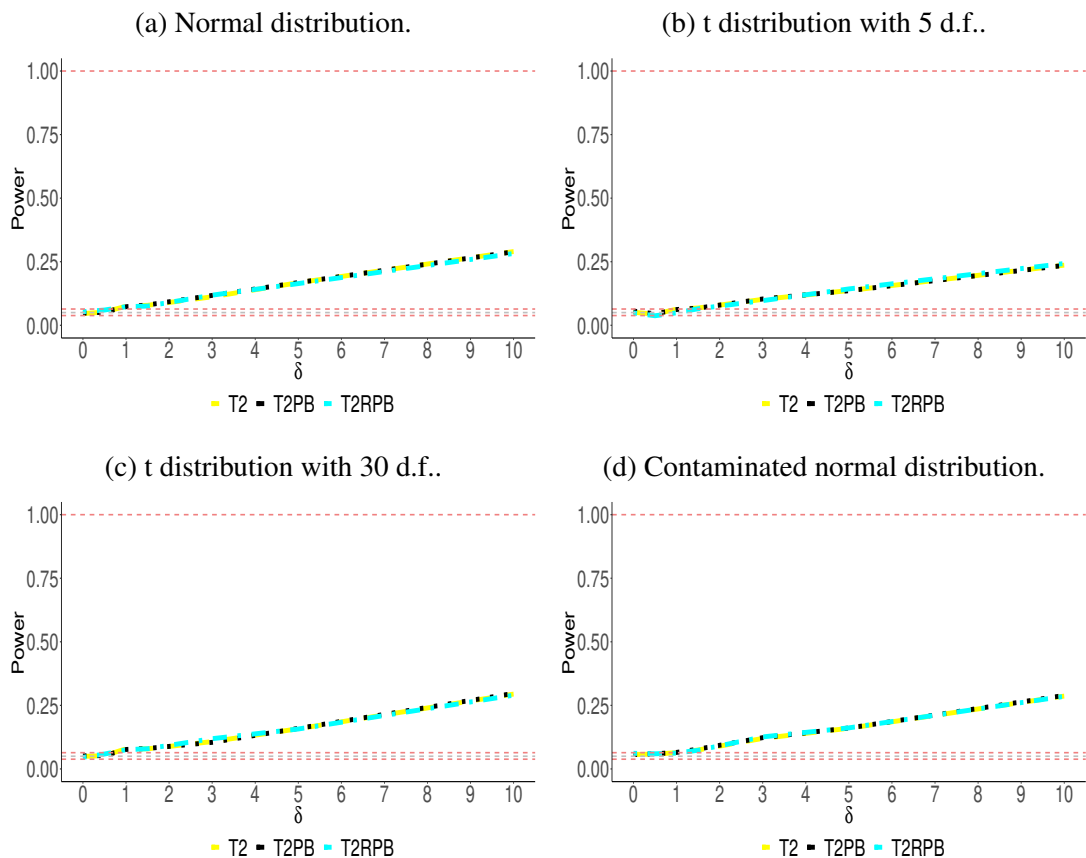


Figura 4.4 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



distributions considered. With increasing sample sizes n and keep the number of variable fixed at $p = 5$, the power performance of the $T2$ and $T2PB$ tests stands out. The same did not occur with the $T2RPB$ test. For the multivariate Student t -distribution with 5 degrees of freedom (Figures 4.6(b), 4.7(b), 4.8(b), 4.9(b), 4.10(b)), the power performance for all tests does not change. A curious fact is that for the multivariate contaminated normal distribution (Figures 4.6(d), 4.7(d), 4.8(d), 4.9(d), 4.10(d)), the multivariate Student t -distribution with 30 degrees of freedom (Figures 4.6(c), 4.7(c), 4.8(c), 4.9(c), 4.10(c)) and multivariate normal distribution (Figures 4.6(a), 4.7(a), 4.8(a), 4.9(a), 4.10(a)) the performance of $T2$, $T2PB$ are also very close. This fact was expected, except for normal distribution (Figures 4.6(a), 4.7(a), 4.8(a), 4.9(a), 4.10(a)) where the $T2$ test was expected to have better power performance, but this did not occur.

Figura 4.6 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



In general, the $T2PB$ is recommend for testing the null hypothesis in (1.1) under any conditions, where assumptions for applying the traditional $T2$ test were violated. If the results of the actual work were compared with those of Willems et al. (2002), Srivastava e Du (2008),

Figura 4.8 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

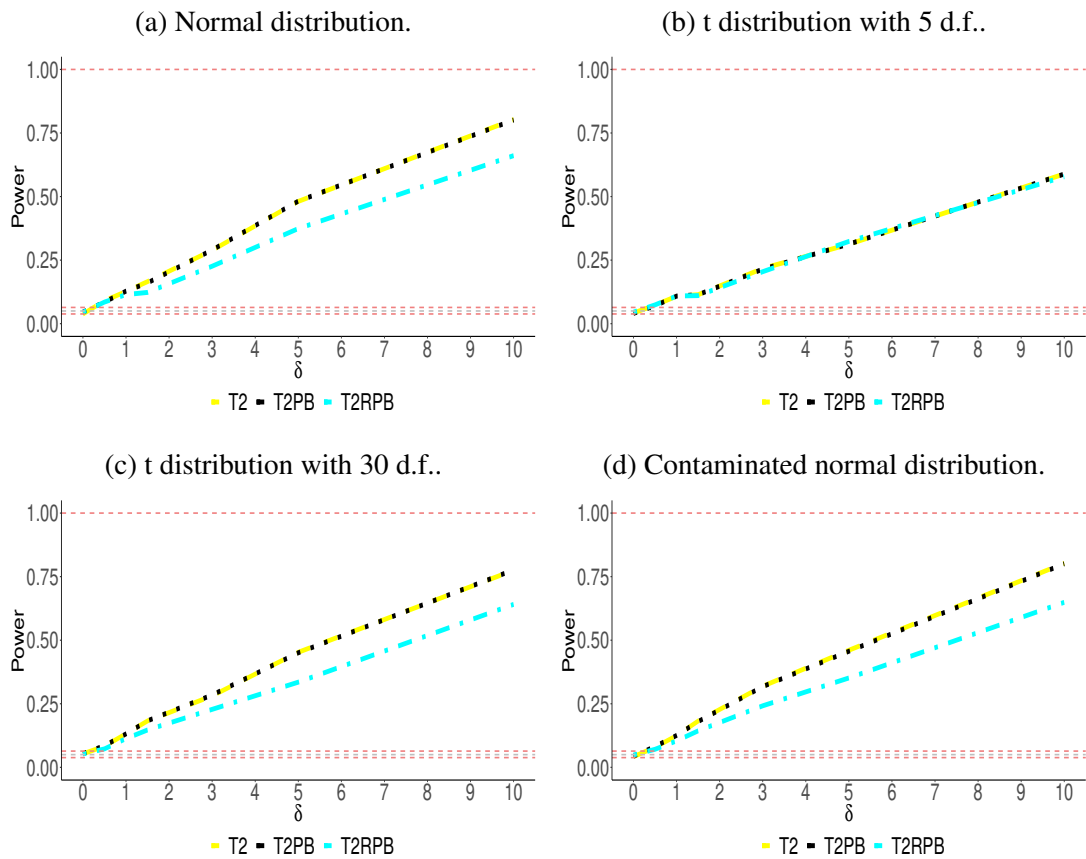
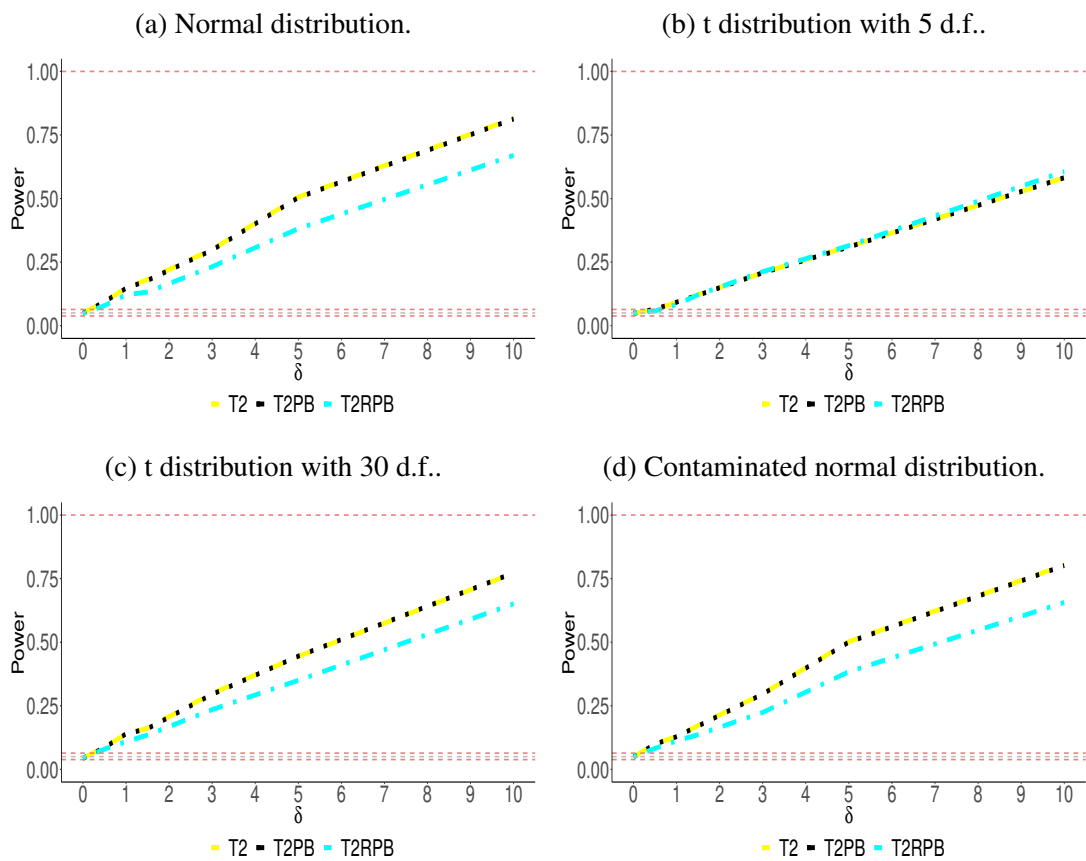


Figura 4.9 – Test power as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



Srivastava (2009) it will be seen that our proposal tests have performed better. Besides the better power performance, the $T2PB$ is simple, easy to implement and computationally fast.

5 CONCLUSION

In this paper we conclude that the $T2PB$ is recommended for testing hypothesis about a p -variate population mean vector. In addition, we also conclude that the use of the robust comedian covariance estimator was not effective in inferential problems in the presence of outliers and with multivariate normality assumption violations.

REFERÊNCIAS

- BAI, Z.; SARANADASA, H. Effect of high dimension: by an example of a two sample problem. **Statistica Sinica**, JSTOR, v. 6, p. 311–329, 1996.
- FALK, M. On mad and comedians. **Annals of the Institute of Statistical Mathematics**, Springer, v. 49, n. 4, p. 615–644, 1997.
- MARONNA, R. A.; ZAMAR, R. H. Robust estimates of location and dispersion for high-dimensional datasets. **Technometrics**, Taylor & Francis, v. 44, n. 4, p. 307–317, 2002.
- MUDHOLKAR, G. S.; SRIVASTAVA, D. K. A class of robust stepwise alternatives to hotelling's T^2 tests. **Journal of Applied Statistics**, Taylor & Francis, v. 27, n. 5, p. 599–619, 2000.
- OLIVEIRA, I. R. Cardoso de; FERREIRA, D. F. Multivariate extension of chi-squared univariate normality test. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 80, n. 5, p. 513–526, 2010.
- DI PALMA, M.A.; SRINIVASAN, M.R.; SAJESH, T.A. Detecting compositional outliers using comedian method. **Advances in Latent Variables-Methods, Models and Applications**, Citeseer, 2013.
- PAN, G. M.; ZHOU, W. Central limit theorem for hotelling's T^2 statistic under large dimension. **The Annals of Applied Probability**, JSTOR, p. 1860–1910, 2011.
- SAJESH, T. A.; SRINIVASAN, M. R. Outlier detection for high dimensional data using the comedian approach. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 82, n. 5, p. 745–757, 2012.
- SRIVASTAVA, M. S. A test for the mean vector with fewer observations than the dimension under non-normality. **Journal of Multivariate Analysis**, Elsevier, v. 100, n. 3, p. 518–532, 2009.
- SRIVASTAVA, M. S.; DU, M. A test for the mean vector with fewer observations than the dimension. **Journal of Multivariate Analysis**, Elsevier, p. 386–402, 2008.
- TIKU, M. L. Robust statistics for testing equality of means or variances. **Communications in Statistics-Theory and Methods**, Taylor and Francis, v. 11, n. 22, p. 2543–2558, 1982.

WILLEMS, G. et al. A robust hotelling test. **Metrika**, Springer, p. 125–138, 2002.

**ARTIGO 2 - PROPOSITION OF NEW ROBUST TESTS TO THE MULTIVARIATE
NORMAL MEAN VECTOR FOR HIGH-DIMENSIONAL DATA.**

**Submetido para avaliação ao journal Communications in Statistics - Simulation and Com-
putation**

ABSTRACT

New alternative tests to the Hotelling T^2 test and the likelihood ratio test (LRT) for the p -variate normal and non-normal population mean vector were proposed here. These new tests are based on the ordinary and robust comedian covariance matrix estimator. Also, the new adapted LRT overcomes the high dimensional issue that occurs with both the classical T^2 and the likelihood ratio tests. The asymptotic and parametric bootstrap distributions for the null distribution of the test statistics were used and the performance of these new tests on normal and non-normal distributions was evaluated through Monte Carlo simulations. Contaminated normal p -variate populations were also considered to evaluate the effects of outliers on test performances. Type I error rates and power in all Monte Carlo simulations were computed using the R system. The nonrobust parametric bootstrap version of the LRT performed better and is recommended since it is easy to implement and computationally fast.

Keywords: power. type I error. parametric bootstrap. LRT . T^2 . robust.

1 INTRODUCTION

A big challenge in statistics is based on verifying if a p -variate normal mean vector $\boldsymbol{\mu}$ is equal to a known vector $\boldsymbol{\mu}_0$ when the dimensionality p is greater than the sample size n . Consider the problem of testing the following null hypothesis:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{against} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0. \quad (1.1)$$

For this, let $\mathbf{X}_j = [X_{j1}, \dots, X_{jp}]^\top$ with $j = 1, 2, \dots, n$ be a random sample of size n from a p -variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Here, n refers to the number of observations, and p refers to the number of variables (dimensionality) in each random vector. In general, the p components in the random vectors are correlated variables, where its covariance matrix $\boldsymbol{\Sigma}$ ($p \times p$) are positively defined.

The Hotelling's T^2 test is widely used to test the null hypothesis in (1.1) and the test statistic is given by

$$T_c^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0). \quad (1.2)$$

where $\bar{\mathbf{X}} = \sum_{j=1}^n \mathbf{X}_j / n$ is the sample mean vector, $\mathbf{S} = \sum_{j=1}^n (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})^\top / (n - 1)$ is the sample covariance matrix, and n is the sample size. Under H_0 and with the assumption of normality and homoscedastic covariance matrix, the T_c^2 in (1.2) follows a Hotelling T^2 distribution given by $(n - 1)pF_{p, n-p} / (n - p)$, where $F_{p, n-p}$ is the F distribution with p and $n - p$ degrees of freedom.

However, under non-normal distributions or in the presence of outliers, the use of the Hotelling's T^2 test is not recommended. Firstly, the statistics in (1.2) was built under multivariate normality. Secondly, even under normality, since statistics in (1.2) considers the sample mean vector $\bar{\mathbf{X}}$ and covariance matrix \mathbf{S} that are strongly influenced by outliers, as in the univariate case (WILLEMS et al., 2002). Thirdly, T_c^2 in (1.2) cannot be calculated when $p \geq n$ since \mathbf{S} is a singular matrix (BAI; SARANADASA, 1996). Bai e Saranadasa (1996) have noticed that the power of the test based on the statistic in (1.2) has low power when $p < n$, but p/n is close to 1, where the sample covariance \mathbf{S} is quasi-singular matrix, as also shown in (PAN; ZHOU, 2011).

Considering the hypotheses given in (1.1), another statistical test used is the likelihood ratio test (LRT) (FERREIRA, 2018). Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is unknown. Therefore, the

likelihood ratio test statistic ($-2\log(\Lambda)$) is given by the following expression

$$-2\ln(\Lambda) = n[\ln|\mathbf{S} + \mathbf{H}| - \ln|\mathbf{S}|], \quad (1.3)$$

where $\mathbf{H} = (\bar{\mathbf{X}} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\top$ and \ln is the natural logarithm. Consider $\Omega \in \mathbb{R}^s$ the unrestricted parametric space and $\Omega_0 \subseteq \mathbb{R}^r$ the restricted parametric space, with $\Omega_0 \subset \Omega$. In general, under certain conditions of regularity, (1.3) follows an asymptotic chi-square distribution with $(s-r)$ degrees of freedom (FERREIRA, 2018) under the null hypothesis H_0 . So the rejection region of H_0 , R , as in (1.2), is given by

$$R = \{\mathbf{X} | \lambda = -2\log(\Lambda(\mathbf{X})) > \chi_{\alpha, s-r}^2\}, \quad (1.4)$$

where α is the nominal significance level and $\chi_{\alpha, s-r}^2$ is the $100\alpha\%$ upper quantile of a chi-square distribution with $s - r$ degrees of freedom. In this case, for testing hypothesis about normal mean vector, the degrees of freedom $s - r$ is equal to p .

Also, for testing the null hypothesis given in (1.1) becomes a challenge under non-normal asymmetric distributions or in the presence of outliers. The Hotelling's or the likelihood ratio tests showed in (1.2) and (1.3) consider at their expressions the $\bar{\mathbf{X}}$ and \mathbf{S} sample estimators that are highly influenced by outliers. Some robust testing proposals can be found in Tiku (1982), Mudholkar e Srivastava (2000), Willems et al. (2002). In contrast, (SRIVASTAVA; DU, 2008; SRIVASTAVA, 2009; CHEN; QIN et al., 2010; LEE et al., 2012; SRIVASTAVA; KATAYAMA; KANO, 2013; WANG; PENG; QI, 2013; MAROZZI; MARCO, 2015; MAROZZI; MARCO, 2016) proposed nonparametric alternative tests to the LRT in non-normal populations considering the cases where $p > n$.

In this paper new statistical tests for the null hypothesis given in (1.1) are proposed. The idea is to get robust adaptations of the T^2 (1.2) and the LRT (1.3) statistics using robust comedian estimators (SAJESH; SRINIVASAN, 2012) for the mean and covariance matrix. The fundamental conception is to replace the ordinary estimators $\bar{\mathbf{X}}$ and \mathbf{S} by their robust respectively comedian estimators and provide accurate tests for the mean vector considering the parametric bootstrap distribution under H_0 . Also, for the LRT statistics with the original and robust comedian estimator of the covariance matrix, the determinants (generalized variances) are replaced by the trace operator, which represents the total sample variance. These new tests are potentially more advantageous than the above-mentioned adaptation tests since they can per-

form better under non-normality and in the presence of outliers. Also, they are computationally easy to implement and apply.

The performance of these proposed new tests was evaluated by Monte Carlo simulations calculating the type I error rates and power of the tests. In the 2 section the proposed new tests were introduced. The results regarding the type I error rate and power of the tests are shown in section 3. The exact binomial test proposed by (OLIVEIRA; FERREIRA, 2010) is applied to deal with the randomness of the Monte Carlo method. In section 4 the conclusions are presented.

2 METHODS

In this section, the proposed new tests based on the modifications of the T^2 (1.2) and LRT (1.3) statistics are shown. Also, the adopted Monte Carlo simulation procedure to assess their performance was described. For this, consider a sample of size n from the normal p -variate distribution with a mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ to test the null hypothesis H_0 given in (1.1). In all cases, the original and transformed LRT given by (1.3) were applied, consider the trace operator replacing the determinant operator and the robust estimator replacing the traditional estimator of the covariance matrix. Only in the cases where $p < n$, the T^2 test and its modifications that use the comedian estimators. The theoretical justification can be seen in section 1. To evaluate test performance, first, type I error rates were calculated by generating sample sizes from populations under H_0 , with a mean vector of $\boldsymbol{\mu}_0$. Second, random samples were generated under H_1 , with $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. In both cases, samples from normal and non-normal populations were generated. The p -variate Student t -distributions with 5 degrees of freedom for the non-normal distribution case. We also consider contaminated normal populations for generating outliers. Some factorial combinations of the number of variables p and sample size n were considered.

Without loss of generality, the population covariance matrix $\boldsymbol{\Sigma}$ with the compound symmetry structure given by

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} = \sigma^2[(1 - \rho)\mathbf{I} + \rho\mathbf{J}] \quad (2.1)$$

was considered, where \mathbf{J} is a matrix ($p \times p$) with all entries equal to 1 and \mathbf{I} is an identity array of the same order as \mathbf{J} . Also, without loss of generality, $\sigma^2 = 1$ and $\rho = 0.9$ since the test statistics are invariant under the true covariance structure. The p -variate Student t -distribution was used to pick a distribution that has heavier tails than the multivariate normal distribution and violate the assumptions of the T^2 and LR (likelihood Ratio) test statistics. The p -variate contaminated normal distribution is given by: $wN_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) + (1 - w)N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$. Again, without loss of generality, $w = 0.9$, $\boldsymbol{\Sigma}_1$ was defined by (2.1) and $\boldsymbol{\Sigma}_2$ is constructed using the following constraint: $|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1| = \Delta$, and thus, $\boldsymbol{\Sigma}_2 = \Delta^{1/p}\boldsymbol{\Sigma}_1$, where $\Delta = 2$. Sample sizes were $n = 10, 50, 70, 100$ and 200 and the nominal significance level α was $\alpha = 5\%$. The number of variables

was $p = 2, 5$ and 200 and $2,000$ Monte Carlo simulations to evaluate the empirical estimates of the type I error rates and power of each test were considered. The parametric bootstrap null distribution were generated with $2,000$ resamples from a $N(\mathbf{0}, \mathbf{S}^*)$ distribution, where the null hypothesis H_0 was imposed by considering $\boldsymbol{\mu} = \mathbf{0}$ to generate the null distribution of the statistic. Also, the covariance matrix used to generate the bootstrap null distribution, given \mathbf{S}^* , is the sample covariance matrix computed by using the traditional or robust comedian estimator in the original sample for traditional and robust bootstrap tests, respectively.

Without loss of generality, to evaluate the type I error rates the $\boldsymbol{\mu}_0$ vector under the null hypothesis given in (1.1) was the p -dimensional $\mathbf{0}$ null vector and the true vector of population mean was also $\boldsymbol{\mu} = \mathbf{0}$. Under the alternative hypothesis H_1 in (1.1), for the power study, the true population mean vector $\boldsymbol{\mu}$ was chosen considering a fixed generalized Mahalanobis distance $\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0)$ between $\boldsymbol{\mu}$ e $\boldsymbol{\mu}_0$ given by

$$\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0). \quad (2.2)$$

In this case, $\boldsymbol{\mu}_0 = \mathbf{0}$, and the true population mean vector was calculated by trial and error in (2.2) by taken a fixed δ value and the final value was used as a parameter in each of the population distributions considered under H_1 . The chosen values from δ were $0, 0.5, 1, 1.5, 3, 5, 10$. Therefore, since the values of the population mean vector change when the sample sizes change, keeping fixed the value of the distance of Mahalanobis $\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, it is expected the power does not show changes with increasing the sample size n .

Thus, five new tests based on the traditional T^2 (1.2) and LR (1.3) tests were proposed, including the traditional and robust versions that uses the comedian mean vector and covariance matrix estimators (FALK, 1997; MARONNA; ZAMAR, 2002; SAJESH; SRINIVASAN, 2012). Some tests were based on parametric bootstrap versions as well as the asymptotic chi-square distribution. But some asymptotic chi-square tests have not been shown since they did not control the type I error rates. The performance of these new tests was evaluated by Monte Carlo simulations. Below each of the proposed new tests for testing the null hypothesis $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ was described. It was considered that the new LRT has a chi-square asymptotic distribution with p degrees of freedom, as the original LRT.

Two special cases were considered regarding the distributions and for some values of n , p and δ . The p -variate t_1 distribution (multivariate Cauchy) that a heavy tail distribution and a shifted zero mean exponential distribution with parameter $\boldsymbol{\lambda} = \mathbf{1}_p$, where $\mathbf{1}_p$ is a p -dimensional

vector with 1 in all entries. The latter is a case of a skewed distribution. For the exponential distribution, a p -dimension random vector \mathbf{Z} is generated from a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. A p -dimensional random vector \mathbf{Y} from this distribution is obtained considering for the i th entry the following random variable

$$Y_i = F^{-1}(\Phi(Z_i); \lambda_i) - \frac{1}{\lambda_i}, \quad i = 1, 2, \dots, p,$$

where $F^{-1}(x; \lambda_i)$ is the quantile function of the exponential λ_i distribution evaluated in x and $\Phi(x)$ is the cumulative distribution function of the standard normal distribution evaluated in x .

2.1 The parametric bootstrap T^2 test

We construct the parametric bootstrap T^2 test, called T_{PB}^2 ($T2PB$), where PB stands for parametric bootstrap, adopting the following steps:

1. from the original sample, the parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ were estimated, respectively, by \mathbf{S}^* and $\bar{\mathbf{X}}^*$, where \mathbf{S}^* and $\bar{\mathbf{X}}^*$ are the traditional sample covariance matrix and vector mean, respectively. The test statistic is computed by

$$T^{*2} = n(\bar{\mathbf{X}}^* - \boldsymbol{\mu}_0)^\top \mathbf{S}^{*-1}(\bar{\mathbf{X}}^* - \boldsymbol{\mu}_0). \quad (2.3)$$

2. using the original covariance estimates \mathbf{S}^* , a random sample of size n is generated from a p -variate normal distribution imposing H_0 , that is, by setting $\boldsymbol{\mu} = \boldsymbol{\mu}_0$. Also $\boldsymbol{\Sigma} = \mathbf{S}^*$. Therefore a sample of size n is generated from a $N(\boldsymbol{\mu}_0, \mathbf{S}^*)$ distribution;
3. in each parametric bootstrap sample the sample mean $\bar{\mathbf{X}}_{PB}$ and the sample covariance matrix \mathbf{S}_{PB} were estimated;
4. compute the test statistic by

$$T_{PB}^2 = n(\bar{\mathbf{X}}_{PB} - \boldsymbol{\mu}_0)^\top \mathbf{S}_{PB}^{-1}(\bar{\mathbf{X}}_{PB} - \boldsymbol{\mu}_0); \quad (2.4)$$

5. steps (2) to (4) are repeated B times and a set of size $B + 1$ was constructed with the B test statistic values computed in (2.4) and the original value computed in (2.3). The null

distribution of the parametric bootstrap test was constituted by this set. Therefore, if the i th member of this set was represented by T_i^2 , $i = 1, 2, \dots, B + 1$, then the p -value was computed by

$$p\text{-value} = \frac{\sum_{i=1}^{B+1} I(T_i^2 \geq T^{*2})}{B + 1}, \quad (2.5)$$

where $I(T_i^2 \geq T^{*2})$ is the indicator function.

6. The null hypothesis in (1.1) should be rejected at the significance level α if the p -value $\leq \alpha$.

It should be noted here that the traditional T^2 test was also considered, by computing the p -value direct from the Hotelling's T^2 distribution of the test statistic value obtained in (2.3). This was named by $T2$ and is considered the benchmark test.

2.2 The robust parametric bootstrap T^2 test

The robust parametric bootstrap T^2 test, called T_{RPB}^2 ($T2RPB$), in which RPB stands for robust parametric bootstrap, are performed by adopting the same steps described for the previous test. However, some of them are modified as in the following sequence. In step 1, the estimators \mathbf{S}^* and $\bar{\mathbf{X}}^*$ are replaced by comedian estimators \mathbf{S}_R and $\bar{\mathbf{X}}_R$ the test statistic in the original sample is computed by

$$T^{*2} = n(\bar{\mathbf{X}}_R - \boldsymbol{\mu}_0)^\top \mathbf{S}_R^{-1} (\bar{\mathbf{X}}_R - \boldsymbol{\mu}_0). \quad (2.6)$$

In step 2, the sample of size n is generated from a $N(\boldsymbol{\mu}_0, \mathbf{S}_R)$ distribution, where again, the null hypothesis is imposed by considering the multivariate normal mean equal to $\boldsymbol{\mu}_0$, the null value of the population mean. In step 3, the mean and the sample covariance in each parametric bootstrap sample were denoted, respectively, by $\bar{\mathbf{X}}_{RPB}$ and \mathbf{S}_{RPB} . In step 4, the test statistic is computed by

$$T_{RPB}^2 = n(\bar{\mathbf{X}}_{RPB} - \boldsymbol{\mu}_0)^\top \mathbf{S}_{RPB}^{-1} (\bar{\mathbf{X}}_{RPB} - \boldsymbol{\mu}_0). \quad (2.7)$$

The steps 5 and 6 are identically as described in the previous test, with T_i^2 replaced now by the i th value from the bootstrap null distribution of T_{RPB}^2 . In this case, the asymptotic chi-square distribution with p degrees of freedom was not considered as an alternative test, since

the corresponding robust T^2 test did not control the type I error rate (results not shown). More details on the performance of the above tests can be seen at (ALVES; FERREIRA, 2019).

We have shown in section 1, in addition to the problems presented for their use in data following non-normal distributions, that the traditional T^2 test is not valid for high dimensional data ($p > n$) due to the singularity of the sample covariance matrix \mathbf{S} . The likelihood ratio test (LRT) has the same limitations of the T^2 to be implemented in high dimensional data sets and non-normal circumstances. Considering the (LEDOIT; WOLF, 2002) paper as reference, we propose an alternative test to the LRT that is based on replacing the determinants of the matrices \mathbf{S} and $\mathbf{S} + \mathbf{H}$ for their respective traces. Here, $\mathbf{H} = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\top$. In this way, we have obtained a new test that applies to high-dimensional ($p > n$) data sets and that maintains the same distributional properties of test LRT . The validity of the asymptotic null distribution of the new test with traditional sample estimators of the mean vector and covariance matrix was evaluated by Monte Carlo simulations. Also, the parametric bootstrap and robust parametric bootstrap versions for this latest proposal were built, as we have done for the T^2 test. In the following subsections, we present the new procedures.

2.3 The asymptotic LR trace test

The asymptotic trace version of the LR test, denominated $ATLRT$ (Asymptotic Trace Likelihood Ratio Test), was obtained by directly replacing the determinant given in the expression of the LRT by the trace operator tr . Let the null hypothesis be $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$, then the test statistic $ATLRT$ is given by

$$T_{ATLRT}^2 = n \{ \log[tr(\mathbf{S}^* + \mathbf{H})] - tr(\mathbf{S}^*) \}, \quad (2.8)$$

where $\mathbf{H} = (\bar{\mathbf{X}}^* - \boldsymbol{\mu}_0)(\bar{\mathbf{X}}^* - \boldsymbol{\mu}_0)^\top$, that under the null hypothesis H_0 and normality has a chi-square distribution with p degrees of freedom, since the $plim$ (probability limit) of \mathbf{H} is $\mathbf{0}$ as $n \rightarrow \infty$, still under the null H_0 . The null hypothesis should be rejected if the $T_{ATLRT}^2 \geq \chi_{\alpha;p}^2$, where $\chi_{\alpha;p}^2$ is the 100 α % upper quantile of the chi-square distribution with p degrees of freedom.

2.4 The *TLR* parametric bootstrap test

There is no guarantee the T_{ATLRT}^2 has an asymptotic chi-square distribution with p degrees of freedom under H_0 and multivariate normality. To overcome this issue we proposed the *TLR* parametric bootstrap test, called *TLRPBT* (Trace Likelihood Ratio Bootstrap Parametric Test). The steps to apply for this test are the same as described previously for T_{PB}^2 , except for some details explained as follows.

In step 1, the estimators \mathbf{S}^* and $\bar{\mathbf{X}}^*$ are computed in the original sample and the test statistic is computed by

$$T^{*2} = n \{ \log[tr(\mathbf{S}^* + \mathbf{H}^*)] - tr(\mathbf{S}^*) \}, \quad (2.9)$$

where $\mathbf{H}^* = (\bar{\mathbf{X}}^* - \boldsymbol{\mu}_0)(\bar{\mathbf{X}}^* - \boldsymbol{\mu}_0)^\top$

In step 2, the sample of size n is generated from a $N(\boldsymbol{\mu}_0, \mathbf{S}^*)$ distribution, where the null hypothesis is imposed by considering the multivariate normal mean equal to $\boldsymbol{\mu}_0$, the null value of the population mean. In step 3, the mean and the sample covariance in each parametric bootstrap sample were denoted respectively by $\bar{\mathbf{X}}_{TLRPB}$ and \mathbf{S}_{TLRPB} . In step 4, the test statistic is computed by

$$T_{TLRPB}^2 = n \{ \log[tr(\mathbf{S}_{TLRPB} + \mathbf{H}_{TLRPB})] - tr(\mathbf{S}_{TLRPB}) \}, \quad (2.10)$$

where $\mathbf{H}_{TLRPB} = (\bar{\mathbf{X}}_{TLRPB} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}}_{TLRPB} - \boldsymbol{\mu}_0)^\top$. The steps 5 and 6 are identically as described in the previous test, with T_i^2 replaced now by the i th value from the bootstrap null distribution of T_{TLRPB}^2 .

2.5 The robust *TLR* parametric bootstrap test

For overcoming problems with outliers the robust parametric bootstrap version of the previous *TLRPB*, called *RTLRPBT* (Robust Trace Likelihood Ratio Bootstrap Parametric Test), was constructed. The steps necessary for this test to be applied are the same as the previous steps described for the T_{PB}^2 , except for some details explained below.

In step 1, the comedian estimators \mathbf{S}_R and $\bar{\mathbf{X}}_R$ are computed in the original sample and the test statistic is given by

$$T^{*2} = n \{ \log[tr(\mathbf{S}_R + \mathbf{H}_R)] - tr(\mathbf{S}_R) \}, \quad (2.11)$$

where $\mathbf{H}_R = (\bar{\mathbf{X}}_R - \boldsymbol{\mu}_0)(\bar{\mathbf{X}}_R - \boldsymbol{\mu}_0)^\top$

In step 2, the sample of size n is generated from the $N(\boldsymbol{\mu}_0, \mathbf{S}_R)$ null distribution. In step 3, the comedian sample mean and sample covariance in each parametric bootstrap sample were denoted respectively by $\bar{\mathbf{X}}_{RTL RP B}$ and $\mathbf{S}_{RTL RP B}$. In step 4, the test statistic is computed by

$$T_{RTL RP B}^2 = n \{ \log[tr(\mathbf{S}_{RTL RP B} + \mathbf{H}_{RTL RP B})] - tr(\mathbf{S}_{RTL RP B}) \}, \quad (2.12)$$

where $\mathbf{H}_{RTL RP B} = (\bar{\mathbf{X}}_{RTL RP B} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}}_{RTL RP B} - \boldsymbol{\mu}_0)^\top$. The steps 5 and 6 are identically as described in the previous test, with T_i^2 replaced now by the i th value from the bootstrap null distribution of $T_{RTL RP B}^2$.

2.6 The exact binomial test

The test type I error rates were evaluated by Monte Carlo simulations and according to (OLIVEIRA; FERREIRA, 2010) these estimates are not error-free. Therefore, an exact binomial test is used to decide whether each of the modified or the original statistical test is considered accurate, liberal or conservative. In this sense, considering a nominal level of significance of 1%, the following null hypothesis is tested

$$H_0 : \alpha = 5\% \quad \text{against} \quad H_1 : \alpha \neq 5\%. \quad (2.13)$$

The statistic of the exact binomial test is given by

$$F_c = \left(\frac{z+1}{N-z} \right) \left(\frac{1-\alpha}{\alpha} \right), \quad (2.14)$$

where z is the number of rejection of the null hypothesis accounted by one of the tests considering the nominal significance level of α and N is the number of Monte Carlo simulations performed. Under the null hypothesis (2.13) the F_c statistic follows a F distribution with $v_1 = 2(N-z)$ and $v_2 = 2(z+1)$ degrees of freedom. If the null hypothesis is rejected and the type

I error is considered significantly lower than the nominal level adopted of the 5%, the test can be considered conservative; if the null hypothesis is rejected and the type I error is considered significantly higher than the nominal level adopted of the 5%, the test can be considered liberal; and if the null hypothesis is not rejected the test can be considered accurate. In the next section, we present our main results.

3 RESULTS AND DISCUSSION

The performance results for the proposed new tests are presented in two stages. First, the results regarding type I error rate control and power for cases where $p = 2$ and $p = 5$ were shown. Second, the results for the special case of high dimension ($p = 200$) were shown. The performance of these new tests was evaluated considering the multivariate normal, the multivariate Student t with 5 degrees of freedom and contaminated normal distributions.

3.1 Type I error rates

The type I error rates for the five new tests proposed via Monte Carlo simulations are shown in Table 4.1, considering the dimension $p = 2$ at the significance level of $\alpha = 0.05$. The exact binomial test was used to classify these tests as exact, liberal, or conservative (see section 2). The traditional and ordinary Hotelling's T^2 test was also applied in each circumstance and it was invoked as a benchmark test. We note that for all n sample sizes considered, as well as for all evaluated multivariate distributions, the proposed tests $T2$, $T2PB$, and $T2RPB$ are exact since they showed test size equal to the nominal significance level α . The traditional T^2 test constituted an exception when considering the bivariate Normal distribution with a sample size of 50. In this case, this test is conservative but still acceptable. The same is not the case for LRT adaptations. The $ATLRT$ is conservative on all distributions and sample sizes considered. For this test, a substantial loss of power is expected to occur, which is an important fact to be taken into account. The $TLRPBT$ and $RTLRPBT$ (see section 2), in all the evaluated scenarios, are all accurate. A single exception occurs for the $TLRPBT$, considering the multivariate normal distribution with $n = 50$, where it shows a liberal performance in the control of the type I error rate. In practice, a test is considered reliable if it has an exact size. Otherwise, if it is conservative, it can be considered acceptable. However, if it is a liberal test, then it must be discarded. It does not appear the case for the $TLRPBT$, once it showed a unique exception. Therefore, since in circumstances where the normality assumption was violated, the performance regarding the type I error rate control of the proposed new tests is acceptable.

To evaluate if the patterns presented in Table 4.1 are maintained we decided to increase the dimensionality p and keep the sample sizes n fixed. Table 4.2 presents the results for the empirical type I error rates considering the dimension $p = 5$. Similar behavioral pattern to the control of the type I error rates showed for the case with $p = 2$ holds. The atypical situation (conservative) for the T^2 test no longer occurs. We note that the $T2RPB$ for the multivariate

Tabela 4.1 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate normal (N), multivariate Student t with 5 degrees of freedom (t_5) degrees of freedom and contaminated normal (CN) distributions.

Test	Model	Sample sizes n				
		10	50	70	100	200
T2	N	0.0525	0.0380 ⁻	0.0505	0.0485	0.0460
	t_5	0.0450	0.0495	0.0420	0.0485	0.0510
	CN	0.0520	0.0450	0.0420	0.0490	0.0570
T2PB	N	0.0510	0.0475	0.0455	0.0485	0.0475
	t_5	0.0510	0.0430	0.0405	0.0485	0.0560
	CN	0.0535	0.0435	0.0440	0.0430	0.0515
T2RPB	N	0.0430	0.0480	0.0525	0.0450	0.0535
	t_5	0.0430	0.0390	0.0520	0.0500	0.0485
	CN	0.0530	0.0455	0.0525	0.0490	0.0615
ATLRT	N	0.0195 ⁻	0.0085 ⁻	0.0120 ⁻	0.0110 ⁻	0.0130 ⁻
	t_5	0.0275 ⁻	0.0085 ⁻	0.0105 ⁻	0.0140 ⁻	0.0145 ⁻
	CN	0.0200 ⁻	0.0125 ⁻	0.0120 ⁻	0.0140 ⁻	0.0120 ⁻
TLRPBT	N	0.0535	0.0940 ⁺	0.0555	0.0445	0.0535
	t_5	0.0500	0.0680	0.0460	0.0520	0.0580
	CN	0.0515	0.0520	0.0485	0.0445	0.0470
RTLPT	N	0.0560	0.0490	0.0445	0.0430	0.0530
	t_5	0.0415	0.0415	0.0555	0.0510	0.0555
	CN	0.0550	0.0505	0.0480	0.0435	0.0570

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5%.

⁺: significantly ($P < 0.01$) greater than the nominal significance level of 5%.

Student t -distribution with 5 degrees of freedom and $n = 10$ and 200 was conservative in this case. The $TLRPBT$ and $RTLTPBT$ are considered exact tests. The $ATLRT$ remains conservative and therefore acceptable. Likewise, it expects that this test substantially loses power for this dimension.

Tabela 4.2 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate Normal (N), multivariate Student t with 5 degrees of freedom (t_5) and contaminated normal (CN) distributions.

Test	Model	Sample sizes n				
		10	50	70	100	200
T2	N	0.0525	0.0545	0.0430	0.0500	0.0565
	t_5	0.0420	0.0545	0.0435	0.0440	0.0445
	CN	0.0475	0.0535	0.0465	0.0475	0.0520
T2PB	N	0.0520	0.0545	0.0430	0.0500	0.0560
	t_5	0.0545	0.0505	0.0440	0.0440	0.0465
	CN	0.0475	0.0540	0.0555	0.0470	0.0515
T2RPB	N	0.0505	0.0595	0.0465	0.0460	0.0605
	t_5	0.0315 ⁻	0.0385	0.0430	0.0405	0.0365 ⁻
	CN	0.0430	0.0525	0.0460	0.0475	0.0500
ATLRT	N	0.0050 ⁻	0.0000 ⁻	0.0005 ⁻	0.0005 ⁻	0.0005 ⁻
	t_5	0.0000 ⁻	0.0005 ⁻	0.0000 ⁻	0.0005 ⁻	0.0010 ⁻
	CN	0.0005 ⁻	0.0005 ⁻	0.0005 ⁻	0.0000 ⁻	0.0105 ⁻
TLRPBT	N	0.0555	0.0550	0.0490	0.0585	0.0570
	t_5	0.0445	0.0455	0.0610	0.0495	0.0530
	CN	0.0510	0.0575	0.0465	0.0545	0.0525
RTLTPBT	N	0.0555	0.0550	0.0490	0.0585	0.0570
	t_5	0.0490	0.0460	0.0535	0.0455	0.0405
	CN	0.0495	0.0550	0.0425	0.0450	0.0520

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5%.

The next section presents the results of the power of the tests for these same cases.

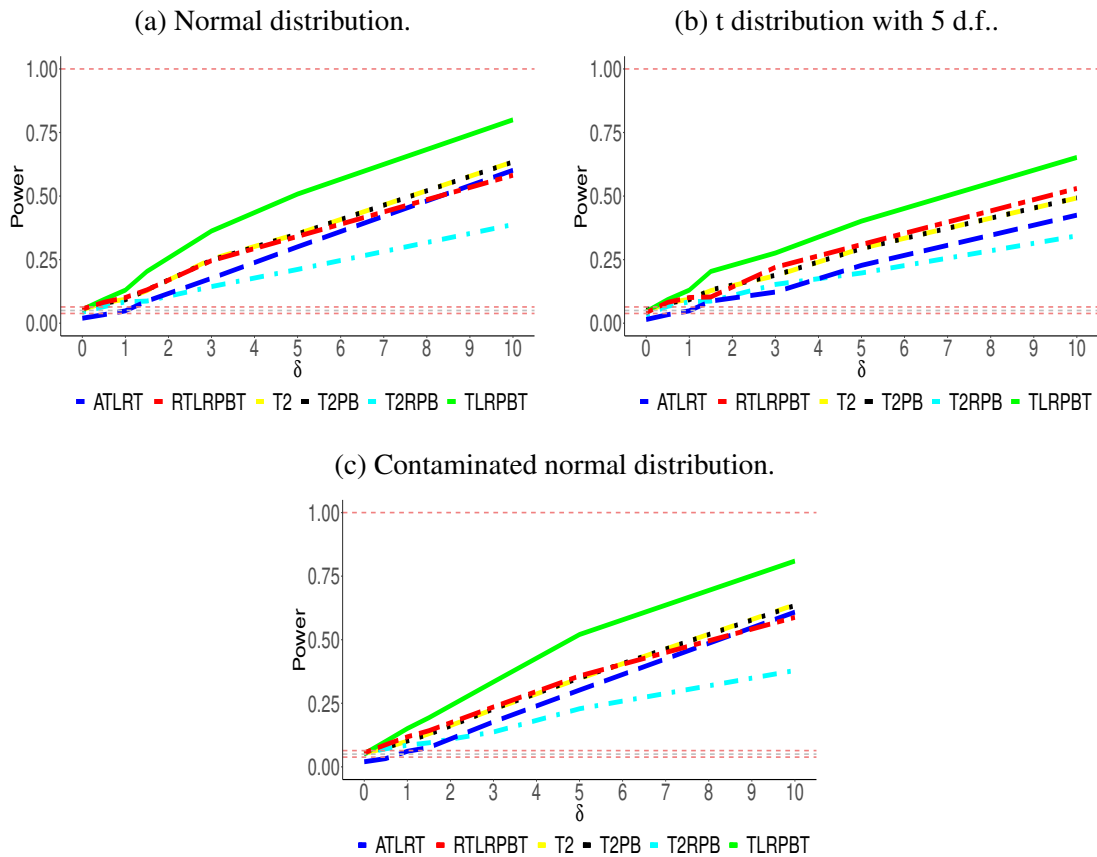
3.2 Power

The powerful performance of the proposed new tests was evaluated in the same cases used to evaluate type I error rates: distributions, sample sizes, and dimensions. In all circumstances, the power curves were plotted against Mahalanobis distances (δ), given in 2.5, between the true population vector $\boldsymbol{\mu}$ and the hypothetical mean vector $\boldsymbol{\mu}_0$. These distances have been fixed (see section 2) to establish the true value of the population mean vector.

It was noticed in Figure 4.11 that the $TLRPBT$ showed the best performance among all evaluated tests with $p = 2$. Under non-normality or in the presence of outliers, the performance

of this test showed a loss of power only when the multivariate Student t -distribution with 5 degrees of freedom was considered (Figure 4.11 (b)). We also noticed that the asymptotic version $ATLRT$ has lower power as expected (see Tables (4.1, 4.2)).

Figura 4.11 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



In order to verify if the patterns observed in Figure 4.11 remain the dimensionality was fixed in $p = 2$ and the multivariate distributions were the same, but the number of observations varied in $n = 50, 70, 100$ and 200 for generating the power curves by Monte Carlo simulations. Under these circumstances, the behavioral power patterns shown in Figure 4.11 remain the same (Figures 4.12, 4.13, 4.14 and 4.15). The only exceptions are for the multivariate Student t -distribution with 5 degrees of freedom (Figures 4.12 (b), 4.13 (b), 4.14 (b) and 4.15(B)). In this case, the performance of the $TLRPBT$ and $RTLRPBT$ is equivalent and higher than the performance of the other tests.

In general, for this dimension ($p = 2$), the parametric bootstrap version ($TLRPBT$) performed better when compared to the other tests. Also, all tests show a very robust behavioral

Figura 4.12 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

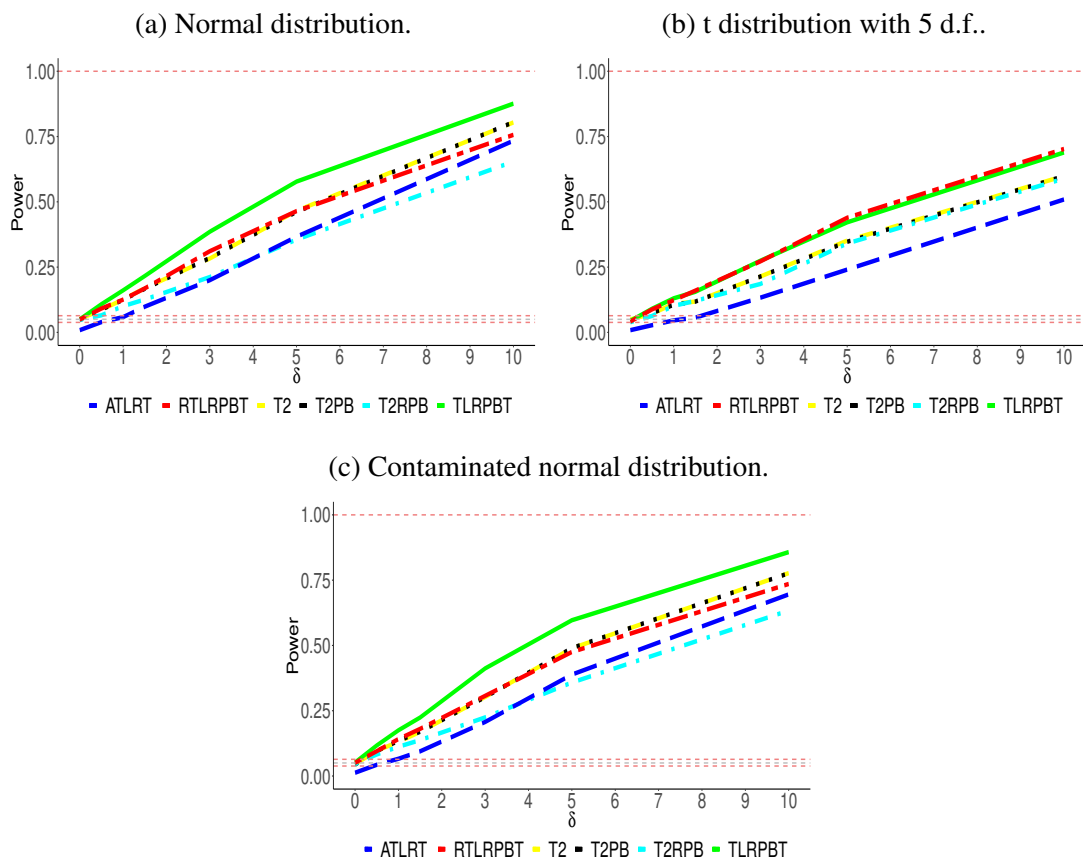


Figura 4.13 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

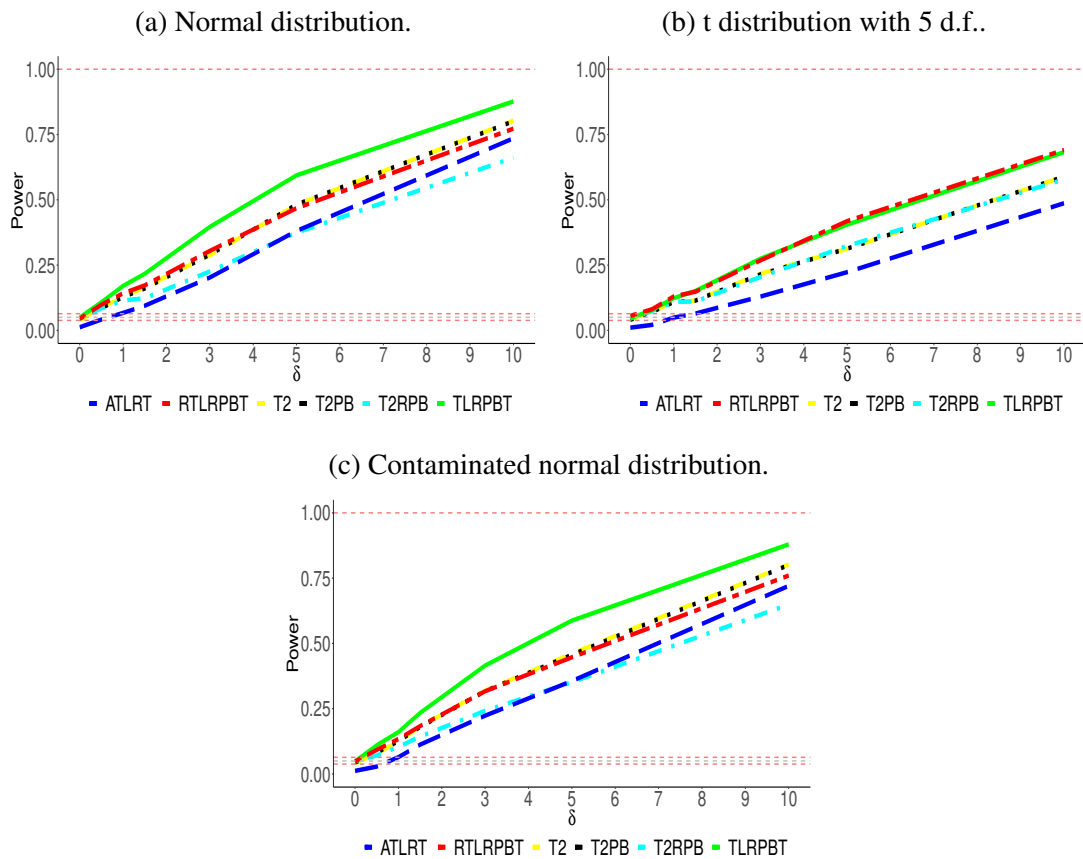


Figura 4.14 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

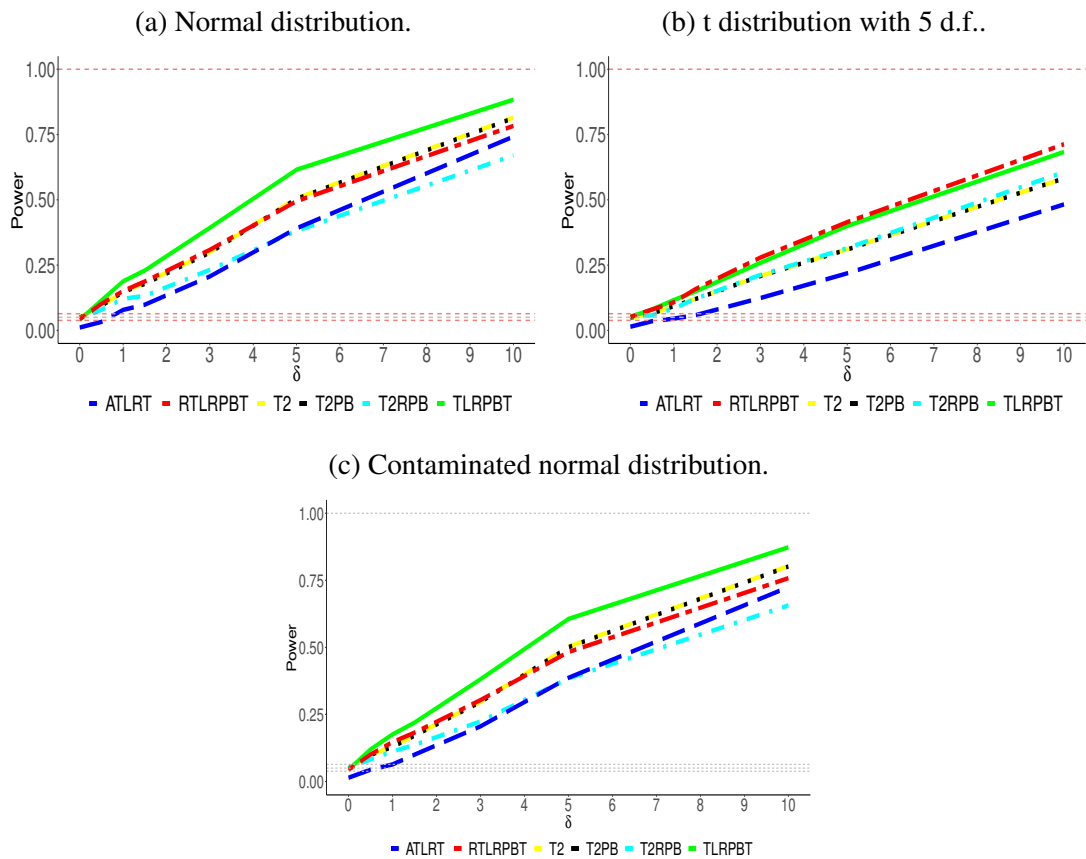
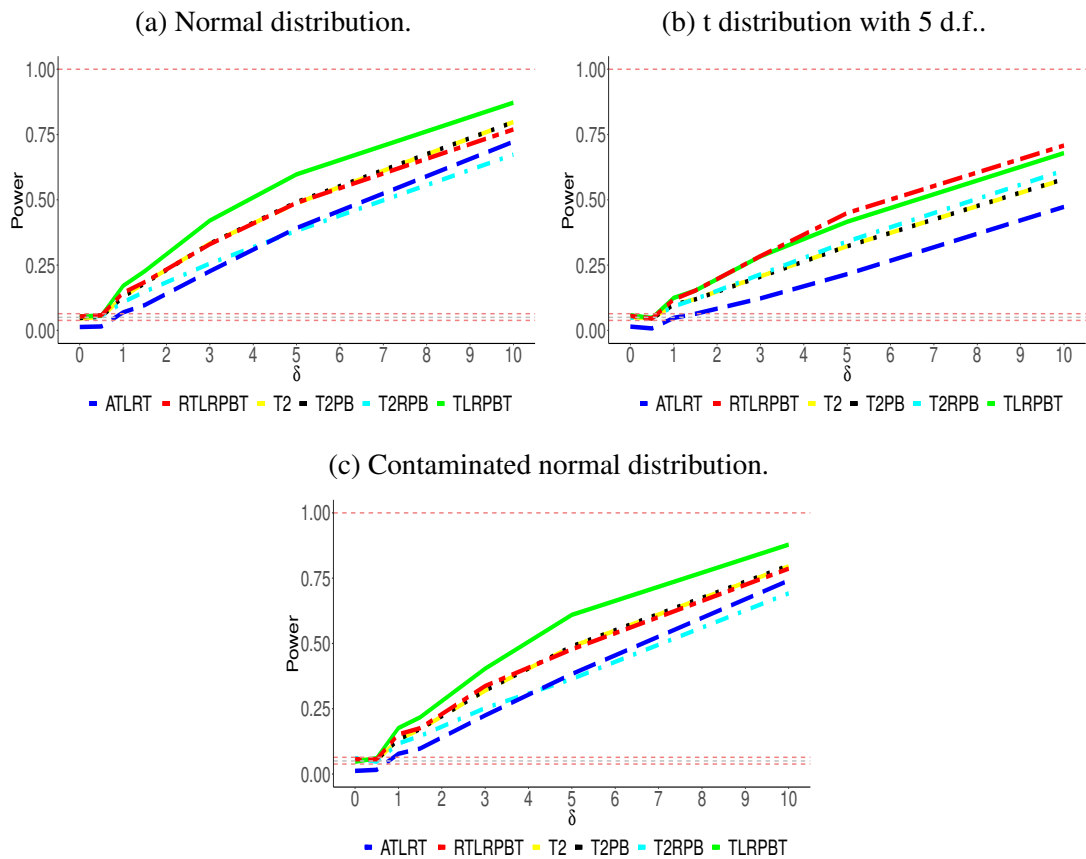


Figura 4.15 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



pattern, since they control the type I error rates and show higher power when compared with the multivariate normal case.

It was decided to increase the dimension to $p = 5$ and to maintain the same sample sizes n and distributions. Similar performance of all tests that presented at $p = 2$ remains. Considering the dimension $p = 5$, Figures 4.16, 4.17, 4.18, 4.19 and 4.20 show the results obtained for the power of the tests. Figures 4.16 (b), 4.17 (b), 4.18 (b), 4.19 (b) and 4.20 (b) show that power of the parametric bootstrap version (*TLRPBT*) and the robust parametric bootstrap version (*RTLRPBT*) are equivalent and higher than the other tests in the multivariate Student t -distribution with 5 degrees of freedom. In general, the *TLRPBT* performs better. The *ATLRT* continues to show substantial low power, once it has shown to be a conservative test (see Tables (4.1, 4.2)) this pattern of behavior was expected.

Figura 4.16 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

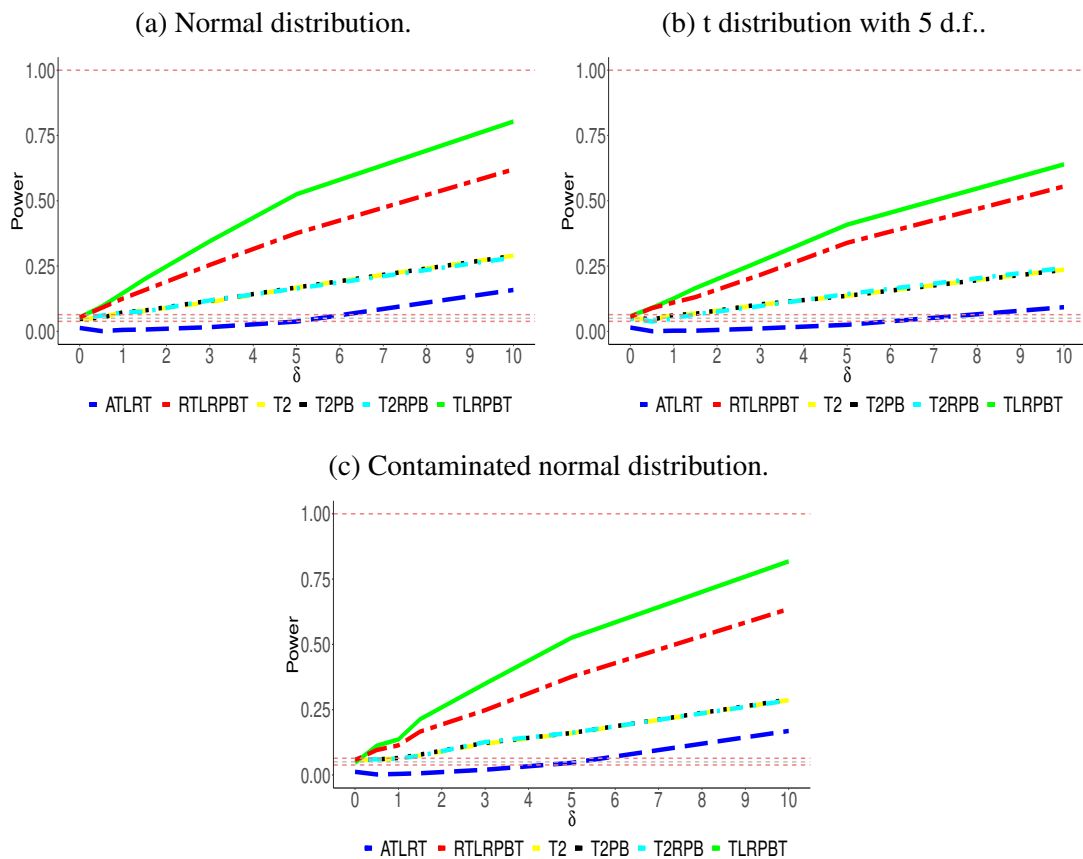


Figura 4.17 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

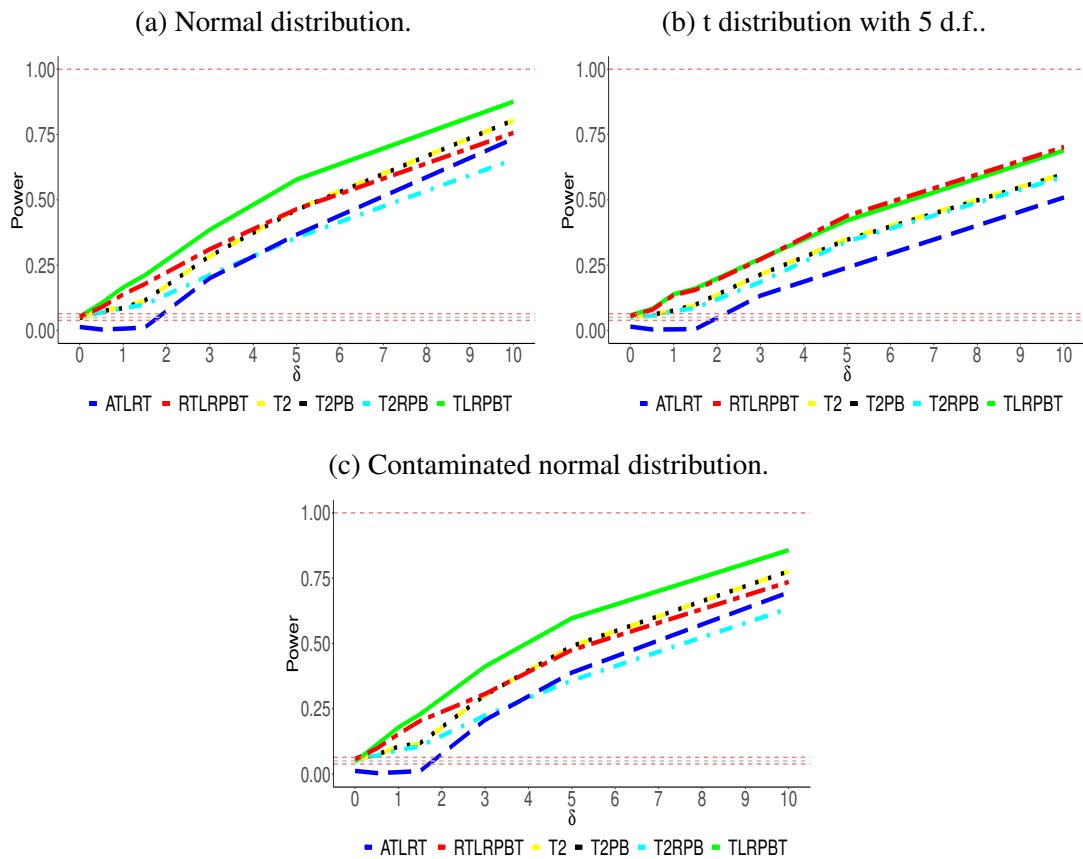


Figura 4.18 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

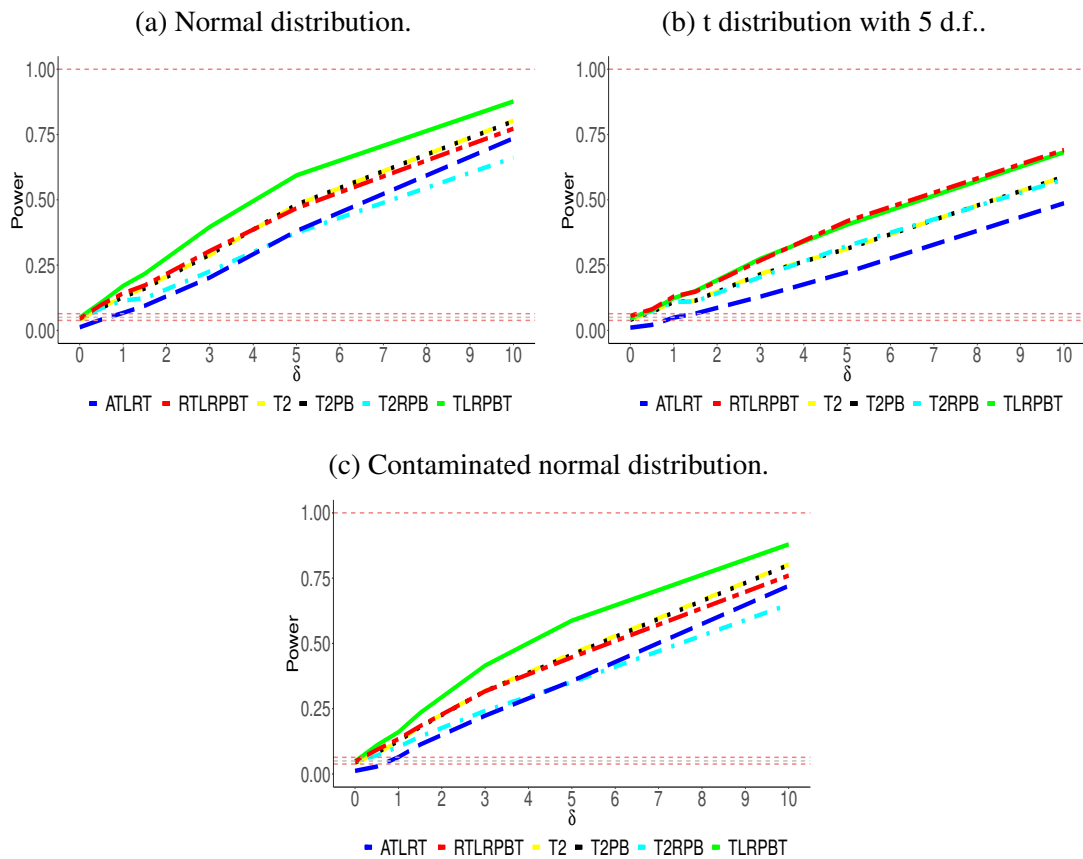
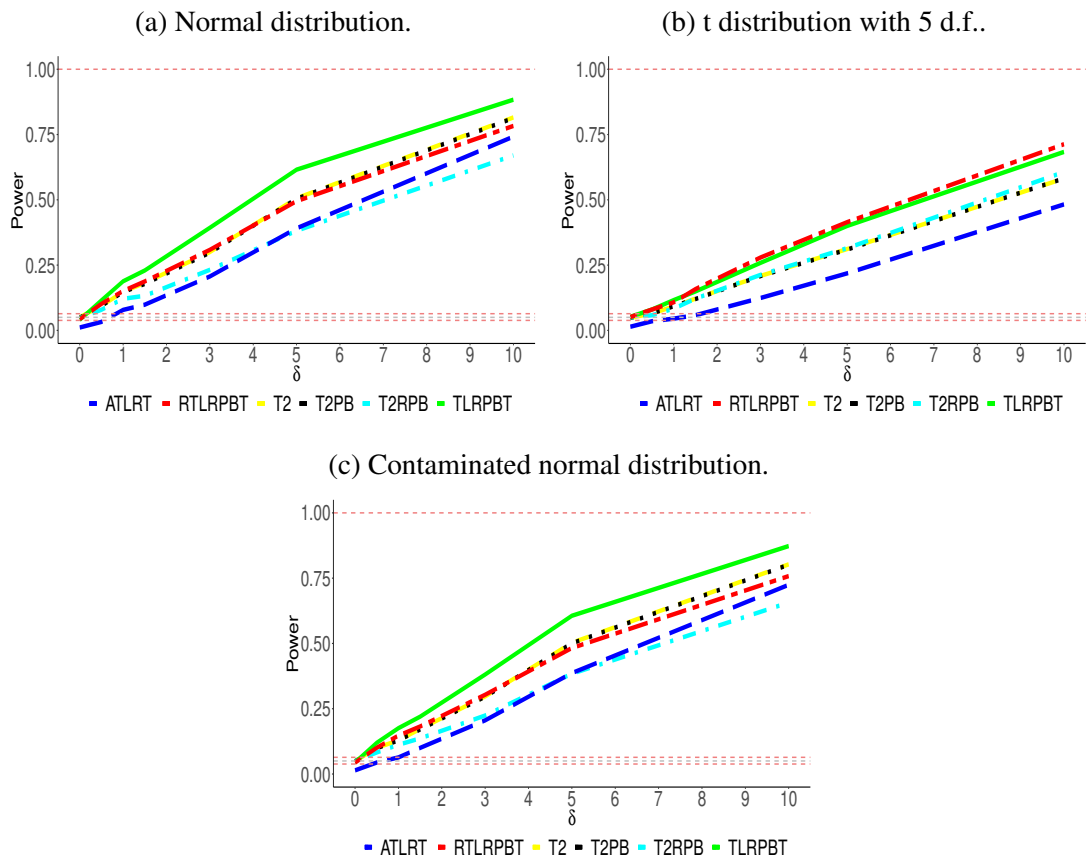


Figura 4.19 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



3.3 Special case of $p = 200$: type I error rates and power

Table 4.3 show the results considering this case of high dimensionality ($p = 200$). It was considering sample sizes of $n = 10, 50, 70, 100$, and 200 and therefore deal with the cases where $p \geq n$. For this circumstance, the traditional T^2 test and the adapted $T2PB$, $T2RPB$ can not be applied due to the high dimension issue (see section 1). It was noticed that the $ATLRT$ was conservative in all scenarios considered. It was expected that this test shows substantial low power. It was also noticed that the other adapted tests, $TLRPBT$ and $RTLRPBT$, are exact, according to the binomial test (see section 2).

Tabela 4.3 – Type I error rates of the three tests with $\alpha = 5\%$ and $p = 200$, considering the multivariate Normal (N), multivariate Student t with 5 degrees of freedom (t_5) and multivariate contaminated normal (CN) distributions.

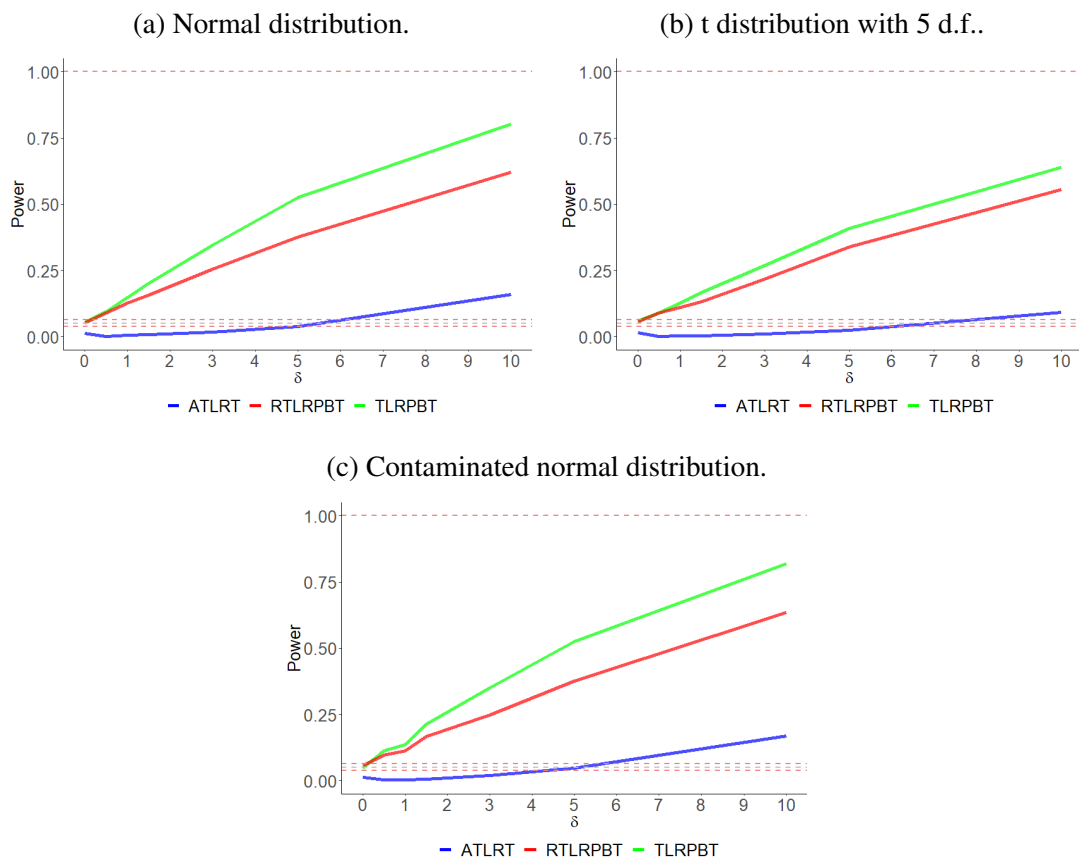
Test	Model	Sample sizes n				
		10	50	70	100	200
ATLRT	N	0.0130 ⁻	0.0130 ⁻	0.0120 ⁻	0.0110 ⁻	0.0130 ⁻
	t_5	0.0145 ⁻	0.0145 ⁻	0.0105 ⁻	0.0140 ⁻	0.0145 ⁻
	CN	0.0120 ⁻	0.0120 ⁻	0.0120 ⁻	0.0140 ⁻	0.0120 ⁻
TLRPBT	N	0.0535	0.0535	0.0495	0.0450	0.0535
	t_5	0.0580	0.0580	0.0460	0.0520	0.0580
	CN	0.0470	0.0470	0.0485	0.0445	0.0470
RTLRPBT	N	0.0530	0.0530	0.0445	0.0430	0.0530
	t_5	0.0555	0.0555	0.0555	0.0510	0.0555
	CN	0.0570	0.0570	0.0480	0.0435	0.0570

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5%.

The power results for $p = 200$ are shown in Figures 4.21 to 4.25. Considering $n = 10$, Figure 4.21 shows that the $TLRPBT$ outperformed the other proposed tests. It was also noticed that, unlike the other dimensions considered ($p = 2$ and $p = 5$), the $TLRTBP$ and $RTLRPBT$ perform similarly with the multivariate Student t -distribution with 5 degrees of freedom (Figure 4.21 (b)). It was also noticed for this scenario that the asymptotic version ($ATLRT$) shows substantial low power. This fact was already expected, as it did control the type I error rates in a conservative way (Table 4.3).

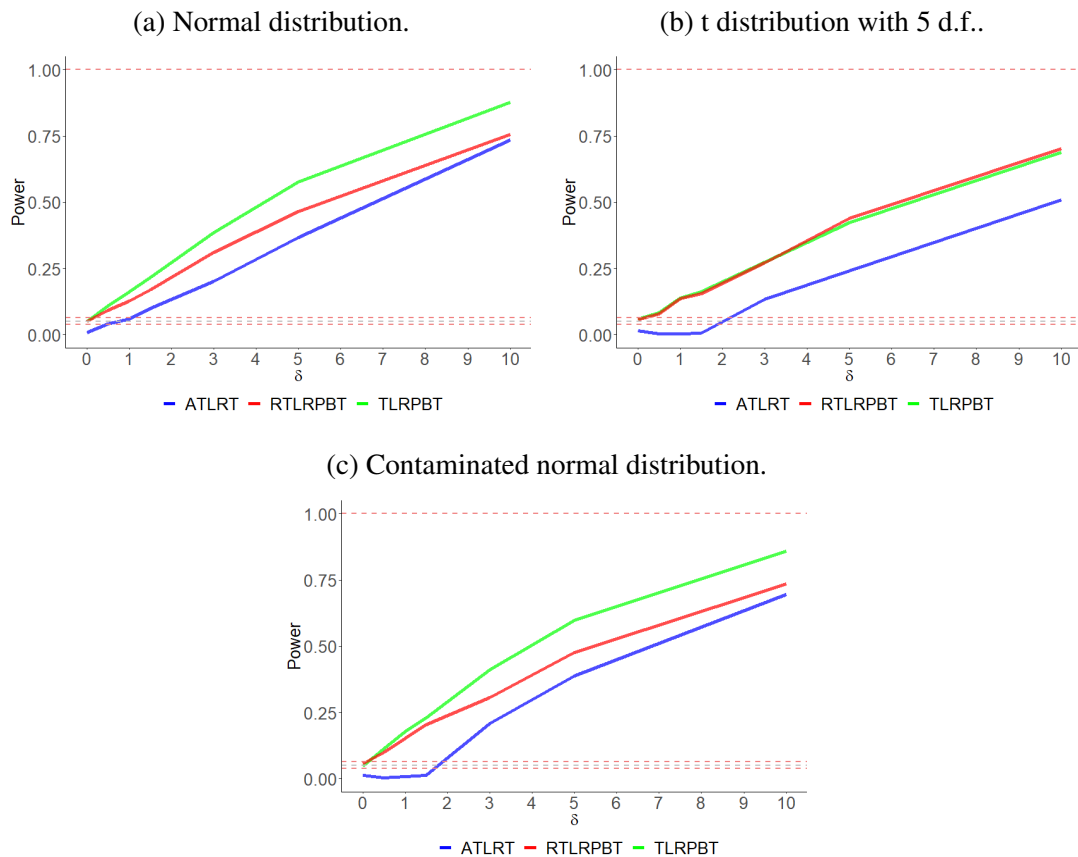
When the sample size n was increased, it was noticed that the $TLRPBT$ continues to perform better (Figures 4.22, 4.23, 4.24 and 4.25). For the multivariate Student t -distribution with 5 degrees of freedom (Figures 4.22 (b), 4.23 (b), 4.24 (b) and 4.25 (b)), the $TLRPBT$ and $RTLRPBT$ had similar performance. This pattern is identical to those for the dimensions of $p = 2$ and $p = 5$. In contrast, the $ATLRT$ has substantial power gain for this dimension

Figura 4.21 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 10$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



($p = 200$). This fact can be noticed in Figures 4.23, 4.24 and 4.25. In general, the *TLRPBT* (parametric bootstrap version test) outperformed in power when confronted with other tests.

Figura 4.22 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 50$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



In a more general analysis, considering all scenarios evaluated ($n = 10, 50, 70, 100$, and 200 and $p = 2, 5, 200$), the Monte Carlo simulations for type I error rate and power showed that the *TLRPBT* performed better. We recommend using this test as it is not hard to implement and computationally fast.

Willems et al. (2002) concluded that his proposed new test T_R^2 showed power losses when compared to the traditional Hotelling's T^2 test for several configurations of n and p ($n > p$). Under contaminated normal p -variate populations, with 10% of contamination, the T_R^2 has also less power than the traditional Hotelling's T^2 test. (DONG et al., 2016) concluded that his proposed new test for high dimensional data, based on a shrinkage process of the traditional Hotelling's T^2 statistic test, showed high power under p -variate normal and Student t with 4 degrees of freedom distributions, considering different values of μ .

Figura 4.23 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 70$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

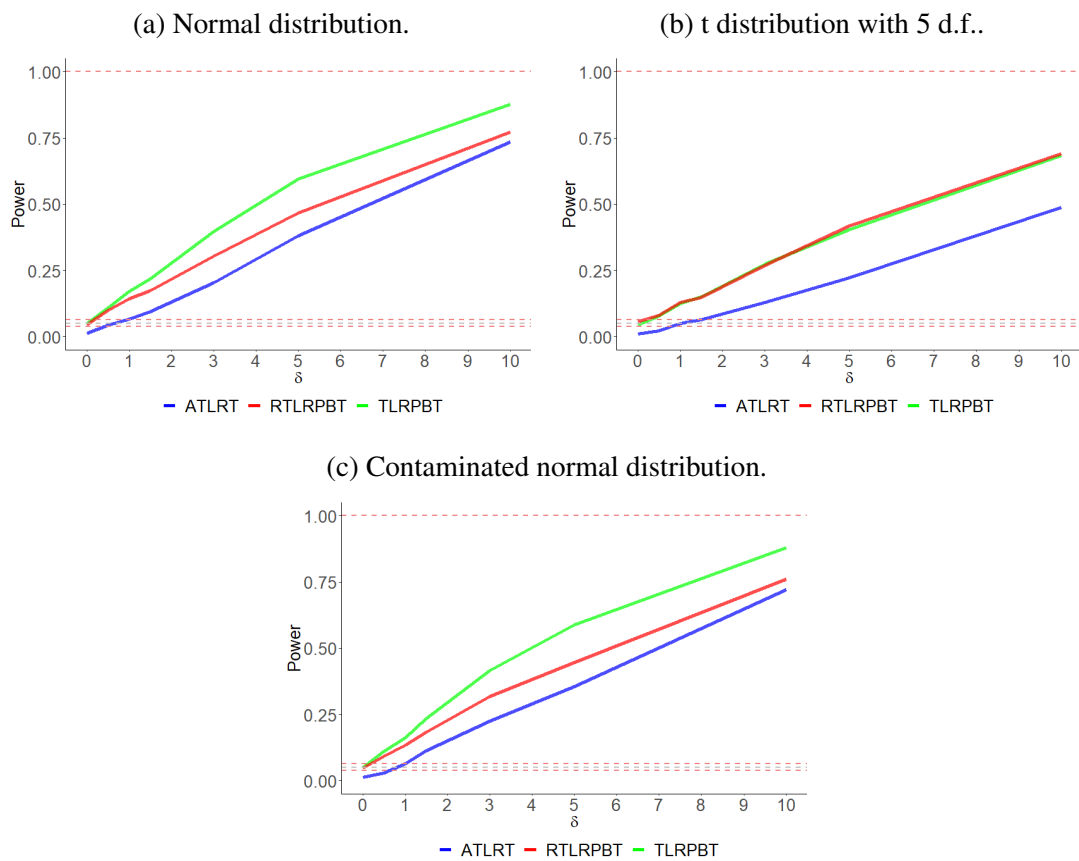


Figura 4.24 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 100$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

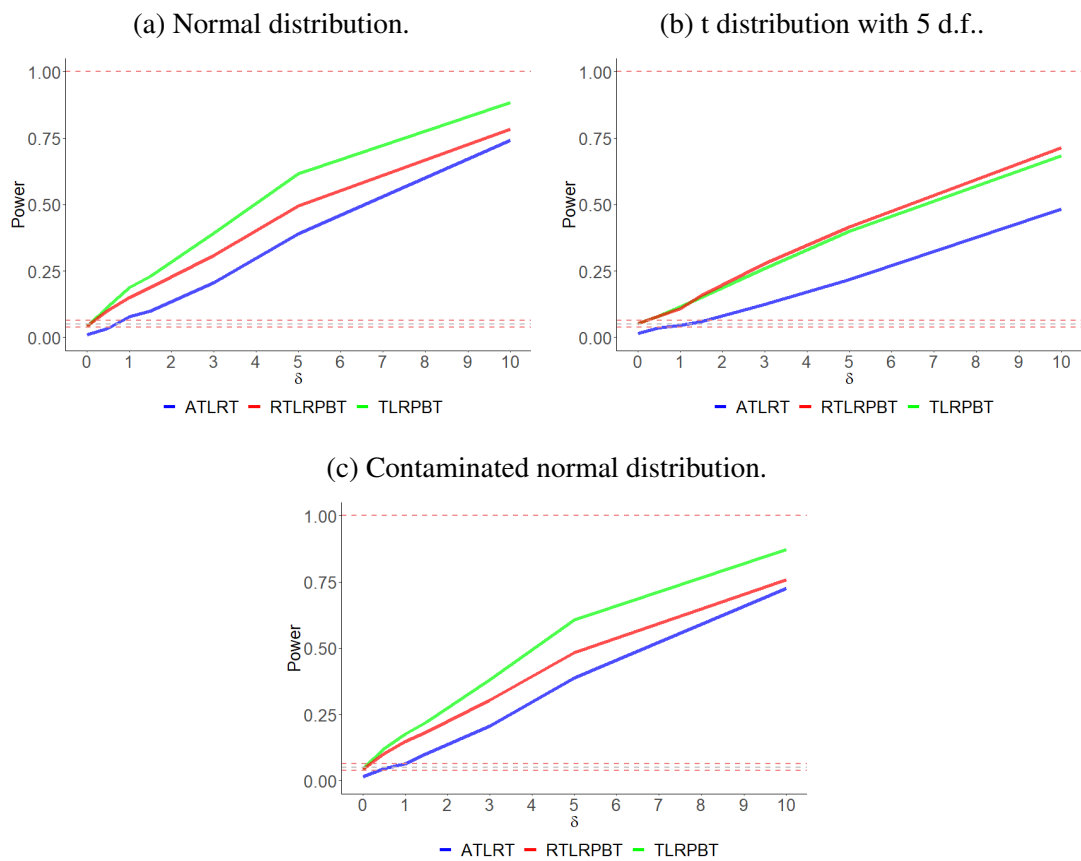
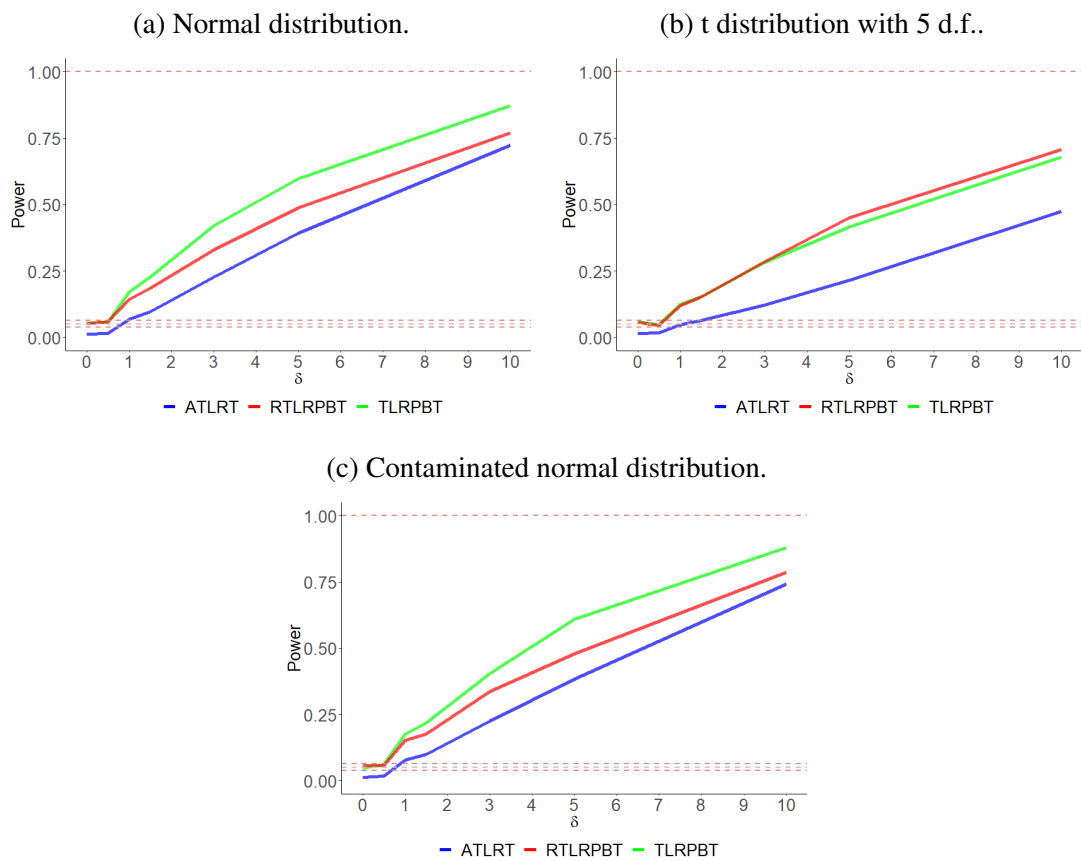


Figura 4.25 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with $n = 200$ and $p = 200$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.



In both tests (WILLEMS et al., 2002; DONG et al., 2016), the effect of the sample size n influenced the power, which is an expected fact. In the results of the present work it does not occur, as presented before, because the values of the population means change when the sample sizes change, keeping fixed the value of the distance of Mahalanobis $\delta(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$. To clarify, note that Willems et al. (2002) and Dong et al. (2016) kept the population mean vector $\boldsymbol{\mu}$ fixed and, therefore, the value of Mahalanobis distance δ increases with increasing sample size n . Thus the power grew with an increase of n . On the contrary, in the present work $\boldsymbol{\mu}$ changed in each simulation when n varied, keeping the value of δ fixed, which keeps the power practically constant, as can be seen from Figures 4.11 to 4.25. Differences in the power values found under identical configuration, but at different values of n , were attributed to the Monte Carlo error.

3.4 Special cases

Marco Marozzi (2015) proposes an alternative multivariate test class for case-control studies for high dimensional data, considering heavy tails or skewed distributions. The proposed tests were based on the combination of tests on interpoint distances. The Euclidean distance was utilized. These tests were exact, unbiased and consistent. The results showed that the proposed tests are very powerful under normality, heavy tails, and skewed distributions. Marco Marozzi (2016) applied these same tests to magnetic resonance data which are usually with few observations and many variables, that is, high-dimensional data.

We decided to verify the behavior of our proposed tests regarding the type I error control and power considering heavy tails and skewed distributions. For this, we consider the multivariate Student t with 1 degree of freedom and the multivariate exponential with parameter $\boldsymbol{\lambda} = \mathbf{1}$ distributions. In the latter case, data were shifted to zero mean by subtracting the population exponential mean $\boldsymbol{\mu} = \mathbf{1}$. We considered in our simulations only the dimensions $p = 2$ and 5 and the same sample sizes n of section 2.

3.4.1 Multivariate t_1 distribution

The tests, generally, controlled the type I error in an exact or conservative way for the multivariate Student t -distribution with 1 degree of freedom, for dimensions of $p = 2$ and 5, with different sample sizes n (see Tables 4.4 and 4.5) at the nominal significance level of 5%. The same response pattern occurred for the nominal significance levels of 1% and 10% (results

not shown). The $T2BP$ test with $n = 10$ and 50 for $p = 2$ and also with $n = 50$ for $p = 5$ was liberal, although with test sizes not expressively much higher than the nominal significance level. This same test for $n \geq 70$ with $p = 2$ and 5 and $n = 10$ with $p = 5$ was exact.

Tabela 4.4 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate Student t -distribution with 1 degrees of freedom (t_1).

Test	Model	Sample sizes n				
		10	50	70	100	200
T2		0.0190 ⁻	0.0190 ⁻	0.0109 ⁻	0.0189 ⁻	0.0149 ⁻
T2BP		0.0815 ⁺	0.0815 ⁺	0.0558	0.0578	0.0508
T2RBP		0.0190 ⁻	0.0190 ⁻	0.0189 ⁻	0.0229 ⁻	0.0169 ⁻
ATLRT	t_1	0.0030 ⁻	0.0030 ⁻	0.0029 ⁻	0.0019 ⁻	0.0019 ⁻
TLRPBT		0.0220 ⁻	0.0220 ⁻	0.0179 ⁻	0.0209 ⁻	0.0179 ⁻
RTLRPBT		0.0200 ⁻	0.0200 ⁻	0.0199 ⁻	0.0229 ⁻	0.0269 ⁻

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5%.

⁺: significantly ($P < 0.01$) greater than the nominal significance level of 5%.

(MAROZZI; MARCO, 2016) applying their interpoint distance and permutation-based tests found exact results in almost all cases for the multivariate Student t -distribution with $\nu = 1$ degrees of freedom with $p = 2$ and $p = 4$. The HT (Hotelling T^2) test was the only one that yielded conservative results such as in the present work ($T2$). Note that (MAROZZI; MARCO, 2016)'s tests are only appropriate for the case-control studies because they are permutation-based tests and are not suitable for the one-sample case. It is interesting to note that the conservative results found here may be associated with the non-existence of the multivariate Cauchy distribution mean (multivariate Student t -distribution with 1 degree of freedom). Moreover, the proposed tests are based on parametric bootstrap considering the multivariate normal distribution for generating the null distribution of the test statistics. Therefore, low robustness tests are expected when data depart from normal distribution.

Regarding power, the $T2BP$, $T2RBP$, and $RTLRPBT$ tests showed highlighted performance (see Figures 4.26 and 4.27). For $p = 2$ (see Figure 4.26), the $T2BP$ test was, generally, the most powerful. However, with $n = 10$ and 50 , its superiority should not be considered, since it has been considered a liberal test (test size greater than the significance nominal level). The performance of this test was accompanied by the $RTLRPBT$ and $T2RBP$ tests in this order. For $p = 5$, the $RTLRPBT$ performed better. This test was consistently among the best tests in the previously other simulated cases, where the $TLRPBT$ that uses the non-robust version of the covariance estimator stood out as the best. In the present case, with a heavier tail distribution, the test with the robust version performed better, as expected.

Tabela 4.5 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate Student t -distribution with 1 degrees of freedom (t_1).

Test	Model	Sample sizes n				
		10	50	70	100	200
T2		0.0259	0.0199	0.0119	0.0119	0.0119
T2BP		0.0448	0.0638 ⁺	0.0548	0.0548	0.0548
T2RBP		0.0209 ⁻	0.0149 ⁻	0.0169 ⁻	0.0169 ⁻	0.0169 ⁻
ATLRT	t_1	0.0000 ⁻	0.0000 ⁻	0.0000 ⁻	0.0000 ⁻	0.0000 ⁻
TLRPBT		0.0209 ⁻	0.0269 ⁻	0.0149 ⁻	0.0149 ⁻	0.0149 ⁻
RTLRPBT		0.0199 ⁻	0.0249 ⁻	0.0229 ⁻	0.0229 ⁻	0.0229 ⁻

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5%.

⁺: significantly ($P < 0.01$) greater than the nominal significance level of 5%.

3.4.2 Multivariate exponential distribution

To the multivariate exponential distribution, in general, the tests were very liberal. The exception occurred for *ATLRT* and *TLRPBT*. The *ATLRT* was exact for $n = 10$ and conservative for $n \geq 50$ with $p = 2$ or $p = 5$ (see Tables 4.6 and 4.7). The *TLRPBT* did not control the type I error for $n = 10$ and $n = 100$ with $p = 2$ and $p = 5$, showing a liberal behavior (see Tables 4.6 and 4.7, though with no expressive difference from the nominal significance level of 5%. In the other cases, it was exact. For large enough n , say $n = 200$, the *T2* and *T2BP* tests showed either type I error rates control or inexpressive liberal behavior, although significant (see Tables 4.6 and 4.7).

The power of the tests for exponential distribution with $n = 200$ and $p = 2$ and $p = 5$ can be seen in Figures 4.28 (a) and (b). Only the *T2*, *ATLRT*, and *TLRPBT* tests should be considered in the comparison, as they were those who controlled the Type I error rates. Again, the *TLRPBT* test was the most powerful, especially with $\delta = 10$, followed by the *T2* and *ATLRT* tests, in this order.

The Table 4.8 shows the results of type I error rates for multivariate exponential distribution. We consider the dimension $p = 200$ and the sample size $n = 200$, with the nominal significance level of 5%, i.e., an extreme case. We realized that only the *ATLRT* and *TLRPBT* tests controlled the type I error, being the first conservative and the second, exact. The margin of error of the Monte Carlo simulations is also presented (see Table 4.8). The *RTLRPBT* test showed all type I errors close to 1 and was extremely liberal in this high-dimensional case.

Figura 4.26 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with t distribution with 1 d.f. and $p = 2$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

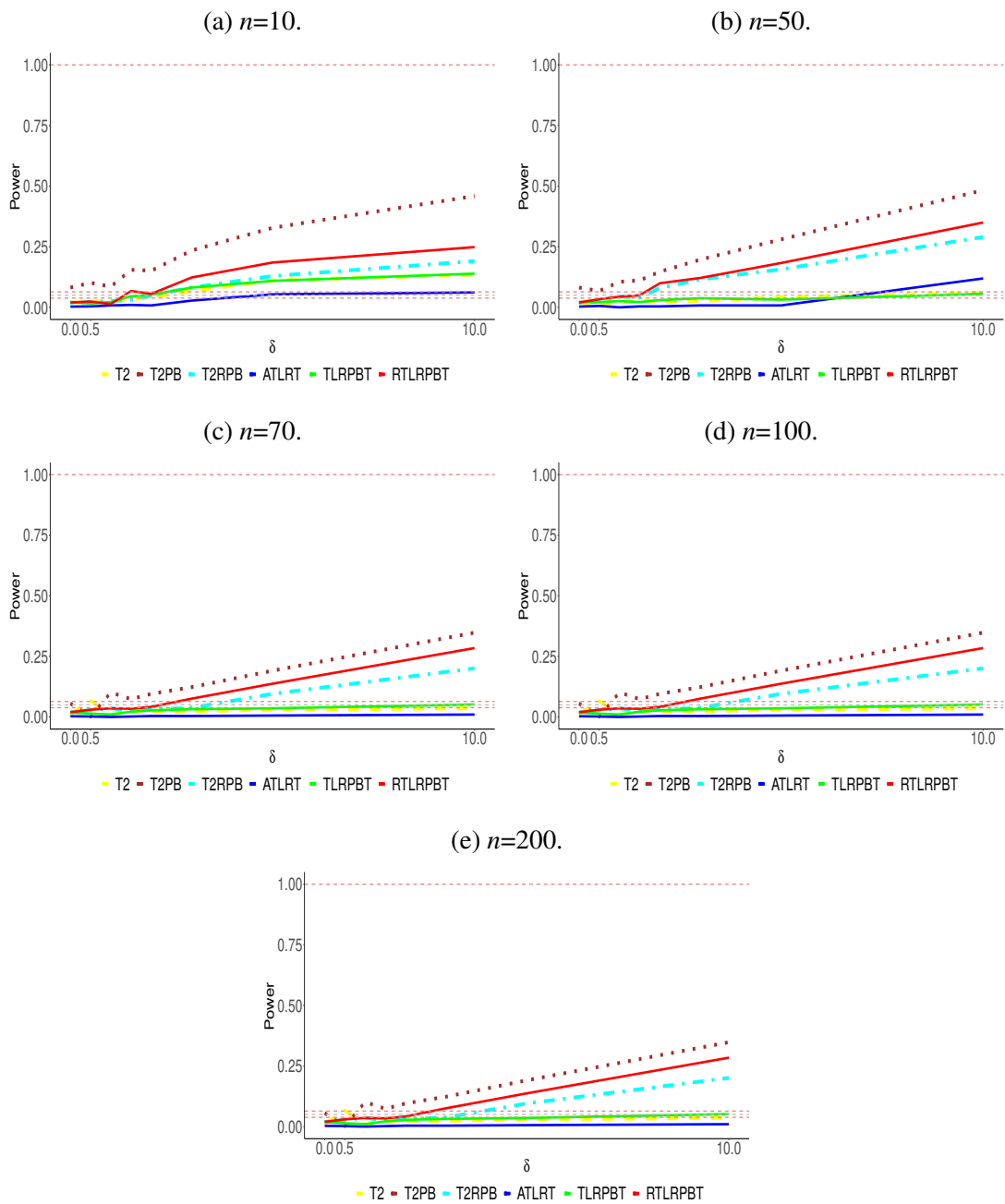


Figura 4.27 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with t distribution with 1 d.f. and $p = 5$, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

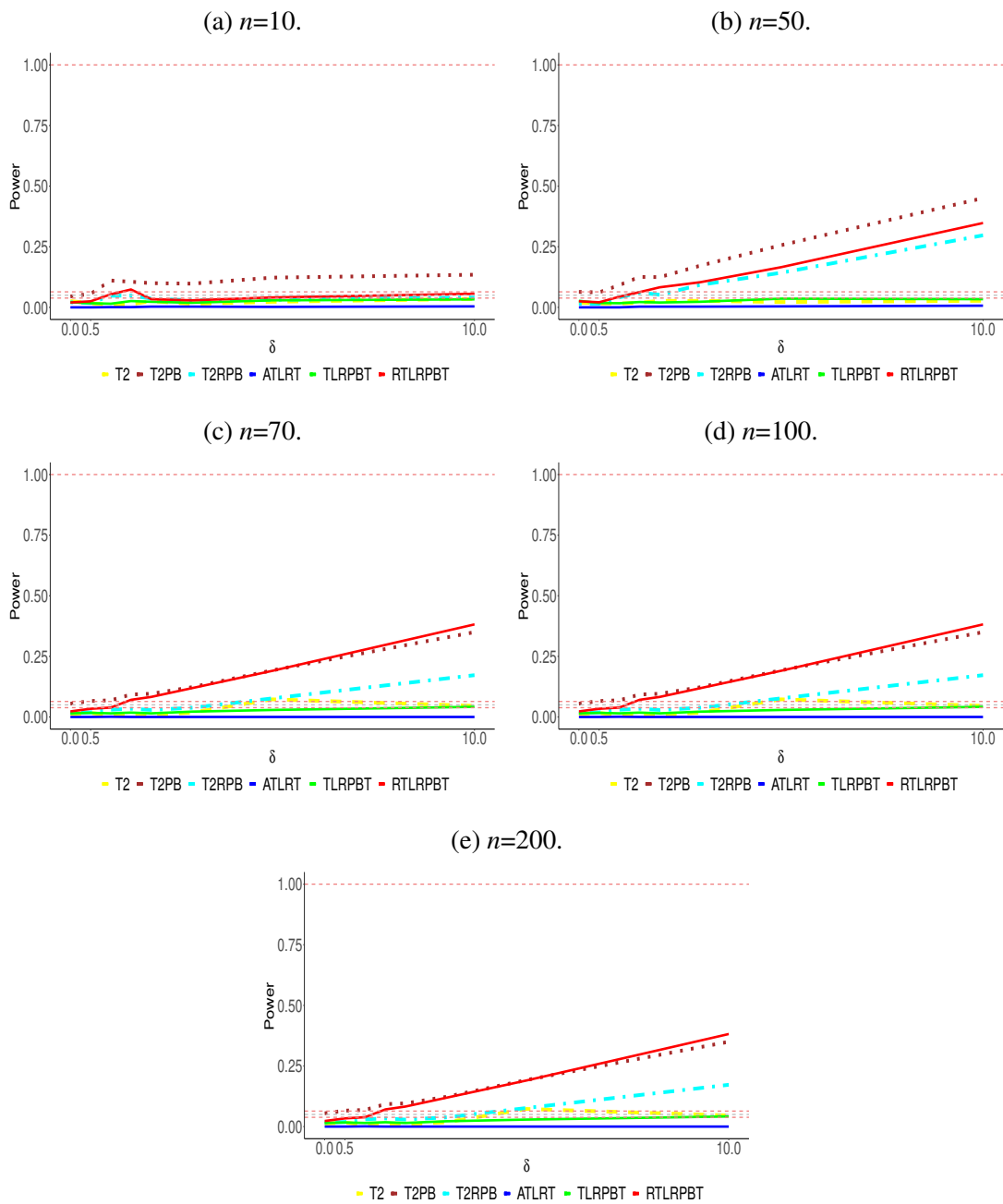


Tabela 4.6 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 2$, considering the multivariate exponential distribution.

Test	Model	Sample sizes n				
		10	50	70	100	200
T2		0.1296 ⁺	0.0778 ⁺	0.0648 ⁺	0.0658 ⁺	0.0558
T2BP		0.5414 ⁺	0.9651 ⁺	0.0648 ⁺	0.0658 ⁺	0.0558
T2RBP		0.3470 ⁺	0.9412 ⁺	0.9821 ⁺	1.0000 ⁺	1.0000 ⁺
ATLRT	exponential	0.0459	0.0269 ⁻	0.0160 ⁻	0.0199 ⁻	0.0070 ⁻
TLRPBT		0.0927 ⁺	0.0608	0.0439	0.0658 ⁺	0.0489
RTLRPBT		0.2393 ⁺	0.6112 ⁺	0.7149 ⁺	0.8335 ⁺	0.9811 ⁺

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5%.

⁺: significantly ($P < 0.01$) greater than the nominal significance level of 5%.

Tabela 4.7 – Type I error rates of the six tests with $\alpha = 5\%$ and $p = 5$, considering the multivariate exponential distribution.

Test	Model	Sample sizes n				
		10	50	70	100	200
T2		0.1296 ⁺	0.0778 ⁺	0.0648 ⁺	0.0658 ⁺	0.0558
T2BP		0.5517 ⁺	0.9753 ⁺	0.0676 ⁺	0.0759 ⁺	0.0658 ⁺
T2RBP		0.3490 ⁺	0.9423 ⁺	0.9827 ⁺	1.0000 ⁺	1.0000 ⁺
ATLRT	exponential	0.0449	0.0267 ⁻	0.0180 ⁻	0.0299 ⁻	0.0080 ⁻
TLRPBT		0.0929 ⁺	0.0618	0.0459	0.0755 ⁺	0.0495
RTLRPBT		0.2397 ⁺	0.6125 ⁺	0.7193 ⁺	0.8435 ⁺	0.9831 ⁺

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5%.

⁺: significantly ($P < 0.01$) greater than the nominal significance level of 5%.

Figura 4.28 – Power of the tests as a function of the generalized Mahalanobis distance δ between the parametric and hypothetical vector means, with multivariate exponential distribution, where the gray lines represent the lower and upper limits of the exact binomial interval and the maximum power.

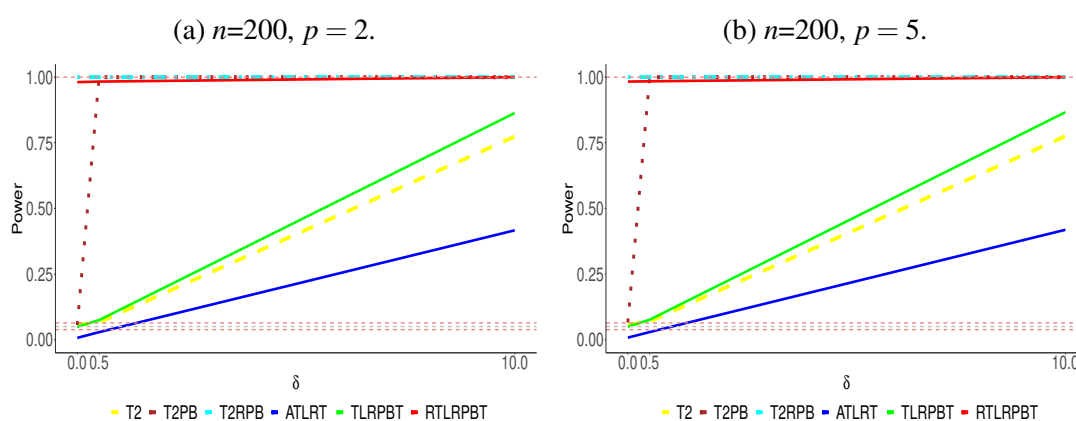


Tabela 4.8 – Type I error rates ($\delta = 0$) and power ($\delta = 0.5$ and 10) of the three tests with $\alpha = 5\%$, $p = 200$ and $n = 200$, considering the multivariate exponential distribution, where the values in parenthesis express the margin of error of Monte Carlo simulations.

Test	Model	δ		
		0	0.5	10
ATLRT		0.0069 ⁻ (0.0051)	0.0000(0.0036)	0.0000(0.0036)
TLRPBT	exponential	0.0488 (0.0133)	0.0473(0.0131)	0.9909(0.0058)
RTLRPBT		0.9810 ⁺ (0.0084)	0.9673(0.0110)	1.0000(0.0036)

⁻: significantly ($P < 0.01$) lower than the nominal significance level of 5%.

⁺: significantly ($P < 0.01$) greater than the nominal significance level of 5%.

4 CONCLUSION

The *TLRPBT* is recommended for testing hypothesis about a p -variate population mean vector of normal and non-normal populations, including the presence of outliers. It is possible to conclude that the use of robust comedian mean and covariance estimators is not helpful for testing hypotheses on a population mean vector. Special attention should be paid to the *ATLRT* test, as it was conservative in all scenarios and, despite presenting substantial loss of power in some scenarios, it resisted regarding the control of type I error even considering distributions that present extreme asymmetry (Cauchy and Exponential). In that case, this test should be considered as having a good performance.

REFERÊNCIAS

- ALVES, H. JP, FERREIRA, D. F. Proposition of new alternative tests adapted to the traditional T^2 test. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, p. 1–14, 2019.
- BAI, Z.; SARANADASA, H. Effect of high dimension: by an example of a two sample problem. **Statistica Sinica**, JSTOR, p. 311–329, 1996.
- CHEN, S. X.; QIN, Y.-L. et al. A two-sample test for high-dimensional data with applications to gene-set testing. **The Annals of Statistics**, Institute of Mathematical Statistics, p. 808–835, 2010.
- DONG, K. et al. Shrinkage-based diagonal hotelling's tests for high-dimensional small sample size data. **Journal of Multivariate Analysis**, Elsevier, v. 143, p. 127–142, 2016.
- FALK, M. On mad and comedians. **Annals of the Institute of Statistical Mathematics**, Springer, v. 49, n. 4, p. 615–644, 1997.
- FERREIRA, D. F. **Estatística multivariada**. 3. ed. Universidade Federal de Lavras: Editora UFLA, 2018. 624 p.
- LEDOIT, O.; WOLF, M. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 30, n. 4, p. 1081–1102, 2002.
- LEE, S. H. et al. Order test for high-dimensional two-sample means. **Journal of Statistical Planning and Inference**, Elsevier, v. 142, n. 9, p. 2719–2725, 2012.
- MARONNA, R. A.; ZAMAR, R. H. Robust estimates of location and dispersion for high-dimensional datasets. **Technometrics**, Taylor & Francis, v. 44, n. 4, p. 307–317, 2002.
- MAROZZI, MARCO. Multivariate multidistance tests for high-dimensional low sample size case-control studies. **Statistics in medicine**, Wiley Online Library, v. 34, n. 9, p. 1511–1526, 2015.
- MAROZZI, MARCO. Multivariate tests based on interpoint distances with application to magnetic resonance imaging. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 25, n. 6, p. 2593–2610, 2016.

MUDHOLKAR, G. S.; SRIVASTAVA, D. K. A class of robust stepwise alternatives to hotelling's T^2 tests. **Journal of Applied Statistics**, Taylor & Francis, v. 27, n. 5, p. 599–619, 2000.

OLIVEIRA, I. R. Cardoso de; FERREIRA, D. F. Multivariate extension of chi-squared univariate normality test. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 80, n. 5, p. 513–526, 2010.

PAN, G. M.; ZHOU, W. Central limit theorem for hotelling's T^2 statistic under large dimension. **The Annals of Applied Probability**, JSTOR, p. 1860–1910, 2011.

SAJESH, T. A.; SRINIVASAN, M. R. Outlier detection for high dimensional data using the comedian approach. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 82, n. 5, p. 745–757, 2012.

SRIVASTAVA, M. S. A test for the mean vector with fewer observations than the dimension under non-normality. **Journal of Multivariate Analysis**, Elsevier, v. 100, n. 3, p. 518–532, 2009.

SRIVASTAVA, M. S.; DU, M. A test for the mean vector with fewer observations than the dimension. **Journal of Multivariate Analysis**, Elsevier, p. 386–402, 2008.

SRIVASTAVA, M. S.; KATAYAMA, S.; KANO, Y. A two sample test in high dimensional data. **Journal of Multivariate Analysis**, Elsevier, v. 114, p. 349–358, 2013.

TIKU, M. L. Robust statistics for testing equality of means or variances. **Communications in Statistics-Theory and Methods**, Taylor and Francis, v. 11, n. 22, p. 2543–2558, 1982.

WANG, R.; PENG, L.; QI, Y. Jackknife empirical likelihood test for equality of two high dimensional means. **Statistica Sinica**, JSTOR, p. 667–690, 2013.

WILLEMS, G. et al. A robust hotelling test. **Metrika**, Springer, p. 125–138, 2002.