ESTIMADORES DA PROBABILIDADE TOTAL DE CLASSIFICAÇÃO INCORRETA NA ANÁLISE DISCRIMINANTE

ALTEMIR DA SILVA BRAGA

2008

ALTEMIR DA SILVA BRAGA

ESTIMADORES DA PROBABILIDADE TOTAL DE CLASSIFICAÇÃO INCORRETA NA ANÁLISE DISCRIMINANTE

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

Orientador

Prof. Dr. Daniel Furtado Ferreira

LAVRAS

MINAS GERAIS-BRASIL

2008

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Braga, Altemir Silva.
Estimadores da probabilidade total de classificação incorreta na análise discriminante / Altemir Silva Braga. – Lavras : UFLA, 2008.
65 p. : il.

Dissertação (Mestrado) – Universidade Federal de Lavras, 2008.
Orientador: Daniel Furtado Ferreira.
Bibliografia.

1. Multivariada. 2. Método. 3. Classificar. 4. Taxa de erros. 5. Jackknife. I. Universidade Federal de Lavras. II. Título.

CDD – 519.2

ALTEMIR DA SILVA BRAGA

ESTIMADORES DA PROBABILIDADE TOTAL DE CLASSIFICAÇÃO INCORRETA NA ANÁLISE DISCRIMINANTE

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

APROVADA em 30 de outubro de 2008.

Prof. Dr. Eric Batista Ferreira	UNIFAL
Prof. Dr. Júlio Sílvio de S. Bueno Filho	UFLA
Prof. Dr. Marcelo Angelo Cirillo	UFLA

Prof. Dr. Daniel Furtado Ferreira
UFLA
(Orientador)

LAVRAS
MINAS GERAIS - BRASIL

Aos meus pais, Sebastião e Maria Lucy
e a minha esposa Carmem

Dedico.

AGRADECIMENTOS

A Deus, por tudo.

A minha esposa Carmem, por ser uma mulher maravilhosa e especial na minha vida, por acreditar no meu trabalho, ser uma ótima pessoa e uma bela mãe para nossos filhos, Tays, Eric e Caio.

Ao professor professor Daniel Furtado Ferreira, pela responsabilidade que tem com seu trabalho, pela confiança que me concedeu, pelos ensinamentos e orientações, muitas vezes tirando dúvidas, outros vezes dando sugestões, criticando, e fazendo as coisas acontecerem.

Aos meus pais, Sebastião e Maria Lucy, pela confiança, compreensão e carinho.

Aos meus tios, Raimundo e Nair, pela confiança, compreensão e a oportunidade que me deram de concluir o ensino fundamental e o médio.

Aos meus irmãos, Adalzemir, Luciana, Geane, Leane, Vanuza, Nailton, Railton, Natalício e Lorinha, pelos momentos alegres que passamos.

Ao meu amigo Eustáquio, a minha amiga Francisca, a minha amiga Dalvinha, pelo apoio, amizade, carinho que tiveram com meus filhos e minha esposa, e pelo respeito e confiança que têm com minha pessoa, a amizade de vocês é que nem leite, é do peito.

Aos meus amigos Ângelo, José Roberto, Reginaldo, Luciano, Jairo, Itamar, Emanuel, Cleilton, Cruz, Soriano, Sarquiz, pela amizade, respeito e carinho.

A todos os colegas de mestrado e doutorado em Estatística, em especial a minha turma Edcarlos, Ricardo, Paulo, Patrícia, Tânia, Augusto, Ana Paula, Denise, Richardson, Isabel, Iron, Stephânia e Vanderley pela amizade.

Ao amigo Edcarlos (Chapolin), pelas horas de estudos, sugestões, críticas, apoio e amizade.

A todos os meus amigos do CEBRB, em especial, à professora Neide que acredita em meu trabalho.

À dona Alda e ao Paulo, pela confiança e amizade.

À dona Bebé, Eliana, Elisandra, Disney, Gleide e Jorge, pela boa amizade, respeito e carinho.

À dona Graça, que Deus a tenha.

Ao Devanil pelas aulas de probabilidades, inferência e pela amizade.

Ao meu avô Natanael e meus tios, João, Messias, Lucimar, Iris e Luciano, pela amizade e carinho.

A minha tia Cleuza e ao meu tio Valdo.

Aos professores Antônio Carlos e José Marques, pelas horas de estudo e amizade.

A minha co-orientadora Patrícia de Siqueira Ramos, pelos ensinamentos e respeito.

A todos os meus professores da graduação e da especialização.

À dona Eulália, Cinthia, Débora, Ducarmo e Dionathan, pela amizade e carinho.

À galera da Bahia: Wal, Vitória, Cleilton, Laine, Elma e Gabriel, pela amizade.

Ao trio parada dura (swat), Edcarlos, Paulo e Ricardo (Bebê) pela amizade sincera e harmoniosa.

Ao casal Naje e Maílcia, pela amizade e carinho.

Aos professores do Departamento de Ciências Exatas, pelos ensinamentos prestados.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas, e todos os funcionários, em especial, a Josi, Joyce, Edila e Selminha.

A SEE/AC, por ter acreditado na realização deste trabalho.

Aos demais que, direta ou indiretamente, contribuíram para a elaboração deste trabalho.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Braga, Altemir da Silva. **Estimadores da probabilidade total de classificação incorreta na análise discriminante** 2008. 65 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, MG. *

A análise discriminante faz parte de um conjunto de técnicas de estatística multivariada e o seu princípio básico consiste em classificar novos indivíduos com várias características em uma de diferentes populações definidas *a priori*. Assim, diversos estimadores da probabilidade total de classificação incorreta paramétrica ($PTCI$) foram propostos utilizando algum método de amostragem e avaliados por meio de simulação Monte Carlo. Neste trabalho, foram comparados os desempenhos dos estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ para duas populações normais multivariadas homocedásticas, considerando-se os custos para cada classificação incorreta e probabilidades *a priori* iguais. O primeiro estimador é o de Lachenbrusch & Mickey (1968) que é baseado na metodologia *jackknife*, o segundo foi derivado utilizando desvio padrão comum no estimador de Lachenbruch & Mickey (1968). O terceiro e o quarto estimadores são parte da proposta do presente trabalho, em que modificou-se o método de Lachenbruch & Mickey (1968), combinando a função linear de Fisher com a metodologia *jackknife*. O quinto e o sexto estimadores foram derivados utilizando o mesmo raciocínio anterior, porém, fixando-se o vetor de combinações lineares $\mathbf{\Gamma}_1$ das variáveis e aplicando-se o *jackknife* para a constante da função linear de Fisher. Os desempenhos foram avaliados por meio dos vieses e dos erros quadráticos médios estimados. Dessa forma, o vetor de médias da população π_1 foi fixado em $\mathbf{0}$ ($\boldsymbol{\mu}_1 = \mathbf{0}$) e o vetor de médias $\boldsymbol{\mu}_2$ da população π_2 foi simulado em função da distância de Mahalanobis Δ^2. A busca de $\boldsymbol{\mu}_2$ aproximada para estabelecer o valor, fixado Δ^2, foi realizada por tentativa e erro. Os tamanhos amostrais da população π_1 foram $n_1 =$ 10, 50, 100 e os da população π_2 foram $n_2 =$ 10, 50, 100 combinados fatorialmente com $p = 2, 10$ variáveis e coeficiente de correlação $\rho = 0$, $\rho =$ 0,5 e $\rho =$ 0,9. Os estimadores $PTCI_5$ e $PTCI_6$ subestimaram a $PTCI$, enquanto que os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$ e $PTCI_4$ superestimaram. Os estimadores $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ foram mais eficientes do que o estimador $PTCI_1$, original de Lachenbruch & Mickey (1968). O estimador $PTCI_3$, que considerou desvios padrões heterogêneos foi determinado ótimo, porque apresentou menor viés positivo.

* **Comitê Orientador:** Daniel Furtado Ferreira (orientador) - UFLA e Patrícia de Siqueira Ramos (Co-orientadora).

ABSTRACT

Braga, Altemir da Silva. **Estimators of the overall misclassification probability in discriminant analysis.** 2008. 65 p. Dissertation (Master of Statistics and mixed-farming experimentation) Federal University of Lavras, Lavras, MG.*

Discriminant analysis is one of the multivariate statistics techniques, which idea consists in classifying new individuals in one of several populations known *a priori*. Thus, several estimators for the parametric overall of misclassification probability ($PTCI$) were proposed, using sampling methods and assessed through Monte Carlo simulation. In the present work, the performance of $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ and $PTCI_6$ methods was compared for two homocedastic multivariate normal populations, considering the same costs of misclassification and *a priori* probabilities. The first one is Lachenbruch & Mickey's method (1968), based on *Jackknife* methods, the second one was derived from Lachenbruch & Mickey's method (1968), using common variance into the function which estimates $PTCI$. Third and fourth methods were proposed in the present work, in which Lachenbruch & Mickey's method (1968) was been modified, associating Fisher's linear function with *Jackknife* methodology. Fifth and sixth methods were derived using the same previous reasoning, setting the linear combination vector $\mathbf{\Gamma}_1$ of the variates and applying the *Jackknife* for the constant of the Fisher's linear combination. The performance was assessed through bias and quadratic mean square estimator. Thus, the mean vector from population π_1 was set to $\mathbf{0}$ ($\boldsymbol{\mu}_1 = \mathbf{0}$). The approximate search of $\boldsymbol{\mu}_2$ from population π_2, for a settled value of the Mahalanobis distance Δ^2, was accomplished by trial and error. For population π_1, the sampling sizes were $n_1 =$ 10, 50, 100 and for π_2, $n_2 =$ 10, 50, 100 that were factorially combined with $p = 2, 10$ variates and correlation coefficient $\rho = 0$, $\rho = 0.5$ and $\rho = 0.9$. The estimators $PTCI_5$ and $PTCI_6$ underestimated $PTCI$, whereas $PTCI_1$, $PTCI_2$, $PTCI_3$ and $PTCI_4$ superestimated it. The $PTCI_3$, $PTCI_4$, $PTCI_5$ and $PTCI_6$ estimators were more efficient than $PTCI_1$ one, originally from Lachenbruch & Mickey's method (1968). The $PTCI_3$ estimator with heterogeneous variances was considered optimum, due to the smallest positive bias.

*Guindance Committee: Daniel Furtado Ferreira - UFLA (Adviser) and Patrícia de Siqueira Ramos (Co-adviser).

1 INTRODUÇÃO

A análise discriminante faz parte de um conjunto de métodos e instrumentos de estatística multivariada e o seu princípio básico consiste em discriminar ou classificar novas observações ou novos indivíduos em uma de diferentes populações definidas *a priori*. Tal conhecimento permite a elaboração de uma função matemática denominada na literatura por função discriminante ou regra de classificação que é utilizada para alocar os indivíduos. Essa função foi proposta por Fisher (1938) para classificar observações em uma de duas populações distintas sem assumir normalidade, porém considerando que as populações fossem homocedásticas.

A idéia dada por Fisher (1938) foi baseada em uma transformação linear que pudesse encontrar uma função de maior discriminação entre as populações. Assim, por exemplo, supondo-se que se tenha n_1 elementos amostrais procedentes, com probabilidade p_1, da população π_1 e n_2 elementos amostrais procedentes, com probabilidade p_2, da população π_2, e dos $n_1 + n_2 = n$ elementos amostrais tenham sido mensuradas *p*-variáveis aleatórias. Pode-se fazer uma análise estatística do comportamento das medidas das *p*-características para identificar o perfil geral de cada população. Assim, se houver uma nova observação amostral, não pertencente a nenhuma das duas amostrais anteriores, e cuja origem é duvidosa, seria possível compará-lo de algum modo com o perfil geral das populações π_1 e π_2 e classificá-lo como pertencente a população cujo perfil geral fosse mais semelhante ao dele. Após classificar os n elementos amostrais em uma das populações π_1 ou π_2 os elementos que foram classificados de forma incorreta servem para estimar a probabilidade total de classificação incorreta paramétrica (*PTCI*).

Os métodos utilizados para estimar a *PTCI* podem ser divididos em duas classes: aqueles que utilizam uma amostra para avaliar uma dada função discriminante chamados de estimadores teóricos e os que utilizam as propriedades da distribuição normal para a sua validação. Alguns desses estimadores podem ser encontrados em Giri (2004), Jonhson & Wichern (1998), Mingoti (2005), Ferreira (2008), entre outros autores. Em Ferreira (2008), por exemplo, pode-se encontrar o estimador da ressubstituição, o da ressubstituição com divisão amostral, o das probabilidades de classificação incorretas estimadas, o pseudo-*jackknife* e o segundo

estimador de Lachenbruch & Mickey (1968). Destes estimadores o da ressubstituição é o que tem pior desempenho e o de Lachenbruch & Mickey (1968) melhor desempenho, justificando o porquê de este estimador ter sido escolhido para a realização deste trabalho.

Giri (2004) sugeriu que se avaliasse esse estimador considerando desvio padrão comum, visto que no ensaio original, Lachenbruch & Mickey (1968) utilizaram desvios padrões heterogêneos no estimador da probabilidade total de classificação incorreta. Essa avaliação foi realizada por Oliveira & Ferreira (2008), sendo que o estimador com as modificações sugeridas, apresentou resultados piores. A motivação para a realização desse trabalho decorreu dessa proposta e da constatação de que esses estimadores apresentaram vieses, principalmente, se as populações avaliadas estivessem a uma distância de Mahalanobis menor do que 2.

Assim, conduziu-se este trabalho, com o objetivo de propor modificações no estimador da probabilidade total de classificação incorreta de Lachenbruch & Mickey (1968) e avaliar seus desempenhos por meio de simulação Monte Carlo, considerando duas populações normais multivariadas homocedásticas.

2 REFERENCIAL TEÓRICO

2.1 Função de densidade normal multivariada

O termo multivariada ou multidimensional é usado em virtude de, sob o ponto de vista matemático, cada variável poder ser vista como uma dimensão. Nos estudos em análise multivariada é importante a generalização da densidade normal univariada para mais de duas dimensões, decorrente do fato de muitas técnicas estatísticas multivariadas basearem-se na pressuposição de que os dados amostrais são obtidos de uma população com distribuição normal multivariada (Johnson & Wichern, 1998).

Analogamente ao caso univariado, todo vetor aleatório *p*-variado tem seus valores gerados por um mecanismo probabilístico (Anderson, 2003). Sabe-se que existem várias distribuições de probabilidades multivariadas, mas, sem dúvida, a mais importante é a normal *p*-variada (Johnson, 1987).

De acordo com Ferreira (2008), a distribuição normal multivariada é muito difícil de ser gerada. E pode ser mostrado que a função de densidade normal multivariada é mostrada como sendo uma generalização da distribuição normal univariada.

Para Rencher (2002), se $\boldsymbol{X}$ tem distribuição normal multivariada com vetor de média $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$, então, sua função densidade é dada por,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-1} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \tag{2.1}$$

em que *p* representa o número de variáveis.

2.2 Distância de Mahalanobis

Para o caso univariado, a distância euclidiana entre dois pontos P_1 e P_2 é simplesmente o módulo da diferença entre eles, que para estudos na área de estatística não é muito informativa (Rencher, 2002).

Um aspecto que é considerado importante das técnicas de análise multivariada diz respeito a muitos métodos de estimação e de decisão serem baseados no conceito de distância. Pode-se citar que o expoente da distribuição normal

multivariada é a distância quadrática entre a realização da variável aleatória e o centróide da distribuição.

O caso mais geral é quando consideram-se as variâncias e as covariâncias amostrais para estimar a distância entre os vetores aleatórios $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$. Neste caso, a métrica é definida por $\boldsymbol{S}^{-1}$ e a distância quadrática por:

$$d^2(\boldsymbol{x}_1,\boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2). \tag{2.2}$$

Essa distância quadrática é conhecida por distância generalizada de Mahalanobis (Ferreira, 2008).

2.3 Maximização de razão de formas quadráticas

Se **A** e **B** são matrizes simétricas $(p \times p)$ e **B** é positiva defindida, então o máximo de $\lambda(\boldsymbol{a}) = (\boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{a})/(\boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a})$ sob a restrição $(\boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a} = 1)$ é dado pelo maior autovalor (λ_i) de$\boldsymbol{S}_B^{-1}\boldsymbol{A}(\boldsymbol{S}_B^{-1})^\top$, (*i* = 1, . . . , *p)* e pelo autovetor correspondente $\boldsymbol{a}_i = (\boldsymbol{S}_B^{-1})^\top \boldsymbol{z}_i$, que é dado pela solução do sistema de equações homogêneas:

$$(\boldsymbol{A} - \boldsymbol{\lambda}_i \boldsymbol{B})\boldsymbol{a}_i = \boldsymbol{0}, \tag{2.3}$$

em que $\boldsymbol{S}_B$ é o fator de Cholesky de $\boldsymbol{B}$, $\boldsymbol{S}_B^{-1}$ é a sua matriz inversa, $\boldsymbol{z}_i$ é o i-ésimo autovetor de $\boldsymbol{S}_B^{-1}\boldsymbol{A}(\boldsymbol{S}_B^{-1})^\top$. Segue ainda, que toda matriz $\boldsymbol{B}$ pode ser fatorada por:

$$\boldsymbol{B} = \boldsymbol{S}\boldsymbol{S}^\top,$$

em que $\boldsymbol{S}$ é o fator de Cholesky da matriz $\boldsymbol{B}$, sendo não-singular e triangular inferior (Ferreira, 2008).

2.4 Erro quadrático médio (*EQM*)

O erro quadrático médio *EQM* de um estimador $\hat{\theta}$ do parâmetro θ é dado por:

$$EQM[\hat{\theta}] = E[(\hat{\theta} - \theta)^2], \tag{2.4}$$

desenvolvendo, tem-se que,

$$EQM[\hat{\theta}] = Var[\hat{\theta}] + [E(\hat{\theta}) - \theta]^2, \tag{2.5}$$

em que $E(\hat{\theta}) - \theta$ é denominado o viés do estimador $\hat{\theta}$. Diz-se que um estimador $\hat{\theta}$ é não viciado para θ se,

$$E[\hat{\theta}] = \theta, \tag{2.6}$$

para todo θ pertencente ao espaço paramétrico Θ. No caso em que $\hat{\theta}$ é um estimador não viciado para θ, tem -se que:

$$EQM[\hat{\theta}] = Var[\hat{\theta}], \tag{2.7}$$

ou seja, o erro quadrático médio do estimador $\hat{\theta}$ fica reduzido à variância do estimador (Bolfarine & Sandoval, 2001).

2.5 Simulação Monte Carlo

A simulação é usada para servir como uma primeira avaliação de um sistema para gerar novas estratégias de ação e regras de decisões antes de se correr o risco de experimentá-las no sistema real. Esse procedimento já era usado pelo homem em épocas remotas (Naylor et al. 1971).

A simulação Monte Carlo é utilizada com os recursos e as técnicas computacionais em que amostras são geradas de acordo com determinadas distribuições teóricas conhecidas, visando estudar o comportamento de diferentes técnicas estatísticas que poderiam ser empregadas num dado problema (Dachs, 1988).

O nome Monte Carlo está relacionado com a cidade de mesmo nome, no

Principado de Mônaco. O nome é originário, principalmente, em razão dos jogos de azar, decorrentes da roleta, que é um mecanismo simples para gerar números aleatórios. Esses números eram gerados manualmente ou mecanicamente. Modernamente são usados computadores para gerá-los. Esses números na verdade são números pseudo-aleatórios (Morettin e Bussab, 2003).

2.6 Estimação das funções discriminantes

Fisher (1938) propôs uma função discriminante para classificar uma observação $\boldsymbol{x}$ em uma de duas populações distintas considerando apenas que a pressuposição de homocedásticidade fosse satisfeita. Seu procedimento foi baseado na transformação linear do vetor de observações multivariadas $\boldsymbol{x}$ em uma observação univariada de tal forma que houvesse uma máxima separação entre as duas populações.

Welch (1939)também propôs um critério para determinar uma função discriminante supondo populações normais. Esta função classifica $\boldsymbol{x}$ na população π_1 se:

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > k,$$

e na população π_2 caso contrário, sendo $f_1(\boldsymbol{x})$ e $f_2(\boldsymbol{x})$ definidas pela função dada pela expressão 2.1. A escolha de *k* dependeria dos custos relativos a cada classificação incorreta.

Para Maroco (2007), baseado nas transformações lineares definidas por Fisher (1938), uma função discriminante pode ser definida da seguinte forma: dadas *p*-variáveis e *k* populações é possível estabelecer uma quantidade *m* = min(*k*-1, *p*) de funções discriminantes que são combinação linear das *p*-variáveis:

$$\boldsymbol{\xi}_i = w_{i1}\boldsymbol{X}_1 + w_{i2}\boldsymbol{X}_2 + \ldots + w_{ip}\boldsymbol{X}_p \;\; (i = 1, \ldots, m),$$

em que os pesos $w_{i1}, w_{i2}, \ldots, w_{ip}$ são estimados de modo que a variabilidade dos escores da função discriminante seja máxima entre as populações e mínima dentro,

isto é, que a razão:

$$\lambda_i = \frac{SQB(\boldsymbol{\xi}_i)}{SQW(\boldsymbol{\xi}_i)},$$

seja máxima; em que $SQB(\boldsymbol{\xi}_i)$ é uma forma quadrática que representa a soma de quadrados entre as populações e $SQW(\boldsymbol{\xi}_i)$ é a forma quadrática que representa a soma de quadrados dentro das populações.

Após a dedução da primeira função discriminante, os pesos das funções seguintes são obtidos sobre a restrição adicional de que os escores das funções discriminantes não estejam correlacionados, isto é, que $Cov(\boldsymbol{\xi}_i,\boldsymbol{\xi}_j) = 0$. Sendo $\boldsymbol{T} = \boldsymbol{A} + \boldsymbol{B}$ em que $\boldsymbol{A}$ e $\boldsymbol{B}$ são repectivamente as matrizes de somas e produtos dos quadrados entre e dentro das populações e $\boldsymbol{T}$ é a matriz de somas e produtos dos quadrados totais (Sharma, 1996). Assim, em termos matriciais, a função discriminante é dada por:

$$\boldsymbol{\xi}_i = {}_{\boldsymbol{n}}\boldsymbol{X}_{\boldsymbol{p}}\boldsymbol{a}_i,$$

em que ${}_{\boldsymbol{n}}\boldsymbol{X}_{\boldsymbol{p}}$ $(p \times p)$ é a matriz com as *p*-variáveis e $\boldsymbol{a}_i$ é o vetor dos pesos. A soma dos totais para os escores de $\boldsymbol{\xi}$ é dado por $\boldsymbol{\xi}^\top\boldsymbol{\xi} = (\boldsymbol{X}\boldsymbol{a})^\top(\boldsymbol{X}^\top\boldsymbol{a})$, então $\boldsymbol{\xi}^\top\boldsymbol{\xi} = \boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a}$ (Sharma, 1996). Dessa forma, tem-se que $\boldsymbol{\xi}^\top\boldsymbol{\xi} = \boldsymbol{a}^\top\boldsymbol{T}\boldsymbol{a}$, logo, $\boldsymbol{\xi}^\top\boldsymbol{\xi} = \boldsymbol{a}^\top(A + B)\boldsymbol{a} = \boldsymbol{a}^\top\boldsymbol{A}\boldsymbol{a} + \boldsymbol{a}^\top\boldsymbol{B}\boldsymbol{a}$. Em que, $\boldsymbol{a}^\top\boldsymbol{A}\boldsymbol{a}$ e $\boldsymbol{a}^\top\boldsymbol{B}\boldsymbol{a}$ são respectivamente, a soma dos quadrados e produtos entre e dentro das populações para a função discriminante $\boldsymbol{\xi}$, a obtenção da função discriminante resume-se a encontrar o vetor $\boldsymbol{a}$ tal que,

$$\lambda = \frac{\boldsymbol{a}^\top\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^\top\boldsymbol{B}\boldsymbol{a}},$$

seja máxima.

A solução desse problema de maximização foi definida na seção 2.3, ou seja, é a solução do sistema de equações homogêneas,

$$(\boldsymbol{A} - \boldsymbol{\lambda}_i\boldsymbol{B})\boldsymbol{a}_i = \boldsymbol{0}.$$

Esse problema tem *m* = min(*k*-1, *p*) soluções correspondentes aos autova-

lores da matriz $\boldsymbol{S}_B^{-1}\boldsymbol{A}(\boldsymbol{S}_B^{-1})^\top$. O maior autovalor (λ_1) corresponde à primeira função discriminante, o segundo autovalor (λ_2) corresponde à segunda função discriminante, e assim por diante.

A Figura 2.1 representa a idéia do procedimento de Fisher (1938) em utilizar uma transformação de *p*-dimensões em uma única dimensão para o caso de $k = 2$ populações homocedásticas. Neste procedimento, Fisher (1938) definiu uma combinação linear que maximizaria a distância estatística quadrática entre as médias dos escores. A figura apresenta uma região de interseção entre as populações π_1 e π_2 que representa a probabilidade total de classificação incorreta paramétrica e duas projeções de escores uma sobre o eixo $\boldsymbol{Z}$ e outra sobre o eixo $\boldsymbol{Z}'$ que representam as transformações lineares de Fisher (1938).

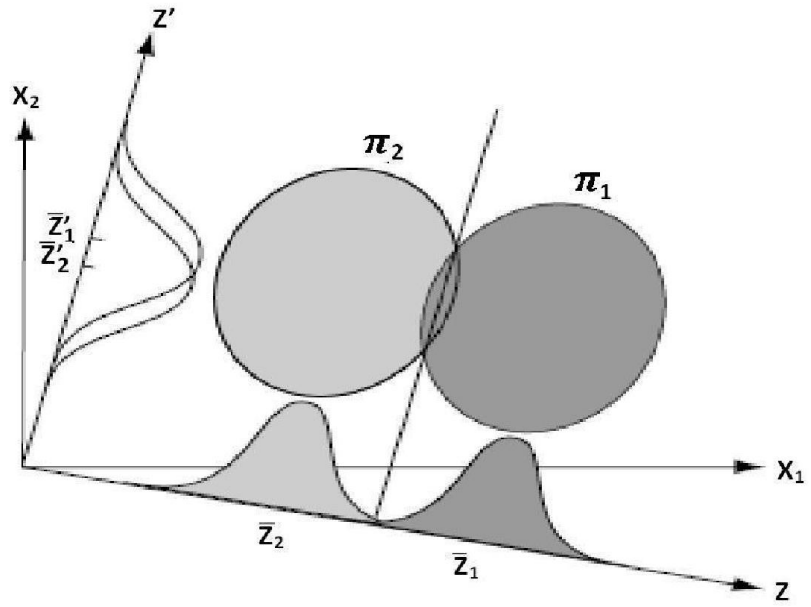

FIGURA 2.1: Representação da região de interseção entre as duas populações homocedásticas π_1 e π_2 e das transformações lineares de Fisher (1938).

2.7 Classificação em uma de duas populações normais

Härdle & Simar (2007) definem o custo esperado para cada classificação incorreta pela expressão:

$$CEI = p_1 C(2|1) p_{21} + p_2 C(1|2) p_{12}, \tag{2.8}$$

em que:

- $C(2|1)$ é custo incidente quando se classifica o elemento amostral como sendo da população π_2, quando na realidade, ele pertence à população π_1;
- $C(1|2)$ é custo incidente quando se classifica o elemento amostral como sendo da população π_1, quando na realidade, ele pertence à população π_2;
- p_{21} é a probabilidade de classificar um elemento na população π_2 dado que o elemento é da população π_1;
- p_{12} é a probabilidade de classificar um elemento na população π_1 dado que o elemento é da população π_2;
- CEI é o custo esperados para cada classificação incorreta;
- p_1 é a probabilidade *a priori* da população π_1;
- p_2 é a probabilidade *a priori* da população π_2;

De acordo com Mingoti (2005) as probabilidades p_{21} e p_{12} podem ser definidas, respectivamente, por:

- Erro 1: o elemento amostral pertence à população π_1, mas a função discriminante o classifica como sendo proveniente da população π_2;
- Erro 2: o elemento amostral pertence à população π_2, mas a função discriminante o classifica como sendo proveniente da população π_1;

As probabilidades de ocorrência destes erros são dados, respectivamente, por

$$\text{Prob(Erro1)} = p_{21} \qquad \text{e} \qquad \text{Prob(Erro2)} = p_{12}.$$

De acordo Johnson & Wichern (1998), a função discriminante que minimiza CEI para duas populações normais é dada por:

$$R_1 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant \left[\frac{C(1|2)}{C(2|1)} \right] \left(\frac{p_2}{p_1} \right) \right\} \tag{2.9}$$

$$R_2 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left[\frac{C(1|2)}{C(2|1)} \right] \left(\frac{p_2}{p_1} \right) \right\}, \tag{2.10}$$

em que R_1 e R_2 são as regiões que minimizam o custo esperado para cada observação classificada incorretamente.

Assim, de acordo com a desigualdade (2.9), deve-se classificar a observação $\boldsymbol{x}$ na região R_1 se :

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant \left[\frac{C(1|2)}{C(2|1)} \right] \left(\frac{p_2}{p_1} \right). \tag{2.11}$$

Se substituir as densidades $f_1(\boldsymbol{x})$ e $f_2(\boldsymbol{x})$, definidas em (2.1) pela densidade normal multivariada correspondente tem-se:

$$\frac{(2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_1})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu_1}) \right\}}{(2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_2})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu_2}) \right\}} \geqslant \left[\frac{C(1|2)}{C(2|1)} \right] \left(\frac{p_2}{p_1} \right).$$

Assim classifica-se $\boldsymbol{x}$ em R_1 se:

$$(\boldsymbol{\mu}_1 \text{ - } \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} (\boldsymbol{\mu}_1 \text{ - } \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geqslant \ln \left\{ \left[\frac{\mathrm{C(1|2)}}{\mathrm{C(2|1)}} \right] \left(\frac{p_2}{p_1} \right) \right\}$$

e em R_2, caso contrário.

Para simplificar a expressão denominou-se o vetor $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ por $\boldsymbol{a}$ assim a região R_1 ficou definida por:

$$R_1 : \left\{ \boldsymbol{x} | a^\top \boldsymbol{x} - \frac{1}{2} \left(\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2 \right) \geqslant \ln \left\{ \left[\frac{C(2|1)}{C(1|2)} \right] \left(\frac{p_2}{p_1} \right) \right\} \right\}, \tag{2.12}$$

e a região R_2 por,

$$R_2 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2} \left(\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2 \right) < ln \left\{ \left[\frac{C(2|1)}{C(1|2)} \right] \left(\frac{p_2}{p_1} \right) \right\} \right\}, \quad (2.13)$$

em que R_1 e R_2 são as regiões que minimizam o custo esperado para cada observação classificação incorretamente.

2.7.1 Casos especiais para o custo mínimo esperado

Pode-se simplificar as desigualdades definidas em (2.12) e (2.13) considerando algumas restrições para os custos e para as probabilidades *a priori* da população π_i. Assim:

a) probabilidades *a priori* iguais, ou seja, $\frac{p_1}{p_2} = 1$:

$$R_1 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2} (\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2) \geqslant \ln \left[\frac{C(2|1)}{C(1|2)} \right] \right\}$$

$$R_2 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2} (\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2) < \ln \left[\frac{C(2|1)}{C(1|2)} \right] \right\}.$$

b) custos de classificações incorretas iguais, ou seja, $\frac{C(2|1)}{C(1|2)} = 1$:

$$R_1 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2} \left(\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2 \right) \geqslant \ln \left(\frac{\pi_2}{\pi_1} \right) \right\}$$

$$R_2 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2} \left(\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2 \right) < \ln \left(\frac{\pi_2}{\pi_1} \right) \right\}.$$

c) probabilidades *a priori* iguais e custos de classificações incorretas iguais, ou seja, $\frac{p_1}{p_2} = 1$ e $\frac{C(2|1)}{C(1|2)} = 1$

$$R_1 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2} (\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2) \geqslant 0 \right\} \quad (2.14)$$

$$R_2 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2}(\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2) < 0 \right\}, \tag{2.15}$$

nota-se que se $R_1 > 0$, o valor de $f_1(\boldsymbol{x})$ é maior do que o valor de $f_2(\boldsymbol{x})$, assim, pelo princípio da máxima verossimilhança é razoável classificar o elemento $\boldsymbol{x}$, que gerou $R_1 > 0$, como um provável elemento da população π_1. Por outro lado, se $R_1 < 0$, é razoável classificar o elemento $\boldsymbol{x}$ na população π_2.

Wald (1944) e Anderson (2003), sugeriram substituir os parâmetros desconhecidos pelos seus estimadores. As estimativas são obtidas de uma amostra de treinamento. Sejam n_1 observações p-variadas $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ amostradas da população π_1 e n_2 observações $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ amostradas da população π_2 com $n_1 + n_2 - 2 \geqslant p$, então:

$$\bar{\boldsymbol{X}}_1 = \frac{\sum_{j=1}^{n} \boldsymbol{X}_{1j}}{n_1} \quad \text{e} \quad \boldsymbol{S}_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n} (\boldsymbol{X}_{1j} - \bar{\boldsymbol{X}}_1)(\boldsymbol{X}_{1j} - \bar{\boldsymbol{X}}_1)^\top; \tag{2.16}$$

$$\bar{\boldsymbol{X}}_2 = \frac{\sum_{j=1}^{n} \boldsymbol{X}_{2j}}{n_2} \quad \text{e} \quad \boldsymbol{S}_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\boldsymbol{X}_{2j} - \bar{\boldsymbol{X}}_2)(\boldsymbol{X}_{2j} - \bar{\boldsymbol{X}}_2)^\top. \tag{2.17}$$

O melhor estimador não viesado da matriz de covariância comum ($\boldsymbol{\Sigma}$) de ambas as populações é dado por:

$$\boldsymbol{S}_p = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2}. \tag{2.18}$$

Assim, a função discriminante estimada pode ser obtida substituindo $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}$ por $\bar{\boldsymbol{X}}_i$ e $\boldsymbol{S}_p$ nas inequações definidas em (2.14) e (2.15) quando se tem os custos de classificação incorreta e as probabilidades *a priori* são iguais . Então deve-se alocar $\boldsymbol{x}$ na população π_1 se:

$$R_1 : \left\{ \boldsymbol{x} | \hat{\boldsymbol{a}}^\top \boldsymbol{x} - \frac{1}{2}(\hat{\boldsymbol{a}}^\top \bar{\boldsymbol{X}}_1 + \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{X}}_2) \geqslant 0 \right\} \tag{2.19}$$

e na população π_2 ,

$$R_2 : \left\{ \boldsymbol{x} | \hat{\boldsymbol{a}}^\top \boldsymbol{x} - \frac{1}{2}(\hat{\boldsymbol{a}}^\top \bar{\boldsymbol{X}}_1 + \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{X}}_2) < 0 \right\}, \tag{2.20}$$

em que $\hat{\boldsymbol{a}} = \boldsymbol{S}_p^{-1}(\bar{X}_1 - \bar{X}_2)$.

2.8 Avaliação de uma função discriminante

Para Mingoti (2005), após a construção de uma função discriminante é necessário avaliar a sua qualidade. Assim, para cada elemento amostral das populações π_1 e π_2, calcula-se o escore numérico da função discriminante construída, e a análise destes permitirá que se faça uma avaliação da qualidade da função em termos de erros de classificação e capacidade de discriminação.

Ferreira (2008) menciona que o desempenho de uma função discriminante pode ser mensurado pelas taxas de erros ou probabilidades de classificação incorretas. Essa probabilidade de classificação incorreta é dada pela soma das taxas de classificação incorreta ponderadas pelas probabilidades *a priori* de cada população e pode ser obtida por:

$$PTCI = p_1 P(2|1, \xi) + p_2 P(1|2, \xi)$$

$$PTCI = p_1 \int_{R_1} f_1(\boldsymbol{x}) dx + p_2 \int_{R_2} f_2(\boldsymbol{x}) dx, \tag{2.21}$$

em que R_1, R_2 e ξ são, respectivamente, as regiões 1, 2 e a função discriminante que minimiza o custo esperado para cada classificação incorreta.

Deve-se determinar uma função discriminante ξ de maneira que se tenha uma região mínima de classificação incorreta. Para o caso particular de duas densidades normais multivariadas conhecidas e homocedásticas, com $\frac{p_1}{p_2} = 1$ e $\frac{C(2|1)}{C(1|2)} = 1$, determina-se ótimas funções de classificação, como definido em (2.14) e (2.15) dadas pelas regiões:

$$R_1 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2}(\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2) \geqslant 0 \right\}$$

$$R_2 : \left\{ \boldsymbol{x} | \boldsymbol{a}^\top \boldsymbol{x} - \frac{1}{2}(\boldsymbol{a}^\top \boldsymbol{\mu}_1 + \boldsymbol{a}^\top \boldsymbol{\mu}_2) < 0 \right\},$$

sendo $\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Considerando $y = \boldsymbol{a}^\top \boldsymbol{x}$, tem-se, conforme Ferreira (2008), que y tem distribuição normal multivariada. Então, a esperança condicional de Y é dada por:

$$E(Y|\pi_i) = \boldsymbol{a}^\top E(\boldsymbol{X}|\pi_i) = \boldsymbol{a}^\top \boldsymbol{\mu}_i = \mu_{iY}$$

para *i* = 1, 2 e a variância de Y por:

$$Var(Y|\pi_i) = \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}.$$

Mas $\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, então,

$$Var(Y|\pi_i) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$Var(Y|\pi_i) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Assim, a variância de Y é a distância de Mahalanobis entre os centróides das duas populações normais multivariadas, que será representada por Δ^2. Dessa forma, se $\boldsymbol{x}$ pertencer a π_1, então, a distribuição de Y será normal univariada com média $\boldsymbol{\mu}_{1Y} = \boldsymbol{a}^\top \boldsymbol{\mu}_1$ e variância Δ^2 e será denominada por $f_1(y)$. Por outro lado, se $\boldsymbol{x}$ pertencer a π_2, então, a distribuição de Y será normal univariada com média $\boldsymbol{\mu}_{2Y} = \boldsymbol{a}^\top \boldsymbol{\mu}_2$ e variância Δ^2 e será denominada por $f_2(y)$. Assim, voltando-se a expressão (2.21) tem-se que a probabilidade total de classificação incorreta ($PTCI$) será dada por:

$$PTCI = p_1 \int_{-\infty}^{0{,}5\boldsymbol{a}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)} f_1(\boldsymbol{y})dy + p_1 \int_{0{,}5\boldsymbol{a}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)}^{\infty} f_2(\boldsymbol{y})dy.$$

Sendo as densidades $f_1(y)$ e $f_2(y)$ normais univariadas. Assim, pode-se representar a probabilidade total de classificação incorreta, conforme apresentado na Figura 2.2.

Como $\Delta^2 > 0$ pode-se mostrar que $\mu_{1y} > \mu_{2y}$. Então, calcula-se a probabilidade total de classificação incorreta (*PTCI*), padronizando o limite de integração $y_i = 0{,}5\boldsymbol{a}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$, considerando as densidades condicionais da variável normal Y. Assim, os limites padronizados são:

$$z_1 = \frac{0{,}5\boldsymbol{a}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \mu_{1Y}}{\boldsymbol{\Delta}}$$

$$z_1 = \frac{0{,}5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}{\Delta}$$

$$z_1 = -\frac{\Delta}{2}.$$

De maneira análoga, tem-se:

$$z_2 = \frac{0{,}5\boldsymbol{a}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \mu_{2Y}}{\boldsymbol{\Delta}}$$

$$z_2 = \frac{0{,}5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2}{\Delta}$$

$$z_2 = \frac{\Delta}{2}.$$

Portanto, a probabilidade total de classificação incorreta paramétrica pode ser de-

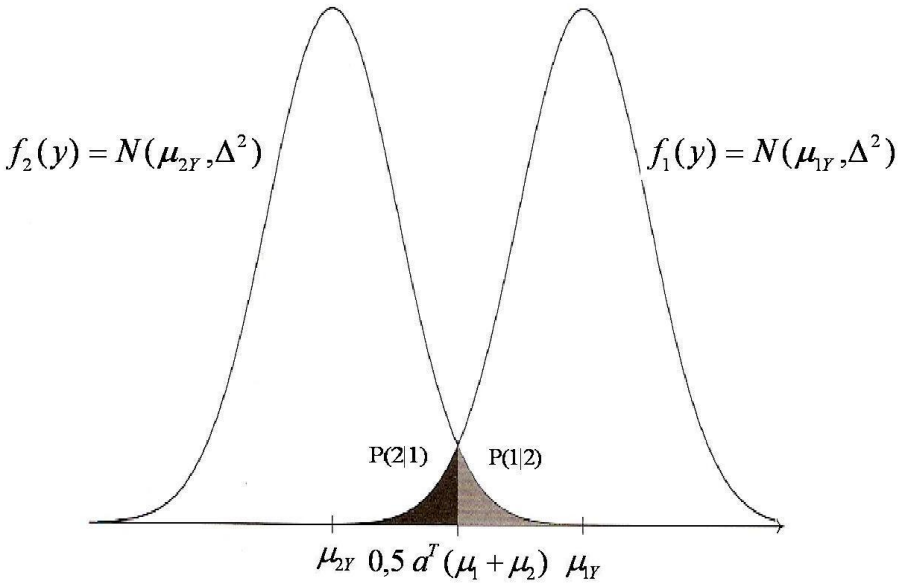

FIGURA 2.2: Probabilidades de classificação incorreta de uma observação $\boldsymbol{x}$ em duas populações normais multivariadas, baseadas no ponto de corte $0{,}5\boldsymbol{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, em $p_1 = p_2 = 1/2$ e na transformação linear de $\boldsymbol{X}$ para uma variável univariada, com distribuição condicional normal.

terminada da seguinte forma:

$$PTCI = \frac{1}{2}\int_{-\infty}^{-0{,}5\Delta} \phi(z)dz + \frac{1}{2}\int_{0{,}5\Delta}^{\infty} f_2(\boldsymbol{x})dy$$

$$PTCI = \frac{1}{2}\Phi(-0{,}5\Delta) + \frac{1}{2}(1 - \Phi(0{,}5\Delta)$$

$$PTCI = \Phi(-0{,}5\Delta), \tag{2.22}$$

em que $\phi(z)$ e $\Phi(z)$ são as funções densidade e distribuição da normal padrão e

são dadas, respectivamente, por:

$$\phi(z) = (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{z^2}{2}\right\} \tag{2.23}$$

$$\Phi(z) = \int_{-\infty}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{t^2}{2}\right\} dt. \tag{2.24}$$

2.8.1 Estimador da ressubstituição

Este estimador foi sugerido por Smith (1947), e a idéia consiste em utilizar duas amostras aleatórias das populações normais multivariadas π_1 e π_2, ou seja, $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ são observações *p*-variadas amostradas da população π_1 e $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ são observações *p*-variadas amostradas da população π_2, com $n_1 + n_2 - 2 \geq p$ para estimar os parâmetros e a função discriminante. Depois, as mesmas observações são utilizadas para estimar a probabilidade total de classificação incorreta paramétrica . O resultado é resumido numa matriz denominada de matriz de confusão, apresentada a seguir:

	População Classificada		
População real	π_1	π_2	Total
π_1	n_{11}	n_{12}	n_1
π_2	n_{21}	n_{22}	n_2
			$n = n_1 + n_2$

Nesta matriz n_i representa o tamanho da amostra obtida na *i*-ésima população, $n_{ij} \neq j$, representa o número de observações da *i*-ésima população classificadas de forma incorreta na *j*-ésima população e n_{ii}, é o número de observações da *i*-ésima população classificadas corretamente na própria população, sendo $i = 1, 2$. A proporção de observações classificadas de forma incorreta representa a probabilidade total classificação incorreta estimada.

Johnson & Wichern (1998) mencionam que esse tipo de medida não depende da forma da distribuição das populações amostradas e a probabilidade total de classificação incorreta que pode ser calculada para qualquer função de classificação.

2.8.2 Estimador de Holdout

Nesse estimador, a amostra conjunta de $n = n_1 + n_2$ elementos é repartida em duas partes, uma que vai servir para a construção da função discriminante (amostra de treinamento) e outra que vai ser utilizada para a estimação das probabilidades de classificação incorretas (amostra de validação). Inicialmente, selecionam-se aleatoriamente alguns indivíduos das amostras das populações π_1 e π_2, deixando-os à parte da amostra original de $n = n_1 + n_2$ elementos. Para cada um destes elementos, sabe-se de qual população ele é proveniente e, portanto, o conjunto de pontos omitidos servirá para testar a função discriminante construída. A função discriminante estimada é utilizada para classificar os elementos que foram colocados à parte inicialmente, e as proporções de classificações incorretas são calculadas da mesma forma como descrita da ressubstituição (Mingoti, 2005).

2.8.3 Estimador pseudo-*Jackknife*

De acordo com Mingoti (2005), este método é também conhecido como validação cruzada. A validação consiste nos seguintes passos:

- passo 1: retira-se um vetor de observações de amostra conjunta e utilizam-se os $n_1 + n_2 - 1$ elementos amostrais restantes para construir a função de discriminação;

- passo 2: utiliza-se a função discriminante construída no passo 1 para classificar o elemento que ficou à parte da construção da função discriminante, determina-se se houve acerto ou não nessa população;

- passo 3: retorna-se o elemento amostral que foi retirado no passo 1 à amostra original e retira-se um outro elemento amostral diferente do primeiro. Os passos 1 e 2 são repetidos.

Os passos 1 e 2 devem ser repetidos para todos os $n = n_1 + n_2$ elementos da amostra conjunta. O total de erros em relação ao tamanho amostral n é uma estimativa da probabilidade total de classificação incorreta paramétrica.

2.8.4 Estimador das taxas de erros estimadas

Conforme Ferreira (2008), para se utilizar este estimador supõe-se que os parâmetros populacionais sejam conhecidos, e que a pressuposição de normalidade multivariada seja verificada. Assim, pode-se determinar a probabilidade total de classificação incorreta paramétrica (*PTCI*) pela expressão definida em (2.22). Dessa forma, estimam-se as probabilidades substituindo-se Δ pelo seu estimador dado por:

$$\hat{\Delta} = \sqrt{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \boldsymbol{S}_p^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)}.$$

A partir da esperança da distribuição F não-central e da relação da distribuição $\hat{\Delta}^2$ com T^2 de Hotelling não-central sob normalidade dada por:

$$\frac{n_1 n_2}{n_1 + n_2} \hat{\Delta}^2 \; \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - 1 - p)} F(p, n_1 + n_2 - 1 - p, \delta^2),$$

sendo $\delta^2 = n_1 n_2 / (n_1 + n_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ o parâmetro de não-centralidade, então pode-se mostrar que a esperança matemática de $\hat{\Delta}^2$ é

$$\frac{n_1 n_2}{n_1 + n_2} E(\hat{\Delta}^2) = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - 1 - p)} E[F(p, n_1 + n_2 - 1 - p, \delta^2)]$$

$$E(\hat{\Delta}^2) = \frac{n_1 + n_2 - 2}{n_1 + n_2 - p + 3} \left(\Delta^2 + \frac{p(n_1 + n_2)}{n_1 n_2} \right).$$

Assim pode-se estimar Δ^2 pelo estimador não-viesado dado por:

$$\tilde{\Delta^2} = \frac{n_1 + n_2 - 3 - p}{n_1 + n_2 - 2} \hat{\Delta}^2 - \frac{p(n_1 + n_2)}{n_1 . n_2}. \tag{2.25}$$

A probabilidade total de classificação incorreta estimada é dada por:

$$ETEE = \phi \left(-\frac{\tilde{\Delta}}{2} \right), \tag{2.26}$$

em que *ETEE* representa o estimador das taxas de erros estimadas.

2.8.5 Segundo estimador de Lachenbruch & Mickey (1968)

De acordo com Ferreira (2008), este estimador combina a técnica de amostragem *jackknife* e o estimador das taxas de erros estimadas. A idéia deste método é omitir das $n_1 + n_2$ observações amostrais a $\boldsymbol{x}_{ij}$ da *i*-ésima população referente a *j*-ésima unidade amostral, sendo *i* = 1, 2 e *j* = 1,..., n .

Sejam duas amostras aleatórias normais multivariadas, sabendo que a amostra $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ pertence à população π_1 e $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ à população π_2, com $n_1 + n_2 - 2 \geq p$. Deve-se estimar as médias $(\bar{\boldsymbol{X}}_1^{-(ij)}, \bar{\boldsymbol{X}}_2^{-(ij)})$ das amostras das populações π_1 e π_2 e a matriz de covariância comum $(\boldsymbol{S}_p^*)$, excluindo a observação $\boldsymbol{x}_{ij}$, utilizando-se as expressões:

$$\bar{\boldsymbol{X}}_1 = \frac{\sum_{j=1}^{n_1^*} \boldsymbol{X}_{1j}}{n_1^*} \quad \text{e} \quad \boldsymbol{S}_1 = \frac{1}{n_1^* - 1} \sum_{j=1}^{n_1^*} (\boldsymbol{X}_{1j} - \bar{\boldsymbol{X}}_1)(\boldsymbol{X}_{1j} - \bar{\boldsymbol{X}}_1)^\top; \quad (2.27)$$

$$\bar{\boldsymbol{X}}_2 = \frac{\sum_{j=1}^{n_2^*} \boldsymbol{X}_{2j}}{n_2^*} \quad \text{e} \quad \boldsymbol{S}_2 = \frac{1}{n_2^* - 1} \sum_{j=1}^{n_2^*} (\boldsymbol{X}_{2j} - \bar{\boldsymbol{X}}_2)(\boldsymbol{X}_{2j} - \bar{\boldsymbol{X}}_2)^\top, \quad (2.28)$$

em que n_1^* e n_2^* ora representarão os tamanhos originais, se não houver perda da observação na amostra específica e ora representarão os tamanhos amostrais descontados de 1, correspondente a observação ignorada.

O estimador comum $(\boldsymbol{S}_p^*)$ não viesado da matriz de covariância comum $(\boldsymbol{\Sigma})$ é dado por:

$$\boldsymbol{S}_p^* = \frac{(n_1^* - 1)\boldsymbol{S}_1 + (n_2^* - 1)\boldsymbol{S}_2}{n_1^* + n_2^* - 2}. \quad (2.29)$$

Para a observação omitida calcula-se o valor y_{ij} por:

$$y_{ij} = \boldsymbol{\Gamma}\boldsymbol{x}_{ij} - \frac{1}{2}(\boldsymbol{\Gamma}\bar{\boldsymbol{X}}_1^{-(ij)} + \boldsymbol{\Gamma}\bar{\boldsymbol{X}}_2^{-(ij)}), \quad (2.30)$$

em que $\boldsymbol{\Gamma} = (\bar{\boldsymbol{X}}_1^{-(ij)} - \bar{\boldsymbol{X}}_2^{-(ij)})^\top \boldsymbol{S}_p^{*-1}$.

Repete-se o processo para todos os valores de *i* e *j*, omitindo somente a observação $\boldsymbol{x}_{ij}$ em cada etapa e determina-se y_{ij} pela expressão (2.30). Assim,

tem-se uma amostra realizada $y_{11}, y_{12}, \ldots, y_{1n_1}$ da população π_1 e uma outra $y_{21}, y_{22}, \ldots, y_{2n_2}$ da população π_2.

Calcular as médias e as variâncias de cada amostra por:

$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \tag{2.31}$$

$$S_i^2 = \frac{1}{n_i - 1} \left[\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{n_i} y_{ij} \right)^2}{n_i} \right] \tag{2.32}$$

para *i* = 1, 2.

Lachenbruch & Mickey (1968) propuseram estimar a probabilidade total de classificação incorreta por:

$$PTCI_1 = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_1}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_2}\right), \tag{2.33}$$

em que $PTCI_1$ representa a probabilidade total de classificação incorreta do estimador de Lachenbruch & Mickey (1968).

3 METODOLOGIA

Neste estudo foram comparados, por meio de simulação Monte Carlo, os desempenhos de seis estimadores utilizados para estimar a probabilidade total de classificação incorreta. O primeiro estimador de Lachenbruch & Mickey (1968) é baseado na metodologia *jackknife*, conforme descrito em na seção 2.8.5 e o segundo foi derivado do estimador de Lachenbruch & Mickey (1968), utilizando-se o desvio padrão comum na função que estima a $PTCI_1$. O terceiro e o quarto estimadores são parte da proposta do presente trabalho, em que modificou-se o estimador de Lachenbruch & Mickey (1968), combinando a função linear de Fisher com a metodologia *jackknife*, assim, fixou-se a constante da função linear de Fisher que foi denominada por Γ_2 e fez-se o *jackknife* para estimar o vetor de combinações lineares das variáveis $\mathbf{\Gamma}_1$. O quinto e o sexto estimadores foram derivados, utilizando o mesmo raciocínio anterior, porém, fixando-se o vetor de combinações lineares das variáveis $\mathbf{\Gamma}_1$ e aplicando-se o *jackknife* para a constante Γ_2. O desempenho foi mensurado pela qualidade dos estimadores da probabilidade total de classificação incorreta, avaliando-se os vieses e os erros quadráticos médios (*EQM*s).

3.1 Estimador de Lachenbruch & Mickey (1968)

Sejam duas amostras normais p-variadas dadas por $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ da população π_1 e $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ da população π_2, com $n_1+n_2-2 \geq p$, estimaram-se as médias $(\bar{\boldsymbol{X}}_1^{-(ij)}, \bar{\boldsymbol{X}}_2^{-(ij)})$ das amostras das populações π_1 e π_2 e a matriz de covariância comum $(\boldsymbol{S}_p^*)$, utilizando-se as expressões definidas em (2.27), (2.28) e (2.29), e excluindo a observação $\boldsymbol{x}_{ij}$, calculou-se o valor y_{ij} por:

$$y_{ij} = \mathbf{\Gamma}\boldsymbol{x}_{ij} - \frac{1}{2}(\mathbf{\Gamma}\bar{\boldsymbol{X}}_1^{-(ij)} + \mathbf{\Gamma}\bar{\boldsymbol{X}}_2^{-(ij)}),$$

em que $\mathbf{\Gamma} = (\bar{\boldsymbol{X}}_1^{-(ij)} - \bar{\boldsymbol{X}}_2^{-(ij)})^\top \boldsymbol{S}_p^{*-1}$.

Repetiu-se o processo para a obtenção de y_{ij}, considerando a combinação de todos os valores de *i* e *j*, omitindo-se a observação $\boldsymbol{x}_{ij}$ selecionado em cada etapa. Assim, obteve-se uma amostra $y_{11}, y_{12}, \ldots, y_{1n_1}$ realizada da população π_1 e outra $y_{21}, y_{22}, \ldots, y_{2n_2}$ da população π_2. Em seguida, calcularam-se as médias

e as variâncias da *i*-ésima amostra por :

$$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad e \quad S_i^2 = \frac{1}{n_i - 1}\left[\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{n_i} y_{ij}\right)^2}{n_i}\right]. \tag{3.1}$$

A probabilidade total de classificação incorreta foi estimada da seguinte forma:

$$PTCI_1 = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_1}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_2}\right), \tag{3.2}$$

este estimador foi denominado de método 1 heterogêneo (M1HE), para indicar que foi feito *jackknife* tanto na combinação linear envolvendo a observação $\boldsymbol{x}_{ij}$, quanto na parte constante e que os estimadores S_1 e S_2 da variável nas amostras 1 e 2 não foram combinadas em um único estimador.

A segunda alternativa foi estimar a probabilidade total de classificação incorreta considerando desvio padrão comum para as duas populações homogêneas, conforme sugerido por Giri (2004). O estimador dessa probabilidade foi dado por:

$$PTCI_2 = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_p}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_p}\right), \tag{3.3}$$

em que o desvio padrão comum S_p é definido por:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}, \tag{3.4}$$

este segundo estimador foi denominado de método 2 homogêneo (M2HO).

3.2 Estimadores propostos

Os quatro estimadores propostos a seguir diferiram do estimador de Lachenbruch & Mickey (1968) na forma que foi idealizado o *jackknife* para estimar a probabilidade total de classificação incorreta.

Considerando-se duas amostras aleatórias normais multivariadas, sendo que

$\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ pertencem à população π_1 e $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ à população π_2, com $n_1 + n_2 - 2 \geq p$. Estimaram-se as médias das amostras das populações π_1 e π_2 e a matriz de covariância comum e calculou-se a constante Γ_2 por:

$$\Gamma_2 = \frac{1}{2}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \boldsymbol{S}_p^{-1}(\bar{\boldsymbol{X}}_1 + \bar{\boldsymbol{X}}_2) \tag{3.5}$$

e deteminou $\boldsymbol{\Gamma}_1$ pela expressão:

$$\boldsymbol{\Gamma}_1 = \frac{1}{2}(\bar{\boldsymbol{X}}_1^{-(ij)} - \bar{\boldsymbol{X}}_2^{-(ij)})^T \boldsymbol{S}_p^{*-1}. \tag{3.6}$$

Fixou-se o valor da expressão (3.5) e, considerando a exclusão da observação $\boldsymbol{x}_{ij}$, determinou-se o valor y_{ij} por:

$$y_{ij} = \boldsymbol{\Gamma}_1 \boldsymbol{x}_{ij} - \Gamma_2. \tag{3.7}$$

Repetiu-se o procedimento conforme mencionado na subseção (3.1) para todos os valores de *i* e *j*, omitindo a observação $\boldsymbol{x}_{ij}$. Assim, tem-se uma amostra $y_{11}, y_{12}, \ldots, y_{1n_1}$ realizada da população π_1 e outra $y_{21}, y_{22}, \ldots, y_{2n_2}$ da população π_2. Em seguida, calcularam-se as médias e as variâncias de cada amostra utilizando-se expressões definidas em (3.1).

Estimou-se a probabilidade total de classificação incorreta por meio de dois estimadores um considerando desvios padrões heterogêneos e o outro considerando desvio padrão comum. Esses estimadores foram representados por $PTCI_3$ e $PTCI_4$ dados por:

$$PTCI_3 = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_1}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_2}\right) \tag{3.8}$$

$$PTCI_4 = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_p}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_p}\right). \tag{3.9}$$

O estimador definido pela expressão (3.8) foi denominado de método 3 heterogêneo (M3HE) e o definido pela expressão (3.9) de método 4 homogêneo (M4HO).

Nos dois últimos estimadores propostos, realizaram-se alterações na constante Γ_2. Assim, fixou-se $\mathbf{\Gamma}_1$ definido pela expressão:

$$\mathbf{\Gamma}_1 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \boldsymbol{S}_p^{-1} \tag{3.10}$$

e omitindo-se cada observação $\boldsymbol{x}_{ij}$ calculou-se,

$$\Gamma_2 = \frac{1}{2}(\bar{\boldsymbol{X}}_1^{-(ij)} - \bar{\boldsymbol{X}}_2^{-(ij)})^T \boldsymbol{S}_p^{*-1}(\bar{\boldsymbol{X}}_1^{-(ij)} + \bar{\boldsymbol{X}}_2^{-(ij)}), \tag{3.11}$$

e para cada combinação de i e de j obteve-se y_{ij} utilizando a expressão:

$$y_{ij} = \mathbf{\Gamma}_1 \boldsymbol{x}_{ij} - \Gamma_2. \tag{3.12}$$

Novamente, repetiu-se procedimento mencionado na seção (3.1) para todos os valores de *i* e *j*, omitindo-se a observação $\boldsymbol{x}_{ij}$ e determinou-se uma amostra $y_{11}, y_{12}, \ldots, y_{1n_1}$ realizada da população π_1 e outra $y_{21}, y_{22}, \ldots, y_{2n_2}$ da população π_2. Em seguida calcularam-se as médias e as variâncias de cada amostra, utilizando-se expressões definidas em (3.1).

Dessa maneira, estimou-se a probabilidade total de classificação incorreta considerando um estimador com desvios padrões heterogêneos e o outro estimador considerando desvio padrão comum. Esses estimadores foram denominados por $PTCI_5$ e $PTCI_6$ e representados pelas expressões:

$$PTCI_5 = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_1}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_2}\right) \tag{3.13}$$

$$PTCI_6 = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_p}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_p}\right). \tag{3.14}$$

O estimador definido em (3.13) foi denominado de método 5 heterogêneo (M5HE) e o definido em (3.14) de método 6 homogêneo (M6HO).

3.3 Simulação Monte Carlo

Foram geradas amostras de duas populações normais multivariadas homocedásticas, considerando custos de classificação incorreta e probabilidade *a priori* iguais nas duas populações. O vetor de médias da população π_1 foi fixado em $\mathbf{0}$ ($\boldsymbol{\mu}_1 = \mathbf{0}$) e o vetor de médias $\boldsymbol{\mu}_2$ da população π_2 foi simulado em função da distância de Mahalanobis dada por $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, que foi considerada igual a 0, 2, 4, 8, 16 e 32. A busca de $\boldsymbol{\mu}_2$ aproximada para estabelecer o vetor fixado Δ^2, foi realizada por tentativa e erro, conforme especificado na função *R* denominada *buscamu2* (ver apêndice).

Os tamanhos amostrais da população π_1 foram $n_1 =$ 10, 50, 100 e os da população π_2 $n_2 =$ 10, 50, 100 combinados fatorialmente com $p = 2, 10$ variáveis. Para simular as amostras de cada população, foi considerada uma estrutura equicorrelação entre as variáveis, definido Σ por:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix},$$

em que σ^2 foi fixado em 1 sem perda de generalidade e $\rho = 0$, $\rho = 0{,}5$ e $\rho = 0{,}9$.

Em cada uma das simulações a probabilidade total de classificação incorreta paramétrica foi estimada utilizando-se os seis métodos descritos nas seções anteriores. Como o vetores de médias $\boldsymbol{\mu}_i$ dessas populações eram conhecidos, foi possível determinar a probabilidade total de classificação incorreta paramétrica. Assim, pode-se comparar o desempenho dos métodos de estimação avaliando-se os vieses e os erros quadráticos médios, utilizando os dados obtidos nas simulações. Para isso, foram utilizadas N = 2000 simulações Monte Carlo.

A probabilidade total de classificação incorreta paramétrica que foi utilizada em cada configuração simulada é dada por:

$$PTCI = \Phi(-0{,}5\Delta)$$

3.4 Viés e erro quadrático médio (*EQM*) estimados

O viés e o erro quadrático médio (*EQM*) dos estimadores da probabilidade total de classificação incorreta (θ) foram computados nas N=2000 simulações, e apresentados a seguir. Seja $\hat{\theta}$ um dos estimadores ($PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$, $PTCI_6$) de θ, então, o viés foi determinado por:

$$Viés(\hat{\theta}) = \frac{\sum_{m=1}^{N} \hat{\theta}_m}{N} - \theta, \tag{3.15}$$

e o erro quadrático médio (*EQM*) de $\hat{\theta}$ por:

$$EQM(\hat{\theta}) = \frac{\sum_{m=1}^{N} (\hat{\theta}_m - \theta)^2}{N}, \tag{3.16}$$

em que $\hat{\theta}_m$ é a estimativa de θ na *m*-ésima simulação Monte Carlo e θ é a probabilidade total de classificação incorreta paramétrica $(PTCI)$.

Todas as simulações foram feitos a partir de rotinas desenvolvidas no software R (R Development Core Team, 2007) (Apêndice).

4 RESULTADOS E DISCUSSÃO

Na seqüência estão descritos os resultados da simulação Monte Carlo para os seis métodos utilizados para estimar a probabilidade total de classificação incorreta paramétrica $PTCI$. Para cada combinação utilizada na simulação, os resultados estimados dos vieses e dos erros quadráticos médios foram apresentados em formas de figuras e de tabelas. Como os resultados apresentados nas simulações considerando coeficiente de correlação $\rho = 0$, $\rho = 0{,}5$ e $\rho = 0{,}9$ foram semelhantes, então, foram discutidos apenas os resultados em que $p = 2$, $p = 10$ e $\rho = 0{,}5$, sendo que os demais resultados podem ser encontrados em apêndice. Nas Figuras 4.1 e 4.2 foram apresentados os vieses para 2 variáveis e na Tabela 4.1 foram mostrados os vieses para 10 variáveis. Nas Figuras 4.3 e 4.4 foram apresentados os resultados simulados para os erros quadráticos médios para duas 2 variáveis e na Tabela 4.2 foram apresentados os resultados para 10 variáveis.

Para a discussão dos resultados, foram realizadas análises descritivas e nas comparações realizadas entre estimadores, os erros de Monte Carlo foram considerados desprezíveis. Como os estimadores $PTCI_5$ e $PTCI_6$ mostraram resultados semelhantes e foram os que apresentaram valores negativos formaram o grupo 1 e, os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, por mostrarem comportamento similares, e apresentarem vieses positivos formaram o grupo 2.

4.1 Vieses estimados

Na Figura 4.1, foram apresentados os resultados dos vieses dos estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função dos tamanhos das amostras $n_1 = 10$ e $n_2 = 10, 50, 100$ e da distância de Mahalanobis Δ^2. Pelos resultados apresentados, observou-se que à medida que a distância Δ^2 entre as duas populações foi aumentado, os valores estimados dos vieses, considerando valores amostrais n_1 e n_2 fixos, foram, em módulo, diminuindo. A situação em que $n_1 = 10$ e $n_2 = 10$ foi a que apresentou os piores resultados, por isso, foi considerada a mais crítica. Para $\Delta^2 < 8$ o grupo 1, em geral, apresentou valores negativos, subestimando, assim o valor paramétrico, enquanto que, o grupo 2 superestimou, em razão do fato de terem apresentado valores positivos de vieses. E para valores de $\Delta^2 \geq 8$ o grupo 1 passou a apresentar estimativas de vieses

positivas e, ainda, menores do que o grupo 2. Assim, nesta situação recomendam-se os estimadores do grupo 2 para distâncias menores que 8 e os estimadores do grupo 1 para distâncias maiores. Na situação em que $n_1 = 10$ e $n_2 = 50$ os estimadores tanto do grupo 1, quanto do grupo 2 tornaram-se mais semelhantes e apresentaram resultados menores do que os observados na situação com $n_1 = 10$ e $n_2 = 10$. Nesta situação, destacam-se os resultados dos estimadores do grupo 1, principalmente para $\Delta^2 \geq 4$, em que estes apresentaram valores, praticamente, nulos. Assim, para este caso, recomenda-se os estimadores $PTCI_5$ e $PTCI_6$ e para $\Delta^2 < 4$ os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$ e $PTCI_4$. Como na situação em que $n_1 = 10$ e $n_2 = 100$ os resultados mostrados, graficamente, são muito parecidos com a situação descrita anteriormente, recomendaram-se os mesmos estimadores nas mesmas situações.

Na Figura 4.2, foram apresentados os resultados para os tamanhos amostrais $n_1 = 50$ e $n_2 = 50, 100$, $n_1 = 100$, $n_2 = 100$. Em razão do fato de os 3 gráficos terem sido semelhantes, então descreveram-se os resultados para os 3 casos, simultaneamente. Na situação em que Δ^2 está compreendido entre 2 e 16, em que os vieses são próximos zero, os estimadores do grupo 1 apresentaram vieses, em módulo, menores do que o grupo 2. E, para distâncias maiores que 16, todos os estimadores foram não viesados, assintoticamente, apresentando vieses, praticamente zero. Dessa forma, para $\Delta^2 \leq 2$ recomendaram-se os estimadores do grupo 2, para distâncias compreendidas entre 2 e 16 os estimadores dos grupos 1 e 2, e para distâncias maiores qualquer um dos estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$.

Na Tabela 4.1, foram apresentados os resultados estimados dos vieses considerando $p = 10$ variáveis e $n_1 = 10$, $n_2 = 10, 50, 100$ e $n_1 = 50$, $n_2 = 50, 100$ e $n_1 = 100$, $n_2 = 100$. Para as distâncias 0, 2, 4, 8, 16, e 32, entre populações, os valores das probabilidades totais de classificação incorretas foram 0,5, 0,2397501, 0,1586553, 0,0786496, 0,0227501 e 0,0023389, respectivamente. Observou-se, da mesma forma que ocorreu para $p = 2$ variáveis, que a distância Δ^2 entre as duas populações foi a característica que mais contribuiu na função discriminante, ou seja, quanto maior a distância entre as duas populações menor foi a $PTCI$, chegando a mensurar um valor máximo de 50%, considerando $\Delta^2 = 0$, e de 0,2% para a $\Delta^2 = 32$. Vale ainda ressaltar, que para os diferentes tamanhos de amostras

n_1 e n_2 os estimadores $PTCI_5$ e $PTCI_6$ apresentaram vieses negativos o que, na maioria das vezes, não é bom, em razão do fato de subestimarem os valores paramétricos, enquanto que os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$ e $PTCI_4$ apresentaram valores positivos, e assim, superestimando a $PTCI$, que é uma situação melhor. Em diferentes situações (Tabela 4.1), principalmente naquelas em que Δ^2 está compreendido entre 2 e 16, os valores estimados do grupo 1 foram quase sempre em módulo inferiores aos valores do grupo 2. No entanto, subestimação do real valor é pior que a superestimação. Isso ocorre em função de o resultado desse estimador induzir o pesquisador a ter uma idéia de que a regra de classificação é a melhor, por possuir menor probabilidade total de classificação incorreta do que de fato ela é. Em pequenas amostras os valores dos vieses em módulo são parecidos entre os estimadores dos dois grupos, mas em grandes amostras os estimadores do grupo 1 apresentaram vieses, apesar de negativos, de magnitude muito inferior aos do grupo2. Na situação em que $n_1 = 50$, $n_2 = 50, 100$ e $n_1 = 100$ e $\Delta^2 > 16$ todos os estimadores apresentaram vieses com valores menores do que 2,5%. Assim, para distâncias $\Delta^2 \leq 16$ recomendaram-se os estimadores do segundo grupo 2, foram recomendados por superestimarem os valores paramétricos e, para distâncias maiores, qualquer um dos estimadores pode ser recomendado.

O uso do estimador comum do desvio padrão populacional, não trouxe benefício na redução do viés dos estimadores. Ao contrário, os estimadores foram mais viesados do que os que consideraram estimadores independentes dos desvios padrões populacionais. Somente para grandes amostras e os estimadores $PTCI_5$ e $PTCI_6$ é que esse fato não foi confirmado de forma consistente. Assim, o estimador $PTCI_3$ é considerado o melhor entre todos, pois os vieses são positivos e de maneira geral apresentou menor viés que seus concorrentes com vieses positivos. Para diversas outras situações, esses resultados, observados e comentados anteriormente, são verificados em apêndice.

TABELA 4.1: Vieses estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 10$ variáveis, e com tamanhos amostrais $n_1 = 10, 50, 100$, $n_2 = 10, 50, 100$, e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 0$	$\Delta^2 = 2$	$\Delta^2 = 4$
10	10	M1HE	0,0342221	0,1746757	0,1821685
		M2HO	0,0360334	0,1850422	0,1966245
		M3HE	0,0250156	0,1620951	0,1695017
		M4HO	0,0277363	0,1735919	0,1853488
		M5HE	-0,2947567	-0,1103909	-0,0698673
		M6HO	-0,2860471	-0,1027291	-0,0649020
10	50	M1HE	0,0234364	0,0898258	0,0732629
		M2HO	0,0271800	0,0942019	0,0802215
		M3HE	0,0202811	0,0805586	0,0653644
		M4HO	0,0190319	0,0860052	0,0727767
		M5HE	-0,2050925	-0,0634507	-0,0425478
		M6HO	-0,2045568	-0,0601262	-0,0381903
10	100	M1HE	0,0211055	0,0774732	0,0573546
		M2HO	0,0252545	0,0820166	0,0631155
		M3HE	0,0193532	0,0685578	0,0502877
		M4HO	0,0167835	0,0746182	0,0571237
		M5HE	-0,1881810	-0,0518095	-0,0323748
		M6HO	-0,1890291	-0,0477920	-0,0291364
50	50	M1HE	0,0163095	0,0355857	0,0294313
		M2HO	0,0163674	0,0371095	0,0309658
		M3HE	0,0151531	0,0331477	0,0267780
		M4HO	0,0152202	0,0347096	0,0283279
		M5HE	-0,1247083	-0,0293579	-0,0201849
		M6HO	-0,1237852	-0,0279975	-0,0190677
50	100	M1HE	0,0103260	0,0272170	0,0203374
		M2HO	0,0104142	0,0283542	0,0216096
		M3HE	0,0096453	0,0258151	0,0189229
		M4HO	0,0095468	0,0267259	0,0198474
		M5HE	-0,1086158	-0,0202881	-0,0148167
		M6HO	-0,1080689	-0,0189028	-0,0133988
100	100	M1HE	0,0113631	0,0169003	0,0141220
		M2HO	0,0113834	0,0176457	0,0147919
		M3HE	0,0109423	0,0157297	0,0128672
		M4HO	0,0109636	0,0164708	0,0135408
		M5HE	-0,0873277	-0,0157059	-0,0103933
		M6HO	-0,0870078	-0,0149851	-0,0098184

...continua...

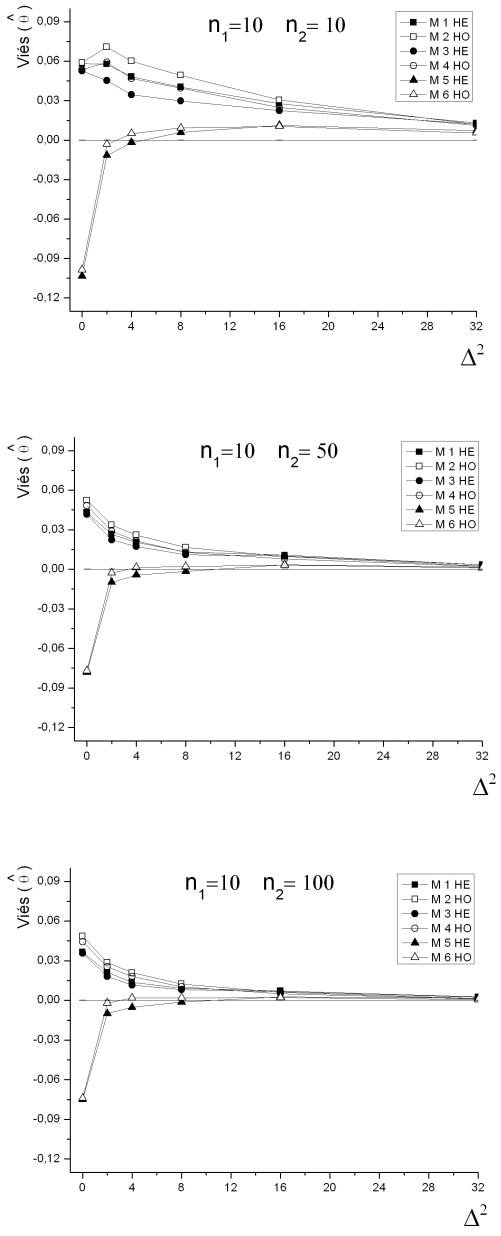

FIGURA 4.1: Vieses estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais n_1, n_2, e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

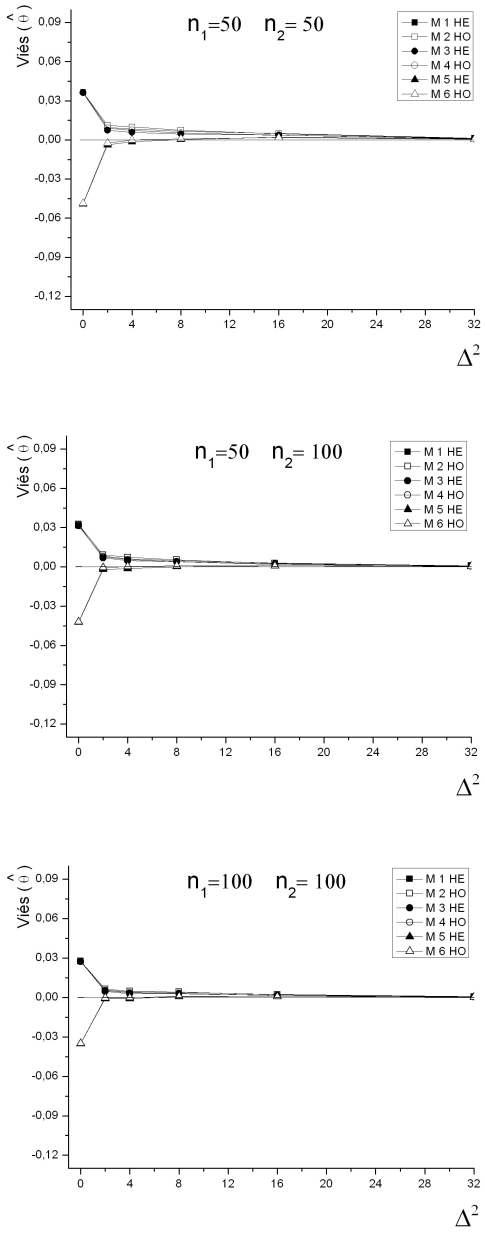

FIGURA 4.2: Vieses estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais n_1, n_2, e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

Tabela 4.1 - Continuação.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 8$	$\Delta^2 = 16$	$\Delta^2 = 32$
10	10	M1HE	0,1678423	0,1342155	0,0791898
		M2HO	0,1842583	0,1508937	0,0900639
		M3HE	0,1550620	0,1251579	0,0773381
		M4HO	0,1731971	0,1432624	0,0893420
		M5HE	-0,0260281	0,0077544	0,0147874
		M6HO	-0,0245249	0,0064969	0,0122905
10	50	M1HE	0,0487324	0,0274778	0,0087463
		M2HO	0,0538165	0,0290008	0,0072653
		M3HE	0,0441074	0,0255532	0,0086449
		M4HO	0,0484350	0,0260258	0,0067251
		M5HE	-0,0228066	-0,0057238	-0,0000281
		M6HO	-0,0213979	-0,0058882	-0,0005799
10	100	M1HE	0,0352620	0,0166555	0,0051516
		M2HO	0,0386349	0,0164588	0,0035842
		M3HE	0,0314475	0,0154176	0,0050244
		M4HO	0,0347179	0,0146677	0,0032138
		M5HE	-0,0155199	-0,0034805	0,0004724
		M6HO	-0,0147315	-0,0044297	-0,0003029
50	50	M1HE	0,0213803	0,0113439	0,0029556
		M2HO	0,0222107	0,0112358	0,0026257
		M3HE	0,0190733	0,0100417	0,0026462
		M4HO	0,0198960	0,0099102	0,0023111
		M5HE	-0,0114401	-0,0037456	-0,0002498
		M6HO	-0,0110997	-0,0040555	-0,0004660
50	100	M1HE	0,0142534	0,0074832	0,0018511
		M2HO	0,0148224	0,0073258	0,0015866
		M3HE	0,0131294	0,0068825	0,0017168
		M4HO	0,0133316	0,0065043	0,0014021
		M5HE	-0,0084336	-0,0026663	-0,0001561
		M6HO	-0,0077987	-0,0027411	-0,0003089
100	100	M1HE	0,0104109	0,0051795	0,0013429
		M2HO	0,0107744	0,0050832	0,0011931
		M3HE	0,0092957	0,0045805	0,0012109
		M4HO	0,0096542	0,0044793	0,0010612
		M5HE	-0,0055979	-0,0019373	-0,0000538
		M6HO	-0,0053621	-0,0020788	-0,0001730

4.2 Erros quadráticos médios estimados (*EQM*s)

Na Figura 4.3, foram mostrados os resultados dos erros quadráticos médios para os tamanhos amostrais $n_1 = 10$, $n_2 = 10, 50, 100$. Pode-se observar, de forma geral, que para distâncias maiores do que 16 todos os estimadores apresentaram erros quadráticos médios muito pequenos e idênticos quase, nulos, podendo-se dizer que nesta situação os estimadores foram muito semelhantes. Comprovou-se, ainda, que nas situações que foram consideradas as mais críticas na seção 4.1, pequenas amostras e Δ^2 pequenos, os resultados dos *EQMs* foram os mais elevados. Isso ocorreu em razão de o *EQM* ser função do viés ao quadrado e da variância do estimador, assim, é esperado que seu valor diminua com o aumento dos tamanhos amostrais, pois, são estimadores consistentes. Como o viés dos estimadores tomados ao quadrado foram quantidades inexpressivas para grandes valores de Δ^2 e para grandes amostrais, então, o *EQM* se tornou praticamente função da variância dos estimadores. Com isso, inferiu-se que as variâncias dos estimadores foram similares. Na Figura 4.4 foram mostrados os valores simulados dos *EQMs* para os tamanhos amostrais $n_1 = 50$, $n_2 = 50, 100$ e $n_1 = 100$. Nesta situação, poucas mudanças foram observadas nos valores dos erros quadráticos médios em relação ao padrão observado para os demais tamanhos amostrais (Figura 4.3). Observou-se que os estimadores apresentaram resultados muito semelhantes. Isso ocorreu em razão de os vieses terem sido menores, em função dos tamanhos amostrais n_1, n_2. Na situação em que $\Delta^2 > 2$, os estimadores apresentaram valores muito próximos de zero, indicando que a precisão é alta e viés pequeno, para todos eles.

Nas situações de pequenas amostras e Δ^2 menores, os estimadores apresentaram diferentes *EQMs*. Nesses casos, destacaram-se os estimadores do grupo 1, consistentemente. No entanto, essa superioridade desses estimadores deve ser vista com ressalva, uma vez que os vieses foram negativos. Dentre os estimadores do grupo 2, destaca-se o $PTCI_3$ que apresentou *EQMs* menores, ou, no máximo, iguais a de seus concorrentes diretos. Assim, em situações de pequeno número de variáveis, esse estimador se destacou e pode ser recomendado.

Na Tabela 4.2, foram apresentados os resultados estimados dos *EQMs* considerando $p = 10$ variáveis e $n_1 = 10$, $n_2 = 10, 50, 100$ e $n_1 = 50$, $n_2 = 50, 100$ e $n_1 = 10$, $n_2 = 100$. Sabe-se que bons estimadores apresentam boas propriedades.

Verificou-se que os estimadores, tanto do grupo 1 quanto do grupo 2, são consistentes, pois, conforme os valores amostrais foram aumentando, os resultados dos *EQMs* foram diminuindo, tornando-se, praticamente nulos, principalmente para grandes valores de Δ^2.

Para todos os tamanhos amostrais, os estimadores do grupo 1 apresentaram menores valores de *EQM* para $\Delta^2 > 0$ do que os do grupo 2, mas para $\Delta^2 = 0$ o oposto foi verificado. Isso aconteceu em razão de os valores dos vieses serem menores no grupo 1 que no grupo 2 para $\Delta^2 > 0$. Um fato importante é que, apesar do grupo 1 ter apresentado valores menores de *EQMs*, sabe-se que seus vieses foram negativos, o que vem sendo considerado uma desvantagem, pois o pesquisador pode ter uma expectativa sepervalorizada da regra de classificação. Vale a pena ressaltar que para $\Delta^2 > 4$ os estimadores apresentaram *EQMs* muito pequenos, em razão de os resultados dos vieses terem sido próximos de zero, principalmente, os do grupo 1.

Como aconteceu no caso de $p = 2$ variáveis, o estimador $PTCI_3$, do grupo 2, foi considerado melhor $p = 10$, uma vez que apresentou *EQMs* menores em praticamente todas as situações, exceto quando comparado aos valores de *EQMs* do grupo 1. Mas, como tem-se optado por estimadores cujo viés é positivo, para não provocar expectativas falso-positivas nos pesquisadores que pretendem utilizar a regra de classificação baseada na combinação linear de Fisher, então o estimador $PTCI_3$, que considera variâncias heterogêneas foi determinado como ótimo. Isso porque apresentou menor viés positivo e foi considerado mais eficiente ou de menor acurácia. Finalmente, vale a pena ressaltar que os estimadores $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ foram mais eficientes do que o estimador $PTCI_1$, original de Lachenbruch & Mickey (1968) e do que o método modificado por Oliveira e Ferreira (2008).

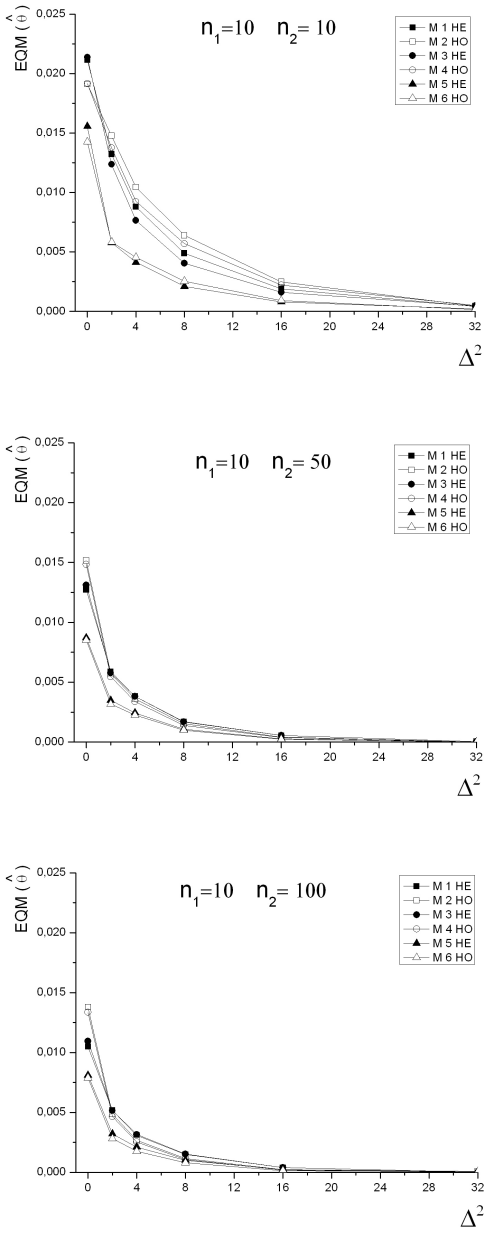

FIGURA 4.3: Erro quadrático médios estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais n_1, n_2, e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

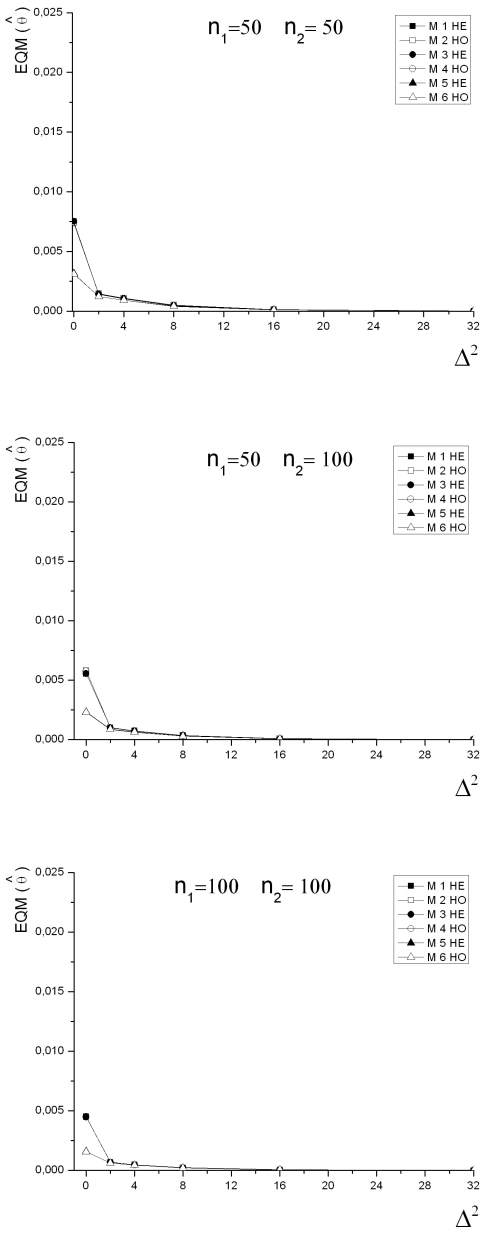

FIGURA 4.4: Erro quadrático médio estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais n_1, n_2, e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

TABELA 4.2: Erro quadrático médio estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 10$ variáveis, e com tamanhos amostrais $n_1 = 10, 50, 100$, $n_2 = 10, 50, 100$, e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 0$	$\Delta^2 = 2$	$\Delta^2 = 4$
10	10	M1HE	0,0148745	0,0443117	0,0466549
		M2HO	0,0132653	0,0468063	0,0515558
		M3HE	0,0151242	0,0407170	0,0424097
		M4HO	0,0133863	0,0434191	0,0476695
		M5HE	0,0941500	0,0177125	0,0088282
		M6HO	0,0895933	0,0172278	0,0092929
10	50	M1HE	0,0087204	0,0154973	0,0107270
		M2HO	0,0091411	0,0157855	0,0112353
		M3HE	0,0091505	0,0139798	0,0095902
		M4HO	0,0088757	0,0142291	0,0099895
		M5HE	0,0449002	0,0067537	0,0037269
		M6HO	0,0444298	0,0061775	0,0032022
10	100	M1HE	0,0074349	0,0127370	0,0078257
		M2HO	0,0080104	0,0126431	0,0078779
		M3HE	0,0079039	0,0114713	0,0070415
		M4HO	0,0076993	0,0113257	0,0070296
		M5HE	0,0377876	0,0052821	0,0028147
		M6HO	0,0378863	0,0044118	0,0023098
50	50	M1HE	0,0037282	0,0030194	0,0021105
		M2HO	0,0036775	0,0031236	0,0022095
		M3HE	0,0037256	0,0028659	0,0019606
		M4HO	0,0036742	0,0029664	0,0020549
		M5HE	0,0163989	0,0020219	0,0012493
		M6HO	0,0161580	0,0019473	0,0012157
50	100	M1HE	0,0027265	0,0019392	0,0012539
		M2HO	0,0027176	0,0019709	0,0012846
		M3HE	0,0027376	0,0018733	0,0012026
		M4HO	0,0027128	0,0018879	0,0012120
		M5HE	0,0124275	0,0012703	0,0008442
		M6HO	0,0122996	0,0011929	0,0007903
100	100	M1HE	0,0020273	0,0010385	0,0006983
		M2HO	0,0020146	0,0010614	0,0007177
		M3HE	0,0020267	0,0010033	0,0006643
		M4HO	0,0020138	0,0010243	0,0006821
		M5HE	0,0080529	0,0008480	0,0005187
		M6HO	0,0079938	0,0008246	0,0005083

...continua...

Tabela 4.2- Continuação.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 8$	$\Delta^2 = 16$	$\Delta^2 = 32$
10	10	M1HE	0,0383786	0,0250923	0,0098026
		M2HO	0,0447850	0,0312780	0,0131672
		M3HE	0,0339981	0,0224134	0,0095009
		M4HO	0,0409161	0,0292419	0,0133454
		M5HE	0,0028735	0,0014265	0,0010654
		M6HO	0,0037331	0,0020685	0,0013016
10	50	M1HE	0,0052306	0,0016042	0,0001865
		M2HO	0,0053056	0,0015705	0,0001151
		M3HE	0,0047911	0,0015049	0,0001980
		M4HO	0,0046767	0,0013595	0,0001042
		M5HE	0,0013150	0,0001875	0,0000087
		M6HO	0,0011929	0,0001755	0,0000048
10	100	M1HE	0,0035375	0,0008444	0,0000868
		M2HO	0,0032209	0,0006123	0,0000323
		M3HE	0,0032609	0,0008087	0,0000890
		M4HO	0,0028580	0,0005353	0,0000280
		M5HE	0,0010584	0,0001741	0,0000112
		M6HO	0,0008437	0,0001195	0,0000034
50	50	M1HE	0,0010767	0,0002930	0,0000184
		M2HO	0,0011264	0,0002954	0,0000159
		M3HE	0,0009734	0,0002581	0,0000159
		M4HO	0,0010197	0,0002601	0,0000136
		M5HE	0,0005069	0,0000854	0,0000231
		M6HO	0,0005090	0,0000890	0,0000021
50	100	M1HE	0,0006121	0,0001531	0,0000089
		M2HO	0,0006074	0,0001442	0,0000071
		M3HE	0,0005816	0,0001437	0,0000083
		M4HO	0,0005623	0,0000597	0,0000062
		M5HO	0,0003528	0,0000597	0,0000019
		M6HE	0,0003316	0,0000577	0,0000017
100	100	M1HE	0,0003692	0,0000855	0,0000045
		M2HO	0,0003805	0,0000852	0,0000040
		M3HE	0,0003447	0,0000781	0,0000040
		M4HO	0,0003552	0,0000779	0,0000036
		M5HE	0,0002346	0,0000424	0,0000013
		M6HO	0,0002351	0,0000433	0,0000012

5 CONCLUSÕES

Os resultados obtidos pelo presente estudo permitem concluir que:

1. As modificações do estimador da probabilidade total de classificação incorreta foram propostas com sucesso;

2. O uso do estimador comum do desvio padrão populacional nas amostragens *jackknife* tiveram menor acurácia, quando comparados com o uso de estimadores independentes do desvio padrão para cada população;

3. Os métodos propostos foram mais eficientes que a modificação proposta por Oliveira e Ferreira (2008);

4. O estimador que aplica o procedimento de *jackknife* apenas na obtenção da combinação linear de Fisher, estimador $PTCI_3$, foi considerado ótimo e é recomendado para ser utilizado para se estimar a qualidade das regras de classificação linear de Fisher;

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. 3. ed. New York: J. Wiley, 2003. 752p.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. São Paulo: Sociedade Brasileira de Matemática, 2000. 125p.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. 5. ed. São Paulo: Atual, 2003. 526p.

DACHS, J. N. W. **Estatística computacional**: uma introdução em turbo pascal. Rio de Janeiro, 1988. 236p.

FERREIRA, D. F. **Estatística multivariada**. Lavras: UFLA, 2008. 642p.

FISHER. R. A. The statistical utilization of multiple measurements. **Annals Eugenics**, London, v. 8, p. 376-386, 1938.

GIRI, N. C. **Multivariate statistical analysis**. 2. ed. New York: Marcel Dekker, 2004. 558p.

HÄRDLE, W.; SIMAR, L. **Applied multivariate statistical analysis**. New York: Springer Verlag, 2007. 486p.

JOHNSON, M. E. **Multivariate statistical simulation**: a guide to selecting and generating continuous multivariate distributions. New York: J. Wiley, 1987. 240p.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 4. ed. New Jersey: Prentice Hall, 1998. 816p.

LACHENBRUSC, P. A. ; MICKEY, M. R. Estimation of error rates in discriminant analysis. **Technometrics**, Washington, v. 10, n. 1, p. 1-11, Feb. 1968.

MAROCO, J. **Análise estatística com utilização de SPSS**. 3. ed. Lisboa: Sílabo, 2007. 825p.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: UFMG, 2005. 295p.

NAYLOR, T. H.; BALINTFY, J. L.; BURDICK, D. S.; CHU, K. **Técnicas de simulação em computadores**. Petrópolis: Vozes, 1971. 402p.

OLIVEIRA, I. R. C.; FERREIRA, D. F. Avaliação da probabilidade de classificação incorreta em análise discriminante para duas populações normais. In: Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, 53., 2008. Lavras. **Anais**...Lavras: DEX/UFLA, 2008. v. 1. p. 36-36.

R DEVELOPMENT CORE TEAM. A. Language nd Environment for Statistical Computing. Vienna, Austria. **R**: Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: 20 dez. 2007.

RENCHER, A. C. **Methods of multivariate analysis**. 2. ed. New York: J. Wiley, 2002. 708p.

SHARMA, S. **Applied multivariate techniques**. New York: J. Wiley, 1996. 295p.

SMITH, C. A. B. Some example of discrimination. **Annals of Eugenics**, Hert, v. 13, p. 272-282, 1947.

WALD, A. On a statistical problem arising in the classification of an individual into one of two groups. Ann Arbor. **Annals of Mathematical statistics**, California, v. 15, n. 2, p. 145-162, 1944.

WELCH, B. L.. Note on discrimination function. **Biometrika**, London, v. 31, n. 1-2, p. 218, Jul. 1939.

APÊNDICE

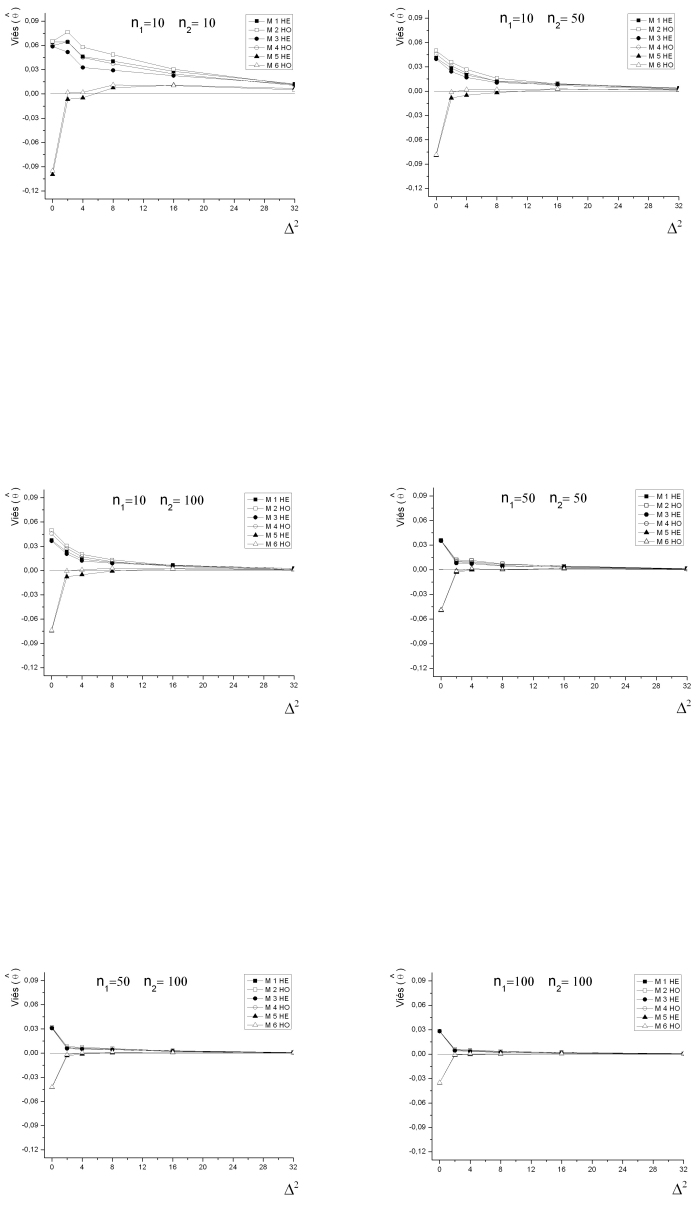

FIGURA 5.1: Vieses estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais n_1, n_2, e correlação $\rho = 0$ em 2000 simulações Monte Carlo.

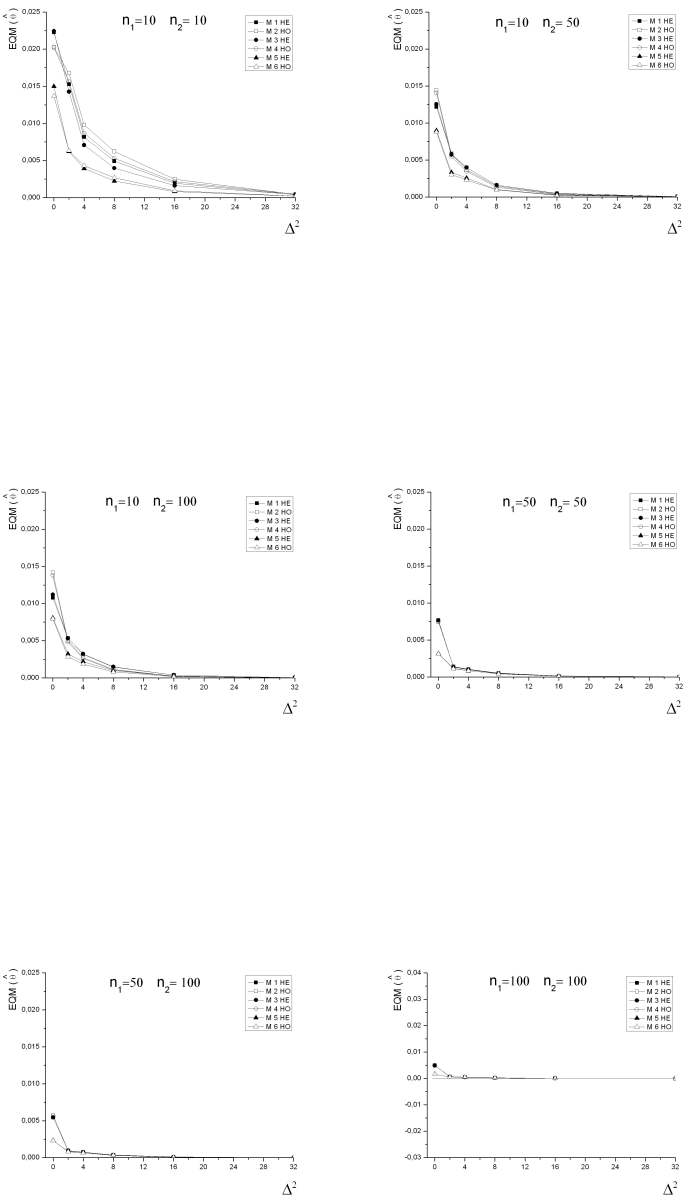

FIGURA 5.2: Erros quadráticos médios estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais n_1, n_2, e correlação $\rho = 0$ em 2000 simulações Monte Carlo.

TABELA 5.1: Vieses estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 10$ variáveis, e com tamanhos amostrais $n_1 = 10, 50, 100$, $n_2 = 10, 50, 100$, e correlação $\rho = 0$ em 2000 simulações Monte Carlo.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 0$	$\Delta^2 = 2$	$\Delta^2 = 4$
10	10	M1HE	0,0383519	0,1754185	0,1835757
		M2HO	0,0396456	0,1859552	0,1982646
		M3HE	0,0292161	0,1628300	0,1705271
		M4HO	0,0313861	0,1746059	0,1867134
		M5HE	-0,2907026	-0,1109814	-0,0683810
		M6HO	-0,2814393	-0,1030539	-0,0630164
10	50	M1HE	0,0248444	0,0908728	0,0718505
		M2HO	0,0289599	0,0958912	0,0792961
		M3HE	0,0218277	0,0816021	0,0640922
		M4HO	0,0208621	0,0876919	0,0720045
		M5HE	-0,2038393	-0,0630743	-0,0429249
		M6HO	-0,2034375	-0,0594354	-0,0387728
10	100	M1HE	0,0182348	0,0774065	0,0589398
		M2HO	0,0230973	0,0818301	0,0644307
		M3HE	0,0163492	0,0686476	0,0518172
		M4HO	0,0145813	0,0744291	0,0583894
		M5HE	-0,1891369	-0,0514627	-0,0310744
		M6HO	-0,1906014	-0,0482596	-0,0282508
50	50	M1HE	0,0139605	0,0368795	0,0283497
		M2HO	0,0140728	0,0384691	0,0298787
		M3HE	0,0127923	0,0344545	0,0257097
		M4HO	0,0129148	0,0360760	0,0272592
		M5HE	-0,1259783	-0,0282639	-0,0211044
		M6HO	-0,1250835	-0,0268572	-0,0200110
50	100	M1HE	0,0110998	0,0260705	0,0212837
		M2HO	0,0112829	0,0273865	0,0222740
		M3HE	0,0104254	0,0246558	0,0199044
		M4HO	0,0104124	0,0257473	0,0205561
		M5HE	-0,1082349	-0,0213053	-0,0141129
		M6HO	-0,1077800	-0,0197154	-0,0129487
100	100	M1HE	0,0103195	0,0177908	0,0145360
		M2HO	0,0103405	0,0184866	0,0151749
		M3HE	0,0098986	0,0166286	0,0132595
		M4HO	0,0099205	0,0173342	0,0139012
		M5HE	-0,0878227	-0,0149492	-0,0099780
		M6HO	-0,0875270	-0,0142852	-0,0094246

...continua...

Tabela 5.1 - Continuação.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 8$	$\Delta^2 = 16$	$\Delta^2 = 32$
10	10	M1HE	0,1702651	0,1367370	0,0792321
		M2HO	0,1875146	0,1531319	0,0895770
		M3HE	0,1574366	0,1275417	0,0777579
		M4HO	0,1760125	0,1452061	0,0889335
		M5HE	-0,0264747	0,0068379	0,0153166
		M6HO	-0,0247919	0,0054596	0,0130474
10	50	M1HE	0,0498210	0,0274499	0,0086011
		M2HO	0,0553809	0,0280370	0,0072712
		M3HE	0,0449728	0,0256221	0,0085122
		M4HO	0,0498451	0,0252273	0,0066913
		M5HE	-0,0219342	-0,0057283	-0,0000576
		M6HO	-0,0202933	-0,0064494	-0,0005656
10	100	M1HE	0,0343790	0,0165751	0,0048456
		M2HO	0,0385610	0,0171521	0,0032347
		M3HE	0,0305216	0,0153393	0,0047243
		M4HO	0,0345708	0,0153409	0,0028922
		M5HE	-0,0159237	-0,0035666	0,0003391
		M6HO	-0,0146606	-0,0040377	-0,0004588
50	50	M1HE	0,0223391	0,0116799	0,0032149
		M2HO	0,0232255	0,0115697	0,0028700
		M3HE	0,0200562	0,0103980	0,0028853
		M4HO	0,0209292	0,0102625	0,0025364
		M5HE	-0,0108235	-0,0035523	-0,0001161
		M6HO	-0,0104554	-0,0038664	-0,0003458
50	100	M1HE	0,0137849	0,0074964	0,0018934
		M2HO	0,0145098	0,0073471	0,0015966
		M3HE	0,0126555	0,0068890	0,0017609
		M4HO	0,0130222	0,0065179	0,0014110
		M5HE	-0,0087937	-0,0026399	-0,0001373
		M6HO	-0,0080551	-0,0027202	-0,0003030
100	100	M1HE	0,0097073	0,0052721	0,0013202
		M2HO	0,0100548	0,0051658	0,0011706
		M3HE	0,0086226	0,0046817	0,0011850
		M4HO	0,0089677	0,0045694	0,0010363
		M5HE	-0,0062521	-0,0018719	-0,0000680
		M6HO	-0,0060300	-0,0020258	-0,0001877

TABELA 5.2: Erros quadráticos médios estimados, para os métodos $M1HE$, $M2HO$, $M3HE$, $M4HO$, $M5HE$ e $M6HO$, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 10$ variáveis, e com tamanhos amostrais $n_1 = 10, 50, 100$ e $n_2 = 10, 50, 100$, e correlação $\rho = 0$ em 2000 simulações Monte Carlo.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 0$	$\Delta^2 = 2$	$\Delta^2 = 4$
10	10	M1HE	0,0171824	0,0447845	0,0466138
		M2HO	0,0152870	0,0472860	0,0515750
		M3HE	0,0174533	0,0412192	0,0422066
		M4HO	0,0154349	0,0439475	0,0475248
		M5HE	0,0926258	0,0177284	0,0085821
		M6HO	0,0878563	0,0172050	0,0090652
10	50	M1HE	0,0090706	0,0159094	0,0105517
		M2HO	0,0095546	0,0163459	0,0110460
		M3HE	0,0095390	0,0143843	0,0094871
		M4HO	0,0092460	0,0147641	0,0098560
		M5HE	0,0445882	0,0068607	0,0037698
		M6HO	0,0441052	0,0061476	0,0032253
10	100	M1HE	0,0071888	0,0122224	0,0079354
		M2HO	0,0078055	0,0124375	0,0079293
		M3HE	0,0076331	0,0110097	0,0071267
		M4HO	0,0075364	0,0111309	0,0070661
		M5HE	0,0381120	0,0051513	0,0027370
		M6HO	0,0384547	0,0043665	0,0022395
50	50	M1HE	0,0037750	0,0031267	0,0019618
		M2HO	0,0037267	0,0032342	0,0020597
		M3HE	0,0037787	0,0029680	0,0018194
		M4HO	0,0037299	0,0030712	0,0019115
		M5HE	0,0167651	0,0019513	0,0012326
		M6HO	0,0165248	0,0018723	0,0011992
50	100	M1HE	0,0028605	0,0018024	0,0012606
		M2HO	0,0028601	0,0018372	0,0012817
		M3HE	0,0028727	0,0017386	0,0012083
		M4HO	0,0028546	0,0017563	0,0012088
		M5HE	0,0123652	0,0012590	0,0007978
		M6HO	0,0122605	0,0011664	0,0007536
100	100	M1HE	0,0019919	0,0010464	0,0007528
		M2HO	0,0019794	0,0010691	0,0007748
		M3HE	0,0019921	0,0010096	0,0007171
		M4HO	0,0019792	0,0010316	0,0007379
		M5HE	0,0081404	0,0008037	0,0005455
		M6HO	0,0080854	0,0007827	0,0005380

...continua...

Tabela 5.2 - Continuação.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 8$	$\Delta^2 = 16$	$\Delta^2 = 32$
10	10	M1HE	0,0393575	0,0256135	0,0100369
		M2HO	0,0459852	0,0316926	0,0132520
		M3HE	0,0350656	0,0228954	0,0096803
		M4HO	0,0419917	0,0295232	0,0133373
		M5HE	0,0027473	0,0013206	0,0011475
		M6HO	0,0035696	0,0017385	0,0013814
10	50	M1HE	0,0051972	0,0016003	0,0001875
		M2HO	0,0054459	0,0014436	0,0001134
		M3HE	0,0047210	0,0015161	0,0002010
		M4HO	0,0047773	0,0012603	0,0001028
		M5HE	0,0012712	0,0001854	0,0000085
		M6HO	0,0011494	0,0001639	0,0000044
10	100	M1HE	0,0033524	0,0008080	0,0000806
		M2HO	0,0031888	0,0006341	0,0000263
		M3HE	0,0030825	0,0007717	0,0000821
		M4HO	0,0028212	0,0005529	0,0000228
		M5HE	0,0010337	0,0001583	0,0000106
		M6HO	0,0008432	0,0001167	2,8035370
50	50	M1HE	0,0011183	0,0003058	0,0000213
		M2HO	0,0011718	0,0003080	0,0000187
		M3HE	0,0010147	0,0002707	0,0000185
		M4HO	0,0010650	0,0002725	0,0000160
		M5HE	0,0004893	0,0000866	0,0000026
		M6HO	0,0004913	0,0000899	0,0000024
50	100	M1HE	0,0005893	0,0001511	0,0000091
		M2HO	0,0005912	0,0001430	0,0000070
		M3HE	0,0005594	0,0001413	0,0000085
		M4HO	0,0005465	0,0001286	0,0000062
		M5HE	0,0003519	0,0000189	0,0000018
		M6HO	0,0003313	0,0000584	0,0000016
100	100	M1HE	0,0003542	0,0000870	0,0000091
		M2HO	0,0003640	0,0000868	0,0000070
		M3HE	0,0003318	0,0000797	0,0000085
		M4HO	0,0003409	0,0000795	0,0000062
		M5HE	0,0002409	0,0000423	0,0000018
		M6HO	0,0002408	0,0000433	0,0000016

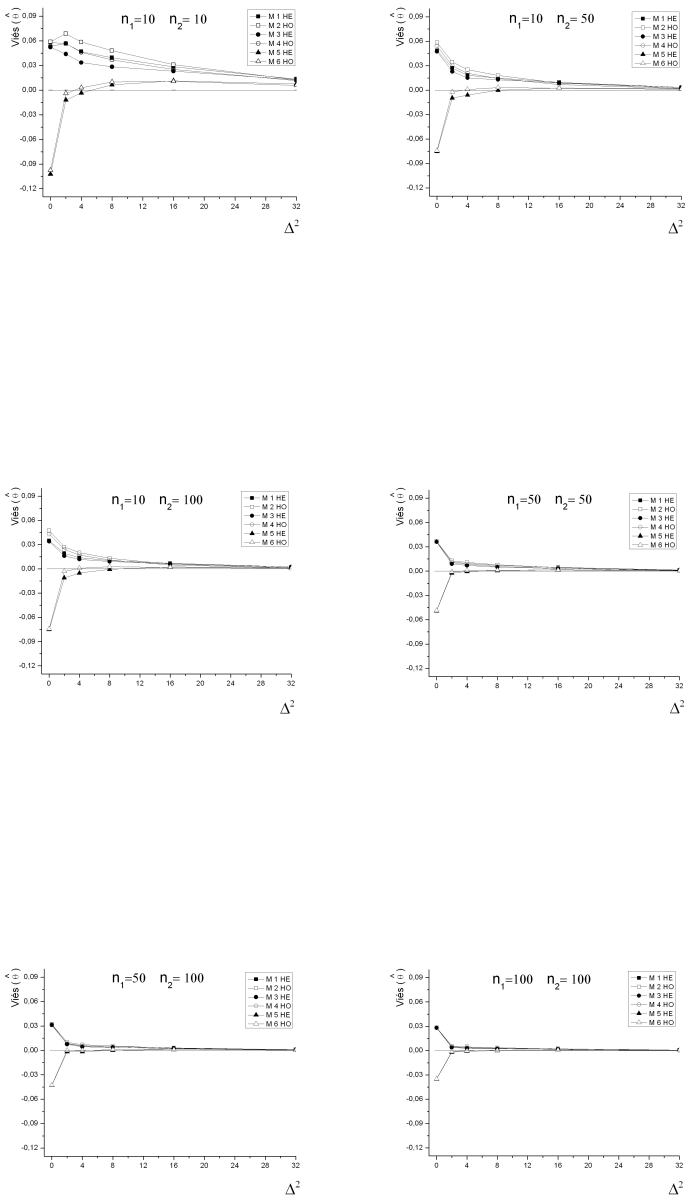

FIGURA 5.3: Vieses estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais n_1, n_2, e correlação $\rho = 0{,}9$ em 2000 simulações Monte Carlo.

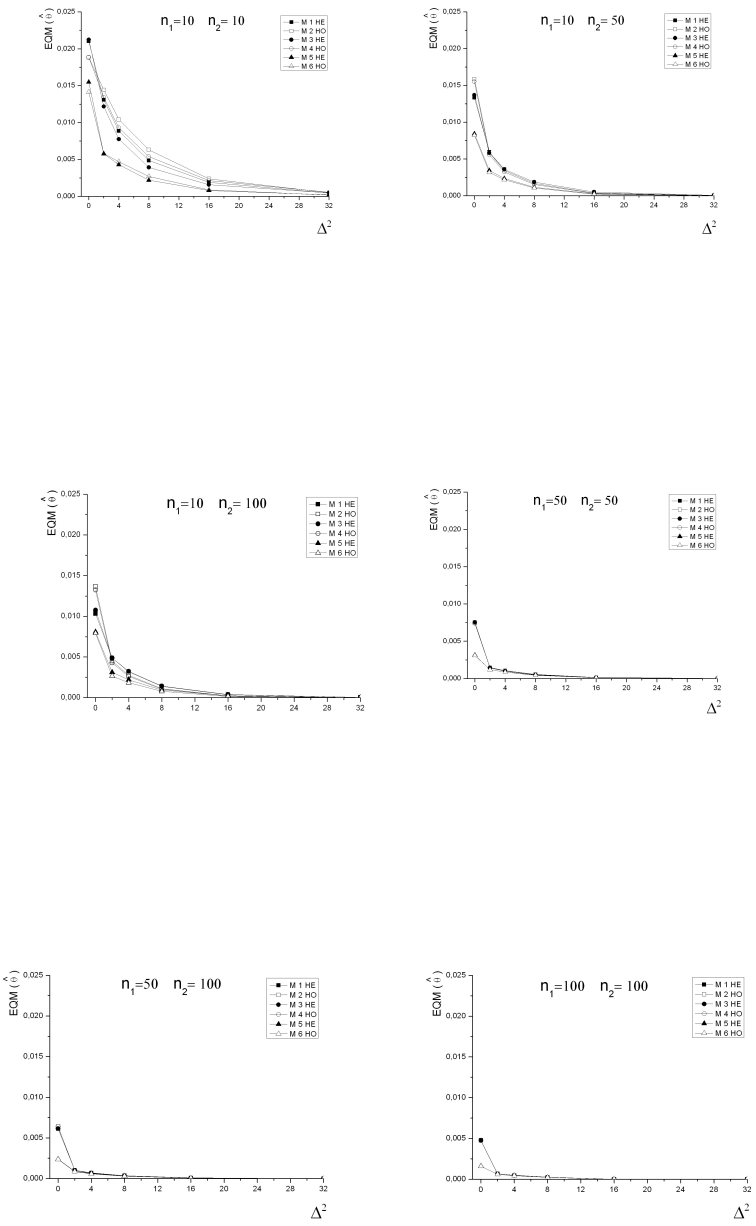

FIGURA 5.4: Erros quadráticos médios estimados, para os métodos M1HE, M2HO, M3HE, M4HO, M5HE e M6HO, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais n_1, n_2, e correlação $\rho = 0{,}9$ em 2000 simulações Monte Carlo.

TABELA 5.3: Vieses estimados, para os métodos $M1HE$, $M2HO$, $M3HE$, $M4HO$, $M5HE$ e $M6HO$, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 10$ variáveis, e com tamanhos amostrais $n_1 = 10, 50, 100$, $n_2 = 10, 50, 100$, e correlação $\rho = 0{,}9$ em 2000 simulações Monte Carlo.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 0$	$\Delta^2 = 2$	$\Delta^2 = 4$
10	10	M1HE	0,0345577	0,1775484	0,1786016
		M2HO	0,0362634	0,1879250	0,1932095
		M3HE	0,0255979	0,1657626	0,1653316
		M4HO	0,0281950	0,1774696	0,1811480
		M5HE	-0,2934379	-0,1105523	-0,0709739
		M6HO	-0,2840474	-0,1036152	-0,0658210
10	50	M1HE	0,0252412	0,0900193	0,0726368
		M2HO	0,0289897	0,0947022	0,0790749
		M3HE	0,0221423	0,0806372	0,0648102
		M4HO	0,0207937	0,0864403	0,0717658
		M5HE	-0,2040576	-0,0638787	-0,0420175
		M6HO	-0,2036940	-0,0597583	-0,0386399
10	100	M1HE	0,0162706	0,0753220	0,0549027
		M2HO	0,0203707	0,0796733	0,0614294
		M3HE	0,0143272	0,0663936	0,0478903
		M4HO	0,0118994	0,0723087	0,0554953
		M5HE	-0,1915363	-0,0524209	-0,0330960
		M6HO	-0,1919463	-0,0491236	-0,0299958
50	50	M1HE	0,0140940	0,0355548	0,0298609
		M2HO	0,0141265	0,0370971	0,0314542
		M3HE	0,0129226	0,0331140	0,0272056
		M4HO	0,0129648	0,0346953	0,0288019
		M5HE	-0,1257533	-0,0292019	-0,0199051
		M6HO	-0,1248144	-0,0278079	-0,0187207
50	100	M1HE	0,0126211	0,0248806	0,0204682
		M2HO	0,0128919	0,0261064	0,0214043
		M3HE	0,0119543	0,0234633	0,0190953
		M4HO	0,0120272	0,0244670	0,0196908
		M5HE	-0,1074487	-0,0223077	-0,0147739
		M6HO	-0,1069909	-0,0208267	-0,0136875
100	100	M1HE	0,0102131	0,0182466	0,0149726
		M2HO	0,0102254	0,0189391	0,0156507
		M3HE	0,0097874	0,0170838	0,0137057
		M4HO	0,0098018	0,0177829	0,0143856
		M5HE	-0,0878682	-0,0145123	-0,0096160
		M6HO	-0,0875583	-0,0138501	-0,0090301

...continua...

Tabela 5.3 - Continuação.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 8$	$\Delta^2 = 16$	$\Delta^2 = 32$
10	10	M1HE	0,1706701	0,1340717	0,0793262
		M2HO	0,1881218	0,1498883	0,0903615
		M3HE	0,1575036	0,1266620	0,0776188
		M4HO	0,1764997	0,1435606	0,0890061
		M5HE	-0,0262178	0,0047667	0,0153052
		M6HO	-0,0246863	0,0035077	0,0131796
10	50	M1HE	0,0523420	0,0267650	0,0084936
		M2HO	0,0568340	0,0278679	0,0072536
		M3HE	0,0474359	0,0250348	0,0083131
		M4HO	0,0511904	0,0250823	0,0066351
		M5HE	-0,0207707	-0,0060056	-0,0000790
		M6HO	-0,0196649	-0,0065037	-0,0000526
10	100	M1HE	0,0352238	0,0171040	0,0047682
		M2HO	0,0387684	0,0167339	0,0033639
		M3HE	0,0313943	0,0158232	0,0046276
		M4HO	0,0348199	0,0149395	0,0030090
		M5HE	-0,0156669	-0,0031597	0,0003312
		M6HO	-0,0146521	-0,0042945	-0,0004018
50	50	M1HE	0,0223147	0,0109004	0,0029679
		M2HO	0,0231468	0,0107789	0,0026422
		M3HE	0,0200355	0,0096546	0,0026607
		M4HO	0,0208641	0,0095076	0,0023270
		M5HE	-0,0108555	-0,0040332	-0,0002297
		M6HO	-0,0105157	-0,0043559	-0,0004452
50	100	M1HE	0,0150359	0,0077972	0,0018204
		M2HO	0,0156312	0,0077006	0,0015643
		M3HE	0,0138954	0,0071837	0,0016925
		M4HO	0,0141279	0,0068563	0,0013823
		M5HE	-0,0077706	-0,0024012	-0,0001728
		M6HO	-0,0071221	-0,0024259	-0,0003193
100	100	M1HE	0,0099371	0,0054785	0,0012602
		M2HO	0,0102527	0,0053795	0,0011099
		M3HE	0,0088423	0,0048920	0,0011274
		M4HO	0,0091535	0,0047863	0,0009778
		M5HE	-0,0060228	-0,0017062	-0,0001079
		M6HO	-0,0058201	-0,0018509	-0,0002287

TABELA 5.4: Erros quadráticos médios estimados, para os métodos $M1HE$, $M2HO$, $M3HE$, $M4HO$, $M5HE$ e $M6HO$, em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 10$ variáveis, e com tamanhos amostrais $n_1 = 10, 50, 100$, $n_2 = 10, 50, 100$, e correlação $\rho = 0{,}9$ em 2000 simulações Monte Carlo.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 0$	$\Delta^2 = 2$	$\Delta^2 = 4$
10	10	M1HE	0,0154478	0,0466973	0,0452989
		M2HO	0,0138490	0,0491808	0,0500993
		M3HE	0,0157089	0,0433483	0,0411262
		M4HO	0,0140550	0,0460259	0,0461157
		M5HE	0,0938497	0,0181651	0,0088633
		M6HO	0,0889881	0,0176902	0,0093620
10	50	M1HE	0,0087455	0,0159056	0,0108034
		M2HO	0,0091827	0,0160119	0,0112646
		M3HE	0,0091747	0,0143347	0,0096942
		M4HO	0,0088871	0,0144311	0,0100589
		M5HE	0,0444999	0,0068471	0,0037593
		M6HO	0,0440858	0,0061395	0,0033383
10	100	M1HE	0,0072240	0,0120864	0,0077547
		M2HO	0,0076644	0,0122952	0,0076786
		M3HE	0,0077308	0,0108798	0,0070245
		M4HO	0,0074094	0,0110196	0,0068625
		M5HE	0,0391715	0,0052381	0,0029977
		M6HO	0,0389638	0,0044997	0,0023797
50	50	M1HE	0,0035420	0,0029188	0,0020355
		M2HO	0,0034785	0,0030180	0,0021341
		M3HE	0,0035436	0,0027644	0,0018831
		M4HO	0,0034794	0,0028606	0,0019743
		M5HE	0,0166506	0,0019412	0,0011724
		M6HO	0,0164053	0,0018598	0,0011356
50	100	M1HE	0,0029553	0,0018200	0,0012535
		M2HO	0,0029752	0,0018521	0,0012733
		M3HE	0,0029649	0,0017594	0,0012033
		M4HO	0,0029672	0,0017750	0,0012033
		M5HE	0,0121804	0,0013617	0,0008412
		M6HO	0,0120737	0,0012750	0,0007999
100	100	M1HE	0,0019843	0,0011057	0,0007390
		M2HO	0,0019698	0,0011294	0,0007612
		M3HE	0,0019843	0,0010680	0,0007024
		M4HO	0,0019699	0,0010906	0,0007232
		M5HE	0,0081402	0,0008234	0,0005145
		M6HO	0,0080830	0,0008040	0,0005056

...continua...

Tabela 5.4 - Continuação.

			Distância de Mahalanobis		
n_1	n_2	Métodos	$\Delta^2 = 8$	$\Delta^2 = 16$	$\Delta^2 = 32$
10	10	M1HE	0,0396317	0,0249615	0,0097100
		M2HO	0,0465178	0,0308206	0,0130468
		M3HE	0,0352763	0,0226143	0,0093921
		M4HO	0,0427944	0,0289757	0,0130055
		M5HE	0,0027481	0,0011361	0,0011548
		M6HO	0,0035234	0,0015650	0,0013560
10	50	M1HE	0,0056589	0,0015890	0,0001808
		M2HO	0,0057110	0,0014724	0,0001204
		M3HE	0,0051524	0,0015026	0,0001864
		M4HO	0,0050218	0,0012787	0,0001077
		M5HE	0,0012728	0,0001940	0,0000084
		M6HO	0,0011385	0,0001731	0,0000053
50	100	M1HE	0,0035023	0,0008392	0,0000737
		M2HO	0,0032614	0,0006163	0,0000282
		M3HE	0,0032262	0,0007960	0,0000750
		M4HO	0,0028939	0,0005375	0,0000245
		M5HE	0,0010448	0,0001713	0,0000093
		M6HO	0,0008524	0,0001174	0,0000029
50	50	M1HE	0,0011510	0,0002670	0,0000194
		M2HO	0,0012033	0,0002686	0,0000170
		M3HE	0,0010435	0,0002363	0,0000169
		M4HO	0,0010925	0,0002375	0,0000146
		M5HE	0,0005117	0,0000803	0,0000025
		M6HO	0,0005151	0,0000840	0,0000024
50	100	M1HE	0,0006408	0,0001593	0,0000087
		M2HO	0,0006423	0,0001520	0,0000069
		M3HE	0,0006073	0,0001491	0,0000082
		M4HO	0,0005929	0,0001368	0,0000061
		M5HE	0,0003445	0,0000059	0,0000019
		M6HO	0,0003284	0,0000057	0,0000017
100	100	M1HE	0,0003526	0,0000909	0,0000043
		M2HO	0,0003613	0,0000906	0,0000038
		M3HE	0,0003295	0,0000834	0,0000038
		M4HO	0,0003376	0,0000832	0,0000033
		M5HE	0,0002336	0,0000428	0,0000013
		M6HO	0,0002335	0,0000437	0,0000012

PROGRAMA A: Rotina para análise dos dados.

```
#### Programa  para determinar a média da população ##
## 2 (mu2) dado a distância de Mahalanobis (Delta2)###
#### realizado por tentativa e erro considerando duas#
#### populações normais p-variadas homocedásticas#####
#### library mvtnorm ################################
buscamu2 = function(Delta2,p,mu1,Sigma,precis = 1e-11)
{  if (Delta2==0) return(list(mu2=matrix(0,p,1)
,Delta2c=Delta2)) else
{
  if (Delta2<=0.5) c=matrix(Delta2,p,1) else
         c = matrix(1,p,1)
       mu2 = mu1+c
    SigmaI = solve(Sigma)
   Delta2c = t(mu1-mu2)%*%SigmaI%*%(mu1-mu2)
if (abs(SigmaI[1,2]-0.0)<precis)
mu2=(Delta2/p)^0.5*c else
while (abs(Delta2c-Delta2)>precis)
{
if ((Delta2c>Delta2))
{
       aux = (Delta2c-Delta2)/Delta2
if (Delta2<=10) aux=aux*runif(1)
      if (aux>1) aux=runif(1)
       mu2 = mu2 - as.numeric(aux)*c
} else
{
       aux = (Delta2-Delta2c)/Delta2
      if (Delta2<=10) aux=aux*runif(1)
       mu2 = mu2 + as.numeric(aux)*c
}
   Delta2c = t(mu1-mu2)%*%SigmaI%*%(mu1-mu2)
   Delta2c;mu2;aux
}
   Delta2c = t(mu1-mu2)%*%SigmaI%*%(mu1-mu2)
}
   return(list(mu2=mu2,Delta2c=Delta2c))
}
## Função para determinar a probabilidade total ######
# de classificação incorreta (PCTI), considerando ####
#a probabilidade a priori das populações 1 e 2########
#iguais a 0,5, ou seja, p1=p2=0,5 ####################
```

```
PCTI = function(psi,Delta2)
 {
   Delta=Delta2^0.5
   if (Delta==0) return(0.5) else
   return(0.5*pnorm(-Delta/2+psi/Delta)
+0.5*(1-pnorm(Delta/2+psi/Delta)))
}
### Função para simular amostrais normais ############
#### multivariadas n1 e n2: pacote MASS, mvrnorm #####
### função para fazer amostragens jackknife ##########
## de Lachenbruch e Mickey (1968) ####################
### essa função retorna uma lista ####################
### y: y$y1 e y$y2 com n1 e n2 elementos #############
## das amostragens jackknifes ########################
LM1P1 = function(X1,X2)
{
     p = ncol(X1);n1=nrow(X1);n2=nrow(X2)
    y1 = matrix(0,n1,1)
    S2 = var(X2)
   Xb2 = apply(X2,2,mean)
   for (i in 1:n1)
{
     X1ij = X1[i!=(1:n1),1:p]
      xij = X1[i==(1:n1),1:p]
       S1 = var(X1ij)
      Xb1 = apply(X1ij,2,mean)
       Sp = (S2*(n2-1)+S1*(n1-2))/(n1+n2-3)
      aux = t(Xb1-Xb2)%*%solve(Sp)
    y1[i] = aux %*% xij-0.5*aux %*% (Xb1+Xb2)
}
       y2 = matrix(0,n2,1)
       S1 = var(X1)
      Xb1 = apply(X1,2,mean)
   for (i in 1:n2)
{
     X2ij = X2[i!=(1:n2),1:p]
      xij = X2[i==(1:n2),1:p]
       S2 = var(X2ij)
      Xb2 = apply(X2ij,2,mean)
       Sp = (S2*(n2-2)+S1*(n1-1))/(n1+n2-3)
      aux = t(Xb1-Xb2)%*%solve(Sp)
    y2[i] = aux %*% xij-0.5*aux %*% (Xb1+Xb2)
```

```
}
   return(list(y1=y1,y2=y2))
}
### Função para fazer as amostragens jackknifes ######
### de Lachenbruch e Mickey (1968), considerando #####
### variâncias heterogêneas retorna a lista y:y$y1 ###
## e y$y2 com n1 e n2 elementos dos jackknifes #######
LMM1P1 = function(X1,X2)
{
      p = ncol(X1);n1=nrow(X1);n2=nrow(X2)
     y1 = matrix(0,n1,1)
     S2 = var(X2)
    Xb2 = apply(X2,2,mean)
     S1 = var(X1)
    Xb1 = apply(X1,2,mean)
     Sp = (S2*(n2-1)+S1*(n1-1))/(n1+n2-2)
   aux2 = 0.5*t(Xb1-Xb2)%*%solve(Sp)%*%(Xb1+Xb2)
   for (i in 1:n1)
{
  X1ij = X1[i!=(1:n1),1:p]
   xij = X1[i==(1:n1),1:p]
   S1j = var(X1ij)
  Xb1j = apply(X1ij,2,mean)
   Spj = (S2*(n2-1)+S1j*(n1-2))/(n1+n2-3)
   aux = t(Xb1j-Xb2)%*%solve(Spj)
 y1[i] = aux %*% xij - aux2
}
   y2  = matrix(0,n2,1)
  for (i in 1:n2)
{
  X2ij = X2[i!=(1:n2),1:p]
   xij = X2[i==(1:n2),1:p]
  S2j  = var(X2ij)
  Xb2j = apply(X2ij,2,mean)
   Spj = (S2j*(n2-2)+S1*(n1-1))/(n1+n2-3)
   aux = t(Xb1-Xb2j)%*%solve(Spj)
 y2[i] = aux %*% xij - aux2
}
   return(list(y1=y1,y2=y2))
}
### Função para fazer as amostragens jackknifes ######
### de Lachenbruch e Mickey (1968), considerando #####
```

```
### variâncias homogêneas e retorna a lista y:y$y1 ###
## e y$y2 com n1 e n2 elementos dos jackknifes #######
LMM2P1 = function(X1,X2)
{
     p = ncol(X1);n1=nrow(X1);n2=nrow(X2)
    y1 = matrix(0,n1,1)
    S2 = var(X2)
   Xb2 = apply(X2,2,mean)
    S1 = var(X1)
   Xb1 = apply(X1,2,mean)
    Sp = (S2*(n2-1)+S1*(n1-1))/(n1+n2-2)
   aux2 = t(Xb1-Xb2)%*%solve(Sp)
   for (i in 1:n1)
{
  X1ij = X1[i!=(1:n1),1:p]
   xij = X1[i==(1:n1),1:p]
   S1j = var(X1ij)
  Xb1j = apply(X1ij,2,mean)
   Spj = (S2*(n2-1)+S1j*(n1-2))/(n1+n2-3)
   aux = t(Xb1j-Xb2)%*%solve(Spj)
 y1[i] = aux2 %*% xij-0.5*aux %*% (Xb1j+Xb2)
}
   y2 = matrix(0,n2,1)
   for (i in 1:n2)
{
 X2ij = X2[i!=(1:n2),1:p]
  xij = X2[i==(1:n2),1:p]
  S2j = var(X2ij)
 Xb2j = apply(X2ij,2,mean)
  Spj = (S2j*(n2-2)+S1*(n1-1))/(n1+n2-3)
  aux = t(Xb1-Xb2j)%*%solve(Spj)
y2[i] = aux2 %*% xij-0.5*aux %*% (Xb1+Xb2j)
}
   return(list(y1=y1,y2=y2))
}
### retorna a taxa de erro estimada (TAE)#############
### retorna a TEA para o estimador PTCI1 #############
### considerando variâncias heterogêneas #############
LM1P2 = function(y)
{
 yb1 = mean(y$y1);S21=var(y$y1)
  yb2 = mean(y$y2);S22=var(y$y2)
```

```
  P21 = pnorm(-yb1/S21^0.5)
  P12 = pnorm(yb2/S22^0.5)
  TEA = 0.5*P21+0.5*P12
   return(TEA)
}
### retorna a TEA para o estimador PTCI2 #############
### considerando variâncias homogêneas ###############
LM2P2 = function(y)
{
  n1 = length(y$y1);  n2 = length(y$y2)
 yb1 = mean(y$y1);S21=var(y$y1)
 yb2 = mean(y$y2);S22=var(y$y2)
 S2p = ((n1-1)*S21+(n2-1)*S22)/(n1+n2-2)
 P21 = pnorm(-yb1/S2p^0.5)
 P12 = pnorm(yb2/S2p^0.5)
 TEA = 0.5*P21+0.5*P12
   return(TEA)
}
### retorna a TEA para o estimador PTCI3 #############
### considerando variâncias heterogêneas #############
LM1P3 = function(y3)
{
 yb1 = mean(y3$y1);S21=var(y3$y1)
 yb2 = mean(y3$y2);S22=var(y3$y2)
 P21 = pnorm(-yb1/S21^0.5)
 P12 = pnorm(yb2/S22^0.5)
 TEA = 0.5*P21+0.5*P12
   return(TEA)
}
### retorna a TEA para o estimador PTCI4 #############
### considerando variâncias homogêneas ###############
LM2P3 = function(y3)
{  n1 = length(y3$y1);  n2=length(y3$y2)
yb1 = mean(y3$y1);S21=var(y3$y1)
yb2 =mean(y3$y2);S22=var(y3$y2)
S2p =((n1-1)*S21+(n2-1)*S22)/(n1+n2-2)
P21 = pnorm(-yb1/S2p^0.5)
P12 = pnorm(yb2/S2p^0.5)
TEA = 0.5*P21+0.5*P12
return(TEA)
}
### retorna a TEA para o estimador PTCI5 #############
```

```
### considerando variâncias heterogêneas #############
 LM1P4 = function(y4)
{
yb1 = mean(y4$y1);S21=var(y4$y1)
yb2 = mean(y4$y2);S22=var(y4$y2)
P21 = pnorm(-yb1/S21^0.5)
P12 = pnorm(yb2/S22^0.5)
TEA = 0.5*P21+0.5*P12
   return(TEA)
}
### retorna a TEA para o estimador PTCI6 #############
### considerando variâncias homogêneas ###############
 LM2P4 = function(y4)
 {
 n1 = length(y4$y1);  n2 = length(y4$y2)
yb1 = mean(y4$y1);S21=var(y4$y1)
yb2 = mean(y4$y2);S22=var(y4$y2)
S2p = ((n1-1)*S21+(n2-1)*S22)/(n1+n2-2)
P21 = pnorm(-yb1/S2p^0.5)
P12 = pnorm(yb2/S2p^0.5)
TEA = 0.5*P21+0.5*P12
   return(TEA)
}
# EQMs e os vieses em N simulações Monte Carlo########
 SMC  = function(N,n1,n2,mu1,mu2,Sigma,PCTIp1)
 {
   Bias1 = 0
   EQM1  = 0
   Bias2 = 0
   EQM2  = 0
   Bias3 = 0
   EQM3  = 0
   Bias4 = 0
   EQM4  = 0
   Bias5 = 0
   EQM5  = 0
   Bias6 = 0
   EQM6  = 0
  Saida  = matrix(0,N,6)
   for (i in 1:N)
{
      X1 = mvrnorm(n1, mu1,Sigma)
```

```
      X2 = mvrnorm(n2, mu2,Sigma)
       y = LM1P1(X1,X2)
    TEA1 = LM1P2(y)
    TEA2 = LM2P2(y)
      y3 = LMM1P1(X1,X2)
    TEA3 = LM1P3(y3)
    TEA4 = LM2P3(y3)
      y4 = LMM2P1(X1,X2)
    TEA5 = LM1P4(y4)
    TEA6 = LM2P4(y4)
   Bias1 = Bias1 + TEA1/N
   Bias2 = Bias2 + TEA2/N
   Bias3 = Bias3 + TEA3/N
   Bias4 = Bias4 + TEA4/N
   Bias5 = Bias5 + TEA5/N
   Bias6 = Bias6 + TEA6/N
   EQM1  = EQM1  + (TEA1 - PCTIp1)^2/N
   EQM2  = EQM2  + (TEA2 - PCTIp1)^2/N
   EQM3  = EQM3  + (TEA3 - PCTIp1)^2/N
   EQM4  = EQM4  + (TEA4 - PCTIp1)^2/N
   EQM5  = EQM5  + (TEA5 - PCTIp1)^2/N
   EQM6  = EQM6  + (TEA6 - PCTIp1)^2/N
Saida[i,1] = TEA1
Saida[i,2] = TEA2
Saida[i,3] = TEA3
Saida[i,4] = TEA4
Saida[i,5] = TEA5
Saida[i,6] = TEA6
#print(TEA1);print(TEA2)
}
     Bias1 = Bias1 - PCTIp1
     Bias2 = Bias2 - PCTIp1
     Bias3 = Bias3 - PCTIp1
     Bias4 = Bias4 - PCTIp1
     Bias5 = Bias5 - PCTIp1
     Bias6 = Bias6 - PCTIp1
return(list(Vies1 = Bias1, Vies2 = Bias2
, Vies3 = Bias3,
Vies4 = Bias4, Vies5 = Bias5, Vies6 = Bias6
 EQM1 = EQM1
EQM2 = EQM2, EQM3 = EQM3, EQM4 = EQM4
, EQM5 = EQM5, EQM6 = EQM6,
```

```
Saida=Saida))
}
#########################################################
#########################################################
```