ESTIMADORES DA PROBABILIDADE TOTAL DE CLASSIFICAÇÃO INCORRETA NA ANÁLISE DISCRIMINANTE

Altemir da Silva BRAGA[1]
Daniel Furtado FERREIRA[1]
Patrícia de Siqueira RAMOS[1]

▪ RESUMO: 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*. Dessa forma, torna-se um desafio interessante desenvolver novas metodologias que possam minimizar o total de classificações incorretas. Uma das alternativas existentes é a função linear de Fisher (1938) que é baseada em uma transformação linear que maximiza a variabilidade dos escores da função discriminante das populações. Com base nessa combinação vários estimadores da probabilidade total de classificação incorreta ($PTCI$) foram propostos na literatura, entre eles destaca-se o estimador de Lachenbruch & Mickey (1968), denominado, neste estudo, por $PTCI_1$. Assim, neste trabalho, foram propostos 4 estimadores ($PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$) da $PTCI$ a partir de modificações no estimador ($PTCI_1$). Utilizando-se simulação Monte Carlo os seus desempenhos foram avaliados e comparados entre si e com os desempenhos dos estimadores original ($PTCI_1$) e de Oliveira & Ferreira (2008), denominado por $PTCI_2$, que foi sugerido por Giri(2004). Dentre eles o estimador $PTCI_3$ foi considerado de melhor desempenho.

▪ PALAVRAS-CHAVE: Multivariada; método; classificação; simulação; *Jackknife*.

1 Introdução

Classificar corretamente novas observações ou novos indivíduos em uma de diferentes populações definidas *a priori* e estimar a quantidade de classificações realizadas de forma incorreta é o principal objetivo da análise discriminante. É

[1]Departamento de Ciências Exatas – DEX, Universidade Federal de Lavras – UFLA, Caixa Postal 3037, CEP: 37200000, Lavras, MG, Brasil. E-mail:*altemirbraga@hotmail.com* / *danielff@dex.ufla.br*

possível elaborar uma função matemática denominada na literatura por regra de classificação 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 e minimizasse a probabilidade de classificação incorreta dos indivíduos em suas respectivas populações. Situações práticas reais como classificar um indivíduo em doente ou não-doente, um sítio florestal em recuperado e não-recuperado ou classificar um animal em um grupo de alto ou baixo risco anestésico com base em várias variáveis mensuradas nos mesmos são exemplos de utilização da técnica.

Existem na literatura (Ferreira, 2008; Giri, 2004; Mingoti, 2005; Johnson & Wichern, 1998; Lachenbruch & Mickey, 1968; entre outros autores) diversos estimadores da probabilidade total de classificação incorreta ($PTCI$). Dentre eles, Ferreira (2008) descreve os estimadores da PTCI denominados de estimadores da ressubstituição, da ressubstituição com divisão amostral, das probabilidades de classificação incorretas estimadas, pseudo-*jackknife* e o estimador de Lachenbruch & Mickey (1968). Destes estimadores o da ressubstituição e o de Lachenbruch & Mickey (1968) possuem o pior e melhor desempenhos, respectivamente, justificando o porquê de este estimador ter sido escolhido. Giri (2004) sugeriu que se avaliasse esse estimador, considerando desvio padrão comum no estimador de Lachenbruch & Mickey (1968), haja vista que, no método original, os autores consideraram desvios padrões heterogêneos, embora tenham assumido populações homocedásticas.

A avaliação do desempenho dessa proposta foi realizada por Oliveira & Ferreira (2008), sendo que o estimador com as modificações sugeridas por Giri (2004), apresentou resultados piores. A motivação para a realização desse trabalho decorreu dessa proposta e da constatação de que esses estimadores apresentaram vieses.

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

2 Metodologia

Neste estudo foram comparados, por meio de simulação Monte Carlo, os desempenhos de seis estimadores propostos para estimar a probabilidade total de classificação incorreta. O desempenho da qualidade dos estimadores da $PTCI$, foi avaliado por intermédio dos vieses e dos erros quadráticos médios estimados (EQM).

2.1 Estimador de Lachenbruch & Mickey e estimadores propostos

Considerando-se 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 $n_1 + n_2 - 2 \geq p$, os estimadores $(\bar{\boldsymbol{X}}_1^{-(ij)}, \bar{\boldsymbol{X}}_2^{-(ij)})$ dos vetores de médias populacionais e os estimadores

$\boldsymbol{S}_i$, $i = 1, 2$, das matrizes de covariâncias populacionais excluindo a observação $\boldsymbol{x}_{ij}$, são:

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

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

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 e $\bar{\boldsymbol{X}}_k^{-(ij)}$ representa a média amostral da k-ésima população, $k = 1$ ou 2, quando a observação x_{ij} foi omitida da amostra.

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

$$\boldsymbol{S}_p^* = \frac{(n_1^* - 1)\boldsymbol{S}_1 + (n_2^* - 1)\boldsymbol{S}_2}{n_1^* + n_2^* - 2}. \tag{3}$$

Assim, para a observação $\boldsymbol{x}_{ij}$ omitida, calcula-se o valor y_{ij} dado pela expressão:

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

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

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 :

$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad e \quad S_i^2 = \frac{1}{n_i - 1} \left[\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{n_i} y_{ij} \right)^2}{n_i} \right]. \tag{4}$$

Os estimadores da probabilidade total de classificação incorreta, para esse e para os demais casos contemplados neste trabalho, são dados pela expressão geral:

$$PTCI_k = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_1}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_2}\right), \tag{5}$$

sendo que a situação em que $k = 1$ representa esse primeiro estimador, que utiliza estimadores independentes para as variâncias de cada população. Para $k = 3$ e 5, os estimadores resultantes utilizam a mesma expressão, com estimadores independentes para as variâncias de cada população, mas utilizam diferentes estratégias dos procedimentos *jackknife*. Esses diferentes procedimentos serão descritos na sequência. A função $\Phi(z)$ corresponde à função de distribuição normal padrão avaliada no ponto z correspondente ao seu argumento.

A segunda opção foi estimar a probabilidade total de classificação incorreta considerando um estimador do desvio padrão comum as duas populações homogêneas, conforme sugerido por Giri (2004). Essa opção possibilitou a obtenção de três novos estimadores da probabilidade total de classificação incorreta, cuja expressão geral é dada por:

$$PTCI_\ell = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_p}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_p}\right), \tag{6}$$

sendo que o caso de $\ell = 2$ representa o segundo estimador da PTCI considerando os procedimentos descritos anteriormente para a determinação de y_{ij} e os casos de $\ell = 4$ e 6, dois novos estimadores que serão descritos a seguir, considerando também o estimador do desvio padrão comum às populações consideradas, sendo este estimador do desvio padrão comum S_p definido por:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}. \tag{7}$$

Convém salientar que os demais estimadores da PTCI diferem apenas na forma de definir as amostras *jackknife*, mas com os mesmas definições para as quantidades $\bar{y}_1$, $\bar{y}_2$, S_1^2, S_2^2 e S_p^2, uma vez obtidas essas amostras. A razão de tal proposta é calcada no fato de que os estimadores PTC_1 e PTC_2 são viesados.

Os estimadores PTC_3 e PTC_4 foram propostos da seguinte forma. Inicialmente, 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

$$\Gamma_2 = \frac{1}{2}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \boldsymbol{S}_p^{-1}(\bar{\boldsymbol{X}}_1 + \bar{\boldsymbol{X}}_2) \tag{8}$$

e determinou $\boldsymbol{\Gamma}_1$, considerando a exclusão da observação x_{ij}, por

$$\boldsymbol{\Gamma}_1 = \frac{1}{2}(\bar{\boldsymbol{X}}_1^{-(ij)} - \bar{\boldsymbol{X}}_2^{-(ij)})^\top \boldsymbol{S}_p^{*-1}. \tag{9}$$

Fixou-se o valor da expressão (8) e, considerando a exclusão da observação $\boldsymbol{x}_{ij}$, determinou-se o valor y_{ij} por:

$$y_{ij} = \boldsymbol{\Gamma}_1 \boldsymbol{x}_{ij} - \Gamma_2. \tag{10}$$

Repetiu-se o procedimento para todos os valores de i e j, omitindo-se a observação $\boldsymbol{x}_{ij}$. Assim, têm-se amostras *jackknifes* $y_{11}, y_{12}, \ldots, y_{1n_1}$ da população π_1 e $y_{21}, y_{22}, \ldots, y_{2n_2}$, da população π_2. Em seguida, calcularam-se as médias e as variâncias de cada amostra utilizando-se expressões definidas em (4). Estimou-se a probabilidade total de classificação incorreta por meio de dois estimadores: um considerando estimadores separados dos desvios padrões das populações amostradas e outro, considerando um desvio padrão comum às duas populações. Esses estimadores foram representados por $PTCI_3$ e $PTCI_4$ e calculados pelas expressões (5) e (6), respectivamente. Reitera-se que a alteração realizada nesses estimadores em relação aos estimadores PTC_1 e PTC_2 ocorreu no processo *jackknife*.

Nos dois últimos estimadores propostos (PTC_5 e PTC_6), realizaram-se alterações na constante Γ_2. Assim, fixou-se $\boldsymbol{\Gamma}_1$, dada por

$$\boldsymbol{\Gamma}_1 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \boldsymbol{S}_p^{-1} \tag{11}$$

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

$$\Gamma_2 = \frac{1}{2}(\bar{\boldsymbol{X}}_1^{-(ij)} - \bar{\boldsymbol{X}}_2^{-(ij)})^\top \boldsymbol{S}_p^{*-1}(\bar{\boldsymbol{X}}_1^{-(ij)} + \bar{\boldsymbol{X}}_2^{-(ij)}), \tag{12}$$

para cada combinação de i e de j, obtendo-se y_{ij} por

$$y_{ij} = \boldsymbol{\Gamma}_1 \boldsymbol{x}_{ij} - \Gamma_2. \tag{13}$$

Novamente, repetiu-se procedimento para todos os valores de i e de j, omitindo a observação $\boldsymbol{x}_{ij}$ e determinando as amostras $y_{11}, y_{12}, \ldots, y_{1n_1}$, da população π_1 e $y_{21}, y_{22}, \ldots, y_{2n_2}$, da população π_2. Em seguida calcularam-se as médias e as variâncias de cada amostra, utilizando-se expressões definidas em (4). Dessa maneira, os estimadores denominados $PTCI_5$ e $PTCI_6$ foram obtidos por intermédio das expressões (5) e (6), respectivamente.

2.2 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 $\boldsymbol{\mu}_2$ foi feita por tentativa e erro

de modo que o valor Δ^2 fixado *a priori* fosse obtido, considerando um margem de erro pré-fixada.

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 $p = 2$ e 10 variáveis. Para simular as amostras de cada população, foi considerada uma estrutura equicorrelação entre as variáveis, em que a matriz $\mathbf{\Sigma}$ foi definida por

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

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 foi estimada a probabilidade total de classificação incorreta paramétrica utilizando cada um dos seis estimadores descritos na seção anterior. Como os 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 estimadores, 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 para cada configuração simulada é dada por:

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

2.3 Viés e erro quadrático médio (*EQM*)

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 ($PTCI$), então o viés foi determinado por

$$Viés(\hat{\theta}) = \frac{\sum_{m=1}^{N} \hat{\theta}_m}{N} - \theta, \tag{14}$$

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

$$EQM(\hat{\theta}) = \frac{\sum_{m=1}^{N} (\hat{\theta}_m - \theta)^2}{N}, \tag{15}$$

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 ($PTCI$), sendo $m = 1, 2, \dots, N$.

2.4 Software

Todas as simulações foram feitos a partir de rotinas desenvolvidas no *software* R (R Development Core Team, 2007) e podem ser obtidas por requisição aos autores por intermédio de *e-mail*.

3 Resultados e discussão

Para a discussão dos resultados, foram realizadas análises descritivas. Nas comparações realizadas entre estimadores, os erros de Monte Carlo foram considerados desprezíveis. Como os estimadores $PTCI_5$ e $PTCI_6$ apresentaram, em geral, valores negativos para os vieses estimados, foram chamados de grupo 1 e, os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, por mostrarem comportamento similares, e apresentarem vieses positivos foram chamados de grupo 2.

3.1 Vieses estimados

Na Figura 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 aumentou, os valores estimados dos vieses, considerando valores amostrais n_1 e n_2 fixos, diminuíram em módulo. 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. Para valores de $\Delta^2 \geq 8$, o grupo 1 passou a apresentar estimativas positivas de vieses e, ainda, menores do que as do 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 $n_2 = 10$. Neste caso, 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. Convém salientar que Δ^2 não é conhecido nas situações reais, assim, deve-se utilizar uma estimativa que pode ser obtida por

$$\hat{\Delta}^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_p^- (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2),$$

para possibilitar o direcionamento da escolha do estimador apropriado.

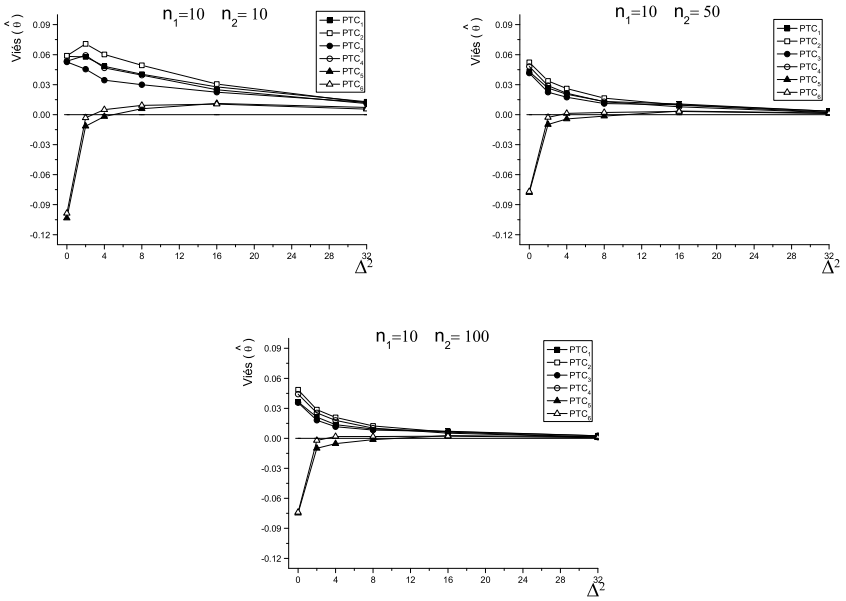

Figura 1 - Vieses estimados para os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, e com tamanhos amostrais $n_1 = 10$ e $n_2 = 10$, 50 e 100 e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

Na Figura 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. 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 deles.

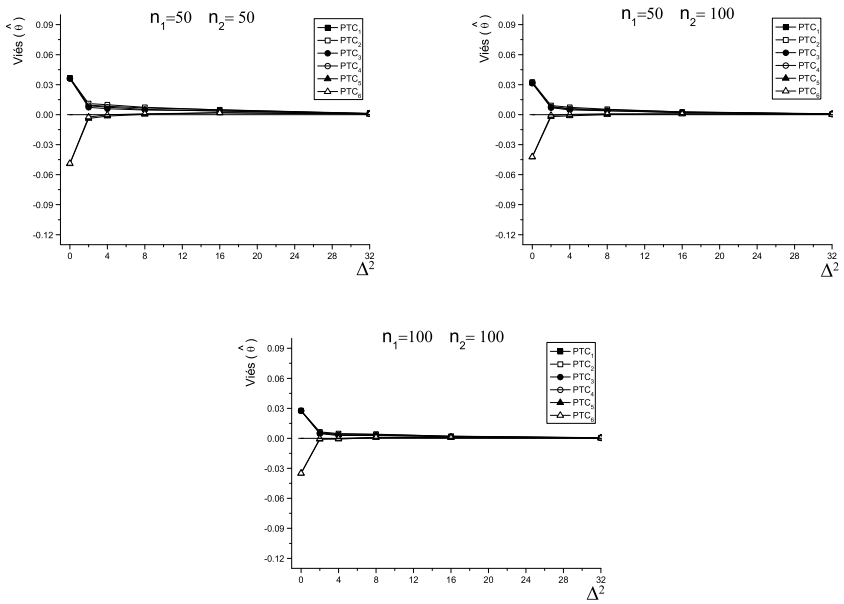

Figura 2 - Vieses estimados para os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis, com tamanhos amostrais$n_1 = 50$ e 100 e $n_2 = 50$ e 100 e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

Nas Tabelas 1 e 2 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 paramétricos das probabilidades totais de classificação incorretas foram 0,5, 0,2397501, 0,1586553, 0,0786496, 0,0227501 e 0,0023389, respectivamente.

Observa-se nessas Tabelas, 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 tanto menor foi a $PTCI$. A $PTCI$ variou de 50%, para $\Delta^2 = 0$ a 0,2%, para $\Delta^2 = 32$. Vale a pena 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 menos pior. No entanto, a subestimação do real valor é pior que a superestimação. Isso ocorre em função de o resultado desse estimador induzir o

Tabela 1 - Vieses estimados para os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função da distância de Mahalanobis $\Delta^2 \leq 4$ entre as médias de 2 populações normais multivariadas com $p = 10$ variáveis com tamanhos amostrais $n_1 = 10, 50, 100$ e $n_2 = 10, 50, 100$ e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo

			Distância de Mahalanobis		
n_1	n_2	Estimadores	$\Delta^2 = 0$	$\Delta^2 = 2$	$\Delta^2 = 4$
10	10	PCT_1	0,0342221	0,1746757	0,1821685
		PCT_2	0,0360334	0,1850422	0,1966245
		PCT_3	0,0250156	0,1620951	0,1695017
		PCT_4	0,0277363	0,1735919	0,1853488
		PCT_5	-0,2947567	-0,1103909	-0,0698673
		PCT_6	-0,2860471	-0,1027291	-0,0649020
10	50	PCT_1	0,0234364	0,0898258	0,0732629
		PCT_2	0,0271800	0,0942019	0,0802215
		PCT_3	0,0202811	0,0805586	0,0653644
		PCT_4	0,0190319	0,0860052	0,0727767
		PCT_5	-0,2050925	-0,0634507	-0,0425478
		PCT_6	-0,2045568	-0,0601262	-0,0381903
10	100	PCT_1	0,0211055	0,0774732	0,0573546
		PCT_2	0,0252545	0,0820166	0,0631155
		PCT_3	0,0193532	0,0685578	0,0502877
		PCT_4	0,0167835	0,0746182	0,0571237
		PCT_5	-0,1881810	-0,0518095	-0,0323748
		PCT_6	-0,1890291	-0,0477920	-0,0291364
50	50	PCT_1	0,0163095	0,0355857	0,0294313
		PCT_2	0,0163674	0,0371095	0,0309658
		PCT_3	0,0151531	0,0331477	0,0267780
		PCT_4	0,0152202	0,0347096	0,0283279
		PCT_5	-0,1247083	-0,0293579	-0,0201849
		PCT_6	-0,1237852	-0,0279975	-0,0190677
50	100	PCT_1	0,0103260	0,0272170	0,0203374
		PCT_2	0,0104142	0,0283542	0,0216096
		PCT_3	0,0096453	0,0258151	0,0189229
		PCT_4	0,0095468	0,0267259	0,0198474
		PCT_5	-0,1086158	-0,0202881	-0,0148167
		PCT_6	-0,1080689	-0,0189028	-0,0133988
100	100	PCT_1	0,0113631	0,0169003	0,0141220
		PCT_2	0,0113834	0,0176457	0,0147919
		PCT_3	0,0109423	0,0157297	0,0128672
		PCT_4	0,0109636	0,0164708	0,0135408
		PCT_5	-0,0873277	-0,0157059	-0,0103933
		PCT_6	-0,0870078	-0,0149851	-0,0098184

Tabela 2 - Vieses estimados para os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função da distância de Mahalanobis $\Delta^2 \geq 8$ entre as médias de 2 populações normais multivariadas com $p = 10$ variáveis, com tamanhos amostrais $n_1 = 10, 50, 100$ e $n_2 = 10, 50, 100$ e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo

			Distância de Mahalanobis		
n_1	n_2	Estimadores	$\Delta^2 = 8$	$\Delta^2 = 16$	$\Delta^2 = 32$
10	10	PCT_1	0,1678423	0,1342155	0,0791898
		PCT_2	0,1842583	0,1508937	0,0900639
		PCT_3	0,1550620	0,1251579	0,0773381
		PCT_4	0,1731971	0,1432624	0,0893420
		PCT_5	-0,0260281	0,0077544	0,0147874
		PCT_6	-0,0245249	0,0064969	0,0122905
10	50	PCT_1	0,0487324	0,0274778	0,0087463
		PCT_2	0,0538165	0,0290008	0,0072653
		PCT_3	0,0441074	0,0255532	0,0086449
		PCT_4	0,0484350	0,0260258	0,0067251
		PCT_5	-0,0228066	-0,0057238	-0,0000281
		PCT_6	-0,0213979	-0,0058882	-0,0005799
10	100	PCT_1	0,0352620	0,0166555	0,0051516
		PCT_2	0,0386349	0,0164588	0,0035842
		PCT_3	0,0314475	0,0154176	0,0050244
		PCT_4	0,0347179	0,0146677	0,0032138
		PCT_5	-0,0155199	-0,0034805	0,0004724
		PCT_6	-0,0147315	-0,0044297	-0,0003029
50	50	PCT_1	0,0213803	0,0113439	0,0029556
		PCT_2	0,0222107	0,0112358	0,0026257
		PCT_3	0,0190733	0,0100417	0,0026462
		PCT_4	0,0198960	0,0099102	0,0023111
		PCT_5	-0,0114401	-0,0037456	-0,0002498
		PCT_6	-0,0110997	-0,0040555	-0,0004660
50	100	PCT_1	0,0142534	0,0074832	0,0018511
		PCT_2	0,0148224	0,0073258	0,0015866
		PCT_3	0,0131294	0,0068825	0,0017168
		PCT_4	0,0133316	0,0065043	0,0014021
		PCT_5	-0,0084336	-0,0026663	-0,0001561
		PCT_6	-0,0077987	-0,0027411	-0,0003089
100	100	PCT_1	0,0104109	0,0051795	0,0013429
		PCT_2	0,0107744	0,0050832	0,0011931
		PCT_3	0,0092957	0,0045805	0,0012109
		PCT_4	0,0096542	0,0044793	0,0010612
		PCT_5	-0,0055979	-0,0019373	-0,0000538
		PCT_6	-0,0053621	-0,0020788	-0,0001730

pesquisador a ter uma idéia de que a regra de classificação é melhor, por possuir menor probabilidade total de classificação incorreta, do que de fato 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 grupo 2. Nas situações em que $n_1 = 50$, $n_2 = 50, 100$ e $n_1 = n_2 = 100$ 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 grupo 2 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 em grandes amostras e para 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.

3.2 Erros quadráticos médios (*EQM*) estimados

Na Figura 3 são mostrados os resultados dos erros quadráticos médios para os tamanhos amostrais $n_1 = 10$ e $n_2 = 10, 50, 100$. Pode-se observar, de forma geral, que para distâncias maiores do que 16 todos os estimadores apresentaram erros quadráticos médios praticamente idênticos e 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 3.1, 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 foram apresentados os valores simulados dos *EQMs* para os tamanhos amostrais $n_1 = 50$ e $n_2 = 50, 100$ e $n_1 = n_2 = 100$. Nesta situação, poucas mudanças foram observadas nos valores dos erros quadráticos médios em relação ao padrão observado para os demais tamanhos amostrais (Figura 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 e n_2. Na situação em que $\Delta^2 > 2$, os estimadores apresentaram valores muito próximos de zero, indicando que a precisão é alta e o 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, a 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

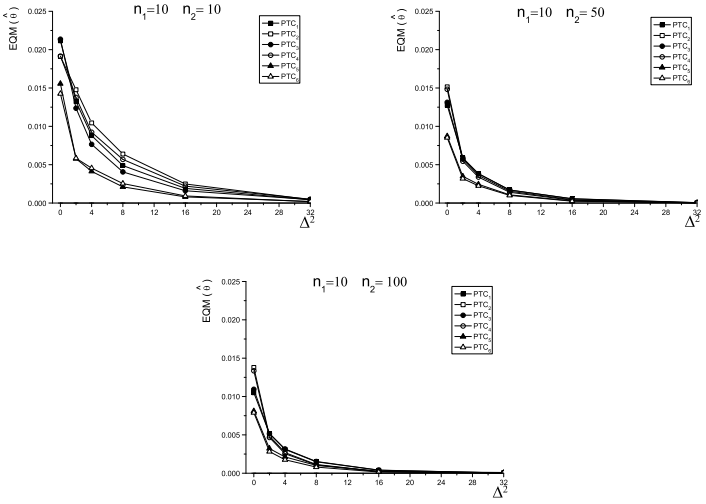

Figura 3 - Erros quadráticos médios estimados para os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis e com tamanhos amostrais $n_1 = 10$ e $n_2 = 10$, 50 e 100 e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

a de seus concorrentes diretos. Assim, em situações de pequeno número de variáveis, esse estimador se destacou e pode ser recomendado.

Nas Tabelas 3 e 4 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, na medida que os tamanhos amostrais aumentaram, os resultados dos *EQMs* diminuíram, 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 supervalorizada 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,

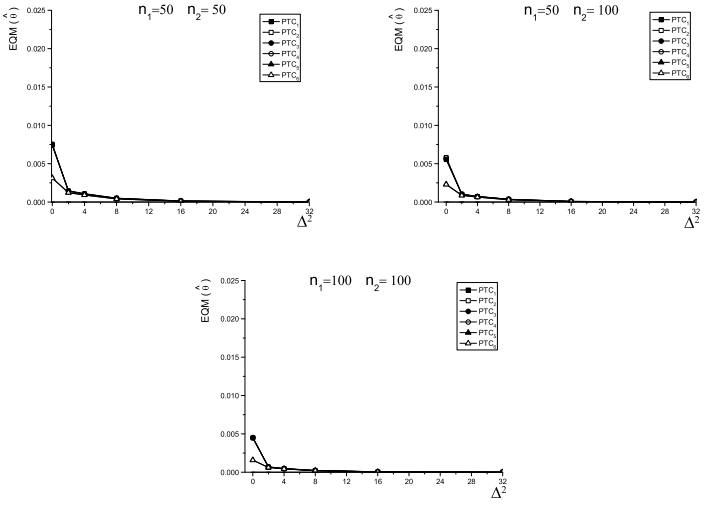

Figura 4 - Erros quadráticos médios estimados para os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função da distância de Mahalanobis Δ^2 entre as médias de 2 populações normais multivariadas com $p = 2$ variáveis e com tamanhos amostrais $n_1 = 50$ e 100 e $n_2 = 50$ e 100 e correlação $\rho = 0{,}5$ em 2000 simulações Monte Carlo.

foi considerado melhor para $p = 10$, 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). Em trabalhos futuros a robustez desses estimadores poderia ser avaliada em situações de populações heterocedásticas. Novos estimadores, acoplando a natureza não-linear, quadrática na verdade, da função discriminante poderiam ser propostos a partir das idéias apresentadas nesse trabalho.

Tabela 3 - Erro quadrático médios estimados para os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função da distância de Mahalanobis $\Delta^2 \leq 4$ 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 $\rho = 0{,}5$ em 2000 simulações Monte Carlo

			Distância de Mahalanobis		
n_1	n_2	Estimadores	$\Delta^2 = 0$	$\Delta^2 = 2$	$\Delta^2 = 4$
10	10	PCT_1	0,0148745	0,0443117	0,0466549
		PCT_2	0,0132653	0,0468063	0,0515558
		PCT_3	0,0151242	0,0407170	0,0424097
		PCT_4	0,0133863	0,0434191	0,0476695
		PCT_5	0,0941500	0,0177125	0,0088282
		PCT_6	0,0895933	0,0172278	0,0092929
10	50	PCT_1	0,0087204	0,0154973	0,0107270
		PCT_2	0,0091411	0,0157855	0,0112353
		PCT_3	0,0091505	0,0139798	0,0095902
		PCT_4	0,0088757	0,0142291	0,0099895
		PCT_5	0,0449002	0,0067537	0,0037269
		PCT_6	0,0444298	0,0061775	0,0032022
10	100	PCT_1	0,0074349	0,0127370	0,0078257
		PCT_2	0,0080104	0,0126431	0,0078779
		PCT_3	0,0079039	0,0114713	0,0070415
		PCT_4	0,0076993	0,0113257	0,0070296
		PCT_5	0,0377876	0,0052821	0,0028147
		PCT_6	0,0378863	0,0044118	0,0023098
50	50	PCT_1	0,0037282	0,0030194	0,0021105
		PCT_2	0,0036775	0,0031236	0,0022095
		PCT_3	0,0037256	0,0028659	0,0019606
		PCT_4	0,0036742	0,0029664	0,0020549
		PCT_5	0,0163989	0,0020219	0,0012493
		PCT_6	0,0161580	0,0019473	0,0012157
50	100	PCT_1	0,0027265	0,0019392	0,0012539
		PCT_2	0,0027176	0,0019709	0,0012846
		PCT_3	0,0027376	0,0018733	0,0012026
		PCT_4	0,0027128	0,0018879	0,0012120
		PCT_5	0,0124275	0,0012703	0,0008442
		PCT_6	0,0122996	0,0011929	0,0007903
100	100	PCT_1	0,0020273	0,0010385	0,0006983
		PCT_2	0,0020146	0,0010614	0,0007177
		PCT_3	0,0020267	0,0010033	0,0006643
		PCT_4	0,0020138	0,0010243	0,0006821
		PCT_5	0,0080529	0,0008480	0,0005187
		PCT_6	0,0079938	0,0008246	0,0005083

Tabela 4 - Erro quadrático médios estimados para os estimadores $PTCI_1$, $PTCI_2$, $PTCI_3$, $PTCI_4$, $PTCI_5$ e $PTCI_6$ em função da distância de Mahalanobis $\Delta^2 \geq 8$ 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 $\rho = 0{,}5$ em 2000 simulações Monte Carlo

			Distância de Mahalanobis		
n_1	n_2	Estimadores	$\Delta^2 = 8$	$\Delta^2 = 16$	$\Delta^2 = 32$
10	10	PCT_1	0,0383786	0,0250923	0,0098026
		PCT_2	0,0447850	0,0312780	0,0131672
		PCT_3	0,0339981	0,0224134	0,0095009
		PCT_4	0,0409161	0,0292419	0,0133454
		PCT_5	0,0028735	0,0014265	0,0010654
		PCT_6	0,0037331	0,0020685	0,0013016
10	50	PCT_1	0,0052306	0,0016042	0,0001865
		PCT_2	0,0053056	0,0015705	0,0001151
		PCT_3	0,0047911	0,0015049	0,0001980
		PCT_4	0,0046767	0,0013595	0,0001042
		PCT_5	0,0013150	0,0001875	0,0000087
		PCT_6	0,0011929	0,0001755	0,0000048
10	100	PCT_1	0,0035375	0,0008444	0,0000868
		PCT_2	0,0032209	0,0006123	0,0000323
		PCT_3	0,0032609	0,0008087	0,0000890
		PCT_4	0,0028580	0,0005353	0,0000280
		PCT_5	0,0010584	0,0001741	0,0000112
		PCT_6	0,0008437	0,0001195	0,0000034
50	50	PCT_1	0,0010767	0,0002930	0,0000184
		PCT_2	0,0011264	0,0002954	0,0000159
		PCT_3	0,0009734	0,0002581	0,0000159
		PCT_4	0,0010197	0,0002601	0,0000136
		PCT_5	0,0005069	0,0000854	0,0000231
		PCT_6	0,0005090	0,0000890	0,0000021
50	100	PCT_1	0,0006121	0,0001531	0,0000089
		PCT_2	0,0006074	0,0001442	0,0000071
		PCT_3	0,0005816	0,0001437	0,0000083
		PCT_4	0,0005623	0,0000597	0,0000062
		PCT_5	0,0003528	0,0000597	0,0000019
		PCT_6	0,0003316	0,0000577	0,0000017
100	100	PCT_1	0,0003692	0,0000855	0,0000045
		PCT_2	0,0003805	0,0000852	0,0000040
		PCT_3	0,0003447	0,0000781	0,0000040
		PCT_4	0,0003552	0,0000779	0,0000036
		PCT_5	0,0002346	0,0000424	0,0000013
		PCT_6	0,0002351	0,0000433	0,0000012

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 Estimadores propostos foram mais eficientes que a modificação proposta por Oliveira e Ferreira (2008);
4. O método 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;

Agradecimentos

Agradecemos à Secretaria Estadual de Educação do Acre (SEE/AC) e ao CNPq pelo suporte financeiro concedido ao primeiro e segundo autores, respectivamente.

BRAGA, A. da S.; FERREIRA, D. F.; RAMOS, P. S. Estimators of the overall misclassification probability in discriminant analysis. *Rev. Mat. Estat.,* São Paulo, v.27, n.2, p.179-196, 2009.

▪ *ABSTRACT: 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 misclassification probability (OMP) were proposed, using jackknifing methods and whose performance was assessed through Monte Carlo simulation. In the present work, the performance of OMP_1, OMP_2, OMP_3, OMP_4, OMP_5 and OMP_6 estimators was compared for two homoscedastic multivariate normal populations, considering the same costs of misclassification and* 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 a common variance estimator into the function which estimates OMP. 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 Γ_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* $p = 2$ *and* 10 *variates and correlation coefficient* $\rho = 0$*,* $\rho = 0.5$ *and* $\rho = 0.9$*. The estimators* OMP_5 *and* OMP_6 *underestimated* OMP*, whereas* OMP_1*,* OMP_2*,* OMP_3 *and* OMP_4 *overestimated it. The* OMP_3*,* OMP_4*,* OMP_5 *and* OMP_6 *estimators were more efficient than* OMP_1 *one, originally proposed by Lachenbruch & Mickey (1968). The* OMP_3 *estimator with heterogeneous variance estimators was considered optimum, due to the smallest positive bias.*

▪ *KEYWORDS: Multivariate; estimator; classifying; simulation;* Jackknife.

Referências

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

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

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

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

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

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

OLIVEIRA, I. R. C. de; 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. *R: a language and enviroment for statistical computing*. Vienna: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: dez. 2007.

Recebido em 22/12/2008.

Aprovado após revisão em 13/07/2009.