CIÊNCIA e NATURA

UFSM

Ci. e nat., Santa Maria, v. 42
Commemorative Edition: Statistic, e14, 2020
DOI: 10.5902/2179460X40384
ISSN 2179-460X

Statistics

Submissão 26/09/19 aprovação 30/09/19 publicação 30/11/20

Análise discriminante aplicada no estudo dos escores de classificação do concurso vestibular 2007 na UFSM

Discriminant analysis applied in the study of 2007 entrance exam classification scores at UFSM

Paulo Roberto Machado Calil[I], Denis Altieri de Oliveira Moraes[II], Ivanor Müller[III], Fernando de Jesus Moreira Junior[IV], Angela Pelegrin Ansuj[V]

RESUMO

O objetivo deste artigo é avaliar o grau de discriminação entre três grupos de candidatos ao Concurso Vestibular da Universidade Federal de Santa Maria (UFSM): não selecionados, selecionados e classificados, dos cursos de Medicina, Direito e Música por meio da Análise Discriminante. O poder de discriminação entre os candidatos pode ser entendido como um índice, o qual é o valor da função discriminante calculado para cada candidato, baseado em suas notas de provas objetivas e redação. A partir dessas notas, foram obtidas duas funções discriminantes e três funções de classificação para os três grupos (não-selecionados, selecionados e classificados), e uma função discriminante e duas funções de classificação para dois grupos (selecionados e classificados). O valor da função de classificação determina a alocação do candidato em uma das três possíveis classes. Conclui-se que a Análise Discriminante possui grande capacidade de discriminação quando os grupos são compostos por amostras de proporções grandes e medianas em relação ao número de parâmetros estimados.

Palavras-chave: Discriminação, Índice Discriminante, Funções de Classificação, Processo Seletivo

ABSTRACT

The main of this paper is to evaluate the degree of discrimination between three groups of candidates to Entrance Exam for the Federal University of Santa Maria (UFSM): non-selected, selected and classified, from the courses of Medicine, Law and Music through Discriminant Analysis. The power of discrimination between candidates can be understood as an index, which is the value of the discriminant function calculated for each candidate based on their objective test scores and wording. From these grades, two discriminant functions and three classification functions were obtained for the three groups (unselected, selected and classified), and one discriminant function and two classification functions for two groups (selected and classified). The rank function value determines the candidate's allocation in one of three possible classes. It is concluded that the Discriminant Analysis has great discrimination capacity when the groups are composed of samples of large and median proportions in relation to the number of estimated parameters.

Keywords: Discrimination, Discriminant Index, Classification Functions,. Selective Process.

[I] Universidade Federal de Santa Maria, Brasil; e-mail: paulo-calil@susepe.rs.gov.br;
[II] Universidade Federal de Santa Maria. Brasil. e-mail: denis.altieri@gmail.com;
[III] Universidade Federal de Santa Maria, Brasil; e-mail: ivanormuller@smail.ufsm.br;
[IV] Universidade Federal de Santa Maria. Brasil. e-mail: fmjunior777@yahoo.com.br;
[V] Universidade Federal de Santa Maria, Brasil; e-mail: angelaansuj@yahoo.com

1 INTRODUÇÃO

O ingresso ao ensino superior no Brasil sempre foi regido por meio da aplicação de exames eliminatórios devido a poucas vagas disponibilizadas pelas Instituições de Ensino Superior (IES) (ANHAIA; MORCHE, 2007). O processo de seleção para ingresso ao ensino superior tem caráter institucional proveniente de políticas federais que visam o reconhecimento do mérito acadêmico (BRASIL, 1999). O processo consiste de provas escritas que abrangem várias matérias denominado Concurso Vestibular (CV), onde o rendimento obtido pelo candidato determina o preenchimento do número de vagas oferecido pela instituição. No entanto, de acordo com Schlichting et al. (2004), o exame vestibular tem sofrido críticas por ocorrer em um único momento e quando os candidatos estão concluindo o segundo grau, onde eles têm um período de tempo muito curto para se preparar. Nesse sentido, surge outro questionamento: o concurso vestibular é adequado para classificar os candidatos qualificados e não qualificados? A Universidade Federal de Santa Maria (UFSM), através da Comissão Permanente de Vestibular (COPERVES), utilizava o Concurso Vestibular como uma forma de ingresso no ensino superior. Assim, havia a necessidade de avaliar a metodologia de classificação dos candidatos nesse processo seletivo.

Esse trabalho teve por objetivo avaliar o grau de discriminação entre três grupos de candidatos: não selecionados, selecionados e classificados, dos cursos de Medicina, Direito e Música. Os dados foram obtidos junto a COPERVES, os quais representam os escores individuais dos candidatos por disciplina. A técnica utilizada para a classificação dos candidatos foi a Análise Discriminante.

2. ANÁLISE DISCRIMINANTE

Segundo Santos e Milioni (2005), a Análise Discriminante “consiste em separar classes de objetos e prever a pertinência do novo objeto a uma classe”. Hair et al. (2005) afirmam que a Análise Discriminante (AD) tem como princípio básico, a estimativa de relação entre uma variável dependente não-métrica (categórica) e um

conjunto de variáveis independentes métricas, cuja forma geral é dada pela expressão:

$$\underset{(não-métrica)}{Y_1} = \underset{(métricas)}{X_1 + X_2 + \ldots + X_n} \qquad (1)$$

A AD determina um conjunto de variáveis transformadas a partir da combinação linear de duas ou mais variáveis independentes, as quais separam os grupos estudados de maneira eficiente. A discriminação é obtida estabelecendo os pesos das variáveis para maximizar a variância entre grupos e minimizar a variância dentro dos grupos. Essa combinação linear é conhecida como função discriminante, representada da seguinte forma:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \ldots + W_n X_{nk} \qquad (2)$$

Onde:

Z_{jk} = escore Z discriminante da função discriminante *j* para o indivíduo *k*;

a = intercepto;

W_i = peso discriminante para a variável independente *i*;

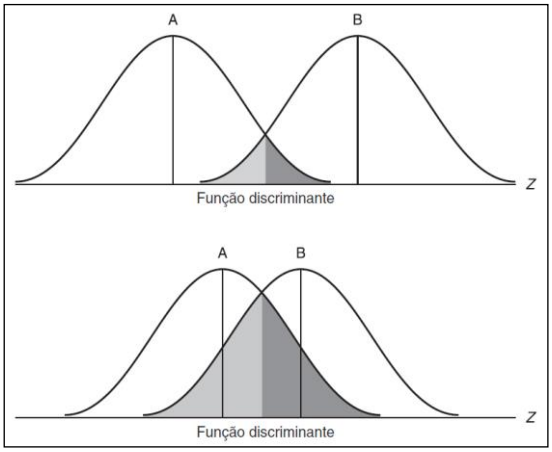
X_{ik} = variável independente *i* para o indivíduo *k*;

i =1,2,...,n.

n = tamanho da amostra.

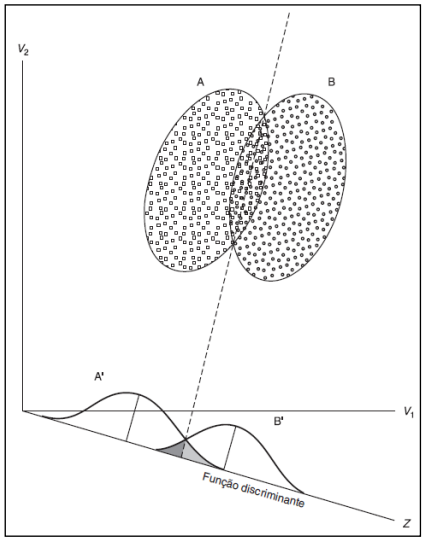
O teste de significância para a função discriminante se dá através da distância entre os centróides (médias) dos grupos, comparando as distribuições dos escores discriminantes dos mesmos. Na Figura 1, há duas distribuições de escores discriminantes. No primeiro diagrama a sobreposição das distribuições é pequena, significando que a função discriminante separa bem os grupos (A e B), enquanto que no segundo a sobreposição é grande, significando que a função discriminante Z não é um bom discriminador entre o grupo A e B.

Figura 1 - Representação univariada de escores Z discriminantes (Fonte: Hair et al., 2005).



Na Figura 2, há uma representação geométrica da função discriminante para dois grupos. Esse tipo de ilustração gráfica possibilita entender melhor a natureza da AD, mostrando o que realmente acontece com uma função. Supondo-se que existam dois grupos, A e B, e duas medidas, V1 e V2, para cada membro dos dois grupos, os pontos pequenos representam as medidas das variáveis do grupo B, enquanto que os pontos grandes representam o grupo A.

Figura 2 - Ilustração gráfica de análise discriminante de dois grupos (Fonte: Hair et al., 2005).



2.1 Processo de Decisão para Análise Discriminante

Segundo Hair et al. (2005), o primeiro passo da análise é definir os objetivos. A seguir abordar questões específicas de planejamento e certificar-se de que as

suposições inerentes estão sendo atendidas. A partir dessas premissas, realizar a dedução da função discriminante e determinar se uma função estatisticamente significativa pode ou não ser obtida para separar dois (ou mais) grupos.

Para implementar a técnica de AD a um conjunto de dados, é necessário averiguar se quatro objetivos são satisfeitos:

I. Na análise, precisa-se de no mínimo dois grupos, para identificar se há diferença estatisticamente significativa entre os perfis de escore médio no conjunto das variáveis estudadas;
II. Definir as variáveis independentes que explicam o máximo de diferenças nos perfis de escore médio dos dois ou mais grupos;
III. Definir os procedimentos para classificar indivíduos em grupos, tendo como base seus escores em um conjunto de variáveis independentes;
IV. Por fim, deve-se definir o número e a composição das dimensões de discriminação dos grupos formados, baseando-se no conjunto de variáveis independentes.

2.2 Tamanho amostral

Para que a AD tenha um bom resultado, deve ser levada em consideração a seleção de variáveis dependentes e independentes e o tamanho da amostra para se estimar as funções discriminantes e a sua divisão. Em muitos casos a amostra é dividida em duas sub-amostras, uma usada para a estimação da função discriminante e a outra para fins de validação. É fundamental que cada uma delas, tenha o tamanho adequado. O procedimento padrão divide a amostra total aleatoriamente em dois grupos: em amostra de análise, ou treinamento; e em amostra de teste. A primeira amostra é usada para desenvolver a função discriminante, ou treinar o classificador. A segunda é necessária para testar a adequação da função discriminante. Esse procedimento de validação da função é chamado de validação cruzada ou partição da amostra.

Um outro método de validação cruzada, bastante utilizado na prática e também usado em diversos pacotes estatísticos é conhecido como *Leave-One-Out* (LOO). Neste

método, apenas um indivíduo é removido do grupo total de cada vez. Após isso, este indivíduo que foi separado é testado nas funções discriminantes para verificar se ele é classificado corretamente ou não.

Por último, é também usado o método conhecido como resubstituição, o qual utiliza todos os indivíduos para calcular os coeficientes das funções discriminantes. Após esta etapa, toda a amostra é novamente utilizada para testar o grau de acurácia, isto é, classificação correta, obtido pela AD.

2.3 Suposições básicas

As suposições necessárias para determinar a função discriminante: normalidade multivariada das variáveis independentes; matrizes de variância e covariância desconhecida (neste trabalho, consideradas supostamente iguais). Matrizes de covariância desiguais podem afetar negativamente o processo de classificação. A não-adequação dos dados quanto à normalidade também pode causar problemas na estimação das funções discriminantes. A falta de multicolinearidade entre as variáveis independentes pode ser outro fator a afetar os resultados.

Hair et al. (2005) e Teixeira (2006), afirmam que o teste de MANOVA é necessário para verificar se as variáveis independentes são significativas. Caso elas não apresentem significância estatística, são eliminadas do estudo.

2.4 Estimação das funções discriminantes

Segundo Hair et al. (2005), para se definir uma função discriminante, deve-se decidir o método de estimação e, conseqüentemente, determinar o número de funções. Com as funções estimadas, o ajuste geral do modelo pode ser avaliado de diversas maneiras. O escore *Z* discriminante, também chamado de escore *Z*, pode ser calculado para cada indivíduo. A comparação das médias dos grupos nos escores *Z* fornece uma medida de discriminação entre grupos.

Dois métodos computacionais são usados para determinar a função discriminante: o método simultâneo e o método seqüencial. O primeiro método inclui todas as variáveis independentes na análise. O segundo envolve a inclusão das

variáveis independentes na função discriminante, uma por vez, com base em seu poder discriminatório (HAIR et al., 2005).

Muitos critérios são empregados para determinar a significância da variável. Quando é aplicado o método seqüencial, utilizam-se as medidas D^2 de *Mahalanobis* e *V* de *Rao*. A primeira baseia-se na distância euclidiana quadrada generalizada, a qual se adapta a variâncias desiguais. O nível de significância convencional é de 0,05 ou acima. Então, assim que as funções discriminantes são definidas, deve-se verificar o ajuste geral das funções encontradas. Primeiramente, calcula-se o escore *Z* discriminante para cada observação. A seguir avaliam-se as diferenças de grupos nos escores *Z* discriminantes, e, por fim, avalia-se a precisão de previsão de pertinência aos grupos. Os escores *Z* para qualquer função discriminante podem ser calculados para cada indivíduo, conforme a equação (2). Depois que a função é ajustada, deve-se avaliar as diferenças entre os grupos. A avaliação da precisão preditiva de pertinência ao grupo é necessária para verificar se cada observação foi corretamente classificada (HAIR et al., 2005).

As matrizes de classificação, também conhecidas como matrizes de confusão, são importantes para determinar a habilidade preditiva de uma função discriminante, ou seja, para mostrar uma avaliação mais precisa do poder discriminatório da função, ou acurácia de classificação (HAIR et al., 2005). Conforme Hair et al. (2005), a determinação do escore de corte determina o grupo em que o indivíduo deve ser classificado. A construção das matrizes de classificação define se o escore de corte é ótimo ou não. Esse corte depende do tamanho dos grupos, se eles são iguais ou diferentes. O escore para dois grupos de mesmo tamanho é definido conforme a equação (3), e de tamanhos diferentes conforme a equação (4):

$$Z_{CE} = \frac{Z_A + Z_B}{2} \qquad (3)$$

em que:

Z_{ce} = valor do escore crítico para grupos de mesmo tamanho;

Z_A = centróide de grupo A;

Z_B = centróide de grupo B.

$$Z_{CU} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B} \qquad (4)$$

em que:

Z_{CU} = valor do escore crítico para grupos com tamanhos diferentes;

N_A = número no grupo A;

N_B = número no grupo B;

Z_A = centróide para o grupo A;

Z_B = centróide para o grupo B.

As Figuras 3 e 4 mostram os escores de corte ótimo para grupos iguais e diferentes.

Figura 3 - Escore de corte ótimo com iguais tamanhos de amostra (Fonte: Hair et al., 2005).

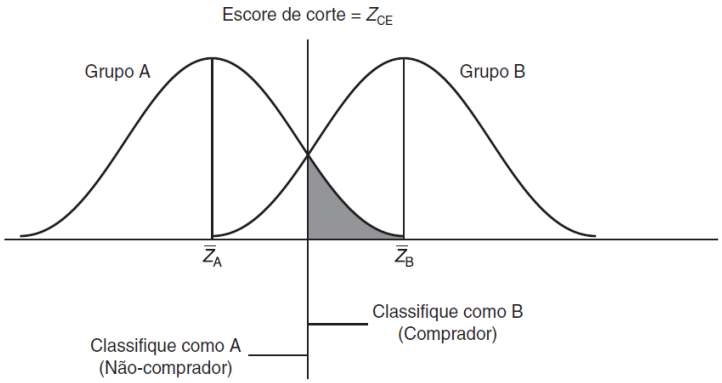
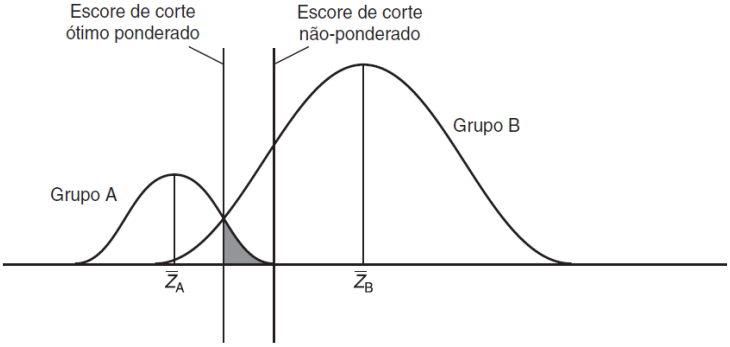


Figura 4 - Escore de corte ótimo com tamanhos diferentes de amostra (Fonte: Hair et al., 2005).



Para validar a função discriminante, devem-se obter amostras aleatórias, criando-se dois grupos: Um grupo (amostra de análise) é utilizado para a obtenção da função discriminante; e o outro (amostra de teste) para validar a função, criando a matriz de classificação. O critério de classificação de cada indivíduo no grupo é definido como:

Classifique um indivíduo no grupo A se $Z_n < Z_{ct}$ Classifique um indivíduo no grupo B se $Z_n > Z_{ct}$

em que:

Z_n : escore Z discriminante para o n-ésimo indivíduo

Z_{ct}: valor do escore de corte crítico

O teste "t" para o procedimento de classificação é definido conforme a Eq. (5):

$$t = \frac{p - 0,5}{\sqrt{\frac{0,5 \cdot 0,5}{n}}} \qquad (5)$$

em que:

p = proporção corretamente classificada

n = tamanho da amostra

2.5 Interpretação dos resultados

O pesquisador deve determinar a importância relativa de cada variável independente na discriminação entre os grupos. Para isso, utiliza os pesos discriminantes padronizados, cargas discriminantes ou contribuições das variáveis preditoras, a cada função separadamente e valores F parciais. No primeiro caso, analisa-se o sinal e a magnitude do peso discriminante das funções discriminantes. No segundo, mede-se a correlação linear simples entre cada variável independente e

a função discriminante. E, no terceiro, aponta-se o nível associado de significância para cada variável.

3 METODOLOGIA

Para o presente estudo foi utilizado o banco de dados dos escores das provas objetivas e da prova de redação dos candidatos ao Concurso Vestibular 2007 da UFSM, fornecidos pela Comissão Permanente de Vestibular (COPERVES). Dos 21.053 candidatos inscritos compareceram 18.020 candidatos durante todo o concurso.

As variáveis independentes são representadas pelo número de questões corretas em cada uma das onze provas objetivas que constituem o vestibular, a saber: Biologia, Física, Química, Geografia, História, Literatura Brasileira, Língua Estrangeira, Língua Portuguesa, Matemática, Filosofia. Além destas, há também a prova de Redação, a qual é avaliada somente para o grupo de candidatos "selecionados", os quais serão posteriormente divididos entre apenas "selecionados" e aqueles que serão considerados como "classificados".

A variável dependente é definida como situação, ou seja, o grupo dos candidatos não-selecionados[1], selecionados[2] e classificados[3], representados por 0, 1 e 2, respectivamente. Dos sessenta e seis cursos existentes em 2007 na UFSM, foram escolhidos apenas três cursos, sendo um com um número grande de inscritos (Medicina), outro médio (Direito) e outro com poucos inscritos (Música). A Tabela 1 apresenta as variáveis do estudo.

Foram utilizadas a estatística descritiva, da Análise de Variância Multivariada (MANOVA) e **o** teste t e a técnica da Análise Discriminante para a análise dos dados com o auxílio do Software Statistica. O nível de significância utilizado nos testes estatísticos foi de 5%.

[1]Candidatos não-selecionados para a correção da prova de redação.
[2]Candidatos selecionados para a correção da prova de redação, mas não classificados no Vestibular.
[3]Candidatos classificados na prova de redação e, portanto, no Vestibular.

Tabela 1 - Variáveis estudas na pesquisa.

Variável	Tipo da variável	
Código do curso	identificadores	Numérico
Nome do curso	identificadores	*String*
Inscritos	identificadores	1-18020
Sexo	identificadores	*Dummy*
Idade	identificadores	0-107
Bio	independente	0-15
Fis	independente	0-15
Qui	independente	0-15
Geo	independente	0-15
His	independente	0-15
Lit	independente	0-15
Lin Est	independente	0-15
LP	independente	0-15
Mat	independente	0-15
Fil	independente	0-15
Red	independente	0-7,5
Sit	dependente	0, 1, 2
Class	identificadores	*Rank*

Fonte: Setor de Estatística e Informática da COPERVES.

4. RESULTADOS

4.1 Análise Preliminar

Apresenta-se a seguir os resultados da estatística descritiva, da Análise de Variância Multivariada (MANOVA) e do teste t. para os três cursos analisados.

4.1.1 Medicina

O curso de Medicina apresentou freqüência de 2.015 candidatos. Deste total, 1.176 são do sexo feminino e 839 do masculino. A idade média dos candidatos ao curso de Medicina foi de 20,3 anos, mediana de 20,0 anos e desvio padrão de 3,0 anos; freqüência de 1.826 (90,6%) candidatos não-selecionados, 109 (5,4%) selecionados e 80 (4%) classificados.

As notas médias por prova entre os três grupos, foram inicialmente comparadas através da Análise de Variância Multivariada (MANOVA). Aplicando o teste *Lambda* de *Wilks* ao conjunto de dados do curso de Medicina, verificou-se a significância do fator situação (valor-p < 0,05). Deste modo, existe diferença significativa entre as notas médias por prova para pelo menos um dos três grupos.

Para identificar qual dos grupos apresenta média estatisticamente diferente dos demais foi utilizado o teste da diferença mínima significativa (DMS).

Comparando-se os três grupos nas provas de Biologia, Física, Química, Geografia, História, Literatura, Língua Estrangeira, Língua Portuguesa e Matemática para o curso de Medicina, constata-se que o grupo dos não-selecionados possui notas médias inferiores ao grupo dos selecionados e dos classificados.

Assim, considerando a nota média geral dos candidatos por prova nos três grupos, conclui-se que o grupo dos não-selecionados apresenta notas médias inferiores em todas as provas, se comparados aos grupos dos selecionados e dos classificados. Entretanto, o grupo dos selecionados e dos classificados apresenta notas médias estatisticamente iguais em quase todas as provas. Apenas na prova de Filosofia o grupo dos selecionados possui nota média estatisticamente inferior ao grupo dos classificados (valor-p < 0,01).

Prosseguindo a análise, foi utilizado o teste "*t*" para a comparação entre as notas médias de redação para os grupos selecionados e classificados, já que o grupo dos não-selecionados não possui a redação. Novamente, a nota média para o fator situação neste caso é significativa (valor-p < 0,01), portanto, a nota média de redação do grupo dos selecionados é inferior à nota média dos classificados. A nota da prova de redação discrimina o grupo dos selecionados dos classificados, já que o grupo dos selecionados possui 9,3 pontos em nota média a menos que o grupo dos classificados.

4.1.2 Direito

O curso de Direito (diurno) apresentou freqüência de 483 candidatos. Deste total, 281 são do sexo feminino e 202 do masculino. Este curso apresenta idade média de 20,7 anos, mediana de 19,0 anos e desvio padrão de 6,8 anos; freqüência de 419 (86,7%) candidatos não-selecionados, 32 (6,6%) selecionados e 32 (6,6%) classificados. A estatística *Lambda* de *Wilks* no curso de Direito constatou significância para o fator situação. Houve também significância estatística através da análise de variância para todas as provas, rejeitando-se a hipótese nula de que a nota média por

prova dos três grupos é igual, a favor da hipótese alternativa de que pelo menos uma das notas médias dos três grupos difere das demais.

Comparando-se as notas médias das provas objetivas para os três grupos, observou-se que há diferença significativa entre o grupo dos não-selecionados e dos selecionados. As notas médias do grupo dos não-selecionados são inferiores às notas médias do grupo dos selecionados e dos classificados (valor-p < 0,05). Comparando-se o grupo dos selecionados com o dos classificados, constata-se uma diferença significativa para as provas de Biologia (valor-p < 0,01), Física (valor-p < 0,001), Química (p < 0,03) e Matemática (p < 0,001), e uma diferença não significativa para as provas de Literatura (p = 0,636), Língua Estrangeira (p = 0,509), Língua Portuguesa (p = 0,483) e Filosofia (p = 0,567).

O teste "*t*" foi utilizado para comparar as notas médias das redações entre os dois grupos com a prova de redação. Com nota média de 53,09 para os selecionados e 68,03 para os classificados, foi identificado diferença significativa nas notas de redação. Desse modo, a prova de redação tem alto poder de discriminação entre os dois grupos (valor-p < 0,01), já que o grupo dos selecionados apresenta nota média 14,94 pontos a menos que o grupo dos classificados.

4.1.3 Música

O curso de Música (Licenciatura Plena) foi selecionado para participar das análises por apresentar baixa densidade de candidatos, observando-se apenas 37 candidatos inscritos. Deste total, 9 são do sexo feminino e 28 do masculino. A idade média é de 22,7 anos, mediana de 21,0 anos e desvio padrão de 8,0 anos, freqüência de 13 (35,1%) candidatos não-selecionados, 12 (32,4%) selecionados e 12 (32,4%) classificados. Apesar disto, este curso também apresentou significância estatística na análise de variância para todas as provas. A estatística *Lambda* de *Wilks* para o curso de Música é significativa para o fator situação.

Comparando-se os três grupos para todas as provas quanto à nota média, também foi observado que há diferença significativa entre o grupo dos não-selecionados e dos selecionados. Portanto, as notas médias do grupo dos não-

selecionados são inferiores às notas médias do grupo dos selecionados e dos classificados. Comparando-se o grupo dos selecionados com o dos classificados com relação às provas objetivas, observa-se uma diferença significativa para nas provas de Literatura (valor-p < 0,05), Língua Estrangeira (valor-p < 0,05), Língua Portuguesa (valor-p < 0,05) e Filosofia (valor-p < 0,05), e uma diferença não significativa para as provas de Biologia (valor-p = 0,336), Física (valor-p = 0,183), Química (valor-p = 0,219), Geografia (valor-p = 0,180), História (valor-p = 0,062) e Matemática (valor-p = 0,229).

O teste "*t*" para grupo dos selecionados e dos classificados, apresentou diferença de 8,92 pontos na prova de Redação, a qual não foi significativa a 5% (valor-p = 0,097), apesar do valor de significância exato do teste levar a crer que existe uma tendência de que os classificados apresentem notas em média superiores.

4.2 Análise Discriminante

Dos três métodos de classificação que podem ser usados, resubstituição, validação cruzada por *Leave-One-Out* e validação cruzada com 50% das amostras para análise e 50% para teste, neste trabalho serão aplicados os dois primeiros.

4.2.1 Medicina

Considerando apenas as provas objetivas, a AD para os três grupos estimou duas funções discriminantes, sendo a primeira função discriminante responsável por 98,6% da variabilidade total dos casos.

As duas funções discriminantes para o curso de Medicina, estimadas com os três grupos, foram as seguintes:

$$Z_{1k} = -3{,}236 - 0{,}045\text{Bio} + 0{,}111\text{Fis} + 0{,}082\text{Qui} - 0{,}009\text{Geo} + 0{,}048\text{His} + 0{,}116\text{Lit} + 0{,}007\text{LE} - 0{,}034\text{LP} + 0{,}064\text{Mat} + 0{,}110\text{Fil}$$

$$Z_{2k} = -0{,}941 - 0{,}147\text{Bio} - 0{,}040\text{Fis} + 0{,}072\text{Qui} - 0{,}022\text{Geo} + 0{,}071\text{His} - 0{,}057\text{Lit} + 0{,}076\text{LE} - 0{,}030\text{LP} - 0{,}206\text{Mat} + 0{,}438\text{Fil}$$

em que:

Z_{1k}: primeira função discriminante para o *k*-ésimo indivíduo

Z_{2k}: segunda função discriminante para o *k*-ésimo indivíduo

A função de classificação é definida pelas notas das provas de cada candidato que será aplicada na função de classificação do grupo dos não-selecionados, dos selecionados e dos classificados. Portanto, a função que mais atribuir probabilidade ao candidato será o grupo ao qual pertencerá. A primeira função refere-se ao grupo dos não-selecionados (0), a segunda ao dos selecionados (1) e a terceira ao dos classificados (2):

$$Z_k(0) = -10{,}440 + 0{,}950Bio - 0{,}041Fis - 0{,}016Qui + 0{,}373Geo + 0{,}069His + 0{,}059Lit + 0{,}081LE + 0{,}420LP - 0{,}559Mat + 0{,}730Fil$$

$$Z_k(1) = -17{,}350 + 0{,}899Bio + 0{,}166Fis + 0{,}115Qui + 0{,}362Geo + 0{,}140His + 0{,}279Lit + 0{,}079LE + 0{,}365LP - 0{,}402Mat + 0{,}837Fil$$

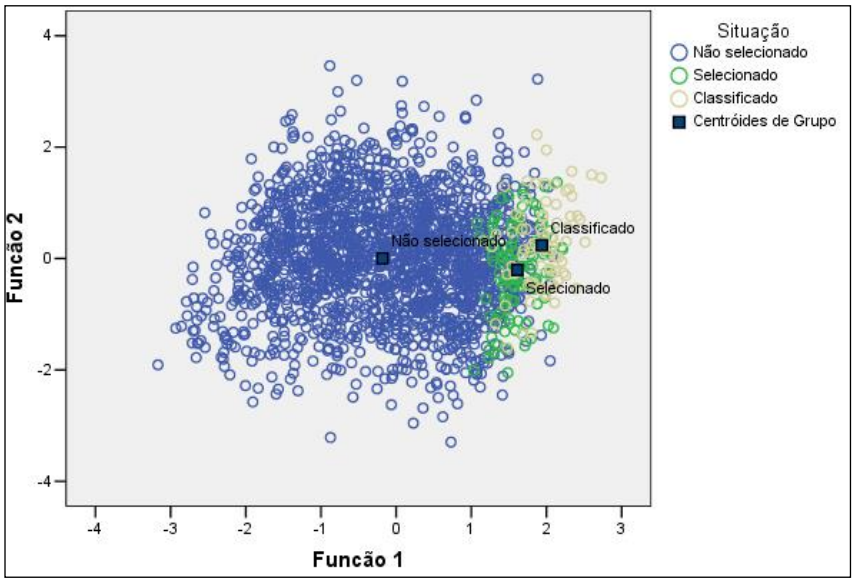
$$Z_k(2) = -19{,}415 + 0{,}819Bio + 0{,}185Fis + 0{,}174Qui + 0{,}349Geo + 0{,}187His + 0{,}292Lit + 0{,}115LE + 0{,}341LP - 0{,}473Mat + 1{,}067Fil$$

em que:

$Z_k(0)$: função de classificação do grupo 0
$Z_k(1)$: função de classificação do grupo 1
$Z_k(2)$: função de classificação do grupo 2

A ilustração gráfica da Figura 5 permite a visualização dos candidatos em torno do centróide para os três grupos conjuntamente.

Figura 5 – Distribuição dos grupos nas funções discriminantes para o curso de Medicina.



São apresentados na Tabela 2 os resultados de classificação obtidos na matriz de confusão, resumindo os dois métodos empregados na pesquisa: resubstituição e *Leave-One-Out*, observando-se os valores da sua diagonal principal.

Tabela 2 – Matriz de confusão: Medicina (provas objetivas).

			Grupo Predito			
Método		Situação	Não selecionado	Selecionado	Classificado	Total
Resubstituição	Nº de casos	Não selecionado	1384	331	111	1826
		Selecionado	0	72	37	109
		Classificado	0	28	52	80
	%	Não selecionado	75,8	18,1	6,1	100
		Selecionado	0	66,1	33,9	100
		Classificado	0	35	65	100
Leave-One-Out	Nº de casos	Não selecionado	1384	331	111	1826
		Selecionado	0	65	44	109
		Classificado	0	34	46	80
	%	Não selecionado	75,8	18,1	6,1	100
		Selecionado	0	59,6	40,4	100
		Classificado	0	42,5	57,5	100

O primeiro método classifica corretamente 1.384 (75,8%) dos 1.826 candidatos do grupo dos não-selecionados, 72 (66,1%) dos 109 candidatos do grupo dos selecionados e 52 (65%) dos 80 candidatos do grupo dos classificados, com média geral de 74,8% para os três grupos. O grau de acurácia obtido pelo método *LOO* foi de 1.384 candidatos (75,8%) para o grupo dos não-selecionados, 65 (59,6%) para o dos selecionados e 46 (57,5%) para o dos classificados. No segundo método (LOO), o percentual médio de classificação correto é menor (74,2%) do que aquele obtido pelo método da resubstituição. Apesar disto, foi realizado um teste para comparar a média geral entre os dois métodos e concluiu-se que não há diferença significativa ($\alpha < 0,05$).

Analisando o método da resubstituição através da matriz de classificação, observa-se que 33,9% dos candidatos selecionados poderiam ser enquadrados como classificados, e por outro lado, 35% dos candidatos classificados poderiam ser enquadrados como selecionados. Tal confusão pode ser explicada por dois motivos: (i) pelo empate na pontuação final (somatório do número de acertos nas provas de múltipla escolha com o número de acertos da prova de redação); (ii) pela mínima diferença entre os escores de cada candidato. No entanto, percebe-se que a diferença mínima influenciará somente o curso de Medicina, observada entre as posições 79, 80

e 81 (mesma pontuação final). Neste caso, a COPERVES utiliza diversos critérios de desempate, os quais não são baseados nas notas das provas e não serão descritos neste trabalho.

Considerando também a prova de Redação para os dois grupos (candidatos selecionados e classificados), a Análise Discriminante gera apenas uma função discriminante.

$$Z_{1k} = -48{,}755 + 0{,}399\text{Bio} + 0{,}346\text{Fis} + 0{,}306\text{Qui} + 0{,}409\text{Geo} + 0{,}424\text{His} + 0{,}351\text{Lit} + 0{,}410\text{LE} + 0{,}479\text{LP} + 0{,}385\text{Mat} + 0{,}442\text{Fil} + 0{,}027\text{Red}$$

As duas funções de classificação, incluindo também a prova de redação, são definidas pelas equações:

$$Z_k(0) = -1232{,}415 + 31{,}024\text{Bio} + 18{,}393\text{Fis} + 14{,}264\text{Qui} + 21{,}141\text{Geo} + 19{,}009\text{His} + 18{,}942\text{Lit} + 20{,}716\text{LE} + 24{,}146\text{LP} + 20{,}510\text{Mat} + 16{,}880\text{Fil} + 0{,}908\text{Red}$$

$$Z_k(1) = -1367{,}700 + 32{,}127\text{Bio} + 19{,}350\text{Fis} + 15{,}110\text{Qui} + 22{,}272\text{Geo} + 20{,}181\text{His} + 19{,}912\text{Lit} + 21{,}847\text{LE} + 25{,}469\text{LP} + 21{,}575\text{Mat} + 18{,}102\text{Fil} + 0{,}982\text{Red}$$

Tabela 3 - Matriz de confusão: Medicina (provas objetivas + redação).

Método	Situação		Grupo Predito		Total
			Selecionado	Classificado	
Resubstituição	Nº de casos	Selecionado	108	1	109
		Classificado	4	76	80
	%	Selecionado	99,1	0,9	100
		Classificado	5	95	100
Leave-One-Out	Nº de casos	Selecionado	107	2	109
		Classificado	9	71	80
	%	Selecionado	98,2	1,8	100
		Classificado	11,25	88,75	100

Nesta fase do processo de classificação, o método da resubstituição classificou corretamente 108 dos 109 (99,1%) candidatos selecionados, ou seja, apenas um candidato foi mal alocado, e dos 80 candidatos do grupo dos classificados, 76 (95%) foram corretamente classificados. Na média geral, tem-se 97,4% de acurácia pelo método da resubstituição e 94,2% pelo método *LOO*. Destaca-se que, para os candidatos ao curso de Medicina no Vestibular de 2007, a única prova objetiva que

apresentou número médio de questões significativamente superior para os candidatos classificados foi a prova de Filosofia. Desse modo, o curso de Medicina foi satisfatoriamente modelado pelo método de classificação da Análise Discriminante, pois obteve alto grau de acurácia. Este bom resultado de classificação deve-se principalmente à alta densidade de candidatos ao curso, isto é, ao tamanho amostral suficiente para estimar corretamente os parâmetros das funções discriminantes.

4.2.2 Direito (diurno)

Considerando a amostra de tamanho mediano formada pelos 483 candidatos ao curso de Direito, a primeira função discriminante foi responsável por 97,5% da variabilidade total, enquanto a segunda função discriminante é responsável por 2,5%. As funções discriminantes para os candidatos ao curso de Direito são as seguintes:

$$Z_{1k} = -3{,}911 + 0{,}130\text{Bio} + 0{,}157\text{Fis} + 0{,}131\text{Qui} + 0{,}016\text{Geo} + 0{,}108\text{His} + 0{,}056\text{Lit} + 0{,}006\text{LE} - 0{,}053\text{LP} + 0{,}104\text{Mat} + 0{,}034\text{Fil}$$

$$Z_{2k} = -1{,}011 - 0{,}199\text{Bio} - 0{,}121\text{Fis} + 0{,}148\text{Qui} + 0{,}165\text{Geo} - 0{,}039\text{His} + 0{,}228\text{Lit} + 0{,}027\text{LE} + 0{,}141\text{LP} - 0{,}306\text{Mat} + 0{,}014\text{Fil}$$

As funções de classificação foram definidas da seguinte forma:

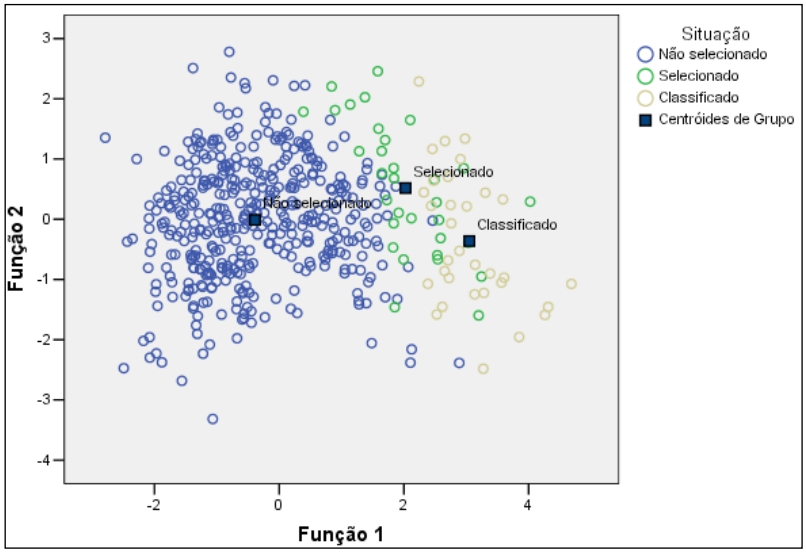
$$Z_k(0) = -9{,}695 + 0{,}929\text{Bio} + 0{,}223\text{Fis} + 0{,}271\text{Qui} + 0{,}344\text{Geo} + 0{,}164\text{His} + 0{,}156\text{Lit} + 0{,}082\text{LE} - 0{,}015\text{LP} - 0{,}110\text{Mat} + 0{,}705\text{Fil}$$

$$Z_k(1) = -21{,}783 + 1{,}136\text{Bio} + 0{,}537\text{Fis} + 0{,}666\text{Qui} + 0{,}470\text{Geo} + 0{,}405\text{His} + 0{,}412\text{Lit} + 0{,}111\text{LE} - 0{,}070\text{LP} - 0{,}022\text{Mat} + 0{,}794\text{Fil}$$

$$Z_k(2) = -27{,}427 + 1{,}444\text{Bio} + 0{,}804\text{Fis} + 0{,}670\text{Qui} + 0{,}342\text{Geo} + 0{,}550\text{His} + 0{,}269\text{Lit} + 0{,}093\text{LE} - 0{,}248\text{LP} + 0{,}353\text{Mat} + 0{,}817\text{Fil}$$

A Figura 6 representa o mapa territorial dos candidatos para os três grupos conjuntamente.

Figura 6 – Distribuição dos grupos nas funções discriminantes para o curso de Direito.



A acurácia de classificação obtido pelo método de resubstituição foi de 360 (85,9%) dos 419 candidatos do grupo dos não-selecionados, 21 (65,6%) dos 32 candidatos do grupo dos selecionados e 23 (71,9%) dos 32 candidatos do grupo dos classificados. A acurácia média para os três grupos neste método foi 83,6%.

Tabela 4 - Matriz de confusão: Direito (provas objetivas).

Método		Situação	Grupo Predito			Total
			Não selecionado	Selecionado	Classificado	
Resubstituição	Nº de casos	Não selecionado	360	49	10	419
		Selecionado	1	21	10	32
		Classificado	0	9	23	32
	%	Não selecionado	85,9	11,7	2,4	100
		Selecionado	3,1	65,6	31,3	100
		Classificado	0	28,1	71,9	100
Leave-One-Out	Nº de casos	Não selecionado	353	56	10	419
		Selecionado	1	20	11	32
		Classificado	0	10	22	32
	%	Não selecionado	84,2	13,4	2,4	100
		Selecionado	3,125	62,5	34,4	100
		Classificado	0	31,3	68,8	100

O método *LOO* classifica corretamente 353 (84,2%) dos 419 candidatos do grupo dos não-selecionados, 20 (62,5%) dos 32 candidatos do grupo dos selecionados e 22 (68,8%) dos 32 candidatos do grupo dos classificados, com média geral de 81,8%

para os três grupos. Novamente neste caso não se observou diferença significativa entre as acurácias médias de classificação obtidas pelos dois métodos.

Tabela 5 - Matriz de confusão: Direito (provas objetivas + redação).

Método	Situação		Grupo Predito		Total
			Selecionado	Classificado	
Resubstituição	Nº de casos	Selecionado	28	4	32
		Classificado	2	30	32
	%	Selecionado	87,5	12,5	100
		Classificado	6,3	93,8	100
Leave-One-Out	Nº de casos	Selecionado	25	7	32
		Classificado	3	29	32
	%	Selecionado	78,1	21,9	100
		Classificado	9,4	90,6	100

Incluindo a prova de Redação para os candidatos selecionados e classificados, a função discriminante extraída é:

$$Z_{1k} = -20,018 + 0,339Bio + 0,156Fis + 0,072Qui + 0,118Geo + 0,317His + 0,145Lit + 0,169LE + 0,063LP + 0,185Mat + 0,142Fil + 0,039Red$$

E as duas funções de classificação:

$$Z_k(0) = -242,836 + 7,660Bio + 0,258Fis + 2,222Qui + 2,913Geo + 6,683His + 8,103Lit + 4,209LE + 5,502LP + 2,453Mat + 4,415Fil + 0,580Red$$

$$Z_k(1) = -294,250 + 8,531Bio + 0,659Fis + 2,408Qui + 3,215Geo + 7,499His + 8,475Lit + 4,643LE + 5,665LP + 2,927Mat + 4,779Fil + 0,679Red$$

Nota-se ainda neste caso que o curso de Direito, por apresentar um tamanho amostral mediano, apresentou também acurácia final de classificação próximo de 90%, para ambos os métodos de teste empregados (resubtituição e *LOO*).

4.2.3 Música

Para os 37 candidatos ao curso de Música, os quais compõem o menor tamanho amostral testado neste trabalho, a primeira função discriminante extraída é responsável por 92,4% da variabilidade total, enquanto a segunda por 7,6%. As funções discriminantes para o curso de Música são as seguintes:

$$Z_{1k} = - 7,532 + 0,192Bio + 0,391Fis - 0,267Qui + 0,066Geo + 0,400His + 0,040Lit + 0,514LE - 0,256LP + 0,227Mat - 0,064Fil$$

$$Z_{2k} = - 2,828 - 0,284Bio + 0,096Fis + 0,323Qui - 0,033Geo + 0,026His - 0,085Lit - 0,200LE + 0,346LP - 0,057Mat + 0,406Fil$$

E as funções de classificação:

$$Z_k(0) = - 21,356 + 0,284Bio + 1,881Fis - 0,425Qui + 0,633Geo + 2,086His - 0,542Lit + 2,051LE + 0,109LP + 1,561Mat + 0,829Fil$$

$$Z_k(1) = - 35,219 + 1,042Bio + 2,760Fis - 1,409Qui + 0,829Geo + 3,056His - 0,358Lit + 3,527LE - 0,870LP + 2,184Mat + 0,269Fil$$

$$Z_k(2) = - 53,341 + 1,050Bio + 3,546Fis - 1,504Qui + 0,905Geo + 3,777His - 0,387Lit + 4,188LE - 0,920LP + 2,513Mat + 0,619Fil$$

A Figura 7 representa o mapa territorial para os candidatos ao curso de Música. Nota-se em primeiro lugar a baixa densidade dos candidatos compondo os três grupos.

A Tabela 6 apresenta a matriz de confusão para o curso de Música e os dois métodos testados. Comparando-se a média geral do primeiro método (89,2%) e do segundo (59,5%), observa-se nitidamente que há diferença significativa quando se utiliza o método de resubtituição e *LOO*. Isto porque como a amostra possui poucos elementos, a retirada de um candidato da amostra influencia grandemente os parâmetros das funções discriminantes e consequentemente as funções de classificação. Desse modo, o método da resubtituição apresenta um grau de acurácia relativamente elevado enquanto o método *LOO* apresenta acurácia bem inferior, pois as funções de classificação não são apropriadas para discriminar indivíduos "estranhos", ou seja, indivíduos que não participaram no processo de treinamento do classificador.

Figura 7 – Distribuição dos grupos nas funções discriminantes para o curso de Música.

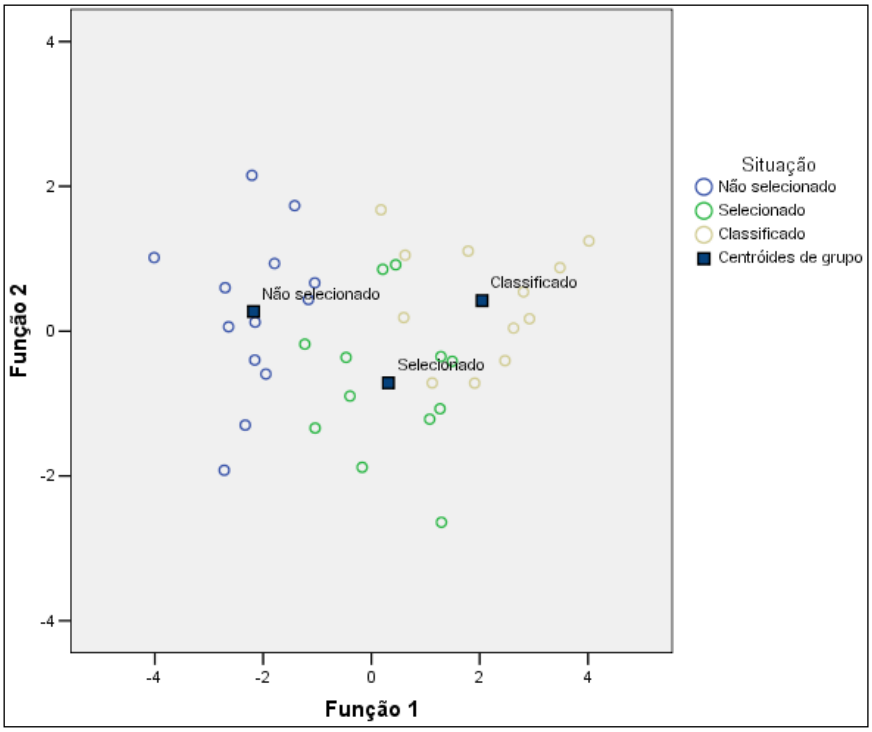


Tabela 6 - Matriz de confusão: Música (provas objetivas).

Método		Situação	Grupo Predito			Total
			Não selecionado	Selecionado	Classificado	
Resubstituição	N° de casos	Não selecionado	13	0	0	13
		Selecionado	1	10	1	12
		Classificado	0	2	10	12
	%	Não selecionado	100	0	0	100
		Selecionado	8,3	83,3	8,3	100
		Classificado	0	16,7	83,3	100
Leave-One-Out	N° de Casos	Não selecionado	11	1	1	13
		Selecionado	2	5	5	12
		Classificado	1	5	6	12
	%	Não selecionado	84,6	7,7	7,7	100
		Selecionado	16,7	41,7	41,7	100
		Classificado	8,3	41,7	50	100

Incluindo a prova de Redação para os selecionados e classificados, a função discriminante estimada é:

$$Z_{1k} = - 13,630 + 0,100Bio + 1,127Fis - 0,782Qui - 0,258Geo + 0,083His + 0,300Lit + 0,338LE - 0,240LP + 0,326Mat + 0,504Fil + 0,109Red$$

E as duas funções de classificação:

$$Z_k(0) = - 77,403 + 1,995Bio + 13,028Fis - 10,259Qui - 2,698Geo + 1,191His + 3,775\ Lit + 5,424LE - 4,121LP + 5,072Mat + 4,818Fil + 1,424Red$$

$$Z_k(1) = - 122,069 + 2,323Bio + 16,721Fis - 12,821Qui - 3,545Geo + 1,463His + 4,758Lit + 6,532LE - 4,907LP + 6,141Mat + 6,468Fil + 1,782Red$$

Tabela 7 - Matriz de confusão: Música (provas objetivas + redação).

Método	Situação		Predicted Group Membership		Total
			Selecionado	Classificado	
Resubstituição	Count	Selecionado	11	1	12
		Classificado	0	12	12
	%	Selecionado	91,7	8,3	100
		Classificado	0	100	100
Leave-One- Out	Count	Selecionado	7	5	12
		Classificado	5	7	12
	%	Selecionado	58,3	41,7	100
		Classificado	41,7	58,3	100

Também neste caso, devido ao tamanho bastante reduzido da amostra, novamente percebe-se um forte decréscimo na acurácia de classificação obtida pelo método de teste através de resubstituição (95,8%) e o método *LOO* (58,3%).

5 CONCLUSÃO

Considerando apenas os escores das provas objetivas, foram obtidas duas funções discriminantes e três funções de classificação para os três grupos (não-selecionados, selecionados e classificados). Ao incluir a prova de Redação, apenas uma função discriminante e duas funções de classificação foram estimadas, as quais discriminam apenas dois grupos (selecionados e classificados).

O modelo ajustado para os candidatos ao curso de Medicina pelo método de classificação da Análise Discriminante obteve alto grau de acurácia, sendo 97,4% de pelo método da resubstituição e 94,2% pelo método *LOO,* devido ao tamanho amostral suficiente para estimar corretamente os parâmetros das funções discriminantes.

O modelo ajustado para os candidatos ao curso de Direito apresentou uma acurácia de classificação próximo de 90% para ambos os métodos de teste empregados (resubtituição e *LOO*), devido ao tamanho amostral mediano.

O modelo ajustado para os candidatos ao curso de Música apresentou uma baixa acurácia, de 58,3% quando é empregado o método *Leave-One-Out*, devido ao baixo número de candidatos. Porém quando se utiliza o método de resubstituição a acurácia se eleva para 95,8%.

Conclui-se que o método de classificação baseada na Análise Discriminante possui um bom poder de classificação no caso do concurso vestibular apenas quando os grupos são compostos de amostras de proporções grandes e medianas em relação ao número de parâmetros estimados. Por outro lado, o método fornece uma baixa acurácia de classificação quando os grupos são formados por amostras pequenas, pois estas acarretam em baixa confiabilidade na estimação dos parâmetros das funções discriminantes e consequentemente um baixo poder discriminatório.

REFERÊNCIAS

ANHAIA BCD, MORCHE B. **Acesso e eqüidade na educação superior: formas de ingresso.** In: Livro de resumos do Salão de Iniciação Científica; 2007; Porto Alegre, Brasil.

BRASIL. Conselho Nacional de Educação - Conselho Pleno. **Parecer n. CP 98/99, aprovado em 06 de julho de 1999.** Regulamentação de Processo Seletivo para acesso a Cursos de graduação de Universidades, Centros Universitários e Instituições Isoladas de Ensino Superior. Brasília: Conselho Nacional de Educação;1999.

HAIR JF, ANDERSON RE, TATHAM RL, BLACK WC. **Análise multivariada de dados.** 5. ed. Porto Alegre: Bookman, 2005.

SANTOS OJS, MILIONI AZ. **Composição de especialistas para classificação de dados**. Inv. Op. 2005, 25(1) 105-121.

SCHLICHTING AM, SOARES DHP, BIANCHETTI L. **Vestibular seriado - análise de uma experiência em Santa Catarina.** Biologia & Sociedade 2004; 16(2):114-26.

TEIXEIRA LL. **O uso de técnica de estatística multivariada no prognóstico de desistência de alunos em IES privados: um estudo de caso na cidade de Foz do Iguaçu-PR** Dissertação (mestrado) - Universidade Federal do Paraná, Setor de Tecnologia e Setor de Ciências Exatas, Programa de pós-graduação em Métodos Numéricos em Engenharia. Defesa: Curitiba, 2006.