LEVI LOPES TEIXEIRA

O USO DE TÉCNICAS DE ESTATÍSTICA MULTIVARIADA NO PROGNÓSTICO DE DESISTÊNCIA DE ALUNOS EM IES PRIVADAS: UM ESTUDO DE CASO NA CIDADE DE FOZ DO IGUAÇU-PR.

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências, Curso de Pós-Graduação em Métodos Numéricos em Engenharia – Programação Matemática, Setores de Tecnologia e Ciências Exatas, Universidade Federal do Paraná.

Orientador: Prof. Dr. Celso Carnieri.

CURITIBA-2006

TERMO DE APROVAÇÃO

Levi Lopes Teixeira

O Uso de Técnicas de Estatística Multivariada no Prognóstico de Desistência de Alunos em IES Privadas: Um Estudo de Caso na Cidade de Foz do Iguaçu-Pr

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Curso de Pós-Graduação em Métodos Numéricos em Engenharia – Área de Concentração em Programação Matemática, Setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador:

_____

Prof. Celso Carnieri, D. Eng.
Programa de Pós-graduação em Métodos Numéricos em Engenharia - UFPR

_____

Prof. Jair Mendes Marques, D. Sc.
Programa de Pós-graduação em Métodos Numéricos em Engenharia - UFPR

_____

Profª. Angela Olandoski Barboza, D. Sc.
Departamento de Matemática da Universidade Tecnológica Federal do Paraná - UTFPR

_____

Profª. Neida Maria Patias Volpi, D. Eng.
Departamento de Matemática da Universidade Federal do Paraná - UFPR

Curitiba, 13 de dezembro de 2006.

Dedico este trabalho à minha mãe Antônia "*gerou-me, educou-me e soube imprimir em meu coração o sentido maior da vida – o amor"*

AGRADECIMENTOS

Ao professor Celso Carnieri pelas orientações, conhecimentos repassados, e por ser um exemplo, seja na pesquisa, na educação ou como pessoa possuidora de valores imprescindíveis ao ser humano.

Ao professor Anselmo Chaves Neto, pelos ensinamentos, incentivos e pelos momentos de descontração nesta árdua caminhada.

Ao professor Jair Mendes Marques, pela sua didática, clareza, e ensinamentos que contribuíram de forma significativa para o desenvolvimento deste trabalho.

À professora Maria Terezinha Arns Steiner pela maneira de ser e ensinar.

Aos professores Liliana Madalena Gramani Cumin, Arinei Carlos Lindbeck da Silva, Neida Maria Patias Volpi e Volmir Eugênio Wilhem, pelos ensinamentos.

À FECILCAM e Universidade Federal do Paraná que juntas viabilizaram a realização do Mestrado.

À minha família, especialmente à minha esposa Lúcia Celina, aos meus filhos Vitória, João Levi e Raquel por entenderem a minha ausência durante o curso.

Aos amigos e colegas que, de alguma maneira, contribuíram com a minha caminhada.

“Jamais considere seus estudos como uma obrigação, mas como uma oportunidade invejável para aprender a conhecer a influência libertadora da beleza do espírito, para seu próprio prazer pessoal e para proveito da comunidade à qual seu futuro trabalho pertencer”.

Alberto Einstein

SUMÁRIO

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

IES	Instituição de Ensino Superior.
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais.
FDL	Função Discriminante Linear

RESUMO

As Instituições de Ensino Superior (IES) privadas e públicas do Brasil têm pela frente um grande desafio: diminuir a evasão escolar. Este trabalho procura contribuir com as discussões que envolvem esse problema tão complexo, dadas as circunstâncias nas quais ele está inserido . Investigou-se variáveis que pudessem discriminar dois grupos, um deles formado por alunos que possuem maiores chances de saírem da IES com a titulação e, o outro, formado por alunos que deixaram a IES sem a titulação. Na busca deste objetivo, aplicou-se um questionário a fim de extrair possíveis diferenças entre os grupos e, para determinar quais variáveis possuíam diferenças significativas. Para tanto, foram utilizadas técnicas estatísticas como a MANOVA. Determinadas as variáveis discriminantes, passou-se para a classificação de indivíduos em um dos dois grupos, com a aplicação da Função Discriminante Linear de Fisher e da Regressão Logística. Por melhor que sejam esses métodos, sempre ocorrem erros de classificação, os quais foram estimados. Um dos métodos aplicados foi o de Lachenbruch. Além das variáveis discriminantes, as funções classificatórias também foram alimentadas por escores fatoriais, comparando-se os resultados obtidos com os dois tipos de dados. A pesquisa foi realizada em uma IES de Foz do Iguaçu – Pr de tamanho médio em comparação com outras Instituições da cidade e, mesmo investigando um universo reduzido, muitos dos resultados obtidos podem ser expandidos para outras instituições. Escores fatoriais e variáveis discriminantes, tais como: tempo para estudos e decepção com o curso escolhido, permitiram a construção de funções para predizer em que grupo um novo indivíduo vai pertencer.

Palavras Chaves: Evasão Escolar, Função Discriminante Linear de Fisher, Regressão Logística, Lachenbruch e Escores Fatoriais.

ABSTRACT

The Brazilian private and public colleges and universities (IES) have a great challenge ahead: diminishing the university evasion. This paper tries to contribute to the discussions which implicate such a complex problem, given the circumstances in which it is inserted. It was investigated variables which could discriminate two groups, one of them formed by students who have better chances to graduate and, the other one, formed by students who left the IES without graduating. In search of this goal, a questionnaire was applied in order to extract possible differences between the groups and, to determine which variables had significant differences. For so much statistical techniques were used, such as the MANOVA. After determining the discriminating variables, happened the individual classification in one of the groups and, with the application of the Fisher Linear Discriminating Function and the Logistics Regression. As good as these methods might be, classification mistakes always happen, which were estimated. One of the methods used was the Lachenbruch's. Besides the discriminating variables, the classificatory functions were also fed on factorial scores, comparing the results obtained with the two kinds of data. The research was performed at one medium size IES in Foz do Iguaçu – PR, compared to other institutions in the city and, even investigating a reduced universe, a lot of the results obtained can be expanded to other institutions. Factorial scores and discriminating variables, such as: time to study and deception with the chosen course, let us build functions to predict to which group a new individual will belong.

Key words: School Evasion, Fisher Linear Discriminating Function, Logistics Regression, Lachenbruch and Factorial Scores.

1 INTRODUÇÃO

Há alguns anos, o Brasil era chamado de “país do futuro”. Pois bem, o futuro chegou e o Brasil não conseguiu diminuir os seus problemas sociais. Especialistas afirmam que um dos componentes desta problemática é a baixa escolaridade do povo brasileiro. Governo e educadores procuram estratégias que diminuam a evasão escolar. Tentou-se a amenização do problema promovendo automaticamente o aluno para a série seguinte, reduzindo a média de aprovação e transferindo toda a culpa da reprovação para o professor. As altas taxas de evasão escolar não são provocadas exclusivamente pela suposta ineficiência do ensino oferecido pelas escolas brasileiras. Aspectos sociais, psicológicos e outros, também devem ser considerados. As Instituições de Ensino Superior, particularmente as privadas, vem encontrando grandes dificuldades nesta questão que, muitas vezes, acarretam o fechamento de cursos.

A estatística é uma ferramenta de grande valor nesta discussão, dado o elevado número de elementos e informações das populações envolvidas. Garimpar dados e analisá-los com métodos apropriados é, com certeza, uma forma de contribuir com as discussões desta problemática.

1.1 O PROBLEMA

É inequívoca a existência dos problemas educacionais no Brasil, haja vista o último teste do PISA[1] no qual o país ficou com a última classificação. O teste avaliou alunos do Ensino Fundamental, mostrando a precariedade do nosso sistema de ensino. Como não poderia deixar de ser, o problema chegou às universidades. As públicas tentam se preservar a partir dos testes seletivos, já as privadas se vêem sem saída para o problema; ou aceitam alunos despreparados ou não pagam as contas no final do mês. O resultado deste despreparo e de outros fatores é um alto índice de desistência. A discussão deste problema passa, entre outras análises, pela identificação dos alunos mais propensos à desistência. Uma questão que se apresenta é a possibilidade de identificar alunos que deixarão o curso sem titulação. Diante deste problema, buscou-se analisar se métodos estatísticos podem estimar,

[1] Programa Internacional de Avaliação Comparada.

com uma certa margem de erro, o número de alunos que se incluirão no grupo dos desistentes?

1.2 OBJETIVOS

OBJETIVO GERAL

Identificar alunos de IES privadas com propensão à desistência, proporcionando aos administradores elementos auxiliadores nas eventuais tomadas de decisão.

OBJETIVOS ESPECÍFICOS

- Aplicar técnicas de Estatística Multivariada.
- Estudar a situação, no que tange as desistências, das IES privadas.
- Apresentar alternativas que contribuam com a minimização do problema de desistências nas IES privadas.
- Analisar o caso de uma IES da região de Foz do Iguaçu-Pr.

1.3 JUSTIFICATIVA

Nos últimos dez anos, houve no Brasil um aumento significativo no número de IES privadas. Particularmente na região de Foz do Iguaçu-Pr, que de duas, passou a ter sete, um aumento de 250%. Um aumento desta magnitude implica em distorções, pois há sinais de que o mercado não exigia tal aumento e uma das evidências é a dificuldade na formação de turmas. Muitas IES possuem cursos autorizados pelo MEC, mas que não funcionam devido a não formação de turmas. São várias as estratégias usadas na montagem das turmas iniciais, acarretando, muitas vezes, na matrícula de alunos que não sabem o que estão fazendo naquela IES e naquele curso. Os discentes foram, de certa forma, levados pelas estratégias de *marketing.* Não é prudente condenar as IES privadas, este é um processo natural pela sobrevivência, mas é um procedimento que, quando agressivo, agrava as desistências. O aluno, em um primeiro momento, coloca-se como possuidor das

capacidades necessárias para ingressar naquela instituição e naquele curso. Com o passar dos meses, ele começa a constatar as suas deficiências na formação escolar, certifica-se das dificuldades financeiras, conclui que não tem afinidade com o curso e que a instituição não lhe oferece elementos que possam convencê-lo do contrário.

A desistência é um fator a ser analisado e minimizado, pois o problema da evasão escolar nas IES privadas não está associado apenas às questões financeiras, no âmbito mais restrito das instituições, mas também à qualidade dos profissionais formados. Diminuir as desistências é um objetivo, mas o uso de estratégias erradas, tais como distribuição de notas – diminuindo as reprovações e em conseqüência as desistências - provoca a diminuição da qualidade dos cursos e implica na formação de profissionais que não estão preparados para o mercado de trabalho. Novamente voltamos à questão econômica, já que este é um custo que será arcado pela sociedade.

Portanto, fazer a predição de alunos propensos à desistência, com o uso de técnicas de Estatística Multivariada é uma maneira de oferecer subsídios para a diminuição da evasão escolar – um problema com várias implicações negativas.

1.4 ESTRUTURA DO TRABALHO

Este tópico abordará as distribuições dos capítulos que comporão o trabalho .

Primeiro capítulo: formado pela descrição do problema, objetivos gerais e específicos, justificativas e a atual seção.

Segundo capítulo: neste capítulo faz-se uma revisão bibliográfica, abordando técnicas estatísticas e as possíveis fontes de pesquisa.

Terceiro capítulo: trata-se dos métodos utilizados no trabalho.

Quarto capítulo: delimita-se a área da pesquisa, descreve-se a população alvo e a metodologia.

Quinto capítulo: usam-se as técnicas apresentadas no capítulo três para o tratamento dos dados e análises.

Sexto capítulo: conclusões e sugestões para trabalhos futuros.

2 REVISÃO DE LITERATURA

A seguir encontra-se uma revisão bibliográfica dos conteúdos tratados neste trabalho, tais como: Educação Superior no Brasil, Evasão Escolar e Estatística.

2.1 EDUCAÇÃO SUPERIOR NO BRASIL

Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) – Censo de 2004 - existem 2.013 instituições de ensino superior no Brasil , sendo 224 públicas e 1.789 privadas. O número de vagas oferecidas nas públicas é de 308.492, já as privadas oferecem um total de 2.011.929, mostrando uma predominância com relação ao número de vagas do ensino superior privado sobre o público. Informa o censo que o número de candidatos inscritos no processo seletivo das instituições públicas é 2.431.388, sendo assim, a relação candidatos por vaga é igual a 7,9 aproximadamente. As instituições privadas tiveram 2.622.604 candidatos inscritos nos vestibulares, ficando a relação candidatos por vaga na ordem de 1,3 . O número de matrículas efetivadas nas públicas é igual a 287.242, isto significa que 93% das vagas ofertadas foram preenchidas. Já nas instituições privadas, o número de ingressos foi de 1.015.868, tendo sido preenchidas 50% das vagas.

Os números mostram a dificuldade das instituições privadas na formação de turmas iniciais, pois apenas 50% das vagas são preenchidas e subtraindo deste número as desistências que ocorrerão nos primeiros períodos, deslumbra-se uma situação no mínimo desconfortável. As IES privadas que sofrem os efeitos desta realidade, procuram minimizar custos com prejuízo na qualidade de seus cursos.

SCHWARTZMAN (1999), conclui que a crescente demanda por vagas no Ensino Superior está sendo atendida pelo setor privado, já que as públicas estão estagnadas neste quesito. Observa também, as dificuldades em promover um ensino de qualidade em massa; um ensino de qualidade exigiria uma taxa relativamente alta de professores por aluno.

SCHWARTMAN (2003) escreve que o crescimento do setor privado é fundamental para o atendimento da demanda e será decisivo para se atingir as metas do Plano Decenal de Educação de prover até o final desta década, educação superior para pelo menos 30% da população na faixa etária de 18 a 24 anos. Isto porque não se espera investimento significativo do setor público federal e estadual , seja pela crise fiscal por que passam, seja pelas insuficiências ainda existentes no

ensino médio e no pré-escolar. A provisão de crédito educativo e outras formas de ajuda a alunos carentes serão decisivas para se atingir a meta para o sistema. Os novos estudantes serão, cada vez mais, oriundos das classes econômicas mais baixas e não poderão arcar com as mensalidades vigentes.

2.2 EVASÃO ESCOLAR

A evasão escolar é um problema que tem preocupado os profissionais ligados à educação de todo o mundo, mormente no Brasil, onde as mazelas sociais e sistema de ensino deficiente agravam o problema. As razões da evasão escolar são as mais diversas segundo os estudiosos da área, desde motivos econômicos até os psicológicos.

BAKER e SIRYK (1989), identificaram quatro dimensões relacionadas à integração do estudante à universidade: (a) o ajustamento acadêmico ; (b) o ajustamento relacional-social; (c) o ajustamento pessoal-emocional; e (d) o comprometimento com a instituição/aderência.

DIAZ (1996) e GONÇALVES (1997), citados por GAIOSO (2005), fundamentados no modelo teórico de TINTO (1975), afirmam ser possível identificar cinco categorias de causas da evasão: as psicológicas, as sociológicas, as organizacionais, as interacionais e as econômicas. As psicológicas, resultantes das condições individuais como imaturidade, rebeldia, dentre outras, desconsideram o impacto que fatores externos podem ter sobre a ‘personalidade’, ocasionando uma predisposição à evasão.

GAIOSO (2005), escreve que a maioria dos estudos consultados sobre o referido tema se refere às causas da evasão. Tais estudos podem ser agrupados, conforme as principais razões apontadas pelos autores, como as responsáveis pela evasão, tais como: a repetência; a desistência do curso em uma IES por haver conquistado nova vaga na mesma ou em outra instituição, através de vestibular; a falta de orientação educacional no ensino médio; o desprestígio da profissão; a (des)motivação e o horário de trabalho incompatível com o do estudo.

2.3 MINERAÇÃO DE DADOS OU *DATA MINING*

É a obtenção de informações desconhecidas de grandes bancos de dados. Para tanto são usadas diversas técnicas, tais como: ferramentas estatísticas multivariadas , árvores de decisão, redes neurais, entre outras.

O armazenamento e mineração de dados passam a ser mais valorizados a partir da expansão da automação e informatização. A crescente quantidade de informações que as organizações têm a sua disposição não seria de utilidade sem o uso das técnicas de armazenamento e mineração de dados.

HAIR *et al.,* (2005) , explicam que armazenamento e mineração de dados são elementos complementares no melhoramento do acesso a dados para tomadas de decisões. Armazenamento de dados é o mecanismo facilitador para sistemas de apoio a decisões, guardando os dados de uma organização em uma única e integrada base de dados e fornecendo uma perspectiva histórica. Mineração de dados, também conhecida como descoberta do conhecimento em bases de dados, é a busca por relações e padrões em grandes bases de dados. Como sugere o termo, mineração de dados tem uma orientação exploratória de busca por conhecimento obscurecido pelos complexos padrões de associação e grandes quantias de dados.

São diversas as técnicas de mineração de dados. A seguir apresentamos algumas delas, tais como: análise fatorial, análise discriminante e regressão logística.

2.4 ANÁLISE FATORIAL

Tem a finalidade de descrever, se possível, as relações de covariância entre diversas variáveis em função de poucas, não observáveis, chamadas de fatores. Com a análise fatorial pode-se resumir um conjunto de variáveis observáveis em um conjunto menor, com uma pequena perda de informações.

POLYDORO, PRIMI, *et al.,* (2001), fizeram um estudo da evasão escolar no âmbito psicológico, desenvolvendo uma escala de integração ao Ensino Superior. A pesquisa envolveu 46 itens e para agrupá-los foi usada a análise fatorial. O estudo apontou a existência de dois grandes fatores relacionados à integração ao ensino superior: (a) o primeiro associado sobretudo aos aspectos externos do indivíduo, relacionados ao ambiente universitário, de satisfação com o curso e, portanto, aderência ao mesmo e, (b) o segundo, sobretudo, os aspectos internos do indivíduo,

de capacidade de enfrentamento, reações físicas psicossomáticas e estado de humor.

CUNICO (2005), usou análise fatorial na predição da satisfação dos funcionários de uma grande rede varejista onde procurou, a partir da análise dos componentes principais identificar as variáveis mais importantes para a sua pesquisa. Tais variáveis foram usadas em funções classificatórias, possibilitando a identificação dos funcionários satisfeitos e dos insatisfeitos.

PEREIRA (2003), em sua tese de doutorado estudou a evasão de alunos e os custos ocultos para as Instituições de Ensino Superior. A análise fatorial indicou os motivos que mais influenciaram na escolha do curso e na desistência do mesmo. Notou-se que os fatores que influenciam a decisão do aluno em abandonar o curso e a IES consistem de fatores internos à instituição (infra-estrutura deficitária, acervo desatualizado, métodos de avaliação docente, deficiência didático pedagógica dos professores) e inerentes ao estudante (dificuldades financeiras, escolha equivocada do curso, falta de base para acompanhar o curso escolhido e o fato de ser admitido em um curso que não foi a sua primeira opção).

2.5 ANÁLISE DISCRIMINANTE

Com a Análise Discriminante procura-se classificar objetos em populações previamente definidas. Em primeiro lugar, é importante determinar as variáveis que diferenciam as populações, para em seguida utilizar a Função Discriminante de Fisher para alocar, com uma certa margem de erro, o indivíduo na população com características mais próximas a dele. Para resolver o problema proposto, foi usada a MANOVA para determinar as variáveis que discriminam as populações de desistentes e não-desistentes. Em seguida, com as variáveis identificadas, aplicou-se a Função Linear Discriminante de Fisher para classificar os indivíduos em uma destas populações.

CUNICO (2005), trabalhou, entre outras técnicas, com a Análise Discriminante para a classificação de funcionários de uma loja varejista em satisfeitos e insatisfeitos. A partir da aplicação de questionários, ele levantou as populações de satisfeitos e insatisfeitos e, em seguida, utilizou a Função Discriminante de Fisher para alocar um novo indivíduo em uma das populações – satisfeito ou insatisfeito.

PIZZOL (2004) , discutiu um método de tipificação de sistemas de produção dividido em duas etapas. Na primeira, foram usados grupos focais e, na segunda, empregou-se a Análise Discriminante para validar os resultados obtidos nas entrevistas em grupos. O método foi aplicado na identificação de sistemas de produção de café para a região de Marília, no Estado de São Paulo.

MARTEL *et al.,* (2003), estudou as pupunheiras ao longo dos rios Amazonas e Solimões. Nesse estudo, foram aplicadas técnicas estatísticas multivariadas a 15 descritores morfológicos numa tentativa de caracterizar, morfometricamente, três raças existentes ao longo da Bacia desses rios. As três análises em conjunto permitiram uma discriminação das raças, mostrando os descritores mais importantes.

2.6 REGRESSÃO LOGÍSTICA

A relação entre variáveis pode ser descrita por métodos de regressão, sendo esses os mais diversos – regressão linear, quadrática, exponencial, entre outras. A regressão logística é caracterizada por possuir variável resposta (dependente) binária ou dicotômica. Neste trabalho, a variável resposta – dicotômica – indica a desistência ou a não-desistência. Este método foi usado em paralelo com a FDL de Fisher e os resultados comparados.

SANTOS, *et al.* (2005), usou regressão logística e redes neurais para a predição da soroprevalência da Hepatite A. O desempenho de tais modelos foi medido através da taxa de classificação incorreta em uma amostra do município de Duque de Caxias, Rio de Janeiro. Resultados mostram que o modelo neural, aplicado sobre a informação relevante extraída do modelo de regressão logística, apresenta um bom desempenho, alcançando uma eficiência de classificação geral acima de 88%.

CUNICO (2005), buscando classificar funcionários de uma rede varejista em satisfeitos ou insatisfeitos, comparou os resultados obtidos a partir da FDL de Fisher e a regressão logística. Concluiu que para o problema estudado a regressão logística mostrou-se mais eficiente do que FDL de Fisher. Os resultados obtidos com a regressão logística foram bastante próximos tanto para o treinamento, quanto para teste, sendo alcançado um percentual de acerto de 71,4% .

GIMENO e SOUZA (1995), utilizaram a análise multivariada por estratificação e com regressão logística, utilizando dados de um estudo de caso-controle sobre câncer de esôfago. Oitenta e cinco casos e 292 controles foram classificados segundo sexo, idade e os hábitos de beber e de fumar. As estimativas mostraram que as duas técnicas são complementares.

3 TÉCNICAS ESTATÍSTICAS UTILIZADAS

Este capítulo tratará de forma mais específica dos métodos que conduziram à execução do trabalho proposto.

3.1 COMPARAÇÃO ENTRE VETORES MÉDIOS DE 2 POPULAÇÕES

No caso univariado, utiliza-se uma estatística com distribuição t de Student para testar a igualdade de médias entre duas populações, para o caso multivariado é possível desenvolver uma estatística com distribuição T^2 de Hotelling para testar a igualdade de dois vetores de médias. Esta estatística é apropriada para testar a igualdade de dois vetores de médias quando alguns pressupostos são verificados.

Considerem-se duas amostras aleatórias de dimensões n_1 e n_2 retiradas de duas populações 1 e 2 , respectivamente. Para as observações sobre p variáveis podem ser calculadas as estatísticas média $\underline{\overline{X}}$ e matriz de covariância S, segundo as fórmulas :

$\underline{\overline{X}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\underline{X}_{1j}$ e $S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\underline{X}_{1j} - \underline{\overline{X}}_1)(\underline{X}_{1j} - \underline{\overline{X}}_1)'$ para uma amostra $\underline{X}_{11}, \underline{X}_{12}, ..., \underline{X}_{1n_1}$ obtida a partir da população 1; $\underline{\overline{X}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\underline{X}_{2j}$ e $S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(\underline{X}_{2j} - \underline{\overline{X}}_2)(\underline{X}_{2j} - \underline{\overline{X}}_2)'$ para uma amostra $\underline{X}_{21}, \underline{X}_{22}, ..., \underline{X}_{2n_2}$ retirada da população 2. Os valores $\underline{\overline{X}}_1$ e $\underline{\overline{X}}_2$ são os vetores de médias das amostras provenientes das populações 1 e 2 . S_1 e S_2 são as matrizes de covariâncias amostrais. Em $\underline{X}_{11}, \underline{X}_{12}, ..., \underline{X}_{1n_1}$ e $\underline{X}_{21}, \underline{X}_{22}, ..., \underline{X}_{2n_2}$ o primeiro subscrito indica a população de onde foi retirada a amostra e o segundo a observação, sendo que cada observação possui p variáveis.

Inferências acerca das médias das populações ($\underline{\mu}_1$ e $\underline{\mu}_2$), deverão ser efetuadas para verificar se $\underline{\mu}_1 = \underline{\mu}_2$. Nesta discussão, os pressupostos a seguir, relativos à estrutura dos dados devem ser observados.

1- A amostra $\underline{X}_{11}, \underline{X}_{12}, ..., \underline{X}_{1n_1}$ é uma amostra aleatória de dimensão n_1 retirada de uma população com vetor de médias $\underline{\mu}_1$ e matriz de covariância $\sum_1$.

2- A amostra $\underline{X}_{21}, \underline{X}_{22}, ..., \underline{X}_{2n_2}$ é uma amostra aleatória de dimensão n_2 retirada de uma população com vetor de médias $\underline{\mu}_2$ e matriz de covariância $\sum_2$.

3- As duas amostras $\underline{X}_{11}, \underline{X}_{12}, ..., \underline{X}_{1n_1}$ e $\underline{X}_{21}, \underline{X}_{22}, ..., \underline{X}_{2n_2}$ são independentes.

Para pequenas amostras, é necessário acrescentar os seguintes pressupostos:

4- As duas populações seguem uma distribuição normal multivariada.

5- As matrizes de covariância das duas populações são iguais ($\sum_1 = \sum_2$).

Quando as duas matrizes $\sum_1$ e $\sum_2$ são desconhecidas, mas se pressupõe serem iguais a $\sum$, é necessário combinar as matrizes amostrais S_1 e S_2 para encontrar um estimador S_p para $\sum$.

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Para testar as hipóteses H_0: $\underline{\mu}_1 - \underline{\mu}_2 = 0$ e H_1: $\underline{\mu}_1 - \underline{\mu}_2 \neq 0$, usa-se a estatística a seguir:

$$T^2 = (\overline{\underline{X}}_1 - \overline{\underline{X}}_2)' . \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) . S_p \right]^{-1} . (\overline{\underline{X}}_1 - \overline{\underline{X}}_2)$$

sendo $\overline{\underline{X}}_1$ e $\overline{\underline{X}}_2$ os estimadores de $\underline{\mu}_1$ e $\underline{\mu}_2$. A hipótese H_0 será rejeitada se:

$$T^2 . \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2).p} > F_{p, n_1 + n_2 - p - 1}(\alpha)$$

onde $F_{p, n_1 + n_2 - p - 1}(\alpha)$ é obtido da distribuição F de Snedecor com p e $(n_1 + n_2 - p - 1)$ graus de liberdade e nível de significância igual a α.

Quando os vetores de médias $\overline{\underline{X}}_1$ e $\overline{\underline{X}}_2$ são considerados diferentes, pode-se determinar quais as componentes desses vetores apresentam diferenças significativas. Segundo JOHNSON e WICHERN (1998), a comparação das componentes de $\overline{\underline{X}}_1$ e $\overline{\underline{X}}_2$ pode ser feita a partir dos intervalos:

$$(\overline{X}_{1i} - \overline{X}_{2i}) \pm c . \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) . s_{ii}} \qquad \text{para } i = 1,2,...,p$$

onde s_{ii} pertence a diagonal principal da matriz S_p e

$$c^2 = \frac{(n_1 + n_2 - 2).p}{n_1 + n_2 - p - 1} . F_{p, n_1 + n_2 - p - 1}(\alpha)$$

3.1.1 Matrizes de Covariâncias Diferentes $\Sigma_1 \neq \Sigma_2$

Sejam $\underline{\boldsymbol{\mu}}_1$ e $\underline{\boldsymbol{\mu}}_2$ as médias das populações 1 e 2, respectivamente. Deseja-se testar as hipóteses H_0: $\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_2 = 0$ e H_1: $\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_2 \neq 0$, considerando as matrizes de covariâncias diferentes.

Considere amostras de tamanhos n_1 e n_2 com p variáveis tais que $n_1 - p$ e $n_2 - p$ sejam grandes. Segundo JOHNSON e WICHERN (1998), devemos rejeitar a hipóteses H_0 se:

$$(\underline{\overline{\boldsymbol{X}}}_1 - \underline{\overline{\boldsymbol{X}}}_2)' \cdot \left(\frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} \cdot (\underline{\overline{\boldsymbol{X}}}_1 - \underline{\overline{\boldsymbol{X}}}_2) > \chi_p^2(\alpha)$$

onde $\chi_p^2(\alpha)$ é proveniente da distribuição qui-quadrado com graus de liberdade p e nível de significância α. Observando-se que tanto S_1 como S_2 são matrizes do tipo pxp e $\underline{\overline{\boldsymbol{X}}}_1 - \underline{\overline{\boldsymbol{X}}}_2$ um vetor coluna $px1$.

Para a comparação entre as componentes dos vetores de médias $\underline{\overline{\boldsymbol{X}}}_1$ e $\underline{\overline{\boldsymbol{X}}}_2$, obtém-se os intervalos:

$$(\overline{X}_{1i} - \overline{X}_{2i}) \pm \sqrt{\chi_p^2(\alpha)} \sqrt{s_{ii}} \qquad \text{para } i = 1,2,...,p \ ,$$

sendo s_{ii} o elemento da diagonal principal da matriz $\left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]$, $\overline{X}_{1i}$ e $\overline{X}_{2i}$ representam o i-ésimo elemento dos vetores $\underline{\overline{\boldsymbol{X}}}_1$ e $\underline{\overline{\boldsymbol{X}}}_2$. As componentes serão consideradas diferentes caso os extremos dos intervalos apresentem sinais iguais. Intervalos com sinais diferentes nos extremos indicam que as diferenças entre as componentes não são significativas.

3.2 MANOVA

A análise de variância multivariada (MANOVA) faz a comparação entre médias para diferentes variáveis simultaneamente. Utilizam-se dois passos seqüenciais: no primeiro, testa-se a hipótese de igualdade de médias entre os grupos; no segundo passo, se o resultado do passo anterior apresentar diferenças significativas entre as médias, utilizam-se testes adicionais no sentido de explicar as diferenças entre os grupos.

A hipótese nula de igualdade de médias é testada para um conjunto de *p* variáveis simultaneamente. A hipótese nula a ser testada na MANOVA é a seguinte:

H_0: $\underline{\boldsymbol{\mu}}_1 = \underline{\boldsymbol{\mu}}_2 = ... = \underline{\boldsymbol{\mu}}_g$ com $\underline{\boldsymbol{\mu}}_j = \begin{bmatrix} \mu_{1j} \\ \mu_{2j} \\ ... \\ \mu_{pj} \end{bmatrix}$ j = 1, 2,..., g isto é, as médias populacionais dos g grupos são todas iguais.

Suposições para o uso da MANOVA:

(1) Independência (as amostras aleatórias devem ser independentes)

(2) Homocedasticidade (todas as populações devem ter mesma matriz covariância Σ)

(3) Todas as populações devem ser normalmente distribuídas.

A condição (3) tem relevância diminuída quando as amostras são de grande dimensão.

O teste de hipóteses segue a forma:

H_0: $\underline{\boldsymbol{\mu}}_1 = \underline{\boldsymbol{\mu}}_2 = ... = \underline{\boldsymbol{\mu}}_g$

H_1: pelo menos uma das médias $\underline{\boldsymbol{\mu}}_i$ (i = 1, 2, ... , g) é diferente das demais.

Onde $\underline{\boldsymbol{\mu}}_1 = \underline{\boldsymbol{\mu}}_2 = ... = \underline{\boldsymbol{\mu}}_g$ são as médias das populações 1, 2, ..., g.

Para $n = \sum_{i=1}^{g} n_i$ grande, rejeita-se a hipótese H_0 ao nível de significância α se

$$-\left(n - 1 - \frac{p+g}{2}\right).\ln\left(\frac{\det(W)}{\det(B+W)}\right) > \chi^2_{p(g-1)}(\alpha)$$

com $B = \sum_{i=1}^{g} n_i(\underline{\overline{X}}_i - \underline{\overline{X}}).(\underline{\overline{X}}_i - \underline{\overline{X}})'$ e $W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underline{X}_{ij} - \underline{\overline{X}}_i).(\underline{X}_{ij} - \underline{\overline{X}}_i)'$

onde: $\underline{X}_{ij}$ = j-ésima observação da i-ésima amostra (ou i-ésimo tratamento)

$\underline{\overline{X}}_i$ = média da i-ésima amostra (ou i-ésimo tratamento)

$\underline{\overline{X}}$ = média global (todas as amostras)

Quando a hipótese H_0 é rejeitada, pode-se identificar qual ou quais componentes dos vetores de médias diferem significativamente dos demais.

Seja $n = \sum_{i=1}^{g} n_i$. Para o modelo de MANOVA descrito, com confiança de no mínimo (1- α), μ_{kj} - μ_{lj} pertence ao intervalo

$$(\overline{X}_{kj} - \overline{X}_{lj}) \pm t_{n-g}\left[\frac{\alpha}{pg(g-1)}\right] \cdot \sqrt{\frac{w_{jj}}{n-g}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}$$

para todas as componentes j = 1, 2, ... , p e todas as diferenças $l < k$ = 1, 2,..., g. Aqui w_{jj} é o j-ésimo elemento da diagonal de W. Caso os extremos do intervalo apresentem sinais diferentes, descarta-se a hipótese de igualdade entre as componentes.

3.3 ANÁLISE DISCRIMINANTE - MODELOS

De acordo com HAIR *et al.,* (2005), a análise discriminante é aplicável a qualquer pesquisa com o objetivo de entender a pertinência a grupos, seja de indivíduos (p.ex., clientes *versus* não-clientes) , empresas (p. ex., lucrativas *versus* não-lucrativas), produtos (p. ex., de sucesso *versus* sem sucesso) ou qualquer outro objeto que possa ser avaliado em uma série de variáveis independentes.

A função discriminante constitui em uma combinação linear de variáveis independentes, sendo os seus principais pressupostos a normalidade multivariada e a igualdade das matrizes de covariâncias. Pode-se construir uma função discriminante a partir das características de dois grupos de indivíduos e com essa função classificar um novo indivíduo em um dos grupos.

3.3.1 Função Discriminante de Fisher.

Dentro da análise discriminante, um tópico de grande relevância é a função discriminante linear de Fisher , apresentada a seguir.

Segundo JOHNSON e WICHERN (1998), a idéia de Fisher foi transformar as observações multivariadas $\underline{\boldsymbol{X}}$ nas observações univariadas Y tal que os Y's nas populações π_1 e π_2 fossem separadas tanto quanto possível.

A FDL de Fisher é dada pela combinação linear $Y = \underline{\boldsymbol{a}}'.\underline{\boldsymbol{X}}$ Considerando os estimadores S e $\overline{\underline{\boldsymbol{X}}}$ de Σ e $\underline{\boldsymbol{\mu}}$, respectivamente, a função discriminante de Fisher estimada para dois grupos é dada por:

$$\hat{Y} = \underline{\hat{\boldsymbol{a}}}'.\underline{\boldsymbol{X}}$$

onde $\hat{\underline{a}}' = (\overline{\underline{X}}_1 - \overline{\underline{X}}_2)S_p^{-1}$ e $S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$, $\overline{\underline{X}}_1$ é a média amostral do primeiro grupo e $\overline{\underline{X}}_2$ do segundo grupo.

A regra para alocação de uma observação $\underline{X}_0$ é a seguinte:

● Aloca-se $\underline{X}_0$ no grupo 1 se:

$$\hat{Y}_0 = \hat{\underline{a}}' \underline{X}_0 \geq m = \frac{1}{2}\hat{\underline{a}}'(\overline{\underline{X}}_1 + \overline{\underline{X}}_2) = \frac{\overline{Y}_1 + \overline{Y}_2}{2}$$

● Aloca-se $\underline{X}_0$ no grupo 2 se:

$$\hat{Y}_0 < m$$

3.3.2 Estimando a Probabilidade de Erro na Classificação.

São dois tipos de erros que podem ocorrer quando se trabalha com duas populações. Quando o elemento amostral pertence à população 1, mas a função discriminante o classifica como sendo da população 2, tem-se o erro tipo 1. Já o erro tipo 2 deriva da classificação de um elemento amostral como sendo da população 1, quando este é proveniente da população 2. Denominando p(2/1) e p(1/2) as probabilidades de ocorrência dos erros 1 e 2 , respectivamente. Logo:

p(2/1) = a probabilidade de classificar erradamente um elemento em 2 quando ele é de 1;

p(1/2) = a probabilidade de classificar erradamente um elemento em 1 quando ele é de 2.

Para a diminuição do erro na classificação de um indivíduo em uma das populações, é importante que estas probabilidades sejam a menor possível.

O método de estimação das probabilidades p(2/1) e p(1/2) que será visto a seguir é denominado de Método da Resubstituição. MINGOTI (2005), escreve que neste método, os escores de cada elemento amostral observado das populações 1 e 2 são calculados, sendo a regra de discriminação utilizada para classificar os $n = n_1 + n_2$ elementos da amostra conjunta. Quando a função discriminante é de boa qualidade, espera-se que ela apresente uma grande porcentagem de acerto na classificação dos elementos amostrais em relação à população a que de fato pertencem. Portanto, neste método, os mesmos elementos amostrais participam da estimação da regra de classificação e de estimação dos erros de classificação. As

freqüências de classificações corretas e incorretas podem ser sumarizadas em uma matriz de confusão, como mostra o quadro 1.

		População classificada pela regra		
		1	**2**	**Total**
População de origem	1	n_{11}	n_{12}	N_1
	2	n_{21}	n_{22}	N_2

Quadro 1. Matriz de Confusão Genérica

sendo n_{ij} o número de elementos pertencentes à população de origem i e que são classificados pela função discriminante como pertencentes à população j. Quando $i = j$, tem-se o número de classificações corretas, e quando $i \neq j$, tem-se o número de classificações incorretas. Com base nesses dados, as estimativas das probabilidades de ocorrência dos erros 1 e 2 são dados respectivamente por

$$\hat{p}(2/1) = \frac{n_{12}}{n_1} \qquad \text{e} \qquad \hat{p}(1/2) = \frac{n_{21}}{n_2}$$

Este procedimento de estimação do erro aparente de classificação (APER) é consistente, mas viciado (Johson; Wichern, 1998), e tende a subestimar os verdadeiros valores de p(2/1) e p(1/2) para elementos que não pertencem à amostra conjunta utilizada para a construção da função discriminante, isto é, novos elementos amostrais.

3.3.3 Método de Lachenbruch

É uma forma de avaliar a eficiência da regra de classificação. Esta técnica segue os passos apresentados abaixo:

(1) Escolher um dos grupos (amostras).

(2) Descartar uma observação do grupo.

(3) Construir uma função discriminante para as (n_1 – 1) observações restantes do grupo escolhido e para as n_2 observações do segundo grupo, ou seja, para (n_1 – 1 + n_2) observações.

(4) Classificar a observação descartada usando a função obtida anteriormente.
(5) Realocar a observação descartada e repetir os passos 1 e 2 para todas as observações do primeiro grupo.
(6) Repetir os passos 1 a 5 para o segundo grupo.

Pode-se obter então: $\hat{p}(2/1) = \frac{n_{12}}{n_1}$, $\hat{p}(1/2) = \frac{n_{21}}{n_2}$ e $\hat{E}(APER) = \frac{n_{12} + n_{21}}{n_1 + n_2}$

que é a proporção total esperada de erro.

Desta forma obtém-se uma regra de reconhecimento e classificação construída com as n observações amostrais e testada com todas as referidas observações, mas sempre com a observação em teste fora do ajuste. Isto equivale a ter um grupo com n observações para o ajuste e outro grupo, também de tamanho n, para testar a eficiência do procedimento.

3.4 REGRESSÃO LOGÍSTICA - MODELO

Neste tipo de regressão a variável dependente é dicotômica ou binária e de maneira geral se assemelha à regressão linear. O modelo da regressão logística é exponencial. Para que a função obtida tenha propriedades da regressão linear, aplica-se a transformação denominada de logit. Como a variável resposta na regressão logística é dicotômica, podemos utilizá-la na classificação de objetos em duas populações distintas, semelhante à função discriminante de Fisher para duas populações. A curva logística tem a forma de um S e segundo HAIR *et al.,* (2005), a forma em S é não-linear porque a probabilidade de um evento deve se aproximar de 0 e 1, porém jamais ser maior. Assim, à medida que as probabilidades se aproximam dos limites inferior e superior de probabilidade (0 e 1), elas devem se “amenizar” e ficar assintóticas nesses limites. A taxa de aproximação de zero é igual à taxa de aproximação de 1. A curva em forma de S pode ser observada no gráfico 1 , onde x são as observações e π a probabilidade.

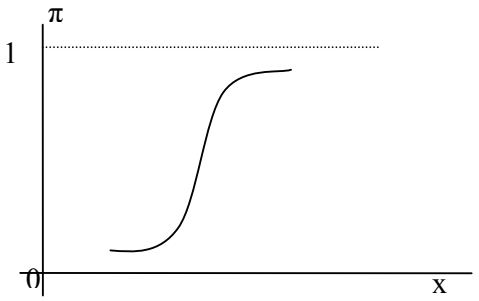

Gráfico 1. Curva Logit

3.4.1 O Modelo Logit

Em diversos problemas a variável resposta apresenta apenas duas categorias. Em especial, pode-se citar o diagnóstico de uma doença, onde os casos observados podem ser classificados como “sucesso” ou “fracasso”.

Uma variável aleatória Y tem uma distribuição de Bernoulli com parâmetro π quando assume apenas os valores 1 e 0 com probabilidade π e (1 – π), respectivamente. O número 1, em geral, representa “sucesso”. Para Y = 1 e Y = 0 temos as probabilidades P(Y = 1) = π e P(Y = 0) = 1 – π. Quando Y_i tem distribuição de Bernoulli com parâmetro π_i , a função de probabilidade é dada por:

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} .$$

Usando a propriedade dos logaritmos: $a^{\log_a N} = N$, pode-se escrever

$$f(y_i;\pi_i) = (1-\pi_i).e^{y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)}$$

para y_i = 0 e 1. A função $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ é chamada de logit de π (AGRESTI, 1990).

Na distribuição de Bernoulli a esperança matemática de Y é

E(Y) = P(Y = 1) = $\pi(x)$,

representando a dependência da variável explicativa X = (x_1, ..., x_p).

Para resposta binária o modelo de probabilidade linear é dado por

E(Y) = $\pi(x)$ = α + βx.

Para as curvas em forma de S, que é o caso da curva Logit, a função mais apropriada é dada por

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} ,$$

chamada de função regressão logística.

Aplicando logaritmo em $\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$ e fazendo as devidas transformações, encontra-se a função logit

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \quad .$$

Para p variáveis explicativas, tem-se:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

3.4.2 Estimação o Parâmetro β

Para p variáveis explicativas x_i e n observações, com i= 1,...,n, o modelo de regressão logística pode ser escrito na forma:

$$\pi(x_i) = \frac{e^{\sum_{j=0}^{p}\beta_j x_{ij}}}{1+e^{\sum_{j=0}^{p}\beta_j x_{ij}}} \quad \text{e} \quad 1-\pi(x_i) = \frac{1}{1+e^{\sum_{j=0}^{p}\beta_j x_{ij}}} \quad ,$$

com $\beta_0 = \alpha$ e $x_{i0} = 1$.

Considere a classificação binária , com os grupos G_i = 1 e G_i = 2. Se G_i = 1, denota-se y_i = 1 e se G_i = 2, denota-se y_i = 0. Fazendo $\pi_1(x) = \pi(x)$, tem-se $\pi_2(x) = 1 - \pi_1(x) = 1 - \pi(x)$.

Se y_i = 1 e G_i = 1, então $\ln \pi_{Gi}(x) = \ln \pi_1(x) = 1.\ln \pi(x) = y_i.\ln \pi(x)$. Se y_i = 0 e G_i = 2, então $\ln \pi_{Gi}(x) = \ln \pi_2(x) = 1.\ln (1- \pi(x)) = (1- y_i).\ln (1- \pi(x))$. Sendo y_i = 0 ou 1 - y_i = 0, tem-se $\ln \pi_{Gi}(x) = y_i.\ln \pi(x) + (1- y_i).\ln (1- \pi(x))$.

Seja uma amostra de tamanho n, para obter-se β é necessário maximizar a função de verossimilhança:

$$l(\beta) = \sum_{i=1}^{n} \ln(\pi_{Gi}(x_i) = \sum_{i=1}^{n} [y_i.\ln(\pi(x_i)) + (1-y_i).\ln(1-\pi(x_i))] \text{, onde } \beta \text{ possui } p\text{+1}$$

parâmetros $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{pmatrix}$ e $x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ ... \\ x_p \end{pmatrix}$. Como $\pi(x) = \frac{e^{\beta^T x}}{1+e^{\beta^T x}}$ e $1-\pi(x) = \frac{1}{1+e^{\beta^T x}}$,

substituindo em $l(\beta)$ vem:

$$l(\beta) = \sum_{i=1}^{n} [y_i \beta^T x_i - \ln(1 + e^{\beta^T x_i})].$$

Para a maximização de $l(\beta)$ faz-se as derivadas parciais, resultando em:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i (y_i - \pi(x_i)).$$

Para resolver o conjunto de p + 1 equações não-lineares $\frac{\partial l(\beta)}{\partial \beta_j} = 0$, j = 0, 1,...,p, usa-se o algoritmo de Newton-Raphson, que possibilita o cálculo de β. Na forma matricial e considerando um β inicial, tem-se:

$$\beta = \beta + (X\widetilde{X})^{-1} X(Y - P)$$

onde: $\widetilde{X} = \begin{pmatrix} \pi(x_1)(1-\pi(x_1))x_1^T \\ \pi(x_2)(1-\pi(x_2))x_2^T \\ \pi(x_3)(1-\pi(x_3))x_3^T \\ \dots \\ \pi(x_n)(1-\pi(x_n))x_n^T \end{pmatrix}$ e $P = \begin{pmatrix} \pi(x_1) \\ \pi(x_2) \\ \pi(x_3) \\ \dots \\ \pi(x_n) \end{pmatrix}$

Tendo os valores de β, constrói-se a função logit

$$\hat{g}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p .$$

A regra para classificar um indivíduo em um dos grupos é a seguinte:

- se $\hat{g}(x) \geq 0$, então x pertence ao grupo 1, ou seja y = 1;
- se $\hat{g}(x) < 0$, então x pertence ao grupo 2, ou seja y = 0.

3.4.3 Algoritmo Para o Cálculo de β

Passo 1: fazer β = 0.

Passo 2: Calcular os elementos de Y , onde:

- y_i = 1 , se G_i = 1 (grupo1);
- y_i = 0 , se G_i = 0 (grupo 2).

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

Passo 3: Calcular os elementos de P, sendo $\pi(x_i) = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$, x_i é um vetor coluna com (p+1) linhas.

Passo 4: Calcular a matriz $\widetilde{X}$ de ordem *n* x (*p*+1) fazendo a multiplicação da i-ésima linha de X (matriz de entrada) por $\pi(x_i)(1-\pi(x_i))$, i = 1,2,...,*n*.

$$X^T = \begin{pmatrix} x_1^T \\ x_2^T \\ x_3^T \\ \dots \\ x_n^T \end{pmatrix} \qquad \widetilde{X} = \begin{pmatrix} \pi(x_1)(1-\pi(x_1))x_1^T \\ \pi(x_2)(1-\pi(x_2))x_2^T \\ \pi(x_3)(1-\pi(x_3))x_3^T \\ \dots \\ \pi(x_n)(1-\pi(x_n))x_n^T \end{pmatrix}$$

Passo 5: $\beta \leftarrow \beta + (X\widetilde{X})^{-1}X(Y-P)$,sendo

$$(\beta_{(p+1)x1}),(X_{(p+1)xn}),(\widetilde{X}_{nx(p+1)}),(Y_{nx1})e(P_{nx1})$$

Passo 6: Se o critério de parada estiver satisfeito, parar. Caso contrário voltar ao passo 3.

A análise da qualidade de ajuste do modelo logístico é feita de forma similar ao que foi apresentado na seção 3.3.2, usando-se a matriz de confusão.

3.5 ANÁLISE FATORIAL - MODELOS

O objetivo da análise fatorial é representar um número de variáveis iniciais observáveis em um número menor de variáveis hipotéticas não observáveis, denominadas de fatores. A partir do momento em que os fatores são identificados, seus valores numéricos, chamados de escores, podem ser utilizados em outras análises, como por exemplo a análise de regressão.

3.5.1 Modelo de Análise Fatorial

Seja o vetor aleatório $\underline{\boldsymbol{X}}_{\boldsymbol{px1}}$ com vetor de médias $\underline{\boldsymbol{\mu}}$. Usando notação matricial o modelo pode ser expresso por:

$$D(\underline{\boldsymbol{X}} - \underline{\boldsymbol{\mu}}) = L\underline{\boldsymbol{F}} + \underline{\boldsymbol{\varepsilon}} \quad , \quad \text{onde}$$

$$(\underline{\boldsymbol{X}} - \underline{\boldsymbol{\mu}})_{px1} = \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \dots \\ X_p - \mu_p \end{bmatrix} , \quad \underline{\boldsymbol{\varepsilon}}_{\boldsymbol{px1}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_p \end{bmatrix} , \quad \underline{\boldsymbol{F}}_{\boldsymbol{mx1}} = \begin{bmatrix} F_1 \\ F_2 \\ \dots \\ F_m \end{bmatrix} , \quad L_{pxm} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \dots & \dots & \dots & \dots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{bmatrix} \text{ e}$$

$$D_{pxp} = \begin{bmatrix} 1/\sigma_1 & 0 & 0 & ... & 0 \\ 0 & 1/\sigma_2 & 0 & ... & 0 \\ 0 & 0 & 1/\sigma_3 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & 1/\sigma_p \end{bmatrix}$$, sendo:

- $\underline{\boldsymbol{F}}$ = vetor aleatório contendo m fatores, com $1 \leq m \leq p$ e p o número de variáveis iniciais.
- L_{pxm} = matriz de parâmetros que precisam ser estimados.
- l_{ij} = peso ou carregamento na i-ésima variável X_i do j-ésimo fator F_j
- $\underline{\boldsymbol{\varepsilon}}_{\boldsymbol{px1}}$ = vetor de erros aleatórios

3.5.2 Modelo de Fatores Ortogonais

Para a apresentação do modelo ortogonal algumas suposições fazem-se necessárias. São elas:

- os fatores têm média igual a zero;
- os fatores não são correlacionados e têm variâncias iguais a 1;
- os erros não são correlacionados entre si e não necessariamente tem a mesma variância, sendo que a variância de ε é dada pela matriz

$$\psi_{pxp} = \begin{bmatrix} \psi_1 & 0 & ... & 0 \\ 0 & \psi_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & \psi_p \end{bmatrix};$$

- os vetores $\underline{\boldsymbol{F}}$ e $\underline{\boldsymbol{\varepsilon}}$ são independentes, ou seja, cov ($\underline{\boldsymbol{\varepsilon}}$, $\underline{\boldsymbol{F}}$) = 0

A partir destas suposições e do modelo $(\underline{\boldsymbol{X}} - \underline{\boldsymbol{\mu}}) = L\underline{\boldsymbol{F}} + \underline{\boldsymbol{\varepsilon}}$ tem-se o modelo ortogonal:

- Cov($\underline{\boldsymbol{X}}$) = $LL' + \psi$
- $V(X_i) = l_{i1}^2 + l_{i2}^2 + ... + l_{1m}^2 + \psi_i$
- $Cov(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + ... + l_{im}l_{km}$
- $Cov(X_i, F_j) = l_{ij}$

3.5.3 Estimando o Número *m* de Fatores

Para a obtenção de m, devemos extrair os autovalores (λ) da matriz de correlação a fim de determinar quais os autovalores são mais importantes. Pode-se seguir um dos seguintes critérios:

- escolher os autovalores que representam maiores proporções da variância total (λ_i/p , i=1,2,...p). Assim, m (número de fatores) é igual ao número de autovalores escolhidos;
- o valor de m será igual ao número de autovalores maiores ou iguais a 1, critério proposto por Kaiser (1958).

3.5.4 Estimando as Matrizes L_{pxm} e ψ_{pxp}

Supondo a variável X padronizada, substitui-se S (matriz de covariância) pela matriz de correlação R. O método das componentes principais consiste em para cada autovalor λ_i , i = 1,2,...,m retido na estimação de m, encontra-se o autovetor normalizado correspondente $\hat{\underline{e}}_i$, onde $\hat{\underline{e}}_i = (\hat{e}_{i1} \ldots \ \hat{e}_{ip})'$. Desta forma, as matrizes L_{pxm} e ψ_{pxp} são definidas por:

$$\hat{L}_{pxm} = [\sqrt{\hat{\lambda}_1}\,\hat{\underline{e}}_1 \quad \sqrt{\hat{\lambda}_2}\,\hat{\underline{e}}_2 \quad \ldots \quad \sqrt{\hat{\lambda}_m}\,\hat{\underline{e}}_m]$$

$\hat{\psi}_{pxp} = diag(R_{pxp} - \hat{L}_{pxm}\hat{L}_{mxp})$, onde $\hat{\psi}_{pxp}$ é uma matriz diagonal.

3.5.5 Estimação dos Escores Fatoriais

Após identificar e interpretar os fatores F_j , j=1,2,...,m, relacionados com as variáveis padronizadas Z_i , i=1,2,...,p, é necessário calcular os escores (valores numéricos) para cada elemento amostral, de modo a utilizar esses valores para outras análises. Para cada elemento amostral k, k = 1,2,...,n, o seu escore no fator F_j é dado por:

$$\hat{F}_{jk} = w_{j1}Z_{1k} + w_{j2}Z_{2k} + \ldots + w_{jp}Z_{pk}$$

onde (Z_{1k} Z_{2k} ... Z_{pk}) são os valores observados das variáveis padronizadas Z_i, para o k-ésimo elemento amostral e os coeficientes w_{ji} , i=1,2,...,p são os pesos de ponderação de cada variável Z_i no Fator F_j. Em um dos métodos usados para a determinação do escore fatorial, tem-se que $\hat{F}_{jk}$ é dado por:

$$\hat{F}_{jk} = (\hat{L}^{`}_{mxp}\hat{\psi}^{-1}_{pxp}\hat{L}_{pxm})^{-1}\hat{L}^{`}_{mxp}\psi^{-1}_{pxp}Z_k = W_{mxp}Z_k$$

onde W_{mxp} é a matriz de ponderação que gera os coeficientes w_{ji} , j=1,2,...,m, i=1,2,...,p

4 DELIMITAÇÃO DA PESQUISA E METODOLOGIA

Neste capítulo, procurou-se descrever a região, e os elementos que compõem a pesquisa. Também será mostrado como os dados foram levantados.

4.1 ÁREA DE ABRANGÊNCIA

O projeto foi executado em uma IES da cidade de Foz do Iguaçu-Pr, situada no extremo-oeste do Estado, fronteira com Paraguai e Argentina, tendo como divisa os rios Paraná e Iguaçu. Os limites de Foz do Iguaçu são: ao norte com o município de Itaipulândia; ao sul com a Argentina; a leste com os municípios de Santa Terezinha de Itaipu, São Miguel do Iguaçu e Medianeira e a oeste com o Paraguai. A vocação econômica de Foz do Iguaçu é o turismo, motivado pelas Cataratas do Iguaçu, compras na Argentina e Paraguai, Itaipu e grandes hotéis que propiciam o turismo de eventos. Os municípios próximos a Foz do Iguaçu-Pr tem como atividade principal a agricultura e algumas indústrias, estas com maior concentração em Medianeira-Pr.

Segundo o IBGE, estimativa populacional de 2004, Foz do Iguaçu-Pr possui 279.620 habitantes; destes, mais de 117 mil tem idade superior a 19 anos e 21.380 estão concluindo ou já possuem o ensino médio, enquanto que as matrículas no Ensino Superior não ultrapassam 8.000 ingressos. O motivo de tal diferença seria uma opção ou um impedimento – por exemplo de ordem econômica – pois dizer que a oferta não está suprindo a demanda não é verdade, já que as faculdades possuem inúmeras vagas sem preenchimento. Talvez seja um problema de ordem econômica ou os cursos oferecidos não são de interesse para a região. Essas questões, entre outras, devem ser discutidas pelo setor do ensino superior privado na busca do crescimento. Com relação à renda familiar da população economicamente ativa, o IBGE -2004 - informa que apenas 10,4% das famílias possuem renda acima de 10 salários mínimos.

4.2 POPULAÇÃO PESQUISADA

São sete as IES privadas existentes em Foz do Iguaçu e cidades próximas, sendo que cinco delas estão localizadas em Foz do Iguaçu-Pr, uma em São Miguel do Iguaçu-Pr e uma em Medianeira-Pr. Por razões operacionais, esta pesquisa desenrolou-se em uma única IES de Foz do Iguaçu-Pr.

As tabelas seguintes, 1, 2, 3, 4 e 5, mostram a distribuição dos cursos e número de alunos matriculados para cada IES de Foz do Iguaçu.

Tabela 1. CESUFOZ: Alunos matriculados por curso

CURSOS	2003
Administração	232
Ciência da Computação	167
Ciências Contábeis	41
Ciências Econômicas	70
Educação Física	253
Direito	93
Processamento de Dados	160
Total	1.016

Fonte: CESUFOZ – Centro de Ensino Superior de Foz do Iguaçu

Tabela 2. UDC : Alunos matriculados por curso

CURSOS	2003
Administração – Gestão de Qualidade	171
Administração – Negócios e *Marketing* Internacional	219
Administração Pública	48
Arquitetura e Urbanismo	201
Comunicação Social (Jornalismo)	185
Comunicação Social (Publicidade e Propaganda)	165
Comunicação Social (Relações Públicas)	28
Direito	263
Engenharia Civil	115
Letras	114
Normal Superior	46
Pedagogia	248
Sistemas de Informação	150
Turismo	187
Total	2.140

Fonte: UDC – Faculdade União Dinâmica Cataratas

Tabela 3. UNIAMÉRICA : Alunos matriculados por curso

CURSOS	2003
Administração em Finanças	104
Administração em *Marketing*	118
Biomedicina	68
Ciências Biológicas	177
Educação Física	99
Enfermagem	243
Engenharia Ambiental	76
Fisioterapia	196
História	113
Nutrição	158
Secretariado Executivo Trilíngue	68
Serviço Social	119
Psicologia	67
Normal Superior	19
Total	1.625

Fonte: UNIAMÉRICA – Faculdade União das Américas

Tabela 4. Anglo-Americano: Alunos matriculados por curso.

CURSOS	2003
Administração/Gestão de Negócios	44
Fisioterapia	66
Normal Superior	67
Total	177

Fonte: Faculdade Anglo-Americano

Tabela 5. UNIFOZ : Alunos matriculados por curso

CURSOS	2003
Administração / Comércio Exterior	162
Direito	867
Tecnologia em Hotelaria	62
Total	1.091

Fonte: UNIFOZ – Faculdades Unificadas de Foz do Iguaçu

4.3 DADOS

Os dados levantados por esta pesquisa têm por finalidade o estudo de características dos elementos formadores de duas populações: uma constituída por alunos que permanecem na faculdade até a titulação e, a outra, por alunos que deixam a faculdade antes de alcançarem a titulação. Foram levantados dados relativos a sexo, renda familiar, nota no vestibular, nível de instrução dos pais, idade, estado civil e níveis de satisfação com relação ao curso e IES escolhida, objetivando encontrar as variáveis que diferenciam as duas populações para, então, classificar um novo indivíduo em uma delas.

4.4 INSTRUMENTO DE COLETA DE DADOS.

Um dos procedimentos para a coleta de informações foi a consulta ao banco de dados da IES, buscando o cadastro dos ex-alunos (alunos desistentes – sem titulação). O banco de dados das IES oferece o endereço , o telefone e a nota obtida no vestibular do aluno evadido.

De posse destas primeiras informações, passou-se à segunda fase da coleta de dados, que foi feita através da aplicação de um questionário (ver Apêndice F) elaborado com base nos objetivos desta pesquisa. Esse questionário foi aplicado aos alunos e ex-alunos.

Os alunos matriculados foram solicitados a responder o questionário durante o período de aulas em suas respectivas salas. A amostra formada pelos alunos foi denominada de grupo 1. Esses alunos estavam cursando um dos últimos quatro períodos de seu curso. Pois a experiência e estatísticas da IES pesquisada mostram que alunos dos últimos períodos raramente abandonam o curso antes da titulação. A amostra formada pelos ex-alunos, aqueles que se afastam do curso sem a titulação, foi denominada de grupo 2. Para formar este grupo , foi feita uma pesquisa nos arquivos da IES, de forma a encontrar os alunos que cancelaram matrícula, abandonaram o curso, transferiram-se para outra instituição ou trancaram matrícula nos últimos 3 anos. Os elementos do grupo 2 receberam os questionários em suas casas via correio tradicional ou eletrônico e, uma outra alternativa usada, foi o contato telefônico. Inicialmente deu-se preferência para o uso dos correios, a fim

de que dessa forma o ex-aluno fosse o mais verdadeiro possível em suas respostas, não ficando assim constrangido pela figura do pesquisador. Outro fator que pesou na escolha dos correios foi o econômico. Infelizmente, dos 148 questionários enviados pelos correios (tradicional e eletrônico), somente 41 questionários retornaram preenchidos e 16 devolvidos pela não localização do destinatário. Usou-se então o contato telefônico, mas, mesmo assim, muitos não foram localizados devido à mudança do número do telefone e não divulgação do novo número. Assim conseguiu-se formar o grupo 1 com 172 elementos e o grupo 2 com 109, todos originários dos cursos Ciência da Computação, Administração, Educação Física e Ciências Contábeis.

4.5 CARACTERIZAÇÃO DOS GRUPOS

Como já mencionado anteriormente, foram pesquisados para formar o grupo 1, 172 alunos dos últimos quatro períodos, estes foram extraídos de um total de 320 alunos para os quatro cursos pesquisados. Dispensou-se a estimativa para o tamanho da amostra, já que o grupo 1 foi formado com 54% da população definida. Da mesma forma, procedeu-se para o grupo 2. Nos últimos 3 anos, os arquivos da IES registravam 268 ex-alunos, de onde extraiu-se o grupo 2, com 109 elementos, que corresponde a 40% da população definida. A definição da população dos ex-alunos a partir dos últimos 3 anos (desde 2003) é conseqüência de vários fatores. Um deles foi a dificuldade em localizar estes ex-alunos; outro, as possíveis mudanças nas características do aluno desistente. Estas por sua vez, podem ser determinadas pela evolução da IES, surgimento de concorrentes e transformações sociais e econômicas.

4.6 IDENTIFICAÇÃO DAS VARIÁVEIS

No quadro 2 estão as correspondências entre as perguntas do questionário do Apêndice F e as respectivas variáveis. A vigésima primeira variável (VAR21) foi extraída dos arquivos da IES pesquisada.

Pergunta	Descrição	Variável
4	Qual é o seu sexo?	VAR1
5	Qual é a sua idade?	VAR2
6	Qual é o seu estado civil?	VAR3
7	Com relação a sua moradia?	VAR4
8	Você tem trabalho remunerado?	VAR5
9	Qual é o nível de instrução do seu pai?	VAR6
10	Qual é o nível de instrução da sua mãe?	VAR7
11	Qual é a renda familiar?	VAR8
12	Com relação ao curso escolhido.....	VAR9
13	Onde você fez o ensino médio?	VAR10
14	Em qual turno você fez o ensino médio?	VAR11
15	Indique a sua principal razão na escolha da faculdade.	VAR12
16	Indique o seu principal motivo na escolha do curso.	VAR13
17	Classifique o seu relacionamento afetivo c/ os colegas.	VAR14
18	Qual o seu principal motivo ao fazer um curso superior?	VAR15
19	Indique o nível de satisfação c/ a infra-estrutura da IES.	VAR16
20	Dê sua opinião c/ relação a capacidade dos professores	VAR17
21	Classifique o atendimento oferecido pelos setores da IES	VAR18
22	Indique a sua satisfação c/ relação ao curso escolhido	VAR19
23	Com relação ao tempo destinado aos estudos.....	VAR20
	Nota média no vestibular	VAR21

Quadro 2. Variáveis

5 RESULTADOS OBTIDOS

A seguir, serão aplicadas as técnicas apresentadas no capítulo 3 para analisar os dados levantados para os grupos 1 e 2.

5.1 COMPARAÇÃO ENTRE AS MÉDIAS DOS GRUPOS 1 E 2

Inicialmente introduziu-se no *software Excel* duas matrizes, uma para o grupo 1(n_1 x p) e outra para o grupo 2 (n_2 x p), sendo $n_1 = 172$, $n_2 = 109$ e $p = 21$, onde n_1 e n_2 representam as observações e p o total de variáveis. Do questionário do Apêndice F, retiram-se p-1 variáveis, iniciando pela questão 04. Assim a primeira variável representa o sexo, a segunda indica a idade, até a questão 23 (tempo destinado para os estudos). A vigésima primeira variável representa a nota média no vestibular, obtida no banco de dados da IES.

Na segunda etapa, são usados dois métodos para verificar se os vetores de médias provenientes dos grupos 1 e 2 são estatisticamente diferentes. O primeiro método considera matrizes de covariâncias diferentes($\Sigma_1 \neq \Sigma_2$) e o segundo matrizes iguais ($\Sigma_1 = \Sigma_2$), neste caso o método é a MANOVA.

5.1.1 Comparação Entre Médias Para $\Sigma_1 \neq \Sigma_2$

Na seção 3.1.1 foi apresentado o método aqui utilizado e no Apêndice E pode-se encontrar o programa PGR05 que efetua os cálculos. Executando o programa PGR05 para os grupos 1 e 2 (matrizes: 172x21 e 109x21) determinou-se que os vetores de médias dos grupos 1 e 2 apresentam diferenças significativas. O mesmo programa determina quais as componentes dos vetores que diferem significativamente e ele indicou as variáveis: VAR9, VAR16, VAR18 e VAR20. Obtidas a partir das questões 12, 19, 21 e 23 do questionário do Apêndice F, sendo o nível de significância de 5%. A questão 12 do questionário (Apêndice F) , procura captar a impressão do aluno ou ex-aluno com relação ao curso escolhido, a questão 19 trata da satisfação do aluno ou ex-aluno com relação a infra-estrutura oferecida pela IES, a questão 21 busca classificar o atendimento oferecido pelos setores da IES e a questão 23 procura uma impressão dos alunos ou ex-alunos com relação ao tempo destinado aos estudos.

5.1.2 Comparação entre médias – MANOVA

O segundo método utilizado para comparar vetores de médias é a MANOVA, método executado pelo programa PGR04 (Apêndice D). O método, através de PGR04, mostrou que os vetores de médias dos grupos 1 e 2 são estatisticamente diferentes, sendo sete as componentes dos vetores com diferenças significativas ao nível de 5% . Das sete, quatro já foram devidamente identificadas na seção anterior. Entre as outras três variáveis, duas representam as questões 18 (VAR15) e 22 (VAR19) do questionário do Apêndice F e a outra (VAR21) carrega a nota média no vestibular. A questão 18 aborda o objetivo de se fazer um curso superior, já a questão 22 trata da satisfação do aluno ou ex-aluno com relação ao curso escolhido.

5.2 ESTATÍSTICAS DESCRITIVAS

A seguir serão descritas as variáveis com diferenças significativas determinadas nas duas últimas seções.

A variável que representa a questão 12 do questionário (Apêndice F) apresenta mediana e moda igual a 2 para ambos os grupos, indicando que a resposta “era o que você esperava” foi a que apresentou maior freqüência. Observando a tabela 6 pode-se verificar que 33,9% dos entrevistados do grupo 2 ficaram decepcionados com o curso escolhido contra 17,4% do grupo 1. Estar decepcionado com o curso escolhido é, com certeza, um fator muito importante na decisão do aluno abandonar ou não um curso de graduação.

Tabela 6. Com relação ao curso escolhido

RESPOSTAS	GRUPO 1		GRUPO 2	
	Freqüência(%)	Acumulada(%)	Freqüência(%)	Acumulada(%)
1-Você ficou decepcionado	30(17,4%)	30(17,4%)	37(33,9%)	37(33,9%)
2-Era o que você esperava	87(50,6%)	117(68,0%)	67(61,5%)	104(95,4%)
3-Superou as suas expectativas	21(12,2%)	138(80,2%)	3(2,8%)	107(98,2%)
4-Não sabe dizer	34(19,8%)	172(100%)	2(1,8%)	109(100%)
TOTAL	172(100%)		109(100%)	

Fonte: Autor

No gráfico 2 visualizam-se as diferenças. O setor que representa a resposta “você ficou decepcionado” é maior no grupo 2 , já o setor que representa a resposta “superou as suas expectativas” é maior no grupo 1.

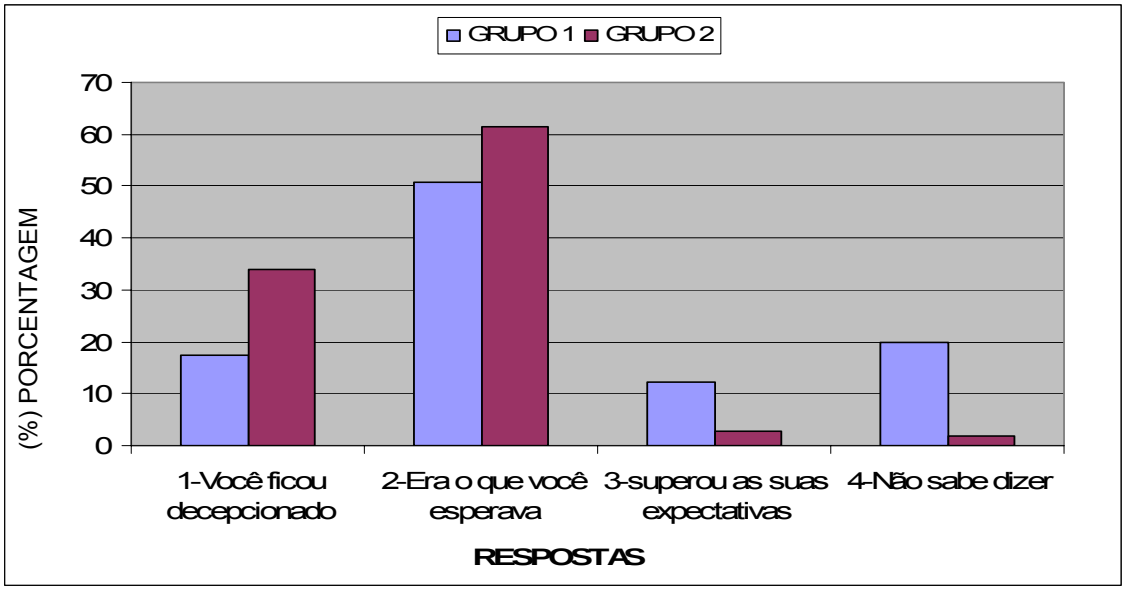

Gráfico 2. Com relação ao curso escolhido

A tabela 7 apresenta as respostas para a pergunta “qual o seu principal objetivo ao fazer um curso superior?” . Analisados os dados foi constatado que as respostas “emprego”, “aumento salarial” e “diploma de nível superior” acumulam porcentagem de 22,1% para o grupo 1 e 42,2% para o grupo 2, enquanto “formação profissional” totaliza 75,6% no grupo 1 e 56,9% no grupo 2, mostrando que o grupo dos ex-alunos tem forte preferência pelas respostas 1, 2 e 3 , já os alunos (grupo 1) estão mais voltados para a resposta 5. Os resultados apontam para uma clara diferença entre os grupos, mostrando, por exemplo, que o objetivo “diploma de nível superior” não é suficientemente forte para manter o estudante em um curso que não lhe satisfaz. Os resultados podem ser conferidos na tabela 7 apresentada a seguir.

Tabela 7. Objetivo ao fazer um curso superior

RESPOSTAS	GRUPO 1		GRUPO 2	
	Freqüência(%)	Acumulada(%)	Freqüência(%)	Acumulada(%)
1-Emprego	14(8,10%)	14(8,10%)	21(19,3%)	21(19,3%)
2-Aumento salarial	11(6,40%)	25(14,5%)	8(7,30%)	29(26,6%)
3-Diploma de nível superior	13(7,60%)	38(22,1%)	17(15,6%)	46(42,2%)
4-Formação teórica	4(2,30%)	42(24,4%)	1(0,90%)	47(43,1%)
5-Formação profissional	130(75,6%)	172(100%)	62(56,9%)	109(100%)
TOTAL	172(100%)		109(100%)	

Fonte: autor

A próxima variável a ser descrita indica o nível de satisfação do aluno ou ex-aluno com relação à infra-estrutura da IES. Estranhamente os entrevistados do grupo 2 se mostraram mais satisfeitos com a estrutura da IES do que os entrevistados do grupo 1. Algumas conjecturas poderiam ser levantadas com o objetivo de explicar este resultado. Talvez, o reduzido tempo de permanência dos desistentes na IES, o fato de muitos se afastarem do curso ainda no primeiro período. Já os indivíduos do grupo 1, que pretendem concluir o curso, lutam por melhorias e fazem questão de declarar a sua insatisfação. Conforme ilustra o gráfico 3.

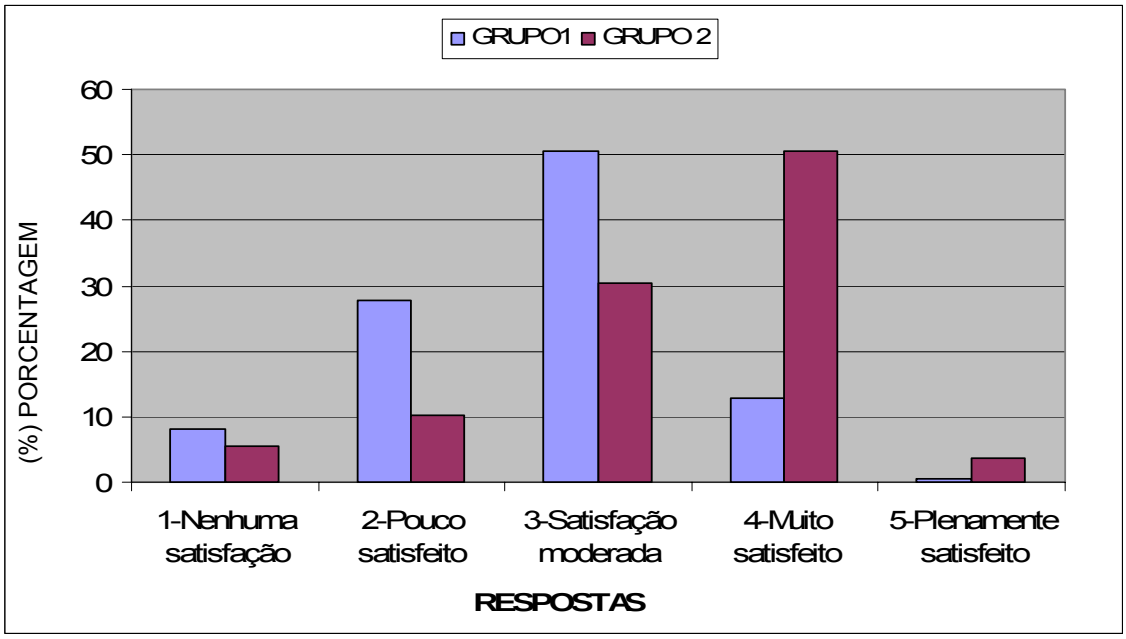

Gráfico 3. Satisfação com relação à infra-estrutura

A variável que representa a classificação do atendimento oferecido pelas coordenações e secretarias se apresenta com resultado semelhante à variável discutida anteriormente, ou seja, os elementos do grupo 2 mostraram-se mais satisfeitos com a qualidade dos serviços prestados por estes setores da IES do que os elementos do grupo 1. O acumulado das respostas “desinteressados”, “fraco” e “moderado” é de 61% para o grupo dos alunos e 30,3% para o grupo dos ex-alunos.

Não serão descritos os resultados para a questão 22, por ser uma complementação da questão 12, já descrita no início desta seção. Para melhor entendimento dos resultados relativos à questão “com relação ao tempo destinado

para os seus estudos, você diria que...", tem-se a tabela 8. A resposta "tempo satisfatório" tem 34,3% no grupo 1 e 22,9% no grupo 2, mas a maior diferença se encontra na porcentagem acumulada para as respostas 1 e 2, totalizando 15,7% para o primeiro grupo e 45,9% para o segundo grupo, conforme a tabela 8.

Tabela 8. Tempo destinado para os estudos

RESPOSTAS	GRUPO 1		GRUPO 2	
	Freqüência(%)	Acumulada(%)	Freqüência(%)	Acumulada(%)
1-Não há tempo	4(2,30%)	4(2,30%)	19(17,4%)	19(17,4%)
2-Tempo insuficiente	23(13,4%)	27(15,7%)	31(28,4%)	50(45,9%)
3-Há pouco tempo	66(38,4%)	93(54,1%)	33(30,3%)	83(76,1%)
4-Tempo satisfatório	59(34,3%)	152(88,4%)	25(22,9%)	108(99,1%)
5-Tempo mais que satisfatório	20(11,6%)	172(100%)	1(0,90%)	109(100%)
TOTAL	172(100%)		109(100%)	

Fonte: autor

Para a nota média no vestibular, foram calculadas as estatísticas média, mediana, desvio padrão e moda. Para o grupo 1, foi encontrada média igual a 3,69, mediana 3,71, moda 4 e desvio padrão 0,70. No grupo 2 obteve-se média igual a 3,15, mediana 3,14, moda 3,28 e desvio padrão 0,96. No princípio, esperava-se uma maior diferença para as médias dos grupos 1 e 2 em relação à nota no vestibular, expectativa que não se confirmou. Mesmo assim, um dos testes de inferência ao nível de significância de 5% classificou a nota média no vestibular como uma variável discriminante. A distribuição das notas pode ser observada na tabela 9.

Tabela 9. Nota média no vestibular

NOTAS	GRUPO 1		GRUPO 2	
	Freqüência(%)	Acumulada(%)	Freqüência(%)	Acumulada(%)
1 ⊢ 2	1(0,58%)	1(0,580%)	13(11,9%)	13(11,9%)
2 ⊢ 3	24(14,0%)	25(14,58%)	37(34,0%)	50(45,9%)
3 ⊢ 4	81(47,1%)	106(61,68%)	36(33,0%)	86(78,9%)
4 ⊢ 5	58(33,7%)	164(95,38%)	17(15,6%)	103(94,6%)
5 ⊢ 6	8(4,62%)	172(100%)	6(5,50%)	109(100%)
TOTAL	172(100%)		109(100%)	

Fonte: autor

Analisando a tabela 9, nota-se semelhanças entre os grupos, mostrando que os candidatos estão, de maneira geral, nivelados.Os resultados também mostram o estado lastimável da educação no Brasil. Visto que, numa prova de nível básico os postulantes a uma vaga em uma IES privada, em sua maioria, não alcançam nota média igual a 5,0 pontos - numa escala de zero a 10,0.

A questão 24 do questionário (Apêndice F) foi direcionada apenas aos ex-alunos com o intuito de identificar os motivos do afastamento do curso. A alternativa com maior freqüência é a número 2 “dificuldades financeiras” , seguido por 1 “escolha errada do curso”. O motivo 2 não foi identificado pelos questionários como uma variável discriminante. Pensou-se, inicialmente, que a renda familiar poderia auxiliar neste quesito. Ocorre que os entrevistados ficam constrangidos diante da questão 11 – “qual é a renda familiar?”. Muitos relutam em responder e quando o fazem passam a impressão de que não estão sendo verdadeiros. Ainda deve-se considerar que “dificuldades financeiras” são determinadas por várias componentes, não somente pela “renda familiar”. A tabela 10 apresenta os resultados obtidos para a questão 24.

Tabela 10. Grupo 2 - Motivo do afastamento do curso.

Respostas	Freqüência	Porcentagem(%)	(%) Acumulada
1-Escolha errada do curso.	19	17,4	17,4
2-Dificuldades financeiras.	29	26,6	44,0
3-Não ter conseguido conciliar estudo e trabalho.	12	11,0	55,0
4-Dificuldades p/ acompanhar o curso.	11	10,1	65,1
5-Professores despreparados.	1	0,9	66,1
6-Infra-estrutura da IES é deficiente.	13	11,9	66,1
7-Mudança de cidade.	12	11,0	78,0
8-Outra.	12	11,0	89,0
TOTAL	109	100,0	100,00

Fonte: autor

5.3 CLASSIFICAÇÃO DE FISHER PARA OS GRUPOS 1 E 2

Selecionadas as variáveis discriminantes, foi construída a partir destas uma função que classifique os indivíduos em um dos dois grupos. Para tanto foi utilizou-se a FDL de Fisher e em seguida as probabilidades de classificações incorretas foram calculadas através do método de Lachenbruch.

5.3.1 Construção da FDL de Fisher

Nesta seção foi construída a FDL de Fisher para classificar indivíduos no grupo 1 ou 2. De início foram formados dois grupos de controle: para criá-los utilizou-se a amostragem sistemática iniciando pela primeira linha da matriz (primeira observação) e, a cada 5 linhas (cinco observações), uma era retirada para compor o grupo de controle. Desta forma, conseguiu-se um grupo de controle de alunos com 34 elementos e um grupo de controle de ex-alunos com 21 elementos. Conseqüentemente, os grupos 1 e 2 tiveram uma redução no número de indivíduos, passando para 138 e 88, respectivamente. Para classificar os indivíduos segundo a FDL de Fisher, foi utilizado o programa PGR01 – Apêndice A. Esse programa utiliza três matrizes de entrada (x1, x2 e xc), sendo as duas primeiras obtidas com a redução dos grupos 1 e 2 e a terceira formada pelo grupo de controle originário do grupo 1 ou 2.

Nos itens 5.1.1 e 5.1.2, foram aplicados dois métodos para a obtenção das variáveis que discriminam os dois grupos; no primeiro, foram encontradas quatro e no segundo sete variáveis discriminantes. O primeiro conjunto de variáveis será denominado de A e o segundo de B. O programa PGR01 – (Apêndice A) foi alimentado com as matrizes x1, x2 e xc. Considerando o conjunto A, as matrizes de entrada são $x1_{138x4}$, $x2_{88x4}$ e xc_{34x4} (grupo de controle de alunos). Num segundo momento, a matriz xc_{34x4} foi substituída por xc_{21x4} (grupo de controle de ex-alunos). Após a execução do aplicativo, foram obtidos os resultados:

Tabela 11. Resultados da FDL de Fisher para o Conjunto A

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de controle	1 (xc_{34x4})	26	8	34
	2 (xc_{21x4})	4	17	21

Fonte: Autor

Ou seja, 24% (8/34) dos indivíduos pertencentes ao grupo 1 foram classificados erradamente no grupo 2 e 19% (4/21) dos indivíduos pertencentes ao grupo 2 foram classificados erradamente no grupo 1. A FDL de Fisher foi obtida a partir das matrizes $x1_{138x4}$ e $x2_{88x4}$, a função encontrada é expressa por:

$$y = \begin{pmatrix} 0{,}8459 & -0{,}7131 & -0{,}5912 & 0{,}7850 \end{pmatrix} \begin{pmatrix} VAR9 \\ VAR16 \\ VAR18 \\ VAR20 \end{pmatrix}$$, com média univariada(ym) igual a -0,0966. Assim, se $y \geq ym$, aloca-se a observação $\begin{pmatrix} VAR9 \\ VAR16 \\ VAR18 \\ VAR20 \end{pmatrix}$ no grupo 1. Caso contrário, aloca-se no grupo 2.

O mesmo procedimento foi aplicado ao conjunto de variáveis B, sendo $x1_{138x7}$, $x2_{88x7}$ e xc_{34x7} ou xc_{21x7}. A tabela 12 exibe os resultados obtidos .

Tabela 12. Resultados da FDL de Fisher para o conjunto B

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de controle	1 (xc_{34x7})	27	7	34
	2 (xc_{21x7})	6	15	21

Fonte: Autor

Neste caso, os elementos classificados erradamente no grupo 2 sendo de 1 foram de 21%(7/34) e os classificados erradamente em 1 sendo de 2, 29%(6/21). A FDL de Fisher obtida é a seguinte:

$$y = \begin{pmatrix} 0{,}5934 & 0{,}4823 & -1{,}4537 & -0{,}7235 & 1{,}0263 & 0{,}7653 & 0{,}8888 \end{pmatrix} \begin{pmatrix} VAR9 \\ VAR15 \\ VAR16 \\ VAR18 \\ VAR19 \\ VAR20 \\ VAR21 \end{pmatrix}$$

sendo a média univariada(ym) = 4,7763. Desta forma, se $y \geq ym$, aloca-se a observação $\begin{pmatrix} VAR9 \\ VAR15 \\ VAR16 \\ VAR18 \\ VAR19 \\ VAR20 \\ VAR21 \end{pmatrix}$ no grupo 1. Caso contrário, aloca-se no grupo 2.

5.3.2 Porcentagem de Classificações Incorretas - Método de Lachenbruch

O método de Lachenbruch é um dos métodos que pode ser utilizado para estimar a porcentagem de classificar erradamente um indivíduo. Esse método foi discutido na seção 3.3.3, e será executado pelo programa PGR02 (Apêndice B). Para o conjunto A, obteve-se a seguinte matriz de confusão:

Tabela 13. Matriz de Confusão – Lachenbruch (Conjunto A)

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de origem	1 ($x1_{172x4}$)	127	45	172
	2 ($x2_{109x4}$)	22	87	109

Fonte: Autor

Ou seja:

- Porcentagem de classificar erradamente um elemento no grupo 2, sendo ele de 1 é igual a 26%(45/172).
- Porcentagem de classificar erradamente um elemento no grupo 1, sendo ele de 2 é igual a 20%(22/109).

A matriz de confusão obtida para o conjunto B é:

Tabela 14. Matriz de Confusão – Lachenbruch (Conjunto B)

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de origem	1 ($x1_{172x7}$)	144	28	172
	2 ($x2_{109x7}$)	19	90	109

Fonte: Autor

Ou seja:

- Porcentagem de classificar erradamente um elemento no grupo 2, sendo ele de 1 é igual a 16%(28/172).
- Porcentagem de classificar erradamente um elemento no grupo 1, sendo ele de 2 é igual a 17%(19/109).

Comparando-se os resultados obtidos para o conjunto A e B, observa-se que a classificação apresentou um erro menor quando foi utilizado as variáveis do conjunto B.

5.4 CLASSIFICAÇÃO A PARTIR DA REGRESSÃO LOGÍSTICA

Este trabalho aborda duas técnicas para classificação de indivíduos. A primeira, já aplicada é a FDL de Fisher. A segunda, é a técnica denominada de regressão logística, apresentada na seção 3.4.

5.4.1 Resultados Para o Conjunto A

Para calcular os parâmetros da função logit e classificar um indivíduo em um dos dois grupos, foi aplicado o programa PGR03 (Apêndice C), que utiliza três matrizes de entrada (x1, x2 e xc), sendo xc a matriz para os grupos de controle. No caso do conjunto A, foram fornecidas ao programa PGR03 as matrizes $x1_{138x4}$, $x2_{88x4}$ e xc_{34x4} (grupo de controle dos alunos) e numa segunda rodada a matriz xc_{34x4} foi substituída por xc_{21x4} (grupo de controle dos ex-alunos) . Na tabela 15, os indivíduos classificados são originários das matrizes x1 e x2.

Tabela 15. Matriz de confusão – Logit (Conjunto A)

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de origem	1 ($x1_{138x4}$)	119	19	138
	2 ($x2_{88x4}$)	36	52	88

Fonte: Autor

Ou seja:

- Porcentagem de classificar erradamente um elemento no grupo 2, sendo ele de 1 é igual a 14%(19/138).

- Porcentagem de classificar erradamente um elemento no grupo 1, sendo ele de 2 é igual a 41%(36/88).

A tabela 16 mostra os acertos e erros quando se utiliza a função logit, obtida a partir de x1 e x2 para classificar um indivíduo proveniente de um dos grupos de controle (xc_{34x4} e xc_{21x4}).

Tabela 16. Resultados da Logit para o conjunto A

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de controle	1 (xc_{34x4})	29	5	34
	2 (xc_{21x4})	8	13	21

Fonte: Autor

Ou seja:

- Porcentagem de classificar erradamente um elemento no grupo 2, sendo ele de 1 é igual a 15%(5/34).
- Probabilidade de classificar erradamente um elemento no grupo 1, sendo ele de 2 é igual a 38%(8/21).

Nota-se que os erros de classificação obtidos com os elementos dos grupos de controle (tabela 16), estão próximos dos erros da matriz de confusão – tabela 15, salientando que a matriz de confusão mencionada foi obtida com a classificação dos elementos de x1 e x2. Utilizando-se as matrizes $x1_{138x4}$ e $x2_{88x4}$, obteve-se a função logit $g(x)$, sendo:

$$g(x) = -0{,}8233 - 1{,}2435VAR9 + 0{,}8053VAR16 + 0{,}8052VAR18 - 0{,}8064VAR20\,.$$

5.4.2 Resultados Para o Conjunto B

Usa-se para o conjunto B, um procedimento semelhante ao utilizado para A. Agora com $x1_{138x7}$, $x2_{88x7}$, xc_{34x7} e xc_{21x7}, os resultados obtidos foram:

Tabela 17. Matriz de confusão – Logit (Conjunto B)

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de origem	1 ($x1_{138x7}$)	123	15	138
	2 ($x2_{88x7}$)	16	72	88

Fonte: Autor

Ou seja:

- Porcentagem de classificar erradamente um elemento no grupo 2, sendo ele de 1 é igual a 11%(15/138).
- Porcentagem de classificar erradamente um elemento no grupo 1, sendo ele de 2 é igual a 18%(16/88).

Analisando as tabelas 17 e 18 observa-se que a probabilidade de classificar erradamente um elemento no grupo 1, sendo ele de 2 é muito maior para os dados provenientes do grupo de controle. Na matriz de confusão (tabela 17), o percentual de classificação é de 18% enquanto a classificação dos elementos de xc_{21x7} (tabela 18) observa-se um percentual de 38%. Observa-se também que a função logit obtida, tanto para o conjunto A como para B, é mais apropriada para classificar elementos provenientes do grupo 1.

Tabela 18. Resultados da Logit para o Conjunto B.

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de controle	1 (xc_{34x7})	30	4	34
	2 (xc_{21x7})	8	13	21

Fonte: Autor

Ou seja:

- Porcentagem de classificar erradamente um elemento no grupo 2, sendo ele de 1 é igual a 12%(4/34).
- Porcentagem de classificar erradamente um elemento no grupo 1, sendo ele de 2 é igual a 38%(8/21).

No caso do conjunto B, obteve-se a seguinte função logit:

$g(x)$ = 2,3454 – 0,6628VAR9 – 0,5095VAR15 + 1,5227VAR16 + 1,1272VAR18 +

+ (-1,1170)VAR19 – 0,6733VAR20 – 0,7232VAR21

5.5 CLASSIFICAÇÃO A PARTIR DE ESCORES FATORIAIS

A eficiência das variáveis discriminantes obtidas pode ser questionada quando se analisa alguns dos resultados de classificação, como por exemplo, o obtido com a função logit onde o erro na classificação dos indivíduos provenientes do grupo 2 é de 38%. Desta forma, optou-se nesta seção por abandonar as variáveis

discriminantes e passar a alimentar a FDL de Fisher e a função logit com escores fatoriais, em busca de melhores resultados.

5.5.1 Determinação dos Escores Fatoriais

Para o cálculo dos escores fatoriais, utilizou-se o aplicativo computacional *Statistica* com as opções “extração por componentes principais” e “rotação varimax normalizada” . O aplicativo foi alimentado com a matriz de dados amostrais, retornando nove autovalores maiores que 1. A tabela 19 mostra os autovalores e a variância explicada.

Tabela 19. Autovalores e Variância Explicada

	Autovalores	Variância Explicada	Autovalores Acumulados	Variância Acumulada
1	2,830211	13,47719	2,830211	13,47719
2	2,067723	9,846298	4,897933	23,32349
3	1,740386	8,287550	6,638319	31,61104
4	1,515300	7,215715	8,153619	38,82676
5	1,347728	6,417750	9,591346	45,24451
6	1,296416	6,173411	10,79776	51,41792
7	1,202485	5,726118	12,00025	57,14404
8	1,035672	4,931773	13,03592	62,07581
9	1,011296	4,815696	14,04722	66,89151

Fonte: Autor

Desta forma, as 21 variáveis iniciais foram substituídas por 9 fatores. Assim foram formadas duas matrizes de escores fatoriais, uma representando o grupo de alunos (grupo 1) de dimensões 172x9 e outra representando o grupo de ex-alunos (grupo 2) de dimensões 109x9. Analisando-se a matriz de pesos, obtida a partir da análise fatorial, podem-se distinguir alguns grupos de variáveis com correlações altas entre si, na tabela 20 estão listados os pesos.

Tabela 20. Matriz de Pesos

	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5	Fator 6	Fator 7	Fator 8	Fator 9
VAR1	0,014456	0,533181	0,02582	0,133206	-0,32978	-0,1524	0,260128	-0,25506	0,033055
VAR2	0,808322	-0,02276	0,167738	0,117692	-0,03835	-0,08623	0,147058	-0,04057	0,077277
VAR3	0,796409	-0,15477	-0,06122	-0,13082	0,020841	0,031941	-0,019	-0,05225	0,082565
VAR4	0,704123	0,113146	-0,04229	0,007601	-0,15362	-0,16478	-0,1537	0,233218	0,039687
VAR5	-0,16307	0,696639	0,109107	-0,24818	0,11547	0,066896	-0,09123	0,053621	-0,13847
VAR6	-0,34338	0,025562	-0,05752	-0,16621	0,729317	0,021613	0,036849	0,091253	-0,03937
VAR7	-0,40279	0,184666	-0,10713	-0,15864	0,585616	0,003516	0,073629	0,120151	0,14199
VAR8	0,281567	0,071119	0,048955	0,19175	0,735758	-0,04602	0,033457	-0,12823	0,002872
VAR9	0,05875	0,088276	-0,21095	-0,18613	0,041763	0,151424	0,014776	-0,12607	0,755414
VAR10	0,147799	0,037936	0,026617	0,049826	0,274936	0,079112	0,694912	-0,13376	0,065308

Tabela 20. Matriz de Pesos. Continuação.

	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5	Fator 6	Fator 7	Fator 8	Fator 9
VAR11	-0,06846	-0,00638	-0,11798	-0,01397	-0,06584	0,004537	0,794885	0,14445	-0,01587
VAR12	-0,09815	0,227267	-0,13287	-0,04207	-0,03782	0,710921	0,093336	0,211393	-0,00309
VAR13	-0,06829	-0,1094	0,076717	0,053476	0,015861	0,799179	-0,01573	-0,20293	-0,01567
VAR14	0,020998	0,026894	0,058719	0,05631	0,018857	-0,04398	0,046378	0,856085	0,005011
VAR15	0,008863	-0,03573	0,0067	0,901729	-0,03444	0,022416	0,020733	0,043351	-0,00084
VAR16	-0,00618	-0,16687	0,833916	0,036707	0,04977	-0,07294	-0,04723	0,097434	0,14171
VAR17	-0,00374	-0,08861	0,448854	0,107916	-0,06679	-0,01716	0,031674	0,259006	0,557166
VAR18	0,105376	0,145864	0,766619	-0,07378	-0,10482	0,080091	-0,05736	-0,07773	-0,08426
VAR19	0,129958	0,081812	0,247576	0,134401	-0,00307	-0,14757	0,014586	0,041959	0,783747
VAR20	-0,0047	0,777229	-0,16496	0,055823	0,142025	0,038072	0,019307	0,061798	0,23335
VAR21	-0,04847	-0,01251	-0,22286	0,36402	0,221791	-0,05158	-0,25344	0,0897	0,364913

Fonte: Autor

Considerando-se pesos iguais ou superiores a 0,7, observam-se na tabela 20 grupos de variáveis correlacionadas, por exemplo: o fator 1 que pode representar as variáveis 2, 3 e 4.

Da mesma forma, como em análises anteriores, foram construídos dois grupos de controle. Para formá-los, utilizou-se a amostragem sistemática iniciando pela primeira linha da matriz e, a cada 5 linhas, retirou-se uma linha para compor o grupo de controle. Assim, conseguiu-se um grupo de controle de alunos com 34 elementos e um grupo de controle de ex-alunos com 21 elementos. Conseqüentemente os grupos 1 e 2 tiveram uma redução no número de indivíduos, passando para 138 e 88, respectivamente.

5.5.2 Escores Fatoriais na FDL de Fisher

O programa PGR01 – Apêndice A, foi alimentado com as matrizes $x1_{138x9}$, $x2_{88x9}$ e o grupo de controle formado por alunos xc_{34x9}. Num segundo momento, a matriz xc_{34x9} foi substituída por xc_{21x9} (grupo de controle de ex-alunos). Com isto, 14,7% dos indivíduos que formam o grupo de controle de alunos foram classificados erradamente no grupo 2 . Já o grupo de controle formado por ex-alunos teve 23,8% dos seus elementos classificados erradamente no grupo 1, conforme mostra a tabela 21.

Tabela 21. Resultados da FDL de Fisher para Escores Fatoriais

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de controle	1 (xc_{34x9})	29	5	34
	2 (xc_{21x9})	5	16	21

Fonte: Autor.

Ou seja:

- Porcentagem de classificar erradamente um elemento no grupo 2, sendo ele de 1 é igual a 14,7%(5/34).
- Porcentagem de classificar erradamente um elemento no grupo 1, sendo ele de 2 é igual a 23,8%(5/21).

Utilizando-se as matrizes $x1_{138x9}$ e $x2_{88x9}$ obteve-se a FDL de Fisher, apresentada a seguir:

$$y = \begin{pmatrix} 0,5072 & 1,8842 & -1,7113 & 0,6564 & 0,1469 & -0,4509 & 0,2872 & -0,2035 & 1,1231 \end{pmatrix} \underline{F}'$$

sendo $\underline{F} = \begin{pmatrix} F_1 & F_2 & \dots & F_9 \end{pmatrix}$ o vetor de escores fatoriais e a média univariada igual a -0,5234.

5.5.3 Escores Fatoriais na Função Logit

Para calcular os parâmetros da função logit e com ela classificar um indivíduo em um dos dois grupos foi utilizado o programa PGR03 – Apêndice C, este programa foi alimentado com as matrizes $x1_{138x9}$, $x2_{88x9}$, xc_{34x9} e xc_{21x9}. Os resultados obtidos estão apresentados na tabela 22, onde observam-se que 8,8% dos elementos do grupo de controle de alunos foram classificados erradamente no grupo 2, enquanto 23,8% dos indivíduos pertencentes ao grupo de controle de ex-alunos foram classificados erradamente no grupo 1.

Tabela 22. Resultados da Função Logit para Escores Fatoriais

		Grupo classificado pela regra		
		1	**2**	**Total**
Grupo de controle	1 (xc_{34x9})	31	3	34
	2 (xc_{21x9})	5	16	21

Fonte: Autor.

Ou seja:

- Porcentagem de classificar erradamente um elemento no grupo 2, sendo ele de 1 é igual a 8,8%(3/34).
- Porcentagem de classificar erradamente um elemento no grupo 1, sendo ele de 2 é igual a 23,8%(5/21).

A partir das matrizes $x1_{138x9}$ e $x2_{88x9}$ obteve-se a função logit $g(x)$, sendo:

$g(x) = -0,9403 - 0,5157F_1 - 2,0801F_2 + 1,9853F_3 – 0,6394F_4 – 0,0543F_5 +$

$+ 0,7009F_6 – 0,2466F_7 + 0,1210F_8 – 1,2935F_9$

5.6 COMPARAÇÃO ENTRE OS MÉTODOS

Os resultados obtidos com a classificação de elementos oriundos dos grupos de treinamento ($x1_{138xj}$ e $x2_{88xj}$, sendo que j pode assumir os valores 4, 7 ou 9) foram, de maneira geral satisfatórios, seja com a utilização das variáveis discriminantes ou escores fatoriais. Em compensação, alguns dos resultados obtidos a partir dos grupos de controle, apresentaram grandes diferenças em relação aos obtidos com os grupos de treinamento quando as variáveis discriminantes foram utilizadas. Com o uso de escores fatoriais obteve-se uma melhora nas porcentagens de acerto, especialmente com a função logit. O quadro 3 resume os resultados obtidos com a classificação de indivíduos provenientes dos grupos de controle.

DADOS	Método / Classificação errada	FDL de Fisher	Função Logit
Variável Discriminante (Conjunto A)	Grupo 2, sendo de 1	24%	15%
	Grupo 1, sendo de 2	19%	38%
Variável Discriminante (Conjunto B)	Grupo 2, sendo de 1	21%	12%
	Grupo 1, sendo de 2	29%	38%
Escores Fatoriais	Grupo 2, sendo de 1	14,7%	8,8%
	Grupo 1, sendo de 2	23,8%	23,8%

Quadro 3. Porcentagem de classificação errada

Observando o quadro 3 verifica-se que, de forma geral, ao utilizarem-se escores fatoriais obteve-se melhores resultados. Já a função logit foi mais eficiente que a FDL de Fisher, na classificação dos indivíduos pertencentes ao grupo 1, cuja porcentagem de acerto foi de 91,2%.

5.7 CLASSIFICAÇÃO DE UM NOVO INDIVÍDUO

Considerando os dados levantados nesta pesquisa, as variáveis do conjunto A se mostraram mais estáveis na discriminação dos grupos 1 e 2, sendo a função logit mais indicada para classificar elementos do grupo 1. A FDL de Fisher mostrou-se mais eficiente na classificação de indivíduos de ambos os grupos, mostrando certa vantagem na classificação de elementos do grupo 2. Os programas PGR01 e PGR03 podem ser usados para a classificação segundo a FDL de Fisher e função logit, respectivamente. Caso sejam conhecidos os coeficientes da FDL de Fisher e a média univariada , podemos fazer algumas mudanças em PGR01 para que este possa então classificar o novo indivíduo, com o devido fornecimento dos valores das variáveis do conjunto A para este indivíduo. A coleta destas informações não pode ocorrer antes que o novo aluno tenha contato com a estrutura da IES e do curso. É necessário que esteja consciente do dia-a-dia que enfrentará para estar apto a responder as questões que permitirão a sua classificação. No caso da função logit, caso sejam conhecidos os parâmetros (β), o programa PGR03 também se presta à classificação bastando, para isso, fazer algumas mudanças em sua estrutura. Vimos que os escores fatoriais possibilitaram uma diminuição no erro de classificação. Na classificação de indivíduos do grupo 2 o erro baixou de 38% para 23,8% e por este motivo, é importante levar em consideração a citada opção de classificação. Neste caso, os dados iniciais devem ser preparados antes de serem fornecidos aos programas PGR01 e PGR03. Tendo uma observação para $\underline{X}$, faz-se a padronização de $\underline{X}$ e em seguida obtém-se os escores fatoriais como foi mostrado na seção 3.5.5 .

6 CONCLUSÕES

Verificou-se ao longo deste trabalho, a diversidade de técnicas classificatórias oferecidas pela estatística multivariada. Caberá ao pesquisador fazer a escolha daquela que mais se adapta aos dados coletados e objetivos pretendidos. Vários foram os testes realizados no sentido de se obter os melhores resultados e, acredita-se que este objetivo tenha sido alcançado. Constatou-se as dificuldades encontradas pela IES estudada para o preenchimento das vagas ofertadas, pois o vestibular de 2006 apresentou 32 opções de cursos de graduação e somente 7 deles receberam matrículas, as quais implicaram em 346 novos alunos. Não se sabe quantas outras IES apresentam situações parecidas, mas, com certeza, a IES pesquisada não é um caso a parte.

Falhou-se parcialmente na determinação de variáveis discriminantes, porque as obtidas não se mostraram muito eficientes na discriminação dos dois grupos. O questionário usado não conseguiu captar elementos essenciais na decisão de um acadêmico continuar ou não os seus estudos, como por exemplo: o financeiro e o desagrado com o ensino ofertado pela instituição. Neste estudo, os escores fatoriais mostraram um desempenho superior ao das variáveis discriminantes na tarefa de classificar indivíduos em suas respectivas populações. Acredita-se que, de maneira geral, os resultados obtidos foram satisfatórios. Há cursos na IES pesquisada onde as desistências no primeiro período são da ordem de 40%. Classificar erradamente um elemento que permanecerá no curso entre aqueles que desistirão não é problema, principalmente quando esta classificação errada é somente de 8%, obtida a partir da função logit. Indiscutivelmente o número de variáveis pode ser aumentado e a forma de captá-las melhorada, de maneira a aumentar o número e a qualidade das variáveis discriminantes.

Os estudos aqui desenvolvidos podem, se usados adequadamente, contribuir com a redução da evasão escolar. A aplicação das regras de classificação junto aos novos alunos da instituição, antecipará o conhecimento dos possíveis desistentes. Assim, a instituição poderá fazer um acompanhamento destes alunos e incentivá-los na participação de programas que visem a redução da evasão escolar. Melhorar currículos na busca da interdisciplinaridade, tornar o curso mais atraente e promover a integração do acadêmico à instituição, conceder descontos na mensalidade, promover a participação do aluno com problemas financeiros em programas de

crédito do governo federal, viabilizar a integração social, acadêmica, cultural e profissional do estudante através da participação em semanas acadêmicas, empresa Júnior e projetos comunitários e oferecer orientação psicológica aos que necessitam são medidas que podem ajudar na redução da evasão escolar.

6.1 SUGESTÕES PARA TRABALHOS FUTUROS

Em trabalhos futuros, que abordem o mesmo tema, é interessante uma melhoria no questionário usado para a coleta de dados. A inserção de um número maior de questões, poderá melhor traçar o perfil dos elementos integrantes de cada grupo e delinear fatores que, certamente, são os responsáveis pela evasão escolar no ensino superior. A construção da maioria das questões deverá seguir o conceito da escala de Likert. Neste tipo de escala os pesquisados são solicitados a concordarem ou discordarem das afirmações, segundo uma escala que vai de 1(discordo totalmente) até 5 (concordo totalmente). Com relação ao desempenho escolar do aluno, uma opção seria uma análise mais aprofundada do histórico do ensino médio. Outro ponto importante é trabalhar com um universo maior, de forma a proporcionar uma maior segurança nos resultados obtidos. Sugere-se a aplicação de redes neurais como uma técnica complementar. E, por fim, é recomendável pesquisar instituições que apliquem questionários sócio-educacionais. O pesquisador já teria, de início, uma grande base de dados o que facilitaria, em muito, o seu trabalho.

REFERÊNCIAS

AGRESTI, A. **Categorical Data Analysis**. New York: John Wiley, 1990

ARAÚJO, E. C. de. **Algoritmos Fundamentos e Prática**. Florianópolis: VisualBooks, 2003.

BAKER, R. W. e SIRYK, B. S. **Student Adaptation To College Questionaire: Manual**. Los Angeles:Western Psychological Services, 1989.

CUNICO, L. H. B. **Técnicas em *data mining* aplicadas na predição de satisfação de uma rede de lojas do comércio varejista.** Dissertação de mestrado em métodos numéricos em engenharia UFPR. Curitiba, 2005

GAIOSO, N. P. L. **O fenômeno da evasão escolar educação superior no Brasil.** Brasília: UCB, 2005.

GIMENO, S. G. A . e SOUZA, J. M. P. **Utilização de estratificação e modelo de regressão logística na análise de estudos caso-controle.** Rev. Saúde Pública vol. 29 n. 4. São Paulo, ago 1995.

HAIR , J.F. ; ANDERSON, R.E. ; TATHAM, R.L. e BLAC, W.C. **Análise multivariada de dados**. Porto Alegre: Bookman, 2005

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em www.inep.gov.br

JOHNSON, R.A. e WICHERN, D.W. **Applied multivariate statistical analysis**. Londres:Prentice-Hall, 1998.

MINGOTI, S. A .. **Análise de Dados Através de Métodos de Estatística Multivariada**. Belo Horizonte: Editora UFMG, 2005

PEREIRA, F. C. B. **Determinantes da evasão de alunos e os custos ocultos para instituições de ensino superior: uma apliacação na Universidade Sul Catarinennse.** Tese de doutorado em engenharia de produção UFSC. Florianópolis, 2003

PIZZOL, S. J. S. de. **Combinação de grupos focais e aálise discriminante: um método para tipificação de sistemas de produção agropecuária.** *Ver. Econ. Social. Rural*, 2004, Vol 42, n.3, p.451-468. ISSN 0103-2003.

POLYDORO, S. A. J. , PRIMI, R. *et al* . **Desenvolvimento de uma escala de integração ao ensino superior.** *Psico-USF,* Jan/Jun. 2001, vol. 6, n.1, p.11-17.

REIS, E. **Estatística Multivariada Aplicada**. Lisboa: Edições Sílabo, 2001.

RODRIGUES, A . M., **Técnicas de *Data Mining* classificadas do ponto de vista do usuário**. Dissertação de mestrado em engenharia de produção da UFRJ . Rio de Janeiro, 2000.

SANTOS, A. M. dos, SEIXAS, J. M. de, PEREIRA, B. de B. *et al* . **Using artificial neural networks and logistic regression in prediction of Hepatitis A.** *Rev. bras. epidemiol.,* June 2005, vol 8, n.2, p. 117-126. ISSN 1415-790X

SCHWARTZMAN, J. **O financiamento das Instituições de Ensino Superior no Brasil**. UNESCO, 2003.

SCHWARTZMAN, S. **O ensino superior no Brasil.** Brasília: INEP, 1999

APÊNDICE A – PROGRAMA PARA A FDL DE FISHER

O programa PGR01 desenvolvido no *MATLAB* calcula os coeficientes da FDL de fisher e classifica um novo indivíduo.

```
%PROGRAMA :   PGR01
%O PROGRAMA CALCULA OS COEFICIENTES DA FDL DE FISHER
% x1= MATRIZ DE DADOS DA POPULAÇÃO 1
% x2 = MATRIZ DE DAOS DA POPULAÇÃO 2
% n1 = N. DE OBSERVAÇÕES POP 1
% n2 = N. DE OBSERVAÇÕES POP 2
% p = N. DE VARIÁVEIS
x1=input('MATRIZ DE DADOS POPULAÇÃO 1, x1=');
x2=input('MATRIZ DE DADOS POPULAÇÃO 2,  x2=');
[n1,p]=size(x1);
[n2,p]=size(x2);
xm1=mean(x1);
xm2=mean(x2);
s1=cov(x1);
s2=cov(x2);
sp=((n1-1)*s1+(n2-1)*s2)/(n1+n2-2);
spi=inv(sp);
xm=xm1-xm2;
c=xm*spi;
ym1=c*xm1';
ym2=c*xm2';
ym=(ym1+ym2)/2;
p1=0;
p2=0;
xc=input('MATRIZ DE DADOS DE INDIVÍDUOS CLASSIFICÁVEIS NA POP 1 OU POP2,
xc=');
[nc,p]=size(xc);
xt=xc';
for i=1:nc
y=c*xt(:,i);
if y > ym
p(i)=1;
p1=p1+1;
else
p(i)=2;
p2=p2+1;
end
end
disp(' ')
disp('########################################################################')
disp(' ')
disp('Os indivíduos classificados na pop 2 (DESISTENTES), são:')
for i=1:nc
if p(i)==2
disp(i)
end
end
disp('elementos classificados na pop 1')
disp(p1)
disp('elementos classificados na pop 2')
disp(p2)
```

APÊNDICE B – MÉTODO DE LACHENBRUCH – PROGRAMA

O programa PGR02 desenvolvido no *MATLAB* , tem como objetivo a resolução do método de Lachenbruch.

```
%PROGRAMA :  PGR02
%O PROGRAMA EXECUTA O MÉTODO DE LANCHENBRUCH
% x1= MATRIZ DE DADOS DA POPULAÇÃO 1
% x2 = MATRIZ DE DAOS DA POPULAÇÃO 2
% n1 = N. DE OBSERVAÇÕES POP 1
% n2 = N. DE OBSERVAÇÕES POP 2
% p = N. DE VARIÁVEIS
x1=input('MATRIZ DE DADOS POPULAÇÃO 1, x1=');
x2=input('MATRIZ DE DADOS POPULAÇÃO 2,  x2=');
[n1,p]=size(x1);
[n2,p]=size(x2);
xm1=mean(x1);
xm2=mean(x2);
s1=cov(x1);
s2=cov(x2);
p1=0;
np1=0;
p2=0;
np2=0;
k=1
while k<(n1+1)
L=1;
x0=x1(k,:);
for i=1:n1
if i~=k
for j=1:p
x1r(L,j)=x1(i,j);
end
L=L+1;
end
end
s1r=cov(x1r);
sp=((n1-2)*s1r+(n2-1)*s2)/(n1+n2-3);
spi=inv(sp);
xm1r=mean(x1r);
xm=xm1r-xm2;
c=xm*spi;
ym1=c*xm1r';
ym2=c*xm2';
ym=(ym1+ym2)/2;
y0=c*x0';
if y0 > ym
p1=p1+1;
```

```
else
np1=np1+1;
end
k=k+1;
end
k=1;
while k<(n2+1)
L=1;
x0=x2(k,:);
for i=1:n2
if i~=k
for j=1:p
x2r(L,j)=x2(i,j);
end
L=L+1;
end
end
s2r=cov(x2r);
sp=((n1-1)*s1+(n2-2)*s2r)/(n1+n2-3);
spi=inv(sp);
xm2r=mean(x2r);
xm=xm1-xm2r;
c=xm*spi;
ym1=c*xm1';
ym2=c*xm2r';
ym=(ym1+ym2)/2;
y0=c*x0';
if y0 < ym
p2=p2+1;
else
np2=np2+1;
end
k=k+1;
end
disp('################################################')
disp('NÚMERO DE ELEMENTOS DA POPULAÇÃO 1')
disp('CLASSIFICADOS ERRADAMENTE NA POPULAÇÃO 2:')
disp(np1)
disp('NÚMERO DE ELEMENTOS DA POPULAÇÃO 1')
disp('CLASSIFICADOS CORRETAMENTE EM 1:')
disp(p1)
disp('NÚMERO DE ELEMENTOS DA POPULAÇÃO 2')
disp('CLASSIFICADOS ERRADAMENTE NA POPULAÇÃO 1:')
disp(np2)
disp('NÚMERO DE ELEMENTOS DA POPULAÇÃO 2')
disp('CLASSIFICADOS CORRETAMENTE EM 2:')
disp(p2)
disp('################################################')
disp('PROBABILIDADE DE CLASSIFICAR ERRADAMENTE')
disp('UM ELEMENTO NA POPULACAO 2, SENDO ELE DE 1:')
p21=np1/n1;
```

```
disp(p21)
disp('PROBABILIDADE DE CLASSIFICAR ERRADAMENTE')
disp('UM ELEMENTO NA POPULAÇÃO 1, SENDO ELE DE 2:')
p12=np2/n2;
disp(p12)
M=[p1 np1;np2 p2];
disp('  M  A  T  R  I  Z    D  E    C  O  N  F  U  S  Ã  O')
disp(' ')
disp(M)
```

APÊNDICE C – PARÂMETROS DA FUNÇÃO LOGIT – PROGRAMA

O programa PGR03 escrito no *MATLAB* calcula os parâmetros da função logit e classifica novos indivíduos.

```
%PROGRAMA : PGR03
%O PROGRAMA DETERMINA OS PARÂMETROS DA REGRESSÃO LOGISTICA
%CALCULA AS PROBABBILIDADES DE CLASSIFICAÇÃO CORRETA E ERRADA
%FORNECE A MATRIZ DE CONFUSÃO
% x1=MATRIZ DE DADOS DO GRUPO 1, GRUPO CUJA VARIÁVEL RESPOSTA É
Y = 0
% x2=MATRIZ DE DADOS DO GRUPO 2, GRUPO CUJA VARIÁVEL RESPOSTA É
Y = 1
x1=input('MATRIZ DE DADOS DO GRUPO 1 (n1xp), VARIÁVEL RESPOSTA
Y=0, x1=');
x2=input('MATRIZ DE DADOS DO GRUPO 2 (n2xp), VARIÁVEL RESPOSTA
Y=1, x2=');
[n1,p]=size(x1);
[n2,p]=size(x2);
n=n1+n2;
for i=1:(p+1)
b(i)=0;
beta=b';
end
for i=1:n1
y0(i)=0;
end
for i=1:n2
y1(i)=1;
end
y=[y0';y1'];
A=[x1;x2];
for i=1:n
u(i)=1;
end
x=[u' A];
X=x';
j=0;
while j < 10
for i=1:n
E=beta'*X(:,i);
p1(i)=(exp(E))/(1+exp(E));
p2=p1(i)*(1-p1(i));
B(i,:)=p2*x(i,:);
end
P=X*B;
I=inv(P);
beta=beta+I*X*(y-p1');
j=j+1;
```

```
end
disp(' ')
disp('############################################################
####')
disp('COEFICENTES DA FUNÇÃO LOGIT: G(X)= B0 + B1X1 + B2X2 +
...+ BpXp')
disp(' ')
u=['NESTE CASO TEMOS p=' num2str(p) ];
disp(' ')
disp(u)
disp(beta')
g1=0; g2=0; ng1=0; ng2=0;
for i=1:n1
r=beta'*X(:,i);
if r < 0
g1=g1+1;
else
ng1=ng1+1;
end
end
for i=(n1+1):n
r=beta'*X(:,i);
if r < 0
ng2=ng2+1;
else
g2=g2+1;
end
end
disp(' ')
disp('############################################')
disp('NÚMERO DE ELEMENTOS DA POPULAÇÃO 1(VARIÁVEL RESPOSTA É
Y=0)')
disp('CLASSIFICADOS ERRADAMENTE NA POPULAÇÃO 2 (VARIÁVEL
RESPOSTA EH Y=1):')
disp(ng1)
disp('NÚMERO DE ELEMENTOS DA POPULAÇÃO 1 (y=0)')
disp('CLASSIFICADOS CORRETAMENTE EM 1:')
disp(g1)
disp('NÚMERO DE ELEMENTOS DA POPULAÇÃO 2 (y=1)')
disp('CLASSIFICADOS ERRADAMENTE NA POPULAÇÃO 1:')
disp(ng2)
disp('NÚMERO DE ELEMENTOS DA POPULAÇÃO 2 (y=1)')
disp('CLASSIFICADOS CORRETAMENTE EM 2:')
disp(g2)
disp('############################################')
disp('PROBABILIDADE DE CLASSIFICAR ERRADAMENTE')
disp('UM ELEMENTO NA POPULAÇÃO 2 (y=1), SENDO ELE DE 1:')
p21=ng1/n1;
disp(p21)
disp('PROBABILIDADE DE CLASSIFICAR ERRADAMENTE')
disp('UM ELEMENTO NA POPULAÇÃO 1 (y=0), SENDO ELE DE 2:')
```

```
p12=ng2/n2;
disp(p12)
M=[g1 ng1;ng2 g2];
disp('  M  A  T  R  I  Z    D  E    C  O  N  F  U  S  Ã  O')
disp(' ')
disp(M)
pause
disp(' ')
xc=input('MATRIZ DE DADOS DOS INDIVÍDUOS CLASSIFICÁVEIS NA POP
1 OU POP 2, xc=');
[nc,p]=size(xc);
for i=1:nc
uc(i)=1;
end
C1=[uc;xc'];
for i=1:nc
rc=beta'*C1(:,i);
if rc < 0
g(i)=1;
else
g(i)=2;
end
end
disp('#############################################################
####################')
disp('Os indivíduos classificados no grupo 2 (DESISTENTES)
são:')
for i=1:nc
if g(i)==2
disp(i)
end
end
```

APÊNDICE D – MANOVA PARA 2 GRUPOS – PROGRAMA

O programa PGR04 desenvolvido no MATLAB executa a MANOVA para dois grupos.

```
%PROGRAMA : PGR04
%O PROGRAMA FAZ A ANÁLISE MANOVA PARA DOIS GRUPOS
% x1 = MATRIZ DE DADOS (n1xp) DO GRUPO 1
% x2 = MATRIZ DE DADOS (n2xp) DO GRUPO 2
% alfa = NÍVEL DE SIGNIFICÂNCIA
x1=input('MATRIZ DE DADOS (n1xp) DO GRUPO 1,  x1=');
x2=input('MATRIZ DE DADOS (n2xp) DO GRUPO 2,  x2=');
a=input('NÍVEL  DE SIGNIFICÂNCIA, alfa=');
[n1,p]=size(x1);
[n2,p]=size(x2);
n=n1+n2;
g=2;
alfa=a/100;
m1=mean(x1);
m2=mean(x2);
for i=1:p
m(i)=(n1*m1(i)+n2*m2(i))/(n1+n2);
end
B=n1*((m1-m)'*(m1-m))+n2*((m2-m)'*(m2-m));
w1=0; w2=0;
ww1=0;  ww2=0;
for i=1:n1
ww1=(x1(i,:)-m1)'*(x1(i,:)-m1);
w1=w1+ww1;
end
for i=1:n2
ww2=(x2(1,:)-m2)'*(x2(i,:)-m2);
w2=w2+ww2;
end
w=w1+w2;
% LÂMBDA DE WILKS
L=det(w)/det(B+w);
% ESTATÍSTICA DO TESTE
qui2=-(n-1-((p+g)/2))*log(L);
Q=chi2inv(1-alfa,p*(g-1));
if qui2 < Q
disp('Os vetores de médias dos grupos 1 e 2  NÃO apresentam
diferenças siginificativas')
else
disp('Os vetores de médias dos grupos 1 e 2 APRESENTAM
diferenças significativas')
%COMPARAÇÃO ENTRE AS COMPONENTES
for i=1:p
```

```
L1=(m1(i)-m2(i))+(tinv(alfa/(p*g*(g-1)),n-g))*sqrt((w(i,i)/(n-
g))*(1/n1+1/n2));
L2=(m1(i)-m2(i))-(tinv(alfa/(p*g*(g-1)),n-g))*sqrt((w(i,i)/(n-
g))*(1/n1+1/n2));
sinal=L1*L2;
if sinal < 0
R(i)=0;
else
R(i)=1;
end
end
disp('##########################################################
####################')
disp('As variáveis que apresentam diferenças significativas')
disp('entre os grupos 1 e 2, são as variáveis das colunas:')
for i=1:p
if R(i)==1
disp(i)
end
end
end
```

APÊNDICE E – PROGRAMA PARA INFERÊNCIA SOBRE MÉDIAS

O programa PGR05 desenvolvido no *MATLAB*, compara vetores de médias provenientes de duas populações com matrizes de covariâncias diferentes. Caso os vetores sejam considerados diferentes, o programa determina as componentes que diferem significativamente.

```
%PROGRAMA : PGR05
%COMPARA VETORES DE MÉDIAS PROVENIENTES DE 2 POPULAÇÕES
% x1=MATRIZ DE DADOS (n1xp) DA POPULAÇÃO 1
% x2=MATRIZ DE DADOS (n2xp) DA POPULAÇÃO 2
% alfa = NÍVEL DE SIGNIFICÂNCIA
x1=input('MATRIZ DE DADOS (n1xp) DA POPULAÇÃO 1, x1=')
x2=input('MATRIZ DE DADOS (n2xp) DA POPULAÇÃO 2, x2=')
alfa=input('NÍVEL DE SIGNIFICÂNCIA, alfa=')
[n1,p]=size(x1);
[n2,p]=size(x2);
x=[x1;x2];
[n,p]=size(x);
m1=mean(x1);
m2=mean(x2);
s1=cov(x1);
s2=cov(x2);
T2=(m1-m2)*(inv(s1/n1+s2/n2))*(m1-m2)';
%cálculo do qui-quadrado teórico - Q'
q1=chi2inv(1-alfa/100,p);
if T2 > q1
disp('Os vetores de médias dos grupos 1 e 2 APRESENTAM
diferenças significativas')
%COMPARAÇÃO ENTRE AS COMPONENTES
s=s1/n1+s2/n2;
for i=1:p
L1=(m1(i)-m2(i))+sqrt(q1)*sqrt(s(i,i));
L2=(m1(i)-m2(i))-sqrt(q1)*sqrt(s(i,i));
sinal=L1*L2;
if sinal < 0
R(i)=0;
else
R(i)=1;
end
end
disp('##########################################################
####################')
disp('As variáveis que apresentam diferenças significativas')
disp('entre os grupos 1 e 2, são as variáveis das colunas:')
for i=1:p
if R(i)==1
disp(i)
end
```

```
end
else
disp('Os vetores de médias dos grupos 1 e 2  NAO apresentam
diferenças significativas')
end
```

APÊNDICE F - QUESTIONÁRIO

01) Qual é o seu curso?.....

02) Situação atual : ( ) Aluno ( ) Ex-aluno

03) Se ex-aluno, assinale a sua situação:( ) Curso trancado ( ) Curso abandonado ( )Transferência ( ) Matrícula cancelada

04)	**Qual é o seu sexo?**	**09)**	**Nível de instrução do seu pai?**
Resp	Descrição	Resposta	Descrição
1	Masculino	1	Sem escolaridade
2	Feminino	2	Ensino Fundamental incompleto
		3	Ensino Fundamental completo
05)	**Qual é a sua idade?**	4	Ensino Médio incompleto
Resp	Descrição	5	Ensino Médio completo
1	Entre 18 e 23 anos(inclusive)	6	Ensino Superior incompleto
2	Entre 23 e 28 anos(inclusive)	7	Ensino Superior completo
3	Entre 28 e 33 anos(inclusive)		
4	Entre 33 e 38 anos(inclusive)	**10)**	**Nível de instrução da mãe?**
5	Entre 38 e 43 anos(inclusive)	Resposta	Descrição
6	Entre 43 e 48 anos(inclusive)	1	Sem escolaridade
7	Acima de 48 anos	2	Ensino Fundamental incompleto
		3	Ensino Fundamental completo
06)	**Qual é o seu estado civil?**	4	Ensino Médio incompleto
Resp	Descrição	5	Ensino Médio completo
1	Solteiro	6	Ensino Superior incompleto
2	Casado	7	Ensino Superior completo
3	Amasiado		
4	Divorciado	**11)**	**Qual é a renda familiar?**
5	Outro	Resposta	Descrição
		1	Até R$ 360,00
07)	**Com relação a sua moradia ?**	2	De R$ 361,00 a R$ 600,00
Resp	Descrição	3	De R$ 601,00 a R$ 1.000,00
1	Mora em casa própria dos pais	4	De R$ 1001,00 a R$ 1.500,00
2	Mora em casa dos pais, alugada	5	De R$ 1.501,00 a R$ 2.000,00
3	Mora em casa própria	6	De R$ 2.001,00 a R$ 2.500,00
4	Mora em casa alugada	7	De R$ 2.501,00 a R$ 3.000,00
5	Mora em república ou pensão	8	De R$ 3.001,00 a R$ 4.000,00
6	Mora em casa de parentes	9	De R$ 4.001,00 a R$ 5.000,00
7	Outro	10	Acima de R$ 5.000,00
08)	**Você tem trabalho remunerado?**	**12)**	**Com relação ao curso escolhido...**
Resp	Descrição	Resposta	Descrição
1	Sim	1	Você ficou decepcionado
2	Não	2	Era o que você esperava
3	Às vezes	3	Superou as sua expectativas
		4	Não sabe dizer

13)	**Onde você fez o Ensino Médio?**		
Resp	Descrição	**18)**	**Qual o seu principal objetivo ao fazer um curso superior?**
1	Integralmente em escola pública		
2	Totalmente em escola particular	Resposta	Descrição
3	Maior parte em escola pública	1	Emprego
4	Maior parte em escola particular	2	Aumento salarial
5	Outro	3	Diploma de nível superior
		4	Formação teórica
14)	**Em qual turno?**	5	Formação Profissional
Resp	Descrição	As questões 19) , 20) e 21) são referentes à faculdade onde você é aluno ou ex-aluno	
1	Integralmente no noturno		
2	Integralmente no diurno	**19)**	**Indique o nível da sua <u>satisfação</u> com relação a infra-estrutura (laboratórios, biblioteca,etc) oferecida pela faculdade**
3	Maior parte no noturno		
4	Maior parte no diurno		
15)	**Indique a sua principal razão na escolha da faculdade**	Resposta	Descrição
		1	Nenhuma satisfação
Resp	Descrição	2	Pouco satisfeito
1	Qualidade do ensino	3	Satisfação moderada
2	Localização	4	Muito satisfeito
3	Oferecer o curso pretendido	5	Plenamente satisfeito
4	Horário do curso		
5	Instalações	**20)**	**Informe a sua opinião com relação à <u>capacidade</u> dos professores na transmissão do conhecimento**
6	É a mais conhecida		
7	Outra		
16)	**Indique o seu principal motivo na escolha do curso**	Resposta	Descrição
		1	Nenhuma capacidade
Resp	Descrição	2	Pouca capacidade
1	Possibilidades salariais	3	Capacidade regular
2	Realização pessoal	4	Boa capacidade
3	Gosta das matérias do curso	5	Ótima capacidade
4	Baixa concorrência pelas vagas		
5	Permite conciliar aula e trabalho	**21)**	**Classifique o atendimento oferecido pelas coordenações e secretarias.**
6	Mercado de trabalho		
7	Outro		
		Resposta	Descrição
17)	**Durante o curso, você classificaria o seu relacionamento afetivo com os colegas como sendo...**	1	Desinteressados pelo problema do aluno
		2	Fraco, mescla de baixo interesse com reduzida eficiência.
Resp	Descrição	3	Moderado interesse e relativa eficiência
1	Inexistente		
2	Fraco, baixo envolvimento	4	Bom, mostram interesse e eficiência.
3	Moderado		
4	Bom	5	Ótimo, são interessados, educados e competentes
5	Ótimo		

ANEXO I - TEOREMA DA DECOMPOSIÇÃO ESPECTRAL

Seja Σ_{pxp} uma matriz de covariâncias. Então, existe uma matriz ortogonal O_{pxp} , isto é, O'O = OO' = I_{pxp} , tal que:

$$O'\sum O = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_p \end{bmatrix} = \Lambda \quad , \quad \text{onde} \quad \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \quad \text{são os autovalores}$$

ordenados em ordem decrescente da matriz Σ_{pxp} . Nesse caso, dizemos que a matriz Σ_{pxp} é similar à matriz Λ_{pxp} , o que implica em dizer que:

(i) $\det(\Sigma_{pxp}) = \left|\sum_{pxp}\right| = \left|\Lambda_{pxp}\right| = \prod_{i=1}^{p} \lambda_i$

(ii) traço(Σ_{pxp}) = traço(Λ_{pxp}) = $\lambda_1 + \lambda_2 + ... + \lambda_p$.

A i-ésima coluna da matriz O_{pxp} é o autovetor normalizado e_1 correspondente ao autovalor λ_i , i= 1, 2, ..., p , que é denotado por:

$e_1 = \begin{bmatrix} e_{i1} \\ e_{i2} \\ ... \\ e_{ip} \end{bmatrix}$. Então, a matriz O é dada por $O = \begin{bmatrix} e_1 & e_2 & ... & e_p \end{bmatrix}$ e pelo teorema da decomposição espectral tem-se que a seguinte igualdade é válida:

$\sum_{pxp} = O\Lambda O' = \sum_{i=1}^{p} \lambda_i e_i e_i'$, sendo e_1 um vetor de comprimento igual a 1, isto é, $\left\|e_i\right\| = (e_{i1}^2 + e_{i2}^2 + ... + e_{ip}^2)^{1/2} = 1$ e $e_i' e_j = 0$, $\forall i \neq j$, pela ortogonalidade da matriz O_{pxp} .

ANEXO II - ESTIMADORES DE MÁXIMA VEROSIMILHANÇA

Conhecendo n observações para X, (x_1, x_2, ..., x_n) , pretende-se estimar os parâmetros θ_i e para tal é necessário definir qual o melhor estimador $\hat{\theta}_j$. No método de máxima verossimilhança o estimador é encontrado a partir da maximização de uma função, a função de verossimilhança. A probabilidade de ocorrência de uma amostra aleatória de n observações [$L(\theta_1,..., \theta_k)$] é dada pela função densidade de probabilidade conjunta dos n elementos da amostra aleatória:

$$[L(\theta_1,...,\theta_k)] = f(x_1,x_2,...,x_n;\theta_1,...,\theta_k) = \prod_{i=1}^{n} f(x_i;\theta_1,...,\theta_k)$$

A função de verossimilhança é uma medida relativa da probabilidade de ocorrência de uma amostra específica de n elementos (x_1, x_2, ..., x_n). O método de verossimilhança permite encontrar estimadores para os parâmetros de tal modo que seja maximizada a função para uma amostra específica. Para encontrar os estimadores calculam-se os máximos da função de verossimilhança depois de logaritmizada $l(\theta_1,...\theta_k) = \ln[L(\theta_1,...,\theta_k)$ isto é, calculando as primeiras derivadas parciais em ordem a cada um dos parâmetros e igualando-os a zero, e verificando-se ainda que as segundas derivadas parciais são negativas.

ANEXO III – IGUALDADE ENTRE MATRIZES DE COVARIÂNCIAS

Em REIS(2001) encontra-se o método definido por BOX(1950), que consiste em testar a hipóteses:

H_0: $\Sigma_1 = \Sigma_2 = ... = \Sigma_g$ com $\hat{\Sigma} = \frac{W}{n-g} = S$ e

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underline{\mathbf{x}}_{\mathbf{ij}} - \underline{\bar{\mathbf{x}}}_{\mathbf{i}})(\underline{\mathbf{x}}_{\mathbf{ij}} - \underline{\bar{\mathbf{x}}}_{\mathbf{i}})'$$

H_1: $\Sigma_i = \Sigma_j$ $i \neq j$

Seja n a dimensão total da amostra, $v_i = n_i - 1$ os graus de liberdade associados a cada grupo, S_i a matriz de covariância do grupo i e S a matriz de covariância total. O teste M de Box define-se do seguinte modo:

$$M = (n-g)\ln|S| - \sum_{i=1}^{g} v_i \ln|S_i|$$

Box sugeriu duas aproximações para o seu teste: a distribuição do χ^2 e a distribuição F. Quando as dimensões dos grupos são superiores a 20, o número de variáveis e de grupos inferior a 6, a aproximação χ^2 é a indicada; em todas as outras situações deve-se optar pela aproximação F.

Aproximação à distribuição do χ^2 :

$$M.C \sim \chi^2_{\frac{1}{2}p(p+1)(g-1)} \quad \text{sendo} \quad C = 1 - \frac{2p^2+3p-1}{6(p+1)(g-1)}\left(\sum_{i=1}^{g}\frac{1}{v_i} - \frac{1}{n-g}\right)$$

Aproximação à distribuição F:

$$a_1 = 1 - C \qquad a_2 = \frac{(p-1)(p+2)}{6(g-1)}\left[\sum_{i=1}^{g}\frac{1}{v_i^2} - \frac{1}{(n-g)^2}\right]$$

$$v = \frac{p(p+1)(g-1)}{2} \quad \text{e} \quad v_0 = \frac{v+2}{a_2 - a_1^2}, \text{ então} \quad \frac{M\left(1 - a_1 - \frac{v}{v_0}\right)}{v} \sim F_{v,v_0}$$