UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Cibele Emília Cassiano

UMA ANÁLISE ESTATÍSTICA DESCRITIVA SOBRE O DESEMPENHO FINAL DOS GRADUANDOS DA UNIFEI EM FUNÇÃO DE SEUS DADOS DE APROVAÇÃO NO VESTIBULAR

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para obtenção do Título de ***Mestre em Engenharia de Produção.***

Área de Concentração: Qualidade e Produto

Orientador: Prof. Pedro Paulo Balestrassi, Dr.

Maio de 2009

Itajubá- MG

UNIVERSIDADE FEDERAL DE ITAJUBÁ

Cibele Emília Cassiano

UMA ANÁLISE ESTATÍSTICA DESCRITIVA SOBRE O DESEMPENHO FINAL DOS GRADUANDOS DA UNIFEI EM FUNÇÃO DE SEUS DADOS DE APROVAÇÃO NO VESTIBULAR

Dissertação aprovada por banca examinadora em 29 de maio de 2009, conferindo ao autor o título de ***Mestre em Engenharia de Produção***

Banca Examinadora:

Prof. Dr. Pedro Paulo Balestrassi (Orientador)

Prof. Dra. Mabel Scianni Morais - GESis

Prof. Dra. Rita de Cássia Magalhães Trindade Stano - UNIFEI

Maio de 2009

Itajubá- MG

DEDICATÓRIA

Dedico este trabalho a minha amada mãe, que sempre esteve comigo e que pelo seu exemplo de luta pela vida, me fez buscar forças para superar todos os obstáculos e atingir este objetivo.

AGRADECIMENTOS

Ao Prof. Pedro Paulo Balestrassi, pelas orientações neste trabalho.

Aos meus colegas e amigos do Grupo GESis, pelo companheirismo e amizade cultivada nestes anos, em especial a Maria Carolina, Rafael, Adriano, Natália, Sandra e Francisco.

Ao Prof. Antônio Carlos Zambroni de Souza e sua esposa Marli Zambroni, pelo incentivo e amizade.

A minha amiga Cristina Ross, pelas conversas e conselhos.

Ao meu pai, pela forma como me criou, me dando tudo o que eu precisava e me ensinando, mesmo distante, a buscar o que eu desejava.

De uma forma muito especial, agradeço ao meu amigo José Wanderley, pela amizade que só acrescentou coisas boas em minha vida. Obrigada pelas exigências, pelas palavras muitas vezes duras, porém sinceras, hoje eu compreendo que tudo foi feito para o meu bem. Obrigada pelos seus ensinamentos que foram fundamentais para meu crescimento pessoal e profissional. Gostaria de agradecer de coração a preocupação e o carinho que tem comigo.

Ao meu Anjo da Guarda, por realmente ter acreditado em mim desde o início e nunca ter me deixado sozinha. Agradeço pela amizade e companheirismo, por ter me ajudado a enfrentar os muitos obstáculos que eu encontrei e mais do que nunca obrigada por ter me feito voltar a acreditar que eu conseguiria chegar ao fim deste trabalho. A sua presença é fundamental em minha vida.

E por fim, agradeço à Deus por ter se manifestado em minha vida através de todas as pessoas acima citadas. Obrigada por nunca ter largado minha mão e ter me dado forças para continuar esta caminhada.

RESUMO

A busca na melhoria da educação tem feito com que as instituições de ensino superior busquem admitir estudantes de alta qualidade. Com o processo de admissão as instituições passam a conhecer o perfil de seus novos estudantes e assim passam a obter uma idéia de quais estudantes atingirão seus objetivos educacionais, ou seja, de obterem um bom desempenho durante o curso e conseqüentemente se graduarem. Usando as ferramentas e dados disponíveis, instituições podem criar modelos preditivos designados a prever o desempenho de seus alunos. O presente estudo busca, com as notas do ENEM (Exame Nacional do Ensino Médio) e das disciplinas dos vestibulares da Universidade Federal de Itajubá, UNIFEI, nos anos de 2000, 2001, 2002 e 2003, estabelecer modelos preditivos designados a prever a probabilidade de graduação do aluno, bem como o de predizer o seu coeficiente final, utilizando para isso técnicas de Regressão Logística Binária e Regressão Linear Múltipla. Os modelos encontrados indicam que, além das notas dos vestibulares, outras variáveis também são importantes para melhor previsão do desempenho dos alunos e da probabilidade de sucesso em se graduarem.

Palavras – chave: Regressão Logística, Regressão Linear Múltipla, Previsão e Desempenho.

ABSTRACT

In order to improve education, universities have looked for admit high quality students. The admission process permits universities to get new students´ profile, and which student will achieve its educational goals, i.e., getting a good performance during the course, and, consequently, graduate. Using available tools and data, institutions may create predictive models designed to forecast their students' performance. This study attempts, with the results of ENEM[1], vestibular and courses offered by University of Itajubá (UNIFEI) during 2000, 2001, 2002 and 2003, to establish predictive models designed to forecast both probability of the student graduation and their final coefficient, using Binary Logistic Regression and Multiple Linear Regression techniques. Models obtained suggest that, besides the admission test results, other variables are also important to get better results in forecasting both the students' performance and their probability of success in graduating.

Keywords - Logistic Regression, Multiple Linear Regression, Predict And Performance

[1] ENEM (High School National Exam in Portugese) is a Brazilian test for all high school students. It is a kind of SAT (Scholastic Assessment Test) in USA

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO 1 – INTRODUÇÃO

1.1 Tema e Problema da Pesquisa

O esforço para o aumento da eficácia na educação tem exercido uma pressão sobre as instituições de ensino superior. Esta pressão tem forçado as instituições a examinar com mais cautela como anda o processo educacional em seus programas, bem como as políticas de admissão. O papel de um programa de educação é estabelecer um processo que proporcione ao estudante habilidades e competências para chegar aos seus objetivos**.**

É através do processo de admissão que as instituições passam a ter uma idéia do perfil de seus novos estudantes e podem assim obter uma probabilidade de quais estudantes atingirão seus objetivos educacionais, ou seja, obterem um bom desempenho durante o curso e conseqüentemente se graduarem.

Avanços tecnológicos têm permitido aumento de acesso a dados, isto faz com que se possam obter novas técnicas de análise de dados. Metodologias antes difíceis e caras, hoje são acessíveis, com isso as organizações conseguem desenvolver sistemas inteligentes de suporte de decisão e de previsão.

Usando as ferramentas e dados disponíveis, instituições podem criar modelos preditivos designados a prever o desempenho de seus alunos. Decisões de admissão representam dedução sobre as chances de o estudante ter sucesso em uma determinada instituição.

1.2 Objetivos

Algumas perguntas foram motivadoras e norteiam os objetivos desse trabalho:

- Será que as notas de entrada de um aluno através do vestibular determinam o seu êxito acadêmico?
- Será que os melhores alunos em termos de notas de ingresso possuem os melhores resultados finais ao se graduarem?

O presente estudo tem, portanto, como objetivo principal desenvolver *uma análise estatística descritiva sobre o desempenho final dos graduandos da UNIFEI em função de seus dados de aprovação no vestibular*. Os seguintes objetivos secundários podem ser ainda listados:

- Procurar um possível modelo de regressão entre os dados de entrada do vestibular e os dados de saída dos graduandos;
- Desenvolver uma revisão sobre os diversos fatores determinantes no desempenho dos alunos graduandos;

1.3 Justificativa

O número de vagas nas instituições públicas é limitado, havendo assim uma grande concorrência. Para solucionar este problema, as instituições utilizam mecanismo de seleção como é o caso dos vestibulares. Este mecanismo procura selecionar os alunos mais bem capacitados e qualificados para ingressarem na universidade.

Dados sobre o desempenho do aluno nos vestibulares e no ENEM são adquiridos no processo da admissão. Estes dados serão utilizados como variáveis independentes e como variáveis dependentes serão utilizadas a situação final do aluno, assim como seu coeficiente final. Com o uso de modelos quantitativos e de técnicas de regressão linear e logística será possível prever o desempenho acadêmico dos estudantes.

Resultados deste estudo podem, além de melhorar a habilidade das universidades na previsão do desempenho dos alunos e na previsão da probabilidade de sucesso em se graduarem, possibilitar um questionamento sobre o vestibular como ferramenta de seleção.

Este estudo foi conduzido dentro da Universidade Federal de Itajubá, UNIFEI. Todas as variáveis usadas neste estudo foram disponibilizadas pela universidade. Este estudo utiliza dados de novos alunos matriculados nos anos de 2000, 2001, 2002 e 2003.

Apesar de muitos esforços para admitir estudantes que poderão ter sucesso em se graduarem, ainda é preocupante o número de alunos que não chegam até o fim de seu curso. A evasão varia pelo tipo de instituição. A tendência é que em instituições privadas o número de evasão seja maior do que as universidades públicas, onde o processo de admissão é mais rigoroso.

A constatação, através deste estudo, de um número elevado de evasão ou retenção, possibilitará também, uma discussão acerca dos currículos dos cursos, com o objetivo de melhorar os índices de graduação.

1.4 Métodos de Pesquisa

Nesta seção será descrita a classificação do presente trabalho. Portanto quanto a sua natureza, a pesquisa apresenta um enfoque prático, pois busca soluções para problemas reais do cotidiano de uma universidade.

Gil (1993) propõe que as pesquisas ou trabalhos científicos podem ser divididos e classificados de acordo com seus objetivos em três tipos básicos de estudo voltados a buscar uma resposta ou solução para um determinado problema. Estas pesquisas são baseadas em métodos descritivos, exploratórios ou explicativos. Pelo aspecto de análise exposto como objetivo principal do trabalho, o método explicativo é o mais adequado ao desenvolvimento desta pesquisa. O método explicativo identifica os fatores que determinam ou contribuem para a ocorrência dos fenômenos (XAVIER, 2008).

Quanto à abordagem de pesquisa o estudo será quantitativo e requer o uso de recursos e de técnicas estatísticas como, percentagem, média, mediana, desvio-padrão, coeficiente de correlação, análise de regressão, etc.

Do ponto de vista dos procedimentos técnicos a pesquisa utilizará o estudo de caso, com foco na descrição do método e análise da aplicação de técnicas regressão linear e logística na Universidade Federal de Itajubá - UNIFEI. O estudo de caso é recomendado quando se busca o estudo profundo e exaustivo de um ou poucos objetos de maneira que se permita o seu amplo e detalhado conhecimento (MENEZES e SILVA, 2005).

1.5 Estrutura do Trabalho

O trabalho proposto está estruturado em cinco capítulos, cujos conteúdos resumidos encontram-se a seguir:

Capítulo 1 – Refere-se à introdução do trabalho, fazendo uma apresentação, a justificativa de sua realização, o objetivo do trabalho, bem como sua estrutura;

Capítulo 2 – Trata da revisão bibliográfica, apresentando as formas de admissão das universidades federais brasileiras, os modelos teóricos de evasão do estudante e uma breve apresentação da técnica estatística utilizada no trabalho;

Capítulo 3 – Expõe de forma detalhada os métodos de Regressão Linear e Logística que serão utilizados para análise dos dados deste trabalho;

Capítulo 4 – Apresenta toda a análise dos dados colhidos, informando de forma sucinta os resultados gerados pelas regressões;

Capítulo 5 – Traz as conclusões e considerações finais sobre o trabalho realizado, além de recomendações para trabalhos futuros.

CAPÍTULO 2 - REVISÃO BIBLIOGRÁFICA

2.1 Introdução

O gerenciamento das matrículas nas instituições de ensino superior envolve mais do que meramente admitir um adequado número de novos estudantes para compor o primeiro ano de um curso. Tem se tornado cada vez mais importante que estudantes admitidos para uma universidade obtenham seus objetivos educacionais e contribuam para os objetivos da universidade. A experiência sobre o ensino superior tem encorajado pesquisas relacionadas com a persistência e a evasão de estudantes.

Primi et. (2002) exploraram correlações existentes entre medidas de *inteligência fluida*, capacidade geral de relacionar idéias complexas, e *inteligência cristalizada*, capacidade de derivar conhecimento a partir de esquemas organizados de informações sobre disciplinas específicas, com o desempenho acadêmico buscando investigar a importância relativa destas medidas na sua previsão. As correlações encontradas indicam que o desempenho acadêmico está associado a diferentes perfis de habilidades cognitivas.

A revisão de três tópicos neste capítulo é necessária para este estudo. O primeiro tópico se refere em como as universidades admitem os estudantes. Este tópico é importante para mostrar qual o tipo de informação é tipicamente usado no processo de admissão. O segundo tópico da revisão é a pesquisa referente às variáveis independentes que podem ser importantes na previsão do sucesso do estudante no ensino superior. O terceiro tópico examinará o uso dos modelos quantitativos para previsão do desempenho acadêmico.

2.2 Admissão das Universidades Federais Brasileiras

O número existente de vagas no ensino superior da rede pública é muito menor do que a quantidade de cidadãos que deseja obter uma formação universitária. A solução que os gestores e educadores universitários encontraram para equacionar o problema foi a utilização de mecanismos de seleção como é o caso dos vestibulares (PRAXEDES, 2003).

A principal forma de acesso as universidades do Brasil é o vestibular. É uma prova que testa os conhecimentos do estudante adquiridos no ensino fundamental e médio. No Brasil, os vestibulares das universidades públicas são os mais concorridos, pelo fato dos estudos serem gratuitos e por se tratar de instituições com alto nível de ensino. O número de vagas nestas instituições é limitado e a concorrência é enorme. O governo brasileiro tomou

medidas a incentivar que as notas do Exame Nacional do Ensino Médio (ENEM) fossem levadas em consideração para a colocação do candidato nas provas dos vestibulares.

Concorreram para as vagas totais do ensino superior, 5.191.760 candidatos no ano de 2007. A maior razão de candidatos por vagas foi observada nas IES federais, com média de 8,3 candidatos por vaga em todo Brasil, conforme a *Tabela 2.1*(INEP, 2009a).

Ano	Total	%Δ	Pública						Privada	%Δ
			Federal	%Δ	Estadual	%Δ	Municipal	%Δ		
2002	2,8	_	9,9	_	9,9	_	2,0	_	1,6	_
2003	2,4	-14,9	10,5	5,0	9,1	-9,7	1,7	-14,8	1,5	-8,4
2004	2,2	-12,3	10,4	-0,6	8,0	-12,8	1,6	-8,6	1,3	-12,9
2005	2,1	-4,8	10,0	-4,1	7,4	-8,8	1,5	-10,3	1,3	-0,5
2006	2,0	-5,4	8,9	-12,5	7,8	5,7	1,4	-6,5	1,2	-5,3
2007	1,8	-7,2	8,3	-6,5	8,1	3,2	1,3	-5,0	1,2	-5,9

Tabela 2.1 – Evolução da relação Candidato/Vaga – 2002/2007
Fonte: INEP - 2009a

O ENEM é um exame individual, voluntário, oferecido aos estudantes que estão concluindo ou que já concluíram o ensino médio. O principal objetivo do ENEM é avaliar o desempenho do aluno ao término da escolaridade básica. O exame foi pensado também como modalidade alternativa ou complementar aos exames de acesso ao ensino superior. Este objetivo vem sendo atingido um pouco mais a cada ano, graças ao esforço do Ministério da Educação na sensibilização e convencimento das instituições de ensino superior (IES) para o uso dos resultados do ENEM como componente dos seus processos seletivos (INEP, 2009b).

A estrutura das provas do vestibular varia de acordo com a instituição responsável pela prova. Abaixo uma breve descrição do processo seletivo 2009 de algumas universidades federais representativas de cada região:

- UNIR - Fundação Universidade Federal de Rondônia **(Região Norte)**

  O Processo Seletivo 2009 da Universidade Federal de Rondônia foi realizado em duas fases, ambas de caráter eliminatório e classificatório:

  Primeira Fase - Constituída de uma Prova Objetiva com 80 (oitenta) questões de múltipla escolha.

  Segunda Fase - Prova Discursiva composta por duas Partes: Parte 1- 1 (uma) redação dissertativa e Parte 2 - 2 (duas) questões discursivas sobre história e geografia regionais (EDITAL UNIR VESTIBULAR 2009).

- UFMA - Fundação Universidade Federal do Maranhão **(Região Nordeste)**

  O Processo Seletivo 2009 da Universidade Federal do Maranhão foi realizado em duas etapas de caráter eliminatório:

  Primeira Etapa - Prova de Conhecimentos Básicos, com questões objetivas.
  Segunda Etapa - Prova de Conhecimentos Específicos, com questões analítico-discursivas e uma Prova de Redação (EDITAL UFMA VESTIBULAR 2009).

- UFMT - Fundação Universidade Federal de Mato Grosso (**Região Centro-Oeste)**

  O Processo Seletivo para ingresso nos cursos de graduação da Universidade Federal de Mato Grosso foi realizado em duas modalidades:
  a) Concurso Vestibular Unificado, com peso de 100%;
  b) Concurso Vestibular Unificado, com peso de 80%, mais resultado do Exame Nacional do Ensino Médio – ENEM, com peso de 20%, de acordo com o que estabelece o Regulamento do Processo Seletivo para Ingresso nos Cursos de Graduação da Universidade Federal de Mato Grosso, aprovado pela Resolução CONSEPE N.o 41, de 19 de maio de 2003.
  O Concurso Vestibular Unificado de 2009 da Universidade Federal de Mato Grosso foi desenvolvido em duas fases, ambas de caráter eliminatório e classificatório.
  Primeira Fase - Prova com 80 (oitenta) questões objetivas de múltipla escolha (Prova Objetiva).
  Segunda Fase - Prova Discursiva/Redação composta de duas partes. A Parte I constituída de 05 (cinco) questões discursivas relativas à leitura de textos de diferentes gêneros veiculados socialmente; a Parte II constituída de uma Redação em Língua Portuguesa (EDITAL UFMT VESTIBULAR 2009).

- UFSCar - Fundação Universidade Federal de São Carlos (**Região Sudeste)**

  O Processo Seletivo 2009 da Universidade Federal de São Carlos foi organizada em uma única fase, realizado em três dias consecutivos, com a aplicação de nove provas (Língua Inglesa, Língua Portuguesa, Redação, Química, Matemática, História, Biologia, Física e Geografia). Foram reservadas 20% (vinte por cento) das vagas de cada curso para candidatos egressos do ensino público. São considerados candidatos egressos do ensino público aqueles que tenham cursado o ensino médio, integralmente, na rede pública de ensino no Brasil (municipal, estadual, federal). Das vagas reservadas, 35% (trinta e cinco por cento) serão destinadas a candidatos negros

(pretos e pardos) que venham a ser aprovados no processo seletivo (EDITAL UFSCAR VESTIBULAR 2009).

- UFSC - Universidade Federal de Santa Catarina (**Região Sul)**

  O Processo Seletivo 2009 da Universidade Federal de Santa Catarina foi realizado em uma única etapa. As provas (Língua Portuguesa e Literatura Brasileira; Língua Estrangeira; Redação, Biologia, Geografia E Matemática, Física, História e Química) foram aplicadas em 3 (três dias).

  I - 20% (vinte por cento) das vagas de cada curso foram destinadas para candidatos que tenham cursado integralmente o ensino fundamental e médio em instituições públicas de ensino;

  II - 10% (dez por cento) das vagas de cada curso foram destinadas para candidatos autodeclarados negros, que tenham cursado integralmente o ensino fundamental e médio em instituições públicas de ensino;

  III - 6 (seis) vagas foram destinadas a candidatos autodeclarados indígenas (EDITAL UFSC VESTIBULAR 2009).

A *Tabela 2.2* apresenta todas as universidades federais do Brasil.

UNIVERSIDADES FEDERAIS	
Região Norte	
• **Instituição**	**Sigla**
Fundação Universidade Federal do Acre	UFAC
Fundação Universidade Federal do Amapá	UNIFAP
Universidade Federal do Amazonas	UFAM
Universidade Federal do Pará	UFPA
Universidade Federal Rural da Amazônia	UFRA
Fundação Universidade Federal de Rondônia	UNIR
Fundação Universidade Federal de Roraima	UFRR
Universidade Federal de Tocantins	UFT
Região Nordeste	
• **Instituição**	**Sigla**
Fundação Universidade Federal do Vale do São Francisco	UNIVASF
Universidade Federal do Recôncavo da Bahia	UFRB
Universidade Federal de Alagoas	UFAL
Universidade Federal da Bahia	UFBA
Universidade Federal do Ceará	UFC
Fundação Universidade Federal do Maranhão	UFMA
Universidade Federal da Paraíba	UFPB

Tabela 2.2 – Universidades Federais
Fonte: MEC (2009)

Universidade Federal de Pernambuco	UFPE
Universidade Federal Rural de Pernambuco	UFRPE
Fundação Universidade Federal do Piauí	UFPI
Universidade Federal do Rio Grande do Norte	UFRN
Fundação Universidade Federal de Sergipe	UFS
Universidade Federal de Campina Grande	UFCG
Universidade Federal Rural do Semi-Árido	UFERSA
Região Centro-Oeste	
• **Instituição**	**Sigla**
Universidade Federal da Grande Dourados	UFGD
Fundação Universidade de Brasília	UnB
Universidade Federal de Goiás	UFG
Fundação Universidade Federal de Mato Grosso	UFMT
Fundação Universidade Federal de Mato Grosso do Sul	UFMS
Região Sudeste	
• **Instituição**	**Sigla**
Fundação Universidade Federal do ABC	UFABC
Universidade Federal do Espírito Santo	UFES
Universidade Federal Fluminense	UFF
Universidade Federal de Juiz de Fora	UFJF
Universidade Federal de Lavras	UFLA
Universidade Federal de Minas Gerais	UFMG
Fundação Universidade Federal de Ouro Preto	UFOP
Universidade Federal Rural do Rio de Janeiro	UFRRJ
Fundação Universidade Federal de São Carlos	UFSCar
Universidade Federal de São Paulo	UNIFESP
Fundação Universidade Federal de Uberlândia	UFU
Fundação Universidade Federal de Viçosa	UFV
Universidade Federal do Estado do Rio de Janeiro	UNIRIO
Universidade Federal do Rio de Janeiro	UFRJ
Universidade Federal de Itajubá	UNIFEI
Fundação Universidade Federal de São João del Rei	UFSJ
Universidade Federal de Alfenas	UNIFAL
Universidade Federal do Triângulo Mineiro	UFTM
Universidade Federal dos Vales do Jequitinhonha e Mucuri	UFVJM
Região Sul	
• **Instituição**	**Sigla**
Fundação Universidade Federal do Rio Grande	FURG
Universidade Federal do Rio Grande do Sul	UFRGS
Universidade Federal de Santa Catarina	UFSC
Universidade Federal do Paraná	UFPR
Fundação Universidade Federal de Pelotas	UFPEL
Universidade Federal de Santa Maria	UFSM
Universidade Tecnológica Federal do Paraná	UTFPR

Tabela 2.2 – Continuação

Pesquisas realizadas sobre os mecanismos de avaliação e de seleção como vestibular e provas, levaram à constatação da existência de um perfil geral de alunos que ingressam nas escolas consideradas de melhor nível, públicas ou privadas, que não evadem e que alcançam um melhor desempenho. Tal perfil é decorrente, entre outros fatores, das seguintes variáveis:

- Escolaridade dos pais;
- Renda familiar;
- Local cidade e região de domicílio;
- Freqüência à escola privada nos ensinos fundamental e médio;
- Abundância de material de escrita no cotidiano;
- Acesso e participação em atividades culturais;
- Desenho, leitura e música incorporados ao lazer infantil;
- Seleção de programas educativos na televisão;
- Hábito de argumentação racional no cotidiano;
- Utilização da linguagem culta no cotidiano;
- Disciplina de estudo e concentração;
- Hábito de estudo autônomo;
- Expectativa de bom desempenho na escola;
- Valorização do sucesso escolar;
- Exemplos de pessoas próximas bem sucedidas no sistema escolar.

Observa-se assim, que a formação educacional, o acesso a bens culturais, e as condições sócio-econômicas estão entre os fatores que influenciam no ensino superior (PRAXEDES, 2003).

O vestibular nas últimas décadas sofreu algumas modificações como pode ser visto abaixo:

- **Década de 60** - O crescimento demográfico e a urbanização contribuíram para o aumento do ensino superior. Diversas fórmulas e iniciativas na tentativa de racionalizar e aprimorar o processo seletivo para o ensino superior foram adotadas pelo poder público. O Ministério da Educação e Cultura – MEC – passou a atuar junto às instituições de ensino superior, com vistas à realização de concursos vestibulares unificados em âmbito regional. Foi instituído o vestibular classificatório, pelo qual o candidato seria admitido até o número total de vagas, especificado no edital, independentemente da nota mínima. Com esse diploma legal, desaparecia a figura do "excedente", candidato aprovado que não conseguia matricular-se por falta de vaga.

- **Década de 70 e 80** - Outras práticas foram experimentadas, algumas permanecem até os dias de hoje com pequenas alterações. Destacam-se a introdução de provas de habilidade específica para os cursos de Educação Física, Música, Artes, Arquitetura, entre outros, a inclusão de redação, o vestibular por etapas, a fixação de pesos diferentes para cada prova, considerando-se a carreira pretendida e a inclusão de questões que envolvem conhecimentos regionais. Na década de 1980, devido à crise da economia brasileira ocorreu uma retração de demanda para o ensino superior.
- **Década de 90** - Com a globalização a busca por uma vaga no ensino superior voltou a intensificar-se. A oferta de cursos também se ampliou consideravelmente, o que não implicou a distribuição uniforme de candidatos por cursos. A década de 1990 inicia-se com uma mudança significativa no que tange à seleção do ensino superior. Delegou-se aos estabelecimentos de ensino superior a competência para a realização do concurso vestibular, nos termos da lei e de seus estatutos e regimentos.
- **Novo Milênio** - A expansão da educação superior impulsionou a adoção de formas alternativas de selecionar candidatos, na tentativa de substituir o vestibular tradicional. Embora ainda persista a antiga forma de selecionar os futuros universitários, outras formas têm sido experimentadas e avaliadas, apresentando resultados satisfatórios, quer na instituição pública, quer na privada (BORGES e CARNIELLI, 2005).

2.3 Modelos Teóricos de Evasão de Estudantes

Modelos teóricos fornecem uma estrutura conceitual para o entendimento do problema relacionado à evasão do estudante na universidade.

Bean (1980) sugere que muitos dos modelos desenvolvidos desde 1970 têm fundamento na pesquisa de Durkheim realizada no início da década de 60 sobre suicídio. Segundo Bean (1980), Durkheim examinou a importância dos valores compartilhados em grupos e o apoio dos amigos na redução do suicídio. Pascarella (1980), Spady (1970), Tinto (1975, 1993) e outros expandiram estas idéias em teorias sobre a evasão dos estudantes.

Estas teorias descrevem a evasão como um processo longitudinal no qual as características de origem do estudante interagem com outras variáveis para afetar sua integração acadêmica e social dentro da universidade. As dificuldades de se integrar em um ambiente acadêmico e social de uma universidade faz com que aumente a probabilidade de evasão do estudante (ASTIN, 1997; PASCARELLA, 1980; TINTO, 1993).

O modelo de Spady (1970) foi o primeiro modelo teórico "completamente desenvolvido" sobre a evasão de estudante (BEAN, 1980). Este modelo sugere que o apoio de amigos, o bom desempenho acadêmico e o desenvolvimento intelectual conduzem para a integração social. A integração social teoricamente aumenta a satisfação do estudante e diminui a evasão.

Os modelos de Tinto (1975, 1987, 1993) são similares ao modelo de Spady. O modelo de Tinto sugere que o comprometimento em se formar e o comprometimento com a universidade, influencia a integração acadêmica e social do estudante no ambiente institucional. Segundo a teoria de Tinto, o grau de aptidão do estudante em integrar-se com o ambiente do campus tem um efeito direto sobre sua propensão em persistir. Tinto sugere que uma integração positiva serve para aumentar os objetivos de alguém e fortalecer o seu comprometimento com seus objetivos e com a instituição. Estudantes que são incapazes de integrar-se dentro das comunidades sociais e acadêmicas da universidade têm mais probabilidade de se sentirem isolados, aumentando assim a probabilidade de deixar o programa antes de completá-lo (TINTO, 1993).

O modelo de Pascarella (1980) enfatizou o estudante e os fatores organizacionais que afetam o nível de contato informal entre estudantes e professores, e o efeito que esse contato tem sobre a persistência do estudante (BEAN, 1980). Pascarella e Terenzini (1977) sugerem que os contatos fora da sala de aula entre o estudante e professor são mais prováveis de influenciar positivamente a integração institucional do estudante e sua decisão em persistir.

A intenção desta revisão não é apresentar uma discussão mais aprofundada sobre estes modelos, mas sim investigar o problema da evasão do estudante e identificar fatores que estão associados com a persistência. Para este fim, é importante entender que a maioria das pesquisas sobre a evasão do estudante está baseada nestes e em outros modelos teóricos.

Lenning (1982) agrupa alguns fatores que afetam a evasão dos estudantes. Entre estes fatores estão:

- Sexo;
- Renda Familiar;
- Nível de Educação dos Pais;
- Ajuda Financeira;
- Outros Fatores Demográficos (estado civil, idade, emprego);
- Aspiração Inicial do Estudante e Variáveis de Motivação;
- Variáveis de Interação;

- Variáveis Institucionais.

Fatores demográficos tais como sexo e renda familiar podem afetar a probabilidade de persistência de um estudante. Estes fatores podem afetar a motivação, suas chances de manter um trabalho enquanto estiver na escola, e sua integração social e acadêmica dentro da comunidade da universidade (LENNING, 1982; PASCARELLA, 1980; TINTO, 1993).

2.3.1 Sexo

Diferenças de sexo na obtenção de educação têm diminuído nas últimas décadas. Tinto (1993) mostra que o índice de mulheres que freqüentam o ensino superior nos Estados Unidos é maior do que os homens. Mulheres também têm maior probabilidade de obter um diploma em relação aos homens (ASTIN, 1997; TINTO, 1993).

Pesquisas mostram que as mulheres dominam as estatísticas de matrículas nas universidades do Brasil. De modo geral, elas são maioria nas instituições públicas e privadas. Os dados mais recentes do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) mostram que, dos 4.163.733 estudantes matriculados no ensino superior, 2.346.516 são do sexo feminino (56,35%). Em lugar de cuidar exclusivamente da família e da casa, tarefas que durante muitos anos as mães assumiram sozinhas, as brasileiras estão se tornando cada vez mais qualificadas e prontas para encarar o mercado de trabalho em pé de igualdade com os homens (UNIVERSIA, 2006). Um outro aspecto é que a mulher tem buscado sua independência e realização profissional.

A pesquisa de Maccoby e Jacklin (1974) concluiu que mulheres têm maiores habilidades verbais do que homens, mas homens têm maiores habilidades matemáticas.

As mulheres são maioria nos cursos de Pedagogia, Enfermagem, Psicologia, Letras e Biologia e estão virando maioria em Matemática, Jornalismo e Odontologia. O número de alunos e alunas é praticamente o mesmo em Medicina, Direito, Engenharia Química, Veterinária, Química. Uma mudança que merece atenção é que as mulheres nas universidades estão se destacando em cursos antes dominados pelos homens como cursos de Engenharia Mecânica, Civil, Elétrica, Administração entre outros.

Apesar deste crescimento, alguns estudiosos afirmam que diferenças entre sexos ainda existem. Apesar das diferenças terem diminuído muito, homens e mulheres ainda se diferem na escolha de carreiras. A probabilidade da mulher se formar em cursos como Ciência da Computação, Engenharias ou Ciências Físicas é ainda menor do que para os homens.

Embora o sexo não seja o maior fator em prever a evasão**,** homens e mulheres freqüentemente deixam o ensino superior por diferentes razões (PASCARELLA *et al*. 1986). Homens têm maior probabilidade de abandonar por razões acadêmicas, enquanto as mulheres por razões não acadêmicas (LENNING, 1982).

Lenning (1982) sugere que diferenças de sexo na evasão é primeiramente um reflexo de diferenças na motivação, status socio-econômico, estado civil e assim por diante.

2.3.2 Renda Familiar

Estudantes de uma camada socioeconômica mais baixa são menos prováveis de persistir quando comparados com estudantes com classe social mais elevada (LENNING, 1982). Trabalhar fora da universidade, especialmente o dia todo, está tipicamente associado com a baixa proporção de persistência (TINTO, 1993).

Para a maioria dos estudantes brasileiros de baixa renda, manter-se na universidade é quase tão difícil quanto ser aprovado no vestibular. Em cada região do país, as instituições públicas de ensino de nível superior, federais ou estaduais, oferecem benefícios diferenciados para os alunos, umas mais, outras menos. As bolsas-trabalho, moradia, alimentação e transporte estão entre as necessidades mais comuns, especialmente para aqueles que são de outros estados ou cidades. Nesse processo de disputa, faltam vagas e milhares de estudantes não recebem auxílio (COM CIÊNCIA, 2003).

A renda familiar é altamente correlacionada com o nível educacional dos pais, que pode influenciar as aspirações educacionais do estudante (TINTO, 1993)**.** Isto é uma combinação de renda familiar e educação dos pais, que influencia o número de livros em casa, as oportunidades de viagem, melhores colégios, tipos de conversas ao redor da mesa de jantar, e em geral, o envolvimento dos pais na educação de seus filhos (BOWEN e BOK, 1998).

2.3.3 Nível de Educação dos Pais

O nível educacional dos pais dos estudantes é positivamente correlacionado com a persistência do graduando e o sucesso do estudante (LENNING, 1982; TINTO, 1993).
Estudantes com pais que possuam alto nível de educação são mais propensos a continuar seus estudos e menos propensos de se casar enquanto ainda estão na universidade. Estudantes são mais propensos a atingir suas metas de entrar no mercado de trabalho, quando seus pais já estão empregados no mesmo campo em que desejam trabalhar.

O modelo de Tinto (1993) de desistência institucional sugere que o nível educacional dos pais pode afetar os objetivos individuais e comprometimento com a instituição.

Estudantes com pais que fizeram faculdade, que estão bem financeiramente e apóiam as decisões de seus filhos em freqüentar a faculdade estão mais propensos a persistir.

Bowen e Bok (1998) sugerem que a obtenção educacional dos pais é altamente correlacionada com a renda, saúde, e classe social da família, e que estes fatores, na média, melhoram a preparação acadêmica e realização de seus filhos.

2.3.4 Ajuda Financeira

Ajuda financeira fornece alívio econômico para estudantes de baixa renda e aumenta a proporção de persistência (STAMPEN e CABRERA, 1986, 1988; TINTO, 1993). Pesquisas sugerem que o tipo de suporte financeiro que o estudante recebe durante a universidade influencia sua persistência. Estudantes que recebem bolsa de estudos ou bolsa de estágio são mais propensos a persistir do que estudantes que recebem outros tipos de ajuda tal como empréstimos. A bolsa de estágio pode ajudar a integrar estudantes dentro da instituição através de contatos com os professores (TINTO, 1993).

O Ministério da Educação (MEC) deixa sob a responsabilidade de cada instituição pública de ensino superior gerar programas que possibilitem aos universitários pobres continuar os seus estudos. Espalhadas pelas cinco regiões brasileiras - Norte, Nordeste, Centro-Oeste, Sul e Sudeste, as universidades públicas apresentam várias discrepâncias no atendimento aos estudantes dependentes de ajuda financeira (COM CIÊNCIA, 2003).

O Plano Nacional de Assistência Estudantil (Pnaes) apóia a permanência de estudantes de baixa renda matriculados em cursos de graduação presencial das instituições federais de ensino superior. O objetivo é viabilizar a igualdade de oportunidades entre todos os estudantes e contribuir para a melhoria do desempenho acadêmico, a partir de medidas que buscam combater situações de repetência e evasão. O Pnaes oferece assistência à moradia estudantil, alimentação, ao transporte, à saúde, inclusão digital, cultura, ao esporte, creche e apoio pedagógico. As ações são executadas pela própria instituição de ensino, que deve acompanhar e avaliar o desenvolvimento do programa. Os critérios de seleção dos estudantes levam em conta o perfil socioeconômico dos alunos, além de critérios estabelecidos de acordo com a realidade de cada instituição. (MEC, 2009).

2.3.5 Outros Fatores Demográficos (estado civil, idade, emprego)

Outros fatores demográficos podem ter um efeito sobre o desempenho e persistência dos estudantes. Estado civil, idade, emprego enquanto freqüenta a escola, e características de sua cidade natal podem afetar o comprometimento com o objetivo dos estudantes, o

comprometimento com a instituição, e a integração na comunidade do campus (LENNING, 1982; TINTO, 1993).

Estudantes mais velhos tendem a sofrer uma maior influência de fatores externos que podem afetar sua persistência. Lenning (1982) sugere que embora estudantes mais velhos possam estar com suas habilidades acadêmicas fora de prática, sua maturidade e motivação podem compensar suas deficiências acadêmicas, e que estudantes maduros são mais propensos a perceber uma conexão entre sua persistência e seus objetivos.

Os estudantes mais velhos têm maior probabilidade de serem casados, terem crianças e trabalharem em tempo integral (LENNING, 1982; TINTO, 1993). Estes fatores afetam a quantidade de tempo que estes estudantes podem passar dentro da universidade e pode inibir sua integração dentro da comunidade do campus. Responsabilidades familiares associadas com casamento e filhos diminuem a persistência entre as mulheres, enquanto aumenta a persistência entre homens (TINTO, 1993).

As famílias podem fornecer apoio para estudantes do sexo masculino, enquanto que os do sexo feminino são freqüentemente restritas por suas responsabilidades no lar (TINTO, 1993). Estudantes de zona rural e de cidades pequenas podem apresentar índices de evasão mais elevados em algumas instituições (LENNING, 1982)**.** Estudantes de origem rural podem ter dificuldade de integração dentro de grandes e urbanas universidades.

2.3.6 Aspiração Inicial do Estudante e Variáveis de Motivação

Estudantes variam em suas habilidades acadêmicas. Suas habilidades em atingir um elevado nível educacional é em parte uma função de sua aptidão intelectual e suas conquistas acadêmicas anteriores. Contudo, muitos fatores não intelectuais afetam a preparação acadêmica dos estudantes para a universidade, sua satisfação com a experiência na universidade, e sua verdadeira realização.

Lenning (1982) e Tinto (1993) sugerem que o comprometimento dos estudantes em obter um diploma, seu comprometimento com a universidade que freqüentam e a influência da família, amigos, e patrões contribuem significativamente na sua motivação em persistir no programa acadêmico.

Brown (1994) constatou que a combinação de fatores de conhecimento, interesse e personalidade contribuem para a persistência do aluno.

2.3.7 Variáveis de Interação

Pascarella e Terenzini (1977) e Pascarella (1980) exploraram a importância das interações entre alunos e membros da universidade nas desistências. Estes estudos sugerem que interações informais entre alunos e membros da universidade contribuem significativamente com a permanência dos alunos na universidade. O contato do aluno com os professores ligados mais diretamente ao seu curso parece ter um efeito mais positivo. Estes autores acham que as instituições podem afetar a freqüência das interações entre alunos e professores, e que as interações ocorrendo no início do curso do estudante podem afetar a disposição do estudante em procurar contato com o professor fora da sala de aula.

Rotter (1998) afirma que estudantes que não persistiram eram significativamente menos satisfeitos com suas oportunidades de interagir informalmente com professor. Pesquisas parecem consistentemente achar que a interação informal aluno-universidade é importante na prevenção da desistência do aluno. Como verificado anteriormente, fatores demográficos dos estudantes podem afetar estas interações. Casamento, idade, trabalho, podem reduzir a oportunidade de interação dos estudantes com os professores fora da sala de aula.

2.3.8 Variáveis Institucionais

Características institucionais podem ter um efeito nas taxas de permanência dos estudantes, tais como tamanho da instituição, tipo de instituição, sua seletividade e localização geográfica (TINTO, 1993; WHITE e MOSELY, 1994).

O tamanho de uma instituição não afeta todos os estudantes da mesma maneira. Grandes instituições podem ajudar alguns estudantes porque elas oferecem diversos tipos de ambientes. Por outro lado, grandes instituições podem intimidar alguns estudantes e aumentar a probabilidade que estes se isolem. A transição do ensino médio para a universidade é difícil para muitos alunos. Em grandes universidades, estudantes podem ficar perdidos, invisíveis para o sistema até serem identificados como estatística de evasão.

Pequenas instituições podem ter ambientes mais acolhedores. Se um estudante se ajusta dentro do ambiente, a probabilidade de persistência pode aumentar. Contudo, se um estudante não se ajusta, a probabilidade de persistência pode diminuir.

> *Das dez melhores universidades do Brasil, nove são instituições federais. É o que aponta o índice Geral de Cursos (IGC) divulgado pelo MEC. Num total de 1.837 instituições superiores cadastradas (universidades, centros universitários e faculdades), 78,8% tiveram IGC calculado pelo Inep. Esta é a primeira edição do IGC, indicador de qualidade das instituições de educação superior, onde estão*

> *sintetizadas, para cada instituição, a qualidade de todos os seus cursos de graduação, mestrado e doutorado, distribuídos nos campus e municípios onde a instituição atua. Com o novo indicador, torna-se possível fazer análises comparativas de desempenho por organização acadêmica, por UF e região geográfica e por categoria administrativa (federais, estaduais, municipais e privadas). Para o cálculo do IGC, o Inep utiliza a média dos conceitos preliminares dos cursos (CPC) da instituição, componente relativo à graduação, e o conceito fixado pela Capes, para a pós graduação. A média dos conceitos dos cursos é ponderada pela distribuição dos alunos entre os diferentes níveis de ensino (graduação, mestrado e doutorado). Nesta edição foram utilizados os CPCs do Exame Nacional de Desempenho dos Estudantes (Enade) no período de 2005 a 2007. A nota da Capes é a referente à avaliação do triênio 2004-2006. Em sua composição, o IGC também considera infra-estrutura, instalações, recursos didático-pedagógicos e corpo docente da universidade* (ANDIFES, 2008).

A *Tabela 2.3* apresenta o *ranking* das instituições federais:

IES	Sigla	UF	Dependência Administrativa
Fundação U. Federal do Vale do São Francisco	UNIVASF	PE	FEDERAL
Universidade Federal de São Paulo	UNIFESP	SP	FEDERAL
F. Univ. F. de Ciênc.da Saúde de Porto Alegre	UFCSPA	RS	FEDERAL
Fundação Universidade Federal de Viçosa	UFV	MG	FEDERAL
Universidade Federal de Minas Gerais	UFMG	MG	FEDERAL
Universidade Federal do Rio Grande do Sul	UFRGS	RS	FEDERAL
Universidade Federal do Triângulo Mineiro	UFTM	MG	FEDERAL
Universidade Federal do Rio de Janeiro	UFRJ	RJ	FEDERAL
Universidade Federal de São Carlos	UFSCAR	SP	FEDERAL
Universidade Federal de Itajubá - Unifei	**UNIFEI**	**MG**	**FEDERAL**
Universidade de Brasília	UnB	DF	FEDERAL
Universidade Federal de Santa Catarina	UFSC	SC	FEDERAL
Universidade Federal de Lavras	UFLA	MG	FEDERAL
Universidade Federal de Alfenas	UNIFAL	MG	FEDERAL
Universidade Federal de Santa Maria	UFSM	RS	FEDERAL
Universidade Federal de Ouro Preto	UFOP	MG	FEDERAL
Universidade Federal de Juiz de Fora	UFJF	MG	FEDERAL
Universidade Federal de Pernambuco	UFPE	PE	FEDERAL
Universidade Federal de Uberlândia	UFU	MG	FEDERAL
Universidade Federal de São João Del Rei	UFSJ	MG	FEDERAL
Universidade Federal do Rio Grande do Norte	UFRN	RN	FEDERAL
Universidade Federal do Estado do Rio de Janeiro	UNIRIO	RJ	FEDERAL
Univ. F. dos Vales do Jequitinhonha e Mucuri	UFVJM	MG	FEDERAL
Fundação Universidade Federal do Rio Grande	FURG	RS	FEDERAL
Universidade Federal de Goiás	UFG	GO	FEDERAL
Universidade Federal da Bahia	UFBA	BA	FEDERAL
Universidade Federal do Ceará	UFC	CE	FEDERAL
Fundação Univ. Federal da Grande Dourados	UFGD	MS	FEDERAL
Universidade Federal Rural do Rio de Janeiro	UFRRJ	RJ	FEDERAL
Universidade Federal do Paraná	UFPR	PR	FEDERAL
Universidade Federal de Mato Grosso do Sul	UFMS	MS	FEDERAL
Universidade Federal de Pelotas	UFPel	RS	FEDERAL
Universidade Federal de Campina Grande	UFCG	PB	FEDERAL
Centro Federal de Educação Tecnológica da Bahia	CEFET/BA	BA	FEDERAL
Universidade Federal Fluminense	UFF	RJ	FEDERAL
Universidade Federal da Paraíba	UFPB	PB	FEDERAL
Universidade Tecnológica Federal do Paraná	UTFPR	PR	FEDERAL
Universidade Federal de Mato Grosso	UFMT	MT	FEDERAL

Tabela 2.3 - Ranking das Instituições Federais
Fonte: ANDIFES - 2008

Universidade Federal de Sergipe	UFS	SE	FEDERAL
Universidade Federal do Espírito Santo	UFES	ES	FEDERAL
Universidade Federal do Piauí	UFPI	PI	FEDERAL
Fundação Universidade Federal de Rondônia	UNIR	RO	FEDERAL
Centro F. de Edu.Tecn. Celso Suckow da Fonseca	CEFET/RJ	RJ	FEDERAL
Universidade Federal do Amazonas	UFAM	AM	FEDERAL
Universidade Federal Rural de Pernambuco	UFRPE	PE	FEDERAL
Universidade Federal do Maranhão	UFMA	MA	FEDERAL
Universidade Federal Rural da Amazônia	UFRA	PA	FEDERAL
Universidade Federal Rural do Semi-Árido	UFERSA	RN	FEDERAL
Universidade Federal do Acre	UFAC	AC	FEDERAL
Universidade Federal de Roraima	UFRR	RR	FEDERAL
Universidade Federal do Pará	UFPA	PA	FEDERAL
Universidade Federal de Alagoas	UFAL	AL	FEDERAL
Fundação Universidade Federal do Tocantins	UFT	TO	FEDERAL
Centro F. de Edu. Tecnológica de Minas Gerais	CEFET	MG	FEDERAL
Centro Federal de Educação Tecn. do Maranhão	CEFET	MA	FEDERAL
Universidade Federal do Amapá	UNIFAP	AP	FEDERAL
Universidade Federal do Recôncavo da Bahia	UFRB	BA	FEDERAL

Tabela 2.3 - Continuação

A expansão do sistema público federal de educação superior deve estar associada a reestruturações acadêmicas e curriculares que proporcionem maior mobilidade estudantil, trajetórias de formação flexíveis, redução das taxas de evasão, utilização adequada dos recursos humanos e materiais colocados à disposição das universidades federais. O Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI, lançado pelo Governo Federal, busca criar condições para a ampliação do acesso e permanência na educação superior, no nível de graduação, para o aumento da qualidade dos cursos e pelo melhor aproveitamento da estrutura física e de recursos humanos existentes nas universidades federais, respeitadas as características particulares de cada instituição e estimulada a diversidade do sistema de ensino superior (REUNI, 2007).

A Universidade Federal de Itajubá, recentemente classificada como a décima melhor universidade do país conforme mostra a *Tabela 2.3*, será o objeto de estudo deste trabalho.

A UNIFEI especializou-se por muito tempo em Engenharia, por esse motivo é considerada uma referência na área, principalmente nas abordagens elétrica e mecânica. Nos últimos anos diversos outros cursos foram surgindo conforme as necessidades do mercado e o melhoramento de sua estrutura interna. Esse processo ocorreu sem desfocar as ciências exatas, apesar de fornecer outras disciplinas além das engenharias.

Abaixo uma breve descrição do processo seletivo 2009 da Universidade Federal de Itajubá (UNIFEI):

O Processo Seletivo foi realizado em uma única etapa, com três provas realizadas em dois dias.

A Prova 1 - Múltipla escolha. Conhecimentos em Biologia, Física, Geografia, História, Língua Inglesa, Língua Portuguesa, Matemática e Química.

A Prova 2 - Dividida em três redações curtas relacionadas com as obras literárias indicadas para leitura.
A Prova 3 - Discursiva. Para o curso de Administração – Conhecimentos em Língua Inglesa e Matemática. Para os demais cursos – Conhecimentos em Física e Matemática.

- A nota do ENEM foi utilizada para compor a Nota Final do candidato (EDITAL UNIFEI VESTIBULAR 2009).

2.4 Modelo Quantitativo para Previsão do Desempenho Acadêmico

Pesquisadores têm usado uma variedade de técnicas estatísticas para examinar a importância de variáveis acadêmicas em predizer o sucesso do estudante. Correlação múltipla e regressão, análise de variância, análise multivariadas de variância, análise discriminante, regressão logística e outros métodos são todos representados em pesquisa de desistência. A técnica apropriada para uso depende da variedade de fatores incluindo a natureza do estudo, a intenção da análise e o julgamento do pesquisador.

A regressão logística prediz a probabilidade da variável de saída assumir um dado valor. Uma vez que a regressão logística prediz probabilidades, sua resposta é limitada pelos valores 0 e 1. Modelos de regressão linear, por outro lado, são inerentemente ilimitados, ou seja, sua resposta vai de menos infinito à mais infinito. O uso de modelo de regressão linear em problemas de predição em que a variável dependente tenha natureza dicotômica, resultará em valores negativos ou valores maiores que 1 os quais seriam difíceis de interpretar.

Os métodos de regressão para Souza (2006), Montgomery e Runger (2003) e Tsuchiya (2002) têm sido um importante componente em várias análises de dados, que trata do relacionamento entre a variável resposta e uma ou mais variáveis explicativas.

O próximo capítulo mostrará de forma detalhada os modelos de regressão linear e logística utilizados nesse trabalho.

CAPÍTULO 3 – ANÁLISE DE REGRESSÃO

Usada na determinação do coeficiente do graduando em função das variáveis de entrada no vestibular, a regressão é o método padrão de análise nesse contexto e será discutido nesse capítulo. Similarmente, para explorar o efeito da desistência estudantil, a Regressão Logística será também discutida.

3.1 Regressão Linear

Muitas vezes a posição dos pontos experimentais em um diagrama de dispersão sugere a existência de uma relação funcional entre duas variáveis. Surge então o problema de se determinar uma função que exprima esse relacionamento. Esse é o problema da regressão.

Assim, se os pontos experimentais se apresentarem como na *Figura 3.1* admite-se a existência de um relacionamento funcional entre os valores y e x, responsável pelo aspecto do diagrama, e que explica grande parte da variação de y com x, ou vice-versa. Este relacionamento funcional corresponderia à linha existente na *Figura 3.1*, que seria a "linha de regressão". Uma parcela da variação, entretanto, permanece em geral sem ser explicada, e será atribuída ao acaso. Em outras palavras, admite-se existir uma função que justifica, em média, a variação de uma das variáveis com a outra.

Na prática, os pontos experimentais terão uma variação em torno da linha representativa dessa função, devido à existência de uma variação aleatória adicional chamada de variação residual.

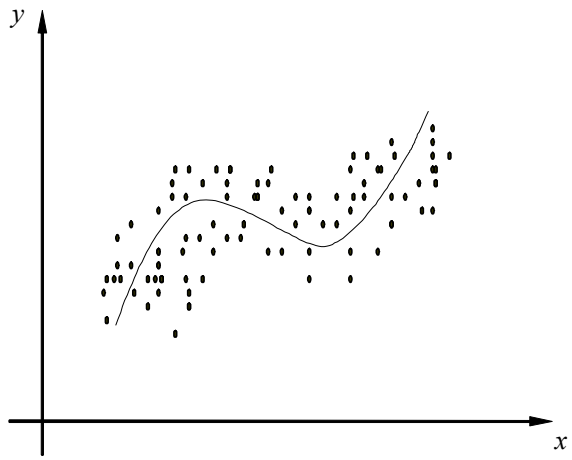


Figura 3.1 - Linha de regressão

Essa função de regressão, portanto, exprime o valor médio de uma das variáveis em função da outra. Posto dessa forma, o problema a ser examinado será, dados os pontos

experimentais, o de realizar uma indução quanto à expressão matemática da função de regressão. Evidentemente, tudo se simplificará se a forma da linha de regressão for supostamente conhecida. O problema, então, se reduzirá apenas à estimação de seus parâmetros. Esse caso ocorrerá se existirem razões teóricas que permitam saber de antemão qual o modelo que rege o comportamento de uma variável em função da outra. Pode também ocorrer o caso em que a forma da linha fica evidente da própria análise do diagrama de dispersão.

Caso a forma da linha de regressão não seja conhecida de antemão, ela deverá ser inferida juntamente com seus parâmetros. Tem-se, então, além do problema de estimação dos parâmetros do modelo da linha de regressão, a dificuldade adicional de especificar a forma do modelo. Em muitos casos, partindo-se de uma relação de forma supostamente conhecida, testes estatísticos são feitos para validar o modelo encontrado.

Quando a linha de regressão aproxima-se de uma reta, supõe-se que o problema seja de regressão linear, que será simples, quando o problema apresenta duas variáveis, ou múltipla, no caso em que mais de duas variáveis são envolvidas. Obtém-se nesse caso, uma equação para prever valores de uma variável dependente em função de duas ou mais variáveis independentes. Nos dois casos a idéia e os princípios fundamentais serão os mesmos exemplificados em seguida.

Adota-se a hipótese que a variável x seja suposta sem erro, ou seja, não-aleatória, enquanto que a variável y apresenta uma parcela de variação residual, a qual é responsável pela dispersão dos pontos experimentais em torno da linha de regressão. Essa suposição permite utilizar um modelo que simplifica a solução do problema, e é justificável porque muitos casos práticos se aproximam dele. Encontram-se, na prática, muitos casos em que a variável x pode ser medida com precisão muito maior do que y, o que coloca o problema praticamente nas condições supostas.

A situação descrita corresponde, muitas vezes, a experimentos em que os valores de x são pré-determinados ou pré-escolhidos pelo experimentador, já que a variável x é suposta não-aleatória. No entanto, os valores de y, sendo aleatórios, não podem ser exatamente previstos, e serão determinados experimentalmente. Pode-se, por exemplo, medir as temperaturas de um forno em aquecimento de 5 em 5 min., a partir de um instante 0. A menos de pequenas imprecisões, totalmente desprezíveis, os tempos (valores de x) estão bem determinados, ao passo que as temperaturas deverão ser verificadas no decurso do experimento. Vemos que, neste exemplo, os valores de x independem dos de y, pois foram simplesmente arbitrados, enquanto que os valores de y dependerão dos de x desde que exista

regressão. Por essa razão, a variável x é dita variável independente, enquanto y é dita variável dependente.

O modelo acima descrito, portanto, considera que os valores da variável aleatória y dependerão do(s) valor(es) assumido(s) pela(s) variável (is) independente(s) e também do acaso, isto é, estarão sujeitos a uma variação aleatória que se sobrepõe à variação explicada pela função de regressão. Isso pode ser expresso sob a forma

$$y = \varphi(x) + \psi, \tag{3.1}$$

onde φ denota a função de regressão e ψ a componente aleatória da variação de y. No caso de regressão múltipla, x deverá ser interpretado como um vetor de valores das variáveis independentes.

É perfeitamente coerente com a idéia contida no modelo admitir-se que a variável aleatória ψ tenha média 0, a fim de que toda a variação explicada de y fique concentrada em $\varphi(x)$. Isso significa que a função de regressão fornece a média de y para cada x considerado, conforme já mencionado.

Se a linha teórica de regressão é uma reta e deseja - se estabelecer a regressão de y em função de x, a função desejada é da forma:

$$y = \alpha + \beta x \tag{3.2}$$

onde α é o intercepto e β a inclinação da reta.

Estima-se os parâmetros α e β da reta teórica através dos pontos experimentais fornecidos pela amostra, obtendo uma reta estimativa na forma:

$$\hat{y} = a + bx, \tag{3.3}$$

onde a é a estimativa do parâmetro α, e b, também chamado coeficiente de regressão linear, é a estimativa do parâmetro β. O símbolo $\hat{y}$ é utilizado para uma conveniente distinção dos valores dados pela reta estimativa, das ordenadas dos pontos experimentalmente obtidos.

Existem diversos métodos para a obtenção da reta desejada. O mais simples de todos, que pode ser chamado de "método do ajuste visual", consiste simplesmente em traçar diretamente a reta, com auxílio de uma régua, no diagrama de dispersão, procurando fazer, da melhor forma possível, com que essa reta passe por entre os pontos. Esse procedimento, entretanto, somente será razoável se a correlação linear for muito forte, caso contrário levará a resultados subjetivos. Acima de tudo, ademais, merece a crítica de ser um procedimento nem um pouco científico.

Por outro lado, a aplicação do *princípio de máxima verossimilhança* leva, nas condições admitidas, ao chamado procedimento de *mínimos quadrados*, segundo o qual a reta a ser adotada deverá ser aquela que torna mínima a soma dos quadrados das distâncias da reta aos pontos experimentais, medidas no sentido da variação aleatória. Ou seja, deve-se procurar a reta para a qual se consiga minimizar $\Sigma_{i=1}^{n} d_i^2$, sendo as distâncias d_i as indicadas na *Figura 3.2*. A idéia central desse procedimento é simplesmente a de minimizar a variação residual em torno da reta estimativa.

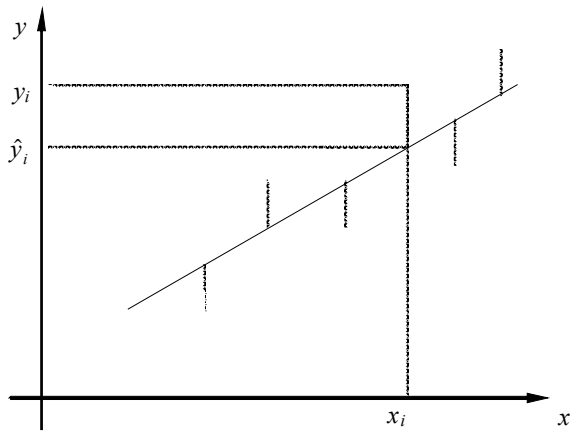


Figura 3.2 - Distâncias cuja soma dos quadrados deve ser minimizada.

Tendo em vista a expressão $\hat{y} = a + bx$, deve-se, portanto, impor a condição

$$\min \Sigma_{i=1}^{n} d_i^2 = \min \Sigma_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \min \Sigma_{i=1}^{n} \left(y_i - a - bx_i\right)^2 \tag{3.4}$$

Os valores *a* e *b* que minimizam essa expressão serão aqueles que anulam as derivadas parciais dessa expressão. Ou seja, deve-se ter

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} d_i^2 = 0 \text{ e } \frac{\partial}{\partial b} \sum_{i=1}^{n} d_i^2 = 0. \tag{3.5}$$

Considerando a última forma da expressão, chega-se às expressões

$$\begin{aligned} -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0, \\ -2\sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0, \end{aligned} \tag{3.6}$$

as quais imediatamente fornecem o seguinte sistema de duas equações a duas incógnitas:

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i, \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{cases} \tag{3.7}$$

Os pontos experimentais fornecem os elementos para a montagem desse sistema, cuja solução forneceria os coeficientes *a* e *b*. Entretanto é mais fácil considerar de uma vez a solução analítica do sistema, a qual fornece expressões que dão diretamente os coeficientes *a* e *b* que se deseja obter.

$$\begin{cases} b = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}, \\ a = \bar{y} - b\bar{x}, \end{cases} \tag{3.8}$$

Um exemplo: Deseja-se obter a equação de regressão para os seguintes pontos experimentais:

x	1	2	3	4	5	6	7	8
y	0,5	0,6	0,9	0,8	1,2	1,5	1,7	2,0

Valores para o ajuste da reta:

$$\Sigma x_i = 36$$

$$\Sigma y_i = 9{,}2$$

$$\bar{x} = 4{,}5$$

$$\bar{y} = 1{,}5$$

$$\boldsymbol{S}_{xy} = 9{,}1$$

$$\boldsymbol{S}_{xx} = 42$$

Resolvendo a *Equação 3.8* são encontrados os valores necessários à determinação dos coeficientes da reta.

$b = \boldsymbol{S}_{xy} / \boldsymbol{S}_{xx} = 9{,}1 / 42 = 0{,}217$

$a = \bar{y} - b\bar{x} = 1{,}15 - (0{,}217).4{,}5 = 0{,}1735$

Logo, a função de regressão para a *Figura 3.3* é:

$$\hat{y} = 0{,}174 + 0{,}217\,x \tag{3.9}$$

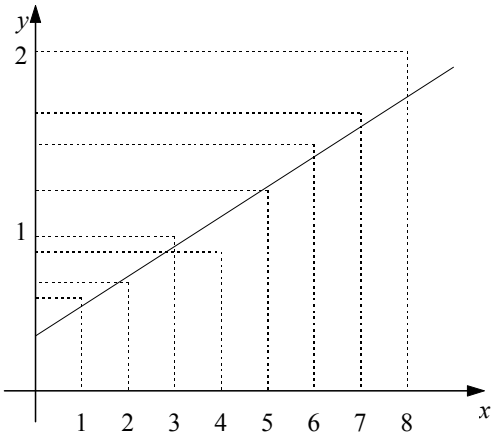


Figura 3.3 - Reta de Mínimos Quadrados

A regressão múltipla, que envolve uma variável dependente e duas ou mais variáveis independentes, tem a finalidade de melhorar a capacidade de predição comparando-se com a regressão linear simples.

A equação da regressão múltipla tem a seguinte forma:

$$y = a + b_1x_1 + b_2x_2 + ... + b_kx_k, \tag{3.10}$$

onde:

a = intercepto do eixo y;

b_i = coeficiente angular da i-ésima variável;

k = número de variáveis independentes.

Também na regressão múltipla, as estimativas dos mínimos quadrados são obtidas pela escolha dos estimadores que minimizam a soma dos quadrados dos desvios entre os valores observados e os valores ajustados.

Na regressão simples:

b = aumento em y, decorrente de um aumento unitário em x.

Na regressão múltipla:

b_i = aumento em y se x_i for aumentado de 1 unidade, mantendo-se constantes todas as demais variáveis.

Com os coeficientes estimados, os valores das variáveis do modelo são testados. A análise dos resultados tem como principais componentes o *Valor P, valor dos coeficientes, o Coeficiente de Determinação* R^2 *e O Coeficiente de Determinação ajustado* R^2.

O Valor P testa a hipótese nula de que os coeficientes são iguais a zero versus estes sendo diferentes de zero. Quando o Valor P for <= 0, 05 (denominado de nível de

significância), a hipótese nula será rejeitada. Quando o Valor P for > 0, 05, a hipótese nula não será rejeitada. De uma forma sucinta as seguintes hipóteses são estabelecidas:

- H_0: todos os coeficientes são iguais a zero (não há regressão);
- H_1: pelo menos um dos coeficientes é diferente de zero (há regressão).

Chama-se Coeficiente de Determinação R^2 a medida do grau de ajustamento da equação de regressão aos dados amostrais. É uma medida descritiva da proporção da variação de y que pode ser explicada por x. Um ajuste perfeito resultaria em R^2=1. Um ajuste muito bom acarreta um valor próximo de 1. E um ajuste fraco ocasiona um valor de R^2 próximo de zero.

O Coeficiente de Determinação é uma medida da aderência da equação de regressão aos dados amostrais, mas apresenta o problema de aumentar na medida em que se incluem mais variáveis. Conseqüentemente, é melhor usar o coeficiente de determinação ajustado R^2adj porque ele ajusta o valor de R^2 com base no número de variáveis e no tamanho da amostra.

3.2 Regressão Logística

A regressão logística é um modelo probabilístico que descreve a relação entre uma ou mais variáveis explicativas e uma variável resposta quando esta é binária ou dicotômica. Essa relação determina a probabilidade de ocorrência de um evento em presença de um conjunto de variáveis independentes ou explicativas (TSUCHIYA, 2002).

Segundo Hosmer e Lemeshov (2000) métodos de regressão têm sido importantes na análise de dados que descrevem a relação entre uma variável resposta e uma ou mais variáveis explicativas. Assim como nos casos de regressão linear, em que a variável resposta é contínua, isto é também freqüente nos casos em que a variável é discreta, podendo assumir dois ou mais valores possíveis. O modelo de regressão logística tornou-se um método padrão de análise dessa situação. O objetivo da regressão logística é encontrar o melhor relacionamento entre uma variável resposta discreta e um conjunto de variáveis independentes.

Os métodos usados na análise de regressão logística seguem os mesmos princípios usados em regressão linear (SCHUSTER, 2000; HOSMER e LEMESHOW, 2000). A diferença entre as técnicas de regressão se deve ao fato de que na regressão logística as variáveis dependentes ou resposta são expressas por meio de uma probabilidade de ocorrência, enquanto na regressão linear obtém-se um valor numérico**.** A abordagem

probabilística é uma das vantagens da regressão logística, pois permite estimar a chance de um evento ocorrer a partir de um conjunto de variáveis independentes.

Segundo Bittencourt e Clarke (2001), com o modelo logístico é possível trabalhar com um número pequeno de parâmetros, com isso passa ser necessário um número menor de amostras de treinamento para o processo de estimação de parâmetros.

De acordo com Gimeno e Souza (1995), a análise logística controla grande número de variáveis simultaneamente, permitindo que os dados sejam utilizados de forma mais eficiente**.** O modelo logístico é mais flexível, com maior poder de exploração de variáveis. A existência de programas de microcomputador e de uso livre torna cada vez mais conhecida e popular a análise logística.

O exemplo a seguir procura mostrar o uso de regressão logística na área acadêmica. A análise desenvolvida é similar para o caso dos alunos desistentes ou não, considerados no capítulo seguinte:

Um exemplo ilustrativo:

Uma pesquisa tenta relacionar a porcentagem de doutores com conceitos 3 e 5 de 77 programas de pós-graduação. A *Tabela 3.1* apresenta os dados colhidos dos programas de estudados. É possível perceber que a porcentagem de doutores dos diversos programas varia entre 5% e 99%. Dentro da amostra, há 37 programas com conceito 3 e 40 com conceito 5. Os valores 3 e 5 foram codificados para os valores binários 0 e 1, respectivamente, representando assim as respostas de uma *regressão logística binária*.

Neste exemplo a variável resposta só pode assumir dois valores, ou seja, conceito 3 ou 5. Pode-se perceber que existe uma tendência dos programas com conceito 5 terem uma maior porcentagem de doutores e verifica-se também que há uma grande sobreposição de dados. Para o estudo, a variável resposta teve seus valores, conceitos 3 e 5, substituídos por 0 e 1 respectivamente, como mostra a *Tabela 3.1*

Porcentagem de Doutores	**Conceitos**	**Porcentagem de Doutores**	**Conceitos**
5	0	57	0
4	0	57	1
6	0	59	1
8	0	60	1
9	0	61	0
10	0	64	1
10	0	65	0
11	0	66	1

Tabela 3.1 – Distribuição da porcentagem de doutores x conceito dos programas.
(Os conceitos 3 e 5 foram codificados para 0 e 1, respectivamente)

13	0	67	1
15	0	68	1
17	0	70	1
17	0	73	0
18	1	75	1
22	0	76	1
24	0	77	1
25	0	78	1
27	0	79	1
28	0	81	0
29	1	81	1
30	0	83	1
30	0	84	1
33	0	86	1
36	0	87	1
37	0	88	1
39	0	88	1
40	1	89	1
43	0	90	1
45	1	91	0
45	0	92	1
47	1	92	1
47	0	94	1
48	0	95	1
52	0	95	1
53	0	97	1
53	0	98	1
55	0	98	1
56	1	98	1
56	1	99	1
		99	1

Tabela 3.1 – Continuação

O *Figura 3.4* mostra a distribuição do percentual de doutores entre as respostas 0 e 1.

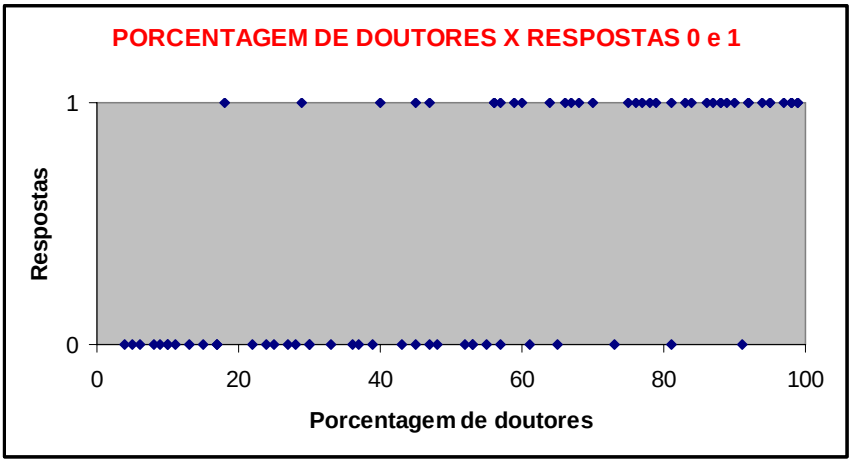


Figura 3.4 - Porcentagem de Doutores x Respostas 0 e 1

A *Figura 3.4* apresenta apenas duas linhas horizontais, o que dificulta uma interpretação mais detalhada do relacionamento entre as variáveis. Para uma melhor análise, a variável independente *"porcentagem de doutores"* foi dividida em 9 faixas, conforme apresentado na *Tabela 3.2*. Nesta tabela é possível perceber a quantidade de programas com conceito 3 e 5 em cada faixa de porcentagem. A divisão em faixas permite que seja feito um agrupamento, representando cada faixa pelo seu ponto médio sendo possível verificar a proporção do conceito 5 em cada faixa.

Faixas %	n	Conceito 3	Conceito 5	Proporção do Conceito 5	Ponto Médio
0-20	13	12	1	0,08	10
21-30	8	7	1	0,13	25,5
31-40	5	4	1	0,20	35,5
41-50	6	4	2	0,33	45,5
51-60	10	5	5	0,50	55,5
61-70	7	2	5	0,71	65,5
71-80	6	1	5	0,83	75,5
81-90	10	1	9	0,90	85,5
91-100	12	1	11	0,92	95,5
TOTAL	**77**	**37**	**40**	**0,52**	

Tabela 3.2 – Faixa com a porcentagem de doutores

O agrupamento realizado é mostrado no *Figura 3.5*, que permite uma melhor visualização do relacionamento entre o ponto médio de cada faixa de porcentagem de doutores e a proporção dos programas com conceito 5. Verifica-se que o relacionamento entre as variáveis não é linear, mas um relacionamento em "forma de S".

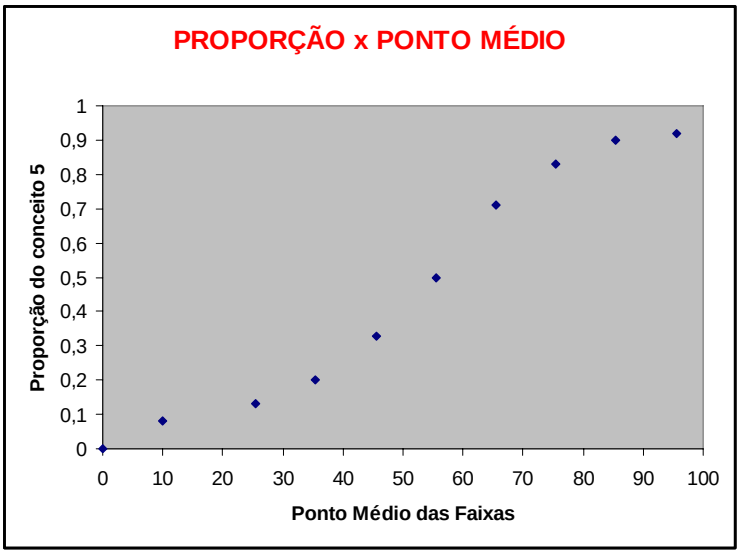


Figura 3.5 - Proporção x Ponto Médio

Muitas funções de distribuição têm sido propostas para o uso na análise de variáveis dicotômicas. A escolha da distribuição logística para este caso se deve ao fato de que a função logística é extremamente flexível, fácil de ser usada e interpretada e permite uma interpretação intrínseca à natureza do problema. (HOSMER e LEMESHOW, 2000).

Conhecendo-se a proporção de programas com conceito 5, torna-se necessário estudar o relacionamento entre a variável independente e a probabilidade de ocorrência do evento, π(x), tido como a probabilidade de "sucesso".

No caso apresentado, o evento é definido pela ocorrência de conceito 5 e a variável independente é a porcentagem de doutores nos programas. Se a probabilidade *π(x)* fosse tratada como uma função linear $\pi(x) = \alpha + \beta x$ surgiria os seguintes problemas:

- A probabilidade *π(x)*, interpretada como a resposta *y* para um dado valor *x*, deve estar entre 0 e 1 quando a variável é dicotômica. A função linear permite uma resposta em toda a reta real.
- O desvio padrão de *Y* para uma variável dependente de distribuição binomial é $V(y) = \pi(x)[1 - \pi(x)]$, o que não é adequado ao método dos mínimos quadrados usado para o ajuste da curva em modelos lineares.

Diante dos problemas descritos a função que descreve a forma "S" do modelo logístico é a função logística dada pela *Equação 3.11*:

$$\pi(x) = \frac{e^{(\alpha+\beta x)}}{1 + e^{(\alpha+\beta x)}} \tag{3.11}$$

Na equação acima é possível notar o relacionamento da probabilidade de sucesso com os dois parâmetros, α e β . O α está relacionado com o posicionamento da curva no eixo horizontal e o β com a mudança entre sucesso e fracasso, ou seja, quão abrupta é a curva "S". Se $\beta < 0$, a probabilidade de sucesso diminui com *x*. Se $\beta = 0$, a probabilidade de resposta é independente de *x*. É através desta análise que surgem as hipóteses nula e alternativa do modelo de regressão logística:

H_0: $\beta = 0$ (todos os β são zero): não há regressão logística

H_1: $\beta \neq 0$ (pelo menos um β é diferente de zero): há regressão logística

Ao se realizar o ajuste de uma curva de regressão deve-se testar as hipóteses nula e alternativa do modelo. A *Figura 3.6* mostra a forma "S" do modelo logístico.

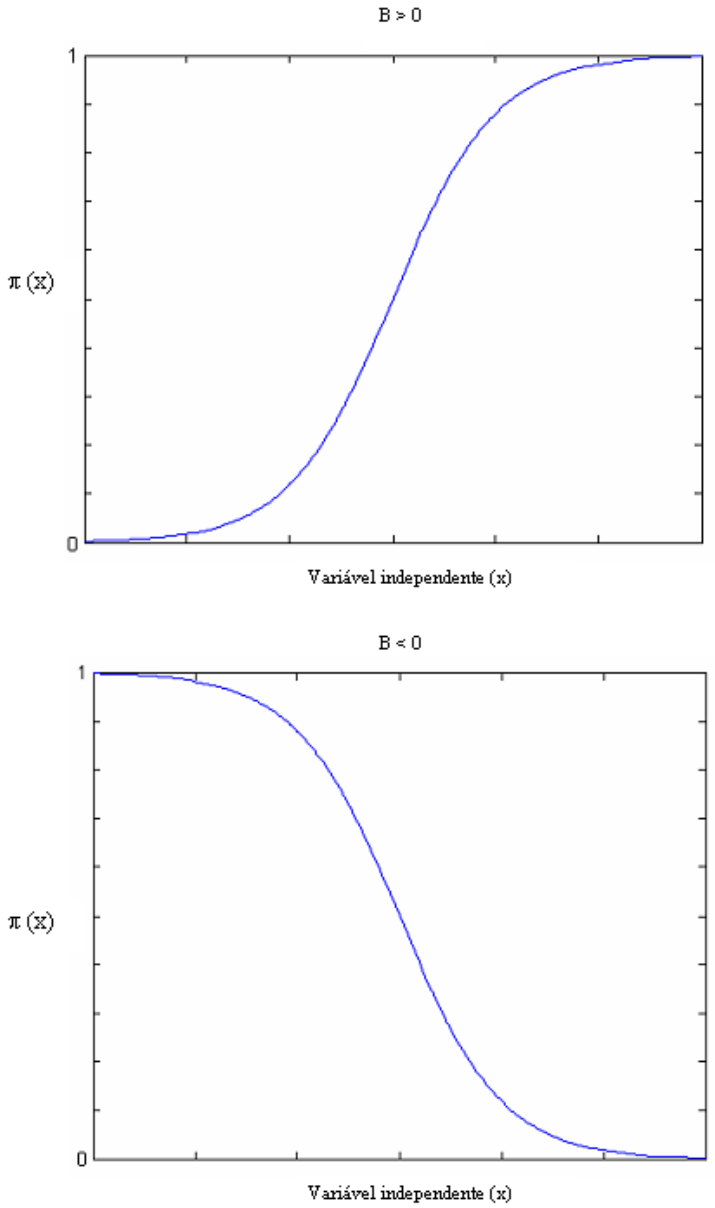


Figura 3.6 – Forma "S" do modelo Logística
Fonte: Adaptado de Pereira (2005)

3.2.1 Modelos de Regressão Logística

3.2.1.1 Modelo de Regressão Logística Simples

Os métodos de regressão têm como objetivo descrever as relações entre a variável resposta (Y) e a variável explicativa (X). Na Regressão Logística, a variável resposta (Y) é dicotômica, isto é, atribui-se o valor 1 para o acontecimento de interesse ("sucesso") e o valor 0 para o acontecimento complementar ("fracasso") com probabilidades $\pi_i = P\ (\ Y = 1 \mid X = x_i\ )$ e $1 - \pi_i = P\ (\ Y = 0 \mid\ X = x_i\ )$, respectivamente (SOUZA, 2006).

A probabilidade de sucesso do modelo logístico simples é definida pela *Equação 3.12* como:

$$\pi_i = \pi(x_i) = P(Y = 1 \mid X_i) = \frac{e^{(\alpha+\beta\, x_i)}}{1+e^{(\alpha+\beta\, x_i)}} \tag{3.12}$$

e a probabilidade de fracasso pela *Equação 3.13*:

$$1-\pi_i = 1-\pi(x_i) = P(Y = 0 \mid X_i) = \frac{1}{1+e^{(\alpha+\beta\, x_i)}} \tag{3.13}$$

em que $(\alpha, \beta)^T$ é o vetor de parâmetros desconhecidos.

3.2.1.2 Modelo de Regressão Logística Múltipla

Segundo Hosmer e Lemeshow (2000) e Souza (2006) podemos denominar modelo de regressão logística múltipla, quando o modelo possue mais de uma variável independente, podendo considerar um vetor $x = (x_1, x_{2,\ldots,} x_p)$. A função de ligação é expressa pela equação logística:

$$g(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3.14}$$

No caso o modelo de regressão logística é:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \tag{3.15}$$

A probabilidade de sucesso do modelo de regressão logística múltipla é definida como:

$$\pi_i = \pi(x_i) = P(Y_i = 1 \mid X_i) = \frac{e^{(\alpha+\beta_1 x_i+\cdots+\beta_p x_p)}}{1+e^{(\alpha+\beta_1 x_i+\cdots+\beta_p x_p)}} \tag{3.16}$$

e a probabilidade de fracasso:

$$1-\pi_i = 1-\pi(x_i) = P(Y_i = 0 \mid X_i) = \frac{1}{1+e^{(\alpha+\beta_1 x_i+\cdots+\beta_p x_p)}} \tag{3.17}$$

A regressão logística, como qualquer tipo de regressão, busca ajustar uma curva aos pontos experimentais. Para se encontrar a regressão logística dos pontos experimentais, é necessário definir critérios de aproximação entre as curvas teóricas e as observações medidas (PEREIRA, 2005).

Suponha que (x_i,y_i) seja uma amostra com n pares de observações, onde y_i, representa o valor da variável resposta dicotômica e x_i o valor da variável independente da i-ésima

observação em que *i = 1,2,...n*. Para o ajuste do modelo de regressão logística é necessário estimar os parâmetros.

Para estimar os parâmetros desconhecidos utiliza-se em regressão logística o método de máxima verossimilhança (SOUZA, 2006; BITTENCOURT e CLARKE, 2001; HOSMER e LEMESHOW, 2000; NELSON, 1999). Este método estima os valores dos parâmetros α e β através da maximização da probabilidade de se obter o conjunto de dados observados. Esta probabilidade é descrita por uma função, chamada de função de verossimilhança. Quando se maximiza esta função encontram-se os parâmetros que definem onde os dados têm probabilidade máxima de ocorrência.

A função de probabilidade indica a chance de um evento ocorrer. Um evento pode assumir apenas dois valores, sucesso ou fracasso. Esta função é dada por:

$$f = \pi^{y}\left(1-\pi\right)^{1-y} \tag{3.18}$$

A probabilidade de sucesso π depende dos valores de *x*. Assim a função de probabilidade é especificada em função de *x*.

$$f\left(x\right) = \pi\left(x\right)^{y}\left[1-\pi\left(x\right)\right]^{1-y} \tag{3.19}$$

Considerando vários pares ordenados *(x,y)*, ou seja, várias observações, adiciona-se o índice *i* para cada par, ou seja, para cada observação. Assim escreve-se:

$$f\left(x_i\right) = \pi\left(x_i\right)^{y_i}\left[1-\pi\left(x_i\right)\right]^{1-y_i} \tag{3.20}$$

onde $y_i \in \left[0,1\right]$

Neste estudo, Y pode assumir os valores 1 para "sucesso" ou 0 para "fracasso".
Substituindo Y = 1 na função de probabilidade

$$f\left(x_i\right) = \pi\left(x_i\right)^{y_i}\left[1-\pi\left(x_i\right)\right]^{1-y_i}$$

$$f\left(x_i\right) = \pi\left(x_i\right)^{1}\left[1-\pi\left(x_i\right)\right]^{1-1}$$

$$f\left(x_i\right) = \pi\left(x_i\right)\left[1-\pi\left(x_i\right)\right]^{0}$$

$$f\left(x_i\right) = \pi\left(x_i\right) = sucesso$$

Substituindo y = 0 na função de probabilidade

$$f\left(x_i\right) = \pi\left(x_i\right)^{y_i}\left[1-\pi\left(x_i\right)\right]^{1-y_i}$$

$$f\left(x_i\right) = \pi\left(x_i\right)^{0}\left[1-\pi\left(x_i\right)\right]^{1-0}$$

$$f\left(x_i\right) = \left[1-\pi\left(x_i\right)\right] \; = fracasso$$

Pode-se perceber que, para os pares (x_i,y_i) em que $y_i = 1$, a contribuição para a função verossimilhança é $\pi(x_i)$ e para os pares em que $y_i = 0$, a contribuição é 1 - $\pi(x_i)$.

A regressão logística supõe que cada evento ocorre de forma independente. Como a probabilidade de dois eventos independentes ocorrerem é o produto das probabilidades individuais, a probabilidade de ocorrência de todos os eventos $f(x_i) = \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}$, é o produto de todas as funções $f(x_i)$.

Assim, a função de verossimilhança é dada por:

$$l(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi)^{1-y_i} \tag{3.21}$$

O símbolo $\prod$ é o produtório, indica o produto das funções. O método de máxima verossimilhança escolhe valores de β que maximizam a função. Para facilitar o tratamento matemático da expressão acima, trabalha-se com o logaritmo da função de verossimilhança, que transformará o produtório em somatório, devido à propriedade de o logaritmo de um produto ser a soma dos logaritmos dos fatores. É utilizada também a propriedade da multiplicação do logaritmo pelo expoente do logaritmando.

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \tag{3.22}$$

Para encontrar o valor de β que maximiza $L(\beta)$ é calculada a derivada de $L(\beta)$ em relação a α e β e o resultado de cada uma igualada a zero (HOSMER e LEMESHOW, 2000). Então têm-se as chamadas equações de verossimilhança que são:

$$\sum[y_i - \pi(x_i)] = 0 \tag{3.23}$$

$$\sum x_i[y_i - \pi(x_i)] = 0 \tag{3.24}$$

Devido à complexidade de seus cálculos, métodos numéricos e iterativos são necessários para se chegar a uma resolução. Estes métodos iterativos estão disponíveis em diversos programas computacionais. No presente estudo será utilizado o software Minitab.

Utilizando-se do exemplo ilustrativo apresentado e fazendo uso do software Minitab para a estimação dos parâmetros e ajuste da curva de regressão logística binária, foram obtidos os seguintes parâmetros ajustados:

α: -3,993
β_1: 0,072
Portanto, o modelo logístico para os dados colhidos é:

$$\pi(x) = \frac{e^{(-3,993+0,072x)}}{1+e^{(-3,993+0,072x)}}$$

Os resultados para a probabilidade de sucesso em função da porcentagem de doutores são apresentados na *Tabela 3.3:*

Porcentagem de Doutores	Probabilidade	Porcentagem de Doutores	Probabilidade
5	0,025	57	0,539
4	0,024	57	0,539
6	0,027	59	0,575
8	0,031	60	0,593
9	0,034	61	0,610
10	0,036	64	0,661
10	0,036	65	0,677
11	0,039	66	0,692
13	0,045	67	0,708
15	0,052	68	0,723
17	0,059	70	0,751
17	0,059	73	0,789
18	0,064	75	0,812
22	0,083	76	0,823
24	0,095	77	0,834
25	0,102	78	0,843
27	0,116	79	0,853
28	0,124	81	0,870
29	0,132	81	0,870
30	0,140	83	0,886
30	0,140	84	0,893
33	0,169	86	0,906
36	0,202	87	0,912
37	0,214	88	0,918
39	0,239	88	0,918
40	0,253	89	0,923
43	0,297	90	0,928
45	0,328	91	0,933
45	0,328	92	0,937
47	0,361	92	0,937
47	0,361	94	0,945
48	0,378	95	0,949
52	0,448	95	0,949
53	0,466	97	0,955
53	0,466	98	0,958
55	0,503	98	0,958
56	0,521	98	0,958
56	0,521	99	0,961
		99	0,961

Tabela 3.3 – Probabilidade de sucesso

A *Figura 3.7* representa a tabela acima, mostrando que a probabilidade de um programa de pós-graduação obter conceito 5 cresce com a porcentagem de doutores no programa.

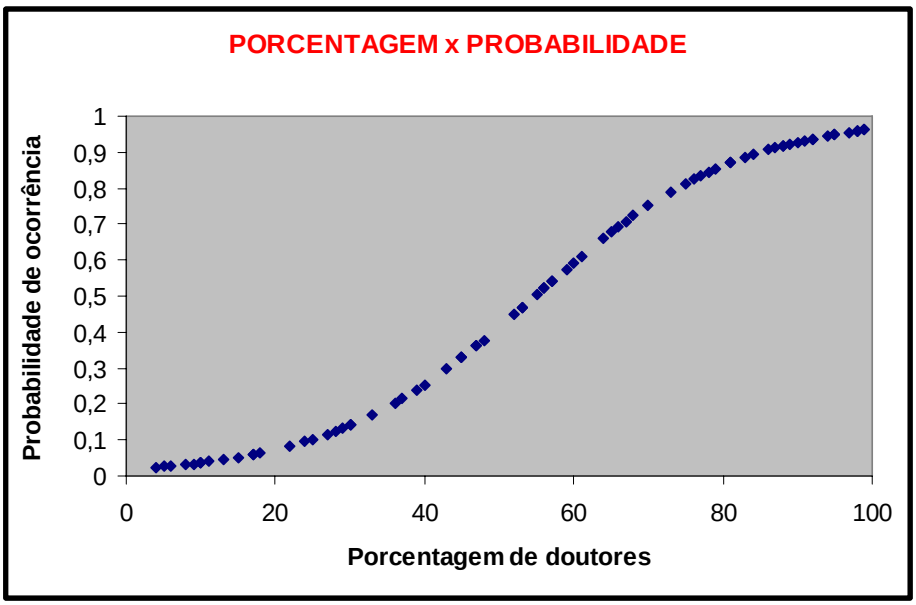


Figura 3.7 – Porcentagem x Probabilidade

Uma importante propriedade da função logística é que ela pode ser linearizada. Utilizando-se desta propriedade para a interpretação dos resultados é feita uma transformação na função logística conhecida como transformação logit.

Fazendo a transformação logit da probabilidade $\pi(x)$ obtém-se:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x \qquad (3.25)$$

Observa-se que esta transformação leva a uma equação linear em x, com α sendo o coeficiente linear e β o coeficiente angular da reta. A razão $\frac{\pi(x)}{1-\pi(x)}$ na transformação logit é chamada de ODDS (chance), a função resposta $g(x)$ é denominada função resposta logit e pode-se dizer que $-\infty \leq g(x) \leq +\infty$. A interpretação dos resultados é feita por meio da Razão de Chances (OR – Odds Ratio).

Considere a situação para a qual a variável independente é também dicotômica (SOUZA, 2006). Neste caso, a variável x será codificada como 0 e 1. Tem-se, então, a chance da resposta quando x = 0 definida por $\frac{\pi(0)}{1-\pi(0)}$ e a chance da resposta quando $x = 1$ definida por $\frac{\pi(1)}{1-\pi(1)}$.

A razão das chances (OR) é definida pela *Equação 3.26***.**

$$OR = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} = \frac{\dfrac{\left(\dfrac{e^{\alpha-\beta_1}}{1+e^{\alpha-\beta_1}}\right)}{\left(\dfrac{1}{1+e^{\alpha-\beta_1}}\right)}}{\dfrac{\left(\dfrac{e^{\alpha}}{1+e^{\alpha}}\right)}{\left(\dfrac{1}{1+e^{\alpha}}\right)}} \tag{3.26}$$

Rearranjando a *Equação 3.26* chega-se em:

$$OR = \frac{e^{\alpha+\beta_1}}{e^{\alpha}} = e^{\beta_1} \tag{3.27}$$

Aplicando o logaritmo na razão de chance obtém-se:

$$\ln e^{\beta_1} = \beta_1 \tag{3.28}$$

Assim, o parâmetro ajustado β_1 é o logaritmo natural da razão de chances. A razão de chance representa o número de vezes que a resposta Y_1 é mais provável que a resposta Y_2. Outra interpretação que se pode fazer é que a cada unidade variada em x, a resposta 1 torna-se mais provável e^{β_1}, se $\beta_1 > 0$ e menos provável e^{β_1}, se $\beta_1 < 0$.

Exemplo:

No caso estudado, obteve-se $\beta_1 = 0{,}0728$.

Achando a razão de chances pela Equação 3.27, obtém-se:

$OR = e^{0{,}072} = 1{,}075$

Ou seja, a cada unidade aumentada na porcentagem de doutores, a chance de sucesso aumenta 1,075 vezes. Isto pode ser visto, tomando a probabilidade de sucesso para 90% de doutores na *Tabela 3.3* e resolvendo a Odds para este valor de x.

$$Odds(90\%) = \frac{\pi(90\%)}{1-\pi(90\%)} = \frac{0{,}928391}{0{,}071609} = 12{,}9647$$

Portanto, a probabilidade de sucesso (obter conceito 5) para 90% de doutores é 12,9647 vezes maior a probabilidade de fracasso. Aumentando-se a porcentagem de doutores para 91%, a probabilidade de sucesso aumentará em 1,0756 vezes, ou seja, será 13,944 vezes maior que a probabilidade de fracasso.

Para a regressão logística múltipla, como pode ser visto na *Equação 3.14,* tem-se para cada variável independente x_i o seu respectivo β_i e sua correspondente Odds e OR.

Segundo Tsuchiya (2002) com os coeficientes estimados, os valores das variáveis do modelo são testados. Neste processo é envolvido o teste de hipóteses estatísticas, onde se determina quais variáveis independentes no modelo estão significantemente relacionadas com a variável resposta.

Em regressão logística há uma série de testes de ajuste e outras medidas para assegurar a validade do modelo. Estas estatísticas permitem identificar dentre as diversas variáveis do estudo quais não se ajustam bem, ou que têm forte influência sobre a estimação dos parâmetros.

Alguns testes estatísticos:

Valor P- Testa a hipótese nula de que os coeficientes são iguais a zero versus estes sendo diferentes de zero. Quando o Valor P for <= ao nível de significância 0,05(nível de significância adotado), a hipótese nula será rejeitada, quando o Valor P for > 0,05, a hipótese nula não será rejeitada.

- H_0: todos os coeficientes são iguais a zero (não há regressão logística);
- H_1: pelo menos um dos coeficientes é diferente de zero (há regressão logística).

Pearson e Deviance - São tipos de resíduos para modelos logísticos. Eles são usualmente medidos para avaliar quão bem é o ajuste do modelo selecionado de dados. Quanto maior o Valor P, melhor é o ajuste do modelo de dados. O método de Pearson compara as diferenças entre as freqüências observadas e esperadas para um certo evento.

Teste de Hosmer e Lemeshow - Tem a finalidade de avaliar o modelo ajustado comparando as freqüências observadas e as esperadas. Considerando-se Y como o valor real da variável e $\hat{Y}$ como o valor previsto, o teste é feito com intuito de medir a proximidade de ambos. A hipótese nula (hipótese de teste) é que não existe diferença significativa entre o valor real e o valor previsto, ou seja, equivale a dizer que o modelo tem bom poder de ajuste. Quanto menor é o valor da diferença entre Y e $\hat{Y}$, mais os valores previstos se aproximam dos reais e,

portanto, melhor desempenho preditivo tem o modelo. Desta forma, um fator positivo a favor do modelo é quando se aceita a seguinte hipótese nula: $H_0 : Y = \hat{Y}$ ou $H_0: Y - \hat{Y} = 0$ (ZANINI, 2007).

Os testes de Person, Deviance e Hosmer-Leshow testam a hipótese nula de que o ajuste dos dados é bom versus o ajuste sendo ruim. Quanto maior o Valor P, melhor é a qualidade de ajuste do modelo.

- H_0: Valor P > 0,05 – O ajuste dos dados é bom;
- H_1: Valor P <= 0,05 – O ajuste dos dados não é bom;

CAPÍTULO 4 – ANÁLISE DESCRITIVA DOS DADOS

Este capítulo faz a análise gráfica e descritiva do banco de dados disponível, relacionando todas as variáveis dependentes e independentes. Modelos de regressão para o coeficiente final do aluno e para a situação atual do aluno ao fim de seu período de curso serão também explorados. A pesquisa foi realizada na Universidade Federal de Itajubá – UNIFEI, que por muitos anos especializou-se em Engenharia, sendo uma referência na área. Conforme a necessidade do mercado, outros cursos foram surgindo e hoje a universidade é composta por 11 cursos de graduação: Administração de Empresas (ADM), Ciência da Computação (CCO), Engenharia Ambiental (EAM), Engenharia de Controle e Automação (ECA), Engenharia da Computação (ECO), Engenharia Elétrica (EEL), Engenharia Hídrica (EHD), Engenharia Mecânica (EME), Engenharia de Produção (EPR), Física – Bacharelado (FBA), Física – Licenciatura (FLI). Foi analisado o desempenho de 1489 alunos da UNIFEI nos vestibulares de 2000, 2001, 2002 e 2003.

4.1 Descrição do Banco de Dados

O banco de dados deste trabalho se refere às informações dos vestibulares nos anos de 2000, 2001, 2002 e 2003 da Universidade Federal de Itajubá - UNIFEI. O banco de dados é composto pelos seguintes itens: Cursos de graduação avaliados; variáveis independentes e variáveis dependentes.

- ✓ **Curso de Graduação:**

Os cursos de graduação que compõem este banco são apresentados abaixo na *Tabela 4.1:*

CURSO	SIGLA	PERÍODO DE CONCLUSÃO
Administração de Empresas	ADM	5 ANOS
Ciência da Computação	CCO	5 ANOS
Engenharia Ambiental	EAM	5 ANOS
Engenharia de Controle e Automação	ECA	5 ANOS
Engenharia da Computação	ECO	5 ANOS
Engenharia Elétrica	EEL	5 ANOS
Engenharia Hídrica	EHD	5 ANOS
Engenharia Mecânica	EME	5 ANOS
Engenharia de Produção	EPR	5 ANOS
Física - Bacharelado	FBA	4 ANOS
Física - Licenciatura	FLI	4 ANOS

Tabela 4.1 – Curso de Graduação

- ✓ **Variáveis Independentes (X):**

  As seguintes variáveis independentes estão presentes no banco de dados:

  - Curso Matriculado;
  - Nota do ENEM;
  - Notas do Vestibular (Valor entre 0 e 100):
    - Matemática
    - Física
    - Português
    - Química
    - Inglês
    - História
    - Geografia
    - Biologia
    - Redação
  - Nota final do aluno (Nota Final do Vestibular).

- ✓ **Variáveis Dependentes (Y):**

  Como variáveis de resposta, as seguintes situações foram obtidas:

  - Coeficiente Final do aluno (Valor entre 0 e 100);
  - Situação Final do aluno, de acordo com os seguintes níveis:
    - Ativo (aluno que ainda frequenta o curso, após o seu período previsto para término)
    - Colou Grau (aluno que se formou no tempo previsto)
    - Desistiu a Pedido (aluno que solicitou o desligamento)
    - Desistiu por Processo (aluno que sofreu desligamento do curso por razões como, número de reprovações, tempo excessivo, etc...)
    - Faleceu
    - Mudou de curso por ter feito novo vestibular em outra universidade
    - Mudou de curso por transferência interna
    - Transferiu-se para outra instituição

Para a análise e conclusões referentes às informações do banco de dados, foram processados 1489 casos. A *Tabela 4.2* apresenta de forma detalhada o número de casos

processados para cada ano. No ano de 2000 e 2001 não foram realizados vestibulares para os cursos de Física-Bacharelado e Física-Licenciatura.

ANOS	CURSOS AVALIADOS											TOTAL DE CASOS
	ADM	CCO	EAM	ECA	ECO	EEL	EHD	EME	EPR	FBA	FLI	
2000	30	30	24	34	39	70	20	59	31	x	x	337
2001	30	30	30	35	49	70	20	58	30	x	x	352
2002	30	30	30	40	50	70	20	60	30	20	20	400
2003	31	30	30	40	50	70	20	60	30	19	20	400
TOTAL	121	120	114	149	188	280	80	237	121	39	40	1489

Tabela 4.2 – Dados Processados

4.2 Caracterizações para os dados dos alunos ingressos no período 2000-2003

Este item faz uso de boxplots para representar tanto as notas dos diversos cursos nos vestibulares como para representar o coeficiente final de cada curso. Utiliza também gráficos de setores para mostrar a situação final dos cursos. A análise será feita primeiramente considerando cada ano individualmente e ao final, é feita a análise conjunta para todo o período em questão.

Assim, para cada ano, a análise será dividida em quatro partes:

- Boxplots com o resumo das notas do vestibular por curso e por disciplina;
- Boxplots do coeficiente final dos alunos em cada curso;
- Gráficos de setores representando a situação final dos cursos;
- Correlações entre as variáveis independentes.

4.2.1 Caracterização para os dados dos alunos ingressos em 2000

4.2.1.1 Dados Expressos por Boxplot

Os dados serão representados por diagramas de caixa (Boxplots), que são utilizados para visualizar a distribuição dos dados e a presença de *outliers* (valores extremos). O Boxplot é um gráfico de dados que consiste em uma reta que se prolonga do menor ao maior valor, e um retângulo com retas traçadas no primeiro quartil $Q1$, na mediana e no terceiro quartil $Q3$.

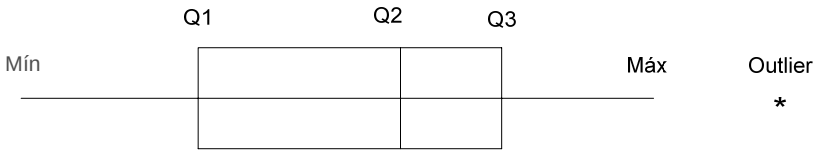


Figura 4.1 – Representação Boxplot

A mediana de um conjunto de valores é o valor do meio do conjunto, quando os valores estão dispostos em ordem crescente ou decrescente. Se o número de valores é ímpar, a mediana é o número localizado exatamente no meio da lista. Se o número é par, a mediana é a média dos dois valores do meio.

Os quartis são definidos como sendo os valores que dividem os dados ordenados em ordem crescente em quatro partes iguais. Assim, por definição tem-se:

Q1 – primeiro quartil, separa os 25% inferiores dos 75% superiores dos valores ordenados;

Q2 – segundo quartil, é a mediana;

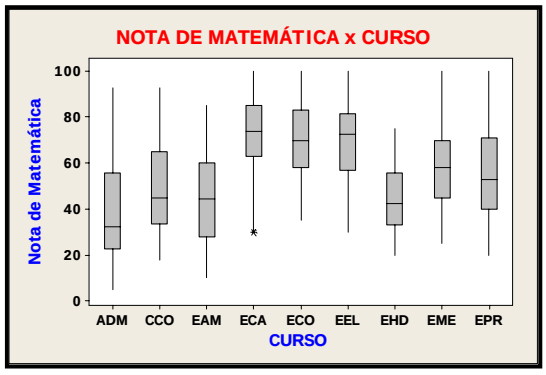
Q3 – terceiro quartil, separa os 25% superiores dos 75% inferiores dos valores ordenados.

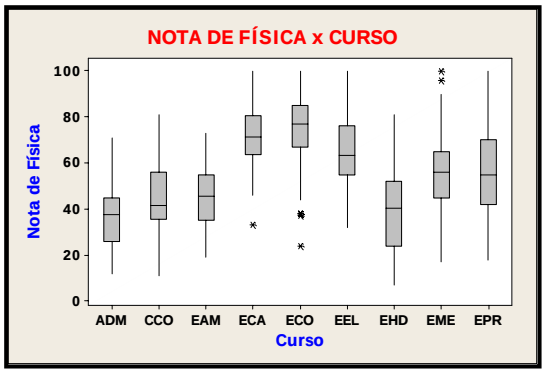
Outliers ou valores extremos - são valores raros, no sentido de que estão muito afastados da maioria dos dados. O cálculo dos *outliers* é definido pela seguinte expressão:

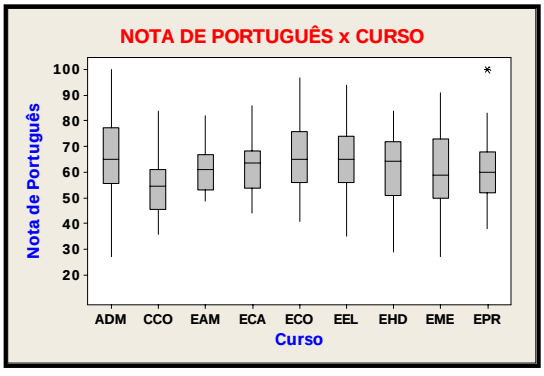
- *Outlier* superior = Q3 – 1.5D
- *Outlier* inferior = Q1 – 1.5D

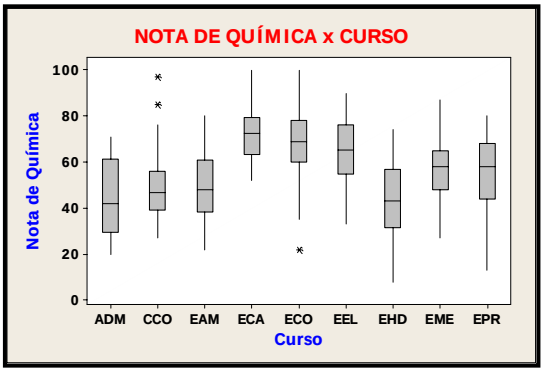
Onde: **D = (Q3-Q1)**, chamado de Intervalo Interquartil

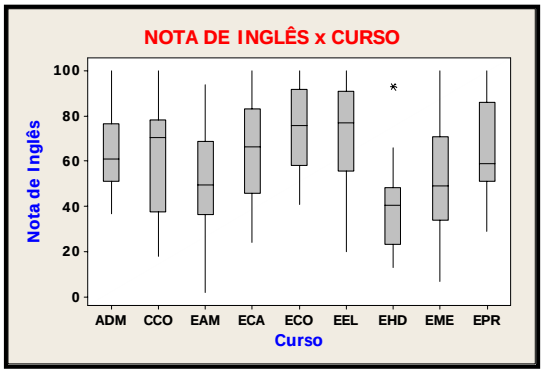
A seguir são apresentados os Boxplots com o resumo das notas de todas as disciplinas e cursos do Vestibular do ano de 2000.

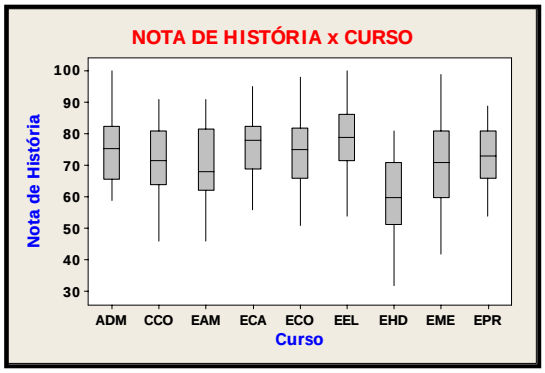


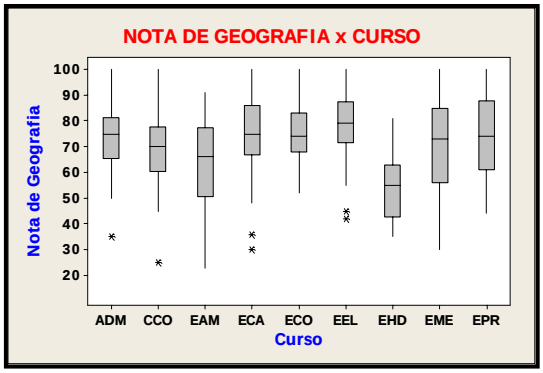


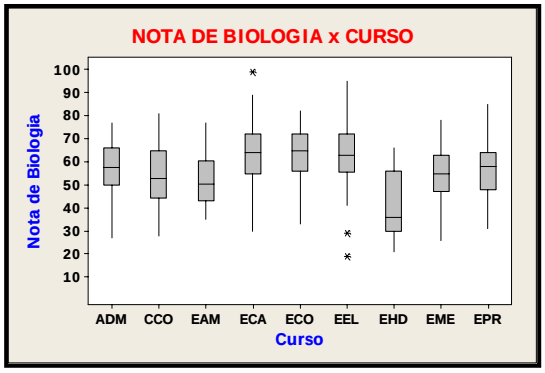


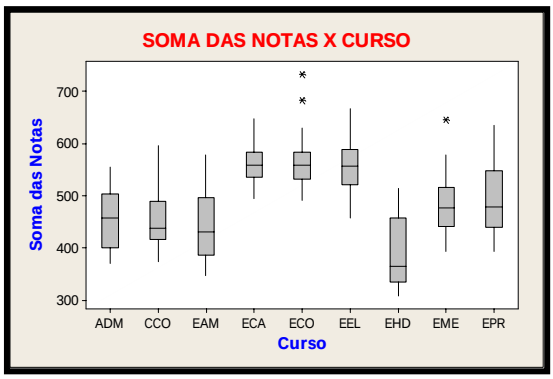












Os cursos ECA/ECO/EEL mantiveram-se regulares em todas as disciplinas em relação aos demais cursos. Destaque para as disciplinas de Matemática, Física e Química. A maioria dos cursos apresentou desempenhos parecidos nas disciplinas de Português, História, Geografia e Biologia. O curso de ADM não obteve bom desempenho em Matemática, Física e Química, porém destacou-se nas disciplinas de Inglês, Português, História e Geografia. O curso EHD em relação aos demais cursos apresentou baixo desempenho nas disciplinas de Biologia, Geografia, História e Inglês.

De uma forma geral, obtiveram o mesmo desempenho os cursos de:

- ECA/ECO/EEL – alto desempenho.
- ADM/CCO/EAM
- EME/EPR

- EHD - baixo desempenho.

Para uma interpretação mais clara, será analisado na *Figura 4.2* o boxplot *"Nota de Matemática x Curso"*.

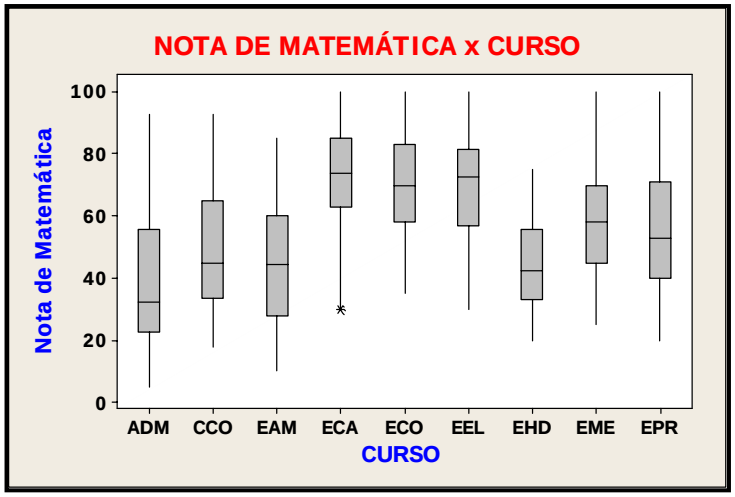


Figura 4.2 - Boxplot – Nota De Matemática

A *Tabela 4.3* mostra o resumo dos 5 valores apresentados pelo boxplot para as notas de Matemática obtidas por cada curso.

Nota de Matemática x Curso									
Medidas	**ADM**	**CCO**	**EAM**	**ECA**	**ECO**	**EEL**	**EHD**	**EME**	**EPR**
Valor Mínimo	5,0	18,0	10,0	30,0	35,0	30,0	22,0	25,0	20,0
Q1	22,5	33,75	28,0	63,0	58,0	56,75	33,0	45,0	40,0
Mediana	32,5	45,0	44,5	74,0	70,0	72,50	42,50	58,0	53,0
Q3	55,75	65,0	60,0	85,0	83,0	81,50	55,75	70,0	71,0
Valor Máximo	93,0	93,0	85,0	100	100	100	75,0	100	100
Outlier Inferior	-	-	-	30,00	-	-	-	-	-
Outlier Superior	-	-	-	-	-	-	-	-	-

Tabela 4.3 – Valores Boxplot Nota de Matemática x Curso

Tomando como exemplo o curso EPR, os resultados são interpretados da seguinte forma:

- Nota mínima: 20,0.
- Q1: 25% das notas de Matemática para o curso EPR estão abaixo de 40,0 e 75% estão acima de 40,0.
- Q2 ou mediana: 50% das notas estão abaixo de 53,0 e 50% estão acima deste valor.
- Q3: 25% das notas estão acima de 71,0 e 75% estão abaixo deste valor.
- Nota máxima: 100,0
- Outlier: não possui.

4.2.1.2– Situação Final expressa por Gráficos de Setores

Através de Gráficos de Setores, utilizados para ilustrar dados qualitativos em que os dados são divididos em fatias com as devidas proporções, é apresentada na *Figura 4.3* abaixo a situação final de cada curso referente ao ano de 2000 com as respectivas proporções de cada nível da variável dependente descrito no item 4.1.

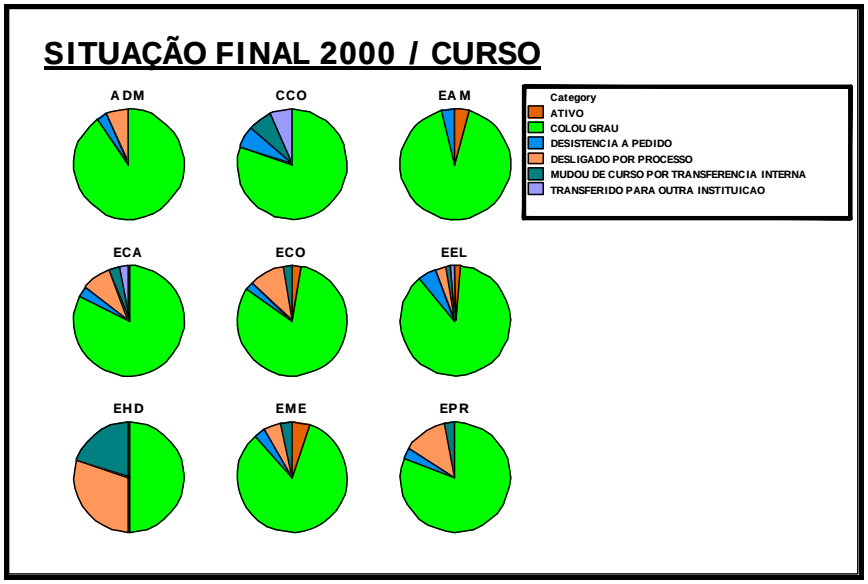


Figura 4.3 – Situação Final Por Curso

CURSOS	VARIÁVEIS					
	Ativo	Colou Grau	Desist. a ped.	Deslig. por proc.	Trans. Int.	Trans. Ext.
ADM		90,0%	3,3%	6,7%		
CCO		80,0%	6,7%		6,7%	6,6%
EAM	4,2%	91,7%	4,1%			
ECA		82,5%	2,9%	8,8%	2,9%	2,9%
ECO	2,5%	82,0%	2,5%	10,5%	2,5%	
EEL	1,5%	87,0%	5,7%	3,0%	1,4%	1,4%
EHD		50,0%		30,0%	20,0%	
EME	5,1%	83,1%	3,4%	5,0%	3,4%	
EPR		80,6%	3,2%	13,0%	3,2%	

Tabela 4.4 – Porcentagem dos níveis- Ano 2000

Considerando a totalidade de cursos, tem-se o gráfico de setores na *Figura 4.4* apresentando os seguintes valores:

- 1,8% dos alunos permanecem ativos;
- 82,5% colaram grau;
- 3,9% dos alunos tiveram desistência a pedido;
- 7,0% dos alunos foram desligados por processo;
- 3,6% dos alunos mudaram de curso por transferência interna;
- 1,2% dos alunos transferiram-se para outra instituição;

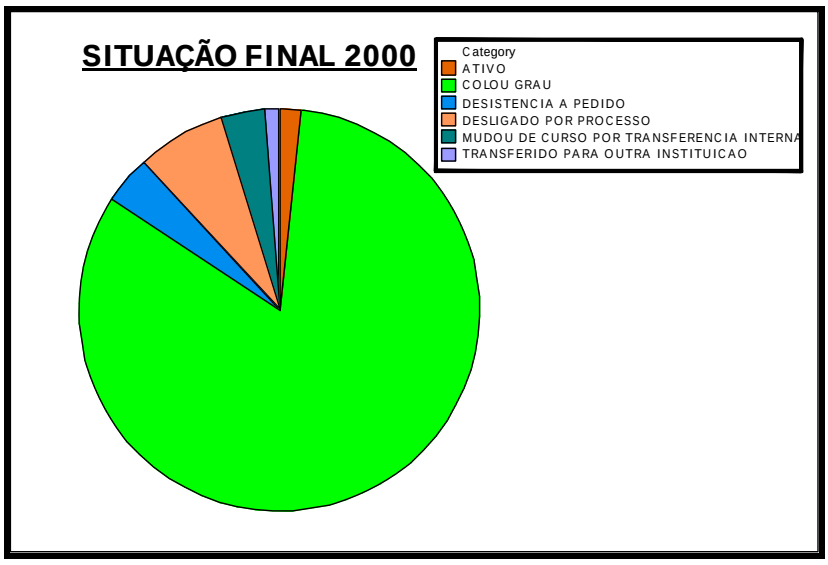


Figura 4.4 – Situação Final Total

4.2.1.3 – Coeficiente Final expresso por Boxplot

Após a análise do Coeficiente Final dos alunos ingressos em 2000, por curso, obtiveram-se os Boxplots apresentados na *Figura 4.5*. *A Tabela 4.4* mostra o resumo dos valores dos Boxplots da *Figura 4.5* e mostra a quantidade de coeficiente (N) avaliado para curso.

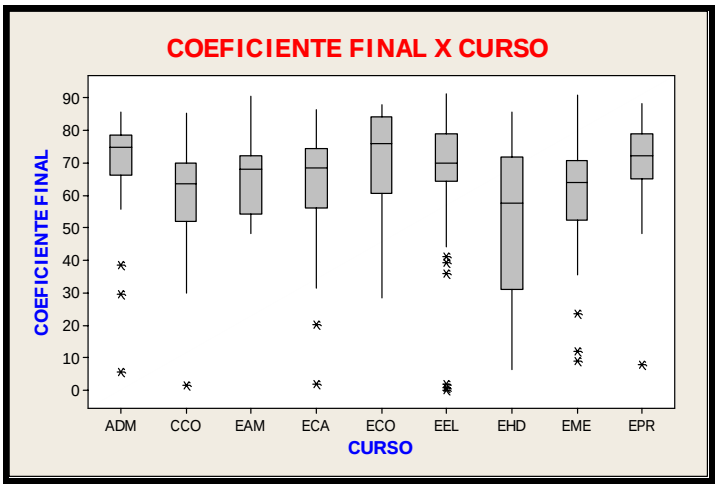


Figura 4.5 – Coeficiente Final por Curso

Curso	Média	Desvio Padrão	Q1	Mediana	Q3	Nº de Coeficientes
ADM	69.1187	17.2359	66.2700	74.900	78.8525	30
CCO	58.2627	18.0849	51.9650	63.610	69.9775	30
EAM	65.7063	11.3337	54.5250	68.240	72.5100	24
ECA	63.5506	18.5067	56.2000	68.455	74.7625	34
ECO	70.0354	17.7080	60.9300	76.220	84.3600	39
EEL	67.5340	18.3349	64.3125	70.285	79.1400	70
EHD	52.8695	22.3869	31.0425	57.915	71.8450	20
EME	61.3275	16.1954	52.5700	64.020	70.7900	59
EPR	68.2635	15.5232	65.3200	72.200	79.1800	31

Tabela 4.5 - Resumo dos valores dos Boxplots Coeficiente Final x Curso

Nota-se na *Figura 4.5*, a presença de *outliers* na maioria dos cursos. Como definido anteriormente, os *outliers* representam pontos extremos, o que significa que são pontos muito distantes dos demais valores observados. Os *outliers*, assim como podem ser valores errados no conjunto de dados, podem também revelar informações importantes. Um conjunto de dados pode conter um ou mais *outlier*. Tomando-se como exemplo o curso EPR, observa-se a presença de um *outlier* abaixo do valor mínimo. A obtenção de um coeficiente muito distante do valor mínimo pode representar uma situação de desistência a pedido, desistência por processo, uma mudança de curso ou uma transferência para outra instituição, ou até mesmo por ainda não ter concluído o curso, o que descarta a possibilidade desse ponto ser um dado errado.

4.2.1.4 – Correlação Entre as Variáveis Independentes

A correlação é a medida padronizada da relação entre duas variáveis. O grau de correlação entre as variáveis é medido pelo Coeficiente de Correlação Linear (r), também chamado de Coeficiente de Correlação Momento-Produto de Pearson. O coeficiente de correlação r deve estar entre -1 e 1.

O Coeficiente de Correlação pode ser obtido através da *Equação 4.1*.

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}} \tag{4.1}$$

Uma correlação positiva (r >0) indica que as duas variáveis movem-se juntas na mesma direção, e a correlação é forte quanto mais r se aproxima de 1. A representação gráfica da relação entre duas variáveis é feita através do diagrama de dispersão, como apresentado nas *Figuras 4.6, 4.7 e 4.8.*

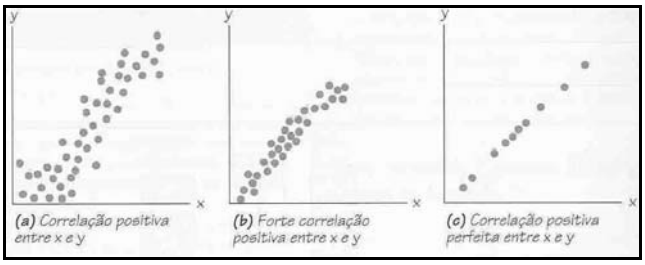


Figura 4.6 - Diagrama de Dispersão

Uma correlação negativa (r<0) indica que as duas variáveis movem-se em direções opostas, e que a correlação também fica mais forte quanto mais próxima de -1 r ficar.

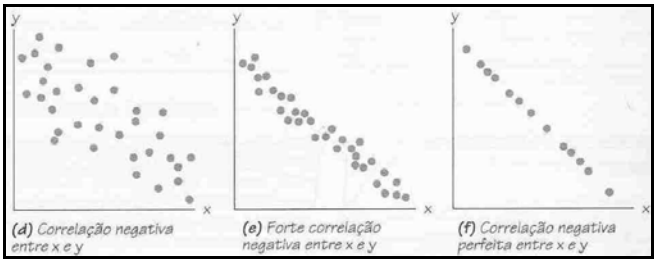


Figura 4.7 - Diagrama de Dispersão

Variáveis que estão perfeitamente correlacionadas positivamente (r = 1) movem-se essencialmente em perfeita proporção na mesma direção, enquanto dois conjuntos que estão perfeitamente correlacionados negativamente (r=-1) movem-se em perfeita proporção em direções opostas.

Se o valor de r está próximo de 0, conclui-se que não há correlação significativa entre as variáveis e se (r = 0) conclui-se que as duas variáveis não estão relacionadas entre si.

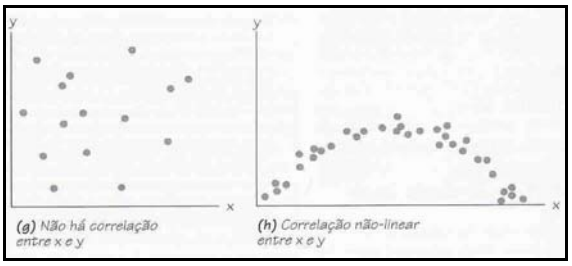


Figura 4.8 - Diagrama de Dispersão

A *Figura 4.9* apresenta os coeficientes de correlação de Pearson existentes entre as notas do vestibular e o Valor P.

O Valor P informa se o coeficiente de correlação é significativamente diferente de 0.

- Se o Valor P é <= 0,05 (nível de significância) então conclui-se que a correlação é diferente de 0.
- Se o Valor P é > 0,05 (nível de significância), então não conclui-se que a correlação é diferente de 0.

	MAT	FIS	POR	QUI	ING	HIS	GEO
FIS	0,610 0,000						
POR	0,008 0,885	-0,067 0,219					
QUI	0,409 0,000	0,496 0,000	0,072 0,189				
ING	0,074 0,174	0,070 0,197	0,086 0,114	0,051 0,351			
HIS	0,093 0,087	0,026 0,634	0,286 0,000	0,167 0,002	0,210 0,000		
GEO	0,092 0,090	0,101 0,064	0,139 0,010	0,219 0,000	0,223 0,000	0,427 0,000	
BIO	0,193 0,000	0,229 0,000	0,222 0,000	0,338 0,000	0,164 0,003	0,343 0,000	0,239 0,000

Cell Contents: Pearson correlation
P-Value

Figura 4.9 – Correlação entre as disciplinas

Apresenta-se também a existência de uma forte correlação positiva r = 0,610 entre as variáveis Física e Matemática. Ou seja, um aumento na nota de Matemática ou Física é, na grande maioria dos casos, acompanhado de um aumento na nota da disciplina correlacionada. O Valor P desta relação é 0,000, comprovando que a correlação existe. Percebe-se também que existe uma correlação, porém não muito forte entre as variáveis Química/Matemática r = 0,409 / Química/Física r = 0,496 / Geografia/História r = 0,427, todos com Valor P 0,000, comprovando a existência das correlações. Todas as correlações entre as variáveis são apresentadas por gráficos através da *Figura 4.10.*

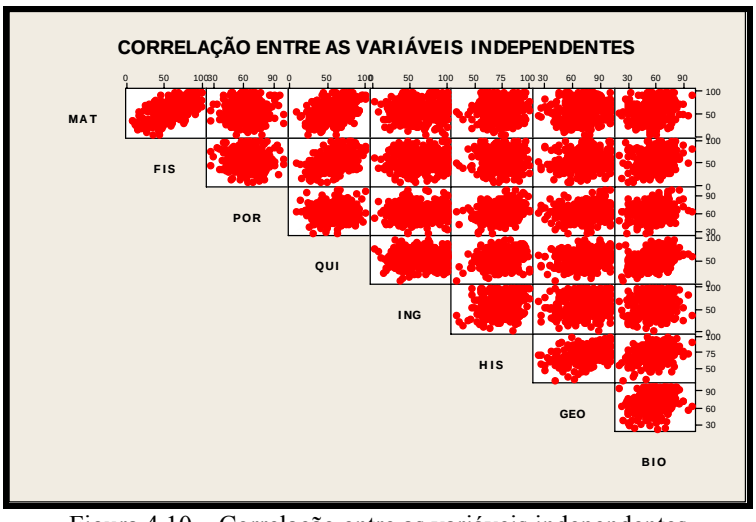


Figura 4.10 – Correlação entre as variáveis independentes

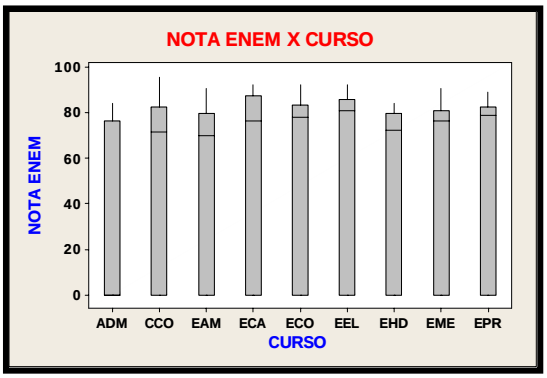
Na *Figura 4.10* é possível perceber que os pontos entre as variáveis Matemática/Física, possuem o formato de uma reta comprovando a forte correlação existente entre elas.

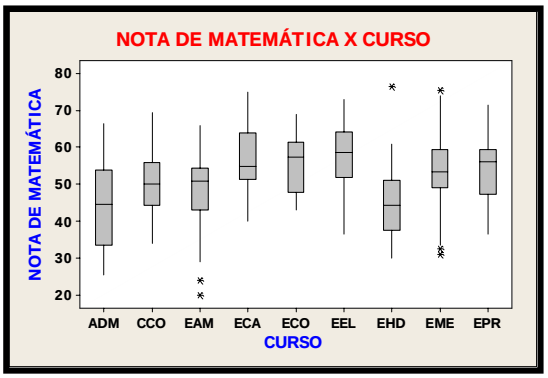
Para os anos de 2001 a 2003 os resultados são apresentados a seguir, obtidos de forma análoga à apresentada anteriormente para o ano de 2000.

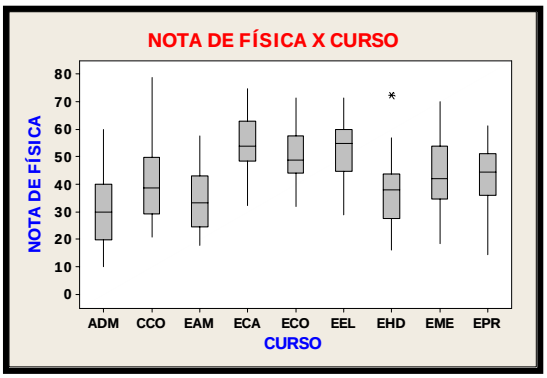
4.2.2 Caracterização para os dados dos alunos ingressos em 2001

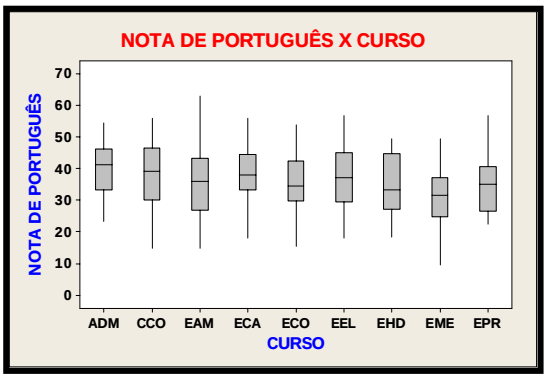
4.2.2.1 Dados Expressos por Boxplot

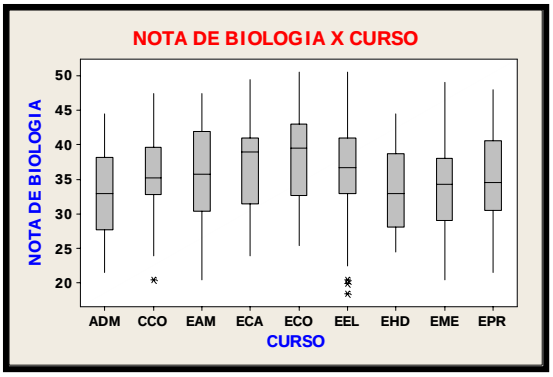
Nos vestibulares de 2001 à 2003 foi também considerada a nota obtida no ENEM pelo aluno. Os dados serão apresentados da mesma forma em que foram representados dos dados do ano de 2000.

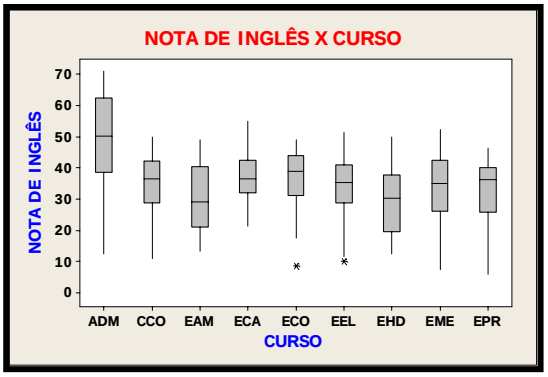


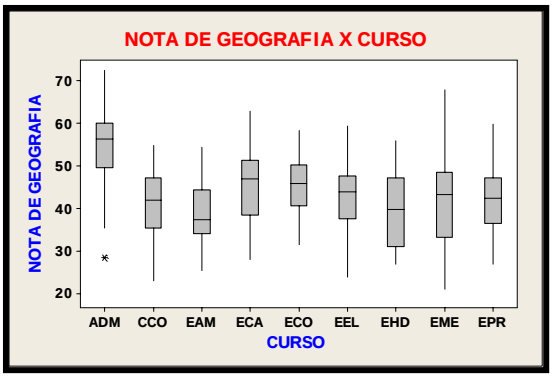


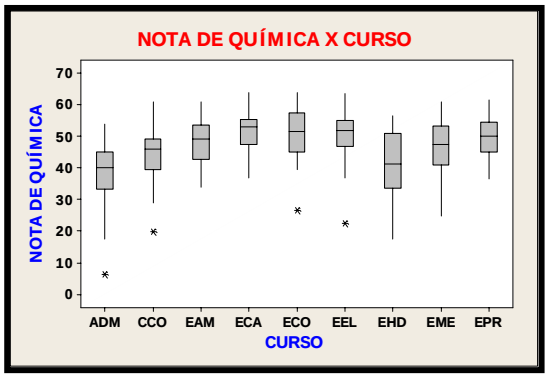


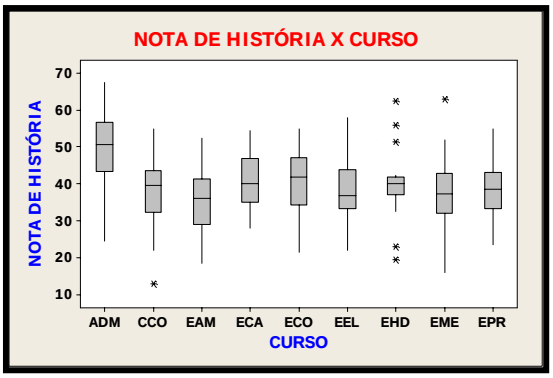


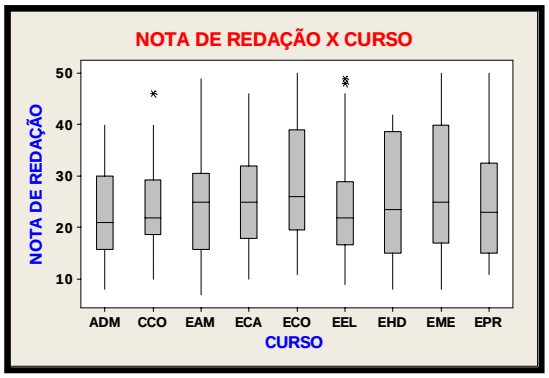


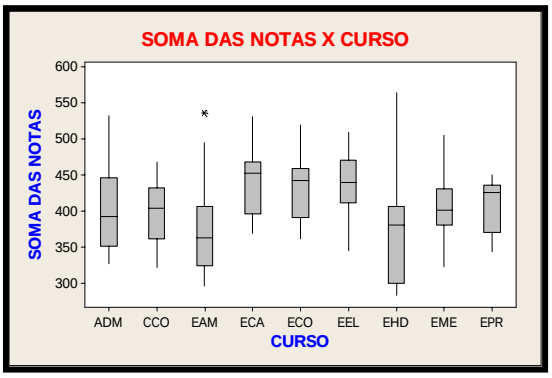












Os cursos ECA/ECO/EEL mantiveram-se regulares em todas as disciplinas em relação aos demais cursos. Destaque para as disciplinas de Matemática, Física e Química. A maioria dos cursos apresentou desempenhos parecidos nas notas do ENEM e nas disciplinas de Português, Biologia, Inglês e Redação. O curso de ADM não obteve bom desempenho em Física, porém destacou-se nas disciplinas de Inglês, História e Geografia.

De uma forma geral, obtiveram o mesmo desempenho os cursos de:

- ECA/ECO/EEL – alto desempenho.
- ADM/CCO
- EHD - baixo desempenho.

4.2.2.2 – Situação Final expressa por Gráficos de Setores

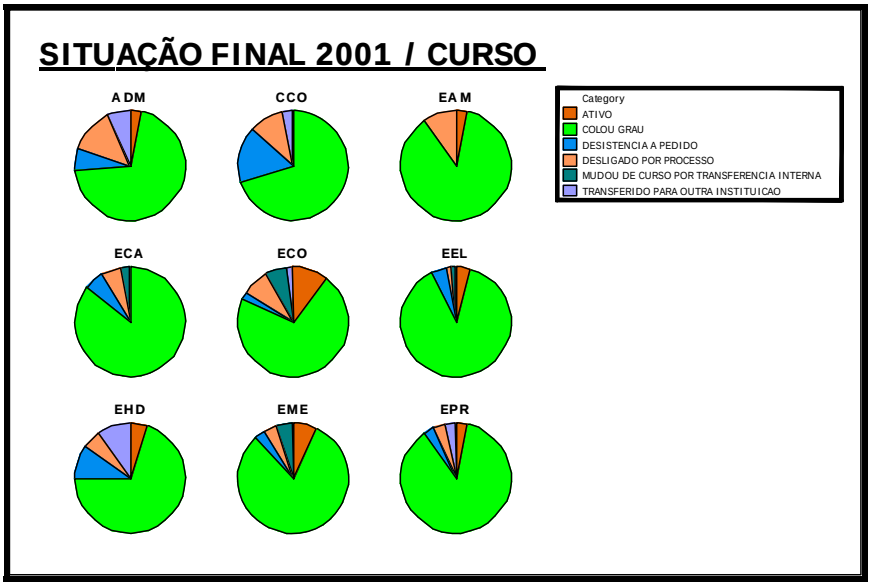


Figura 4.11 – Situação Final Por Curso

CURSOS	VARIÁVEIS					
	Ativo	Colou Grau	Desist. a ped.	Deslig por proc.	Trans. Int.	Trans. Ext.
ADM	3,3%	70,0%	6,7%	13,3%		6,7%
CCO		70,0%	16,7%	10,0%		3,3%
EAM	3,3%	86,7%		10,0%		
ECA		85,7%	5,7%	5,7%	2,9%	
ECO	10,2%	71,5%	2,0%	8,2%	6,1%	2,0%
EEL	4,3%	88,6%	4,3%	1,4%	1,4%	
EHD	5,0%	70,0%	10,0%	5,0%		10,0%
EME	7,0%	81,0%	3,4%	3,4%	5,2%	
EPR	3,3%	86,8%	3,3%	3,3%		3,3%

Tabela 4.6 – Porcentagem dos níveis- Ano 2001

Considerando a totalidade de cursos, tem-se o gráfico de setores na *Figura 4.12* apresentando os seguintes valores:

- 4,5 % dos alunos permanecem ativos;
- 80,0 % colaram grau;
- 5,0 % dos alunos tiveram desistência a pedido;
- 6,0 % dos alunos foram desligados por processo;
- 2,3 % dos alunos mudaram de curso por transferência interna;
- 2,2 % dos alunos transferiram-se para outra instituição;

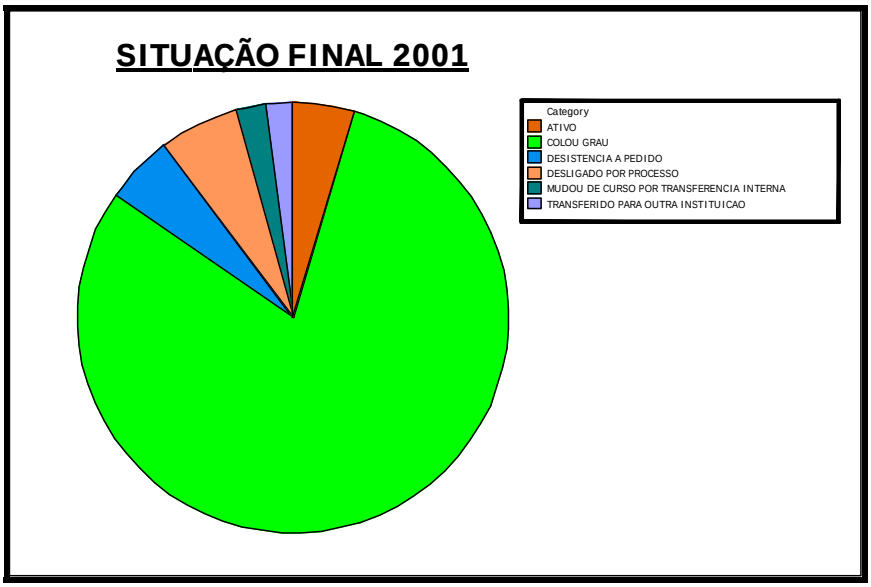


Figura 4.12 – Situação Final Total

4.2.2.3 – Coeficiente Final expresso por Boxplot

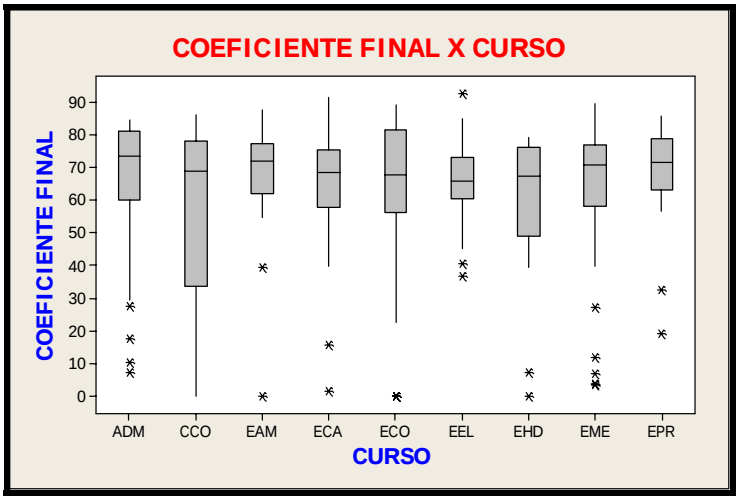


Figura 4.13 – Coeficiente Final por Curso

A *Tabela 4.5* mostra o resumo dos valores dos Boxplots da *Figura 4.13*.

Curso	Média	DesvPad	Q1	Mediana	Q3	N de Coeficientes
ADM	65,399	23,306	60,223	73,620	81,035	30
CCO	56,410	28,646	33,788	68,820	78,223	30
EAM	67,866	16,693	62,150	72,085	77,470	30
ECA	65,007	17,727	57,670	68,490	75,530	35
ECO	64,383	22,106	56,255	67,890	81,515	49
EEL	65,947	10,886	60,590	66,005	73,128	70
EHD	58,975	22,810	49,033	67,480	76,273	20
EME	64,909	19,805	58,198	70,895	89,800	58
EPR	69,496	14,576	63,038	71,660	78,775	30

Tabela 4.7 - Resumo dos valores dos Boxplots Coeficiente Final x Curso

4.2.2.4 – Correlação Entre as Variáveis

	MAT	FIS	POR	QUI	ING	BIO	GEO	HIS
FIS	0,430 0,000							
POR	-0,067 0,208	-0,009 0,861						
QUI	0,202 0,000	0,374 0,000	0,062 0,243					
ING	-0,077 0,150	-0,105 0,050	0,043 0,420	-0,126 0,018				
BIO	0,009 0,860	0,157 0,003	0,035 0,510	0,166 0,002	-0,026 0,628			
GEO	-0,063 0,235	0,027 0,620	0,052 0,327	-0,043 0,423	0,184 0,001	0,059 0,269		
HIS	-0,117 0,028	-0,030 0,572	0,144 0,007	-0,108 0,044	0,065 0,222	0,076 0,157	0,575 0,000	
RED	-0,037 0,489	-0,050 0,350	-0,085 0,111	-0,033 0,536	-0,013 0,806	-0,078 0,143	-0,002 0,978	-0,069 0,199

Cell Contents: Pearson correlation
P-Value

Figura 4.14 – Correlação entre as disciplinas

A *Figura 4.14* mostra a existência de uma forte correlação positiva r = 0,575 entre as variáveis História e Geografia. O Valor P desta relação é 0,000, comprovando a existência da correlação. A *Figura 4.14* também apresenta a existência das correlações positivas, mas não muito fortes entre as variáveis Física e Matemática r = 0,374 e Química/Física, ambas com Valor P 0,000. Como dito anteriormente, com Valor P maior que 0,05 e muito alto, a correlação entre as variáveis quase não existe, um exemplo é a correlação quase inexistente entre Biologia/Matemática, com r= 0,009 e Valor P 0,860.

Todas as correlações entre as variáveis são apresentadas por gráficos através da *Figura 4.15*.

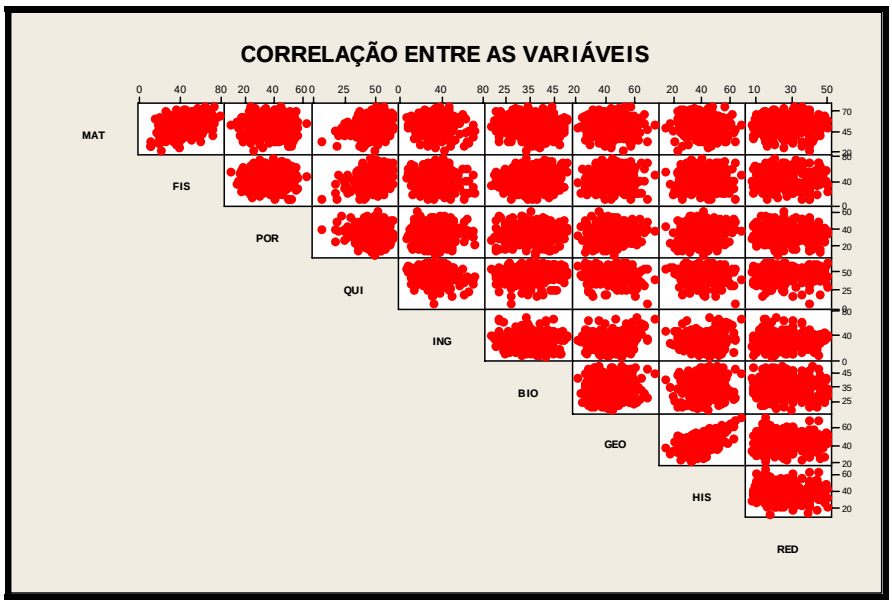
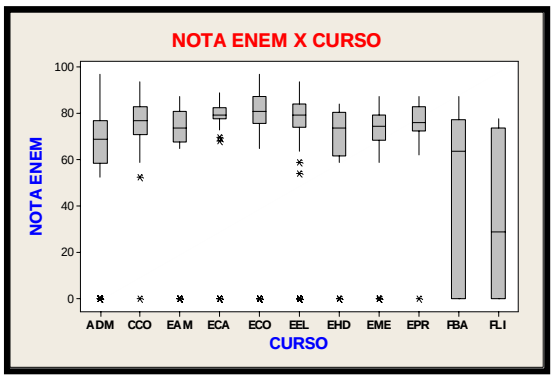


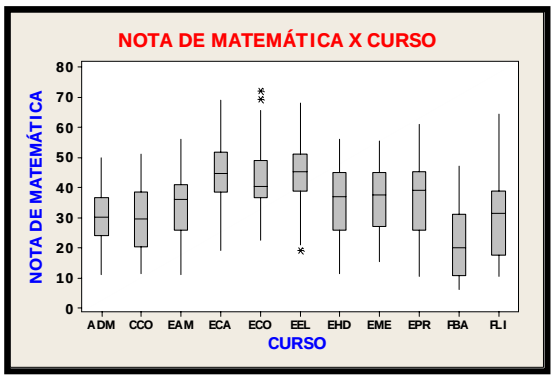
Figura 4.15 – Correlação entre as variáveis independentes

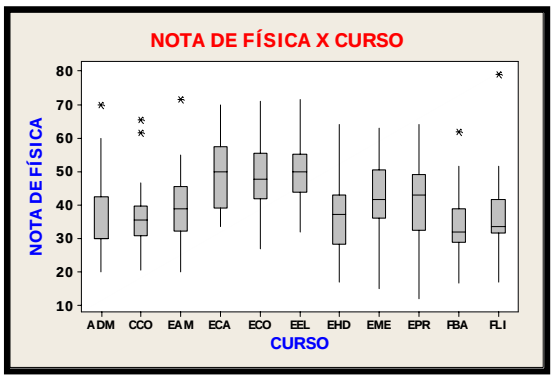
Na *Figura 4.15* é possível perceber que os pontos entre as variáveis História/Geografia, possuem o formato de uma reta comprovando a forte correlação existente entre elas.

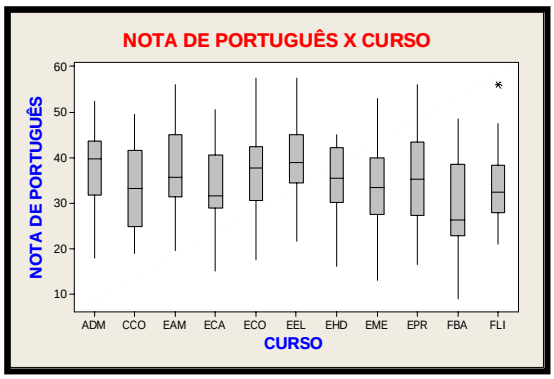
4.2.3 - Caracterização para os dados dos alunos ingressos em 2002

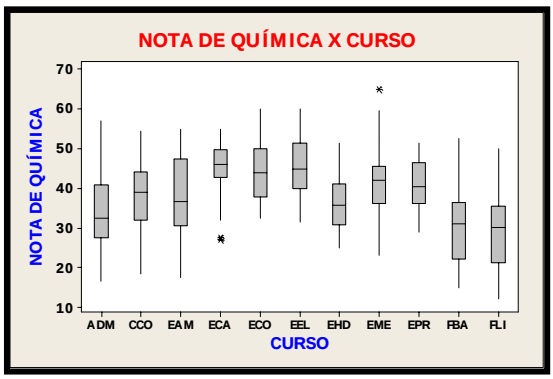
4.2.3.1 - Dados Expressos por Boxplot

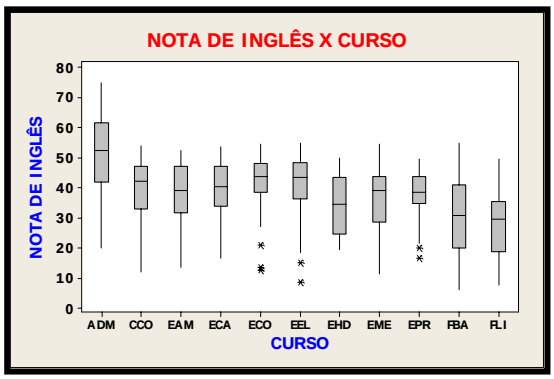


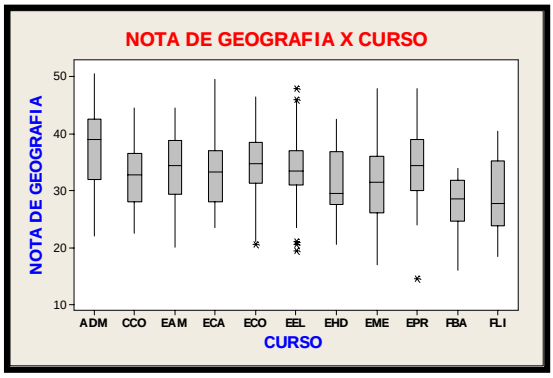


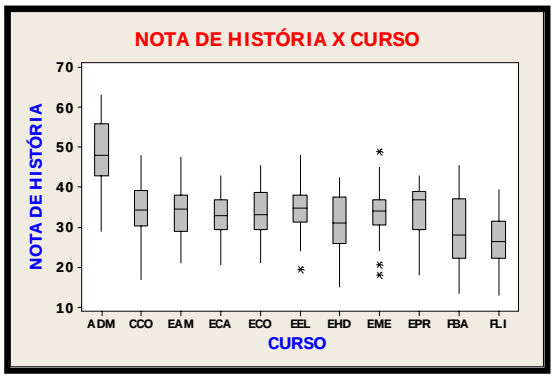


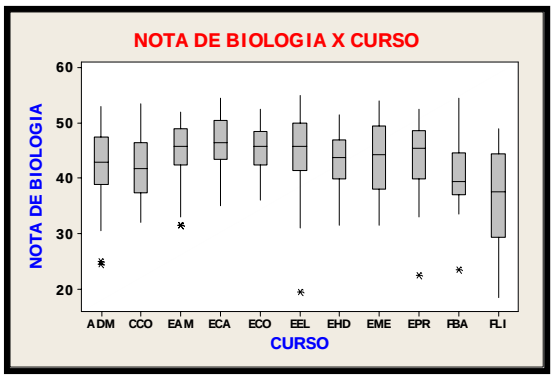


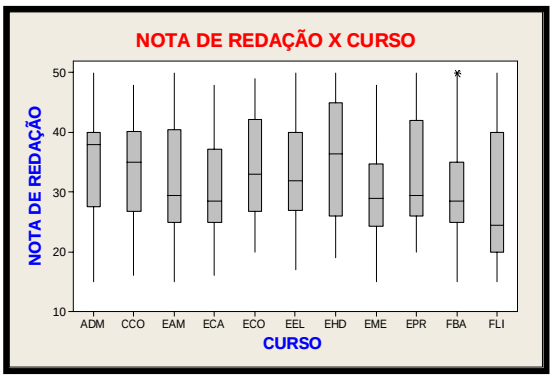


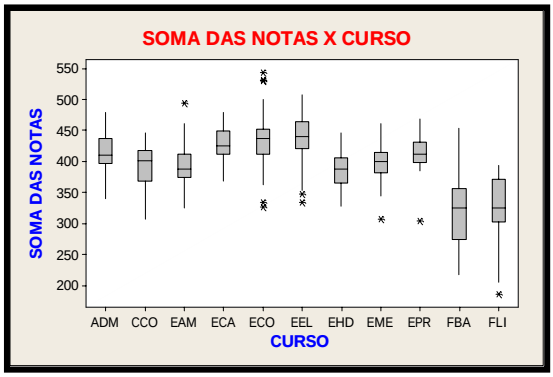












Os cursos ECA/ECO/EEL mantiveram-se regulares em todas as disciplinas em relação aos demais cursos. Destaque para as disciplinas de Matemática, Física e Química. A maioria dos cursos apresentou desempenhos parecidos nas disciplinas de Português, Inglês, História, Geografia, Biologia e Redação. Com relação às notas do ENEM, os cursos de FBA/FLI apresentaram baixo desempenho em relação aos demais cursos. O curso de ADM não obteve bom desempenho em Matemática, Física e Química, porém destacou-se nas disciplinas de Inglês, História e Geografia. O curso EHD em relação aos demais cursos apresentou baixo desempenho nas disciplinas de Biologia, Geografia, História e Inglês.

De uma forma geral, obtiveram o mesmo desempenho os cursos de:

- ECA/ECO/EEL – alto desempenho.
- FBA/FLI - baixo desempenho.

4.2.3.2 – Situação Final expressa por Gráficos de Setores

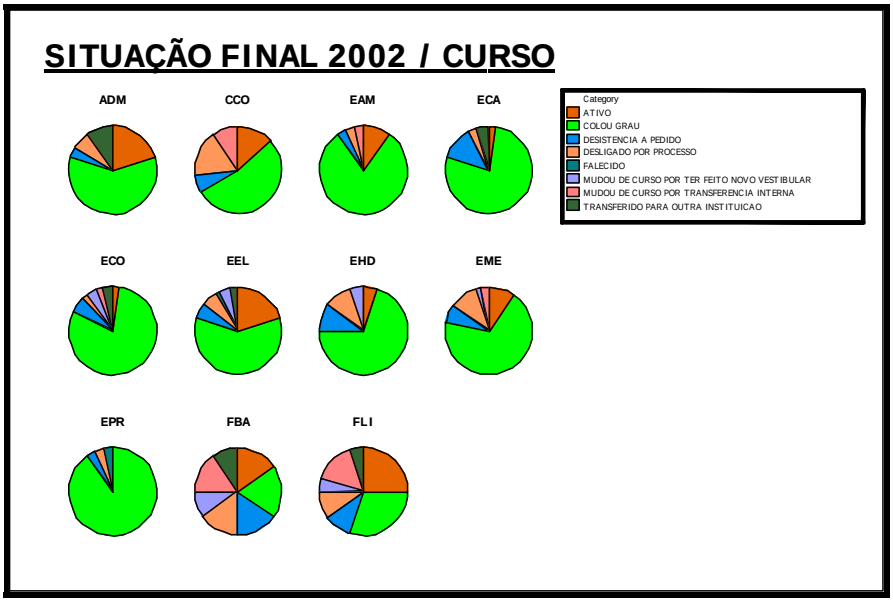


Figura 4.16 – Situação Final Por Curso

CURSOS	VARIÁVEIS							
	Ativo	Colou Grau	Desist. a ped.	Deslig. por proc.	Falecido	Novo Vest.	Trans. Int.	Trans. Ext.
ADM	20,0%	60,0%	3,3%	6,7%				10,0%
CCO	13,3%	53,3%	6,7%	16,7%			10,0%	
EAM	10,0%	80,0%	3,4%	3,3%			3,3%	
ECA	2,5%	77,5%	12,5%	2,5%				5,0%
ECO	2,0%	80,0%	6,0%	2,0%		4,0%	2,0%	4,0%
EEL	20,0%	60,0%	5,7%	5,7%	1,4%	4,3%		2,9%
EHD	5,0%	70,0%	10,0%	10,0%		5,0%		
EME	10,0%	68,3%	6,7%	10,0%		1,7%	3,3%	
EPR		90,0%	3,4%	3,3%	3,3%			
FBA	15,0%	20,0%	15,0%	15,0%		10,0%	15,0%	10,0%
FLI	25,0%	30,0%	10,0%	10,0%		5,0%	15,0%	5,0%

Tabela 4.8 – Porcentagem dos níveis- Ano 2002

Considerando a totalidade de cursos, tem-se o gráfico de setores na *Figura 4.17* apresentando os seguintes valores:

- 11,0 % dos alunos permanecem ativos;
- 65,8 % colaram grau;
- 7,0 % dos alunos tiveram desistência a pedido;
- 7,0 % dos alunos foram desligados por processo;
- 0,5 % dos alunos faleceram;
- 2,5 % dos alunos mudaram de curso por terem feito novo vestibular;
- 3,2 % dos alunos mudaram de curso por transferência interna;
- 3,0 % dos alunos transferiram-se para outra instituição;

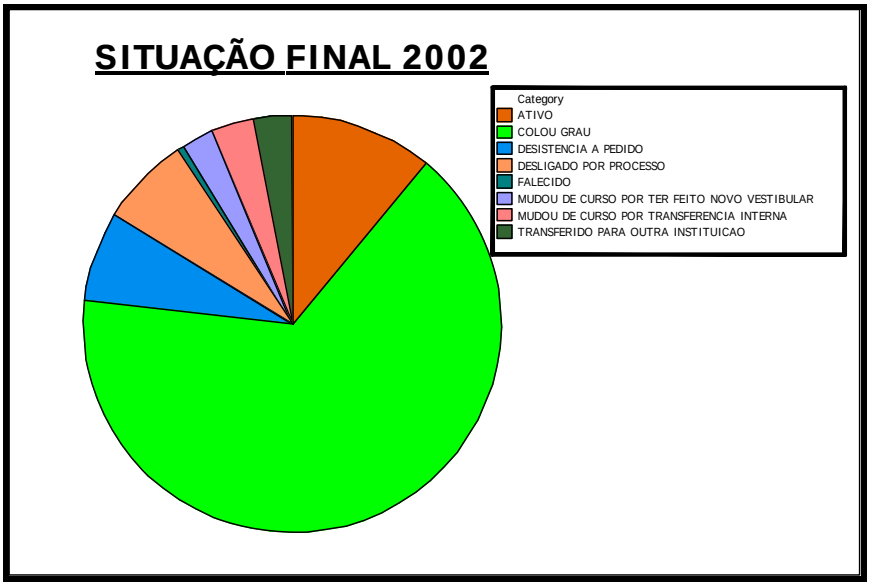


Figura 4.17 – Situação Final Total

4.2.3.3 – Coeficiente Final expresso por Boxplot

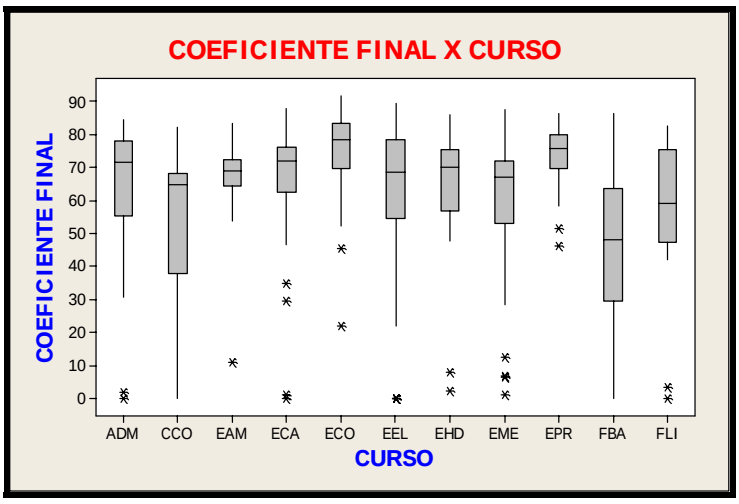


Figura 4.18 – Coeficiente Final por Curso

A *Tabela 4.6* mostra o resumo dos valores dos Boxplots da *Figura 4.18*.

Curso	Média	DesvPad	Q1	Mediana	Q3	N de Coeficientes
ADM	63,748	20,931	55,400	71,675	77,945	30
CCO	55,258	21,425	37,848	64,885	68,228	30
EAM	67,242	12,626	64,365	68,835	72,325	30
ECA	65,655	19,396	62,680	72,005	76,050	40
ECO	74,840	13,325	69,753	78,380	83,358	50
EEL	63,563	21,211	54,383	68,615	78,388	70
EHD	62,244	21,708	56,930	69,980	75,545	20
EME	60,216	20,155	52,935	66,945	71,988	60
EPR	73,309	9,648	69,895	75,900	80,070	30
FBA	47,586	21,738	29,628	48,250	63,570	20
FLI	57,149	23,201	47,140	59,015	75,410	20

Tabela 4.9 - Resumo dos valores dos Boxplots Coeficiente Final x Curso

4.2.3.4 – Correlação Entre as Variáveis

	MAT	FIS	POR	QUI	ING	BIO	GEO	HIS
FIS	0,492 0,000							
POR	0,003 0,959	-0,021 0,670						
QUI	0,374 0,000	0,376 0,000	0,024 0,638					
ING	-0,006 0,905	0,005 0,915	0,252 0,000	0,014 0,780				
BIO	0,122 0,015	0,175 0,000	0,150 0,003	0,260 0,000	0,159 0,001			
GEO	-0,020 0,685	0,022 0,656	0,198 0,000	0,108 0,030	0,140 0,005	0,215 0,000		
HIS	-0,075 0,133	-0,027 0,597	0,200 0,000	0,032 0,526	0,294 0,000	0,146 0,003	0,344 0,000	
RED	-0,089 0,075	-0,094 0,060	0,170 0,001	-0,074 0,142	0,152 0,002	0,063 0,208	0,018 0,718	0,039 0,438

Cell Contents: Pearson correlation
P-Value

Figura 4.19 - Correlação entre as disciplinas

A *Figura 4.19* mostra a existência de correlações entre as variáveis:

Forte correlação:

- Física/Matemática r: 0,492 / Valor P: 0,000
- Química/Matemática r: 0,374 / Valor P: 0,000
- Química/Física: r: 0,376 / Valor P: 0,000

Fraca correlação:

- Português/ Matemática: r:0,003 / Valor P: 0,959

Todas as correlações entre as variáveis são apresentadas por gráficos através d*a Figura 4.20.*

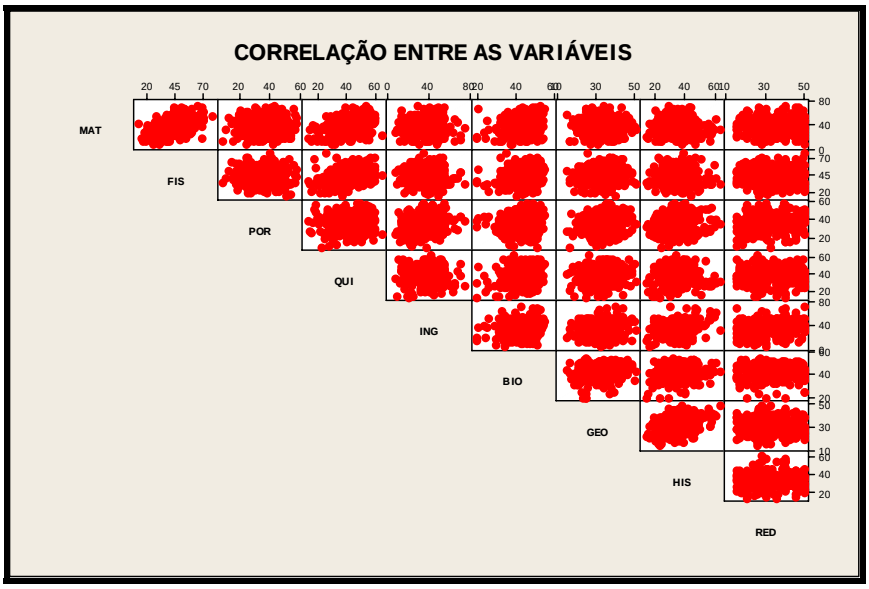
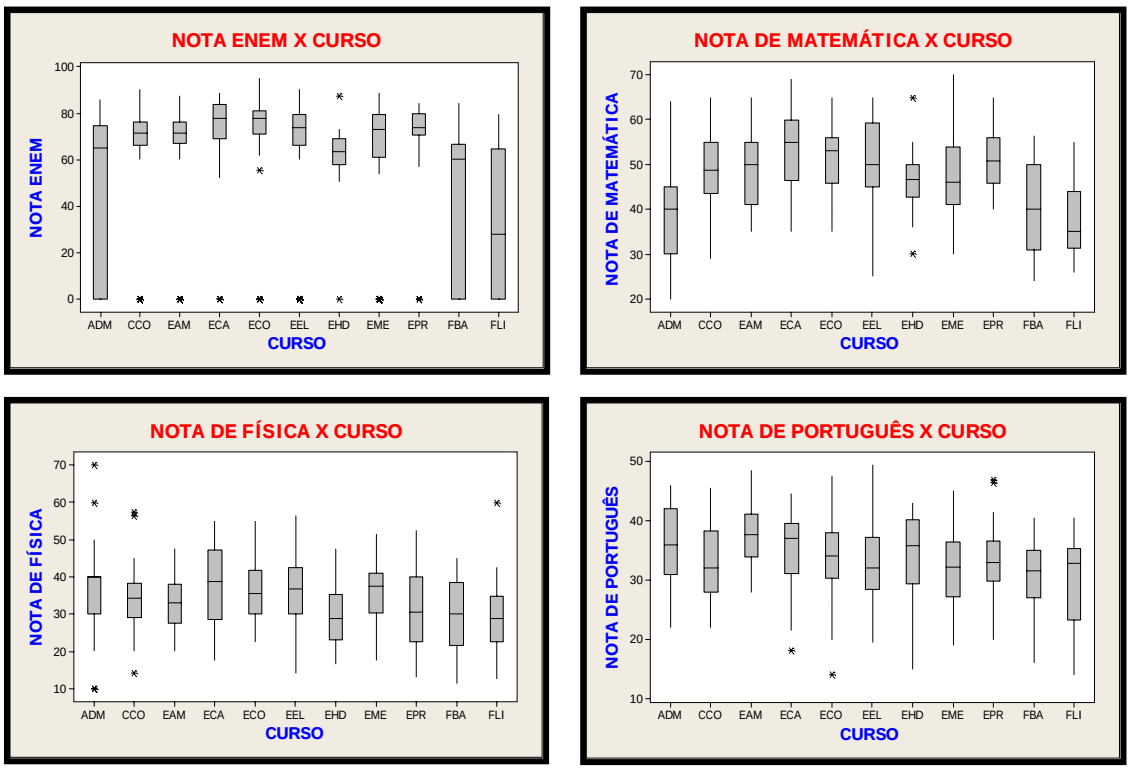


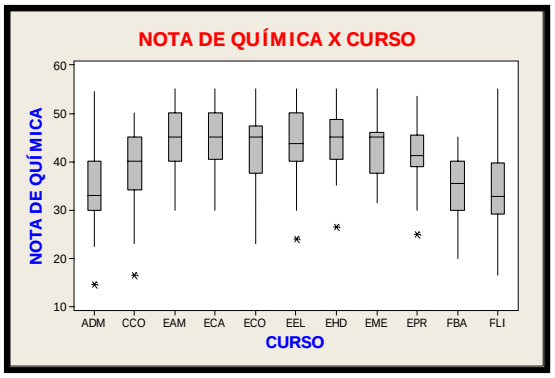
Figura 4.20 – Correlação entre as variáveis independentes

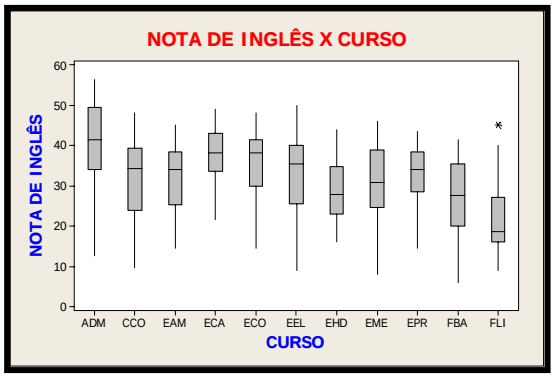
Através do gráfico é possível perceber que os pontos entre as variáveis Física/Matemática, possuem o formato de uma reta comprovando a forte correlação existente entre elas.

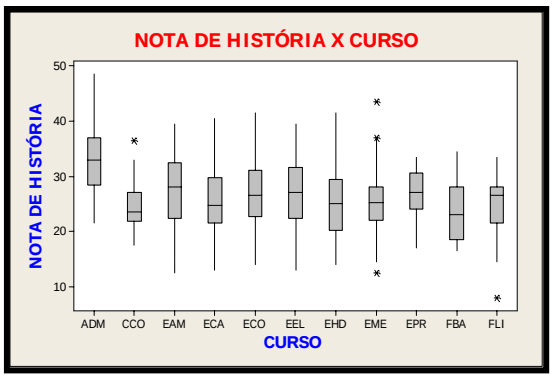
4.2.4 Caracterização para os dados dos alunos ingressos em 2003

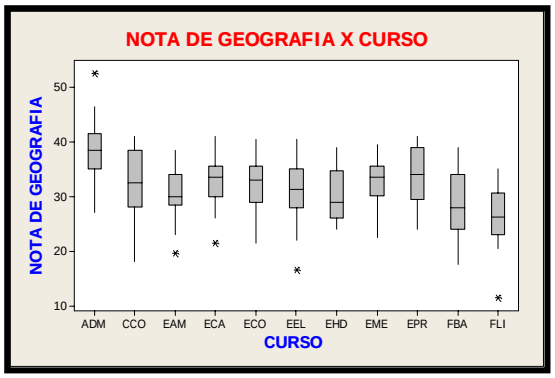
4.2.4.1 – Dados Expressos por Boxplot

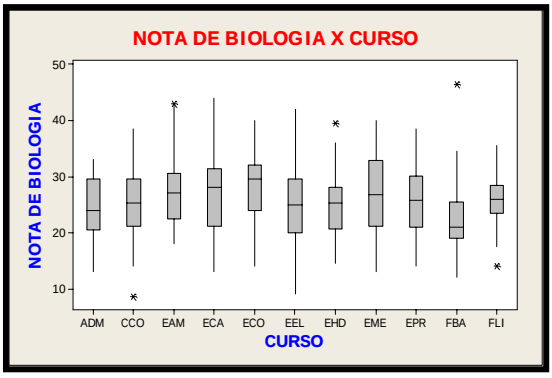


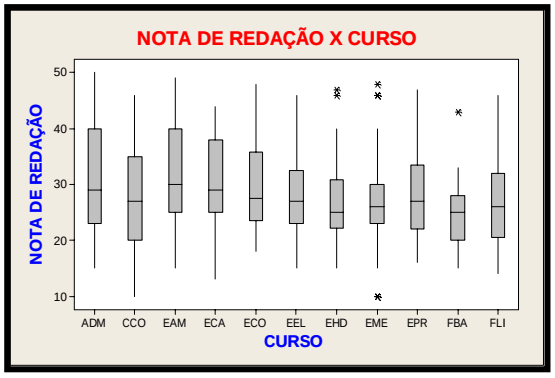


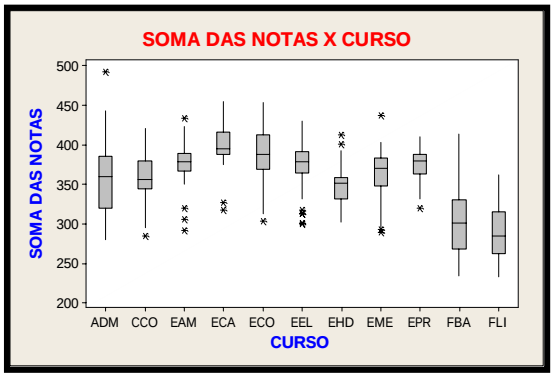












Os cursos de uma forma geral mantiveram-se regulares em todas as disciplinas, destaque para o curso ECA. Com relação às notas do ENEM, os cursos de ECA/ECO/EEL/EPR apresentaram bom desempenho em relação aos demais cursos. O curso de ADM não obteve bom desempenho em Matemática, Física e Química, porém destacou-se nas disciplinas de Inglês, História e Geografia.

De uma forma geral, obtiveram o mesmo desempenho os cursos de:

- ECA/ECO – alto desempenho.
- FBA/FLI - baixo desempenho.

4.2.4.2 Situação Final expressa por Gráficos de Setores

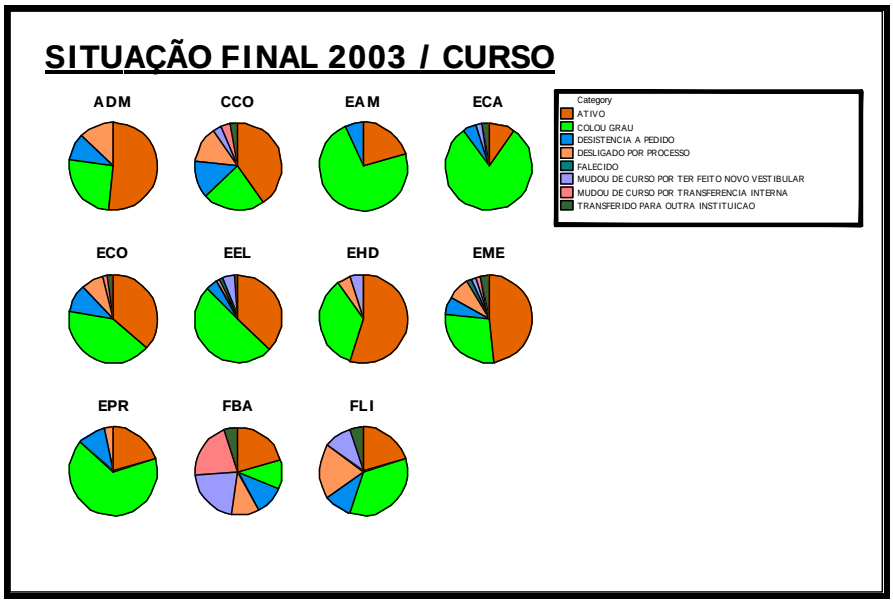


Figura 4.21 – Situação Final Por Curso

CURSOS	VARIÁVEIS							
	Ativo	Colou Grau	Desist. a pedido	Deslig. por Proc.	Falecido	Novo Vest.	Trans. Int.	Transf. Ext.
ADM	51,6%	25,8%	9,7%	12,9%				
CCO	40,0%	23,3%	13,4%	13,4%		3,3%	3,3%	3,3%
EAM	20,0%	73,3%	6,7%					
ECA	10,0%	80,0%	5,0%			2,5%		2,5%
ECO	36,0%	42,0%	10,0%	8,0%			2,0%	2,0%
EEL	37,1%	50,0%	4,3%	1,4%	1,4%	4,4%		1,4%
EHD	55,0%	35,0%		5,0%		5,0%		
EME	48,3%	28,3%	6,7%	8,3%	1,7%	1,7%	1,7%	3,3%
EPR	20,0%	66,7%	10,0%	3,3%				
FBA	21,1%	10,5%	10,5%	10,5%		21,0%	21,1%	5,3%
FLI	20,0%	35,0%	10,0%	20,0%		10,0%		5,0%

Tabela 4.10 – Porcentagem dos níveis- Ano 2003

Considerando a totalidade de cursos, tem-se o gráfico de setores na *Figura 4.22* apresentando os seguintes valores:

- 34 % dos alunos permanecem ativos;
- 44,5 % colaram grau;
- 7,5 % dos alunos tiveram desistência a pedido;
- 6,5 % dos alunos foram desligados por processo;
- 0,5 % dos alunos faleceram;
- 3,3 % dos alunos mudaram de curso por terem feito novo vestibular;
- 1,7 % dos alunos mudaram de curso por transferência interna;
- 2,0 % dos alunos transferiram-se para outra instituição.

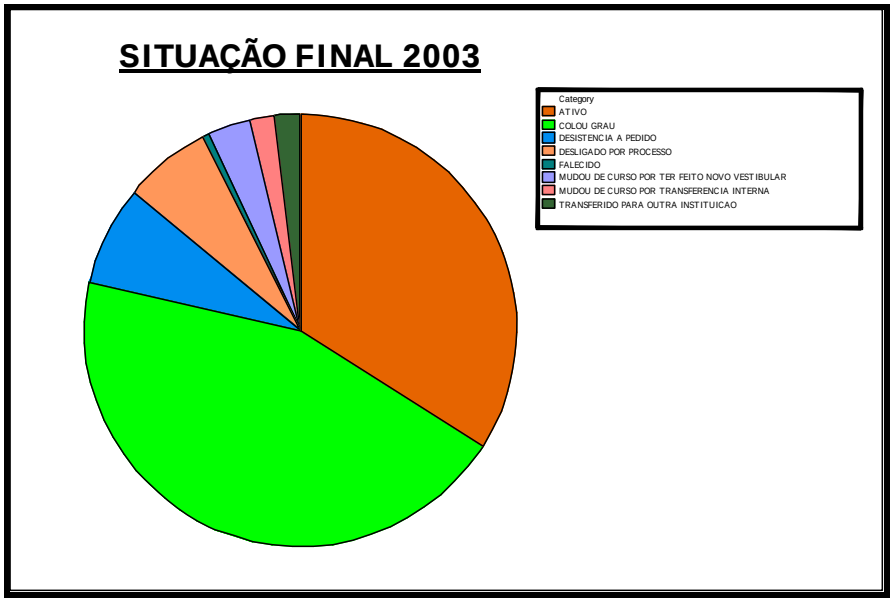


Figura 4.22 – Situação Final Total

4.2.4.3 – Coeficiente Final expresso por Boxplot

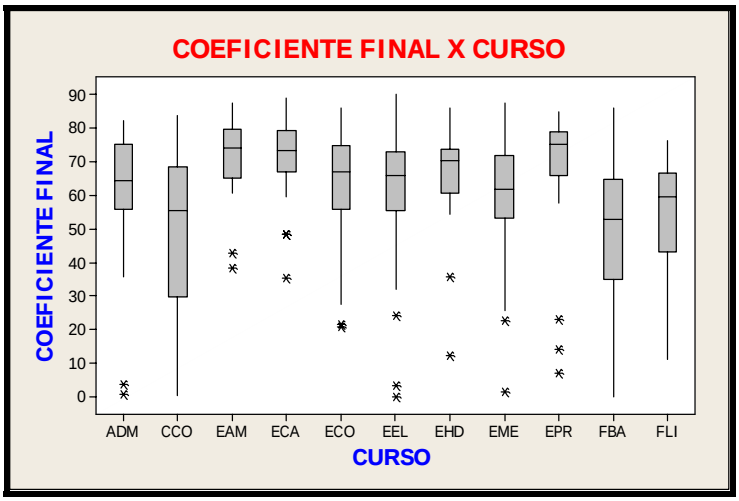


Figura 4.23 – Coeficiente Final Por Curso

A *Tabela 4.7* mostra o resumo dos valores dos Boxplots da *Figura 4.23.*

Curso	Média	DesvPad	Q1	Mediana	Q3	N de Coeficientes
ADM	61,852	19,390	55,950	64,310	75,170	31
CCO	50,898	23,509	29,913	55,400	68,533	30
EAM	71,753	11,126	65,128	74,375	79,855	30
ECA	71,817	11,037	66,995	73,640	79,315	40
ECO	63,497	15,836	55,888	67,295	74,900	50
EEL	62,681	16,390	55,688	66,070	73,080	70
EHD	66,028	17,152	60,915	70,350	86,250	20
EME	59,591	16,885	53,135	61,865	71,955	60
EPR	68,397	19,729	66,155	75,445	79,045	30
FBA	49,710	21,218	35,120	52,890	64,990	19
FLI	55,954	16,078	43,065	59,525	66,595	20

Tabela 4.11 – Resumo dos valores dos Boxplots Coeficiente Final x Curso

4.2.4.4 – Correlação Entre as Variáveis

	MAT	FIS	POR	QUI	BIO	ING	GEO	HIS
FIS	0,262 0,000							
POR	-0,016 0,752	-0,015 0,770						
QUI	0,313 0,000	0,152 0,002	0,029 0,570					
BIO	-0,009 0,865	-0,055 0,272	0,084 0,094	0,081 0,105				
ING	0,019 0,703	-0,027 0,586	0,151 0,002	0,004 0,933	0,001 0,980			
GEO	-0,041 0,410	-0,012 0,804	0,166 0,001	0,004 0,935	-0,010 0,839	0,130 0,009		
HIS	0,008 0,869	-0,032 0,529	0,113 0,024	-0,027 0,589	0,060 0,234	0,090 0,073	0,199 0,000	
RED	-0,024 0,638	-0,015 0,761	0,162 0,001	-0,031 0,538	0,022 0,665	-0,007 0,890	0,053 0,294	0,111 0,027

Cell Contents: Pearson correlation
P-Value

Figura 4.24 – Correlação entre as disciplinas

A *Figura 4.24* mostra a existência de correlações entre as variáveis:

Correlação não muito forte:

- Física / Matemática r: 0,262 / Valor P : 0,000
- Química/Matemática r: 0,313 / Valor P: 0,000

Fraca correlação:

- Inglês/ Química r:0,004 / Valor P: 0,933

Todas as correlações entre as variáveis são apresentadas por gráficos através da *Figura 4.25.*

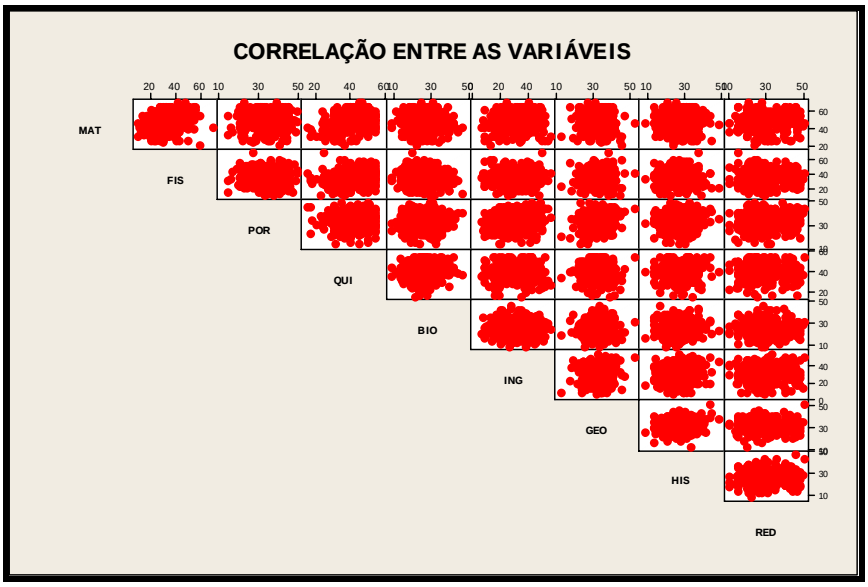
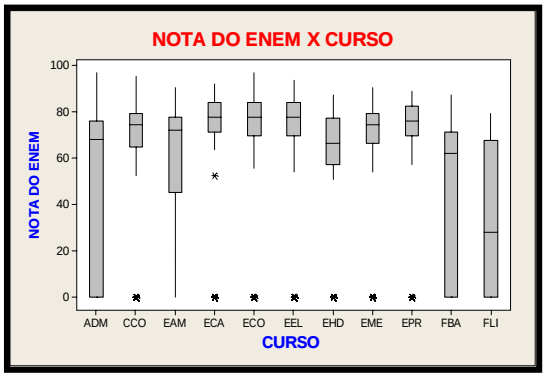


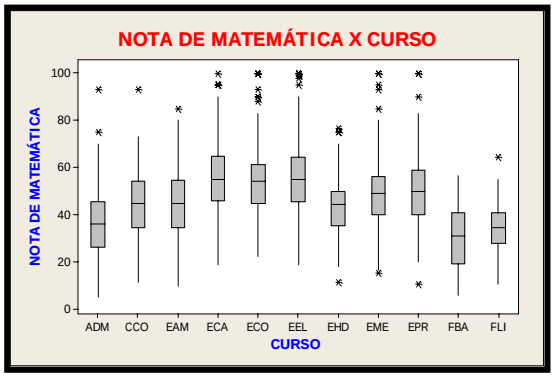
Figura 4.25 – Correlação entre as variáveis independentes

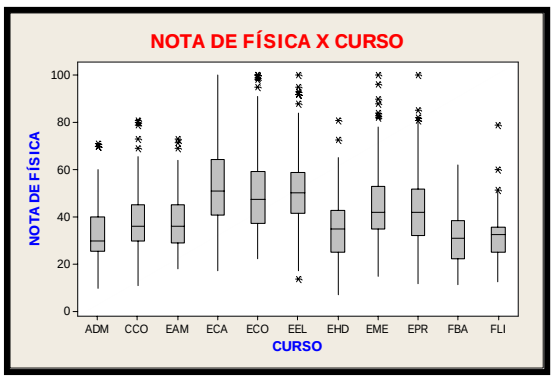
4.2.5 Caracterização para os dados dos alunos ingressos de 2000 a 2003

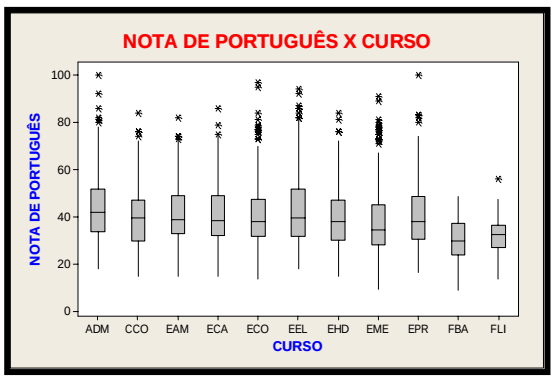
Abaixo apresenta-se a análise realizada para todo o banco de dados, envolvendo as notas dos vestibulares no período de 2000 à 2003.

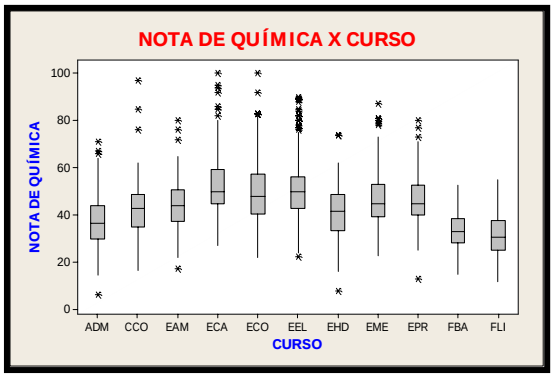
4.2.5.1 - Dados Expressos por Boxplot

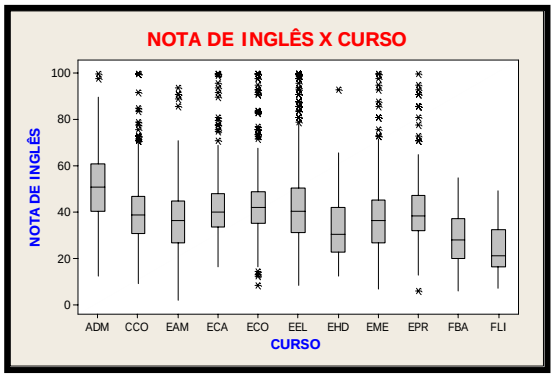


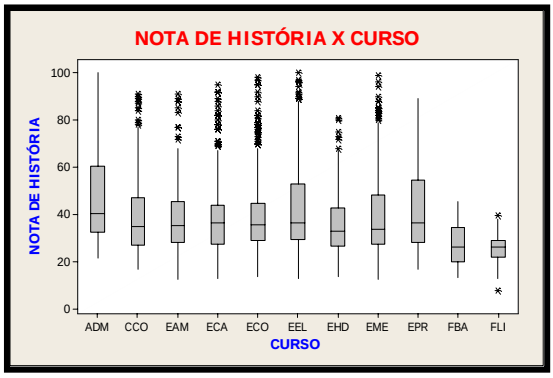


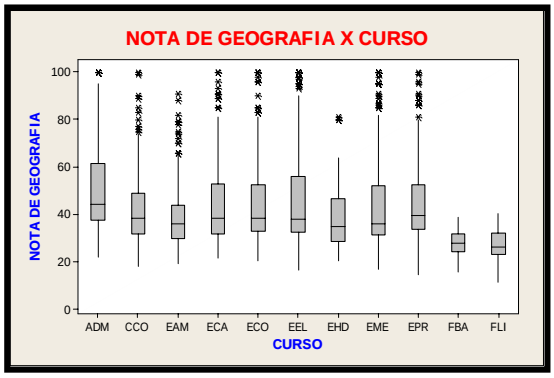


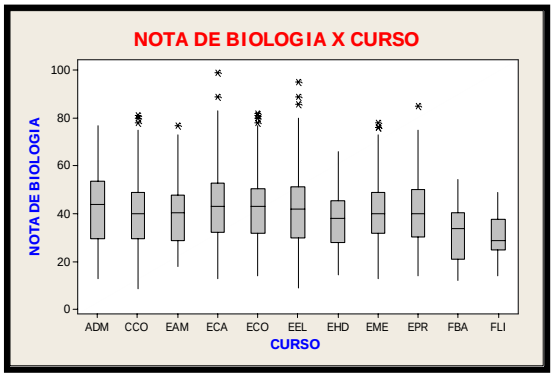


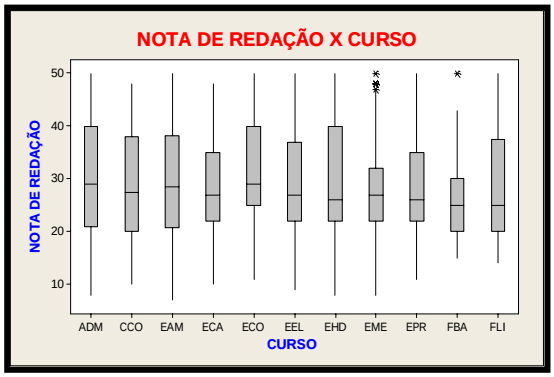


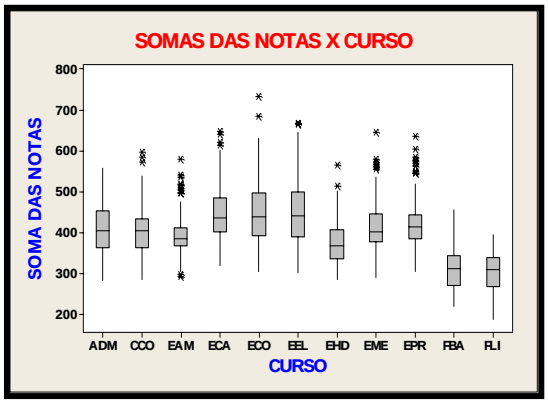












As notas dos cursos referentes ao ENEM foram elevadas. Os cursos de ECA/ECO/EEL apresentaram bom desempenho em relação aos demais cursos nas disciplinas de Matemática, Física e Química. De uma forma geral os cursos obtiveram desempenhos parecidos nas disciplinas Redação, Biologia, História, Geografia, Inglês e Português. O curso de ADM não obteve bom desempenho em Matemática, Física e Química, porém destacou-se

nas disciplinas de Inglês, História e Geografia. Os cursos de FBA/FLI não tiveram desempenho favorável em relação aos demais nas disciplinas História e Geografia.

De uma forma geral, os cursos obtiveram desempenhos parecidos. Destaque para:

- EEL – alto desempenho.
- FBA/FLI - baixo desempenho.

4.2.5.2 Situação Final expressa por Gráficos de Setores

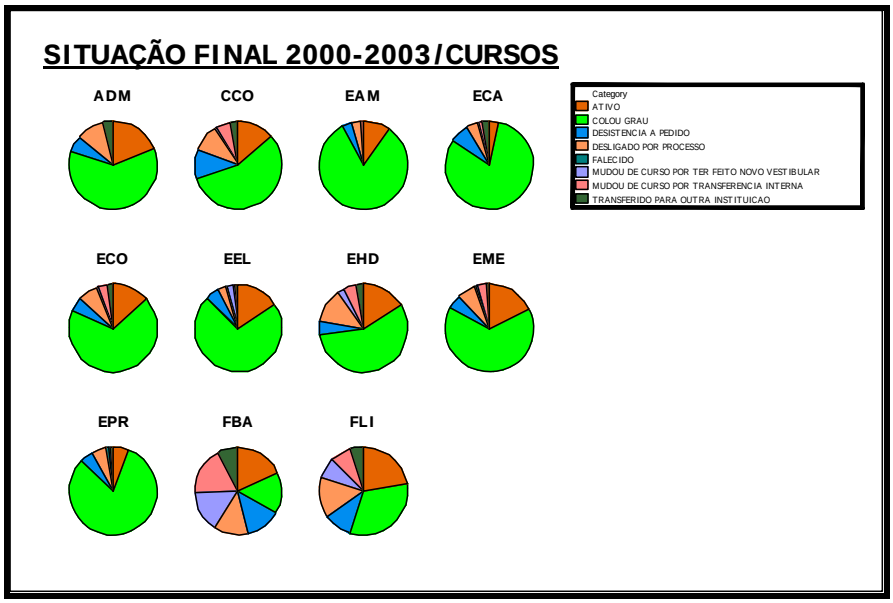


Figura 4.26 – Situação Final Por Curso

CURSOS	VARIÁVEIS							
	Ativo	Colou Grau	Desist. a Ped.	Deslig. por proc.	Falecido	Novo Vest.	Trans. Int.	Trans. Ext.
ADM	19,0%	61,2%	5,8%	9,9%				4,1%
CCO	13,3%	56,8%	10,8%	10,0%		0,8%	5,0%	3,3%
EAM	9,6%	82,5%	3,5%	3,5%			0,9%	
ECA	3,5%	81,2%	7,0%	4,0%			1,3%	3,0%
ECO	13,3%	68,1%	5,3%	6,9%		1,1%	3,2%	2,1%
EEL	15,7%	71,4%	5,0%	3,0%	0,7%	2,1%	0,7%	1,4%
EHD	16,3%	56,2%	5,0%	12,5%		2,5%	5,0%	2,5%
EME	17,7%	65,0%	5,1%	6,8%	0,4%	0,8%	3,4%	0,8%
EPR	5,8%	81,0%	5,0%	5,8%	0,8%		0,8%	0,8%
FBA	17,9%	15,4%	12,8%	12,8%		15,4%	18,0%	7,7%
FLI	22,5%	32,5%	10,0%	15,0%		7,5%	7,5%	5,0%

Tabela 4.12 – Porcentagem dos níveis- Ano 2000-2003

Considerando a totalidade de cursos, tem-se o gráfico de setores na *Figura 4.27* apresentando os seguintes valores:

- 13,6 % dos alunos permanecem ativos;
- 67,2 % dos alunos colaram grau;
- 6,0 % dos alunos tiveram desistência a pedido;
- 6,6 % dos alunos foram desligados por processo;
- 0,3 % dos alunos faleceram;
- 1,5 % dos alunos mudaram de curso por terem feito novo vestibular;
- 2,7 % dos alunos mudaram de curso por transferência interna;
- 2,1 % dos alunos transferiram-se para outra instituição;

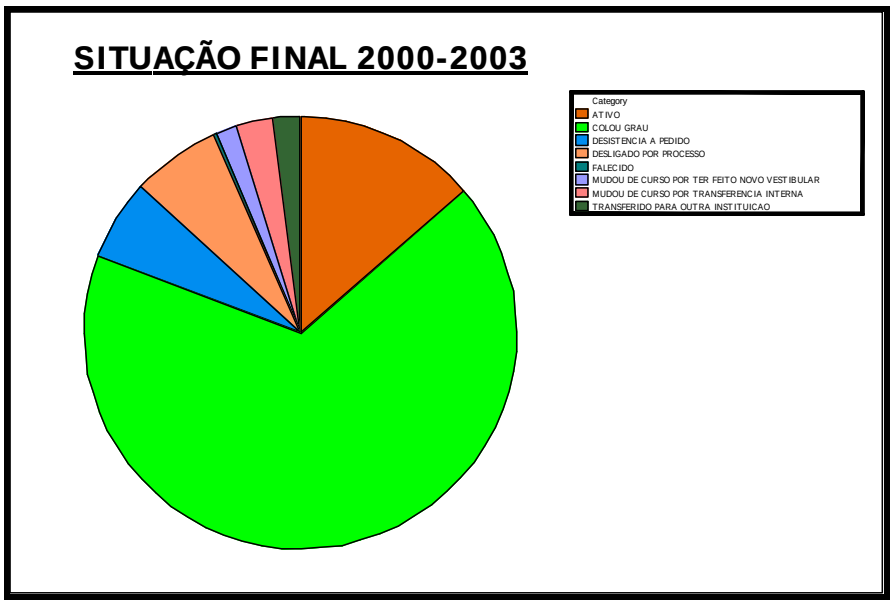


Figura – 4.27 – Situação Final Total

4.2.5.3 4.2.5.3 – Coeficiente Final expresso por Boxplot

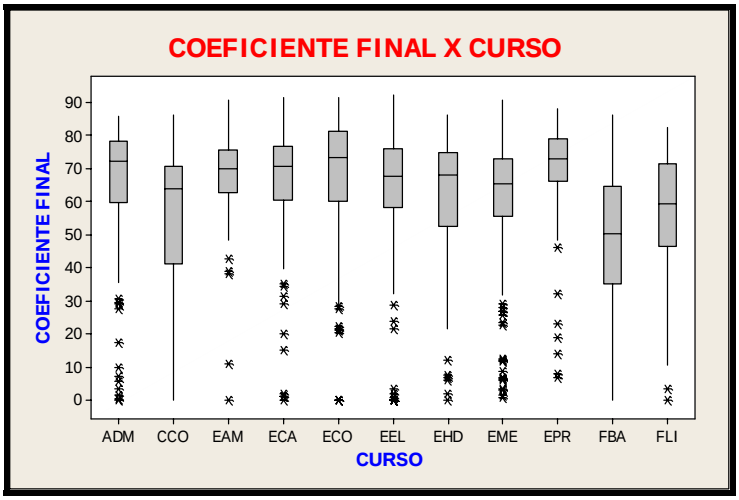


Figura 4.28 - Coeficiente Final por Curso

A *Tabela 4.8* mostra o resumo dos valores dos Boxplots da *Figura 4.28.*

Curso	Média	Desvio Padrão	Q1	Mediana	Q3	N de Coeficientes
ADM	65,003	20,252	59,750	72,380	78,330	121
CCO	55,207	23,101	41,288	64,180	70,750	120
EAM	68,270	13,239	62,790	69,990	75,640	114
ECA	66,677	17,033	60,655	71,000	77,080	149
ECO	68,101	17,989	60,275	73,420	81,300	188
EEL	64,931	17,142	58,285	67,860	76,263	280
EHD	60,029	21,292	52,770	68,145	74,938	80
EME	61,483	18,337	55,670	65,680	73,085	237
EPR	69,853	15,243	66,305	72,970	79,060	121
FBA	48,621	21,229	35,120	50,430	64,650	39
FLI	56,551	19,712	46,553	59,525	71,763	40

Tabela 4.13 – Resumo dos valores dos Boxplots Coeficiente Final x Curso

4.2.5.4 Correlação Entre as Variáveis

	MAT	FIS	POR	QUI	ING	HIS	GEO	BIO
FIS	0,520							
	0,000							
POR	0,253	0,314						
	0,000	0,000						
QUI	0,494	0,529	0,391					
	0,000	0,000	0,000					
ING	0,195	0,285	0,530	0,293				
	0,000	0,000	0,000	0,000				
HIS	0,306	0,426	0,742	0,485	0,612			
	0,000	0,000	0,000	0,000	0,000			
GEO	0,351	0,406	0,674	0,502	0,571	0,817		
	0,000	0,000	0,000	0,000	0,000	0,000		
BIO	0,191	0,433	0,572	0,424	0,490	0,687	0,635	
	0,000	0,000	0,000	0,000	0,000	0,000	0,000	
RED	-0,191	-0,057	0,068	-0,134	0,090	0,011	-0,121	0,060
	0,000	0,052	0,021	0,000	0,002	0,708	0,000	0,041

Cell Contents: Pearson correlation
P-Value

Figura 4.29 – Correlação entre as disciplinas

A *Figura 4.29* mostra a existência de correlações entre as variáveis:

Correlação muito forte:

- Geografia/História r: 0,817 / Valor P : 0,000
- História/Português r: 0,742 / Valor P: 0,000
- Geografia/Português r: 674 / Valor P: 0,000

Fraca correlação:

- Redação/ História r:0,011 / Valor P: 0,708

Todas as correlações entre as variáveis são apresentadas por gráficos através da *Figura 4.30.*

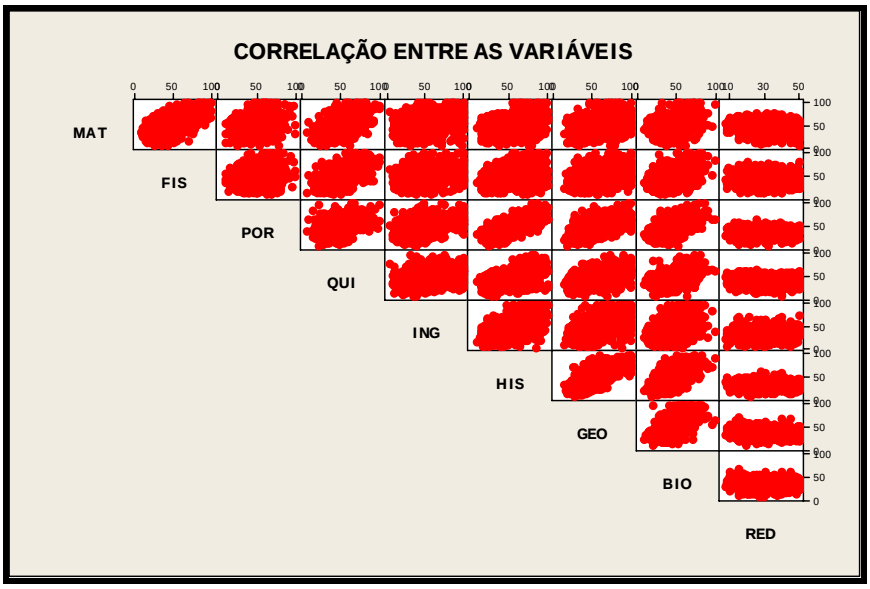


Figura 4.30 – Correlação entre as variáveis independentes

Através do gráfico é possível perceber que os pontos entre as variáveis Português/História e História/Geografia possuem o formato de uma reta comprovando a forte correlação existente entre elas.

4.3 Análise de Regressão para o Coeficiente Final

Após visualizar a correlação entre as variáveis independentes através dos diagramas de dispersão, busca-se uma relação funcional entre as variáveis independentes e a variação dependente do problema. Admitindo um relacionamento linear entre as notas obtidas no vestibular e no ENEM e o coeficiente final do aluno, tem-se um problema de regressão linear. Considerando as notas do ENEM e de todas as disciplinas como variáveis independentes o problema é de regressão linear múltipla.

A análise dos resultados tem como principais componentes o ***Valor P, valor dos coeficientes, o Coeficiente de Determinação R^2 e O Coeficiente de Determinação ajustado R^2***. Conforme mostrado no Capítulo 3, o Valor P testa a hipótese nula de que os coeficientes são iguais a zero versus estes sendo diferentes de zero. Quando o Valor P for <= ao nível de significância 0,05, a hipótese nula será rejeitada, quando o Valor P for > ao nível de significância 0,05 , a hipótese nula não será rejeitada.

- H_0: todos os coeficientes são iguais a zero (não há regressão);
- H_1: pelo menos um dos coeficientes é diferente de zero (há regressão).

O *Coeficiente de Determinação* R^2 mede o grau de ajustamento da equação de regressão aos dados amostrais. É uma medida descritiva da proporção da variação de Y que

pode ser explicada por X. Um ajuste perfeito resultaria em R^2=1. Um ajuste muito bom acarreta um valor próximo de 1. E um ajuste fraco ocasiona um valor de R^2 próximo de zero. O Coeficiente de Determinação é uma medida da aderência da equação de regressão aos dados amostrais, mas apresenta o problema de aumentar na medida em que se incluem mais variáveis. Conseqüentemente, é melhor usar o *Coeficiente de Determinação ajustado* R^2 porque ele ajusta o valor de R^2 com base no número de variáveis e no tamanho da amostra.

Utilizando o software Minitab, chegou-se aos modelos de regressão para o Coeficiente Final mostrados nas tabelas abaixo.

ADM

The regression equation is :

COEF. = 61,7 + 0,133 ENEM + 0,009 MAT - 0,332 FIS + 0,293 POR + 0,386 QUI
+ 0,096 ING - 0,174 HIS - 0,401 GEO + 0,152 BIO - 0,166 RED

Predictor	Coef	P
Constant	61,72	0,010
ENEM	0,13310	**0,037**
MAT	0,0087	0,968
FIS	-0,3324	**0,048**
POR	0,2932	0,289
QUI	0,3857	0,114
ING	0,0963	0,554
HIS	-0,1738	0,501
GEO	-0,4007	0,204
BIO	0,1525	0,531
RED	-0,1658	0,468

R-Sq = **16,3%** **R-Sq(adj)** = **5,8%**

Tabela 4.14 – Curso de Administração (ADM)

Apenas as variáveis ENEM e FÍS, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. De R-Sq nota-se que somente 16,3% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 5,8%. Conclui-se que apesar do modelo possuir variáveis independentes que influenciam de forma significativa no resultado, este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Administração.

CCO

The regression equation is:

COEF. = 44,8 + 0,0570 ENEM + 0,039 MAT + 0,115 FIS + 0,525 POR + 0,122 QUI - 0,702 ING + 0,454 HIS - 0,656 GEO + 0,130 BIO + 0,210 RED

Predictor	Coef	P
Constant	44,81	0,145
ENEM	0,05696	0,555
MAT	0,0395	0,873
FIS	0,1146	0,673
POR	0,5253	0,113
QUI	0,1219	0,704
ING	-0,7019	**0,012**
HIS	0,4542	0,299
GEO	-0,6558	0,096
BIO	0,1300	0,674
RED	0,2098	0,480

R-Sq = **13,0%** **R-Sq(adj)** = **2,0%**

Tabela 4.15 – Curso Ciência da Computação (CCO)

Apenas a variável ING, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. De R-Sq nota-se que somente 13,0% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 2%. Conclui-se que apesar do modelo possuir uma variável independente que influencia de forma significativa no resultado, este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Ciência da Computação.

EAM

The regression equation is

COEF. = 48,2 + 0,0626 ENEM + 0,018 MAT + 0,432 FIS + 0,182 POR + 0,136 QUI - 0,091 ING + 0,027 HIS - 0,154 GEO - 0,098 BIO - 0,015 RED

Predictor	Coef	P
Constant	48,17	0,004
ENEM	0,06261	0,180
MAT	0,0177	0,904
FIS	0,4320	**0,008**
POR	0,1824	0,302
QUI	0,1360	0,475
ING	-0,0911	0,549
HIS	0,0266	0,906
GEO	-0,1539	0,474
BIO	-0,0979	0,622
RED	-0,0150	0,919

R-Sq = **15,0%** **R-Sq(adj)** = **4,2%**

Tabela 4.16 – Curso de Engenharia Ambiental (EAM)

Apenas a variável FIS, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. De R-Sq nota-se que somente 15,0% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 4,2% Conclui-se que apesar do modelo possuir uma variável independente que influencia de forma significativa no

resultado, este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Engenharia Ambiental.

ECA

The regression equation is

COEF. = 64,0 - 0,0367 ENEM + 0,324 MAT + 0,023 FIS + 0,348 POR - 0,124 QUI - 0,192 ING - 0,086 HIS - 0,326 GEO + 0,071 BIO + 0,037 RED

Predictor	Coef	P
Constant	64,02	0,007
ENEM	-0,03674	0,535
MAT	0,3235	0,053
FIS	0,0228	0,868
POR	0,3482	0,096
QUI	-0,1238	0,642
ING	-0,1920	0,332
HIS	-0,0860	0,754
GEO	-0,3258	0,128
BIO	0,0705	0,700
RED	0,0368	0,855

R-Sq = **10,0%** **R-Sq(adj)** = **1,3%**

Tabela 4.17 – Curso de Engenharia de Controle e Automação (ECA)

Este modelo não possui nenhuma variável que tenha influencia significativa sobre a resposta. De R-Sq nota-se que somente 10,0% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 1,3%. Conclui-se que este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Engenharia de Controle e Automação.

ECO

The regression equation is

COEF. = 41,9 + 0,0326 ENEM - 0,253 MAT + 0,306 FIS + 0,394 POR + 0,049 QUI + 0,023 ING + 0,033 HIS - 0,040 GEO + 0,032 BIO + 0,149 RED

Predictor	Coef	P
Constant	41,93	0,014
ENEM	0,03257	0,504
MAT	-0,2532	0,146
FIS	0,3058	**0,048**
POR	0,3935	**0,039**
QUI	0,0486	0,801
ING	0,0234	0,888
HIS	0,0335	0,889
GEO	-0,0402	0,863
BIO	0,0321	0,872
RED	0,1495	0,329

R-Sq = **11,6%** **R-Sq(adj)** = **5,2%**

Tabela 4.18 – Curso de Engenharia da Computação (ECO)

Apenas as variáveis FÍS e POR, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. De R-Sq nota-se que somente 11,6% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 5,2%. Conclui-se que apesar do modelo possuir variáveis independentes que influenciam de forma

significativa no resultado, este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Engenharia da Computação.

EEL
The regression equation is
COEF. = 10,0 + 0,0412 ENEM + 0,297 MAT + 0,055 FIS + 0,196 POR + 0,400 QUI + 0,141 ING - 0,174 HIS + 0,299 GEO - 0,165 BIO + 0,118 RED

Predictor	Coef	P
Constant	10,03	0,442
ENEM	0,04117	0,285
MAT	0,2969	**0,013**
FIS	0,0550	0,630
POR	0,1961	0,151
QUI	0,3999	**0,016**
ING	0,1406	0,223
HIS	-0,1737	0,312
GEO	0,2985	0,069
BIO	-0,1647	0,223
RED	0,1180	0,343

R-Sq = **11,7%** **R-Sq(adj)** = **7,3%**

Tabela 4.19– Curso de Engenharia Elétrica (EEL)

Apenas as variáveis MAT e QUI, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. De R-Sq nota-se que somente 11,7% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 7,3%. Conclui-se que apesar do modelo possuir variáveis independentes que influenciam de forma significativa no resultado, este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Engenharia Elétrica.

EHD
The regression equation is

COEF. = 41,7 + 0,0074 ENEM + 0,376 MAT + 0,133 FIS + 0,351 POR + 0,048 QUI - 0,668 ING + 0,260 HIS - 0,182 GEO + 0,066 BIO + 0,060 RED

Predictor	Coef	P
Constant	41,69	0,120
ENEM	0,00736	0,940
MAT	0,3756	0,190
FIS	0,1331	0,607
POR	0,3515	0,322
QUI	0,0480	0,885
ING	-0,6680	**0,027**
HIS	0,2603	0,519
GEO	-0,1823	0,642
BIO	0,0660	0,834
RED	0,0600	0,827

R-Sq = **21,8%** **R-Sq(adj)** = **5,9%**

Tabela 4.20 – Curso de Engenharia Hídrica (EHD)

Apenas a variável ING, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. De R-Sq nota-se que somente 21,8% da variação em Y pode ser explicada pelas variáveis independente, com o ajuste R-Sq(adj) essa variação cai para 5,9%. Conclui-se que

apesar do modelo possuir uma variável independente que influencia de forma significativa no resultado, este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Engenharia Hidráulica.

EME

The regression equation is

COEF. = 3,9 + 0,0309 ENEM + 0,270 MAT + 0,391 FIS + 0,180 POR + 0,436 QUI - 0,212 ING + 0,060 HIS - 0,217 GEO + 0,113 BIO + 0,407 RED

Predictor	Coef	P
Constant	3,92	0,810
ENEM	0,03088	0,531
MAT	0,2704	0,051
FIS	0,3905	**0,004**
POR	0,1804	0,309
QUI	0,4356	**0,016**
ING	-0,2120	0,097
HIS	0,0601	0,773
GEO	-0,2168	0,207
BIO	0,1131	0,487
RED	0,4072	**0,006**

R-Sq = **19,6%** **R-Sq(adj)** = **14,8%**

Tabela 4.21 – Curso de Engenharia Mecânica (EME)

Apenas as variáveis FIS, QUI e RED, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. De R-Sq nota-se que somente 19,6% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 14,8%.Conclui-se que apesar do modelo possuir variáveis independentes que influenciam de forma significativa no resultado, este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Engenharia Mecânica.

EPR

The regression equation is

COEF. = 18,8 - 0,0120 ENEM + 0,235 MAT - 0,011 FIS + 0,171 POR + 0,375 QUI + 0,078 ING + 0,207 HIS - 0,314 GEO + 0,319 BIO + 0,340 RED

Predictor	Coef	P
Constant	18,78	0,431
ENEM	-0,01203	0,850
MAT	0,2351	0,127
FIS	-0,0111	0,944
POR	0,1706	0,383
QUI	0,3751	0,143
ING	0,0781	0,698
HIS	0,2068	0,427
GEO	-0,3144	0,209
BIO	0,3191	0,105
RED	0,3405	0,058

R-Sq = **14,0%** **R-Sq(adj)** = **3,1%**

Tabela 4.22 – Curso de Engenharia de Produção (EPR)

Este modelo não possui nenhuma variável que tenha influencia significativa sobre a resposta. De R-Sq nota-se que somente 14,0% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 3,1%. Conclui-se que este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Engenharia de Produção.

FBA
The regression equation is
COEF. = *44,3 + 0,030 ENEM + 0,499 MAT + 0,426 FIS + 0,534 POR - 0,519 QUI - 0,503 ING + 0,977 HIS - 0,681 GEO - 0,253 BIO - 0,353 RED*

Predictor	Coef	P
Constant	44,28	0,165
ENEM	0,0301	0,800
MAT	0,4992	0,149
FIS	0,4265	0,247
POR	0,5342	0,205
QUI	-0,5190	0,282
ING	-0,5032	0,150
HIS	0,9768	0,066
GEO	-0,6810	0,321
BIO	-0,2528	0,525
RED	-0,3530	0,515

R-Sq = **31,9%** **R-Sq(adj)** = **7,6%**

Tabela 4.23 – Curso de Física - Bacharelado (FBA)

Este modelo não possui nenhuma variável que tenha influencia significativa sobre a resposta. De R-Sq nota-se que somente 31,9% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 7,6% . Conclui-se que este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Física – Bacharelado.

FLI
The regression equation is
COEF. = *27,1 + 0,083 ENEM - 0,171 MAT - 0,102 FIS + 0,753 POR + 0,636 QUI - 0,566 ING - 0,496 HIS + 0,141 GEO + 0,722 BIO - 0,287 RED*

Predictor	Coef	P
Constant	27,09	0,309
ENEM	0,0835	0,481
MAT	-0,1712	0,643
FIS	-0,1024	0,774
POR	0,7526	0,209
QUI	0,6359	0,066
ING	-0,5665	0,203
HIS	-0,4956	0,382
GEO	0,1408	0,803
BIO	0,7216	0,112
RED	-0,2875	0,454

R-Sq = **30,0%** **R-Sq(adj)** = **5,9%**

Tabela 4.24 – Curso de Física - Licenciatura (FLI)

Este modelo não possui nenhuma variável que tenha influencia significativa sobre a resposta. De R-Sq nota-se que somente 30,0% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 5,9%. Conclui-se que este modelo não possui um bom ajuste e não pode ser utilizado para predizer o coeficiente final de um aluno do curso de Física – Licenciatura.

Além de analisar o modelo de cada curso, procurou-se analisar os dados de todos os cursos juntos, com o objetivo de encontrar um modelo que pudesse ser utilizado para predizer o coeficiente final de cada aluno, independente de seu curso. O resultado encontrado é apresentado na *Tabela 4.25*.

(GLOBAL)
The regression equation is
COEF. = *23,8 + 0,0399 ENEM + 0,184 MAT + 0,0665 FIS + 0,289 POR + 0,263 QUI - 0,0714 ING + 0,0580 HIS - 0,0626 GEO + 0,0900 BIO + 0,157 RED*

Predictor	Coef	P
Constant	23,808	0,000
ENEM	0,03990	**0,022**
MAT	0,18419	**0,000**
FIS	0,06646	0,177
POR	0,28863	**0,000**
QUI	0,26261	**0,000**
ING	-0,07136	0,157
HIS	0,05799	0,455
GEO	-0,06258	0,380
BIO	0,08996	0,146
RED	0,15679	**0,007**

R-Sq = **7,9%** **R-Sq(adj)** = **7,1%**

Tabela 4.25 – Análise Global

As variáveis ENEM, MAT, POR, QUI e RED com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. De R-Sq nota-se que somente 7,9% da variação em Y pode ser explicada pelas variáveis independentes, com o ajuste R-Sq(adj) essa variação cai para 7,1%. Conclui-se que este modelo não pode ser utilizado para predizer o coeficiente final de um aluno, independente de seu curso.

4.4 Análise de Regressão Logística Binária para a Situação Final

Em alguns modelos de regressão as variáveis dependentes são dicotômicas, apresentando, portanto, apenas dois valores possíveis, como sim ou não, 0 ou 1, e a decisão por uma delas depende da variável independente. Como visto no capítulo 3, a melhor forma de manipular estes dados é estimar através de um modelo, a probabilidade de ocorrência para cada uma das respostas. O modelo utilizado neste caso é o modelo de Regressão Logística Binária.

Tem-se para a variável dependente Situação Final, um caso de variável dicotômica, portanto, o modelo utilizado será o da Regressão Logística Binária. As variáveis

independentes da equação serão as notas do ENEM, e as notas das disciplinas dos vestibulares, Matemática, Física, Português, Química, Inglês, História, Geografia, Biologia e Redação, logo o modelo será de Regressão Logística Múltipla. A resposta "Situação Final" será 0 ou 1. Atribui-se o valor 1 para o evento de sucesso "Colou Grau" e o valor 0 para o acontecimento de fracasso "Não Colou Grau". O evento "Não Colou Grau" engloba os níveis Ativo, Desistência a Pedido, Desistência por Processo, Falecido, Mudou de curso por ter feito novo vestibular, Mudou de curso por transferência interna, Transferido para outra instituição.

A análise dos resultados tem como principais componentes o ***Valor P***, valor dos ***coeficientes***, ***as odds ratio*** e ***os teste de ajustes (Pearson, Deviance e Hosmer-Leshow)***. Conforme mostrado no Capítulo 3, o Valor P testa a hipótese nula de que os coeficientes são iguais a zero versus estes sendo diferentes de zero. Quando o Valor P for <= ao nível de significância 0,05, a hipótese nula será rejeitada, quando o Valor P for > que o ao nível de significância 0,05, a hipótese nula não será rejeitada.

- H_0: todos os coeficientes são iguais a zero (não há regressão logística);
- H_1: pelo menos um dos coeficientes é diferente de zero (há regressão logística).

Os testes de Pearson, Deviance e Hosmer-Leshow testam a hipótese nula de que o ajuste dos dados é bom versus o ajuste sendo ruim. Quanto maior o Valor P, melhor é a qualidade de ajuste do modelo.

- H_0: Valor P > 0,05 – O ajuste dos dados é bom;
- H_1: Valor P <= 0,05 – O ajuste dos dados não é bom;

A razão de chances (odds-ratio) é a razão entre a probabilidade do evento ocorrer e a probabilidade do evento não ocorrer. No caso, o evento é quando o valor da variável "Situação Final" é 1, isto é, quando o aluno colou grau.

Tomando como dados de entrada as notas do ENEM e das disciplinas do vestibular no período de 2000/2003 obteve-se no programa Minitab os modelos apresentados nas *Tabelas 4.26 a 4.37.*

Logistic Regression Table (ADM)

Predictor	Coef	P	Odds Ratio
Constant	-3,30538	0,203	
ENEM	0,0187263	**0,012**	**1,02**
MAT	0,0027988	0,909	1,00
FIS	-0,0333900	0,107	0,97
POR	0,0621809	0,054	1,06
QUI	0,0523066	0,065	1,05
ING	0,0056144	0,754	1,01
HIS	-0,0098501	0,731	0,99
GEO	-0,0432361	0,215	0,96
BIO	0,0619428	**0,029**	**1,06**
RED	-0,0372289	0,168	0,96

Goodness-of-Fit Tests

Method	P
Pearson	0,294
Deviance	0,065
Hosmer-Lemeshow	0,140

Tabela 4.26 – Curso de Administração (ADM)

Somente as variáveis ENEM e BIO, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. Para a variável ENEM, por exemplo, a razão de chances é 1,02, ou seja, a cada um ponto aumentado na variável, as chances do aluno colar grau aumentam em 2% em relação as chances dele não colar grau. Os testes de ajustes apresentam Deviance com Valor P muito pequeno (0,065). Dessa maneira apesar do modelo possuir variáveis que influenciam de forma significativa a resposta, ele não apresenta um bom ajuste. Conclui-se que este modelo não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Administração.

Logistic Regression Table (CCO)

Predictor	Coef	P	Odds Ratio
Constant	-4,40770	0,106	
ENEM	0,0055683	0,523	1,01
MAT	0,0171326	0,436	1,02
FIS	-0,0036667	0,875	1,00
POR	0,0343554	0,244	1,03
QUI	0,0160303	0,569	1,02
ING	-0,0312107	0,208	0,97
HIS	0,0713037	0,069	1,07
GEO	-0,0199097	0,566	0,98
BIO	0,0373710	0,181	1,04
RED	-0,0053850	0,836	0,99

Goodness-of-Fit Tests

Method	P
Pearson	0,170
Deviance	0,009
Hosmer-Lemeshow	0,568

Tabela 4.27 – Curso Ciência da Computação (CCO)

Neste modelo nenhuma variável apresenta Valor P abaixo de 0,05, ou seja, nenhuma tem influência significativa sobre a resposta. Os testes de ajustes apresentam Deviance com Valor P muito pequeno (0,009). O modelo não possui variáveis que influenciam de forma significativa a resposta, além de não apresentar um bom ajuste. Conclui-se que este modelo não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Ciência da Computação.

Logistic Regression Table (EAM)

Predictor	Coef	P	Odds Ratio
Constant	2,80423	0,383	
ENEM	0,0067984	0,484	1,01
MAT	-0,0192656	0,537	0,98
FIS	-0,0298296	0,334	0,97
POR	-0,0233179	0,537	0,98
QUI	0,0652677	0,086	1,07
ING	-0,0495342	0,141	0,95
HIS	0,0329696	0,491	1,03
GEO	-0,0471166	0,292	0,95
BIO	0,0469926	0,245	1,05
RED	-0,0349344	0,271	0,97

Goodness-of-Fit Tests

Method	P
Pearson	0,152
Deviance	0,540
Hosmer-Lemeshow	0,735

Tabela 4.28 – Curso Engenharia Ambiental (EAM)

Neste modelo nenhuma variável apresenta Valor P abaixo de 0,05, ou seja, nenhuma tem influência significativa sobre a resposta. O Valor P dos teste de qualidade de ajuste, Pearson, Deviance e Hosmer-Lemeshow variam entre 0,152 a 0,735, indicando que há evidências insuficientes para afirmar que o ajuste do modelo não é bom. Conclui-se que este modelo apesar de apresentar um bom ajuste, não possui variáveis que influenciam de forma significativa a resposta, não existindo regressão logística. Dessa forma o modelo não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Engenharia Ambiental.

Logistic Regression Table (ECA)

Predictor	Coef	P	Odds Ratio
Constant	-7,88646	0,056	
ENEM	0,0076076	0,476	1,01
MAT	0,0642755	**0,025**	**1,07**
FIS	0,0267862	0,248	1,03
POR	0,0378401	0,299	1,04
QUI	0,0268017	0,532	1,03
ING	-0,0028421	0,928	1,00
HIS	0,0548490	0,225	1,06
GEO	-0,0082138	0,821	0,99
BIO	-0,0148262	0,642	0,99
RED	0,0381487	0,266	1,04

Goodness-of-Fit Tests

Method	P
Pearson	0,117
Deviance	0,615
Hosmer-Lemeshow	0,628

Tabela 4.29 – Curso de Engenharia de Controle e Automação (ECA)

Somente a variável MAT, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. Sua razão de chances é 1,07, ou seja, a cada um ponto aumentado na variável, as chances do aluno colar grau aumentam em 7% em relação as chances dele não colar grau. O Valor P dos teste de qualidade de ajuste, Pearson, Deviance e Hosmer-Lemeshow variam entre 0,117 a 0,628, indicando que há evidências insuficientes para afirmar que o ajuste do modelo não é bom. Conclui-se que este modelo pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Engenharia de Controle e Automação.

Logistic Regression Table (ECO)

Predictor	Coef	P	Odds Ratio
Constant	-3,35151	0,130	
ENEM	-0,0061347	0,350	0,99
MAT	-0,0046612	0,838	1,00
FIS	0,0183895	0,342	1,02
POR	0,0131825	0,577	1,01
QUI	0,0021604	0,929	1,00
ING	0,0303678	0,140	1,03
HIS	0,0164118	0,581	1,02
GEO	0,0034130	0,909	1,00
BIO	0,0327694	0,188	1,03
RED	0,0062319	0,750	1,01

Goodness-of-Fit Tests

Method	P
Pearson	0,174
Deviance	0,013
Hosmer-Lemeshow	0,102

Tabela 4.30 – Curso de Engenharia da Computação (ECO)

Neste modelo nenhuma variável apresenta Valor P abaixo de 0,05, ou seja, nenhuma tem influência significativa sobre a resposta. Os testes de ajustes apresentam Deviance com Valor P muito pequeno (0,013). O modelo não possui variáveis que influenciam de forma significativa a resposta, não existindo regressão logística, além de não apresentar um bom ajuste. Conclui-se que este modelo não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Engenharia da Computação.

Logistic Regression Table (EEL)

Predictor	Coef	P	Odds Ratio
Constant	-8,03820	0,000	
ENEM	-0,0074544	0,262	0,99
MAT	0,0487118	**0,008**	**1,05**
FIS	0,0077988	0,642	1,01
POR	0,0371345	0,084	1,04
QUI	0,0661587	**0,006**	**1,07**
ING	0,0234970	0,172	1,02
HIS	0,0003217	0,990	1,00
GEO	0,0593782	**0,024**	**1,06**
BIO	-0,0075705	0,704	0,99
RED	-0,0212459	0,261	0,98

Goodness-of-Fit Tests

Method	P
Pearson	0,253
Deviance	0,125
Hosmer-Lemeshow	0,384

Tabela 4.31 – Curso de Engenharia Elétrica (EEL)

Somente as variáveis MAT, QUI e GEO, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. Para a variável MAT, por exemplo, a razão de chances é 1,05, ou seja, a cada um ponto aumentado na variável MAT, as chances do aluno colar grau aumentam em 5% em relação as chances dele não colar grau. O Valor P dos teste de qualidade de ajuste, Pearson, Deviance e Hosmer-Lemeshow variam entre 0,125 a 0,384, indicando que há evidências insuficientes para afirmar que o ajuste do modelo não é bom. Conclui-se que este modelo pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Engenharia Elétrica.

Logistic Regression Table (EHD)

Predictor	Coef	P	Odds Ratio
Constant	-4,22385	0,291	
ENEM	0,0200941	0,189	1,02
MAT	0,0607291	0,117	1,06
FIS	0,0333305	0,323	1,03
POR	-0,0768674	0,142	0,93
QUI	-0,0515509	0,264	0,95
ING	-0,112635	**0,028**	**0,89**
HIS	0,0357911	0,487	1,04
GEO	0,0820470	0,204	1,09
BIO	0,0991180	**0,032**	**1,10**
RED	0,0264541	0,472	1,03

Goodness-of-Fit Tests

Method	P
Pearson	0,000
Deviance	0,263
Hosmer-Lemeshow	0,008

Tabela 4.32 – Curso de Engenharia Hídrica (EHD)

Somente as variáveis ING e BIO, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. Para a variável ING, por exemplo, a razão de chances é 0,89, ou seja, a cada um ponto diminuído na variável, as chances do aluno não colar grau aumentam em 11% em relação as chances dele colar grau. Os testes de ajustes apresentam Pearson e Hosmer-Lemeshow com Valor P (0,000) e (0,008). Dessa maneira apesar do modelo possuir variáveis que influenciam de forma significativa a resposta, ele não apresenta um bom ajuste. Conclui-se que este modelo não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Engenharia Hídrica.

Logistic Regression Table (EME)

Predictor	Coef	P	Odds Ratio
Constant	-8,99755	0,000	
ENEM	-0,0008324	0,901	1,00
MAT	0,0247503	0,191	1,03
FIS	0,0350730	0,055	1,04
POR	-0,0276035	0,258	0,97
QUI	0,0675856	**0,006**	**1,07**
ING	-0,0009598	0,955	1,00
HIS	0,0319494	0,244	1,03
GEO	0,0149643	0,541	1,02
BIO	0,0659215	**0,003**	**1,07**
RED	0,0387839	0,057	1,04

Goodness-of-Fit Tests

Method	P
Pearson	0,493
Deviance	0,069
Hosmer-Lemeshow	0,515

Tabela 4.33 – Curso de Engenharia Mecânica (EME)

Somente as variáveis QUI e BIO, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. Para a variável QUI, por exemplo, a razão de chances é 1,07, ou seja, a cada um ponto aumentado na variável, as chances do aluno colar grau aumentam em 7% em relação as chances dele não colar grau. Os testes de ajustes apresentam Deviance com Valor P muito pequeno (0,069). Dessa maneira apesar do modelo possuir variáveis que influenciam de forma significativa a resposta, ele não apresenta um bom ajuste. Conclui-se que este modelo não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Engenharia Mecânica.

Logistic Regression Table (EPR)

Predictor	Coef	P	Odds Ratio
Constant	-5,16095	0,312	
ENEM	-0,0022680	0,874	1,00
MAT	-0,0203623	0,550	0,98
FIS	0,0187573	0,491	1,02
POR	0,0194792	0,632	1,02
QUI	0,0386085	0,436	1,04
ING	-0,0420867	0,358	0,96
HIS	-0,0103225	0,839	0,99
GEO	0,0371408	0,488	1,04
BIO	0,0787246	**0,043**	**1,08**
RED	0,0960664	**0,031**	**1,10**

Goodness-of-Fit Tests

Method	P
Pearson	0,367
Deviance	0,773
Hosmer-Lemeshow	0,500

Tabela 4.34 – Curso de Engenharia de Produção (EPR)

Somente as variáveis BIO e RED, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. Para a variável BIO, por exemplo, a razão de chances é 1,08, ou seja, a cada um ponto aumentado na variável, as chances do aluno colar grau aumentam em 8% em relação as chances dele não colar grau. O Valor P dos teste de qualidade de ajuste, Pearson, Deviance e Hosmer-Lemeshow variam entre 0,367 a 0,773, indicando que há evidências insuficientes para afirmar que o ajuste do modelo não é bom. Conclui-se que este modelo pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Engenharia de Produção.

Logistic Regression Table (FBA)

Predictor	Coef	P	Odds Ratio
Constant	6,51583	0,544	
ENEM	0,0692085	0,144	1,07
MAT	-0,0684534	0,395	0,93
FIS	0,368298	0,070	1,45
POR	0,302917	0,066	1,35
QUI	-0,237693	0,234	0,79
ING	-0,383404	0,061	0,68
HIS	0,255275	0,100	1,29
GEO	-0,606012	0,112	0,55
BIO	-0,0179434	0,875	0,98
RED	-0,178438	0,246	0,84

Goodness-of-Fit Tests

Method	P
Pearson	0,888
Deviance	0,932
Hosmer-Lemeshow	0,963

Tabela 4.35 – Curso de Física - Bacharelado (FBA)

Neste modelo nenhuma variável apresenta Valor P abaixo de 0,05, ou seja, nenhuma têm influência significativa sobre a resposta. O Valor P dos teste de qualidade de ajuste, Pearson, Deviance e Hosmer-Lemeshow variam entre 0,888 a 0,932, indicando que há evidências insuficientes para afirmar que o ajuste do modelo não é bom. Conclui-se que este modelo apesar de apresenta um bom ajuste, não possuir variáveis que influenciam de forma significativa a resposta, não existindo regressão logística. Dessa forma o modelo não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Física - Bacharelado.

Logistic Regression Table (FLI)

Predictor	Coef	P	Odds Ratio
Constant	-8,38102	0,044	
ENEM	0,0266928	0,124	1,03
MAT	0,0106544	0,834	1,01
FIS	-0,0235643	0,647	0,98
POR	0,213258	**0,029**	**1,24**
QUI	0,0518493	0,231	1,05
ING	-0,100408	0,120	0,90
HIS	-0,0600208	0,410	0,94
GEO	0,0452308	0,576	1,05
BIO	0,0172587	0,773	1,02
RED	0,0212535	0,657	1,02

Goodness-of-Fit Tests

Method	P
Pearson	0,122
Deviance	0,101
Hosmer-Lemeshow	0,139

Tabela 4.36 – Curso de Física - Licenciatura (FLI)

Somente a variável POR, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. Sua razão de chances é 1,24, ou seja, a cada um ponto aumentado na variável POR, as chances do aluno colar grau aumentam em 24% em relação as chances dele não colar grau. O Valor P dos teste de qualidade de ajuste, Pearson, Deviance e Hosmer-Lemeshow variam entre 0,101 a 0,139, indicando que há evidências insuficientes para afirmar que o ajuste do modelo não é bom. Conclui-se que este modelo pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno de Física - Licenciatura.

Além de analisar o modelo de cada curso, procurou-se analisar os dados de todos os cursos juntos, com o objetivo de encontrar um modelo que pudesse ser utilizado para predizer a probabilidade de conclusão de um aluno, independente de seu curso. O resultado encontrado é apresentado na *Tabela 4.37*.

Logistic Regression Table (GLOBAL)

Predictor	Coef	P	Odds Ratio
Constant	-5,71949	0,000	
ENEM	0,0031774	0,135	1,00
MAT	0,0223819	**0,000**	**1,02**
FIS	0,0056170	0,343	1,01
POR	0,0181983	**0,025**	**1,02**
QUI	0,0432277	**0,000**	**1,04**
ING	-0,0039898	0,517	1,00
HIS	0,0185577	**0,046**	**1,02**
GEO	0,0128430	0,146	1,01
BIO	0,0332903	**0,000**	**1,03**
RED	0,0085714	0,222	1,01

Goodness-of-Fit Tests

Method	P
Pearson	0,097
Deviance	0,000
Hosmer-Lemeshow	0,073

Tabela 4.37 – Análise Global

As variáveis MAT, POR, QUI, HIS e BIO, com Valor P abaixo de 0,05, têm influência significativa sobre a resposta. Para a variável MAT, por exemplo, a razão de chances é 1,02, ou seja, a cada um ponto aumentado na variável, as chances do aluno colar grau aumentam em 2% em relação as chances dele não colar grau. Os testes de ajustes apresentam Valor P muito pequenos. Dessa maneira apesar do modelo possuir variáveis que influenciam de forma significativa a resposta, ele não apresenta um bom ajuste. Conclui-se que este modelo não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno.

Diante de toda a análise tanto da parte de Regressão Linear, como da parte de Regressão Logística pode-se concluir que somente a nota do vestibular não é suficiente para prever o desempenho do aluno ao final do curso. Outros fatores são importantes, tais como os citados no Capítulo 2.

CAPÍTULO 5 – CONCLUSÃO

A pressão que atualmente as instituições de ensino superior têm sofrido no sentido de aumentarem a eficácia na educação e melhorarem as taxas de graduação tem feito com que elas olhem com mais cautela suas políticas de admissão.

É através do processo de admissão que as instituições passam a ter uma idéia do perfil de seus novos estudantes e podem assim obter uma probabilidade de quais atingirão seus objetivos educacionais, ou seja, de obterem um bom desempenho durante o curso e conseqüentemente se graduarem.

Os indicadores de sucesso do estudante têm se tornado cada vez mais importantes para o bem estar financeiro de uma instituição. Os índices de graduação afetam a percepção pública de uma instituição e podem afetar os fundos governamentais da mesma. A experiência sobre o ensino superior tem encorajado pesquisas relacionadas com a persistência e evasão de estudantes.

Estudos mostraram que alguns fatores como sexo do aluno; renda familiar; nível de educação dos pais; ajuda financeira; estado civil, idade, emprego, motivação, interação com a instituição podem afetar seu desempenho e sua probabilidade de persistência.

Pesquisadores têm usado uma variedade de técnicas estatísticas para examinar a importância de variáveis acadêmicas em predizer o sucesso do estudante.

Usando as ferramentas e dados disponíveis, instituições podem criar modelos preditivos designados a prever o desempenho de seus alunos.

O presente estudo buscou, com as notas do ENEM e das disciplinas dos vestibulares, da Universidade Federal de Itajubá, UNIFEI, nos anos de 2000, 2001, 2002 e 2003, estabelecer modelos preditivos designados a prever a probabilidade de graduação do aluno, bem como o de predizer o seu coeficiente final, utilizando para isso técnicas de Regressão Logística Binária e Regressão Linear Múltipla.

Realizou-se uma análise gráfica e descritiva do banco de dados disponível, relacionando todas as variáveis dependentes, coeficiente final e situação final do aluno, e variáveis independentes, notas do ENEM e as notas das disciplinas dos vestibulares no período de 2000 a 2003.

Admitindo um relacionamento linear entre as notas obtidas no vestibular e no ENEM, e o coeficiente final do aluno, obteve-se um problema de regressão. O modelo de regressão linear múltipla foi utilizado para previsão do coeficiente final do aluno.

Para a variável dependente "Situação Final" utilizou-se o modelo da Regressão Logística Binária, para predizer a probabilidade do aluno se graduar ou não ao fim de seu período de curso. Atribui-se o valor 1 para o evento de sucesso "Colou Grau" e o valor 0 para o acontecimento de fracasso "Não Colou Grau".

O evento "Não Colou Grau" englobou os níveis: *Ativo*, *Desistência a Pedido*, *Desistência por Processo*, *Falecido*, *Mudou de curso por ter feito novo vestibular*, *Mudou de curso por transferência interna*, *Transferido para outra instituição*.

Tomando-se como dados de entrada as notas do ENEM e das disciplinas do vestibular no período de 2000/2003 obteve-se os modelos de Regressão Linear Múltipla e Regressão Logística Binária. Realizou-se uma análise de dados para cada curso e depois uma análise global.

Utilizou-se também os *boxplots* para representar as notas dos vestibulares de cada disciplina em cada curso, além de utilizá-los para representar o coeficiente final dos cursos. Os gráficos de setores foram necessários para mostrar a situação final de cada curso e de todos de uma forma geral.

Os principais achados foram:

1. Analisando de uma forma geral a soma das notas do vestibular, verificou-se uma queda com o passar dos anos, o mesmo acontecendo com a porcentagem de alunos que colaram grau. Isto não significa que haja uma correlação entre o resultado do vestibular e a porcentagem de graduação, levando-se em conta que vários motivos levam o aluno a não colar grau, tais como: alunos ainda ativos, desistência a pedido, desistência por processo, falecimento, mudança de curso por ter feito novo vestibular, mudança de curso por transferência interna, transferência para outra instituição. Um menor desempenho no vestibular não deve ser sempre associado a um menor desempenho no curso;
2. Não foi encontrado nenhum modelo que possa ser utilizado para predizer o coeficiente final de um aluno, independente de seu curso. Dessa forma, um aluno que obtém as melhores notas no vestibular não será necessariamente um aluno de ótimo desempenho, tal como dito acima. O contrário também não se verifica. Existe sim uma regressão para a média indicando que o processo não tolera os extremos. Dessa forma, os valores discrepantes são ajustados à média;
3. Cursos que se destacaram pelas notas do vestibular de seus candidatos não conseguiram manter o mesmo nível no coeficiente final e na porcentagem de

graduação por outras razões como: o vestibular de uma forma geral é o mesmo para os diversos cursos, porém a forma de ensino e o grau de dificuldade dos cursos se diferem;

4. A grande maioria dos cursos não apresentou modelos que possam ser usados para predizer a probabilidade de sucesso de um aluno em colar grau. O modelo da análise global, envolvendo todos os cursos não pode ser utilizado para predizer a probabilidade de conclusão de curso de um aluno;
5. Foram encontrados poucos modelos que podem ser utilizados para predizer a probabilidade de sucesso de um aluno. Pode-se concluir que estes cursos, na maioria das vezes, conseguem fazer com que o aluno admitido, permaneça de uma forma geral, com o mesmo desempenho do vestibular até o fim de seu curso. Uma das maneiras de se obter esse sucesso pode ser através da atenção atribuída às variáveis de sucesso citadas no Capítulo 2;
6. Cada curso apresenta uma característica própria. Cursos se diferem por suas estruturas, qualificação de professores, formas de avaliação, apoio aos alunos, relacionamento extra-classe e assim por diante. Isto é um dos motivos pelos quais muitos cursos se destacam quando relacionados aos outros cursos da mesma instituição. Pelo fato de cada curso possuir suas características não é possível assim encontrar um modelo único que possa predizer a probabilidade de sucesso de um aluno;
7. Para se prever o desempenho de um aluno, quer seja seu desempenho final ou a probabilidade de graduar-se é necessário que seja analisado diversos fatores. O vestibular é uma ferramenta importante para a seleção de candidatos a um curso superior, visto que, as escolas possuem número limitado de vagas. Porém, ele não é o fator principal quando se deseja prever o desempenho destes alunos num período futuro, isto é um engano que não só as instituições possuem como toda a sociedade.
8. A previsão é uma medida importante para qualquer área de estudo e para que seus resultados sejam precisos é necessário muito cuidado antes de se criar um modelo para este fim. Através deste trabalho, conclui-se principalmente, que para se prever o desempenho de um aluno, a instituição precisa conhecê-lo como um todo, saber sua origem, conhecer sua personalidade, seus planos, suas experiências de vida, suas necessidades, etc. Além de tudo isso é necessário que a instituição também se auto-avalie sempre, verificando a

satisfação dos alunos quanto ao ensino, instalações físicas, relacionamento com o corpo docente e funcionários em geral e outras variáveis que possam estimular o aluno a permanecer no curso e na instituição e a manter um bom desempenho. Outro aspecto é a necessidade de padronização dos cálculos do coeficiente do aluno para todos os cursos;

9. Após serem analisadas as relações entre as notas do vestibular e o coeficiente final do aluno e a sua situação final através da regressão linear múltipla e regressão logística respectivamente, pode-se concluir que somente a nota do vestibular não é suficiente para prever o desempenho do aluno ao final do curso o que sugere que variáveis tais como as citadas no Capítulo 2 também exercem influência no desempenho do aluno durante o curso.

Sugestão para Trabalhos Futuros

Realização de um modelo preditivo que contemple não somente as notas do exame de seleção adotado pela instituição, mas também outras variáveis (sexo, renda familiar, nível de educação dos pais, ajuda financeira, estado civil, idade, emprego, aspiração inicial do estudante, variáveis de motivação, variáveis de interação, variáveis institucionais) apresentadas no Capítulo 2. Para tanto, é necessário que a instituição disponha de uma base de dados. Na ausência da mesma, propõe-se que seja formulado um questionário estruturado para ser aplicado junto aos alunos matriculados. Desta forma, será possível obter um conjunto de dados mais completo para o desenvolvimento de modelos que permitam predizer o desempenho do aluno ao final do período previsto.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDIFES - *Indicador do MEC coloca universidades federais entre as 10 melhores do país.* Publicado em set/2008. Disponível em: http://www.andifes.org.br. Acesso em: jan/2009.

ASTIN, A.W. *How Good is Your Institution's Retention Rate?* Research in Higher Education, v.38, n.6, p. 647-658, 1997.

BEAN, J. P. *Drop-outs and turnover: The synthesis and test of a causal model of student attrition.* Research in Higher Education, v.12, n.2, p. 155-187, 1980.

BITTENCOURT, H.R.; CLARKE,R.T. *Um Classificador Baseado na Discriminação Logística: Vantagens e Desvantagens.* Anais X SBSR, INPE, p. 1217-1223, Workshops, Foz do Iguaçu, 2001.

BORGES, J.L.G; CARNIELLI, B.L. Educação e Estratificação Social no Acesso à Universidade Pública. Cardernos de Pesquisa, v.35, n.124, p. 113-139, 2005.

BOWEN*, W. G.*; *BOK*, D. *The shape of the river: long-term consequences of considering race in college and university admissions.* Princeton, NJ. Princeton University Press, 1998.

BROWN, N.W. *Cognitive*, *interest*, and personality variables predicting first-semester GPA. Psychological Reports, v. 74, n. 2, p. 605-606, 1994.

COM CIÊNCIA UNIVERSIDADES - *Carência econômica de estudantes supera o limite de ajuda financeira das universidades.* fev/2003. Disponível em: http://www.comciencia.br. Acesso em: fev/2009.

EDITAL UFMA - Universidade Federal do Maranhão - PRÓ-REITORIA DE ENSINO - EDITAL No. 95/2008 – PROEN - Processo Seletivo Vestibular 2009.

EDITAL UFMT - Universidade Federal de Mato Grosso - Edital n.º 01/2008 - CEV/UFMT - Processo Seletivo de 2009 - (Publicado no DOU dia 16/07/2008, Seção 3, páginas 29 a 34.) - (alterado pelo Edital Complementar 01 de 29.08.2008).

EDITAL UFSC – Universidade Federal de Santa Catarina - EDITAL 04/COPERVE/2008. Processo Seletivo 2009.

EDITAL UFSCAR – Universidade Federal de São Carlos RESOLUÇÃO CEPE Nº 586, de 14 de agosto de 2008. Processo Seletivo 2009.

EDITAL UNIFEI – Universidade Federal de Itajubá - EDITAL DO VESTIBULAR 2009 – Campus Itajubá - Aprovado pela Câmara de Graduação em 05/09/2008 – 17ª Reunião Extraordinária.

EDITAL UNIR - Universidade Federal de Rondônia - Processo Seletivo 2009 - EDITAL N.º 010, 18 de agosto de 2008.

GIL, A.C., *Como Elaborar Projeto de Pesquisa*, 3ª ed., São Paulo, Atlas, 1993.

GIMENO, S.G.A.; SOUZA, J.M.P. *Utilização de estratificação e modelo de regressão logística na análise de dados de estudos caso-controle.* Revista Saúde Pública, v.29, n.4, p 283-290, 1995.

HOSMER, D.W.; LEMESHOW, S. *Applied Logistic Regression.* 2. ed. New York: John Wiley & Sons, 2000.

INEP a – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Ministério da Educação. *Resumo Técnico – Censo da Educação Superior 2007*. Brasília, 2009.

INEP b – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Ministério da Educação. *ENEM: Um exame diferente*. Disponível em: http://www.enem.inep.gov.br. Acesso em: jan/2009.

LENNING, O. T. *Variable-selection and measurement concerns*. In E. Pascarella. (Ed.), Studying Student Attrition. New Directions for Institutional Research, 36. San Francisco: Jossey – Bass, 1982.

MACCOBY E. E.; JACKLIN, C. N. *The psychology of sex differences*. Stanford, California: Stanford University Press, 1974.

MEC- Ministério da Educação – Secretaria da Educação Superior. Disponível em: http://portal.mec.gov.br . Acesso em jan/2009.

MENEZES, E. M.; SILVA, E. L. *Metodologia da pesquisa e elaboração de dissertação.* Florianópolis: UFSC, 2005.

MONTGOMERY, D.C & RUNGER, G.C. *Estabilidade Aplicada e Probabilidade para Engenheiros.* 2. ed. John Wiley & Sons, 2003.

NELSON, L.S. *A Comparison of Classification Methods for Trauma Scoring and Prediction Outcome.* Dissertation (Doctor of Philosophy Statistics). The University of New Mexico. Albuquerque, New Mexico, July 1999.

PASCARELLA, E. T. *Student-Faculty Informal Contact and College Outcomes.* Review of Educational Research, v.50, n.4, p. 545-595, 1980.

PASCARELLA, E. T.; TERENZINI. P. T. *Patterns* of student-faculty informal interaction beyond the classroom and voluntary freshman attrition. The Journal of Higher Education, v. 48, n.5, p. 540-552, 1977.

PASCARELLA, E., SMART, J., & ETHINGTON, C.A. *Long-term* persistence of two year college students. Research in Higher Education, v.24, n.1, p. 47-71, 1986.

PEREIRA, G.G. *Avaliação da CAPES – Abordagem Quantitativa Multivariada dos Programas de Administração*. Dissertação (Mestrado) - Universidade de São Paulo – Faculdade de Economia, Administração e Contabilidade – Departamento de Administração – Programa de Pós Graduação em Administração — São Paulo, 2005.

PRAXEDES, W. *Fatores que influenciam o desempenho no Ensino Superior e a proposta de cotas para alunos negros.* REVISTA ESPAÇO ACADÊMICO – N^0 31 – dez/2003 – Mensal – ISSN 1519.6186. Disponível em: http://www.espacoacademico.com.br. Acesso em: jan/2009.

PRIMI, R.; SANTOS, A.A.A; VENDRAMINI, C.M. *Habilidades Básicas e Desempenho Acadêmico em Universitários Ingressantes*. Estudos De Psicologia, v.7, n.1, p. 47-55, 2002.

REUNI – Reestruturação e Expansão das Universidades Federais – Diretrizes Gerais / Plano de Desenvolvimento da Educação, Agosto 2007.

ROTTER, N.G. *Student attrition in a technological university: academic lifestyle*. College Student Journal, n. 22, p. 241-248, 1998.

SCHUSTER, C. *Logistic Regression Theory With Applicatios*. Thesis (Master of Science) Departament of Mathematical Sciences. The University of Texas at El Paso. USA, 2000.

SOUZA*, E.C. Análise de influência local no modelo de regressão logística. Dissertação* (Mestrado em Agronomia) Área de concentração: Estatística e Experimentação Agronômica, Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, SP, 2006.

SPADY, W.G. *Drop-outs from higher education: An interdisciplinary review and synthesis.* Interchange, v.1, n.1, p. 64-85, 1970.

STAMPEN*,* J. O.; CABRERA*,* A. F. *Exploring the effects of student aid on attrition.* The Journal of Financial Aid, n.16, p. 28-40, 1986.

STAMPEN*,* J. O.; CABRERA*,* A. F. *The targeting and packaging of student aid and its effects on attrition.* Economics of Education Review, n.17, p. 29-46, 1988.

TINTO, V. *Drop-out from higher education: A theoretical synthesis of recent research.* Review of Educational Research, v.45, n.1, p. 89-125, 1975.

TINTO, V. *Leaving college: Rethinking the causes and cures of student attrition.* Chicago: The University of Chicago Press, 1987.

TINTO, V. *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.), Chicago: The University of Chicago Press, 1993.

TSUCHIYA, I. *Regressão Logística Aplicada na Análise Espacial de Dados Arqueológicos*. Dissertação (Mestrado em Ciências Cartográficas). Universidade Estadual Paulista. Presidente Prudente, SP, 2002.

UNIVERSIA BRASIL – *Mulheres na Universidade.* Jun/2006. Disponível em http://www.universia.com.br. Acesso em Jan /2009.

WHITE,W.F.; MOSELY, D. *Twelve year pattern of retention and attrition in a commuter type university.* Education, v.115, n.3, p. 400-402, 1994.

XAVIER, S. S. *Medição de desempenho da cadeia de suprimentos: um estudo de caso em uma empresa fornecedora do setor elétrico.* Dissertação (Mestrado) – Universidade Federal de Itajubá, Itajubá, Minas Gerais, 2008.

ZANINI. A. *Regressão Logística E Redes Neurais Artificiais: Um Problema De Estrutura De Preferência Do Consumidor E Classificação De Perfis De Consumo - TD*. Mestrado em Economia Aplicada - FEA/UFJF 007/2007 - Juiz de Fora - 2007.