MARCELO RIBEIRO DA LUZ
MARCOS KUFNER

Projeto de Trabalho de Conclusão de Curso

Trabalho apresentado para a disciplina de Laboratório de Estatística II do curso de graduação em Estatística da Universidade Federal do Paraná.
Orientadora: Profa. Sonia Isoldi M. Muller

CURITIBA
2008

Comparação entre Reconhecimento de Padrões e Classificação e Regressão Logística como método de melhor resultado na verificação de fatores que influenciam a evasão de alunos do curso de estatística da UFPR .

MARCELO RIBEIRO DA LUZ
MARCOS KUFNER

Curso de Estatística
Universidade Federal do Paraná

2008

RESUMO

O estudo que é apresentado neste trabalho tem como objetivo comparar duas técnicas estatísticas, uma técnica sendo a de Reconhecimento de Padrões e Classificação e a outra a Regressão Logística, como sendo a melhor técnica para verificar os fatores que influenciam na evasão de alunos do curso de Estatística. Neste estudo foram coletadas informações de 163 alunos que ingressaram no curso de Estatística da Universidade Federal do Paraná entre os anos de 1998 e 2000, sendo que foi tomado como base nota e freqüência nas disciplinas cursadas no 1° semestre do curso, e variáveis como gênero, idade, sexo, estado civil entre outras. Para esta análise será utilizada as técnicas de Regressão Logística e Reconhecimento de Padrões e Classificação Como a variável resposta é dicotômica, ou seja, apresenta as categorias desistência do curso ou a não desistência do curso, pode-se aplicar estas duas análises neste estudo.

Palavras-chaves: Regressão Logística, Reconhecimento de Padrões, Evasão.

1. INTRODUÇÃO

1.1 O PROBLEMA

No Brasil, evasão escolar entende-se como a interrupção do ciclo de estudos, o que é uma realidade em todas as IES (Instituição de Ensino Superior) do país. Esse abandono trás prejuízos tanto para o aluno que não terminou o curso e não terá em seu currículo o título de formação, quanto para as instituições que perdem em prestigio externo ou internamente e também a sociedade com investimentos mal aproveitados.

Por que um jovem ou uma jovem que, por meio de todos os esforços possíveis, conseguiu uma vaga universitária abandona a escola? A desistência na educação superior, segundo CASTRO (1994), é relacionada grande diversidade do sistema e à especificidade de cada instituição.

Na busca de respostas para as causas desse fenômeno há que se analisar o que está sendo efetivamente implementado para favorecer as condições acadêmicas do aluno e, conseqüentemente, melhorar o sistema de ensino nacional. Conforme enfatiza a UNESCO (2004), a evasão é sempre processo individual, se bem que pode constituir-se em fenômeno coletivo a ser estudado como associado à eficiência do sistema.

Pode haver decepções, também, quanto às expectativas levantadas em relação à vida universitária, à estrutura e metodologia do trabalho acadêmico e ao excesso de aulas teóricas nos primeiros semestres, quando o aluno, mesmo com o pouco conhecimento específico, almeja o exercício da profissão.

Candidatos à educação superior, em decorrência de suas condições sociais e financeiras, desistem desde o início, da tentativa de ingressar em um curso mais concorrido, portanto, de mais difícil acesso, e optam por outro menos procurado, mesmo com pouco interesse em exercer a profissão correspondente. Esperam que a opção por áreas menos concorridas possibilite o ingresso a um nível educacional, cujo título poderá facilitar a ascensão social.

Neste aspecto, SCHIEFELBEIN (1974) observa que a universidade construiu em seu interior um sistema semelhante ao resto do sistema escolar. Algumas carreiras fazem

parte do sonho da maioria dos candidatos, e chegam a selecionar os mais preparados, seja qual for o critério de seleção. Com o significativo crescimento da iniciativa privada na educação superior brasileira, fatores econômicos ligados ao trabalho e ao estudo podem ser mais decisivos que a qualidade.

Conforme SGANZERLA (2001), as raízes das diferentes formas de abandono são distintas e as ações preventivas para tratarem desses comportamentos também devem ser diferentes. Antes de iniciar programas de manutenção dos estudantes na universidade, é indispensável conhecer as formas de evasão. Não basta saber quem e quantos abandonam, mas o porquê da decisão e avaliar o grau de integração universitário, a fim de buscar o desenvolvimento dos sistemas.

Inúmeros estudos, teses de mestrados, doutorado tentam entender os aspectos comuns entre os estudantes que evadem os cursos superiores, na tentativa de criar uma ferramenta que possa identificar características visando auxiliar a IES a desenvolver programas que ajudem a reduzir os números da evasão.

Neste estudo vamos analisar diferentes variáveis de alunos matriculados no ano de 1998/1999/2000 no curso de Estatística da Universidade Federal do Paraná, notas e freqüências nas 5 disciplinas do 1º semestre do curso, idade, gênero, estado civil, escore do vestibular e outras possíveis. Tentaremos através das informações encontrar o perfil dos estudantes que concluíram ou evadiram o curso, para embasamento de um próximo trabalho que vise diminuir a evasão escolar.

1.2 HIPÓTESES

- O procedimento estatístico de Reconhecimento de Padrões e Classificação é mais eficiente do que o de Regressão Logística
- As co-variáveis nota e frequencia das disciplinas são mais importantes na evasão que as demais.

2. OBJETIVOS

2.1OBJETIVO GERAL

O objetivo geral do trabalho é verificar qual das metodologias estatísticas, neste caso o reconhecimento de Padrões e Classificação e a Regressão Logística têm o melhor desempenho na verificação dos fatores que influenciam na evasão dos alunos do curso de Estatística da UFPR.

2.2OBJETIVO ESPECÍFICO

Verificar quais são as co-variáveis significativas na detecção de alunos que ingressam no curso de Estatística e que já cursaram o 1° semestre venham a desistir ou não do curso.

3. JUSTIFICATIVA:

O curso de estatística da UFPR historicamente tem um número elevado de alunos que não concluem o curso, muitos deles já desistem logo após completarem o 1° semestre, sendo assim este estudo visa tentar detectar as possíveis razões desta evasão, e que num futuro próximo seja possível trabalhar com os aspectos que mais influenciam na evasão para que este tipo de situação ocorra o menor numero de vezes. Como na estatística temos vários métodos disponíveis, escolhemos fazer uma comparação entre a análise multivariada e a regressão logística para ter um método que nos conduza a uma avaliação mais apurado sobre este assunto.

4. MATERIAL E MÉTODOS

4.1 DADOS COLETADOS

Este estudo foi realizado com 163 alunos que ingressaram no curso e estatística da universidade federal do Paraná nos anos de 1998, 1999 e 2000, tendo como objetivo investigar quais as covariáveis apresentam-se significativas para a variável resposta: evasão (54 observações) ou não evasão do curso (109 observações).

4.2 METODOLOGIA ESTATÍSTICA

4.2.1 Regressão Logística

A escolha da técnica estatística a ser utilizada segundo GIOLO (2004) deve ser levada em conta com relação à natureza da variável resposta, neste caso a variável resposta é dicotômica, ou seja, apresenta as categorias evasão do curso ou não evasão do curso, e ainda do ponto de vista matemático, fácil de ser usada e de interpretação bem simples. Sendo assim optou-se pela utilização da regressão logística, na tentativa de encontrar um modelo explicativo da variável resposta em função das variáveis explicativas.

A regressão logística parte da função de distribuição logística que é dada por:

$$F(x)= \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}, \text{ para } x = -\infty,\ldots,+\infty \qquad (1)$$

A função de distribuição logística toma valores entre zero e um, assume valor zero em uma parte do domínio das variáveis explicativas, um em outra parte do domínio e cresce suavemente na parte intermediária possuindo uma particular curva em forma de "S".

O modelo de regressão logística é expresso por:

$$\theta(\underline{x}) = P(Y=1 \mid \underline{x}) = \frac{\exp\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\}}{1+\exp\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\}} \qquad (2)$$

Para descrever a variação entre os $\theta(\underline{x}) = E(Y|\underline{x})$ foi, então, foi proposto a utilização do modelo acima citado, onde $Y_i = 1$ significa a presença da resposta, **$\underline{x}$ é** o vetor que representa as covariáveis (fatores de risco), isto é, $\underline{x} = (x_1, x_2, ..., x_p)$. O parâmetro β_0 é o intercepto e β_k (k=1,..., p) são os p parâmetros da regressão. Nota-se que este modelo retornará uma estimativa da probabilidade do indivíduo ter a resposta dado que o mesmo possui, ou não, determinados fatores de risco. Conseqüentemente,

$$1 - \theta(\underline{x}) = \frac{1}{1 + \exp\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\}} \qquad (3)$$

retornará uma estimativa de probabilidade do indivíduo não ter a resposta dado que o mesmo possui ou não determinados fatores de risco.

Observe, ainda, que fazendo-se:

$$\text{LOG}\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \sum_{K=1}^{P} \beta_k x_k$$

Tem-se um modelo linear para seus parâmetros, e dependendo da variação de $\underline{x}$, pode ser contínuo e variar de $-\infty$ a $+\infty$. A estimação dos parâmetros em regressão logística geralmente é feita pelo método da máxima verossimilhança. Para a aplicação, deste método é necessário construir inicialmente a função de verossimilhança a qual expressa à probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Os estimadores de máxima verossimilhança dos parâmetros serão os valores que maximizam esta função.

Para encontrar esses valores no modelo de regressão logística, considera-se a variável resposta Y codificada como 0 ou 1. Da expressão (2) pode-se obter a probabilidade condicional de que Y=1 dado x, Isto é, $\theta(x) = P(Y=1|x)$ e, que $1 - \theta(x)$ fornece a probabilidade condicional de que Y=0 dado x. Assim, $\theta(x)$ será a contribuição para a função de verossimilhança dos pares (Y_i, x_i) em que $Y_i = 1$ e $1 - \theta(x_i)$, a contribuição dos pares em que $Y_i=0$.

Assumindo então que as observações são independentes tem-se a seguinte expressão: $L(\underline{\beta}) = \prod (\theta(x_i))^{Y_i} (1 - \theta(x_i))^{1-y_i}$ (4)

As estimativas de $\underline{\beta}$ serão os valores que maximizam a função de verossimilhança dada em (4). Algebricamente é mais fácil trabalhar com o logaritmo desta função, isto é, com:

$$l(\underline{\beta}) = \log L(\underline{\beta}) = +\sum_{i=1}^{n} y_i \log(\theta(x_i)) + (1 - y_i) \log(1 - \theta(x_i)) \qquad (5)$$

Para obter os valores de $\underline{\beta}$ que maximizam $l(\underline{\beta})$ basta diferenciar a respectiva função com respeito a cada parâmetro β_j ($j = 0, 1, ..., p$) obtendo-se, assim, o sistema de p+1 equações,

$$\sum_{i=1}^{n}(y_i - \theta(x_i)) = 0$$

$$\sum_{i=1}^{n} x_{ij}(y_i - \theta(x_i)) = 0 \quad j = 1, ..., p$$

que, quando igualadas a zero, produzem como solução as estimativas de máxima verossimilhança de β. Os valores ajustados para o modelo de regressão logístico são, portanto, obtidos substituindo-se as estimativas de β em (2).

4.2.1.1 Estatísticas Qui-quadrado

Nesta estatística os totais marginais n_{+1} e n_{+2} são fixos e, portanto, sob a hipótese nula H_0, de não existência de significância para o estudo, a distribuição de probabilidade associada é a hipergeométrica. Assim o valor esperado de nij é:

$$E(N_{ij} \mid H_0) = \frac{(n_{i+})(n_{+j})}{n} = m_{ij}$$

e a variância:

$$V(N_{ij} \mid H_0) = \frac{(n_{1+})(n_{2+})(n_{+1})(n_{+2})}{n^2(n-1)} = v_{ij}.$$

Para uma amostra suficientemente grande, n_{11} tem aproximadamente uma distribuição normal, o que implica que:

$$Q = \frac{(n_{11} - m_{11})^2}{v_{11}}$$

tem aproximadamente uma distribuição qui-quadrado com um grau de liberdade. Não importa como as linhas e colunas sejam arranjadas, Q assumirá sempre o mesmo valor, uma vez que:

$$|n_{11} - m_{11}| = |n_{ij} - m_{ij}| = \frac{|n_{11}n_{22} - n_{12}n_{21}|}{n}$$

Uma estatística relacionada a Q é a estatística de Pearson dada por:

$$Q_P = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \frac{n}{(n-1)} Q.$$

Se as contagens (freqüências) nas caselas forem suficientemente grandes, Q_P segue uma distribuição qui-quadrado com um grau de liberdade. Ainda, quando n cresce, Q_P e Q convergem. Uma regra útil para determinar o tamanho amostral ade-quado para Q e Q_P é que o valor esperado m_{ij} seja maior do que 5 para todas as caselas.

4.2.1.2 Sensibilidade e Especificidade

Estas medidas determinam a eficiência do modelo selecionado detectar a verdade. A sensibilidade é definida como a proporção de resultados positivos que o estudo apresenta, quando realizado em sujeitos conhecidos terem a doença, ou seja, é a proporção de verdadeiros positivos. A especificidade, por outro lado, é definida como a proporção de resultados negativos que o estudo apresenta, quando realizado em sujeitos conhecidos estarem livres da doença (proporção de verdadeiros negativos). O desejado de um exame (teste) é que ele tenha, simultaneamente, alta sensibilidade e especificidade.

4.2.1.3 Poder Preditivo do Modelo

O poder preditivo do modelo pode também ser obtido com a finalidade de avaliar a qualidade do modelo ajustado. Para isso, faz-se necessário estabelecer uma probabilidade, denominada “ponto de corte", a partir da qual se estabeleça que:

- a variável resposta receba o valor 1, isto _e, Y = 1 para probabilidades estimadas pelo modelo que sejam maiores ou iguais a esse ponto de corte e, ainda, que
- a variável resposta receba o valor 0, isto _e, Y = 0 para probabilidades estimadas pelo modelo que sejam menores do que esse ponto de corte.

4.2.1.4 *Deviance* Residual e Resíduos de Pearson

As estatísticas Q_p e Q_L, são usadas para verificar a qualidade de ajuste do modelo de regressão logística, fornecem um único número o qual resume a concordância entre os

valores observados e os ajustados. PREGIBON (1981) estendeu os métodos de diagnóstico de regressão linear para a regressão logística e argumenta que, como as estatísticas qui-quadrado de Pearson (Q_p) e *deviance* (Q_L) são duas medidas usadas para verificar a qualidade do modelo ajustado, faz sentido analisar os componentes individuais dessas estatísticas, uma vez que estes componentes são funções dos valores observados e preditos pelo modelo.

Assim, se em uma tabela de contingência s x 2, tem-se para cada uma das s linhas n_{i+} sujeitos dos quais n_{i1} apresentam a resposta de interesse (sucesso) e θ_{i1} denota a probabilidade predita de sucesso para a i-ésima linha (grupo), define-se o i-ésimo resíduo por:

$$c_i = \frac{n_{i1} - ((n_{i+})\,\theta_{i1})}{\sqrt{(n_{i+})\,\underline{\theta}_{i1}\,(1 - \underline{\theta}_{i1})}} \quad i = 1; \ldots s.$$

Esses resíduos são conhecidos como resíduos de Pearson, uma vez que a soma deles ao quadrado resulta em Q_P. O exame dos valores residuais c_i auxiliam a determinar quão bem o modelo se ajusta aos grupos individuais.

Freqüentemente, resíduos excedendo o valor |2,0| (ou |2,5|) indicam falta de ajuste. Similarmente, a *deviance* residual é um componente da estatística *deviance* e é expressa por:

$d_i = \text{sinal}(n_{i1} - y_{i1})\,[\,2\,n_{i1}\log(n_{i1}/y_{i1}) + 2(_{ni+} - n_{i1})\log((n_{i+} - n_{i1})/(n_{i+} - y_{i1}))]^{1/2}$, em que $y_{i1} = (n_i+)\,\underline{\theta}_{i1}$. A soma das *deviances* residuais ao quadrado resulta na estatística deviance Q_L. A partir do exame dos resíduos *deviance* pode-se observar a presença de resíduos não usuais (demasiadamente grandes), bem como a presença de outliers ou, ainda, padrões sistemáticos de variação indicando, possivelmente, a escolha de um modelo não muito adequado.

As estatísticas de diagnóstico apresentadas permitem, ao analista, identificar padrões de covariáveis que estão com um ajuste pobre. Após estes padrões serem identificados, pode-se, então, avaliar a importância que eles têm na análise.

4.2.1.5 Gráfico Q-Qplot com Envelope Simulado

No caso em que a variável resposta é assumida ser normalmente distribuída, é comum que afastamentos sérios da distribuição normal sejam verificados por meio do gráfico de probabilidades normal dos resíduos. No contexto de modelos lineares generalizados, em que distribuições diferentes da normal são também consideradas,

gráficos similares com envelopes simulados podem ser também construídos com os resíduos gerados a partir do modelo ajustado. A inclusão do envelope simulado no Q-Qplot auxilia a decidir se os pontos diferem significativamente de uma linha reta, (GIOLO, 2006) apresenta códigos em linguagem Splus, que podem ser utilizados no pacote estatístico R, para gerar tais gráficos em: regressão gama, logística, Poisson e binomial negativa, além da normal. Para que o modelo ajustado seja considerado satisfatório, faz-se necessário que as *deviances* residuais caiam dentro do envelope simulado.

4.2.2 Reconhecimento de Padrões e Classificação

O reconhecimento de padrões, segundo SONKA, HLAVAC & BOYLE (1993), baseia-se na atribuição de classes para os pixels através do processo chamado Análise Discriminante. De acordo com JOHNSON & WICHERN (1988), análise discriminante são técnicas multivariadas interessadas com a separação de uma coleção de objetos (ou observações) distintos e que alocam novos objetos em grupos previamente definidos. A análise discriminante quando empregada como procedimento de classificação não é uma técnica exploratória, uma vez que ela conduz a regras bem distribuídas, as quais podem ser utilizadas para classificação de novos objetos.

As técnicas estatísticas de discriminação e classificação estão incorporadas num contexto mais amplo, que é o do reconhecimento de padrões. Participa junto com técnicas de programação matemática e redes neurais na formação do conjunto de procedimentos usados no reconhecimento e classificação de objetos e indivíduos. Algumas *máquinas inteligentes* são exemplos do que vem a ser reconhecimento de padrões, são elas:

- míssil que escolhe por onde entrar em um abrigo (Guerra do Golfo);
- carro que se desloca sozinho em um *campus* universitário;
- máquina que classifica tábuas de madeira pela sua tonalidade de cor;
- etc.

Estes exemplos refletem o emprego da chamada *inteligência artificial* que consiste, entre outras, de aplicações de técnicas de reconhecimento de padrões usando tecnologia adequada como a câmera de televisão para *visão* e um processador eletrônico como *cérebro*.

Os objetivos imediatos da técnica quando usada para discriminação e classificação são, respectivamente, os seguintes:

1. Descrever algebricamente ou graficamente as características diferenciais dos objetos (observações) de várias populações conhecidas, no sentido de achar “discriminantes”

cujos valores numéricos sejam tais que as populações possam ser separadas tanto quanto possível.

2. Grupar os objetos (observações) dentro de duas ou mais classes determinadas. Tenta-se encontrar uma regra que possa ser usada na alocação ótima de um novo objeto (observação) nas classes consideradas.

Uma função que separa, pode servir como alocadora, e da mesma forma uma regra alocadora, pode sugerir um procedimento discriminatório. Na prática, os objetivos 1 e 2, freqüentemente, sobrepõem-se e a distinção entre separação e alocação torna-se confusa.

A terminologia de "discriminar" e "classificar" foi introduzida por FISCHER (1936) no primeiro tratamento moderno dos problemas de separação.

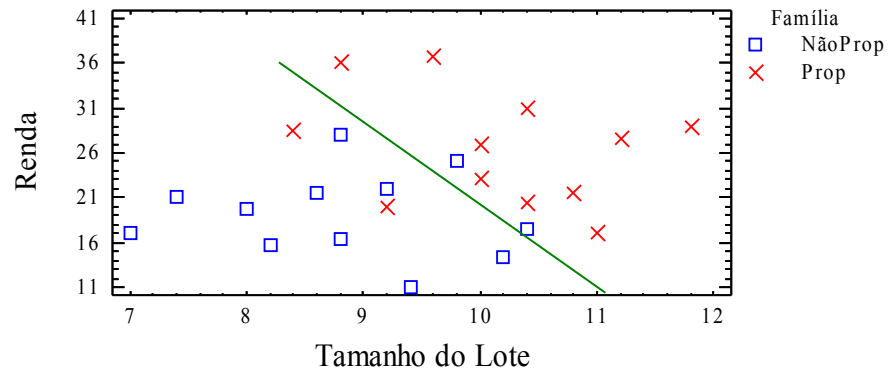
4.2.2.1 Problema Geral de Classificação

Em seu livro, JOHNSON & WICHERN (1988) considera dois grupos em uma cidade, proprietários de certo equipamento e não-proprietários desse equipamento. Afim de identificar o melhor tipo de campanha de vendas, o fabricante do equipamento está interessado em classificar famílias como futuros compradores do equipamento ou não, com base em x_1 = renda e x_2 = tamanho do lote de moradia. Amostras aleatórias de $n_1 = 12$ proprietário e $n_2 = 12$ não-proprietários produziram os dados abaixo.

Tabela 4.1: Amostras de Famílias Proprietárias e Não-Proprietárias com Base na Renda e Tamanho do Lote de suas Moradias.

Π_1 = Proprietários		Π_2 = Não-Proprietários	
x_1	x_2	x_1	x_2
20.0	9.2	25.0	9.8
28.5	8.4	17.6	10.4
21.6	10.8	21.6	8.6
20.5	10.4	14.4	10.2
29.0	11.8	28.0	8.8
36.7	9.6	16.4	8.8
36.0	8.8	19.8	8.0
27.6	11.2	22.0	9.2
23.0	10.0	15.8	8.2
31.0	10.4	11.0	9.4
17.0	11.0	17.0	7.0
27.0	10.0	21.0	7.4

Figura 4.1- Representação Gráfica da Tabela 4.1 no Espaço Discriminante



Observa-se na figura 4.1 que:

1) proprietários tendem a ter maiores rendas e maiores lotes;
2) renda parece discriminar melhor que lote
3) existem mistura entre grupos.

Dado que existe mistura e conseqüentemente classificações erradas, a idéia é criar uma regra (regiões R_1 e R_2) que minimize a chance de fazer esta mistura. Um bom procedimento resultará pouca mistura de elementos grupais. Pode ocorrer que de uma classe ou população exista maior probabilidade de ocorrência do que de outra classe. Uma regra de classificação ótima deve levar em conta as probabilidades de ocorrência a "priori". Outro aspecto da classificação é o custo. Suponha que classificar um item em Π_1 quando na verdade ele pertencente a Π_2 represente um erro mais sério do que classificar em Π_2 quando o item pertencente a Π_1. Então deve-se levar isso em conta.

Seja $f_1(\underline{x})$ e $f_2(\underline{x})$ as f.d.p.'s associadas com o vetor aleatório $\underline{X}$ de dimensão p das populações Π_1 e Π_2, respectivamente. Um objeto, com as medidas $\underset{\sim}{x}$, deve ser reconhecido como de Π_1 ou de Π_2. Seja Ω o espaço amostral, isto é, o conjunto de todas as possíveis observações $\underline{x}$. Seja R_1 o conjunto de valores $\underline{x}$ para os quais nós classificamos o objeto como Π_1 e $R_2 = \Omega - R_1$ os remanescentes valores $\underset{\sim}{x}$ para os quais nós classificaremos os objetos como Π_2. Os conjuntos R_1 e R_2 são mutuamente exclusivos.

Para p = 2, podemos ter a figura:

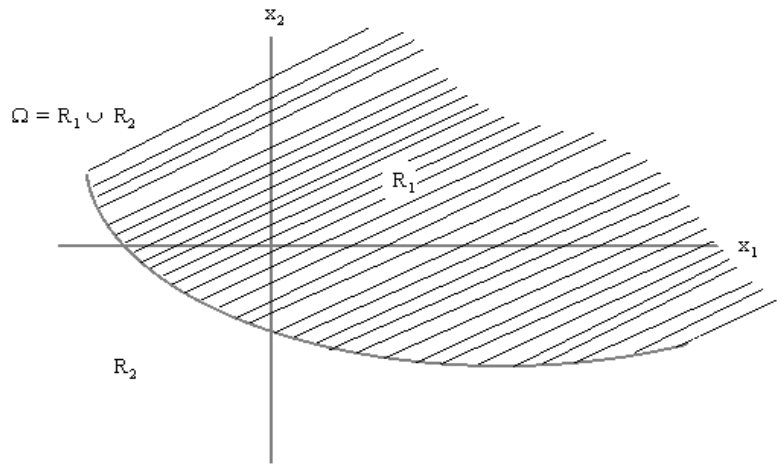


Figura 4.2: Regiões de classificação para duas populações

A probabilidade condicional, de reconhecer um objeto como de Π_2 quando na verdade ele é de Π_1 é: $P(2|1) = P(\underline{X} \in R_2 | \Pi_1) = \int_{R_2 = \Omega - R_1} f_1(\underline{x}) d\underline{x}$

Da mesma forma: $P(1|2) = P(\underline{X} \in R_1 | \Pi_2) = \int_{R_1} f_2(\underline{x}) d\underline{x}$

P(2|1) representa o volume formado pela f.d.p. $f_1(\underline{x})$ na região R_2.

Sendo p = 1 (caso univariado) tem-se:

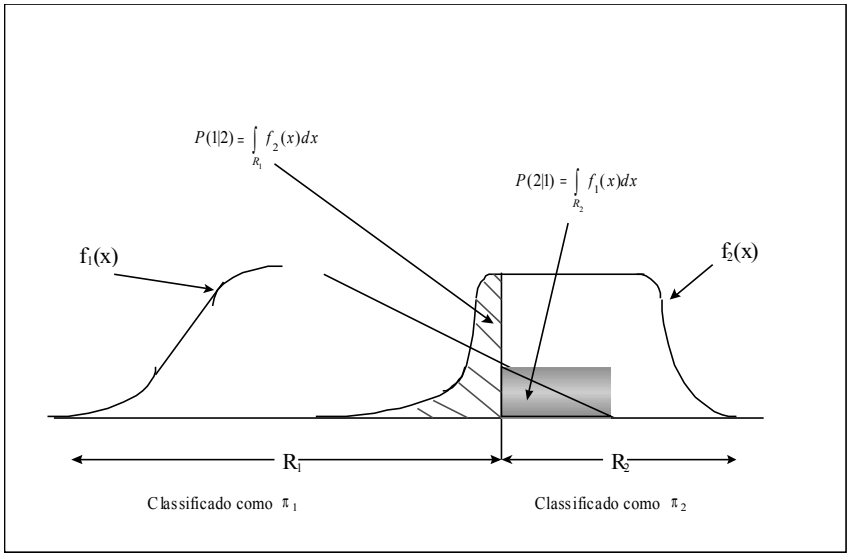


Figura 4.3: Classificação das Regiões para Duas Populações.

Seja p_1 a probabilidade a "priori" de Π_1 e p_2 ser a probabilidade a "priori" de Π_2, onde $p_1 + p_2 = 1$. As probabilidades de reconhecer corretamente ou incorretamente são dados por:

P(rec. correta/te. como Π_1) = P($\underset{\sim}{X} \in \Pi_1$ e é rec. correta/te como Π_1) =

$$= P(\underset{\sim}{X} \in R_1 | \Pi_1)P(\Pi_1) = P(1|1)p_1$$

P(rec. incorreta/te como Π_1) = P($\underset{\sim}{X} \in \Pi_2$ e é rec. incorreta/te como Π_1) =

$$= P(\underset{\sim}{X} \in R_1 | \Pi_2)P(\Pi_2) = P(1|2)p_2$$

P(rec. correta/te como Π_2) = P($\underset{\sim}{X} \in \Pi_2$ e é rec. correta/te como Π_2) = P($\underset{\sim}{X} \in R_2 | \Pi_2$)P($\Pi_2$) = P(2|2)$p_2$

P(rec. incorreta/te como Π_2) = P($\underset{\sim}{X} \in \Pi_1$ e é rec. incorreta/te como Π_2) = P($\underset{\sim}{X} \in R_2 | \Pi_1$)P($\Pi_1$) = P(2|1)$p_1$

Regras de reconhecimento são freqüentemente avaliados em termos de suas probabilidades de reconhecimento errado.

Tabela 4.2: Matriz do Custo de Reconhecimento Errado

		Reconhecimento como	
		Π_1	Π_2
População verdadeira	Π_1	0	c(2|1)
	Π_2	c(1|2)	0

Para qualquer regra, a média, ou o custo esperado de reconhecimento (classificação) errado é dado pela soma dos produtos dos elementos fora da diagonal principal pelas respectivas probabilidades:

ECM = c(2|1)P(2|1)p(1) + c(1|2)P(1|2)p(2)

Uma regra razoável de reconhecimento deve ter ECM muito baixa, tanto quanto possível.

As regiões R_1 e R_2 que minimizam o ECM são definidas pelos valores de $\underset{\sim}{x}$ tal que valem as desigualdades:

$$R_1 = \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq \left[\frac{c(1|2)}{c(2|1)}\right] \cdot \left[\frac{p_2}{p1}\right]$$

$$\left[\begin{matrix}\text{Razão das} \\ \text{densidades}\end{matrix}\right] \geq \left[\begin{matrix}\text{Razão dos} \\ \text{custos}\end{matrix}\right] \cdot \left[\begin{matrix}\text{Razão das} \\ \text{probabilidades `a priori}\end{matrix}\right]$$

$$R_2 = \frac{f_1(\underline{x})}{f_2(\underline{x})} < \left[\frac{c(1|2)}{c(2|1)}\right] \cdot \left[\frac{p_2}{pl}\right]$$

$$\left[\begin{array}{c}\text{Razão das}\\ \text{densidades}\end{array}\right] < \left[\begin{array}{c}\text{Razão dos}\\ \text{custos}\end{array}\right] \cdot \left[\begin{array}{c}\text{Razão das}\\ \text{probabilidades `a priori}\end{array}\right]$$

4.2.2.2. Critério TPM

Outro critério, além do ECM, pode ser usado para construir procedimentos ótimos. Assim, pode-se ignorar o ECM e escolher R_1 e R_2 que minimizam a probabilidade total de erro de classificação (TPM).

TPM = P($\underset{\sim}{x} \in \Pi_1$ e é classificada errada) + P($\underset{\sim}{x} \in \Pi_2$ e é classificada errada)

$$\text{TPM} = p_1 \int_{R_2} f_1(\underset{\sim}{x}) d\,\underset{\sim}{x} + p_2 \int_{R_1} f_2(\underset{\sim}{x}) d\,\underset{\sim}{x}$$

Matematicamente, isto é equivalente a minimizar ECM quando os custos de classificação errada são iguais. Assim, podemos alocar uma nova observação $\underset{\sim}{x_0}$ para a população com a maior probabilidade posteriori $P(\Pi_i|\underset{\sim}{x_0})$, onde

$$P(\Pi_1|\underset{\sim}{x_0}) = \frac{P(\Pi_1 \text{ ocore e observa - se } \underset{\sim}{x_0})}{P(\text{observa - se } \underset{\sim}{x_0})}$$

$$= \frac{P(\text{observa - se } \underset{\sim}{x_0}|\Pi_1)P(\Pi_1)}{P(\text{observa - se } \underset{\sim}{x_0}|\Pi_1)p(\Pi_1) + P(\text{observa - se } \underset{\sim}{x_0}|\Pi_2)p(\Pi_2)}$$

$$= \frac{p_1 f_1(\underset{\sim}{x_0})}{p_1 f_1(\underset{\sim}{x_0}) + p_2 f_2(\underset{\sim}{x_0})}$$

e $P(\Pi_2|\underset{\sim}{x_0}) = 1 - P(\Pi_1|\underset{\sim}{x_0}) = \frac{p_2 f_2(\underset{\sim}{x_0})}{p_1 f_1(\underset{\sim}{x_0}) + p_2 f_2(\underset{\sim}{x_0})}$

e classifica-se $\underset{\sim}{x_0}$ em Π_1 quando $P(\Pi_1|\underset{\sim}{x_0}) > P(\Pi_2|\underset{\sim}{x_0})$

4.2.2.3. Classificação com Duas Populações Normais Multivariadas

Assume-se que $f_1(\underset{\sim}{x})$ e $f_2(\underset{\sim}{x})$ são densidades normais multivariadas, a primeira com $\underset{\sim}{\mu_1}$ Σ_1 e a segunda com $\underset{\sim}{\mu_2}$ Σ_2, então supondo $\Sigma_1 = \Sigma_2$, a F.D.L. de Fisher pode ser usada para classificação e corresponde a um caso particular da regra de classificação com base em ECM.

Assim, Seja $\underset{\sim}{X}' = [\, X_1, X_2, ..., X_p \,]$ para populações Π_1 e Π_2 e

$$f_i(\underset{\sim}{x}): \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\underset{\sim}{x} - \underset{\sim}{\mu_i} \right)' \Sigma^{-1} \left(\underset{\sim}{x} - \underset{\sim}{\mu_i} \right) \right] \qquad \text{para i=1,2}$$

Suponha que os parâmetros Π_1 e Π_2 e Σ, são conhecidos, tem-se as regiões de mínimo ECM.

$$R_1: \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = \frac{\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (\underset{\sim}{x} - \underset{\sim}{\mu_i})' \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu_i}) \right]}{\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (\underset{\sim}{x} - \underset{\sim}{\mu_i})' \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu_i}) \right]} =$$

$$= \exp\left[-\frac{1}{2} \left(\underset{\sim}{x} - \underset{\sim}{\mu_1} \right)' \Sigma^{-1} \left(\underset{\sim}{x} - \underset{\sim}{\mu_1} \right) + \frac{1}{2} \left(\underset{\sim}{x} - \underset{\sim}{\mu_2} \right)' \Sigma^{-1} \left(\underset{\sim}{x} - \underset{\sim}{\mu_2} \right) \right] \geq \left[\frac{c(1\,2)}{c(2\,1)} \right] \left[\frac{p_2}{p_1} \right]$$

$$R_2: \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} =$$

$$= \exp\left[-\frac{1}{2} \left(\underset{\sim}{x} - \underset{\sim}{\mu_1} \right)' \Sigma^{-1} \left(\underset{\sim}{x} - \underset{\sim}{\mu_1} \right) + \frac{1}{2} \left(\underset{\sim}{x} - \underset{\sim}{\mu_2} \right)' \Sigma^{-1} \left(\underset{\sim}{x} - \underset{\sim}{\mu_2} \right) \right] < \left[\frac{c(1\,2)}{c(2\,1)} \right] \left[\frac{p_2}{p_1} \right]$$

Sejam as populações Π_1 e Π_2 normais multivariadas. A regra de reconhecimento que minimiza ECM é dada por: reconhecer $\underset{\sim}{x_0}$ como sendo de Π_1 se

$$\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\underset{\sim}{x}_0 - \frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\left(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2\right) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

e $\underset{\sim}{x}_0$ como sendo de Π_2 em caso contrário.

Em situações em que $\underset{\sim}{\mu}_i$ i=1,2 são desconhecidas e Σ também, a regra deve ser modificada. Segundo Anderson (1984) tem-se a seguinte regra do ECM mínimo para duas populações normais (regra amostral).

Alocar $\underset{\sim}{x}_0$ em Π_1 se

$$\left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)'S_p^{-1}\underset{\sim}{x}_0 - \frac{1}{2}\left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)'S_p^{-1}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2\right) \geq \ln\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]$$

Alocar $\underset{\sim}{x}_0$ em Π_2 em caso contrário.

O primeiro termo da regra de classificação e reconhecimento, $\left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)'S_p^{-1}\underset{\sim}{x}$, é a função linear obtida por Fisher que maximiza a variabilidade univariada entre as amostras relativamente a variabilidade dentro das amostras. A expressão inteira

$$w = \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)'S_p^{-1}\underset{\sim}{x} - \frac{1}{2}\left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)'S_p^{-1}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2\right) = \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)'S_p^{-1}\left[\underset{\sim}{x} - \frac{1}{2}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2\right)\right]$$ é

conhecida como função de classificação de Anderson(1984).

Quando $\sum_1 \neq \sum_2$, tem-se a classificação quadrática.

Supondo as matrizes de covariância $\sum_1$ para $\underset{\sim}{x} \in \Pi_1$ e Σ_2 para $\underset{\sim}{x} \in \Pi_2$ em $\sum_1 \neq \sum_2$, as regras de reconhecimento de padrões tornam-se mais complicadas.

Seja então $\underset{\sim}{x} \sim N_p(\underset{\sim}{\mu}_i, \Sigma_i)$ i=1,2 com $\underset{\sim}{\mu}_1 \neq \mu_2$ e $\underset{\sim}{\Sigma}_1 \neq \Sigma_2$. A probabilidade total de reconhecimento errada (TPM) e o custo esperado de reconhecimento errado dependem da razão de densidades

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})}$$

ou, equivalentemente, do logaritmo das razões das densidades

$$\ln\left[\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})}\right] = \ln\left[f_1(\underset{\sim}{x})\right] - \ln\left[f_2(\underset{\sim}{x})\right]$$

Sejam as populações Π_1 e Π_2 descritas por densidades normais multivariadas $N_p(\underset{\sim}{\mu}_1, \Sigma_1)$ e $N_p(\underset{\sim}{\mu}_2, \Sigma_2)$. Então a regra de reconhecimento que minimiza o ECM é dado por: $R_1 = -\frac{1}{2}\underset{\sim}{x}_0'\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\underset{\sim}{x}_0 + \left(\underset{\sim}{\mu}_1'\Sigma_1^{-1} - \underset{\sim}{\mu}_2'\Sigma_2^{-1}\right)\underset{\sim}{x}_0 - k$.

Substituindo as expressões das densidades normais multivariadas tem-se:

$$R_1 = \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = \frac{\frac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}}\exp\left[-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)\right]}{\frac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}}\exp\left[-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right]} \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Na prática, a regra de reconhecimento estabelecida é implementadas substituindo-se os parâmetros $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \Sigma_1$ e Σ_2 pelas suas estimativas $\underset{\sim}{\bar{x}}_1, \underset{\sim}{\bar{x}}_2, S_1$ e S_2, tal que:

Alocamos $\underset{\sim}{x}_0$ em Π_1 se:

$$-\frac{1}{2}\underset{\sim}{x}_0'\left(S_1^{-1} - S_2^{-1}\right)\underset{\sim}{x}_0 + \left(\underset{\sim}{\bar{x}}_1'S_1^{-1} - \underset{\sim}{\bar{x}}_2'S_2^{-1}\right)\underset{\sim}{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Alocamos $\underset{\sim}{x}_0$ em Π_2, caso contrário.

4.2.2.4 Discriminação e Classificação entre Duas Populações - Método de Fischer

Basicamente, o problema consiste em separar duas classes de objetos ou fixar um novo objeto em uma das duas classes. Deste modo, é interessante alguma exemplificação.

A tabela I a seguir mostra diversas situações onde a análise discriminante pode ser empregada. É comum denominar as classes (populações) de Π_1 e Π_2, e os objetos separados ou classificados com base nas medidas de p variáveis aleatórias são associadas com vetores do tipo:

$$\underset{\sim}{X}' = [X_1, X_2, \ldots, X_P],$$

onde as variáveis X_i, i = 1 , 2 , ... , p, são as medidas das características investigadas nos objetos.

Os valores observados de $\underset{\sim}{X}$ podem diferir de uma classe para outra, sendo que a totalidade dos valores da 1ª classe é a população dos valores $\underset{\sim}{x}$ para Π_1 e aqueles da 2ª classe é a população dos valores de $\underset{\sim}{X}$ para Π_2. Assim, estas populações podem ser descritas pelas funções densidade de probabilidade $f_1(\underset{\sim}{x})$ e $f_2(\underset{\sim}{x})$.

Tabela 4.3: Situações - Exemplos

Populações Π_1 e Π_2	Variáveis medidas (componentes de $\underset{\sim}{x}$)
1. Sucesso ou insucesso de estudantes na Universidade.	-Notas no vestibular, notas no curso, número de disciplinas no curso.
2. Machos e fêmeas adultos.	-Altura, peso, perímetro do bíceps, perímetro do tórax, perímetro do quadril.
3. Comprador de um novo produto e não comprador de um novo produto.	-Educação, renda, tamanho da família.
4. Artigos jornalísticos escritos por Paulo Francis e Carlos Castelo Branco.	-Freqüência de diferentes palavras, comprimento das sentenças.
5. Pessoa de alto risco no crédito e pessoa de baixo risco.	-Renda, idade, número de cartões de crédito, tamanho da família.
6. Duas espécies de planta semelhantes.	-Comprimento da pétala, profundidade da fenda da pétala, diâmetro do pólen.

A idéia de Fischer foi transformar as observações multivariadas $\underset{\sim}{X}$'s nas observações univariadas y's tal que os y's das populações Π_1 e Π_2 sejam separadas tanto quanto possível. Fischer teve a idéia de tomar combinações lineares de $\underset{\sim}{X}$ para criar os y's, dado que as combinações lineares são funções de $\underset{\sim}{X}$ e por outro lado são de fácil cálculo

matemático.

Seja μ_{1y} a média dos y's obtidos dos $\underset{\sim}{X}$'s pertencentes a Π_1 e μ_{2y} a média dos y's obtidos dos $\underset{\sim}{X}$'s pertencentes a Π_2, então Fischer selecionou a combinação linear que maximiza a distância quadrática entre μ_{1y} e μ_{2y} relativamente à variabilidade dos y's. Assim, seja:

$\underset{\sim}{\mu}_1 = E(\underset{\sim}{X}|\Pi_1)$ = valor esperado de uma observação multivariada de Π_1.

$\underset{\sim}{\mu}_2 = E(\underset{\sim}{X}|\Pi_2)$ = valor esperado de uma observação multivariada de Π_2.

e supondo a matriz de covariância

$$\Sigma = E(\underset{\sim}{X} - \underset{\sim}{\mu}_i)(\underset{\sim}{X} - \underset{\sim}{\mu}_i)' \quad i = 1, 2$$

como sendo a mesma para ambas as populações, e considerando a C.L.

$$\underset{1x1}{Y} = \underset{1xp}{\underset{\sim}{c}'} \underset{px1}{\underset{\sim}{X}}$$

tem-se

$$\mu_{1y} = E(Y|\Pi_1) = E(\underset{\sim}{c}'\underset{\sim}{X}|\Pi_1) = \underset{\sim}{c}'E(\underset{\sim}{X}|\Pi_1) = \underset{\sim}{c}'\underset{\sim}{\mu}_1 \quad ,$$

$$\mu_{2y} = E(Y|\Pi_2) = E(\underset{\sim}{c}'\underset{\sim}{X}|\Pi_2) = \underset{\sim}{c}'E(\underset{\sim}{X}|\Pi_2) = \underset{\sim}{c}'\underset{\sim}{\mu}_2 \quad ,$$

e

$$V(Y) = \sigma_y 2 = V(\underset{\sim}{c}'\underset{\sim}{X}) = \underset{\sim}{c}'V(\underset{\sim}{X})\underset{\sim}{c} = \underset{\sim}{c}'\Sigma\underset{\sim}{c} \quad ,$$

que são a mesma para ambas as populações. Segundo Fischer, a melhor combinação linear é a derivada da razão entre o "quadrado da distância entre as médias" e a "variância de Y".

$$\frac{(\mu_{1y} - \mu_{2y})^2}{\sigma_y^2} = \frac{(\underset{\sim}{c}'\underset{\sim}{\mu}_1 - \underset{\sim}{c}'\underset{\sim}{\mu}_2)^2}{\underset{\sim}{c}'\sum\underset{\sim}{c}} = \frac{\underset{\sim}{c}'(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\underset{\sim}{c}}{\underset{\sim}{c}'\sum\underset{\sim}{c}} = \frac{(\underset{\sim}{c}'\underset{\sim}{\delta})^2}{\underset{\sim}{c}'\sum\underset{\sim}{c}}$$

onde $\underset{\sim}{\delta} = \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$.

Seja $\underset{\sim}{\delta} = \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ e $Y' = \underset{\sim}{c}'\underset{\sim}{X}$, então $\frac{(\underset{\sim}{c}'\underset{\sim}{\delta})^2}{\underset{\sim}{c}'\sum\underset{\sim}{c}}$ é maximizada por $\underset{\sim}{c} = k\,\Sigma\text{-}1\,\underset{\sim}{\delta} = k$

$\Sigma\,1\,(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ para qualquer $k \neq 0$. Escolhendo $k = 1$ tem-se $\underset{\sim}{c} = \Sigma\text{-}1\,(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ e

$$Y = \underset{\sim}{c}'\underset{\sim}{X} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma\text{-}1\,\underset{\sim}{X} \quad ,$$

que é conhecida como ***função discriminante linear de Fischer***.

A função discriminante linear de Fischer transforma as populações multivariadas Π_1 e Π_2 em populações univariadas, tais que as médias das populações univariadas correspondentes sejam separadas tanto quanto possível relativamente a variância populacional, considerada comum.

Assim tomando-se

$$y_0 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\,\underset{\sim}{x}_0$$

como o valor da função discriminante de Fischer para uma nova observação $\underset{\sim}{x}_0$, e considerando o ponto médio entre as médias das duas populações univariadas,

$$m = \frac{1}{2}(\mu_{1y} - \mu_{2y}),$$

como

$$m = \frac{1}{2}(\underset{\sim}{c}'_1\,\underset{\sim}{\mu}_1 + \underset{\sim}{c}'_2\,\underset{\sim}{\mu}_2)$$

$$m = \frac{1}{2}\left[(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{\mu}_1 + (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{\mu}_2\right]$$

$$m = \frac{1}{2}\left[(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)\right] \quad ,$$

tem-se que:

$$E(Y_0|\Pi_1) - m \geq 0$$

$$E(Y_0|\Pi_2) - m < 0 \ ,$$

ou seja, se $\underset{\sim}{X}_0$ pertence a Π_1 , se espera que Y_0 seja igual ou maior do que o ponto médio. Por outro lado se $\underset{\sim}{X}_0$ pertence a Π_2, o valor esperado de Y_0 será menor que o ponto médio. Desta forma a regra de classificação é :

- alocar $\underset{\sim}{x}_0$ em Π_1 se $y_0 - m \geq 0$
- alocar $\underset{\sim}{x}_0$ em Π_2 se $y_0 - m < 0$

Geralmente, os parâmetros $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ e Σ são desconhecidos, então supondo que se tenha n_1 observações da v.a. multivariada X_1 da população Π_1 e n_2 observações da v.a. multivariada X_2 da população Π_2, então os resultados amostrais para aquelas quantidades são:

$$\underset{\sim}{\bar{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\underset{\sim i1}{x}\ ; S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(\underset{\sim i1}{x} - \underset{\sim}{\bar{x}}_1)(\underset{\sim i1}{x} - \underset{\sim}{\bar{x}}_1)'$$

$$\underset{\sim}{\bar{x}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \underset{\sim}{x}_{i2} \; ; S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} (\underset{\sim}{x}_{i2} - \underset{\sim}{\bar{x}}_2)(\underset{\sim}{x}_{i2} - \underset{\sim}{\bar{x}}_2)'$$

mas uma vez que se assuma que as populações sejam assemelhadas é natural considerar a variância como a mesma, daí estima-se a matriz de covariância comum Σ por:

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

que é um estimador não-viciado daquele parâmetro.

conseqüentemente, a função discriminante linear de Fischer amostral é dada por:

$$y = \underset{\sim}{\hat{c}}' \underset{\sim}{x} = (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S_p^{-1} \underset{\sim}{x}$$

a estimativa do ponto médio entre as duas médias amostrais univariadas $\bar{y}_1 = \underset{\sim}{\hat{c}}' \underset{\sim}{\bar{x}}_1$ e

$\bar{y}_2 = \underset{\sim}{\hat{c}}' \underset{\sim}{\bar{x}}_2$ é dada por:

$$\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}\left[(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S_p^{-1} \underset{\sim}{\bar{x}}_1 + (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S_p^{-1} \underset{\sim}{\bar{x}}_2 \right]$$

$$\hat{m} = \frac{1}{2}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S_p^{-1} (\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2)$$

e finalmente a regra de classificação é a seguinte:

- alocar $\underset{\sim}{x}_0$ em Π_1 se $y_0 = (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S_p^{-1} \underset{\sim}{x}_0 \geq \hat{m}$
- alocar $\underset{\sim}{x}_0$ em Π_2 se $y_0 = (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S_p^{-1} \underset{\sim}{x}_0 < \hat{m}$

ou melhor se $y_0 - \hat{m} \geq 0$ $\underset{\sim}{x}_0$ é alocado em Π_1

$y_0 - \hat{m} < 0$ $\underset{\sim}{x}_0$ é alocado em Π_2

A combinação linear particular $y = \underset{\sim}{\hat{c}}' \underset{\sim}{x} = (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S_p^{-1} \underset{\sim}{x}$ maximiza a razão:

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{S_y^2} = \frac{(\underset{\sim}{\hat{c}}' \underset{\sim}{\bar{x}}_1 - \underset{\sim}{\hat{c}}' \underset{\sim}{\bar{x}}_2)^2}{\underset{\sim}{\hat{c}}' S_p \underset{\sim}{\hat{c}}} = \frac{(\underset{\sim}{\hat{c}}' \underset{\sim}{d})^2}{\underset{\sim}{\hat{c}}' S_p \underset{\sim}{\hat{c}}}$$

onde $\underset{\sim}{d} = \underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2$ e $S_y^2 = \dfrac{\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{i2} - \bar{y}_2)^2}{n_1 - n_2 - 2}$

4.3 RECURSOS COMPUTACIONAIS

A análise dos dados será feita através dos softwares R e Statgraphics Centurion, sendo que serão utilizadas: regressão logística,estatística qui-quadrado, *deviance* residual, resíduos de Pearson, *Q-Qplot* com envelope simulado, sensibilidade, especificidade e poder preditivo do modelo, reconhecimento de padrões e classificação.

5. CRONOGRAMA DE ATIVIDADES

	Março / 2008																														
Atividades	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Definição de tema			x	x	x	x	x																								
Revisão de literatura								x	x	x	x	x	x	x	x	x	x	x													
Redação Projeto de Monografia													x	x	x	x	x	x	x	x	x	x	x	x							
Tabulação																						x	x	x	x	x	x	x			
Análise de Regressão Logistica																													x	x	x

	Abril / 2008																														
Atividades	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
Análise de Regressão Logistica	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	xx												
Análise Reconhec. Padrões				x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Redação Apresentação preliminar																												x	x	x	

	Maio / 2008																															
Atividades	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Análise Reconhec. Padrões	x	x	x																												
Redação Apresentação preliminar	x	x	x	x	x	x	x	x	x																						
Apresentação Preliminar												x	x	x	x	x															
Redação da Monografia																	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

	Junho / 2008																														
Atividades	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
Redação da Monografia	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x					
Entrega Monografia																															

6. REFERÊNCIAS BIBLIOGRÁFICAS

CASTRO, O fenômeno da evasão escolar na educação superior no Brasil, 2005.

Disponível em http://www.iesalc.unesco.org.ve/programas/Deserci%C3%B3n/Informe%20Deserci%C3%B3n%20Brasil%20-%20D%C3%A9bora%20Niquini.pdf acessado 06/03/07 as 16:10 min

FISHER, R.A., The statistical utilizacion of multiple measurements, **Annals os Eugenics**. 8 (1938), 376-386.

GIOLO, Suely Ruiz. Apostila de Análise de Regressão, 2003.

GIOLO, Suely Ruiz. Apostila de Análise de Dados Discretos, 2004.

GIOLO, Suely Ruiz. Introdução a Análise de Dados Categóricos, 2006.

JOHNSON, R. A., WICHERN, D.W. **Applied Multivariate Statistical Analysis.** Prentice Hall International, Inc. New Jersey, 1988.

MULLER, Sonia I. M. Gama. Sistema Integrado de Avaliação com Aplicação na Engenharia, 2007.

NETO, Anselmo Chaves. Apostila de Análise Multivariada II, 2005.

PREGIBON, D. Logistic regression diagnostics, Annals of Statistics, v.9, 1981.

SONKA, Milan; HLAVAC, Vaclav & BYLE, Roger. **Image Processing, Analysis and Machine Vision**. Ed. Chapman & Hall, Cambridge, 1993.