ANÁLISE DISCRIMINANTE MÚLTIPLA
O que é, para que serve e como se faz.

Autores: **Istvan Karoly Kasznar, *PhD***

Professor Titular da FGV e

Presidente da IBCI

Bento Mario Lages Gonçalves, *MSc*
Consultor Sênior da IBCI

IBCI – Institutional Business Consultoria Internacional
Rua da Alfândega, 25, sala 805/806, Centro, Rio de Janeiro, RJ.
Telefax: (21)2263-7017 / 2233-8552 / 2263-0563
e-mail: ibci@ibci.com.br
Home Page: www.ibci.com.br

ANÁLISE DISCRIMINANTE MÚLTIPLA

1- Introdução

As relações e funções resultantes não possuem características inéditas; pelo contrário, trata-se de metodologia estatística multivariada consagrada que busca exclusivamente interpretar as relações entre inúmeras variáveis (observações) ao longo do tempo. A utilização extensiva de medidas estatísticas e derivações matemáticas devem-se exclusivamente à :

- Necessidade de se buscar um conjunto reduzido de variáveis explicativas que possa introduzir uma redução estrutural do modelo.

- Necessidade de se ordenar e agrupar um conjunto de variáveis visando a sua classificação em grupos homogêneos.

- Investigar o grau de dependência entre as variáveis.

2 - Conceito de Análise Discriminante

A análise discriminante é uma "ferramenta" estatística utilizada para classificar um determinado elemento (*E*) num determinado grupo de variáveis; entre os diversos grupos existentes π_1, π_2, π_3,.... π_i. Para tal é necessário que o elemento (*E*) a ser classificado pertença realmente a um dos *i* grupos, e que sejam conhecidas as características dos elementos dos diversos grupos. Essas características são especificadas a partir de *n* variáveis aleatórias (X_1 , X_2 , X_3 ,..., X_n). No processo de classificação consideram-se os custos decorrentes de eventuais erros de classificação, bem como as probabilidades "a priori" de que o elemento pertença a cada um dos grupos.

Como exemplo, considere-se uma agência que queremos classificar. Inicialmente desconhecemos a sua condição de complexidade. Para fins de exemplo, vamos supor que os únicos indicadores financeiros existentes sejam os de *inadimplência* [Créditos em Liquidação (CL) / Total de Operações de Crédito (OC)] e de *rentabilidade* [Resultado Financeiro (RF) / Ativo Total (AT)]. Dessa forma calculamos os índices de inadimplência e de rentabilidade para a agência que desejamos classificar e comparamos com um conjunto de agências com índices de elevada inadimplência e outro conjunto de agências com baixos índices de

inadimplência, com a finalidade de discriminar a agência através dos índices, classificando-a num dos dois grupos. A análise discriminante múltipla consiste em

estabelecer o melhor critério de classificação, tendo em vista minimizar as consequências do erro de discriminação, isto é, evitar que uma agência com baixa inadimplência seja classificada como de alta inadimplência e vice-versa.

Cabe enfatizar que uma das vantagens do uso de análise discriminante múltipla é que os pesos a serem atribuídos aos índices – ou coeficientes técnicos (α, β, ou ω) são determinados por cálculos e processos estatísticos, o que exclui a subjetividade ou mesmo o estado de espírito do analista no instante da análise.

Assim, como já citado em nosso exemplo, estamos considerando dois grupos de agências : um composto de agências com elevado grau de inadimplência e outro, com agências com baixos índices de inadimplência. Cada um desses grupos constitui uma população que denominamos π_1e π_2. De cada população, tomamos uma amostra, conforme as Tabelas A.1 e A2.

Tabela A.1 – Amostra da População de Agências com Baixa Inadimplência (π_1)

Agências E_{ie}	X_1 – CL/OC	X_2 – RF/AT
E_{11}	1,34	0,24
E_{12}	1,21	0,20
E_{13}	1,48	0,36
E_{14}	0,81	0,15
E_{15}	1,15	0,21
E_{16}	0,66	0,20
E_{17}	0,73	0,17
E_{18}	0,69	0,29
E_{19}	1,53	0,17
E_{110}	0,30	0,12
$\sum$	**9,90**	**2,11**

Tabela A.2 – Amostra da População de Agências com Elevada Inadimplência (π_2)

Agências E_{ie}	X_1 – CL/OC	X_2 – RF/AT
E_{21}	7,45	-0,14
E_{22}	3,21	-0,02
E_{23}	4,27	0,06
E_{24}	1,85	-0,08
E_{25}	1,45	0,11
E_{26}	9,25	-0,62
E_{27}	2,76	0,25
E_{28}	3,54	0,01
E_{29}	4,88	0,25
E_{210}	4,41	0,08
$\sum$	**43,07**	**-0,10**

Conforme se pode observar nas Tabelas A.1 e A.2 cada amostra das populações π_1e π_2 é composta por 10 agências (E_{ie}), onde i = 1,2 identifica a população, enquanto que e = 1,.2,....,10 identifica o indivíduo, isto é, a agência dentro da amostra.

A cada agência estão associados dois índicadores de complexidade : X_1, que representa a *inadimplência*, e X_2, que representa a *rentabilidade*.

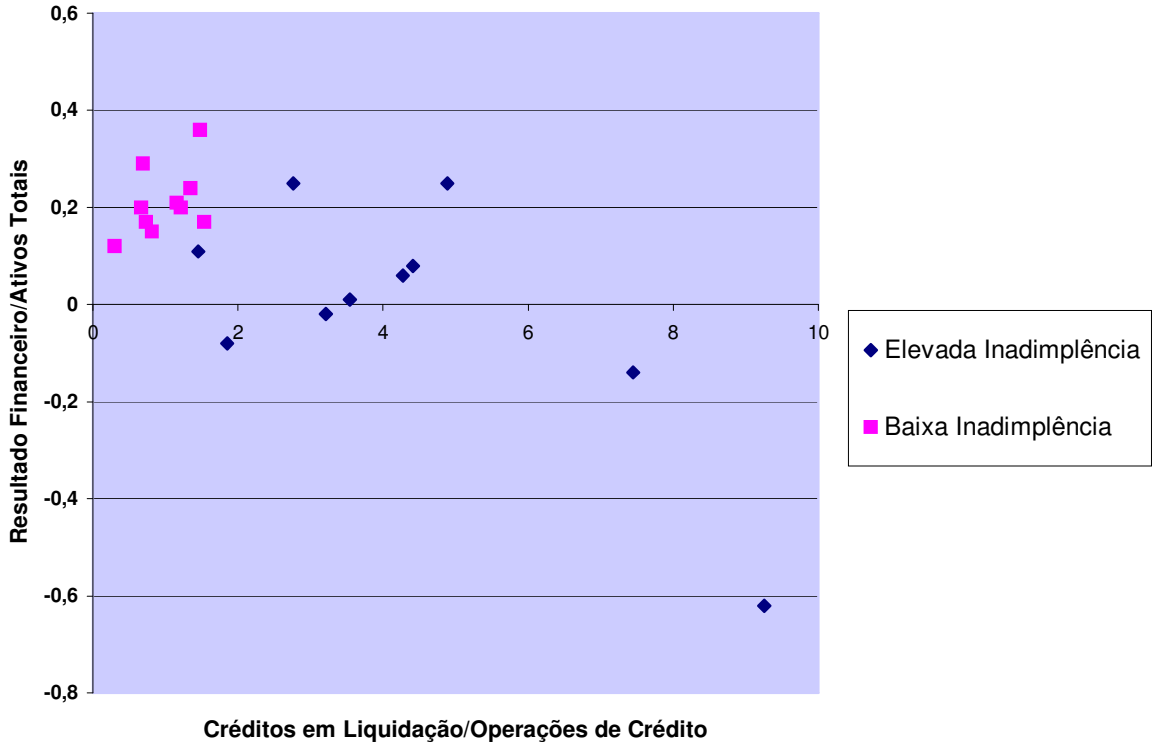
Pelas Tabelas A.1 e A.2, observa-se que os indicadores de inadimplência apresentam valores maiores para as agências da amostra de elevada inadimplência, enquanto que os índices de rentabilidade, de uma forma geral, são maiores para as agências com baixa inadimplência. Apesar da aparência óbvia, esse comportamento dos índices, no exemplo, caracteriza uma forma de discriminação.

Cabe enfatizar também, que o exemplo considera apenas dois grupos de agências, mas, como se trata de uma análise multivariada, pode-se considerar um número superior de grupos.

O ponto de partida do foi um longo e rigoroso processo de identificação, análise qualitativa e quantitativa, e seleção operacional básica de dados, referente aos 1027 municipios do Brasil (universo amostral). A ADM, neste caso, serve como instrumental para a determinação de quais variáveis podem ser utilizadas para que sejam as mais representativas no que se refere à caracterização de Municipios (Praças) bancárias.

Neste âmbito, um grupo técnico normalmente identifica o que deveria ser visto como dado relevante para caracterizar os dados de pesquisa. Neste caso particular, um grupo de fato debateu, selecionou e verificou (em certos casos ainda que em caráter preliminar), quais variáveis poderiam, em sendo disponíveis no Banco de Dados, prever melhor as características de "semelhança" entre grupos.

Ao plotar-se os dados do nosso exemplo, incluídos nas Tabelas A 1 e A.2, num Gráfico de Dispersão (*Scatter Graphic*) obtém-se a seguinte representação gráfica :



Do gráfico resultante, observa-se que as agências dos grupos π_1e π_2 se situam em regiões distintas, enquanto as agências com um nível reduzido de inadimplência se localizam numa região caracterizada por baixa inadimplência e rentabilidade elevada; a situação do outro grupo é exatamente inversa.

3 – A Função Discriminante de Fisher

A função discriminante de Fisher é tida como a primeira solução específica para o problema da discriminação, assim como a própria análise discriminante durante muito tempo se resumiu ao uso dessa função.

Para as situações de discriminação entre duas populações normais de mesma matriz de Covariância, a função discriminante de Fisher apresenta propriedades ótimas. Para o escopo deste trabalho, uma breve apresentação da função discriminante de Fisher, cujo artigo original data de 1936, é o suficiente para uma idéia genérica do que seja esse instrumento estatístico. Cabe, no entanto, frizar que após Fisher a análise discriminante evoluiu com a contribuição de outros trabalhos.

A idéia básica de Fisher foi transformar observações multivariadas *X* em observações univariadas *Y* derivadas das populações π_1 e π_2 aonde estas apresentassem o maior grau de separação (Desvio Padrão) possível. Fisher sugere tomar-se combinações lineares de *X* para criar-se *Y*'s porque tais combinações podem ser facilmente manipuladas, não justificando o porque da escolha de uma função discriminante linear.

Usando nosso exemplo de agências e índices de desempenho, podemos dizer que a função discriminante é uma combinação linear dos índices de inadimplência (X_1) e de rentabilidade (X_2), isto é :

$$Z = aX_1 + bX_2$$

Onde *a* e *b* são determinados de forma a maximizar o quosciente entre a diferença ao quadrado entre os valores de *Z* calculados para a média das amostras (π_1 e π_2) e a variância de *Z* estimada dentro das amostras, o que é equivalente a :

$$\frac{(\overline{Z}_1 - \overline{Z}_2)^2}{\sum_i (Z_{1i} - \overline{Z}_1)^2 + \sum_i (Z_{2i} - \overline{Z}_2)^2}$$

Daí o que se procura é uma função *Z* que maximize a "distância" entre as populações π_1 e π_2. A maximização deste quosciente leva à resolução de um sistema de equações lineares em *a* e *b*. A solução (*a,b*) deste sistema define a função $Z = aX_1 + bX_2$ que atenda ao objetivo. O sistema é :

$$\begin{cases} aS_{11} + bS_{12} = D_1 \\ aS_{12} + bS_{22} = D_2 \end{cases}$$

Sendo *X* o valor da variável *X* (no caso, inadimplência ou rentabilidade) associada a um e elemento (agência) da amostra da população (de agências com "baixo" ou "alto" grau de inadimplência), temos :

i = 1,2 (Variável)
j = 1,2 (População)
e = 1,.....,10 (Agência)

Dessa forma, temos :

S_{11} = Soma, das somas dos quadrados dos desvios em relação à média; para o índice de endividamento, isto é,

$$S_{11} = S_{11}^1 + S_{11}^2$$

Aonde :

$$S_{11}^1 = \sum_{e=1}^{10} (X_{11e} - \overline{X}_{11})^2$$

$$S_{11}^2 = \sum_{e=1}^{10} (X_{12e} - \overline{X}_{12})^2$$

S_{22} = Soma, das somas dos quadrados dos desvios em relação à média; em cada uma das amostras das populações, para o índice de rentabilidade, isto é,

$$S_{22} = S_{22}^1 + S_{22}^2$$

Aonde :

$$S_{22}^1 = \sum_{e=1}^{10} (X_{21e} - \overline{X}_{21})^2$$

$$S_{22}^2 = \sum_{e=1}^{10} (X_{22e} - \overline{X}_{22})^2$$

S_{12} = Soma, das somas dos produtos dos desvios em relação às médias; em cada uma das amostras das populações, para os índices de inadimplência e rentabilidade, isto é,

$$S_{12} = S_{12}^{1} + S_{12}^{2}$$

Aonde :

$$S_{12}^{1} = \sum_{e=1}^{10} (X_{11e} - \overline{X}_{11})(X_{21e} - \overline{X}_{21})$$

$$S_{12}^{1} = \sum_{e=1}^{10} (X_{12e} - \overline{X}_{12})(X_{22e} - \overline{X}_{22})$$

D_1 = Diferença entre as médias do índice de inadimplência nas duas amostras das populações, isto é,

$$D_1 = (\overline{X}_{11} - \overline{X}_{21})$$

D_2 = Diferença entre as médias do índice de rentabilidade nas duas amostras das populações, isto é,

$$D_2 = (\overline{X}_{12} - \overline{X}_{22})$$

Os cálculos inerentes à obtenção dos coeficientes *a* e *b*, de X_1 e X_2, respectivamente, estão nas Tabelas A.3 a A.6, evidentemente a partir dos dados constantes das Tabelas A.1 e A.2.

Tabela A.3 – Dados para cálculos das médias e desvios
(Agências com Baixa Inadimplência)

Agências com Baixa Inadimplência	Inadimplência			Rentabilidade			
	A	B	C	D	E	F	G
	CL/OC (X_1)	($X_{11e} - \bar{X}_{11}$)	($X_{11e} - \bar{X}_{11}$)^2	RF/AT (X_2)	($X_{12e} - \bar{X}_{12}$)	($X_{12e} - \bar{X}_{12}$)^2	(B.E)
E_{11}	1,34	0,350	0,1225	0,24	0,029	0,0008	0,0102
E_{12}	1,21	0,220	0,0484	0,20	-0,011	0,0001	-0,0024
E_{13}	1,48	0,490	0,2401	0,36	0,149	0,0222	0,0730
E_{14}	0,81	-0,180	0,0324	0,15	-0,061	0,0037	0,0110
E_{15}	1,15	0,160	0,0256	0,21	-0,001	0,0000	-0,0002
E_{16}	0,56	-0,330	0,1089	0,20	-0,011	0,0001	0,0036
E_{17}	0,73	-0,260	0,0676	0,17	-0,041	0,0017	0,0107
E_{18}	0,69	-0,300	0,0900	0,29	0,079	0,0062	-0,0237
E_{19}	0,53	0,540	0,2916	0,17	-0,041	0,0017	-0,0221
E_{110}	0,30	-0,690	0,4761	0,12	-0,091	0,0083	0,0628
$\sum$	9,90	0,000	1,422	2,11	0,000	0,0448	0,1229
Médias	$\bar{X}_1 = \frac{\sum X_1}{n}$			$\bar{X}_2 = \frac{\sum X_2}{N}$			0
	$\bar{X}_1 = \frac{9,90}{10} = 0,99$			$\bar{X}_2 = \frac{2,11}{10} = 0,211$			

IBCI – Institutional Business Consultoria Internacional
Rua da Alfândega, 25, sala 805/806, Centro, Rio de Janeiro, RJ.
Telefax: (21)2263-7017 / 2233-8552 / 2263-0563
e-mail: ibci@ibci.com.br
Home Page: www.ibci.com.br

Tabela A.4 – Dados para cálculos das médias e desvios
(Agências com Elevada Inadimplência)

Agências com Elevada Inadimplência	Inadimplência			Rentabilidade			
	A	B	C	D	E	F	G
	CL/OC (X_1)	($X_{21e} - \bar{X}_{21}$)	($X_{21e} - \bar{X}_{21}$)^2	RF/AT (X_2)	($X_{22e} - \bar{X}_{22}$)	($X_{22e} - \bar{X}_{22}$)^2	(B.E)
E_{21}	7,45	3,143	9,8784	-0,14	-0,130	0,0169	-0,4086
E_{22}	3,21	-1,097	1,2034	-0,02	-0,010	0,0001	0,0109
E_{23}	4,27	-0,037	0,0014	0,06	0,070	0,0049	-0,0026
E_{24}	1,85	2,457	6,0368	-0,08	-0,070	0,0049	-0,1720
E_{25}	1,45	-2,857	8,1624	0,11	0,120	0,0144	-0,3428
E_{26}	9,25	4,943	24,4332	-0,62	-0,610	0,3721	-3,0152
E_{27}	2,76	-1,547	2,3932	0,25	0,260	0,0676	-0,4022
E_{28}	3,54	-0,767	0,5883	0,01	0,020	0,0004	-0,0153
E_{29}	4,88	0,573	0,3283	0,25	0,260	0,0676	0,1490
E_{210}	4,41	0,103	0,0106	0,08	0,090	0,0081	0,0093
Σ	**43,07**	**0,000**	**53,0360**	**-0,10**	**0,000**	**0,5570**	**-4,1895**
Médias	$\bar{X}_1 = \frac{\Sigma X_1}{N}$ $\bar{X}_1 = \frac{43,07}{10} = 4,307$			$\bar{X}_2 = \frac{\Sigma X_2}{N}$ $\bar{X}_2 = \frac{-0,10}{10} = -0,01$			0

Tabela A.5 - Médias e Diferenças entre Médias

	Baixa Inadimplência	Elevada Inadimplência	Diferença
Média dos Índices de Inadimplência (X_1)	0,990	4,307	-3,317 (D_1)
Média dos Índices de Rentabilidade (X_2)	0,211	-0,010	0,221 (D_2)

Tabela A.6 - Dados para Matriz de Covariância

	Baixa Inadimplência	Elevada Inadimplência	Diferença
Soma dos Quadrados (desvios) da Inadimplência = $\Sigma (X_{1e} - X_1)^2$	1,422	53,0360	54,4582 (S_{11})
Soma dos Quadrados (desvios) da	0,0488	0,5570	0,6018 (S_{22})

IBCI – Institutional Business Consultoria Internacional
Rua da Alfândega, 25, sala 805/806, Centro, Rio de Janeiro, RJ.
Telefax: (21)2263-7017 / 2233-8552 / 2263-0563
e-mail: ibci@ibci.com.br
Home Page: www.ibci.com.br

Rentabilidade = $\Sigma (X_{2e} - X_2)^2$			
Soma dos Produtos (desvios) entre Inadimplência e Rentabilidade = $\Sigma (X_{1e} - X_1)(X_{2e} - X_2)$	*0,1229*	*-4,1895*	*-4,0596 (S_{12})*

Dada a função :

$$Z = aX_1 + bX_2$$

Obtém-se o seguinte sistema :

$$\begin{cases} S_{11}a + S_{12}b = D_1 \\ S_{12}a + S_{22}b = D_2 \end{cases}$$

$$\begin{cases} 54,4582\,a + (-4,0596)\,b = -3,317 \\ -4,0596\,a + 0,601\,b = 0,221 \end{cases}$$

Resolvendo esse sistema de equações, encontramos os coeficientes de X_1 e X_2, isto é, os valores de a e b, respectivamente :

$$a = -0,06745$$
$$b = -0,08779$$

o que nos dá :

$$Z = -0,06745\,X_1 - 0,08779\,X_2$$

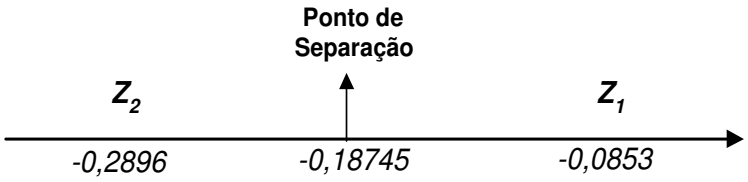
4 – Interpretação e Uso da Função Discriminante

Uma vez conhecidos os coeficientes de X_1 e X_2, podemos calcular os valores médios para cada Z em cada uma das amostras das populações π_1 e π_2. Assim temos :

$$Z_1 = -0,06745\,(0,99) - 0,08779\,(0,211) = -0,0853$$

$$Z_2 = -0,06745\,(4,307) - 0,08779\,(-0,01) = -0,2896$$

Dessa forma Z_1 representa o valor da função linear Z para a média da amostra das agências com baixa inadimplência, enquanto que Z_2 representa o valor da função linear Z para média das agências com elevada inadimplência. Gráficamente tem-se :



Vale acrescentar que quando usamos a função discriminante de Fisher, que assume duas populações de mesma matriz de covariância, o ponto de separação entre as duas populações é o ponto médio entre os valores que representam as funções para as médias das amostras das duas populações.

A aplicação da função discriminante $Z = -0,06745\ X_1 - 0,08779\ X_2$ para uma agência em particular, a qual desconhecemos suas condições de inadimplência, levará a um resultado que deverá ser comparado com os valores das funções que representam as médias das amostras das duas populações. Se o valor encontrado for maior que *–0,18745*, a agência será classificada como de baixa inadimplência; se for menor, sofrerá classificação inversa.

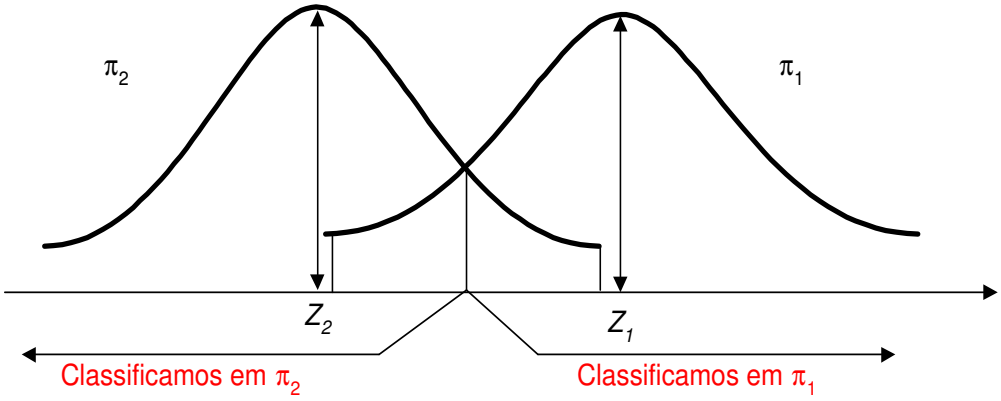
Na Tabela A.7 apresentamos os valores Z para as 20 agências que estamos utilizando na montagem de nosso exemplo.

Tabela A7 – Valores de Z para as Vinte Agências Analisadas

Baixa Inadimplência		**Alta Inadimplência**	
Agências E_{ie}	**Z_1**	**Agências E_{ie}**	**Z_2**
E_{11}	-0,111453	E_{21}	-0,490212
E_{12}	-0,099173	E_{22}	-0,214759
E_{13}	-0,131430	E_{23}	-0,293279
E_{14}	-0,067803	E_{24}	-0,117760
E_{15}	-0,096003	E_{25}	-0,107460
E_{16}	-0,062075	E_{26}	-0,569483
E_{17}	-0,064163	E_{27}	-0,208110
E_{18}	-0,072000	E_{28}	-0,239651
E_{19}	-0,116123	E_{29}	-0,351104
E_{110}	-0,030770	E_{210}	-0,304478

Observando as dez primeiras agências, da Tabela A.7, E_{11} *a* E_{110}, que constituem a amostra de agências com baixa inadimplência, nota-se que todas elas apresentaram um *Z* maior que *–0,18745*, o que corresponde a uma classificação corrente de 100% desse grupo. Quanto as agências pertencentes à amostra de população de alta inadimplência, nota-se que, em dez, oito apresentaram *Z* menor que o ponto de separação; apenas duas, E_{24} e E_{25}, têm *Z* acima de *–0,18745*. Para o grupo de elevada inadimplência, o erro de classificação seria de 20%; no geral, nas 20 agências o erro seria de 10%.

Para duas populações (π_1 e π_2) normais, com a mesma matriz de covariância, temos a seguinte representação gráfica :



Note-se a existência de uma área de superposição na qual temos :

α = probabilidade de classificar em π_1 um elemento pertencente a π_2

β = probabilidade de classificar em π_2 um elemento pertencente a π_1

Classificar uma agência de baixa inadimplência (de π_1) como de elevada inadimplência (de π_2) pode trazer consequências. Se estivermos em um ciclo de expansão do crédito de varejo, aonde a extensão da rede de agências de baixa inadimplência e do número de clientes seja fator determinante para uma alocação de recursos de empréstimo eficiente (rentabilidade), a intensidade do efeito do erro de classificação pode ser significativa; o *custo de oportunidade* de abrir novas agências pode inviabilizar uma política de crédito expansionista. Por outro lado, classificar uma agência de elevada inadimplência (de π_2) como de baixa inadimplência (de π_1) pode, diante do mesmo cenário expansionista, implicar no crescimento dos créditos

de qualidade duvidosa e de outros custos adicionais como cobrança, recuperação de crédito, etc.

Quanto a ponto de separação, a média entre Z_1 e Z_2 pode não ser a melhor forma de minimizar o risco de erro de classificação, uma vez que depende das probabilidades *a priori* e dos custos decorrentes do erro de classificação. Se assumirmos que o custo de classificar uma agência de alta inadimplência como de baixa inadimplência é o mesmo de classificarmos uma agência de baixa inadimplência como de alta inadimplência, assim teremos iguais probabilidades a priori (0,5), o que torna a regra de classificação ótima.

5 – Separação e Classificação de Observações/Populações

A separação e classificação de observações/populações não é tarefa simples; ainda mais em um projeto aonde deveremos lidar com um universo amostral significativo. Dada a *Lei dos Grandes Números*, ao lidarmos com 1.027 Municípios e seus dados nos aproximaremos em diversos casos de dados distribuídos possivelmente de forma similar a uma curva normal (curva de Gauss). Embora isto nem sempre ocorra, pois há assimetrias sistemáticas como as geradas pelo sistema de distribuição de renda e de PIB , é de bom alvitre utilizar a aproximação da distribuição normal na gestão das funções básicas.

Um bom procedimento de classificação deve resultar em poucas desclassificações. Em outras palavras, as chances, ou probabilidades, de desclassificação devem ser reduzidas uma vez que o custo da desclassificação em um universo amostral maior podem comprometer a análise. Em um sistema simples de duas classes, como o nosso exemplo, a inserção de um dado evento/observação numa determinada classe da população/amostra em detrimento de outra, pode ter probabilidades diferenciadas em função das diferenças de tamanho da amostra. Assim, a classificação tida como *ótima* deve levar em consideração as probabilidades *a priori* de ocorrência dos eventos/observações.

Outro aspecto a ser considerado na classificação é o custo. Um procedimento classificatório *ótimo* deve, quando possível, considerar o custo da desclassificação.

6 – Teste de Significância da Discriminação

Para conhecermos se a discriminação é boa ou não, Fisher sugere que se faça uma Análise da Variância - ANOVA (***AN****alysis* ***O****f* ***VA****riance*). Este teste é citado por inúmeros autores como *"teste F"* ou *"Estatística F"* em virtude de fazer uso da Distribuição *F* de Snedecor para verificar a significância ou não do poder discriminador das variáveis $X_1, X_2, \ldots, X_n$ consideradas.

Na Análise Discriminante sabemos que as populações são diferentes e o que queremos é construir uma função dicotômica – para o caso de duas populações – que discrimine se um dado elemento pertence a uma ou a outra população.

Na Análise da Variância - ANOVA não se sabe *a priori* se as populações são diferentes, mas queremos testar se o são. Para tal, como na Análise Discriminante, devemos extrair uma amostra de cada uma das populações/grupos e buscar analisar as variações entre grupos e intragrupos. A Variância total é explicada pelas variações dentro dos grupos/populações e entre os grupos/populações.

A Variância (S^2)de uma amostra é definida por :

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$$

Onde X_i é a iésima observação

$\overline{X}$ é a média das observações, ou seja;

$$\overline{X} = \frac{\sum X_i}{n}$$

e n o número de observações/eventos.

Estatísticamente, a Variância S^2 indica a dispersão dos dados X_i em relação a média. Cada desvio da média das observações é obtido através do desmembramento da soma dos quadrados e dividido por *n-1* o que torna a Variância semelhante a uma média dos quadrados das distâncias entre os dados observados. A Variância é particularmente importante quando a distribuição de frequência dos dados aproxima-se da curva normal (Curva de Gauss) pois a Variância e a Média a especificam única e completamente.

Na maioria dos casos, a previsão de dados assume a *"normalidade"*. Isto porque a distribuição amostral dos estimadores pode ser aproximada a curva normal onde *n* possua amplitude suficiente, o que na maior parte ocorre quando *n* é igual a 30. O *Teorema do Limite Central* da estatística permite esta aproximação e torna possível o uso da curva normal na avaliação da dispersão dos dados da amostra em torno do parâmetro central (média). Assim ao calcularmos sua média e variância, a extensão de possíveis erros pode ser avaliada; o que introduz um intervalo de confiança de 30 observações para a variância.

Assim, no caso do nosso exemplo constante das Tabelas A.1 e A.2, podem-se testar se um grupo de agências com baixa inadimplência e outro com alta inadimplência são diferentes quanto a rentabilidade.

A Análise da Variância - ANOVA não fornece uma função que permita classificar elementos em uma ou outra população, que, como já vimos, é o objetivo da Análise Discriminante. Contudo, seu conteúdo informacional está vinculado às diferenças significativas em torno das médias. Assim, a utilização do *Teste de Significância F* pode dizer se de fato uma variável discrimina bem entre dois grupos.

A *Estatística F* é uma razão, uma proporção da variância entre grupos em relação a uma taxa média ponderada de variância intergrupal. Caso a relação entre

estas razões seja pequena, então a razão entre os dois é significativa. Desta forma, existe pelo menos uma diferença notável entre as médias dos grupos.

No caso de estatísticas multi-grupais e multi-funções, que é o que ocorre aonde introduziremos três dimensões, dadas pelas funções FUS – DIPE (**DI**scriminante de **PE**rfil), FUS – DIAT (**DI**scriminante de **AT**ratividade) e FUS – DICO (**DI**scriminante de **CO**mplexidade), e estaremos trabalhando com cerca de159 variáveis para 1.027 Municípios – deveremos nos ater a uma Análise Multivariada da Variância – MANOVA (***M**ultivariate **AN**alysis **O**f **VA**riance*) aonde as interações em torno das médias ocorrem de forma multivariada, e não linear como em ANOVA.

Neste caso, teremos uma matriz de Variâncias totais e de Covariancias. Estas matrizes serão comparadas por meio de dois Testes *F* multivariados. Assim, poderemos definir se há diferenças significativas entre os grupos, em relação a todas as variáveis das funções FUS que trabalharemos.

Para definir o procedimento de corte entre variáveis dependentes e variáveis independentes nas três dimensões selecionadas para este caso específico - Perfil, Atratividade e Complexidade – os analistas vão entender que em cada dimensão há indicadores claros que definem riqueza (força econômica), enquanto outros dirigem-se à sinalização da pobreza (ou de carências municipais).

Desta forma, o corte ocorrerá em dois niveis, bem claros e discriminatórios. O que um Banco deseja é identificar aonde possui reais formas de gerar resultados maximizantes para os acionistas, por Município neste caso.

7 – A Variância como Medida de Risco

Como vimos nos blocos anteriores, os métodos estatísticos de previsão podem utilizar as propriedades estatísticas dos dados observados/populações para construir intervalos de confiança e testar diferentes hipóteses acerca dos dados de uma previsão. Este processo envolve o *Teorema do Limite Central*, o qual permite que a distribuição dos valores previstos sofra uma aproximação à curva normal. A Média da distribuição normal é valor mais esperado, e a Variância é a medida de dispersão de todos os valores em torno da Média. A Variância é uma medida estatística extremamente útil porque sumariza as incertezas e erros na estimativa dos parâmetros de um modelo. Além do que, com estas duas medidas – a Média e a Variância – o intervalo de todos os valores futuros esperados e a sua probabilidade de ocorrência podem ser previstos.

Além do uso destas medidas na construção de um intervalo de confiança para uma previsão, a Variância de uma previsão é uma medida de risco e pode ser utilizada como tal. Funciona como um indicativo do grau de incerteza associado na previsão de uma variável. Ao analisarmos um plano de ação, em qualquer área da atividade empresarial, a Análise da Variancia – ANOVA pode ser utilizada como balizadora do risco envolvido e na preparação de medidas contingenciais para cenários otimistas e/ou pessimistas.

IBCI – Institutional Business Consultoria Internacional
Rua da Alfândega, 25, sala 805/806, Centro, Rio de Janeiro, RJ.
Telefax: (21)2263-7017 / 2233-8552 / 2263-0563
e-mail: ibci@ibci.com.br
Home Page: www.ibci.com.br

Um trabalho volumoso em termos de cálculos estatísticos, e de conteúdo informacional bastante extenso na Análise de Carteiras de Investimento, foi desenvolvido por Markowitz [9], em 1952, envolvendo a Covariância como base para minimização do risco da administração (compra e venda) de papéis e na otimização do retorno da carteira para um dado nível de risco (vide Referências Bibliográficas). De maneira similar, os modelos de previsão podem ser construídos de forma a minimizar os erros de previsão quando um volume significativo de observações ou áreas da administração está envolvido.

Em outras palavras, o risco total de imprecisão nas previsões de um determinado conjunto de fatores multivariados, pode ser examinado a partir da Análise Multivariada da Variância – MANOVA e da Análise Multivariada da Covariância – MANCOVA. Estas técnicas podem reduzir substancialmente os efeitos de super/sub estimação de um conjunto de dados/observações agrupando as previsões de uma forma que a sua Variância e Covariância sejam as menores possíveis.

8 – Recursos Computacionais em Utilização

Definida a amostra, optamos pela utilização do *STATISTICA Release 5* (1997) para ambiente operacional Microsoft Windows 3.11/95/NT da StatSoft Inc. de Tulsa – EUA. Desenvolvido em compiladores Microsoft C/C++ este pacote apresentou a melhor relação custo-benefício relativamente a outros pacotes de software específicos como o *Statistical Analysis System - SAS* e do *Statistical Package for Social Sciences – SPSS* ambos do North Carolina Institute – EUA.

Com ampla utilização internacional em plataformas do tipo IBM-PC e compatíveis, o *STATISTICA*, é conhecido pela sua ampla capacidade de armazenar dados, versatilidade em dialogar com os principais pacotes de software de planilha eletrônica (do tipo Microsoft Excel) e de banco de dados (do tipo Microsoft Access) do mercado, além de enorme capacidade no processamento estatístico interativo de um grande número (tende a infinito) de dados.

Assim, o *STATISTICA* analisará cada um dos dados inseridos no seu módulo de administração de dados (*Data Management*) e verificará qual deles contribui mais ou menos, para a discriminação entre os grupos determinados. A variável de maior relevância será então incluída no modelo, e o sistema procederá à etapa seguinte, da análise interativa dos dados.

Neste procedimento de inclusão das variáveis de alto poder explicativo e exclusão das de baixo poder explicativo, serão mantidas as variáveis mais relevantes. Óbviamente, estas serão as que mais discriminam entre os grupos.

Municipios, praças e logradouros de baixo "potencial", vistos sob as dimensões de "Atratividade" e "Complexidade", indicarão características menos interessantes às atividades de *Banking*; e vice - versa. Naturalmente, o que sería desejável é dispor do maior número de Municipios com alta Atratividade, baixa Complexidade e Perfil bem definido.

9 – Referências Bibliográficas

1. Altman, E.I., *Corporate Financial Distress – A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy*, New York : John Wiley & Sons, 1993.

2. Anderson, T.W., *An Introduction to Multivariate Statistical Methods* (Second Edition), New York : John Wiley & Sons, 1984.

3. Fisher, R.A., "The Statistical Utilization of Multiple Measurements", *Annals of Eugenics*, Vol.8 (1938).

4. Haley, C.W. & Schall, L.D., *The Theory of Financial Decisions* (Second Edition), New York : McGraw-Hill Book Company, 1979.

5. Hand, D.J., *Discrimination and Classification*, New York : John Wiley & Sons, 1981.

6. Johnson, R.A. & Wichern, D.W., *Applied Multivariate Statistical Analysis* (Third Edition) , New Jersey : Prentice Hall, 1992.

7. Kasznar, I.K., *Falências e Concordatas de Empresas – Modelos Teóricos e Estudos Empíricos (1978 – 1982/87)* – Dissertação submetida à Congregação da Escola de Pós-Graduação em Economia (EPGE/FGV) para Obtenção do Grau de Mestre em Economia – Novembro de 1987.

8. Kendall, M.G., *Multivariate Analysis*, New York: Hafner Press, 1975.

9. Markowitz, H., Portfolio Selection, *Journal of Finance*, Vol.7, pp 77-91 (1952).

10. Sharpe, W.F., *Investments*, New Jersey : Prentice Hall, 1982.

11. Wonnacott T.H. & Wonnacott R.J., *Introdutory Statistics for Business and Economics* (Second Edition), New York : John Wiley & Sons, 1979.