UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DE RIBEIRÃO PRETO
DEPARTAMENTO DE ADMINISTRAÇÃO

RAD1509 – Estatística Aplicada à Administração II
Prof. Dr. Evandro Marcos Saidel Ribeiro
RESUMO

1 Matriz de correlação

A matriz de correlação é escrita na forma:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \dots & 1 \end{bmatrix}. \qquad (1)$$

O cálculo de cada elemento desta matriz pode ser obtido da matriz $\mathbf{S}$ por:

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}} . \qquad (2)$$

Sendo i e j duas variáveis, x_1 e x_2. Por exemplo, para, x_1 e x_2 dados por

$$\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 9 & 8 \\ 7 & 5 \\ 3 & 4 \\ 4 & 2 \end{bmatrix}, \qquad (3)$$

A correlação será r_{12} = 0,908025.

A correlação pode ser obtida na HP12C pelo seguinte procedimento:

- Digite os dados: x_2 [enter] x_1 Σ+ (4)
- Obtenha o valor de r: [g] 2 [xy] (5)

2 Matriz de covariância

O cálculo de cada elemento da matriz de covariância pode ser feito a partir da matriz de correlação. Assim, a partir da correlação r_{12}, entre duas variáveis x_1 e x_2, a covariância é dada por:

$$S_{ij} = r_{ij} s_i s_j \qquad (6)$$

Na expressão (11) s_i e s_j são os desvios-padrão das variáveis i e j, respectivamente.

3 Componentes Principais e Análise Fatorial

Na Análise Fatorial (AF) cada variável é associada a Fatores:

$$x_i = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{in}F_n \qquad (7)$$

sendo a_{ij} cargas fatoriais.

As cargas fatoriais são dadas numa "Tabela de Cargas Fatoriais". Por exemplo: Considerando 4 variáveis e dois fatores:

Tabela 1. Cargas Fatoriais

Variável	Fator 1	Fator 2
X1	0,15	0,70
X2	0,02	0,84
X3	-0,85	0,01
X4	0,90	0,05

A soma das cargas fatoriais ao quadrado para uma determinada variável (soma em todos os NF fatores) é igual a comunalidade da variável:

$$h_i = \sum_{j=1}^{NF} a_{ij}^2 \qquad (8)$$

Se a soma for feita para um determinado índice relacionado a um fator (soma para todas as k variáveis) então obtemos o autovalor correspondente ao fator em questão:

$$\lambda_j = \sum_{i=1}^{k} a_{ij}^2 \qquad (9)$$

Para o exemplo na Tabela 1 temos os resultados de autovalores e comunalidades apresentados na Tabela 2.

Tabela 2. Cargas Fatoriais, Comunalidades e Autovalores

Variável	Fator 1	Fator 2	Comunalidade (hi)
X1	0,15	0,70	0,51
X2	0,02	0,84	0,71
X3	-0,85	0,01	0,72
X4	0,90	0,05	0,81
Autovalor (λi)	1,5554	1,1982	
% Variância	38,89	29,96	

Autovalores:

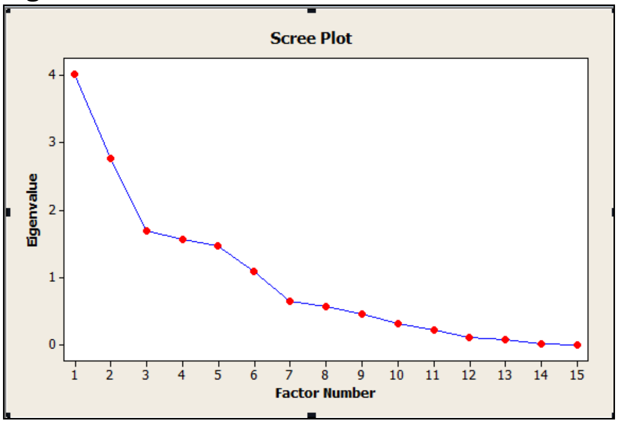
O percentual da variação explicada: Na análise de componentes principais a variação explicada por cada componente é dada pelo autovalor dividido pela soma dos autovalores:

$$CP_i = \frac{\lambda_i}{\sum \lambda} \times 100 \qquad (10)$$

Considerando a matriz de correlação, a soma dos autovalores é igual ao número de variáveis.

Um gráfico de autovalores é o Scree Plot. Para fazr este gráfico deve-se ordenar os autovalores do maior para o menor. No eixo-x são consideradas as componentes (ou índices dos fatores), no eixo-y são considerados os autovalores. Exemplo na Figura 1 abaixo.

Figura 1. Scree Plot com 15 fatores.



Um dos objetivos principais da AF: cada Fator pode ser escrito em termos das variáveis sendo possível assim obter uma função para o Escore Fatorial *j*:

$$F_j = b_{j1}x_1 + b_{j2}x_2 + ... + b_{jk}x_k \qquad (11)$$

O número de fatores a serem considerados pode ser obtido pelo

Critério Kaiser: autovalores > 1 (12)

Os coeficientes para o Escore Fatorial são dados na Tabela de Coeficientes de Escore Fatorial. Por exemplo: Considerando 4 variáveis e dois fatores, um exemplo pode ser visto na Tabela 3.

Tabela 3. Coeficientes dos Escores Fatoriais

Variável	Fator 1	Fator 2
X1	0,35	0,80
X2	0,02	0,95
X3	-0,94	0,10
X4	0,99	0,10

A partir da Tabela 3 e dos valores de cada variável são obtidos os Escores Fatoriais de acordo com a equação (11) para cada caso em análise.

Por Exemplo:

Considerando um caso no qual

(x1; x2; x3; x4) = (1,5; -1,0; 1,0; 2)

O Escore Fatorial 1, dado pela expressão (11) é calculado com os valores de bj da Tabela 3:

$$F_1 = 0{,}35 \times 1{,}5 + 0{,}02 \times (-1{,}0)$$
$$- 0{,}94 \times 1{,}0 + 0{,}99 \times 2{,}0 = 1{,}545$$

4 Análise de Agrupamentos: Hierárquico

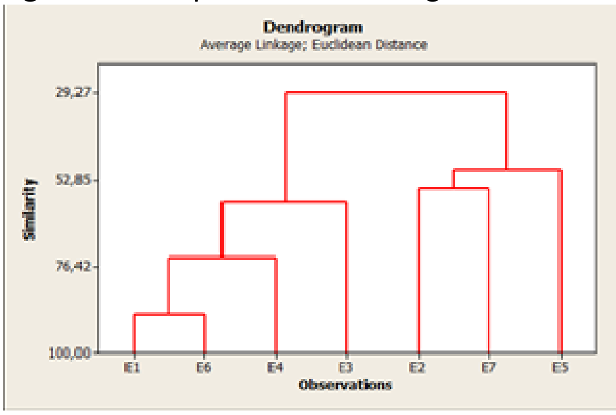
Objetivo dessa técnica é agrupar objetos (itens) semelhantes segundo suas características (variáveis)

Procedimento:

- Padronizar as variáveis: $z_i = (x_i - \bar{x}_i)/s_i$
- Obter uma matriz de distâncias
- Aplicar um método de aglomeração, por exemplo os métodos hierárquicos
  - Ligação Simples (vizinho mais próximo), ou ainda Single Linkage
  - Ligação Completa (vizinho mais distante), ou ainda Complete Linkage
  - Ligação Média (distância média), ou ainda Average linkage
- Considerar a menor distância para formar um grupo.
- Recalcular a matriz de distâncias
- No final, esboçar um Dendrograma

A Figura 2 apresenta um exemplo de Dendrograma resultante da Análise de Agrupamento de sete empresasmvisto em aula (ver slides).

Figura 2. Exemplo de um dendrograma.



5 Análise Discriminante

Utilizada para determinar quais variáveis discriminam entre dois ou mais grupos já conhecidos.

A variável dependente é categórica e indica os grupos já conhecidos. As variáveis independentes podem ser métricas ou categóricas, estas deverão ter poder de discriminar os indivíduos entre dois ou mais grupos.

O primeiro passo é compor para cada grupo (g_i), as funções discriminantes (d_i):

Para g_1: $d_1 = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1n}x_n$

Para g_2: $d_2 = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \ldots + \beta_{2n}x_n$

$$\vdots$$

Para g_k: $d_k = \beta_{k0} + \beta_{k1}x_1 + \beta_{k2}x_2 + \ldots + \beta_{kn}x_n$

Para classificar um indivíduo num determinado grupo devemos substituir as variáveis x_1, x_2, ..., x_n em todas as funções discriminantes d_i. A função que obtiver o maior valor indica a classificação do indivíduo para o grupo correspondente.

Os coeficientes β's são fornecidos numa tabela de coeficientes para Funções Discriminantes. Por exemplo, a Tabela 4 apresenta os coeficientes para as funções discriminantes, considerando 5 variáveis explicativas e uma variável dependente com 3 categorias.

Tabela 4. Coeficientes para Funções Discriminantes para três grupos.

↓	C7	C8	C9
	disc1	disc2	disc3
1	-2,6888E+01	-4,3228E+01	-3,5491E+01
2	3,1581E-06	1,2579E-06	9,7455E-07
3	4,0853E-01	6,5124E-01	6,8182E-01
4	2,3688E+01	4,3778E+01	2,2095E+01
5	3,3481E-01	4,6246E-01	5,8346E-01
6			

O percentual de acerto pode ser obtido a partir de uma tabela de classificação. Nesta Tabela de Classificação (exemplificada na Figura 3) pode-se observar as categorias verdadeiras (coluna: True Group), bem como, as Classificações feitas de acordo com as funções discriminantes (Put into Group). Com isto podemos verificar a proporção de acertos (0,952 = 20/21 no exemplo abaixo).

Figura 3. Tabela de Classificação

```
Summary of classification

                     True Group
Put into Group      1       2       3
1                   6       0       0
2                   0      10       0
3                   0       1       4
Total N             6      11       4
N correct           6      10       4
Proportion      1,000   0,909   1,000

N = 21             N Correct = 20
```

6 Regressão Logística

A variável dependente é categórica. Na Regressão Logística Binária são consideradas duas categorias: 0 e 1. O modelo desenvolvido fornece a probabilidade de um caso (empresa, indivíduo, ...) ser classificado na categoria 1.

$$P(Y=1) = \frac{1}{1+\exp(-g(x))}, \qquad (13)$$

onde, a função g(x) é obtida na forma:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n . \qquad (14)$$

Os valores de β's são dados numa Tabela de coeficientes para regressão logística.

Por exemplo: A tabela 5 fornece os valores de betas para um modelo de regressão logística com 3 variáveis: R, ND e VE, para previsão de inadimplência.

Tabela 5. Coeficientes betas para função g(x)

Variáveis	**Beta**
Constante	1,5
R	-1,9
ND	0,9
VE	2,8

Considere um caso no qual

R = 3, ND = 2 e VE = 0.

A partir da Tabela 5 e dos valores de cada variável calcula-se o valor de g(x) dado pela equação (14)

$$\mathrm{g(x)} = 1{,}5 - 1{,}9 \times 3 + 0{,}9 \times 2 + 2{,}8 \times 0 = \text{-}2{,}4$$

Este valor é substituído na expressão (13):

$$P(Y = 1) = \frac{1}{1 + \exp(2{,}4)} = 0{,}083$$

O valor obtido é a probabilidade de que o caso seja classificado como 1, ou seja, neste caso a probabilidade de ser inadimplente.

Para obter a previsão de classificação o resultado deve ser arredondado para 0 ou 1.

Neste caso: Prev = 0.

O percentual de acerto também pode ser obtido a partir de uma tabela de classificação, de forma análoga à análise discriminante na seção 5, anterior.