CE-704 ANÁLISE MULTIVARIADA APLICADA À PESQUISA

NOTAS DE AULA

Estas **notas de aula** seguem de muito perto os livros referenciados na **BIBLIOGRAFIA** e que na verdade correspondem aos livros textos deste Curso. Sugere-se a sua aquisição. A única finalidade destas notas é facilitar o trabalho do aluno em sala de aula, de modo que não há necessidade de anotar todo o conteúdo apresentado pelo professor. A leitura, consulta e resolução de exercícios do livro é dever do aluno.

Prof. Anselmo Chaves Neto

BIBLIOGRAFIA

- Johnson, R. A. & Wichern, D.W. – Applied Multivariate Statistical Analysis; 4ed. Prentice Hall Inc., Englewood NJ (1998).
- Mardia, K. V. Kent, J. T. & Bibby, J.M. – Multivariate Analysis; Academic Press, New York (1978).
- Morrison, D.F. – Multivariate Statistical Methods - McGraw Hill, N.Y., 1971.
- Hair, Joseph F. Jr. et alii – Multivariate Data Analysis, 5ed., Prentice Hall Inc. Upper Saddle River, N.J. (1998).
- Hair, Joseph F. Jr. et alii – Análise de Dados Multivariados, Prentice Hall Inc., Bookman Edt./Artmédia Edt., Porto Alegre, 2005.

ÍNDICE

ANÁLISE MULTIVARIADA

1. INTRODUÇÃO

1.1 - Conceitos Básicos

ANÁLISE MULTIVARIADA: é um conjunto de técnicas estatísticas que tratam dos dados correspondentes às medidas de muitas variáveis simultaneamente. Basicamente, a Análise Multivariada consiste no estudo estatístico dos problemas relacionados com:

- Inferências sobre médias multivariadas;
- Análise da estrutura de covariância de uma matriz de dados;
- Técnicas de reconhecimento de padrão, classificação e agrupamento.

No estudo de $p \geq 1$ variáveis, geralmente, toma-se n observações de cada variável para obter informações sobre parâmetros, relacionamentos entre variáveis, comparações, etc. Assim, as medidas registradas são x_{ij} com i = 1,2, ... ,n (observações) e j = 1,2, ... ,p (variáveis) que podem ser agrupadas na matriz de dados ${}_nX_p$, com n linhas e p colunas

$$ {}_nX_p = \begin{bmatrix} x_{11}x_{12.}.....x_{1p} \\ x_{21}x_{22}.....x_{2p} \\ ..... \\ x_{n1}x_{n2}.....x_{np} \end{bmatrix} $$

A matriz de dados ${}_nX_p$ contém n observações do vetor aleatório p-dimensional $\underline{X}' = [X_1, X_2, ... , X_P]$.

EXEMPLO 1:
Uma amostra aleatória composta por quatro (4) notas de vendas de livros de uma livraria foi obtida a fim de investigar-se a natureza dos livros vendidos. Cada nota fiscal especifica, entre outras coisas, o número de livros vendidos e o valor de cada venda. Seja a 1ª variável o **total vendido** em reais e a 2ª variável o **número de livros vendidos**. Assim, seja o vetor aleatório $\underline{X}' = [X_1\ X_2]$ cujas componentes são as v.a's: X_1 (valor da venda) e X_2 (número de livros).

A matriz de dados é ${}_nX_p = \begin{bmatrix} 42 & 4 \\ 80 & 5 \\ 48 & 4 \\ 36 & 3 \end{bmatrix}$

1.2 - Estatísticas Descritivas

Muito da informação contida na matriz de dados pode ser dada pelo cálculo de **números sumários** conhecidos como **estatísticas descritivas**.

Vetor médio amostral : $\underline{\bar{x}}' = [\ \bar{x}_1\ \ \bar{x}_2 \ldots\ldots\ \bar{x}_p\ ]$ com $\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$ j = 1,2, ... ,p.

Matriz de covariância amostral: $S = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{bmatrix}$ onde

$s_{jj} = s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_i)^2}{n-1}$ é a variância da v.a. X_j

$s_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n-1}$ j, k = 1,2, ... ,p é a covariância entre X_j e X_k.

Matriz de correlação amostral: $R = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{p1} & r_{p2} & \ldots & 1 \end{bmatrix}$ onde $r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$

EXERCÍCIOS

1) Considere os dados do exemplo 1. Então, calcule:
   a) O vetor médio amostral.
   Solução: Calcule a média de cada variável usando a calculadora ou, então, usando o STATGRAPHICS siga o caminho: DESCRIBE, NUMERIC DATA, MULTIPLE VARIABLES ANALYSIS; entre com as duas variáveis e OK; agora vá no botão amarelo de TABULAR OPTIONS e escolha SUMMARY STATISTICS e o resultado estará na tabela abaixo com várias estatísticas. O vetor médio será $\overline{X}' = [51{,}5\ \ 4]$.

	valor$	qLIVROS
Count	4	4
Average	51.5	4.0
Median	45.0	4.0
Variance	385.0	0.666667
Standard deviation	19.6214	0.816497
Standard error	9.81071	0.408248
Minimum	36.0	3.0
Maximum	80.0	5.0
Range	44.0	2.0
Lower quartile	39.0	3.5
Upper quartile	64.0	4.5
Coeff. of variation	38.0998%	20.4124%

b) a matriz de covariância amostral S.

Solução: Calcule a variância de cada variável, depois calcule a covariância entre elas usando as fórmulas e a calculadora, ou então, use os resultados da tabela anterior para montar a matriz. Lembre que a covariância $s_{12} = \rho_{12}s_1s_2$. A matriz de covariância é:

$$S = \begin{bmatrix} 385 & 14.6667 \\ 14.6667 & 0.66667 \end{bmatrix}$$

Veja que $s_1^2 = 385$ é a variância amostral e estima a verdadeira variância populacional σ_1^2; $s_{12} = 14{,}6667$ é a covariância amostral e $s_2^2 = 0{,}6667$ é a estimativa amostral de σ_2^2. Finalmente, a matriz S é a estatística que estima o verdadeiro parâmetro Σ (matriz de covariância populacional).

c) a matriz de correlação amostral R.

Solução: Calcule o coeficiente de correlação $\hat{\rho}_{12} = r_{12}$ entre as duas variáveis usando as fórmulas e a calculadora, ou então, pegue a matriz diretamente no STATGRAPHICS. A matriz de correlação é:

$$R = \hat{\rho} = \begin{bmatrix} 1 & 0.915475 \\ 0.915475 & 1 \end{bmatrix}$$

Finalmente, a matriz R é a estatística que estima o verdadeiro parâmetro ρ (matriz de correlação populacional).

2) Você sabia que a correlação entre as v.a's X e Y é igual a covariância entre as v.a's X e Y padronizadas? Prove este fato.

Prova:
Por definição a covariância entre duas v.a's é dada por:

$cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ e dividindo pelo produto dos desvios padrões $\sigma_X\,\sigma_Y$ tem-se o quociente:

$$\frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} = E\left[\frac{(X-\mu_X)}{\sigma_X}\frac{(Y-\mu_Y)}{\sigma_Y}\right]$$ que é a covariância entre duas v.a's padronizadas. Então o coeficiente de correlação $\rho = E\left[\frac{(X-\mu_X)}{\sigma_X}\frac{(Y-\mu_Y)}{\sigma_Y}\right]$ é a covariância entre duas v.a's padronizadas.

1.3 - Distância

Várias técnicas estatísticas são baseadas no conceito simples de distância. A distância Euclidiana entre os pontos P e O $\in R^p$, ou seja, do ponto $P(x_1, x_2, \ldots, x_p)$ até a

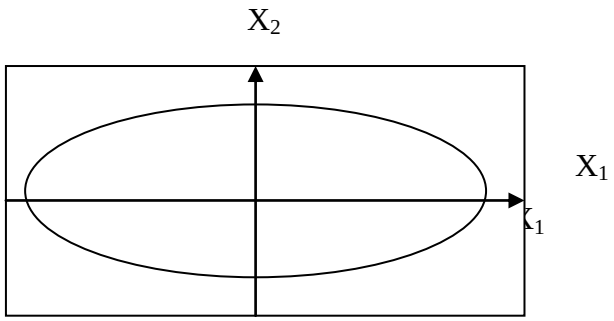
origem O(0, 0, ... , 0) é a distância na linha reta d(PO) dada de acordo com o Teorema de Pitágoras:

$$d(PO) = \sqrt{x_1^2 + x_2^2 + \ldots\ldots + x_p^2}$$

e a distância de $P(x_1, x_2, \ldots\ldots, x_p)$ ao ponto $Q(y_1, y_2, \ldots\ldots, y_p)$ é dada por:

$$d(PQ) = \sqrt{(x_1 - y_1)^2 + \ldots\ldots + (x_p - y_p)^2}$$

Contudo, a distância Euclidiana não é satisfatória em várias propostas estatísticas porque cada coordenada contribui igualmente para o cálculo da distância. Quando as coordenadas são medidas de v.a's de diferentes magnitudes (escalas), (p.ex. x_1 é da ordem de 1; 0,5; 2; 0,1; etc e x_2 é da ordem de 1000; 5652; 15314; etc.), variabilidades fortemente diferenciadas, é preferível ponderar as coordenadas de acordo com as variâncias. Isto produz a chamada **distância estatística.** Na figura a seguir observa-se que a variância da v.a no sentido horizontal é maior que a variância da v.a no sentido vertical $V(X_1) > V(X_2)$



Assim, pondera-se as v.a's dividindo-as pelo seu desvio padrão, ou seja:

$$x_1^* = x_1/\sigma_1 \qquad \text{e} \qquad x_2^* = x_2/\sigma_2$$

E a distância Euclidiana entre o ponto $P^*(X_1, X_2)$ e a origem O(0,0) é:

$$d(P^*O) = d_{ij} = \sqrt{\frac{x_1^2}{s_1^2} + \frac{x_2^2}{s_2^2}}$$ que é conhecida como DISTÂNCIA ESTATÍSTICA.

Considerando $\underline{x}_i'$ e $\underline{y}_j' \in R^3$, com σ_1, σ_2 e σ_3 sendo os desvios padrões das v.a's correspondentes às componentes (direções) 1, 2 e 3, a distância Estatística entre os pontos $\underline{x}_i'$ e $\underline{y}_j'$ é dada por:

$$d_{ij} = \sqrt{\frac{(x_{i1} - y_{j1})^2}{\sigma_1^2} + \frac{(x_{i2} - y_{j2})^2}{\sigma_2^2} + \frac{(x_{i3} - y_{j3})^2}{\sigma_3^2}}$$

É fácil perceber que a diferença entre a distância Euclidiana e a distância Estatística está nos pesos (inversos das variâncias) e que quando as variâncias são iguais usa-se a distância Euclidiana.

MÉTRICA DE MAHALANOBIS

É bastante geral, pois leva em conta os padrões de covariância que pode existir nos dados. Sua expressão para a distância entre os pontos x_i e $x_j \in R^p$, considerando que Σ é a matriz de covariância correspondente a matriz de dados X é:

$$D_{ij}^2 = (\underline{x}_i - \underline{x}_j)'\Sigma^{-1}(\underline{x}_i - \underline{x}_j)$$

A chamada distância de Mahalanobis é a raiz quadrada de D^2.

DISTÂNCIA DE MINKOWSKI ou MÉTRICA DE MINKOWSKI

Existe uma classe geral de distâncias conhecida como Minkowski p-metrics (ou L_p metrics) que é definida pela equação:

$$d_{ij}(p) = [\sum_k | x_{ik} - x_{jk} |^p]^{1/p}$$

Dessa forma a distância Euclidiana é apenas um caso especial da métrica de Minkowski com p = 2, $d_{ij}(2) = [\sum_k | x_{ik} - x_{jk} |^p]^{1/2}$

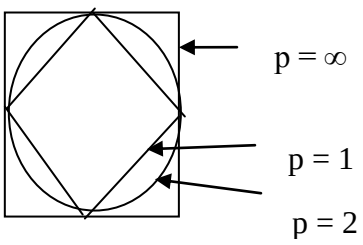
A métrica de Minkowski tem ainda dois casos especiais, que são:

com p = 1 que corresponde a **distância de city-block** $d_{ij}(1) = \sum_k | x_{ik} - x_{jk} |$

com p = ∞ que corresponde a **distância sup-metric**

$d_{ij}(\infty) = \max(|x_{j1} - x_{j1})|, |x_{i2} - x_{j2}|, \ldots , |x_{ip} - x_{jp}|)$

Observe a figura adiante, o quadrado exeterno corresponde a $d_{ij}(\infty)$ sup-metric; o círculo corresponde a $d_{ij}(2)$ distância Euclidiana e o quadrado interno ao círculo corresponde a $d_{ij}(1)$ city-block.



1.4- Medidas de Similaridade

1.4.1- Introdução

Muitas vezes as variáveis estudadas só podem ser medidas na escala nominal e, conseqüentemente, não é adequado calcular uma medida de distância. O procedimento adotado, então, é baseado no pareamento de atributo. Assim, tem-se as medidas de similaridade que consideram atributos comuns.

EXEMPLO 1:

Considere quatro refrigerantes e quatro atributos relacionados na tabela a seguir:

	ATRIBUTO			
REFRIGERANTE	**Sabor cola**	**Cafeína**	**Diet**	**Fabricado pela Coca-cola**
Coca-cola	1	1	0	1
Pepsi-cola	1	1	0	0
Diet Coca	1	1	1	1
Livre de Cafeína e Diet Coca	1	0	1	1

Uma medida de similaridade entre Coca-cola e Pepsi-cola corresponde ao número de empates no total de atributos, ou seja, ¾. Pode-se construir a matriz de similaridade entre os quatro produtos com base nessa medida de similaridade. A seguir tem-se essa matriz de similaridade.

$$S = \begin{array}{c} \\ \text{Coke} \\ \text{Pepsi} \\ \text{DCoca} \\ \text{CFDCoke} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Coke} & \text{Pepsi} & \text{DPepsi} & \text{CFDCoke} \end{array} \\ \begin{bmatrix} 1 & & & \\ 3/4 & 1 & & \\ 3/4 & 2/4 & 1 & \\ 2/4 & 1/4 & 3/4 & 1 \end{bmatrix} \end{array}$$

1.4.2- Coeficientes de Similaridade

O entendimento do conceito de coeficiente de similaridade fica mais claro a partir do próximo exemplo.

EXEMPLO

Seja p = 5 variáveis binárias que indicam a presença (1) ou a ausência (0) de certas características nos objetos A e B, na tabela adiante:

		CARACTERÍSTICAS			
OBJETO	C1	C2	C3	C4	C5
a	1	0	0	1	1
B	1	1	0	1	0

A distância Euclidiana entre A e B ao quadrado $d^2(A,B) = \|\underline{a} - \underline{b}\|^2$ é dada por:

$$d^2(A,B) = 0^2 + (-1)^2 + 0^2 + 0^2 + 1^2 = 2$$

E, $d^2(A,B)$ fornece uma medida do número de **não emparelhamentos** no par de objetos e é claro que um número grande de **não emparelhamentos** indica **uma menor semelhança**.

Fica claro que uma ponderação nos empates (emparelhamentos) em (1-1) e (0-0) é necessária, pois pode ocorrer da presença de uma característica ser mais forte do que a ausência. Por exemplo: se 1 significa "lê grego antigo" é óbvio que o empate em 1-1 é maior indicador de semelhança que o empate 0-0 (não lê grego antigo). Assim é razoável diminuir o número de igualdades 0-0 ou até desconsiderá-las completamente.

Portanto, desse tratamento diferenciado para empates 1-1 e 0-0 surgiram diversos esquemas para definir os coeficientes de similaridades.

Seja a tabela de contingência para os itens i e k:

		item k		
		1	0	TOTAL
item i	1	a	b	a + b
	0	c	d	c + d
TOTAL		a + c	b + d	p = a + b + c + d

onde: a é a freqüência de igualdades 1-1
b é a " " desigualdades 1-0
c é a " " " 0-1
d é a " " igualdades 0-0

Assim, os **coeficientes usuais de similaridade** são dados na tabela adiante:

COEFICIENTE $\tilde{s}(i,k)$	PONDERAÇÃO
1) $\frac{a+d}{p}$	1) Pesos iguais para 1-1 e 0-0.
2) $\frac{2(a+d)}{2(a+d)+b+c}$	2) Pesos em dobro para 1-1 e 0-0.
3) $\frac{a+d}{a+d+2(b+c)}$	3) Pesos em dobro para as desigualdades 1-0 e 0-1.
4) $\frac{a}{p}$	4) Desconsiderando 0-0 no numerador.
5) $\frac{a}{a+b+c}$	5) Desconsiderando 0-0 no numerador e denominador.
6) $\frac{2a}{2a+b+c}$	6) Desconsiderando 0-0 no numerador e denominador e com peso em dobro para 1-1
7) $\frac{a}{a+2(b+c)}$	7) Peso em dobro para as desigualdades
8) $\frac{a}{b+c}$	8) Razão entre as igualdades e desigualdades, excluindo 0-0.

EXERCÍCIO

Suponha que cinco indivíduos possuem as características listadas na tabela adiante:

indivíduo	altura	Peso	cor dos olhos	cor dos cabelos	habilidade manual	Sexo
1	68 pol	140 lb	verde	louro	destro	feminino
2	73 “	185 “	castanho	castanho	“	masculino
3	67 “	165 “	azul	louro	“	masculino
4	64 “	120 “	castanho	castanho	“	feminino
5	76 “	210 “	castanho	castanho	canhoto	masculino

a) Defina variáveis binárias para as características.
b) Monte o quadro considerando as variáveis binárias definidas.
c) Usando o coeficiente de similaridade $\frac{a+d}{p}$ construa a matriz dos coeficientes de similaridades de ordem (5x5) para os n = 5 indivíduos.
d) Faça alguma conclusão com base nos coeficientes de similaridade.
e) Você teria alguma crítica a fazer ao coeficiente de similaridade $\tilde{s}(i,k)=\frac{a+d}{p}$?

1.4.3- Relação entre coeficiente de similaridade e distância

O coeficiente de similaridade entre os objetos i e k pode ser escrito em função da distância entre i e k, ou seja:

$$\tilde{s}\ (i,\ k)=\frac{1}{1+d(i,k)}\text{, onde } 0 \leq \tilde{s}\ (i,k) \leq\ 1$$

Observa-se que diminuindo a distância aumenta a similaridade e vice-versa. É sempre possível construir coeficientes de similaridades a partir das distâncias, contudo não é possível construir as distâncias a partir das similaridades, a não ser que a matriz $\tilde{S}$ seja

não-negativa definida e $\tilde{s}$ (i, i) = 1. Desta forma d(i, k) = $\sqrt{2(1-\tilde{s}(i,k))}$ tem as propriedades de uma distância.

1.4.4- Similaridade e medida de associação para pares de variáveis

Quando as variáveis são binárias, os dados podem ser colocados na forma de uma tabela de contingência. As variáveis, melhor do que os objetos delineiam as categorias. Para cada par de variáveis existem n objetos categorizados na tabela. Assim tem-se:

		Variável k		Total
variável i	1	a	b	a+b
	0	c	d	c+d
Total		a + c	b + d	n

e o coeficiente de correlação amostral, calculado com base na tabela de contingência, é $r = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$ que pode ser tomado como uma medida de similaridade entre i e k.

EXERCÍCIOS

1) Um conjunto de pares de medidas de duas v.a's tem o vetor médio $\underline{\mu}$' = [0, 0] e variâncias $\sigma_1^2 = 4$ e $\sigma_2^2 = 1$. Seja o ponto $\underline{x} \in R^2$ com coordenadas (x_1, x_2). Suponha que as v.a's X_1 e X_2 não sejam correlacionadas.

a) Calcule a distância estatística do ponto $\underline{x}$ de coordenadas (x_1, x_2) à origem.

R.: d(P,O) = $\sqrt{\frac{x_1^2}{4}+\frac{x_2^2}{1}}$

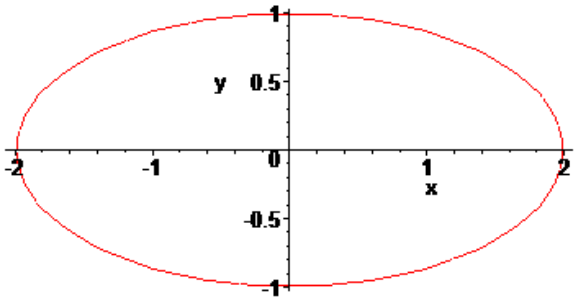
b) Construa o gráfico do lugar geométrico dos pontos cuja distância estatística à origem é 1.

R.: O lugar geométrico especificado é a elipse com equação $\frac{x_1^2}{4}+\frac{x_2^2}{1} = 1$.

c) Escreva também a equação deste lugar geométrico para uma distância c e ainda o gráfico nesta situação genérica.

R.: $\frac{x_1^2}{\sigma_1^2}+\frac{x_2^2}{\sigma_2^2} = c^2$.

d) Faça o gráfico dos pontos cuja distância à origem é 1.



Considerando $x_1 = x$ e $x_2 = y$.

e) Faça o gráfico dos pontos cuja distância à origem é c.
R.: O gráfico é uma elipse e os pontos onde o eixo das abscissas corta a elipse são: $(-c\sigma_1, 0)$ e $(c\sigma_1, 0)$; e os pontos onde o eixo das ordenadas corta a curva são: $(0, -c\sigma_1)$ e $(0, c\sigma_1)$, ou melhor, (-2c, 0) e (2c, 0) na horizontal e (0, -c) e (0, c).

2) Escreva a expressão da distância estatística do ponto P de coordenadas x's ao ponto Q de coordenadas y's, ambos situados no R^p. Sabe-se que cada coordenada distinta tem variância σ_i^2 i = 1,2, ...,p.

R.: $d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2} + \frac{(x_2 - y_2)^2}{\sigma_2^2} + \ldots + \frac{(x_p - y_p)^2}{\sigma_p^2}}$

OBS.: O lugar geométrico dos pontos P que têm a mesma distância ao quadrado do ponto Q jazem sobre um **hiperelipsóide** de centro em Q cujos eixos maior e menor são paralelos aos eixos das coordenadas.

2. ÁLGEBRA MATRICIAL E VETORES ALEATÓRIOS

2.1 - Álgebra Matricial

Um arranjo $\underline{x}$ de números reais $x_1, x_{2, \ldots\ldots,} x_p$ é chamado **vetor** e é escrito como

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_p \end{bmatrix} \qquad \text{ou} \qquad \underline{x}' = [x_1\ x_2\ \ldots\ x_p] \text{ (vetor transposto).}$$

Um vetor pode ter o seu módulo contraído ou aumentado quando é **multiplicado** por uma constante c, $c\underline{x}' = [cx_1\ cx_2\ \ldots\ldots\ cx_p]$ e a **adição** de vetores é feita somando-se os elementos componentes dos vetores (ordenadamente), ou seja:

$$\underline{z} = \underline{x} + \underline{y} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_p \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_p \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \dots\dots\dots \\ x_p + y_p \end{bmatrix}$$

O **produto interno** dos vetores $\underline{x}$ e $\underline{y}$ de dimensão p é definido por $\underline{x}.\underline{y} = \underline{y}.\underline{x} = \underline{x}'\underline{y} = \sum_{i=1}^{p} x_i y_i$ (escalar).

Comprimento ou **norma** de um vetor p-dimensional $\underline{x}$ é definido como a raiz quadrada do produto interno do vetor por ele mesmo, ou seja,

$$\|\underline{x}\| = \sqrt{\underline{x}'.\underline{x}} = \sqrt{x_1^2 + x_2^2 + \dots\dots + x_p^2}$$

EXERCÍCIO

1) Dados os vetores $\underline{x}'$= [10 3 12] e $\underline{y}'$= [-2 1 0], pede-se:
   a) o vetor $3\underline{x}$;
   b) o vetor soma $\underline{x} + \underline{y}$;
   c) o comprimento ou norma de cada um dos vetores;
   d) a norma quadrática de cada um dos vetores;
   e) o ângulo entre os dois vetores.

Matriz: uma matriz A de ordem n x p é um arranjo retangular de números reais formado por n linhas e p colunas. Quando n = p a matriz é dita quadrada,

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}$$

Matriz Transposta: a matriz transposta, A', de A é formada quando se troca as linhas pelas colunas, obtendo-se A' de ordem p x n.

Matriz Simétrica: quando a matriz A é formada de modo que A' = A, então ela é chamada de simétrica.

Matriz Inversa: a matriz quadrada A de ordem pxp admite inversa representada por A^{-1} de ordem pxp se existe uma matriz A^{-1} tal que $AA^{-1} = I$, onde I é a matriz identidade de ordem p com 1's na diagonal principal e zeros fora dela. Assim, $AA^{-1} = A^{-1}A = I$ e $A^{-1} = \frac{1}{\det(A)}\text{adj}(A)$ com adj(A) sendo a matriz dos co-fatores transposta.

A condição técnica para que a inversa exista é que as p colunas da matriz sejam linearmente independentes.

EXERCÍCIOS

1) Verifique se os vetores $\underline{x}$'= [1 2 1], $\underline{y}$' = [1 0 -1] e $\underline{z}$'= [1 -2 1] são linearmente independentes.
2) Verifique se os vetores $\underline{x}$'= [1 1 3] e $\underline{y}$'= [4 4 12] são independentes.
3) Mostre que a matriz A = $\begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix}$ admite inversa.

Matriz Ortogonal: uma matriz quadrada A é chamada de ortogonal quando suas linhas consideradas como vetores são mutuamente perpendiculares e têm comprimentos unitários, isto é: A'A = I e consequentemente A' = $A^{-1.}$

Autovalores e autovetores: uma matriz quadrada A é dita ter um autovalor λ (*eigenvalue*) com correspondente autovetor $\underline{e}' \neq \underline{0}$ (*eigenvector*) se $A\underline{e} = \lambda\underline{e}$.

RESULTADO 2.1
Uma matriz quadrada simétrica A de ordem k x k tem k pares de autovalor e autovetor, ou seja:

$(\lambda_1,\underline{e}_1), (\lambda_2,\underline{e}_2), \ldots, (\lambda_k,\underline{e}_k)$

OBS. Os autovetores podem ser escolhidos de modo a terem o comprimento igual a 1, ou seja, $\underline{e}'.\underline{e} = 1$. Isto chama-se "padronizar os autovetores".

RESULTADO 2.2
Seja A uma matriz quadrada de ordem k x k e I a matriz identidade de ordem kxk, então os escalares $\lambda_1, \lambda_2, \ldots, \lambda_k$ satisfazendo a equação $|A - \lambda I| = 0$ são os autovalores de A.

EXERCÍCIOS:

1) Determine os autovalores e autovetores da matriz $\begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix}$.

R.: Resolva a equação $|A - \lambda I| = 0$ para achar os autovalores e, com eles, use a definição para achar os autovetores. Como a matriz não é simétrica não é possível usar o STATGRAPHICS para achar os autovalores e autovetores. Contudo isto pode ser feito com o MATLAB ou MAPLE, etc.

2) Determine os autovalores e autovetores da matriz $\begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$.

R.: Resolva a $|A - \lambda I| = 0$ para achar os autovalores e, com eles, use a definição para achar os autovetores. Na prática use o STATGRAPHICS seguindo o seguinte caminho: SPECIAL, MULTIVARIATE METHODS, PRINCIPAL COMP0NENTS.

Os autovalores são: λ_1 = 3.41421 e λ_2 = 0.585786, obtidos em Analysis Summary. Já os autovetores são: e_1' = [0.382683 0.92388] e e_2' = [0.92388 -0.382683] obtidos em Component Weights.

3) Dada a matriz $A = \begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix}$ verifique se 6 e $[1/\sqrt{2} \;\; -1/\sqrt{2}]$ formam um dos pares de autovalor/autovetor de A.

R.: Aplique a definição de autovalor e autovetor, ou seja, $A\underline{e} = \lambda\underline{e}$.

Formas Quadráticas: uma forma quadrática $Q(\underline{x})$ nas p variáveis $x_1, x_2, ... ,x_p$ é definida por $Q(\underline{x}) = \underline{x}'A\underline{x}$, onde $\underline{x}' = [x_1, x_2, ... ,x_p]$ e A é uma matriz quadrada de ordem pxp simétrica. Note que a forma quadrática pode ser escrita como

$$Q(\underline{x}) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_i x_j$$

EXERCÍCIO:

Escreva a forma quadrática $Q(\underline{x}) = [x_1 \; x_2]\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ como um polinômio.

R.: $Q(\underline{x}) = x_1^2 + 2x_1x_2 + x_2^2$.

Matriz positiva definida: a matriz A é positiva definida se $\underline{x}'A\underline{x} > 0 \; \forall \underline{x} \neq \underline{0}$.

Matriz positiva semi-definida: a matriz A é positiva semi-definida ou **não negativa** se $\underline{x}'A\underline{x} \geq 0 \; \forall \underline{x} \neq \underline{0}$.

RESULTADO 2.3: **Teorema da Decomposição Espectral (Decomposição de Jordan)**

Qualquer matriz simétrica A de ordem pxp pode ser escrita como

$$A = P\Lambda P' = \sum_{i=1}^{p} \lambda_i \, \underline{e_i}\underline{e_i}'$$

onde Λ é uma matriz diagonal formada com os autovalores de A e P é uma matriz ortogonal (P'P=I) cujas colunas são os autovetores padronizados (normalizados $\underline{e_i}'\underline{e_i} = 1$ e $\underline{e_i}'\underline{e_j} = 0$ $i \neq j$) de A.

EXERCÍCIOS:

1) Escrever a forma quadrática $Q(\underline{x}) = [x_1 \; x_2]\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ como polinômio.

2) Considere a matriz simétrica A = $\begin{bmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{bmatrix}$. Calcule os autovalores e autovetores de A.

R.: Os autovalores e autovetores poderiam ser calculados usando-se as equações: $|A - \lambda I| = 0$ e $A\underline{e} = \lambda\underline{e}$. Mas, dado a ordem 3x3 da matriz isto dá muito trabalho. Então, o que se pode fazer é usar o STATGRAPHICS, usando o caminho:
SPECIAL, MULTIVARIATE METHODS, PRINCIPAL COMP0NENTS.

Os autovalores são: $\lambda_1 = 18$, $\lambda_2 = 9$ e $\lambda_3 = 9$, obtidos em Analysis Summary. Já os autovetores são: $e_1' = [0.667\ -0.667\ 0.333]$, $e_2' = [-0.236\ 0.236\ 0.943]$ e $e_3' = [0.707\ 0.707\ 0.0]$ obtidos em Component Weights.

3) Mostre que a forma quadrática $Q(\underline{x}) = 3x_1^2 + 2x_2^2 - 2\sqrt{2}\, x_1 x_2$ pode ser escrita na forma $\underline{x}'A\underline{x}$.

4) Mostre que a matriz A = $\begin{bmatrix} 3 & -\sqrt{2} \\ -\sqrt{2} & 2 \end{bmatrix}$ é definida não-negativa.

R.: Pelo TDE $A = \sum_{1=1}^{2} \lambda_i \underline{e}_i \underline{e}_i'$ e pré e pós-multiplicando antes por $\underline{x}'$ e $\underline{x}$ tem-se:

$x'Ax = \underline{x}' \sum_{1=1}^{2} \lambda_i \underline{e}_i \underline{e}_i' \underline{x}$ e os autovalores de A são $\lambda_1 = 4$ e $\lambda_2 = 1$. Substituindo os valores 4 e 1 no somatório obtém-se como resultado $4y_1^2 + y_2^2 \geq 0$. Portanto, A é definida não negativa.

5) Verifique se a matriz A = $\begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix}$ é definida não negativa.

6) Determine a inversa da matriz A = $\begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix}$.

Matriz raiz quadrada: a decomposição espectral permite expressar a inversa de uma matriz quadrada em termos dos seus autovalores e autovetores e isto leva a uma matriz muito útil, que é a matriz raiz quadrada (exercício adiante).

Matriz idempotente: a matriz quadrada A de ordem p x p é chamada de idempotente se A A = $A^2 = A$

EXERCÍCIOS:

1) Seja uma matriz quadrada A, simétrica de ordem k x k, determine A^{-1} dada a matriz dos autovalores Λ e a matriz dos autovetores P (ortogonal).

R.: Pelo TDE a matriz $A = P\Lambda P'$ e dado que P é ortogonal tem-se que

$(P\Lambda P')(P\Lambda^{-1}P') = I$ e por identidade, já que $A^{-1}A = I$, $A^{-1} = P\Lambda^{-1}P'$.

2) Explique por que é possível escrever $A^{-1} = P\Lambda^{-1}P' = \sum_{i=1}^{k} \frac{1}{\lambda_i} \underline{e}\underline{e}'$.

R.: Porque a matriz Λ é diagonal e a inversa de uma matriz diagonal tem na sua diagonal principal os inversos dos elementos de Λ e P pode ser considerada como formada por uma linha com os autovetores nessa linha e P' formada por uma coluna tendo os autovetores nessa coluna.

3) Seja uma matriz quadrada A, simétrica de ordem k x k. Determine a matriz raiz quadrada de A, $A^{1/2}$, dada a matriz dos autovalores Λ e a matriz dos autovetores P (ortogonal).

R.: Pelo TDE a matriz $A = P\Lambda P'$ e dado que P é ortogonal tem-se que

$(P\Lambda^{1/2}P')\,(P\Lambda^{1/2}P') = P\Lambda P' = A$ e por identidade, já que $A^{1/2}A^{1/2} = A$, tem-se que a matriz raiz quadrada é $A^{1/2} = P\Lambda^{1/2}P'$.

2.2 - Matriz e Vetor Aleatório

Um **vetor aleatório** é o vetor cujos elementos são v.a's e de modo semelhante uma **matriz aleatória** é a matriz cujas entradas são v.a's.

Seja X uma matriz aleatória de ordem nxp, então:

$$E(X) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & .... & E(X_{1p}) \\ E(X_{21}) & E(X_{22}) & .... & E(X_{2p}) \\ .... & .... & .... & .... \\ E(X_{n1}) & E(X_{n2}) & .... & E(X_{np}) \end{bmatrix} \quad \text{onde } E(X_{ij}) = \int_{-\infty}^{\infty} x_{ij} f_{ij}(x_{ij}) dx_{ij}$$

Propriedades: sejam X e Y matrizes aleatórias de mesmas dimensões e sejam A e B matrizes de constantes (não-aleatórias) de dimensões compatíveis com X e Y. Então:

a) $E(X+Y) = E(X) + E(Y)$
b) $E(AXB) = AE(X)B$

e se $\underline{\mu}$ é $E(\underline{X}) = [\mu_1\ \mu_2\ .....\mu_p]'$ então μ_i é $E(X_i) = \mu_i$

Matriz de Covariância: de um vetor aleatório $\underline{X}$ é definida por,

$\Sigma = V(\underline{X}) = E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})'$

EXERCÍCIOS:

1) Construir a matriz de covariâncias do vetor aleatório $\underline{X} \in R^p$ a partir da definição anterior.

R.: $V(\underline{X}) = \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{bmatrix}$

2) Construir a matriz de correlação do vetor aleatório $\underline{X}$ a partir da matriz de covariância

R.: $\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$, com $\rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$

3) Mostre o resultado $V^{1/2}\rho V^{1/2} = \Sigma$.

R.: Monte a matriz desvio padrão $V^{1/2}$ que é uma matriz diagonal com os desvios padrões na diagonal principal; monte a matriz de correlação ρ considerando que em cada entrada tem-se $\rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$. Agora, multiplique as matrizes.

4) Dada a matriz de covariância a seguir, determine a matriz raiz quadrada $V^{1/2}$ e a matriz de correlação ρ.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{bmatrix}$$

R.: Monte a matriz desvio padrão $V^{1/2}$ que é uma matriz diagonal com os desvios padrões na diagonal principal. Já a matriz de correlação é formada considerando que em cada entrada $\rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$.

5) Faça um quadro que contenha definição, notação e exemplos triviais de:
matriz escalar, vetor coluna, vetor de unidades, matriz retangular, matriz quadrada, matriz diagonal, matriz identidade, matriz simétrica, matriz de unidades, matriz triangular superior, matriz triangular inferior, matriz assimétrica, matriz nula, matriz definida positiva, matriz definida não-negativa e matriz idempotente.

6) Faça um quadro que contenha as definições das seguintes operações com matrizes:
Adição, subtração, multiplicação por escalar, produto interno, multiplicação, traço de uma matriz e determinante.

7) Dadas as matrizes abaixo determine as operações indicadas em seqüência:

$$A = \begin{bmatrix} 1 & 2 & -1 \\ -1 & 3 & -1 \\ 2 & 2 & 4 \end{bmatrix} \quad B = \begin{bmatrix} 3 & 2 & -1 \\ 2 & 3 & 1 \\ -1 & 1 & 3 \end{bmatrix} \quad \text{e } C = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 3 & 2 \end{bmatrix}$$

a) A + B
b) A - B
c) A - 2B
d) A'+ B
e) (A+B)'
f) (3A'-2B)'
g) tr(A)
h) tr(B)
i) AB
j) BC

8) Calcule a matriz inversa de $A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}$

3 - MATRIZ DE DADOS, VETOR DE MÉDIAS E MATRIZ DE COVARIÂNCIA

3.1- Matriz de Dados

Uma matriz de dados com n unidades observacionais e p variáveis pode ser escrita na seguinte forma:

$$\text{unidades observacionais} \begin{cases} 0_1 \\ 0_2 \\ \dots \\ 0_i \\ \dots \\ 0_n \end{cases} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$

i = 1, 2, ... , n
j = 1, 2, ... , p

$$\text{onde } \underline{x}_{(i)} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{ij} \\ \dots \\ x_{ip} \end{bmatrix} \text{ (vetor linha)} \qquad \text{e} \quad \underline{x}_{(j)} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{ij} \\ \dots \\ x_{nj} \end{bmatrix} \text{ (vetor coluna)}$$

EXEMPLO 1

Matriz de dados com 5 estudantes como unidades observacionais e, idade em anos na entrada para a universidade, nota até 100 no exame de fim do 1º ano e sexo como as variáveis, respectivamente, X_1, X_2 e X_3. Veja adiante.

	Variáveis		
	X_1	X_2	X_3
Observações	idade	nota	sexo
1	18,45	70	1
2	18,41	65	0
3	18,39	71	0
4	18,70	72	0
5	18,34	94	1

EXERCÍCIO:

Para os dados do exemplo anterior escreva o vetor linha da 3ª unidade observacional e o vetor coluna da 2ª variável:

3.2- Vetor de Médias

Dada a matriz ${}_nX_p = (x_{ij})$, i = 1, ... , n itens e j = 1, ... , p variáveis, a média amostral da j-ésima variável é dada por:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$${}_nX_p = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \qquad {}_pX'_n = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots \\ x_{1p} & x_{2p} & \dots & x_{np} \end{bmatrix}$$

O vetor de médias amostral é dado por $\underline{\bar{x}}' = [\ \bar{x}_1 \quad \bar{x}_2 \quad \dots \quad \bar{x}_p\ ]$ e representa o centro de gravidade dos pontos amostrais sendo que $\bar{x}_i$ representa o centro de gravidade da amostra da variável X_i.

EXEMPLO 2:
Para a matriz de dados do exemplo 1, o vetor de médias pode ser obtido calculando-se a média de cada variável e montando o vetor médio. Computacionalmente, pode-se usar o STATGRAPHICS seguindo o caminho: DESCRIBE, NUMERIC DATA, MULTIPLE VARIABLE ANALYSIS.

$\underline{\bar{x}}' = [\ 18{,}458\ ;\ 74{,}40\ ;\ 0{,}4\ ]$

O vetor de médias pode ser escrito em notação matricial:

$$\underline{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i = \frac{1}{n} X' \underline{1}_n$$

3.3- Matriz de Covariâncias Amostral e Matriz de Correlação Amostral

- A variância amostral da j-ésima variável é:

$$s_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = s_j^2 \qquad j = 1, 2, \ldots, p \text{ variáveis}$$

- A covariância amostral entre a j-ésima e a k-ésima variável é:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \bar{x}_j\bar{x}_k \qquad i = 1,2,\ldots,n$$

$$k,j = 1,2,\ldots,p$$

- A matriz de ordem pxp, $S = (s_{jk})$, com os elementos dados pelas expressões acima é chamada MATRIZ DE COVARIÂNCIA AMOSTRAL.

EXERCÍCIOS:

1) Verifique a afirmação anterior.

2) Calcule o vetor de médias amostral para a matriz de dados do exemplo 1, usando a notação matricial.

3) Calcule o vetor de médias, a matriz de covariância e a matriz de correlação das variáveis aleatórias observadas segundo a matriz de dados seguinte. Os dados mostram os pesos de depósitos de cascas de 28 árvores (árvore da cortiça) em 4 direções (N,S,L,O).

N	L	S	O	N	L	S	O
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

R.: Computacionalmente, pode-se usar o STATGRAPHICS seguindo o caminho: DESCRIBE, NUMERIC DATA, MULTIPLE VARIABLE ANALYSIS.

Vetor médio amostral: $\overline{\underline{X}}' = [50.5357\ 46.1786\ 49.6786\ 45.1786]$

A matriz de covariância amostral é:

$$S = \begin{bmatrix} 290.406 & 223.753 & 288.438 & 226.271 \\ 223.753 & 219.93 & 229.06 & 171.374 \\ 288.438 & 229.06 & 350.004 & 259.541 \\ 226.271 & 171.374 & 259.541 & 226.004 \end{bmatrix}$$

A matriz de correlação amostral é:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.885367 & 0.904717 & 0.883219 \\ 0.885367 & 1 & 0.8256 & 0.76868 \\ 0.904717 & 0.8256 & 1 & 0.922808 \\ 0.883219 & 0.76868 & 0.922808 & 1 \end{bmatrix}$$

EXERCÍCIO

1) Para os dados do exemplo 3.1:
   a) Estime as variâncias das variáveis X_1, X_2, X_3.
   b) Repita o item a matricialmente.
   c) Estime o coeficiente de correlação ρ entre as variáveis X_1 e X_2 , X_2 e X_3.

3.4- Vetores Aleatórios

DEF 1**:** Um espaço de probabilidade é um trio $(\Omega, \mathcal{A}, P)$ onde :
   a) Ω é um conjunto não vazio (espaço amostral) ;
   b) $\mathcal{A}$ é uma σ-álgebra de subconjuntos de Ω ;
   c) P é uma medida de probabilidade em $\mathcal{A}$.

DEF. 2: Um vetor $\underline{X}' = (X_1 , X_2 , \ldots , X_p)$ cujas componentes são variáveis aleatórias definidas no mesmo espaço de probabilidade $(\Omega, \mathcal{A}, P)$, é chamado vetor aleatório p-dimensional.

DEF. 3: Função de Distribuição de Vetor Aleatório
A função de distribuição $F = F_{\underline{x}} = F_{x_1, x_2, \ldots, x_p}$ de um vetor aleatório
$\underline{X}' = (X_1 , X_2 , \ldots , X_P)$ é definida como :
$F(\underline{x}) = F(x_1,x_2,\ldots,x_P) = P(X_1 \leq x_1, X_2 \leq x_2,\ldots,X_N \leq x_P) \quad \forall (x_1,x_2,\ldots,x_P) \in R^P$

F é também chamada função de distribuição conjunta das variáveis aleatórias $X_1,X_2,\ldots,X_P$.

EXEMPLO:

Uma urna contém 3 bolas numeradas 1,2,3. Duas bolas são retiradas sucessivamente da urna , ao acaso e sem reposição. Seja X o número da 1ª bola retirada e Y o número da 2ª.

a) Escreva o espaço amostral Ω .
b) Escreva a distribuição conjunta de (X,Y).
c) Calcule a P(X<Y).
d) Calcule a F(1,2).

4- ANÁLISE DA ESTRUTURA DE COVARIÂNCIA

4.1- Componentes Principais

4.1.1- Introdução

A Análise de Componentes Principais procura explicar a estrutura de variância-covariância da matriz de dados através de combinações lineares não correlacionadas das p variáveis originais. Embora p componentes sejam necessárias para reproduzir a variabilidade total do sistema, freqüentemente muito dessa variabilidade pode ser explicada por um número pequeno, k, de componentes principais.

Neste caso, existe quase a mesma quantidade de informação nas k componentes que nas p variáveis originais. As k componentes principais podem, então, substituir as p variáveis iniciais e, o conjunto de dados original que consiste de n medidas das p variáveis, é reduzido para um formado por n medidas das k componentes principais.

A.C.P. freqüentemente revela relacionamentos que não são previamente suspeitos e permite interpretações que poderiam não ocorrer ordinariamente.

OBJETIVOS DA A.C.P.

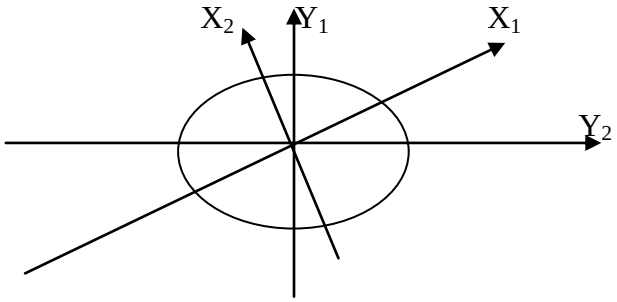
1º) Redução de dados;
2^0) Obtenção de v.a's não correlacionadas;
3º) Interpretação.

4.1.2- Componentes Principais da População

Algebricamente componentes principais são C.L.'s particulares das p variáveis aleatórias X_1, X_2, ... , X_p. Geometricamente estas C.L.'s representam a seleção de um novo sistema de coordenadas obtido por rotação do sistema original com X_1, X_2, ... , X_p como eixos. Os novos eixos Y_1, Y_2, ... ,Y_p representam as direções com variabilidade máxima e fornecem uma descrição mais simples e mais parcimoniosa da estrutura de covariância.

As componentes principais dependem da matriz de covariâncias Σ (ou da matriz de correlação ρ) das v.a's X_1, X_2, ... , X_p. O seu desenvolvimento não necessita da suposição de Gaussianidade. Por outro lado a Análise de Componentes Principais

derivada de populações normais multivariadas tem sua interpretação usual em termos de elipsóides de densidade constante,



Seja o vetor aleatório $\underline{X}' = [X_1, X_2, \ldots, X_p]$ que tem vetor de médias $\underline{\mu} = E(\underline{X})$ e matriz de covariância $\Sigma = V(\underline{X})$ com autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Considere as C.L's

$Y_1 = \underline{c}_1'\underline{X} = c_{11}X_1 + c_{21}X_2 + \ldots + c_{p1}X_p$

$Y_2 = \underline{c}_2'\underline{X} = c_{12}X_1 + c_{22}X_2 + \ldots + c_{p2}X_p$ $\qquad {}_pC_p = \begin{bmatrix} c_{11} & c_{21} & \ldots & c_{p1} \\ c_{12} & c_{22} & \ldots & c_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ c_{1p} & c_{2p} & \ldots & c_{pp} \end{bmatrix}$

.....
.....
$Y_p = \underline{c}_p'\underline{X} = c_{1p}X_1 + c_{2p}X_2 + \ldots + c_{pp}X_p$

Então, tem-se $\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_p \end{bmatrix} = \begin{bmatrix} c_{11} & c_{21} & \ldots & c_{p1} \\ c_{12} & c_{22} & \ldots & c_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ c_{1p} & c_{2p} & \ldots & c_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_p \end{bmatrix} = {}_pC_p\underline{X}$ e

- $V(Y_i) = V(\underline{c}_i'\underline{X}) = \underline{c}_i'V(\underline{X})\underline{c}_i = \underline{c}_i'\Sigma\underline{c}_i$
- $cov(Y_i, Y_k) = \underline{c}_i'\Sigma\underline{c}_k$
- $cov(\underline{Y}_p) = V({}_pC_p\underline{X}) = C\Sigma C'$

As componentes principais são as C.L's não-correlacionadas $Y_1, Y_2, \ldots, Y_p$ cujas variâncias são tão grande quanto possível. Assim:

- a 1ª componente principal é a C.L com variância máxima, isto é, é a C.L. que maximiza $V(\underline{c}_1'\underline{X})$ sujeito a restrição $\underline{c}_1'\underline{c}_1 = 1$ (vetor de comprimento unitário);

- a 2ª componente principal é a C.L. que maximiza $V(\underline{c}^2\underline{X})$ sujeito a restrição $\underline{c}_2'\underline{c}_2 = 1$ e assim sucessivamente.

RESULTADO 4.1

Seja Σ a matriz de covariâncias associada ao vetor aleatório $\underline{X}' = [X_1, X_2, \ldots, X_p]$ e que tem os pares de autovalor-autovetor $(\lambda_1, \underline{e}_1), (\lambda_2, \underline{e}_2), \ldots, (\lambda_p, \underline{e}_p)$ onde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. A i-ésima componente principal é dada por $Y_i = \underline{e}_i'\underline{X}$, que tem $V(Y_i) = \underline{e}_i'\Sigma\underline{e}_i = \lambda_i$ e $cov(Y_i, Y_k) = 0$ $i \neq k$. Se algum λ_i é igual a outro, na escolha do correspondente vetor de coeficientes $\underline{e}_i$, Y_i então não é único.

EXERCÍCIOS

1) Enunciar e demonstrar o resultado sobre "**maximização de formas quadráticas para pontos na esfera unitária**".

R.: "Seja a matriz B de ordem pxp positiva definida com auto valores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ e com os respectivos autovetores padronizados $\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_p$. Então, o $\max_{\underline{x} \neq \underline{0}} \frac{\underline{x}'B\underline{x}}{\underline{x}'\underline{x}} = \lambda_1$ que é alcançado em $\underline{x} = \underline{e}_1$ e o $\min_{\underline{x} \neq \underline{0}} \frac{\underline{x}'B\underline{x}}{\underline{x}'\underline{x}} = \lambda_p$ que é alcançado em $\underline{x} = \underline{e}_p$, além disso considerando os vetores $\underline{x} \perp \underline{e}_1, \underline{e}_2, \ldots, \underline{e}_k$ o $\max \frac{\underline{x}'B\underline{x}}{\underline{x}'\underline{x}} = \lambda_{k+1}$ alcançado em $\underline{x} = \underline{e}_{k+1}$, com $k = 1, 2, \ldots, p-1$".

Prova: Seja P a matriz ortogonal cujas colunas são os autovetores $\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_p$ de B e Λ a matriz diagonal dos autovalores de B. Seja, ainda, $B^{1/2} = P\Lambda^{1/2}P'$ e $\underline{y} = P'\underline{x}$ com dimensões compatíveis. Então, com $\underline{x} \neq \underline{0}$ tem-se $\frac{\underline{x}'B\underline{x}}{\underline{x}'\underline{x}} =$

$$\frac{\underline{x}'B^{1/2}B^{1/2}\underline{x}}{\underline{x}'PP'\underline{x}} = \frac{\underline{x}'P\Lambda^{1/2}P'P\Lambda^{1/2}P'\underline{x}}{\underline{x}'PP'\underline{x}} = \frac{\underline{y}'\Lambda\underline{y}}{\underline{y}'\underline{y}} = \frac{\sum_{i=1}^{p}\lambda_i y_i^2}{\sum_{i=1}^{p} y_i^2} \leq \frac{\sum_{i=1}^{p}\lambda_1 y_i^2}{\sum_{i=1}^{p} y_i^2} = \lambda_1.$$

Agora, fazendo $\underline{x} = \underline{e}_1$ tem-se $\underline{y} = P'\underline{e}_1 = [1\ 0\ 0\ \ldots\ 0]'$ e substituindo este valor em $\frac{\underline{y}'\Lambda\underline{y}}{\underline{y}'\underline{y}}$ obtém-se $\frac{\underline{e}_1'\Lambda\underline{e}_1}{\underline{e}_1'\underline{e}_1} = \lambda_1$.

De modo semelhante prova-se para o menor autovalor λ_p.

Finalmente, com $\underline{x} = P\underline{y} = y_1\underline{e}_1 + \ldots + y_p\underline{e}_p$ e $\underline{x} \perp \underline{e}_1, \ldots, \underline{e}_k$ tem-se:

$0 = \underline{e}_i'\underline{x} = y_1\underline{e}_i'\underline{e}_1 + \ldots + y_p\,\underline{e}_i'\underline{e}_p = y_i \quad i \leq k.$

Conseqüentemente, para $\underline{x}$ perpendicular aos primeiros k autovetores $\underline{e}_i$ tem-se $\frac{\underline{x}'B\underline{x}}{\underline{x}'\underline{x}} = \frac{\sum_{i=1}^{p}\lambda_i y_i^2}{\sum_{i=1}^{p} y_i^2}$ e fazendo $y_{k+1} = 1$ e $y_{k+2} = \ldots\ldots = y_p = 0$ alcança-se o máximo em λ_i.

2) Prove o resultado 4.1, enunciado anteriormente.

Prova:

Do resultado "**maximização de formas quadráticas para pontos na esfera unitária**" tem-se que $\max_{\underline{x}\neq\underline{0}} \frac{\underline{x}'B\underline{x}}{\underline{x}'\underline{x}} = \lambda_1$ e com $B = \Sigma$ resulta $\max_{\underline{e}_1\neq\underline{0}} \frac{\underline{e}_1'\Sigma\underline{e}_1}{\underline{e}_1'\underline{e}_1} = \lambda_1 = V(Y_1)$, pois $Y_1 = \underline{e}_1'\underline{X}$ e $V(Y_1) = V(\underline{e}_1'\underline{X}\underline{e}_1) = \underline{e}_1'\Sigma\underline{e}_1$.

De modo semelhante prova-se para o menor autovalor λ_p.

E, do resultado citado tem-se que para $\underline{x}$ perpendicular aos primeiros k autovetores $\underline{e}_i$ tem-se o máximo de $\frac{\underline{x}'\Sigma\underline{x}}{\underline{x}'\underline{x}}$ em λ_{k+1} k = 1, 2, ... p-1 e com $\underline{x} = \underline{e}_{k+1}$ tem-se:

$\underline{e}_{k+1}'\Sigma\underline{e}_{k+1} = V(Y_{k+1})$

Finalmente, $cov(Y_i, Y_k) = cov(\underline{e}_i'\underline{x}\ ;\ \underline{e}_k'\underline{x}) = E[(\underline{e}_i'\underline{x} - \underline{e}_i'\underline{\mu})(\underline{e}_k'\underline{x} - \underline{e}_k'\underline{\mu})'] = \underline{e}_i'\Sigma\underline{e}_k$

E, pré-multiplicando a expressão da definição $\Sigma\underline{e}_k = \lambda_k\underline{e}_k$ por $\underline{e}_i'$ resulta $\underline{e}_i'\Sigma\underline{e}_k = \underline{e}_i'\lambda_k\underline{e}_k = 0$, pois os autovetores são perpendiculares.

3) Determine a variância da componente principal Y_i e a covariância entre Y_j e Y_k.

R.: $V(Y_i) = V(\underline{e}_i'\underline{x}) = \underline{e}_i'V(\underline{x})\underline{e}_i = \underline{e}_i'\Sigma\underline{e}_i$ e já foi provado que $\max_{\underline{x}\neq\underline{0}} \frac{\underline{x}'B\underline{x}}{\underline{x}'\underline{x}} = \lambda_1$ alcançado em $\underline{x} = \underline{e}_1$. Portanto, substituindo $\underline{x} = \underline{e}_1$ na expressão do máximo resulta em

$\underline{e}_i'\Sigma\underline{e}_i = \lambda_i = V(Y_i)$, pois $\underline{e}_i'\underline{e}_i = 1$, no denominador.

4) Prove que a soma das variâncias das v.a's X_i é igual a soma das variâncias das componentes principais Y_i e que é igual a soma dos autovalores de Σ matriz de covariância das v.a's.

R.: Da matriz de covariância Σ tem-se $\sigma_1^2 = \sigma_2^2 + ... + \sigma_p^2 = tr(\Sigma)$ e do TDE tem-se que $\Sigma = P\Lambda P' = tr(\Lambda P'P) = tr(\Lambda) = \lambda_1 + ... + \lambda_p = V(Y_i) = \sigma_1^2 = \sigma_2^2 + ... + \sigma_p^2$.

5) Determine o coeficiente de correlação entre a componente principal Y_j e a v.a. X_k.

R.: Seja $\underline{c}_k' = [0\ 0\ 0\ ...\ 0\ 1\ 0\ ...\ 0]$ com 1 na posição k, de modo que $X_k = \underline{c}_k'\underline{X}$ e pré-multiplicando a definição $\Sigma\underline{e}_i = \lambda_i\underline{e}_i$ por $\underline{c}_k'$ resulta: $\underline{c}_k'\Sigma\underline{e}_i = \underline{c}_k'\lambda_i\underline{e}_i$. Então, como $cov(X_k, Y_i) = \underline{c}_k'\Sigma\underline{e}_i = \underline{c}_k'\lambda_i\underline{e}_i = \lambda_i\underline{e}_{ki}$ tem-se que:

$$\rho(Y_i, X_k) = \frac{cov(Y_i, X_k)}{\sqrt{V(Y_i)}\sqrt{V(X_k)}} = \frac{\lambda_i e_{ki}}{\sqrt{\lambda_i}\sqrt{\sigma_k^2}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sigma_k}$$

O resultado enunciado e demonstrado no exercício 4 garante que a variância total populacional é igual a soma das variâncias das componentes principais. E,

consequentemente, a proporção da variância total explicada (devido a) pela k-ésima componente principal é:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p} \qquad k = 1, 2, ... , p$$

Por outro lado, se a maior parte da variância populacional pode ser atribuída a uma, duas ou k componentes, então estas k componentes podem substituir as p variáveis originais sem muita perda de informação.

EXERCÍCIO:

Dada a matriz de covariância $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ do vetor aleatório $\underline{X}' = (X_1, X_2, X_3)$:

a) determine os pares de autovalor-autovetor da matriz de covariância;
R.: Digite a matriz Σ na planilha do STATGRAPHICS e siga o seguinte caminho: SPECIAL, MULTIVARIATE METHODS, PRINCIPAL COMPONENTS (não padronize). O resultado é:

Principal Components Analysis

Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	5.82843	72.855	72.855
2	2.0	25.000	97.855
3	0.171573	2.145	100.000

	Component 1	Component 2	Component 3
X_1	-0.382683	0.0	0.92388
X_2	0.92388	0.0	0.382683
X_3	0.0	1.0	0.0

b) escreva as componentes principais;
R.: A 1ª. componente principal é $Y_1 = \underline{e}_1' \underline{X} = -0.382X_1 + 0.923X_2$

c) determine a $V(Y_1)$;
R.: $V(Y_1) = 5{,}82843$

d) escreva a matriz C das combinações lineares referente às componentes principais;
R:

$${}_pC_p = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{12} & c_{22} & \dots & c_{2p} \\ \dots & \dots & \dots & \dots \\ c_{1p} & c_{2p} & \dots & c_{pp} \end{bmatrix}$$

e) calcule a matriz de covariâncias do vetor de componentes principais $\underline{Y}$;

R.: $V(\underline{Y}) = \begin{bmatrix} V(Y_1) & 0 & 0 \\ 0 & V(Y_2) & 0 \\ 0 & 0 & V(Y_3) \end{bmatrix} = \begin{bmatrix} 5{,}82843 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0{,}171573 \end{bmatrix}$

OBS. Fora da diagonal principal são zeros, pois as componentes são não correlacionadas.

f) determine a proporção da variância total que cabe a cada uma das componentes principais;

R.:

Component Number	Percent of Variance	Cumulative Percentage
Y1	72.855	72.855
Y2	25.000	97.855
Y3	2.145	100.000

g) qual a proporção da variância total explicada pelas duas primeiras componentes principais;

R.: 97,855%

h) calcule os coeficientes de correlação entre a componente Y_1 e as v.a's X_1 e X_2;

R.: $\rho(Y_i, X_k) = \dfrac{e_{ki}\sqrt{\lambda_i}}{\sigma_k}$, então

$$\rho(Y_1, X_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sigma_1} = \frac{-0{,}38268\sqrt{5{,}82843}}{1} = -0{,}923$$

$$\rho(Y_1, X_2) = \frac{e_{21}\sqrt{\lambda_1}}{\sigma_2} = \frac{0{,}92388\sqrt{5{,}82843}}{\sqrt{5}} = 0{,}997$$

i) calcule os coeficientes de correlação entre a componente Y_2 e as v.a's X_1, X_2 e X_3 ;

R.: 0, 0, 1.

j) faça alguma conclusão quanto as componentes principais e as v.a's originais.

R.: As componentes Y_1 e Y_2 podem substituir as três variáveis originais com pouca perda de informação. A componente Y_2 é idêntica à variável X_3.

4.1.3 Componentes principais obtidas de v.a's padronizadas

As componentes principais podem ser obtidas, também, de v.a's padronizadas, ou seja, de $Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$ i = 1,2, ... ,p que em notação matricial é $\underline{Z} = (V^{1/2})^{-1}[\underline{X} - \underline{\mu}]$

onde $V^{1/2}$ é a matriz desvio-padrão da forma $V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & 0 & 0 \\ 0 & \sqrt{\sigma_{22}} & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \sqrt{\sigma_{pp}} \end{bmatrix}$.

É claro que $E(\underline{Z}) = \underline{0}$ e é também fácil verificar que $V^{1/2}\rho V^{1/2} = \Sigma$. As componentes principais de $\underline{Z}$ podem ser obtidas dos autovalores e autovetores da matriz de correlação ρ de $\underline{X}$.

RESULTADO 4.2
A i-ésima componente principal das v.a's padronizadas Z_i i = 1, 2, ... , p com matriz de covariância $V(\underline{Z}) = \rho$ é dada por $Y_i = \underline{e_i}'\underline{Z}$ e com a soma das variâncias de Y_i igual a p e a correlação entre Y_i e Z_k, $\rho(Y_i,Z_k) = e_{ki}\sqrt{\lambda_i}$ i,k = 1, 2, ... , p, sendo neste caso que os autovalores e autovetores são obtidos da matriz de correlação.

Assim, pelo resultado anterior vemos que a proporção da variância populacional (padronizada) devido a k-ésima componente principal é dada por:

$$\lambda_k/p \qquad k = 1,2, \dots ,p.$$

EXERCÍCIOS:

1) Prove o resultado 4.2.
R.: Semelhante ao 4.1.

2) Seja a matriz de covariância $\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$ do vetor aleatório $\underline{X}' = [X_1\ X_2]$.

   a) Determine a matriz de correlação do vetor;

   R.: $\rho = \begin{bmatrix} \frac{1}{1.1} & \frac{4}{1.10} \\ \frac{4}{1.10} & \frac{100}{10.10} \end{bmatrix} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$

   b) determine os pares de autovalor-autovetor de Σ;

R.: Principal Components Analysis

Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	100.161	99.170	99.170
2	0.838647	0.830	100.000

	$\underline{e}_1$	$\underline{e}_2$
	0.0403055	0.999187
	0.999187	-0.0403055

c) determine os pares de autovalor-autovetor de ρ;
R.:

Principal Components Analysis

Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	1.4	70.000	70.000
2	0.6	30.000	100.000

	$\underline{e}_1$	$\underline{e}_2$
X_1	0.707107	0.707107
X_2	0.707107	-0.707107

d) determine as componentes principais, por Σ;
R.:

e) determine as componentes principais, por ρ;
R.:

f) determine as variâncias das componentes principais, por Σ e por ρ;

g) determine a proporção da variação total que cabe p/c/ uma das componentes principais;

h) calcule os coeficientes de correlação entre Y_1 e X_1 e X_2 ;

i) calcule os coeficientes de correlação entre Y_1 e Z_1 e Z_2;

5.1.4 Componentes principais a partir da amostra

Suponha que $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ são n observações do vetor aleatório $\underline{X}$ p-dimensional com vetor de médias $\underline{\mu}$ e matriz de covariância Σ. Estes dados produzem o vetor de médias amostral $\overline{\underline{x}}$, a matriz de covariância amostral S e, a matriz de correlação amostral R. Então, as componentes principais da amostra são determinadas a partir dessas matrizes e são definidas como as combinações lineares que maximizam a variância amostral. Sendo assim, com S como a matriz de covariância amostral de ordem pxp

com os pares de autovalor-autovetor $(\hat{\lambda}_1, \underline{\hat{e}}_1)$, $(\hat{\lambda}_2, \underline{\hat{e}}_2)$, ... , $(\hat{\lambda}_p, \underline{\hat{e}}_p)$ onde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$, a i-ésima componente principal amostral é dada por

$$\hat{Y}_i = \underline{\hat{e}}_i' \underline{x} = \hat{e}_{1i}x_1 + \hat{e}_{2i}x_2 + ... + \hat{e}_{pi}x_p \quad i = 1,2, ... ,p$$

A variância amostral de $\hat{Y}_k$ é dada por $\hat{\lambda}_k$ k = 1, 2, ... , p e valem os resultados provados para a situação populacional considerando o contexto amostral.

EXERCÍCIOS:

1) O censo de 1970 forneceu informações sobre 5 variáveis sócio-econômicas de determinada região. Os dados produziram os seguintes resultados:

$$\underline{\bar{x}} = \begin{bmatrix} 4{,}32 \\ 14{,}01 \\ 1{,}95 \\ 2{,}17 \\ 2{,}45 \end{bmatrix}$$

1ª linha - população (mil)
2ª linha - idade escolar média (anos)
3ª linha - total de empregados (mil)
4ª linha - total de empregados em serviços ligados a saúde (100)
5ª linha - valor médio de residências

$$S = \begin{bmatrix} 4{,}308 & 1{,}683 & 1{,}803 & 2{,}155 & -0{,}253 \\ 1{,}683 & 1{,}768 & 0{,}588 & 0{,}177 & 0{,}176 \\ 1{,}803 & 0{,}588 & 0{,}801 & 1{,}065 & -0{,}158 \\ 2{,}155 & 0{,}177 & 1{,}065 & 1{,}970 & -0{,}357 \\ -0{,}253 & 0{,}176 & -0{,}158 & -0{,}357 & 0{,}504 \end{bmatrix}$$

Pergunta-se:

Pode a variação total amostral ser sumariada por 1 ou 2 componentes principais?

2) Em um estudo sobre o relacionamento entre comprimento e forma de tartarugas pintadas, dois cientistas mediram o comprimento da carapaça, largura e altura. Seus dados, sugerem uma análise em termos de logaritmos. O objetivo do trabalho era dar um significado para os conceitos de "tamanho" e "forma". Foram feitas as medidas de X_1 (comprimento), X_2 (largura) e X_3 (altura) em 24 tartarugas machos e os valores estão no arquivo tipo STATGRAPHICS denominado apostila. Pede-se:

   a) Transforme as medidas em logaritmos neperianos;

R.: Para transformar a variável compTURTLE em logaritmo base e marque uma coluna vazia, aperte o botão da direita do mouse, escolha GENERATE DATA, vá na janela OPERATORS, escolha log(?), preencha o ? com a coluna compTURTLE. Repita o procedimento para as outras duas variáveis. As três variáveis logaritmadas aparecerão na planilha de dados do STATGRAPHCS.

b) Calcule a matriz de covariância dos dados transformados;
R.: Aplicando o STATGRAPHICS no caminho DESCRIBE, NUMERIC DATA, MULTIPLE VARIABLE ANALYSIS. O resultado aparecerá na tela covariances. Você pode salvar a matriz apertando o botão do disquete e marcando para salvar covariances.

$$S = \begin{bmatrix} 0.011072 & 0.00801914 & 0.00815965 \\ 0.00801914 & 0.00641673 & 0.00600527 \\ 0.00815965 & 0.00600527 & 0.00677276 \end{bmatrix} = 10^{-3} \begin{bmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8.160 & 6.005 & 6.772 \end{bmatrix}$$

c) Calcule a matriz de correlação dos dados transformados;
R.: Siga o mesmo caminho do item (b), resultado aparece na tela correlations. Você pode salvar a matriz apertando o botão do disquete e marcando para salvar correlation.

$$R = \begin{bmatrix} 1 & 0.951389 & 0.94227 \\ 0.951389 & 1 & 0.910947 \\ 0.94227 & 0.910947 & 1 \end{bmatrix}$$

d) Calcule os autovalores e autovetores de S;
R.: Você pode obtê-los das variáveis transformadas ou da própria matriz S que você salvou, dá no mesmo. Usando a matriz S siga o caminho: SPECIAL, MULTIVARIATE METHODS, PRINCIPAL COMPONENTS. Agora, vá na tela SUMMARY ANALYSIS, tecle o botão da direita do mouse, ANALYSIS OPTIONS e desmarque STANDARTIZED.

Principal Components Analysis

Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	**0.0233033**	96.051	96.051
2	**0.000598307**	2.466	98.517
3	**0.000359836**	1.483	100.000

Autovetores (vieram da tela component weights)

	Component 1	Component 2	Component 3
VMAT_1	0.683102	-0.159468	-0.7127
VMAT_2	0.51022	-0.594022	0.621944
VMAT_3	0.522539	0.788485	0.324414

e) Determine as componentes principais Y_1, Y_2 e Y_3;

R.: $Y_1 = 0{,}683\ln COMP + 0{,}510\ln LARG + 0{,}522\ln ALT$

$Y_2 = -0{,}159\ln COMP - 0{,}594\ln LARG + 0{,}622\ln ALT$

$Y_3 = -0{,}713\ln COMP + 0{,}622\ln LARG + 0{,}324\ln ALT$

f) Qual a porcentagem da variância total explicada por cada uma das componentes principais?

R.:

Principal Components Analysis

Component Number	**Eigenvalue**	Percent of Variance	Cumulative Percentage
1	**0.0233033**	96.051	96.051
2	**0.000598307**	2.466	98.517
3	**0.000359836**	1.483	100.000

g) Calcule os coeficientes de correlação entre as componentes e as v.a's originais;

R.: Basta aplicar a fórmula $\rho(Y_i, X_k) = \dfrac{e_{ki}\sqrt{\lambda_i}}{\sigma_k}$, então

$$\rho(Y_1, X_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sigma_1} = \frac{0.683\ \sqrt{0{,}0233}}{\sqrt{0{,}011072}}$$

As outras correlações são achadas de modo semelhante.

h) Tente interpretar a 1ª. componente principal.

3) Berce e Wilbaux (1935) coletaram medidas de 5 variáveis meteorológicas durante um período de 11 anos. As variáveis são:

ANO	X_1	X_2	X_3	X_4	X_5
1920-21	87.9	19.6	1.0	1661	28.37
1921-22	89.9	15.2	90.1	968	23.77
1922-23	153.0	19.7	56.61	1353	26.04
1923-24	132.1	17.0	91.0	1293	25.74
1924-25	88.8	18.3	93.7	1153	26.68
1925-26	220.9	17.8	106.9	1286	24.29
1926-27	117.7	17.8	65.5	1104	28.00
1927-28	109.0	18.3	41.8	1574	28.37
1928-29	156.1	17.8	57.4	1222	24.96
1929-30	181.5	16.8	140.6	902	21.66
1930-31	181.4	17.0	74.3	1150	24.37

a) Estime com base na amostra de tamanho n = 11 observações do vetor $\underline{X}$ o vetor médio populacional, a matriz de covariância e a matriz de correlação populacional;

b) Calcule os autovetores e os autovalores da matriz de correlação R;

c) Escreva as componentes principais com base na matriz de correlação R;

d) Determine a parcela da variação total explicada por cada uma das componentes principais;

e) Calcule os coeficientes de correlação entre as componentes principais Y_j e as variáveis originais X_j j=1, ... , 5;

f) Usando a matriz de covariâncias S que você obteve no item a determine as componentes principais e determine a proporção da variância total explicada por elas. Compare com o que você obteve anteriormente com base em R. Qual a interpretação é mais significativa.

5.2- Análise Fatorial

5.2.1- Introdução

A Análise Fatorial teve início modernamente no princípio do século XX com K. Pearson e C. Spearman, que estudaram as medidas de inteligência. A dificuldade nos cálculos impediu um desenvolvimento maior da técnica. O advento dos computadores altamente velozes trouxe de novo o interesse nos aspectos teóricos e computacionais da Análise Fatorial. O **objetivo** da A.F. é descrever, se possível, a estrutura de covariância dos relacionamentos entre muitas variáveis em termos de poucas variáveis fundamentais, mas não observáveis (latentes), aleatórias chamadas FATORES.

Suponha que variáveis possam ser agrupadas por suas correlações, isto é, todas as variáveis dentro de um grupo particular são altamente correlacionadas entre si, mas têm correlações relativamente baixas com variáveis de um grupo diferente. É admissível que cada grupo de variáveis represente um FATOR, que é responsável pelas correlações observadas.

Por exemplo:

1) As correlações no grupo das notas dos testes de Inglês, Francês, Matemática e Música sugerem um FATOR FUNDAMENTAL (não observável diretamente), "a inteligência".
2) A correlação alta entre as variáveis sabor e aroma na avaliação de um produto alimentício sugere um FATOR FUNDAMENTAL (não observável diretamente), o "paladar".

5.2.2- Objetivos da Análise Fatorial

O objetivo da análise fatorial é agrupar as informações contidas em um grande número de variáveis originais, em um conjunto menor de fatores com o mínimo de perda de informação. Em GONTIJO & AGUIRRE (1988) encontram-se descritos os objetivos da análise fatorial. São eles:

1. Harmonizar ou condensar um grande número de observações em grupos;
2. Obter o menor número de variáveis a partir do material original e reproduzir toda a informação de forma resumida;
3. Obter os fatores que reproduzam um padrão separado de relações entre as variáveis;
4. Interpretar de forma lógica o padrão de relações entre as variáveis;
5. Identificar as variáveis apropriadas para uma posterior análise de regressão e correlação ou análise discriminante.

Segundo os mesmos autores, existem certos fatores causais gerais na análise fatorial que originam as correlações observadas entre as variáveis, sendo assim pode-se considerar que muitas relações entre as variáveis são derivadas dos mesmos fatores causais gerais, e o número de fatores deverá ser menor que o número de variáveis.

Assim a análise fatorial, por meios de técnicas estatísticas, pode encontra uma forma resumida das informações contida na matriz de dados, transformando as muitas variáveis originais em um conjunto menor de novas variáveis estatísticas (fatores) com perda mínima de informações. Mais especificamente, as técnicas de análise fatorial atendem um entre dois objetivos:

- Identificar uma estrutura por meio do resumo dos dados: ao analisar as correlações entre as variáveis, torna-se possível identificar as relações estruturais existente entre essas variáveis. A análise fatorial, aplicada a um conjunto de variáveis é utilizada para identificar as dimensões latentes (fatores), enquanto a análise fatorial aplicada a uma matriz de correlação de respondentes individuais consiste em um método de agrupamento;

- Redução de Dados: por meio da análise fatorial, é possível identificar as variáveis representativas de um conjunto maior criando um novo conjunto de variáveis, muito menor que o original, que poderá substituir sem muito prejuízo, o conjunto original de variáveis.

Nos dois casos, o propósito é manter a natureza e o caráter das variáveis originais, reduzindo seu número para simplificar a análise multivariada a ser aplicada posteriormente sem comprometer o resultado da análise. PASCHOAL e TAMAYO (2004) sugerem o uso da técnica de análise fatorial como forma de validação de instrumentos de pesquisa, questionários ou coletas de dados, possibilitando o agrupamento dos itens da escala, bem como a identificação das variáveis representativas do conjunto original.

5.2.3- Suposições da Análise Fatorial

A verificação da suposição de Gaussianidade para os dados é necessária somente quando um teste estatístico for aplicado para verificar a significância dos fatores. Devido ao fato de que a análise fatorial identifica e agrupa conjuntos de variáveis inter-relacionadas, há necessidade de que exista um certo grau de multicolinearidade (uma variável pode ser explicada por outra variável) entre as variáveis, e a matriz de dados deve apresentar correlações não-nulas.

TESTE DE ESFERICIDADE DE BARTLETT

O teste de esfericidade de *Bartlett* é um dos meios de se verificar a adequação aplicação da análise fatorial aos dados. O teste identifica a presença de correlações não-nulas entre variáveis. Esse teste serve para testar a hipótese nula de que a matriz de correlação é uma matriz identidade. Se essa hipótese for rejeitada, então a análise fatorial pode ser aplicada (FERREIRA JÚNIOR, 2004). O teste examina a matriz de correlação interna, e fornece a probabilidade estatística de que a matriz de correlações possui correlações estatisticamente significativas entre pelo menos um par de variáveis, sendo que o teste é mais eficiente em detectar as correlações na medida em que se aumenta o tamanho da amostra.

A hipótese de interesse é dada por:

$H_0 : \Sigma = \Sigma_0 = \sigma^2 I$

O teste dessa hipótese é aplicado com base em uma a.a. de uma distribuição normal p-variada com vetor de médias $\underline{\mu}$ e natriz de covariância Σ. Então, a estatística do teste é:

CRITÉRIO DE KAISER-MEYER-OLKIN – KMO

O critério de Kaiser-Meyer-Olkin – KMO é outra forma para identificar se o modelo de análise fatorial que está sendo utilizado está adequadamente ajustado aos dados, isto se dá testando a consistência geral dos dados. O método verifica se a matriz de correlação inversa é próxima da matriz diagonal, consiste em comparar os valores dos coeficientes de correlação linear observados com os valores dos coeficientes de correlação parcial. A medida de adequacidade que fundamenta esse princípio é dada pela seguinte expressão:

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2}$$

em que r_{ij} é o coeficiente de correlação simples entre as variáveis X_i e X_j, e a_{ij} é o coeficiente de correlação parcial entre X_i e X_j, dados os outros $X^{'s}$.

Para interpretação do critério de KMO, os valores vão variar de 0 a 1, pois, pequenos valores de KMO indicam que o uso da análise fatorial não é adequada, e quanto mais próximo de 1, mais adequada é a aplicação da análise fatorial nos dados. Assim existe a seguinte referência conforme TABELA adiante.

TABELA – TABELA DE CRITÉRIO DE KAISER-MEYER-OLKIN – KMO

Tabela de Critério de Kaiser-Meyer-Olkin – KMO

Valor	Grau da Adequação da Amostra
> 0,90	Ótima
de 0,80 a 0,90	Boa
de 0,70 a 0,80	Razoável
de 0,60 a 0,70	Baixa
< 0,60	Inadequada

FONTE: Adaptado de MINGOTI,S. A. – Análise de dados através de métodos de estatística multivariada.
NOTA: Tabela para interpretação do Critério de Kaiser-Meyer-Olkin – KMO.

5.2.4- O Modelo Fatorial Ortogonal

Seja o v.a. observável $\underline{X}$, com p componentes, $\underline{X} \sim \cdot (\mu, \Sigma)$. O modelo fatorial postula que $\underline{X}$ é linearmente dependente sobre algumas variáveis aleatórias não observáveis (latentes) $F_1, F_2, \ldots, F_m$ ($m < p$) chamadas fatores comuns (m fatores extraídos dos p existentes) e p fontes de variação aditivas $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$, chamadas erros ou, algumas vezes, fatores específicos,

$$
\begin{aligned}
X_1 - \mu_1 &= \ell_{11} F_1 + \ell_{12} F_2 + \ldots + \ell_{1m} F_m + \varepsilon_1 \\
X_2 - \mu_2 &= \ell_{21} F_1 + \ell_{22} F_2 + \ldots + \ell_{2m} F_m + \varepsilon_2 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
X_i - \mu_i &= \ell_{i1} F_1 + \ell_{i2} F_2 + \ldots + \ell_{im} F_m + \varepsilon_i \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
X_p - \mu_p &= \ell_{p1} F_1 + \ell_{p2} F_2 + \ldots + \ell_{pm} F_m + \varepsilon_p
\end{aligned}
$$

ou, em notação matricial :

$$\underset{px1}{\underline{X} - \underline{\mu}} = \underset{pxm}{L} \underset{mx1}{\underline{F}} + \underset{px1}{\underline{\varepsilon}}$$

Os coeficientes ℓ_{ij} (entradas da matriz L) são chamados de pesos ou carregamentos e , especificamente, ℓ_{ij} é o carregamento na i-ésima variável do j-ésimo fator, tal que a matriz $\underset{pxm}{L}$ é a matriz de carregamentos (pesos) dos fatores. Note que o fator específico ou erro ε_i é associado somente com a i-ésima variável original X_i. Os desvios $X_1 - \mu_1$, $X_2 - \mu_2$, ... , $X_p - \mu_p$ são expressos em termos de p+m variáveis aleatórias: $F_1, F_2, \ldots, F_m, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ que não são observáveis. Isto distingue o modelo fatorial do modelo de regressão multivariada, cujas variáveis independentes (cujas posições são ocupadas por $\underline{F}$) podem ser observadas.

Agora vamos assumir que:

$E(\underline{F}) = \underset{mx1}{\underline{0}}$, $\quad cov(\underline{F}) = E(\underline{F}\underline{F}') = I_m$

$$E(\underline{\varepsilon}) = \underset{px1}{\underline{0}}, \quad cov(\underline{\varepsilon}) = E(\underline{\varepsilon}\underline{\varepsilon}') = \underset{pxp}{\Psi} = \begin{bmatrix} \Psi_1 & 0 & 0 & \dots & 0 \\ 0 & \Psi_2 & 0 & \dots & 0 \\ 0 & 0 & \Psi_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \Psi_p \end{bmatrix}$$

e que $\underline{F}$ e $\underline{\varepsilon}$ são <u>independentes</u>, assim

$$cov(\underline{\varepsilon},\underline{F}) = E(\underline{\varepsilon},\underline{F}') = \underset{pxp}{0} \quad \text{com } m = p$$

Com estas suposições o relacionamento construído em $\underset{px1}{\underline{X}} - \underset{px1}{\underline{\mu}} = \underset{pxm}{L}\ \underset{mx1}{\underline{F}} + \underset{px1}{\underline{\varepsilon}}$ é chamado <u>modelo fatorial ortogonal</u> e pode ser escrito $\underset{px1}{\underline{X}} = \underset{px1}{\underline{\mu}} + \underset{pxm}{L}\ \underset{mx1}{\underline{F}} + \underset{px1}{\underline{\varepsilon}}$

MATRIZ DE COVARIÂNCIA DO VETOR <u>X</u>:

Considerando a matriz:

$(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = (L\underline{F} + \underline{\varepsilon})(L\underline{F} + \underline{\varepsilon})' = (L\underline{F} + \underline{\varepsilon})((L\underline{F})' + \underline{\varepsilon}')$

$= L\underline{F}(L\underline{F})' + \underline{\varepsilon}(L\underline{F})' + L\underline{F}\underline{\varepsilon}' + \underline{\varepsilon}\underline{\varepsilon}'$

a matriz de covariância de $\underline{X}$ é :

$\Sigma = cov(\underline{X}) = E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = E[L\underline{F}(L\underline{F})' + \underline{\varepsilon}(L\underline{F})' + L\underline{F}\underline{\varepsilon}' + \underline{\varepsilon}\underline{\varepsilon}']$

$\Sigma = LE(\underline{F}\underline{F}')L' + E(\underline{\varepsilon}\underline{F}')L' + LE(\underline{F}\underline{\varepsilon}') + E(\underline{\varepsilon}\underline{\varepsilon}')$

$$= \underset{pxp}{LL'} + \underset{pxp}{0} + \underset{pxp}{0} + \underset{pxp}{\Psi}$$

$\Sigma = LL' + \psi$

Conseqüentemente tem-se $V(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \dots + \ell_{im}^2 + \psi_i \quad i = 1, 2, \dots , p.$

Assim, a matriz de covariância Σ pode ser decomposta em duas partes (matrizes) LL' e ψ. A matriz ψ é chamada de matriz de variâncias específicas; é uma matriz diagonal possuindo na diagonal principal as "variâncias específicas" ψ_i das variáveis originais. Já a matriz produto LL' tem na diagonal principal as comunalidades $h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \dots + \ell_{im}^2 \quad i = 1, 2, \dots , p \quad (j = 1, 2, \dots , m)$.

A covariância entre o vetor das variáveis originais $\underline{X}$ e o vetor dos fatores $\underline{F}$ é:

$cov(\underline{X}, \underline{F}) = E[(\underline{X} - \underline{\mu})(\underline{F} - \underline{0})'] = E[(\underline{X} - \underline{\mu})F'] = E[(\underline{\mu} + L\underline{F} + \underline{\varepsilon} - \underline{\mu})F']$

$= E[(L\underline{F} + \underline{\varepsilon})\underline{F}'] = LE(\underline{F}\underline{F}') + E(\underline{\varepsilon}\underline{F}') = LI_m + 0 = L$

$cov(X_i, X_k) = \ell_{i1}\ell_{k1} + \ell_{i2}\ell_{k2} + \dots + \ell_{im}\ell_{km}$

$\text{cov}(X_i, F_j) = \ell_{ij}$

COMUNALIDADES E VARIÂNCIAS ESPECÍFICAS :

A porção da variância da i-ésima variável aleatória X_i advinda como contribuição dos m fatores comuns (extraídos) é chamada de COMUNALIDADE e a porção da $V(X_i) = \sigma_{ii} = \sigma_i^2$ oriunda do fator específico é a VARIÂNCIA ESPECÍFICA. Assim, tem-se:

$$V(X_i) = V[\mu_i + \ell_{i1}F_1 + \ell_{i2}F_2 + ... + \ell_{im}F_m + \varepsilon_i] =$$
$$= V(\mu_i) + \ell_{i1}^2 V(F_1) + \ell_{i2}^2 V(F_2) + ... + \ell_{im}^2 V(F_m) + V(\varepsilon_i) = 0 + \ell_{i1}^2 \cdot 1 + \ell_{i2}^2 \cdot 1 + ... + \ell_{im}^2 \cdot 1 + \psi_i$$

$$V(X_i) = \underbrace{\ell_{i1}^2 + \ell_{i2}^2 + ... + \ell_{im}^2}_{\text{comunalidade}} + \underbrace{\Psi i}_{\text{variancia-especifica}}$$

$V(X_i) = h_i^2 + \Psi_i \quad i = 1, 2, ..., p \quad \text{com} \quad h_i^2 = \ell_{1i}^2 + \ell_{2i}^2 + ... + \ell_{mi}^2$ sendo a soma de quadrados dos carregamentos na i-ésima variável dos m fatores comuns (extraídos).

5.2.5- Estimação

Dadas as observações $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ de p variáveis geralmente correlacionadas a Análise Fatorial procura responder a pergunta:

"Representará o modelo fatorial os dados adequadamente, com um número m < p (baixo) de fatores?"

A matriz de covariância amostral S é um estimador da matriz de covariâncias populacional desconhecida Σ. Se os elementos fora da diagonal de S são baixos ou equivalentemente na matriz de correlação amostral R eles são praticamente nulos as variáveis não são relacionadas e a Análise Fatorial não é útil. Por outro lado quando Σ é significativamente diferente de uma matriz diagonal, então o modelo fatorial pode ser usado e o problema inicial é o de estimar os carregamentos (pesos) ℓ_{ij} e as variâncias específicas ψ_i. Vamos considerar no nosso estudo a estimação pelo Método das Componentes Principais. Seja Σ a matriz de covariâncias de X, então, dado que Σ seja positiva definida, podemos decompô-la na forma abaixo, segundo a decomposição espectral:

$$\Sigma = \lambda_1 \underline{e}_1 \underline{e}_1' + \lambda_2 \underline{e}_2 \underline{e}_2' + ... + \lambda_p \underline{e}_p \underline{e}_p'$$

$$\sum = [\sqrt{\lambda_1}\underline{e}_1 \sqrt{\lambda_2}\underline{e}_2 ... \sqrt{\lambda_p}\underline{e}_p] \begin{bmatrix} \sqrt{\lambda_1}\underline{e}_1' \\ \sqrt{\lambda_2}\underline{e}_2' \\ ... \\ \sqrt{\lambda_p}\underline{e}_p' \end{bmatrix} = \underset{pxp}{LL'} \qquad \text{se m = p, então, } \psi_i = 0 \ \forall_i$$

Assim, se $\Sigma = LL' + \psi$ tem-se ${}_p\psi_p = {}_p0_p$ no ajuste do modelo fatorial. Exceto pelo escalar $\sqrt{\lambda_j}$, os carregamentos no j-ésimo fator são os coeficientes populacionais na j-ésima componente principal. Embora a representação de $\Sigma = LL' + 0 = LL'$ seja

exata, ela não é particularmente útil, pois tem muitos fatores comuns. É preferível um modelo que explique a estrutura de covariância em termos de poucos fatores comuns. Uma aproximação, quando p - m autovalores são baixos, é negligenciar a contribuição de $\lambda_{m+1}\underline{e}_{m+1}\underline{e}'_{m+1} + \lambda_{m+2}\underline{e}_{m+2}\underline{e}'_{m+2} + ... + \lambda_p\underline{e}_p\underline{e}'_p$ para Σ na decomposição espectral. Assim, tem-se:

$$\Sigma \approx \sqrt{\lambda_1}\,\underline{e}_1\sqrt{\lambda_2}\,\underline{e}_2 \cdots \sqrt{\lambda_m}\,\underline{e}_m \begin{bmatrix} \sqrt{\lambda_1}\,\underline{e}'_1 \\ \sqrt{\lambda_2}\,\underline{e}'_2 \\ \cdots \\ \sqrt{\lambda_m}\,\underline{e}'_m \end{bmatrix} = \quad LL' \qquad \text{de ordem pxp}$$

Esta representação aproximada assume que os fatores específicos $\underline{\varepsilon}$ são de menor importância e podem, também, ser ignorados na fatorização de Σ. Se os fatores específicos $\underline{\varepsilon}$ são incluídos no modelo, suas variâncias são os elementos da diagonal da matriz diferença Σ - LL' e conseqüentemente $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}^2$ para i = 1,2, ... ,p.

Para aplicar esta abordagem aos dados amostrais $\underline{x}_1$, $\underline{x}_2$, ... , $\underline{x}_n$ é usual, primeiro, centrar as observações subtraindo a média amostral $\underline{\bar{x}}$. As observações centradas são:

$$[x_j - \underline{\bar{x}}] = \begin{bmatrix} x_{ij} - \bar{x}_1 \\ x_{2j} - \bar{x}_2 \\ ..... \\ x_{pj} - \bar{x}_p \end{bmatrix} \qquad j = 1,2,3, ..., n$$

Pode-se, também, trabalhar com as variáveis padronizadas,

$$\underline{z}_j = \begin{bmatrix} \dfrac{x_{ij} - \bar{x}_1}{\sqrt{s_{11}}} \\ \dfrac{x_{21} - \bar{x}_2}{\sqrt{s_{22}}} \\ \cdots \\ \dfrac{x_{pj} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \qquad j = 1,2, ..., n$$

cuja matriz de correlação amostral é a matriz de correlação R das observações originais $\underline{x}_1$, $\underline{x}_2$, ... ,$\underline{x}_n$. A representação $\Sigma \approx LL' + \psi$, quando se usa a matriz de covariância S ou, então, a matriz de correlação R, é conhecida como Solução Por Componentes Principais.

RESUMO DA SOLUÇÃO POR COMPONENTES PRINCIPAIS PARA O MODELO FATORIAL

A Análise Fatorial por Componentes Principais da matriz de covariância S é especificada em termos de seus pares de autovalor/autovetor (λ_1, $\underline{e}_1$),(λ_2, $\underline{e}_2$), ... ,(λ_p, $\underline{e}_p$) onde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Seja m < p o número de fatores comuns extraídos. A matriz dos carregamentos estimados $\hat{\ell}_{ij}$ é dada por:

$$\hat{L} = [\sqrt{\hat{\lambda}_1}\underline{\hat{e}}_1 \ \sqrt{\hat{\lambda}_2}\underline{\hat{e}}_2 \ ..... \ \sqrt{\hat{\lambda}_m}\underline{\hat{e}}_m]$$

As variâncias específicas estimadas são dadas pelos elementos da matriz $\hat{\Psi} = S - LL'$,

$$\hat{\Psi} = \begin{bmatrix} \hat{\psi}_1 & 0 & ... & 0 \\ 0 & \hat{\psi}_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \hat{\psi}_p \end{bmatrix} \text{ com } \hat{\psi}_{ii} = s_{ii} - \sum_{i=1}^{m} \hat{\ell}_{ij}^2$$

As comunalidades são estimadas por:

$$\hat{h}_i^2 = \hat{\ell}_{1i}^2 + \hat{\ell}_{2i}^2 + ... + \hat{\ell}_{mi}^2$$

E, para determinar o número m de fatores comuns, o indicado é basear-se na proporção da variância amostral devido a cada fator, que é:

$$\frac{\hat{\lambda}_j}{s_{11} + s_{22} + ... + s_{pp}}$$ para a análise feita a partir de S

$$\frac{\hat{\lambda}_j}{p}$$ para análise feita a partir de R

Considerando a solução por componentes principais partindo-se da matriz S ou R que fornece os pares de autovalores/autovetores ($\hat{\lambda}_1$, $\underline{\hat{e}}_1$), ($\hat{\lambda}_2$, $\underline{\hat{e}}_2$), ... ,($\hat{\lambda}_p$, $\underline{\hat{e}}_p$) onde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$ tem-se a matriz de carregamentos (pesos, *loads*)

$$\underset{pxm}{\hat{L}} = \left[\sqrt{\hat{\lambda}_1}\underline{\hat{e}}_1 \quad \sqrt{\hat{\lambda}_2}\underline{\hat{e}}_2 \quad ... \quad \sqrt{\hat{\lambda}_m}\underline{\hat{e}}_m\right] = \begin{bmatrix} \sqrt{\hat{\lambda}_1}\hat{e}_{11} & \sqrt{\hat{\lambda}_2}\hat{e}_{12} & ... & \sqrt{\hat{\lambda}_m}\hat{e}_{1m} \\ \sqrt{\hat{\lambda}_1}\hat{e}_{21} & \sqrt{\hat{\lambda}_2}\hat{e}_{22} & ... & \sqrt{\hat{\lambda}_m}\hat{e}_{2m} \\ ... & ... & ... & ... \\ \sqrt{\hat{\lambda}_1}\hat{e}_{p1} & \sqrt{\hat{\lambda}_2}\hat{e}_{p2} & ... & \sqrt{\hat{\lambda}_m}\hat{e}_{pm} \end{bmatrix}$$

e a matriz de variâncias específicas é: $\underset{pxp}{\hat{\Psi}} = \begin{bmatrix} \hat{\psi}_1 & 0 & ... & 0 \\ 0 & \hat{\psi}_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \hat{\psi}_p \end{bmatrix}$ com $\hat{\psi}_{ii} = s_{ii} - \sum_{j=1}^{m} \hat{\ell}_{ij}^2$

e onde as comunalidades estimadas são $\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + ... + \hat{\ell}_{im}^2 = \sum_{j=1}^{m} \hat{\ell}_{ij}^2$ e podemos interpretar estes resultados como:

- a contribuição do 1º. fator p/ a variância $s^2_i = s_{ii}$ da v.a. i é $\hat{\ell}^2_{i1}$
- a contribuição do 1º. fator p/ a variância total $s_{11}+s_{22}+ ... +s_{pp} = tr(S)$ é $\sum_{i=1}^{m} \hat{\ell}^2_{i1}$

EXERCÍCIOS:

1) Prove ou disprove que $\sum_{i=1}^{p} \hat{\ell}^2_{i1} = \hat{\lambda}_1$
2) Expresse a proporção da variância amostral devida ao i-ésimo fator em função dos carregamentos (pesos) e das variâncias das variáveis originais.

3) Em uma pesquisa sobre preferência do consumidor, uma a.a. de consumidores foi tomada e perguntou-se sobre os diversos atributos de um novo produto. As respostas, em notas numa escala até 7 pontos, foram tabuladas e a matriz de correlação R construída. As notas se referem a: sabor, preço, aroma, refeição rápida e se o sanduíche é nutritivo nesta ordem.

$$R = \begin{bmatrix} 1.00 & 0.02 & 0.96 & 0.42 & 0.01 \\ 0.02 & 1.00 & 0.13 & 0.71 & 0.85 \\ 0.96 & 0.13 & 1.00 & 0.50 & 0.11 \\ 0.42 & 0.71 & 0.50 & 1.00 & 0.79 \\ 0.01 & 0.85 & 0.11 & 0.79 & 1.00 \end{bmatrix}$$

a) Você distinguiria alguns grupos na matriz?
b) Calcule os autovalores e autovetores de R;
c) Especifique o número de fatores comuns na Análise Fatorial;
d) Estime a matriz dos carregamentos L;
e) Estime as comunalidades;
f) Estime as variâncias específicas;
g) Monte uma tabela com os carregamentos estimados dos fatores, comunalidades e variâncias específicas. Interprete os fatores.

5.2.6- Rotação dos Fatores

Os carregamentos obtidos mediante uma derivação dos carregamentos iniciais, mediante uma transformação ortogonal têm a mesma habilidade para reproduzir a matriz de covariância ou de correlação. Da álgebra matricial, nós sabemos que uma transformação ortogonal corresponde a uma rotação rígida dos eixos coordenados. Se $\hat{L}$ é a matriz estimada dos carregamentos dos fatores, então:

$\hat{L}^* = \hat{L}\,T$, onde $TT' = T'T = I$, T ortogonal,

é a matriz dos carregamentos "rotacionados". Além disso, a matriz de covariância (ou de correlação) permanece intacta, pois:

$$\hat{L}\ \hat{L}\,'+\ \hat{\Psi} = \hat{L}\,TT'\,\hat{L}\,' + \hat{\Psi}\ = \hat{L}^{*}\hat{L}^{*\prime} + \hat{\Psi}$$

e também a matriz dos resíduos

$$S - \hat{L}\ \hat{L}\ - \hat{\Psi}\ = S - \hat{L}^{*}\hat{L}^{*\prime} - \hat{\Psi}$$

permanece intacta e, ainda, as variâncias específicas ψ_i e as comunalidades h_i^2 não se alteram. Portanto, do ponto de vista matemático, não é importante se $\hat{L}$ ou $\hat{L}^{*}$ é definida. Às vezes não é fácil interpretar os carregamentos originais e, então, é usual fazer uma rotação dos carregamentos até que uma "estrutura simples" seja alcançada. Idealmente, nós gostaríamos de ter uma estrutura de cargas tal que cada variável tenha um alto peso em um único fator determinado e baixos ou moderados pesos nos demais fatores. Não é sempre possível obter esta estrutura simples, embora a rotação forneça uma estrutura próxima da ideal. Uma medida analítica da estrutura simples é o conhecido critério VARIMAX, que define $\tilde{e}_{ij}^{*} = \hat{e}_{ij}^{*} / \hat{h}_{ij}$ como sendo os coeficientes escalonados pela raiz quadrada das comunalidades. O critério seleciona a transformação ortogonal T que faz $V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{e}_{ij}^{*4} - (\sum_{i=1}^{p}\tilde{e}_{ij}^{*2})^2 / p\right]$ tão grande quanto possível. Escalonar os coeficientes rotacionados $\hat{e}_{ij}^{*}$ tem o efeito de dar às variáveis com pequenas comunalidades maior peso na determinação da estrutura simples. Após a transformação T ser determinada, os pesos $\hat{e}_{ij}^{*}$ são multiplicados por $\hat{h}_{ij}$ tal que as comunalidades originais sejam preservadas.

Quando m = 2, ou se considerarmos dois fatores comuns de uma vez, a transformação para uma estrutura simples pode freqüentemente ser determinada graficamente. Um gráfico dos pares de carregamentos $(\hat{e}_{1i}, \hat{e}_{2i})$ i = 1, 2, ... , p produzem p pontos, um para cada variável. Então os eixos podem ser rotacionados de um ângulo $\varnothing$, e os novos carregamentos obtidos $\hat{e}_{ij}^{*}$ são determinados: ${}_{p}\hat{L}^{*}{}_{p} = {}_{p}\hat{L}_{22}T_{2.}$

A matriz T é neste caso, $T = \begin{bmatrix} \cos\varphi & \operatorname{sen}\varphi \\ -\operatorname{sen}\varphi & \cos\varphi \end{bmatrix}$ com rotação no sentido horário.

EXERCÍCIOS:

4) Faça uma rotação dos fatores, para os dados do exercício 3.

5) Dados sobre os valores de ações consistem de n = 100 taxas semanais de p = 5 ações. As ações pertencem às empresas: Allied Chemical, Du Pont, Union Carbide, Exxon e Texaco. A matriz R é dada adiante e os dados estão na tabela T8.3. Obtenha os autovalores e autovetores de R. Sabendo, especificamente, que os carregamentos (pesos) estimados são os coeficientes das componentes principais amostrais multiplicados pela raiz quadrada dos autovalores correspondentes calcule os carregamentos estimados dos fatores e as variâncias específicas. Monte uma tabela com os resultados, inclusive proporção da variância amostral (padronizada)

explicada por cada fator para as soluções com m = 1 e m = 2 fatores. Procure interpretar os resultados da Análise Fatorial.

$$R = \begin{bmatrix} 1.000 & 0.577 & 0.509 & 0.387 & 0.462 \\ & 1.000 & 0.599 & 0.389 & 0.322 \\ & & 1.000 & 0.436 & 0.426 \\ & & & 1.000 & 0.523 \\ & & & & 1.000 \end{bmatrix}$$

6) Faça uma rotação nos fatores do problema 5 e construa o quadro completo da Análise Fatorial.

5.2.7- Escores Fatoriais

Na Análise Fatorial, o interesse usual está nos parâmetros do modelo fatorial. Contudo, os valores estimados dos fatores comuns, também chamados **escores fatoriais,** podem ser necessários. Estas quantidades são freqüentemente usadas para diagnosticar propostas da mesma forma que a análise anterior. Escores fatoriais não são estimativas de parâmetros desconhecidos no sentido usual. Mas, na verdade, eles são estimativas de **valores não observáveis** dos vetores de fatores aleatórios $\underline{F}j$**,** j = 1,2, ... , m. Isto é, escore fatorial $\underline{\hat{f}}$ é a estimativa do valor $\underline{f}_j$ assumido por $\underline{F}_j$. A estimação é complicada pelo fato de que as quantidades $\underline{f}_j$ e $\underline{\varepsilon}_j$ superam em número os valores observados $\underline{x}_j$. Para superar essa dificuldade são usadas aproximações para estimar os valores fatoriais. Existem, basicamente, dois métodos que têm dois elementos em comum:

1. Eles tratam os carregamentos estimados $\hat{\ell}_{ij}$ e as variâncias específicas $\hat{\psi}_i$ como se eles fossem os verdadeiros valores;

2. Eles envolvem transformações dos dados originais, padronizados. Tipicamente, os carregamentos rotacionados são melhores do que os carregamentos originais para se calcular os escores fatoriais.

5.2.7.1 Método dos Mínimos Quadrados

Supondo, de início, que o vetor médio $\underline{\mu}$, a matriz de carregamentos L e a matriz de variâncias específicas ψ sejam conhecidos no modelo $\underline{X} - \underline{\mu} = \underline{LF} + \underline{\varepsilon}$, então a soma dos quadrados dos erros, ponderados pelos recíprocos das suas variâncias, é:

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = \underline{\varepsilon}'\psi^{-1}\underline{\varepsilon} = (\underline{X} - \underline{\mu} - \underline{LF})' \, \psi^{-1}((\underline{X} - \underline{\mu} - \underline{LF}),$$

Bartlett propôs escolher os estimadores $\hat{\underline{f}}$ de $\underline{f}$ que minimizam a expressão anterior, resultando nas estimativas dos parâmetros populacionais pelo Método da Máxima Verossimilhança, no estimador:

$$\hat{\underline{f}}_j = (\hat{L}' \hat{\psi}^{-1} \hat{L})^{-1} \hat{L}' \hat{\psi}^{-1} (\underline{x}_j - \bar{\underline{x}}) \qquad j = 1, 2, \ldots\ldots, n$$

ou se a análise é feita a partir da matriz de correlação R

$$\hat{\underline{f}}_j = (\hat{L}_z' \hat{\psi}_z^{-1} \hat{L}_z)^{-1} \hat{L}_z' \hat{\psi}_z^{-1} \underline{z}_j \qquad j = 1,2, \ldots\ldots ,n$$

Mas, quando se usa Componentes Principais para estimar os carregamentos é costume estimar os escores fatoriais usando os Mínimos Quadrados Ordinários. Desta forma, as variâncias específicas ψ_i são consideradas como iguais ou como aproximadamente iguais e os escores são:

$$\hat{\underline{f}}_j = (\hat{L}' \hat{L})^{-1} \hat{L}' (\underline{x}_j - \bar{\underline{x}}) \qquad j = 1,2,3, \ldots\ldots ,n$$

ou se a análise é feita a partir da matriz de correlação R

$$\hat{\underline{f}}_j = (\hat{L}_z' \hat{L}_z)^{-1} \hat{L}_z' \underline{z}_j \qquad j = 1,2, \ldots\ldots ,n$$

EXERCÍCIOS:

7) Em uma Análise Fatorial obteve-se a matriz de carregamentos rotacionados obtidos por Máxima Verossimilhança e a matriz das variâncias específicas correspondente. Tem-se também a matriz de carregamentos e matriz das variâncias específicas obtidas por Componentes Principais. Tem-se, ainda, um vetor de dados (que corresponde a taxas semanais de retorno de investimentos em 5 companhias do setor químico (três primeiras variáveis) e setor petrolífero (duas últimas variáveis)) e o vetor médio para 100 semanas observadas. Obtenha os escores fatoriais correspondentes aos fatores 1 e 2 para a 50$^{a.}$ semana (vetor de dados apresentado). Repita o processo com os dados padronizados e para a observação indicada por z abaixo.

$$\hat{L} = \begin{bmatrix} 0.783 & -0.217 \\ 0.773 & -0.458 \\ 0.794 & -0.234 \\ 0.713 & 0.412 \\ 0.712 & 0.524 \end{bmatrix} \quad \hat{\Psi} = \begin{bmatrix} 0.34 & 0 & 0 & 0 & 0 \\ 0 & 0.19 & 0 & 0 & 0 \\ 0 & 0 & 0.31 & 0 & 0 \\ 0 & 0 & 0 & 0.27 & 0 \\ 0 & 0 & 0 & 0 & 0.22 \end{bmatrix}$$

$$\text{Taxas } \hat{L}_z^* = \begin{bmatrix} 0.601 & 0.377 \\ 0.850 & 0.164 \\ 0.643 & 0.335 \\ 0.365 & 0.507 \\ 0.208 & 0.883 \end{bmatrix} \quad \hat{\Psi}_z = \begin{bmatrix} 0.50 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0.47 & 0 & 0 \\ 0 & 0 & 0 & 0.61 & 0 \\ 0 & 0 & 0 & 0 & 0.18 \end{bmatrix}$$

$\underline{x}'$ = [0.068965 0.014663 0.016360 0.038135 0.063829]
$\underline{\bar{x}}'$ = [0.0054 0.0048 0.0057 0.0063 0.0037]

$\underline{z}'$ = [0.50 -1.40 -0.20 -0.70 1.40]

5.3. Análise de Correlação Canônica

5.3.1. Introdução

A Análise de Correlação Canônica é uma técnica estatística que trata da **identificação e quantificação da associação** entre dois grupos de variáveis. Basicamente, o objetivo dessa técnica é determinar as combinações lineares $\underline{c}_1'\underline{x}$ e $\underline{c}_2'\underline{y}$ tais que tenham a **maior correlação possível.** Tais correlações podem dar discernimento sobre o relacionamento entre os dois conjuntos de variáveis.

A idéia é determinar primeiro o par de c.l's que tenha maior correlação. Em seguida, determina-se o par de c.l's, seguinte, que tenha maior correlação, escolhido entre todos os pares não correlacionados com o primeiro par já selecionado. E assim sucessivamente. Os pares de c.l's são chamados de **variáveis canônicas** e suas correlações são as **correlações canônicas**. Pode-se entender a Análise de Correlação Canônica como uma extensão da Análise de Regressão Múltipla. Na Análise de Regressão Múltipla as variáveis formam o conjunto das covariáveis $\underline{x}$ com p –1 variáveis e o conjunto da variável resposta $\underline{y}$ com uma única variável. A solução do problema de regressão múltipla trata de achar a c.l. $\underline{\beta}'\underline{x}$ que é altamente correlacionada com y. Já na Análise de Correlação Canônica o conjunto $\underline{y}$ contém $p \geq 1$ variáveis e procura-se os vetores $\underline{c}_1$ e $\underline{c}_2$ para os quais a correlação entre $\underline{c}_1'\underline{x}$ e $\underline{c}_2'\underline{y}$ é máxima. Se $\underline{x}$ é interpretado como o *causador* de $\underline{y}$, então $\underline{c}_1'\underline{x}$ pode ser chamado o *melhor preditor* e $\underline{c}_2'\underline{y}$ o *mais provável critério*. As correlações canônicas medem a força da associação entre os dois conjuntos de variáveis. O aspecto de maximização da técnica representa uma tentativa de concentrar um relacionamento dimensionalmente alto entre dois conjuntos de variáveis em uns poucos pares de variáveis canônicas.

EXEMPLO
Suponha o vetor de variáveis $\underline{x}$ que corresponde a resultados de técnicas administrativas e o conjunto de variáveis $\underline{y}$ que correspondem a medidas de variáveis de qualidade. Alguém pode estar interessado em saber se o conjunto de técnicas administrativas é altamente relacionado com o conjunto das variáveis de qualidade e também predizer os resultados de um dos conjuntos em função do outro. A Análise de Correlação Canônica ajuda neste sentido.

5.3.2. Variáveis Canônicas e Correlações Canônicas

Seja um vetor aleatório p-dimensional $\underline{X}_1$ e outro vetor aleatório $\underline{X}_2$ q-dimensional $p \leq q$. Suponha que $\underline{X}_1$ e $\underline{X}_2$ tenham médias $\underline{\mu}_1$ e $\underline{\mu}_2$ e matrizes de covariância $\sum_1$ e $\sum_2$, então:

$V(\underline{X}_1) = E\{(\underline{X}_1 - \underline{\mu}_1)(\underline{X}_1 - \underline{\mu}_1)'\} = \sum_1$
$V(\underline{X}_2) = E\{(\underline{X}_2 - \underline{\mu}_2)(\underline{X}_2 - \underline{\mu}_2)'\} = \sum_2$
$cov(\underline{X}_1,\underline{X}_2) = E\{(\underline{X}_1 - \underline{\mu}_1)(\underline{X}_2 - \underline{\mu}_2)'\} = \Sigma_{12} = \Sigma_{21}'$ (matriz de covariância cruzada)

EXERCÍCIOS

1) Seja $\underline{X}_1$ de dimensão p = 2 e $\underline{X}_2$ de dimensão q = 3 escreva a expressão da matriz de correlação cruzada entre estes vetores.

2) Considere os dois vetores do exercício anterior conjuntamente, ou seja, formando um único vetor de dimensão p + q = 5. Determine:

   a) a esperança do vetor conjunto;
   b) a variância do vetor conjunto ;

3) Repita o exercício anterior considerando as dimensões p e q para os vetores.

4) Seja U = $\underline{c}_1^{'}\underline{x}$ e V = $\underline{c}_2^{'}\underline{y}$ c.l's dos vetores aleatórios $\underline{x}$ e $\underline{y}$. Determine a correlação entre U e V.

Você deve ter observado que a correlação entre U e V depende dos coeficientes $\underline{c}_1$ e $\underline{c}_2$, então quais os valores desses coeficientes que maximizam $\rho(U,V)$? A resposta é maximizar $\underline{c}_1^{'}\sum_{12}\underline{c}_2$ com a restrição de $\underline{c}_1^{'}\sum_1\underline{c}_1 = \underline{c}_2^{'}\sum_2\underline{c}_2 = 1$ para que a correlação não dependa da escala de $\underline{c}_1$ e c_2. Então definiremos como primeiro par de variáveis canônicas o par de c.l's U_1 e V_1 que tendo variâncias unitárias maximizarão a correlação $\rho(U,V)$; o segundo par seria obtido da mesma forma entre todas as escolhas não correlacionados com a primeira escolha e assim sucessivamente. Então, desta forma é enunciado o resultado seguinte:

RESULTADO 5.3.1

Sejam os vetores $\underline{X}$ e $\underline{Y}$ de dimensão p e q e com matrizes de covariâncias $\sum_1$ e $\sum_2$ respectivamente e covariância cruzada $\sum_{12}$ e ainda as c.l's U = $\underline{c}_1^{'}\underline{X}$ e V = $\underline{c}_2^{'}\underline{Y}$. Então a máxima corr(U,V) é alcançada em corr(U,V) = ρ_1^* com $\underline{c}_1^{'} = \underline{e}_1^{'}\sum_1^{-1/2}$ e $\underline{c}_2^{'} = \underline{f}_1^{'}\sum_2^{-1/2}$, onde $\underline{e}_1$ é o autovetor correspondente ao maior autovalor ρ_1^{*2} de $\sum_1^{-1/2}\sum_{12}\sum_2^{-1}\sum_{21}\sum_1^{-1/2}$ que tem p autovalores $\rho_1{}^{*2} \geq \rho_2{}^{*2} \geq .... \geq \rho_p{}^{*2}$ e p autovetores $\underline{e}_i$ i = 1,2, ...,p e $\underline{f}_1$ é o autovetor correspondente ao maior autovalor de $\sum_2^{-1/2}\sum_{21}\sum_1^{-1}\sum_{12}\sum_2^{-1/2}$ que tem q autovetores $\underline{f}_i$ correspondentes aos autovalores $\rho_1{}^{*2} \geq \rho_2{}^{*2} \geq .... \geq \rho_q{}^{*2}$.

EXERCÍCIOS

1) Escreva as expressões do primeiro e do k-ésimo par de variáveis canônicas de acordo com o resultado anterior.

2) Qual o valor de $V(U_k)$, $V(V_k)$, $corr(U_k,U_\ell)$ $k \neq \ell$, $corr(U_k,V_\ell)$ $k \neq \ell$ e $corr(V_k,V_l)$ $k \neq \ell$?

EXERCÍCIOS

1) Seja $\underline{Z}_1' = [Z_{11}\ Z_{21}]$ e $\underline{Z}_2' = [Z_{12}\ Z_{22}]$ vetores aleatórios formados com v.a's padronizadas e seja $\underline{Z}$ o vetor composto pelos dois anteriores tendo matriz de covariância $\rho = V(\underline{Z}) = \begin{bmatrix} 1 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1 & 0.3 & 0.4 \\ 0.5 & 0.3 & 1 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1 \end{bmatrix}$, faça uma análise de correlação canônica.

SOLUÇÃO:

- primeiro par de variáveis canônicas é: $U_1 = 0{,}856\ Z_{11} + 0{,}277 Z_{21}$
  $V_1 = 0{,}545\ Z_{12} + 0{,}737 Z_{22}$

- A correlação entre as variáveis canônicas do $1^{0.}$ par é: $\rho_1^* = \sqrt{\lambda_1} = \sqrt{0{,}5458} = 0{,}74$ indicando uma forte associação entre os dois conjuntos de variáveis, note que o primeiro par é sempre o mais importante;

- A correlação entre as variáveis canônicas do $2^{0.}$ par é: $\rho_2^* = \sqrt{\lambda_2} = \sqrt{0{,}0009} = 0{,}03$ indicando uma fraca associação entre os dois conjuntos de variáveis;

- As correlações entre as variáveis originais do primeiro conjunto, $\underline{Z}_1' = [Z_{11}\ Z_{21}]$ com a variável canônica U_1 são [0,97 0,62] e as correlações entre as variáveis originais do segundo conjunto, $\underline{Z}_2' = [Z_{12}\ Z_{22}]$ com a segunda variável canônica são [0,69 0,85]. Isto indica que as variáveis Z_{11} e Z_{22} são mais importantes do que as outras. Da mesma forma pode-se ter as correlações de U_1 com as variáveis de $\underline{Z}_2$ que são: [0,51 0,63] e de V_1 com $\underline{Z}_1$ que são: [0,71 0,46].

- É possível ainda afirmar que: "se $\underline{Z}_2' = [Z_{21}\ Z_{22}]$ é interpretado como o ***causador*** de $\underline{Z}_1' = [Z_{11}\ Z_{21}]$, então V_1 pode ser interpretado como o ***melhor preditor*** e U_1 o ***mais provável critério***.

5.3.3. Escores e Predição

As variáveis canônicas são em geral artificiais. Isto é, não têm significado físico. Mas estas variáveis podem ser "identificadas" em termos das suas variáveis principais. Esta identificação pode ser feita por meio da correlação entre as variáveis originais e as variáveis canônicas. Contudo esta interpretação deve ser feita com cautela.

O uso da correlação canônica para predição é feito do seguinte modo: sejam os coeficientes $\underline{a}_i$ e $\underline{b}_i$ das variáveis canônicas, ou seja, os vetores de correlação canônica, então os vetores de dimensão n $\underline{a}_i' \underline{X}_1$ e $\underline{b}_i' \underline{X}_2$ denotam os escores dos n indivíduos (dimensão de cada um dos vetores) nas i-ésimas variáveis canônicas, $U_i = \underline{a}_i' \underline{X}_1$ e $V_i = \underline{b}_i' \underline{X}_2$. Então o preditor de U_i dado V_i é dado por

$$\hat{U}_{ji} = \hat{\rho}^{*}_{ji}(V_{ji} - \underline{a}'\underline{\bar{x}}_1) + \underline{b}_i'\underline{\bar{x}}_2 \quad \text{para um j = 1,2, ... ,n}$$

e onde $\underline{\bar{x}}_i = \frac{1}{n}\sum_{j=1}^{n}\underline{x}_{ji}$ i = 1, 2

EXERCÍCIOS

2) Calcule as correlações entre o primeiro par de variáveis canônicas e suas variáveis componentes para a situação do exercício anterior.

3) Suponha 5 variáveis: X_1 e X_2 que correspondem a notas de prova sem consulta e Y_1, Y_2 e Y_3 que correspondem a notas de prova com consulta. Um problema que pode surgir aqui é o interesse em saber "quão altamente correlacionadas estão as habilidades dos estudantes examinados sem consulta com as habilidades dos estudantes examinados com consulta". Por outro lado, alguém pode usar os resultados da prova com consulta para predizer os resultados da prova sem consulta.

Solução: $U_1 = 0{,}0260x_1 + 0{,}0518x_2$ e $V_1 = 0{,}0824y_1 + 0{,}0081y_2 + 0{,}0035y_3$

As correlações canônicas são: $\hat{\rho}_1 = 0{,}6630$ e $\hat{\rho}_2 = 0{,}0412$

Os vetores médios são: [38,9545 50,5909] e [50,6023 46,6818 42,3068]

$$\hat{V}_{ji} = 0{,}6630[0{,}0260x_1 + 0{,}0518x_2 - [0{,}0260\ \ 0{,}0518]\begin{bmatrix}38{,}95\\50{,}59\end{bmatrix} + [0{,}0824\ \ 0{,}0081\ \ 0{,}0035]\begin{bmatrix}50{,}6\\46{,}68\\42{,}31\end{bmatrix} = 2{,}2905 + 0{,}0172x_1 + 0{,}0343x_2$$

De modo que se prevêem os valores de Y dado os valores de X.

6- DISCRIMINAÇÃO, CLASSIFICAÇÃO e RECONHECIMENTO de PADRÕES

6.1.Introdução

A técnica multivariada conhecida como "Análise Discriminante" trata dos problemas relacionados com **SEPARAR** conjuntos distintos de objetos (ou observações) e **FIXAR (alocar)** novos objetos (observações) em conjuntos previamente definidos. A análise discriminante quando empregada como **procedimento de classificação** não é uma técnica exploratória, uma vez que ela

conduz a regras bem definidas, as quais podem ser utilizadas para classificação de outros 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. Mas o que vem a ser Reconhecimento de Padrões? Vamos citar alguns exemplos de *máquinas inteligentes*:

- Míssil que escolhe por onde entrar em um abrigo (Guerra do Golfo);
- Carro que se desloca sozinho em um *campus* universitário;
- Carro que estaciona sozinho (já existe no mercado japonês);
- Máquina que classifica tábuas de madeira pela sua tonalidade de cor;
- Programa que identifica se uma pessoa com icterícia está com câncer ou colédoco entupito;
- 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 câmara de televisão para *visão* e um processador eletrônico como *cérebro*.

Historicamente as principais abordagens de Reconhecimento de Padrões eram: a abordagem estatística e a abordagem sintática (estrutural). Surgiram e ganharam bastante espaço a tecnologia de Redes Neurais e, também, a de métodos de Programação Matemática. Existem três questões importantes em Reconhecimento de Padrões:

- São estas técnicas adequadas ou mesmo aplicáveis para resolver problemas de reconhecimento e classificação?
- É possível desenvolver ou modificar modelos úteis para determinados problemas, determinando os parâmetros do modelo?
- Existem algoritmos que podem ser aplicados e que são computacionalmente práticos nos procedimentos de solução do problema?

O Reconhecimento de Padrões está relacionado com as seguintes áreas:

1. Processamento de sinais (equipamento bio-médico e outros);
2. Reconhecimento de caracteres (manual ou não);
3. Reconhecimento de faces;
4. Identificação de impressões digitais;
5. Análise de sinais eletrocardiográficos;
6. Diagnóstico médico preliminar;

7. Teoria da estimação e otimização;
8. Teoria da automação;
9. Conjuntos nebulosos (fuzzy);
10. Linguagens formais (computador);
11. Problemas de classificação.

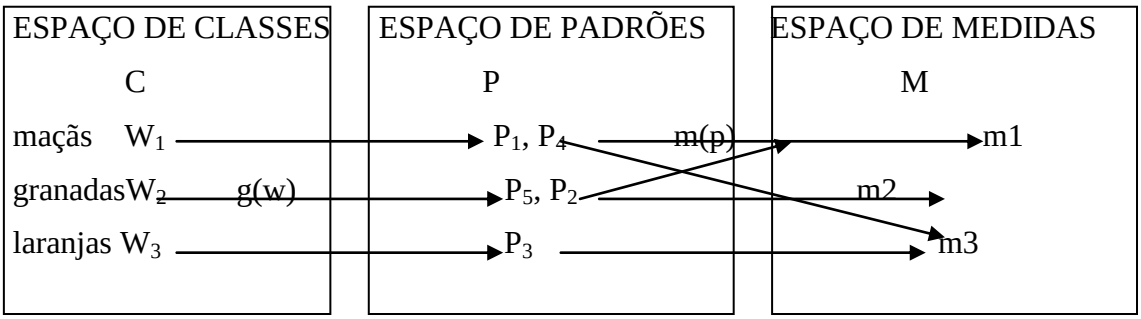
Os objetivos imediatos da técnica quando usada para DISCRIMINAR e CLASSIFICAR são, respectivamente, os seguintes:

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 R. A. Fisher [ref. 1] no primeiro tratamento moderno dos problemas de separação.

No reconhecimento de padrões o problema básico é "dado o vetor de medidas $\underline{m_i}$ deseja-se um método para inverter o mapeamento nas relações g e m, identificando a classe geradora das medidas".



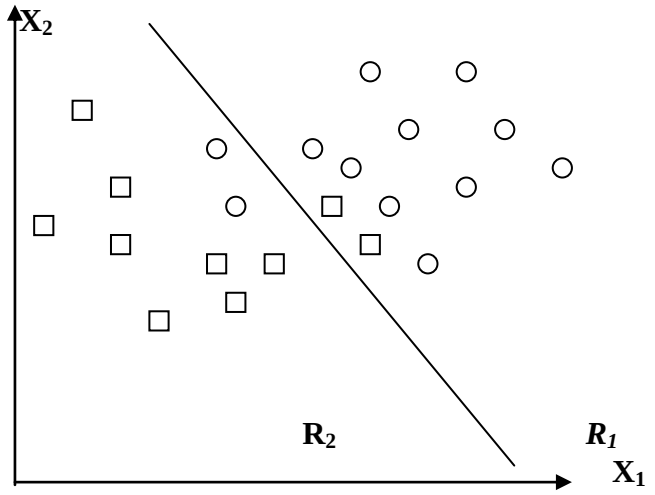
- As maças são diferentes no tamanho, peso e forma;
- O mesmo vale para as classes W_2 e W_3;
- Algumas realizações de granadas e maçãs são similares (P_1 e P_2) nos atributos;
- A distinção entre as classes será feita em função das variáveis, p.ex. o peso é fundamental para diferenciar maçã de granada de mão.

6.2 Problema geral de reconhecimento e classificação

6.2.1. Introdução

EXEMPLO

Considere dois grupos em uma cidade: proprietários de certo equipamento e não proprietários desse equipamento. A fim de identificar o melhor tipo de campanha de vendas, o fabricante do equipamento está interessado em classificar famílias como futuras compradoras do equipamento ou não, com base em X_1 = renda e X_2 = tamanho do lote de moradia. Amostras aleatórias de $n_1 = 12$ proprietários ○ e $n_2 = 12$ não proprietários □ produziram o diagrama de dispersão abaixo:

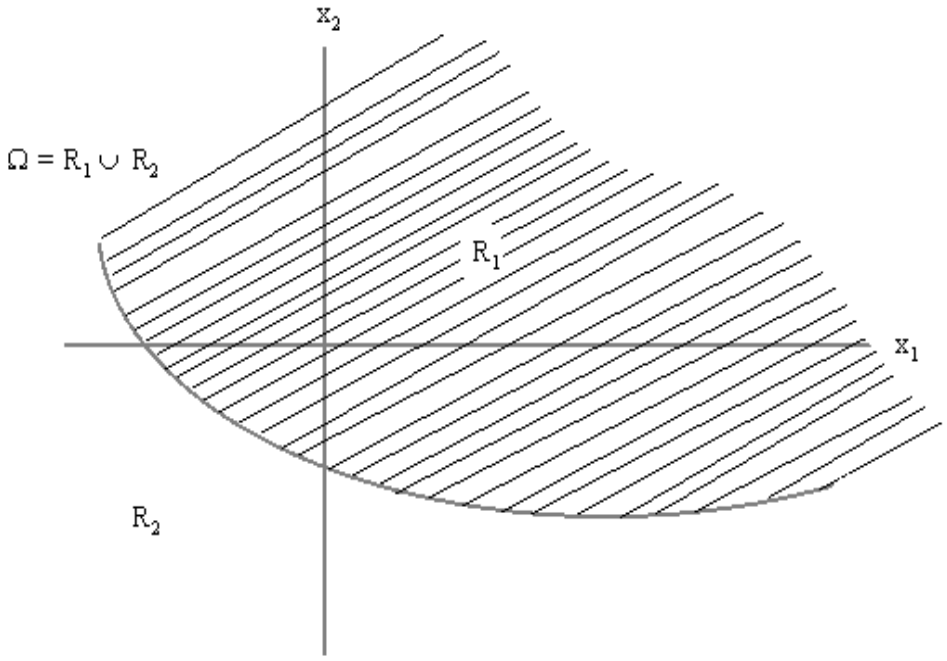


Observa-se que:

1) Proprietários tendem a ter maiores rendas e maiores lotes;
2) Renda parece discriminar melhor do que área do lote;
3) Existe mistura entre os 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 em pouca mistura de elementos grupais. Pode ocorrer que de uma classe ou população exista maior probabilidade de ocorrência do que de outra classe. Por exemplo, existe uma tendência de serem financiáveis empresas sólidas e não empresas em situação pré-falimentar. 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 pertence a Π_2 represente um erro mais sério do que classificá-lo em Π_2 quando ele pertence a Π_1. Então, deve-se levar isto em conta.

Seja $f_1(\underset{\sim}{x})$ e $f_2(\underset{\sim}{x})$ as f.d.p's associadas com o vetor aleatório $\underset{\sim}{X}$ 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 $\underset{\sim}{x}$. Seja R_1 o conjunto de valores $\underset{\sim}{x}$ para os quais nós classificamos o objeto como de Π_1 e $R_2 = \Omega - R_1$ os remanescentes valores $\underset{\sim}{x}$ para os quais nós classificamos os objetos como Π_2. Os conjuntos R_1 e R_2 são mutuamente exclusivos. Para p = 2, podemos ter a figura:

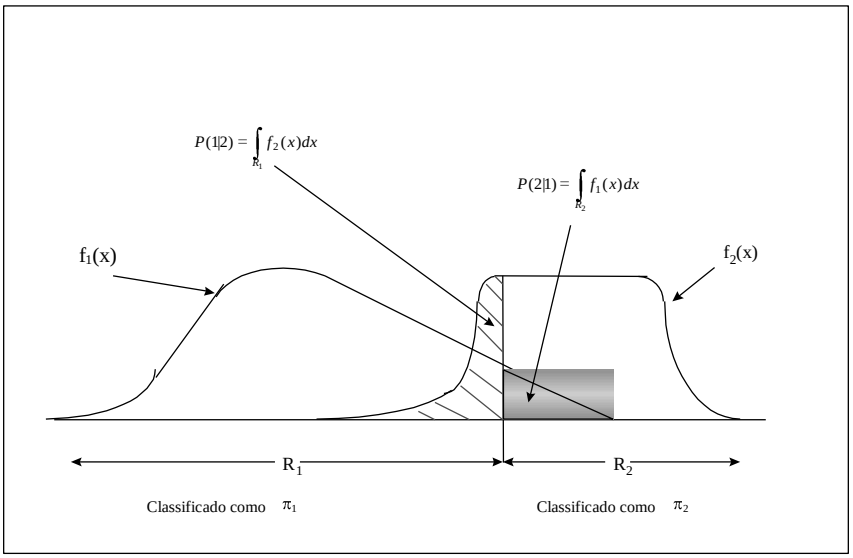


6.2.2. Regiões de classificação para duas populações

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

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

P(2|1) representa o volume formado pela f.d.p. $f_1(\underset{\sim}{x})$ na região R_2 e sendo p = 1 (caso univariado) tem-se:



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

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

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

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

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

P(rec. corr. como Π_2) = P($\underset{\sim}{X} \in \Pi_2$ e é rec. corr. como Π_2)

= P(($\underset{\sim}{X} \in R_2 | \Pi_2$)P($\Pi_2$) = P(2|2)$p_2$

P(rec. incorr. como Π_2) = P($\underset{\sim}{X} \in \Pi_1$ e é rec. incorr. 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 avaliadas em termos de suas probabilidades de reconhecimento errado. Assim, é comum construir a matriz que segue.

6.2.3. Matriz do Custo de Reconhecimento (classif.) Errado e ECM

		Reconhecido 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 <u>regra razoável</u> de reconhecimento deve ter ECM muito baixa, tanto quanto possível.

Resultado 1:
As regiões R_1 e R_2 que minimizam o ECM são definidas pelos valores de $\underset{\sim}{x}$ tal que valem as desigualdades:

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

$$\begin{bmatrix}\text{Razão das}\\ \text{densidades}\end{bmatrix} \geq \begin{bmatrix}\text{Razão dos}\\ \text{custos}\end{bmatrix} \cdot \begin{bmatrix}\text{Razão das}\\ \text{probabilidades à priori}\end{bmatrix}$$

$$R_2 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \left[\frac{c(1|2)}{c(2|1)}\right] \cdot \left[\frac{p_2}{p1}\right]$$

$$\begin{bmatrix}\text{Razão das}\\ \text{densidades}\end{bmatrix} < \begin{bmatrix}\text{Razão dos}\\ \text{custos}\end{bmatrix} \cdot \begin{bmatrix}\text{Razão das}\\ \text{probabilidades à priori}\end{bmatrix}$$

<u>Prova:</u>

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

$$ECM = c(2|1)\,p_1 \int_{R_2} f_1(\underset{\sim}{x})\,d\,\underset{\sim}{x} + c(1|2)\,p_2 \int_{R_1} f_2(\underset{\sim}{x})\,d\,\underset{\sim}{x}$$

e como $\Omega = R_1 \cup R_2$ tem-se:

$$1 = \int_{\Omega} f_1(\underset{\sim}{x})\,d\,\underset{\sim}{x} = \int_{R_1} f_1(\underset{\sim}{x})\,d\,\underset{\sim}{x} + \int_{R_2} f_1(\underset{\sim}{x})\,d\,\underset{\sim}{x}$$

podemos escrever:

$$ECM = c(2|1)\,p_1\left[1 - \int_{R_1} f_1(\underset{\sim}{x})\,d\,\underset{\sim}{x}\right] + c(1|2)\,p_2 \int_{R_1} f_2(\underset{\sim}{x})\,d\,\underset{\sim}{x}$$

e das propriedades de integral (volume).

$$ECM = \int_{R_1}[c(1|2)\,p_2\,f_2(\underset{\sim}{x}) - c(2|1)\,p_1\,f_1(\underset{\sim}{x})]\,d\,\underset{\sim}{x} + c(2|1)\,p_1$$

e como p_1, p_2, c(1|2), e c(2|1) são não-negativos, e ainda $f_1(\underset{\sim}{x})$ e $f_2(\underset{\sim}{x})$ também são não-negativos e são as únicas quantidades de ECM que dependem de $\underset{\sim}{x}$. Assim ECM é minimizado se R_1 <u>inclui esses valores</u> $\underset{\sim}{x}$ tal que $[c(1|2)\,p_2\,f_2(\underset{\sim}{x}) - c(2|1)\,p_1\,f_1(\underset{\sim}{x})] \leq 0$

e exclui aqueles para os quais esta quantidade é positiva. Isto é, R_1 deve ser o conjunto de pontos tal que:

$$c(1|2)\, p_2\, f_2(\underset{\sim}{x}) \le c(2|1)\, p_1\, f_1(\underset{\sim}{x}) \Rightarrow \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \ge \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p1}\right]$$

e dado que R_2 é o complemento de R_1 em Ω, R_2 deve ser o conjunto de pontos $\underset{\sim}{x}$ para os quais:

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p1}\right].$$

Casos Especiais de Regiões de ECM

a) Probabilidades a *priori* iguais: $\frac{p_2}{p_1} = 1$

$$R_1 = \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \ge \frac{c(1|2)}{c(2|1)}; \qquad R_2 = \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \frac{c(1|2)}{c(2|1)}$$

b) Custos de erro de reconhecimento iguais: $\frac{c(1|2)}{c(2|1)} = 1$

$$R_1 = \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \ge \frac{p_2}{p_1}; \qquad R_2 = \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \frac{p_2}{p_1}$$

c) Probabilidades a *priori* iguais e custos de reconhecimento errado iguais:

$$\frac{p_2}{p_1} = \frac{c(1|2)}{c(2|1)} = 1 \text{ ou } \frac{p_2}{p_1} = \frac{1}{c(1|2)\,/\,c(2|1)}$$

$$R_1 = \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \ge 1; \qquad R_2 = \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < 1$$

OBS.1) Quando as probabilidades a priori são desconhecidas, elas são freqüentemente tomadas como iguais e a razão de f.d.p's é comparada com a razão de custos de reconhecimento errado.

2) Se a razão de custo de reconhecimento errado é indeterminada, ela é usualmente tomada como 1 e a razão das f.d.p's é comparada com a razão de probabilidades a prior.

3) Finalmente, quando ambas, razão das probabilidades a prior e razão de custos são unitários ou uma razão é recíproca da outra, então as regiões de

reconhecimento (classificação) ótimo são determinadas comparando-se os valores das f.d.p's. Assim, se $\underset{\sim}{x}_0$ é uma nova observação e $f_1(\underset{\sim}{x}_0)/f_2(\underset{\sim}{x}_0) \geq 1 \Rightarrow f_1(\underset{\sim}{x}_0) \geq f_2(\underset{\sim}{x}_0)$], assumimos que $\underset{\sim}{x}_0 \in \Pi_1$.

EXERCÍCIO

Um pesquisador tem dados disponíveis para estimar a função densidade $f_1(\underset{\sim}{x})$ e $f_2(\underset{\sim}{x})$ associadas a Π_1 e Π_2, respectivamente. Suponha que c(2|1) = 5 unidades e c(1|2) = 10 unidades. Sabe-se, ainda, que 20% de todos os itens para os quais as medidas $\underset{\sim}{x}$ foram registradas pertencem a Π_2.

a) Escreva as probabilidades a prior das populações Π_1 e Π_2.
b) Determine as regiões de classificação (reconhecimento) R_1 e R_2.
c) Suponha que para uma nova observação $\underset{\sim}{x}_0$ tem-se: $f_1(\underset{\sim}{x}_0) = 0.3$ e $f_2(\underset{\sim}{x}_0) = 0.4$. Em qual dos grupos (populações) você classificaria a nova observação?

6.2.4. 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 <u>probabilidade total de erro de classificação</u> (TPM).

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

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

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 <u>maior probabilidade posteriori</u> $P(\Pi_i|\underset{\sim}{x}_0)$, onde

$$P(\Pi_1 | \underset{\sim}{x}_0) = \frac{P(\Pi_1 \text{ ocorre e observa-se } \underset{\sim}{x}_0)}{P(\text{observa-se } \underset{\sim}{x}_0)}$$

$$= \frac{P(\text{observa - se } \underset{\sim}{x}_0 | \Pi_1) P(\Pi_1)}{P(\text{observa - se } \underset{\sim}{x}_0 | \Pi_1) p(\Pi_1) + P(\text{observa - se } \underset{\sim}{x}_0 | \Pi_2) p(\Pi_2)}$$

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

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

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

6.2.5. Classificação com duas populações Normais Multivariadas

Assumindo-se que $f_1(x)$ e $f_2(x)$ são normais multivariadas, a primeira com vetor de médias μ_1 e matriz de covariâncias $\sum_1$ e a segunda com os parâmetros μ_2 e $\sum_2$ e então supondo $\sum_1 = \sum_2$, 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 o vetor aleatório $X' = [X_1, X_2, ..., X_p]$ para populações Π_1 e Π_2 e

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

Suponha que os parâmetros μ_1, μ_2 e $\sum$ sejam conhecidos. Então, a região de mínimo ECM,

$$R_1: \frac{f_1(x)}{f_2(x)} = \frac{\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (x - \mu_1)' \Sigma^{-1} (x - \mu_1) \right]}{\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (x - \mu_2)' \Sigma^{-1} (x - \mu_2) \right]} =$$

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

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

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

Resultado 2: Sejam as populações Π_1 e Π_2 normais multivariadas. A regra de reconhecimento que minimiza ECM é dada por: reconhecer x_0 como sendo de Π_1 se

$$\left(\mu_1 - \mu_2 \right)' \Sigma^{-1} x_0 - \frac{1}{2} \left(\mu_1 - \mu_2 \right)' \Sigma^{-1} \left(\mu_1 + \mu_2 \right) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \right]$$

e $\underline{x}_0$ é reconhecido como de II_2 em caso contrário.

EXERCÍCIO: Prove o resultado enunciado anteriormente.

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 $w = \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)' S_p^{-1} \underset{\sim}{x} - \frac{1}{2}\left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)' S_p^{-1}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2\right) = \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)' S_p^{-1}[\underset{\sim}{x} - \frac{1}{2}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2\right)]$ é conhecida como função de classificação de Anderson.

EXERCÍCIOS:

1) Comparando a expressão da regra de reconhecimento com ECM mínimo e a F.D.L. de Fisher, mostre a condição na qual a regra torna-se a F.D.L. de Fisher, se houver.

2) Seja a situação dos dados de portadores de hemofilia A e não-portadores, pg. 477 do livro do Johnson de 1988 e T11.8. Construa uma regra de reconhecimento de padrões, quando as probabilidades a prior de uma pessoa pertencer a cada um dos grupos são conhecidas e assumindo irrealisticamente que os custos de reconhecimento errado sejam iguais $c(1|2) = c(2|1)$. Suponha que o sangue é extraído de uma pessoa primo em primeiro grau de um hemofílico e os resultados são $x_1 = \log_{10}$(AHF-atividade) = -0.210 e $x_2 = \log_{10}$(antígeno AHF)=-0.044. Contudo, a chance genética de se ter uma portadora e uma prima em primeiro grau portadora seja 0.25. Faça um diagrama que mostre os pontos no sistema de eixos x_1 e x_2 e obtenha a regra de reconhecimento, além de testar a Gaussianidade.

6.2.6. Classificação Quadrática, $\sum_1 \neq \sum_2$

Supondo as matrizes de covariância $\sum_1$ para $\underset{\sim}{x} \in \Pi_1$ e Σ_2 para $\underset{\sim}{x} \in \Pi_2$ com $\sum_1 \neq \sum_2$, 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 $\underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$ e $\underset{\sim}{\Sigma}_1 \neq \Sigma_2$. A probabilidade total de reconhecimento errado (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

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

Resultado 3:
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 e classificação que minimiza o ECM é dado por:

$$R_1 : -\frac{1}{2}\underset{\sim}{x}_0'\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\underset{\sim}{x}_0 + \left(\underset{\sim}{\mu}_1'\Sigma_1^{-1} - \underset{\sim}{\mu}_2'\Sigma_2^{-1}\right)\underset{\sim}{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

R_2 em c/c.

EXERCÍCIO

Prove o resultado anterior.

Na prática, a regra de reconhecimento estabelecida é aplicada substituindo-se os parâmetros $\underline{\mu}_1$, $\underline{\mu}_2$, Σ_1 e Σ_2 pelas suas estimativas $\underline{\bar{x}}_1$, $\underline{\bar{x}}_2$, S_1 e S_2, tal que:

Aloca-se $\underline{x}_0$ em Π_1 se:

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

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

6.3- Discriminação e Classificação entre Populações: Método de Fisher

6.3.1- Função Discriminante Linear de Fisher Para duas Populações

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 :

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

onde as variáveis X_i , i = 1 , 2 , ... , p , são as medidas das características investigadas nos objetos. Os valores observados de $\underline{X}$ podem diferir de uma classe para outra, sendo que a totalidade dos valores da 1ª. classe é a população dos valores $\underline{X}$ para π_1 e aqueles da 2ª. classe são a população dos valores de $\underline{X}$ para π_2.

Assim, estas populações podem ser descritas pelas funções densidade de probalidade $f_1(\underline{x})$ e $f_2(\underline{x})$.

TABELA I - SITUAÇÕES EXEMPLOS

Populações π_1 e π_2	Variáveis medidas (componentes de $\underline{X}$)
1. Sucesso ou insucesso de estudantes na Universidade.	- Nota 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.

Fonte: o autor.

A idéia de Fisher foi transformar as observações multivariadas $\underline{X}$'s nas observações univariadas y's tal que os y's das populações π_1 e π_2 sejam separados tanto quanto possível. Fisher teve a idéia de tomar combinações lineares de $\underline{X}$ para criar os y's, dado que as combinações lineares são funções de $\underline{X}$ e por outro lado são de fácil cálculo matemático.
Seja μ_{1y} a média dos y's obtidos dos $\underline{x}$'s pertencentes a π_1 e μ_{2y} a média dos y's obtidos dos $\underline{x}$'s pertencentes a π_2 , então Fisher selecionou a combinação linear que maximiza a distância quadrática entre μ_{1y} e μ_{2y} relativamente a variabilidade dos y's. Assim, seja:

$\underline{\mu}_1 = E (\underline{X} |\pi_1)$ = valor esperado de uma observação multivariada de π_1.
$\underline{\mu}_2 = E (\underline{X} |\pi_2)$ = valor esperado de uma observação multivariada de π_2.

e supondo a matriz de covariância

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

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

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

tem-se

$$\mu_{1y} = E (Y |\pi_1) = E (\underline{c}'\underline{X} |\pi_1) = \underline{c}'E (\underline{X} |\pi_1) = \underline{c}'\underline{\mu}_1,$$
$$\mu_{2y} = E (Y |\pi_2) = E (\underline{c}'\underline{X} |\pi_2) = \underline{c}'E (\underline{X} |\pi_2) = \underline{c}'\underline{\mu}_2,$$

e

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

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

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

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

RESULTADO 6.3.1:

Seja $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$ e $Y = \underline{c}'\underline{X}$, então $\frac{(\underline{c}'\underline{\delta})^2}{\underline{c}'\sum\underline{c}}$ é maximizada por

$\underline{c} = k\,\Sigma^{-1}\underline{\delta} = k\,\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$ para qualquer $k \neq 0$

Escolhendo k = 1 tem-se $\underline{c} = \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$ e $Y = \underline{c}'\underline{X} = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{X}$,

que é conhecida como ***FUNÇÃO DISCRIMINANTE LINEAR DE FISHER***.

EXERCÍCIO 1

Prove o resultado anterior.

A FUNÇÃO DISCRIMINANTE LINEAR DE FISHER transforma as populações multivariadas π_1 e π_2 em populações univariadas, tais que as médias das populações univariadas correspondentes são separadas tanto quanto possível relativamente a variância populacional, considerada comum. Assim tomando-se

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

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

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

como

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

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

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

e tem-se que:

$$E(Y_0|\pi_1) - m \geq 0$$
$$E(Y_0|\pi_2) - m < 0 \,,$$

ou seja, se $\underline{X}_0$ pertence a π_1, se espera que Y_0 seja igual ou maior do que o ponto médio. Por outro lado se $\underline{X}_0$ pertence a π_2 , o valor esperado de Y_0 será menor que o ponto médio. Desta forma a regra de classificação é :

- alocar $\underline{x}_0$ em π_1 se $y_0 - m \geq 0$
- alocar $\underline{x}_0$ em π_2 se $y_0 - m < 0$

Geralmente, os parâmetros $\underline{\mu}_1$, $\underline{\mu}_2$ e Σ são desconhecidos, então supondo que se tem n_1 observações da v.a. multivariada

$$\underline{X}_1' = [\underline{X}_{11} \vdots \underline{X}_{21} \vdots \ldots \vdots \underline{X}_{p1}]$$

da população π_1 e n_2 observações da v.a. multivariada

$$\underline{X}_2' = [\underline{X}_{12} \vdots \underline{X}_{22} \vdots \ldots \vdots \underline{X}_{p2}]$$

da população π_2 , então os resultados amostrais para aquelas quantidades são :

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

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

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 FISHER AMOSTRAL é dada por:

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

E, a estimativa do ponto médio entre as duas médias amostrais univariadas $\bar{y}_1 = \underline{\hat{c}}'\underline{\bar{x}}_1$ e $\bar{y}_2 = \underline{\hat{c}}'\underline{\bar{x}}_2$ é dado por :

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

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

Finalmente a regra de classificação é a seguinte :

- alocar $\underline{x}_0$ em π_1 se $y_0 = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)' S_p^{-1} \underline{x}_0 \geq \hat{m}$

- alocar $\underline{x}_0$ em π_2 se $y_0 = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)' S_p^{-1} \underline{x}_0 < \hat{m}$

ou melhor se:

$y_0 - \hat{m} \geq 0$ x_0 é alocado em π_1
$y_0 - \hat{m} < 0$ x_0 é alocado em π_2

RESULTADO 6.3.2 :

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

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

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

EXERCÍCIO 2

Seja um estudo onde se pretende detectar portadores de Hemophilia A. A fim de se construir um procedimento para detectar portadores potenciais de Hemophilia A, amostras de sangue são analisadas em dois grupos de mulheres e medidas as variáveis: $X_1 = \log_{10}$(atividade AHF) e $X_2 = \log_{10}$(antígeno AHF). Os dados estão na tabela T11.8. O primeiro grupo é composto por $n_1 = 30$ mulheres selecionadas de uma população de mulheres que não possui portadora do gene da hemofilia. Este grupo é o grupo normal. O segundo grupo de $n_2 = 22$ mulheres foi selecionado de uma população de conhecidas portadoras da deficiência (irmãs de hemofílicos, mães com mais de um filho hemofílico e mães com um filho hemofílico e outro hemofílico relativo). Este é o grupo dos portadores obrigatórios. Os resultados dos dados são:

$$\bar{\underline{x}}_1 = \begin{bmatrix} -0.0065 \\ -0.039 \end{bmatrix} \quad \bar{\underline{x}}_2 = \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} \quad S_p^{-1} = \begin{bmatrix} 131{,}158 & -90{,}423 \\ -90{,}423 & 108{,}147 \end{bmatrix}$$

a) Calcule a F.D.L. de Fisher;
b) Calcule o ponto médio das duas médias amostrais;
c) Sejam as medidas $x_1 = -0{,}210$ e $x_2 = -0{,}044$, verifique em qual das populações é alocado o indivíduo com estas medidas (normal ou portadora).

EXERCÍCIO 3

A comissão de admissão de uma escola deseja classificar um pretendente em uma de duas populações: população dos estudantes que completam com sucesso o curso ou população dos estudantes que não completam o curso. A comissão considera a nota dos pretendentes em p = 2 testes. Seja $\underline{x}' = [x_1\ x_2]$ o vetor das notas. De experiências anteriores é sabido que : $\underline{\mu}_1' = [60\ 57]$, $\underline{\mu}_2' = [42\ 39]$ e $\Sigma = \begin{bmatrix} 100 & 70 \\ 70 & 100 \end{bmatrix}$

a) Encontre a F.D.L. de Fisher;
b) Calcule $E(Y|\pi_1) = \mu_{1y}$, $E(Y|\pi_2) = \mu_{2y}$ e $V(Y) = \sigma_y^2$,
c) Determine o ponto médio entre as médias populacionais univariadas,
d) Daça um gráfico das duas populações, admitindo que $\underline{X} \sim N_p(\underline{\mu}_1,\Sigma)$ ou $\underline{X} \sim N_p(\underline{\mu}_2,\Sigma)$
e) Determine a probabilidade de classificação errônea,
f) Escreva o critério de classificação,
g) Verifique em qual população será alocado o ponto $\underline{x}' = [52\ 45]$

6.3.2- Discriminação entre Diversas Populações

O método de discriminação exposto no item 6.3.1 pode ser estendido para diversas populações. O primeiro objetivo de Fisher com a Análise Discriminante foi o de separar populações. Ele pode, contudo, ser usado também para classificar. Este método não necessita da suposição de que as diversas populações sejam normais multivariadas. Entretanto, assume-se que as matrizes de covariâncias populacionais Σ's são iguais e com posto completo, isto é, $\Sigma_1 = \Sigma_2 = ... = \Sigma_g = \Sigma$. Assim, seja $\underline{\bar{\mu}}$ o vetor médio dos diversos grupos (populações) $\underline{\bar{\mu}} = \frac{1}{g}\sum_{i=1}^{g} \underline{\mu}_i$ e B_0 a matriz "Soma de produtos cruzados entre grupos populacionais" tal que

$$B_0 = \sum_{i=1}^{g} (\underline{\mu}_i - \underline{\bar{\mu}})(\underline{\mu}_i - \underline{\bar{\mu}})'$$

A combinação linear $Y = \underline{c}'\underline{x}$ tem por esperança $E(Y) = \underline{c}'E(\underline{X}|\pi_i) = \underline{c}'\underline{\mu}_i$ para a população π_i e variância

$$V(Y) = \sigma_y^2 = \underline{c}'V(\underline{X})\underline{c} = \underline{c}'\Sigma\underline{c}$$

para todas as populações. Desta forma, o valor esperado $\mu_{iy} = \underline{c}'\underline{\mu}_i$ muda quando a população da qual $\underline{X}$ é selecionado é outra. Vamos então definir a média global

$$\bar{\mu}_y = \frac{1}{g}\sum_{i=1}^{g} \mu_{iy} = \underline{c}'\underline{\bar{\mu}}$$

e a razão entre a "soma dos quadrados das distâncias das populações para a média global e a variância de Y" é:

$$\frac{\underline{c}'B_0\underline{c}}{\underline{c}'\Sigma\underline{c}}$$

que é uma generalização multigrupal do caso de duas populações. Esta razão mede a variabilidade entre grupos de valores (escores) Y relativamente a variabilidade comum dentro dos grupos. Da mesma forma que no caso de duas populações, nós podemos selecionar $\underline{c}$ que maximiza esta razão, e é conveniente normalizar $\underline{c}$ tal que $\underline{c}'\Sigma\underline{c} = 1$.

RESULTADO 6.3.3

Seja $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_s > 0$ os s = min(g-1,p) autovalores não-nulos de $\Sigma^{-1}B_0$ e $\underline{e}_1, \underline{e}_2, ... , \underline{e}_s$ os correspondentes autovetores (escalonados tal que $\underline{e}'\Sigma\underline{e} = 1$). Então o vetor de coeficientes $\underline{c}$ que maximiza a razão $\frac{\underline{c}' B_0 \underline{c}}{\underline{c}'\Sigma\underline{c}}$ é dada por $\underline{c}_1 = \underline{e}_1$. A combinação linear $\underline{c}_1'\underline{X}$ é chamada 1^o. discriminante e da mesma forma podemos generalizar para o k-ésimo discriminante com $\underline{c}_k = \underline{e}_k$ k = 1,2, ... s.

Como, geralmente, Σ e $\underline{\mu}_i$ não são disponíveis, nós tomamos amostras aleatórias de tamanhos n_i das populações π_i, i = 1,2, ..., g e denotando o conjunto de dados (a.a) da população π_i, i = 1,2, ..., g, por $_{ni}X_p$ temos na j-ésima linha o vetor $\underline{x}_{ij}$ e os estimadores dos parâmetros $\underline{\mu}_i$ e $\overline{\underline{\mu}}$ são

$$\overline{\underline{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \underline{x}_{ij}$$

$$\overline{\underline{x}} = \frac{\sum_{i=1}^{g} n_i \overline{\underline{x}}_i}{\sum_{i=1}^{g} n_i} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i} \underline{x}_{ij}}{\sum_{i=1}^{g} n_i}$$

A matriz "Soma de produtos cruzados entre grupos", B_0, é estimada por

$$\hat{B}_0 = \sum_{i=1}^{g} (\overline{\underline{x}}_i - \overline{\underline{x}})(\overline{\underline{x}}_i - \overline{\underline{x}})'$$

e um estimador para Σ pode ser conseguido com base na matriz W

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

Conseqüentemente, $\frac{W}{n_1 + n_2 + ... + n_g - g} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + ... + (n_g - 1)S_g}{n_1 + n_2 + ... + n_g - g} = S_p$

É claro que o mesmo $\hat{\underline{c}}$ que maximiza a razão $\frac{\hat{\underline{c}}'\hat{B}_0\hat{\underline{c}}}{\hat{\underline{c}}'S_p\hat{\underline{c}}}$ também maximiza $\frac{\hat{\underline{c}}'\hat{B}_0\hat{\underline{c}}}{\hat{\underline{c}}'W\hat{\underline{c}}}$.

Assim, apresentaremos o otimizante $\hat{\underline{c}}$ na forma mais usual, que é o autovetor $\hat{\underline{e}}_i$ da matriz $W^{-1}B_0$, porque se $W^{-1}B_0\hat{\underline{e}} = \hat{\lambda}\hat{\underline{e}}$ então $S_p^{-1}\hat{B}_0\hat{\underline{e}} = \hat{\lambda}(n_1 + n_2 + ... + n_g - g)\hat{\underline{e}}$, portanto, concluindo sejam $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_g > 0$ os autovalores não nulos de $W^{-1}B_0$ e $\hat{\underline{e}}_1, \hat{\underline{e}}_2, ..., \hat{\underline{e}}_s$ os correspondentes autovetores, sendo s = min(g-1, p) e $\hat{\underline{e}}_i$ normalizado tal que $\hat{\underline{e}}_i S_p \hat{\underline{e}}_i = 1$; então o vetor de coeficientes que maximiza a razão citada acima é $\hat{\underline{c}}_1 = \hat{\underline{e}}_1$ e a combinação linear $\hat{\underline{e}}_1'\underline{x}$ é chamada 1^o. discriminante amostral. Generalizando, teremos no passo k o k-ésimo discriminante amostral $\hat{\underline{e}}_k'\underline{x}$, $k \leq s$.

EXERCÍCIO 4

Dadas as observações de p = 2 variáveis oriundas de 3 grupos populacionais π_1, π_2 e π_3

$$X_1 = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{bmatrix} \quad X_2 = \begin{bmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix} \quad X_3 = \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}$$

a) Determine os vetores médios amostrais $\underline{\bar{x}}_i$ i = 1, 2, 3.
b) Determine o vetor médio global, amostral $\underline{\bar{x}}$.
c) Dadas as matrizes de covariância amostral $S_1 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$, $S_2 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$ e $S_3 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$ determine as matrizes S_p, $\hat{B}_0$ e W.
d) Determine a matriz inversa W^{-1} e $W^{-1}\hat{B}_0$.
e) Determine os autovalores e autovetores de $W^{-1}\hat{B}_0$.
f) Quais são os discriminantes para os três grupos.
g) Faça um gráfico que represente o "espaço discriminante" nas duas dimensões que você encontrou, represente a amostra no gráfico.
h) Em qual grupo seria classificado o indivíduo $\underline{x}_0' = [1, 3]$.

6.4 Avaliação de funções de reconhecimento e classificação

6.4.1. Critério TPM

Uma maneira de julgar o desempenho de um procedimento de reconhecimento de padrões é calcular a sua taxa de erro no reconhecimento dos padrões. A probabilidade total de erro de reconhecimento, TPM, é dada por:

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

e o menor valor para esta quantidade, obtido pela escolha adequada de R_1 e R_2, é chamado de taxa ótima de erro (OER),

$$\text{OER} = p_1 \int_{R_2} f_1(\underset{\sim}{x}) d\underset{\sim}{x} - p_2 \int_{R_1} f_2(\underset{\sim}{x}) d\underset{\sim}{x}$$

com R_1 e R_2 determinados por $R_1: \dfrac{f_1(\underset{\sim}{x})}{f_1(\underset{\sim}{x})} \geq \dfrac{p_2}{p_1}$ e R_2 em caso contrário.

EXERCÍCIO

Determine a expressão para taxa ótima de erro quando $p_1 = p_2 = 1/2$ e $f_1(\underset{\sim}{x})$ e $f_2(\underset{\sim}{x})$ são densidades normais multivariadas.

Solução:
Sabemos que o ECM e TPM, mínimos, coincidem quando $c(1|2) = c(2|1)$ e devido as probabilidades a priori serem também iguais tem-se:

$$\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1 \Rightarrow \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] = 0 \text{ e}$$

$$R_1: \left[\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\left(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2\right)\right] \geq 0$$

$$R_2: \left[\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\left(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2\right)\right] < 0$$

e fazendo $y = \left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\underset{\sim}{x} = \underset{\sim}{\ell}'\underset{\sim}{x}$ fica:

$$R_1(y): y \geq \frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\left(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2\right)$$

$$R_2(y): y < \frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\left(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2\right)$$

Mas como y é uma combinação linear de v.a's Gaussianas, as f.d.p's de Y, $f_1(y)$ e $f_2(y)$, são Gaussianas univariadas com parâmetros:

$$\mu_{1y} = \underset{\sim}{\ell}'\underset{\sim}{\mu}_1 = \left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\underset{\sim}{\mu}_1$$

$$\mu_{2y} = \underset{\sim}{\ell}'\underset{\sim}{\mu}_2 = \left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\underset{\sim}{\mu}_2$$

$$\sigma_y^2 = V(y) = V(\underset{\sim}{\ell}'\underset{\sim}{x}) = \underset{\sim}{\ell}'y(x)\ell = \underset{\sim}{\ell}'\Sigma\ell = \left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)'\Sigma^{-1}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)$$

e $TPM = \frac{1}{2}\int_{R_2} f_1(\underset{\sim}{x})d\underset{\sim}{x} + \int_{R_1} f_2(\underset{\sim}{x})d\underset{\sim}{x}$

$$= \frac{1}{2} P\left[\underset{\sim}{x} \in R_2 | R_1 \right] + \frac{1}{2} P\left[\underset{\sim}{x} \in R_1 | R_2 \right]$$

e $P\left[\underset{\sim}{x} \in R_2 | R_1 \right] = P(2|1) = P\left[y < \frac{1}{2} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2) \right]$

$$= P\left(\frac{y - \mu_{iy}}{\sigma_y} < \frac{\frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \right)' \Sigma^{-1} \left(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2 \right) - \left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \right)' \Sigma^{-1} \underset{\sim}{\mu}_1}{\sigma_y} \right)$$

$$= P\left(Z < \frac{\frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \right)' \Sigma^{-1} \underset{\sim}{\mu}_2 - \frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \right)' \Sigma^{-1} \underset{\sim}{\mu}_1}{\sigma_y} \right)$$

$$= P\left(Z < \frac{-\frac{1}{2}\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \right)' \Sigma^{-1} \left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \right)}{\sigma_y} \right)$$

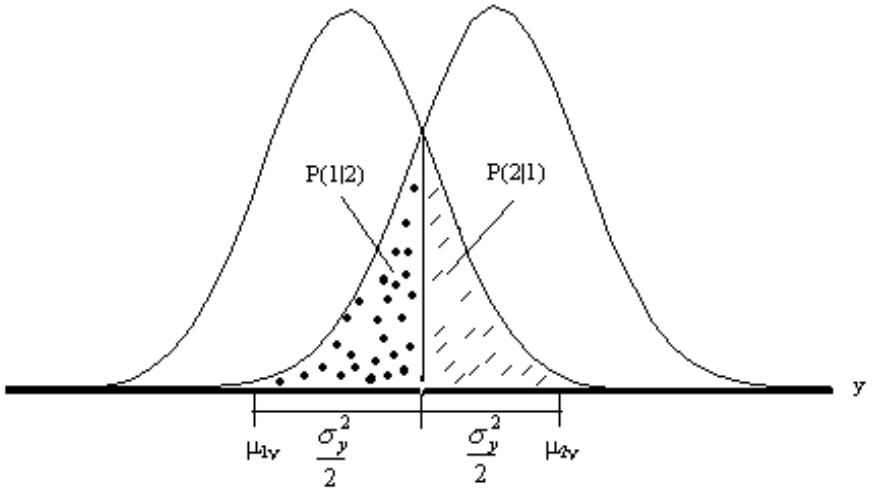
$$P\left[\underset{\sim}{x} \in R_2 \mid R_1 \right] = P(2|1) = P\left[Z < \frac{-\sigma_y^2}{2\sigma_y} \right] = P\left[Z < \frac{-\sigma_y}{2} \right] = \Phi\left(-\frac{\sigma_y}{2} \right)$$

Da mesma forma

$$P\left[\underset{\sim}{x} \in R_1 | R_2 \right] = P(1|2) = P\left[Z \geq \frac{\sigma_y}{2} \right] = 1 - P\left[Z \geq \frac{\sigma_y}{2} \right] = \Phi\left(-\frac{\sigma_y}{2} \right)$$

Logo a TAXA ÓTIMA DE ERRO é:

$$OER = mínimo TPM = \frac{1}{2}\Phi\left(\frac{-\sigma_y}{2} \right) + \frac{1}{2}\Phi\left(\frac{-\sigma_y}{2} \right) = \Phi\left(\frac{-\sigma_y}{2} \right)$$



Supondo, $\sigma_y^2 = (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2}) = 2.56$, tem-se:

$\sigma_y = 1.6$

$\sigma_y /2 = -1.6/2 = 0.8$

ϕ (-0.8) = 0.2119

Assim, a regra ótima de reconhecimento cometerá erros em torno de 21.19%.

Uma medida do desempenho que não depende da forma da distribuição e que pode ser calculada para qualquer procedimento de classificação é a taxa aparente de erro que é definida como a fração das observações no treinamento amostral correspondente a reconhecimento equivocado pela função. A taxa aparente de erro é calculada da matriz de confusão que mostra a situação real das observações nos grupos *versus* o reconhecimento.

Para n_1 observações de Π_1 e n_2 de Π_2 , a matriz de confusão tem a forma:

	PREDITO		
	Π_1	Π_2	
Π_1	n_{1C}	n_{1M}	n_1
ATUAL			
Π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

onde:
n_{1C}: número de itens de Π_1 corretamente reconhecido como de Π_1
n_{1M}: número de itens Π_1 misturados com de Π_2
n_{2C}: número de itens Π_2 corretamente reconhecido como de Π_2
n_{2M}: número de itens Π_2 misturados com de Π_1
A taxa aparente de erro (APER) é dada por:

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

e é entendida como a proporção de itens no conjunto de treinamento que são reconhecidos erroneamente.

6.4.2. Abordagem de Lachenbruch

Uma abordagem que costuma funcionar bem, neste caso, é a de Lachenbruch cujo procedimento operacional é o seguinte:

1. Comece com o grupo da população Π_1. Omita uma observação deste grupo e construa uma função baseada nas n_1-1 e n_2 observações.
2. Reconheça (classifique), usando a função, a observação não incorporada.
3. Repita os passos 1 e 2 até que todas as n_1 observações de Π_1 sejam classificadas. Seja $n_{1M}^{(H)}$ o número de observações reconhecidas erroneamente neste grupo.
4. Repita os passos de 1 a 3 para as n_2 observações de Π_2. Seja $n_{2M}^{(H)}$ o número de observações reconhecidas erroneamente neste grupo.

Então

$$\hat{P}(1) = \frac{n_{1M}^{(H)}}{n_1} \qquad \hat{P}(2) = \frac{n_{2M}^{(H)}}{n_2}$$

e a proporção total esperada de erro é $\hat{E}(\text{AER}) = \dfrac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}$

EXEMPLO
Sejam as matrizes de dados e as estatísticas de v.a bivariada:

$$X_1 = \begin{bmatrix} 2 & 4 & 3 \\ 12 & 10 & 8 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 5 & 3 & 4 \\ 7 & 9 & 5 \end{bmatrix} \qquad \underset{\sim}{\bar{x}}_1 = \begin{bmatrix} 3 \\ 10 \end{bmatrix} \quad \underset{\sim}{\bar{x}}_2 = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} \qquad S_P = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

Para a matriz de confusão de uma função de reconhecimento FDL de Fisher.

	Reconhecida como:	
Pop. verdadeira	Π_1	Π_2
Π_1	2	1
Π_2	1	2

A APER = (1+1)/(3+3) = 2/6 = 0.33 = 33%

Retirando a primeira observação de $x_H^{'}$ = [2 12] de x_1

$$\bar{x}_{1H} = \begin{bmatrix} 3.5 \\ 9 \end{bmatrix} \quad e \quad S_{1H} = \begin{bmatrix} 0.5 & 1 \\ 1 & 2 \end{bmatrix}$$

$$S_{PH} = \frac{1}{3}\left[S_{1H} + 2S_2\right] = \frac{1}{3}\begin{bmatrix} 2.5 & -1 \\ -1 & 10 \end{bmatrix}$$

$$S_{HP}^{-1} = \frac{1}{8}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}$$

Aplicando Lachenbruch teremos:

$$\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} = \frac{2+1}{3+3} = 0.5$$

Logo APER=0.33 parece otimista, contudo o tamanho da amostra deve ser levado em conta.

6.5. Reconhecimento de padrões envolvendo várias populações (grupos)

6.5.1. Introdução

Na teoria, a generalização de procedimentos estatísticos de reconhecimento de padrões e classificação de 2 para g > 2 grupos é direta. Contudo, não se conhece muito sobre as propriedades das funções amostrais e em particular, suas taxas de erro não foram totalmente investigadas.

A robustez da função de reconhecimento de padrões linear para 2 grupos, por exemplo, em matrizes de covariâncias diferentes e distribuições não-normais pode ser estudada através de experimentos gerados por meio de computador. Para mais que duas populações esta abordagem não leva a conclusões gerais, porque as propriedades dependem de onde os grupos estão localizados e existem muitas configurações para serem estudadas convenientemente.

6.5.2 Método do Mínimo Custo Esperado de Mistura

Seja $f_i(x)$ a f.d.p. associada com a população Π_i i = 1,2,3,...,g e seja p_i a probabilidade *a priori* da população Π_i i=2,3,...,g, e c(k | i) = o custo de reconhecer um item como de Π_k quando na realidade ele pertence a Π_i, k,i = 1,2,...,g.
Para k = i, c(i | i) = 0 e finalmente, seja R_k o conjunto dos $x's$ classificados como Π_k e

P(k|i) = P (classificar $\underset{\sim}{x}$ em $\Pi_k | \Pi_i$) = $\int_{R_k} f_i(\underset{\sim}{x}) d\underset{\sim}{x}$

para k,i = 1,2,...,g com P(i | i) = 1- $\sum_{k \neq 1} P($.

O custo esperado de misturar (classificar errado) $\underset{\sim}{x}$ de Π_1 em Π_2, ou Π_3, Π_4, ..., Π_g é dado por:

ECM(1) = P(2 | 1) c (2 | 1) + P(3 | 1) c (3 | 1) + ...+ P(g | 1) c (g | 1) =

$$= \sum_{k=2}^{g} P(k \mid 1)\, c\,(k \mid 1)$$

O custo esperado condicional de classificação errada ocorre com probabilidade p_1, a probabilidade de Π_1. Pode-se obter o custo esperado condicional de classificação errada, ECM(2), ECM(3), ... , ECM(g). E multiplicando cada ECM condicional por sua probabilidade a prior e somando resulta o ECM.

ECM = p_1ECM(1) + p_2 ECM(2) + ... + p_g ECM(g)

$$\text{ECM}=p_1\left(\sum_{k=2}^{g} P(k|1)c(k|1)\right) + p_2\left(\sum_{k\neq 2}^{g} P(k|21)c(k|2)\right)+...+ p_g\left(\sum_{k=1}^{g-1} P(k|g)c(k|g)\right)$$

$$\text{ECM}=\sum_{i=1}^{g} p_i\left(\sum_{k\neq i}^{g} P(k|i)c(k|i)\right)$$

Então o problema de construir um procedimento ótimo de classificação consiste em se escolher as regiões mutuamente exclusivas R_1, R_2, ..., R_g tal que o ECM seja mínimo.

RESULTADO 6.5.1:

As regiões que minimizam o ECM= $\sum_{i=1}^{g} p_i\left(\sum_{\substack{k=1 \\ k\neq i}}^{g} P(k|i)c(k|i)\right)$, são definidas pelo reconhecimento de $\underset{\sim}{x}$ como sendo de Π_k , k=1, 2, ..., g tal que $\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(\underset{\sim}{x})c(k|i)$ seja o menor possível. Se ocorrer um empate, $\underset{\sim}{x}$ será reconhecido para qualquer dos grupos empatados.

Prova

Suponha que os custos de classificação errada sejam iguais (sem perda de generalidade, nós podemos estabelecer iguais a 1) e usando o argumento anterior podemos alocar $\underset{\sim}{x}$ em Π_k , k=1, 2, ..., g, tal que $\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(\underset{\sim}{x})$ seja o menor possível.

Mas $\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(\underset{\sim}{x})$ será baixo quando $p_k f_k(\underset{\sim}{x})$,for grande. Conseqüentemente, quando os custos forem os mesmos, a regra do mínimo custo esperado terá a forma simples:

6.5.3. Regra do mínimo ECM em custos iguais de reconhecimento errado

Reconhecer $\underset{\sim}{x}$ como de Π_k se: $p_k f_k(\underset{\sim}{x}) > p_i f_i(\underset{\sim}{x}), \forall\ i \neq k$

ou, equivalentemente $\underset{\sim}{x}$ como de Π_k se: $\ln[p_k f_k(\underset{\sim}{x})] > \ln[p_i f_i(\underset{\sim}{x})], \forall\ i \neq k.$

É interessante notar que a regra de classificação anterior é idêntica aquela de maximizar a probabilidade a posteriori, $P(\Pi_k|\underset{\sim}{x})$

$$P(\Pi_k \mid \underset{\sim}{x}) = \frac{p_k f_k(\underset{\sim}{x})}{\sum_{i=1}^{g} p_i f_i(\underset{\sim}{x})} = \frac{(priori)x(verossimilhanca)}{\sum(priori)x(verossimilhanca)} \qquad k=1, 2, ..., g$$

Devemos ter em mente que, em geral, a regra de mínimo ECM tem três componentes: probabilidades a prior, custos de reconhecimento errado e funções densidade. Estes componentes devem ser conhecidos (ou estimados) antes da regra ser implementada.

EXEMPLO

Seja reconhecer $\underset{\sim}{x_0}$ como de um dos 3 grupos (g = 3) populações Π_1, Π_2 ou Π_3 , dado as probabilidades a prior (hipotéticas), custos de reconhecimento errado, e f.d.p., usando o procedimento do mínimo ECM.

		VERDADEIRA POPULAÇÃO		
		Π_1	Π_2	Π_3
CLASSIFICAR EM:	Π_1	$c(1\mid 1) = 0$	$c(1\mid 2) = 500$	$c(1\mid 3) = 100$
	Π_2	$c(2\mid 1) = 10$	$c(2\mid 2) = 0$	$c(2\mid 3) = 50$
	Π_3	$c(3\mid 1) = 50$	$c(3\mid 2) = 200$	$c(3\mid 3) = 0$
Probabilidade a prior:		$p_1 = 0.05$	$p_2 = 0.60$	$p_3 = 0.35$
densidades em $\underset{\sim}{x_0}$:		$f_1(\underset{\sim}{x_0}) = 0.01$	$f_2(\underset{\sim}{x_0}) = 0.85$	$f_3(\underset{\sim}{x_0}) = 2$

Os valores de $\sum_{\substack{i=1 \\ i\neq k}}^{3} p_i f_i(\underset{\sim}{x_0}) c(k|i)$ são:

k = 1: $p_2 f_2(\underset{\sim}{x_0})c(1\mid 2) + p_3 f_3(\underset{\sim}{x_0})c(1\mid 3) =$

$= (0.60)(0.85)(500) + (0.35)(2)(100) = 325$

k =2: $p_1 f_1(\underset{\sim}{x}_0)c(2|1) + p_3 f_3(\underset{\sim}{x}_0)c(2|3) =$

$= (0.05)(0.01)(10) + (0.35)(2)(50) = 35.055$

k = 3: $p_1 f_1(\underset{\sim}{x}_0)c(3|1) + p_2 f_2(\underset{\sim}{x}_0)c(3|2) =$

$= (0.05)(0.01)(50) + (0.60)(0.85)(200) = 102.025$

Desde que $\sum_{\substack{i=1 \\ i\neq 1}}^{3} p_i f_i(\underset{\sim}{x}_0)c(k|i)$ é menor para k = 2, nós reconhecemos $\underset{\sim}{x}_0$ em Π_2.

Se todos os custos forem iguais, nós poderíamos classificar $\underset{\sim}{x}_0$ da seguinte forma:

$p_1 f_1(\underset{\sim}{x}_0) = (0.05)(0.01) = 0.0005$

$p_2 f_2(\underset{\sim}{x}_0) = (0.60)(0.85) = 0.510$

$p_3 f_3(\underset{\sim}{x}_0) = (0.35)(2) = 0.700$

e desde que $p_3 f_3(\underset{\sim}{x}_0) = 0.700 \geq p_i f_i(\underset{\sim}{x}_0)$, i =1,2 nós alocaremos $\underset{\sim}{x}_0$ para Π_3.

Equivalentemente, calculando as probabilidades a posteriori,

$$P(\Pi_1|\underset{\sim}{x}_0) = \frac{p_1 f_1(\underset{\sim}{x}_0)}{\sum_{i=1}^{3} p_i f_i(\underset{\sim}{x}_0)} = \frac{(0.05)(0.01)}{(0.05)(0.01)+(0.60)(0.85)-(0.35)(2)} = \frac{0.0005}{1.2105} = 0.0004$$

$$P(\Pi_2|\underset{\sim}{x}_0) = \frac{p_2 f_2(\underset{\sim}{x}_0)}{\sum_{i=1}^{3} p_i f_i(\underset{\sim}{x}_0)} = \frac{(0.60)(0.85)}{1.2105} = \frac{0.510}{1.2105} = 0.421$$

$$P(\Pi_3|\underset{\sim}{x}_0) = \frac{p_3 f_3(\underset{\sim}{x}_0)}{\sum_{i=1}^{3} p_i f_i(\underset{\sim}{x}_0)} = \frac{(0.35)(2)}{1.2105} = \frac{0.700}{1.2105} = 0.578$$

Nós vemos que $\underset{\sim}{x}_0$ será reconhecido como de Π_3, a população com a maior probabilidade a posteriori.

6.6. RECONHECIMENTO DE PADRÕES COM POPS. GAUSSIANAS

Um caso especialmente importante de Reconhecimento de Padrões Estatístico envolvendo várias populações ocorre quando

$$f_i(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}\left(\underset{\sim}{x} - \underset{\sim}{\mu_i}\right)' \Sigma_i^{-1}\left(\underset{\sim}{x} - \underset{\sim}{\mu_i}\right)\right], \quad i = 1, 2, \cdots, g.$$

são densidades normais multivariadas com vetor de médias $\underset{\sim}{\mu_i}$ e matriz de covariância Σ_i. Se além disso c(i | i) = 0, c (k | i) = i, k ≠ i (ou equivalentemente, os custos de reconhecimento errado são iguais), a regra vista anteriormente

$$\ln[p_k f_k(\underset{\sim}{x})] > \ln[p_i f_i(\underset{\sim}{x})],$$

torna-se:

Reconhecer $\underset{\sim}{x}$ em Π_k se :

$$\ln\ p_k f_k(\underset{\sim}{x}) = \ln\ p_k - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu_k})'\Sigma_k^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu_k}) = \max_i\ \ln\ p_i f_i(\underset{\sim}{x})$$

A constante (p/2)ln(2π) pode ser ignorada desde que seja a mesma para todas as populações (grupos). Assim, definimos o <u>escore quadrático</u> para a população i como:

$$d_i^Q(\underset{\sim}{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}\left(\underset{\sim}{x} - \underset{\sim}{\mu_i}\right)'\Sigma_i^{-1}\left(\underset{\sim}{x} - \underset{\sim}{\mu_i}\right) + \ln(p_i). \qquad i = 1,2, ..., g$$

Assim, a regra de reconhecimento com base na probabilidade total mínima de erro de reconhecimento para populações normais é:

Reconhecer $\underset{\sim}{x}$ como de Π_k se: $d_k^Q(\underset{\sim}{x}) > d_i^Q(\underset{\sim}{x})\ \forall$, i = 1, 2, ... , g, k≠i.

Na prática, os $\underset{\sim}{\mu_i}$ e Σ_i são desconhecidos, mas trabalha-se com as estimativas $\underset{\sim}{\bar{x}_i}$ e S_i e, então, tem-se $\hat{d}_k^Q$ com base nestes valores.

Se as matrizes de covariância, $\Sigma_i = \Sigma$, são iguais tem-se uma simplificação dado que as duas primeiras parcelas são iguais em todas as populações, então tem-se um <u>escore linear</u>:

$$d_i(\underset{\sim}{x}) = \underset{\sim}{\mu_i}'\Sigma^{-1}\underset{\sim}{x} - 1/2\ \underset{\sim}{\mu_i}'\Sigma^{-1}\underset{\sim}{\mu_i} + \ell n(p_i)$$

e da mesma forma reconhecer $\underset{\sim}{x}$ como de Π_k se:

$$d_k(\underset{\sim}{x}) > d_i(\underset{\sim}{x}),\ \forall i = 1, 2, ... , g \qquad k \neq i.$$

Na prática $S_p = \dfrac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g}{n_1 + n_2 + \cdots + n_g - g}$ é o estimador de Σ.

EXERCÍCIO

Calcular os escores discriminante lineares dos dados de g = 3 populações, assumindo que elas sigam distribuição normal bivariada com matriz de covariância comum Σ. Amostras aleatórias das três populações Π_1 , Π_2 e Π_3 são dadas abaixo e as probabilidades *a priori* são $p_1 = p_2 = 0.25$ e $p_3 = 0.50$. Faça também o reconhecimento de $\underset{\sim}{x}_0$ '=[-2, -1].

$$\Pi_1 : X_1 = \begin{bmatrix} -2 & 0 & -1 \\ 5 & 3 & 1 \end{bmatrix} \qquad \underset{\sim}{\bar{x}}_1 = [-1 \;\; 3] \qquad S_1 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

será

$$\Pi_2 : X_2 = \begin{bmatrix} 0 & 2 & 1 \\ 6 & 4 & 2 \end{bmatrix} \quad \underset{\sim}{\bar{x}}_2 = [1 \; 4] \qquad S_2 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\Pi_3 : X_3 = \begin{bmatrix} 1 & 0 & -1 \\ -2 & 0 & -4 \end{bmatrix} \qquad \underset{\sim}{\bar{x}}_3 [0 \;\; -2] \qquad S_3 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

$$S_p = \begin{bmatrix} 1 & -1/3 \\ -1/3 & 4 \end{bmatrix}$$

6.7. REGRA DE RECONHECIMENTO PARA VÁRIAS POPS. COM IGUAL VARIÂNCIA BASEADA NA DIST. DE MAHALANOBIS

Dada a expressão $d_i^Q(\underset{\sim}{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}\left(\underset{\sim}{x} - \underset{\sim}{\mu}_i\right)' \Sigma_i^{-1}\left(\underset{\sim}{x} - \underset{\sim}{\mu}_i\right) + \ell n(p_i)$, $i = 1, 2, \cdots, g$, assume-se variâncias iguais para as várias populações. Tem-se que deve-se reconhecer $\underset{\sim}{x}$ como Π_i se $-\frac{1}{2}\left(\underset{\sim}{x} - \underset{\sim}{\mu}_i\right)' \Sigma_i^{-1}\left(\underset{\sim}{x} - \underset{\sim}{\mu}_i\right) + \ln p_i$ tiver o maior valor para ∀i. Se as probabilidades a prior são desconhecidas assume-se $p_i = 1/g$. A expressão $\left(\underset{\sim}{x} - \underset{\sim}{\mu}_i\right)' \Sigma_i^{-1}\left(\underset{\sim}{x} - \underset{\sim}{\mu}_i\right) = D^2$ é conhecida como distância de Mahalanobis do vetor $\underset{\sim}{x}$ ao vetor $\underset{\sim}{\mu}_i$. Então para o caso real (amostral) $\underset{\sim}{x}_0$ será reconhecido como de Π_k se:

$$-\frac{1}{2} D_i^2(\underset{\sim}{x}_0) + \ln p_i$$

for o maior ∀i.

7. REGRESSÃO LOGÍSTICA: MODELO PARA VARIÁVEIS DICOTÔMICAS.

7.1. Introdução

A regressão logística, dentro da Análise Estatística, consiste em relacionar, através de um modelo, uma variável resposta Y, dicotômica, com os fatores (X_1, X_2, ..., X_{p-1}) que influenciam as ocorrências de determinado evento. Por exemplo, em um estudo para se quantificar a influência de certos fatores na ocorrência de doenças do fígado (colestase), a variável resposta Y será dicotômica, isto é, presença de câncer (Y=1) ou presença de cálculo (Y=0) e os fatores serão, por exemplo, x_1 = bilirrubina total, x_2 = fosfatase alcalina, etc. Os fatores são as variáveis explicativas do modelo e correspondem aos níveis quantitativos obtidos nos exames bio-químicos, etc.

Quando a variável resposta é dicotômica (Y = 1 ou 0) o **Modelo Linear Geral** não deve ser aplicado principalmente por duas razões:

- produzirá valores fora do intervalo [0,1]
- as variâncias das observações não serão constantes..

7.2. Modelo Linear Geral

Seja o problema da construção de um modelo para o relacionamento entre a variável resposta Y e p-1 variáveis explicativas (fatores) X_1, X_2, ..., X_{p-1}. Considerando-se n pontos (y_i, x_{1i}, x_{2i}, ..., x_{p-1i}) i = 1,2, ... ,n pode-se ajustar o modelo linear com p parâmetros:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_{p-1} X_{p-1i} + \varepsilon_i, \quad i=1,2,...,n$$

ou em notação matricial tem-se o vetor aleatório $\underset{\sim}{Y}$ de dimensão n,

$$\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

onde a matriz X

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p-11} \\ 1 & x_{12} & x_{22} & \cdots & x_{p-12} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{p-1n} \end{bmatrix}$$

é de ordem nxp e é denominada matriz do modelo, $\underset{\sim}{\beta}' = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \cdots \quad \beta_{p-1}]$ é o vetor de parâmetros de dimensão p e o vetor dos erros de dimensão n, é dado por $\underset{\sim}{\varepsilon}' = [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_n]$. O modelo $\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ é composto por duas componentes:

- a parte determinística (sistemática) $X \underset{\sim}{\beta}$

- a parte estocástica $\underset{\sim}{\varepsilon}$

No ajuste de um modelo deste tipo o que se deve fazer é:

- Estimar os parâmetros β_i i = 0,1,2, ... ,p-1 ($\underset{\sim}{\hat{\beta}}$ é o estimador do vetor de parâmetros);
- Verificar a qualidade de ajuste;
- Fazer uma análise dos resíduos que indique se premissas de aplicação do modelo estão satisfeitas.

O estimador de mínimos quadrados ordinários (MQO) do vetor dos parâmetros é obtido pela minimização da soma dos quadrados dos resíduos, SQR, na suposição de que os erros são v.a's independentes identicamente distribuídas (i.i.d.),

$$\varepsilon_i \sim (0, \sigma^2)$$

$$\underset{\sim}{\varepsilon_i} \sim (\underset{\sim}{0}, \sigma^2 I)$$

Quando a distribuição comum é Gaussiana, o modelo linear é chamado de modelo normal de Gauss-Markov e o estimador obtido nestas condições é "BLUE" (melhor estimador linear não-viciado). Assim, considerando-se a soma dos quadrados dos erros,

$$SQR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - X \underset{\sim}{\beta})^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

$$SQR = \underset{\sim}{\varepsilon}' \underset{\sim}{\varepsilon} = [\underset{\sim}{Y} - X \underset{\sim}{\beta}]'[\underset{\sim}{Y} - X \underset{\sim}{\beta}] = [\underset{\sim}{Y}' - \underset{\sim}{\beta}' X'][\underset{\sim}{Y} - X \underset{\sim}{\beta}]$$

que é um produto interno, logo comutativo e distributivo

$$SQR = \underset{\sim}{Y}' \underset{\sim}{Y} - \underset{\sim}{\beta}' X' \underset{\sim}{Y} - \underset{\sim}{Y}' X \underset{\sim}{\beta} + \underset{\sim}{\beta}' X' X \underset{\sim}{\beta}$$

e visto que $\underset{\sim}{Y}' X \underset{\sim}{\beta}$ é um escalar resulta:

$$\frac{\partial SQR}{\partial \underset{\sim}{\beta}} = -2X'\underset{\sim}{Y} + 2X'X \underset{\sim}{\beta}$$

e as equações normais de mínimos quadrados são:

$$X'X \underset{\sim}{\beta} = X'\underset{\sim}{Y} \Rightarrow \underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\underset{\sim}{Y}$$

e $\underset{\sim}{\hat{\beta}}$ é a solução do sistema, é o estimador de MQO de $\underset{\sim}{\beta}$.

Então, o modelo ajustado é: $\underset{\sim}{\hat{Y}} = X \underset{\sim}{\hat{\beta}}$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_{p-1} X_{p-1i}$$

E, uma estimativa dos resíduos é:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

A variância dos resíduos, estimada, é $s^2 = \frac{1}{n-p}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$.

No Modelo Linear Normal são feitas as suposições, a serem verificadas, de:

- $\varepsilon_i \sim N(0, \sigma^2)$
- ε_i independentes implica em $V(\underline{\varepsilon}) = \sigma^2 I$

Isto deve ser verificado.

Este procedimento, quando aplicado às variáveis dicotômicas (variável que assume somente um de dois valores: Y = 1 ou Y = 0) apresenta algumas limitações:

$$E(Y_i) = P(Y_i = 1) = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ji} = \theta_i$$

$$V(Y_i) = E[Y_i - E(Y_i)]^2 = (1-\theta)^2 P(Y_i = 1) + (0-\theta_i)^2 P(Y_i = 0)$$

$$V(Y_i) = (1-\theta)^2 \theta_i + \theta_i^2 (1-\theta_i) = (1-\theta)\theta_i [1-\theta_i + \theta_i] = \theta_i (1-\theta_i),$$

logo, $V(Y_i) = \theta_i(1-\theta_i)$ não é constante, invalidando os testes de significância usuais com o modelo linear geral e resposta politômica. Outra dificuldade é que o modelo linear geral fornecerá valores de Y_i fora do intervalo [0,1].

7.3. Modelo Logístico Linear Simples

Seja o modelo linear logístico simples (umavariável explicativa) derivado da função matemática

$$f(y) = \frac{1}{1+e^{-y}}, \ y \in R.$$

que varia monotonicamente de 0 a 1, à medida que y cresce, sendo simétrica em torno de y = 1/2.

É claro que:

$$f(y) = \frac{1}{1+e^{-y}} = \frac{e^y}{1+e^y}$$

e ainda a transformação LOGIT $f(y) = \ln\left[\frac{f(y)}{(1-f(y))}\right] = \ln\left[\frac{(1+e^{-y})^{-1}}{1-(1+e^{-y})^{-1}}\right]$

LOGIT $f(y) = \ln\left[\frac{1}{1+e^{-y}} \Big/ (1-\frac{1}{1+e^{-y}})\right] = \ln\left[\frac{1}{1+e^{-y}} \Big/ \frac{1+e^{-y}-1}{1+e^{-y}}\right]$

$$= \ln\left[\frac{1}{1+e^{-y}} \Big/ \frac{e^{-y}}{1+e^{-y}}\right] = -\ln(1+e^{-y}) - (-y) - (-\ln(1+e^{-y}))$$

$$= -\ln(1+e^{-y}) + y + \ln(1+e^{-y})$$
$$= y$$

Então, impondo um Modelo de Regressão Logístico Linear para estimar P(Y = 1) = p(x) (tratando o caso linear simples, somente com uma variável explicativa), tem-se o modelo dado por:

LOGIT $p(x) = \mu = \beta_0 + \beta_1 x$ que é o nosso y.

A aplicação desse modelo para x = 0 resulta:

$p(0) = P(Y = 1 \mid x = 0) = \frac{e^{\beta_0+\beta_1 \cdot 0}}{1+e^{\beta_0+\beta_1 \cdot 0}} = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ da forma $\frac{e^{y}}{1+e^{y}}$.

7.4. Modelo Logístico Linear Múltiplo

Mas, quando o interesse está em se estabelecer a relação entre a variável resposta Y e as diversas variáveis explicativas $X_1, X_2, \ldots, X_{p-1}$ que podem representar fatores de interesse, o Modelo Logístico Linear Múltiplo tem a forma:

$$\text{LOGIT } p(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} = \mu$$

ou

$$p(\underset{\sim}{x}) = p(x_1, x_2, \ldots x_{p-1}) = e^{u} / (1+e^{u}) = \frac{1}{(1+e^{-u})}$$

onde $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} = \underset{\sim}{x}' \underset{\sim}{\beta}$.

8. ANÁLISE DE AGRUPAMENTOS (CLUSTER ANALYSIS)

8.1- Introdução

A análise de agrupamento é distinta dos métodos de classificação discutidos anteriormente. Classificar é concernente com um número de grupos conhecidos e o objetivo operacional é fixar uma nova observação em um destes grupos. Agrupar é uma técnica mais primitiva no sentido de que nenhuma suposição é feita quanto ao número de grupos ou estrutura de agrupamento. O agrupamento é feito na base de similaridades ou distâncias (dissimilaridades).

Existem os "bons" grupos e os "maus" grupos. Existem algoritmos para, computacionalmente, procurar os bons agrupamentos, com base nas medidas de "proximidade" ou "similaridade".

8.2- Medidas de Similaridades

8.2.1- Distâncias e Coeficientes de Similaridades para Pares de Itens

Quando itens (unidades ou casos) são agrupados, a proximidade é usualmente indicada por alguma espécie de distância. Por outro lado, variáveis são usualmente agrupadas com base nos coeficientes de correlação ou outras medidas de avaliação. A distância Euclidiana entre duas observações multivariadas de dimensão p,

$$\underline{x}' = [x_1, x_2, ..., x_p] \qquad \text{e} \qquad \underline{y}' = [y_1, y_2, ..., y_p]$$

é dada por:

$$d(\underline{x},\underline{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} = \sqrt{(\underline{x} - \underline{y})'(\underline{x} - \underline{y})} = \left\| \underline{x} - \underline{y} \right\|$$

A distância estatística entre as mesmas duas observações é da forma:

$$d(\underline{x},\underline{y}) = \sqrt{(\underline{x} - \underline{y})' A^{-1} (\underline{x} - \underline{y})}$$

onde A^{-1} é tal que $d(x,y) \geq 0$. Assim $(\underline{x} - \underline{y})' A^{-1} (\underline{x} - \underline{y})$ é uma forma quadrática e as entradas de A^{-1} são variâncias e covariâncias amostrais. Contudo, sem conhecimento dos grupos distintos, estas quantidades não podem ser calculadas, e então a distância Euclidiana é preferível na Análise de Cluster.

Outra medida de distância é a métrica de Minkowski ,

$$d(\underline{x},\underline{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{\frac{1}{m}}$$

Para m = 2 $d(\underline{x}, \underline{y})$ é a distância Euclidiana. Em geral variando m é feita uma troca de pares dando maiores e menores diferenças. Sempre que possível, é aconselhável usar distâncias verdadeiras, isto é, distâncias satisfazendo as propriedades:

1ª) $d(P,Q) = d(Q,P)$ (simetria)
2ª) $d(P,Q) > 0$ se $P \neq Q$ (positividade)
3ª) $d(P,Q) = 0$ se $P = Q$
4ª) $d(P,Q) \leq d(P,R) + d(R,Q)$ (desigualdade triangular, com R um ponto intermediário)

para agrupar objetos. Por outro lado, vários algoritmos de agrupamento aceitam distâncias que podem não satisfazer, por exemplo, a desigualdade triangular. Quando os objetos não podem ser representados por medidas p-dimensionais, então os pares de objetos são comparados com base na ausência ou presença de certas características e é claro que itens semelhantes têm mais características em comum do que itens não semelhantes.

EXEMPLO 1:

Sejam p = 5 variáveis binárias que indicam presença (1) ou ausência (0) de certas características nos objetos A e B, adiante:

	Variáveis (Características)				
Observações	C1	C2	C3	C4	C_5
O_1	1	0	0	1	1
O_2	1	1	0	1	0

A distância Euclidiana ao quadrado $d^2(A,B) = \|\underline{a} - \underline{b}\|^2$

$$d^2(A,B) = \left| \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right|^2 = 0^2 + (-1)^2 + 0^2 + 0^2 + 1^2 = 2$$ (norma do vetor ao quadrado)

fornece uma medida do número de **não emparelhamentos** em um par de objetos e é claro que um número grande de **não emparelhamentos** indica **uma menor semelhança**. Mas, uma ponderação nos empates (emparelhamentos) em (1-1) e (0-0) é necessária pois pode ocorrer da presença de uma característica ser mais forte do que a ausência. Por exemplo: se 1 significa "lê grego antigo", é óbvio que o empate em 1-1 é maior indicador de semelhança que o empate 0-0 (não lê grego antigo). Assim é razoável diminuir o número de igualdades 0-0 ou até desconsiderá-las completamente. Desse tratamento diferenciado para igualdades 1-1 e 0-0 surgiram diversos esquemas para definir COEFICIENTES DE SIMILARIDADES.

Seja a tabela de contingência para os itens i e k:

		item k		
		1	0	TOTAL
item i	1	a	b	a + b
	0	c	d	c + d
TOTAL		a + c	b + d	p = a + b + c + d

onde : a = freqüência de igualdades 1-1
b = " " desigualdades 1-0
c = " " " 0-1
d = " " igualdades 0-0

Então os coeficientes usuais de similaridade são dados no quadro adiante.

COEFICIENTE $\tilde{s}(i,k)$	PONDERAÇÃO
1) $\frac{a+d}{p}$	1) pesos iguais para 1-1 e 0-0;
2) $\frac{2(a+d)}{2(a+d)+b+c}$	2) " em dobro para 1-1 e 0-0;
3) $\frac{a+d}{a+d+2(b+c)}$	3) " " " para 1-0 e 0-1;
4) $\frac{a}{p}$	4) desconsiderando 0-0 no numerador;
5) $\frac{a}{a+b+c}$	5) " 0-0 no numerador e denominador;
6) $\frac{2a}{2a+b+c}$	6) " " " " e com peso em dobro para 1-1;
7) $\frac{a}{a+2(b+c)}$	7) desconsiderando 0-0 no numerador e denominador e peso em dobro para as desigualdades
8) $\frac{a}{b+c}$	9) razão das igualdades para as desigualdades, excluindo 0-0

EXERCÍCIO:
Suponha que cinco indivíduos possuem as seguintes características registradas na tabela adiante:

indivíduo	altura	Peso	cor dos olhos	cor dos cabelos	habilidade manual	sexo
1	68 pol	140 lb	verde	louro	destro	feminino
2	73 "	185 "	castanho	castanho	"	masculino
3	67 "	165 "	azul	louro	"	masculino
4	64 "	120 "	castanho	castanho	"	feminino
5	76 "	210 "	castanho	castanho	canhoto	masculino

a) Defina as variáveis binárias
b) Monte o quadro considerando as variáveis binárias definidas
c) Usando o coeficiente de similaridade $\frac{a+d}{p}$, construa a matriz dos coeficientes de similaridades de ordem (5x5) para os n = 5 indivíduos
d) Faça alguma conclusão com base nos coeficientes de similaridade
e) Você teria alguma crítica a fazer ao coeficiente de similaridade $\tilde{s}(i,k) = \frac{a+d}{p}$?

8.2.2. Relação entre coeficiente de similaridade e distância

Pode-se escrever $\tilde{s}$ (i,k)= $\frac{1}{1+d(i,k)}$, onde $0 \leq \tilde{s}$ (i,k) ≤ 1

pois diminuindo a distância aumenta a similaridade e vice-versa. É sempre possível construir coeficientes de similaridades a partir das distâncias, contudo não é possível construir as distâncias a partir das similaridades, a não ser que $\tilde{S}$ seja não-negativa definida e $\tilde{s}$ (i,i) = 1. Desta forma d(i,k) = $\sqrt{2(1-\tilde{s}(i,k)}$ tem as propriedades de uma distância.

8.2.3. Similaridade e medida de associação para pares de variáveis

Quando as variáveis são binárias, os dados podem ser colocados na forma de uma tabela de contingência. As variáveis delineiam as categorias. Para cada par de variáveis existem n objetos categorizados na tabela, assim tem-se:

		Variável k		
		1	0	Total
Variável i	1	a	b	a + b
	0	c	d	c + d
	Total	a + c	b + d	n

e o coeficiente de correlação amostral é $r = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$ que pode ser tomado como uma medida de similaridade entre i e k.

EXERCÍCIO

O significado das palavras muda com o curso da história, contudo o significado dos números 1, 2, 3, ... é uma notável exceção. Uma 1ª. comparação de línguas pode ser baseada nos números. A tabela adiante mostra os primeiros 10 números em Inglês, Polonês, Húngaro, e 8 outras línguas modernas européias. Consideram-se línguas que usam o alfabeto romano e omite-se acentos na análise. A matriz com as freqüências de concordância entre as línguas está em seguida.

	Inglês	Norueg.	Dinam.	Holandês	Alemão	Francês	Esp.	Ital.	Pol.	Húng.	Finlandês
I	10										
N	8	10									
D	8	9	10								
H	3	5	4	10							
A	4	6	5	5	10						
F	4	4	4	1	3	10					
E	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
F	1	1	1	1	1	1	1	1	1	1	10

Você poderia agrupar algumas línguas?

DISTÂNCIA DE MAHALANOBIS

Além da distância Euclidiana, já vista, existe a distância de Mahalanobis (estatística) que leva em conta a matriz de variâncias e covariâncias, S, das variáveis aleatórias e que é dada pela expressão: $d^2(i,j) = (\underline{x}_i - \underline{x}_j')'S^{-1}(\underline{x}_i - \underline{x}_j')$ para os pontos de coordenadas $\underline{x}_i$ e $\underline{x}_j$.

8.3- Agrupamento Hierárquico

No método aglomerativo hierárquico, de início existem tantos grupos quanto objetos. Diversos objetos semelhantes são agrupados primeiro, e estes grupos iniciais são fundidos de acordo com suas similaridades. Eventualmente, relaxando no critério de similaridade, todos os sub-grupos são fundidos dentro de um grupo único. No método aglomerativo hierárquico o procedimento é o seguinte:

1) Iniciando com n grupos (existem n objetos) calcula-se a matriz de distâncias (ou de similaridade) de ordem n x n com d_{ij} sendo a distância (ou similaridade) entre i e j ;

$$D = (d_{ij})$$

2) Na matriz D acha-se o par de grupos mais próximos (mais similares) e junta-se esses grupos;
3) O novo grupo formado é denominado (A,B), p.ex., se os grupos primitivos do par são A e B, nova matriz de distâncias é construída, simplesmente apagando-se as linhas e colunas correspondentes aos grupos A e B e adicionando-se a linha e coluna dadas pelas distâncias entre (AB) e os grupos remanescentes;
4) Repete-se os passos 2 e 3 (n-1) vezes, observando-se a identidade dos grupos que são agrupados e os níveis (distâncias ou similaridades) nos quais ocorrem os agrupamentos;

8.4- Ligações

No item 8.3 descreveu-se o método aglomerativo hierárquico, e ali é feito referências a formação dos agrupamentos, mas como isto é feito? Os agrupamentos são feitos por **ligações**.

8.4.1- Ligação Simples (ou vizinho mais próximo)

Na ligação simples o agrupamento é feito juntando-se os dois grupos com menor distância (ou maior similaridade). Uma vez formado o grupo AB, na ligação simples, a distância entre AB e algum outro grupo C é calculada por:

$$d_{(AB)C} = \min\{d_{AC}, d_{BC}\}$$

Os resultados obtidos são dispostos graficamente em um gráfico chamado diagrama em árvore ou DENDROGRAMA, que possui uma escala para se observar os níveis.

EXERCÍCIO 1

Dada a matriz de distâncias abaixo, faça uma análise de agrupamento construindo o dendrograma.

$$D = (d_{ij}) = \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

EXERCÍCIO 2

Considere a matriz de concordâncias do exercício das línguas, logo a matriz das não-concordâncias é um tipo de matriz de distâncias. Construa a matriz das distâncias e a partir dela o dendrograma.

8.4.2- Ligação Completa (vizinho mais longe)

Na ligação completa o procedimento é muito semelhante ao da ligação simples, com uma única exceção. Em cada estágio, a distância (ou similaridade) entre grupos é determinada pela distância (ou similaridade) entre dois elementos, um de cada grupo e, que são os mais distantes. Assim, a ligação completa assegura que todos os elementos no grupo estão dentro de alguma distância máxima (ou mínima similaridade) de cada outro grupo. O algoritmo aglomerativo começa determinando a menor distância $d_{ik,}$ constrói-se a matriz das distâncias $D = (d_{ik})$ e os grupos vão se

juntando. Se A e B são dois grupos de um único elemento tem-se (AB) como novo grupo. A distância entre (AB) e outro grupo C é dada por $d_{(AB)C} = \max \{d_{(AC)} , d_{(BC)}\}$.

EXERCÍCIO 3

No exercício 1 faça uma Análise de Agrupamento, inclusive com dendrograma, adotando a ligação completa.

8.4.3- Ligação Média

No procedimento com ligação média a distância entre dois grupos é a distância média entre todos os pares de itens. Assim, dada a matriz de distâncias $D = (d_{ik})$ tem-se o grupo (AB) formado devido $d_{AB} = \min\{d_{ik}\}\ \forall i,k$ e a distância entre (AB) e C é dada por:

$$d_{(AB)C} = \frac{\sum_i \sum_k d_{ik}}{n_{(AB)} n_C}$$

EXERCÍCIO 4

Faça um esquema gráfico para ilustrar os três tipos de ligações estudadas.

8.4.5- Método de Agrupamento Não-hierárquico

O agrupamento não-hierárquico é uma técnica usada quando se deseja formar k grupos de itens ou objetos, especificamente. Evidentemente, pode-se usar a técnica para outras propostas. O método de agrupamento não-hierárquico mais usual é o das k-médias, cujo algoritmo é o que segue abaixo.

5) Partição dos itens em k grupos iniciais;
6) Prosseguir com a lista de itens, colocando cada item no grupo cuja média (centróide) está mais próximo, usualmente calcula-se a distância usando a Euclidiana com observações padronizadas ou não, sendo o centróide recalculado para o grupo que recebeu um novo item e para o grupo que perdeu o item;
7) Repete-se o segundo passo até que não restem relocações a serem feitas.

EXERCICIO 5

Suponha que nós possamos medir duas variáveis X_1 e X_2 para cada um dos itens A, B, C e D. Os dados estão na tabela abaixo. Agrupe os itens em dois grupos tais que os itens dentro de um grupo estejam mais próximos uns dos outros do que dos itens do outro grupo.

Item	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

9. DISTRIBUIÇÃO NORMAL MULTIVARIADA

9.1 - Introdução

Dizemos que um vetor aleatório tem distribuição Normal Multivariada se possui a mesma distribuição de uma transformação afim de normais padrões independentes. Isto significa que se X_1, X_2, ... ,X_p são i.i.d. N(0,1), então o vetor $\underline{Y}'=[Y_1,Y_2, ... ,Y_p]$, onde $Y_i = \mu_i + a_{1j} X_1 + a_{2j}X_2 + ... + a_{pj} X_p$ para i,j = 1,2, ..., p , possui distribuição Normal p-variada. Na forma matricial temos $\underline{Y} = A'\underline{X} + \underline{\mu}$ onde A é a matriz da transformação, real p x p, e $\underline{\mu}$ é um vetor real p-dimensional. Então dizemos que $\underline{Y}$ tem distribuição Normal p-variada com média $\underline{\mu}$ e matriz de covariâncias $\Sigma = A'A$, ou seja, $\underline{Y}' \sim N(\underline{\mu}, \Sigma)$.

9.2 - A função densidade de probabilidade da Normal p-variada

A densidade do vetor $\underline{Y}$ é dada por:

$$f(y_1, y_2, ... , y_p) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{y}-\underline{\mu})'\Sigma^{-1}(\underline{y}-\underline{\mu})}$$

$\underline{y} \in \Re^p$, $\underline{\mu} \in \Re^p$ e Σ é definida não-negativa.

EXERCÍCIO 1

Seja o vetor aleatório $[Y_1, Y_2]$ com distribuição Normal Bivariada. Escreva:

a) a f.d.p. do vetor $\underline{Y}$;

b) determine as distribuições marginais: $f_{Y1}(y_1)$, de Y_1, e $f_{Y2}(y_2)$ de Y_2;

c) determine a matriz da covariância, Σ, do vetor $\underline{Y}$;

d) a matriz de correlação do vetor $\underline{Y}$;

EXERCÍCIO 2

Sejam o vetor de médias $\underline{\mu}' =[0, 0]$ e a matriz de transformação $A = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$ para a transformação do vetor $\underline{X}' = [X_1, X_2]$ no vetor $\underline{Y}' = [Y_1, Y_2]$ com X_i v.a's i.i.d N(0,1).

a) Escreva a equação da transformação para cada componente do vetor $\underline{Y}$;
b) Quais as distribuições marginais de Y_1 e de Y_2?
c) Qual a distribuição de $W_1 = X_1+X_2$ e a de $W_2 = X_1-X_2$?
d) Qual a matriz de covariâncias de $\underline{Y}$?

e) Qual a f.d.p. (conjunta) de $\underline{Y}$?

EXERCÍCIO 3

Seja o vetor aleatório $[Y_1, Y_2]$ com distribuição Normal Bivariada com $\sigma_1 = \sigma_2$. Escreva:

a) a f.d.p. do vetor $\underline{Y}$;
b) a matriz de covariâncias do vetor $\underline{Y}$, Σ;
c) as densidades marginais de Y_1 e Y_2

9.3 - Contornos (contours) em densidades de probabilidade constante

Na expressão, citada, da Normal p-variada é possível ver que o lugar geométrico dos valores de $\underline{y}$ a uma altura constante no eixo da f.d.p ($f(\underline{y})$) são elipsóides centrados em $\underline{\mu}$, ou seja, são elipsóides definidos por $(\underline{y} - \underline{\mu})'\Sigma^{-1}(\underline{y} - \underline{\mu}) = c^2$. Os eixos de cada elipsóide de densidade constante estão nas direções dos autovetores de Σ^{-1} (e também de Σ) e seus comprimentos são proporcionais aos recíprocos das raízes quadradas dos autovalores de Σ^{-1}. Assim considerando a expressão:

$$f(y_1, y_2, \ldots , y_p) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{y}-\underline{\mu})'\Sigma^{-1}(\underline{y}-\underline{\mu})}$$

os eixos são $\pm \sqrt{\lambda}_i \underline{e}_i$.

EXERCÍCIO 4

Para a situação do exercício 3, pede-se:

a) os autovalores de Σ;
b) os autovetores deΣ;
c) os eixos considerando o *contour* associado a f.d.p. no valor c^2 ;
b) o comprimento de cada eixo e o ângulo que o eixo maior faz com o eixo Y_1.
e) Faça um esboço da figura gerada na solução do problema.

RESULTADO 9.1

Seja o vetor $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ com $|\Sigma| > 0$. Então:

a) $(\underline{X} - \underline{\mu})' \Sigma^{-1}(\underline{X} - \underline{\mu}) \sim \chi^2_p$ (qui-quadrado com p graus de liberdade);

b) a $N_p(\underline{\mu},\Sigma)$ assume probabilidade 1 - α para o elipsóide (sólido) $\{\underline{x} \mid (\underline{x} - \underline{\mu})' \Sigma^{-1}(\underline{x} - \underline{\mu}) \leq \chi^2_{p\,(1-\alpha)}\}$,

onde $\chi^2_{p\,(1-\alpha)}$ denota o 100(1-α) percentil da distribuição χ^2_p. Assim, o elipsóide $\underline{x}$ satisfazendo $(\underline{x} - \underline{\mu})' \Sigma^{-1}(\underline{x} - \underline{\mu}) \leq \chi^2_{p\,(1-\alpha)}$ tem probabilidade 1 - α. (obs. veja ex. 9)

EXERCÍCIO 5

Suponha que $\underline{Y} \sim N_2(\underline{\mu}, \Sigma)$ tal que $\underline{\mu}'=[15, 20]$ e $\sigma_1^2 = \sigma_2^2 = 25$ e $\rho = 0{,}6$.

a) Escreva a expressão da f.d.p. na forma vetorial e na forma clássica;
b) Determine os autovalores de Σ;
c) Determine os autovetores de Σ;
d) Determine os eixos de um 'contour' (curva de nível) associado a constante c^2;
e) Determine o comprimento de cada eixo da curva de nível do item anterior;
f) Determine o ângulo que o eixo maior faz com o semi-eixo positivo das abscissas;

EXERCÍCIO 6

Em quais circunstâncias a curva de nível do exercício anterior é um círculo?

EXERCÍCIO 7

Seja o vetor $\underline{Y} \sim N_2(\underline{\mu}, \Sigma)$ do exercício 5, determine:

a) o valor de χ^2_p tal que $P[(\underline{y} - \underline{\mu})' \Sigma^{-1}(\underline{y} - \underline{\mu}) \leq \chi^2_{p\,(1-\alpha)}] = 1 - \alpha = 0.90$;

b) Descreva como a $N_2(\underline{\mu},\Sigma)$ assume a probabilidade de 0.90 para o elipsóide sólido (cilindro elíptico) e também faça o esboço do elipsóide;
c) Faça a interpretação geométrica dessa região tridimensional;
d) Faça a interpretação estatística dessa região tridimensional;
e) Escreva a equação da elipse que gera o elipsóide de 90% de confiança;
f) Considerando a equação da elipse encontrada no item anterior verifique quais dos pontos seguintes caem dentro da elipse de 90%: $P_1(23, 25)$, $P_2(10, 15)$, $P_3(19, 14.435)$, $P_4(12, 28)$.

EXERCÍCIO 8

Na situação do problema anterior considere a transformação tal que $Y_1 - \mu_1 = y_1$ e $Y_2 - \mu_2 = y_2$, de modo que a equação da elipse torna-se: $\frac{1}{16}[y_1^2 + y_2^2 - 1.2y_1y_2] = c^2$.

a) Explique, geometricamente, o que ocorreu com essa transformação;

b) Considerando que a equação da elipse de 90% é dada por $y_1^2 + y_2^2 - 1.2y_1y_2 = 73.680$, faça um esboço da elipse e marque os 4 pontos do exercício anterior.

EXERCÍCIO 9 (Densidade Condicional da Normal Bivariada)

Seja o vetor $\underline{Y} \sim N_2(\underline{\mu},\Sigma)$, determine a f.d.p. $f(y_1|y_2)$, condicional de Y_1 dado $Y_2 = y_2$.

RESULTADO 9.2

Se Σ é definida positiva tal que Σ^{-1} existe, $\Sigma\underline{e} = \lambda\underline{e}$ implica em $\Sigma^{-1}\underline{e} = (1/\lambda)\underline{e}$ de modo que ao par de autovalor/autovetor $(\lambda, \underline{e})$ de Σ corresponde o par de autovalor/autovetor $(1/\lambda, \underline{e})$ de Σ^{-1} e ainda Σ^{-1} é definida positiva.

EXERCÍCIO 10

Prove o resultado anterior.

RESULTADO 9.3

Dado a matriz B simétrica, positiva definida, de ordem pxp e o escalar b > 0, então $\frac{1}{|\Sigma|^b} e^{-tr(\Sigma^{-1}B)/2} \le \frac{1}{|B|^b}(2b)^{pb} e^{-bp}$ para toda matriz positiva definida Σ com a igualdade valendo somente para $\Sigma = \frac{1}{2b} B$.

EXERCÍCIO 11

Prove o resultado anterior.

RESULTADO 9.4

Seja $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$ uma a.a. de uma população normal p-variada com média $\underline{\mu}$ e matriz de covariância Σ. Então, $\underline{\hat{\mu}} = \underline{\overline{X}}$ e $\hat{\Sigma} = \frac{n-1}{n} S$ são respectivamente os estimadores de máxima verossimilhança dos parâmetros $\underline{\mu}$ e Σ.

EXERCÍCIO 14

Prove o resultado anterior.

9.4 - Estatísticas suficientes

Da expressão da função densidade de probabilidade conjunta

$$f(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n) = \frac{1}{(2\pi)^{\frac{np}{2}} |\Sigma|^{\frac{n}{2}}} e^{-tr[\Sigma^{-1}(\sum_{j=1}^{n}(\underline{x}_j - \underline{\bar{x}})(\underline{x}_j - \underline{\bar{x}})' + n(\underline{\bar{x}} - \underline{\mu})(\underline{\bar{x}} - \underline{\mu})')]/2}$$

observa-se que a $f(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ depende do conjunto das observações somente através da média amostral

$\overline{X}$ e da soma de quadrados e produtos cruzados $\sum_{j=1}^{n}(\underline{x}_j - \underline{\bar{x}})(\underline{x}_j - \underline{\bar{x}})' = (n-1)S$. Isto significa que $\overline{X}$ e (n-1)S são estatísticas suficientes para estimar $\underline{\mu}$ e Σ. Então, dada a a.a. $[\underline{X}_1, \underline{X}_2, .. , \underline{X}_n]$ de uma população normal p-variada com vetor médio $\underline{\mu}$ e matriz de covariância Σ, as estatísticas $\overline{X}$ e S são estatísticas suficientes para estimar os parâmetros, respectivamente.

9.5 – Distribuição amostral de $\overline{X}$ e S

Seja a a.a. $[\underline{X}_1, \underline{X}_2, ... , \underline{X}_n]$ da v.a. $\underline{X} \sim N_p(\underline{\mu},\Sigma)$, então a distribuição de $\overline{X}$ é determinada de forma análoga ao caso univariado e tem-se que $\overline{X} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma)$ e a distribuição amostral de (n – 1)S segue a distribuição de Wishart. Resumindo tem-se:

- $\underline{\bar{X}} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma)$
- (n – 1)S ~ Wishart com n – 1 g.l's
- $\underline{\bar{X}}$ e S são independentes.

A distribuição de Wishart é definida como a soma de produtos independentes de vetores aleatórios normais, ou seja, $W_m(.|\Sigma)$ é a distribuição de Wishart com m g.l's do produto $\sum_{j=1}^{m}\underline{Z}_j\underline{Z}_j'$ onde $\underline{Z}_j \sim N_p(\underline{0}, \Sigma)$.

- $\sqrt{n}\,(\underline{\bar{X}} - \underline{\mu}) \sim N_p(\underline{0}, S)$
- $n(\underline{\bar{X}} - \underline{\mu})'S^{-1}(\underline{\bar{X}} - \underline{\mu}) \sim \chi^2_p$

EXERCÍCIO 12

Enuncie o Teorema Central do Limite para o caso multivariado.

- R. "Seja $[\underline{X}_1, \underline{X}_2, ... , \underline{X}_n]$ observações independentes da v.a. $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$. Então $\sqrt{n}\,(\underline{\bar{X}} - \underline{\mu})$ tem aproximadamente distribuição $N_p(\underline{0}, \Sigma)$ para n grande e ainda a magnitude de n pode ser relativamente a p".

9.6- Testes sobre os parâmetros de locação e de dispersão de distribuições normais multivariadas e regiões de confiança

9.6.1- Testes da Razão de Verossimilhança

A estratégia geral dos Testes da Razão de Verossimilhança é maximizar a função de verossimilhança sob a hipótese H_0 e também maximizar a função de verossimilhança sob a hipótese alternativa H_1.

Def. Se a distribuição do vetor aleatório $\underline{X}' = [X_1, X_2, ... , X_p]$ depende do vetor de parâmetros $\underline{\theta}$ e se $H_0 : \underline{\theta} \in \Theta_0$ e $H_1 : \underline{\theta} \in \Theta_1$ são as hipóteses envolvidas no teste, então a estatística da razão de verossimilhança que testa H_0 contra H_1 é definida por:

$$\lambda(\underline{x}) = L_1^*/L_0^*$$

onde L_i^* é o maior valor que a função de verossimilhança assume na região Θ_i i = 0, 1. Equivalentemente, pode ser usada a estatística:

$$-2\log(\lambda(\underline{x})) = 2(l_1^* - l_0^*)$$

onde $l_i^* = \log(L_i^*)$. No caso de hipóteses simples, onde cada região Θ_i i = 0, 1 contém somente um único ponto, as propriedades ótimas da estatística razão de verossimilhança são provadas pelo bem conhecido Lema de Neyman-Pearson. De uma maneira geral decidiremos a favor de H_1 quando a estatística da razão de verossimilhança é alta e a favor de H_0 quando ela é baixa. Assim, um teste baseado na estatística razão de verossimilhança pode ser definido da seguinte forma:

Def. O teste da razão de verossimilhança de tamanho α para testar H_0 contra H_1 tem região de rejeição $R = \{\underline{x} \mid \lambda(\underline{x}) > c\}$ onde c é determinado tal que $\sup P_{\underline{\theta}}(\underline{x} \in R) = \alpha$ ($\underline{\theta}$ de H_0).

9.6.2- Seja testar a hipótese $H_0: \underline{\mu} = \underline{\mu_0}$ quando Σ é conhecida e $\underline{X} \sim N_p(\underline{\mu},\Sigma)$

Do procedimento do teste da razão de verossimilhança tem-se a estatística do teste:

$$-2\log(\lambda(\underline{x})) = 2(l_1^* - l_0^*) = n\,(\underline{\bar{x}} - \underline{\mu_0})'\Sigma^{-1}(\underline{\bar{x}} - \underline{\mu_0}) \sim \chi^2_p \quad \text{(exata)}$$

Exemplo 1:
Considere a estatística $\underline{\bar{x}}' = [185.72\ 183.84]$ obtida de uma a.a. com tamanho n = 25 tomada de uma população $N_2(\underline{\mu}, \Sigma)$ com $\Sigma = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$.

a) Teste a hipótese nula de que a distribuição (população) tem média $\underline{\mu_0}' = [182\ 182]$.

Resposta: $-2\log(\lambda(\underline{x})) = 4.31 < \chi_2 = 5.99$ aceitamos H_0.

b) Determine a região de confiança para as médias μ_1 e μ_2

Resposta: $25(185.72 - \mu_1 \quad 183.84 - \mu_2)\begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}^{-1}\begin{bmatrix} 185.72-\mu_1 \\ 183.84-\mu_2 \end{bmatrix} < 5.99$

9.6.3- Seja testar a hipótese H_0: $\underline{\mu} = \underline{\mu}_0$ quando Σ é desconhecido e $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$

Neste caso Σ deve ser estimado sob H_0 e sob H_1. Portanto ambas as hipóteses são compostas. Assim,

$-2\log(\lambda(\underline{x})) = 2(l_1^* - l_0^*) = n\log(1+(\bar{\underline{x}} - \underline{\mu}_0)'S^{-1}(\bar{\underline{x}} - \underline{\mu}_0))$ e $[\frac{n-p}{p}](\bar{\underline{x}} - \underline{\mu}_0)'S^{-1}(\bar{\underline{x}} - \underline{\mu}_0) \sim F_{p,n-p}$

Exemplo 2:
Considere as estatísticas $\bar{\underline{x}}' = [185.72 \; 183.84]$ obtido de uma a.a. com tamanho n = 25 tomada de uma população $N_2(\underline{\mu}, \Sigma)$ que também forneceu $S = \begin{bmatrix} 91.481 & 66.875 \\ 66.875 & 96.775 \end{bmatrix}$.

Teste a hipótese nula de que a distribuição (população) tem média $\underline{\mu}_0' = [182 \; 182]$.

Resposta: $[(n-p)/p](\bar{\underline{x}} - \underline{\mu}_0)'S^{-1}(\bar{\underline{x}} - \underline{\mu}_0) = (23/2)\,[3.72 \quad 1.84]\begin{bmatrix} 91.481 & 66.875 \\ 66.875 & 96.775 \end{bmatrix}^{-1}\begin{bmatrix} 3.72 \\ 1.84 \end{bmatrix} = 1.95 < F_{2,23}(0.95) = 3.44$, logo aceitamos H_0.

9.6.4- Seja testar a hipótese H_0: $\Sigma = \Sigma_0$ quando $\underline{\mu}$ é desconhecido e $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$

Os estimadores de máxima verossimilhança de $\underline{\mu}$ e Σ sob H_0 são, respectivamente, $\underline{\bar{X}}$ e Σ_0. Sob H_1 são $\underline{\bar{X}}$ e S, portanto,

$-2\log(\lambda(\underline{x})) = 2(\ell_1^* - \ell_0^*) = n\,tr(\Sigma_0^{-1}S) - n\log|\Sigma_0^{-1}S| - np$

E, esta estatística é função dos autovalores de $\Sigma_0^{-1}S$ e tem-se, ainda, que Σ_0 é aproximada por S quando $-2\log(\lambda(\underline{x}))$ se aproxima de zero. Então,

$-2\log(\lambda(\underline{x})) = 2(\ell_1^* - \ell_0^*) = n.tr(\Sigma_0^{-1}S) - n\log|\Sigma_0^{-1}S| - np = np[a - \log(g) - 1] \dot{\sim} \chi^2_m$ (assintótica).

Onde, **a** é a média aritmética dos autovalores, **g** é a média geométrica e o número de graus de liberdade **m** é igual ao número de parâmetros independentes em Σ, ou seja, p(p+1)/2.

Exemplo 3

Considere as estatísticas $\underline{\bar{x}}'$ = [185,72 183,84] obtida de uma a.a. com tamanho n = 25 tomada de uma população $N_2(\underline{\mu}, \Sigma)$ que também forneceu S = $\begin{bmatrix} 91,481 & 66,875 \\ 66,875 & 96,775 \end{bmatrix}$. Teste a hipótese nula de que a distribuição (população) tem matriz de covariância Σ = $\begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$, ou seja, H_0: Σ = Σ_0 = $\begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$.

SOLUÇÃO:
A matriz

$$\Sigma_0^{-1}S = \begin{bmatrix} 0,01 & 0 \\ 0 & 0,01 \end{bmatrix}\begin{bmatrix} 91,481 & 66,875 \\ 66,875 & 96,775 \end{bmatrix} = \begin{bmatrix} 0,91481 & 0,66875 \\ 0,66875 & 0,96775 \end{bmatrix}$$

tem autovalores iguais a λ_1 = 1,611 e λ_2 = 0,272. Então, a = 0,9413 e g = 0,6619 e, conseqüentemente, -2log(λ($\underline{x}$)) = 17,70. Comparando o valor da estatística com o escore χ_3^2(0,95) = 7,81 rejeita-se a hipótese H_0 ao nível de 5% de significância. Isto é evidente devido a matriz apresentar forte correlação entre as variáveis.

Exemplo 4

Considere as estatísticas $\underline{\bar{x}}'$ = [185,72 183,84] e S = $\begin{bmatrix} 91,481 & 66,875 \\ 66,875 & 96,775 \end{bmatrix}$ obtidas de uma a.a. de tamanho n = 25 tomada de uma população $N_2(\underline{\mu}, \Sigma)$ com parâmetros desconhecidos. Teste a hipótese nula de que a população tem matriz de covariância Σ = Σ_0 = $\begin{bmatrix} 100 & 50 \\ 50 & 100 \end{bmatrix}$, ou seja, H_0: Σ = Σ_0 = $\begin{bmatrix} 100 & 50 \\ 50 & 100 \end{bmatrix}$.

Reposta: a = 0,8092, g = 0,7642 e -2log(λ($\underline{x}$)) = 3,9065; portanto, comparando com χ_3^2(0,95) = 7,81 aceitá-se H_0.

9.6.5- Região de Confiança do vetor de médias $\underline{\mu}$

Seja $\underline{\theta}$ o vetor de parâmetros populacional desconhecido e Θ o espaço paramétrico de $\underline{\theta}$, ou seja, o conjunto de todos os possíveis valores de $\underline{\theta}$. A região R(X), onde X é a matriz com as observações multivariadas da a.a. X = [$\underline{X}_1$, $\underline{X}_2$, ... , $\underline{X}_n$], é dita ser uma região de confiança ao nível de confiança de (1 - α) se,

$$P[R(X) \text{ cobrir o verdadeiro } \underline{\theta}] = 1 - \alpha$$

A *região de confiança* para o vetor de médias $\underline{\mu}$ de uma população normal p-dimensional é aquela que:

$$P[n(\underline{\bar{x}}-\underline{\mu})' S^{-1}(\underline{\bar{x}}-\underline{\mu}) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(1-\alpha)] = 1-\alpha$$

quando os parâmetros $\underline{\mu}$ e Σ são desconhecidos e estimados por $\underline{\bar{x}}$ e S.

EXERCÍCIO 13

O Departamento de Controle de Qualidade de uma indústria de fornos de microondas recebeu a exigência do Governo Federal para controlar a quantidade de radiação emitida quando as portas dos fornos são fechadas. Foram feitos 42 pares de observações da radiação emitida por n = 42 fornos escolhidos ao acaso, sendo 1° com a porta fechada e 2° com a porta aberta. Sejam X_1 e X_2 as variáveis medidas (com a porta fechada e com a porta aberta). Assumindo que essas variáveis seguem a distribuição Gaussiana e que os dados correspondentes aos 42 pares de observações forneceram as estatísticas seguintes:

$$S = \begin{bmatrix} 0.014 & 0.012 \\ 0.012 & 0.015 \end{bmatrix} \text{ e } \underline{\bar{X}} = [0{,}564; 0{,}603]$$

a) Calcule os autovalores e autovetores de S;
b) Calcule a elipse de 95% de confiança para $\underline{\mu}$;
c) Verifique se $\underline{\mu}$'=[0.562 0.589] está na região de confiança;
d) Determine o comprimento dos semi-eixos positivos da elipse de 95% de confiança;
e) Faça um esboço detalhado da elipse de 95%.

9.6.6- Seja testar a hipótese de matrizes de covariâncias iguais, ou seja:

H_0: $\Sigma_1 = \Sigma_2 = ... = \Sigma_g$

G. E. P. Box, em artigos de 1949 -1950, desenvolveu um teste para a hipótese nula enunciada. A estatística do teste é baseada em:

$$M = \prod_{i=1}^{g} [\frac{|S_i|}{|S_p|}]^{(n_i-1)/2}$$

onde S_i é a matriz de covariância do grupo i;
S_p é a matriz de covariância conjunta (grupos);
n_i é o número de observações do grupo i;
p é a dimensão do vetor $\underline{X}$ amostrado;
g é o número de grupos.

Aplicando uma transformação logarítmica em M tem-se como resultado a estatística com distribuição qui-quadrado adiante:

$$B = (1 - c)\{[\sum_{i=1}^{g}(n_i - 1)\,]\,\ell n|S_p| - \sum_{i=1}^{g}[(n_i - 1)\ell n\,|\,S_i\,|]\} \sim \chi^2_{\frac{1}{2}p(p+1)(g-1)}$$

onde

$$C = [\sum_{i=1}^{g}\frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{g}(n_i - 1)}][\frac{2p^2 + 3p - 1}{6(p+1)(g-1)}$$

EXERCÍCIO 14

Verifique por meio de um teste se a hipótese nula H_0: $\Sigma_1 = \Sigma_2$ é verdadeira com base nos dados seguintes:

$g = 2$, $n_1 = 917$, $n_2 = 83$, $S_1 = \begin{bmatrix} 65,73 & 0,239 \\ 0,239 & 0,369 \end{bmatrix}$ e $S_2 = \begin{bmatrix} 65,73 & 0,239 \\ 0,239 & 0,369 \end{bmatrix}$

9.6.7- Verificação da Gaussianidade para distribuições bivariadas

A suposição de Gaussianidade é muito importante em muitas propostas estatísticas. Por razões práticas é usualmente suficiente investigar-se a Gaussianidade das distribuições univariadas e bivariadas. Se as observações foram geradas de uma distribuição normal multivariada, cada distribuição bivariada pode ser normal e as curvas de nível de densidade constante são elipses. Assim, pelo resultado 9.1, tem-se:

$$(\underline{x} - \underline{\mu})'\Sigma^{-1}(\underline{x} - \underline{\mu}) \leq \chi^2_{2\,(1-\alpha)}$$

e é possível esperar grosseiramente que a porcentagem de 100(1-α)% das observações situem-se na elipse de nível (1 - α) de confiança quando usamos o modelo com os parâmetros estimados por $\bar{\underline{x}}$ e S, respectivamente.

EXERCÍCIO 15

Verifique se os dados das empresas listadas na tabela adiante, ou seja, observações das variáveis X_1 e X_2 seguem a distribuição normal bivariada.

empresa	$X_{1\,(capital)}$	$X_{2\,(rend.\ líquido)}$
1	26.7	3.3
2	38.4	2.4
3	19.2	1.7
4	20.6	1.0
5	18.9	0.9
6	14.8	1.0
7	19.0	2.7
8	14.2	0.8
9	13.7	1.1
10	7.7	0.2

EXERCÍCIOS

1) Suponha uma população normal bivariada com matriz de covariâncias $\Sigma = \begin{bmatrix} 16 & 8 \\ 8 & 9 \end{bmatrix}$ e que uma a.a. de n = 25 observações forneceu um centróide de [15.4 , 9.9]. Teste a hipótese nula de que $\underline{\mu}' = [17, 10]$ ao nível de 5% de significância.

2) Suponha uma a.a. $[\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n]$ de uma distribuição normal $N_p(\underline{\mu}, \Sigma)$.

a) Escreva a distribuição de $\sqrt{n}(\underline{\bar{x}} - \underline{\mu})$;

b) Escreva a distribuição de $n(\underline{\bar{x}} - \underline{\mu})' S^{-1}(\underline{\bar{x}} - \underline{\mu})$;

10. COMPARAÇÃO ENTRE VETORES MÉDIOS

10.1- Comparação entre dois vetores médios: teste T^2 de Hotelling

Sejam duas populações P_1 e P_2 das quais foram tomadas amostras de tamanhos n_1 e n_2 respectivamente. Essas amostras forneceram as estatísticas que estimam os parâmetros populacionais $\underline{\mu}_i$ e Σ_i, ou seja: $\underline{\bar{X}}_1$, $\underline{\bar{X}}_2$, S_1 e S_2. Para se testar a hipótese de que os vetores médios são iguais usa-se a estatística:

$T^2 = [(\underline{\bar{x}}_1 - \underline{\bar{x}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)]' \, [(1/n_1 + 1/n_2)S_p]^{-1}[(\underline{\bar{x}}_1 - \underline{\bar{x}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)]$ que tem por

distribuição: $T^2 \sim \dfrac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}$

EXERCÍCIO 1

Cinqüenta barras de sabão são feitas de duas maneiras. Duas características: X_1 (espuma) e X_2 (brancura) são medidas. As estatísticas para as barras produzidas pelos métodos 1 e 2 são: $\underline{\bar{x}}_1' = [8{,}1 \quad 4{,}1]$, $\underline{\bar{x}}_2' = [\,10{,}2 \quad 3{,}9]$, $S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$, $S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$, pede-se:

a) a estimativa da matriz de covariâncias Σ (supondo comum a variância);
b) o teste da hipótese de que os dois processos de fabricação estão centrados no mesmo ponto;
c) os autovalores e autovetores da estimativa da matriz de covariâncias Σ;
d) a elipse de confiança de nível 95% e verifique se o ponto $\underline{\mu}_1 - \underline{\mu}_2 = 0$ pertence à região de confiança.

SOLUÇÃO
a) Estimativa da matriz de convariância comum Σ

$$S_p = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2} = \frac{(50-1)\begin{bmatrix}2 & 1\\ 1 & 6\end{bmatrix}+(50-1)\begin{bmatrix}2 & 1\\ 1 & 4\end{bmatrix}}{50+50-2} = \begin{bmatrix}2 & 1\\ 1 & 5\end{bmatrix}$$

b) o teste da hipótese de que os dois processos de fabricação estão centrados no mesmo ponto;

$H_0 : \underline{\mu}_1 = \underline{\mu}_2$

$$T^2 = [(\underline{\bar{x}}_1 - \underline{\bar{x}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)]'[(\frac{1}{n_1}+\frac{1}{n_2})S_p]^{-1}[(\underline{\bar{x}}_1 - \underline{\bar{x}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)]$$

$$T^2 = [-2{,}1 \;\; 0{,}2][0{,}04\begin{bmatrix}2 & 1\\ 1 & 5\end{bmatrix}^{-1}\begin{bmatrix}-2{,}1\\ 0{,}2\end{bmatrix} = 63{,}8$$

$T^2 \sim \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)}F_{p,n_1+n_2-p-1}$, então $\frac{98x2}{97}F_{2,50+50-2-1} = \frac{98x2}{97}F_{2,50+50-2-1} =$

2,0206x3,09019 = 6,24

E, como, $T^2 = 63{,}8 > 6{,}24$ rejeitá-se H_0, ou seja, os vetores médios não são iguais.

10.2- Comparação entre vários vetores médios: Manova

Freqüentemente mais de duas populações necessitam ser comparadas em relação aos seus valores médios. As a.a's coletadas de k populações (k > 2) fornecem estatísticas usadas para testar a hipótese de que as populações possuem mesmo ponto médio. As suposições quanto à estrutura dos dados são as seguintes:

1ª.) as amostras aleatórias das diferentes populações são independentes;
2ª.) todas as população têm a mesma matriz de covariância Σ;
3ª.) Cada população é Normal Multivariada, sendo que esta condição pode ser relaxada quando os tamanhos das amostras são grandes (Teorema Central do Limite).

EXERCÍCIO 2

A partir da estrutura dos dados enunciada anteriormente escreva o modelo para uma observação multivariada $\underline{X}_{ij}$, decomponha o vetor de observações e monte o quadro da MANOVA, incluindo os valores de λ^* , o lambda de Wilks, da distribuição exata de Wilks. Escreva ainda a expressão de teste devido a M. S. Bartlett (1938).

EXERCICIO 3

Considere as seguintes amostras independentes das populações 1, 2 e 3, que são Normais Bivariadas com mesma matriz de covariância Σ.

Pop1 [9 3], [6 2], [9 7]
Pop2 [0 4], [2 0]
Pop3 [3 8], [1 9], [2 7]

a) Calcule os vetores médios amostrais;
b) Construa a tabela da MANOVA;
c) Calcule o lambda de Wilks;
d) Calcule a estatística de teste;
e) Teste a hipótese H_0 de que as populações têm o mesmo vetor médio ao nível de 99%.

EXERCÍCIO 4

O Departamento de Saúde e Serviços Sociais de certo Estado subsidia os serviços prestados por asilos de idosos (serviços de amparo a velhice). Esse departamento desenvolveu um conjunto de fórmulas para avaliar o subsídio, baseadas em fatores como nível de cuidados, salário mínimo e salário médio no Estado. As entidades podem ser classificadas com base no tipo de estabelecimento (privado, público e sem fins lucrativos) e na qualidade dos serviços prestados (SNF, ICF ou combinação SNF & ICF). Um estudo pretende investigar os efeitos do tipo de estabelecimento ou qualidade dos serviços (ou ambos) nos custos. Quatro despesas, calculadas por cliente/dia e em horas/cliente por dia, foram selecionadas para análise:

X_1 despesa com o trabalho
X_2 " " a dieta
X_3 " de operação e manutenção do sistema
X_4 " doméstica e de lavanderia

Um total de n = 516 observações de p = 4 variáveis foram tomadas e um resumo das estatísticas está abaixo:

Privado $n_1 = 271$ $\bar{x}_1 = [2{,}066\ 0{,}480\ 0{,}082\ 0{,}360]'$
Sem lucro $n_2 = 138$ $\bar{x}_2 = [2{,}167\ 0{,}596\ 0{,}124\ 0{,}418]'$
Público $n_3 = 107$ $\bar{x}_3 = [2{,}273\ 0{,}521\ 0{,}125\ 0{,}383]'$

e as três matrizes de covariância amostral são:

$$S_1 = \begin{bmatrix} 0{,}291 & & & \\ -0{,}001 & 0{,}011 & & \\ 0{,}002 & 0{,}000 & 0{,}001 & \\ 0{,}010 & 0{,}003 & 0{,}000 & 0{,}010 \end{bmatrix} \quad S_2 = \begin{bmatrix} 0{,}561 & & & \\ 0{,}011 & 0{,}025 & & \\ 0{,}001 & 0{,}004 & 0{,}005 & \\ 0{,}037 & 0{,}007 & 0{,}002 & 0{,}019 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} 0{,}261 & & & \\ 0{,}030 & 0{,}017 & & \\ 0{,}003 & 0{,}000 & 0{,}004 & \\ 0{,}018 & 0{,}006 & 0{,}001 & 0{,}013 \end{bmatrix}$$

a) Calcule o vetor médio amostral;
b) Calcule a matriz da SQ entre os tratamentos;
c) Calcule a matriz da SQ residual;
d) Construa a tabela da MANOVA;
e) Calcule o lambda de Wilks;
f) Calcule a estatística de teste para hipótese de populações com mesma média;

BIBLIOGRAFIA

- Johnson, R. A. & Wichern, D.W. – Applied Multivariate Statistical Analysis; Prentice Hall Inc., Englewood NJ, 1998.
- Mardia, K. V. Kent, J. T. & Bibby, J.M. – Multivariate Analysis; Academic Press, New York, 1978.
- Morrison, D.F. – Multivariate Statistical Methods; McGraw Hill, New York, 1971.
- Hair, Joseph F. Jr. et alii – Multivariate Data Analysis, 5ed., Prentice Hall Inc. Upper Saddle River, N.J. (1998).
- Hair Joseph F. Jr. et alii – Análise de Dados Multivariados; Prentice Hall Inc. Bookman/Artimédia, 2005.