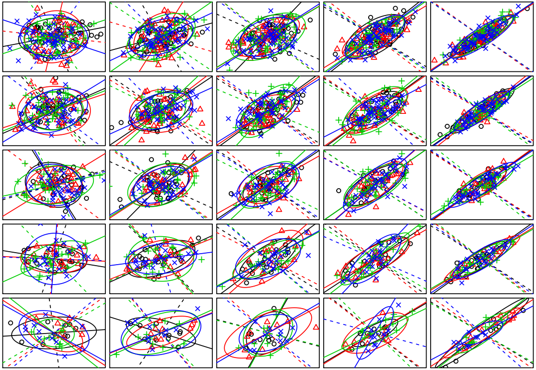
Estatística Multivariada

Slides de apoio às aulas



Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2018/19

Aula 6

Comparação de $g\,(g \geq 2)$ amostras não independentes

1. q medições repetidas, uma amostra aleatória $\rightarrow$ **Testes com matrizes de contrastes**

Elemento	Medições repetidas			
	1	2	$\cdots$	q
1	x_{11}	x_{12}	$\cdots$	x_{1q}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nq}

2. q medições repetidas, duas amostras aleatórias independentes $\rightarrow$ **Análise de perfis**

Amostra	Elemento	Medições repetidas			
		1	2	$\cdots$	q
Amostra 1	1	x_{111}	x_{112}	$\cdots$	x_{11q}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	n_1	$x_{1n_1 1}$	$x_{1n_1 2}$	$\cdots$	$x_{1n_1 q}$
Amostra 2	1	x_{211}	x_{212}	$\cdots$	x_{21q}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	n_2	$x_{2n_2 1}$	$x_{2n_2 2}$	$\cdots$	$x_{2n_2 q}$

- Pretende-se testar as hipóteses:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_q$$

sendo q o número de repetições

- Note-se que a hipótese anterior equivale a:

$$H_0 : \mu_2 - \mu_1 = \mu_3 - \mu_2 = ... = \mu_q - \mu_{q-1} = 0$$

que se pode representar como

$$H_0 : \begin{bmatrix} -1 & 1 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$$

- A matriz $\mathbf{C}_{(q-1)\times q}$ é designada por *matriz de contrastes* representando $q-1$ combinações lineares dos valores médios μ_j $(j = 1, ..., q)$. Cada linha representa um *vetor contraste* cuja soma dos elementos é zero.

- Considerando a amostra aleatória $(\mathbf{x}_1, ..., \mathbf{x}_n)$ proveniente da população $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, tem-se a matriz de médias $\mathbf{C}\bar{\mathbf{x}}$ e a matriz de covariâncias $\mathbf{CSC}'$ e, sob H_0

$$T^2 = n(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{CSC}')^{-1}\mathbf{C}\bar{\mathbf{x}} \frown T^2_{q-1}(n-1)$$

onde

$$T^2_{q-1}(n-1) \stackrel{d}{=} \frac{(n-1)(q-1)}{n-q+1} F_{(q-1,n-q+1)}$$

- Assim, H_0 será rejeitada quando

$$T^2 > \frac{(n-1)(q-1)}{n-q+1} F_{(q-1,n-q+1);1-\alpha}$$

onde $F_{(q-1,n-q+1);1-\alpha}$ representa o quantil de probabilidade $1-\alpha$ da distribuição $F_{(q-1,n-q+1)}$

- É possível mostrar que a estatística T^2 não depende da escolha da matriz $\mathbf{C}$, sendo por isso o mesmo procedimento válido para testar outros contrastes

- Rejeitando H_0 podem testar-se individualmente da um dos $q-1$ contrastes, usando a estatística:

$$T_i^2 = \frac{\sqrt{n}\mathbf{c}_i'\bar{\mathbf{x}}}{\sqrt{\mathbf{c}_i'\mathbf{S}\mathbf{c}_i}} \frown T^2_{q-1}(n-1),\ (i = 1, ..., q-1)$$

sendo $\mathbf{c}_i$ o i-ésimo vetor contraste.

- Assim, ao nível α rejeita-se H_0 quando

$$T^2 > \frac{(n-1)(q-1)}{n-q+1} F_{(q-1,n-q+1);1-\alpha}$$

onde $F_{(q-1,n-q+1);1-\alpha}$ representa o quantil de probabilidade $1-\alpha$ da distribuição $F_{(q-1,n-q+1)}$

- Os IC simultâneos podem determinar-se usando a expressão

$$\mathbf{c}_i'\bar{\mathbf{x}} \pm \sqrt{\frac{(n-1)(q-1)}{n-q+1} F_{(q-1,n-q+1);1-\alpha}} \sqrt{\frac{\mathbf{c}_i'\mathbf{S}\mathbf{c}_i}{n}}\ (i = 1, ..., q-1)$$

Exemplo 1

Considere a seguinte base de dados com observações relativas à velocidade de realização de 2 tarefas (fator A) usando duas marcas de máquinas calculadoras (fator B):

	A1		A2	
Elementos	B1	B2	B1	B2
1	30	21	21	14
2	22	13	22	5
3	20	13	18	17
4	12	7	16	14
5	23	24	23	8

Teste os efeitos dos fatores A e B e interação, considerando $\alpha = 0.05$.

Pretende-se portanto testar $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ que se pode representar pelos seguintes contrastes:

1. $\frac{\mu_1+\mu_2}{2} = \frac{\mu_3+\mu_4}{2}$
2. $\frac{\mu_1+\mu_3}{2} = \frac{\mu_2+\mu_4}{2}$
3. $\frac{\mu_1+\mu_4}{2} = \frac{\mu_2+\mu_3}{2}$

Exemplo (continuação)

- Matriz de contrastes:

$$\mathbf{C}_{3\times 4} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

- Sob $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$

$$T^2 = n(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{x}}) = 21.78$$

Código R:

```
> C<-matrix(c(1,1,1,1,-1,-1,-1,1,-1,-1,-1,1),3,4)
> A1.B1<-c(30,22,20,12,23)
> A1.B2<-c(21,13,13,7,24)
> A2.B1<-c(21,22,18,16,23)
> A2.B2<-c(14,5,17,14,8)
> X<-matrix(c(A1.B1,A1.B2,A2.B1,A2.B2),5,4)
> m<-colMeans(X)
> S=var(X)
> n=nrow(X)
> T2<-n*t(C%*%m)%*%solve(C%*%S%*%t(C))%*%(C%*%m)
> T2

         [,1]
[1,] 21.78032
```

Exemplo (continuação)

- Sob H_0

$$T^2 \frown T_3^2(4) \stackrel{d}{=} \frac{(4)(3)}{2} F_{(3,2)}$$

- Assim, H_0 será rejeitada se

$$T^2 > 6F_{(3,2);0.95} = 114.986$$

Código R:

```
> n=nrow(X)
> q=ncol(X)
> ((n-1)*(q-1))/(n-q+1)*qf(0.95,q-1,n-q+1)

[1] 114.9858
```

Logo, não se rejeita H_0, concluindo-se não existirem evidências de diferenças significativas entre os valores médios.

- Suponhamos agora que se pretende comparar os perfis que se obtêm por ligação linear dos pontos (j, μ_{1j}) e (j, μ_{2j}) $(j = 1, ..., q)$
- Há essencialmente três questões com particular interesse:
  - Serão os perfis paralelos?

$$H_0 : \mu_{1j} - \mu_{1j-1} = \mu_{2j} - \mu_{2j-1} \, (j = 1, ..., q)$$

  - Serão os perfis coincidentes (dado que são paralelos)?

$$H_0 : \mu_{1j} = \mu_{2j} \, (j = 1, ..., q)$$

  - Serão os perfis horizontais (dado que são paralelos e coincidentes)?

$$H_0 : \mu_{11} = ... = \mu_{1q} = \mu_{21} = ... = \mu_{2q} \, (j = 1, ..., q)$$

- O **teste ao paralelismo** dos perfis pode expressar-se pela hipótese

$$H_0 : \mathbf{C}\boldsymbol{\mu}_1 - \mathbf{C}\boldsymbol{\mu}_2 = \mathbf{0}$$

sendo $\boldsymbol{\mu}_1' = (\mu_{11}, ..., \mu_{1q})$, $\boldsymbol{\mu}_2' = (\mu_{21}, ..., \mu_{2q})$ e

$$\mathbf{C}_{(q-1)\times q} = \begin{bmatrix} -1 & 1 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & -1 & 1 \end{bmatrix}$$

- Considerando as amostras aleatórias $(\mathbf{x}_{11}, ..., \mathbf{x}_{1n_1})$ e $(\mathbf{x}_{21}, ..., \mathbf{x}_{2n_2})$ respetivamente provenientes das populações $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ e $N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, tem-se, sob H_0

$$T^2 = (\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{C}\mathbf{S}_{pooled} \mathbf{C}' \right]^{-1} (\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)) \frown T^2_{q-1}(n_1 + n_2 - 2)$$

onde

$$T^2_{q-1}(n_1 + n_2 - 2) \stackrel{d}{=} \frac{(n_1 + n_2 - 2)(q - 1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q)}$$

- Assim, H_0 será rejeitada quando

$$T^2 > \frac{(n_1 + n_2 - 2)(q - 1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q);1-\alpha}$$

onde $F_{(q-1, n_1+n_2-2);1-\alpha}$ representa o quantil de probabilidade $1 - \alpha$ da distribuição $F_{(q-1, n_1+n_2-2)}$

- O **teste à coincidência** dos perfis pode expressar-se pela hipótese

$$H_0 : \frac{\boldsymbol{\mu}_{11} + ... + \boldsymbol{\mu}_{1q}}{q} = \frac{\boldsymbol{\mu}_{21} + ... + \boldsymbol{\mu}_{2q}}{q}$$

equivalente a

$$H_0 : \mathbf{1}'\boldsymbol{\mu}_1 = \mathbf{1}'\boldsymbol{\mu}_2$$

- Considerando as amostras aleatórias $(\mathbf{x}_{11}, ..., \mathbf{x}_{1n_1})$ e $(\mathbf{x}_{21}, ..., \mathbf{x}_{2n_2})$ respetivamente provenientes das populações $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ e $N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, tem-se, sob H_0

$$T^2 = (\mathbf{1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{1}\mathbf{S}_{pooled}\mathbf{1}' \right]^{-1} (\mathbf{1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))$$

equivalente a

$$t = \frac{\mathbf{1}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\sqrt{\mathbf{1}'\mathbf{S}_{pooled}\mathbf{1}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \frown t_{n_1+n_2-2}$$

- Assim, H_0 será rejeitada quando $|t| \geq t_{(n_1+n_2-2);1-\alpha/2}$

- O **teste à horizontalidade** dos perfis pode expressar-se pela hipótese

$$H_0 : \frac{1}{2}(\boldsymbol{\mu}_{11} + \boldsymbol{\mu}_{21}) = \frac{1}{2}(\boldsymbol{\mu}_{12} + \boldsymbol{\mu}_{22}) = \ldots = \frac{1}{2}(\boldsymbol{\mu}_{1q} + \boldsymbol{\mu}_{2q})$$

equivalente a

$$H_0 : \frac{1}{2}\mathbf{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbf{0}$$

- Para estimar $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ usamos $\bar{\mathbf{x}} = \frac{n_1\bar{\mathbf{x}}_1 + n_2\bar{\mathbf{x}}_2}{n_1+n_2}$
- Sob H_0

$$T^2 = (n_1 + n_2)(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{C}\mathbf{S}_{pooled}\mathbf{C})^{-1}\mathbf{C}\bar{\mathbf{x}} \frown T^2_{q-1}(n_1 + n_2 - 2)$$

onde

$$T^2 \stackrel{d}{=} \frac{(n_1 + n_2 - 1)(q - 1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q)}$$

- Assim, H_0 será rejeitada quando

$$T^2 > \frac{(n_1 + n_2 - 2)(q - 1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q);1-\alpha}$$

onde $F_{(q-1,n_1+n_2-q);1-\alpha}$ representa o quantil de probabilidade $1 - \alpha$ da distribuição $F_{(q-1,n_1+n_2-q)}$

Exemplo 2

Considere os dados do ficheiro "data7.xlsx" referentes aos resultados (4 variáveis) de um teste psicológico aplicado a 32 homens (código 1) e 32 mulheres (código 2). Compare os perfis psicológicos das duas populações ($\alpha = 0.01$).

Código R:

```
> dados7<-as.data.frame(readxl::read_xlsx("./Datasets/data7.xlsx", col_names = FALSE))
> colnames(dados7)<-c("sexo","x1","x2","x3","x4")
> q<-ncol(dados7[,-1])
> n1<-sum(dados7$sexo==1)
> n2<-sum(dados7$sexo==2)
> m1<-colMeans(dados7[dados7$sexo==1,-1])
> m1

      x1       x2       x3       x4
15.96875 15.90625 27.18750 22.75000

> m2<-colMeans(dados7[dados7$sexo==2,-1])
> m2

      x1       x2       x3       x4
12.34375 13.90625 16.65625 21.93750

> S1<-var(dados7[dados7$sexo==1,-1]); S2<-var(dados7[dados7$sexo==2,-1])
> Spool<-((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
> Spool

         x1        x2        x3        x4
x1 7.164315  6.047379  5.693044  4.700605
x2 6.047379 15.894153  8.492440  5.855847
x3 5.693044  8.492440 29.356351 13.980847
x4 4.700605  5.855847 13.980847 22.320565

> C<-matrix(c(-1,0,0,1,-1,0,0,1,-1,0,0,1),3,4)
```

- Matriz de contrastes:

$$\mathbf{C}_{3\times 4} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

- Sob $H_0 : \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2$

$$T^2 = (\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{C}\mathbf{S}_{pooled}\mathbf{C}' \right]^{-1} (\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)) = 74.24$$

Código R:

```
> T2<-t(C%*%(m1-m2))%*%solve((1/n1+1/n2)*(C%*%Spool%*%t(C)))%*%(C%*%(m1-m2))
> T2

         [,1]
[1,] 74.24037
```

sendo $T^2 \frown T_3^2(62) \stackrel{d}{=} \frac{(62)(3)}{60} F_{(3,60)}$

- Assim, H_0 será rejeitada quando $T^2 > 3.1F_{(3,60);0.99} = 12.79$

Código R:

```
> (((n1+n2-2)*(q-1))/(n1+n2-q))*qf(0.99,q-1,n1+n2-q)

[1] 12.79026
```

Logo, rejeita-se H_0, concluindo-se que não existem evidências a favor do paralelismo dos perfis.

Análise discriminante

Análise discriminante

- Em termos gerais, a **análise discriminante** visa discriminar grupos a partir de informação recolhida sobre os elementos que constituem esses grupos.

- Objetivos específicos:

  1. Determinar qual ou quais as combinações (lineares) de variáveis — *funções discriminantes* — que maximizam as diferenças entre os grupos → *discriminação*
  2. Predição da pertença de indivíduos não agrupados através do uso das funções discriminantes estimadas → *classificação*

- Exemplos:

  1. Informação sobre p sintomas em n pacientes com g doenças. Pretende-se saber quais os sintomas que melhor discriminam as doenças (discriminação) e dado um novo paciente saber qual a doença mais provável (classificação)
  2. Os potenciais clientes de crédito são agrupados de acordo com o seu comportamento (1 - incumprimento e 0 - não incumprimento). Pretende-se distinguir os dois grupos com base em p variáveis (e.g., idade, estado civil, número de elementos do agregado familiar, salário, número de créditos) e face a um novo cliente prever o seu comportamento.
  3. Informação sobre características geo e bioquímicas de bivalves (concha). Pretende-se saber quais as variáveis que melhor discriminam as regiões de origem e dado uma nova observação alocar a uma origem.

Duas populações

- Sejam $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ os vetores médios de 2 populações multivariadas de dimensão p com matriz de covariâncias $\boldsymbol{\Sigma}$ comum

- Considere-se $\mathbf{x}' = (x_1, ..., x_p)$:

$$\text{População 1: } \mathbf{x}_1' = (x_{11}, ..., x_{1p}) \longrightarrow \text{Amostra 1: } (\mathbf{x}_{11}, ..., \mathbf{x}_{1n_1})$$

$$\text{População 2: } \mathbf{x}_2' = (x_{21}, ..., x_{2p}) \longrightarrow \text{Amostra 2: } (\mathbf{x}_{21}, ..., \mathbf{x}_{2n_2})$$

- *Objetivo geral*: Encontrar a combinação linear das p variáveis — *função discriminante* — que maximiza a distância entre os dois vetores médios $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$

- *Ideia geral*: Transformar as variáveis $(x_1, ..., x_p)$ (originais) numa nova variável y de forma a maximizar a distância entre os vetores médios $(\mathbf{y}_1, \mathbf{y}_2)$ (tantos quantos o número de grupos)

Duas populações

- As variáveis $(x_1, ..., x_p)$ são transformadas (linearmente) usando um vetor de constantes **a** (a estimar):

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \end{bmatrix} = \begin{bmatrix} x_{111} & x_{112} & \cdots & x_{11p} \\ \vdots & \vdots & & \vdots \\ x_{1n_11} & x_{1n_12} & \cdots & x_{1n_1p} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} \Leftrightarrow \mathbf{y}_{1j} = \mathbf{a}'\mathbf{x}_{1j} \, (j = 1, ..., n_1)$$

$$\begin{bmatrix} y_{21} \\ \vdots \\ y_{2n_2} \end{bmatrix} = \begin{bmatrix} x_{211} & x_{212} & \cdots & x_{21p} \\ \vdots & \vdots & & \vdots \\ x_{2n_21} & x_{2n_22} & \cdots & x_{2n_2p} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} \Leftrightarrow \mathbf{y}_{2j} = \mathbf{a}'\mathbf{x}_{2j} \, (j = 1, ..., n_2)$$

- As médias de y em cada amostra são dadas por

$$\bar{y}_1 = \mathbf{a}'\bar{\mathbf{x}}_1 \quad \bar{y}_2 = \mathbf{a}'\bar{\mathbf{x}}_2$$

- A variância comum das novas variáveis (y_1, y_2) é estimada por

$$s_y^2 = \mathbf{a}'\mathbf{S}_{pooled}\mathbf{a}$$

onde

$$\mathbf{S}_{pooled} = \frac{1}{\sum_{i=1}^{g}(n_i - 1)} \left(\sum_{i=1}^{g} (n_i - 1)\mathbf{S}_i \right) .$$

- Expressando a separação entre os dois grupos por

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

pretende-se encontrar o vetor **a** que maximize $(\bar{y}_1 - \bar{y}_2)^2/s_y^2$, ou seja,

$$\frac{\left(\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right)^2}{\mathbf{a}'\mathbf{S}_{pooled}\,\mathbf{a}}$$

ocorrendo o máximo desta razão para

$$\mathbf{a} = \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

- A combinação linear $y = \mathbf{a}'\mathbf{x}$ projeta os pontos $\mathbf{x}$ na reta para a qual $(\bar{y}_1 - \bar{y}_2)^2/s_y^2$ é máxima

Exemplo 3

Considere duas amostras provenientes de populações bivariadas normais com matriz de covariâncias comum, $\mathbf{x} \frown N_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\,(i = 1, 2)$.

Amostra	x_1	x_2
1	33	60
1	36	61
1	35	64
1	38	63
1	40	65
2	35	57
2	36	59
2	38	59
2	39	61
2	41	63
2	43	65
2	41	59

Código R:

```
> x1<-c(33,36,35,38,40,35,36,38,39,41,43,41)
> x2<-c(60,61,64,63,65,57,59,59,61,63,65,59)
> g<-c(rep(1,5),rep(2,7))
> X<-matrix(c(g,x1,x2),12,3)
```

Exemplo (continuação)

Estatísticas amostrais: $\bar{\mathbf{x}}_1' = (36.4, 62.6)$ $\bar{\mathbf{x}}_2' = (39.0, 60.4)$

$$\mathbf{S}_{pooled} = \begin{bmatrix} 7.92 & 5.68 \\ 5.68 & 6.29 \end{bmatrix}$$

Código R:

```
> m1<-colMeans(X[1:5,2:3]); m2<-colMeans(X[6:12,2:3])
> S1<-var(X[1:5,2:3]); S2<-var(X[6:12,2:3])
> Spool<-(4*S1+6*S2)/10
> Spool

     [,1]     [,2]
[1,] 7.92 5.680000
[2,] 5.68 6.291429
```

Logo, $a' = (-1.633, 1.820)$

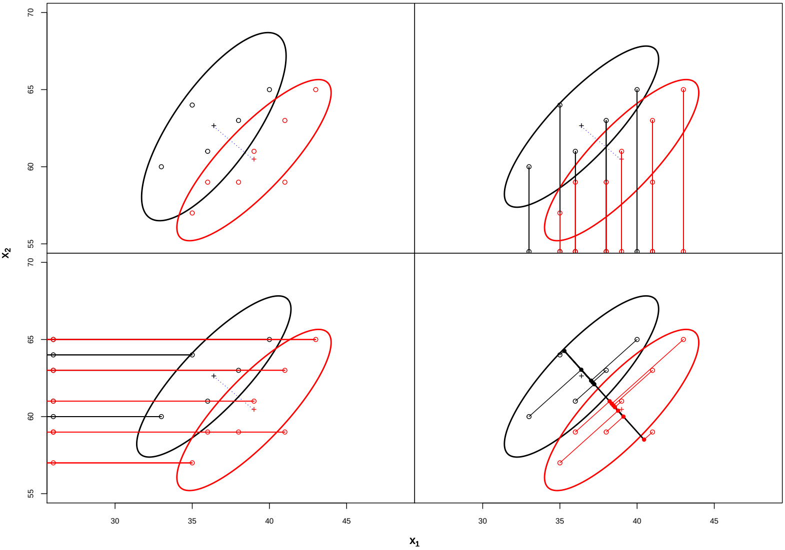
Código R:

```
> a<-solve(Spool)%*%(m1-m2)
> a

          [,1]
[1,] -1.633377
[2,]  1.819779
```

Função discriminante: $y = -1.633x_1 + 1.820x_2$

Exemplo (continuação)



- Os elementos do vetor **a** podem interpretar-se com expressando a contribuição relativa vas variáveis para a função discriminante sse as suas escalas são comparáveis

- Assim, é comum standartizar as variáveis e obter o vetor de constantes $\mathbf{a}^*$ que maximizam a distância entre os vetores médios standartizados:

$$y = a_1^* \frac{x_1 - \bar{x}_1}{s_1} + \ldots + a_p^* \frac{x_1 - \bar{x}_p}{s_p}$$

sendo $a^* = \text{diag}(\mathbf{S}_{pooled})^{1/2}\mathbf{a}$

- No exemplo anterior:

Código R:

```
> std_a<-(diag(Spool)^(1/2))*a
> std_a

         [,1]
[1,] -4.59673
[2,]  4.56450
```

Comente a diferença entre **a** e $\mathbf{a}^*$

- Considere-se g amostras com dimensões $n_i\,(i = 1, ..., g)$ podem obter-se as variáveis transformadas $y_{ij} = \mathbf{a}'\mathbf{x}_{ij}\,(i = 1, ..., g;\; j = 1, ..., n_i)$ com $\bar{y}_i = \mathbf{a}'\bar{\mathbf{x}}_i$
- O vetor de constantes $\mathbf{a}$ pode determinar-se maximizando a razão

$$\frac{\mathbf{a}'\left(\sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})\right)\mathbf{a}}{\mathbf{a}'\left(\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_i)(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_i)\right)\mathbf{a}} = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

  onde $\mathbf{B} = \sum_i n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$ e $\mathbf{W} = \sum_i \sum_j (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$
- Sejam $\ell_1, ..., \ell_s$ os valores próprios não nulos de $\mathbf{W}^{-1}\mathbf{B}$ e $\mathbf{e}_1, ..., \mathbf{e}_s$ os correspondentes vetores próprios (normalizados), então $\mathbf{a}_i = \mathbf{e}_i\,(i = 1, ..., s)$
- A partir dos vetores próprios $\mathbf{e}_i$ podem obter-se as s funções discriminantes

$$y_i = \mathbf{a}_i'\mathbf{x}\,(i = 1, ..., s)$$

- A importância relativa de cada uma das funções é dada por

$$\frac{\ell_i}{\sum_{i=1}^{s} \ell_i}$$

Considere 3 amostras aleatórias relativas a 2 variáveis, selecionadas aleatoriamente a partir de 3 populações com matriz de covariâncias comum $\mathbf{\Sigma}$:

Amostra	x_1	x_2
1	-2	5
1	0	3
1	-1	1
2	0	6
2	2	4
2	1	2
3	1	-2
3	0	0
3	-1	-4

Obtenha as 2 funções discriminantes que maximizam a diferença entre os vetores médios.

Código R:

```
> x1<-c(-2,0,-1,0,2,1,1,0,-1)
> x2<-c(5,3,1,6,4,2,-2,0,-4)
> G<-rep(c(1,2,3),each=3)
> X<-cbind(G,x1,x2)
```

Matrizes **B** e **W**:

$$\mathbf{B} = \begin{bmatrix} 6 & 3 \\ 3 & 62 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} 6 & -2 \\ -2 & 24 \end{bmatrix}$$

Código R:

```
> p=ncol(X[,-1])
> g=length(unique(G))
> n=nrow(X)
> ni<-as.numeric(table(G))
> n1<-ni[1]; n2<-ni[2]; n3<-ni[3]
> S1<-var(X[1:3,-1]); S2<-var(X[4:6,-1]); S3<-var(X[7:9,-1])
> S<-var(X[,-1])
> T<-(n-1)*S
> W<-(n1-1)*S1+(n2-1)*S2+(n3-1)*S3
> B=T-W
> solve(W)%*%B

          x1  x2
x1 1.0714286 1.4
x2 0.2142857 2.7
```

Exemplo (continuação)

Valores próprios não nulos: $\ell_1 = 2.867$ e $\ell_2 = 0.904$
Vetores próprios: $\mathbf{e}_1' = (-0.615, -0.789)$ e $\mathbf{e}_2' = (-0.993, 0.118)$

Código R:

```
> ev<-eigen(solve(W)%*%B)
> ev$values

[1] 2.8670711 0.9043575

> ev$vectors

           [,1]       [,2]
[1,] -0.6148676 -0.9929546
[2,] -0.7886303  0.1184957
```

Sendo as duas funções dadas por:

$$y_1 = -0.615x_1 - 0.789x_2$$

$$y_2 = -0.993x_1 + 0.118x_2$$

com importâncias relativas respetivamente iguais a 0.76 e 0.24.

- Neste contexto, pretende-se saber se o vetor de constantes que maximiza a diferença entre vetores médios é significativamente diferente do vetor nulo

$$H_0 : \boldsymbol{\alpha} = \mathbf{0}$$

sendo $\boldsymbol{\alpha} = \boldsymbol{\Sigma}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

- Note-se que

$$H_0 : \boldsymbol{\alpha} = \mathbf{0} \Leftrightarrow H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

- Assim, sob H_0

$$T^2 = [(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pooled} \right]^{-1} [(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \frown T_p^2(n_1 + n_2 - 2)$$

com

$$T^2 \overset{d}{=} \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{(p, n_1 + n_2 - p - 1)}$$

- Assim, H_0 será rejeitada quando $T^2 > \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{(p,n_1+n_2-p-1);1-\alpha}$ onde $F_{(p,n_1+n_2-p-1);1-\alpha}$ representa o quantil de probabilidade $1 - \alpha$ da distribuição $F_{(p,n_1+n_2-p-1)}$

Testes de significância - g populações

- Vimos anteriormente que $(\mathbf{a}'\mathbf{B}\mathbf{a})/(\mathbf{a}'\mathbf{W}\mathbf{a})$ tem máximo correspondente ao primeiro valor próprio de $\mathbf{W}^{-1}\mathbf{B}$ (correspondendo os restantes às outras funções discriminantes)

- Recorde que estes valores próprios são os mesmos usado na determinação da estatística Lambda de Wilks Λ^*, com distribuição assintótica

$$-\left(n-1-\frac{p+g}{2}\right)\ln\Lambda^* \overset{a}{\sim} \chi^2_{p(g-1)}$$

- Pretende-se testar $H_0 : \lambda_i = 0\,(i = 1, ..., s)$ (sendo s o número de valores próprios não nulos)

- Para o m-ésimo teste, tem-se

$$\Lambda^*_m = \prod_{i=m}^{s} \frac{1}{1+\ell_i} \overset{a}{\sim} \chi^2_{(p-m+1)(g-m)}$$

(Note-se que não basta atender à significância dos valores próprios. Para cada função deve ter-se em conta $\ell_i / \sum_i \ell_i$)

Exemplo (continuação)

Pretende-se testar $H_0^{(1)} : \lambda_1 = 0$ e $H_0^{(2)} : \lambda_2 = 0$. Têm-se as estatísticas de teste: $\Lambda_1^* = \prod_{i=1}^{2} \frac{1}{1+\ell_i} = 0.135$ e $\Lambda_2^* = \prod_{i=2}^{2} \frac{1}{1+\ell_i} = 0.525$, com $\Lambda_1^* \overset{a}{\sim} \chi_2^2$ e $\Lambda_2^* \overset{a}{\sim} \chi_1^2$, sob H_0.

Código R:

```
> p=ncol(X[,-1]); g=length(unique(G))
> ev<-eigen(solve(W)%*%B)
> Lambda1<-prod(1/(1+ev$values[1:2]))
> Lambda1

[1] 0.1357905

> Lambda2<-prod(1/(1+ev$values[2]))
> Lambda2

[1] 0.5251115
```

Quantil 0.95 das distribuições χ_4^2 e χ_1^2: $\chi^2_{(2),0.95} = 9.488$ e $\chi^2_{(1),0.95} = 3.841$

Código R:

```
> qchisq(0.95,p*(g-1))

[1] 9.487729

> qchisq(0.95,(p-2+1)*(g-2))

[1] 3.841459
```

Rejeitam-se as duas hipóteses. Logo as duas funções são significativas na separação dos grupos.

- Na prática, a interpretação das fd corresponde à determinação da contribuição relativa de cada uma das variáveis na discriminação dos grupos

- Existem basicamente 3 métodos:

  1. Análise dos coeficientes das funções discriminantes (standartizadas): Os valores absolutos quantificam a importância de cada variável na separação dos grupos. O sinal de cada coeficiente fornece informação sobre o sentido da associação.

  2. Cálculo da **correlação** linear entre cada variável original e as novas variáveis

  3. Calculo da estatística F-**parcial** para cada variável

Valores F-parcial

- Para cada elemento do vetor $\mathbf{x}$, x_j $(j = 1, ..., p)$ é possivel calcular a estatística F parcial, para estudar a significância da contribuição de x_j em adição às restantes $j-1$ variáveis $(j = 1, ..., p)$

- Para $g = 2$

$$F = (\nu - p + 1)\frac{T_p^2 - T_{p-1}^2}{\nu + T_{p-1}^2} \frown F_{1,(\nu-p+1)}$$

sendo T_p^2 a estatística T^2 de Hotelling para 2 populações considerando as p variáveis e T_{p-1}^2 a mesma estatística considerando $p-1$ variáveis (todas menos a x_j) e $\nu = n_1 + n_2 - 2$

- Para $g > 2$

$$F = \frac{1-\Lambda}{\Lambda}\frac{n-g-p+1}{g-1} \frown F_{(g-1),(n-g-p+1)}$$

com $\Lambda(x_j|x_1, ..., x_{j-1}, x_{j+1}, ..., x_p) = \frac{\Lambda_p^*}{\Lambda_{p-1}^*}$

- Os valores F-parcial constituem um índice global da contribuição de cada variável para a separação dos grupos.

Cálculo de Λ-parciais: $\Lambda^*(x_1)$ e $\Lambda^*(x_2)$:

Código R:

```
> #Lambda partial 1
> s11<-var(X[1:3,2]); s21<-var(X[4:6,2]); s31<-var(X[7:9,2])
> s1<-var(X[,2])
> T<-(n-1)*s1
> W<-(n1-1)*s11+(n2-1)*s21+(n3-1)*s31
> B=T-W
> ev1<-eigen(solve(W)%*%B)
> Lambdap1<-prod(1/(1+ev1$values))
> Lambdap1

[1] 0.5
```

Código R:

```
> #Lambda partial 2
> s12<-var(X[1:3,3]); s22<-var(X[4:6,3]); s32<-var(X[7:9,3])
> s2<-var(X[,3])
> T<-(n-1)*s2
> W<-(n1-1)*s12+(n2-1)*s22+(n3-1)*s32
> B=T-W
> ev2<-eigen(solve(W)%*%B)
> Lambdap2<-prod(1/(1+ev2$values))
> Lambdap2

[1] 0.2790698
```

Exemplo (continuação)

Cálculo de F-parciais:

Código R:

```
> Lambdap<-prod(1/(1+ev$values[1:2]))
> Lambdap
```

```
[1] 0.1357905
```

```
> L1<-Lambdap/Lambdap1
> L2<-Lambdap/Lambdap2
> Fp1<-((1-L1)/L1)*((n-g-p+1)/(g-1))
> Fp2<-((1-L2)/L2)*((n-g-p+1)/(g-1))
> Fp1
```

```
[1] 6.705357
```

```
> Fp2
```

```
[1] 2.637874
```

Classificação/Alocação

- A análise discriminante é essencialmente um processo *descritivo* no qual a discriminação de grupos é feita mediante a caracterização das funções discriminantes
- Estas mesmas funções podem ser usadas numa perspectiva *preditiva*, permitindo alocar observações (cuja classificação é desconhecida) aos grupos
- Para $g = 2$:
  1. Calcular $\hat{y}_0$ correspondente à observação multivariada de classificação desconhecida $\mathbf{x}_0$
  $$\hat{y}_0 = \mathbf{a}'\mathbf{x} = \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{x}_0$$
  2. Calcular o ponto médio
  $$\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$
  3. Alocar $\mathbf{x}_0$ ao grupo 1 se $\hat{y}_0 > \hat{m}$ (ao grupo 2, caso contrário)
- Para $g > 2$:
  1. Comparar $\mathbf{x}$ (observação de classificação desconhecida) com cada $\bar{\mathbf{x}}_i$ ($i = 1, \ldots, g$) usando a função distância
  $$d_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}_{pooled}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)\,(i = 1, \ldots, g)$$
  2. Alocar $\hat{x}$ ao grupo para o qual d_i^2 é mínima.

- Uma vez que este processo analítico é um também um procedimento preditivo, é possível prever a classificação para cada uma das observações

- A construção de uma tabela de contingência cruzando as classificações observada com a prevista permite estimar a *eficácia preditiva* do método

Observado	Previsto	
	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}

Proporção de casos corretamente classificados (taxa de acerto) $= \frac{n_{11}+n_{22}}{n_1+n_2}$

Proporção de casos incorretamente classificados (taxa de erro) $= \frac{n_{12}+n_{21}}{n_1+n_2}$

Exemplo (continuação)

Quais as taxas de acerto e erro com base na análise realizada?

Código R:

```
> m<-aggregate(X[,2:3],list(X[,1]),mean)
> Spool<-((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
> d<-data.frame()
> x<-numeric()
> for (i in 1:3) {
 for (j in 1:9) {
 x<-X[j,-1]
 center<-as.matrix(x-m[,-1])
 d[j,i]<-t(matrix(center[i,],2,1))%*%solve(Spool)%*%matrix(center[i,],2,1)
 }}
> for (j in 1:9) {
 d$min[j]<-min(d[j,1:3])
 d$prev[j]<-which(d[j,1:3]==d$min[j])
 }
> d$real<-G
> tab<-table(d$prev,d$real)
> acerto<-sum(diag(tab))/9
> erro<-1-acerto
> erro

[1] 0.1111111
```

Análise discriminante (LDA) no R

Considere a base de dados que contém as concentrações de 13 substâncias químicas em castas de uva cultivadas, numa mesma região de Itália, provenientes de três cultivares diferentes:

Código R:

```
> library("MASS")
> wine<-read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/wine/
 wine.data",sep=",")
```

Código R:

```
> wine.lda <- lda(wine$V1 ~ wine$V2 + wine$V3 + wine$V4 + wine$V5 + wine$V6 + wine$V7 +
 wine$V8 + wine$V9 + wine$V10 + wine$V11 + wine$V12 + wine$V13 +wine$V14)
> wine.lda$scaling

                 LD1           LD2
wine$V2  -0.403399781  0.8717930699
wine$V3   0.165254596  0.3053797325
wine$V4  -0.369075256  2.3458497486
wine$V5   0.154797889 -0.1463807654
wine$V6  -0.002163496 -0.0004627565
wine$V7   0.618052068 -0.0322128171
wine$V8  -1.661191235 -0.4919980543
wine$V9  -1.495818440 -1.6309537953
wine$V10  0.134092628 -0.3070875776
wine$V11  0.355055710  0.2532306865
wine$V12 -0.818036073 -1.5156344987
wine$V13 -1.157559376  0.0511839665
wine$V14 -0.002691206  0.0028529846

> wine.lda.values <- predict(wine.lda, wine[2:14])
```

Análise discriminante (LDA) no R

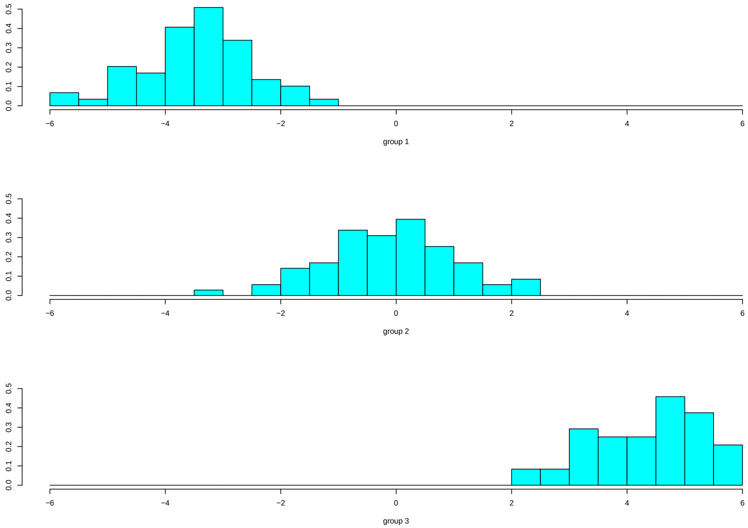
Código R:

```
> ldahist(data = wine.lda.values$x[,1], g=wine$V1)
> ldahist(data = wine.lda.values$x[,2], g=wine$V1)
```

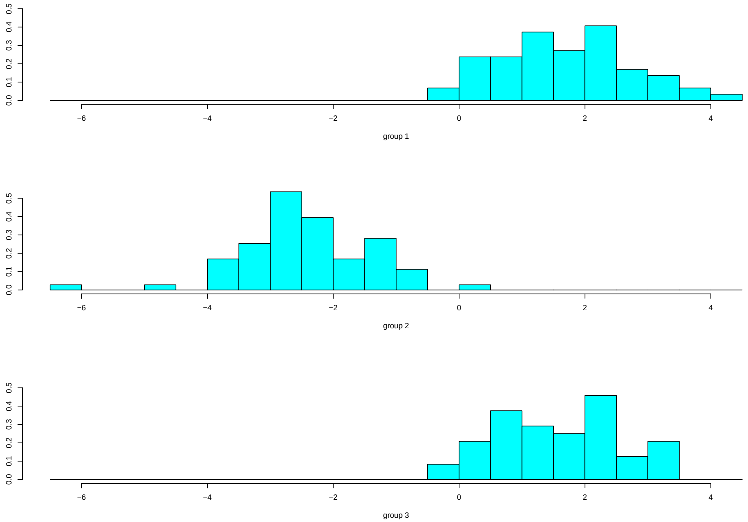
Código R:

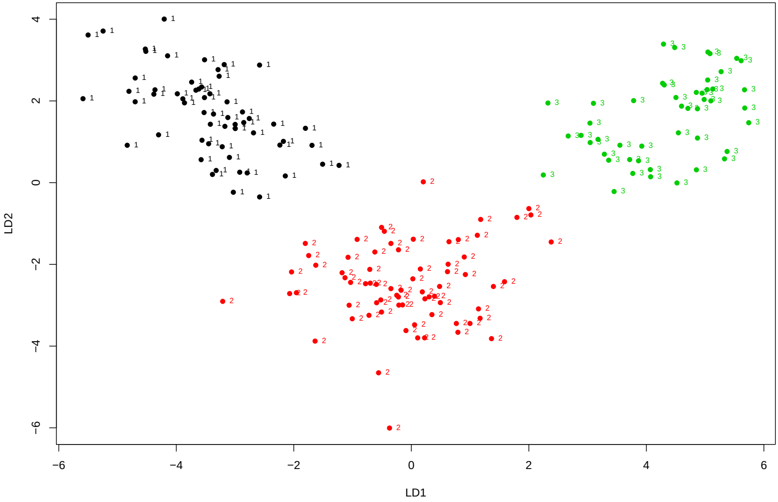
```
> plot(wine.lda.values$x[,1],wine.lda.values$x[,2],xlab = "LD1",ylab = "LD2",
 col=wine$V1,pch=16)
> text(wine.lda.values$x[,1],wine.lda.values$x[,2],wine$V1,
      cex=0.7,pos=4,col=wine$V1)
```

Análise discriminante (LDA) no R



Análise discriminante (LDA) no R




LD1
LD2
−6
−4
−2
0
2
4
6