Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IBI 5086 – Bioinformática - IME/USP

2º Sem/2023

Programa

- Álgebra linear básica: cálculo matricial, determinantes, sistemas lineares, produto interno, norma, ortogonalidade, autovalores e autovetores

- ✓ Estrutura de Dados: variáveis (resposta, explicativa), unidades amostrais e experimentais

- ✓ 1.1. Comparação de 2 ou mais grupos: Testes Clássicos (teste t, Wilcoxon,

  ANOVA), Testes de Aleatorização, Comparações Múltiplas, Efeitos Genéticos

- ✓ 1.2. Análise de Tabelas de Contingência: Testes Qui-Quadrado, Regressão Logística.

2. **Análise Multivariada de Dados**: Componentes Principais, Coordenadas Principais, Análise de Correspondência, **Análise Discriminante e Classificação**, Análise de Agrupamento, Correlação Canônica, Modelos MANOVA

3. Simulação de Monte Carlo, Intervalos de Confiança Bootstrap

Análise Multivariada $\rightarrow$ Redução de Dimensionalidade

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$$Y_{n\times p}; \quad n > p \qquad \Re^p \rightarrow \Re^m, m < p$$

Redução de Dimensionalidade $\Rightarrow$ de ***p*** variáveis, obter ***m*** novas variáveis (m<p)

(são combinações lineares das ***p*** variáveis e atendem a critérios de otimalidade)

$Y_{n\times p}$ ***p*** vetores das respostas para ***n*** indivíduos $\Rightarrow$ Vetores de **Escores** para os n indivíduos (CP)

Vetores de **Cargas** (pesos) às ***p*** variáveis

Ténicas Multivariadas de Redução de Dimensionalidade

Como obter vetores reducionistas de dados?

Depende:

- Estrutura dos Dados
- Objetivo da análise

✓ Análise de Componentes Principais: $Y_{n \times p} \Rightarrow S_{p \times p}, \quad R_{p \times p}$

✓ Escalonamento Multidimensional: $Y_{n \times p} \Rightarrow D_{n \times n}$

✓ Análise de Correspondência: $Y_{n \times p} \Rightarrow [0,1]^{I \times J}$

Análises não supervisionadas

- Análise Discriminante $Y_{n \times (p+1)} \Rightarrow \Re^{p \times p}$ $\Rightarrow$ MANOVA Regressã logística!

Análise supervisionada

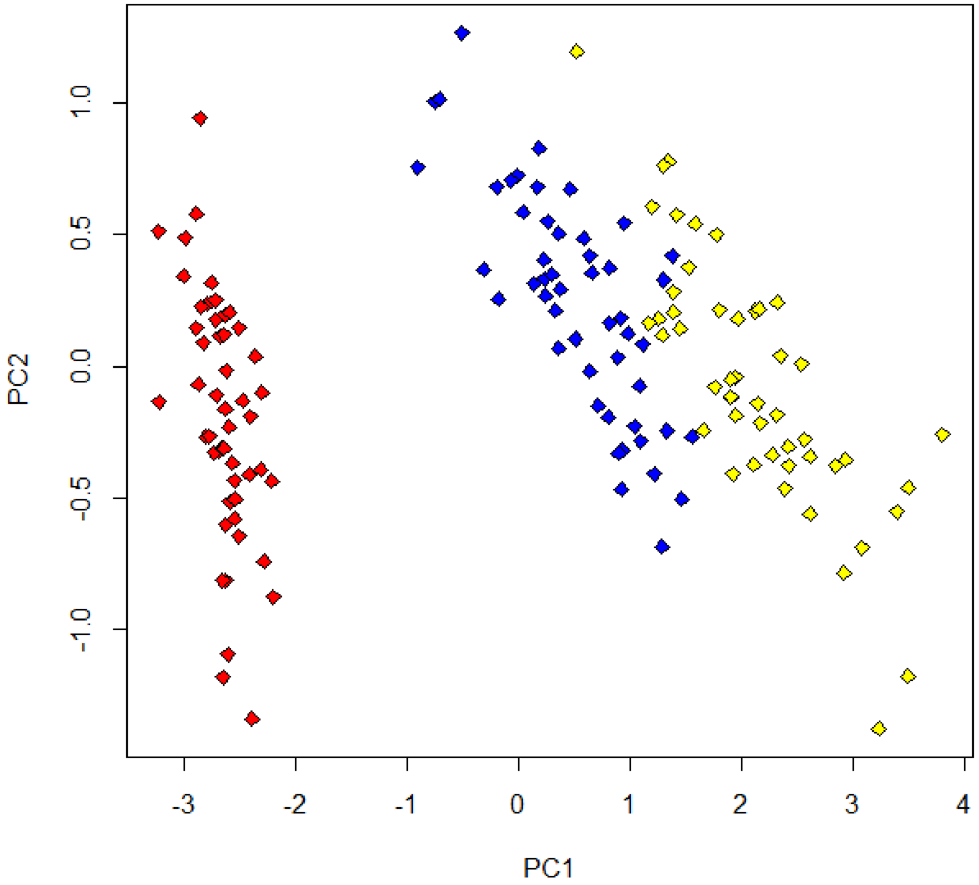
- Análise de Agrupamento: (das n observações)
- Análise de Correlação Canônica: $Y_{n \times (p+q)} \Rightarrow S_{(p+q) \times (p+q)}$

Análises não supervisionadas

Redução de Dimensionalidade

Dados "Iris" do R :Medidas do comprimento e largura da pétala e sépala de 50 flores de íris de três espécies (setosa, versicolor e virginica).

$$Y_{150\times 4} = \begin{pmatrix} Y_{G=1\ 50\times 4} \\ Y_{G=2\ 50\times 4} \\ Y_{G=3\ 50\times 4} \end{pmatrix}$$



Representação dos dados Iris de acordo com os escores dos dois primeiros Componentes Principais (Cp1 e CP2). Indicação das três espécies de flores.

Esta análise discriminou as três espécies?
A construção destes componentes levou em conta a estratificação dos dados?

Análise Discriminante

Análise Clássica

n > p

P Variáveis quantitativas

Observações independentes

e agrupadas *(G=2 ou + grupos)*

Técnicas de Redução de Dimensionalidade
Análise Discriminante

Análises Não-Supervisionadas

- ✓ Componentes Principais
- ✓ Coordenadas Principais
- ✓ An. de Correspondência

$$Y_{n\times p}$$

Reduzir dimensionalidade dos dados, de "p" para "m", m ≤ mínimo(n,p)

$S_{p\times p}$ $D_{n\times n}$

CP **CoP**

Análise Supervisionada de Redução de dimensionalidade

Análise Discriminante: $Y_{n\times p};\quad \boxed{n=\sum_{g=1}^{G} n_g,}\quad g=1,2,\ldots,G$

$$\Re^{p}\rightarrow\Re^{m};\quad m<p$$

- Dados Estratificados (G grupos): Trataremos de situações com G=2 e G>2

$\Rightarrow$ Solução de Fisher

- Solução Probabilística (Regra Discriminante de Bayes)

População Estratificada

τ_1 τ_2 ... τ_G

$E(Y_i \mid \tau_1) = \mu_{1(p\times 1)}$ $E(Y_i \mid \tau_2) = \mu_{2(p\times 1)}$ ... $E(Y_i \mid \tau_G) = \mu_{G(p\times 1)}$

$Cov(Y_i \mid \tau_1) = \Sigma_{1(p\times p)}$ $Cov(Y_i \mid \tau_2) = \Sigma_{2(p\times p)}$ ... $Cov(Y_i \mid \tau_G) = \Sigma_{G(p\times p)}$

n_1 n_2 n_G

Amostra

Grupos	Unidades amostrais	Variáveis 1	2	…	j	…	p
1	1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
	2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
	…	…	…	…	…		…
…	i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
	…	…	…	…	…	…	…
G	n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

Matriz de Dados

$$Y_{n\times p} = \begin{pmatrix} Y_1 \\ \dots \\ Y_G \end{pmatrix}$$

$$Y_{g(n_g\times p)}$$

$$n = \sum_{g=1}^{G} n_g$$

Objetivos da ANÁLISE DISCRIMINANTE

- Obter “m” Variáveis Discriminantes das “p” variáveis

  - Redução de dimensionalidade: $\Re^p \rightarrow \Re^m$; m≤min[n,p,(G-1)]

  - Classificação de “novas” observações (predição de grupos)

Análise Discriminante - Motivação

Exemplo: Dados do desempenho de atletas.

n = 95 p= (5+**1**)

Dados

	V0	Gene	V1_1	V1_2	V2_1	V2_2
1	73.0	1	210	260	27.133	31.398
2	78.0	1	260	320	23.841	26.950
3	70.0	3	220	320	23.755	28.937
...						
94	95.0	2	175	260	25.120	28.605
95	71.0	2	220	330	25.452	29.029

Indicação de Grupos Genotípicos:
(dados desbalanceados)

1	2	3	Total
17	50	28	95

Vetor centróide

V0	V1_1	V1_2	V2_1	V2_2
75.97	261.79	337.58	25.48	29.13

Matriz de Covariância (superior) e **Correlação** (inferior)

	V0	V1_1	V1_2	V2_1	V2_2
V0	**155.24**	317.53	303.47	20.80	19.37
V1_1	0.40	**4032.93**	3815.02	70.70	59.41
V1_2	0.36	0.88	**4640.35**	71.05	65.73
V2_1	0.41	0.27	0.26	**16.40**	17.10
V2_2	0.35	0.21	0.22	0.95	**19.85**

Como as observações estão distribuídas de acordo com os 3 Grupos genotípicos?

É possível predizer o grupo genotípico em função dessas 5 variáveis?

Análise Discriminante - Motivação

1. Medidas biométricas (mm) de Pardais fêmea

(Manly, 2005; Hermon Bumps, 1898).

Pardal	**Sobrev.**	**X1**	**X2**	**X3**	**X4**	**X5**
1	S	156	245	31.6	18.5	20.5
…	…					
21	S	159	236	31.5	18.0	21.5
22	N	155	240	31.4	18.0	20.7
…	…					
49	N	164	248	32.3	18.8	20.9

Como as características corporais dos pardais pode ser usada para predizer os grupos de Sobreviventes e Não-Sobreviventes?

Análise Discriminante - Motivação

⇒ Dados do Transcriptoma: A expressão de “genes” pode ser usada para predizer (caracterizar) diferentes tecidos tumorais (cancer)?

Irizarry, R.A. and Love, M.I.
Data Analysis for the Life Sciences, 2015.

$$Y_{189\times 22.215} = \begin{pmatrix} Y_1' \\ \cdots \\ Y_{189}' \end{pmatrix} = \left(Y_{(1)}, \ldots, Y_{(22.215)}\right)$$

```
library(devtools)
install_github("genomicsclass/tissuesGeneExpression")

library(tissuesGeneExpression)
data(tissuesGeneExpression)
dim(e)  ## e contains the expression data
## [1] 22215 189

table(tissue) ##tissue[i] tells us what tissue is represented by e[,i]
## tissue
## cerebellum colon endometrium hippocampus kidney liver placenta
## 38          34    15          31          39     26    6
```

p=22.215

n=189

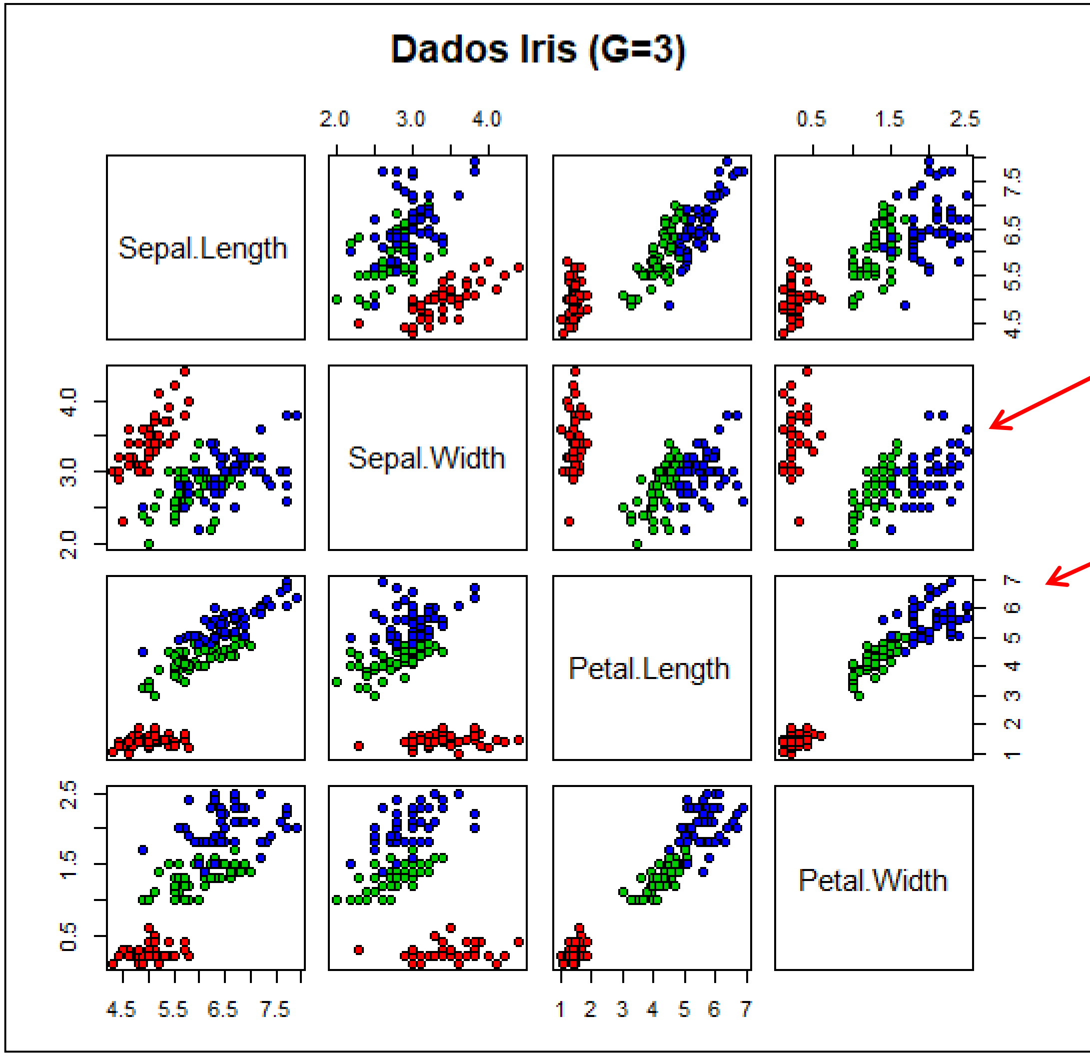
Análise Discriminante - Motivação

Dados "Iris" do R: Medidas do comprimento e largura da pétala e sépala de 50 flores de íris de três espécies (setosa, versicolor e virginica).

E os dados Iris, podem ser melhor representados com o objetivo de discriminar as espécies?
Uma flor com as dimensões (4.5, 2.7, 5.0, 0.8) para as variáveis avaliadas, provavelmente pertence a qual espécie?

	Sepal.Length	**Sepal.Width**	**Petal.Length**	**Petal.Width**	**Species**
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
...					
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
...					
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
...					
150	5.9	3.0	5.1	1.8	virginica

Análise Discriminante



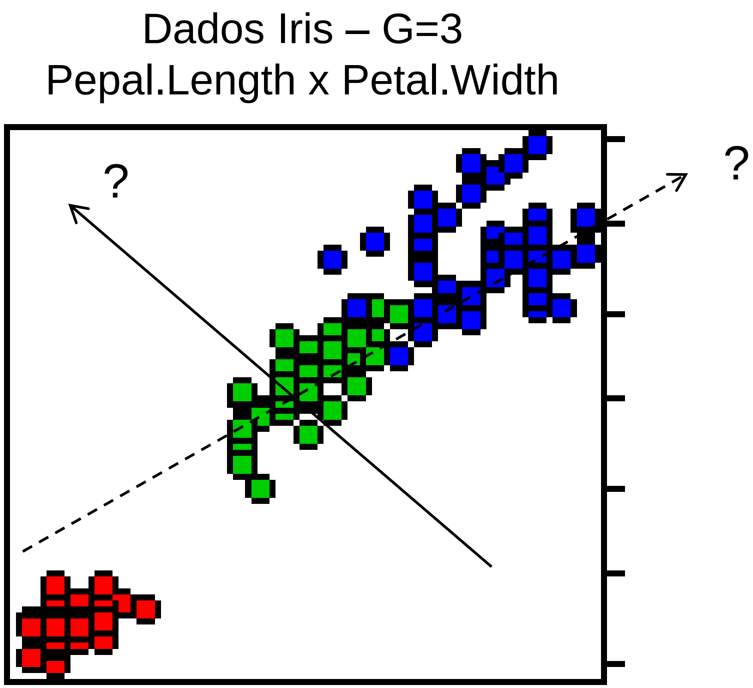
Considerando a dispersão das variáveis:

Sepal.Width x Petal.Width

Petal.Length x Petal.Width

Qual seria uma "direção otima" para discriminar as espécies?

Análise Discriminante



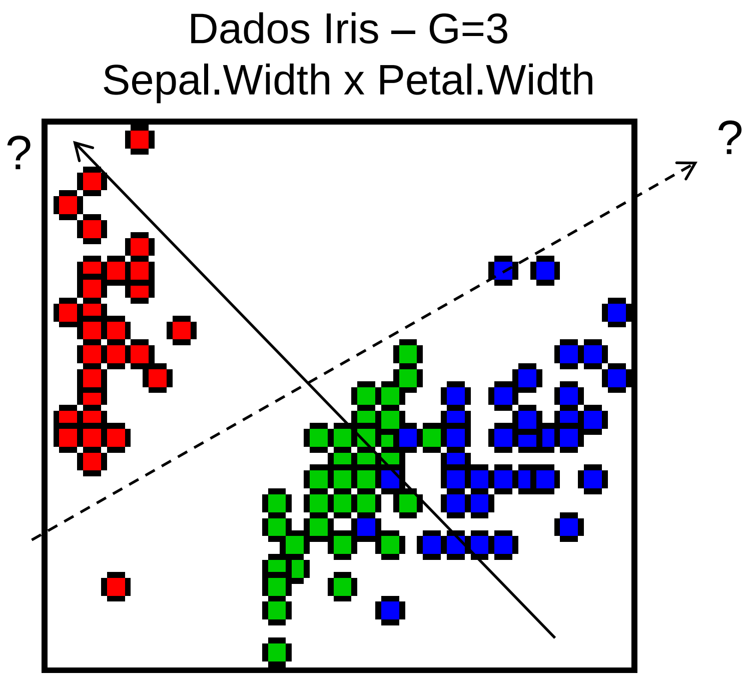
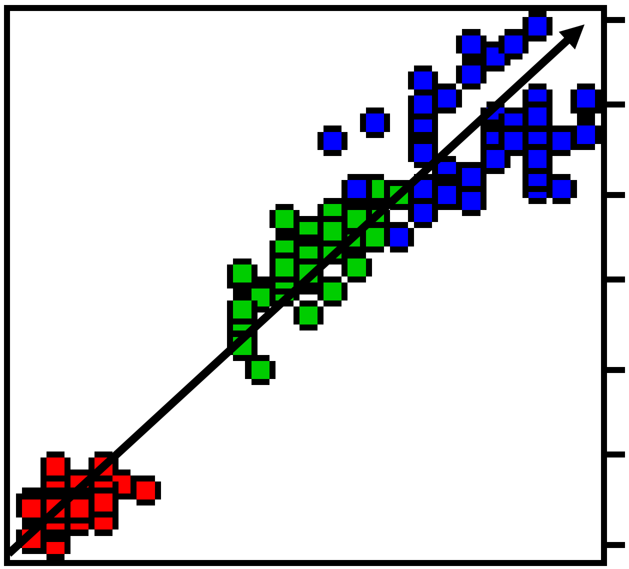


Gráfico de dispersão das observações (em $\Re^2$).

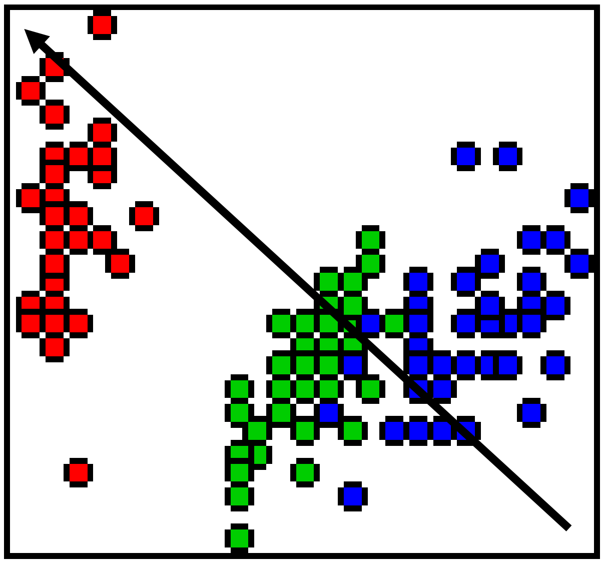
Indique uma direção (uma reta, um terceiro eixo, além de x e y) que defina uma função discriminante linear de separação dos grupos!

Análise Discriminante

Dados Iris – G=3
Pepal.Length x Petal.Width



Dados Iris – G=3
Sepal.Width x Petal.Width



Indicação de um terceiro eixo, $l'Y$ (em preto), que define uma função discriminante linear de separação entre os grupos.

Como obter essa direção discriminante?

$$Y_i \in \Re^p; \quad Y_i = \left(Y_{i1}, Y_{i2}, \ldots, Y_{ip}\right) \quad \rightarrow \quad X = l' Y_i = l_1 Y_{i1} + l_2 Y_{i2} + \ldots + l_p Y_{ip}$$

obter Cargas e os correspondentes Escores

Função Discriminante Linear de Fisher

Formulação de Fisher – Caso de 2 Populações

Considere uma População constituída por observações multivariadas (quantitativas) e estratificadas em dois grupos, tal que:

O grupo ao qual a observação pertence é conhecido.

$$Y_{n\times p}=\begin{bmatrix} Y_{1(n_1\times p)} \\ Y_{2(n_2\times p)} \end{bmatrix} \Rightarrow \begin{cases} Y_i \in \Re^p; \quad E(Y_i \mid \tau_1)=\mu_{1(p\times 1)} \quad Cov(Y_i \mid \tau_1)=\Sigma_{1(p\times p)} \\ \qquad\qquad\quad E(Y_i \mid \tau_2)=\mu_{2(p\times 1)} \quad Cov(Y_i \mid \tau_2)=\Sigma_{2(p\times p)} \end{cases}$$

Suposição $\Rightarrow \quad \Sigma_1=\Sigma_2=\Sigma$ Matrizes de covariâncias homogêneas

Para G=2: A proposta de Fisher é obter combinações lineares, $l'\,Y_i$,das “p” variáveis que maximizem a distância entre os centróides (projetados) dos grupos:

$$Y_i=\left(Y_{1i},Y_{2i},\ldots,Y_{pi}\right) \;\rightarrow\; X_i=l'\,Y_i=\left(l_1Y_{1i}+l_2Y_{2i}+\ldots+l_pY_{pi}\right)$$

$$l=\left(l_1,l_2,\ldots,l_p\right)'\,; \quad \max_l \frac{\left(l'\mu_1-l'\mu_2\right)^2}{l'\Sigma l}$$

Função Discriminante Linear de Fisher

Formulação de Fisher – Caso de 2 Populações

Para G=2: Proposta de Fisher é obter combinações lineares, $l' Y_i$, das "p" variáveis que maximizem a distância entre os centróides (projetados) dos grupos:

$$Y_i = \left(Y_{1i}, Y_{2i}, \ldots, Y_{pi}\right) \quad \rightarrow \quad X_i = l' Y_i = \left(l_1 Y_{1i} + l_2 Y_{2i} + \ldots + l_p Y_{pi}\right)$$

$$l = \left(l_1, l_2, \ldots, l_p\right)'; \quad \max_l \frac{\left(l'\mu_1 - l'\mu_2\right)^2}{l'\Sigma l}$$

$l' Y_i$ Função discriminante

Como obter o vetor de cargas ***l*** ?

Solucão ao problema de otimização:

$$\boxed{l_{p\times 1} = \Sigma^{-1}\left(\mu_1 - \mu_2\right);} \qquad \boxed{X_i = l' Y_i = \left(\mu_1 - \mu_2\right)' \Sigma^{-1} Y_i}$$

$$l = \left(l_1, l_2, \ldots, l_p\right)' \qquad X = \left(X_1, X_2, \ldots, X_n\right)'$$

Cargas (das variáveis) **Escores (dos indivíduos)**

Função Discriminante Linear de Fisher
Estimação

Suposição: Obs. independentes, Matriz de covariâncias homogêneas, prioris dos grupos iguais

$$Y_{n\times p} = \begin{bmatrix} Y_{1(n_1\times p)} \\ Y_{2(n_2\times p)} \end{bmatrix} \Rightarrow Y_i \in \Re^p \to X_i \in \Re; \quad X_i = l'\,Y_i = \left(\mu_1 - \mu_2\right)' \Sigma^{-1} Y_i$$

Função Discriminante calculada para dados amostrais:

$$\hat{\mu}_1,\ \hat{\mu}_2,\ \hat{\Sigma} \quad \downarrow \quad \bar{Y}_1 \quad \bar{Y}_2 \quad S_c$$

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$

Função discriminante

$$S_{c_{p\times p}} = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

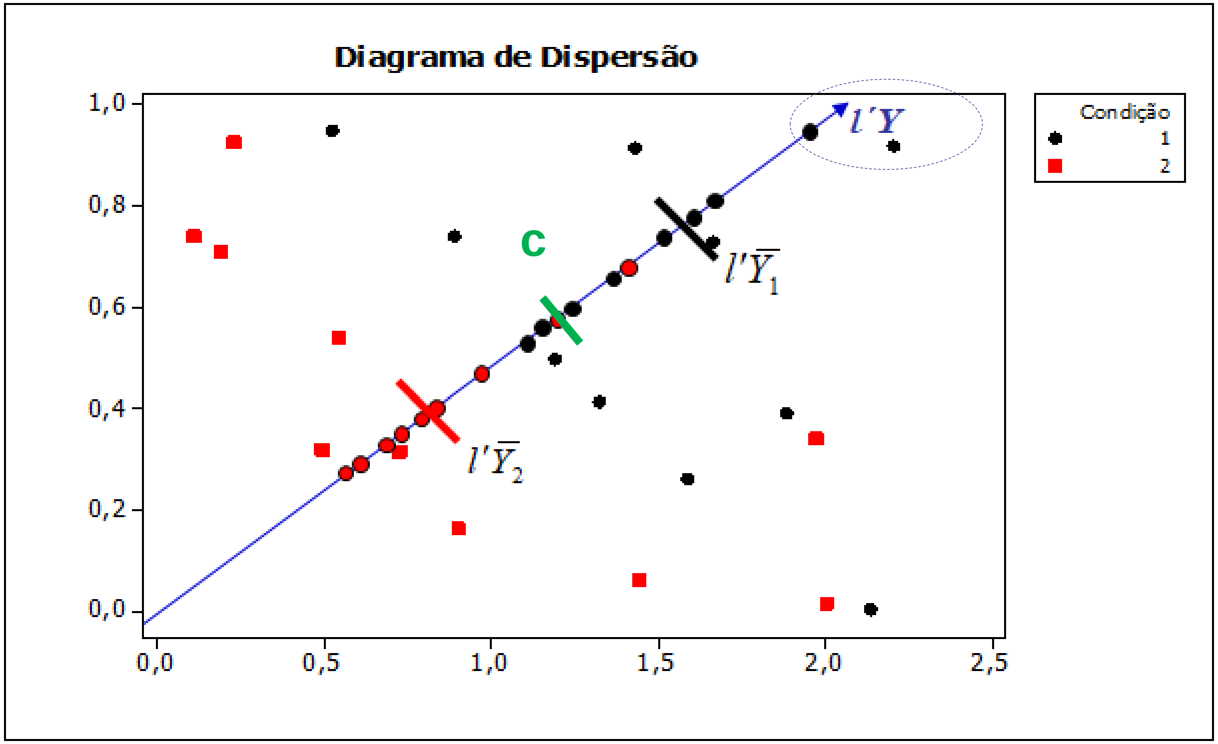
Matriz de covariância comum aos grupos

Regra de Classificação de Observações: Alocar novas observações a um dos Grupos

$$Y_0 = \left(Y_{01}, Y_{02}, \ldots Y_{0p}\right) \,?\, \to \quad X_0 = l'Y_0 \quad \begin{cases} X_0 \geq c \Rightarrow Y_0 \in \tau_1 \\ X_0 < c \Rightarrow Y_0 \in \tau_2 \end{cases} \quad c\,?$$

$$c = \bar{X} = \frac{1}{2}\left(\bar{X}_1 + \bar{X}_2\right) = \frac{1}{2}\left(l'\bar{Y}_1 + l'\bar{Y}_2\right) = \frac{1}{2} l'\left(\bar{Y}_1 + \bar{Y}_2\right) = \frac{1}{2}\left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1}\left(\bar{Y}_1 + \bar{Y}_2\right)$$

Função Discriminante Linear de Fisher
Critério



Dados hipotéticos:
p=2, G=2

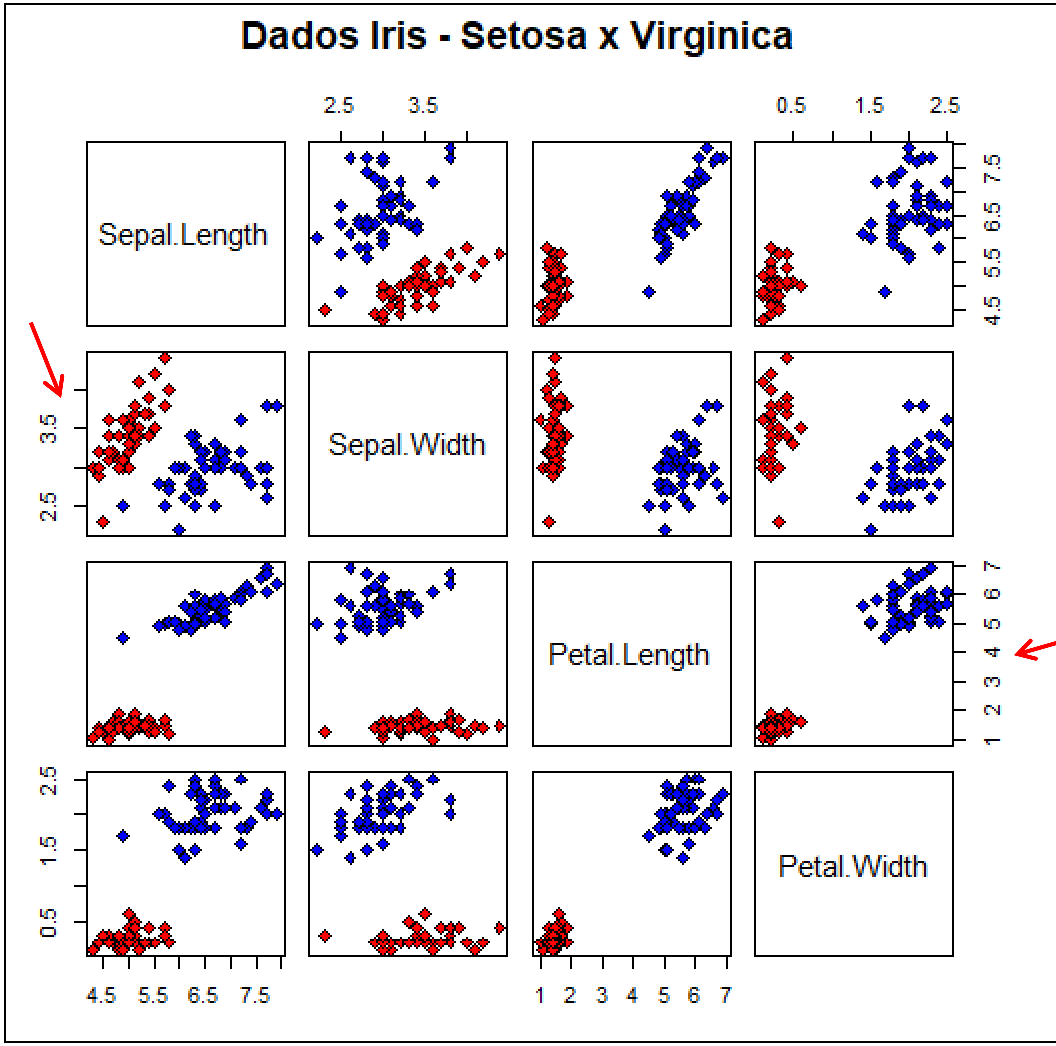
$\boldsymbol{l'Y_i}$

A função discriminante é uma direção na qual os dados serão projetados e os grupos estarão o mais "distante" possível (na comparação entre as medias, em unidades de variância), o que facilitará a predição de uma nova observação.

$\boldsymbol{l'Y_i}$

$l = (l_1, l_2, \ldots, l_p)$ Cargas $\quad Y_0 = (Y_{01}, Y_{02}, \ldots Y_{0p}) \rightarrow X_0 = l'Y_0$ Escores

$$X_0 = l'Y_0 \quad \begin{cases} X_0 \geq c \Rightarrow Y_0 \in \tau_1 \\ X_0 < c \Rightarrow Y_0 \in \tau_2 \end{cases}$$ Regra de Classificação



Obter a Função Discriminante Linear de Fisher nos seguintes casos:

G=2: Setosa x Virginica

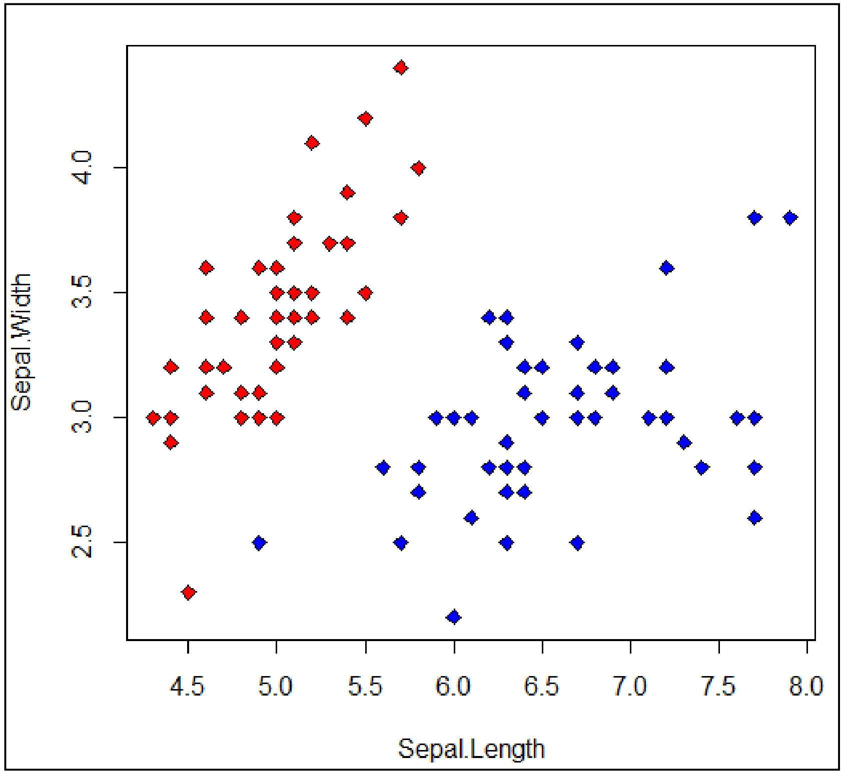
- ✓ p=2: Sepal.Length
  Sepal.Width
- ✓ p=2: Petal.Length
  Petal.Width
- ✓ p=4: Sepal.Length
  Sepal.Width
  Petal.Length
  Petal.Width

Redução de dimensionalidade: O número máximo de funções discriminantes (de Fisher) que podemos obter é mínimo (n, p, G-1)!

Dados Iris, G=2(Setosa x Virginica) n=100

✓ **p=2:** Sepal.Length
Sepal.Width

$$X_i = l'Y_i = \left(\overline{Y}_1 - \overline{Y}_2\right)' S_c^{-1} Y_i$$



Centróides por grupo

	Sepal.Length	Sepal.Width
setosa	5.006	3.428
virginica	6.588	2.974
Dif	**-1.582**	**0.454**

Matriz de Covariância S1: S.setosa

	Sepal.Length	Sepal.Width
Sepal.Length	0.12424898	0.09921633
Sepal.Width	0.09921633	0.14368980

Matriz de Covariância S2: S.virginica

	Sepal.Length	Sepal.Width
Sepal.Length	0.40434286	0.09376327
Sepal.Width	0.09376327	0.10400408

Matriz de Covariância Comum:S.pooled

	Sepal.Length	Sepal.Width
Sepal.Length	**0.2642959**	0.0964898
Sepal.Width	0.0964898	**0.1238469**

Cargas da Função discriminante

	Sepal.Length	Sepal.Width
l'	-10.23536	11.64024

c = -22.07399

Dados Iris, G=2(Setosa[1:50] x Virginica[101:150]) p=2, n=100

$$X_i = l'Y_i = \left(\overline{Y}_1 - \overline{Y}_2\right)' S_c^{-1} Y_i$$

$$X_i = -10.23536 * Sepal.Length_i + 11.640 * Sepal.Width_i \begin{cases} \geq -22.07399 \Rightarrow Setosa \\ < -22.07399 \Rightarrow Virginica \end{cases}$$

	X_i		X_i		X_i		X_i
1	-11.459508	26	-16.256091	101	-26.069989	126	-36.445839
2	-15.232555	27	-11.599996	102	-27.936452	127	-30.866573
3	-10.857435	28	-12.483044	103	-37.750350	128	-27.514989
4	-10.997923	29	-13.647068	104	-30.726085	129	-32.913645
5	-9.271948	30	-10.857435	105	-31.609133	130	-38.773886
6	-9.874021	31	-13.044995	106	-42.868031	131	-43.149006
7	-7.505851	32	-15.694140	107	-21.052674	132	-36.626448
8	-11.599996	33	-5.498901	108	-40.961446	133	-32.913645
9	-11.278898	34	-7.405485	109	-39.476325	134	-31.890109
10	-14.068531	35	-14.068531	110	-31.789743	135	-32.171084
11	-12.202069	36	-13.928043	111	-29.281086	136	-43.891567
12	-9.552923	37	-15.553652	112	-34.077669	137	-24.905966
13	-14.209019	38	-8.248412	113	-34.679742	138	-29.421573
14	-9.091338	39	-10.114874	114	-29.240964	139	-26.491453
15	-12.804141	40	-12.623532	115	-26.772428	140	-34.539254
16	-7.124510	41	-10.435972	116	-28.257549	141	-32.492182
17	-9.874021	42	-19.286577	117	-31.609133	142	-34.539254
18	-11.459508	43	-7.786826	118	-34.579376	143	-27.936452
19	-14.108653	44	-10.435972	119	-48.547663	144	-32.351694
20	-7.967436	45	-7.967436	120	-35.803644	145	-30.164134
21	-15.694140	46	-14.209019	121	-33.375230	146	-33.656206
22	-9.131460	47	-7.967436	122	-24.725356	147	-35.382180
23	-5.177803	48	-9.833899	123	-46.219615	148	-31.609133
24	-13.787556	49	-11.178532	124	-33.054133	149	-23.882429
25	-9.552923	50	-12.764019	125	-30.164134	150	-25.467916

Matriz de classificação
(ou confusão)

	Predito setosa	Predito virginica
setosa	50	0
virginica	1	49

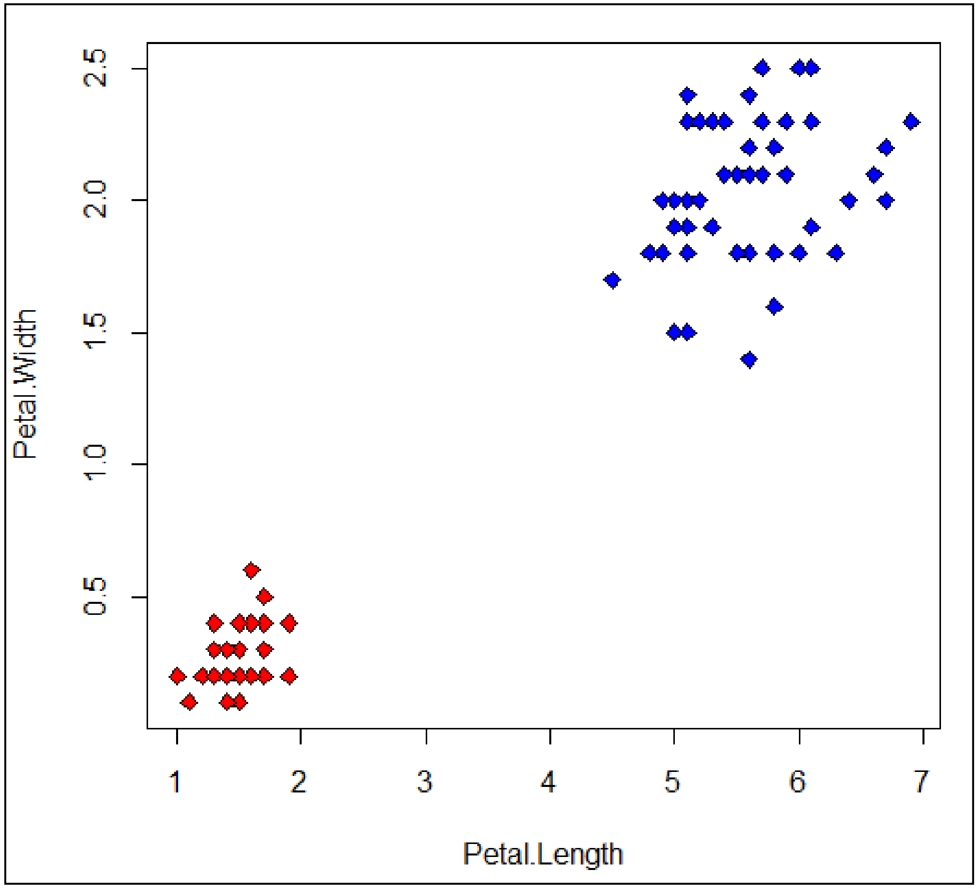
% de classificação correta: 100% Setosa
98% Virginica

Dados Iris, G=2(Setosa x Virginica) n=100

✓ **p=2:** Petal.Length

Petal.Width

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$



Centróides por grupos

	Petal.Length	Petal.Width
setosa	1.462	0.246
virginica	5.552	2.026
Dif	**-4.090**	**-1.78**

S.setosa

	Petal.Length	Petal.Width
Petal.Length	0.030159184	0.006069388
Petal.Width	0.006069388	0.011106122

S.virginica

	Petal.Length	Petal.Width
Petal.Length	0.30458776	0.04882449
Petal.Width	0.04882449	0.07543265

S.pooled

	Petal.Length	Petal.Width
Petal.Length	**0.16737347**	0.02744694
Petal.Width	0.02744694	**0.04326939**

Cargas da Função discriminante

	Petal.Length	Petal.Width
l'	-19.74417	-28.61337

c = -101.7476

Dados Iris, G=2(Setosa[1:50] x Virginica[101:150]) p=2

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$

$$X_i = -19.74417 * Petal.Length_i - 28.6133711.640 * Petal.Width_i \begin{cases} \geq -101.7476 \Rightarrow Setosa \\ < -101.7476 \Rightarrow Virginica \end{cases}$$

	X_i		X_i		X_i		X_i
1	-33.36452	26	-37.31335	101	-189.99845	126	-169.96910
2	-33.36452	27	-43.03602	102	-155.06068	127	-146.27609
3	-31.39010	28	-35.33893	103	-176.57869	128	-148.25051
4	-35.33893	29	-33.36452	104	-162.07143	129	-170.65544
5	-33.36452	30	-37.31335	105	-177.46561	130	-160.29759
6	-45.01044	31	-37.31335	106	-190.39961	131	-174.80485
7	-36.22585	32	-41.06161	107	-137.49150	132	-183.58944
8	-35.33893	33	-32.47760	108	-175.89235	133	-173.51677
9	-33.36452	34	-33.36452	109	-166.02026	134	-143.61533
10	-32.47760	35	-35.33893	110	-191.97287	135	-150.62608
11	-35.33893	36	-29.41568	111	-157.92202	136	-186.25020
12	-37.31335	37	-31.39010	112	-159.00951	137	-179.23945
13	-30.50318	38	-30.50318	113	-168.68102	138	-160.09701
14	-24.57993	39	-31.39010	114	-155.94760	139	-146.27609
15	-29.41568	40	-35.33893	115	-169.36736	140	-166.70660
16	-41.06161	41	-34.25143	116	-170.45486	141	-179.23945
17	-37.11277	42	-34.25143	117	-160.09701	142	-166.50603
18	-36.22585	43	-31.39010	118	-195.23536	143	-155.06068
19	-42.14910	44	-48.75870	119	-202.04554	144	-182.30136
20	-38.20027	45	-48.95927	120	-141.64091	145	-184.07520
21	-39.28777	46	-36.22585	121	-178.35253	146	-168.48044
22	-41.06161	47	-37.31335	122	-153.97318	147	-153.08626
23	-25.46685	48	-33.36452	123	-189.51269	148	-159.89643
24	-47.87178	49	-35.33893	124	-148.25051	149	-172.42928
25	-43.23660	50	-33.36452	125	-172.62986	150	-152.19934

Matriz de classificação
(ou confusão)

	Predito	
	setosa	virginica
setosa	50	0
virginica	0	50

100% de classificação correta!

Dados Iris, G=2 (Setosa x Virginica) p=4

Centróide dos grupos:

	Sepal.L	Sepal.W	Petal.L	Petal.W
setosa	5.006	3.428	1.462	0.246
virginica	6.588	2.974	5.552	2.026

S.comum

0.264	0.096	0.160	0.030
0.096	0.124	0.042	0.028
0.160	0.042	0.167	0.027
0.030	0.028	0.027	0.043

Cargas do discriminante linear

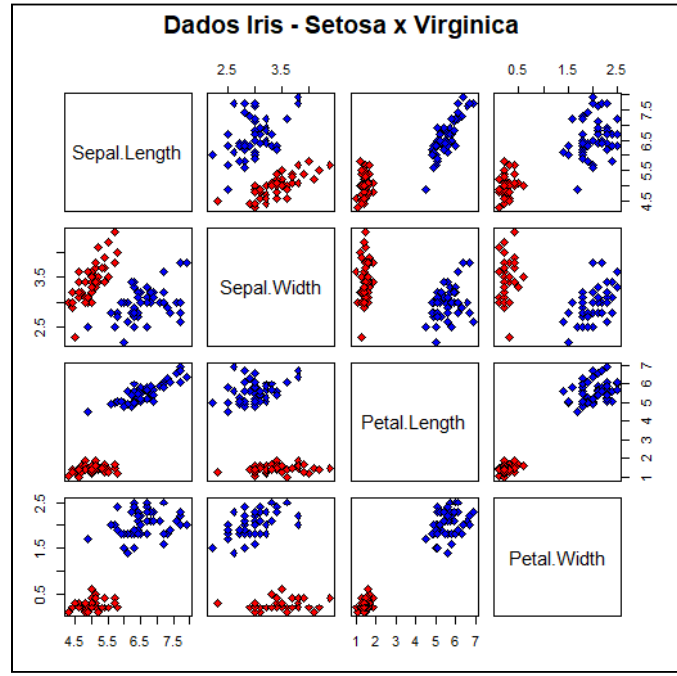
	LD1
Sepal.L	-1.1338828
Sepal.W	-0.8603685
Petal.L	2.6138926
Petal.W	2.6310427

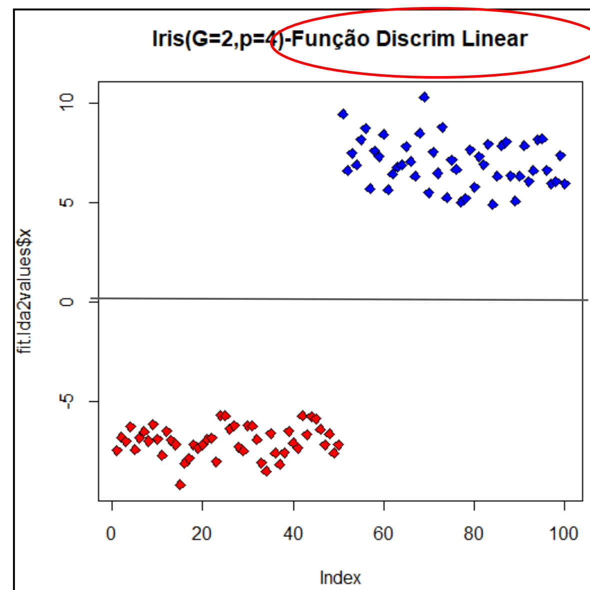
LD1 ≤ 0 Setosa
> 0 Virginica

100% de classificação correta em ambos os grupos!

Escore da função discriminante

	LD1
1	-7.437061
2	-6.780101
3	-6.986787
4	-6.264583
5	-7.409710
6	-6.810997
…	
47	-7.172393
48	-6.612009
49	-7.574522
50	-7.151599
101	9.449657
102	6.601690
103	7.486855
104	6.906517
105	8.168899
106	8.749638
107	5.699715
108	7.602359
…	
148	6.074355
149	7.382464
150	5.967087





Análise Discriminante - Exemplo

	Hemofílicos -			Hemofílicos +	
Grupo	log(AHF ativ)	log(AHF antig)	Grupo	log(AHF ativ)	log(AHF antig)
1	-0,0056	-0,1657	2	-0,3478	0,1151
1	-0,1698	-0,1585	2	-0,3618	-0,2008
1	-0,3496	-0,1879	2	-0,3618	-0,086
1	-0,0894	0,0064	2	-0,4986	-0,2984
1	...	...	2	...	...
1	-0,2228	-0,171	2	-0,1744	0,1892
1	-0,0997	-0,0733	2	-0,4055	-0,2418
1	-0,1972	-0,0607	2	-0,2444	0,1614
1	-0,0867	-0,056	2	-0,4784	0,0282

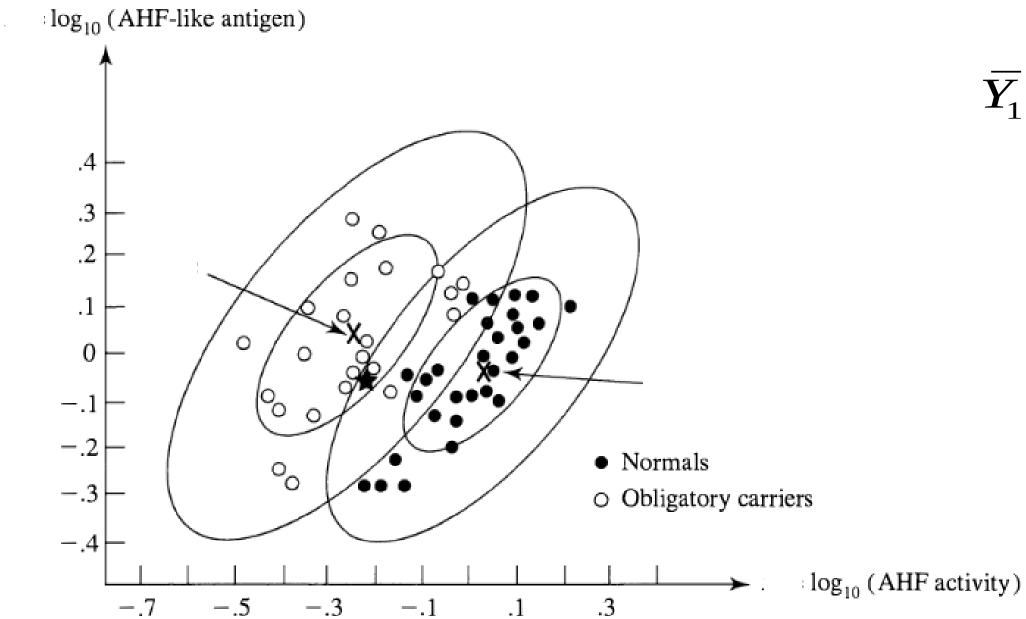
(Johnson and Wichern, 1992, Tabela 11.8)

Objetivo: Obter uma função discriminante para grupos de mulheres que carregam ou não genes da hemofilia com base na atividade da proteína AHF e de seu antígeno.

Com base na função discriminante classifique uma mulher com medidas de AHF iguais a: (-0,11   -0,037)

Análise Discriminante

Exemplo: Dados dos grupos de pacientes Hemofílicos (carregadores do gene) e não Hemofílicos (Normais) .



$$\bar{Y}_1 = \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix} \quad \bar{Y}_2 = \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix}$$

$$S_c^{-1} = \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix}$$

$$l' = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1}$$

$$X = l'Y = 37.61Y_1 - 28.92Y_2$$

$X_i \geq m \Rightarrow$ Normais

$X_i < m \Rightarrow$ Hemof.

$$\bar{X} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} = \begin{pmatrix} l'\bar{Y}_1 \\ l'\bar{Y}_2 \end{pmatrix} = \begin{pmatrix} 0.88 \\ -10.10 \end{pmatrix} \Rightarrow m = \frac{1}{2}\left(\bar{X}_1 + \bar{X}_2\right) = -4.61$$

Análise de Componentes Principais
(Único grupo)

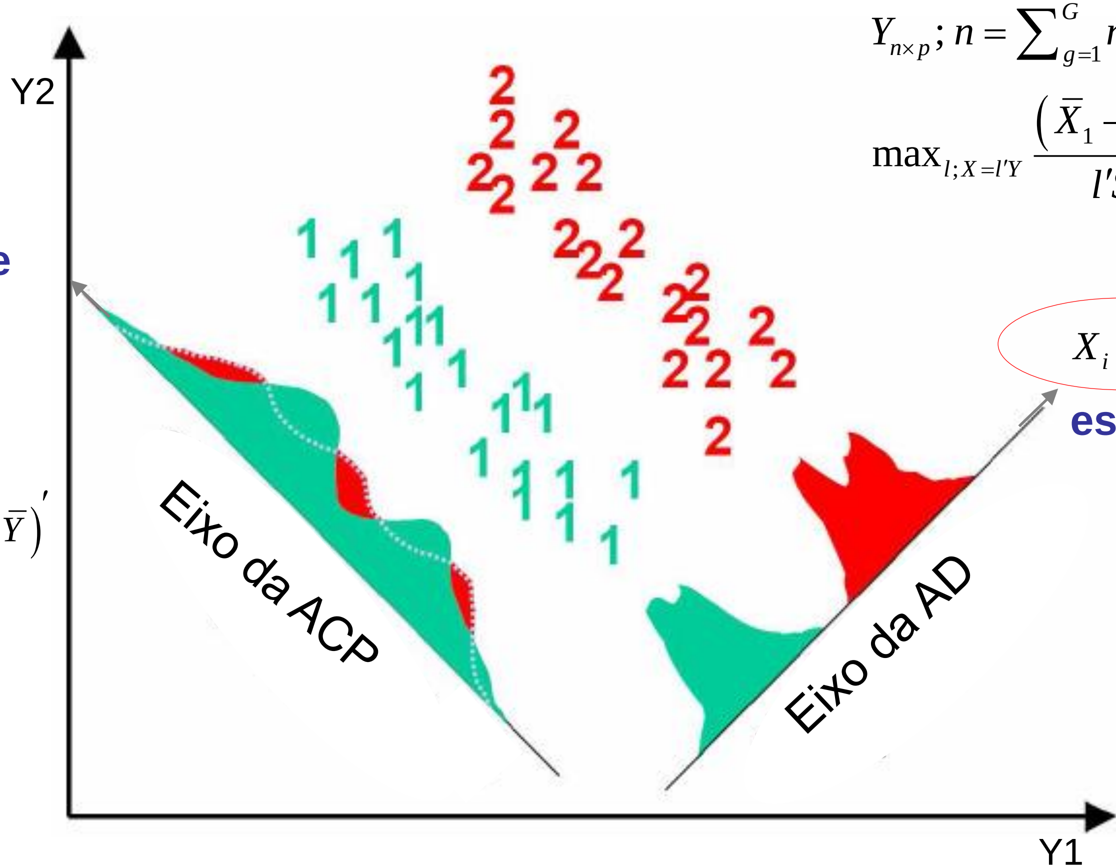
Análise Discriminante
G=2 (ou mais grupos)

$Y_{n\times p} \Rightarrow Z_{n\times m}$

$S = V\Lambda V'$

$Z_{ji} = V_j' Y_i$ **escore**

$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(Y_i - \bar{Y}\right)' = \frac{1}{n-1} S_T$



$Y_{n\times p};\ n = \sum_{g=1}^{G} n_g;\quad Y_{i\ p\times 1} \Rightarrow X_i = l'Y_i$

$\max_{l;X=l'Y} \frac{\left(\bar{X}_1 - \bar{X}_2\right)^2}{l'S_c l}$

$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$

escore

$S_c = \frac{1}{n-G} S_W$

$$\boldsymbol{S_T = S_B + S_W}$$

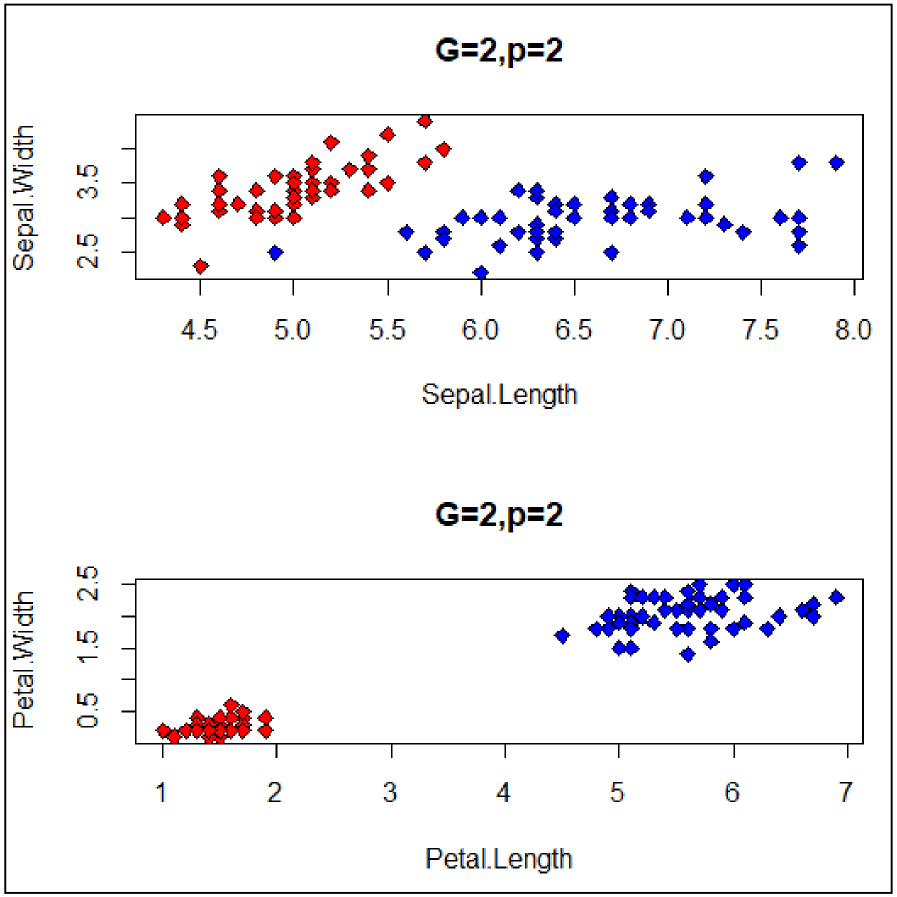
Dados estratificados (em grupos): a variabilidade total ($\boldsymbol{S_T}$) é decomposta nos efeitos Entre ($\boldsymbol{S_B}$) e Dentro ($\boldsymbol{S_W}$) dos grupos

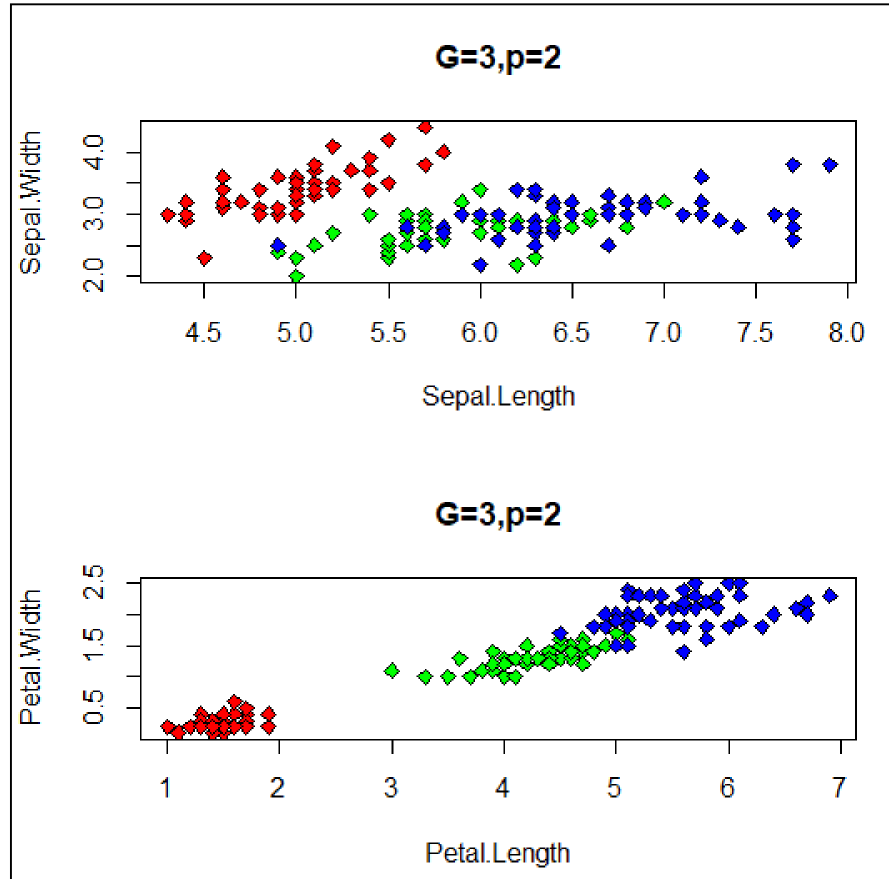
Funções Discriminantes de Fisher

Solução de Fisher: G=2 $\rightarrow$ G>2

Como obter as funções discriminantes?

Número de funções discriminantes: $m \leq \min(n, p, G-1)$





“p” pode ser maior que 2!

Análise Discriminante

Método de Fisher para Muitas Populações

- Re-escrevendo o Critério para Duas Populações:

$$\frac{(\mu_{X1}-\mu_{X2})^2}{\sigma_X^2} = \frac{\sum_{g=1}^{2}\left(\mu_{Xg}-\bar{\mu}_X\right)^2}{\sigma_X^2} = \frac{l'\sum_{g=1}^{2}(\mu_g-\bar{\mu})(\mu_g-\bar{\mu})' l}{l'\Sigma l}$$

- Critério para Muitas Populações (estender o somatório para g grupos):

Maximizar a variabilidade Entre grupos (numerador) relativamente à variabilidade Dentro dos grupos (Σ)

$$\frac{\sum_{g=1}^{G}\left(\mu_{Xg}-\bar{\mu}_X\right)^2}{\sigma_X^2} = \frac{\sum_{g=1}^{G}\left(l'\mu_g - l'\bar{\mu}\right)^2}{\sigma_X^2} = \frac{l'\sum_{g=1}^{G}(\mu_g-\bar{\mu})(\mu_g-\bar{\mu})' l}{l'\Sigma l} = \frac{l'\,B\,l}{l'\Sigma l}$$

Funções Discriminantes de Fisher

Solução de Fisher para Muitas Populações

Para G populações: $\mu_1, \mu_2, \ldots, \mu_G$

$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_G = \Sigma$

Centróides diferentes
Matrizes de covariância homogêneas

$$Y_i \in \Re^p \rightarrow X_i = l'\,Y_i;$$

Funções discriminantes

B: Matriz de covariância ENTRE grupos

$$l; \quad \max_{l;X=l'Y} \frac{\sum_{g=1}^{G} \left(l'\mu_g - l'\bar{\mu}\right)^2}{\sigma_X^2} = \frac{l'\sum_{g=1}^{G}\left(\mu_g - \bar{\mu}\right)\left(\mu_g - \bar{\mu}\right)' l}{l'\Sigma l} = \frac{l'\,B\,l}{l'\Sigma l}$$

Σ : Matriz de covariância DENTRO de grupos

Neste caso, as funções discriminantes, $L_{p\times m} = \left(l_1,\ldots,l_m\right)$, são obtidas a partir dos autovetores da matriz $\Sigma^{-1}\,B$, restritos a $L'\Sigma\,L = I_m$.

Funções Discriminantes de Fisher

Método de Fisher para Muitas Populações

- Dados Amostrais:

Matriz de “Soma de Quadrados e Produtos Cruzados Entre grupos” (SQPC da MANOVA):

$$\hat{B}_{p\times p} = S_B = \sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)'$$

Matriz de “Quadrado Médio e Produtos Cruzados Dentro de grupos” (QMPC da MANOVA):

$$\hat{\Sigma} = S_{c_{p\times p}} = \frac{(n_1 - 1)S_1 + \dots + (n_G - 1)S_G}{n_1 + \dots + n_G - G} = \frac{1}{n-G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)' = \frac{1}{n-G} S_W$$

- Regra de Classificação Amostral:

Alocar a observação Y_0 ($\in \Re^p$) à população τ_k à qual o valor da função discriminante X_0 ($\in \Re^m$) está mais “**próxima**” de seu centróide.

Distância Euclidiana Padronizada

Tabela de ANOVA

$\rightarrow$ MANOVA

Já vimos!

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_G = \mu \in \Re \quad \times \quad H_1 : \exists \text{ pelo menos uma diferença}$$

F.V.	g l	SQ	QM	F	valor-p
ENTRE	G-1	$\sum_{g} n_g (\bar{y}_g - \bar{y})^2$	SQEntre/(G-1)	QME/QMD	$\leftarrow$ Efeito de Grupo
DENTRO	n-G	$\sum_{g,i} (y_{ig} - \bar{y}_g)^2$	SQDentro/(n-G)		$\leftarrow$ Resíduo
TOTAL	n-1	$\sum_{g,i} (y_{ig} - \bar{y})^2$			

$$\mathbf{F} = \frac{QMEntre}{QMDentro} \sim \mathbf{F}_{(G-1,\, n-G)}$$

Sob: $Y_{ig} \overset{iid}{\sim} N_1\left(\mu_g ; \sigma^2\right)$

normalidade
homocedasticidade
independência

Tabela de MANOVA

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_G = \mu$$

Comparar Populações considerando "p" variáveis!

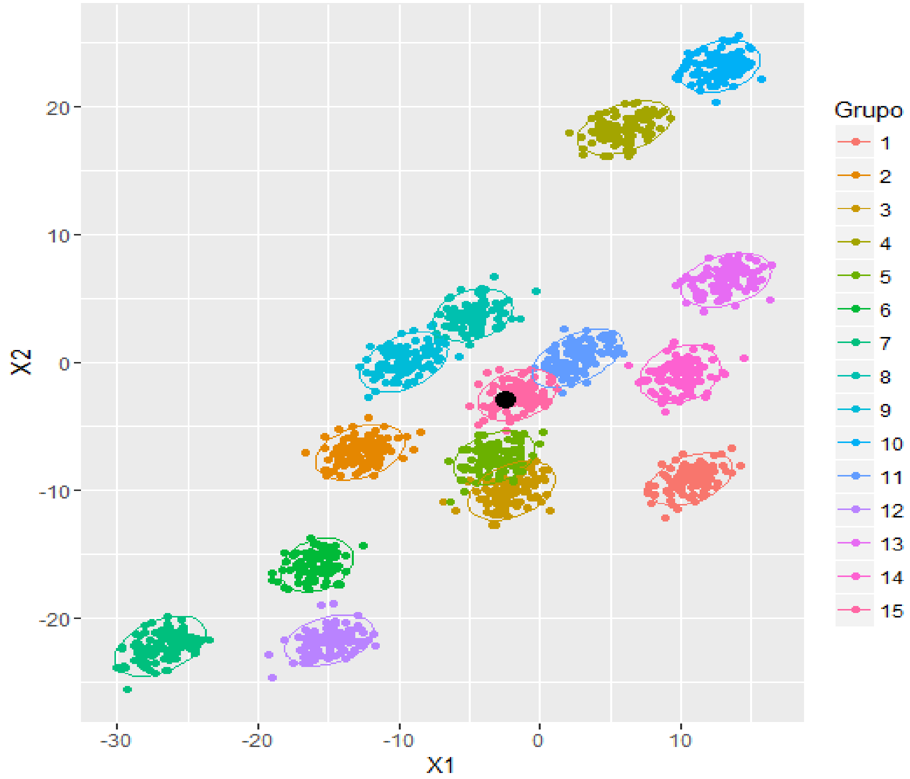
F.V.	g.l.	Matriz de SQPC
Trat	G-1	$B_{p\times p} = \sum_{g=1}^{G} n_g \left(\bar{y}_g - \bar{y}\right)\left(\bar{y}_g - \bar{y}\right)'$ Matriz de Soma de Quadrados e produtos cruzados ENTRE Grupos
Resíduo	n-G	Matriz de Soma de Quadrados e produtos cruzados DENTRO de Grupos (comum) $S_{W\,p\times p} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(y_{ig} - \bar{y}_g\right)\left(y_{ig} - \bar{y}_g\right)'$ $\Rightarrow$ **Sc = Sw/(n-G)**
TOTAL	n-1	$T = \sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)'$

$$\Lambda^* = \frac{|\,S_c\,|}{|\,B + S_c\,|} = \left|I + S_c^{-1}B\right|^{-1}$$

Estatística lambda de Wilks para testar H_0 (sob normalidade multivariada, independência e homocedasticidade)

Funções Discriminantes de Fisher

Populações Estratificadas em Muitos Grupos (G>2)



$$m \le \min\left(n, p, G-1\right)$$

Como realizar a redução dos dados (**p=2, G=15**)? Uma única dimensão ($X = l'Y$) é suficiente para a discriminação dos grupos?

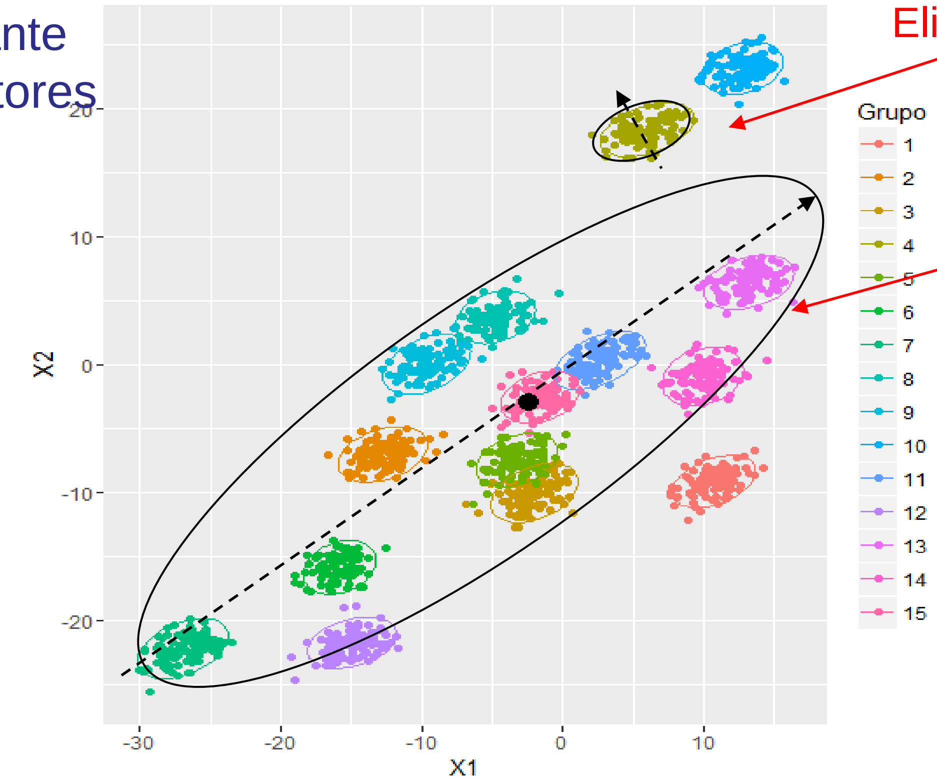
Função Discriminante de Fisher

Populações Estratificadas em Muitos Grupos (G>2)
Entendendo a direção discriminante

A direção discriminante é obtida dos autovetores da matriz $S_c^{-1}\, B$

que equivale a:

$$\max_l \frac{l'\,\hat{B}\,l}{l'\,S_c\,l}$$



Elipse menor: Variabilidade Dentro de grupos (S_c)

Elipse maior: Variabilidade Entre grupos (B)

A direção discriminante ótima é aquela que maximiza B (eixo de variação ENTRE grupos) relativamente a Σ (eixo de variação DENTRO de grupos).

Alocar a observação Y_0 ($\in \Re^p$) à população τ_k em que o valor da função discriminante X_0 ($\in \Re^m$) está mais "próxima" de seu centróide

Dados Iris: G=3 e p=4

Probabilidades a priori dos grupos

setosa	versicolor	virginica
0.3333333	0.3333333	0.3333333

Centróides dos grupos

	Sepal.L	Sepal.W	Petal.L	Petal.W
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

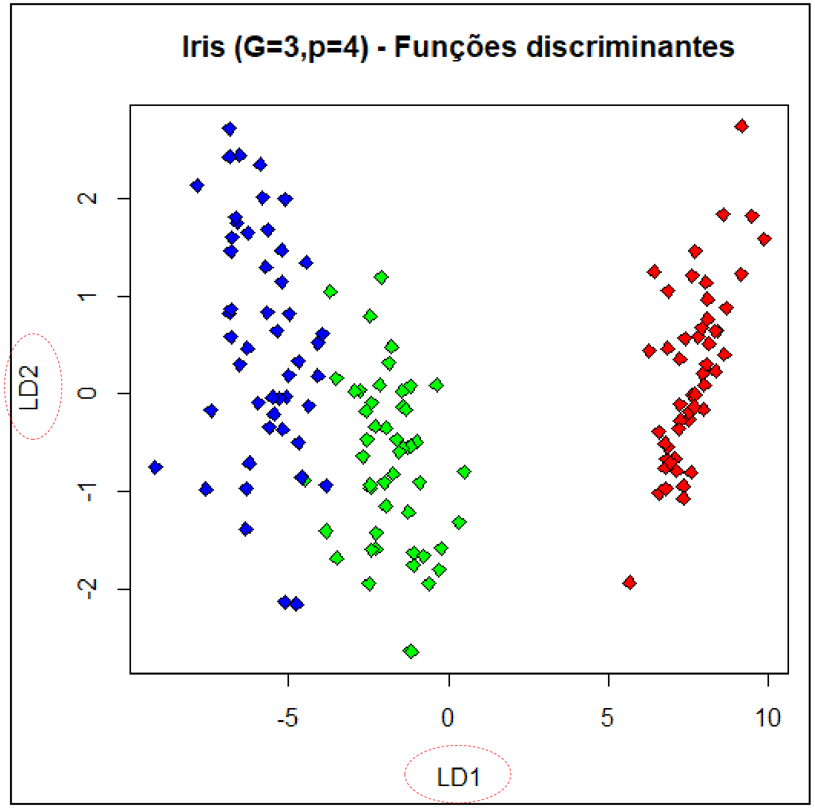
Cargas das Funções discriminantes

	LD1	LD2
Sepal.L	0.8293776	0.02410215
Sepal.W	1.5344731	2.16452123
Petal.L	-2.2012117	-0.93192121
Petal.W	-2.8104603	2.83918785

Matriz de Classificação

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

setosa	versicolor	virginica
1.00	0.96	0.98



Redução de dimensionalidade em
Análise discriminante: m=min(n,p,G-1)=2

Centróide do espaço discriminante

		LD1	LD2
1	setosa	7.607600	0.2151330
2	versicolor	-1.825049	-0.7278996
3	virginica	-5.782550	0.5127666

X_0 é classificada no grupo ao qual possui menor distância (Euclidiana Padronizada) ao centróide

Análise Discriminante

Validação Empírica de um Algoritmo de Classificação Amostral

Métricas de validação via a Matriz de Classificação

Matriz de Classificação (ou de Confusão)

Verdade	Predito τ_1 **+**	τ_2 **-**	
τ_1 **+**	n_{1c} **V+**	n_{1M} **F-**	n_1
τ_2 **-**	n_{2M} **F+**	n_{2c} **V-**	n_2

- Taxa de Erro Aparente (proporção de itens mal classificados):

$$TxErro = \frac{n_{1M} + n_{2M}}{n_1 + n_2} = \frac{F_+ + F_-}{n}$$ Estima Pr(classificação errada)

- Acurácia: $$Acurácia = \frac{n_{1C} + n_{2C}}{n_1 + n_2} = \frac{V_+ + V_-}{n}$$ Estima Pr(classificação correta)

Métricas de Validação via a Matriz de Classificação

Matriz de Classificação (ou de Confusão)

Verdade	Predito τ_1 **+**	τ_2 **-**	
τ_1 **+**	n_{1c} **V+**	n_{1M} **F-**	n_1
τ_2 **-**	n_{2M} **F+**	n_{2c} **V-**	n_2

- Sensibilidade = $\dfrac{V_+}{V_+ + F_-}$ = Pr(classificação + | +)
- Especificidade = $\dfrac{V_-}{F_+ + V_-}$ = Pr(classificação - | -)

Poder Preditivo via a Curva ROC:
Sensibilidade x (1-Especificidade)

- Preditivo Positivo = $\dfrac{V_+}{F_+ + V_+}$

Precisão do classificador

- Preditivo Negativo = $\dfrac{V_-}{F_- + V_-}$

- Escore F1 = $2\dfrac{\Pr ecisão * Sensibilidade}{\Pr ecisão + Sensibilidade}$

Média harmônica da precisão e sensibilidade

Análise Discriminante

Validação de um Algoritmo de Classificação

Matriz de Classificação (ou de Confusão)

Verdade	Predito τ_1	τ_2	
τ_1	n_{1c} **V+**	n_{1M} **F+**	n_1
τ_2	n_{2M} **F-**	n_{2c} **V-**	n_2

TxErro subestima a Probabilidade de erro de classificação (populacional): os mesmos dados são usados para Treinamento e Teste do algoritmo

Alternativas

- Método de Particionamento (*Data Split*): particiona os dados em Amostra de Treinamento e Amostra de Validação (Teste)
- Método de "Validação Cruzada" (Cross-validation)

Análise Discriminante

Validação de um Algoritmo de Classificação Amostral

Validação Cruzada pelo método Leave-One-Out (*Fold*=N)

1. Inicie com as observações de τ_1 . Omita uma obs deste grupo e obtenha a função de classificação baseada nos remanescentes N-1=(n1-1)+n2 observações (supondo G=2)
2. Classifique a obs omitida usando a função calculada no passo 1
3. Repetir os passos 1 e 2 até que todas as obs de τ_1 tenham sido classificadas. Calcule o número de erros de classificação neste grupo
4. Repita os passos de 1 a 3 para as observações do grupo 2.

Taxa de Erro de Classificação esperada é dada por:

$$TxErro = \frac{n_{1M}^{Cross} + n_{2M}^{Cross}}{n_1 + n_2}$$

Algoritmos de CV podem usar *Fold*=k

Análise Discriminante
Normalização de Variáveis

	Unidades Amostrais	Variáveis 1	2	…	j	…	p
G1	1	Y_{111}	Y_{112}		Y_{11j}		Y_{11p}
	2	Y_{121}	Y_{122}		Y_{12j}		Y_{12p}
	…	…	…	…	…		…
	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1j}		Y_{1n1p}
G2	1	Y_{211}	Y_{212}		Y_{21j}		Y_{21p}
	2	Y_{221}	Y_{222}		Y_{22j}		Y_{22p}
	…	…	…	…	…		…
	n_2	Y_{2n21}	Y_{2n22}		Y_{2n2j}		Y_{2n2p}

$\bar{Y}_{1\,p\times 1}$ $S_{1\,p\times p}$

$\bar{Y}_{2\,p\times 1}$ $S_{2\,p\times p}$

$\bar{Y}_{p\times 1}$ $S_{c\,p\times p}$

Na AD a normalização das variáveis é usada com a finalidade de facilitar a interpretação das cargas das variáveis na função discriminante e no cálculo de "*c*". O comando "lda" do R adota a "normalização" das variáveis para calcular as funções discriminantes, mas o "linda" não. A normalização da variável j avaliada no indivíduo i do grupo g é dada por:

$$Y_{gij}^{*} = \left(\frac{Y_{gij} - \bar{Y}_j}{S_{gj}}\right)$$

Média: para cada j independente de grupo

Variância: para cada grupo g e variável j

$$\bar{Y}_j = \frac{1}{n_1 + n_2}\sum_{g=1}^{2}\sum_{i=1}^{n_g} Y_{gij}$$

Média da variável j (j=1,..,p), independente de grupo

$$S_{gj} = \frac{1}{n_g - 1}\sum_{g=1}^{2}\left(Y_{gij} - \bar{Y}_{gj}\right)^2$$

Variância da variável j no grupo g

Análise Discriminante

✓ Regra Discriminante Linear de Fisher

- Métodos Probabilísticos de Análise Discriminante:

    Regra de Classificação de Bayes

    Regressão Logística (já vimos)

- MANOVA: estimar as matrizes de covariância ENTRE (S_B) e DENTRO (S_W) de grupos sob diferentes modelos (ajuste por covariáveis)

MANOVA
Uma Aplicação

MANOVA: Delineamento Completamente Aleatorizado com Um Único Fator

	T6	T12	T18	Trat
1	1.48	2.81	3.56	1
2	1.04	2.07	2.81	2
3	1.48	2.52	3.41	1
4	1.04	1.93	2.89	2
5	1.80	2.15	3.20	1
6	1.50	2.70	3.75	2
…				
23	1.80	2.15	3.90	1
24	1.20	2.25	3.30	2
25	1.78	2.96	4.00	3
26	1.48	2.81	3.85	4
27	1.33	2.52	3.84	3
28	1.03	2.07	2.96	4
29	1.65	3.00	3.98	3
30	1.50	2.85	3.75	4
…				
45	1.65	3.00	4.05	3
46	1.20	2.70	3.90	4
47	1.35	2.55	3.67	3
48	1.20	2.70	3.60	4

MANOVA.RM do R

Considere os seguintes dados de um Delineamento Completmente Aleatorizado com 1 Fator Tratamento em 4 níveis (Trat=1, Trat=2, Trat=3 e Trat=4)

p=3 variáveis (medidas repetidas de O2): T6, T12 e T18

$$Y_{48\times(3+1)}$$

3 variáveis resposta quantitativas e 1 categórica (identificando grupo)

Centróides

	T6	T12	T18
Total(n=48)	1.497500	2.558333	3.664375

Trat	T6	T12	T18
1 (n=9)	1.618333	2.434167	3.526667
2 (n=9)	1.321667	2.430000	3.425000
3 (n=9)	1.655833	2.799167	4.029167
4 (n=9)	1.394167	2.570000	3.676667

Matrizes de Covariância

	T6	T12	T18	Trat
1	1.48	2.81	3.56	1
2	1.04	2.07	2.81	2
3	1.48	2.52	3.41	1
4	1.04	1.93	2.89	2
5	1.80	2.15	3.20	1
6	1.50	2.70	3.75	2
…				
23	1.80	2.15	3.90	1
24	1.20	2.25	3.30	2
25	1.78	2.96	4.00	3
26	1.48	2.81	3.85	4
27	1.33	2.52	3.84	3
28	1.03	2.07	2.96	4
29	1.65	3.00	3.98	3
30	1.50	2.85	3.75	4
…				
45	1.65	3.00	4.05	3
46	1.20	2.70	3.90	4
47	1.35	2.55	3.67	3
48	1.20	2.70	3.60	4

Trat=1	T6	T12	T18
T6	0.02	-0.02	0.00
T12	-0.02	0.09	0.00
T18	0.00	0.00	0.08

Trat=2	T6	T12	T18
T6	0.04	0.04	0.04
T12	0.04	0.07	0.06
T18	0.04	0.06	0.11

Trat=3	T6	T12	T18
T6	0.04	0.01	0.05
T12	0.01	0.11	0.01
T18	0.05	0.01	0.07

Trat=4	T6	T12	T18
T6	0.05	0.02	0.06
T12	0.02	0.06	0.05
T18	0.06	0.05	0.12

Sc	T6	T12	T18
T6	0.04	0.01	0.03
T12	0.01	0.08	0.03
T18	0.03	0.03	0.09

Teste M de Box: teste de homocedasticidade multivariado (testa a homogeneidade das matrizes de covariância)

Chi-Sq=34.61, df = 18, p-value = 0.01058

α=1% $\Rightarrow$ Não há evidência para a rejeição da hipótese de Homocedasticidadeâ

	T6	T12	T18	Trat
1	1.48	2.81	3.56	1
2	1.04	2.07	2.81	2
3	1.48	2.52	3.41	1
4	1.04	1.93	2.89	2
5	1.80	2.15	3.20	1
6	1.50	2.70	3.75	2
…				
23	1.80	2.15	3.90	1
24	1.20	2.25	3.30	2
25	1.78	2.96	4.00	3
26	1.48	2.81	3.85	4
27	1.33	2.52	3.84	3
28	1.03	2.07	2.96	4
29	1.65	3.00	3.98	3
30	1.50	2.85	3.75	4
…				
45	1.65	3.00	4.05	3
46	1.20	2.70	3.90	4
47	1.35	2.55	3.67	3
48	1.20	2.70	3.60	4

$$H_0 : \mu_g = \mu_{3\times1}, \quad g = 1,\dots,4$$

Atenção à decomposição imposta à Matriz de Covariância

$$\mathbf{T} = \mathbf{B} + \mathbf{S_W}$$

Tabela de MANOVA:

FV	no.gl	SQPC
Trat	4-1	$\begin{array}{c|ccc} \textbf{SSB} & T6 & T12 & T18 \\ T6 & \mathbf{0.98} & 0.53 & 0.98 \\ T12 & 0.53 & \mathbf{1.08} & 1.63 \\ T18 & 0.98 & 1.63 & \mathbf{2.51} \end{array}$
Resíduo	48-4	$\begin{array}{c|ccc} \textbf{SSW} & T6 & T12 & T18 \\ T6 & \mathbf{1.67} & 0.55 & 1.53 \\ T12 & 0.55 & \mathbf{3.66} & 1.24 \\ T18 & 1.53 & 1.24 & \mathbf{4.15} \end{array}$
Total	48-1	$\begin{array}{c|ccc} \textbf{SST} & T6 & T12 & T18 \\ T6 & 2.65 & 1.08 & 2.51 \\ T12 & 1.08 & 4.74 & 2.87 \\ T18 & 2.51 & 2.87 & 6.67 \end{array}$

Fonte de Variação ENTRE grupos

Fonte de Variação DENTRO de grupos

	Estat.	approxF	numDf	denDf	Pr(>F)	
Pillai	0.7651	5.0206	9	132	8.062e-06	***
Wilks	0.3807	5.5401	9	102.37	3.354e-06	***
Hotel.-Lawley	1.2444	5.6229	9	122	1.742e-06	***
Roy	0.7004	10.273	3	44	3.013e-05	***

Concl.?

	T6	T12	T18	Trat
1	1.48	2.81	3.56	1
2	1.04	2.07	2.81	2
3	1.48	2.52	3.41	1
4	1.04	1.93	2.89	2
5	1.80	2.15	3.20	1
6	1.50	2.70	3.75	2
…				
23	1.80	2.15	3.90	1
24	1.20	2.25	3.30	2
25	1.78	2.96	4.00	3
26	1.48	2.81	3.85	4
27	1.33	2.52	3.84	3
28	1.03	2.07	2.96	4
29	1.65	3.00	3.98	3
30	1.50	2.85	3.75	4
…				
45	1.65	3.00	4.05	3
46	1.20	2.70	3.90	4
47	1.35	2.55	3.67	3
48	1.20	2.70	3.60	4

Decomposição Espectral de $Sc^{-1}B$:

Autovalores

[,1]	[,2]	[,3]
0.7004284	0.5425676	0.001408624

Atovetores

	[,1]	[,2]	[,3]
T6	0.01632176	0.94481561	-0.08938875
T12	-0.20483386	-0.07142574	-0.50725864
T18	-0.40074216	-0.30089017	0.36425094

Contribuição das variáveis para a discriminação dos Trats!

Modelo estrutural e distribucional adotado:

$$Y_{ig\,3\times1} = \mu_g + e_{ig}; \quad e_{ig} \sim N_3\left(\mu_g; \Sigma\right)$$

$$= \mu + \tau_g + e_{ig}; \quad \sum_{g=1}^{4} \tau_g = 0$$

Parametrização de desvios

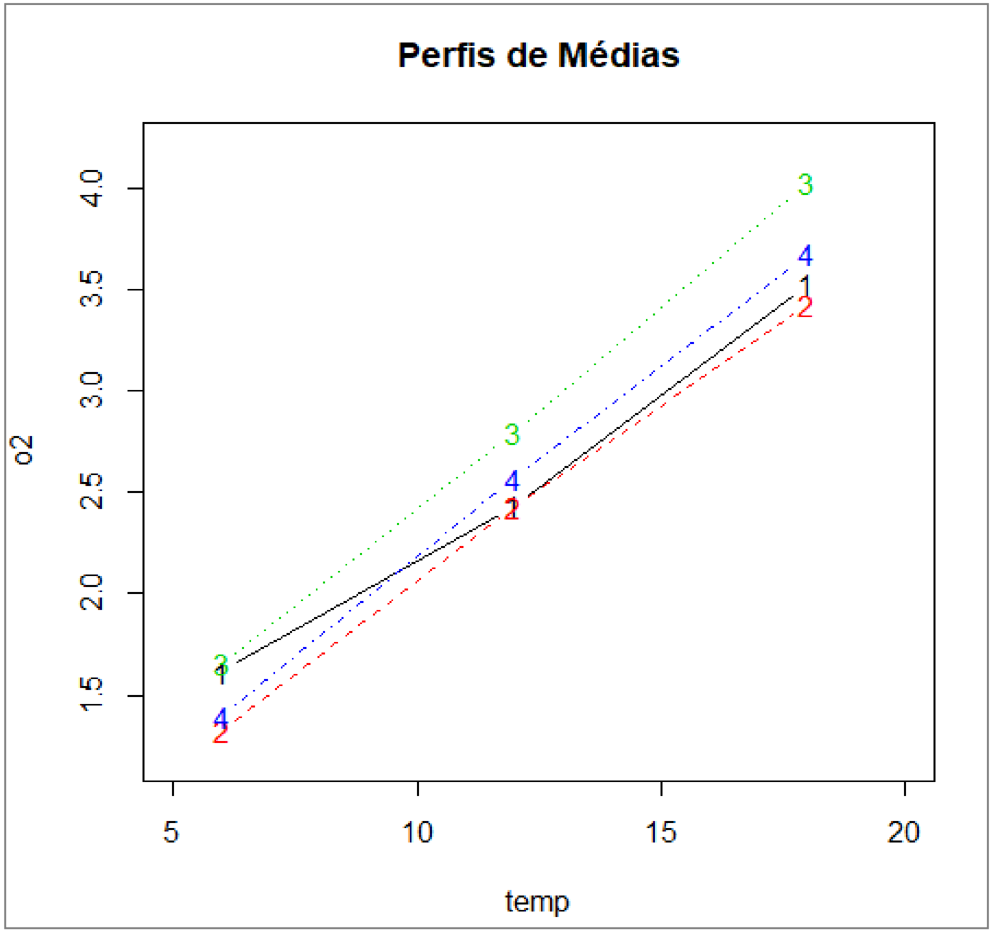
$$= \begin{cases} \mu_1 + e_{i1} \\ \mu_1 + \tau_g + e_{ig}; \quad g = 2,3,4 \end{cases}$$

Parametrização casela de referência

Estimativas:	T6	T12	T18	
(Intercept)	1.6183333	2.434166667	3.5266667	$\rightarrow \hat{\mu}_1$
Trat=2	-0.2966667	-0.004166667	-0.1016667	
Trat=3	0.0375000	0.365000000	0.5025000	$\rightarrow \hat{\tau}_g; \quad g = 2,3,4$
Trat=4	-0.2241667	0.135833333	0.1500000	

Comparações Múltiplas

Intervalos de Confiança de Bonferroni (correção por variável)



		Li	Ls	Conclusão
T6	**[2-1]**	-0.52	-0.08	$\mu_{21} < \mu_{11}$
	[3-1]	-0.18	0.26	$\mu_{31} = \mu_{11}$
	[4-1]	-0.44	0.00	$\mu_{41} = \mu_{11}$
	[3-2]	0.11	0.55	$\mu_{31} > \mu_{21}$
	[4-2]	-0.15	0.29	$\mu_{41} = \mu_{21}$
	[4-3]	-0.48	-0.04	$\mu_{41} < \mu_{31}$
T12	**[2-1]**	-0.33	0.32	$\mu_{22} = \mu_{12}$
	[3-1]	0.04	0.69	$\mu_{32} > \mu_{12}$
	[4-1]	-0.19	0.46	$\mu_{42} = \mu_{12}$
	[3-2]	0.04	0.69	$\mu_{32} > \mu_{22}$
	[4-2]	-0.19	0.47	$\mu_{42} = \mu_{22}$
	[4-3]	-0.55	0.10	$\mu_{42} = \mu_{32}$
T18	**[2-1]**	-0.45	0.24	$\mu_{23} = \mu_{13}$
	[3-1]	0.16	0.85	$\mu_{33} > \mu_{13}$
	[4-1]	-0.20	0.50	$\mu_{43} = \mu_{13}$
	[3-2]	0.26	0.95	$\mu_{33} > \mu_{23}$
	[4-2]	-0.09	0.60	$\mu_{43} = \mu_{23}$
	[4-3]	-0.70	-0.01	$\mu_{43} < \mu_{33}$

$$ICB\left(\mu_{gj}-\mu_{hj}\right)\ a\ (1-\alpha)100\% = \left(\bar{Y}_{g\,j}-\bar{Y}_{h\,j}\right) \pm t_{n-G}\left(\alpha/2K\right)\sqrt{V\left(\bar{Y}_{g\,j}-\bar{Y}_{h\,j}\right)}$$

$$\left(\frac{1}{n_g}+\frac{1}{n_h}\right)\frac{E_{jj}}{n-G}$$