

# Análise Discriminante: parte 1

Prof. Caio Azevedo

# Motivação

- Com o advento (ressurgimento) das áreas de **Data Sciences** e **Big data**, tem aumentado o interesse na construção/utilização de regras de classificação apropriadas.
- **Regras de classificação** são de fundamental importância na **tomada de decisões**.
- Metodologias de **aprendizado** (supervisionado e não supervisionado) de máquina, **aprendizado estatístico** etc, muitas vezes, correspondem a algum método de classificação.
- Existem diversos métodos de classificação: **regressão**, **árvores de decisão**, **análise discriminante**, dentro outros.

# Motivação

- Dispõe-se de uma matriz de dados com várias informações (variáveis resposta e/ou explicativas).
- Entre essas informações, há (pelo menos uma) variável referente à grupos de interesse.
- Exemplo 1: Íris de Fisher (tipo de íris), Exemplo 8: dados sobre cereais (fabricantes).
- No Exemplo 1 temos quatro medidas morfológicas das plantas bem como a espécie à qual cada uma pertence (grupo).
- Em geral usaremos os termos “grupo” e “população” indistintamente, a menos que o contrário seja mencionado.

# Motivação

- A classificação de cada unidade (amostral/experimental) feita originalmente (agrupamento), em geral, foi obtida através de algum método: caro e/ou invasivo e/ou muito custoso (financeiramente, temporalmente etc) e/ou que requer que a unidade experimental seja destruída porém, é (muito) confiável.
- Uma (outra) forma estatística de se criar grupos é através de métodos de [análise de conglomerados](#).

# Motivação

- Algumas vezes, esse método de agrupamento só pode ser utilizado em circunstâncias muito específicas (paleontologia e arqueologia).
- Eventualmente (embora não seja usual) algumas das variáveis disponíveis podem ter sido utilizadas na classificação inicial.
- O objetivo é criar uma regra de classificação estatística utilizando as variáveis disponíveis no banco de dados e a classificação anteriormente feita.

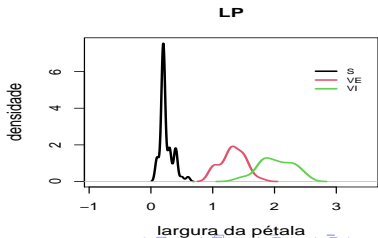
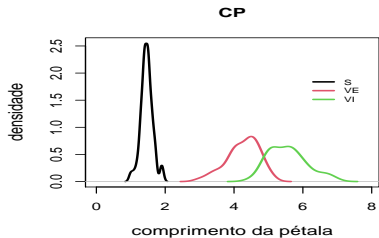
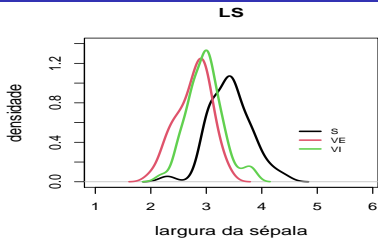
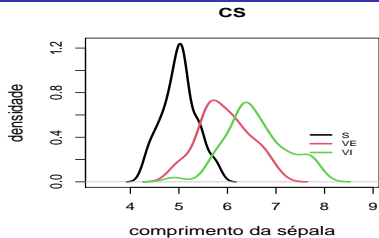
# Motivação

- Tal procedimento consiste em identificar padrões de “comportamento” para cada grupo em relação às variáveis disponíveis.
- Se este procedimento for satisfatório, em termos de classificação, ele pode ser usado em futuros estudos, no lugar do método utilizado inicialmente.
- Em princípio, quanto mais diferentes forem os grupos entre si, com relação às variáveis disponíveis, melhor será a regra de classificação.

# Análise Discriminante (AD) para duas populações

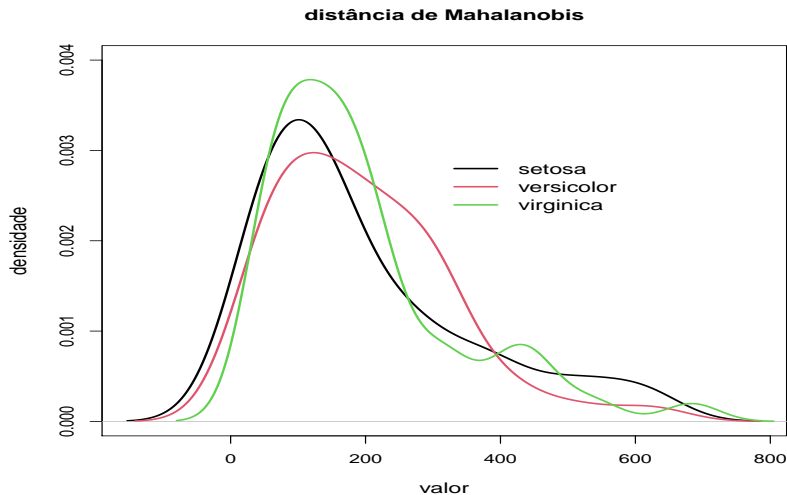
- Suponha que tenhamos duas populações das quais extraímos duas amostras aleatórias (independentes entre populações e entre indivíduos) e delas medimos  $p$  características.
- Cada uma dessas populações (processo de amostragem/experimentação) pode ser representado por uma fdp (discreta, contínua ou mista)  $f_i, i = 1, 2$ .
- Em geral assume-se o mesmo modelo probabilístico (fdp) entre as populações diferindo somente em termos de seus parâmetros. Por exemplo  $f_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, 2$  (normal multivariada).

# Densidades para cada variável em função dos grupos





# Densidade de Mahalanobis



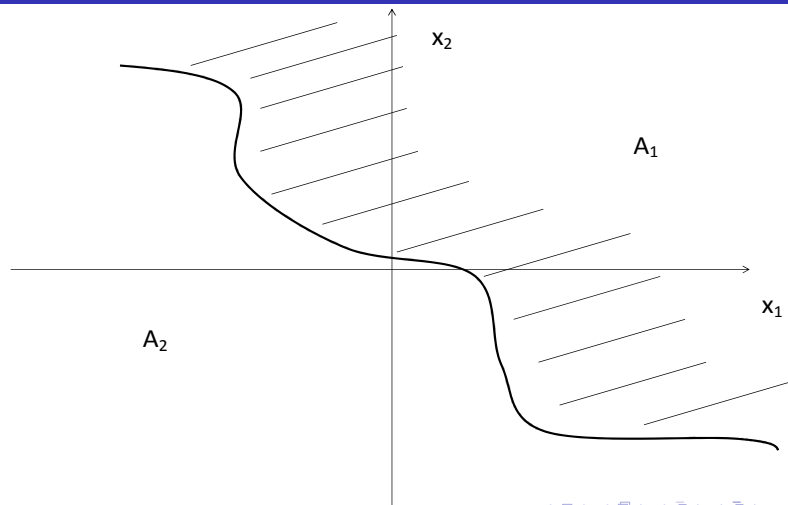
## AD para duas populações

- Suponha que tenhamos uma observação  $\mathbf{x}_0$  (vetor de valores observados para as variáveis de interesse de um determinado indivíduo) e denote por  $g_i; i = 1, 2$  o grupo  $i$ .
- Vamos denotar também por  $\theta_i$  os parâmetros de  $f_i, i = 1, 2$ . Se  $f_i$  for discreta então, uma forma de decidir à qual população pertence essa unidade experimental é: se  $f_1(\mathbf{x}_0) > f_2(\mathbf{x}_0)$ ,  $\mathbf{x}_0 \in g_1$ , caso contrário  $\mathbf{x}_0 \in g_2$ . Note que, nesse caso  $f_i(\mathbf{x}_0) = P_i(\mathbf{X} = \mathbf{x}_0)$  (teoria da verossimilhança).

## AD para duas populações

- O mesmo raciocínio pode ser usado se  $f_i$  for contínua pois, a partir da densidade, calcula-se probabilidades (intervalares) de interesse.
- Defina o suporte dessa distribuição por :  $A = \{\mathbf{x} \in \mathcal{R}^p, f(\mathbf{x}) > 0\}$ .
- A ideia é particionar  $A = A_1 \dot{\cup} A_2$  (união disjunta) de tal forma que, se uma observação, digamos  $\mathbf{x}_0$  é tal que  $\mathbf{x}_0 \in A_1$  alocaremos o indivíduo a população 1, caso contrário, ele será alocado para a população 2.

## Exemplo hipotético de partição para $p=2$



## AD para duas populações

- Defina  $p(i|j)$  a probabilidade de classificar um indivíduo no grupo  $i$  dado que ele pertence ao grupo  $j$ .
- Dessa forma, temos que:  $p(1|2) = P(\mathbf{X} \in A_1|g_2) = \int_{A_1} f_2(\mathbf{x})d\mathbf{x}$ .
- Analogamente, temos que:  $p(2|1) = P(\mathbf{X} \in A_2|g_1) = \int_{A_2} f_1(\mathbf{x})d\mathbf{x}$ .
- Seja  $p_i = P(g_i)$  a **probabilidade à priori** (antes de ser realizada a análise discriminante) de um indivíduo pertencer ao grupo  $i$ .
- Seja  $P(C_i)$  : a probabilidade do indivíduo ter sido corretamente classificado no grupo  $i$ .

# AD para duas populações

- Assim

$$\begin{aligned} P(C_i) &= P(\text{o indivíduo pertence a população } i \\ &\quad \text{e foi corretamente classificado na população } i) = P(C_i \cap g_i) \\ &= P(C_i|g_i)P(g_i) = p(i|i)p_i = p_i \int_{A_i} f_i(\mathbf{x})d\mathbf{x} \end{aligned}$$

# AD para duas populações

- Analogamente

$$\begin{aligned} P(\bar{C}_i) = 1 - P(C_i) &= P(\text{o indivíduo pertence a população } j \\ &\quad \text{e foi incorretamente classificado na população } i) \\ &= P(\bar{C}_i \cap g_j) = P(\bar{C}_i | g_j) P(g_j) = p(i|j) p_j \\ &= p_j \int_{A_i} f_j(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (1)$$

## AD para duas populações

- Um outro aspecto de interesse diz respeito ao fato dos custos (financeiros, logísticos etc) de se classificar indivíduos incorretamente.
- Define-se então a seguinte tabela:

População verdadeira	Classificação	
	$g_1$	$g_2$
$g_1$	0	$c(2 1)$
$g_2$	$c(1 2)$	0



# AD para duas populações

- Dessa forma podemos definir o **custo esperado de classificação errada (CECE)**:

$$CECE = p(1|2)c(1|2)p_2 + p(2|1)c(2|1)p_1$$

- Objetivo: criar uma regra de classificação de modo a minimizar o CECE.

## AD para duas populações

- Pode-se mostrar (exercício 11.3 do livro de Johnson & Wichern, sétima edição) que tal regra (sob a estrutura apresentada) é dada por:

$$\left\{ \begin{array}{l} \text{O indivíduo é classificado na população 1 se } \frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \\ \text{O indivíduo é classificado na população 2 se } \frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} \leq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \end{array} \right. \quad (2)$$

# AD para duas populações

- A probabilidade total de classificação incorreta (**PTCI**) é dada por:

$$\begin{aligned}PTCI &= P(\text{classificação incorreta na população 1 ou na população 2}) \\ &= P(\overline{C}_1 \cup \overline{C}_2) = P(\overline{C}_1) + P(\overline{C}_2) \\ &= p_1 \int_{A_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{A_1} f_2(\mathbf{x}) d\mathbf{x}\end{aligned}$$

(veja equação (1)).

## AD para duas populações sob normalidade multivariada

- Suponha que  $f_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$  (homocedasticidade).
- Sob a suposição acima temos que a regra de classificação (2) transforma-se em: Seja  $\mathbf{x}_0$  uma observação associada à um determinado indivíduo, então classificamos tal indivíduo na população 1 se

$$h(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \geq \ln \left( \frac{c(1|2) p_2}{c(2|1) p_1} \right) \quad (3)$$

em que

$h(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x}_0 - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$  e na população 2, caso contrário (exercício).

## AD para duas populações sob normalidade multivariada

- Defina  $y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x}_0$  e  $m = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ .
- Então a regra de classificação (equação (3) e o complemento) pode ser reescrita como: Seja  $\mathbf{x}_0$  uma observação associada à um determinado indivíduo, então classificamos tal indivíduo na população 1 se

$$y \geq m + \ln \left( \frac{c(1|2) p_2}{c(2|1) p_1} \right) \quad (4)$$

e na população 2, caso contrário.

## AD para duas populações sob normalidade multivariada

- Se os custos de classificação errada forem os mesmos bem como as probabilidades à priori, a regra torna-se ainda mais simples, ou seja:

$$y \geq m$$

- Vamos calcular o PTCI para esta regra de classificação.
- Queremos calcular  $P(\bar{C}_1) = p_1 P(Y_2 \geq m)$  e  $P(\bar{C}_2) = p_2 P(Y_1 < m)$  em que  $Y_i = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{X}_i$ ,  $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ .
- Pode-se demonstrar (exercício) que  $Y_i \sim N_1(\mu_{Y_i}, \Delta^2)$ , em que  $\mu_{Y_i} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i$  e  $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ ,  $i=1,2$ .

## AD para duas populações sob normalidade multivariada

- Note que se  $(Z \sim N(0, 1))$  e  $\Phi(a) = P(Z \leq a)$ , então:

$$\begin{aligned}P(Y_1 < m) &= P\left(Y_1 < \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) \\&= P\left(Z < \left(\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1\right) \frac{1}{\Delta}\right) \\&= P\left(Z < \left(\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right) \frac{1}{\Delta}\right) \\&= P\left(Z < -\frac{1}{2} \frac{\Delta^2}{\Delta}\right) = P\left(Z < -\frac{\Delta}{2}\right) = \Phi\left(-\frac{\Delta}{2}\right)\end{aligned}$$

## AD para duas populações sob normalidade multivariada

- Analogamente pode-se provar que  $P(Y_2 \geq m) = \Phi\left(-\frac{\Delta}{2}\right)$ . Assim  $PTCI = p_1\Phi\left(-\frac{\Delta}{2}\right) + p_2\Phi\left(-\frac{\Delta}{2}\right) = (p_1 + p_2)\Phi\left(-\frac{\Delta}{2}\right) = \Phi\left(-\frac{\Delta}{2}\right)$
- Como o PTCI foi obtido através de um mecanismo de classificação ótimo, dizemos (nesse caso) que o PTCI = TOE (ou taxa ótima de erro).
- Quanto menor for o TOE, melhor será a regra de classificação.



# AD para duas populações sob normalidade multivariada

- Uma outra forma de verificar a qualidade da regra de classificação é a chamada taxa de erro aparente (TEA). Para calculá-la, considere a seguinte tabela:

População verdadeira	Classificação		Total
	$g_1$	$g_2$	
$g_1$	$n_{C_1}$	$n_{E_2}$	$n_1$
$g_2$	$n_{E_1}$	$n_{C_2}$	$n_2$

# AD para duas populações sob normalidade multivariada

- Em que  $n_{C_i}$  : número de indivíduos que foram corretamente classificados no grupo  $i$  e  $n_{E_i}$  : número de indivíduos que foram incorretamente classificados no grupo  $i$ .
- Assim

$$TEA = \frac{n_{E_1} + n_{E_2}}{n_1 + n_2}$$

(quanto menor a TEA, melhor o método de classificação).

## AD para duas populações sob normalidade multivariada

- Na prática, desconhecemos  $(\mu_1, \mu_2)$  e  $\Sigma$ . O que temos é uma matriz de dados com dois grupos, como vista anteriormente.
- Assim, substituímos tais parâmetros por estimadores apropriados, ou seja, utilizamos  $\bar{\mathbf{X}}_1$ ,  $\bar{\mathbf{X}}_2$  e  $\mathbf{S}_P^2 = \frac{1}{n_1+n_2-2} [(n_1-1)\mathbf{S}_1^2 + (n_2-1)\mathbf{S}_2^2]$ , em que  $\mathbf{S}_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$
- Para estimar o TOE podemos utilizar 
$$\tilde{\Delta} = \sqrt{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' (\mathbf{s}_P^2)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}.$$

# AD para duas populações sob normalidade multivariada

- É digno de nota que (Sir) Ronald Fisher chegou à mesma regra de classificação (4) usando um argumento totalmente diferente.
- Ele buscava transformar observações multivariadas ( $\mathbf{x}$ ) em univariadas ( $y$ ) de tal forma que os valores de  $y$  fossem os mais diferentes possíveis entre as duas populações, em que  $y$  é definida como alguma combinação linear de  $\mathbf{X}$ .
- Os desenvolvimentos de Fisher não requerem normalidade multivariada, apenas homocedasticidade. Assim, a regra (4) é válida mesmo se a suposição de normalidade multivariada não for observada.

# AD para duas populações sob normalidade multivariada

- Resumidamente:
  - A metodologia baseada na minimização da CECE não requer normalidade multivariada nem homocedasticidade.
  - A metodologia desenvolvida por Fisher requer apenas homocedasticidade.
  - Sob normalidade, a primeira, equivale a segunda.
  - A metodologia desenvolvida por Fisher está disponível na função “lda”, ou seja, ela é apropriada sob homocedasticidade, sem requerer normalidade.

## AD para duas populações

- Por isso a regra (4) também é chamada de discriminação linear de Fisher e  $y$  é chamada de função discriminante linear de Fisher.
- Resumindo, seja  $\mathbf{x}_0$  uma observação associada à um determinado indivíduo, então classificamos tal indivíduo na população 1 se

$$y \geq m + \ln \left( \frac{c(1|2) p_2}{c(2|1) p_1} \right) \quad (5)$$

em que  $y = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' (\mathbf{s}_P^2)^{-1} \mathbf{x}_0$ ,  $m = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' (\mathbf{s}_P^2)^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$ ,  $\bar{y}_i = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' (\mathbf{s}_P^2)^{-1} \bar{\mathbf{x}}_i$ ,  $i = 1, 2$ , e na população 2, caso contrário.

## AD para duas populações

- Na prática temos um conjunto de observações:  $x_{01}^{(1)}, \dots, x_{0n_1}^{(1)}$  (grupo 1) e  $x_{01}^{(2)}, \dots, x_{0n_2}^{(2)}$  (grupo 2), os quais, quando juntos, compõem a amostra total.
- Usualmente, retiramos uma amostra aleatória de cada grupo e as utilizamos para calcular  $\bar{\mathbf{x}}_1$ ,  $\bar{\mathbf{x}}_2$  e  $\mathbf{s}_p^2$  (amostra treino).
- A amostra teste (o que resta da amostra total, em retirando-se a amostra treino) será usada para testar a regra.
- De posse dessas quantidades calculamos os coeficientes de  $y$   $\left[ (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' (\mathbf{s}_p^2)^{-1} \right]$  e  $m$ .

## AD para duas populações

- Para cada unidade amostral (da amostra treino e da amostra teste) podemos calcular  $y_{0j}^{(i)} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' (\mathbf{s}_p^2)^{-1} \mathbf{x}_{0j}^{(i)}, j = 1, \dots, n_i$  (grupo  $i$ ).
- Cada unidade é classificado de acordo com a regra (5), ou seja se  $y_{0j}^{(i)} \geq m + \ln \left( \frac{c(1|2) p_2}{c(2|1) p_1} \right)$  ela é alocada à população 1, caso contrário, à população 2.
- Como dito anteriormente, usualmente considera-se apenas a amostra teste para se avaliar a qualidade da regra de classificação.
- Pode-se ainda usar os valores estimados das funções discriminantes para se analisar/comparar os grupos.



## Apoio computacional

- A função “lda”, do pacote R, executa a análise discriminante (sob homocedasticidade) via função linear discriminante.
- Os coeficientes da função linear discriminante ( $\mathbf{b} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}$ ) são reescalados (divididos por uma constante, ou seja  $\mathbf{b}^* = \frac{\mathbf{b}}{\mathbf{b}' \boldsymbol{\Sigma} \mathbf{b}}$ ).
- Cuidado deve ser tomado com reescalamentos, se o objetivo for obter interpretações para os coeficientes.
- Em princípio, os coeficientes não têm uma interpretação, embora normalizar as variáveis originais e os coeficientes possa ser útil nesse sentido.

## Exemplo 1: versicolor x virginica

- Versicolor (grupo 1) e virginica (grupo 2).
- Amostra aleatória de 25 plantas de cada uma das espécies acima para gerar a regra de classificação.
- Vetores de médias:

Variável	Versicolor	Virginica
CS	5,94	6,70
LS	2,74	2,94
CP	4,29	5,61
LP	1,32	2,03

■ Matrizes de variância-covariância:

Versicolor

	CS	LS	CP	LP
CS	0,28	0,11	0,15	0,05
LS	0,11	0,12	0,09	0,05
CP	0,15	0,09	0,18	0,06
LP	0,05	0,05	0,06	0,04

Virginica

	CS	LS	CP	LP
CS	0,43	0,05	0,33	0,07
LS	0,05	0,08	0,02	0,04
CP	0,33	0,02	0,34	0,05
LP	0,07	0,04	0,05	0,06

# TEA e TOE

- Resultados da classificação:

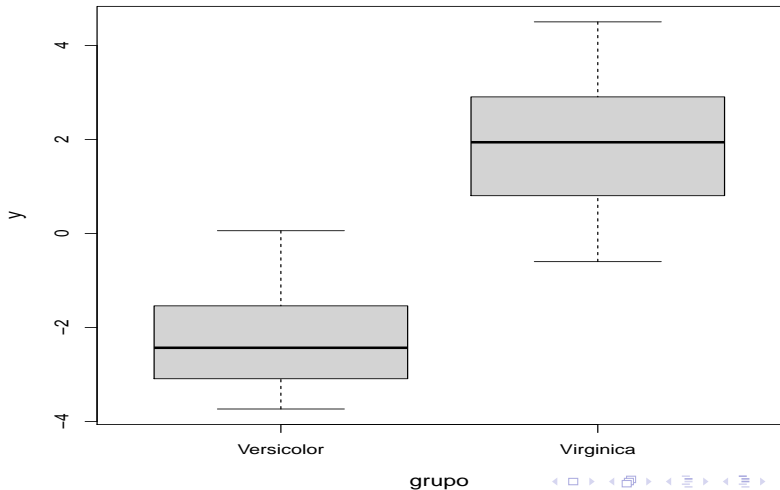
Observado	Classificado	
	VE	VI
VE	24	1
VI	2	23

- TEA (%) : 6,00.
- TOE (%): 1,87.

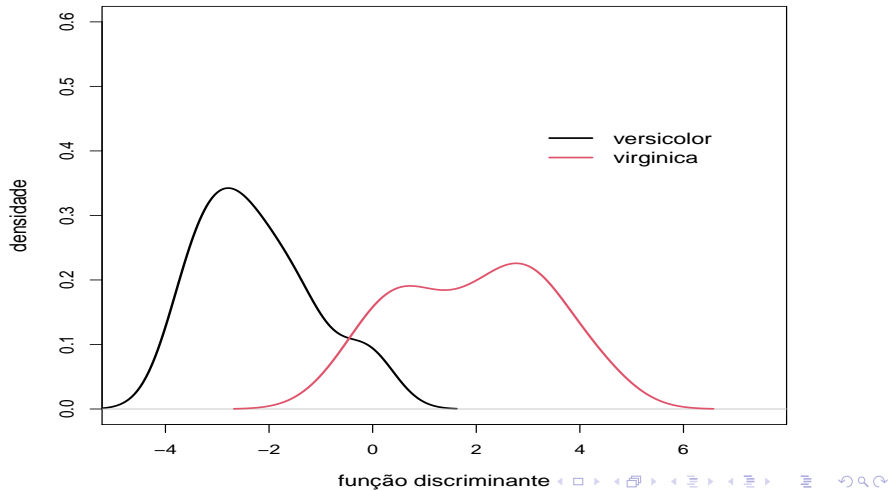
# Medidas resumo das funções discriminantes

Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Versicolor	-2,24	1,10	1,21	-3,73	-2,43	0,06	25
Virginica	1,94	1,46	2,14	-0,60	1,94	4,50	25

# Boxplots das funções discriminantes



# Densidades estimadas das funções discriminantes



# Comentários

- O método de classificação apresentou um ótimo desempenho (embora os critérios de qualidade podem variar consoante o objetivo e/ou problema).
- As distribuições das funções discriminantes mostraram-se razoavelmente diferentes entre si.
- A suposição de normalidade não parece ser razoável para as distribuições das funções discriminantes, para nenhum dos dois grupos.



## Exemplo 1: setosa x virginica

- Setosa (grupo 1) e virginica (grupo 2).
- Amostra aleatória de 25 plantas de cada uma das espécies acima para gerar a regra de classificação.
- Vetores de médias:

	Setosa	Virginica
CS	4,94	6,62
LS	3,36	2,3
CP	1,47	5,60
LP	0,24	1,96

■ Matrizes de covariância:

Setosa

	CS	LS	CP	LP
CS	0,11	0,10	0,01	0,01
LS	0,10	0,15	0,02	0,01
CP	0,01	0,02	0,03	0,01
LP	0,01	0,01	0,01	0,01

Virginica

	CS	LS	CP	LP
CS	0,68	0,16	0,52	0,08
LS	0,16	0,13	0,10	0,04
CP	0,52	0,10	0,47	0,06
LP	0,08	0,04	0,06	0,08

# TEA e TOE

- Resultados da classificação:

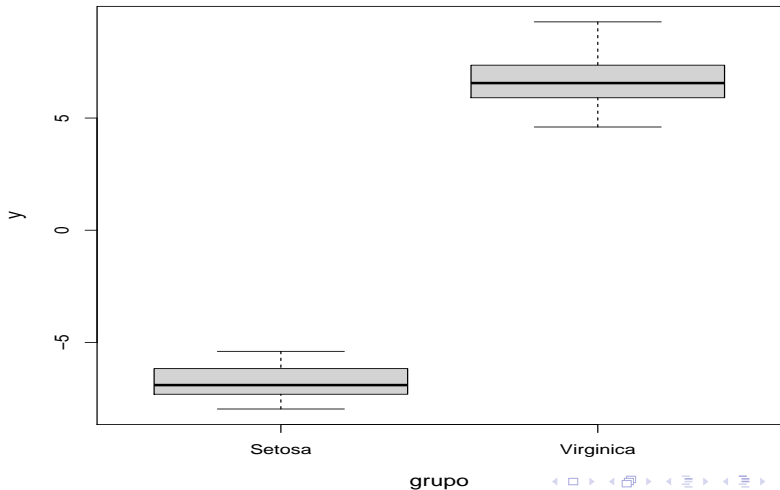
Observado	Classificado	
	VE	VI
VE	25	0
VI	0	25

- TEA (%) : 0,00.
- TOE (%): < 0,01.

## Medidas resumo

Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Setosa	-6,72	0,67	0,45	-7,97	-6,90	-5,40	25
Virginica	6,58	1,11	1,23	4,60	6,56	9,29	25

# Boxplots da função discriminante



# Densidade estimada da função discriminante

