Análise Discriminante

Prof. Caio Azevedo

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.
- 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, em geral, foi obtida através de algum método: caro e/ou invasivo e/ou que requer que a unidade experimental seja destruída porém, é (muito) confiável.
- Algumas vezes, esse método 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.

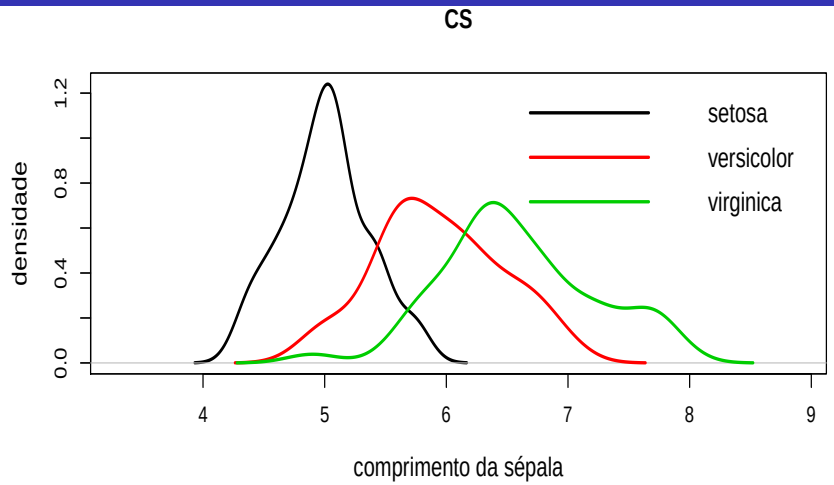
Motivação

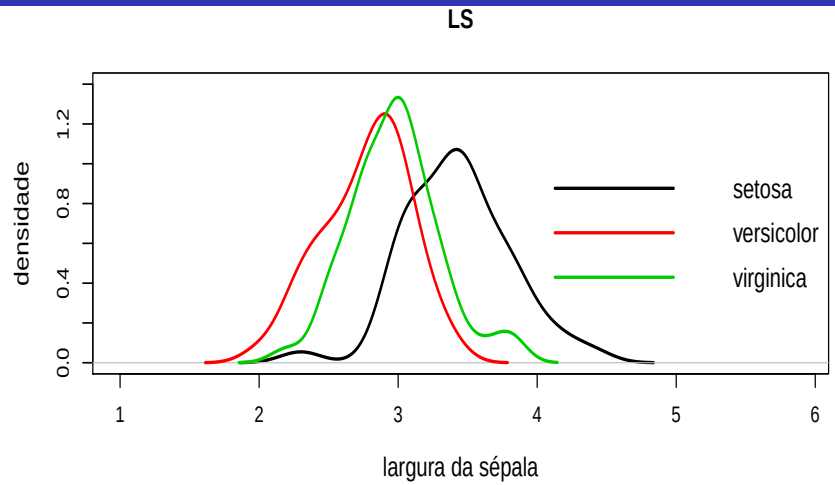
- Tal procedimento consiste em identificar padrões de "comportamento" para cada grupo em relação às variáveis disponíveis.
- Se este procedimento for satisfatório, em termos de classificação, ele pode ser usado em futuros estudos no lugar do método utilizado inicialmente.
- Em princípio, quanto mais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

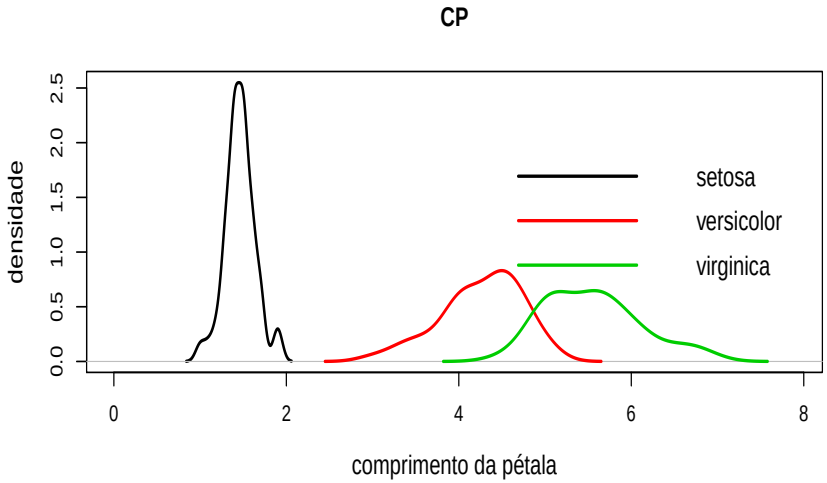
AD para duas populações

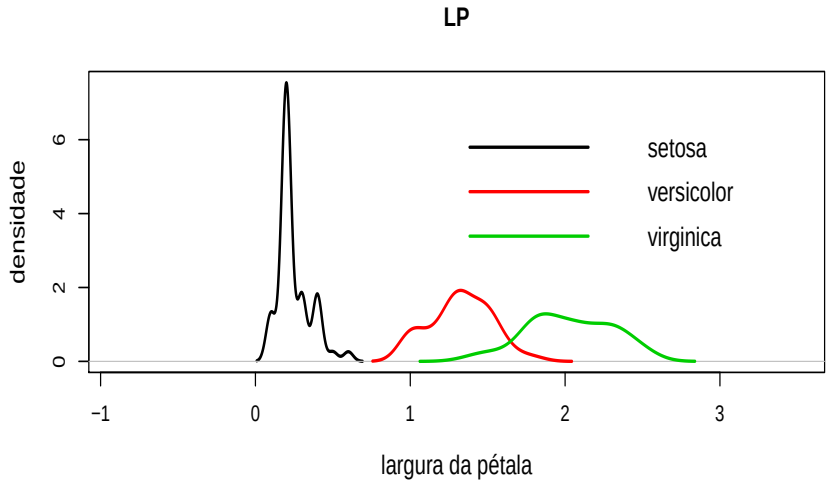
- Suponha que tenhamos duas populações das quais extraímos duas amostras aleatórias (independentes entre populações e entre indivíduos) e delas medimos p características.
- Cada uma dessas populações (processo de amostragem/experimentação) pode ser representado por uma fdp (discreta, contínua ou mista)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

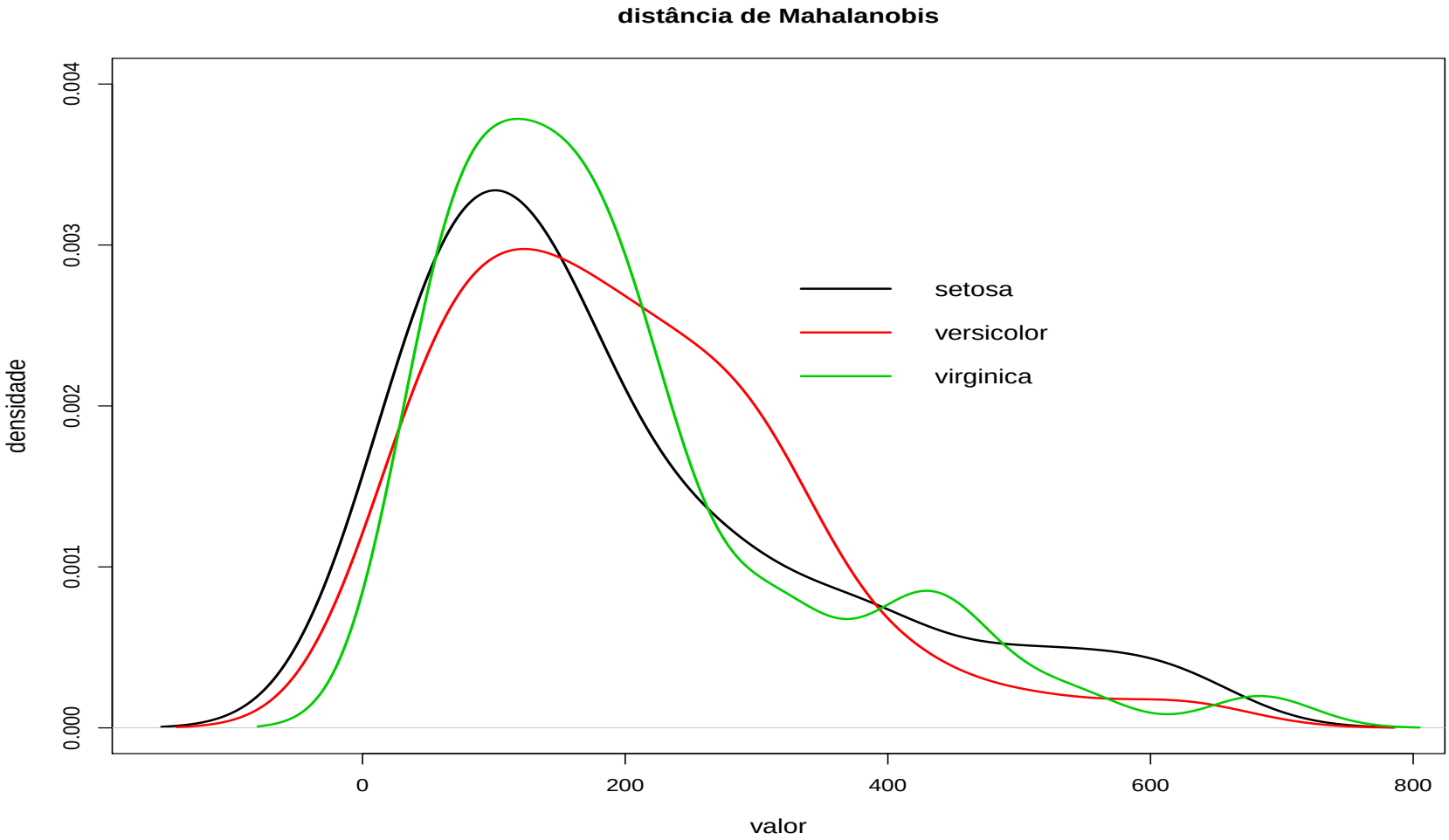
Ex. 1: densidades para cada variável em função dos grupos










distância de Mahalanobis
densidade
0.000
0.001
0.002
0.003
0.004
0
200
400
600
800
valor
setosa
versicolor
virginica

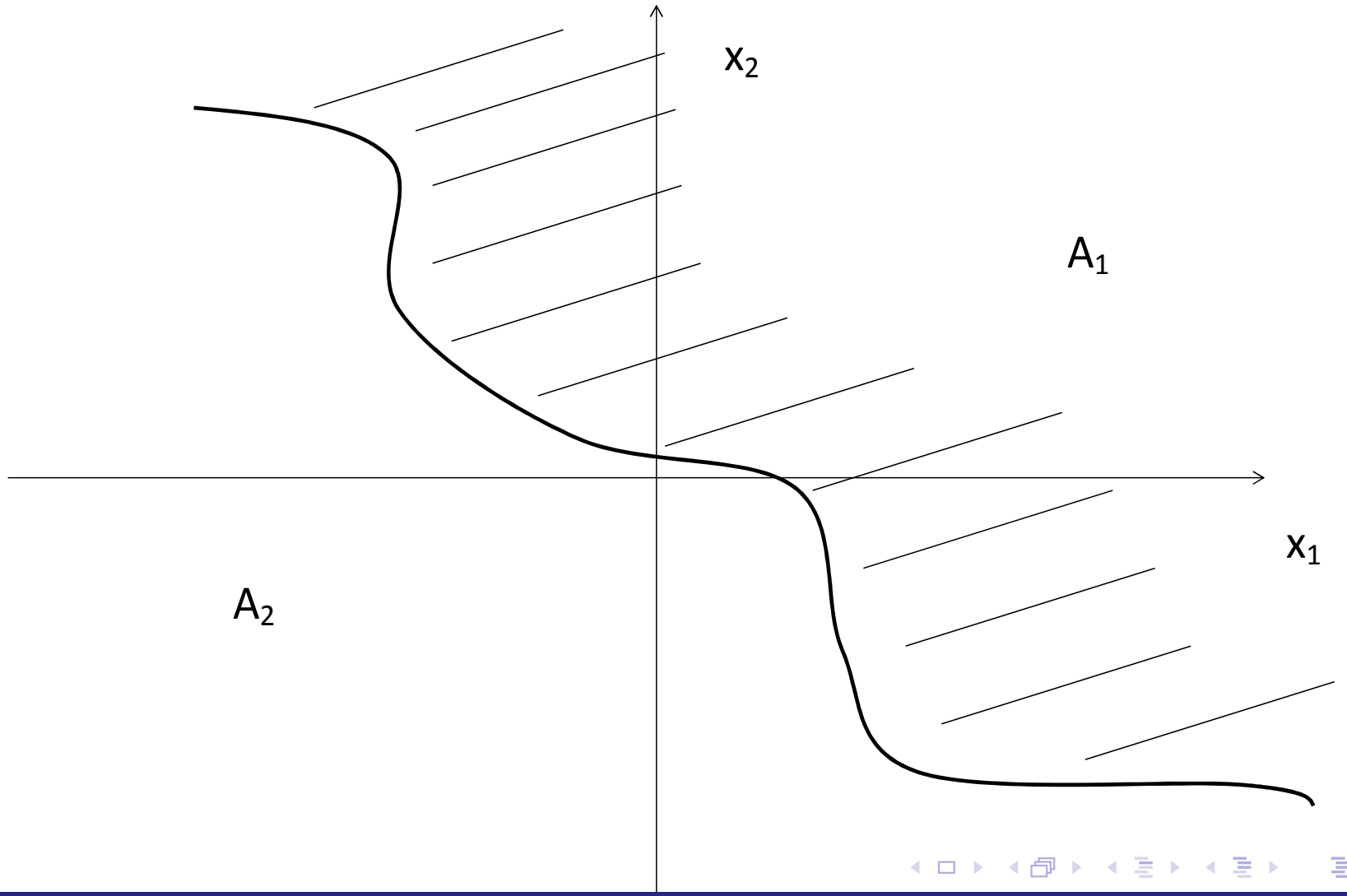
AD para duas populações

- Suponha que tenhamos uma observação $\boldsymbol{x}_0$ (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 $\boldsymbol{\theta}_i$ 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 $f_1(\boldsymbol{x}_0) > f_2(\boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in g_1$, caso contrário $\boldsymbol{x}_0 \in g_2$. Note que, nesse caso $f_i(\boldsymbol{x}_0) = P_i(\boldsymbol{X} = \boldsymbol{x}_0)$.
- O mesmo raciocínio pode ser usado se f_i for contínua pois, a partir da densidade, calcula-se probabilidades de interesse.

AD para duas populações

- Defina o suporte dessa distribuição por : $A = \{\boldsymbol{x} \in \mathcal{R}^p, f(\boldsymbol{x}) > 0\}$.
- A idéia é particionar $A = A_1 \dot{\bigcup} A_2$ (união disjunta) de tal forma que, se uma observação, digamos $\boldsymbol{x}_0$ é tal que $\boldsymbol{x}_0 \in A_1$ alocaremos o indivíduo a população 1, caso contrário, ele será alocado para a população 2.

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: $p(1|2) = P(\boldsymbol{X} \in A_1|g_2) = \int_{A_1} f_2(\boldsymbol{x})d\boldsymbol{x}$.
- Analogamente, temos que: $p(2|1) = P(\boldsymbol{X} \in A_2|g_1) = \int_{A_2} f_1(\boldsymbol{x})d\boldsymbol{x}$.
- 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

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

- Analogamente

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

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:

$$\begin{cases} \text{O indivíduo é classificado na população 1 se } \dfrac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} \geq \dfrac{c(1|2)}{c(2|1)} \dfrac{p_2}{p_1} \\ \text{O indivíduo é classificado na população 2 se } \dfrac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} \leq \dfrac{c(1|2)}{c(2|1)} \dfrac{p_2}{p_1} \end{cases} \tag{2}$$

AD para duas populações

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

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

(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 $\boldsymbol{x}_0$ uma observação associada à um determinado indivíduo, então classificamos tal indivíduo na população 1 se

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

em que

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

AD para duas populações sob normalidade multivariada

- Defina $y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_0$ e $m = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$. Então a regra (equação (3) e o complemento) pode ser reescrito como:

  Seja $\boldsymbol{x}_0$ uma observação associada à um determinado indivíduo, então classificamos tal indivíduo na população 1 se

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

  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 $P(\overline{C}_1) = p_1 P(Y_2 \geq m)$ e $P(\overline{C}_2) = p_2 P(Y_1 < m)$ em que $Y_i = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i$, $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Pode-se demonstrar 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:

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

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 uma 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

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

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 ($\boldsymbol{x}$) 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 $\boldsymbol{X}$.
- Os desenvolvimentos de Fisher não requerem normalidade multivariada, apenas homocedasticidade. Assim, a regra (4) é valida 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, elá 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 $\boldsymbol{x}_0$ uma observação associada à um determinado indivíduo, então classificamos tal indivíduo na população 1 se

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

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

AD para duas populações

- Na prática temos uma conjunto de observações: $x_{01}^{(1)}, \ldots x_{0n_1}^{(1)}$ (grupo 1) e $x_{01}^{(2)}, \ldots x_{0n_2}^{(2)}$ (grupo 2) - amostra total.
- Usualmente, retiramos uma amostra aleatória de cada grupo e as utilizamos para calcular $\overline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_2$ e $\boldsymbol{s}_P^2$ (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 $(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \left(\boldsymbol{s}_P^2\right)^{-1}$ e m.

AD para duas populações

- Para cada unidade amostral (da amostra treino e da amostra teste) podemos calcular $y_{0j}^{(i)} = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \left(\boldsymbol{s}_P^2\right)^{-1} \boldsymbol{x}_{0j}^{(i)}, j = 1, ..., n_i$ (grupo i).
- Cada unidade é classificado de acordo com a regra (5), ou seja se $y_{0j}^{(i)} \geq m + \ln\left(\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}\right)$ ela é classificada na população 1, caso contrário, na 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 ($\boldsymbol{b} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}$) são reescalonados (divididos por uma constante, ou seja $\boldsymbol{b}^* = \frac{\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b}}$).
- Cuidado deve ser tomado com reescalonamentos, 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:

	Versicolor	Virginica
CS	5,99	6,50
LS	2,76	2,92
CP	4,26	5,54
LP	1,32	2,02

- Matrizes de variância-covariância:

Versicolor

	CS	LS	CP	LP
CS	0,31	0,12	0,20	0,07
LS	0,12	0,14	0,10	0,05
CP	0,20	0,10	0,20	0,07
LP	0,07	0,05	0,07	0,04

Virginica

	CS	LS	CP	LP
CS	0,46	0,07	0,31	0,07
LS	0,07	0,09	0,05	0,07
CP	0,31	0,05	0,28	0,06
LP	0,07	0,07	0,06	0,11

TEA e TOE

- Resultados da classificação:

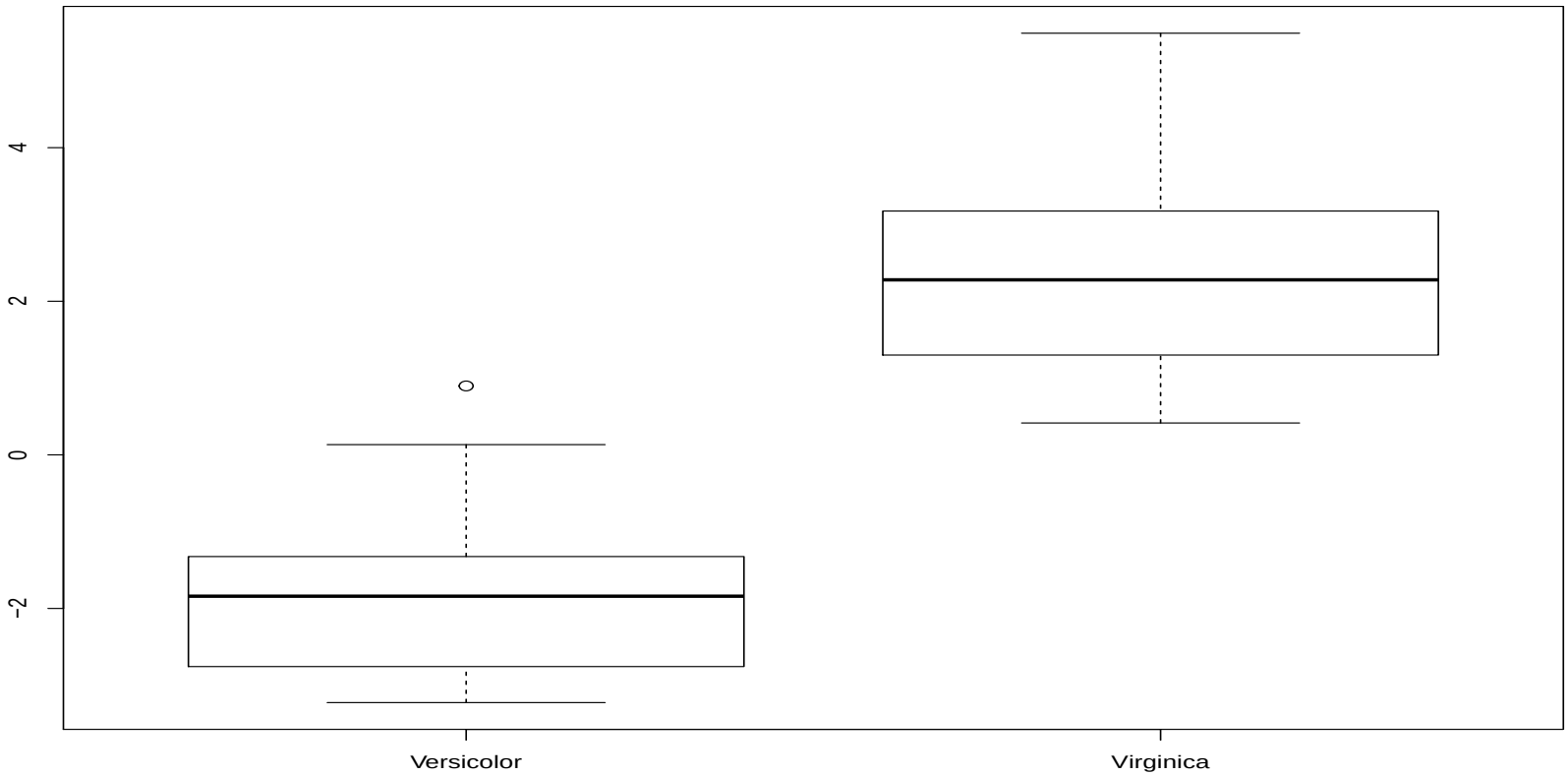
	Classificado	
Observado	VE	VI
VE	22	3
VI	0	25

- TEA (%) : 6,00.
- TOE (%): 2,06.

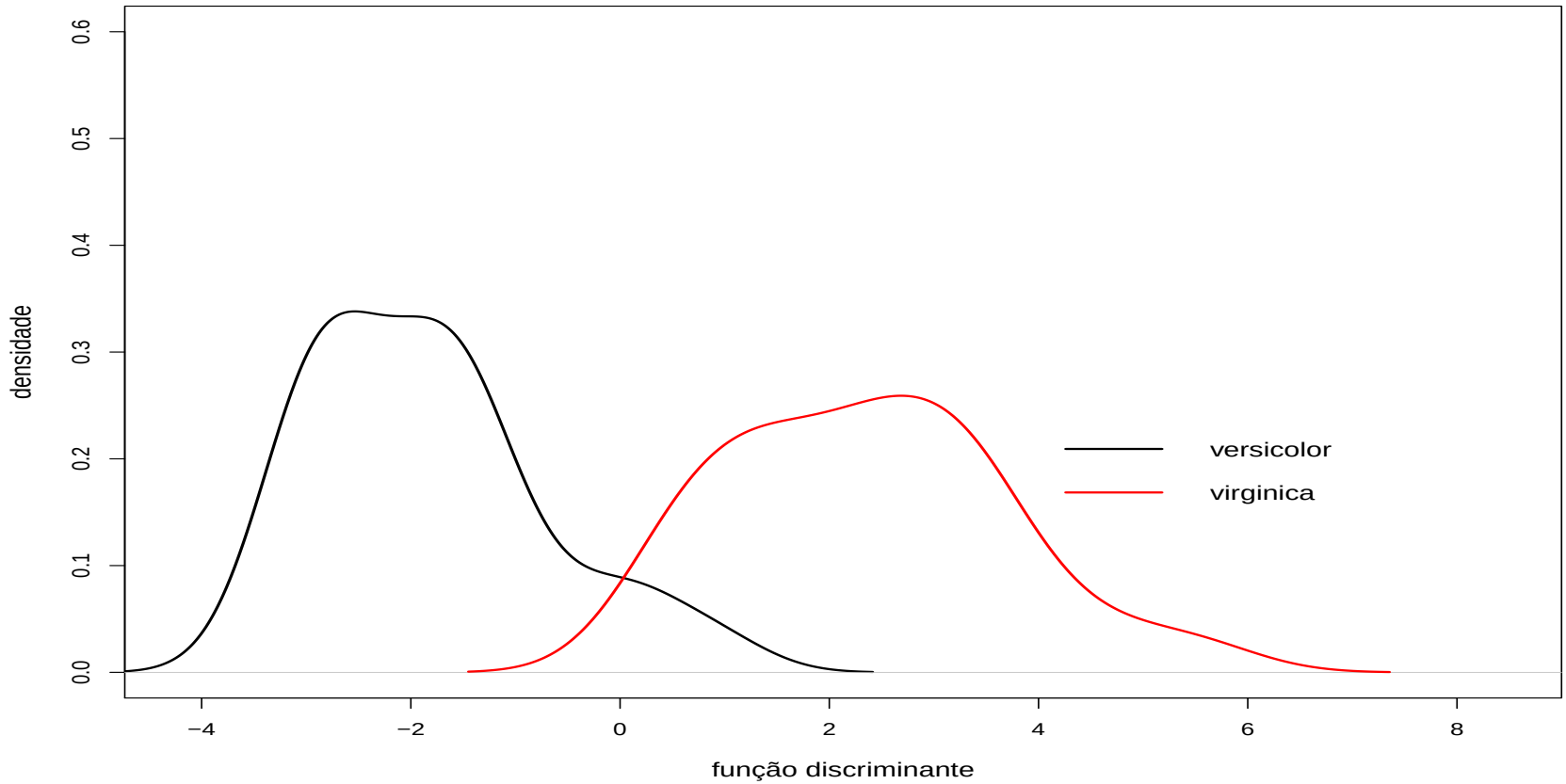
Medidas resumo

Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Versicolor	-1,85	1,10	1,20	-3,23	-1,84	0,90	25
Virginica	2,35	1,31	1,73	0,41	2,28	5,49	25

Ex. 1: boxplots da função discriminante



Ex. 1: densidade estimada da função discriminante



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	5,00	6,51
LS	3,40	2,94
CP	1,45	5,51
LP	0,25	2,04

- Matrizes de covariância:

Setosa

	CS	LS	CP	LP
CS	0,14	0,09	0,02	0,01
LS	0,09	0,12	-0,01	0,02
CP	0,02	-0,01	0,02	0,00
LP	0,01	0,02	0,00	0,01

Virginica

	CS	LS	CP	LP
CS	0,42	0,07	0,27	0,06
LS	0,07	0,09	0,04	0,04
CP	0,27	0,04	0,25	0,04
LP	0,06	0,04	0,04	0,08

TEA e TOE

- Resultados da classificação:

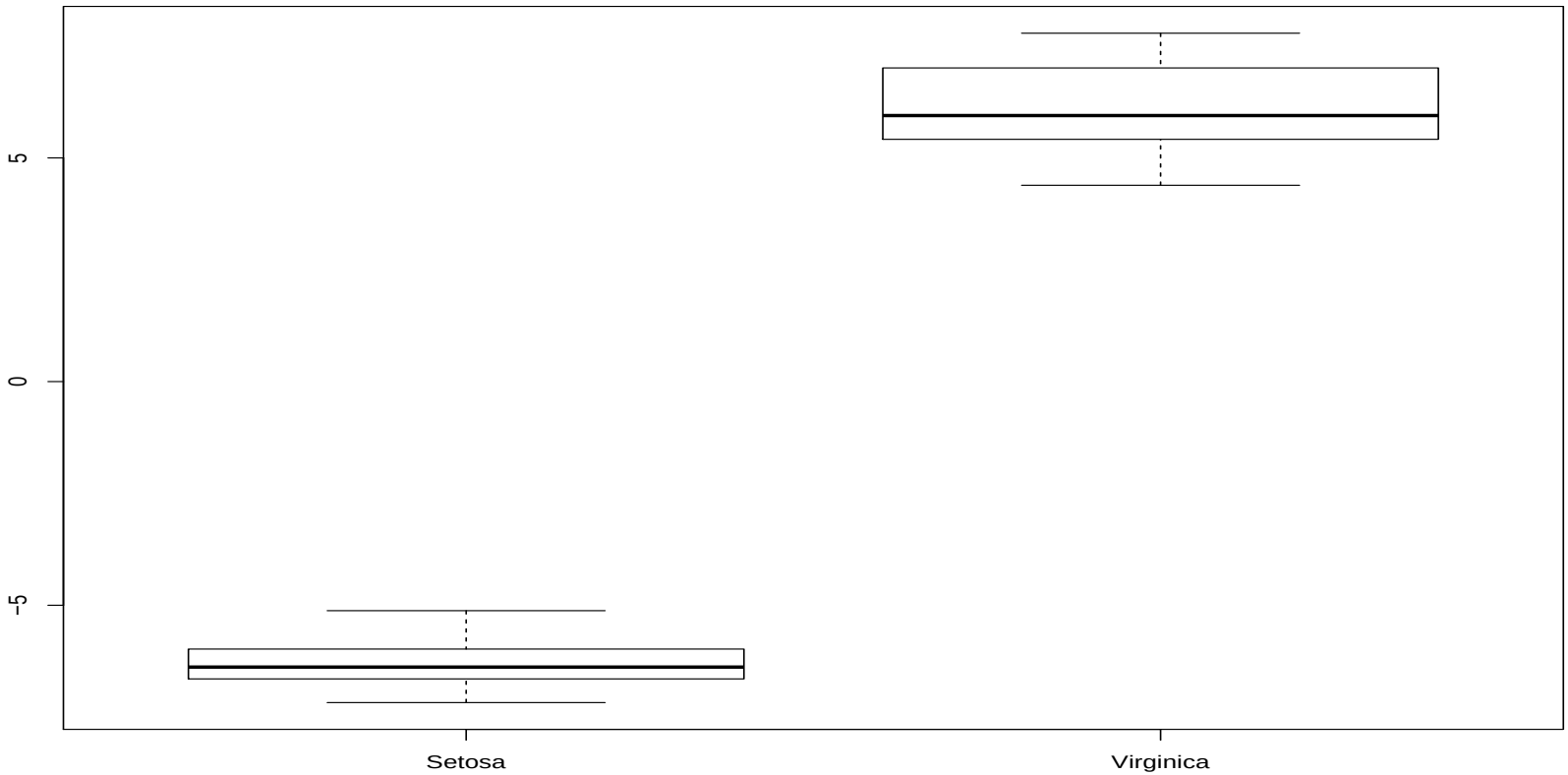
	Classificado	
Observado	VE	VI
VE	25	0
VI	0	25

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

Medidas resumo

Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Setosa	-6,31	0,49	0,24	-7,18	-6,38	-5,12	25
Virginica	6,09	1,04	1,08	4,39	5,95	7,79	25

Ex. 1: boxplots da função discriminante



Ex. 1: densidade estimada da função discriminante