

# Análise Discriminante: parte 2

Prof. Caio Azevedo

# Comentários

- A classificação entre duas populações sob heterocedasticidade ( $\Sigma_1 \neq \Sigma_2$ ) pode ser encontrada em [Johnson and Wichern(2007)].
- A regra de classificação apresentada no supramencionado livro pode gerar resultados indesejados (altas taxas de erro), principalmente quando temos mais de duas variáveis e quando a suposição de normalidade multivariada não é válida.

# Comentários

- Não discutiremos a AD sob heterocedasticidade. Contudo, apresentaremos a situação com vários grupos, sob homocedasticidade.
- Algumas alternativas nesse caso modelos de regressão para **dados binários** (dois grupos) e modelos de regressão para **dados politômicos** (três ou mais grupos).

# Classificação com várias populações, baseada no CECE

- Suponha  $g$  populações de sorte que cada unidade amostral (experimental) pertença a uma e somente uma população. Defina (analogamente ao caso de duas populações):
  - Suporte da distribuição:  $A = \{\mathbf{x} \in \mathcal{R}^p, f(\mathbf{x}) > 0\}$  e a respectiva partição  $A = \dot{\bigcup}_{i=1}^g A_i$ .
  - $p_i$  : probabilidade (a priori) de um determinado indivíduo pertencer à população  $i$ .
  - $c(k|i)$  : custo de classificar o indivíduo na população  $k$  dado que ele pertence à população  $i$ , naturalmente  $c(i|i) = 0, \forall i$ .
  - $P(k|i) = \int_{A_k} f_i(\mathbf{x})d\mathbf{x}$  : probabilidade de classificar um indivíduo na população  $k$  dado que ele pertence à população  $i$ . Além disso  $P(i|i) = 1 - \sum_{k=1, k \neq i}^g P(k|i)$ .

# Classificação com várias populações, baseada no CECE

- O custo esperado condicional (esperança condicional) de classificar equivocadamente uma unidade pertencente à população 1, em uma outra população, (2,3,...,g) é dada por:

$$\begin{aligned} ECE(1) &= P(2|1)c(2|1) + P(3|1)c(3|1) + \dots + P(g|1)c(g|1) \\ &= \sum_{k=2}^g P(k|1)c(k|1) = \sum_{k=1}^g P(k|1)c(k|1) \end{aligned} \quad (1)$$

- Assumi-se que este valor esperado condicional ocorre com probabilidade  $p_1$  (a priori).
- Analogamente podemos definir  $ECE(2), \dots, ECE(g)$  e, além disso, cada uma delas ocorre com probabilidade  $p_k, k = 2, \dots, g$

# Classificação com várias populações, baseada no CECE

- Assim o custo (global) esperado de classificação errada é dado por:

$$\begin{aligned} CECE &= p_1 ECE(1) + p_2 ECE(2) + \dots + p_g ECE(g) \\ &= \sum_{i=1}^g p_i \sum_{k=1, k \neq i}^g P(k|i) c(k|i) \end{aligned} \quad (2)$$

# Classificação com várias populações, baseada no CECE

- A regra de classificação que minimiza o CECE (equação (2)) consiste em classificar uma dada observação ( $\mathbf{x}_0$ ), associada à uma determinada unidade, na população  $k$ ,  $k = 1, 2, \dots, g$ , para o qual

$$\sum_{i=1, i \neq k}^g p_i f_i(\mathbf{x}_0) c(k|i) = \sum_{i=1}^g p_i f_i(\mathbf{x}_0) c(k|i)$$

é mínimo. Para a demonstração veja [[Anderson\(2003\)](#)].

- Note que, neste caso, a suposição de normalidade (de alguma distribuição) é necessária ( $f_i(\mathbf{x}_0)$ ). A suposição de homocedasticidade também é importante (no mesmo sentido apresentado para o caso de duas populações).

# Classificação com várias populações baseado no CECE com custos iguais ( $c(.|.)$ )

- Nesse caso a regra passa a ser: aloca-se a unidade amostral na população  $k$ , com base em um vetor de observações  $\mathbf{x}_0$  se

$$p_k f_k(\mathbf{x}_0) > p_i f_i(\mathbf{x}_0), \forall i \neq k \quad (3)$$

ou de modo equivalente, se

$$\ln(p_k f_k(\mathbf{x}_0)) > \ln(p_i f_i(\mathbf{x}_0)), \forall i \neq k$$



# Classificação com várias populações baseada no CECE com custos iguais ( $c(.|.)$ )

- Um aspecto interessante é que a regra de classificação (3) equivale àquela que maximiza a probabilidade à posteriori  $P(k|\mathbf{x}_0)$  ( $\mathbf{x}_0$  vem da população  $k$  dado que  $\mathbf{x}_0$  foi observado), a qual é dada por:

$$P(k|\mathbf{x}_0) = \frac{p_k f_k(\mathbf{x}_0)}{\sum_{i=1}^g p_i f_i(\mathbf{x}_0)} = \frac{\text{priori} \times \text{verossimilhança}}{\sum [\text{priori} \times \text{verossimilhança}]}$$

- Que é, essencialmente, um resultado do [Teorema de Bayes](#).

# Método de Fisher para AD com várias populações

- Fisher também desenvolveu uma metodologia de AD considerando várias populações.
- A ideia é semelhante ao caso de duas populações, no sentido que ele buscou definir funções univariadas (função discriminante) com base nas observações multivariadas.
- Neste caso, também, a metodologia de Fisher é equivalente à regra do mínimo CECE, sob normalidade multivariada, homocedasticidade, custos de classificação errada e probabilidades à priori iguais (veja [[Johnson and Wichern\(2007\)](#)]).

# Método de Fisher para AD com várias populações

- Características importantes:
  - Permite trabalhar com representações univariadas de dados multivariados.
  - Permite análises gráficas que permitem analisar de modo simples o comportamento das populações de interesse.
  - Não assume normalidade dos dados.
  - Considera homocedasticidade.

# Método de Fisher para AD com várias populações

- A ideia é trabalhar com combinações lineares das observações  $\mathbf{x}$  ( $\mathbf{x}_0$ ), ou seja,  $Y = \mathbf{a}'\mathbf{x}$ . Assuma que  $\mathcal{E}(\mathbf{X}|i) = \boldsymbol{\mu}_i$  e  $\text{Cov}(\mathbf{X}|i) = \boldsymbol{\Sigma}$ .
- Assim  $\mathcal{E}(Y) = \mathbf{a}'\boldsymbol{\mu}_i$  (população  $i$ ) e  $\text{Cov}(Y) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$  (para todas as populações).
- Defina (considerando:  $\bar{\boldsymbol{\mu}} = \frac{1}{g} \sum_{i=1}^g \boldsymbol{\mu}_i$ )

$$\begin{aligned} V &= \frac{\mathbf{a}' \left( \sum_{i=1}^g (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})' \right) \mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} \\ &= \frac{\text{variabilidade entre grupos}}{\text{variabilidade dentro de cada grupo}} \end{aligned}$$

## Método de Fisher para AD com várias populações

- A idéia de Fisher foi tentar “separar” ao máximo as populações em relação à medida  $V$ , ou seja, ele buscou maximizá-la.
- Resultado: defina  $\mathbf{w} = \sum_{i=1}^g (n_i - 1) \mathbf{s}_i^2$  (variabilidade dentro de cada população),  $\mathbf{b} = \sum_{i=1}^g (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$ , em que  $\bar{\mathbf{x}}_i$  é a média amostral,  $\mathbf{s}_i^2$  é a matriz de covariâncias amostral, ambas relativas a população  $i$  e  $\bar{\mathbf{x}} = \frac{1}{g} \sum_{i=1}^g \bar{\mathbf{x}}_i$ .
- Sejam  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_s > 0$ , em que  $s \leq \min(g - 1, p)$  os autovalores diferentes de zero de  $\mathbf{w}^{-1} \mathbf{b}$  e  $\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \dots, \tilde{\mathbf{e}}_s$  os respectivos autovetores, devidamente reescalados (veja [Johnson and Wichern(2007)]).

# Método de Fisher para AD com várias populações

- Então o vetor de coeficientes, digamos  $\tilde{\mathbf{a}}$ , que maximiza a razão

$$\frac{\tilde{\mathbf{a}}' \mathbf{b} \tilde{\mathbf{a}}}{\tilde{\mathbf{a}}' \mathbf{w} \tilde{\mathbf{a}}},$$

é dado por  $\tilde{\mathbf{a}}_1 = \tilde{\mathbf{e}}_1$ .

- Então, a combinação linear  $\tilde{y}_1 = \tilde{\mathbf{a}}_1 \mathbf{x}$  é denominada de primeira função discriminante (amostral).
- Analogamente, a combinação linear  $\tilde{y}_k = \tilde{\mathbf{a}}_k \mathbf{x}$  é denominada  $k$ -ésima função discriminante (amostral).

# Método de Fisher para AD com várias populações

- Regra de classificação: aloca-se a observação  $\mathbf{x}_0$  (ou seja, a unidade amostral), à população  $k$ , com base em  $s$  funções discriminantes, de sorte que:

$$\sum_{j=1}^s (\tilde{y}_j - \bar{y}_{kj})^2 = \sum_{j=1}^s [\tilde{\mathbf{a}}_j'(\mathbf{x}_0 - \bar{\mathbf{x}}_k)]^2 \leq \sum_{j=1}^s [\tilde{\mathbf{a}}_j'(\mathbf{x}_0 - \bar{\mathbf{x}}_i)]^2, \forall i \neq k$$

- Nota: assim como no caso de duas populações, a regra de classificação de Fisher coincide com a regra do CECE (sob normalidade e homocedasticidade), sob custos iguais e probabilidades de classificação iguais.

## Voltando ao Exemplo 1 (considerando os três grupos)

- Resultados da classificação:

	S	VE	VI
S	25	0	0
VE	0	23	2
VI	0	0	25

- TEA (%) = 2,67.



# Medidas resumo

## Função discriminante 1

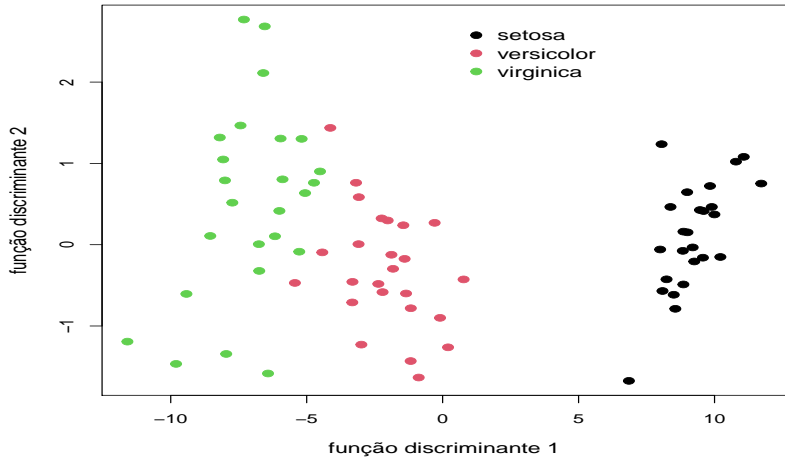
Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Setosa	9,19	1,08	1,16	6,85	9,00	11,72	25
Versicolor	-2,09	1,49	2,22	-5,43	-2,02	0,77	25
Virginica	-7,04	1,69	2,86	-11,59	-6,74	-4,51	25

# Medidas resumo

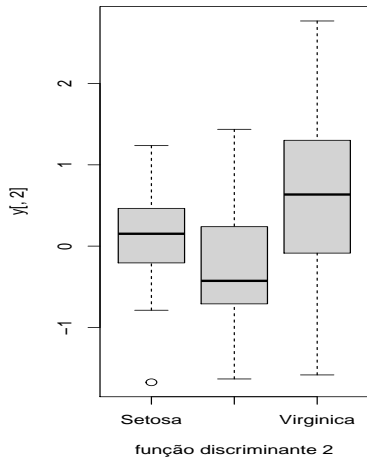
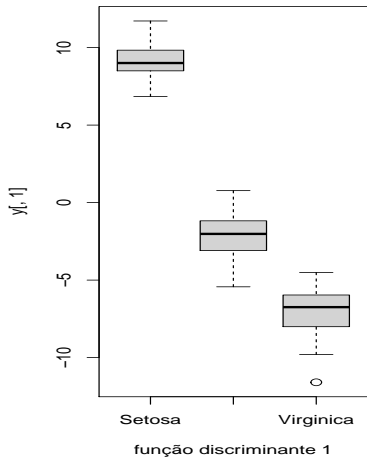
## Função discriminante 2

Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Setosa	0,11	0,66	0,44	-1,67	0,15	1,24	25
Versicolor	-0,31	0,72	0,51	-1,63	-0,43	1,44	25
Virginica	0,50	1,18	1,39	-1,58	0,63	2,77	25

# Dispersões entre as funções discriminantes

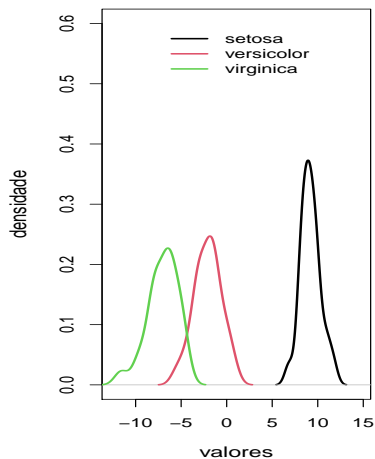


# Ex. 1: boxplots da função discriminante

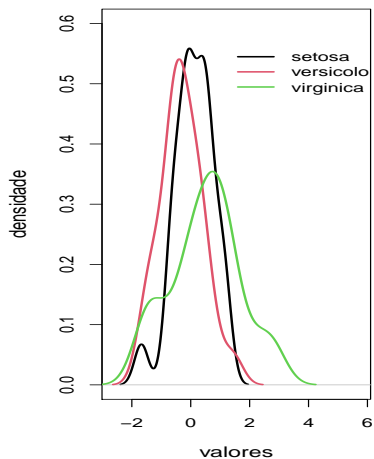


# Ex. 1: densidade estimada da função discriminante

Função discriminante 1



Função discriminante 2



# Comentários

- A função discriminante 1, como esperado, consegue distinguir melhor os grupos.
- Os grupos versicolor e virgínica são bem diferentes do grupo setosa e menos diferentes entre si.
- Em relação a função discriminante 1, o grupo setosa apresenta uma distribuição essencialmente positiva, enquanto que os dois outros grupos, negativa.
- A suposição de normalidade não parece ser razoável, para nenhuma das funções discriminantes e nenhum dos grupos.

# Referências I



T. W. Anderson.

*An introduction to multivariate statistical analysis.*

John Wiley, New York, 2003.



R. A. Johnson and D. W. Wichern.

*Applied multivariate statistical analysis.*

Prentice Hall, Upper Saddle River, 6 edition, 2007.