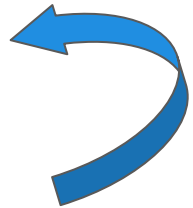
FIOCRUZ

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

# Análise Discriminante

Fernanda Lopes
Iasmim Almeida
Rafael França

# Análise Discriminante X Regressão Logística



Se ⇒ Variáveis independentes categóricas ou não multivariadas normais

E se ⇒ Grupos são inicialmente desconhecidos

Agrupar observações ou Agrupar variáveis

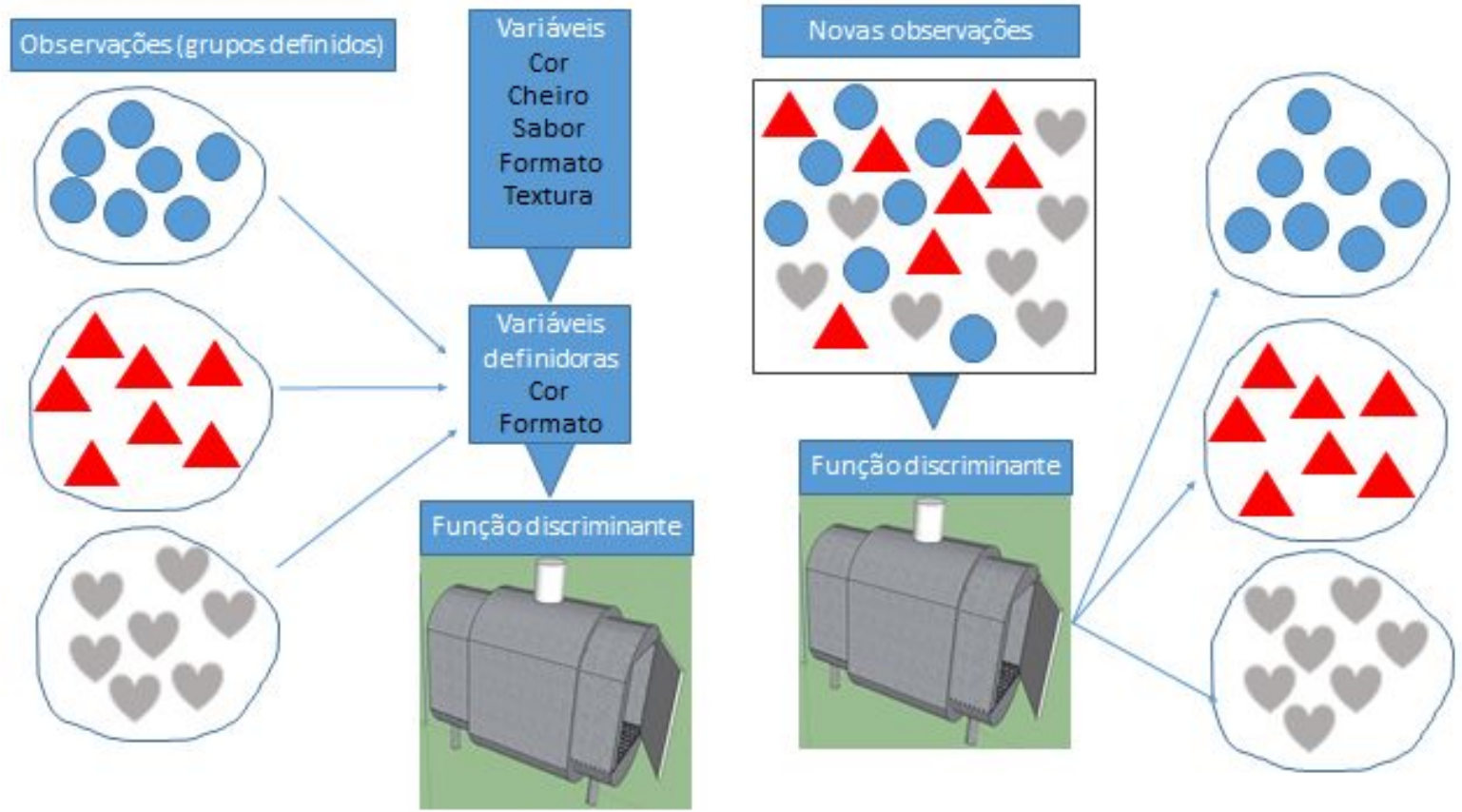
Então se ⇒ Variável dependente não métrica +
Grupos previamente conhecidos +
Variáveis independentes multivariadas normais +
Variáveis independentes métricas

Análise Discriminante

# Objetivos

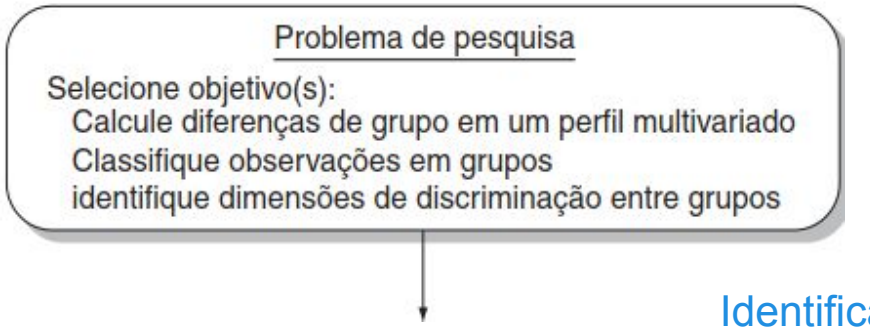
- Determinar o grau de precisão com que as observações são classificadas nos grupos conhecidos
- Avaliar como as variáveis preditoras distinguem os grupos
- Predizer os grupos para observações que têm grupos desconhecidos
- Obter uma função das variáveis que melhor faça a predição dos grupos (seleção de variáveis mais discriminatórias)

# Ilustrando


Observações (grupos definidos)
Variáveis
Cor
Cheiro
Sabor
Formato
Textura
Variáveis definidoras
Cor
Formato
Função discriminante
Novas observações
Função discriminante

# Estagio 1

O que realmente desejamos encontrar?




## Diferenças de grupo?

Determinar se existem diferenças no escore médio para dois (ou mais) grupos definidos a priori.

## Classificar observações em grupos?

Estabelecer procedimentos para classificar objetos em grupos com base em seus escores

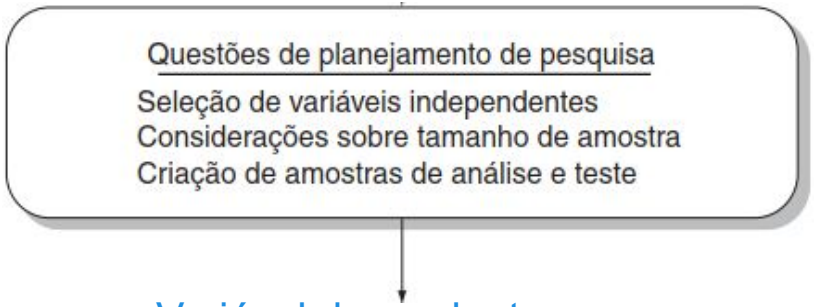
## Identificar dimensões de discriminação entre grupos?

Identificar dimensões de discriminação entre grupos

Quais variáveis independentes explicam o máximo de diferenças nos perfis de escore médio entre os grupos.

Estabelecer o número e a composição das dimensões de discriminação entre grupos

# Estagio 2




## Variável dependente

→ dois ou mais grupos (menor número possível)

→ mutuamente excludentes

→ conversão de variáveis métricas (Escala métrica e Extremos polares)

## Variável independente

→ pesquisa prévia ou modelo teórico

→ intuição

## Tamanho da amostra

Geral → Proporção de 20 observações para cada variável preditora (mínimo 5)

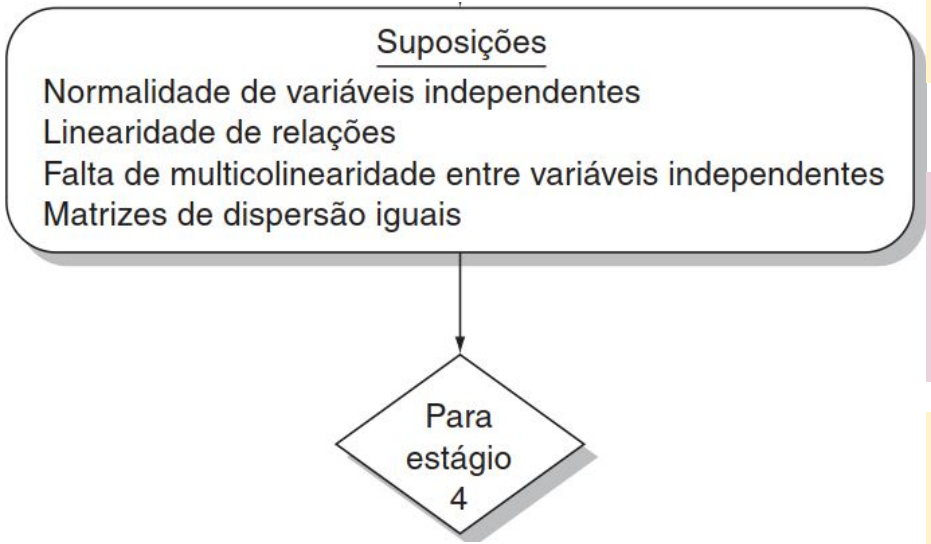
Categoria → Cada categoria deve ter no mínimo 20 observações (Exceder número de variáveis independentes).

Equilíbrio entre os tamanhos relativos das categorias

## Sub-amostras

Uma para estimação da função discriminante e outra para validação

# Estagio 3



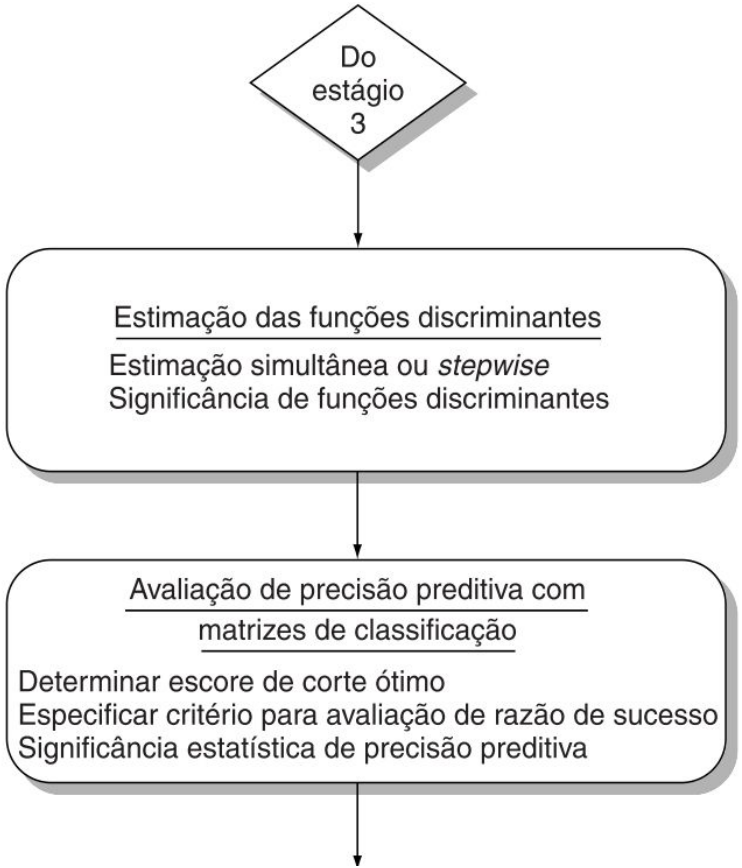

→ Normalidade - Impacta na estimação da função discriminante.

→ Linearidade - A não-linearidade impacta na interpretação dos resultados, pois as relações não são refletidas na função discriminante.

→ Ausência de Multicolinearidade - impacta na interpretação. Principalmente no uso de stepwise.

→ Matrizes de dispersão: suposição mais importante que afeta tanto a estimação quanto a classificação. (**M DE BOX**)

# Estagio 4




→ Estimação Simultânea - Variáveis independentes consideradas juntas

**Teste de Significância**: Lambda de Wilks, o traço de Hotelling e o critério de Pillai.

Raiz Característica de Roy (1ª F.D.)

→ Estimação Stepwise - inclusão com base no poder discriminatório.

**Teste de Significância:**

D2 de Mahalanobis e V de Rao

# Análise Discriminante - Geral

Função discriminante ⇒ Regra de classificação

Equação matemática

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk}$$

- $Z_{jk}$: Variável dependente
- $a$: Intercepto
- $W_1X_{1k}$: Variável independente 1
- $W_2X_{2k}$: Peso (coeficiente discriminante)
- $W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk}$: Escore

→ Avaliação do Ajuste Geral

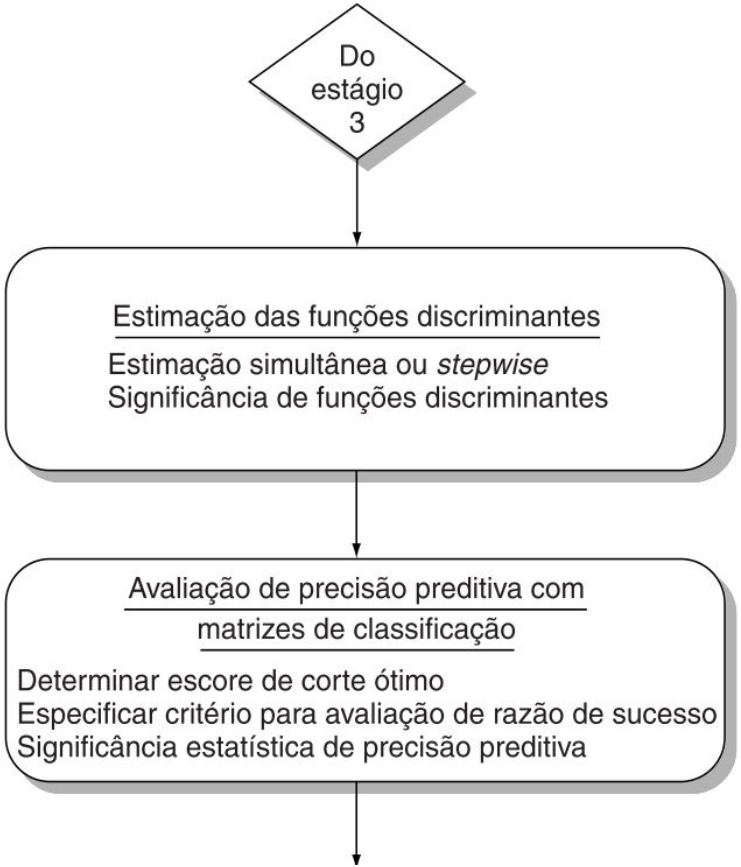
→ Calcular escores Z discriminantes para cada observação.

→ Calcular diferenças de grupos nos escores Z.

→ Avaliar a precisão de previsão de pertinência a grupos.

# Estagio 4

→ Valor Z crítico - centróide de dois grupos e tamanho relativo dos grupos



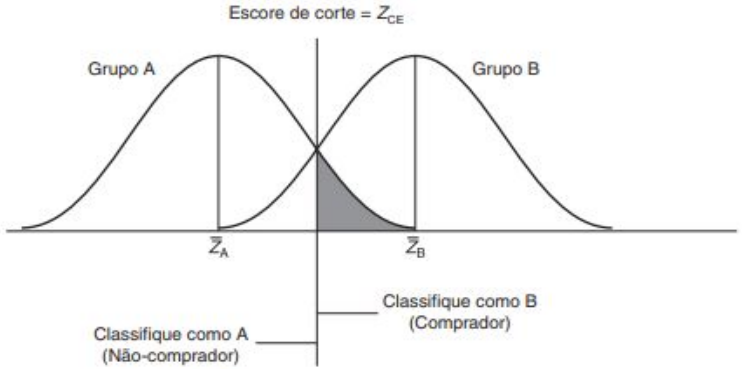




**FIGURA 5-7** Escore de corte ótimo com amostras de tamanhos iguais.

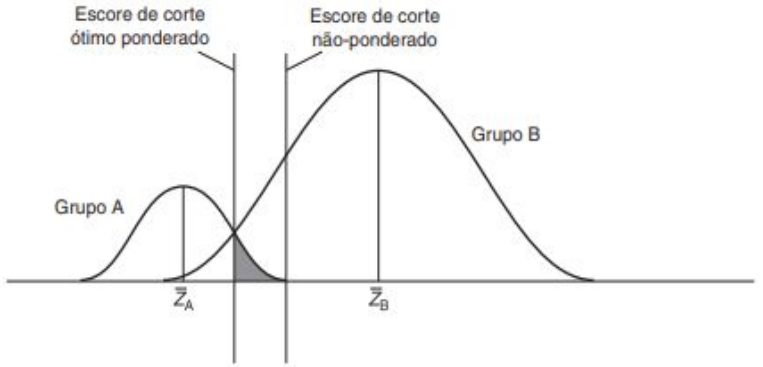


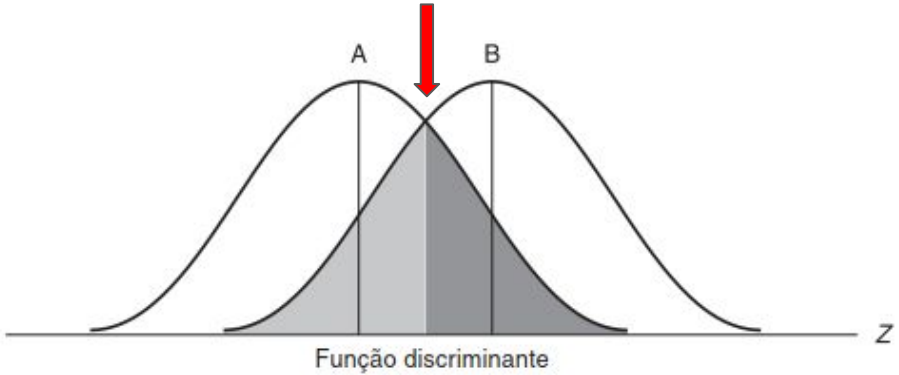

**FIGURA 5-8** Escore de corte ótimo com tamanhos desiguais de amostras.

# Análise Discriminante - Ilustrando

Teste de significância estatística ⇒ medida generalizada da distância entre os centróides de grupos.

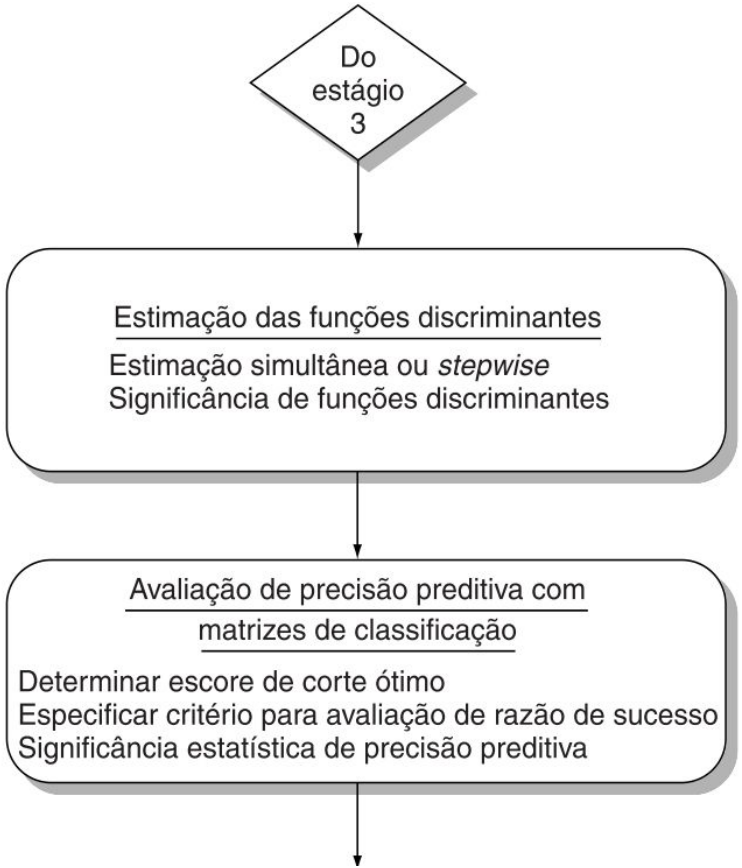
Compara-se as distribuições dos escores discriminantes para os grupos


Sobreposição pequena ⇒ função discriminante separa bem os grupos.




Sobreposição grande ⇒ função é um discriminador pobre entre os grupos.

# Estagio 4



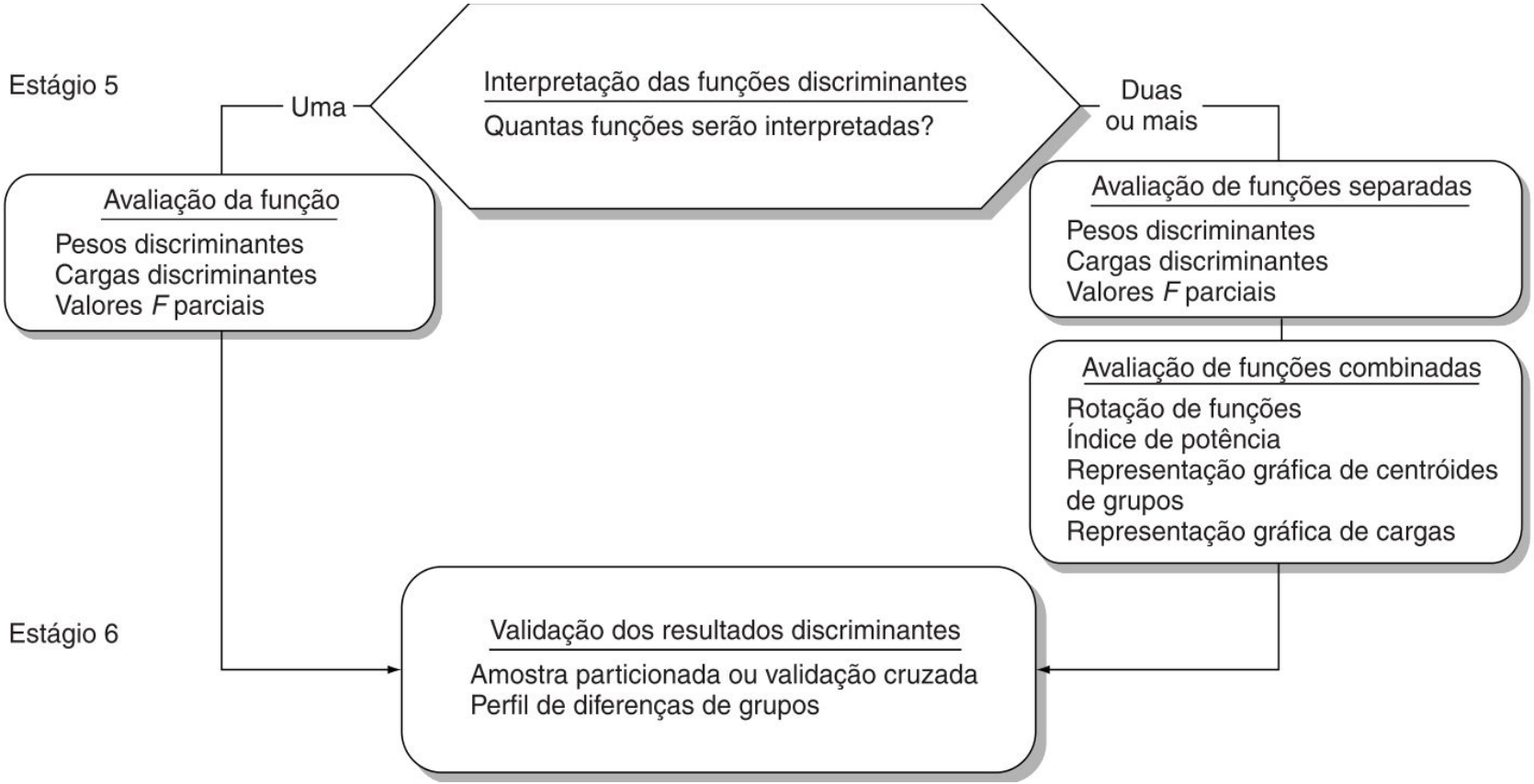

→ Construção de Matriz de Classificação.(amostra teste).

**TABELA 5-4** Matriz de classificação para análise discriminante de dois grupos

| | *Grupo previsto* | | | |
|---|---|---|---|---|
| **Grupo real** | **1** | **2** | **Tamanho do grupo real** | **Percentual corretamente classificado** |
| 1 | 22 | 3 | 25 | 88 |
| 2 | 5 | 20 | 25 | 80 |
| Tamanho previsto do grupo | 27 | 23 | 50 | 84[a] |

[a]Percentual corretamente classificado = (Número corretamente classificado/Número total de observações) x 100
= [(22 + 20)/50] × 100
= 84%

# Estagio 5 e 6


Estágio 5
Uma
Interpretação das funções discriminantes
Quantas funções serão interpretadas?
Duas ou mais
Avaliação da função
Pesos discriminantes
Cargas discriminantes
Valores F parciais
Avaliação de funções separadas
Pesos discriminantes
Cargas discriminantes
Valores F parciais
Avaliação de funções combinadas
Rotação de funções
Índice de potência
Representação gráfica de centróides de grupos
Representação gráfica de cargas
Estágio 6
Validação dos resultados discriminantes
Amostra particionada ou validação cruzada
Perfil de diferenças de grupos

# Análise discriminante no R

# Bibliotecas

```
## Carregando Biblioteca

```{r}
library(htmltools)
library(klaR) #para análise discriminante
library(psych) #para o gráfico de dispersão
library(MASS) #para a análise discriminante
library(devtools)
library(ggplot2)
library(ggord)
```
```

# Banco Iris

**Total de 150 observações e 5 variáveis contidas.**

- 4 variáveis são numéricas;
- 1 categórica as Species (setosa, versicolor e virginica).


**Iris Versicolor**

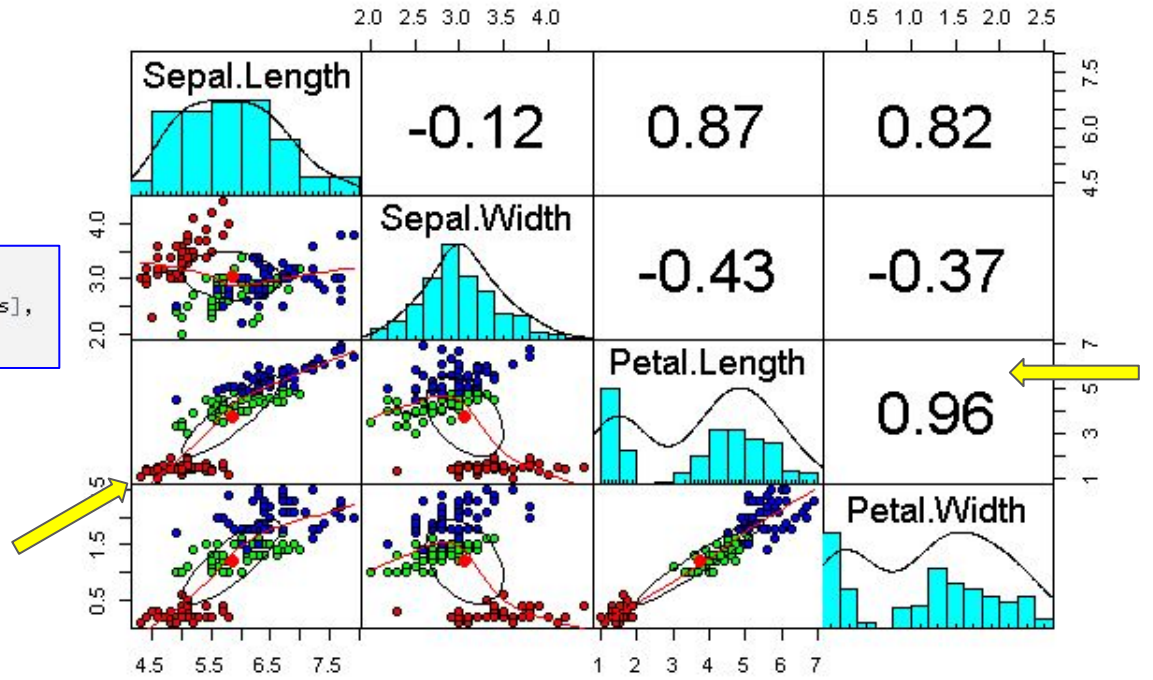
**Iris Setosa**

**Iris Virginica**

```
> str ( iris )
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

# 1. Observar os pressupostos de normalidade e linearidade dos dados.

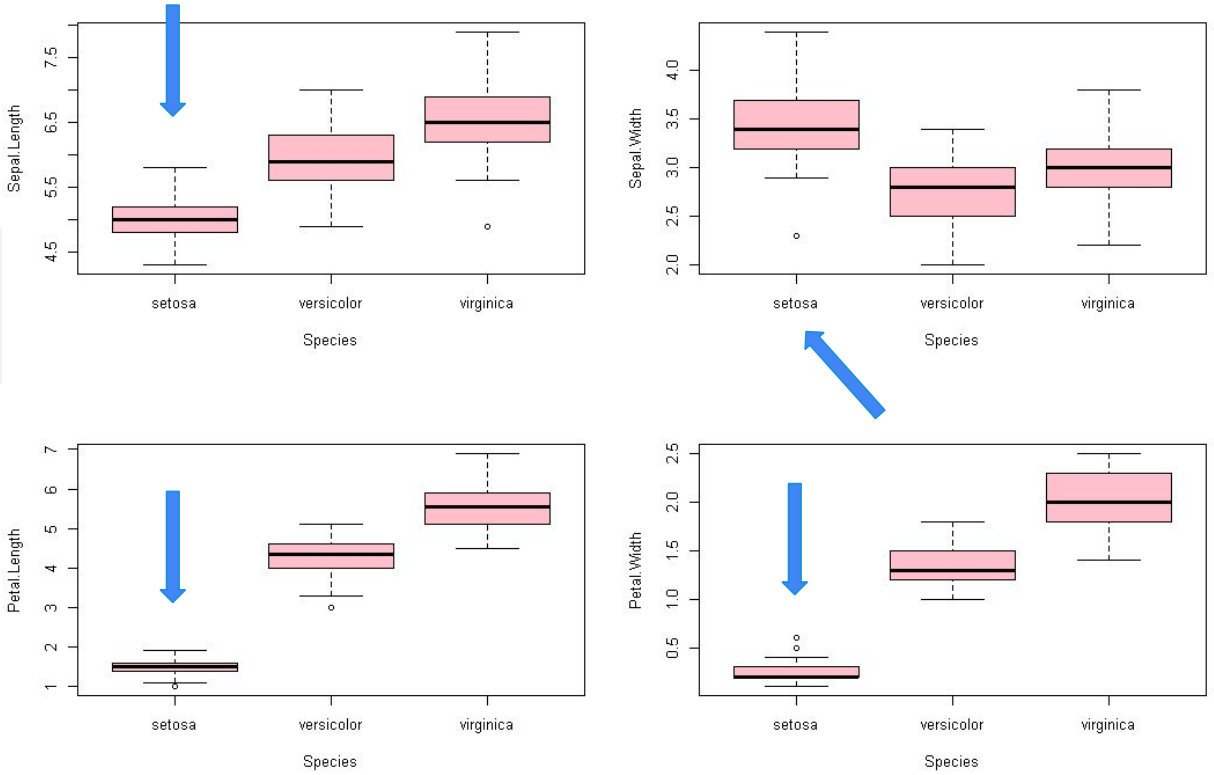
```
pairs.panels(iris[1:4],
             gap = 0,
             bg = c("red", "green", "blue")[iris$Species],
             pch = 21)
```



# 1. Observar os pressupostos de normalidade e linearidade dos dados.

```
par(mfrow = c(2, 2))

boxplot(Sepal.Length ~ Species, data = iris, col = "pink")
boxplot(Sepal.Width ~ Species, data = iris, col = "pink")
boxplot(Petal.Length ~ Species, data = iris, col = "pink")
boxplot(Petal.Width  ~ Species, data = iris, col = "pink")
```




Boxplot

# 2. Construção do modelo discriminante - Partição dos dados

```
set.seed(555)
ind <- sample(2, nrow(iris),
              replace = TRUE,
              prob = c(0.6, 0.4))
training <- iris[ind==1,]
testing <- iris[ind==2,]
```

| | |
|---|---|
| testing | 64 obs. of 5 variables |
| training | 86 obs. of 5 variables |

# 2. Construção do modelo discriminante

A probabilidade de um indivíduo desse banco ser classificado como:

- Setosa é de 38%,
- Versicolor é de 31%,
- Virgínica é de 30%.

É a média de cada variável em cada categoria da variável desfecho.

Mostra as variáveis que são mais relevantes para discriminar um indivíduo nas 3 categorias.

Componentes no modelo de análise discriminante.

No R

```
Call:
lda(Species ~ ., data = training)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3837209  0.3139535  0.3023256

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         4.975758    3.357576     1.472727   0.2454545
versicolor     5.974074    2.751852     4.281481   1.3407407
virginica      6.580769    2.946154     5.553846   1.9807692

Coefficients of linear discriminants:
                    LD1        LD2
Sepal.Length  1.252207 -0.1229923
Sepal.Width   1.115823  2.2711963
Petal.Length -2.616277 -0.7924520
Petal.Width  -2.156489  2.6956343

Proportion of trace:
   LD1    LD2
0.9937 0.0063
>
> attributes(linear)
$names
 [1] "prior"   "counts"  "means"   "scaling" "lev"     "svd"     "N"       "call"    "terms"   "xlevels"

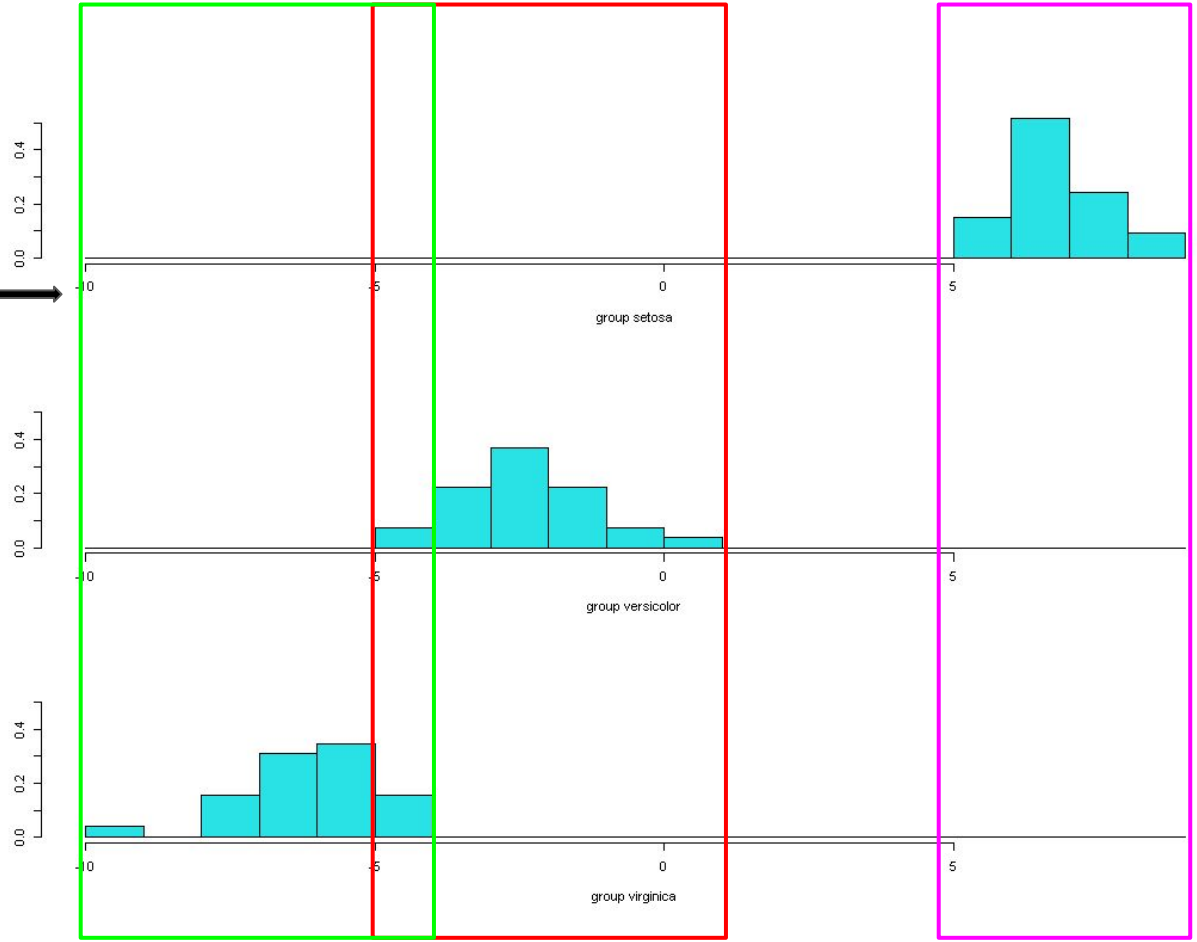
$class
[1] "lda"
```

A primeira função discriminante consegue discriminar ou separar 99,3% dos dados dentro das 3 categorias.

# 3. Analisar as dimensões de discriminação do grupo - LD1

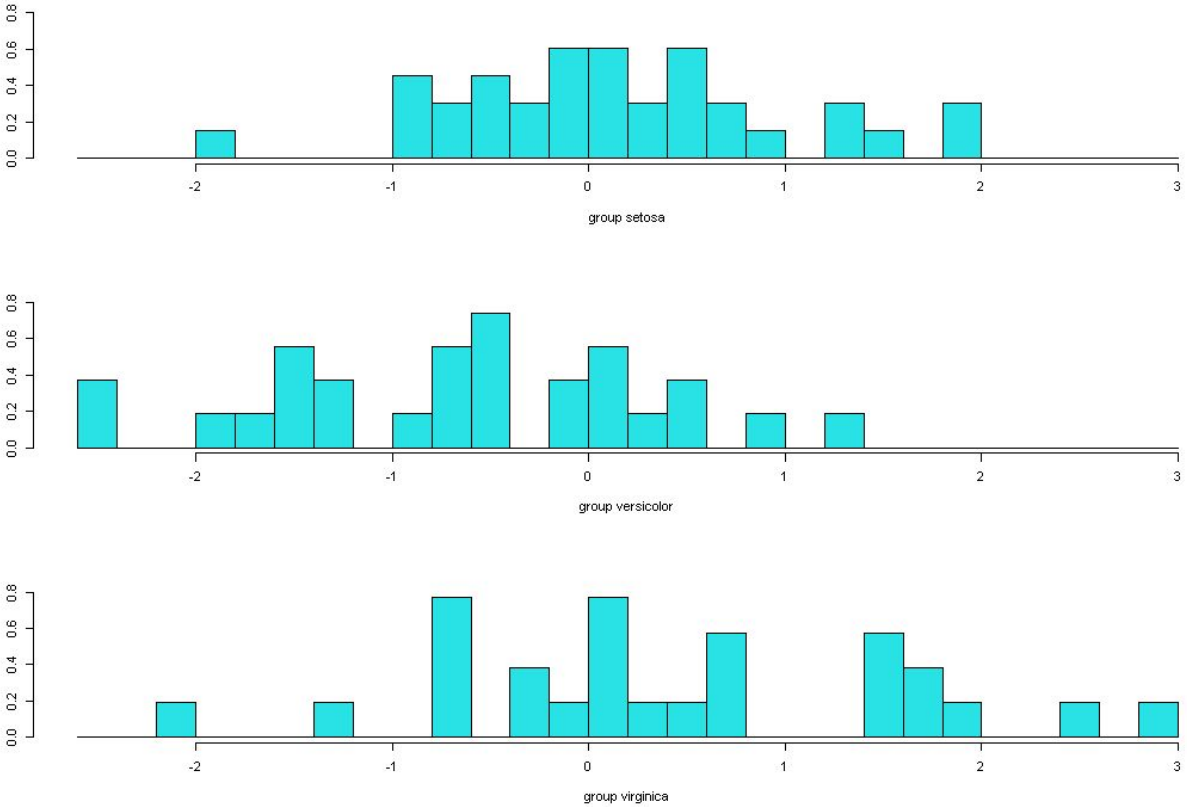
**Score discriminante**

```
p <- predict(linear, training)
ldahist(data = p$x[,1], g = training$Species)
```



# 3. Analisar as dimensões de discriminação do grupo - LD2

```
ldahist(data = p$x[,2], g = training$Species)
```



4. Observar a distribuição das categorias da variável desfecho na função discriminante.

5 .Observar as variáveis que discriminam melhor cada grupo.

**Biplot e Partition plot**

# Biplot

```
Call:
lda(Species ~ ., data = training)

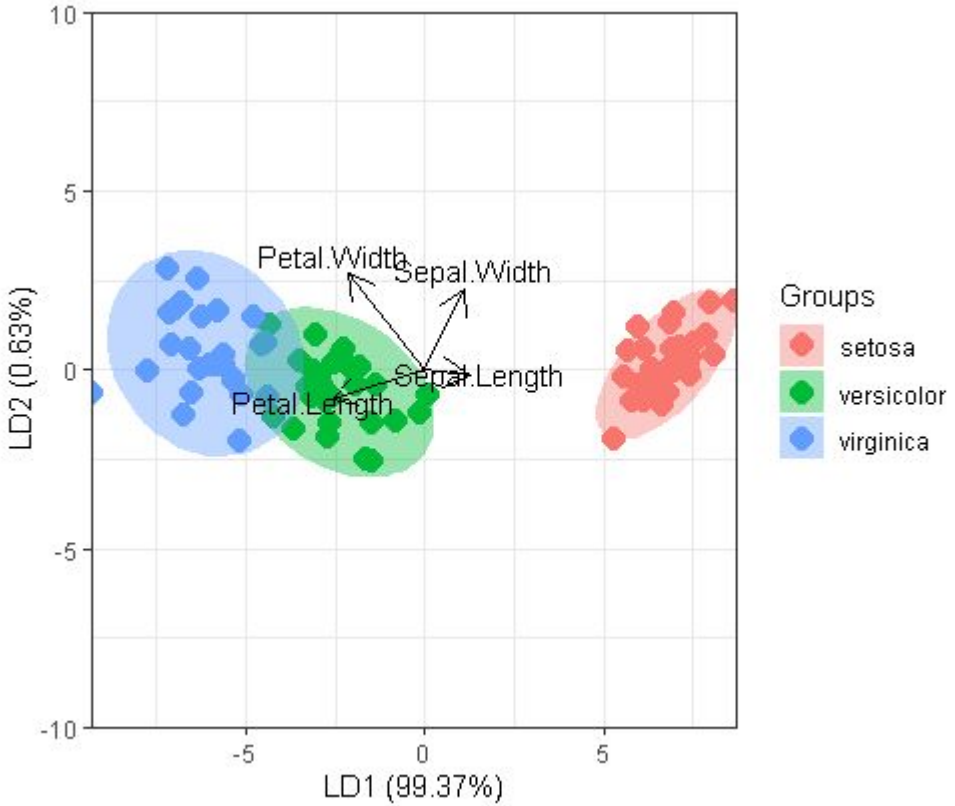
Prior probabilities of groups:
    setosa versicolor  virginica
 0.3837209  0.3139535  0.3023256

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         4.975758    3.357576     1.472727   0.2454545
versicolor     5.974074    2.751852     4.281481   1.3407407
virginica      6.580769    2.946154     5.553846   1.9807692

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  1.252207 -0.1229923
Sepal.Width   1.115823  2.2711963
Petal.Length -2.616277 -0.7924520
Petal.Width  -2.156489  2.6956343

Proportion of trace:
   LD1    LD2
0.9937 0.0063
```

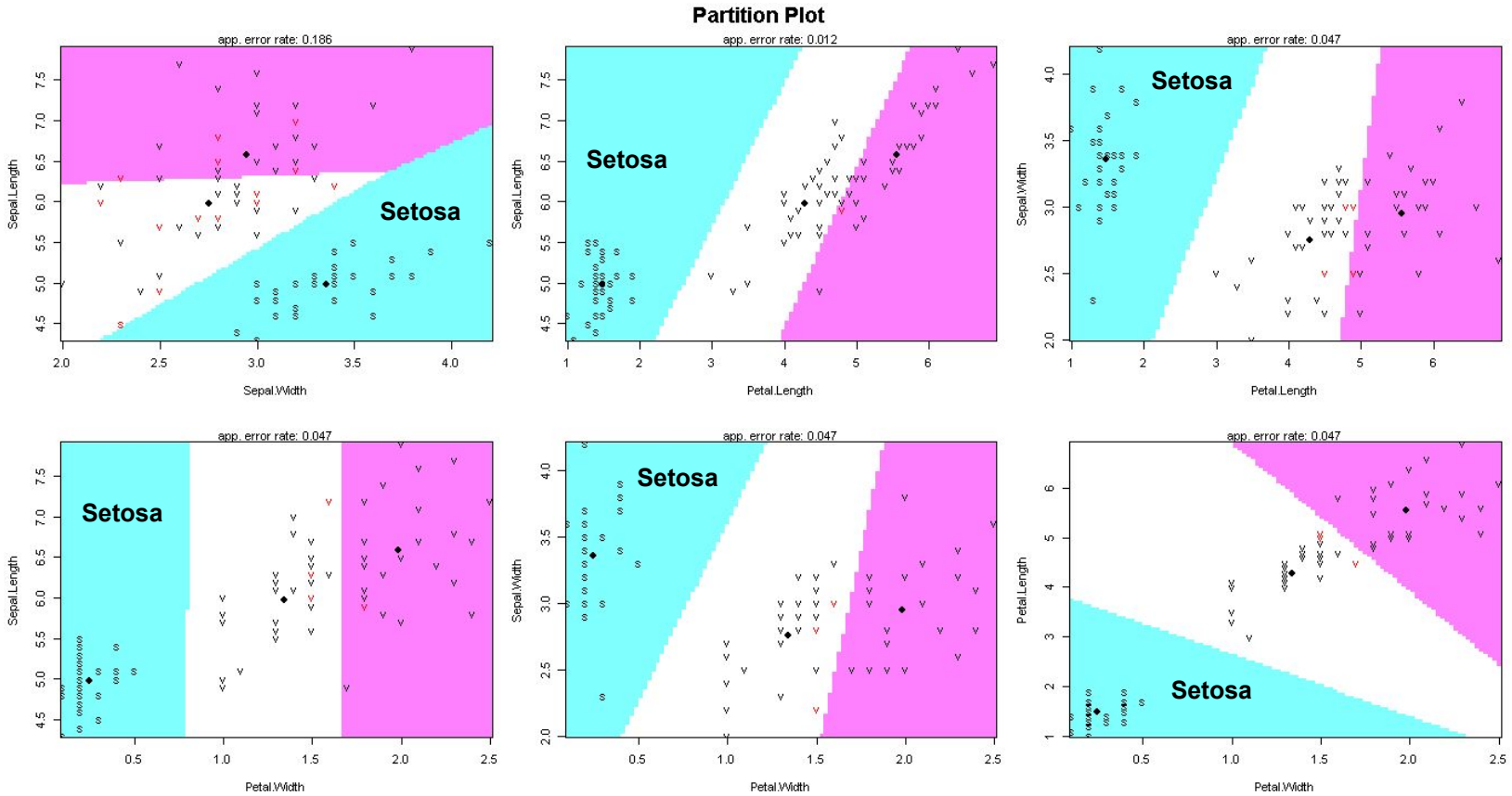
```
ggord(linear, training$Species, ylim = c(-10, 10))
```



# Partition plot

**Matriz de partição**

```
partimat(Species~., data = training, method = "lda")
```



# 6. Validar a discriminação das categorias da variável desfecho - Matriz de Confusão

**Treinamento**

```
p1 <- predict(linear, training)$class
tab <- table(Predicted = p1, Actual = training$Species)
tab

sum(diag(tab))/sum(tab)
```

```
            Actual
Predicted    setosa versicolor virginica
  setosa         33          0         0
  versicolor      0         26         1
  virginica       0          1        25
>
> sum(diag(tab))/sum(tab)
[1] 0.9767442
```

**Antes: 86**
**Agora: 84**

**Teste**

```
p2 <- predict(linear, testing)$class
tab1 <- table(Predicted = p2, Actual = testing$Species)
tab1
sum(diag(tab1))/sum(tab1)
```

```
            Actual
Predicted    setosa versicolor virginica
  setosa         17          0         0
  versicolor      0         22         0
  virginica       0          1        24
> sum(diag(tab1))/sum(tab1)
[1] 0.984375
```

**Antes: 64**
**Agora: 63**

# Artigo

# Tendências da epidemia de AIDS entre subgrupos sob maior risco no Brasil, 1980-2004

Trends in the AIDS epidemic in groups at highest risk in Brazil, 1980-2004

*Aristides Barbosa Júnior* [1]
*Célia Landmann Szwarcwald* [2]
*Ana Roberta Pati Pascom* [1]
*Paulo Borges de Souza Júnior* [2]

# Objetivo

Apresentar as tendências das taxas de incidência de AIDS entre os UDI e heterossexuais masculinos, os HSH e as mulheres, por meio das informações disponíveis no SINAN e das estimativas do tamanho destes subgrupos populacionais, realizadas a partir de inquérito de base populacional conduzido no país em 2004.

# Metodologia

**Dados → Duas fontes**

- **Inquérito** → para estimar a população sob risco

Pesquisa de Conhecimento, Atitudes e Práticas (PCAP-BR, 2004), inquérito de base populacional para investigação do conhecimento, práticas e comportamentos de risco relacionados à infecção pelo HIV e outras doenças sexualmente transmissíveis na população brasileira de 15 a 54 anos. (2004).

- **SINAN** → para avaliar as taxas de incidência de AIDS ao longo do tempo

Foram estabelecidas as incidências de AIDS entre os HSH, UDI do sexo masculino, heterossexuais masculinos e mulheres de 15 a 49 anos.

Os casos foram avaliados por sexo e período de diagnóstico (1980-1988, 1989-1992, 1993-1996, 1997-2000, 2001-2004). Os casos masculinos foram agrupados por categoria de exposição: HSH; heterossexuais; UDI; **categoria de exposição ignorada;** outros.

Percentual grande de categoria de exposição ignorada ⇒ tratamento dessa variável pela Análise discriminante.

# Metodologia - Análise Discriminante

- Aplicativo estatístico SPSS 13.0 (SPSS Inc., Chicago, Estados Unidos).
- São geradas duas funções discriminantes, combinações lineares das variáveis independentes que fornecem a melhor discriminação entre os grupos.
- Foram utilizadas as seguintes variáveis independentes:

(a) anos de estudo (0-3; 4-7; 8-11; 12+; ignorado);
(b) idade (15-24; 25- 29; 30-39; 40-49);
(c) múltiplas parcerias (sim; não; ignorado);
(d) parceiro tem múltiplas parcerias (sim; não; ignorado);
(e) parceiro é UDI (sim; não; ignorado);
(f) parceiro tem relações sexuais somente com homens (sim; não; ignorado);
(g) parceiro tem relações sexuais somente com mulheres (sim; não; ignorado);

(h) parceiro tem relações sexuais com homens e mulheres (sim; não; ignorado);
(i) período de diagnóstico (1980-1988; 1989-1992; 1993-1996; 1997-2000; 2001-2004);
(j) tamanho da população do município de residência (1-50.000; 50.001-200.000; 200.001-500.000; 500.001+);
(l) presença de sintomas: sarcoma de Kaposi, tuberculose (pulmonar ou disseminada), tosse;
(m) critério de definição de caso (Caracas; CDC; óbito); e
(n) sobrevida (< 5 meses; 5 meses +)

# Resultados

**Percentual de acerto 59,7%**

## Sem AD

Tabela 2

Incidência de AIDS e percentual do total de casos por sexo, período de diagnóstico e categoria de exposição. Brasil, 1980-2004.

| Categoria de exposição | Período de diagnóstico | | | | | Total |
|---|---|---|---|---|---|---|
| | 1980-1988 | 1989-1992 | 1993-1996 | 1997-2000 | 2001-2004 | |
| Sexo masculino | | | | | | |
| HSH | | | | | | |
| n | 4.570 | 12.684 | 16.427 | 19.151 | 17.303 | 70.135 |
| % | 54,6 | 33,0 | 22,3 | 19,3 | 17,6 | 22,1 |
| Heterossexual | | | | | | |
| n | 513 | 3.845 | 10.965 | 19.683 | 23.366 | 58.372 |
| % | 6,1 | 10,0 | 14,9 | 19,8 | 23,8 | 18,4 |
| UDI | | | | | | |
| n | 1.321 | 10.206 | 15.432 | 14.013 | 8.517 | 49.489 |
| % | 15,8 | 26,6 | 21,0 | 14,1 | 8,7 | 15,6 |
| Outra | | | | | | |
| n | 276 | 596 | 612 | 229 | 150 | 1.863 |
| % | 3,3 | 1,6 | 0,8 | 0,2 | 0,2 | 0,6 |
| Ignorada | | | | | | |
| n | 850 | 4.581 | 11.574 | 12.160 | 10.820 | 39.985 |
| % | 10,2 | 11,9 | 15,7 | 12,2 | 11,0 | 12,6 |
| Sexo feminino | | | | | | |
| n | 839 | 6.490 | 18.605 | 34.201 | 37.976 | 98.111 |
| % | 10,0 | 16,9 | 25,3 | 34,4 | 38,7 | 30,9 |
| **Total** | | | | | | |
| N | 8.369 | 38.402 | 73.615 | 99.437 | 98.132 | 317.955 |
| % | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |

HSH: homens que fazem sexo com homens; UDI: usuários de drogas injetáveis.

## Com AD

Tabela 3

Incidência de AIDS e percentual do total de casos por sexo, período de diagnóstico e categoria de exposição após reclassificação dos casos com categoria de exposição ignorada pela análise discriminante. Brasil, 1980-2004.

| Categoria de exposição | Período de diagnóstico | | | | | Total |
|---|---|---|---|---|---|---|
| | 1980-1988 | 1989-1992 | 1993-1996 | 1997-2000 | 2001-2004 | |
| Sexo masculino | | | | | | |
| HSH | | | | | | |
| n | 5.319 | 14.634 | 20.459 | 21.679 | 18.004 | 80.095 |
| % | 63,6 | 38,1 | 27,8 | 21,8 | 18,3 | 25,2 |
| Heterossexual | | | | | | |
| n | 555 | 4.162 | 12.956 | 24.267 | 32.636 | 74.576 |
| % | 6,6 | 10,8 | 17,6 | 24,4 | 33,3 | 23,5 |
| UDI | | | | | | |
| n | 1.380 | 12.520 | 20.983 | 19.061 | 9.366 | 63.310 |
| % | 16,5 | 32,6 | 28,5 | 19,2 | 9,5 | 19,9 |
| Outra | | | | | | |
| n | 276 | 596 | 612 | 229 | 150 | 1863 |
| % | 3,3 | 1,6 | 0,8 | 0,2 | 0,2 | 0,6 |
| Sexo feminino | | | | | | |
| n | 839 | 6.490 | 18.605 | 34.201 | 37.976 | 98.111 |
| % | 10,0 | 16,9 | 25,3 | 34,4 | 38,7 | 30,9 |
| **Total** | | | | | | |
| N | 8.369 | 38.402 | 73.615 | 99.437 | 98.132 | 317.955 |
| % | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |

HSH: homens que fazem sexo com homens; UDI: usuários de drogas injetáveis.

# Resultados

Variáveis discriminantes

- **Categoria UDI -**

  Ter parceiro UDI, períodos de diagnóstico 1989-1992 e 1993-1996, parceiro não tem múltiplas parcerias e ter tuberculose.

- **Categoria HSH -**

  Parceiro tem relações com homens e mulheres, escolaridade alta (ensino médio completo ou superior), períodos de diagnóstico 1980-1988 e 1989-1992, ter sarcoma de Kaposi, residir em municípios com 500 mil habitantes ou mais e grupos etários 30-39, 40-49.

- **Categoria Heterossexuais** -

  Múltiplas parcerias, parceiro só tem parceiros homens e morar em cidades pequenas (1-50 mil habitantes).

# Conclusões

- No período 1980-1988, os casos homossexuais ou bissexuais masculinos correspondiam a 63,6% dos casos, e a proporção de mulheres era de 10%.
- A análise da dinâmica da epidemia de AIDS no Brasil mostra a importância dos grupos HSH e UDI masculinos enquanto grupos de risco diferenciado.


Obrigadx!