**Serviço Público Federal**

**Universidade Federal do Pará**

**Instituto de Ciências Exatas e Naturais**

**Faculdade de Estatística**

Dionisio Alves da Silva Neto

Matrícula: 202007840008

**Atividade 4 de Análise Multivariada II**:

Análise do Discriminante Linear

Belém, PA

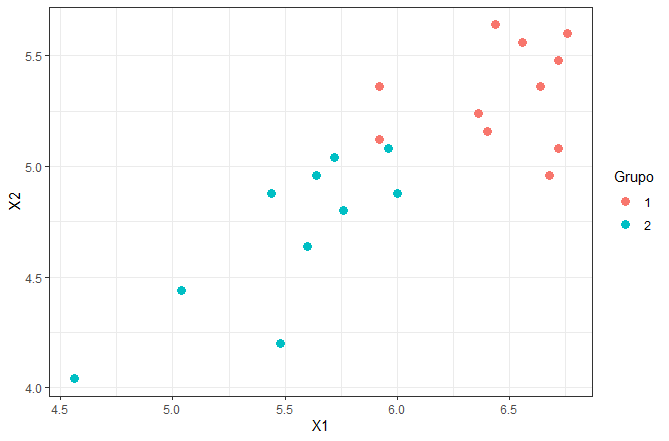
2022

1. **Realização da Análise de Discriminante Linear para os dados de Inseto**

**1. 1 Exploração dos dados**

Em uma análise descritiva inicial, a **Figura 1** aborda a dispersão entre as variáveis X1 e X2 do banco de dados. Desse modo, percebe-se a possibilidade de aplicar a técnica de Análise de Discriminante Linear por haver uma separação por uma reta entre os grupos 1 e 2.

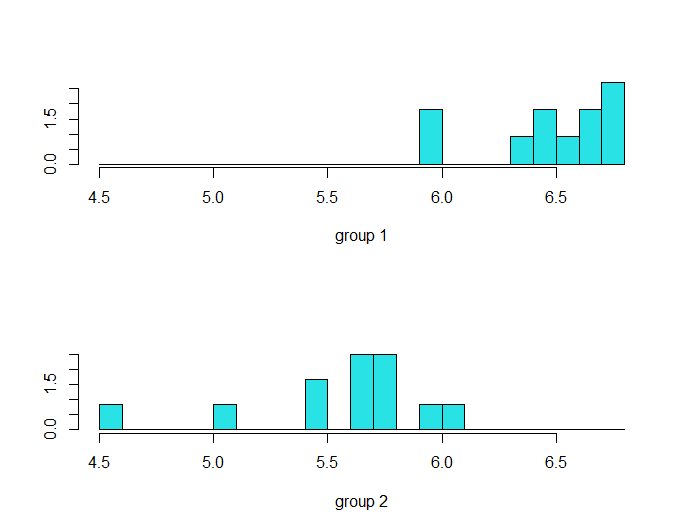
**Figura 1**: Dispersão entre as variáveis X1 e X2 do banco Inseto.



Fonte: Construído pelo autor, 2022.

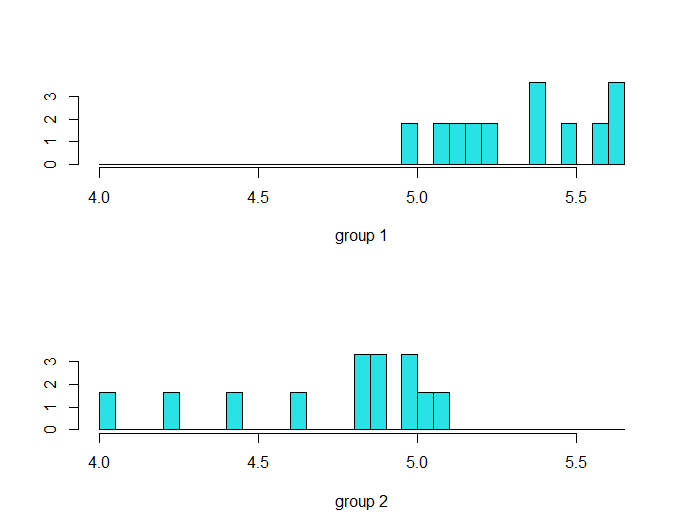
Nas **Figuras 2** e **3**, podemos visualizar o poder de separação entre os grupos 1 e 2 para as variáveis X1 e X2, respectivamente. Nesta perspectiva, para X1, na **Figura 2**, percebe-se distintamente que os valores para o grupo 1 estão acima de 6 e para o grupo 2 estão mais concentrados em valores abaixo de 6. Por outro lado, **na Figura 3**, para a variável X2, também percebe-se uma divisão entre as variáveis, para o grupo 1 existe uma maior agregação em valores acima de 5 e para o grupo 2, uma maior concentração em valores abaixo de 5.

**Figura 2**: Distribuição da variável X1, por grupo.



Fonte: Construído pelo autor, 2022.

**Figura 3**: Distribuição da variável X2, por grupo.



Fonte: Construído pelo autor, 2022.

**1.2 Construção do modelo de Discriminante Linear para o banco de dados Inseto**

No ajuste do modelo de Discriminante Linear para os dados foram geradas as probabilidades a priori de cada observação pertencer aos grupos 1 e 2, disponíveis na Tabela 1. Vale ressaltar que este cálculo é apenas uma definição frequentista de probabilidade, a qual baseia-se no balanceamento das classes alvo do método.

**Tabela 1:** Probabilidades a priori das observações pertencerem às classes 1 e 2.

| Grupo | Probabilidade a priori |
| --- | --- |
| 1 | 0,4783 |
| 2 | 0,5217 |

Fonte: Construído pelo autor, 2022.

Ademais, temos que as médias para as variáveis X1 e X2 foram calculadas no intuito de verificar se era plausível aplicar a técnica do Discriminante Linear. Logo, por haver uma certa distância entre os valores médios para cada variável entre os grupos, concluímos que é possível construir um separador linear entres os pares de variáveis.

**Tabela 2:** Médias das variáveis X1 e X2, por grupo.

| Grupo | X1 | X2 |
| --- | --- | --- |
| 1 | 6,4665 | 5,3236 |
| 2 | 5,5500 | 4,7267 |

Fonte: Construído pelo autor, 2022.

De forma complementar, foram calculadas as matrizes de variâncias covariâncias para cada grupo, de forma a observar o comportamento da variabilidade entre as variáveis do estudo em cada grupo.

**=**

|  |  | **X1** | **X2** |  |
| --- | --- | --- | --- | --- |
| **X1** |  | **0,09128727** | **0,01125818** |  |
| **X2** |  | **0,01125818** | **0,05262545** |  |

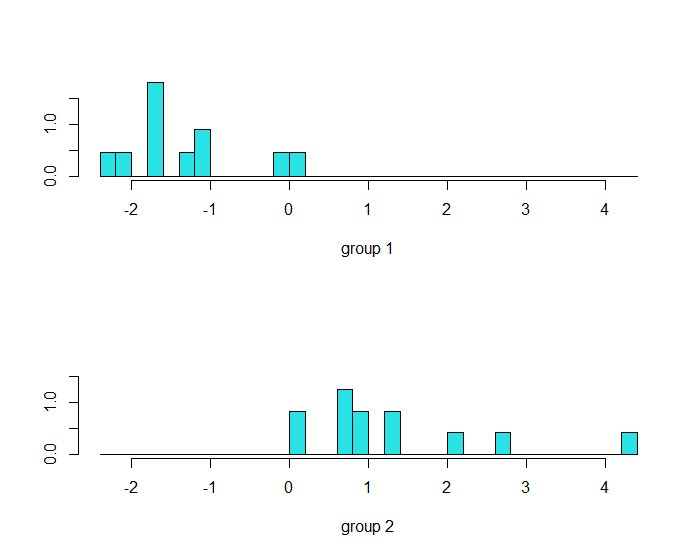
**=**

|  |  | **X1** | **X2** |  |
| --- | --- | --- | --- | --- |
| **X1** |  | **0,1603273** | **0,1074182** |  |
| **X2** |  | **0,1074182** | **0,1116606** |  |

Em resultado do ajuste, temos a equação abaixo, a qual exemplifica a equação da reta que separa as classes 1 e 2 em um espaço bidimensional para as variáveis X1 e X2. É preciso ter em mente que os valores dos coeficientes estão padronizados e, por isso, deve-se padronizar uma nova observação para a predição.

Um dos objetivos da análise do discriminante Linear, é poder representar o espaço de variáveis originais em um novo conjunto. Nesta visão, a Figura 4 aborda os histogramas da distribuição do Discriminante Linear construído. Portanto, podemos perceber que mesmo em outra escala, as classes 1 e 2 parecem estar divididas pela equação desenvolvida, evidenciando a qualidade do ajuste do modelo.

**Figura 4**: Distribuição do escore do discriminante, por grupo.

****

Fonte: Construído pelo autor, 2022.

1. **Realização da Análise de Discriminante Linear para os dados de sítios**

Para o banco de dados sítios, não foi possível encontrar uma maneira de realizar o método de seleção stepwise considerando a Distância de Mahalanobis como critério de seleção. Dentro do pacote *klaR* , é apenas viável aplicar o método stepwise considerando o critério do Lambda de Wilks. Entretanto, a análise por este outro meio será de suma importância, pois poderemos comparar os resultados gerados pelos critérios Lambda de Wilks e Distância de Mahalanobis.

De acordo com a **Tabela 3**, percebemos que as variáveis selecionadas pelo método stepwise, considerando o critério Lambdas de Wilks foram as variáveis: Altura, KiB, KiA, Areia B, SiO2A, KrA e ProfA.

**Tabela 3:** Resultados da seleção de variáveis pelo método stepwise, considerando o critério Lambda de Wilks.

| Posição | Variável | Lambda de Wilks | Estatística F | P-valor | Estatística F parcial | P-valor parcial |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Altura | 0.084975657 | 96.91269 | 2.310206e-10 | 96.912685 | 2.310206e-10 |
| 2 | KiB | 0.062969729 | 25.37295 | 8.519604e-10 | 2.970481 | 7.676763e-02 |
| 3 | KiA | 0.047932390 | 19.02705 | 2.687851e-09 | 2.509758 | 1.109046e-01 |
| 4 | Areia B | 0.038015717 | 15.48312 | 9.969476e-09 | 1.956429 | 1.737365e-01 |
| 5 | SiO2A | 0.015247825 | 19.87535 | 3.594665e-10 | 10.452326 | 1.435668e-03 |
| 6 | KrA | 0.007580025 | 22.71943 | 9.286617e-11 | 6.575269 | 9.692755e-03 |
| 7 | ProfA | 0.005325707 | 21.77634 | 2.788774e-10 | 2.539740 | 1.171988e-01 |

Fonte: Construído pelo autor, 2022.

Posteriormente à seleção de variáveis, foi aplicado o método de Análise de Discriminante Linear, considerando apenas as variáveis selecionadas.

Resultados do ajuste do Discriminante Linear para as variáveis selecionadas:

**Tabela 4:** Probabilidades a priori de cada classe.

| Classe 1 | Classe 2 | Classe 3 |
| --- | --- | --- |
| 0,2380952 | 0,3809524 | 0,3809524 |

Fonte: Construído pelo autor, 2022.

**Tabela 5:** Valores médios de cada variável, por grupo.

| Grupo | Altura | KiB | KiA | AreiaB | SiO2A | KrA | ProfA |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 10.0400 | 1.720 | 1.7600 | 38.400 | 12.6400 | 1.3200 | 49.000 |
| 2 | 15.2875 | 0.875 | 1.1375 | 19.250 | 10.2875 | 0.8625 | 49.750 |
| 3 | 19.5750 | 1.200 | 1.2625 | 29.625 | 12.2625 | 0.8375 | 46.125 |

Fonte: Construído pelo autor, 2022.

**Equação do Primeiro Discriminante Linear**

**Equação do Segundo Discriminante Linear**

Proporção de variabilidade explicada por cada Discriminate:

| Discriminante | Proporção de variância explicada |
| --- | --- |
| 1 | 0,9518 |
| 2 | 0,0482 |

**Anexo I**: Script R desenvolvido

## ------

## Analise Multivariada

## Analise de Discriminate Linear

## ------

## ---

## pacotes

## ---

if(!require(pacman)) install.packages("pacman"); library(pacman)

p\_load(readxl, MASS, caret, e1071, ggplot2, dplyr, ggord, GGally)

## ---

## Banco de dados 1

## Insetos

## ---

setwd('D:\\MUlti2-Discrim\\')

insetos = read\_excel("insetos.xlsx")

head(insetos)

str(insetos)

insetos$Grupo = as.factor(insetos$Grupo) ## target

table(insetos$Grupo)

plot(insetos$X1, insetos$X2)#, col = insetos$Grupo["deeppink", "steelblue"])

## visualizando a dispersao entre as variaveis

ggplot(data = insetos, aes(x = X1, y = X2, color = Grupo)) +

geom\_point(size = 3) +

theme\_bw()

## vetor de medias para cada grupo

insetos %>% filter(Grupo == 1) %>% summary()

insetos %>% filter(Grupo == 1) %>% var(.)

insetos %>% filter(Grupo == 2) %>% summary()

insetos %>% filter(Grupo == 2) %>% var(.)

## Verificar o quao bom cada variavel e bom para discriminar a variavel target

## X1

ldahist(data = insetos$X1, g = insetos$Grupo)

## X2

ldahist(data = insetos$X2, g = insetos$Grupo)

## Ajuste da funcao discriminante

disc.linear = lda(Grupo ~ (X1 + X2), data = insetos)

summary(disc.linear)

## visualizando o escore do escriminate para as duas populacaoes

## quanto mais distantes, maior estarao discriminadas

plot(disc.linear)

disc.linear

## probbilidades a priori

## valores de medias para cada variavel

## coeficientes de cada variavel para o discriminante linear

predicoes = predict(object = disc.linear, newdata = insetos)

confusionMatrix(predicoes$class, insetos$Grupo)

# exemplo

#classificar um novo individuo que apresenta numero

#medio de cerdas primordiais e distais de 6,21 e 5,31,

#respectivamente.

pred.disc = predict(disc.linear, newdata = list(X1 = 6.21, X2 = 5.31))

pred.disc$x

## ---

## Banco de dados 2

## Sitio

## ---

p\_load(klaR, biotools)

sitio = read\_excel("Sitio.xlsx")

head(sitio)

#View(sitio)

## Target

sitio$Classe = as.factor(sitio$Classe)

ggpairs(sitio, aes(color = Classe, # Color by group (cat. variable)

alpha = 0.5),

upper = list(na = "na"))

all.disc.linear = greedy.wilks(Classe~., data = sitio, niveau = 0.15)

all.disc.linear

step.disc.linear2 = greedy.wilks(Classe~., data = sitio, method = "moment")

## aplicacao do lda nos dados

step.var = step.disc.linear2$results[,1]

lda.sitio = lda(Classe ~ (Altura + KiB + KiA + AreiaB +

SiO2A + KrA + ProfA), data = sitio)

lda.sitio

pred.sitio = predict(lda.sitio)

ggplot(data = pred.sitio$x, aes(x = LD1, y = LD2)) +

geom\_point(size = 3) +

theme\_bw()