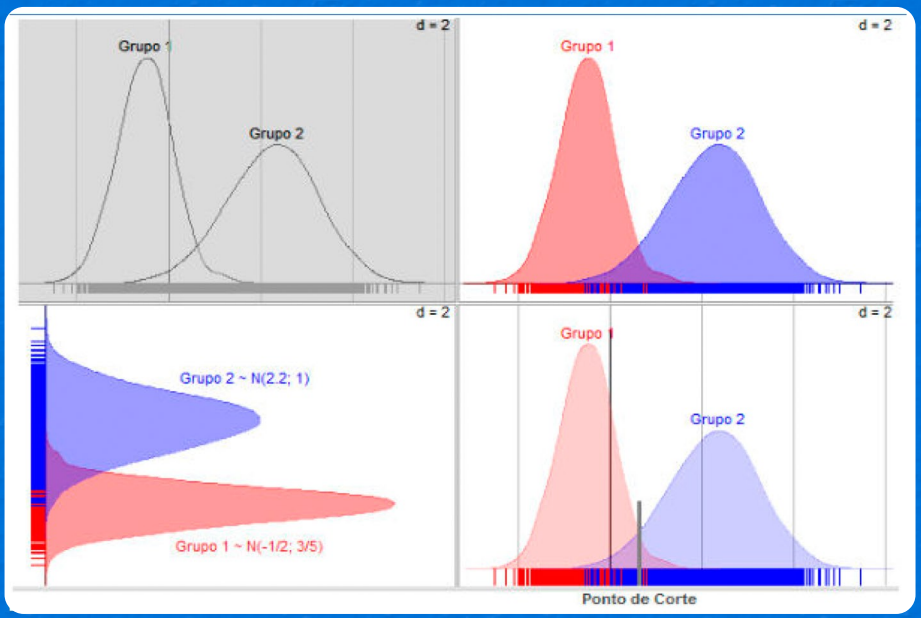
Estudo dos Principais Pressupostos de Análise Discriminante Simples

com aplicação em R

Edwirde Luiz Silva Camêlo | Andrés González Carmona
Ramón Gutiérrez Sánchez | Dalila Camêlo Aguiar



eduepb

EDITORA DA UNIVERSIDADE ESTADUAL DA PARAÍBA
Rua Baraúnas, 351 - Bairro Universitário - Campina Grande-PB - CEP 58429-500
Fone/Fax: (83) 3315-3381 - http://eduepb.uepb.edu.br - email: eduepb@uepb.edu.br

Edwirde Luiz Silva Camêlo
Ramón Gutiérrez Sánchez
Andrés González Carmona
Dalila Camêlo Aguiar

Estudo dos Principais Pressupostos de Análise Discriminante Simples
com aplicação em R

eduepb
Campina Grande - PB
2020

Editora da Universidade Estadual da Paraíba

Luciano Nascimento Silva | *Diretor*
Antonio Roberto Faustino da Costa | *Editor Assistente*
Cidoval Morais de Sousa | *Editor Assistente*

Expediente EDUEPB

Erick Ferreira Cabral | *Design Gráfico e Editoração*
Jefferson Ricardo Lima Araujo Nunes | *Design Gráfico e Editoração*
Leonardo Ramos Araujo | *Design Gráfico e Editoração*
Elizete Amaral de Medeiros | *Revisão Linguística*
Antonio de Brito Freire | *Revisão Linguística*
Danielle Correia Gomes | Divulgação

Depósito legal na Biblioteca Nacional, conforme decreto nº 1.825, de 20 de dezembro de 1907.

P957 Estudo dos principais pressupostos de análise discriminante simples.[Livro eletrônico]./Edwirde Luiz Silva Câmelo...[et al.]. Campina Grande: EDUEPB, 2020.
1200 Kb. - 151 p.: il. color.

ISBN 978-85-7879-600-6

1. Análise discriminante simples. 2. Pressupostos básico. 3. Aplicação dos conceitos. 4. Aplicações em R. I. Camêlo, Edwirde Luiz Silva. II. Sánchez, Ramón Gutiérrez. III. Aguiar, Dalila Camêlo. IV. Carmona, Andrés González

21 ed. CDD 512.2

Ficha catalográfica elaborada por Heliane Maria Idalino Silva – CRB-15ª/368

Lista de Figuras

Capítulo 1

Introdução ao *R*

O *R* é um software estatístico, com ambiente integrado e com linguagem de programação especialmente desenvolvida para a análise de dados, cálculos estatísticos e representações gráficas.

É uma linguagem de programação muito simples, disponibilizada para diferentes plataformas (*Unix, MacOS, Windows*) e de fácil instalação. O melhor de tudo isso, é que se trata de um software gratuito e amplamente utilizado na pesquisa científica

Primeiro, você precisa instalar o *R* e, para isso, faça o *download* do site oficial http://cran.at.r-project.org/, de preferência a última versão estável, 2.15.0, clicando no *Windows* e depois no link *base* e, a partir daí, baixamos *R-2.15.0-win.exe*. E agora é só seguir os passos solicitados, em caso de dúvidas, pode consultar a instalação passo a passo em https://cran.r-project.org/doc/contrib/Itano-installation.pdf.

Após a instalação do *R*, instala-se o *RStudio*, que é um software livre de ambiente de desenvolvimento integrado ao *R*. É um ambiente mais amigável do que *R* para se trabalhar. Sua instalação

também simples e recomenda-se baixar a versão mais estável. O download pode ser feito em

https://www.rstudio.com/products/rstudio/download/.

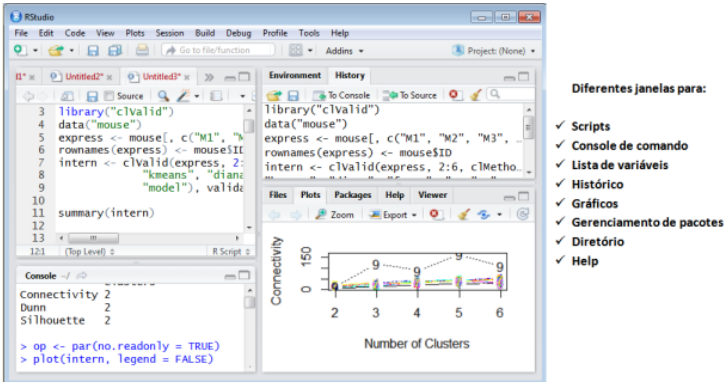


Figura 1.1: Interface de usuário para *RStudio*

Algumas vantagens do *RStudio*:

- Uma interface que permite uma simples manipulação, incluindo a visualização ou modificação dos dados brutos.
- Uma interface entre o usuário e os dados que podem executar várias operações ou aplicar vários testes para dados, além de apresentar resultados de diversas maneiras (gráficos, tabelas, etc).
- Para evitar a multiplicação de aplicações de software para a obtenção de dados utilizados na análise estatística.
- Para dados em 3D os dispositivos gráficos devem permitir alguma interatividade. Uma vez que a tela é plana, é preciso observar e girar dados 3D de uma forma convivial.

- Para obter resultados e armazená-los sob a forma de gráficos e tabelas.
- Para se adaptar e converter arquivos que foram tratados em outros softwares e arquivos de exportação para usuários que ainda relutam em usar o programa, assim como criar funções personalizadas.

Pode-se obter ajuda do próprio programa digitando o comando `help()`. Uma breve introdução ao uso do programa *R* pode ser encontrada em (LANDEIRO, 2011) no site https://cran.r-project.org/ em contribuição. Outra opção é ler o manual *R for Beginners* (PARADIS, 2005).

Para usar o *R* é necessário conhecer e digitar comandos. Em seguida apresentaremos alguns tópicos necessários.

1.1 Objetos

Os objetos são caracterizados por seus atributos. O modo e duração são atributos de todos os objetos em *R*. Se os elementos são dados, eles podem ter quatro modos diferentes: numérico, caráter, complexo e lógica (`FALSE` ou `TRUE`, alternativamente digitado como `F` ou `T`). O comprimento é o número de elementos do objeto e é devolvido digitando comprimento `lenght`(*objeto*). Finalmente, a função `str` mostra a estrutura interna de um objeto.

Exemplo 1.1.1. *Numérico, complexo, categórico e lógico*

```
a<-777; b<-3+4i; c<-"CD"; d<-TRUE
mode(a); mode(b); mode(c);  mode(d)
```

Também é possível verificar e coagir o modo de objetos usando as funções: `is.numeric`, `is.complex`, `is.character`, `is.logical` e `as.numeric`, `as.complex`, `as.character`, `as.logical`. O `"is"` antes dos modos significa uma pergunta, por exemplo o número dois é numérico? (`is.numeric(2)`), a resposta será `TRUE` (verdadeiro). E o `"as"` antes dos modos em algumas vezes significa transformar um modo, por exemplo, como caracter, `as.character(3)`, embora o 3 seja um numérico se torna aqui como um caracter, e para identificar entre aspas, "3".

```
#SE workspace está vazio?
ls()
# Se seu diretório de trabalho é o desejado?
verifique que está vazio, com o comando:
dir()
#Salve-o usando o comando:
save.image()
```

1.2 Encontrando uma função no *R*

Através do web site (http://rseek.org/) pode-se encontrar qualquer função no *R*. No quadro abaixo tem-se alguns prováveis links de ajuda no *R*.

http://www.r-project.org	*R* web site
http://www.cran.r-project.org	Downloads
http://www.rseek.org	Buscador de função
http://www.cran.r-project.org/web/views	Organizado por tarefa
www.tolstoy.newcastle.edu.au/R/	Discussão sobre o *R*

Através do comando `"apropos"`, é possível encontrar algumas funções que foram carregadas com os pacotes instalados. Também é possível pesquisar a documentação entre todos os pacotes instalados usando o comando `"help.search"`.

apropos("read")	Funções que se iniciam com *read*
apropos("mult")	Funções que se iniciam com *mult*
help.search(".matrix")	Funções que se iniciam com *matrix*
apropos(".test")	Busca as funções que terminan com .test

```
apropos(plot)
help.search(field="title","skew")
example(mean)
args(chisq.test) #lembrar do argumento da função
```

1.3 Instalando pacotes no *R*

Pacotes são conjuntos de funcionalidades (funções, dados e exemplos) distribuídos em conjunto para realizar tarefas específicas. Por exemplo, o pacote base carrega na sua área de trabalho (deixa disponível para uso) um conjunto de ferramentas básicas no *R*. É necessário entender as diferenças entre baixar (*download*) o pacote do repositório e carregar em sua área de trabalho. Para baixar algum pacote disponível no repositório CRAN do *R* é necessário utilizar o comando `install.packages`("pacote"). O *R* possui muitos pacotes (http://www.r-project.org/).

1.4 Operações básicas em vetores e matrizes

É muito fácil realizar operações básicas no *R*. A Tabela 1.1 mostra alguns exemplos.

```
A<- 3 + 5 + 3; A
B<- 3-8-7; B
C<- 3*4
D<- 9/7
2 + 6 # forma direta e o resultado
```

Tabela 1.1: Algumas operações básicas no *R*.

Log	$log_{base}(x)$	$log_{10}(2)$	log(2,base =10)=0.301
Raiz	sqrt(x)	$\sqrt{4}$	sqrt(4) =2
Log exp	$\log_e(2)$	$\log(2, exp(1))$	$\log(2, base = exp(1)) = 0.693$
sen(x)	$seno(\pi/4)$	$sen(\pi/4)$	$\sin(pi/4) = 0,707$

O *R* tem uma sintaxe de expressão aritmética convencional com aritmética habitual e operadores condicionais.

```
help(Arithmetic)
help(Comparison)
help(Syntax)
```

Os operadores aritméticos e condicionais são:

a + b	soma	a == b	a é igual a b?
a - b	divisão	a ! = b	a não é iqual a b?
a*b	multiplicação	a < b	a é menor que b?
a/b	divisão	a <= b	a é menor e igual a b?
a^b	Potenciação	a > b	a é maior que b?
-a	negação	a >= b	a é maior ou igual que b?

```
a=c(0,2,3,4,5,6,7)
b=c(1,5,2,8,12,10,10)
a+b #somando os elementos de a com os de b
[1]  1  7  5 12 17 16 17
a<b #  Testando os correspondentes elementos
TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
# Apenas o terceiro é maior que b
```

Na matriz aritmética teremos:

```
x=matrix(1:9,nrow=3, ncol=3)
# 9 elementos de 1 a 9 em 3 linhas e 3 colunas
xl=matrix(c(1,2,3,4,5,6,7,8,9), ncol=3,
byrow=TRUE)
# 3 colunas, leitura por linha
xc=matrix(c(1,2,3,4,5,6,7,8,9), ncol=3,
byrow=TRUE)
# 3 colunas, leitura por colunas igual a x
```

Uma variável de tipo caracter que mantém como valor uma cadeia de valores de dígitos pode ser manipulado por meio de funções especiais, vejamos um exemplo.

- `paste(..., sep="")`. Concatena vetores depois de transformá-los em caracteres; o `sep=""` indica como as séries serão separadas (a definição padrão é um espaço em branco).

```
(nth<-paste0(1:5,c("ind","ind","ind",
rep("Novo",2))))
#[1] "1ind" "2ind" "3ind" "4Novo" "5Novo"
```

1.4.1 Matrizes

Uma matriz é uma coleção de elementos de dados dispostos em um layout retangular bidimensional. O seguinte exemplo é de uma matriz com 2 linhas e 3 colunas.

```
H = matrix(
c(7, 7, 3, 2, 4, 2),# os elementos da matriz
nrow=2,             # número de linhas
ncol=3,             # número de colunas
byrow = TRUE)       # lendo os dados por linha
H                   # "print" a matriz
#       [,1]  [,2] [,3]
# [1,]    2    4    3
# [2,]    1    5    7
```

Pode-se separar alguns elementos da matriz usando a expressão $H[m,n]$. Por exemplo:

```
H[2, 3] # Elemento da 2ª linha e 3ª coluna
# [1] 7
H[ ,c(1,3)] # 1ª e 3ª coluna
# [,1] [,2]
# [1,]    7    3
# [2,]    2    2
dimnames(H) = list(
    c("row1", "row2"), # nomes das linhas
    c("col1", "col2", "col3")) #das colunas
H
#       col1 col2 col3
# row1    7    7    3
# row2    2    4    2
A<-H[ ,c(1,3)]
#       col1 col3
# row1    7    3
# row2    2    2
```

Algumas funções de matrizes são apresentadas abaixo.

Funções	Significados	Comandos
t	transposta	$t(H)$
$diag$	diagonal	$diag(H)$
$\%*\%$	multiplicação	$H\%*\%H$
det	determinante	$det(H)$
$solve$	inversa	$solve(H)$
$eigen$	autovalores	$eigen(H)\$values$
$eigen$	autovetores	$eigen(H)\$vectors$
svd	decomposição de valores singulares	$svd(H)$
qr	descomposição QR	$qr(H)$
$chol$	decomposição Choleski	$chol(H)$

1.5 Criando fatores

Fatores representam variáveis categóricas e são usados também como indicadores de grupos. Usa-se a função "as.factor()" para criar um fator de um vetor e a função "as.numeric" e "levels" para ter os códigos internos dos fatores e legendas. Por exemplo:

```
x<-as.factor(c("Pulmonar","ExtraPulmonar",
"Urinária",
"Outras","Pulmonar","ExtraPulmonar"))
as.numeric(x)
levels(x)
table(x)


x<-c("Pulmonar","ExtraPulmonar","Urinária",
"Outras",
"Pulmonar","ExtraPulmonar")
as.factor(x)
x<-rep(6:1, each=2)
as.factor(x)
# Agrupando fatores
gl(4,3,labels=c("1", "2","3", "4"))
```

1.6 Operação de arredondamento e truncamento

O sistema utiliza basicamente 4 funções:

- floor(x), arredonda o valor passado no argumento para o próximo menor. *Floor*=piso.

- `trunc(x)`, trunca o valor eliminando a componente decimal.
- `round(x, digits=0)`, arredonda para o número inteiro mais próximo. O arredondamento vem efetuado ao número decimal considerado.
- `ceiling(x)`, arredonda para o próximo superior. *Ceiling*=teto.

A Tabela 1.2 ilustra estes arredondamentos e truncamentos.

Tabela 1.2: Arredondamentos e truncamentos.

Valor	floor(x)	trumc(x)	round(x)	round(x,3)	ceiling(x)
7.4955	7	7	7	7.495	8
-7.4955	-7	-7	-7	-7.295	-7
7.5	7	7	7	7.5	8
-7.511	-8	-7	-8	-7.511	-7

1.7 Criando função no *R*

Para criar uma função no *R* usa-se a seguinte sintaxe: *Nome*<-*function(argumento/i)corpo da função*. Vamos encontrar as raízes de uma equação do segundo grau `x^2-3x+2=0`.

```
zero.funcao2<-function(a,b,c){
   delta<-b^2-4*a*c
   x1<-(-b+sqrt(delta))/(2*a)
   x2<-(-b-sqrt(delta))/(2*a)
   return(c(x1,x2))}
zero.funcao2(1,-3,2)
# [1] 2 1
```

```
prima.função<-function()
{cat("pi grego = "pi, "\n")}
Binomio<-function(a,b)
  {(a+b)/(a-b)}
Binomio(3,1)
# [1] 2  # (3+1)/(3-1)=2
```

1.8 Valores pré-determinados pelo sistema

O sistema utiliza valores pré-estabelecidos. Através da função "options" pode-se observar:

```
oldOp<-options() # armazena as configurações
                 # originais
oldOp #configurações presentes no sistema
ls(oldOp) # Lista todas as configurações
options(digits=4)
# [1] 3.142   valor pi com 4 dígitos
options(oldOp) # Configuração original
pi
# [1] 3.141593
```

1.9 Vetores

Um vetor é um conjunto ordenado de dados associados a um objeto. Se quisermos criar um vetor chamado de a com os valores 10, 20 e 30, usamos:

```
a<-c(10,20,30)
```

```
a
# [1] 10 20 30
```

Observe abaixo alguns vetores não métricos.

```
Nomes<-c("Jonas", "Ana")
Idades<-c(40,23,8)
Gêneros<-c("M", "F")
Irmãos<-c("Ivo", "Adan","Eva", "Edson",
          "Pedro", "Luisa")
Amigos<-c("Jorge","André", "José")
Todos<-c(Nomes, Irmãos, Amigos)
Todos
data.entry(Todos)
Todos
# [1] "Jonas" "Ana"   "Ivo"   "Adan"   "Eva"
# [6] "Edson"    "Pedro"    "Luisa"
# [9] "Jorge"    "André"    "José"
```

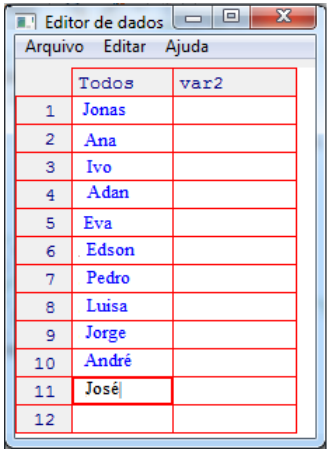


Figura 1.2: Acrescentando o nome José na tabela do editor de texto

Antes de voltar ao programa, não esquecer de fechar a tabela `data.entra()`.

1.10 Tratando dados perdidos

Dados perdidos no *R* é indicado por `NA`. Criando um vetor numérico `A` com três valores perdidos, a função `mean()` retorna `NA`, indicando que o valor faltante não permite o cálculo da média.

```
A<-c(1,3,4,NA,1,0,NA,8,1,NA); A
# [1]  1  3  4 NA  1  0 NA  8  1 NA
mean(A)
# NA
```

Informando `na.rm=TRUE` no *R*, ignora-se os `NA` e calcula-se a média com os números restantes. Por exemplo:

```
mean(A, na.rm=TRUE)
# 2.571429
(1+3+4+1+0+8+1)/7
# 2.571429
```

Usando a função `cbind` (colunas (`c`) vinculadas (`bind`)), junta as colunas `x` e `y` no mesmo objeto `z` sem os `NA`.

```
x <- c(160, NA, 175, NA, 180)
y <- c(NA, NA, 65, 80, 70)
z<- cbind(x = x[!is.na(x) & !is.na(y)],
y = y[!is.na(x) & !is.na(y)]); z
#      x  y
# [1,] 175 65
# [2,] 180 70
```

1.11 *Looping* - fazer um laço

Quando se precisa executar uma operação que se repete, usa-se um `Looping` (laço), informando a ideia de um círculo que se repete até uma determinada conclusão.

1.11.1 *Conditional Statements and Branching* - As declarações condicionais e ramos

Utilizam-se as seguintes instruções:

- `switch(<expr:stat>,<expr:casel> = <codl>,<expr:case2> = <cod2>, etc).`

Na declaração acima `<expr:test>` é um número ou string (valor que liga a outros). Esta declaração retorna: `<codl> if <expr:test> values \verb<expr:casel>`", `<cod2> if <expr:test> values <expr:case2>`, etc. Se `<expr:test>` não é igual a qualquer um dos `<expr:case>`, a função `switch()` retorna `NULL`. Por exemplo:

```
r<-rnorm(10,0,1)
seja <- "valores"
switch(seja,valores = mean(x),mediane
         = median(x))
seja <- "mediane"
switch(seja,valores = mean(x),mediane
         = median(x))
seja <- "sd"
switch(seja,valores = mean(x),mediane
         = median(x))
```

- Instruções `if` e `else` (se, caso de outra maneira). A instrução `if` condicional é usado nas duas formas seguintes:

```
if "cond" <expr:vrai>
ou
if "cond" <expr:vrai> el se <expr:faux>
```

O parâmetro "`cond`" deve ser, portanto, uma lógica que leva a um dos valores `TRUE` ou `FALSE`. Se `cond` for `TRUE`, a instrução será executada. Mas se é falso, nada acontece. Em outras palavras, a condicional é realizada com `if` e `else`, tem-se a seguinte sintaxe:

```
if (test) {
executes something
} else {
executes something else
}
Gênero<-c("M", "F" ,"F", "M", "F", "M")
ifelse(Gênero=="M", "Masculino", "Feminino")
# [1] "Masculino" "Feminino"
# [3] "Feminino" "Masculino"
# [5] "Feminino" "Masculino"
ifelse(Gênero == "F","Feminino",
       "Masculino")
# [1] "Masculino" "Feminino"
# [3] "Feminino" "Masculino"
# [5] "Feminino""Masculino"
```

```
x<- 1:50
xt<-ifelse(x%%7==0, NA,x)
ris<-rt[!is.na(xt)]
ris[1:20]
# [1]  1  2  3  4  5  6  8  9 10 11 12 13
# 15 16 17 18 19 20 22 23
# Observa-se que os números múltiplos de 7
# foram eliminados
```

- Instrução `stop` e `return`

```
x <- TRUE
if(x) y <- 1 else y <- 0; y
# [1] 1
###### Fatorial de número positivo
Fatorial<-function(n){
if(n<0)
stop("Argumento negativo")
else{
if(n==0)
return(1)
else
return(prod(1:n))}
}
Fatorial(0)
# [1] 1
Fatorial(4)
# [1] 24
Fatorial(-1)
```

```
# Error in Fatorial(-1) : Argumento negativo
```

- Instruções `for`

```
for(n in (1:3)) print(n)
# [1] 1
# [1] 2
# [1] 3
for(n in seq(0,6, by=2)) print(n)
# [1] 0
# [1] 2
# [1] 4
# [1] 6
for(x in c("Jonas", "Ana")) print(x)
# [1] "Jonas"
# [1] "Ana"
for(f in c(log,log2,log10)) print(f(25))
# [1] 3.218876
# [1] 4.643856
# [1] 1.39794
```

- Instruções `return`

```
zero.funcao2<-function(a,b,c){
   delta<-b^2-4*a*c
   x1<-(-b+sqrt(delta))/(2*a)
   x2<-(-b-sqrt(delta))/(2*a)
   return(c(x1,x2))}
```

```
zero.funcao2(1,-3,2)
# [1] 2 1
```

- instrução `while`

A sintaxe da instrução é a seguinte: `while` (<`condition`>) <`expression`>.

A lógica "`while`" é que enquanto houver número inteiro e positivo se calcula o fatorial (`n!`).

```
 fatorial2<-function(n){
if(n<0)
stop("Argumento negativo")
ra<-1
while(n>0){
ra<-ra*n
n<-n-1
}
return(a)
}
 fatorial2(5)
# [1] 120
 fatotoril2(0)
# [1] 1
 prod(1:5) # 5!
# [1] 120


H <- 0
while (H < 1){
```

```
        H <- rnorm(1)
         cat(H,"\n")
      }
# -1.114971
# -0.7465893
# 1.740903
```

1.11.2 Instrução *repeat* (repetir) e *break* (interromper)

Observa-se abaixo que foram gerados números somado com 1 até encontrar o 4, e quando encontrou houve um `break` (stop).

```
j <- 0
repeat{
j<-j+1
if (j==4) break}; j
# [1] 4
```

1.12 Operação de leitura de dados

Os data.frame apresenta estrutura semelhante à de uma matriz, embora suporte valores numéricos e alfanuméricos. Normalmente, quando um estudo estatístico é realizado sobre os sujeitos ou objetos de uma amostra, a informação se organiza precisamente em um dataframe: uma folha de dados, em que cada linha corresponde a uma sujeito e cada coluna a uma variável.

1.12.1 Lendo um data.frame de um arquivo de texto

Usa-se a função `read.table()` para ler dados em formato de texto. Considere os seguintes arquivo no bloco de notas (arquivo

do tipo `.txt` ou `.dat`).

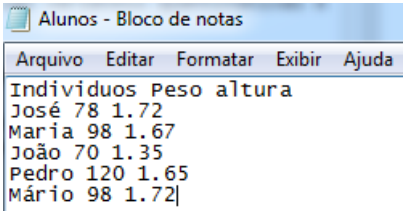


Figura 1.3: Nomes de alunos, peso e altura dos mesmos.

As colunas foram separadas por espaço no bloco de notas, se os espaços fossem separados por (;) teríamos que escrever `sep=";"`.

```
setwd("C:/LIVRO") # Criar um arquivo em C
                  # chamado LIVRO
getwd() # confirmar a mudança
#salvar o arquivo Alunos.txt
# dentro desse arquivo.
dados<-read.table("Alunos.txt",
header=TRUE, sep="")
dados
# Individuos Peso altura
# 1       José   78   1.72
# 2      Maria   98   1.67
# 3       João   70   1.35
# 4      Pedro  120   1.65
# 5      Mário   98   1.72
```

1.12.2 Lendo um data.frame de um arquivo de Excel

A abreviação `.csv` *comma-separated values* significa (valores separados por vírgula). O programa *Excel* também ler este tipo de arquivo. É um arquivo de apenas uma planilha que contém dados separados por um sinal de pontuação, como vírgula, ponto e vírgula ou ponto.

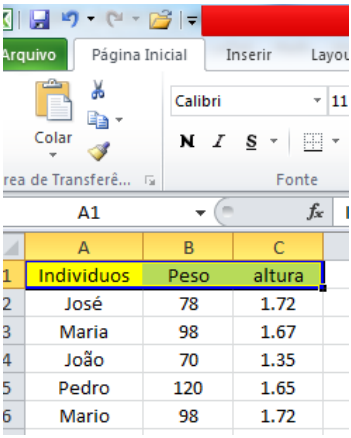


Figura 1.4: Alunos com peso e altura

```
setwd("C:/LIVRO")
dados<-read.csv("AlunosExcel.csv",
       header=TRUE, sep=";")
dados
# Individuos Peso altura
# 1       José   78   1.72
# 2      Maria   98   1.67
# 3       João   70   1.35
# 4      Pedro  120   1.65
```

```
# 5       Mário    98    1.72
```

Existe uma pequena diferença entre os comandos `read.csv()` e `read.csv2()`. O *default* (valor padrão) desses dois argumentos são: `sep` (separador de colunas) e `dec` (separador de casas decimais). Na função `read.csv()` o default é `sep=","` e `dec="."`. Já em `read.csv2()` tem-se `sep=";"` e `dec=","`.

1.12.3 Lendo arquivos que usam um formato fixo

Neste tipo usa-se a função `read.fwf()`. Observa-se que os dados são separados em tamanho 6 (6 caracteres da palavra *ProdBB*), 4 para os anos, 2 código e 3 para índice (o número 5, o ponto e o número 7) considerando a primeira linha.

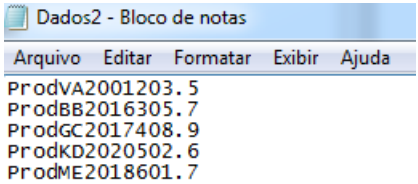


Figura 1.5: Produção, ano e códigos.

```
dados1<-read.fwf("Dados2.txt",
        widths=c(6,4,2,3), skip=1)
# Warning message:
# In readLines(file, n = thisblock) :
# incomplete final line found on 'Dados2.txt'
dados1
# V1 V2 V3 V4
# 1 ProdBB 2016 30 5.7
```

```
# 2 ProdGC 2017 40 8.9
# 3 ProdKD 2020 50 2.6
# 4 ProdME 2018 60 1.7
```

1.12.4 Leitura de arquivo direto da internet

Para ler um arquivo direto da internet, primeiro se deve clicar com o botão direito sobre o arquivo e selecionar a opção "Copiar endereço do link" para que possa ser escrito dentro da função *read.csv*2(). No *R*, basta colar o link do arquivo no local onde você normalmente colocaria o diretório de um arquivo do seu computador, assim observe um exemplo fictício:

```
read.csv2("http://www.estatistica777.com/
arquivos/dados.csv",sep=",", dec=".")
```

1.12.5 Como importar qualquer arquivo no *R*

O pacote *rio* serve para facilitar a importação de arquivos para o *R*. Com a função, *import*(), se pode detectar qual arquivo o usuário deseja abrir. O primeiro passo para usá-lo é instalar o pacote, que está no *CRAN*:

```
# Instalando o pacote 'rio'
install.packages(rio, dependencies = TRUE)
```

Após instalar o pacote com todas as dependências (outros pacotes que são necessários para que ele funcione), fica fácil carregar qualquer arquivo de dados, basta usar a função `import()` do pacote `rio` (CHAN *et al*., 2016).

```
# Carrega um arquivo em .csv
Dados <- import(file = "AlunosExcel.csv")

# Carrega um arquivo em .txt
Dados <- import(file = "Dados2.txt")

# Carrega um arquivo em .dta (Stata)
Dados <- import(file = "dados.dta")

# Carrega um arquivo em .sav (SPSS)
Dados <- import(file = "dados.sav")
```

vejamos alguns exemplos abaixo:

```
Dados_txt <- import(file = "dados3.txt")
Dados_txt
# G X1 X4
# 1   1  1  4
# 2   2  1  2
# 3   3  1  6
# 4   4  1  4
# 5   5  1  6
# 6   6  1  8
# 7   7  2  6
# 8   8  2  8
# 9   9  2  8
# 10 10  2  2
# 11 11  2  5
Dados_csv <- import(file = "AlunosExcel.csv")
Dados_csv
Indivíduos Peso altura
```

```
# 1       José   78   1.72
# 2      Maria   98   1.67
# 3       João   70   1.35
# 4      Pedro  120   1.65
# 5      Mário   98   1.72
```

1.13 Criando medidas repetidas

Considera-se dados multivariados de diferente natureza, ou seja, os dados resultantes das medições repetidas na mesma variável em cada unidade no conjunto de dados. Para entender o procedimento considere um pequeno conjunto de dados.

```
Dados<-read.table("MedidasRepetidas.txt",
         header=TRUE, sep="")
Dados
# Id G T.1 T.2 T.5 T.7
# 1 1 1 15 15 10 7
# 2 2 1 10 9 11 12
# 3 3 1 8 7 6 9
# 4 4 2 11 8 13 7
# 5 5 2 11 12 11 11
# 6 6 2 12 12 6 10
Rep<-reshape(Dados, direction="long",
idvar="Id", varying=colnames(Dados)[-(1:2)])
Rep
# ID G time T
# 1.1 1 1 1 15
# 2.1 2 1 1 10
```

```
# 3.1 3 1 1 8
# ...
# 5.7 5 2 7 11
# 6.7 6 2 7 10
```

1.14 Acrescentando colunas, linhas e grupos

Considere os mesmos dados trabalhado anteriormente.

```
dados<-read.table("dados3.txt", header=TRUE,
                  sep="")
dados
# G X1 X4
# 1   1  1  4
# 2   2  1  2
# 3   3  1  6
# 4   4  1  4
# 5   5  1  6
# 6   6  1  8
# 7   7  2  6
# 8   8  2  8
# 9   9  2  8
# 10 10  2  2
# 11 11  2  5
dadosU<- rep("G1",11) # repetindo G1
                      # onze vezes
dadosU# Acrescentando G1 na quarta coluna
# [1] "G1" "G1" "G1" "G1" "G1" "G1" "G1" "G1"
# [9] "G1" "G1" "G1"
```

```
dados[,4]<-dadosU[1]
# Alterando os nomes das colunas
colnames(dados)<-c("X2","X3","X4","G")
# Acrescentando elementos na 1a linha
dados[1,]<-c(10,22,47,"G1")
dados
colnames(dados, do.NULL=FALSE)
# [1] "X2" "X3" "X4" "G"
colnames(dados)<- c("X2","X3","X4","G")
dados<- cbind(1, dados)
colnames(dados)<- c("X1","X2","X3","X4","G")
dados
# X1 X2 X3 X4  G
# 1   1 10 22 47 G1
# 2   1  2  1  2 G1
# 3   1  3  1  6 G1
# 4   1  4  1  4 G1
# 5   1  5  1  6 G1
# 6   1  6  1  8 G1
# 7   1  7  2  6 G1
# 8   1  8  2  8 G1
# 9   1  9  2  8 G1
# 10  1 10  2  2 G1
# 11  1 11  2  5 G1
rownames(dados)<- rownames(dados[,1],
do.NULL = FALSE, prefix = "Ind.")
dados
# X1 X2 X3 X4 G
# Ind.1 1 10 22 47 G1
```

```
# Ind.2 1 2 1 2 G1
# Ind.3 1 3 1 6 G1
# Ind.4 1 4 1 4 G1
# Ind.5 1 5 1 6 G1
# Ind.6 1 6 1 8 G1
# Ind.7 1 7 2 6 G1
# Ind.8 1 8 2 8 G1
# Ind.9 1 9 2 8 G1
# Ind.10 1 10 2 2 G1
# Ind.11 1 11 2 5 G1
```

1.15 Janelas gráficas

Todos os gráficos criados em *R* são exibidos em janelas especiais, distinto do console, chamado "*R graphies: Device numero-device*", em que "número-device" é um número inteiro da janela (ou dispositivo). As diferentes instruções de *R* para realizar gráficos, que formam partes dos diferentes pacotes, se podem dividir em:

- Funções gráficas: Permitem realizar diferentes tipos de gráficos e têm seus próprios argumentos específicos.
- Gráficos complementares: São também funções que permitem acrescentar aos gráficos linhas, textos, flechas, legendas, etiquetas, etc., e também tem seus próprios argumentos específicos.
- Argumentos gerais: São argumentos que se podem usar nas funções e complementos gráficos anteriores.

Um pacote muito conhecido é o `graphics` (MURRELL, 2005). Para obter uma lista completa de funções com páginas de ajuda individuais, use `library(help="graphics")`, além de consultar o menu de ajuda da função:

```
library(graphics)
help(base)
```

Através do comando `demo(graphics)` se mostram bons exemplos de gráficos com seu correspondente código.

1.15.1 Principais *Scripts* e arquivos para geração de gráficos

abline: acrescenta linhas. Argumentos: `abline(a=NULL, b=NULL, h=NULL, v=NULL, reg=NULL, coef=NULL, untf=FALSE, ...)`.

- a e b: intersecção e inclinação da reta.
- h: valor de y em uma linha horizontal.
- v: valor de x em uma linha vertical.
- `reg`: um vetor do tipo $c(a,b)$ com a intersecção e inclinação da reta.
- `coef`: especifica-se uma regressão, $coef(lm(c(x \sim y)))$.
- `untf`: `FALSE` ou `TRUE`, se um eixo tem transformação `log`, com `TRUE`, representa a linha sem transformação.

arrows: representa flechas. Argumentos: `arrows(x0, y0, x1=x0, y1=y0, length=0.25, angle=30, code=2, col=par("fg"), lty=par("lty"), lwd=par("lwd"), ...)`.

- x_o e y_o: coordenadas da origem.
- x_1 e y_1: coordenadas finais.
- `lenght`: longitude da ponta da flecha.
- `angle`: ângulo em relação a ponta da flecha.
- `code`: tipo de flecha.

box: caixa de texto no gráfico. Argumentos: `box(which= "plot", lty="solid", ...)`. *Which* indica onde representa o quadro e pode ser: `plot, figure, "inner"` ou `"outer"`.

legend: acrescenta uma legenda. Principais Argumentos: `legend(x, y=NULL, legend, fill=NULL, col=par("col"), lty, lwd, pch, border="black", angle=45, density=NULL, bty="o", bg=par("bg"))`

- *x* e *y*: coordenadas da legenda.
- `legend`: texto da legenda.
- `fill`: cor do preenchimento.
- `ncol`: número de colunas.
- `horiz`: se for verdadeiro (`TRUE`) os textos são horizontais.
- A posição da legenda é especificada com coordenadas ou com palavras-chave:

`bottomright`: abaixo à direita.

`bottomleft`: abaixo à esquerda.

`topleft`: no topo à esquerda.

`topright`: no topo à direita.

`left`: à esquerda.

`right`: à direita.

`center`: no centro.

- `bty`: define o quadro da legenda: `"o"` com o quadro e `"n"` sem o quadro.

lines: acrescenta linhas ou pontos com linhas. Principais Argumentos: `lines(x, y=NULL, type="l", ...)`.

- *x* e *y*: coordenada dos pontos a unir mediante linhas.
- type: tipos de linhas: `"p"` para pontos, `"l"` para linhas, `"b"` desenha pontos ligados por curvas, `"c"` para as linhas da parte isolada de `"b"`, `"o"` desenha pontos com curvas sobrepostas `"overplotted"`, `"h"` para `"histograma"` com (ou "alta densidade") nas linhas verticais, `"s"` para passos de escada, `"S"` para outras etapas e `"n"` não desenha o gráfico, mas apresenta os eixos cujas coordenadas são determinadas de acordo com os dados.

locator: com auxílio do mouse, permite localizar a posição do objeto no painel gráfico e devolve as coordenadas. Argumentos: `locator(n=512, type="n", ...)`.

- `n`: o número máximo de objetos que se deseja localizar.
- `type`: com `"n"` localiza qualquer objeto, com `"p"` representa símbolos e com `"I"` localiza a linha unida aos pontos para o qual `n` deve ser maior que 1.

polygon: representa polígonos. Argumentos: `polygon(x, y=NULL, density=NULL, angle=45, border=NULL, col=NA, lty=par("lty"), ..., fill Odd Even=FALSE)`

- `x` e `y`: coordenadas dos vértices do polígono.

- `density`: densidade do sombreamento.
- `angle`: ângulo das linhas do sombreamento.
- `border`: cor do bordados, `NA` sem borda e `NULL` com a cor (preta) por *default*.

segments: desenha segmentos entre pontos. Argumentos:
`segments(x0, y0, x1=x0, y1=y0, col=par("fg"), lty=par("lty"), lwd=par("lwd"), ...)`

- `x_{o}, y_{o}, x_{1}, y_{1}`: coordenadas de origem e final da linha.

text: permite acrescentar etiquetas nas coordenadas de um gráfico. Argumentos: `text(x, y=NULL, labels=seqalong(xx), adj=NULL, pos=NULL, offset=0.5, vfont=NULL, cex=1, col=NULL, font=NULL, ...)`

- `x` ou `y`: vetores numéricos de coordenadas onde os rótulos de texto devem ser escritos.
- `labels`: texto que deseja-se escrever ou variáveis com as etiquetas do texto dos pontos.
- `srt`: ângulo do texto, ou seja, o texto pode ser rotacionado usando este argumento.
- `offset`: posição do texto no eixo vertical em relação às coordenadas.
- `pos`: eixo em que se localizará o texto: abaixo (1), esquerda (2), acima (3) e direita (4).

title: permite acrescentar etiquetas as coordenadas de um gráfico. Argumentos: `title(main=NULL, sub=NULL, xlab=NULL, ylab=NULL, line=NA, outer=FALSE, ...)`

- `main`: texto do título do gráfico.

- sub: subtítulo.
- xlab: legenda do eixo x.
- ylab: legenda do eixo y.
- line: valor numérico que define a separação do texto em relação ao gráfico.

A Figura 1.6 mostra os valores de 1 a 25 que definem os símbolos do argumento "pch".

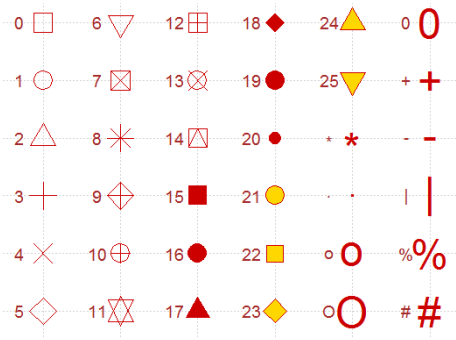


Figura 1.6: Valores numéricos e símbolos da função do argumento "pch".

A função "colorTable()" do pacote fBasics (WUERTZ, 2011) permite observar os números com diferentes cores.

1.16 Uso da função %>% (*pipe*)

O operador %>% (*pipe*) é usado para inserir um argumento em uma função. A ideia é usar o valor resultante da expressão do lado esquerdo como primeiro argumento da função do lado direito. Para utilizar o *pipe*, carregue o pacote margrittr (BACHE, 2014) utilizando o comando library(magrittr). Um exemplo abaixo

mostra que o operador do lado esquerdo é o primeiro argumento (1 a 20) e a função no lado direito, em seguida, o desvio padrão.

```
1:20 %>% sd
# [1] 5.91608 # é equivalente a
sd(1:20) # desvio padrão
# [1] 5.91608
```

Usa-se também para filtrar funções específicas dentro de pacotes existentes, como é o caso da função `cor.test` (um teste de correlação). Para o exemplo abaixo utilize o pacote `dplyr` (HADLEY *et al.*, 2018).

```
library(magrittr)
library(dplyr)
mtcars %>% filter(wt>2) %$% cor.test(hp, mpg)
```

A função `show.colors()` do pacote `DAAG` (MAINDONALD; BRAUN, 2015) permite ver os nomes dos diferentes tipos de cores: `"singles"` (simples) que não têm diferentes intensidades e `"shades"` (tons) apresenta diferentes tonalidades da cor cinza `"gray"`.

```
install.packages("fBasics")
library(fBasics)
colorTable(cex=1)
```

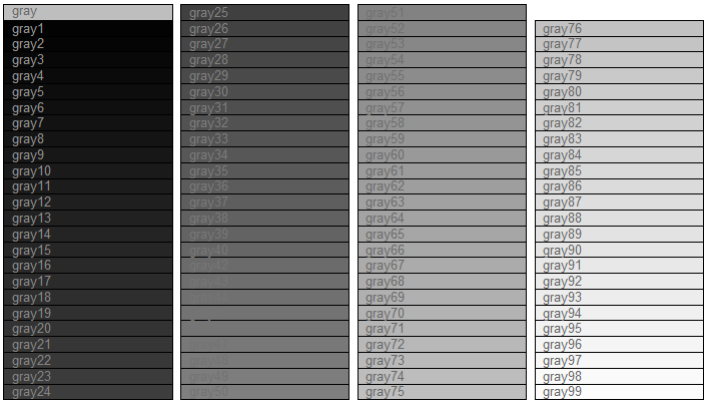


Figura 1.7: Tons de cinza.

```
show.colors("gray")
show.colors("singles")
show.colors("shades")
```

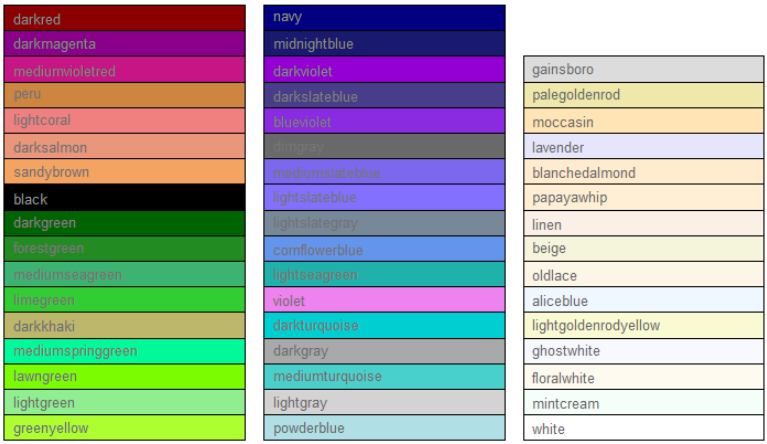


Figura 1.8: Cores sem tonalidades.

A Figura 1.9 mostra cores de diferentes tonalidades em cinco palhetas.

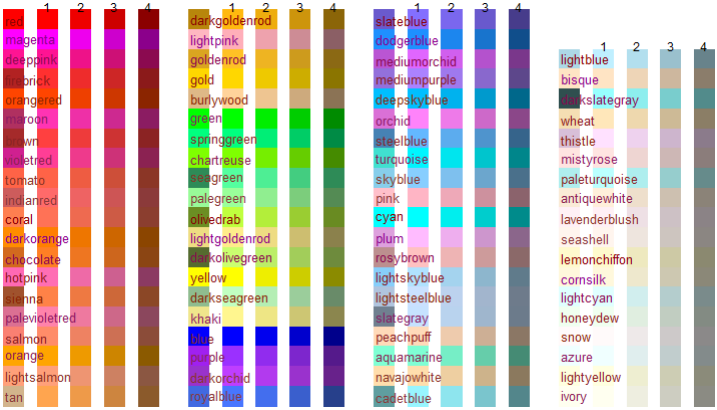


Figura 1.9: Cores que possuem 4 ou 5 tons.

1.17 Fórmulas matemáticas e caracteres especiais

Acrescenta fórmulas matemáticas ou textos com caracteres especiais nos gráficos gerados, como exemplo:

```
x<-seq(13.547,46.453, length=100)
plot(x,dnorm(x,mean=30, sd=5), xlab="x",
ylab="Densidade", main=expression(paste("y=",
frac(1,sigma*sqrt(2*pi))*e^{-frac(sum((x[i]
-mu))^2, 2*sigma^2)})), type="l",
sub=expression(paste(N(mu,sigma^2))))
```

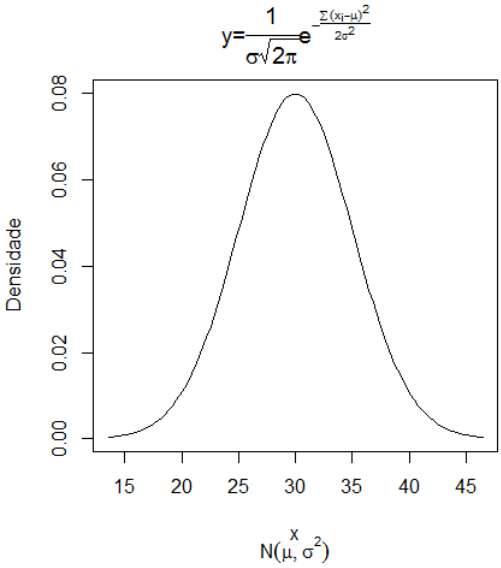


Figura 1.10: Normal com média zero ($\mu = 0$) e variância (σ^2)

```
plot(0,0,type="n",bty="n", xaxt="n", yaxt="n",
     xlab="", ylab="")
text(0,0.9,expression(chi^2==sum(sum(frac(((
O[mc] - E[mc]) - frac(1,2))^2, E[mc]), c-1,
i), m-1, j)))
```

$$\chi^2 = \sum_{m-1}^{j} \sum_{c-1}^{i} \frac{((O_{mc} - E_{mc}) - \frac{1}{2})^2}{E_{mc}}$$

Figura 1.11: Normal com média zero, $\mu = 0$ e variância, σ^2

Usaremos algumas fórmulas para exemplificar.

```
plot(0,0,type="n", bty="n", xaxt="n", yaxt="n",
```

```
xlab="",ylab="")
text(0,0.9,expression(L[t]==L[infinity](1
    -e^-k(t[f]-t[0]))))
text(0,0.5,expression(y[i] == sqrt(a[i]^2
+b[i]^2)))
text(0,0.1,expression("r"==paste(frac(
   paste(mu[max]*"S"), paste("K"[s]+"S")))))
text(0,-0.4,expression(bar(x) == frac(sum(
   x[i], n, i==1), n)))
```

Respectivamente, têm-se as seguintes fórmulas:

$$L_t = L_\infty(1 - e^{-k(t_f - t_0)})$$

$$y_i = \sqrt{a_i^2 + b_i^2}$$

$$r = \frac{\mu_{max} S}{K_s + S}$$

$$\bar{x} = \frac{\sum_{n}^{i=1} x_i}{n}$$

Capítulo 2

Análise discriminante simples (ADS)

Definição 2.0.1 (Análise discriminante)**.** *É uma técnica que pode ser utilizada para classificação de elementos de uma amostra ou população. Para a sua aplicação, é necessário que os grupos para os quais cada elemento amostral pode ser classificado sejam predefinidos, ou seja, conhecidos a priori considerando-se suas características gerais. Este conhecimento permite uma elaboração de uma função matemática chamada de regra de classificação, ou discriminação, que é utilizada para classificar novos elementos amostrais nos grupos já existentes, portanto, o número de grupos é conhecido a priori.*

2.1 É importante na análise discriminante

a) Conhecer em que contexto se aplica esta técnica.

b) Conhecer as pressupostos básicos e limitações da análise discriminante.

c) Diferenciar entre análise discriminante das outras técnicas multivariadas.

d) Conhecer que tipo de perguntas de pesquisa permite resolver com análise discriminante.

e) Identificar os passos necessários para análise discriminante.

f) Diferenciar análise discriminante linear da função de classificação linear.

g) Saber avaliar o poder classificatório de análise discriminante.

h) Interpretar os resultados obtidos. Identificar as variáveis que melhor discriminam dois ou mais grupos. Interpretar a matriz confusão e mapa perceptual.

O número de variáveis discriminantes deve ser menor que o número de indivíduos menos dois, ou seja, $(X_1, \ldots, X_p)$, em que $p < (n-2)$ e n é o número de indivíduos (casos ou objetos).

Nenhuma das variáveis discriminantes poderá ser combinação linear das restantes. O número de observações em cada grupo é pelo menos dois $(n_G \geq 2)$. Em que G representa a quantidade de categorias(grupos) da variável dependente.

A Tabela 2.1 mostra a estrutura ideal da matriz de dados que se aplica à análise discriminante, como se consideram os indivíduos (linhas) e as p variáveis (colunas) para uma análise discriminante. O tamanho de cada grupo são $a, b, \ldots, n_G$, respectivamente.

Um conjunto de dados formado por amostras aleatórias obtidas de cada um dos grupos distintos. Os n vetores de indivíduos se subdividem em $n_1 = a$ do grupo 1, $n_2 = b$ do grupo 2 e $n_G = G$ do grupo G. As variáveis quantitativas são representadas por $X_1, \cdots, X_p$, em que $i = 1, \ldots, p$.

Tabela 2.1: Caso geral.

Indivíduos/ variáveis	X_1	X_2	...	X_p	
1	X_{111}	X_{112}	...	X_{11p}	
2	X_{211}	X_{212}	...	X_{21p}	G_1
⋮	⋮	⋮		⋮	
a	X_{a11}	X_{a12}	...	X_{a1p}	
1	X_{121}	X_{122}	...	X_{12p}	
2	X_{221}	X_{222}	...	X_{22p}	G_2
⋮	⋮	⋮		⋮	
b	X_{b21}	X_{b22}	...	X_{b2p}	
⋮	⋮	⋮	...	⋮	
1	X_{1G1}	X_{1G2}	...	X_{1Gp}	
2	X_{2G1}	X_{2G2}	...	X_{2Gp}	G_G
⋮	⋮	⋮		⋮	
n_G	X_{n_GK1}	X_{n_GG2}	...	X_{n_GGp}	

2.2 Relação com outras técnicas

A Tabela 2.2 mostra as principais diferenças e semelhanças entre análise discriminante com outras técnicas estatísticas.

Tabela 2.2: Semelhança e diferenças entre ANOVA, regressão e AD

Variáveis	Regressão	ANOVA	Discriminante
Número	Similaridade	Similaridades	Similitude
Dependente	Uma	Uma	Uma
Independente	Múltiplas	Múltiplas	Múltiplas
Natureza	Diferenças	Diferenças	Diferenças
Dependente	Métrica	Métrica	Categórica
Independente	Categórica	Métrica	Métrica

2.3 Tipos de análise discriminante

As técnicas de AD podem ser classificadas segundo o número de grupos na variável dependente.

- Análise discriminante simples: a variável dependente tem apenas dois grupos
- Análise discriminante múltiplo: a variável dependente tem mais de dois grupos.

procedimentos que, em geral, vão além do uso de distâncias matemáticas.

Há dois objetivos principais:

- Discriminação: Use as informações de um conjunto de observações etiquetadas para construir um classificador (ou regra de classificação) que irá separar as classes predefinidas, tanto quanto possível.
- Classificação: Dado um conjunto de medidas em uma nova observação não rotulada, use o classificador para prever a classe dessa observação.

2.4 Utilidade da análise discriminante

Dependendo do objeto da pesquisa a AD emprega-se com as seguintes finalidades:

1. Explicativa: quantifica a contribuição relativa de cada uma das variáveis independentes na classificação correto dos indivíduos considerados dentro dos distintos grupos. Tenta provar o poder discriminante de cada uma destas variáveis, em muitos casos com a finalidade de selecionar o subconjunto que melhor discrimina os grupos predeterminado a priori.

2. Predicativa: classificar um novo indivíduo pertencente a priori a partir dos valores das variáveis independentes quantitativas. É bastante interessante classificar indivíduos em grupos, particularmente quando desconhecemos seu grupo de pertinência.

3. Reclassificadores: Para este caso muitas vezes se realiza uma análise de agrupamento antes da análise discriminante.

2.5 Significado geométrico das funções discriminantes

Se n é o número de indivíduos e p o número de variáveis independentes, os dados poderão ser elementos de uma matriz de dados de n linhas e p colunas, cujo formato se observa na Tabela 2.3 abaixo:

Tabela 2.3: Formato simplificado da matriz de dados correspondentes às variáveis discriminantes.

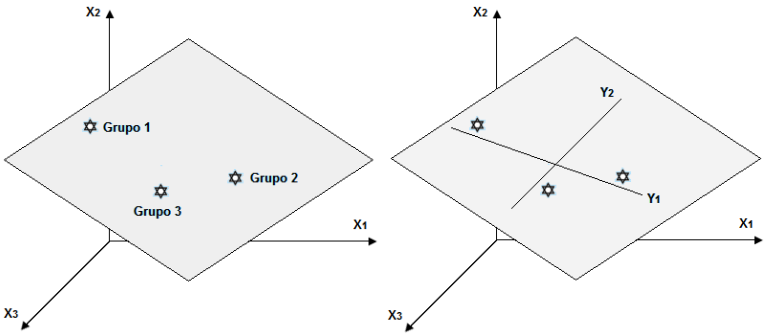
	X_1	X_2	...	X_3
Indivíduo 1	x_{11}	x_{12}	...	x_{1p}
Indivíduo 2	x_{21}	x_{22}	...	x_{2p}
⋮	⋮	⋮	⋮	⋮
Indivíduo n	x_{n1}	x_{n2}	...	x_{np}

Para cada linha em que se observa os indivíduos pode ser considerado um ponto do espaço p-dimensional definido ao tomar as variáveis discriminantes como eixo do dito espaço.

A posição do grupo no espaço pode ser caracterizada por seu centróide, definidos como o ponto que se obtém ao considerar como coordenadas os valores médios que o grupo de indivíduos representa em cada uma das variáveis.

Na Figura 2.1 observa-se que no primeiro eixo os centróides aparecem mais separados, por isso que a primeira função discriminante tem

sempre um significado especial quando se interpreta a diferença entre grupos, pois esse eixo capta a máxima dispersão em relação às demais dimensões, como se observa na Figura (b).



(a) Centróides dos grupos (b) Funções discriminantes

Figura 2.1: Plano determinado por três centróides de grupo no espaço tridimensional definido pelas variáveis X_1, X_2 e X_3.

2.6 Processo da análise discriminante

Os passos para obtenção das funções discriminantes se encontram no fluxograma abaixo.

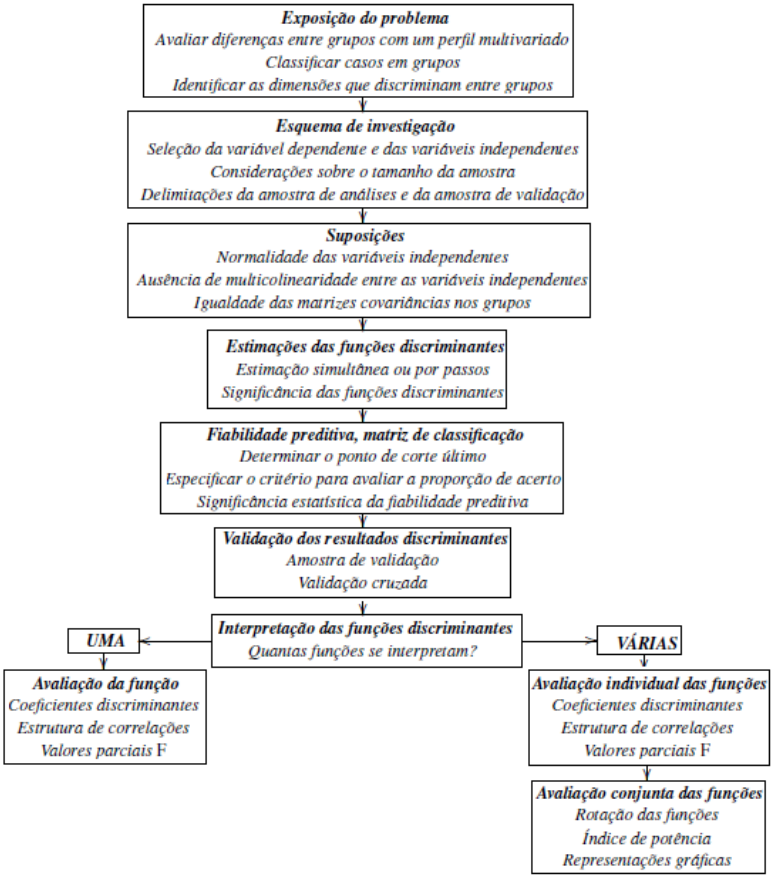


Figura 2.2: Sequência de investigação orientada para análise discriminante. Fonte: Adaptado de Hair et al., (1995).

Os resultados devem contextualizar teoricamente, o que podem dar lugar a novos problemas de investigação que podem requerer ou não da análise discriminante.

2.7 Variáveis classificadoras para obter a função discriminante de Fisher

Agora realizaremos uma exposição formalizada utilizando variáveis classificadoras para obter a função discriminante de Fisher. A Figura 2.3 mostra a representação das elipses de concentração dos dados correspondentes a duas distribuições de frequências bivariadas, nas que as variáveis $\mathbf{X}_1$ e $\mathbf{X}_2$ estão correlacionadas positivamente.

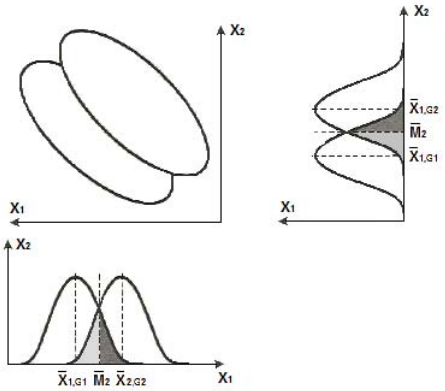


Figura 2.3: Elipse de concentração das distribuições de frequências e sua projeção sobre os eixos $\mathbf{X}_1$ e $\mathbf{X}_2$.

A Figura 2.3 observa-se que as distribuições dos dois grupos estão entrelaçadas. As duas elipses têm o mesmo tamanho, apenas diferem em seu centro. Debaixo do eixo $\mathbf{X}_1$ se tem representado a projeção das distribuições de frequências bivariadas sobre este eixo. Esta projeção oferece as distribuições univariadas marginais de $\mathbf{X}_1$.

Sobre o eixo $\mathbf{X}_2$ tem-se projetado igualmente as distribuições de frequências bivariadas, obtendo-se as correspondentes distribuições de frequências marginais. Também neste caso, as distribuições marginais aparecem muito entrelaçadas. Como temos visto, quanto maior seja o grau de en-

trelaçamento, maior será o percentual de indivíduos classificados erroneamente.

Na Figura 2.4, o eixo que aparece o menor entrelaçamento será o ótimo. A este eixo denomina-se eixo discriminante e as projeções dos valores das variáveis $\mathbf{X}_1$ e $\mathbf{X}_2$ são os escores discriminantes. A variável obtida na projeção, que será designada por D é a função discriminante. Portanto, às distribuições de frequência que aparecem sobre o eixo discriminante são as correspondentes funções discriminantes desejadas.

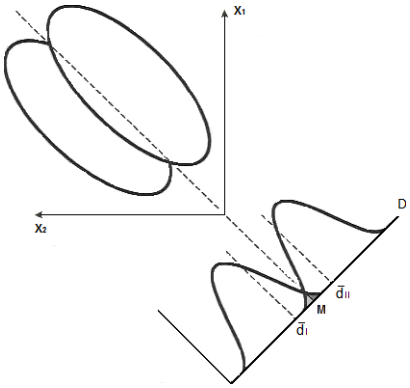


Figura 2.4: Elipse de concentração das distribuições de frequências e sua projeção sobre o eixo discriminante. Fonte: Adaptado de Jímenes e Manzano, (2005).

Comparando as distribuições de frequências da Figura 2.5 com as distribuições de frequências da variável **X**, observa-se que o entrelaçamento da função discriminante é inferior ao representado na Figura 2.4. Assim, o percentual de indivíduos bem classificados seria maior.

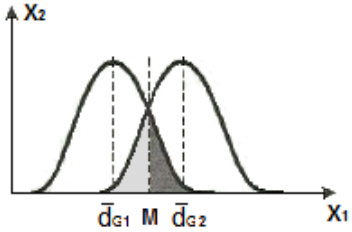


Figura 2.5: Funções de distribuição de frequências das pontuações sobre o eixo discriminante. Fonte: Adaptado de Jímenes e Manzano, (2005).

2.7.1 Centro de gravidade

O ponto de corte médio discriminante M se calcula mediante $\bar{D}_I$ e $\bar{D}_{II}$ da seguinte forma:

$$M = \frac{\bar{D}_I + \bar{D}_{II}}{2} = \frac{\mu_I + \mu_{II}}{2} \tag{2.1}$$

O centróide em função do número de elementos de cada amostra pode ser calculado através da seguinte fórmula:

$$M = \frac{n_1\bar{D}_I + n_2\bar{D}_{II}}{n_1 + n_2} \tag{2.2}$$

Em que, n_1 é o número de observação do G_I, e n_2 é o número de observação do G_{II}.

Quando um grupo tem um número de amostra maior que o outro, as probabilidades de retirarem uma amostra dos grupos são diferentes. O critério para classificar o indivíduo i é:

Se $D_i < M$, classifica-se o indivíduo i no G_I.

Se $D_i > M$, classifica-se o indivíduo i no G_{II}.

Em geral, quando se aplica a análise discriminante se resta o valor de M para a função. Desta forma, a função discriminante vem dada por:

$$D - M = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \cdots + a_p\mathbf{X}_p - M \quad (2.3)$$

Assim, classifica-se um indivíduo no G_I se $D_x - M < 0$, e em G_{II} se $D_x - M > 0$. Igualando a zero o segundo membro da Equação 2.3, para o caso $p = 2$ variáveis, obtém-se a seguinte equação da reta:

$$a_1\mathbf{X}_1 + a_2\mathbf{X}_2 - M = 0 \quad (2.4)$$

Em que a_1, a_2 e M são números reais.

Capítulo 3

Critério para obtenção da função discriminante de Fisher

Seja $F(\mathbf{X})$ função discriminante de Fisher como função linear de p variáveis explicativas $\mathbf{X}$.

$$F(\mathbf{X}) \quad = \quad a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \cdots + a_p\mathbf{X}_p$$

Precisa-se obter os coeficientes de ponderação a_i, para isso considere que existem n observações, assim podemos expressar uma função discriminante teórica para as n observações da seguinte forma:

$$F(\mathbf{X}) = a_1\mathbf{X}_{1i} + a_2\mathbf{X}_{2i} + \cdots + a_p\mathbf{X}_{ni} \qquad i = 1,2,\ldots,p$$

Escrevendo em forma de matriz, tem-se:

$$\begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

Numa notação matricial mais compacta será: $F = \mathbf{X}a$. Multiplicando $F'F$, tem-se:

$$F'F = a'Ba + a'Wa = a'Ta$$

Em que B e W são as matrizes de soma de quadrados e produtos cruzados entre os grupos e dentro dos grupos respectivamente. A soma dessas duas matrizes é indicada por T, que é a soma de quadrados e produtos cruzados total. Para estimar os coeficientes a_i, Fisher utilizou o seguinte critério:

$$\frac{\text{Maximização de Variabilidade entre-grupos } (B)}{\text{Variabilidade dentro-grupos } (W)}$$

Sendo autovalores altos resultam em boas funções discriminantes. Segundo Urial e Aldás (2005) esse critério determina o discriminante de forma que as distribuições projetadas sobre o eixo sejam as mais separadas possíveis entre si (maior variabilidade entre os grupos) e, ao mesmo tempo, que cada uma das distribuições estejam menos dispersas (menor variabilidade dentro dos grupos). O critério de Fisher para maximizar λ, será então:

$$\lambda = \frac{a'Ba}{a'Wa}$$

Como podemos ver, se trata de que o primeiro término *entre-grupos* seja o maior em detrimento do segundo término *intragrupos*. Em outras

palavras o critério é o seguinte:

$$a_{max}[a^{'}\mu_1 - a^{'}\mu_2]^2/a^{-1}\Sigma a$$

Neste caso a solução de a é dada por: $a = k\Sigma^{-1}(\mu_1 - \mu_2)$. Em que k é uma constante arbitrária. Fazendo a condição $a^{'}\Sigma a = 1$, o valor de k será dado por:

$$k = [(\mu_1 - \mu_2)^{'}\Sigma^{-1}(\mu_1 - \mu_2)]^{-1/2}$$

Maximizando a razão de variabilidade entre grupo e variabilidade dentro do grupo, obtém-se o primeiro eixo discriminante.

$$\lambda_1 = \frac{a_1' B a_1}{a_1' W a_1}$$

Derivando λ_1 em relação a a_1 e igualando a zero, ou seja,

$$\frac{\delta \lambda_1}{\delta a_1} = 0$$

Obtém-se então,

$$\frac{\delta \lambda_1}{\delta a_1} = \frac{2Ba_1(a_1' W a_1) - 2Wa_1(a_1' B a_1)}{(a_1' W a_1)^2} = 0$$

$$2Fa_1(a_1' W a_1) = 2Wa_1(a_1' B a_1)$$

Considerando a maximização de λ_1 e operando a igualdade anterior, tem-se:

$$\frac{2Ba_1}{Wa_1} = \frac{a_1' Ba_1}{a_1' Wa_1} \tag{3.1}$$

O segundo membro da Equação 3.1 é justamente igual a maximização de λ_1, ou seja, $\lambda_1 = \frac{a_1' Ba_1}{a_1' Wa_1}$. Assim, obtém-se:

$$Ba_1 = Wa\lambda_1 \tag{3.2}$$

Pré-multiplicando ambos os membros de 3.2 por W^{-1}, sob a suposição que W é uma matriz não singular. Obtém-se então o primeiro eixo discriminante.

$$W^{-1}Ba_1 = W^{-1}W\lambda_1 a_1 \rightarrow Ba_1 = \lambda_1 a_1 \tag{3.3}$$

A obtenção do vetor a_1 é, pois, um problema de cálculo de um vetor característico associado à matriz não simétrica $W^{-1}B$. Das raízes características obtidas ao resolver a equação:

$$Ba_1 - \lambda_1 a_1 = 0 \rightarrow (B - \lambda_1 W)a_1 = 0 \tag{3.4}$$

Ou equivalente,

$$Ba_1 - \lambda_1 a_1 = 0 \rightarrow (B - \lambda_1)a_1 = 0 \tag{3.5}$$

$$(B - \lambda_1 W)v = 0 \tag{3.6}$$

ou equivalente,

$$(W^{-1}W - \lambda_1 I)v = 0 \tag{3.7}$$

Resolvendo a equação, retira-se a maior raiz, que é λ_1, pois se pretende maximizar. No R facilmente como por exemplo, calculam-se os vetores característicos associados à cada raiz característica da seguinte forma:

```
# Raízes e autovetores associados a cada
S<- matrix(c(1,2,4,2), ncol=2, byrow=TRUE)
S
Cal<-eigen(S)
Raizes<-Cal$values
Raizes
VetoresAssociados<- Cal$vectors
#VetoresAssociados
# Verifica-se que estão normalizados
N1<-sum(VetoresAssociados[,1]^2)
N1  #  tamanho 1
N2<-sum(VetoresAssociados[,2]^2)
N2  #  tamanho 1
```

Para dois grupos e duais variáveis a equação que delimita no plano será $(\mathbf{X}_1, \mathbf{X}_2) - M = 0$.

Capítulo 4

Classificação com dois grupos e uma variável classificadora

Considere os dados abaixo relativos a uma amostra de (n=16) indivíduos e (p=1) variável X para dois grupos. Com as informações sobre estas variáveis trata-se de encontrar uma função discriminante que classifique com menor erro possível os grupos I e II. Se tivermos uma boa classificação dos grupos, num passo posterior utiliza-se uma função discriminante para 'encaixar' novos indivíduos nos grupos pré-determinados. O problema agora é classificar cada um dos oito indivíduos nos grupos.

Grupo I	1	2	3	4	5	6	7	8
X	1.3	3.7	5.0	5.9	7.1	4.0	7.9	5.1
Grupo II	1	2	3	4	5	6	7	8
X	5.2	9.8	9.0	12	6.3	8.7	11.1	9.9

O ponto de corte discriminante seria:

$$M = \frac{\bar{\mathbf{X}}_I + \bar{\mathbf{X}}_{II}}{2}$$

Parece razoável tomar o seguinte critério na classificação do indivíduo j:

$X_x < M$, classifica o indivíduo i no G_I.

$X_x > M$, classifica o indivíduo i no G_{II}.

Aplicando este critério comete-se erros de classificação, como se pode comprovar ao examinar a Figura 2.5. Assim, a área pintada à direita de M inclui indivíduos pertencentes ao grupo I, porém nos que $X_i > M$, os indivíduos do grupo I estão mal classificados no grupo II. Reciprocamente, a área mais suave existente à esquerda de M leva os indivíduos ao grupo II porém, nos que $X_i < M$, os indivíduos do grupo II estão mal classificados no grupo I.

As médias amostrais do grupo I é $\bar{X}_I = 5$ e do grupo II, $\bar{X}_{II} = 9$. O ponto de corte será então:

$$M_I = \frac{\bar{\mathbf{X}}_I + \bar{\mathbf{X}}_{II}}{2} = \frac{5+9}{2} = 7$$

O ponto de corte M_I é utilizado para os indivíduos nos grupos pre-estabelecidos a priori. Se a variável $\mathbf{X}_1$ for menor que 7, classifica-se o indivíduo como pertencente ao grupo I e é classificado no grupo II, se a variável $\mathbf{X}_I$ for maior que este valor. Com este critério, pergunta-se: quantos indivíduos foram classificados incorretamente? O exame da coluna da variável $\mathbf{X}_I$ na Tabela 4.1 4.1 permite dar uma resposta imediata a este questionamento.

A Tabela 4.1 também mostra a porcentagem correta e incorreta de classificação de X_1 em cada um dos dois grupos.

Tabela 4.1: Porcentagem de classificação correta e incorreta utilizando apenas a variável $\mathbf{X}_1$.

	Classificado como		
Situação real	Grupo I	Grupo II	Total
Grupo I	6 (75%)	2 (25%)	8 (100%)
Grupo II	2 (25%)	6 (75%)	8 (100%)

De um total de 16 indivíduos, apenas 10 foram classificados, o que equivale a 75% do total. Em concreto, foi classificado incorretamente no grupo *II* os indivíduos 5 e 7, já no grupo *I* foram classificados incorretamente os indivíduos 9 e 13.

4.1 Modelo e Hipóteses

No modelo de referência na análise de variância multivariada, um fator é a variável dependente. Considere a seguinte expressão:

$$y_g = \mu_g + \varepsilon_g \tag{4.1}$$

em que $g = 1, \ldots, G$.

As hipóteses estatísticas sobre a população são as seguintes:

- Os vetores de médias para cada população são diferentes.
- A matriz de covariância de todas as populações é igual Σ (hipótese de homocedasticidade).
- Cada uma das populações tem uma distribuição normal multivariada.

Para testar estas hipóteses considera-se:

$$y_g \sim N(\mu_g, \Sigma)$$

Sendo assim, a hipótese sob o processo de obtenção da amostra facilita a realização do processo de inferência a partir da informação disponível. Deseja-se testar se estas médias são significativamente diferentes. Se as médias forem iguais, não faz sentido discriminar os indivíduos nos grupos.

Supondo que se tenha extraído uma amostra aleatória multivariada independente em cada uma das g grupos. No modelo da Equação 4.1 a hipótese nula e alternativa a contrastar são as seguintes:

$$\begin{cases} H_o: & \mu_1 = \mu_2 \\ H_a: & \text{Nem todas as } \mu_g \text{ são iguais} \end{cases}$$

No próximo exemplo, mostraremos o cálculo do teste desse teste de hipótese.

4.2 Decomposição da matriz de covariância

Supondo que se tem extraído uma amostra em cada grupo, e mediante agregação, obtemos o total da amostra designadas por $n = n_1 + n_2 + \ldots + n_G$. O vetor das médias amostrais globais $\bar{y}$ (ou seja, de todos os grupos) se obtém somando para cada variável todos os valores da amostra e dividindo pelo total da amostra. Dentro de cada grupo pode-se obter o correspondente vetor de médias amostrais $\bar{y}_g$. Tem-se:

Fontes de variação	Graus de liberdade	Matriz de soma de quadrado
Between (B)	G -1	$B = n_g \sum_g (\bar{X}_g - \bar{X})(\bar{X}_g - \bar{X})'$
Within (W)	n - G	$W = \sum_g \sum_u (\bar{X}_{gu} - \bar{X}_g)(\bar{X}_{gu} - \bar{X}_g)'$
Total	T	$T = \sum_g \sum_u (\bar{X}_{gu} - \bar{X})(\bar{X}_{gu} - \bar{X})'$

A matriz da soma dos quadrados e produtos cruzados entre os grupos (B, do inglês *between*), também denominada de matriz da soma de quadrados e produtos cruzados do fatos, se deve a influência do fator. Já a matriz da soma dos quadrados e produtos cruzados dentro dos grupos (W, do inglês *within*). A componente intra grupo é a matriz da soma de quadrado e produto cruzados dos desvios entre cada dado e a média do seu grupo. Chama-se de soma de quadrados e produtos cruzados residual.

A matriz W pode obter por agregação das matrizes da soma de quadrado e produtos cruzados calculados para cada grupo: $W = W_1 + W_2 + \ldots + W_G$, sendo que W_g é a soma de quadrado e produto cruzado no grupo $g = 1, \ldots, G$. A matriz B também pode ser obtida por agregação das matrizes $B = B_1 + B_2 + \ldots, B_G$. Assim, a decomposição da soma de quadrado e produto cruzado total, ou matriz T, é o somatório dessas duas matrizes, ou seja:

$$T = B + W \tag{4.2}$$

Exemplo 4.2.1. *Com objetivo de exemplificar como se calculam as matrizes W, B e T, considere duas variáveis Y_1 e Y_2 e três grupos retirados de tamanhos aleatórios: $n_1 = 4$, $n_2 = 3$ e $n_3 = 5$, conforme se observa na Tabela 4.2.*

Tabela 4.2: Dados hipotéticos de duas variáveis em três grupo.

Grupo I		Grupo II		Grupo III	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
4	5	6	6	11	8
2	3	8	8	9	7
6	3	8	3	10	12
4	5	2	9	8	6
6	2	5	9	7	11
8	9			10	8
				9	9.5
$\bar{y}_{11} = 5$	$\bar{y}_{21} = 4.5$	$\bar{y}_{12} = 5.8$	$\bar{y}_{22} = 7$	$\bar{y}_{13} = 9.1428$	$\bar{y}_{23} = 8.7857$

```
manova.data<-data.frame(group = as.factor(
            rep(1:3, c(6, 5, 7))))
X1<-c(4,2,6,4,6,8,6,8,8,2,5,11,9,10,8,7,
      10,9)
X4<-c(5,3,3,5,2,9,6,8,3,9,9,8,7,12,6,11,
      8,9.5)
with(manova.data, tapply(X1, group, mean))
with(manova.data, tapply(X4, group, mean))
```

Observa-se a dispersão das variáveis de cada grupo na Figura 4.1.

```
par(mfrow = c(2, 1))
boxplot(y1~group, manova.data, main=
        "y1 Boxplot", horizontal = T)
boxplot(y2~group, manova.data, main=
        "y2 Boxplot", horizontal = T)
```

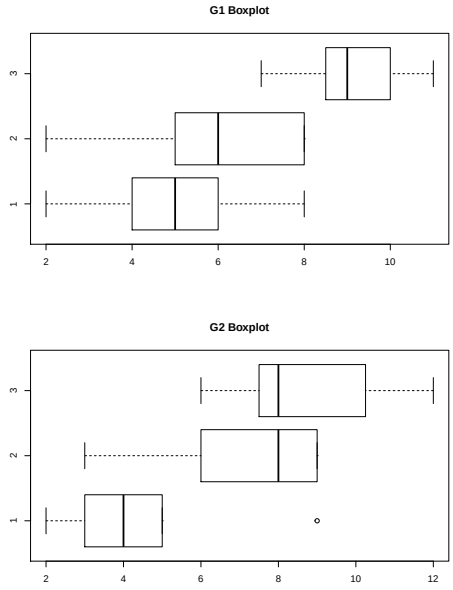


Figura 4.1: Boxplot para os três grupos.

Em que testa as hipóteses:

$$\begin{cases} H_o: & \mu_1 = \mu_2 = \mu_3 \\ H_a: & \mu_i \neq \mu_j, \quad i \neq j. \end{cases}$$

Todos os testes MANOVA atuais são feitas em $A = E^{-1}H$. Feito este cálculo existem quatro diferentes testes multivariados em $E^{-1}H$. Cada um destes testes tem sua própria razão de F associada, em alguns casos exatos em outros aproximados. Os quatro testes clássicos são:

```
Saída<-manova(cbind(G1,G2)~group,manova.data)
manova(cbind(X1, X4) ~ group, manova.data)
summary(Saída, test = "Wilks")
summary(Saída, test="Pillai")
summary(Saída, test="Hotelling-Lawley")
summary(Saída, test="Roy")
```

- Cálculo da soma de quadrado e produto cruzado intragrupo (residual).

No caso de três grupos, tem-se:

$$W = W_1 + W_2 + W_3$$

sendo W_g com $g = 1,2,3$, a soma de quadrados e produtos cruzados para os 3 grupos, respectivamente.

$$W_1 = \sum_{i=1}^{6} (Y_{1g_i} - \bar{Y}_1)(Y_{1g_i} - \bar{Y}_1)'$$

Então temos:

$$\begin{aligned} W_1 = & \begin{bmatrix} 4-5 \\ 5-4.5 \end{bmatrix} \begin{bmatrix} 4-5 & 5-4.5 \end{bmatrix} + \begin{bmatrix} 2-5 \\ 3-4.5 \end{bmatrix} \begin{bmatrix} 2-5 & 3-4.5 \end{bmatrix} \\ & + \begin{bmatrix} 6-5 \\ 3-4.5 \end{bmatrix} \begin{bmatrix} 6-5 & 3-4.5 \end{bmatrix} + \begin{bmatrix} 4-5 \\ 5-4.5 \end{bmatrix} \begin{bmatrix} 4-5 & 5-4.5 \end{bmatrix} \\ & + \begin{bmatrix} 6-5 \\ 2-4.5 \end{bmatrix} \begin{bmatrix} 6-5 & 2-4.5 \end{bmatrix} + \begin{bmatrix} 8-5 \\ 9-4.5 \end{bmatrix} \begin{bmatrix} 8-5 & 9-4.5 \end{bmatrix} \\ W_1 = & \begin{bmatrix} 22 & 13 \\ 13 & 31.5 \end{bmatrix} \end{aligned}$$

No *R*, tem-se:

```
Sa<-matrix(c(4-5, 5-4.5), ncol=2, byrow=TRUE)
Sa<-t(Sa) %*% Sa
Sb<-matrix(c(2-5, 3-4.5), ncol=2, byrow=TRUE)
Sb<-t(Sb) %*% Sb
```

```
Sc<-matrix(c(6-5, 3-4.5), ncol=2, byrow=TRUE)
Sc<-t(Sc) %*% Sc
Sd<-matrix(c(4-5, 5-4.5), ncol=2, byrow=TRUE)
Sd<-t(Sd) %*% Sd
Se<-matrix(c(6-5, 2-4.5), ncol=2, byrow=TRUE)
Se<-t(Se) %*% Se
Sf<-matrix(c(8-5, 9-4.5), ncol=2, byrow=TRUE)
Sf<-t(Sf) %*% Sf
S<- Sa+Sb+Sc+Sd+Se+Sf
S

# S
#      [,1] [,2]
# [1,]   22 13.0
# [2,]   13 31.5
```

Calculando as variâncias e covariâncias do grupo 1, teremos:

```
S11<-(4-5)^2 + (2-5)^2 + (6-5)^2 +
     (4-5)^2 +(6-5)^2 + (8-5)^2
S11
# [1] 22
S22<-(5-4.5)^2 + (3-4.5)^2 + (3-4.5)^2
     +(5-4.5)^2 + (2-4.5)^2 + (9-4.5)^2
S22
# [1] 31.5
SS21<-(4-5)*(5-4.5)+(2-5)*(3-4.5)+(6-5)
      *(3-4.5)+(4-5)*(5-4.5)+(6-5)*
      (2-4.5)+(8-5)*(9-4.5)
SS21
# [1] 13
```

Semelhantemente, calcular-se as matrizes W_2 e W_3:

```
S21<-(6-5.8)^2 + (8-5.8)^2 + (8-5.8)^2 +
```

```
        (2-5.8)^2 + (5-5.8)^2); S21
# [1] 24.8
S21<-(6-5.8)^2 + (8-5.8)^2 + (8-5.8)^2 +
        (2-5.8)^2 + (5-5.8)^2; S21
[1] 24.8
S22<-(6-7)^2+(8-7)^2+(3-7)^2+(9-7)^2+(9-7)^2
S22 # [1] 26
SS22<-(4-5.8)*(6-7)+(8-5.8)*(8-7)+(8-5.8)*(3-7)
      +(2-5.8)*(9-7)+(5-5.8)*(9-7); SS22
[1] -14
SS22<-(6-5.8)*(6-7)+(8-5.8)*(8-7)+(8-5.8)*(3-7)
      +(2-5.8)*(9-7)+(5-5.8)*(9-7); SS22
# [1] -16
```

Encontrando W_2 e W_3, respectivamente tem-se:

$$W_2 = \begin{bmatrix} 24,8 & -16 \\ -16 & 26 \end{bmatrix} \quad \text{e} \quad W_2 = \begin{bmatrix} 24,8 & -16 \\ -16 & 26 \end{bmatrix}$$

Finalmente a matriz agregada de soma de quadrados e produtos cruzados dentro dos grupos (residual) será:

$$W = W_1 + W_2 + W_3 = \begin{bmatrix} 57.6571 & -3.7857 \\ -3.7857 & 85,4285 \end{bmatrix}$$

- Cálculo da soma de quadrados e produtos cruzados entre grupos.

O cálculo da matriz B é feito da seguinte forma:

$$B = \sum_{g=1}^{G} n_g(\bar{Y}_{1g} - \bar{Y}_1)(\bar{Y}_{pg} - \bar{Y}_p)'$$

O vetor de médias será:

$$\bar{Y}_1 = 123/18 = 6,8333 \quad \text{e} \quad \bar{Y}_2 = 123.5/18 = 6.8611$$

$$\text{ou} \quad \bar{Y} = \begin{bmatrix} 6,8333 \\ 6.8611 \end{bmatrix}$$

Pode-se facilmente encontrar os vetores de médias de cada grupo da seguinte forma:

```
dados<-data.frame(group = as.factor(rep(1:3,
      c(6, 5, 7))),
X1<-c(4,2,6,4,6,8,6,8,8,2,5,11,9,10,
      8,7,10,9),
X4<-c(5,3,3,5,2,9,6,8,3,9,9,8,7,12,
      6,11,8,9.5))
with(dados, tapply(X1, group, mean))
#        1        2        3
# 5.000000 5.800000 9.142857
with(dados, tapply(X4, group, mean))
#        1        2        3
# 4.500000 7.000000 8.785714
```

A matriz *B* pressupõe o conhecimento das médias globais das variáveis 1 e 2 de cada grupo, então:

```
b1<-matrix(c(5-6.8333, 4.5-6.8611),ncol=1,
           byrow=TRUE)
b2<-matrix(c(5.8-6.8333, 7-6.8611),ncol=1,
           byrow=TRUE)
b3<-matrix(c(9.1428-6.8333, 8.7857-6.8611),
           ncol=1, byrow=TRUE)
B<-6*(b1%*%t(b1))+5*(b2%*% t(b2))+7*(b3%*%t(b3))
```

```
B
#           [,1]     [,2]
# [1,] 62.84101 56.36805
# [2,] 56.36805 59.47382
```

Portanto, a matriz B será:

$$\begin{aligned} B &= 6\begin{bmatrix} 5-6.8333 \\ 4.5-6.8611 \end{bmatrix}\begin{bmatrix} 5-6.8333 & 4.5-6.8611 \end{bmatrix} + \\ &\quad 5\begin{bmatrix} 5.8-6.8333 \\ 7-6.8611 \end{bmatrix}\begin{bmatrix} 5.8-6.8333 & 7-6.8611 \end{bmatrix} + \\ &\quad 7\begin{bmatrix} 9.1428-6.8333 \\ 8.7857-6.8611 \end{bmatrix}\begin{bmatrix} 9.1428-6.8333 & 8.7857-6.8611 \end{bmatrix} \\ B &= \begin{bmatrix} 62.84101 & 56.36805 \\ 56.36805 & 59.47382 \end{bmatrix} \end{aligned}$$

- Assim a matriz de soma de quadrado e produto cruzado total (T) será:

$$T = B + W = \begin{bmatrix} 62.84101 & 56.36805 \\ 56.36805 & 59.47382 \end{bmatrix} + \begin{bmatrix} 57.6571 & -3.7857 \\ -3.7857 & 85.4285 \end{bmatrix}$$

$$T = \begin{bmatrix} 120.49811 & 52.58235 \\ 52.58235 & 144.90232 \end{bmatrix}$$

Através do seguinte algoritmo será possível encontrar a média e o desvio padrão dos grupos.

```
Media_Desvio_Grupos<-function(
  Variaveis,GruposVariaveis)
{
  # Nomes das variáveis em todos os grupos
  NomesVariaveis <-names(GruposVariaveis),
  names(as.data.frame(Variaveis)))
```

```
  # cada variavel dentro do grupo
  GruposVariaveis <- GruposVariaveis[,1]
  means<-aggregate(as.matrix(Variaveis)
  ~ GruposVariaveis, FUN = mean)
  names(means) <- NomesVariaveis
  print(paste("Means:"))
  print(means)
  # Matriz dentro do grupos.
  # Encontra cada desvio padrão
  DPs<-aggregate(as.matrix(Variaveis) ~
  GruposVariaveis, FUN = sd)
  names(DPs) <- NomesVariaveis
  print(paste("Desvio padrão:"))
  print(DPs)
  # Dentro de cada grupo se tem o tamanho
  # da amostra
  samplesizes<-aggregate(as.matrix(Variaveis)
              ~GruposVariaveis, FUN = length)
  names(samplesizes)<-NomesVariaveis
  print(paste("Tamanho da amostra:"))
  print(samplesizes)
}
colnames(dados)
Media_Desvio_Grupos(dados[2:3],dados[1])
```

Cálculo de variância dentro do grupo (w).

```
Variancia_Dentro_Grupo<-function(
         variavel,GrupoVariavel)
{
  GrupoVariavel2<-
  as.factor(GrupoVariavel[[1]])
```

```
  levels<-levels(GrupoVariavel2)
  Num_níves<-length(levels)
  # Média e desvio de cada grupo
  Num_total<-0
  denomtotal<-0
  for (i in 1: Num_nives)
  {
  leveli<-levels[i]
  levelidata<-variavel[GrupoVariavel==leveli,]
  levelilength<-length(levelidata)
  # Encontrando o desvio padrão do grupo i
  Sdi<-sd(levelidata)
  Numi<-(levelilength - 1)*(Sdi * Sdi)
  Denomi<-levelilength
  Num_total<-Num_total + Numi
  denomtotal<-denomtotal + Denomi
  }
  # Cálculo da variância dentro dos grupos
  Var_dentro<-Num_total / (denomtotal -
              Num_níves)
  return(Var_dentro)
}
Variancia_Dentro_Grupo(dados[2],dados[1])
Variancia_Dentro_Grupo(dados[3],dados[1])
```

Encontrando a variância para o caso de duas variáveis.

```
Calculo_Var_dentro_Grupos<-function(
           variavel1,variavel2,GrupoVariavel)
 {
GrupoVariavel2<-as.factor(GrupoVariavel[[1]])
levels<-levels(GrupoVariavel2)
levels
```

```
Num_niveis<-length(levels)
Num_niveis
# Encontrando a variância 1 e 2 de cada grupo
Covw<-0
 for (i in 1:Num_niveis)
  {
  leveli<-levels[i]
  levelidata1<-variavel1[GrupoVariavel ==
                         leveli,]
  levelidata2<-variavel2[GrupoVariavel ==
                         leveli,]
   media1<-mean(levelidata1)
   media2<-mean(levelidata2)
   levelilength<-length(levelidata1)
   # obter a covariância para este grupo:
   term1<-0
   for (j in 1:levelilength)
   {
   term1<-term1 + ((levelidata1[j] - media1)*
    (levelidata2[j] - media2))
   }
   Cov_groupi<-term1 # covariância para
                     # este grupo
   Covw <- Covw + Cov_groupi
 }
 totallength<-nrow(variavel1)
 Covw<-Covw / (totallength - Num_niveis)
 return(Covw)
}
Calculo_Var_dentro_Grupos(dados[2],dados[3],
                          dados[1])
```

Calculando a variância entre grupos.

```
Calculo_Covariancia_dentro_Grupos<-
function(variavel1,variavel2,GrupoVariavel)
{
  GrupoVariavel2<-as.factor(GrupoVariavel[[1]])
  levels<-levels(GrupoVariavel2)
  Num_niveis<-length(levels)
  # Calculo das médias
  mediavariavel1<-mean(variavel1)
  mediavariavel2<-mean(variavel2)
  # Calculando a matriz de covariância
  Covb<-0
  for (i in 1:Num_niveis)
{
  leveli<-levels[i]
 levelidata1<-variable1[GrupoVariavel ==
              leveli,]
  levelidata2<-variable2[GrupoVariavel ==
               leveli,]
  media1<-mean(levelidata1)
  media2<-mean(levelidata2)
  levelilength<-length(levelidata1)
  term1<-(media1 - mediavariavel1)
  *(media2 - mediavariavel2)*(levelilength)
  Covb<-Covb + term1
}
  Covb<-Covb / (Num_niveis - 1)
  Covb<-Covb[[1]]
  return(Covb)
}
attach(dados)
dados
```

```
Calculo_Covariancia_dentro_Grupos(X1,X4,G)
```

4.3 Encontrando a estatística Λ de Wilks

$$\Lambda = \frac{|W|}{|B+W|} = \frac{\begin{vmatrix} 57.6571 & -3.7857 \\ -3.7857 & 85.4285 \end{vmatrix}}{\begin{vmatrix} 120.4981 & 52.5823 \\ 52.58235 & 144.9023 \end{vmatrix}}$$

$$\Lambda = \frac{57.6571 \cdot (85.4285) - 3.7857^2}{120.4981 \cdot (144.9023) - 52.5823^2} = 0.3341$$

No *R*, teremos:

```
Wilks<- det(W)/det(B+W)
Wilks<- det(W)/det(T)
Wilks
# [1] 0.3341983
round(Wilks,4)
# [1] 0.3341
```

```
Manova <- data.frame(group = as.factor(rep(1:3,
+ c(6, 5, 7))),
X1<-c(4, 2, 6, 4, 6, 8, 6, 8, 8, 2, 5, 11, 9,
     10, 8, 7, 10, 9),
X4<-c(5, 3, 3, 5, 2, 9, 6, 8, 3, 9, 9, 8, 7,
     12, 6, 11, 8, 9.5))
with(Manova, tapply(X1, group, mean))
```

```
with(Manova, tapply(X4, group, mean))
(m1 <- manova(cbind(X1, X4)~group, Manova))
summary(m1, test = "Wilks")
#          Df  Wilks approx F num Df den Df Pr(>F)
# group     2 0.33419 5.1087 4   28   0.003216 **
# Residuals 15
# ---
# Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05
# '.' 0.1 ' ' 1
```

- Aproximação à distribuição F de Snedecor.

Uma aproximação melhor foi sugerida por Rao (1951) e é utilizada pela maioria dos pacotes estatísticos. Sendo:

$$r = n - 1 - \frac{1}{2}(p+G)$$

Sabendo que:

$$t = \sqrt{\frac{p^2 \cdot (G-1)^2 - 4}{p^2 + (G-1)^2 - 5}}$$

Então:

$$V = \frac{r \cdot t - \frac{1}{2} p \cdot (G-1) + 1}{p \cdot (G-1)} \cdot \frac{1 - \Lambda^{\frac{1}{t}}}{\Lambda^{\frac{1}{t}}} \sim F_{\left[(p \cdot (G-1)); (r \cdot t - \frac{1}{2} p \cdot (G-1) + 1)\right]}$$

Aplicando o exemplo, tem-se:

$$r = 18 - 1 - \frac{1}{2}(2+3) = 14.5 \quad \text{e} \quad t = \sqrt{\frac{2^2 \cdot (3-1)^2 - 4}{2^2 + (3-1)^2 - 5}} = \sqrt{\frac{12}{3}} = 2$$

$$\text{então,} \quad V = \frac{14.5 \cdot 2 - \frac{1}{2}2 \cdot (3-1)+1}{2 \cdot (3-1)} \cdot \frac{1-(0.3341)^{\frac{1}{2}}}{(0.3341)^{\frac{1}{2}}} = 5.1088$$

Sendo $n = n_1 + n_2 + n_3 = 18$, teremos em *R*.

```
n<-18
r<-n-1-(0.5)*(p+G)
r
A<-(p^2)*(G-1)^2-4
B<-p^2+(G-1)^2-5
t<-sqrt(A/B)
AproxF<-(r*t-(0.5)*p*(G-1)+1)/
         (p*(G-1))*(1-L^(1/t))/
         (L^(1/t))
AproxF
# [1] 5.108806
```

Sendo os graus de liberdade `numDf`$= (2 \cdot (3-1)) = 4$ e `denDF`$= 14.5 \cdot 2 - \frac{1}{2}2 \cdot (3-1)+1 = 28$.

```
v1<- 2*(G-1)
v2<-r*t-(0.5)*p*(G-1)+1
v1
# [1] 4
v2
# [1] 28
```

```
AproxF<-(r*t-(0.5)*p*(G-1)+1)/(p*(G-1))*
           (1-L^(1/t))/(L^(1/t)); AproxF
# [1] 5.108806
```

```
1-pf(AproxF, numDf, denDF)
# [1] 0.003215332
```

$$AproxF \sim F_{[numDf;\quad denDF]} = 2.7140$$

```
qf(0.95,numDf,denDF)
# [1] 2.714076
```

O valor-*p* associado ao valor da estatística de teste `Aprox F` é 0.00321, sendo que a hipótese nula deve ser rejeitada ao nível de significância de 5%.

- Aproximação à χ^2: $-\left[(n_1+n_2+n_3-1)-\frac{1}{2}(p+G)\right] ln(\Lambda) \sim \chi^2_{(p(G-1))}$.

Substituindo-se os valores, tem-se:

$$-\left[(18-1)-\frac{1}{2}(2+3)\right] ln(0.3314) = 15.8926 \quad \text{com 4 graus de liberdade}$$

```
# Aprox Qui-quadrada
n1<-6
n2<-5
n3<-7
n<-n1+n2+n3
n
p<- 2
k<- 3
L<-  0.33419
QuiQ<- -((n1+n2+n3-1)-(0.5)*(p+G))*log(L)
QuiQ
# [1] 15.89266
```

Observando os graus de liberdade da $\chi^2_{2(3-1),5\%} = 9,49 =$ `(qchisq`(0.95, 4)= 09,49).

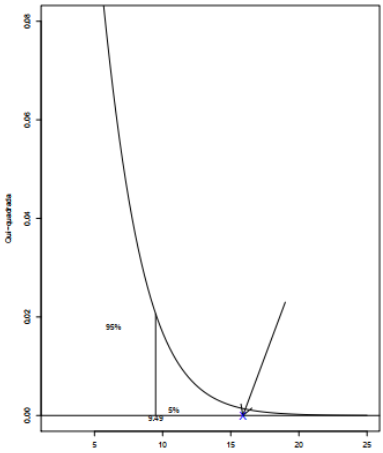


Figura 4.2: Densidades da χ^2 com 4 graus de liberdade.

4.3.1 Teste de hipóteses de várias médias (MANOVA) usando a razão de verossimilhança

Aqui considera o contraste da razão de verossimilhança. Novamente a hipótese a ser testada será:

Em que testa as hipóteses:

$$\begin{cases} H_o: & \mu_1 = \mu_2 = \mu_3 = \mu \\ H_a: & \mu_i \neq \mu_j, \quad i \neq j. \end{cases}$$

A função de verossimilhança sob a hipótese nula, H_o de uma amostra normal homogênea para o máximo alcance da razão de verossimilhança é $\bar{x} = x$) e $\hat{V} = S$, em que S é a matriz de covariância total.

A verossimilhança $L(V|S)$ será máxima se todos os autovalores de $V^{-1}S$ forem iguais à unidade, o que implica que $V^{-1}S = I$. Logo isso se consegue tomando como estimador de máxima verossimilhança (MV) de

V, $\hat{V} = S$. Sabe-se que os estimadores de MV de μ e V são $\bar{x}$ e S.

Vamos encontrar a função suporte para receber estes estimadores.

$$f(X|\mu,V) = \Pi_{i=1}^{n}|V|^{-1/2}(2\pi)^{-p/2}\exp(-(1/2)(x-\mu)'V^{-1}(x-\mu))$$

Desprezando os constantes, a função suporte será:

$$L(\mu,V|X) = -\frac{n}{2}\log|V| - \frac{1}{2}\sum_{i=1}^{n}(x-\mu)'V^{-1}(x-\mu)$$

Substituindo a média amostral $\bar{x}$ e a variância amostral S em 4.3, tem-se:

$$L(\mu,V|X) = -\frac{n}{2}\log|V| - \frac{n}{2}trV^{-1}S - \frac{n}{2}(x-\mu)'V^{-1}(x-\mu)$$

A expressão 4.3 será utilizada para suporte dos parâmetros em amostra normal multivariada. A função apenas depende da amostra através dos valores $\bar{x}$ e S, que serão, portanto, estimadores suficientes de μ e S.

Para obter o estimador do vetor de média na população, utilizamos que, por ser V^{-1} definida positiva, $(x-\mu)'V^{-1}(x-\mu) \geq 0$.

Se diz que Λ segue uma distribuição de Wilks com parâmetros $\Lambda(p,\alpha,\beta)$. Para $\beta = 1$ e 2 esta distribuição pode ser expressa em função da distribuição F de Snedecor (MARDIA et al., 1979). Em outros casos utiliza-se:

$$-\left(\alpha-\frac{1}{2}(p-\beta+1)\right)-\log\Lambda(p,\alpha,\beta)\sim\chi^2_{\alpha p}$$

Se aproxima a distribuição qui-quadrada com αp graus de liberdade. Se tomamos $\alpha = n-1$ e $\beta = G-1$ obtém-se:

$$\lambda_o = J\log\frac{|T|}{|W|} = J\log\frac{|W+B|}{|W|} = J\log|I+W^{-1}B|$$

Sabendo que $|I+B| = \Pi(1-\lambda_i)$, em que λ_i são os autovetores da matriz $W^{-1}B$.

Da distribuição qui-quadrada se obtém da diferença entre ambos espaços paramétricos, ou seja, sob a hipótese nula o número de parâmetros será: $p+p(p+1)/2$, no caso $p=2$, 5 parâmetros, ou seja duas médias mais 2 variâncias e mais uma covariância. Sob a hipóteses H_1 estima-se G vetores de médias mais a matriz de covariância, o que supõem $Gp-p(p+1)/2$. Como exemplo, para o caso de 2 variáveis e 3 grupos têm-se: 6 médias (duas de cada grupo), 6 variâncias e 3 covariâncias, fazendo um total de 15, pois $Gp-p(p+1)/2 = 3.2-2(2+1)/2 = 15$ parâmetros. A diferença será então:

$$\begin{aligned} D_g &= dim(H_1-H_o) \\ &= Gp+p(p+1)/2-(p+p(p+1)/2 \\ &= Gp-p = p(G-1) \end{aligned}$$

A aproximação a distribuição $\chi^2_{D_g}$ do quociente da verossimilhança pode ser melhor para tamanho de amostras pequenas.

$$\lambda_o = J \log \frac{|S|}{|S_w|}$$

O subíndice w indica que é a matriz de covariância dentro do grupo (*within*). Em que $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$, e $J = (n-1) - (p+G)/2$, segue uma distribuição χ^2_{Dg}, em que D_g vem dado pela Equação 4.3 e a distribuição é melhor quando tomamos $J = n$ em pequenas amostras.

Em que, V é definida positiva e idêntica aos 3 grupos considerados. A hipótese alternativa significa que nem todas as médias são iguais com a mesma condição para V. O máximo alcance para razão de verossimilhança será $(\bar{x} = x)$ e $\hat{V} = S$.

Antes de aplicar no *R*, vamos explicar o que significa variância efetiva. PEÑA (2002) propõem como medida global de variabilidade a variância efetiva, dada por:

$$VE = |S|^{1/p}$$

Tem a vantagem de que quando todas as variáveis têm as mesmas dimensões tem as unidades da variância. Para matrizes diagonais, esta medida é simplesmente a média geométrica das variâncias. Semelhante, pode-se definir o desvio-padrão efetivo mediante

$$DpE = |S|^{1/2p}$$

```
SW<-matrix(c(57.6571,-3.7857,-3.7857,85.4285)
```

```
           ,ncol=2)
SW
ST<-matrix(c(120.4981,52.5823,52.5823,
             144.9023),ncol=2); ST
# Variância efetiva para ambas
# as hipóteses Ho e H1
VEHo<-det(ST)^(1/p); VEHo
VEH1<-det(SW)^(1/p); VEH1
n1<-6; n2<-5; n3<-7
n<-n1+n2+n3
p<-2 # 2 variáveis
G<-3 # 3 grupos
Lo<-p*((n-1)-(p+G)/2)*(log(VEHo) - log(VEH1))
Lo
pchisq(Lo, p*(G-1)) # pvalor=0.6622463 > 0,05
```

Considerando este simples exemplo, observa-se que não existe evidências estatística para firmar que as médias são diferentes. Devemos ter atenção a este resultado, por ser decorrente de uma amostra muito pequena de duas variáveis.

4.4 Bondade da análise e lambda de Wilks

A matriz confusão é uma tabela genérica de frequências de acertos e erros de classificação, ou seja, aparecem os resultados dos indivíduos bem classificados e também os que não foram classificados corretamente. Também pode-se observar nesta matriz os percentuais tanto de cada grupo como no total, junto com o número de indivíduos que se tem classificado em cada grupo. Isso serve também para interpretar a bondade da análise. Compara-se a máxima probabilidade a priori de pertencer a um grupo, sem considerar a informação proporcionada pelas variáveis selecionadas,

com a porcentagem dos que estejam bem classificados nos não selecionados depois de ter em conta a análise discriminante. Se este último aumenta muito, se tem obtido uma melhora e poderíamos falar de uma classificação boa mediante a análise discriminante.

Não apenas a porcentagem de casos que estejam bem classificados também serve para avaliar a eficácia das funções discriminantes, existem outras estatísticas para esse fim, dentre elas, temos:

a) Lambda de Wilks das funções discriminantes.

O lambda de Wilks mede os desvios das pontuações discriminantes dentro dos grupos em relação aos desvios totais sem distinguir os grupos. Assim, se o valor é grande, tende a 1, a dispersão é devido a diferença entre grupos e, portanto, estarão pouco separadas. O primeiro valor de lambda coincide com o total do conjunto de variáveis selecionadas e este é o valor para o conjunto formado por todas as funções discriminantes, sem eliminar nenhuma delas.

A extração das funções discriminantes se faz com um critério de cada uma, em que se vai aportando menos informações que a anterior, o seguinte valor tem que ser maior (valor que temos que contrastar), já que é o valor de eliminar o efeito da primeira função, ou seja, considerando a informação da segunda função. E isto se repete com cada uma das funções na ordem até chegar à última apenas.

A estatística de teste qui-quadrada permite contrastar a hipótese nula H_o: (Centro dos grupos iguais), ou mesmo, a função ou funções não separam os grupos. Para uma significação menor que α rejeita-se a hipótese nula e concluímos que a função discriminante separa os grupos. Isto se faz para cada uma das funções.

Pergunta-se: é importante a variabilidade total explicada pelo fator?

Para responder esta pergunta, utiliza-se como medida de bondade do ajuste um estatístico. A razão é a proporção da soma de quadrado generalizada total não explicada pelo fator. O valor de $1-\Lambda$ será a proporção explicada pelo fator. Esta proporção, a que denomina-se também (η^2) (eta quadrado), toma-se como medida de bondade do ajuste, sendo uma generalização do coeficiente de determinação ao caso multivariado. A expressão é dada por:

$$\text{Eta quadrado} = \eta^2 = 1 - \Lambda = 1 - \frac{|W|}{|T|} \tag{4.3}$$

De modo que os valores próximos à unidade indica que a maior parte da variabilidade total pode atribuir-se ao fator, enquanto que um valor próximo a zero significa que o fator explica muito pouco essa variabilidade total.

b) Correlação canônica e autovalores.

São valores atribuídos a cada uma das funções discriminantes extraídas. A correlação canônica mede os desvios das pontuações discriminantes entre grupos em relação aos desvios totais sem distinguir grupos. Se seu valor é alto, próximo de 1, a dispersão será devida as diferenças entre grupos. Já o autovalor mede os desvios das pontuações discriminantes entre os grupos em relação aos desvios dentro dos grupos. Interpreta-se como parte da variabilidade total da nuvens de pontos projetadas sobre o conjunto de todas as funções atribuídas à função correspondente. Se seu valor é grande, a função também discriminará muito.

Como é lógico, os valores da correlação canônica decrescem. Sempre tem que discriminar mais a primeira componente que a segunda, a segunda que terceira, etc. Isso acontece o mesmo com os autovalores. Estes autovalores podem ser interpretados como variabilidade total explicada. A primeira função explica um total de variabilidade (% de variância) maior que a segunda, a segunda maior que a terceira e assim sucessivamente até explicar tudo (% acumulado). A primeira sempre dar praticamente a classificação.

c) Médias dos grupos para as funções.

Se as médias dos grupos em cada função são muito parecidas, a função discriminante não discrimina os grupos, isto é, o que acabamos de ver com a qui-quadrada para a significação de lambda da função discriminante.

Capítulo 5

Principais pressupostos da função discriminante e outros testes

A técnica consiste em encontrar combinações lineares das variáveis independentes para descriminar indivíduos pertencentes a diferentes grupos, é considerada "a melhor"quando permite a minimização dos erros de incorreta classificação. A investigação científica trata de estabelecer afirmações de natureza casual quando se analisa uma variável e se medem seus efeitos sobre outra variável. Quando as variáveis têm "manipulado"de forma direta os indivíduos, sendo indicados aleatoriamente aos grupos, a garantia de casualidade é maior. No entanto este fato só é verdadeiro quando se verificam seguintes limitações:

1. Normalidade multivariada.
2. Homogeneidade de variância-covariância.
3. Linearidade.
4. Multicolinearidade.

5. Pontos extremos.
6. Tamanho da amostra.

Em outras palavras:

- Os grupos deverão ser retirados de populações normais que seguem uma distribuição normal multivariada para as p variáveis discriminantes.
- Dentro dos grupos a variabilidade deverá ser idêntica, isto é, as matrizes de variância e covariância devem ser aproximadamente iguais para todos os grupos.
- O suposto de linearidade implica que os preditores (as variáveis independentes) estejam relacionados linearmente entre si em cada grupo. Sua ausência reduz a potência de classificação.
- Se os preditores estão altamente relacionados podem produzir multicolinearidade e impedir a análise uma vez que a matriz de erro da soma dos quadrados e produtos cruzados carecerá de inversa. Quando isso ocorre o melhor é eliminar as variáveis que produzem a multicolinearidade
- Pontos extremos podem ser uma observação que apresenta um grande afastamento dos demais pontos dentro desta variável.
- É necessário que existam pelo menos dois grupos (caso univariado) e mais de dois grupos (caso multivariado) e para cada grupo exista dois ou mais casos.

Estes suposições representam na análise discriminante em especial condições que garantirá a existência de um isomorfismo entre a situação real e o modelo estatístico da análise. Basicamente, os pressupostos são normalidade multivariada, homogeneidade de matriz de covariância, linearidade, ausência de multicolinearidade e pontos extremos. Vamos estudar cada um individualmente.

5.1 Normalidade bivariada

Esta distribuição é um caso particular da distribuição normal p-dimensional para $p = 2$. Para exemplificar considere 4 diferentes matrizes de covariâncias. Os subscritos de σ informam os valores das covariâncias consideradas. Observa-se que com pequena correlação o gráfico da normal bidimensional fica mais pontiagudo.

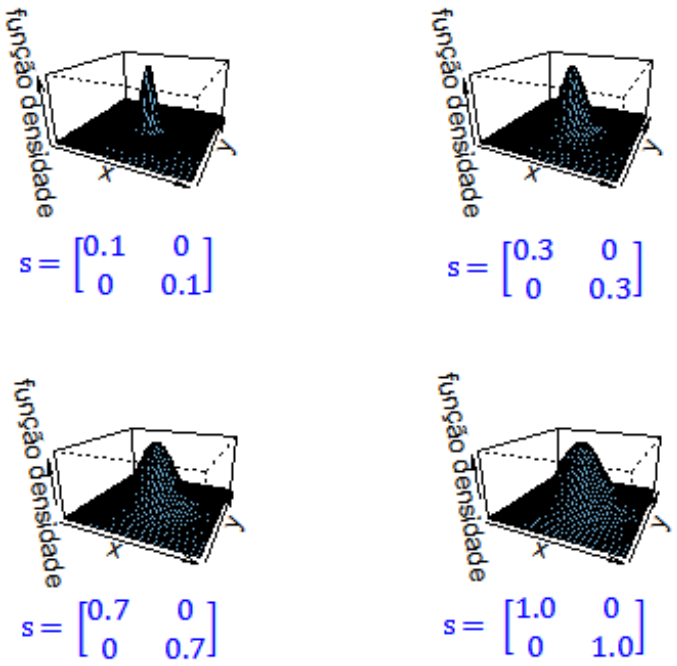


Figura 5.1: Exemplos de normal bivariada com diferentes variâncias.

Os passos no *R* para construção destas figuras se encontram abaixo:

```
# sigma<-matrix(c(0.1,0,0,0.1)
library(mvtnorm)
n = 50
x = seq(-3, 3, length = n)
y=x
z = matrix(0,n,n)
sigma<-matrix(c(0.1,0,0,0.1), ncol=2, byrow=TRUE)
for (i in 1:n)
for (j in 1:n)
z[i,j]=dmvnorm(c(x[i],y[j]),c(0,0), sigma)
```

```
end
end
persp(x,y,z,theta=25,phi=20,zlab="density
         function", expand=0.5, col="#6BAED6")
# sigma<-matrix(c(0.3,0,0,0.3)
n = 50
x = seq(-3, 3, length = n)
y=x
z = matrix(0,n,n)
sigma<-matrix(c(0.3,0,0,0.3), ncol=2, byrow=TRUE)
for (i in 1:n)
for (j in 1:n)
z[i,j] = dmvnorm(c(x[i],y[j]),c(0,0), sigma)
end
end
persp(x,y,z,theta=25,phi=20,zlab="density
      function", expand=0.5, col="#6BAED6")
# sigma<-matrix(c(0.7,0,0,0.7)
n = 50
x = seq(-3, 3, length = n)
y=x
z = matrix(0,n,n)
sigma<-matrix(c(0.7,0,0,0.7), ncol=2, byrow=TRUE)
for (i in 1:n)
for (j in 1:n)
z[i,j] = dmvnorm(c(x[i],y[j]),c(0,0), sigma)
end
end
persp(x,y,z,theta=25,phi=20,zlab="density
      function", expand=0.5, col="#6BAED6")
# sigma<-matrix(c(1,0,0,1)
n = 50
x = seq(-3, 3, length = n)
```

```
y=x
z = matrix(0,n,n)
sigma<-matrix(c(1,0,0,1), ncol=2, byrow=TRUE)
for (i in 1:n)
for (j in 1:n)
z[i,j]=dmvnorm(c(x[i],y[j]),c(0,0), sigma)
end
end
persp(x,y,z,theta=25,phi=20,zlab="density
      function", expand=0.5, col="#6BAED6")
```

A análise das distribuições marginais univariadas e bivariadas auxiliam na verificação da suposição de normalidade multivariada (Looney, 1995). Para cada variável do conjunto de dados amostrais, a suposição de normalidade univariada pode ser verificada através de gráficos como o de probabilidade normal, histograma ou ramo e folhas (Johnson; Bhattacharyya, 2001), ou através de hipóteses apropriadas como Ryan Joiner (1976), Shapiro-Wilks (1965), Anderson-Darling (1952) e Zhang (1999). Vamos estudar o teste de Shapiro-Wilks, conforme se observa na Equação 5.1 por ser bem conhecido e bem aceito pela comunidade estatística.

$$V = \frac{\left(\sum_{i=1}^{G} a_{i,n}\left(X_{n-i+1} - X_{(i)}\right)\right)^2}{(n-1)S^2}$$

Em que:

(a) $k = \frac{n}{2}$ se n é par, e $k = \frac{n-1}{2}$ se n é ímpar.

(b) $\mathbf{X}_i$ é a estatística de ordem i-ésimo.

(c) $a_{i,n}$ coeficientes tabulados segundo o tamanho amostral.

A estatística V de Shapiro-Wilks se usa para contrastar:

$$\begin{cases} H_o: & \text{X é normal} \\ H_1: & \text{X não é normal} \end{cases}$$

Propriedades:

- Verifica-se que dando a igualdade a 1 apenas no caso de que a amostra seja a replica de uma normal.
- Os valores do estatística de Shapiro-Wilks encontram-se tabulados. Na tabela aparecem os pontos críticos $(V_{n,\alpha})$, tal que $P[V \leq V_{n,\alpha}] = \alpha$.

Apresenta-se, então, um exemplo para exemplificar do teste de normalidade de Shapiro-Wilks. Considere a seguinte amostra:

$$X = [19.8, 20.5, 19.7, 17.6, 19.2, 18.4, 18.1, 19.1, 17.9, 17.3, 20]$$

Contrasta-se a hipótese de normalidade da população da qual se extraiu a amostra X usando o Teste de Shapiro-Wilks. Para isso se observa os seguintes passos.

1. Se ordena os 11 valores amostrais em sentido decrescente: $(20.5, 20, 19.8, 19.7, 19.2, 19.1, 18.4, 18.1, 17.9, 17.6, 17.3)$.
2. Obtém-se as diferenças entre os valores equidistante do centro: $[20.5 - 17.3 \quad 20 - 17.6 \quad 19.8 - 17.9 \quad 19.7 - 18.1 \quad 19.2 - 18.4]$.
3. Multiplica-se cada uma dessas diferenças pelo correspondente coeficiente apresentados por colunas como se observa abaixo:

	n									
i	2	3	4	5	6	7	8	9	10	11
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739	0,5601
2		0	0,1677	0,2413	0,2806	0,3031	0,3031	0,3244	0,3291	0,3315
3				0	0,0875	0,1401	0,1743	0,1976	0,2141	0,226
4						0	0,0561	0,0947	0,1224	0,1429
5								0	0,0399	0,0695
6										0

Como $n = 11$ temos,

$x_{(n-i+1)-x_{(i)}}$	$a_{i,n}$	$a_{i,n}((n-i+1)-x_{(i)})$
20,5-17,3=3,2	0,5601	$0,5601 \cdot (3,2) = 1,7923$
20,0-17,6=2,4	0,3315	0,7956
19,8-17,9=1,9	0,2260	0,4294
19,7-18,1=1,6	0,1429	0,2286
19,2-18,4= 0,8	0,0695	0,0556

4. Observa-se o valor calculado da estatística V. A variância amostral será:

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{11} x_i^2 - \frac{\left(\sum_{i=1}^{11} x_i\right)^2}{n}\right]$$

$$S^2 = \frac{1}{10}\left[19,8^2 + \ldots + 20^2 - \frac{(19,8 + \ldots + 20)^2}{11}\right] = 1,14$$

Logo, V será:

$$V = \frac{\left(\sum_{i=1}^{k} a_{i,n}\left(X_{n-i+1} - X_{(i)}\right)\right)^2}{(n-1)S^2}$$

$$V = \frac{\left(0,5601 \cdot (3,2) + \ldots + 0,0695 \cdot (0,8)^2\right)}{10 \cdot (1,14)} = 0,9493$$

Usando a Tabela Shapiro-Wilks vemos que $V_{11,5\%} = 0,85$.

5. Logo, $v_o = 0,9493 > V_{11,5\%} = 0,85$, ao nível de 5%, assim não há evidências estatística para rejeitar a hipótese de que a amostra tenha sido extraída de uma população normal. Assim, X foi extraído de uma população normal, com 5% de incerteza.

Ou simplesmente, usando o *R* podemos encontrar diversos testes de normalidade.

```
 x<-c(19.8,20.5,19.7,17.6,19.2,18.4,
          18.1,19.1,17.9,17.3,20)
 shapiro.test(x)
Shapiro-Wilk normality test
data:  x
W = 0.94971, p-value = 0.6402
ksnormTest(x)
shapiroTest(x)
```

```
pchiTest(x)
jarqueberaTest(x)
```

Usando também o pacote *nortest* (GROSS e LIGGES, 2015) para outros testes de normalidade.

```
library(nortest)
S<-adTest(x)
S
    # Title:
# Anderson - Darling Normality Test
# Test Results:
# STATISTIC:
# A: 0.2598
# P VALUE:
# 0.6369
cvmTest(x)
# Title:
# Cramer - von Mises Normality Test
# Test Results:
# STATISTIC:
# W: 0.043
# P VALUE:
# 0.5918
lillieTest(x)
sfTest(x)
```

5.1.1 Distribuição normal bivariada

A distribuição normal é um modelo mais importante para variáveis contínuas unidimensionais. Para variáveis aleatórias contínuas bidimensionais condicionadas é possível encontrar os estimadores de uma regressão linear simples. Encontrando a curva de regressão e a reta de regressão observa-se que eles coincidem, assim pode-se usar os componentes de

uma distribuição normal bivariada condicionada para encontrar os parâmetros estimados desta regressão linear simples. Exemplificou-se a teoria proposta através de duas variáveis quantitativas x e y com finalidade de estimar os parâmetros de um modelo linear simples usando as componentes de uma normal bivariada condicional.

A distribuição normal é um importante modelo de distribuição contínua: a distribuição normal p dimensional. Anteriormente, desenvolveram-se o caso de uma variável e agora veremos o caso bidimensional. A distribuição normal bidimensional é um caso particular da distribuição p dimensional. Quando se modelam fenômenos aleatórios bivariados, assume-se que as variáveis tenham distribuições normais. Finalmente, seu desenvolvimento resulta ilustrativo, já que podem escrever sem recorrer à notação matricial, evidenciando, de forma explícita, o conteúdo de ditos resultados.

Seja $X_1,...,X_p$, em que p é a dimensão da variáveis aleatórias independentes e igualmente distribuídos com distribuição $N(0,1)$. A variável aleatória de dimensão p.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$$

Diz-se que segue uma distribuição normal p dimensional padronizada. Posto que as variáveis $X_1,...,X_p$ são independentes, a função de densidade da variável aleatória p dimensional X é:

$$\begin{aligned} f_{\mathbf{X}}(X) &= f_{X_1}(X_1)\cdots f_{X_p}(X_p) \\ &= \frac{1}{(2\pi)^{p/2}} e^{\left[-\frac{1}{2}\left(X_1^2,\ldots,X_p^p\right)\right]} \\ &= \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\left(X'X\right)} \end{aligned}$$

Em que $-\infty < f_X(X) < \infty$. Define-se agora uma distribuição p dimensional qualquer:

Definição 5.1.1. *Seja* $\mathbf{X}$ *uma variável aleatória com distribuição normal* p *dimensional padronizada,* $\mathbf{V} = (a_{ij})$ *uma matriz quadrada não nula de ordem* p *com determinante não nulo, e* $\mu = (\mu_1, \cdots, \mu_p)^{'}$ *uma matriz coluna* $p \times 1$. *Seja* Σ *a matriz quadrada de ordem* p, *e* $\Sigma = \mathbf{V}.\mathbf{V}^{'}$.

$$\mathbf{Y} = \mathbf{V}\mathbf{X} + \mu \tag{5.1}$$

Logo, se diz que a variável p dimensional $\mathbf{Y} \sim N(\mu, \Sigma)$. Evidentemente, se X segue uma distribuição normal p dimensional de parâmetros padronizados, $V = I$, a matriz identidade de dimensão $p \times p$ e $\mu = 0$ matriz coluna p dimensional formada por zeros. Portanto, os parâmetros de uma distribuição normal p dimensional padronizado tem média zero e variância unitária, denotando-se de forma abreviada $\mathbf{X} \sim N_p(0, I)$.

Proposição 5.1.1. *Seja* Z *uma variável aleatória com distribuição* $N(0,1)$. *A variável* $Y = \sigma Z + \mu$, *onde* μ *e* σ *são dois números reais qualquer, com* $\sigma > 0$, *segue uma distribuição* $N(\mu, \sigma)$.

Este proposição indica que, a partir de uma variável padronizada, Z, é possível obter uma variável normal $N(\mu, \sigma)$, fazendo a transformação linear, $X = \sigma Z + \mu$.

As variáveis normais p dimensionais se obtém, também, mediante transformações lineares de uma normal p dimensional padronizado. Desenvolvendo a transformação anterior tem-se:

$$\begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_p \end{pmatrix} = \begin{pmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \vdots & \vdots \\ v_{p1} & \cdots & v_{pp} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_p \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

Fazendo o produto de matrizes e somando com a média, têm-se:

$$\begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_p \end{pmatrix} = \begin{pmatrix} v_{11}X_1 + & \cdots + & v_{1p}X_p + \mu_1 \\ \vdots & \vdots & \vdots \\ v_{p1}X_1 + & \cdots + & v_{pp}X_p + \mu_p \end{pmatrix}$$

Definitivamente, para $i = 1, \ldots, p$. Sendo $Y_i = v_{i1}X_1 + \cdots + v_{ip}X_p + \mu_i$, ou seja, cada componente de um normal p dimensional é uma combinação linear de p variáveis com distribuição $N(0,1)$ e independente. O fato de que a matriz $\mathbf{V}$ seja não singular, $|V| \neq 0$, implica que as componentes $Y_1, \cdots, Y_p$ são linearmente independentes, ou seja, que nenhuma delas pode ser combinação linear do resto. Calcula-se a função densidade de $\mathbf{Y}$. Como a matriz $\mathbf{V}$ é não singular, existe inversa, assim:

$$\mathbf{X} = \mathbf{V}^{-1}(\mathbf{Y} - \mu) \tag{5.2}$$

A mudança da variável de ordem R^p em R^p induzido pela transformação Jacobiano da Equação 5.2 é:

$$\mathbf{x} = \mathbf{V}^{-1}(\mathbf{y} - \mu) \tag{5.3}$$

Chamando c_{ij} $(i = 1, \ldots, p; j = 1, \ldots, p)$ aos elementos de $\mathbf{V}^{-1}$ e desenvolvendo a expressão anterior tem-se:

$$y_i = c_{i1}(y_1 - \mu_1) + \cdots + c_{ip}(y_p - \mu_p) \tag{5.4}$$

Portanto, $\dfrac{\delta_{y_i}}{\delta_{x_i}} = c_{ij}$.

O Jacobiano da transformação é o determinante da matriz $\mathbf{V}^{-1}$, logo:

$$Jacob = |\mathbf{V}^{-1}| \tag{5.5}$$

Assim, $f_{\mathbf{Y}}(x) = f_{\mathbf{X}}(y(x)).|Jacob|$

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\left[\mathbf{V}^{-1}(Y-\mu)'\left(\mathbf{V}^{-1}(Y-\mu)\right)\right]} \parallel \mathbf{V}^{-1} \parallel \tag{5.6}$$

Simplificando as expressões que aparecem na função densidade Equação 5.6. Como $\Sigma = \mathbf{V}\mathbf{V}'$, deduz-se que: $\Sigma^{-1} = (\mathbf{V}^{-1})'\mathbf{V}^{-1}$.

Logo, $|\Sigma^{-1}| = |(\mathbf{V}^{-1})'||\mathbf{V}^{-1}| = (|\mathbf{V}^{-}|)^2$.

A função densidade de $\mathbf{Y}$ em função da matriz de covariância, Σ será:

$$f = \frac{1}{\sqrt{|\Sigma|}(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}(\mathbf{X}-\mu)'\Sigma^{-1}(\mathbf{X}-\mu)} \tag{5.7}$$

Em que $-\infty < x_i < +\infty$ para $i = 1, \cdots, p$.

Conhecendo a matriz de covariância das variáveis, no caso bidimensional, resulta habitual escrever a matriz de covariância em função do coeficiente de correlação entre as variáveis. Considere a matriz de covariância:

$$\Sigma_{p\times p} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{5.8}$$

Como $\sigma_{12} = \dfrac{\sigma_{12}}{\sigma_1\sigma_2}$, temos:

$$\Sigma_{p\times p} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma 2 \\ \rho_{12}\sigma_1\sigma 2 & \sigma_2^2 \end{pmatrix} \tag{5.9}$$

Assim, o determinante de Σ será:

$$|\Sigma| = \begin{vmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma 2 \\ \rho_{12}\sigma_1\sigma 2 & \sigma_2^2 \end{vmatrix} = \sigma_1^2\sigma_2^2(1-\rho_{12}^2)$$

O valor do inverso da raiz quadrada do determinante será:

$$\frac{1}{\sqrt{|\Sigma|}} = \frac{1}{\sigma_1\sigma_2\sqrt{(1-\rho_{12}^2)}} \tag{5.10}$$

Calculando a inversa da matriz de covariância obtém-se:

$$\Sigma^{-1} = \frac{1}{\sigma_1\sigma_2\sqrt{(1-\rho_{12}^2)}} \begin{pmatrix} \sigma_1^2 & -\rho_{12}\sigma_1\sigma 2 \\ -\rho_{12}\sigma_1\sigma 2 & \sigma_2^2 \end{pmatrix} \tag{5.11}$$

Juntando 5.10 e 5.11, tem-se:

$$(x_1-\mu_1, x_2-\mu_2)\Sigma^{-1} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} =$$
$$\frac{1}{\sigma_1^2\sigma_2^2(1-\rho_{12}^2)}(x_1-\mu_1, x_2-\mu_2) \begin{pmatrix} \sigma_1^2 & -\rho_{12}\sigma_1\sigma 2 \\ -\rho_{12}\sigma_1\sigma 2 & \sigma_2^2 \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} =$$
$$\frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_1-\mu_1)^2}{\sigma_1^2}\right]$$

Observa-se que a densidade depende de cinco parâmetros: as médias μ_1 e μ_2 que podem assumir quaisquer valores reais, as variâncias σ_1^2 e σ_2^2, que devem ser positivas e o coeficiente de correlação ρ entre as duas variáveis que deve satisfazer $-1 < \rho < 1$. Assim, encontra-se a função densidade de normal bidimensional:

$$f_{(x_1,x_2)}(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2(\sqrt{1-\rho_{12}^2})}$$
$$\exp\left[\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right]$$

em que $-\infty < x_i < \infty, \forall i = 1,2$.

A Figura 5.3 representa a gráfica da distribuição normal bivariada com médias e covariâncias especificadas na própria figura. No *R* teremos:

```
install.packages("scatterplot3d")
library(scatterplot3d)
install.packages("mvtnorm")
library("mvtnorm")
x1 <- x2 <- seq(-10, 10, length = 51)
dens<- matrix(dmvnorm(expand.grid(x1, x2),
sigma=rbind(c(3, 2), c(2, 3))), #Matriz de covariância
               ncol = length(x1))
s3d <- scatterplot3d(x1, x2,
seq(min(dens), max(dens), length = length(x1)),
type = "n", grid = FALSE, angle = 70,
zlab = expression(f(x[1], x[2])),
xlab = expression(x[1]), ylab = expression(x[2]),
main = "")
text(s3d$xyz.convert(-1, 10, 0.07),
     labels=expression(f(x)==frac(1, sqrt((2*pi)^p*
phantom(".")*det(Sigma[X])))*phantom(".")*exp*{
bgroup("(", - scriptstyle(frac(1, 2)*phantom("."))*
(x - mu)^T * Sigma[X]^-1 * (x - mu),")")}))
text(s3d$xyz.convert(1.5, 10, 0.05),
     labels=expression("with" * phantom("m")*
```

```
      mu==bgroup("(",atop(0, 0),")")*phantom(".")*","*
      phantom(0)*
      {Sigma[X] == bgroup("(", atop(5*phantom(0)*2,
      2*phantom(0)*5), ")")}))
for(i in length(x1):1)
  s3d$points3d(rep(x1[i], length(x2)), x2,
  dens[i,], type = "l")
for(i in length(x2):1)
  s3d$points3d(x1, rep(x2[i], length(x1)),
  dens[,i], type = "l")
```

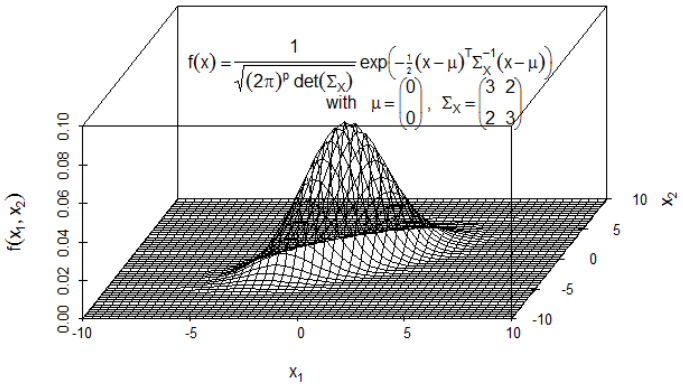


Figura 5.2: Representação gráfica da função densidade bidimensional.

Considerando o caso de uma maior dispersão com variâncias igual a 5, covariâncias 2 e vetor de média $\mu = (0,0)$:

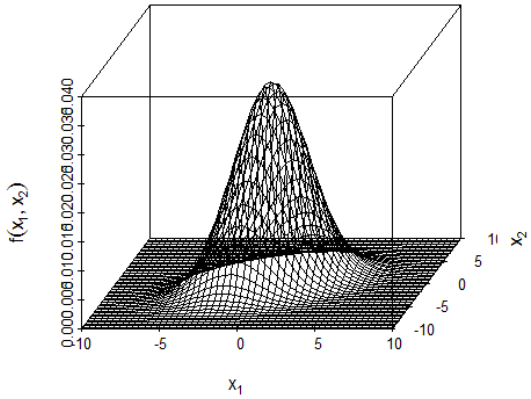


Figura 5.3: Representação gráfica da função densidade bidimensional.

5.2 Normalidade multivariada

Para o caso multivariado, o gráfico $Q-Q$ tem por base o conceito de distância generalizada (ou distância de Mahalanobis) entre dois vetores de observações, neste caso particular, a distância generalizada entre cada vetor de observação. Se observa que quando a normalidade p-variada se ajusta bem aos dados amostrais, este gráfico se parece uma nuvem de pontos próximo de uma reta diagonal. Se tiverem curvas acima ou abaixo da reta diagonal padrão revela ausência de normalidade.

$$d_u^2 = (X_u - \bar{X})' S^{-1} (X_u - \bar{X}) \tag{5.12}$$

Quando a população de onde são retiradas as n observações ($n \geq 25$) é normal multivariada, as distâncias anteriores têm um comportamento semelhante ao de uma variável com distribuição de χ^2 com p graus de liberdade. Considere a Tabela 5.1 para ilustrar o gráfico $Q-Qplot$.

Tabela 5.1: Vinte indivíduos com três características hipotéticas.

Indiv	X_1	X_2	X_3	Indiv	X_1	X_2	X_3
1	20.0	19.5	19.8	11	23.0	20.1	22.1
2	21.1	14.0	18.0	12	22.0	22.0	22.4
3	24.0	24.5	25.0	13	16.0	15.8	14.8
4	20.0	18.0	19.0	14	16.1	16.1	15.1
5	22.0	25.0	21.0	15	21.7	25.3	24.3
6	16.0	15.0	17.0	16	21.9	21.9	19.9
7	20.0	24.0	23.0	17	24.1	24.1	24.7
8	20.0	24.6	25.7	18	23.0	23.0	25.0
9	18.5	21.6	19.5	19	20.1	18.3	16.3
10	17.0	18.0	18.6	20	19.5	24.0	23.0

A construção do gráfico $Q-Q$ segue os mesmos passos do caso univariado. Os passos para encontrar os valores das últimas quatro colunas da Tabela 5.2 são os seguintes:

1. Ordena os indivíduos i em ordem crescente.
2. Encontra o vetor de média e a matriz de covariância, ou seja:

$$\bar{X} = \begin{bmatrix} 20.30 & 20.74 & 20.71 \end{bmatrix} \quad \text{e} \quad S = \begin{bmatrix} 6.4968 & 5.9542 & 6.4747 \\ 5.9542 & 13.501 & 10.9174 \\ 6.4747 & 10.9174 & 11.7451 \end{bmatrix}$$

3. Encontra a distância de Mahalanobis.

$$d_i^2 = (X_i - \bar{X})' S_{pxp}^{-1} (X_i - \bar{X}) \quad \text{com } i = 1, \ldots, n.$$

4. Ordena a distância de Mahalanobis: $d_{(1)}^2 \leq d_{(2)}^2 \leq \ldots \leq d_{(n)}^2$, em que d_i^2 representa a i-ésima estatística de ordem.
5. Ordena $\left(\frac{(i-0.5)}{20}\right)$.
6. Ordena a χ^2, $\chi^2_{\left(d_i^2, \left(\frac{(i-0.5)}{20}\right)\right)}$.

7. Encontra o gráfico (d_i^2, χ^2).

Fazendo o produto das matrizes, tem-se:

$$d_1^2 = \begin{bmatrix} 20.0-20.30 & 19.5-20.74 & 19.8-20.71 \end{bmatrix} \begin{bmatrix} 6.4968 & 5.9542 & 6.4747 \\ 5.9542 & 13.501 & 10.9174 \\ 6.4747 & 10.9174 & 11.7451 \end{bmatrix} \begin{bmatrix} 20.0-20.30 \\ 19.5-20.74 \\ 19.8-20.71 \end{bmatrix}$$

Para o primeiro indivíduo teremos:

$$d_1^2=(X_1-\bar{X})' S(X_1-\bar{X})$$

Assim para o indivíduo 1, tem-se que $d_1^2 = 0.1297$ como se observa na Tabela 5.2.

Tabela 5.2: Vinte indivíduos e suas três características hipotéticas

i	X_1	X_2	X_3	d_i^2	d_i^2 ordenadas	$\left(\frac{(i-0.5)}{20}\right)$	$10(\chi^2)$ ordenadas
1	20.0	19.5	19.8	0.1297	0.1297	0.025	0.000
2	21.1	14.0	18.0	7.6120	0.4706	0.075	0.004
3	24.0	24.5	25.0	2.1875	0.7753	0.125	0.019
4	20.0	18.0	19.0	0.7753	1.6394	0.175	0.038
5	22.0	25.0	21.0	5.6487	1.6488	0.225	0.447
6	16.0	15.0	17.0	4.5434	1.7090	0.275	0.647
7	20.0	24.0	23.0	1.6394	2.1420	0.325	0.717
8	20.0	24.6	25.7	5.5125	2.1875	0.375	1.014
9	18.5	21.6	19.5	1.7090	2.2513	0.425	1.427
10	17.0	18.0	18.6	2.1420	2.2597	0.475	1.842
11	23.0	20.1	22.1	2.5070	2.4882	0.525	1.970
12	22.0	22.0	22.4	0.4706	2.5070	0.575	2.240
13	16.0	15.8	14.8	3.4287	2.6300	0.625	2.503
14	16.1	16.1	15.1	3.1888	3.1888	0.675	2.554
15	21.7	25.3	24.3	1.6488	3.4287	0.725	2.801
16	21.9	21.9	19.9	2.6300	4.2266	0.775	3.277
17	24.1	24.1	24.7	2.2597	4.5434	0.825	3.943
18	23.0	23.0	25.0	2.4882	5.5125	0.875	4.275
19	20.1	18.3	16.3	4.2266	5.6487	0.925	7.781
20	19.5	24.0	23.0	2.2513	7.6120	0.975	7.784

Quando a normalidade p-variada é coerente com os dados amostrais, este gráfico deve resultar em algo próximo a uma reta. Curvas diferentes da reta indicam falta de normalidade. Este gráfico também é útil para a identificação de valores discrepantes.

O gráfico de pares apresenta a ordenada do percentil de ordem $100\left(\frac{(i-\frac{1}{2})}{n}\right)$ da distribuição χ_p^2, ou seja,

$$P\left[\chi_p^2 \leq \chi_p^2\left(\frac{i-\frac{1}{2}}{n}\right)\right] = \frac{\left(\frac{i-\frac{1}{2}}{n}\right)}{n}$$

Com os dados da Tabela 5.2 constrói-se a Figura 5.4.

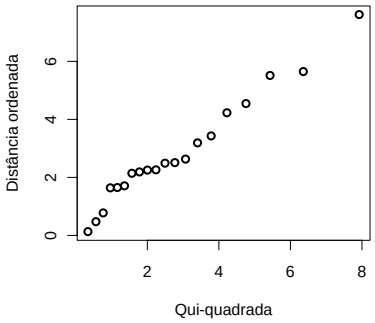


Figura 5.4: Gráfico χ^2 $(Q-Qplot)$ para as três variáveis.

Os pontos não estão totalmente próximos de uma reta considerando as três variáveis, o que sugere alguma evidência contra a normalidade dessas três variáveis.

```
chisplot<- function(x){
if(!is.matrix(x))) stop("x não é uma matriz")
n<-nrow(x); p<_ncol(x) # Num linhas e colunas
xbar<-apply(x,2,mean)
# Médias das variáveis (2 = por colunas)
S<-var(x) # Matriz de covariância
S<-solve(S) # Inversa na matriz de covariância
Indices<-(1:n)/(n+1)
H<-t(t(x)-xbar)
# t transposta de x e média de x (xbar)
Disti<-apply(H,1,function(x,S) x%*%S%*% x, S)
# Multiplicação de matrizes usa-se: %*%
A<-qchsiq(Indices, p)
qqplot(A, sort(Disti), ylab="Ordem das distâncias",
xlab="Qui-quadrado", lwd=2)
abline(0.1,col="black") # Inserindo a reta diagonal
title("Q-Q plot")
locator()}  # Indica onde será o título
```

O esquema dos cinco números, como se observa na Figura 5.5 de uma distribuição ajuda a identificar normalidade nas variáveis.

```
boxplot(x, ylab="Esquema dos cinco números")
```

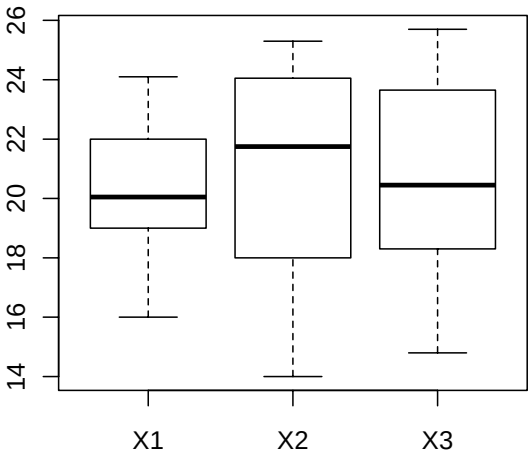


Figura 5.5: Função de distribuição de frequência hipotética de 2 grupos.

5.3 Teste de assimetria

O teste de hipótese de assimetria contrasta a normalidade da distribuição da que se tem extraído os dados mediante a consideração do coeficiente de assimetria amostral. Pretende-se realizar o seguinte teste de hipótese:

$$\begin{cases} H_o: & X \text{ é normal} \\ H_a: & X \text{ não é normal} \end{cases}$$

A estatística de teste do coeficiente de assimetria são definidos por:

$$ACA = \frac{\frac{1}{n}\sum_{n=1}^{n}(X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{n=1}^{n}(X_i - \bar{X})^2\right)^{3/2}}$$

Se a hipótese nula (H_o) for aceita, a estatística α tem uma distribuição assintótica normal de média zero e variância $\sigma^2 = \frac{6}{n}$.

$$ACA \sim N\left(0, \frac{6}{n}\right)$$

Assim, o contraste pode ser expresso da seguinte forma:

$$\begin{cases} H_o: & X \text{ segue uma simetria normal (assimetria=0)} \\ H_a: & X \text{ não tem assimetria normal} \end{cases}$$

Se a hipótese nula (H_o) é correta, o coeficiente de assimetria amostral estima um parâmetro da população que é zero, ou seja, o coeficiente de assimetria de uma distribuição normal é zero.

Rejeita H_o a um nível α para grandes valores da estatística *ACA*. Pode-se também resolver o contraste mediante o coeficiente de assimetria amostral padronizado definido por:

$$ACA_{padronizado} = \frac{ACA}{\sqrt{\frac{6}{n}}}$$

Neste caso, rejeita H_o a um nível α para grandes valores da estatística $ACA_{padronizado}$ não está no intervalo [-2,2].

Exemplo 5.3.1. *Usaremos o mesmo exemplo do teste de normalidade de Shapiro-Wilks para exemplificar os passos do teste de assimetria. Considere as seguintes amostras:*

$$X = [19.8, 20.5, 19.7, 17.6, 19.2, 18.4, 18.1, 19.1, 17.9, 17.3, 20]$$

```
X<-c(19.8, 20.5, 19.7, 17.6, 19.2, 18.4,
     18.1, 19.1, 17.9, 17.3, 20)
ACA<- function(x){
 n<-length(X)
 v<-var(X)
 m<-mean(X)
 terceiro.momento<-(1/(n))*sum((X-m)^3)
 terceiro.momento/(var(X)^(3/2))
 }
 ACA(X)
# [1] -0.0212

A_padronizado<-ACA(X)/sqrt(6/n)
A_padronizado
```

```
# [1] -0.0867
```

Portanto, o coeficiente de assimetria amostral padronizado foi $-0.0867 \in [-2,2]$, logo aceita a hipótese de normalidade. Se o valor estivesse fora deste intervalo rejeitaria a hipótese nula.

5.4 Teste de curtose

O teste de hipótese de curtose contrasta a normalidade da distribuição da que se tem extraído os dados mediante a consideração do coeficiente de curtose amostral. Pretende-se realizar o seguinte teste de hipótese.

$$\begin{cases} H_o: & X \text{ é normal} \\ H_a: & X \text{ não é normal} \end{cases}$$

A estatística de teste do coeficiente de curtose é definido por:

$$TC = \frac{\frac{1}{n}\sum_{n=1}^{n}(X_i - \bar{X})^4}{\left(\frac{1}{n}\sum_{n=1}^{n}(X_i - \bar{X})^2\right)^3} - 3$$

Se a hipótese nula (H_0) normal for aceita, a estatística α tem uma distribuição assintótica normal de média zero e variância $\sigma^2 = \frac{6}{n}$.

$$TC \sim N\left(0, \frac{24}{n}\right)$$

Assim, o contraste pode ser expresso da seguinte forma:

$$\begin{cases} H_o: & X \text{ curtose normal (curtose=0)} \\ H_a: & X \text{ não tem curtose normal} \end{cases}$$

Se a hipótese nula (H_o) tem curtose normal, rejeita H_0 a nível α para grandes valores de TC. O contraste também pode ser realizado mediante o coeficiente de curtose amostral padronizado definido por: $TC_{padronizado} = \frac{TC}{\frac{24}{n}}$. Neste caso, rejeita H_o a um nível α para grandes valores da estatística $TC_{padronizado}$, ou seja, se $TC_{padronizado}$ não estiver no intervalo [-2,2].

```
X<-c(19.8, 20.5, 19.7, 17.6, 19.2, 18.4,
      18.1, 19.1, 17.9, 17.3, 20)
TC<- function(x){
```

```
 n<-length(X)
 v<-var(X)
 m<-mean(X)
 quarto.momento<-(1/(n))*sum((X-m)^4)
 quarto.momento/(var(X)^(2)) -3
 }
 TC(X)
# [1] -1.63
A_padronizado<-TC(X)/sqrt(24/n)
A_padronizado
# [1] -1.287254
```

Portanto, o coeficiente de curtose amostral padronizado foi -1,28 $\in [-2,2]$, logo aceita a hipótese de normalidade. Se o valor estivesse fora deste intervalo rejeitaria a hipótese nula.

5.5 Igualdade de matriz de covariância

Terá que realizar o contraste de hipótese de homocedasticidade, a medida de bondade de ajuste (para medir a variabilidade total explicada pelo fator) e o contraste de esfericidade para observar se existe ou não uma relação significativa entre as variáveis analisadas.

Enquanto a hipótese de homocedasticidade, a validez da estatística F a partir de Λ está condicionada ao cumprimento, entre outras condições, de que a matriz de covariância seja a mesma para todas as populações ou grupos. Portanto, a hipótese nula e alternativa para contrastar esta hipóteses são as seguintes:

$$\begin{cases} H_o: & \Sigma_1 = \Sigma_2 = \ldots = \Sigma_G \\ H_a: & \text{Nem todas } \Sigma_G \text{ são iguais} \end{cases}$$

Para determinar se a matriz de covariância é a mesma para os distintos grupos pode-se utilizar o contraste de Bartlett-Box [1], que utiliza a estatística M. Esta estatística se define da seguinte forma:

$$M = \frac{\prod_{g=1}^{G} |S_g|^{(n_g-1)/2}}{|S|^{\frac{(n-G)}{2}}} \tag{5.13}$$

Aplicando o logaritmo em ambos os lados da Equação (5.13) tem-se:

[1] Este contraste foi proposto por Bartllet em 1947. Posteriormente, Box em 1949 desenvolveu um procedimento para aproximá-lo a uma estatística F ou a distribuição χ^2.

$$M = (n-G)ln|S| - \sum_{g=1}^{G} ln|S_g|$$

Em que: n é a dimensão total da amostra, n_g é o número de grupos, $v_g = (n_g - 1)$ são os graus de liberdade associados a cada grupo, S_g a matriz de covariância do grupo g e S a matriz de covariância total.

$$S_G = \frac{W_G}{v_G} \quad \text{e} \quad S_G = \frac{\sum_{g=1}^{G} W_g}{n-G} = \frac{\sum_{g=1}^{G} (v_g)S_g}{n-G}$$

Infelizmente, a estatística M não tem uma distribuição exata. Segundo REIS (2001), Box sugeriu duas aproximações para o teste.

1. Aproximação à distribuição do χ^2.

$$M \cdot H \sim \chi^2_{[\frac{1}{2}p(p+1)(G-1)]},$$

sendo,

$$H = 1 - \frac{2p^2+3p-1}{6(p+1)(G-1)} \left(\sum_{g=1}^{G} \frac{1}{v_g} - \frac{1}{n-G} \right)$$

1. Aproximação à distribuição F.

O teste de Box pode ser realizado também utilizando uma aproximação à distribuição F de Snedecor, que quando o tamanho dos grupos é pequeno ou grande número de variáveis grande resulta mais adequada que a aproximação a χ^2. Para esta estatística considere:

$$b_1 = 1 - H \quad \text{e} \quad b_2 = \frac{(p+1)(p-2)}{6(G-1)} \left(\sum_{j=1}^{k} \frac{1}{v_g^2} - \frac{1}{(n-G)^2} \right)$$

$$v_1 = \frac{p(p+1)(G-1)}{2} \quad \text{e} \quad v_2 = \frac{v_1+2}{b_2 - b_1^2}$$

Então:

$$\frac{M\left(1-b_1-\frac{v_1}{v_2}\right)}{v_1} \sim F_{(v_1,v_2)}$$

Para determinar se a matriz de covariância é a mesma para os distintos grupos pode-se utilizar o contraste de Bartlett-Box.

```
Teste_M_BOX=function(Variaveis,grupo,alfa=0.05){
Variaveis=as.matrix(Variaveis)
p=ncol(Variaveis) # Colunas (variáveis)
n=nrow(Variaveis) # Total de linhas do congunto
G=max(grupo) # Número de grupos, G=3
Ng=rep(0,G) #tamanho de cada grupo armazenado em Ng
pame=rep(0,G)
# Armazena o determinante da matriz de cov. de cada grupo
Ng=as.vector(table(grupo))
Sg= Sg1=array(dim=c(p,p,G))
for (g in 1:G) {
Sg[,,g]=cov(Variaveis[grupo==g,])
pame[g]=det(Sg[,,g]) #det. de cada matriz de covariância
Sg1[,,g]=(Ng[g]-1)*cov(Variaveis[grupo==g,]) }
# cálculo da matriz de covariância pooled
    # (congunta dos 3 grupos)
S=apply( Sg1,1:2,sum)
S=S/(n-G)
for (g in 1:G)
pamela=det(S) #det da matriz de covariância congunta
Ts1=sum((Ng-1)*log(pamela/pame))
Ge1=(2*(p^2)+3*p-1)/(6*(p+1)*(G-1))
Ge2=(sum(1/(Ng-1))-1/(n-G))
H=1-Ge1*Ge2
Ts=H*Ts1 # A estatística de teste M
gl=(1/2)*p*(p+1)*(G-1) #Graus de liberdade da qui-quadrada
pvalue=1-pchisq(Ts,gl) #p-value da estatística de teste
crit=qchisq(1-alfa,gl) #Valor crítico da dist. qui-quadrada
list(Teste_M=Ts,Grau_de_liberdade = gl,
Valor_critico=crit,p.value=pvalue)}
```

O resultado aplicado a matriz de dados:

```
Teste_M_BOX(datos[,2:3],dados[,1], alfa=0.05)
$Teste_M
```

```
# [1] 4.407017
$Grau_de_liberdade
# [1] 6
$Valor_critico
# [1] 12.59159
$p.value
# [1] 0.6217732
```

Assim, aceita-se a hipótese de igualdade das matrizes de covariâncias de cada um dos três grupos, pois $p-valor > 0,05$.

5.5.1 Intervalo de confiança para o coeficiente de correlação e teste de hipóteses usando teste de transformação de Fisher

O teste de transformação de Fisher para o coeficiente de correlação é definido por:

$$\hat{z} = \frac{1}{2}\log\frac{1+r}{1-r} \tag{5.14}$$

Com inversa igual a:

$$\frac{\exp(2\hat{z})-1}{\exp(2\hat{z})+1} \tag{5.15}$$

O erro padrão estimado da Equação 5.14 é:

$$\frac{1-r^2}{\sqrt{n-3}} \tag{5.16}$$

Calcula-se o intervalo de confiança baseado na diferença e faz o teste de hipótese para o valor zero apenas. Abaixo calcula-se o intervalo de confiança baseado na equação 5.14, assumindo uma normal assintótica. A função `FTcor` informa verdadeiro se existe correlação (nenhuma correlação ou não apenas correlação zero).

Vamos testar a correlação entre uma normal com média zero e variância 1 e uma t de Student para parâmetro $\lambda = 1$, ambas com tamanho $n = 100$.

```
x<-rnorm(100,0,1)
y<-rt(100,1)
FTcor=function(y,x,a=5/100,rho=0) {
## rho o valor da correlação para hipóteses
```

```
y=as.vector(y); x=as.vector(x)
# x e y duas variáveis métricas
nx=length(x) # Tamanho do vetor x
r=cor(y,x) ## the correlation value
Zho=0.5*log((1+rho)/(1-rho))
# Transf de Fisher para Hipótese nula, Ho
Zh1=0.5*log((1+r)/(1-r))
# Transf de Fisher para Hipótese alternativa, H1
Ep=(1-r^2)/sqrt(nx-3)
# Erro padrão para transf de Fisher de Ho
test=(Zh1-Zho)/Ep  ## Teste estatístico
pvalue=2*(1-pnorm(abs(test))) # p-value
ZL=Zh1-qnorm(1-a/2)*Ep ; ZH=Zh1+qnorm(1-a/2)*Ep
    #inferior do intervalo
fisherL=(exp(2*ZL)-1)/(exp(2*ZL)+1)
fisherH=(exp(2*ZH)-1)/(exp(2*ZH)+1)
    #superior do intervalo
Int_de_Conf=c(fisherL,fisherH)
names(Int_de_Conf)=c("Inferior","Superior")
list(correlação=r,p.value=pvalue,Int_de_Conf
    =Int_de_Conf)}
FTcor(x,y)
FTcor(x,y)
correlação
# [1] -0.1106054
p.value
# [1] 0.2681388
# Int_de_Conf
# Inferior    Superior
# -0.29827897  0.08530208
```

Temos então o valor da correlação de Pearson entre x e y, o intervalo de confiança para correlação (95 %) e o p-valor informando que aceita a hipótese nula de igualdade de correlação dos grupos, pois $0,05 < 0,2681$.

5.6 Teste de hipóteses para igualdade de duas correlações

Seja z_1 e z_2 a transformação de Fisher da equação 5.14, n_1 e n_2 os tamanhos respectivos das duas variáveis correlacionadas. A estatística de teste é definida por:

$$Z = \frac{\hat{z}_1 - \hat{z}_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \tag{5.17}$$

Implementando a Equação 5.17 no *R*, tem-se:

```
x1<-rnorm(100,0,1); y1<-rt(100,1)
x2<-rnorm(100,0,1.1); y2<-rt(100,2)
rg1<-cor(x1,y1); rg2<-cor(x2,y2)
ng1=length(x1)
ng2=length(x2) # Tamanho do vetor do grupo 2
rg1g2=function(rg1,rg2,ng1,ng2){
#rg1 e rg2 - coeficientes de correlação de cada grupo
Zg1=0.5*log((1+rg1)/(1-rg1))  #Transf de Fisher p/ g1
Zg2=0.5*log((1+rg2)/(1-rg2))  #Transf de Fisher p/ g2
Z=(Zg1-Zg2)/sqrt(1/(ng1-3)+1/(ng2-3)) #Teste Z
pvalue=2*(1-pnorm(abs(Z)))  #p-valor
list(Z=Z,p.value=pvalue) }
rg1g2(rg1,rg2,ng1,ng2)
Z
# [1] -0.2874333
p.value
# [1] 0.7737806
```

O resultado do *p*-valor é calculado de uma distribuição normal padronizada, como se observa na função `pnorm()`.

5.7 Pontos extremos (*outliers*)

Os pontos extremos podem afetar as funções discriminantes como as de classificação. É importante verificar se existem e eliminá-los antes de analisar os dados. Usando o pacote *extremevalues* *extremevalues*2010 e fazendo uma simples mudança em: $y < -c(min(y), y, max(y))$ para $y < -c(y)$

Pode-se detectar os *outliers* facilmente. Na Figura 5.6 claramente se observa dois pontos nos novos casos de TB em 2012 e um ponto em 2013.

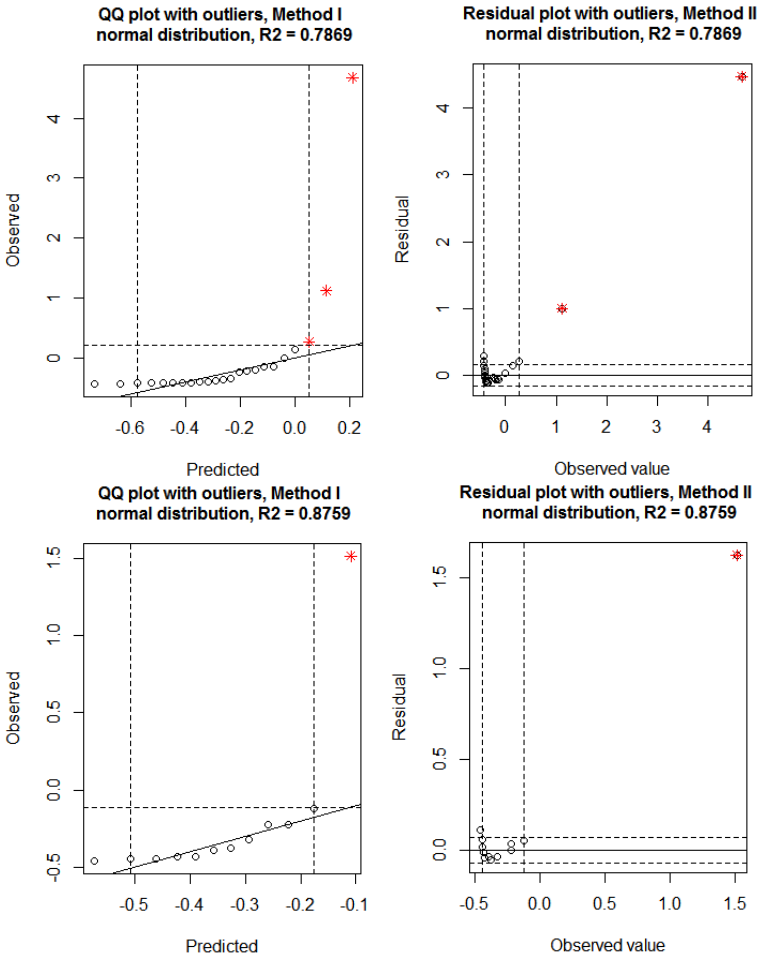


Figura 5.6: Detecção de outliers nas funções discriminantes 1 e 2, respectivamente

```
# Detectar outliers
library(extremevalues)
# Para primeira coluna discriminante
D1<-Ds1; D1
#[1] -0.43889710 -0.15484107 -0.42990798 -0.39035588
```

```
    #    -0.39754717
#[6] -0.01101523 -0.42451451  4.67950485 -0.34720812
    #    -0.22855181
#[12]-0.41912104 -0.41732322 -0.20158447 -0.24113657
    #     1.11801758
#[17]-0.42271669  0.26944515  0.14000190 -0.42091887
    #    -0.39215370
#[23]-0.42631233 -0.14764978 -0.37057982
y1 <- c(D1); y1
#[1] -0.43889710 -0.15484107 -0.42990798 -0.39035588
    #    -0.39754717
#[6] -0.01101523 -0.42451451  4.67950485 -0.34720812
    #    -0.22855181
#[12]-0.41912104 -0.41732322 -0.20158447 -0.24113657
    #     1.11801758
#[17]-0.42271669  0.26944515  0.14000190 -0.42091887
#[22]-0.39215370 -0.42631233 -0.14764978 -0.37057982
K <- getOutliers(y1,method="I",distribution="normal")
L <- getOutliers(y1,method="II",distribution="normal")
par(mfrow=c(1,2))
outlierPlot(y1,K,mode="qq")
outlierPlot(y1,L,mode="residual")
# Para segunda coluna discriminante
D2<-Ds2; D2
#[1] -0.3779492 -0.4605075 -0.2225452 -0.1229896 -0.2249734
#[6] -0.4289411 -0.3245291 -0.4435102 -0.4459384  1.5136082
#[11]-0.4337975 -0.3900901
y2 <- c(D2); y2
#[1] -0.3779492 -0.4605075 -0.2225452 -0.1229896 -0.2249734
#[6] -0.4289411 -0.3245291 -0.4435102 -0.4459384  1.5136082
#[11]-0.4337975 -0.3900901
K <-getOutliers(y2,method="I",distribution="normal")
L <-getOutliers(y2,method="II",distribution="normal")
par(mfrow=c(1,2))
outlierPlot(y2,K,mode="qq")
outlierPlot(y2,L,mode="residual")
```

Os efeitos de um só ponto atípico pode ser grave, pois distorcem as médias e o desvio padrão das variáveis e destroem as relações existentes entre elas. Para ilustrar o problema do valor atípico, suponha que em uma amostra multivariada de tamanho n introduz um valor atípico, **a**, que é um vetor de falsas observações. Denomina-se $\bar{x}$ e S o vetor de médias e matriz de covariância sem o dado atípico e $\bar{x}_k$ e S_k aos da amostra contaminada com este

dado atípico. Assim, a média e a matriz de covariância serão:

$$\bar{x}_c \quad = \quad \bar{x} + \frac{(a-\bar{x})}{n+1} \tag{5.18}$$

$$S_c = \frac{n}{n+1}S + \frac{(a-\bar{x})(a-\bar{x})'}{n+1}\left(\frac{n}{(n+1)}\right) \tag{5.19}$$

As fórmulas 5.18 e 5.19 indicam que apenas um valor atípico pode afetar muito o valor da média e todas as variâncias e covariâncias entre as variáveis. Criamos um exemplo para ilustrar tal situação:

```
n<-4 #indivíduos com 3 variáveis
M<-matrix(c(1,2,4,3,6,5,4,3,4,2,3,5), ncol=3,
     byrow=T);M
# [,1] [,2] [,3]
# [1,]    1    2    4
# [2,]    3    6    5
# [3,]    4    3    4
# [4,]    2    3    5
Media_sem_atipico<-apply(M,2,mean); Media_sem_atípico
# [1] 2.5 3.5 4.5
Msa<- Media_sem_atipico;Msa
# [1] 2.5 3.5 4.5
a<-matrix(c(300,245,700), ncol=3, byrow=T);a
# [,1] [,2] [,3]
# [1,]  300  245  700
Ma<-rbind(M,a); Ma # com atípicos
# [,1] [,2] [,3]
# [1,]    1    2    4
# [2,]    3    6    5
# [3,]    4    3    4
# [4,]    2    3    5
# [5,]  300  245  700
Media_com_atipico<-apply(Ma,2,mean); Media_com_atípico
# [1]  62.0  51.8 143.6
Xc<-Media_contaminada<-Msa+((Ma-Msa)/(n+1));
Xc
# [,1] [,2]  [,3]
# [1,]  2.2  4.0   3.6
# [2,]  3.4  3.2   4.6
```

```
# [3,]  4.4  3.4   2.8
# [4,]  2.4  4.2   3.8
# [5,] 62.8 51.0 143.6
```

Observa-se que o valor atípico influenciam os dados originais. Semelhantemente a matriz de covariância, pois a média influencia também a variância.

5.8 Linearidade

O suposto de linearidade implica que existem relações lineares entre as variáveis dentro de cada grupo, ou seja implica que os preditos estão relacionados linearmente entre si em cada grupo.

Seu descumprimento reduz a potência de explicação no eixo discriminante. A existência de linearidade pode ser representada em uma linha reta. Um distanciamento da relação linear entre duas variáveis pode ser examinado a partir de diagramas de constituídos para representar os valores de ambos.

O contraste de linearidade indica a hipótese de que as médias das distribuições condicionadas próximo à diagonal de uma reta é compatível com a amostra observada.

5.8.1 Contraste de linearidade

Conhecendo as variáveis aleatórias métricas é possível construir um contraste de hipótese de linearidade. Seja $X_i (i = 1, \cdots, p)$ os valores *distintos* da variável aleatória X. Para cada valor de X_i existirá observações v_{ij}, sendo que $j = 1, \ldots, n_i; \quad i = 1, \ldots, p$. Assim, pode-se verificar que a média de v será:

$$\bar{v} = \frac{\sum_i \sum_j v_{ij}}{n} = \sum_i \frac{n_i \bar{v}_i}{n} \tag{5.20}$$

em que $\bar{v}_i = \sum_i \frac{v_{ij}}{n_i}$.

Na Figura 5.7 se observa uma linearidade, visto que as médias $\bar{v}_i$ mostram uma relação linear.

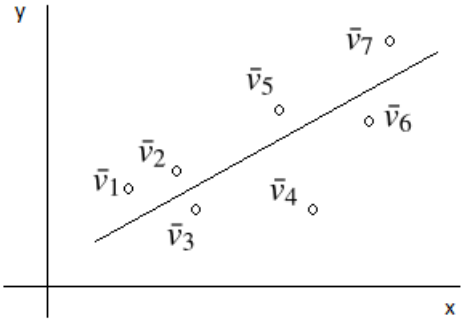


Figura 5.7: Média condicionada e reta de regressão dos pares x e y.

A Figura 5.8 mostra um caso de heterocedasticidade (existência de variâncias diferentes) ocorre frequentemente em dados de corte transversal. Claramente se observa que a variação aumenta na medida que x e y aumentam.

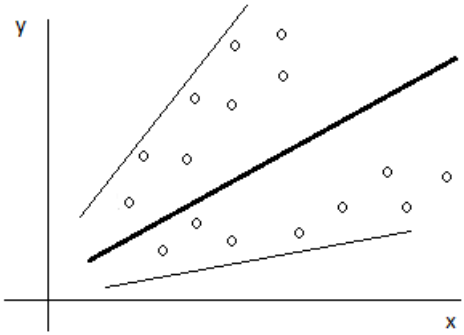


Figura 5.8: Exemplo de heterocedasticidade.

5.8.2 O contraste

Acrescentando e subtraindo $\bar{v}_i$ em ambos os lados da Equação 5.21 os resíduos da regressão serão:

$$e_{ij} = v_{ij} - \hat{v}_i = v_{ij} - \bar{v}_i + \bar{v}_i - \hat{v}_i \tag{5.21}$$

Elevando ao quadrado e somando para todos os pontos obtém-se:

$$VA = \sum n_i(\hat{v}_i - \bar{v}_i)^2 + \sum\sum(v_{ij} - \bar{v}_i)^2$$

O primeiro engloba as diferenças entre as médias observadas e as previstas na reta de regressão que se denomina *termo de falta de ajuste*; o segundo, as diferenças entre os valores observados e suas médias condicionadas, sendo esta uma medida do erro experimental ou pertubação que não depende da reta.

A hipótese de linearidade estabelece que, no posto de valores observados para as variáveis, a média da variável resposta cresce linearmente com a variável independente (explicativa).

Passo a passo mostra-se o teste de hipótese para linearidade (PEÑA, 2002) e posteriormente aplica-se no *R*.

1. O vetor de n componentes $(\hat{v}_i - \bar{v}_i)$ só tem p valores distintos, sendo $d = 7$ (no caso) o número de valores distintos de X_i, que além disso estão ligados pelas duas equações de restrições.

$$\sum n_i(\hat{v}_i - \bar{v}_i) = \sum n_i(\hat{v}_i - \bar{v}_i)X_i = 0 \tag{5.22}$$

Portanto, sua dimensão é $d - 2$.

$$\frac{\sum n_i(\bar{v}_i - \hat{v}_i)^2}{\sigma^2} \sim \chi^2_{p-2} \tag{5.23}$$

1. A dimensão do vetor com componentes $v_{ij} - \bar{v}_i$ é $n - d$, já que tem n componentes distintos mas ligados pelas d equações de restrição $\sum(v_{ij} - \bar{v}_i) = 0$, ou seja que não é certa a hipótese de linearidade.

$$\frac{\sum\sum(v_{ij} - \bar{v}_i)^2}{\sigma^2} \sim \chi^2_{n-d} \tag{5.24}$$

Ambas distribuições correspondem ao quadrado da longitude de vetores ortogonais - são independentes. Chamaremos:

$$\hat{s}^2_{12} = \frac{\sum n_i(\bar{v}_i - \hat{v}_i)^2}{d - 2} \tag{5.25}$$

A variância entre as médias e a reta mede a falta de ajuste.

$$\hat{s}^2_2 = \frac{\sum\sum(y_{ij} - \bar{y}_i)^2}{n - p} \tag{5.26}$$

É a estimação da variância do erro (ou pertubação) sem a hipótese de linearidade. Então, o quociente será:

$$F_{(d-2,n-d)} = \frac{\hat{s}^2_{12}}{\hat{s}^2_2} \tag{5.27}$$

Segue uma distribuição F com $d-2$ e $n-d$ graus de liberdade, apenas quando a hipótese de linearidade seja certa, toma-se valores maiores se esta é falsa. Encontra-se o valor $F_{d-2,n-d}$ graus de liberdade, o valor crítico α_{crtico} tal que $P(F > F^*) = \alpha_{crtico}$ tal que $P(F > F^*) = \alpha_{crtico}$. Se este α_{crtico} é suficientemente pequeno (menor que 0,05 ou 0,01), rejeita a hipóteses de linearidade. Em outro caso conclui-se que não existe evidência significativa contra esta hipótese. A hipótese a testar, H_o é que $E[y_{ij}/x_i]$, ou seja, que todas as médias das distribuições condicionadas encontram-se em uma linha reta. O procedimento é o seguinte:

1. Calcula-se as p médias das distribuições condicionadas.
2. Encontra-se os coeficiente da regressão linear $\hat{\beta}_1$.
3. Após a reta de regressão encontram-se resíduos $e_{ij} = y_{ij} - \hat{y}_i$.
4. Calcula-se as estimativas 5.25 e 5.26. Posteriormente, o quociente: $F_{(p-2,n-p)} = \frac{\hat{s}^2_{12}}{\hat{s}^2_2}$.
5. Calcula-se no R o valor da distribuição $F_{(p-2,n-p)}$ graus de liberdade, indicando o nível de significância tal que $P(F > F^*) = \alpha_{crtico}$.

```
pf(q,df1,df2,ncp, lower.tail=TRUE, log.p=FALSE)
qf(p,df1,df2,ncp, lower.tail=TRUE, log.p=FALSE)
# x e q são vetores quantis
# p é o vetor de probabilidades
# n - number of observations.
# df1 e df2 - são os gl (degrees of freedom)
# ncp - parâmetro não centralizado
```

1. Se α_{crtico} é suficiente pequeno (menor que 0,05 ou 0,01), rejeita-se a hipóteses de linearidade. Caso contrário, conclui-se que não existe evidência significativa contra esta hipótese.

Seja $x_i (i = 1, \ldots, p)$ aos valores distintos que toma a variável X. Para cada valor de x_i existem observações y_{ij} em que $j = 1, \ldots, n_i$ e $i = 1, \ldots, p$.

Exemplo 5.8.1. *A Tabela apresenta os valores de pessoas com idade de 25, 28, 32 e 36 anos que apresentaram casos de Tuberculose em 7 anos em uma determinada região.*

Anos/Idades	20	22	24	26
1999	114	123	101	130
2000	100	129	161	152
2001	111	110	122	170
2002	102	119	121	133
2003	118	131	123	158
2004	109	109	134	126
2005	106	130	128	129

Calculam-se as médias e o desvios padrões das distribuições condicionadas para cada valor de X.

```
Y <- matrix(c(114,123,101,130,
              100,129,161,152,
              111,110,122,170,
              102,119,121,133,
              118,131,123,158,
              109,109,134,126,
              106,130,128,129), ncol=4, byrow=T)
Y
X<-matrix(c(25,28,32,36), ncol=1, byrow=TRUE) #Idades
MediaX<-mean(X);MediaX
Idade25<-matrix(c(114,100,111,102,118,109,106),ncol=1)
Idade28<-matrix(c(123,129,110,119,131,109,130),ncol=1)
Idade32<-matrix(c(101,161,122,121,123,134,128),ncol=1)
Idade36<-matrix(c(130,152,170,133,158,126,129),ncol=1)

Y<-cbind(Idade25,Idade28,Idade32,Idade36)
Y
length(Y) # 28

Med_Anos<-apply(Y,2,mean)
Med_Anos<-as.matrix(Med_Anos)
Med_Anos
DP_Anos<-apply(Y,2,sd); DP_Anos
VAR_Anos<-apply(Y,2,var); VAR_Anos
VAR_Anos
MediaY<-mean(Y); MediaY
nCOVxy<-7*((25-30.25)*(MAnos[1])+(28-30.25)*
```

```
              (MAnos[2])+(32-30.25)*MAnos[3]+
              (36-30.25)*MAnos[4])
nCOVxy #1404
X<-matrix(c(25,28,32,36), ncol=1, byrow=TRUE) #Anos
MediaX<-mean(X)
MediaX # 30.25
nX<-7 # 7 anos
nS2x<-nX*(sum(25-MediaX)^2 + sum(28-MediaX)^2
+ sum(32-MediaX)^2 + sum(36-MediaX)^2)
nS2x  # 481.25
Ymed<-mean(Med_Anos);
Ymed  # 124.9643
Bi<- nCOVxy/nS2x; Bi  # 2.917403

#Yobs = Yprev + resíduos (Y-Yajustado)

Yprev<- Ymed + Bi*(X-MediaX)
Yprev

e1<-Y[,1]-Yprev[1]
e2<-Y[,2]-Yprev[2]
e3<-Y[,3]-Yprev[3]
e4<-Y[,4]-Yprev[4]
e<-rbind(e1,e2,e3,e4)
t(e)
n<-length(e);n
SR2<-sum(e^2)/(n-2)
SR2 # 178.974
DPR2<-sqrt(SR2)
DPR2 # 13.37812
Yprev<- Ymed + Bi*(X-MediaX);
as.matrix(Yprev)
S122a<-(108.5714-109.6479)^2+(121.5714-118.4001)^2
      +(127.1429-130.0697)^2+(142.5714-141.7394)^2
S122<-S122a*(7/2); S122 # 71.66032
VAR_Anos<-apply(Y,2,var); VAR_Anos
S22<-(6/24)*sum(VAR_Anos)
S22
S22<-(41.28571 + 85.95238 +326.47619 +297.95238)
S22*(6/24)  #  187.9167
F<-S122/S22; F  # 0.09533524
```

O modelo de regressão é: $\hat{Y} = 124,96 + 2,91(Idade - 30,25)$. Sendo $\bar{Y} = \frac{\sum_i \bar{Y}_j}{n} = 124,96$. E $\hat{\beta}_1 = \frac{Cov(Idade,Y)}{s^2_{Idade}} = 2.91$. Para testar a hipótese de linearidade, primeiro se calculam as diferenças entre as médias amostrais, $\bar{Y}_i$, e as obtidas mediante o modelo de regressão $\hat{Y}_i$. Os valores estimados são:

$$\begin{aligned}
\hat{Y}_{25} &= 124,96 + 2,91(25 - 30,25) = 109,64 \\
\hat{Y}_{28} &= 124,96 + 2,91(28 - 30,25) = 118,40 \\
\hat{Y}_{32} &= 124,96 + 2,91(32 - 30,25) = 130,06 \\
\hat{Y}_{36} &= 124,96 + 2,91(36 - 30,25) = 141,73
\end{aligned}$$

e para $\hat{Y}^2_{12}$ teremos:

$$\begin{aligned}
\hat{Y}^2_{12} &= \frac{\sum n_i(\bar{Y}_i - \hat{Y}_i)^2}{d-2} \\
&= 7/2[(108,57 - 109,64)^2 + (121,57 - 118,40)^2 + (127,14 - 130,06)^2 \\
&\quad + (142,57 - 141,73)^2] \\
&= 71.66
\end{aligned}$$

O valor de F será $F^*_{2,24} = \frac{187,91}{71,66} = 0,0953$

```
qf(0.09533524,2,24)
# [1] 0.1006103
pf(0.10,2,24)
# [1] 0.09478757
```

Não existe evidências estatísticas para rejeitar a linearidades entre as idades.

5.8.3 Multicolinearidade

No modelo linear $Y = XB + \varepsilon$, considere uma série de hipóteses entre as que se encontram as variáveis $X_1, X_2, \ldots, X_p$ são linearmente independentes, ou seja, não existe relação linear entre elas. Esta hipótese denomina-se hipótese de independência e quando não se cumpre, dizemos que o modelo apresenta multicolinearidade.

Sabe-se que quando tem forte colinearidade existe uma forte associação linear entre as variáveis independentes $X'X$, isso faz com que o determinante seja próximo de zero e não seria calculado $(X'X)^{-1}$ com o que não se podia encontrar o vetor de estimação dos parêmetros $(X'X)^{-1}X'Y$.

Os indícios mais comuns de multicolinearidade são:

1. Valores altos em módulo na matriz de correlação das variáveis explicativas.

2. Pouca significatividade das variáveis em X e o coeficiente de explicação alto.
3. Grande significatividade conjunta do model (grande rejeição de $R^2 = 0$.
4. Influência nas estimações da eliminação de uma observação no conjunto de dados.
5. Fatores de inflação da variância $FIV = \frac{1}{1-R_j^2}$ elevados (> 10), em que R_j^2 é o coeficiente de explicação da regressão auxiliar da variável explicativa j em função das demais variáveis explicativas.
6. Os autovalores λ_i do produto $X'X$ próximo a zero ou índice condicional $\lambda_{max/min}$ maior que 30.
7. Contraste de Farrar-Glauber baseado na estatística $G = -[T - 1(2p+5)/6]\log(\det(R_{xx}))$ que sob a hipótese nula de não multicolinearidade é uma χ^2 com $p(p-1)/2$ graus de liberdade. Sendo T o tamanho amostral, $p-1$ o número de variáveis explicativas de R_{xx} sua matriz de correlação.

As principais soluções para a multicolinearidade são:

- Ampliar a amostra ou transformar as variáveis (por exemplo, razão ou diferenças entre as variáveis).
- Suprimir algumas variáveis como uma boa justificativa estatística.
- Substituição das variáveis explicativas por suas componentes principais mais significativas, e posteriormente realizar uma regressão em suas componentes (que não são correlacionadas e não existe multicolinearidade). Ajusta-se a regressão nas componentes e substituindo cada uma delas por sua combinação linear em função das variáveis iniciais e o modelo resultante já ajustado nas variáveis iniciais. Este método se tem automatizado mediante a regressão de mínimos quadrados parciais.
- Usar a regressão em sequência, que oferece como estimadores dos parâmetros $(X'X + cI)^{-1}Y'X$, sendo c uma constante adequada. A matriz de covariância adota a forma $\sigma^2(X'X+cI)^2X'X(X'X+cI)^{-1})$. Na prática considera o valor de c entre $0,01$ a $0,1$ que faz com que o ajuste seja bom, ou seja o coeficiente de explicação elevado.

Exemplo 5.8.2. *Para ilustrar a existência de multicolinearidade, consideram-se duas variáveis numéricas X_1 e X_2, sendo que $X_2 = 2X_1$. Seja a matriz X definida por:*

$$X = \begin{pmatrix} 2 & 4 \\ 5 & 10 \end{pmatrix}$$

```
X1<-matrix(c(2,4,5,10), ncol=2, byrow=TRUE)
tXX1<- t(X1) %*% X1; tXX1
#         [,1] [,2]
# [1,]   29   58
# [2,]   58  116
det(tXX1)
# [1] 0
solve(tXX1)
# Error in solve.default(tXX1) :
#   Lapack routine dgesv:
# system is exactly singular: U[2,2]=0
X2<-matrix(c(2,4,5,9.9), ncol=2, byrow=TRUE)
tXX2<- t(X2) %*% X2; tXX2
#      [,1]   [,2]
# [1,] 29.0  57.50
# [2,] 57.5 114.01
det(tXX2)
# [1] 0.04
solve(tXX2)
#         [,1]    [,2]
# [1,]  2850.25 -1437.5
# [2,] -1437.50   725.0
```

Em X_1 observa-se que existe uma relação perfeita entre as duas variáveis, logo não é possível calcular a inversa. Já na matriz X_2 substituindo 10 por 9,9, a estimação já pode ser realizada, pois já existe uma inversa. No caso da matriz X_3 observa-se uma grande redução, como por exemplo, o elemento $a_{11} = 2850,25$ passou para 24,25. Assim, a matriz de covariância $(X'X)^{-1}$ é muito influenciada pela grau de multicolinearidade.

5.8.4 Tamanho da amostra

Em relação a amostra é importante observar uma série de recomendações.

- ◇ A amostra deve ser representativa de cada um dos grupos que estejam constituídos na variável dependente. Entretanto, não é necessário que o tamanho da amostra de cada grupo seja o mesmo.
- ◇ As variáveis deverão ser escolhidas de maneira que possam definir e discriminar os grupos a priori. Devem ser independentes uma das outras.
- ◇ Seria bom que tivesse 20 indivíduos em cada grupo da variável dependente, na prática isso nem sempre é possível.

5.8.5 Distância de Mahalanobis

Antes do Teste T^2 de Hotelling para duas amostras, vamos ver no *R* a distância de Mahalanobis. Sendo n_1 e n_2 o tamanho da amostra de cada grupo e D^2 a distância generalizada dada por: $D^2 = (x_1 - x_2)S^2(x_1 - x_2)$. Em que, x_1 e x_2 são médias amostrais e S uma estimativa da matriz de covariância populacional dos três grupos. Para isso considere três variáveis de casos de TB em 20 municípios no Estado de Pernambuco.

```
X1<-c(20,21.1, 24,20,22,16,20,20,18.5,17,23,
22,16,16.1,21.7,21.9,24.1,23,20.1,19.5)
X2<-c(19.5,14, 24.5,18,25,15,24,24.6,21.6,18,
20.1,22,15.8,16.1,25.3,21.9,24.1,23,18.3,24)
X3<-c(19.8,18, 25,19,21,17,23,25.7,19.5,18.6,
22.1,22.4,14.8,15.1,24.3,19.9,24.7,25,16.3,23)
```

Juntando as três variáveis na matriz X e encontrando o vetor de médias e a matriz de covariância de X, tem-se:

```
X<-cbind(X1,X2,X3)
mu=apply(X,2,mean)
sigma1=var(X)
```

Encontrando a distância de Mahalanobis para os 20 indivíduos, tem-se:

```
for(i in 1:20){
d<-matrix(c(X[i,]-mu), ncol=3)
D[i]<-d %*% solve(sigma1)%*% t(d)
D[1:20]
}
```

5.8.6 Teste T^2 de Hotelling para duas amostras

A primeira tarefa antes da análise discriminante é testar a hipótese de que os vetores médios são os mesmos nas duas populações das quais as amostras surgem. Para isso, o teste multivariado é análogo ao teste t de Student de amostras independentes, conhecido como teste T^2 de Hotelling.

A hipótese nula é que a média da variável na primeira população é igual à média das variáveis da segunda população e assim por diante. Suponha que os grupos são três populações. Assim, a hipótese a ser testada será:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

O teste T^2 de Hotelling é definido por:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2$$

Em que n_1 e n_2 são os tamanhos amostrais. Pode-se calcular a matriz de covariância conjunta (*pooled*) da seguinte forma:

$$S^2 = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Sob a hipótese nula H_0 a estatística F é dada por:

$$F = \frac{(n_1 + n_2 - p - 1)T^2}{(n_1 + n_2 - 2)p}$$

Com objetivo de aplicar o teste para duas amostras consideram-se aqui, apenas os dois primeiros grupos, a saber:

```
    dados<-read.table("dados3.txt",
    header=TRUE, sep="")
dados2<-dados[1:11,]
attach(dados2)
dados2
# G X1 X4
# 1  1  4  5
# 2  1  2  3
# 3  1  6  3
# ...
# 9  2  8  3
# 10 2  2  9
# 11 2  5  9
install.packages("rrcov")
library(rrcov)
grp <-as.factor(datos2[,1])
grp
x <- datos2[which(grp==levels(grp)[1]),2:3]
y <- datos2[which(grp==levels(grp)[2]),2:3]
T2.test(x,y)
T2.test(x,y)
# Two-sample Hotelling test
# data:  x and y
```

```
# T2=3.1236,F=1.3882,df1=2,df2=8,p-value=0.3037
# alternative hypothesis: true difference
# in mean vectors is not equal to (0,0)
# sample estimates:
# X1  X4
# mean x-vector 5.0 4.5
# mean y-vector 5.8 7.0
```

Tem distribuição F com p e $n_1 + n_2 - p - 1$ graus de liberdades. Assumindo que as variáveis têm distribuição normal, os três grupos têm a mesma matriz de covariância e que as observações são independentes.

```
dados2
m1<-apply(dados2[G==1,-1],2,mean);m1
m2<-apply(dados2[G==2,-1],2,mean);m2
n1<-length(G[G==1]);n1
n2<-length(G[G==2]);n2
Gx1<-dados2[dados2==1,-1];Gx1
Gx2<-dados2[dados2==2,-1];Gx2
Gx2<-Gx2[-c(6,7,8),];Gx2
S123<-((n1-1)*var(Gx1)+(n2-1)*
    var(Gx2))/(n1+n2-2)
    S123
D2<-t(m1-m2)%*%solve(S123)%*%(m1-m2)
T2<- ((n1*n2)/(n1+n2))*D2;T2
# [1,] 3.123551
p<-2 # 2 variáveis
gl1<- p; gl2<- n1+n2-p-1
1-pf(1.3882,gl1,gl2)
0.3037142
test=as.vector(((n-p-1)*T2)/((n-2)*p))
test
# 1.3882
```

Portanto, aceita-se a hipótese de que os vetores de médias dos dois grupos são iguais, pois 0.3037142 > 0,05, assim não existe evidência estatística para rejeitar a igualdade dos vetores de médias do grupo 1 e 2. O valor crítico é $qf(0.95,2,8) = 4.4589$, assim, $4.4589 > 3.1235$, logo aceita-se H_o.

Capítulo 6

Gráficos de médias e densidades dos grupos

O gráfico de densidades e médias dos dois grupos (eixo das abscissas) contribui para visualizar a dispersão. Cada célula (i, j) de uma matriz de dispersão contém o `scatterplot` da coluna i versus a coluna j da matriz de dados.

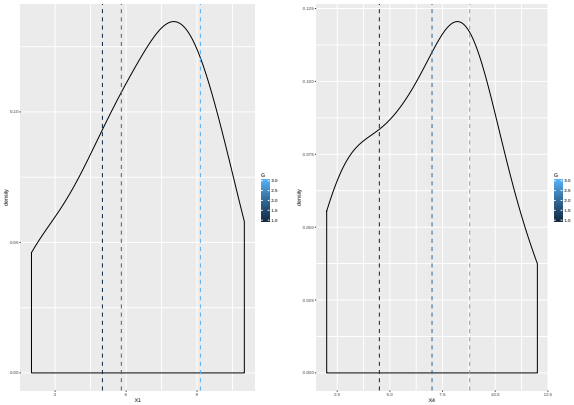


Figura 6.1: Densidades e médias dos dois grupos (eixo das abscissas).

Observa-se no eixo X que as linhas verticais informam os valores das médias de cada variável em cada grupo. Pode-se também usar o seguinte algoritmo para encontrar as médias, desvio padrão e tamanho dos grupos.

```
Med_Desvio_Tamanho<-function(variaveis,
    grupos){
# find the names of the variables
nomesvar<-c(names(grupos),
            names(as.data.frame(variaveis)))
grupos<- grupos[,1]
#Dentro de cada grupo encontra a média
    # de cada variável
medias<-aggregate(as.matrix(variaveis)~grupos,
    FUN=mean)
names(medias) <- nomesvar
print(paste("Médias:"))
print(medias)
# Dentro de cada grupo encontra o
    #desvio de cada variável
dp<- aggregate(as.matrix(variaveis)~grupos,
    FUN=sd)
names(dp)<- nomesvar
print(paste("Desvio padrão:"))
print(dp)
#Dentro de cada grupo o tamanho da amostra
TamAmost<-aggregate(as.matrix(variaveis)~grupos,
FUN = length)
names(TamAmost)<- nomesvar
print(paste("Tamanho da amostra:"))
print(TamAmost)}
Med_Desvio_Tamanho(datos[2:3],datos[1])
# [1] "Médias:"
# G       X1       X4
# 1 1 5.000000 4.500000
# 2 2 5.800000 7.000000
# 3 3 9.142857 8.785714
# [1] "Desvio padrão:"
$ G       X1      X4
# 1 1 2.097618 2.50998
# 2 2 2.489980 2.54951
# 3 3 1.345185 2.15749
# [1] "Tamanho da amostra:"
```

```
# G X1 X4
# 1 1  6  6
# 2 2  5  5
# 3 3  7  7
```

Também se pode separar cada grupo para algumas estatísticas ou algumas operações matemáticas

```
Grupo1 <- datos[datos$G=="1",]; Grupo1
# G X1 X4
# 1 1  4  5
# 2 1  2  3
# 3 1  6  3
# 4 1  4  5
# 5 1  6  2
# 6 1  8  9
sapply(Grupo1[2:3],mean) # médias
# X1  X4
# 5.0 4.5
sapply(Grupo3[2:3],var) #variâncias
# X1       X4
# 1.809524 4.654762
sapply(Grupo3[2:3],sd) #desvio padrão
# X1       X4
# 1.345185 2.157490
    # raiz de cada valor da variável
sapply(Grupo3[2:3],sqrt)
# X1       X4
# [1,] 3.316625 2.828427
# [2,] 3.000000 2.645751
# [3,] 3.162278 3.464102
# [4,] 2.828427 2.449490
# [5,] 2.645751 3.316625
# [6,] 3.162278 2.828427
# [7,] 3.000000 3.082207
Grupo2 <- datos[datos$G=="2",]
Grupo3 <- datos[datos$G=="3",]
sapply(Grupo2[2:3],mean)
sapply(Grupo3[2:3],mean)
```

Para encontrar a correlação de Pearson, intervalo de confiança, valor da estatística de teste t e o valor-p entre as variáveis, utilize:

```
cor.test(datos$X1, datos$X4)
# Pearson's product-moment correlation
# data:  datos$X1 and datos$X4
# t = 1.735, df = 16, p-value = 0.1019
# alternative hypothesis: true correlation is not
    # equal to 0
# 95 percent confidence interval:
# -0.08465992  0.72931292
# sample estimates:
cor
# 0.3979388
```

Referências Bibliográficas

Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 2 edition, 1984.

Ávila, M. J. del M., García, J. M. T. *Técnicas Estadísticas Aplicadas*. Grupo Editorial Universitário, 2006.

Ávila, M. J. del M. *Estadística Matemática*. Grupo Editorial Universitário, 2006.

Bernstein, I. H. *Applied Multivariate Analysis*. Springer - Verlag, 1987.

Corrar, L. J., Paulo, E., Dias Filho, J. M. *Análise Multivariada*. São Paulo: Atlas, 2007.

Dillon, W., Goldstein, M. *Multivariate Analysis*. New York: Wiley, 1984.

Everitt, B., Hothorn, T. *An Introduction to Applied Multivariate Analysis with R*. Springer, 2011.

Faceli, K., Lorena, A. C., Gama, J., Carvalho, A. C. P. L. F. *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC, 2011.

Fernández-Abascal, H., Guijarro, M. M., Rojo, J. L., Sanz, J. A. *Cálculo de probabilidades y estadística*. Barcelona: Editorial Arial, S. A., 1994.

Ferreira, D. F. *Estatística multivariada*. 1. ed. Lavras: Ed. UFLA, 2008.

Flury, B., Riedwyl, H. *Multivariate Statistics*. Chapman and Hall, 1988.

Genz, A., Bretz, F. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics, Vol. 195., Springer-Verlag, Heidelberg, 2009. ISBN 978-3-642-01688-2.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T. *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-5, 2016. http://CRAN.R-project.org/package=mvtnorm

Gross, J., Ligges, U. *nortest: Tests for Normality*. R package version 1.0-4, 2015. http://CRAN.R-project.org/package=nortest

Jaimez, R.G., Carmona, A.G. *Estadística Multivariable. Volumen I: Introducción al Análisis Multivariante*. Universidad de Granada, 1991.

Hair, J. F., Anderson, R. E., Tatham, R. L. y Black, W. C. *Análisis Multivariante*. 5ª edición. Ed. Prentice Hall, 1999.

Härdle, W. K., Simar, L. *Applied Multivariate Statistical Analysis*. 3ª edición. Springer, 2011.

Härdle, W. K., Hlávka, Z. *Multivariate Statistics: Exercises and Solutions*. 2ª edición. Springer, 2015.

Konishi, S., Kitagawa, G. *Information Criteria and Statistical Modeling*. Springer, 2008.

Jiménez, E. U., Manzano, J. A. *Análisis Multivariante Aplicado*. Thomson, 2005.

Lebart, L., Morineau, A., Warwick, K. M. *Multivariate Descriptive Analysis*. New York: Wiley, 1984.

Ligges, U., Mächler, M. *Scatterplot3d - an R Package for Visualizing Multivariate Data*. Journal of Statistical, Software 8(11), 1-20, 2003.

Maindonald, J. H., Braun, W. J. *DAAG: Data Analysis and Graphics Data and Functions*. note = R package version 1.22, 2015. http://CRAN.R-project.org/package=DAAG

McLachan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

Mingoti, S.A. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Belo Horizonte: UFMG, 2005.

Paradis, E. *R for Beginners*. Institut des Sciences de lˆÉvolution, Université Montpellier II, France, 2005. https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

Peña, D. *Análisis de datos multivariantes*. España: McGraw-Hill, 2002.

Peña, D. *Regresión y Diseño de Experimentos*. Alianza Editorial, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna: Austria, 2015. http://www.R-project.org/.

Reis, E. *Estatística Multivariada Aplicada*. Edições Sílabo, Lisboa, 2001.

Rmetrics Core Team., Wuertz, D., Setz, T., Chalabi, Y. *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 3011.87, 2014, http://CRAN.R-project.org/package=fBasics.

Sarkar, D., Andrews, F. *latticeExtra: Extra Graphical Utilities Based on Lattice*. R package version 0.6-28, 2016. http://CRAN.R-project.org/package=latticeExtra

Sharma, S. *Applied Multivariate Techniques*. Jonh Wiley & Sons, New York, 1996.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. *"ROCR: visualizing classifier performance in R"*. Bioinformatics,21(20), pp. 7881. 2005. http://rocr.bioinf.mpi-sb.mpg.de.

Todorov, V., Filzmoser, P. Package *rrcov*. *An Object-Oriented Framework for Robust Multivariate Analysis*. Journal of Statistical Software, 32(3), 1-47. 2009. http://www.jstatsoft.org/v32/i03/.

Jiménez, E.U, Manzano, J.A. *Análisis Multivariante Aplicado*. Thomson, España, 2205.

Van der Loo, M.P.J. *extremevalues, an R package for outlier detection in univariate data*. R package version 2.3, 2010.

Venables, W.N., Ripley, B.D. Package *MASS*. *Modern Applied Statistics with S*. Fourth Edition. Springer, New York, 2002. ISBN 0-387-95457-0

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. Package *klaR*. *Analyzing German Business Cycles*. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335-343, Springer-Verlag, Berlin, 2005.

Whittaker, J. *Graphical Models in Applied Multivariate Statistics*. Wiley, 1990.

Organizador e Autores

Edwirde Luiz Silva Camêlo (Brasil) - (Organizador) Professor Associado da Universidade Estadual da Paraíba (UEPB). Pós-doutorado em *Estadística Aplicada* (2016) e Doutor em *Estadística e Investigación Operativa* (2007) pela *Universidad de Granada*. Mestrado em Biometria e Estatística Aplicada (2001) pela Universidade Federal Rural de Pernambuco (UFRPE). Técnicas de estatística multivariada aplicada em diversas áreas tem sido sua principal linha de pesquisa. E-mail: edwirde@uepb.edu.br

Paulo Lisboa (Inglaterra) - Professor e chefe do Departamento de Matemática Aplicada da *Liverpool John Moores University* (LJMU). Estudou física matemática na Universidade de Liverpool, onde obteve um doutorado em física teórica de partículas em 1983. Foi nomeado para a cátedra de Matemática Industrial na (LJMU) em 1996 e Chefe de Graduação em 2002. PhD em Física de Partículas (1983) e bacharel em Física matemática (1979) pela Universidade de Liverpool. Aplica ciência de dados para medicina personalizada, saúde pública, análise esportiva e marketing digital. Vice-presidente do Grupo Consultivo *Horizon2020* para o Desafio Societário 1: Saúde, Mudança Demográfica e Bem-estar, fornecendo conselhos científicos a um dos maiores programas de pesquisa coordenada do mundo em saúde. Membro do Conselho do Instituto de Matemática e suas Aplicações. Presidente da Equipe de Tarefa de Análise de Dados Médicos no Comitê Técnico de Mineração de Dados do *IEEE*. Presidente do Comitê de Prêmio *JA Lodge* e presidente da Rede Profissional de Tecnologias de Saúde na Instituição de Engenharia e Tecnologia. Assessora o *Group of Performance.Lab at Prozone* e tem papéis de revisão editorial e de pares em várias revistas e órgãos de financiamento da pesquisa, incluindo EPSRC. E-mail: P.J.Lisboa@ljmu.ac.uk

Andrés González Carmona (Espanha) - Professor Titular da *Universidad de Gra-*

nada (UGR), Diretor do *Departamento de Estadística e Investigación Operativa*, Coordenador do Mestrado em *Estadística Aplicada*, Diretor da *Área de Formación del Centro Andaluz de Prospectiva*, Vice-Presidente da *Academia de Ciencias Matemáticas, Físico-Químicas y Naturales de Granada*. E-mail: andresgc@ugr.es

Ramón Gutiérrez Sánchez (Espanha) - Professor da *Universidad de Granada* (UGR), Secretário do Departamento de *Estadística e IO*. Doutor desde 2005, pertence à linha de pesquisa de Análise Multivariada e Processos Estocásticos. Publicou mais de 40 artigos em *JCR* na área de Estatística. Também colabora com o Departamento de Parasitologia da UGR. E-mail: ramongs@ugr.es

Dalila Camêlo Aguiar (Brasil) - Doutoranda em *Estadística Matemática y Aplicada* pela *Universidad de Granada* (UGR), Mestra em *Estadística Aplicada* (2016) pela UGR, Especialista em Estatística Aplicada (2011) pela Fundação de Apoio, Pesquisa e Extensão (FURNE) e Bacharela em Estatística (2010) pela Universidade Estadual da Paraíba (UEPB). Pesquisadora na área de estatística multivariada aplicada. E-mail: dalilacamelo@correo.ugr.es