Universidade Federal de Mato Grosso
Instituto de Ciências Exatas e da Terra
Departamento de Estatística

Introdução ao R

Anderson Castro Soares de Oliveira

2011

SUMÁRIO

1

INTRODUÇÃO

R é uma linguagem e ambiente para computação estatística e gráficos, e começou a ser desenvolvido por Robert Gentleman e Ross Ihaka do Departamento de Estatística da Universidade de Auckland em Nova Zelândia, mais conhecidos por "R & R", apelido do qual originou-se o nome R do programa. O objetivo inicial de "R & R", em 1991, era produzir um software para as suas aulas de laboratório baseado na já revolucionária linguagem S, utilizada pelo software comercial S-Plus criado por Jonh M. Chambers da AT&T que atualmente vem contribuindo para o aperfeiçoamento e ampliação das análises estatísticas do R.

O primeiro relato da distribuição do R foi em 1993, quando algumas cópias foram disponibilizadas no StatLib, um sistema de distribuição de softwares estatísticos. Para saber mais sobre o StatLib, acesse http://lib.stat.cmu.edu/. Com o incentivo de um dos primeiros usuários deste programa, Martin Mächler do ETH Zürich (Instituto Federal de Tecnologia Zurique da Suíça), "R & R", em 1995, lançaram o código fonte do R, disponível por ftp , sobre os termos de Free Software Foundations GNU general license, que seria um tipo de "licença para softwares livres". Para saber mais sobre a organização Free Software Foundation's e o projeto GNU, acessar www.gnu.org.

Em 1997 foi formado um grupo de profissionais que têm acesso ao código fonte do R, possibilitando assim a atualização mais rápida do software, para ver esse grupo acessar http://www.r-project.org/contributors.html. Desde então o R vem ganhando cada vez mais adeptos em todo o mundo, em parte devido ao fato de ser totalmente gratuito e também por ser um programa que exigi do usuário o conhecimento das análises que está fazendo, diminuindo assim as chances de uma interpretação errada dos resultados.

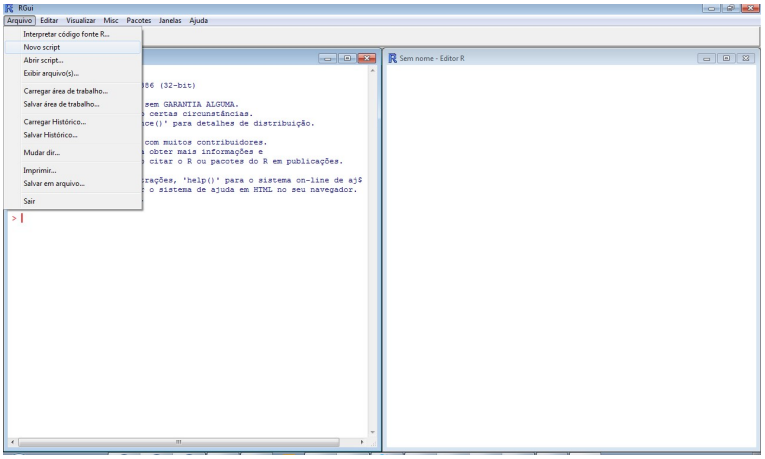
Outro fato importante para a difusão do R é a sua compatibilidade com quase todos os sistemas operacionais. O R está disponível para a maior parte dos MacOS, Windows a partir do 95 e para UNIX e sistemas similares como Linux e FreeBSD.

2

USANDO O R

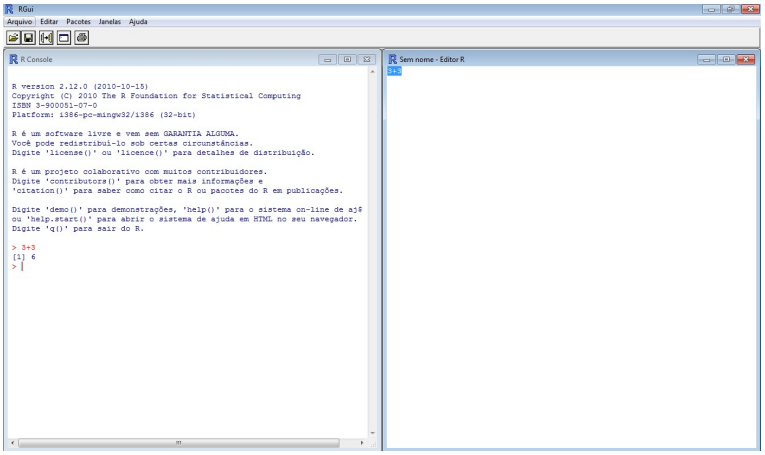
O R não é um programa com muita interatividade, a menos que interfaces gráficas sejam utilizadas. As análises feitas no R são digitadas diretamente na linha de comandos. No Windows o R apresenta uma interface gráfica com barra de ferramentas no topo, que permite realizar as tarefas comuns: abrir/gravar cheiros, cortar/colar texto, instalar pacotes, gerir as janelas, etc.) surge uma janela (R Console) na qual se introduzem os comandos, a seguir ao sinal.

Uma maneira que otimiza o uso do R e que poupa muito tempo é usar um script para digitar seus comandos. Neste caso, os comando não são digitados na linha de comandos e sim em um editor de texto (R editor). Um script é um arquivo, onde você digita todas as análises e comandos, faz alterações e correções, além de salvar o script e poder refazer rapidamente suas análises após algum tempo. Para criar um scrip basta ir na barra de ferramentas `Arquivo`, em seguida em `Novo script`



Para executar os comandos digitados no script, seleciona-se o texto ou linhas de interesse e utiliza-se as teclas `CTRL+R` ou `F5`.

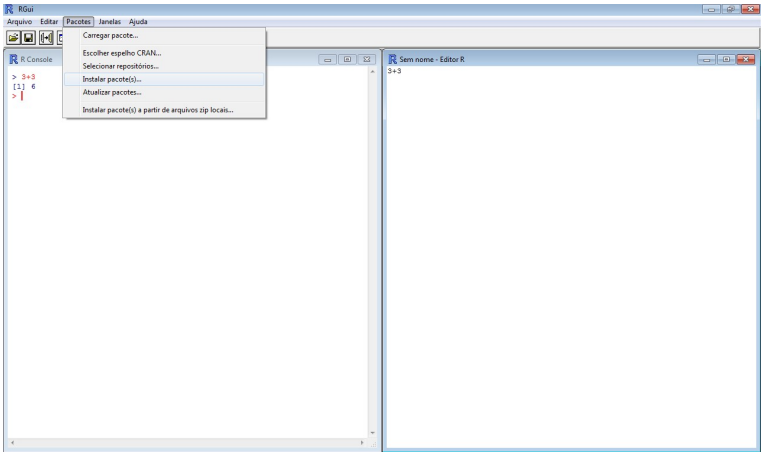
Por exemplo digite 3+3 no script e aperte `CTRL+R`. O comando 3+3 será enviado para a linha de comandos do R e o resultado aparecerá na tela.



2.1 INSTALANDO PACOTES NO R

O R em geral é instalado apenas as configurações mínimas para seu funcionamento básico são instaladas (pacote base). Para realizar tarefas mais complicadas pode ser necessário instalar pacotes adicionais (packages ou library), esses pacotes são escritos por pesquisadores das mais diferentes áreas do conhecimento e profissionais da area de estatística. Qualquer pessoa pode contribuir para o desenvolvimento do programa R mediante criação de pacotes que façam determinadas análises.

Para instalar um pacote, deve estar conectado a internet, assim, vá na barra de ferramentas `Pacotes` em seguida em `Instalar pacote(s)`,



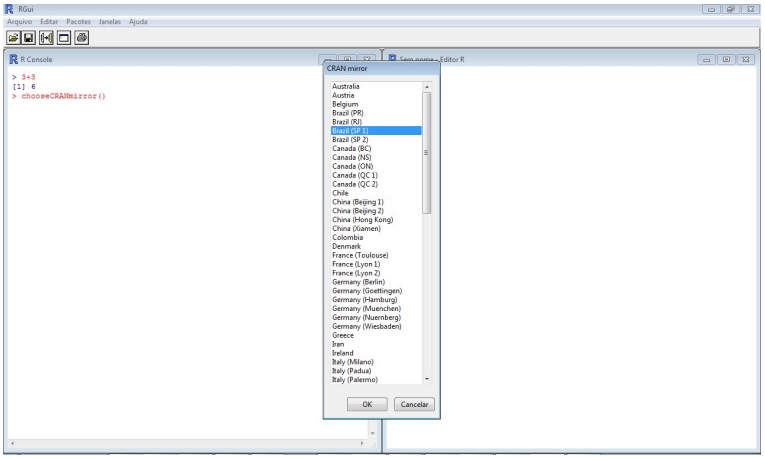
em seguida irá abrir uma janela de espelho do CRAN (base de pacotes).

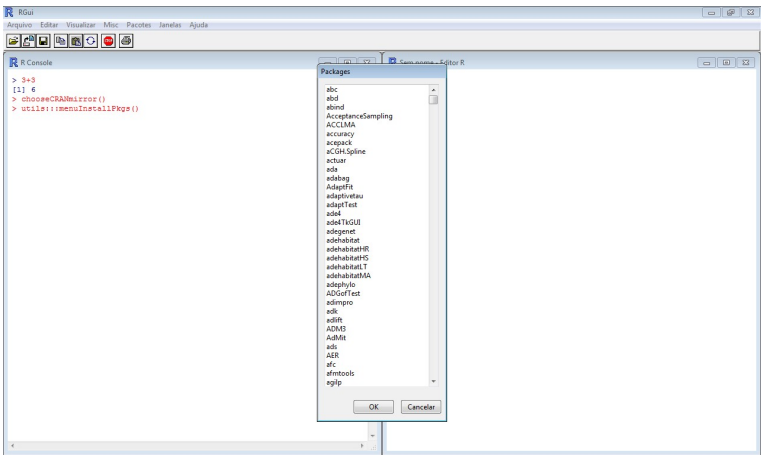
Escolhido o espelho do qual irá baixar o pacote, irá aparecer uma lista de pacotes, assim escolhe qual quer instalar.

Outra maneira de instalar um pacote é digitar a linha de comando:

`install.packages("nome do pacote",dependencies=TRUE)`

Após instalar um pacote, para usa-lo é necessário carrega-lo, para isto utilize uma das linhas de comando: `library(nome do pacote)` ou `require(nome do pacote)`





2.2 AJUDA E DOCUMENTAÇÃO

Existem três formas obter documentação para o R

- sistema de ajuda;
- ajuda online;
- manuais e publicações eletrônicas trabalho sob a forma de livros, etc.

O sistema de ajuda é uma coleção de páginas de manual que descreve cada função para o usuário. Para ver a ajuda relativa a uma basta utilizar os comandos:

```
help("função")
?função
```

no lugar de `função` coloca-se o nome da função que ser quer ajuda, por exemplo ajuda sobre a média, utiliza-se

```
help("mean")
```

Para ver informações relativa a um pacote utiliza-se o comando `help(package ="nome pacote")`

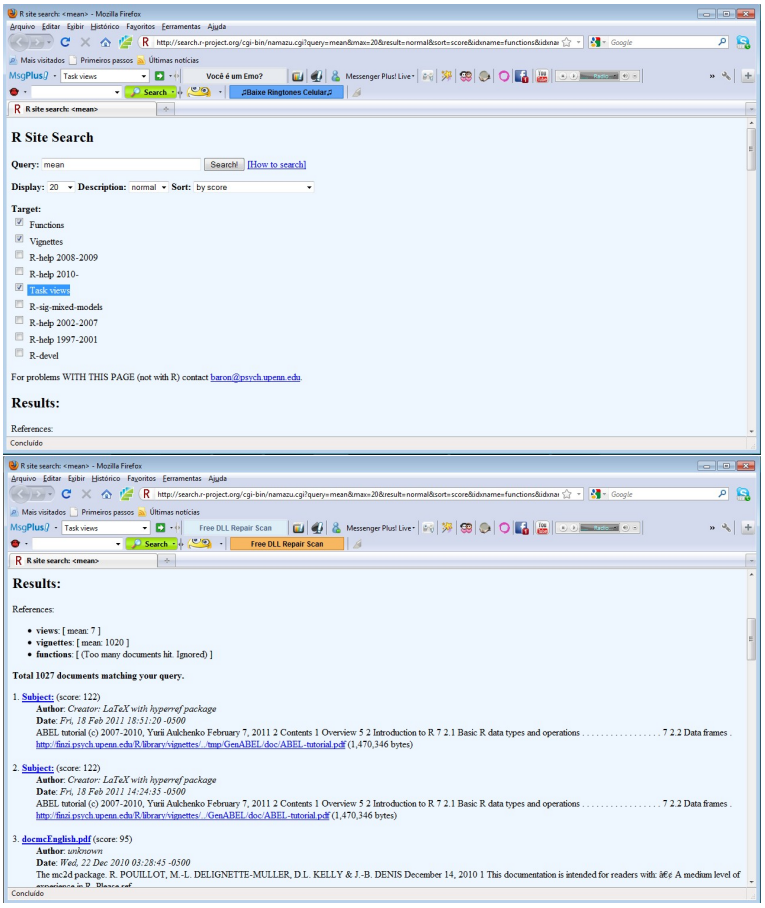
A ajuda online possui diversas informações relativos a funções e perguntas relativas a utilização do R. Para ver ajuda on line deve estar conectado a internet e utilizar o comando:

`RSiteSearch("nome de algo pra ajuda")`

Por exemplo para obter ajuda on line sobre a média, utiliza-se

`RSiteSearch("mean")`

e irá abrir um pagina de consulta on line do R contendo os resultados encontrados de funções e perguntas relativas a função `mean`.



Documentação mais completa está disponível por via electrônica a partir da coleção dos manuais na pagina

`http://CRAN.R-project.org/manuals.html`

2.3 OBJETOS DO R

O R é uma linguagem baseada em objetos. Isto quer dizer que tudo o que for usado no R está guardado na memória do computador sob a forma de um objeto. Todos os objetos em R tem um nome associado e podem armazenar diferentes tipos de objetos no R tais como vetores, matrizes, data frames, listas, funções, expressões e muitos outras.

Para armazenar algo num objeto pode ser usado o operador de atribuição `<-` ou o simbolode igualdade `=`.

Por exemplo:

```
> x=5
```

Para ver o conteúdo de um objeto basta digitar o nome do objeto no prompt do R, carregando em seguida um objeto em Enter. Como resposta o R mostra-nos o conteúdo do objeto que lhe indicamos.

```
> x
[1] 5
```

O `[1]` que aparece antes do número guardado em `x`, significa que esta linha esta mostrando o conteúdo de `x` e começa no primeiro elemento.

Podemos listar quais os objetos que estão na memória do computador usando as função `ls()`. Para remover um objeto, podemos utilizar a função `rm(nome do objeto)`, ou se quiser remover todos os objetos da memoria utiliza-se o comando `rm(list=ls(all=TRUE))`

2.3.1 Funções básicas

O uso mais básico do R é usá-lo como calculadora. Os operadores são:

- `+` soma,
- `-` subtração,
- `*` multiplicação,
- `/` divisão,
- `^` exponenciação.

O R tem diversas funções matemáticas que podemos usar para fazer os cálculos desejados

- `sqrt(numero)` raiz quadra.
- `exp(numero)` exponencial.
- `log(numero)` logaritmo de base natural
- `log(numero,base)` logaritmo para qualquer base, por exemplo 10 coloca-se no lugar de `base`.
- `abs(numero)` modulo
- `factorial(numero)` fatorial
- `max(nome)` máximo
- `min(nome)` mínimo
- `sum(nome)` somatório
- `sin(numero)`, `cos(numero)`, `tan(numero)` funções trigonométricas;
- `asin(numero)`, `acos(numero)`, `atan(numero)` funções trigonométricas inversas;
- `sinh(numero)`, `cosh(numero)`, `tanh(numero)` funções hiperbólicas;
- `asinh(numero)`, `acosh(numero)`, `atanh(numero)` funções hiperbólicasinversas;

Para mais detalhes sobre as funções básicas, consulte

`http://finzi.psych.upenn.edu/R/library/base/html/00Index.html`

As funções básicas podem ser utilizadas separadamente ou por meio de expressões, por exemplo:

```
> x=5
> y=2
> z=sqrt((x+y)^2)/3
> z
[1] 2.333333
in(pi)
> w
[1] 1.224606e-16
```

2.3.2 Vetores

Os vetores são sequencias de valores, que podem ser de vários tipos, por exemplo, lógicos, inteiros, reais e caracteres. O modo mais simples de criar um vetor é utilizar a função de concatenação `c()` separando os elementos que formam o vetor por vírgulas. Para verificar o tamanho do vetor, utiliza-se a função `length(nome do vetor)`.

Por exemplo criar um vetor de tamanho 5, contendo os valores (9, 7.2, 5, 3, 4.2)

```
> x<-c(9,7.2,5,3,4.2)
> x
[1] 9.0 7.2 5.0 3.0 4.2
> length(x)
[1] 5
> mode(x)
[1] "numeric"
```

Os elementos de um vetor podem ser acessados por meio de um índice. Para poder acessar os valores, coloca-se o nome do vetor seguido de colchetes, e dentro do colchete um numero que indica a posição do elemento que se quer acessar.

Assim, se quiséssemos acessar o 4 elemento do vetor `x` usaríamos

```
> x[4]
[1] 3
```

Em alguns casos gerar vetores em forma de sequencias, para isto utilizamos a função `seq`. Por exemplo, podemos criar um vetor com a sequencia de números 1 a 5, utilizando uma variação de 0,5.

```
> x=seq(1,5,by=0.5)
> x
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

Podemos também criar uma sequencia variando de 1 a 5, em que contenha 4 elementos

```
> x=seq(from = 1, to = 5, length = 4)
> x
[1] 1.000000 2.333333 3.666667 5.000000
```

Uma outra função bastante útil para gerar sequências é a função `rep()`. Por exemplo queremos criar uma sequencia de 3 repetidos 10 vezes.

```
> y=rep(3, 10)
> y
 [1] 3 3 3 3 3 3 3 3 3 3
```

Podemos também criar uma sequencia entre, 1 e 3, repetida 5 vezes.

```
> w=rep(1:3, 5)
> w
 [1] 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
```

2.3.2.1 Operações com Vetores

Um dos aspectos mais poderosos da linguagem R é poder realizar com vetores elemento a elemento. Essas operações são realizadas por meio das funções basica.

```
> v1 <- c(2,8,5,7)
> v2 <- c(5,4,2,9)
> v1 + v2
[1]  7 12  7 16
```

Em operação entre vetores, caso um dos vetores, tenha um tamanho menor, o R vai repetindo os seus elementos atpe atingir o tamanho do maior. Por exemplo,

```
> v1 <- c(2,8,5,7)
> v2 <- c(5,4)
> v1 + v2
[1]  7 12 10 11
```

2.4 MATRIZES

As matrizes são coleções de vetores em linhas e colunas, todos os vetores dever ser do mesmo tipo (numérico ou de caracteres). Para criar uma matriz utiliza-se a função
`matrix(elementos,numero de linhas, numero de colunas)`

Por exemplo, vamos criar uma matriz contendo os numero de 1:12, com 4 linha e 3 colunas,

```
> A=matrix(1:12,4,3)
> A
     [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
```

```
[3,]    3    7   11
[4,]    4    8   12
```

Nas matrizes também é possível dar nomes aos elementos das linhas e colunas, per meio das funções `rownames` e `colnames`

```
> rownames(A)=c("linha1","linha2","linha3","linha4")
> colnames(A)=c("coluna1","coluna2","coluna3")
> A
       coluna1 coluna2 coluna3
linha1       1       5       9
linha2       2       6      10
linha3       3       7      11
linha4       4       8      12
```

Os elementos de uma matriz podem ser acessados por meio de índice de linhas e colunas. Para poder acessar os valores, coloca-se o nome da matrix seguido de colchetes, e dentro do colchete um numero que indica a posição da linha e da coluna do elemento que se quer acessar.

Assim, se quiséssemos acessar o elemento que esta na 2 linha e 1 coluna da matrixr `A` usaríamos

```
> A[2,1]
[1] 2
```

As funções `cbind()` e `rbind()` podem ser usadas acrescentar linhas e colunas nas matrizes, respectivamente. Os seguintes exemplos ilustram o seu uso,

```
> x=c(3,4,5,6)
> y=c(6,7,8,9)
> A=cbind(A,y)
> A=rbind(A,x)
> A
       coluna1 coluna2 coluna3 y
linha1       1       5       9 6
linha2       2       6      10 7
linha3       3       7      11 8
linha4       4       8      12 9
x            3       4       5 6
```

As funções `nrow()` `ncol()` `dim()` são utilizadas para ver o numero de linhas, numero de colunas e dimensão da matriz.

```
> nrow(A)
[1] 5
> ncol(A)
[1] 4
> dim(A)
[1] 5 4
```

2.4.1 Operações com matrizes

As operações com matrizes, podem ser feitas diretamente. Para realizar a soma de matrizes, utiliza-se o operador +. Se desejarmos o produto matricial entre duas matrizes, devemos utilizar o operador %*%.O comando `t()` retorna a matriz transpota. O comando `solve( )`, retorna a matriz inversa. A função `eigen()` retorna os autovalores e autovetores da matriz.

```
> A=matrix(c(5,2,6,1,6,3,1,4,4),3,3)
> B=matrix(c(10,12,8,10,17,6,11,10,12),3,3)
> A
     [,1] [,2] [,3]
[1,]    5    1    1
[2,]    2    6    4
[3,]    6    3    4
> B
     [,1] [,2] [,3]
[1,]   10   10   11
[2,]   12   17   10
[3,]    8    6   12
> A+B
     [,1] [,2] [,3]
[1,]   15   11   12
[2,]   14   23   14
[3,]   14    9   16
> A%*%B
     [,1] [,2] [,3]
[1,]   70   73   77
[2,]  124  146  130
[3,]  128  135  144
> t(A)
     [,1] [,2] [,3]
[1,]    5    2    6
[2,]    1    6    3
[3,]    1    4    4
> solve(A)
           [,1]        [,2]        [,3]
[1,]  0.2608696 -0.02173913 -0.04347826
[2,]  0.3478261  0.30434783 -0.39130435
[3,] -0.6521739 -0.19565217  0.60869565
> eigen(A)
$values
[1] 10.107856  3.642881  1.249263
$vectors
```

```
           [,1]       [,2]        [,3]
[1,] -0.2659764 -0.4121283 -0.03663773
[2,] -0.7358580  0.8600578 -0.63457438
[3,] -0.6227115 -0.3007506  0.77199289
```

2.5 ARQUIVOS DE DADOS

Vários são os formatos que podem ser lidos pelo R. É possível ter vários conjuntos de dados abertos ao mesmo tempo. Como resultado, cada conjunto de dados deve ser dado um nome único pelo qual pode ser referido e identificado.

Antes de importar um arquivo de dados é necessário informar o diretório onde está o arquivo. Para isto pode utilizar a função `setwd`.

Por exemplo, se o conjunto de dados está no diretório Meus Documentos,

```
setwd("C:/Users/anderson/Documents")
```

2.5.1 Dados de texto

Para carregar um banco de dados no formato texto basta utilizar a função `read.table()` da seguinte forma:

```
dados=read.table("nome do arquivo.txt",header=TRUE)
```

A opção `header=TRUE` indica que no arquivo de dados a primeira linha contem o nome da variáveis, caso contrário deve-se utilizar `header=FALSE`.

2.5.2 Dados do Excel

Para carregar um banco de dados do Excel basta utilizar a função `read.xlsx()`, disponível no pacote `xlsx`. Esta função é utilizada da seguinte forma:

```
require(xlsx)
dados=read.xlsx("nome do arquivo.xls",sheetIndex,header = TRUE)
```

A opção `sheetIndex` indica qual a planilha encontra-se o dados, por exemplo se esta na primeira planilha, substitui `sheetIndex` por 1. A opção `header=TRUE` indica que no arquivo de dados a primeira linha contem o nome da variáveis, caso contrário deve-se utilizar `header=FALSE`.

2.6 FUNÇÕES DIVERSAS

A função `sample` pode ser utilizada para criar uma amostra de um conjunto de dados, ou vetor. Quando vai se gerar amostras aleatórias é importante estabelecer uma semente de números aleatórios por meio da função `set.seed`.

```
> ##vetor de sequencia de 0 a 200 a cada 5
> x=seq(0,200,by=5)
> x
```

```
 [1]   0   5  10  15  20  25  30  35  40  45  50  55  60  65  70  75  80  85  90
[20]  95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185
[39] 190 195 200
> ##amostra de tamanho 10
> set.seed(12345)
> x1=sample(x,10)
> x1
 [1] 145 175 200 165  80  25  55  85 120 155
```

A função `sort` coloca os valores de um objeto em ordem crescente ou em ordem decrescente.

```
> ##Dados ordenados de x1
> sort(x1)
 [1]  25  55  80  85 120 145 155 165 175 200
> sort(x1,decreasing=TRUE)
 [1] 200 175 165 155 145 120  85  80  55  25
```

A função `order` coloca uma planilha de dados seguindo a ordem de uma de suas variáveis.

```
> ##Ordenar planilha
> x2=sample(x,10)
> y=cbind(x1,x2)
> y
       x1  x2
 [1,] 145   5
 [2,] 175  30
 [3,] 200 140
 [4,] 165   0
 [5,]  80  70
 [6,]  25  80
 [7,]  55  65
 [8,]  85 170
 [9,] 120  25
[10,] 155 150
> ## ordenar y de acordo com os valores de x1,
> y1=y[order(y[,1]),]
> y1
       x1  x2
 [1,]  25  80
 [2,]  55  65
 [3,]  80  70
 [4,]  85 170
 [5,] 120  25
 [6,] 145   5
```

```
 [7,] 155 150
 [8,] 165   0
 [9,] 175  30
[10,] 200 140
```

2.7 EXERCÍCIOS

1) Os dados apresentados a seguir referem-se ao tempo que determinada de duas marcas de transformado A e B, levou para apresentar a primeira falha grave, em anos, obtidos em uma amostra de n = 10 transformadores.Os resultados do tempo de falhas em anos são dados por

Marca A	2,80	3,86	5,23	5,62	5,98	6,51	6,62	7,47	7,80	8,37
Marca B	2,91	4,08	5,27	5,80	6,08	6,52	6,66	7,50	7,81	8,49

a) Criar um vetor com os dados do tempo da primeira falha grave para cada marca;

b) modificar o 3 elemento do vetor da marca A para 5,44;

c) Obter os valores máximos, mínimos e a soma de cada marca;

d) fazer a soma, subtração e multiplicação dos vetores das marcas A e B;

e) Criar uma matriz com a primeira linha para marca A, e a segunda linha com a marca B;

f) Renomear a linhas da matriz tempo marca A e tempo marca B.

2) Criar um vetor com a sequencia de números 2 a 12, utilizando uma variação de 0,5

a) verificar o tamanho do vetor;

b) Obter o 4^o, 7^o e 9^o elemento do vetor.

3) Calcule:

a) $|2^3 - 8^4|$

b) $log(2^3) + ln(4^2)$

c) $sen^2(\pi) + cos^2(\pi/2)$

4) Verificar a na ajuda o que significa `NaN` e `NA`.

3

ESTATÍSTICA DESCRITIVA

Neste capítulo vamos ver algumas funções do R para fazer uma análise descritiva de um conjunto de dados.

Para iniciar uma análise descritiva é importante verificar o tipo de variáveis do conjunto de dados. Para os exemplos deste capítulo considere o conjunto de dados da tabela A.1, em que são apresentadas as variáveis idade, sexo, grau de instrução, estado civil, atividade física, numero de refeições diárias, peso e altura de 29 funcionários de uma empresa.

O comando `summary` pode ser utilizado para uma analise preliminar, ele retorna:

- Variáveis qualitativas - frequências absolutas;
- Variáveis quantitativas - menor valor, primeiro quartil, mediana, média, terceiro quartil e maior valor.

```
> require(xlsx)
> dados=read.xlsx("peso.xls",1)
> attach(dados)
> summary(dados)
     IDADE        SEXO              GI               EC        AT
 Min.   :20.00   F:17   fundamental: 7   Casado    :13   Nao:18
 1st Qu.:24.00   M:12   medio      :14   Divorciado: 4   Sim:11
 Median :28.00          superior   : 8   Solteiro  :12
 Mean   :31.34
 3rd Qu.:39.00
 Max.   :51.00
      NFD              P               P1              A
 Min.   :2.000   Min.   : 51.5   Min.   :50.13   Min.   :1.560
 1st Qu.:2.000   1st Qu.: 66.8   1st Qu.:64.88   1st Qu.:1.650
 Median :3.000   Median : 75.0   Median :72.47   Median :1.690
 Mean   :3.207   Mean   : 76.6   Mean   :71.87   Mean   :1.690
 3rd Qu.:4.000   3rd Qu.: 86.7   3rd Qu.:80.36   3rd Qu.:1.730
 Max.   :5.000   Max.   :102.9   Max.   :95.40   Max.   :1.790
   NA.
 Mode:logical
 NA's:29
```

3.1 TABELAS

3.1.1 Dados Qualitativos

Frequentemente o primeiro passo da descrição de dados qualitativos é criar uma tabela de distribuição frequências. Para obter a frequências absolutas de variáveis qualitativas podemos utilizar a função `table`. E as demais frequências relativa e percentual podem ser obtidas, com operações a partir da frequência absoluta.

```
##Distribuição de frequência para variável Grau de Instrução.
> fa=table(GI) ##frequencia absoluta
> fa
GI
fundamental       medio    superior
          7          14           8
> fr=fa/sum(fa) ##frequencia relativa
> fr
GI
fundamental       medio    superior
  0.2413793   0.4827586   0.2758621
> fp=100*fr ##fp=frequencia percentual
> fp
GI
fundamental       medio    superior
   24.13793    48.27586    27.58621
> dist=cbind(fa,fr,fp) ##distribuição de frequências
> dist
            fa        fr       fp
fundamental  7 0.2413793 24.13793
medio       14 0.4827586 48.27586
superior     8 0.2758621 27.58621
```

A combinação entre duas variáveis com atributos distintos podem ser apresentadas em tabelas de dupla entrada, em que um atributo e disposto na coluna (vertical) e outro na linha (horizontal). Esta tabela é de muita utilidade, podendo resumir uma grande quantidade de dados, que podem ser apresentados ao mesmo tempo, para isso pode-se utilizar a função `table`.

```
> fa=table(SEXO,AT) ##frequencia absoluta
> fa
    AT
SEXO Nao Sim
   F   9   8
   M   9   3
> fr=fa/sum(fa) ##frequencia relativ
> fr
```

```
    AT
SEXO       Nao       Sim
   F 0.3103448 0.2758621
   M 0.3103448 0.1034483
> fp=100*fr ##fp=frequencia percentual
> fp
    AT
SEXO      Nao      Sim
   F 31.03448 27.58621
   M 31.03448 10.34483
```

Na função `table`, pode-se utilizar mais de duas variáveis.

```
> fa=table(SEXO,AT,EC,GI) ##frequencia absoluta
> fa
, , EC = Casado, GI = fundamental

    AT
SEXO Nao Sim
   F   1   2
   M   1   0

, , EC = Divorciado, GI = fundamental

    AT
SEXO Nao Sim
   F   1   1
   M   0   0

, , EC = Solteiro, GI = fundamental

    AT
SEXO Nao Sim
   F   0   0
   M   1   0

, , EC = Casado, GI = medio

    AT
SEXO Nao Sim
   F   1   3
   M   2   3
```

```
, , EC = Divorciado, GI = medio

    AT
SEXO Nao Sim
   F   0   0
   M   0   0

, , EC = Solteiro, GI = medio

    AT
SEXO Nao Sim
   F   3   0
   M   2   0

, , EC = Casado, GI = superior

    AT
SEXO Nao Sim
   F   0   0
   M   0   0

, , EC = Divorciado, GI = superior

    AT
SEXO Nao Sim
   F   1   0
   M   1   0

, , EC = Solteiro, GI = superior

    AT
SEXO Nao Sim
   F   2   2
   M   2   0
```

3.1.2 Dados Quantitativos

A tabela de distribuição de frequências de uma variável discreta é, em geral bastante semelhante à das variáveis qualitativas ordinais, pois os valores inteiros que a variável assume podem ser considerados como "categorias", ou "classes naturais". Assim, para fazer a distribuição de frequência pode-se utilizar a função `\table`.

```
> fa=table(NFD) ##frequencia absoluta
> fa
```

```
NFD
 2  3  4  5
10  8  6  5
> fr=fa/sum(fa) ##frequencia relativa
> fr
NFD
        2         3         4         5
0.3448276 0.2758621 0.2068966 0.1724138
> fp=100*fr ##fp=frequencia percentual
> fp
NFD
       2        3        4        5
34.48276 27.58621 20.68966 17.24138
> dist=cbind(fa,fr,fp) ##distribuição de frequências
> dist
  fa        fr       fp
2 10 0.3448276 34.48276
3  8 0.2758621 27.58621
4  6 0.2068966 20.68966
5  5 0.1724138 17.24138
```

A construção de tabelas de distribuição de frequências para variáveis quantitativas contínuas é feita agrupando os dados em classes e obtendo as frequências observadas em cada classe. Para fazer isso pode-se utilizar a função `fdt` pacote `fdth`. Esta função é definida da seguinte forma:

```
fdt(x, k, start, end, h, breaks=c("Sturges", "Scott", "FD"))
```

em que:

- x - são os dados que devem agrupados;
- k - numero de intervalos de classe;
- start - limite inferior da primeira classe;
- end - limite superior da ultima classe;
- h - amplitude da classe;
- breaks - caso não seja definido `k`, `start`, `end` e `h`, pode-se utilizar um método para isso.

```
> require(fdth)
> x=fdt(IDADE)
> x
 Class limits  f   rf rf(%) cf  cf(%)
  [19.8,25.1) 10 0.34 34.48 10  34.48
  [25.1,30.4)  8 0.28 27.59 18  62.07
```

```
  [30.4,35.7)  2 0.07  6.90 20  68.97
  [35.7,40.9)  3 0.10 10.34 23  79.31
  [40.9,46.2)  2 0.07  6.90 25  86.21
  [46.2,51.5)  4 0.14 13.79 29 100.00
> x1=fdt(IDADE,k=4)
> x1
 Class limits  f   rf rf(%) cf  cf(%)
  [19.8,27.7) 14 0.48 48.28 14  48.28
  [27.7,35.7)  6 0.21 20.69 20  68.97
  [35.7,43.6)  5 0.17 17.24 25  86.21
  [43.6,51.5)  4 0.14 13.79 29 100.00
> x2=fdt(IDADE,start=20,end=50,h=5)
> x2
 Class limits f   rf rf(%) cf cf(%)
      [20,25) 9 0.31 31.03  9 31.03
      [25,30) 8 0.28 27.59 17 58.62
      [30,35) 1 0.03  3.45 18 62.07
      [35,40) 5 0.17 17.24 23 79.31
      [40,45) 2 0.07  6.90 25 86.21
      [45,50) 3 0.10 10.34 28 96.55
> x3=fdt(IDADE,breaks="Scott")
> x3
 Class limits  f   rf rf(%) cf  cf(%)
  [19.8,30.4) 18 0.62 62.07 18  62.07
  [30.4,40.9)  5 0.17 17.24 23  79.31
  [40.9,51.5)  6 0.21 20.69 29 100.00
```

3.2 GRÁFICOS

O R possui muitas funções para produzir gráficos de qualidade. Os gráfico mais comuns são os gráficos de barras, pizza e de pontos (gráfico de dispersão).

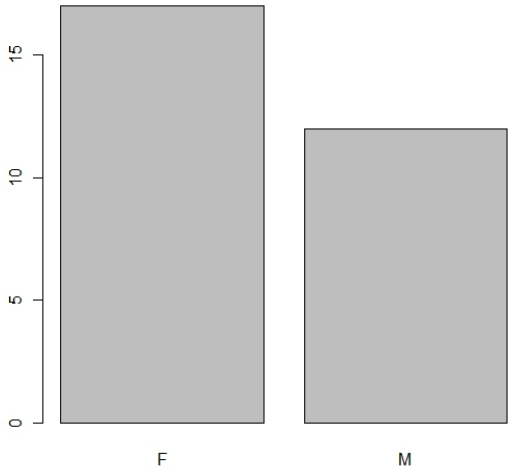
3.2.1 Gráfico de Barras

O Gráfico de Barras é um gráfico formado por retângulos horizontais (verticais) de larguras iguais, onde cada um deles representa a intensidade de uma modalidade ou atributo. O objetivo deste gráfico é de comparar grandezas e é recomendável para variáveis cujas categorias tenham designações extensas.

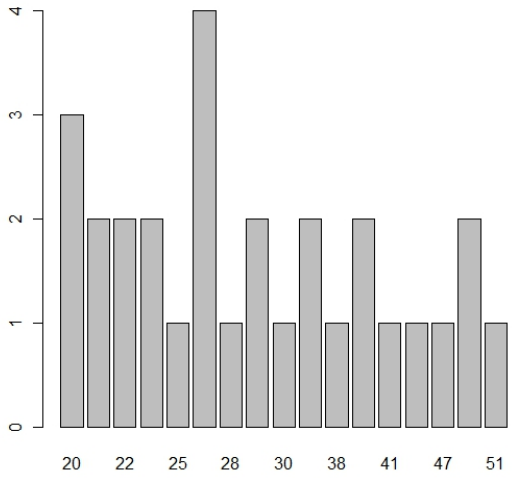
Para fazer um gráfico de barras de dados qualitativos pode-se utilizar duas funções `plot` e `barplot` que produzem resultados semelhantes. Já para variáveis quantitativas utiliza-se apenas a função `barplot`. Para utilizar a função `barplot` é necessário primeiro obter algumas das frequências.

```
##Grafico de barras
> fa=table(SEXO) #frequência absoluta
```

```
> barplot(fa)   #plota gráfico.
> plot(SEXO)  ##plota gráfico
```




```
> fa=table(IDADE) #frequência absoluta
> barplot(fa) #plota gráfico.
```



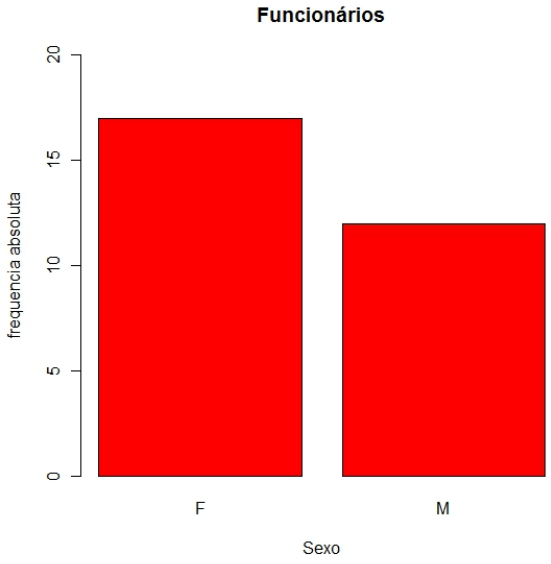
As duas funções possuem diversas opções. As principais são:

- `main` para adicionar titulo;
- `xlab` titulo para o eixo x;
- `ylab` titulo para o eixo y;

- `xlim` delimita os valores de x;
- `ylim` delimita os valores de y;
- `col` para definir cor;
- `horiz` para definir se as barras são horizontais ou verticais.

Por exemplo, colocar o título `Funcionários`, titulo para o eixo x `Sexo`, titulo para o eixo y `frequência absoluta`, os limites do eixo y entre 0 e 20 e mudar a cor da coluna para vermelho, utiliza-se

```
> plot(SEXO,main="Funcionários",xlab="Sexo",ylab="frequencia absoluta",ylim=c(0,20),
+ col="red")
```



Podemos plotar o gráfico para estado civil com barras horizontais.

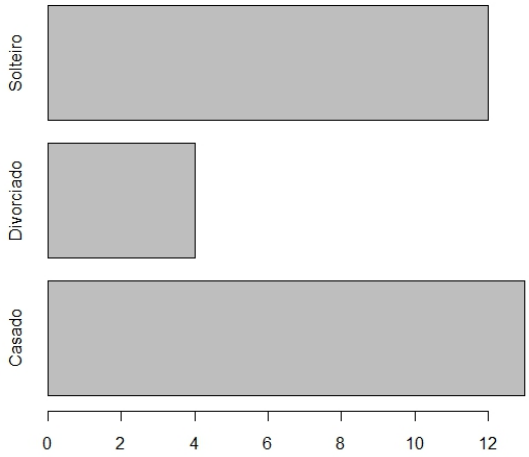
```
> plot(EC,horiz=TRUE)
```

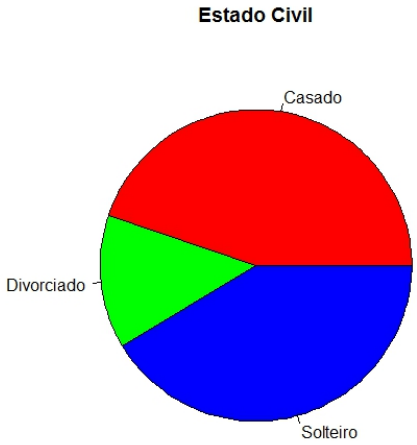
3.2.2 Gráfico de Pizza

O Gráfico de Pizza é um tipo de gráfico onde a variável em estudo é projetada num círculo, de raio arbitrário, dividido em setores com áreas proporcionais às frequências das suas categorias. São indicados quando se deseja comparar cada valor da série com o total. Recomenda-se seu uso para o caso em que o número de categorias não é grande e não obedecem a alguma ordem específica.

Para fazer um gráfico de pizza pode-se utilizar a função `pie`. Para utilizar a função `pie` é necessário primeiro obter algumas das frequências. AS opções `main` e `col`, podem ser utilizadas nesse gráfico.

```
> fa=table(EC)
> fr=fa/sum(fa)
> n=length(fr) ##contar quantos classes
> pie(fr,main="Estado Civil",col=rainbow(n))
```





3.2.3 Gráfico de Dispersão

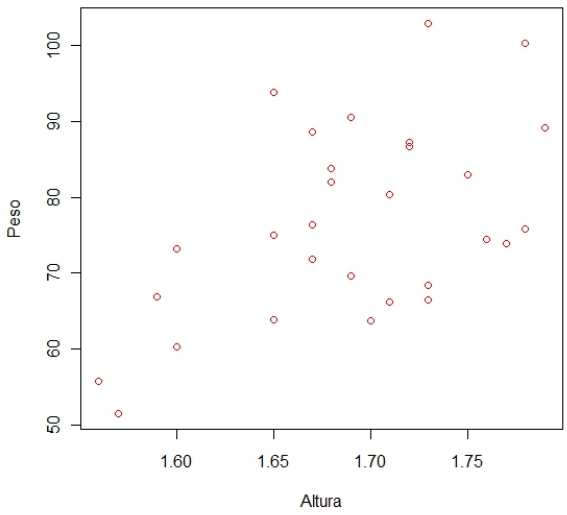
Para fazer um gráfico de dispersão utiliza-se a função `plot()` e precisa de apenas dois argumentos: o primeiro é o nome da variável do eixo `X`, o segundo é o da variável do eixo `Y`.

Por exemplo, plotar o gráfico de altura com peso.

```
> plot(A,P,xlab="Altura",ylab="Peso",col="red")
```

Também é fácil mudar os símbolos do gráfico por meio da opção `pch` , a opção default é a bolinha vazia (`pch=1`). Os símbolos podem assumir valores entre 1 e 25. Por exemplo, se quiser o seja triângulos, basta utilizar (`pch=2`)

```
> plot(A,P,xlab="Altura",ylab="Peso",col="red",pch=2)
```

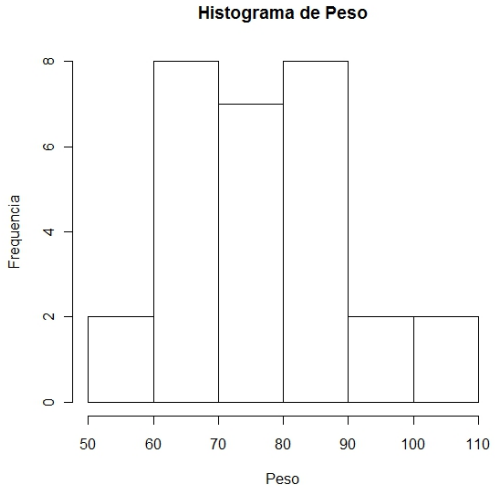


3.2.4 Histograma e Polígono de Frequência

O histograma é semelhante ao gráfico de barras verticais, no eixo vertical pode-se utilizar as frequências ou densidades de frequências e no eixo horizontal as classes. O polígono de frequências é um gráfico de linhas em que no eixo vertical pode-se utilizar as frequências ou densidades de frequências e no eixo horizontal o ponto médio de cada classe.

Para fazer um histograma utiliza-se a função `hist`. Podem ser utilizadas as mesmas opções da função `plot`.

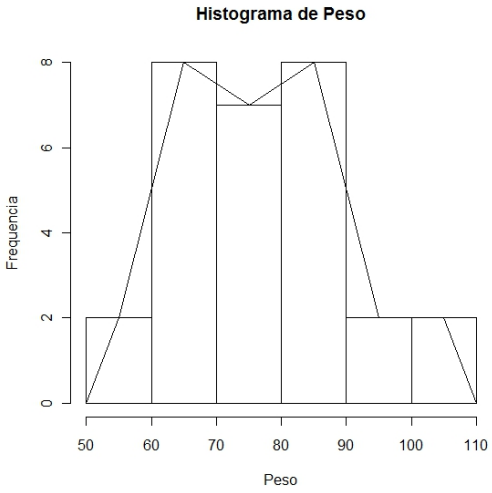
```
#histogram para variável peso.
hist(P,main="Histograma de Peso",xlab="Peso",ylab="Frequencia")
```

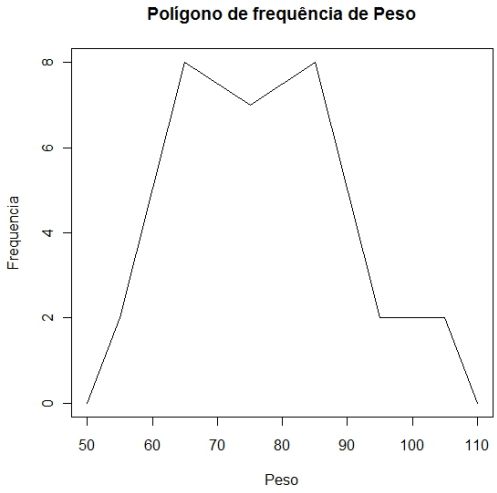


Para plotar o polígono de frequência, pode-se utilizar as funções abaixo.

```
h=hist(P,main="Histograma de Peso",xlab="Peso",ylab="Frequencia")
##polígono de frequência com histograma
```

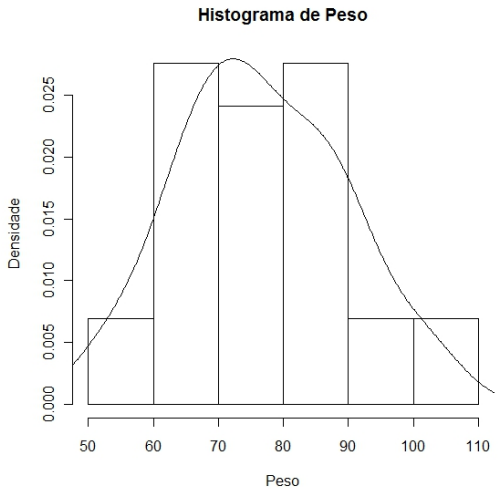
```
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0), type = "l")
 ##polígono de frequência
> plot(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0),
+ type = "l",main="Polígono de frequência de Peso",xlab="Peso",ylab="Frequencia")
```





Pode-se plotar o histograma pela densidade de frequência e utilizar a função `density` para plotar o polígono pela densidade de frequência

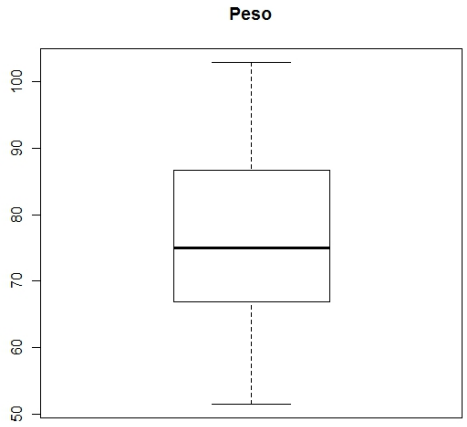
```
##histograma pela densidade de frequência
hist(P,freq=FALSE)
h1=density(P) ##obtendo a densidade dos dados
lines(h1) ##adicinar um gráfico de linhas
```



3.2.5 Boxplot

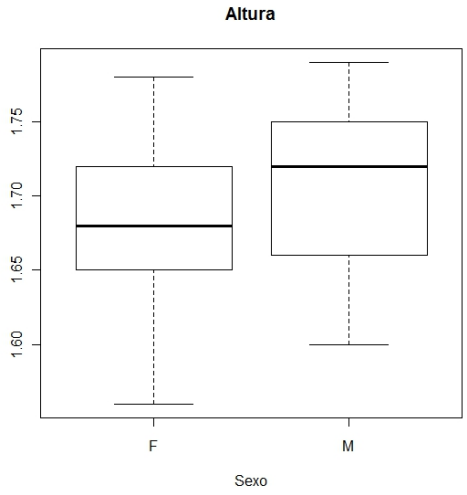
O gráfico Boxplot é uma análise gráfica que oferece a ideia da posição, dispersão, assimetria e dados discrepantes. Para construí-lo, basta utilizar a função `boxplot`.

```
#Boxplot para variável peso.
> boxplot(P,main="Peso")
```



O boxplot pode ser feito por sub-conjuntos ou categorias.

```
#Boxplot para variável altura dentro de cada sexo.
> boxplot(A~SEXO,main="Altura",xlab="Sexo")
```

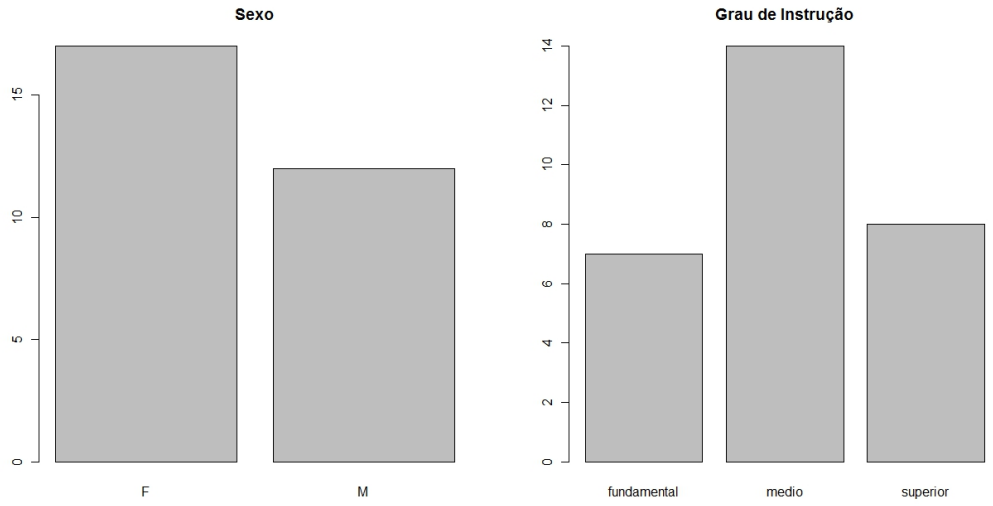


3.2.6 Dividir a janela dos gráficos

Em alguns casos pode-se plotar dois ou mais gráfico, assim é necessário dividir a janela de gráficos do R. A função `par` controla diversas características dos gráficos, para maiores informações consulte o help desta função par.

Para dividir a janela de gráficos o comando é `par(mfrow=c(nl,nc))`, em que `nl` indica o número de linhas e `nc` o número de colunas.

```
> par(mfrow=c(1,2)) ##uma linha e duas colunas
> plot(SEXO,main="Sexo")
> plot(GI,main="Grau de Instrução")
```



3.3 MEDIDAS DE POSIÇÃO

Medidas de Posição são medidas de tendência central, ou seja, representativas do valor central, ao redor do qual se agrupam a maioria dos valores.

3.3.1 Média

A média de uma população ou amostra é a soma de todos os elementos da população (amostra) dividida pelo número de elementos. Esta medida apresenta a mesma unidade dos dados. Para obter a média utiliza-se a função `mean`. Quando existem categorias, pode-se fazer as médias por categorias, para isso utiliza-se a função `tapply`.

```
> ##Média de Peso
> mean(P)
[1] 76.59655
> ##Média de Altura por Sexo
> tapply(A, SEXO, mean)
       F        M
1.678824 1.705000
```

3.3.2 Mediana

Num conjunto de dados ordenados, a mediana é o valor que deixa metade da frequência abaixo dele. Para obter a mediana utiliza-se a função `median`. Quando existem categorias, pode-se fazer as medianas por categorias, para isso utiliza-se a `tapply`.

```
> ##Mediana de Peso
> median(P)
[1] 75
> ##Mediana de Altura por Sexo
> tapply(A, SEXO, median)
   F    M
1.68 1.72
```

3.3.3 Moda

A moda de um conjunto de dados é o valor mais frequente. Para obter a moda de uma variável discreta utiliza-se a função `mfv` do pacote `modeest`.

```
> require(modeest)
> ##Moda de Idade
> mfv(IDADE)
[1] 27
> ##Moda de Idade por Sexo
> tapply(IDADE, SEXO,mfv)
$F
[1] 20 27
$M
[1] 22 39 49
```

Para obter a moda de uma variável contínua utiliza-se função `mlv` do pacote `modeest`. Esta função permite estimar a moda por diferentes métodos de Lientz, Chernoff, Venter, Grenander, HSM, HRM, Kernel de Parzen, Tsybakov, Asselin de Beauville e Vieu.

```
> ##moda de Peso
> mlv(P)
Mode (most likely value): 69.71333
Bickel's modal skewness: 0.3103448
Call: mlv.default(x = P)
Warning message:
In mlv.default(P) :
  argument 'method' is missing. Data are supposed to be continuous. Default method 'shorth' is
> mlv(P, method = "lientz", bw = 0.2)
Mode (most likely value): 69
Bickel's modal skewness: 0.3793103
Call: mlv.default(x = P, bw = 0.2, method = "lientz")
> mlv(P, method = "naive", bw = 1/3)
Mode (most likely value): 66.5
Bickel's modal skewness: 0.5517241
Call: mlv.default(x = P, bw = 1/3, method = "naive")
> mlv(P, method = "venter")
```

```
Mode (most likely value): 68.25
Bickel's modal skewness: 0.4482759
Call: mlv.default(x = P, method = "venter")
> mlv(P, method = "grenander", p=4)
Mode (most likely value): 75.40257
Bickel's modal skewness: -0.03448276
Call: mlv.default(x = P, method = "grenander", p = 4)
> mlv(P, method = "hrm", bw = 0.3)
Mode (most likely value): 75.4
Bickel's modal skewness: -0.03448276
Call: mlv.default(x = P, bw = 0.3, method = "hrm")
> mlv(P, method = "hsm")
Mode (most likely value): 60.2
Bickel's modal skewness: 0.8275862
Call: mlv.default(x = P, method = "hsm")
> mlv(P, method = "parzen")
Mode (most likely value): 74.02429
Bickel's modal skewness: 0.1034483
Call: mlv.default(x = P, method = "parzen")
> mlv(P, method = "tsybakov")
Mode (most likely value): 69.58456
Bickel's modal skewness: 0.3793103
Call: mlv.default(x = P, method = "tsybakov")
> mlv(P, method = "asselin")
Mode (most likely value): 68.4
Bickel's modal skewness: 0.4137931
Call: mlv.default(x = P, method = "asselin")
> mlv(P, method = "vieu")
Mode (most likely value): 74.02429
Bickel's modal skewness: 0.1034483
Call: mlv.default(x = P, method = "vieu")
```

3.4 MEDIDA DE DISPERSÃO

As medidas de posição são importantes para caracterizar um conjunto de dados, mas não são suficientes para caracterizar completamente a distribuição dos dados. Para isso é necessário obter as medidas de dispersão, que medem a sua variabilidade.

3.4.1 Amplitude

Amplitude é a diferença entre o maior e o menor valor da amostra. Essa medida é bastante simples, mas expressa a variabilidade dos dados. Para obter a amplitude utiliza-se a função `range` e função `diff`.

```
> ##Amplitude de Altura
> range(A)
[1] 1.56 1.79
f(range(A))
[1] 0.23
> X=tapply(A, SEXO,range)
> X
$F
[1] 1.56 1.78

$M
[1] 1.60 1.79

> diff(X$F) ##Ampliture Altura Mulheres
[1] 0.22
> diff(X$M) ##Ampliture Altura Homens
[1] 0.19
```

3.4.2 Variância e Desvio Padrão

O variância e o desvio padrão são medidas de dispersão ou variabilidade, a opção do uso de um ou outro, depende da finalidade da informação. A variância é baseada pela quadrado dos desvios dos dados em relação à média. O desvio padrão é a raiz quadrada positiva da variância. Para obter a variância utiliza-se a função `var` e o desvio padrão a função `sd`.

```
> ##Variância de Altura
> var(A)
[1] 0.004010591
> ##Variância de Altura por Sexo
> tapply(A, SEXO,var)
          F           M
0.003886029 0.004118182
> ##Desvio Padrão de Altura
> sd(A)
[1] 0.06332923
> ##Desvio Padrão de Altura por Sexo
> tapply(A, SEXO,sd)
         F          M
0.06233803 0.06417306
```

3.5 EXERCÍCIOS

1) Utilizando o conjunto de dados da tabela A.2 obter:

a) Tabela de distribuição de frequência para cada uma das variáveis.

b) Tabela de distribuição de renda, considerando os intervalos da primeira classe iniciando em 0 e encenrrando em 12 e variando de 2 em 2.

c) Tabela cruzada das variáveis bairro e grau de instrução.

d) Representação gráfica da distribuição de todas as variáveis.

e) As medidas de posição e dispersão para as variáveis número de filhos e renda.

f) As medidas de posição e dispersão para as variáveis número de filhos e renda por bairro.

g) Gráfico boxplot as variáveis número de filhos e renda por bairro.

2) Um questionário foi aplicado aos funcionários de uma empresa fornecendo os dados apresentados na tabela

Funcionário	Curso (completo)	Numero Filhos	Funcionário	Curso (completo)	Numero Filhos
1	Superior	1	11	Superior	4
2	Superior	0	12	Superior	2
3	Médio	1	13	Médio	3
4	Médio	0	14	Fundamental	1
5	Médio	3	15	Fundamental	1
6	Médio	2	16	Superior	0
7	Médio	0	17	Superior	4
8	Médio	2	18	Fundamental	1
9	Fundamental	3	19	Fundamental	4

a) Classifique cada uma das variáveis.

b) Obtenha a distribuição de frequência para cada uma das variáveis.

c) Obtenha a distribuição de frequência absoluta e uma representação gráfica das duas variáveis conjuntamente.

3) Os dados apresentados a seguir referem-se ao tempo que uma bateria levou para apresentar uma falha grave, em anos, ou seja, para descarregar completamente. A amostra de tamanho n = 25 foi obtida com objetivo de caracterizar a robustez das baterias e é dada por:

1,6	4,1	4,6	5,8	7,0
2,2	4,1	5,0	6,2	7,2
2,5	4,2	5,0	6,3	7,6
2,7	4,4	5,0	6,6	8,5
3,7	4,4	5,3	6,9	10,7

a) Obter a média, mediana e moda;

b) Obter o 1º e 3º Quartil;

d) Você identifica algum valor discrepante dentre os que foram observados? Se sim, remova-o(s) e refaça os itens (a) e (b). Comente as diferenças encontradas.

4) A idade dos 20 ingressantes num certo curso de uma universidade foi o seguinte: 22, 22, 22, 22, 23, 23, 24, 24, 24, 24, 25, 25, 26, 26, 26, 26, 27, 28, 35, 40.

a) calcule a média, a moda e a mediana;

b) calcule os quartis.

c) faça um boxplot destes dados;

d) idades atípicas parecem ter ocorrido este ano. Identifique-as e após sua retirada refaça os itens anteriores. Comente as diferenças encontradas.

4

INFERÊNCIA ESTATÍSTICA

4.1 DISTRIBUIÇÕES AMOSTRAIS

4.1.1 Distribuição Amostral da Média $(\overline{X})$

4.1.1.1 Distribuição Normal

Se o tamanho n da amostra for suficientemente grande (≥ 30 elementos), então a média de uma amostra aleatória retirada de uma população terá uma distribuição aproximadamente normal, e seus valores de média e desvios padrão estão relacionados com média $\mu_{\overline{X}} = \mu$ e variância $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.

A distribuição Normal corresponde a mais importante distribuição de variáveis aleatórias contínuas, em razão da sua enorme aplicação nos mais variados campos do conhecimento. A distribuição normal possui dois parâmetros característicos μ e σ^2 que correspondem respectivamente a média e a variância da distribuição.

A distribuição Normal com média $\mu = 0$ e variância $\sigma^2 = 1$ é conhecida como distribuição Normal reduzida ou padronizada. Uma variável aleatória com essa distribuição geralmente é simbolizada pela letra Z.

No R, a distribuição normal pode ser utilizada por meio das funções abaixo, em todas elas pode ser definir a média (`mean`) e o desvio padrão (`sd`):

- `dnorm(x,mean,sd)` - calcula a densidade de probabilidade no ponto x;
- `pnorm(x,mean,sd)` - calcula a função de probabilidade acumulada no ponto x;
- `qnorm(p,mean,sd)` - calcula o quantil correspondente a uma dada probabilidade p;
- `rnomr(n,mean,sd)` - gera uma amostra da distribuição normal de tamanho n;

```
> ##densidade no ponto 1 de uma normal padrão
> dnorm(1,2,3)
[1] 0.1257944
> ##probabilidade x<1 de uma normal padrão
> pnorm(1,2,3)
[1] 0.3694413
> ##quantil de deixa 95% de probabilidade abaixo dele, de uma normal padrão
> qnorm(0.95,2,3)
[1] 6.934561
```

```
> ##gerar uma amostra de 10 elementos, de uma normal padrão
> rnorm(10,2,3)
 [1] -2.01936612  6.38652439 -0.25540386 -0.63521111  6.02160033 -0.04164142
 [7] -4.39936955  4.06665189  0.43857050  0.67846715
> ##densidade no ponto 1,  com média 2 e desvio padrão 3
> dnorm(1,2,3)
[1] 0.1257944
> ##probabilidade x<1, com média 2 e desvio padrão 3
> pnorm(1,2,3)
[1] 0.3694413
> ##quantil de deixa 95% de probabilidade abaixo dele, com média 2 e desvio padrão 3
> qnorm(0.95,2,3)
[1] 6.934561
> ##gerar uma amostra de 10 elementos, com média 2 e desvio padrão 3
> rnorm(10,2,3)
 [1]  9.022310  2.414009  4.557748  7.948228  5.428807  1.533121  8.410468
 [8]  0.959984 -7.135576 -2.158707
```

4.1.1.2 Distribuição t de student

A distribuição t de Student aparece naturalmente no problema de se determinar a média de uma população (que segue a distribuição normal) a partir de uma amostra. Neste problema, não se sabe qual é a média ou o desvio padrão da população, mas ela deve ser normal. Na distribuição t existe uma curva para cada tamanho de amostra (n) e o valor $\nu = n - 1$ (número de graus de liberdade) e todas curvas tem máximo em $t = 0$. A medida que n cresce a distribuição t se aproxima da normal padrão z;

No R, a distribuição t pode ser utilizada por meio das funções abaixo, em todas elas pode ser definir o grau de liberdade `(df)`.

- `dt(x,df)` - calcula a densidade de probabilidade no ponto x;
- `pt(x,df)` - calcula a função de probabilidade acumulada no ponto x;
- `qt(p,df)` - calcula o quantil correspondente a uma dada probabilidade p;
- `rt(n,df)` - gera uma amostra da distribuição t de tamanho n;

```
>##densidade no ponto 2, com grau de liberdade 12
> dt(1,12)
[1] 0.2322303
> ##probabilidade x<2, com grau de liberdade 12
> pt(2,12)
[1] 0.9656725
> ##quantil de deixa 95% de probabilidade abaixo dele, com grau de liberdade 12
> qt(0.95,12)
[1] 1.782288
```

```
> ##gerar uma amostra de 10 elementos, com grau de liberdade 12
> rt(10,12)
 [1]  0.10838734 -0.44992792 -0.35572801  0.01100810 -0.49214326  0.19236174
 [7]  0.26643412 -1.00732027  1.53588875  0.18353544
```

4.1.2 Distribuição amostral para proporção

Quando n é grande $(n > 30)$, a proporção amostral $\hat{p}$ de sucessos em n ensaios de Bernoulli tem distribuição aproximadamente normal com média $\mu = p$ e variância $\sigma^2 = \frac{pq}{n}$, e assim podemos utilizar as funções `dnorm`, `pnorm`, `qnorm` e `rnomr`.

4.1.3 Distribuição Amostral da Variância

4.1.3.1 Distribuição Qui-Quadrado

Retirando-se uma amostra de n elementos de uma população normal com media (μ) e variância (σ^2), então, a distribuição amostral da variância amostral segue uma distribuição de χ^2_{n-1} (qui-quadrado) com n-1 graus de liberdade. Na distribuição χ^2 existe uma curva para cada tamanho de amostra (n) e todas curvas tem inicio em $\chi^2 = 0$.

No R, a distribuição χ^2 pode ser utilizada por meio das funções abaixo, em todas elas pode ser definir o grau de liberdade

- `dchisq(x,df)` - calcula a densidade de probabilidade no ponto x;
- `pchisq(x,df)` - calcula a função de probabilidade acumulada no ponto x;
- `qchisq(p,df)` - calcula o quantil correspondente a uma dada probabilidade p;
- `rchisq(n,df)` - gera uma amostra da distribuição χ^2 de tamanho n;

```
> ##densidade no ponto 5, com grau de liberdade 15
> dchisq(1,12)
[1] 7.897535e-05
> ##probabilidade x<5, com grau de liberdade 15
> pchisq(2,12)
[1] 0.0005941848
> ##quantil de deixa 95% de probabilidade abaixo dele, com grau de liberdade 15
> qchisq(0.95,15)
[1] 24.99579
> ##gerar uma amostra de 10 elementos, com grau de liberdade 15
> rchisq(10,15)
 [1]  7.950830 20.480046  8.987797 14.177976 14.192847 14.740364 11.464687
 [8]  7.752778 22.090825 32.352168
```

4.1.3.2 Distribuição F

A distribuição F está entre as distribuições de probabilidade mais importantes na estatística, tem maior destaque na área de experimentação agrícola. Essa distribuição é definida pela variável resultante

da razão duas variâncias. Neste distribução é necessário observar dois graus de liberdade $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$, o primeiro associado à variância amostral do numerador, e o segundo associado à variância amostral do denominador.

No R, a distribuição F pode ser utilizada por meio das funções abaixo, em todas elas pode ser definir o grau de liberdade

- `df(x,df1,df2)` - calcula a densidade de probabilidade no ponto x;
- `pf(x,df1,df2)` - calcula a função de probabilidade acumulada no ponto x;
- `qf(p,df1,df2)` - calcula o quantil correspondente a uma dada probabilidade p;
- `rf(n,df1,df2)` - gera uma amostra da distribuição F de tamanho n;

```
> ##densidade no ponto 5, com grau de liberdade 12 e 20
> df(5,12,20)
[1] 0.001019420
> ##probabilidade x<5, com grau de liberdade 12 e 20
> pf(5,12,20)
[1] 0.999205
> ##quantil de deixa 95% de probabilidade abaixo dele, com grau de liberdade 12 e 20
> qf(0.95,12,20)
[1] 2.277581
> ##gerar uma amostra de 10 elementos,  com grau de liberdade 12 e 20
> rchisq(10,12,20)
 [1] 40.71985 31.64577 27.40006 43.09402 25.09778 20.52160 33.08932 33.52143
 [9] 19.65764 41.55187
```

4.2 INTERVALO DE CONFIANÇA E TESTE DE HIPÓTESE

Uma maneira de se calcular um estimativa de um parâmetro desconhecido, é construir um intervalo de confiança $[a,b]$ para esse parâmetro com uma probabilidade de $1-\alpha$ (nível de confiança) de que o intervalo contenha o verdadeiro parâmetro, usando as distribuições de amostragem podemos obter expressões do tipo:

$$P(a \leq \mu \leq b) = 1 - \alpha$$

Dessa maneira α será o nível de significância, isto é, o erro que se estará cometendo ao afirmar que o parâmetro está entre o limite inferior e o superior calculado.

O teste de hipótese é uma metodologia estatística que permite tomar decisão sobre uma ou mais populações baseando no conhecimento de informações da amostra. Os processos que habilitam a decidir se aceitam ou rejeitam as hipóteses formuladas, ou determinar se a amostra observada difere, de modo significativo, dos resultados esperados, são denominados de Testes de Hipóteses ou Testes de Significância.

Ao testar uma hipótese estabelecida, a probabilidade máxima com a qual se sujeitaria a correr o risco de um erro do tipo I é denominada de Nível de Significância do Teste e é representada por α.

Os testes de hipóteses são formulados com uma hipótese nula (H_0) e uma hipótese alternativa (H_a) e estabelecendo o nível de significância do Teste que é representada por α. A partir da formulação de (H_0) e (H_a), podemos definir se teste de hipótese é unilateral ou bilateral.

Consideremos θ o parâmetro estudado e θ_0 valor inicialmente suposto para. Podemos formular as seguintes hipóteses:

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases} \quad \text{Teste Bilateral}$$

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta > \theta_0 \end{cases} \quad \text{Teste Unilateral}$$

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta < \theta_0 \end{cases} \quad \text{Teste Unilateral}$$

4.2.1 Proporção

Consideremos uma população cujos elementos podem ser classificados em dois tipos: Sucesso e Insucesso. Pretende-se estimar a proporção p de sucessos na população.

Dada uma amostra de tamanho n, uma estimativa pontual de p da proporção de sucessos é dada por

$$\hat{p} = \frac{x}{n}$$

.

Fixando uma probabilidade de confiança $(1-\alpha)$, o intervalo de confiança para uma proporção pode ser obtido da seguinte forma:

$$IC_{1-\alpha}(p) = \left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}; \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$$

em que: $z_{\frac{\alpha}{2}}$ é o valor da curva normal padrão acima do qual encontramos uma área de $\frac{\alpha}{2}$.

O intervalo de confiança exato para as proporções binomiais deve ser utilizado principalmente se n for pequeno e se p se afastar $1/2$. Este intervalo é baseado na relação da binomial com a distribuição F. O intervalo de confiança exato para as proporções binomiais é dado por:

$$IC_{1-\alpha}(p) = \left[\frac{1}{1 + \dfrac{(n-x+1)F_{\left(\frac{\alpha}{2},2(n-x+1),2x\right)}}{x}}; \frac{1}{1 + \dfrac{n-x}{(x+1)F_{\left(\frac{\alpha}{2},2(x+1),2(n-x)\right)}}}\right]$$

Para obter o intervalo de confiança no R utilizando a distribuição normal é necessário fazer as contas passo a passo.

```
> ##Intervalo de  95%  de confiança para proporção de homens
> fa=table(SEXO)
> fr=fa/sum(fa)
> fr
SEXO
```

```
        F         M
0.5862069 0.4137931
> p=fr[2] ##proporção de homens
> p
        M
0.4137931
> q=1-p
> q
        M
0.5862069
> n=sum(fa)
> p.ic=p+qnorm(c(0.025,0.975))*sqrt(q*p/n)
> p.ic
[1] 0.2345402 0.5930460
```

Utilizando a função `binom.test`, é possível fazer um teste de hipótese e o intervalo de confiança exato para proporção. Esta função apresenta os seguintes argumento:

```
binom.tes(x, n,p = 0.5,alternative=c("two.sided","less","greater"),conf.level = 0.95)
```

sendo `x` o número de sucessos, `n` tamanho da amostra, `alternative` o tipo de hipótese alternativa, `conf.level` o nível de confiança. Caso não seja declarado a opções `alternative` e `conf.level` é considerado teste bilateral e nível de confiança de 95%.

Vamos testar se a proporção de homens pode ser considerada igual a 50% e com hipótese alternativa se ela é diferente de 50%, considerando uma significância de 0,05.

$$\begin{cases} H_0: & p = 0,5 \\ H_1: & p \neq 0,5 \end{cases}$$

```
> binom.test(12,29, p = 0.5, conf.level = 0.95)

        Exact binomial test

data:  12 and 29
number of successes = 12, number of trials = 29, p-value = 0.4583
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2352402 0.6106372
sample estimates:
probability of success
             0.4137931
```

Considerando $\alpha = 0,05$, temos `p-value = 0.4583`, ou seja, `p-value>0,05` então não rejeita-se H_0. Assim a proporção de homens pode ser considerada igual a 50%, ao nível de 5% de significância. E o intervalo de 95%dde confiança exata seria de $[0,23;0,61]$.

Se mudar a hipótese alternativa testar para proporção de homens menor que 50%, considerando uma significância de 0,05, temos:

$$\begin{cases} H_0: & p = 0,5 \\ H_1: & p < 0,5 \end{cases}$$

```
> binom.test(12,29, p = 0.5, conf.level = 0.95, alternative ="less")

        Exact binomial test

data:  12 and 29
number of successes = 12, number of trials = 29, p-value = 0.2291
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.5825361
sample estimates:
probability of success
             0.4137931
```

A interpretação do teste é a mesma `p-value>0,05` então não rejeita-se H_0. Neste caso, não podemos utilizar apresentado no final do teste.

4.2.2 Normalidade

Muitos teste de hipótese pressupõe que os dados tenham distribuição normal, assim antes de aplicar tais teste é necessário verificar a normalidade. Existem várias funções do R para isto, uma delas é a função `shapiro.test`.

$$\begin{cases} H_0: & \text{Dados são normais} \\ H_1: & \text{Dados não são normais} \end{cases}$$

```
> shapiro.test(IDADE)

        Shapiro-Wilk normality test

data:  IDADE
W = 0.8974, p-value = 0.008515

> shapiro.test(P)

        Shapiro-Wilk normality test

data:  P
W = 0.9869, p-value = 0.9684

> shapiro.test(P1)
```

```
        Shapiro-Wilk normality test

data:  P1
W = 0.9841, p-value = 0.9277

> shapiro.test(A)

        Shapiro-Wilk normality test

data:  A
W = 0.9593, p-value = 0.3164
```

Considerando $\alpha = 0,05$, temos que `p-value>0,05` somente para variável Idade `(p-value<0,05)` rejeita-se H_0 ou seja os dados não são normais.

4.2.3 Média

Como apresentado anteriormente, $\overline{X}$ tem distribuição normal de média μ e variância $\frac{\sigma^2}{n}$, assim um intervalo de $(1-\alpha)$ de confiança para μ será dado por:

$$IC_{1-\alpha}(\mu) = \left[\overline{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \overline{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Para aplicar este intervalo de confiança é necessário que se conheça a variância populacional (σ^2). Na prática quando não se conhece a média $\overline{X}$ também não se conhece a variância populacional, nesse caso utilizamos o intervalo de confiança:

$$IC_{1-\alpha}(\mu) = \left[\overline{X} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}; \overline{X} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

Para obter o intervalo de confiança no R utilizando a distribuição normal é necessário fazer as contas passo a passo.

Supondo que o desvio padrão populacional da altura seja 0,5.

```
> ##Intervalo de 95% confiança para altura
> XB=mean(A)
> XB
[1] 1.689655
> Sigma=0.5
> n=length(A)
> n
[1] 29
> XB.ic=XB+qnorm(c(0.025,0.975))*(Sigma/sqrt(n))
> XB.ic
[1] 1.507677 1.871633
```

Utilizando a função `t.test`, é possível fazer um teste de hipótese e o intervalo de confiança para média. Esta função é apropriada somente para o caso de dados normais e apresenta os seguintes argumento:

```
t.tes(x, mu = 0,alternative=c("two.sided","less","greater"),conf.level = 0.95)
```

sendo `x` vetor de dados, `mu` valor da média a ser testado, `alternative` o tipo de hipótese alternativa, `conf.level` o nível de confiança. Caso não seja declarado a opções `alternative` e `conf.level` é considerado teste bilateral e nível de confiança de 95%.

Testar se média de altura é igual a 1,72 e com hipótese alternativa se ela é diferente de 1,72, considerando uma significância de 0,05.

$$\begin{cases} H_0: & mu = 1,72 \\ H_1: & mu \neq 1,72 \end{cases}$$

```
> t.test(A,mu=1.72)

        One Sample t-test

data:  A
t = -2.5804, df = 28, p-value = 0.01541
alternative hypothesis: true mean is not equal to 1.72
95 percent confidence interval:
 1.665566 1.713744
sample estimates:
mean of x
 1.689655
```

Considerando $\alpha = 0,05$, temos que `p-value<0,05` então rejeita-se H_0. Assim idade média pode ser considerada diferente de 1,72 anos, ao nível de 5% de significância. E o intervalo de 95% de confiança seria de $[1,66; 1,71]$.

Se a distribuição dos dados não for normal pode-se utilizar o teste não-parámetrico de Wilcoxon por meio da função `wilcox.test`. Esta função apresenta argumentos semelhantes ao da função `t.test`

```
wilcox.tes(x, mu=0,alternative=c("two.sided","less","greater"),conf.level = 0.95)
```

Como a variável Idade foi considerada não normal, podemos utilizar a função `wilcox.test` para teste de hipótese para média.

Testar se média de idade é igual a 36 anos e com hipótese alternativa se ela é diferente de 36, considerando uma significância de 0,05.

$$\begin{cases} H_0: & mu = 36 \\ H_1: & mu \neq 36 \end{cases}$$

```
> wilcox.test(IDADE,mu=36)

        Wilcoxon signed rank test with continuity correction
```

```
data:  IDADE
V = 109.5, p-value = 0.02000
alternative hypothesis: true location is not equal to 36
```

Considerando $\alpha = 0,05$, temos que `p-value<0,05` então rejeita-se H_0.Assim idade média pode ser considerada diferente de 36 anos, ao nível de 5% de significância.

4.2.4 Diferenças de Médias

A funções `t.test` e `wilcox.test` também pode ser utilizada para comparação entre médias dois grupo, para isso utiliza-se na função

```
t.tes(x,y, mu = 0,alternative=c("two.sided","less","greater"),
      var.equal = FALSE,paired = FALSE,conf.level = 0.95)
wilcox.tes(x,y, mu = 0,alternative=c("two.sided","less","greater"),
                  paired = FALSE,conf.level = 0.95)
```

sendo `x` e `y` vetor de dados de cada grupo, `mu` valor da diferença da média dos grupos a ser testado, `alternative` o tipo de hipótese alternativa, `conf.level` o nível de confiança, `var.equal` se a variância entre os grupos são iguais, `paired` se o dados são pareados ou não.

Os dados pareado quando é medido a mesma variável sobre o mesmo individuo após um tempo ou um tratamento. No caso da variável peso, temos um exemplo de dados pareados, nesse caso poderíamos estar interessados em ver teve diferença no peso antes e depois do estudo. Nesse caso teríamos um teste de hipótese da seguinte forma.

$$\begin{cases} H_0: & \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu \neq \mu_2 \Rightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$$

```
> t.test(P,P1,m=0,paired=TRUE)

        Paired t-test

data:  P and P1
t = 3.9659, df = 28, p-value = 0.0004604
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.286957 7.173043
sample estimates:
mean of the differences
                   4.73
> mean(P)
[1] 76.59655
> mean(P1)
[1] 71.86655
```

Considerando que $\alpha = 0,05$, temos que `p-value<0,05` então rejeita-se H_0. Assim o peso médio depois do estudo é inferior ao inicio do estudo, ao nível de 5% de significância.

Vamos verificar se a peso inicial de homens e mulheres são iguais. Antes de aplicar o teste t, é necessário verificar se as variâncias entre os grupos são iguais.

$$\begin{cases} H_0: & \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu \neq \mu_2 \Rightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$$

```
> Grupos
$F
 [1] 51.5 55.7 63.8 66.8 69.6 71.8 74.5 75.0 75.8 76.3 80.3 82.0 83.0 83.8 86.7
[16] 87.2 90.6

$M
 [1]  60.2  63.7  66.2  66.5  68.4  73.2  73.9  88.6  89.2  93.8 100.3 102.9

> ##valores de idade para mulheres
> x=Grupos$F
> x
 [1] 51.5 55.7 63.8 66.8 69.6 71.8 74.5 75.0 75.8 76.3 80.3 82.0 83.0 83.8 86.7
[16] 87.2 90.6
> ##valores de idade para homens
> y=Grupos$M
> y
 [1]  60.2  63.7  66.2  66.5  68.4  73.2  73.9  88.6  89.2  93.8 100.3 102.9
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 0.5203, num df = 16, denom df = 11, p-value = 0.2281
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1574663 1.5264942
sample estimates:
ratio of variances
         0.5203309
```

Considerando que $\alpha = 0,05$, temos que `p-value>0,05` então não rejeita-se H_0. Assim as variâncias são iguais.

```
> t.test(x,y,m=0,var.equal=TRUE)

        Two Sample t-test
```

```
data:  x and y
t = -0.8169, df = 27, p-value = 0.4211
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.84925   5.96200
sample estimates:
mean of x mean of y
 74.96471  78.90833
```

Considerando que $\alpha = 0,05$, temos que `p-value>0,05` então não rejeita-se H_0. Assim o peso inicial médio de homens e mulheres são iguais, ao nível de 5% de significância.

```
> ##Obter os valores da idade por sexo
> Grupos=tapply(IDADE,SEXO,sort)
> Grupos
$F
 [1] 20 20 20 21 21 24 24 25 27 27 27 29 35 38 41 43 47

$M
 [1] 22 22 27 28 29 30 35 39 39 49 49 51

> ##valores de idade para mulheres
> x=Grupos$F
> x
 [1] 20 20 20 21 21 24 24 25 27 27 27 29 35 38 41 43 47
> ##valores de idade para homens
> y=Grupos$M
> y
 [1] 22 22 27 28 29 30 35 39 39 49 49 51
> wilcox.test(x,y,m=0)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 58.5, p-value = 0.05626
alternative hypothesis: true location shift is not equal to 0
```

Considerando que $\alpha = 0,05$, temos que `p-value>0,05` então não rejeita-se H_0. Assim a idade médio de homens e mulheres são iguais, ao nível de 5% de significância.

4.3 EXERCÍCIOS

1) Gerar amostras de uma distribuição normal padrão e t-student de tamanhos $n = 10$, $n = 20$, $n = 30$, $n = 50$ e $n = 100$. Utiliza a função `density` e plote os gráficos e compare-os.

2) Utilizando os dados da tabela A.3

a) Calcule as medidas de posição e dispersão para Calcio e Zinco dentro de cada grupo;

b) Obtenha os intervalos intervalos de 95% confiança para Calcio e Zinco dentro de cada grupo;

c) Faça um teste de hipótese para comparar se a concentração de Calcio e Zinco são iguais entre os grupos.

3) Utilizando os dados da tabela A.2

a) Obtenha o intervalo de 95% confiança para a proporção de analfabeto; ensino fundamental completo e ensino médio completo.

b) Faça um teste de hipótese para comparar a renda de analfabetos e pessoas com o ensino médio;

c) Faça um teste de hipótese para comparar o numero de filhos de analfabetos e de pessoas com o ensino fundamental.

4) Para estimar a diferença de tempos médios de vida (em anos) entre fumantes e não fumantes, foram recolhidas obtidos as seguintes observações

Fumantes											
52,4	55,0	55,2	55,2	55,5	56,2	57,0	57,4	58,3	58,4	59,2	59,3
59,6	59,7	60,0	60,5	60,6	61,2	61,6	61,9	62,1	62,2	62,4	62,7
63,5	64,1	64,7	64,8	64,9	65,0	65,4	66,0	66,9	69,1	69,2	69,8
Não Fumantes											
63,8	65,7	66,2	66,2	66,2	66,8	67,5	67,7	67,9	68,0	68,1	68,3
68,6	68,6	68,7	68,8	68,8	69,2	69,3	69,4	69,5	70,1	70,1	70,2
70,2	70,3	70,4	70,7	70,8	70,8	71,0	71,4	71,5	71,6	72,7	72,7
72,9	73,3	73,3	73,9	74,1	75,8	75,9	77,5				

a) Obtenha o intervalo de confiança a 95% para tempo médio de vida fumantes e não fumantes

b) Por meio do intervalo de confiança é possível dizer que não fumantes vivem mais que fumantes?

c) Obtenha o intervalo de confiança a 95% para proporção de fumantes e não fumantes

d) Por meio do intervalo de confiança é possível dizer que a proporção de fumantes e não fumantes são iguais?

5) Os pesos dos bois de uma fazenda na idade de abate apresentam os seguintes dados

545,92	532,29	472,92	533,68	479,48	476,38	497,51	504,02	543,66	522,27
502,90	532,45	538,84	548,86	506,67	508,43	520,23	466,14	521,67	482,56
429,72	476,62	509,82	501,61	477,28	530,99	485,24	515,24	544,60	460,53

a) Obtenha a distribuição de frequência;

b) Calcule a média, mediana e moda.

c) Obtenha o gráfico de boxplot;

0,008	0,045	0,024	0,034	0,027	0,032	0,018	0,032	0,012	0,008	0,004
0,025	0,008	0,031	0,009	0,014	0,004	0,033	0,032	0,004	0,023	0,028

c) Encontre o intervalo de confiança a 95% para a média.

6) fator K em $(MJmm)^{-1}$ (erodibilidade do solo em relação a quantidade de solo perdido em uma dada área por unidade do índice de erosividade) de n = 22 unidades amostrais de solos brasileiros com horizonte B textural (Bt) estão apresentados a seguir: a) Calcule a média, mediana e moda.

b)Testar a hipótese de que a média brasileira do fator K é igual a de um outro país sul americano dada por 0,074.

7) Os produtores de um programa de televisão pretendem modificá-lo se for assistido regularmente por menos de um quarto dos possuidores de televisão. Uma pesquisa encomendada a uma empresa especializada mostrou que, de 400 famílias entrevistadas 80 assistiam ao programa regularmente. Com base nos resultados, utilize um teste unilateral com nível de significância de 1%, para ver qual deve ser a decisão dos produtores?

8) Uma empresa petroquímica diz que se em 80% de uma determinada região for encontrado petróleo então existe viabilidade econômica da exploração de petróleo. Para verificar a viabilidade econômica da exploração foram tomadas 30 amostras de uma determinada região e em 21 foram encontradas petróleo. Utilizando um teste bilateral com nível de significância de 5%, verifique se existe viabilidade econômica nessa região.

5

REGRESSÃO E CORRELAÇÃO

Para estudar a relação entre duas variáveis quantitativas utilizamos a análise de regressão e correlação destas variáveis.

Correlação é um número entre -1 e 1 que mede o grau relacionamento entre duas variáveis quantitativas

Regressão é o estudo que busca ajustar uma equação a um conjunto de dados de forma que a relação entre duas variáveis quantitativas possa ser expressa matematicamente.

5.1 CORRELAÇÃO

A análise de correlação permite verificar o grau de associação entre duas variáveis. Utilizando a função `cor.test`, é possível fazer um teste de hipótese e o intervalo de confiança para correlação. Esta função apresenta os seguintes argumento:

```
cor.test(x, y, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95,), method = c("pearson", "kendall", "spearman")
```

sendo `x` e `y` vetor de dados, `alternative` o tipo de hipótese alternativa, `conf.level` o nível de confiança. Caso não seja declarado a opções `alternative` e `conf.level` é considerado teste bilateral e nível de confiança de 95%. O argumento `method` especifica o método utilizado para calcular a correlação. O método de Pearson deve ser utilizado quando as variáveis apresenta normalidade e os métodos de Kendall e Spearman quando as variáveis não apresentam normalidade (teste não-paramétricos). Caso não seja especifica do método o calculo será realizado pelo método de Pearson.

$$\begin{cases} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{cases}$$

```
> ##Correlação entre Peso e Altura
> cor.test(P,A)

        Pearson's product-moment correlation

data:  P and A
t = 3.3135, df = 27, p-value = 0.002629
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
 0.2131712 0.7553306
sample estimates:
      cor
0.5376696

> ##Correlação entre Peso e Idade
> cor.test(P,IDADE,method ="spearman")

        Spearman's rank correlation rho

data:  P and IDADE
S = 4015.886, p-value = 0.9554
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01086558
```

Considerando que $\alpha = 0,05$, temos que existe correlação entre Peso e Altura `p-value<0,05` e esta correlação é positiva `0.5376696`. Não existe correlação entre Peso e Idade `p-valu><0,05`.

5.2 REGRESSÃO LINEAR SIMPLES

Regressão linear simples é um modelo de relação entre uma variável aleatória dependente Y e uma variável independente X A função que expressa a relação linear entre X e Y é dada por

$$y = a + bx + \varepsilon$$

em que:

- a é coeficiente linear, interpretado como o valor da variável de dependente quando a variável indenpente é igual a 0;
- b é coeficiente de regressão, interpretado como acréscimo na variável dependente para a variação de uma unidade na variável.
- ε são os erros aleatórios de uma população normal, com média 0 e variância constante.

Uma medida importante ao ajustar o modelo de regressão é o coeficiente de determinação (r^2), que representa a porcentagem da variação total que é explicada pela equação de regressão, quanto maior o seu valor melhor.

Para fazer uma regressão no R utiliza-se a função `lm(Y~X)`.

```
> ##modelo de regressão entre Peso e Altura
> model=lm(P~A)
```

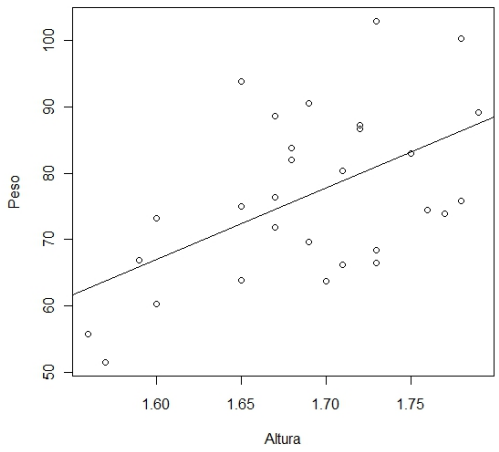
```
> model

Call:
lm(formula = P ~ A)

Coefficients:
(Intercept)            A
     -106.0         108.1
```

O coeficiente `Intercept` corresponde ao `a` e coeficiente `A (variável independente)` corresponde ao b corresponde da reta de regressão. Podemos plotar os dados e adicionar a reta de regressão no mesmo gráfico por meio da função `abline`.

```
> plot(A,P,xlab="Altura",ylab="Peso",)
> abline(model)
```



Após um modelo de regressão é necessário verificar se os parâmetros são significativos. Por meio da função `summary`, pode-se obter um teste t para cada parâmetro do modelo. Além disso, esta função também retorna o coeficiente de determinação `(Multiple R-squared)` informação relativas ao resíduo com valores máximo e mínimo, quartis e mediana,

```
> summary(model)

Call:
lm(formula = P ~ A)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4563  -9.6981   0.9724   6.8243  21.9437

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -105.99      55.14  -1.922  0.06519 .
A             108.06      32.61   3.314  0.00263 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.93 on 27 degrees of freedom
Multiple R-squared: 0.2891,     Adjusted R-squared: 0.2628
F-statistic: 10.98 on 1 and 27 DF,  p-value: 0.002629
```

Como resultado verifica-se que o intercepto é não significativo `valor-p=0.06519` e que o coeficiente b é significativo `valor-p=0.00263`. Pode-se observar também o $r^2 = 0.2891$.

Como no teste t verificou-se que o intercepto é não significativo pode-se ajustar um novo modelo sem o mesmo. Para isso, basta adicionar `-1` ao modelo.

```
#modelo sem intercepto
> model1=lm(P~A-1)
> summary(model1)

Call:
lm(formula = P ~ A - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-19.8057 -10.1725   0.0609   7.9025  24.3275

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
A   45.418      1.257   36.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.44 on 28 degrees of freedom
Multiple R-squared: 0.979,      Adjusted R-squared: 0.9783
F-statistic:  1306 on 1 and 28 DF,  p-value: < 2.2e-16
```

A análise de variância é uma técnica utilizada para se testar o ajuste da equação como um todo, ou seja, um teste para verificar se a equação de regressão obtida é significativa ou não, esta pode ser feita por meio da função `anova`

```
> ##análise de variância para o modelo ajustado
> anova(model)
Analysis of Variance Table
```

```
Response: P
          Df Sum Sq Mean Sq F value   Pr(>F)
A          1 1311.3 1311.33  10.979 0.002629 **
Residuals 27 3224.7  119.43
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O valor $Pr(>F)$ indica o valor-p, assim como como $valor-p < 0,05$, considera-se o modelo significativo ao nível nominal de 5% de significância, ou seja há contribuição da variável Altura para o modelo.

Ao ajustar o modelo de regressão é necessário verificar a pressuposição de que são os erros aleatórios de uma população normal. Para gerar o erros do modelo pode-se utilizar a função `resid`

```
> ##gerar o residuos
> erro=resid(model1)
> ##teste de normalidade
> shapiro.test(erro)

        Shapiro-Wilk normality test

data:  erro
W = 0.9679, p-value = 0.5055
```

Para verificar os valores preditos pelo modelo de regressão, pode-se utiliza a função `predict`. Esta função tem várias opções,

```
predict(object,  se.fit = FALSE, interval = c("none", "confidence", "predic-
                            tion"),level = 0.95)
```

sendo:

- `object` o modelo ajustado
- `se.fit` indica se quer calcular o erro padrão (`TRUE`) ou não (`FALSE`)
- `interval` para calcular o intervalo de confiança dos valores preditos. `none` não calcula e `confidence`, `prediction` calculam o intervalo
- `level` o nível de confiança do intervalo.

```
> ##calcula a previsão e seu intervalo de confiança
> prev=predict(model,interval = "confidence")
> plot(A,P,xlab="Altura",ylab="Peso")
> previsao=cbind(A,prev) ##adicionar altura ao conjunto
> previsao=previsao[order(previsao[,1]),] ##ordenar os dados pela altura
> previsao
      A      fit      lwr      upr
```

```
16 1.56 62.58578 52.96239 72.20918
3  1.57 63.66640 54.64163 72.69117
27 1.59 65.82764 57.96591 73.68936
8  1.60 66.90825 59.60551 74.21100
24 1.60 66.90825 59.60551 74.21100
12 1.65 72.31134 67.37374 77.24894
21 1.65 72.31134 67.37374 77.24894
28 1.65 72.31134 67.37374 77.24894
9  1.67 74.47258 70.10583 78.83933
17 1.67 74.47258 70.10583 78.83933
25 1.67 74.47258 70.10583 78.83933
6  1.68 75.55320 71.33940 79.76700
26 1.68 75.55320 71.33940 79.76700
10 1.69 76.63381 72.46977 80.79785
13 1.69 76.63381 72.46977 80.79785
20 1.70 77.71443 73.49331 81.93555
19 1.71 78.79505 74.41418 83.17592
29 1.71 78.79505 74.41418 83.17592
18 1.72 79.87567 75.24299 84.50835
22 1.72 79.87567 75.24299 84.50835
2  1.73 80.95629 75.99373 85.91884
5  1.73 80.95629 75.99373 85.91884
7  1.73 80.95629 75.99373 85.91884
15 1.75 83.11752 77.31718 88.91786
14 1.76 84.19814 77.91357 90.48270
11 1.77 85.27876 78.47853 92.07898
1  1.78 86.35937 79.01867 93.70008
4  1.78 86.35937 79.01867 93.70008
23 1.79 87.43999 79.53909 95.34090
> lines(previsao[,1:2]) ##plotar a previsao
> lines(previsao[,1],previsao[,3],lty=2) ##plotar limite inferior
> lines(previsao[,1],previsao[,4],lty=2) ##plotar limite superior
```

5.3 EXERCÍCIOS

1) Utilizando as variáveis numero de filhos e renda do conjunto de dados da tabela A.2

a) Obtenha o gráfico de dispersão;

b) Calcule a correlação;

c) Ajuste um modelo de regressão linear e verifique o ajuste.

2) Utilizando as variáveis calcio e zinco do conjunto de dados da tabela A.3

a) Obtenha o gráfico de dispersão dentro de cada grupo;

b) Calcule a correlação dentro de cada grupo;

c) Ajuste um modelo de regressão linear dentro de cada grupo e verifique o ajuste.

3) Oito programas foram monitorados para estudar a demanda por recursos. Neste trabalho, a variável resposta (dependente Y) é o tempo de CPU, e a variável independente (X) é o número de acessos ao disco (disk I/O)

X	14	15	23	31	38	40	53	51
Y	2,0	4,6	5,7	7,3	9,8	10,9	12,6	13,2

a) Faça o diagrama de dispersão dos dados;

b) Obtenha o coeficiente de correlação e verifique se a correlação é significativa;

c) Estime a equação de regressão que se ajusta aos dados;

d) Teste o ajuste da equação de regressão;

e) Faça um gráfico da equação de regressão junto com diagrama de dispersão;

f) Calcule o coeficiente de determinação e interprete.

g) Obtenha a previsão para o tempo de CPU.

4) É esperado que a massa muscular de uma pessoa diminua com a idade. Para estudar essa relação, uma nutricionista selecionou 18 mulheres, com idade entre 40 e 79 anos, e observou em cada uma delas a idade (X) e a massa muscular (Y).

Massa muscular (Y)	Idade (X)
82.0	71.0
91.0	64.0
100.0	43.0
68.0	67.0
87.0	56.0
73.0	73.0
78.0	68.0
80.0	56.0
65.0	76.0
84.0	65.0
116.0	45.0
76.0	58.0
97.0	45.0
100.0	53.0
105.0	49.0
77.0	78.0
73.0	73.0
78.0	68.0

a) Faça o diagrama de dispersão dos dados;

b) Obtenha o coeficiente de correlação e verifique se a correlação é significativa;

c) Estime a equação de regressão que se ajusta aos dados;

d) Teste o ajuste da equação de regressão;

e) Faça um gráfico da equação de regressão junto com diagrama de dispersão;

f) Calcule o coeficiente de determinação e interprete.

A

CONJUNTOS DE DADOS

Tabela A.1: Dados referentes a 29 empregados de uma empresa.

IDADE	SEXO	GI	EC	AT	NFD	P	P1	A
29	M	médio	Casado	Sim	5	100,3	80,26	1,78
22	M	médio	Solteiro	Não	5	102,9	95,4	1,73
21	F	médio	Casado	Não	3	51,5	50,84	1,57
20	F	médio	Solteiro	Não	2	75,8	76,56	1,78
35	M	fundamental	Solteiro	Não	4	68,4	54,72	1,73
21	F	médio	Solteiro	Não	3	83,8	67,04	1,68
27	M	superior	Solteiro	Não	2	66,5	65,84	1,73
39	M	superior	Solteiro	Não	2	73,2	72,47	1,60
39	M	médio	Casado	Não	4	88,6	87,71	1,67
20	F	médio	Solteiro	Não	2	69,6	68,9	1,69
51	M	médio	Solteiro	Não	3	73,9	73,16	1,77
20	F	superior	Solteiro	Não	2	63,8	62,52	1,65
38	F	médio	Casado	Sim	4	90,6	86,07	1,69
43	F	fundamental	Casado	Não	3	74,5	70,78	1,76
35	F	médio	Casado	Sim	5	83,0	58,1	1,75
27	F	fundamental	Casado	Sim	2	55,7	50,13	1,56
27	F	superior	Solteiro	Sim	5	71,8	71,08	1,67
24	F	superior	Solteiro	Não	2	86,7	83,23	1,72
24	F	superior	Solteiro	Sim	5	80,3	78,69	1,71
28	M	fundamental	Casado	Não	2	63,7	61,15	1,70
22	M	superior	Divorciado	Não	3	93,8	84,42	1,65
29	F	médio	Casado	Sim	4	87,2	82,84	1,72
49	M	médio	Casado	Não	3	89,2	82,06	1,79
30	M	médio	Casado	Sim	2	60,2	59	1,60
25	F	fundamental	Casado	Sim	3	76,3	75,54	1,67
27	F	superior	Divorciado	Não	4	82,0	80,36	1,68
47	F	fundamental	Divorciado	Não	2	66,8	66,13	1,59
41	F	fundamental	Divorciado	Sim	3	75,0	74,25	1,65
49	M	médio	Casado	Sim	4	66,2	64,88	1,71

- GI - grau de instrução;
- EC - Estado Civil;
- AT - faz atividade física (sim ou não);
- NFD - Numero de refeições diárias;
- P - Peso no inicio do estudo;
- P1 - Peso no final do estudo;
- A - Altura

Tabela A.2: Dados referentes a um levantamento feito com 90 famílias.

BAIRRO	GI	NF	RENDA	BAIRRO	GI	NF	RENDA	BAIRRO	GI	NF	RENDA
1	2	2	8,7	1	2	3	6,1	3	2	2	7,7
1	0	2	4,7	1	1	0	7,2	3	2	0	4,9
1	0	2	5,6	1	2	3	10,8	3	2	1	4,6
1	1	2	2,6	1	1	1	6,1	3	1	3	4,4
1	1	2	6,3	1	1	1	7,0	3	0	1	4,4
1	2	2	5,4	2	1	1	4,2	3	0	2	5,0
1	0	1	8,9	2	1	2	6,5	3	2	2	7,5
1	1	1	6,3	2	1	2	6,7	3	0	3	9,4
1	2	0	7,3	2	0	2	5,4	3	0	2	7,4
1	1	4	2,5	2	1	3	7,9	3	2	2	3,9
1	2	2	6,2	2	0	1	4,0	3	1	1	5,3
1	1	2	5,8	2	2	1	6,4	3	1	2	11,2
1	1	2	9,4	2	0	1	3,1	3	0	2	4,8
1	2	1	7,1	2	2	1	4,2	3	2	2	3,1
1	2	0	5,6	2	2	2	5,6	3	3	2	6,9
1	2	1	5,3	2	0	2	6,1	3	1	1	5,3
1	2	1	1,8	2	0	3	6,9	3	1	1	7,1
1	1	2	5,9	2	1	3	7,9	3	1	0	7,0
1	2	2	6,6	2	1	0	7,1	3	1	2	6,7
1	0	2	8,5	2	2	3	9,6	3	1	1	8,2
1	0	0	8,2	2	1	2	7,1	3	1	3	8,1
1	1	1	4,7	2	0	3	5,7	3	0	2	7,5
1	1	2	8,9	2	1	1	6,5	3	1	1	3,6
1	2	1	4,2	2	0	1	9,2	3	0	1	4,5
1	0	1	12,0	2	3	0	8,7	3	1	1	3,0
1	2	2	7,4	2	1	0	8,6	3	0	1	4,9
1	1	1	6,6	2	1	3	6,2	3	1	1	8,5
1	2	1	8,4	2	2	1	4,3	3	2	1	4,9
1	2	1	5,6	2	3	1	4,2	3	0	1	5,0

- GI - grau de instrução do chefe da casa (0 = analfabeto; 1 = ensino fundamental completo; 2 = ensino médio completo);
- NF - numero de filhos;
- RENDA: renda familiar mensal, em salários mínimo

Tabela A.3: Concentração de minerais no leite materno, dados de 53 mães do Hospital Maternidade Odete Valadares em Belo Horizonte. As mães foram divididas em dois grupos, segundo o período de lactação: colostro e leite maduro.

Amostra	grupo	Calcio	Zinco	Amostra	grupo	Calcio	Zinco
1	Maduro	181	0,78	28	Maduro	238	1,60
2	Maduro	213	1,28	29	Colostro	241	5,82
3	Maduro	217	1,40	30	Maduro	256	2,27
4	Maduro	238	1,55	31	Maduro	264	2,57
5	Colostro	256	6,13	32	Colostro	303	6,82
6	Maduro	263	2,52	33	Colostro	334	7,45
7	Colostro	323	7,43	34	Maduro	175	0,60
8	Maduro	344	5,50	35	Maduro	206	1,23
9	Colostro	437	9,54	36	Maduro	259	2,40
10	Colostro	163	3,13	37	Maduro	260	2,41
11	Colostro	254	6,07	38	Maduro	275	2,60
12	Maduro	281	3,02	39	Colostro	275	6,50
13	Maduro	293	3,05	40	Maduro	279	2,92
14	Colostro	296	6,82	41	Maduro	303	3,57
15	Colostro	311	6,90	42	Colostro	312	7,23
16	Colostro	163	1,30	43	Colostro	313	7,42
17	Maduro	214	1,31	44	Maduro	314	3,88
18	Maduro	242	1,92	45	Colostro	325	7,43
19	Colostro	372	8,40	46	Colostro	145	1,20
20	Colostro	113	1,07	47	Colostro	167	3,20
21	Maduro	159	0,52	48	Colostro	181	3,70
22	Maduro	188	0,86	49	Colostro	221	4,40
23	Maduro	200	1,08	50	Maduro	231	1,55
24	Colostro	225	4,57	51	Maduro	244	2,15
25	Colostro	231	5,20	52	Maduro	277	2,63
26	Maduro	394	7,38	53	Colostro	344	8,25
27	Colostro	375	8,77				