

Estatística Computacional

Pós- Graduação em Ciência dos Dados

Prof. Dr. Roberta A. de A. Fagundes
roberta.fagundes@upe.br

Homepage
robertafagundes.wix.com/raaf



O que veremos na aula

- Softwares Estatísticos
- Linguagem R
 - Instalação do R
 - Conhecendo o R
 - Estatística Descritiva



Softwares Estatístico

- Atualmente existem dezenas de softwares estatísticos;
- É praticamente impossível, imaginar “a vida” de um analista de dados sem os recursos computacionais atuais;
- Um cientista de dados deve conhecer o máximo de softwares de análises de dados;



Softwares Estatístico



Softwares Estatístico

- Linguagem de programação especializada em computação de dados;
 - É um software gratuito;
 - Multiplataforma (Win, Linux, Mac...);
 - Grande quantidade de bibliotecas (pacotes);



Linguagem R

- Foi criado por Ross Ihaka e Robert Gentleman;
- Departamento de Estatística da universidade de Auckland, Nova Zelândia;
- O nome foi inspirado nas iniciais dos autores;
- Foi baseado na linguagem S.



Linguagem R

■ Vantagens

- Grande variedade de pacotes disponíveis gratuitamente;
- Controle total sobre o processo de análise;
- Possibilidade de integração com outras linguagens;
- Grande comunidade de desenvolvedores;
- Muita documentação grátis;
- Excelente para a simulação, programação, análises intensivas de computado.

Linguagem R

■ Desvantagens

- Não há suporte comercial;
- Trabalhando com grandes conjuntos de dados é limitada pela RAM
- Fácil cometer erros se não conhecer bem a linguagem;
- Preparação e limpeza de dados pode ser mais confusa e mais propenso erro em R.

Linguagem R

■ Grande quantidade de pacotes:

- sqldf - pacote que permite realizar queries SQL em dataframes no R;
- lm- regressão linear
- plyr - dividir uma estrutura de dados em grupos;
- stringr - manipulação de strings;
- database drivers - RMongo, RODBC, RMySQL;
- ggplot2 - visualização de dados
- caret - pacote para Machine Learning;
✓ quase 9.000 pacotes (<https://cran.r-project.org>)

Instalação do R

- Linguagem e ambiente
- Software livre
- Multiplataforma
 - <http://www.r-project.org/>
- RStudio
 - <https://www.rstudio.com/products/rstudio/download/>

Instalação do R

[\[Home\]](#)[Download](#)[CRAN](#)[R Project](#)[About R](#)[Contributors](#)[What's New?](#)[Mailing Lists](#)[Bug Tracking](#)[Conferences](#)[Search](#)[R Foundation](#)[Foundation](#)[Board](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- [R version 3.2.2 \(Fire Safety\)](#) has been released on 2015-08-14.
- [The R Journal Volume 7/1](#) is available.
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-09.
- [useR! 2015](#), will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- [useR! 2014](#), took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.



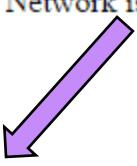
Instalação do R

CRAN MIRRORS

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the [windows old release](#).

0-Cloud

<https://cran.rstudio.com/>
<http://cran.rstudio.com/>



Algeria

<http://cran.usthb.dz/>

Argentina

<http://mirror.fcaglp.unlp.edu.ar/CRAN/>

Australia

<http://cran.csiro.au/>
<http://cran.ms.unimelb.edu.au/>

Austria

<https://cran.r-project.org/>
<http://cran.at.r-project.org/>

Belgium

<http://www.freestatistics.org/cran/>
<http://lib.ugent.be/CRAN/>

Rstudio, automatic redirection to servers worldwide

Rstudio, automatic redirection to servers worldwide

University of Science and Technology Houari Boumediene

Universidad Nacional de La Plata

CSIRO

University of Melbourne

Wirtschaftsuniversität Wien

Wirtschaftsuniversität Wien

K.U.Leuven Association

Ghent University Library

Instalação do R



[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)

[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)



R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

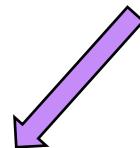
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2015-08-14, Fire Safety) [R-3.2.2.tar.gz](#), read [what's new](#) in the latest version.

Instalação do R

[CRAN](#)[Mirrors](#)[What's new?](#)[Task Views](#)[Search](#)[About R](#)

R for Windows



Subdirectories:

[base](#)

Binaries for base distribution (managed by Duncan Murdoch). This is what you want to [install R for the first time](#).

[contrib](#)

Binaries of contributed packages (managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows services and corresponding environment and make variables.

[Rtools](#)

Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions to Windows binaries.

Conhecendo o Ambiente R

- Para solicitar uma tarefa do R podemos digitar uma linha de comando ou, se a tarefa é complexa, digitar várias linhas de comando, respeitando-se a sintaxe do R.
- Esta sucessão de comandos é chamada um programa ou código ou função. Tal programa ou função, pode conter apenas uma linha com uma única tarefa a ser executada como conter várias páginas com comandos a serem executados.

Conhecendo o Ambiente R

- Os programas em R, bem como os dados a serem explorados, podem ser armazenados em arquivos de texto (extensão .txt);

- Os dados também podem ser armazenados em uma planilha de cálculo e depois, salvos como arquivo texto para que possam ser lidos no R.

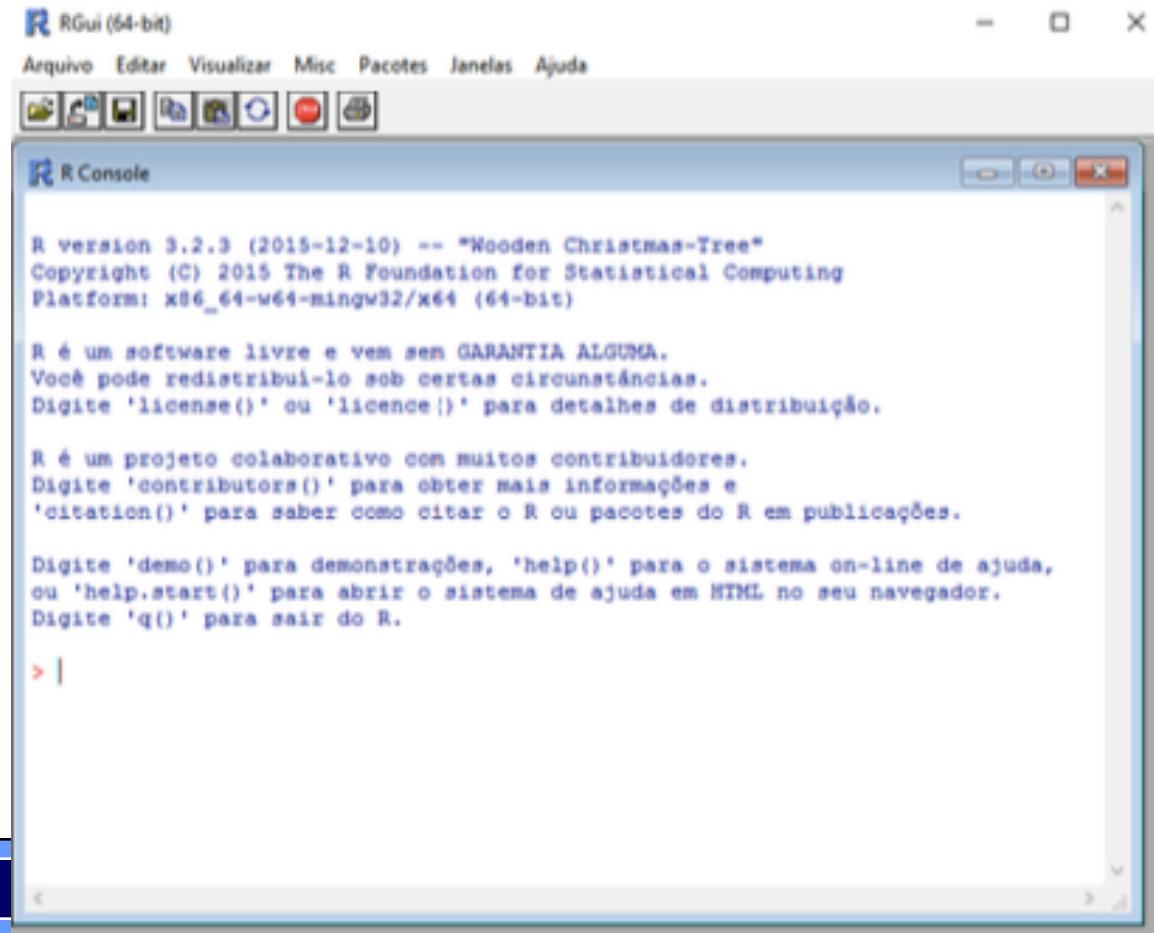
Conhecendo o Ambiente R

- Todas as funções do R devem ser digitadas em letras minúsculas pois o R é sensível a letras maiúsculas e minúsculas. Todas as palavras-chaves do R estão em letras minúsculas;

- O R não reconhecerá, por exemplo, o comando *MEAN(x)*, pedindo para calcular a média de uma base de dados x. O correto será digitar *mean(x)*, que retornará a média de uma base de dados x.

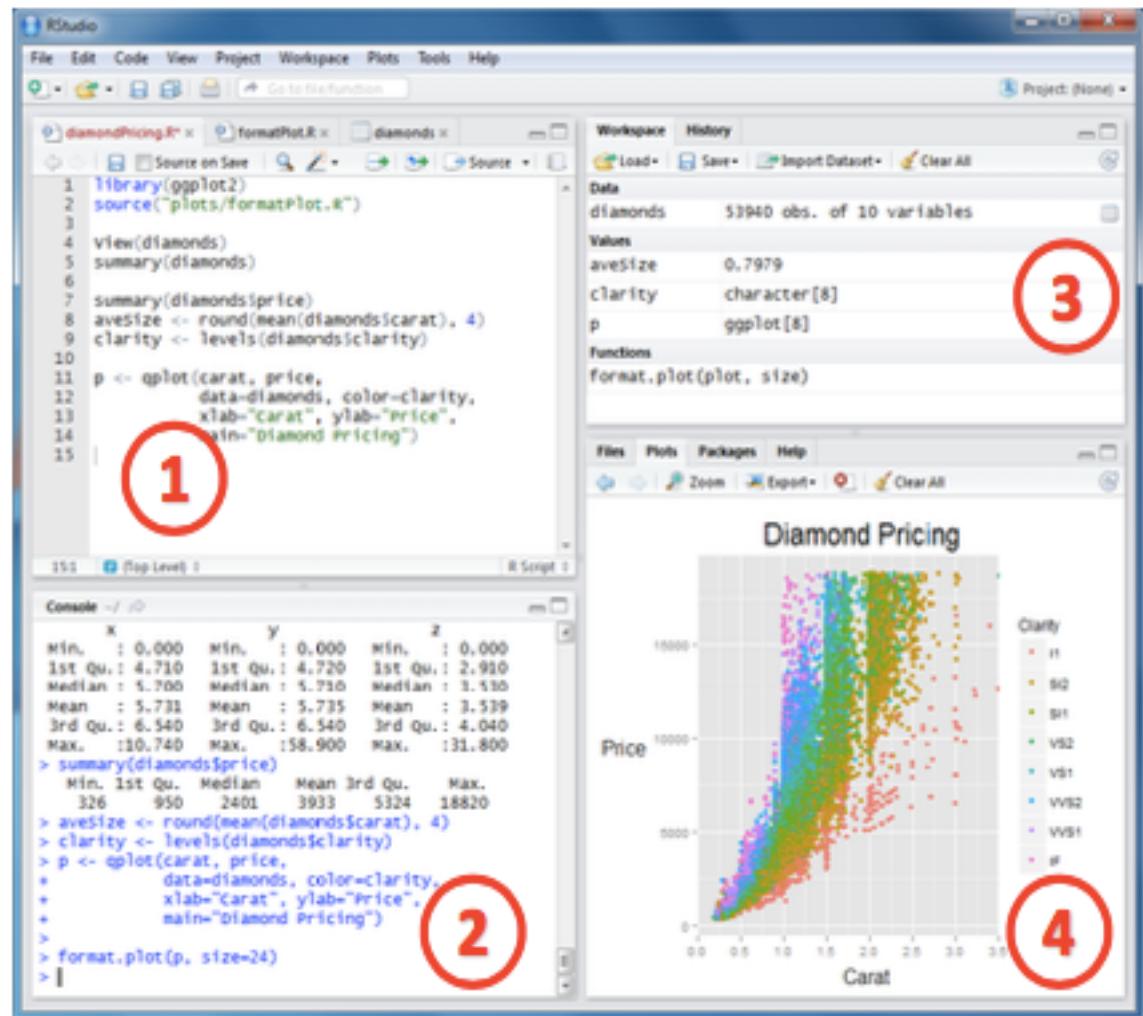
Conhecendo o Ambiente R

- Console básico do R;
 - O console pode ser improdutivo;
 - Não fornece funcionalidades para codificação;



Conhecendo o Ambiente R

- IDE – RStudio;
 - Disponível gratuitamente;
 - Função autocomplete;
- Quatro áreas básicas:
 - Codificação (1);
 - Console (2);
 - Status(3);
 - Output(4).



Conhecendo o Ambiente R

■ Para começar veremos como é a forma de um comando de atribuição no R. Ou seja, como fazemos para atribuir a alguma variável, x (por exemplo), o valor 4.

- $x=4$

■ Experimente atribuir a uma variável x1 um valor qualquer, a outra variável x2 outro valor qualquer e finalmente, atribua à z o valor da soma de x1 e x2.

- $x1=4$
- $x2=6$
- $z=-x1+x2$

Linguagem R

■ Operações Básicas

■ soma: +

■ subtração: -

■ divisão: /

■ multiplicação: *

■ potenciação: ** ou ^

■ Resto da divisão: %%

```
3 # Soma  
4 4 + 4  
5 # Subtracao  
6 4 - 4  
7 # Multiplicacao  
8 4 * 4  
9 # Divisao  
10 4 / 4  
11 # Potencia  
12 4^2  
13 4**2  
14 # Modulo  
15 14 %% 3
```

Linguagem R

<i>Operador</i>	<i>Descrição</i>
<	<i>menor que</i>
>	<i>maior que</i>
<=	<i>menor ou igual</i>
>=	<i>maior ou igual</i>
==	<i>igual</i>
!=	<i>Diferente</i>
&	<i>and</i>
	<i>or</i>

```

3 a = 7
4 b = 5
5 # Operadores
6 a > 8
7 a < 8
8 a <= 8
9 a >= 8
10 a == 8
11 a != 8
12 # Operadores logicos
13 # And
14 (a==8) & (b==6)
15 # Or
16 (a==8) | (b>5)

```

Linguagem R

Tipos de dados e objetos

- **Tipos de Dados**

- ✓ **Numérico;**
- ✓ **Character;**
- ✓ **Logic;**
- ✓ **Vetor.**

nome = "EU"

Nome

letra = 'A'

Letra

x=1

y=2

z = x>y

z

vetor = c(1,2,3,4)

vetor

var = c("EU","TU", "ELE")

var[1]

Linguagem R

Operações com Vetores

Função	Descrição	Exemplos
c	combina valores (qualquer tipo)	c(1,3,2,6) c("sim", "não")
rep	repete valores (qualquer tipo)	rep(c(1,2), 3) x<-rep('a',5)
:	sequências numéricas	1:5 1:-1
seq	sequências numéricas	x<-seq(-1,1,0.4) x<-seq(1, by=2, length=10)

➤ **help("seq")**
 ➤ **? seq()**

Linguagem R

Operações com Vetores

■ Subconjunto de vetores

■ `x<-c(0, 8, 9, 7, 4, 2, 10, 0, 2, 1)`

- `xa = x[x > 4]`
- `xb = x[x > 2 & x <= 8]`
- `x<-seq(-2, 2, by=0.5)`
- `x>=-1`
- `x>=-1 & x<=1`
- `x<=-1 | x>=1`

Linguagem R

Operações com Vetores

■ Vetores alfanuméricos

```
x=paste(c("X"), 1:10, sep="-")
```

```
[1] "X-1" "X-2" "X-3" "X-4" "X-5" "X-6" "X-7"
```

```
[8] "X-8" "X-9" "X-10"
```

```
xy=paste(c("X","Y"), 1:10, sep="-")
```

```
[1] "X-1" "Y-2" "X-3" "Y-4" "X-5" "Y-6" "X-7"
```

```
[8] "Y-8" "X-9" "Y-10"
```

Linguagem R

Operações com Vetores

■ Subconjunto de vetores

```
x = c(0, 8, 9, 7, 4, 2, 10, 0, 2, 1)
```

```
x1 = x[6]
```

```
x2 = x[2:6]
```

```
x3 = x[c(2, 4, 8)]
```

```
frutas = c(5, 10, 1, 20)
```

```
names(frutas)<-c("laranja", "banana", "maçã",  
"pera")
```

```
jantar<-frutas[c("maçã", "laranja")]
```

Linguagem R

Operações com Vetores

■ Exemplo vetores:

- `x = c(10.4, 5.6, 3.1, 6.4, 21.7)`
- `fruta = c("banana", "laranja", "uva")`
- `length(x), length(y)`
- `mode(x), mode(y)`

■ Listar objetos: `ls()`

- *Eliminar os objetos x e y: `rm(x,y)`*
- *Eliminar todos os objetos: `rm()`*

Linguagem R

Operações com Vetores

- Vetores numéricos e lógicos
- Operações elemento a elemento, caso tenham a mesma dimensão.

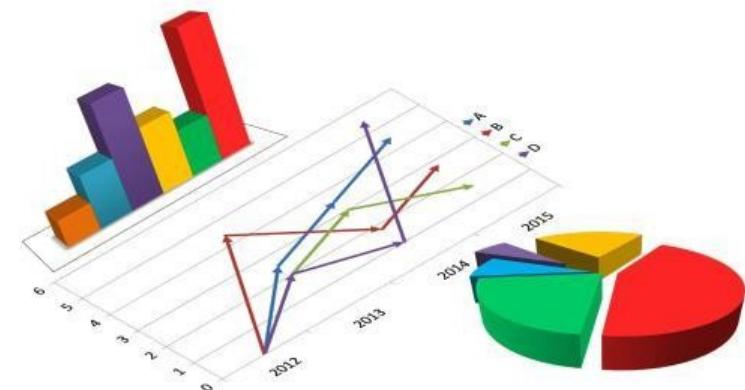
```
> peso = c(62, 70, 52, 98, 90, 70)  
> altura = c(1.70, 1.82, 1.75, 1.94, 1.84, 1.61)
```

- Calcular IMC para essas pessoas:
 - $i.m.c = peso/altura^2$
 - i.m.c
 - [1] 21.45329 21.13271 16.97959 26.03890
26.58318 27.00513

Linguagem R

Sumarização Descritiva

- É importante sempre aplicar medidas descritivas antes de qualquer análise:
 - Medidas de tendência central;
 - Medidas de dispersão;



Linguagem R

Sumarização Descritiva

■ Funções

- $\text{max}(\text{peso}) = 98$
- $\text{min}(\text{peso}) = 52$
- $\text{range}(\text{peso}) = 52 \text{ } 98$
- $\text{mean}(\text{peso}) = 73.66667$
- $\text{median}(\text{peso}) = 70$
- $\text{var}(\text{peso}) = 298.2667$
- $\text{sd}(\text{peso}) = 17.2704$

sum(x)/length(x)
73.66667

sum((x-mean(x))^2)/(length(x)-1)
298.2667

Linguagem R

Sumarização Descritiva

■ Quantis

- $q = c(48,49,51,50,49)$
- $\text{quantile}(q)$

■ Percentis

- $\text{percentis} = \text{seq(.01,.99,.01)}$
- $\text{quantile}(q, \text{percentis})$

■ Dercis

- $d = c(48,49,51,50,49)$
- $\text{quantile}(d, \text{seq}(0.10,0.9,0.1))$

Linguagem R

Sumarização Descritiva

- $v = c(10,11,9,10,10,9,11)$
 - $CV = 100 * (sd(v)/mean(v))$
 - CV
 - [1] 8.164966 #em torno de 8%
- $\sqrt{64} = 8$
- $abs(-2) = 2$
- $z = c(5,2,6,9,10,13,15)$
 - `summary(z)`

Exercício

- Um artigo no Journal of Structural Engineering (Vol. 115, 1989) descreve um experimento para testar a resistência resultante em tubos circulares com calotas soldadas nas extremidades. Os primeiros resultados são: 96; 96; 102; 102; 102; 104; 104; 108; 126; 126; 128; 128; 140; 156; 160; 160; 164 e 170. Pede-se:
 - a) Calcule a média e mediana da amostra e dê uma interpretação.
 - b) Calcule os percentis 9%, 25%, 5% e 69%.
 - c) Calcule o segundo quartil ou mediana.
 - d) Calcule a amplitude da amostra.
 - e) Calcule a variância e o desvio padrão da amostra.
 - f) Qual a fonte de maior variabilidade deste experimento.

Exercício _ Resposta

- A) Com este valor podemos concluir que a resistência da solda das calotas circulares se concentra, na maioria dos testes, em torno do valor médio. Isto é, se pegarmos aleatoriamente uma calota soldada é de se esperar que a resistência da solda se concentre em torno (e próximo) da média.
- F) Como a estatística se preocupa com a variabilidade dos dados amostrais, devemos apontar suas causas. Neste exemplo, podemos apontar como possíveis causas de variabilidade os erros de medição da resistência da solda, soldagem feita por soldadores diferentes (caso não seja automatizado), etc. Enfim, devemos reduzir a variabilidade para termos garantias de qualidade e, num cenário ideal, eliminá-la.

Dúvidas

