

# Análise de Dados com [R]

Prof. Diego Mello da Silva

IFMG - Campus Formiga

20 de fevereiro de 2018

# Conteúdo

## 1 Dia 01 - Manipulando Dados

- Ambiente [R]
- Matrizes, Vetores, Dataframes, Listas e Arrays
- Datasets
- Manipulando Dataframes
- Importando Datasets como Dataframes

## 2 Dia 2 - Analisando Dados

- Gráficos
- Análise Descritiva
- Histogramas e Boxplots
- Regressão Linear

## 3 Anexo - Revisão Teórica

## 4 Referências Bibliográficas

# Ambiente R

# Ambiente R

- Ambiente de software livre para computação científica, disponível em plataformas Unix, Windows e MacOS
- Criada por Ross Ihaka e por Robert Gentleman no Departamento de Estatística da Universidade de Auckland, Nova Zelândia. É vista como uma implementação derivada da linguagem 'S', da AT&T.
- Possui implementação de diversas técnicas gráficas e estatísticas para modelagem linear e não-linear, testes estatísticos, análise de séries temporais, classificação, clusterização, dentre outros
- Expansível com uso de pacotes, disponíveis no CRAN (*Comprehensive R Archive Network*), rede de servidores que armazenam cópias do código e documentação do R.
- Plataforma aberta: criação e hospedagem de pacotes no CRAN  
<https://cran.r-project.org/web/packages/policies.html>
- URL do CRAN  
<https://cran.r-project.org/>
- Pacotes por ordem alfabética  
[https://cran.r-project.org/web/packages/available\\_packages\\_by](https://cran.r-project.org/web/packages/available_packages_by)

■ Disponível em: <https://www.r-project.org/>

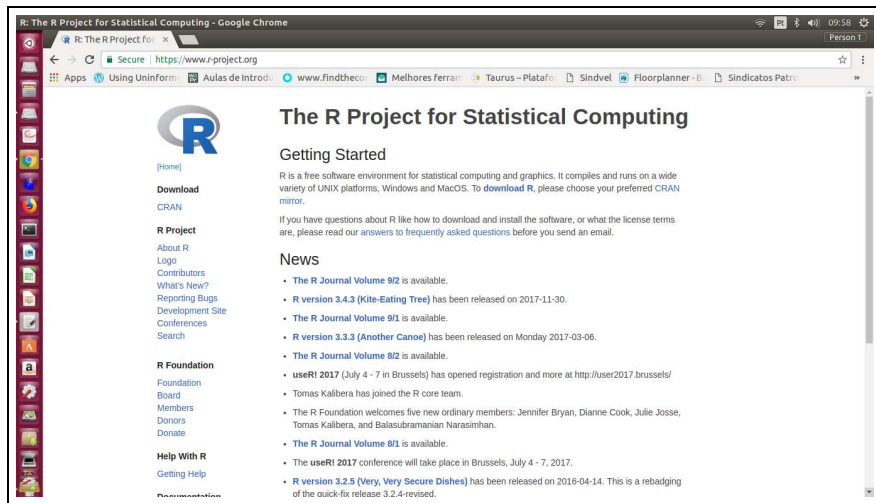


Figura: Site Web do R Project

# Ambiente R

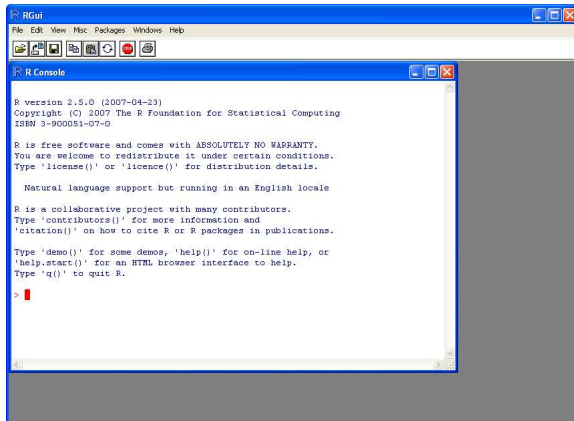
- Pode-se utilizar o interpretador da linguagem R, que é instalado junto com o ambiente, linha de comando

```
diego@Hermes: ~  
diego@Hermes:~$ R  
  
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"  
Copyright (C) 2017 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> █
```

- Nele comandos são inseridos diretamente no console, em linha de comando.

# Ambiente R

- Pode-se utilizar o interpretador da linguagem R, que é instalado junto com o ambiente, linha de comando



The screenshot shows the RGui application window. The title bar reads 'RGui'. The menu bar includes 'File', 'Edit', 'View', 'Misc', 'Packages', 'Windows', and 'Help'. Below the menu bar is a toolbar with icons for file operations and execution. The main window is titled 'R Console' and contains the following text:

```
R version 2.5.0 (2007-04-23)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

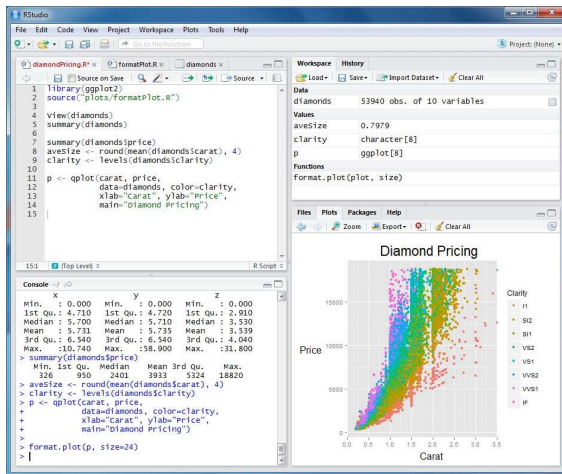
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> █
```

- Nele comandos são inseridos diretamente no console, em linha de comando.

# Ambiente R

## ■ R Studio: uma das IDEs mais usadas com o ambiente R



## ■ Disponível em: <https://www.rstudio.com/products/RStudio/>



# Matrizes, Vetores, Dataframes, Listas, Arrays

# Criando Vetores no R

- Para criar um vetor no R a partir de uma lista de números

```
> <vetor> <- c(<lista de números separados por vírgula>)
```

- Para criar um vetor de com números repetidos no R

```
> <vetor> <- rep(<quantidade de números>, valor)
```

- Para criar um vetor de números em uma sequência, com intervalo unitário

```
> <vetor> <- <min>:<max>
```

- Para criar um vetor de números em uma sequência, com intervalo informado

```
> <vetor> <- seq(<min>, <max>, by=intervalo)
```

- Para criar um vetor de números a partir de outro

```
> <vetor> <- c(<outro.vetor>, <lista de números>)
```

# Operações com Vetores no R

- Somar vetores existentes no R, digite:

```
> <resultado> <- <vetor1> + <vetor2> + ... + <vetorN>
```

- Produto de número real por vetor no R

```
> <resultado> <- <real> * <vetor>
```

- Produto escalar de dois vetores no R, digite

```
> <resultado> <- <vetor1> %*% <vetor2>
```

- Calcular o norma (intensidade) de um vetor no R, digite

```
> <resultado> <- sqrt (sum (<vetor>^2))
```

# Matrizes no R

- Criar uma matriz a partir de um vetor de elementos

```
> <resultado> <- matrix(<vetor>, <lin>, <col>)
```

- Criar uma matriz a partir de um vetor de elementos, listador por linha

```
> <resultado> <- matrix(<vetor>, <lin>, <col>, byrow=T)
```

- Criar uma matriz a partir de vetores coluna

```
> <resultado> <- cbind(<vetor1>, <vetor2>, ..., <vetorN>)
```

- Atribuindo nomes às colunas da matriz

```
> colnames(<matriz>) <- c('<lab1>', '<lab2>', ..., '<labN>')
```

# Operações com Matrizes no R

- Calcular a transposta de uma matriz no R

```
> <resultado> <- t(<matriz> )
```

- Calcular a inversa de uma matriz no R, digite

```
> <resultado> <- solve(<matriz> )
```

- Calcular o produto de duas matrizes no R, digite:

```
> <resultado> <- <matriz1> %*% <matriz2>
```

# Acessando elementos de matrizes no R

- Acessar elemento específico em uma matriz

```
> <matriz> [<linha>, <coluna>]
```

- Acessar linha específica da matriz

```
> <matriz> [<linha>, ]
```

- Acessar coluna específica da matriz

```
> <matriz> [, <coluna>]
```

- Acessar linhas ou colunas específicas da matriz

```
> <matriz> [<intervL>, <intervC>]
```

# Criando dataframes no R

Um **dataframe** pode conter colunas heterogêneas, cada qual de um tipo de dados diferente

- Criando um dataframe a partir de vetores

```
> <frame> <- data.frame(<V1>=<vetor>, ..., <VN>=<vetor>)
```

- Criando um dataframe a partir da edição de valores

```
> <frame> <- edit(data.frame())
```

- Editando e atualizando o conteúdo de um dataframe

```
> <frame> <- edit(<frame>)
```

- Obtendo o número de linhas (observações) e colunas (variáveis) de um dataframe

```
> nrow(<frame>)  
> ncol(<frame>)
```

# Criando listas no R

Uma **lista** permite armazenar objetos no R de maneira genérica e flexível. Pode ter diferentes tamanhos de linhas e colunas, com diferentes tipos de dados.

## ■ Criando uma lista a partir de objetos

```
> <lista> <- list(<V1>=<objeto>, ..., <VN>=<objeto>)
```

## ■ Criando uma lista vazia

```
> <lista> <- list()
```

## ■ Acessando objeto específico da lista

```
> <lista>[[<Indice>]]
```

## ■ Atribuindo/acrescentando um objeto na lista

```
> <lista>[[<Indice>]] <- <objeto>
```



# Criando arrays no R

Um **array** estende o conceito da matriz para mais dimensões

## ■ Criando um array a partir de objetos

```
> <obj.array> <- array(<vetor>, dim=c(<dim1>, ..., <dimN>))
```

## ■ Acessando elementos de um array

```
> <obj.array> [<Ind1>, <Ind2>, ..., <IndN>]
```

# Datasets

# Pacote datasets do [R]<sup>1</sup>

Dataset	Descrição
ability.cov	Ability and Intelligence Tests
airmiles	Passenger Miles on Commercial US Airlines, 1937-1960
AirPassengers	Monthly Airline Passenger Numbers 1949-1960
airquality	New York Air Quality Measurements
anscombe	Anscombe's Quartet of 'Identical' Simple Linear Regressions
attenu	The Joyner-Boore Attenuation Data
attitude	The Chatterjee-Price Attitude Data
austres	Quarterly Time Series of the Number of Australian Residents
beaver1	Body Temperature Series of Two Beavers
beaver2	Body Temperature Series of Two Beavers
beavers	Body Temperature Series of Two Beavers
BJsales	Sales Data with Leading Indicator
BJsales.lead	Sales Data with Leading Indicator
BOD	Biochemical Oxygen Demand
cars	Speed and Stopping Distances of Cars
ChickWeight	Weight versus age of chicks on different diets
chickwts	Chicken Weights by Feed Type
CO2	Carbon Dioxide Uptake in Grass Plants
co2	Mauna Loa Atmospheric CO2 Concentration
crimtab	Student's 3000 Criminals Data

<sup>1</sup> <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>

# Pacote datasets do [R]

Dataset	Descrição
discoveries	Yearly Numbers of Important Discoveries
DNase	Elisa assay of DNase
esoph	Smoking, Alcohol and (O)esophageal Cancer
euro	Conversion Rates of Euro Currencies
euro.cross	Conversion Rates of Euro Currencies
eurodist	Distances Between European Cities and Between US Cities
EuStockMarkets	Daily Closing Prices of Major European Stock Indices, 1991-1998
faithful	Old Faithful Geyser Data
fdeaths	Monthly Deaths from Lung Diseases in the UK
Formaldehyde	Determination of Formaldehyde
freeny	Freeny's Revenue Data
freeny.x	Freeny's Revenue Data
freeny.y	Freeny's Revenue Data
HairEyeColor	Hair and Eye Color of Statistics Students
Harman23.cor	Harman Example 2.3
Harman74.cor	Harman Example 7.4
Indometh	Pharmacokinetics of Indomethacin
infert	Infertility after Spontaneous and Induced Abortion
InsectSprays	Effectiveness of Insect Sprays
iris	Edgar Anderson's Iris Data
iris3	Edgar Anderson's Iris Data
islands	Areas of the World's Major Landmasses

# Pacote datasets do [R]

<b>Dataset</b>	<b>Descrição</b>
JohnsonJohnson	Quarterly Earnings per Johnson & Johnson Share
LakeHuron	Level of Lake Huron 1875-1972
Ideaths	Monthly Deaths from Lung Diseases in the UK
lh	Luteinizing Hormone in Blood Samples
LifeCycleSavings	Intercountry Life-Cycle Savings Data
Loblolly	Growth of Loblolly pine trees
longley	Longley's Economic Regression Data
lynx	Annual Canadian Lynx trappings 1821-1934
mdeaths	Monthly Deaths from Lung Diseases in the UK
morley	Michelson Speed of Light Data
mtcars	Motor Trend Car Road Tests
nhtemp	Average Yearly Temperatures in New Haven
Nile	Flow of the River Nile
nottem	Average Monthly Temperatures at Nottingham, 1920-1939
npk	Classical N, P, K Factorial Experiment
occupationalStatus	Occupational Status of Fathers and their Sons
Orange	Growth of Orange Trees
OrchardSprays	Potency of Orchard Sprays
quakes	Locations of Earthquakes off Fiji
randu	Random Numbers from Congruential Generator RANDU
rivers	Lengths of Major North American Rivers
rock	Measurements on Petroleum Rock Samples

# Pacote datasets do [R]

Dataset	Descrição
PlantGrowth	Results from an Experiment on Plant Growth
precip	Annual Precipitation in US Cities
presidents	Quarterly Approval Ratings of US Presidents
pressure	Vapor Pressure of Mercury as a Function of Temperature
Puromycin	Reaction Velocity of an Enzymatic Reaction
Seatbelts	Road Casualties in Great Britain 1969-84
sleep	Student's Sleep Data
stack.loss	Brownlee's Stack Loss Plant Data
stack.x	Brownlee's Stack Loss Plant Data
stackloss	Brownlee's Stack Loss Plant Data
state	US State Facts and Figures
state.abb	US State Facts and Figures
state.area	US State Facts and Figures
state.center	US State Facts and Figures
state.division	US State Facts and Figures
state.name	US State Facts and Figures
state.region	US State Facts and Figures
state.x77	US State Facts and Figures
sunspot.month	Monthly Sunspot Data, from 1749 to "Present"
sunspot.year	Yearly Sunspot Data, 1700-1988
sunspots	Monthly Sunspot Numbers, 1749-1983
swiss	Swiss Fertility and Socioeconomic Indicators (1888) Data

# Pacote datasets do [R]

Dataset	Descrição
Theoph	Pharmacokinetics of Theophylline
Titanic	Survival of passengers on the Titanic
ToothGrowth	The Effect of Vitamin C on Tooth Growth in Guinea Pigs
treering	Yearly Treering Data, -6000-1979
trees	Girth, Height and Volume for Black Cherry Trees
UCBAdmissions	Student Admissions at UC Berkeley
UKDriverDeaths	Road Casualties in Great Britain 1969-84
UKgas	UK Quarterly Gas Consumption
UKLungDeaths	Monthly Deaths from Lung Diseases in the UK
USAccDeaths	Accidental Deaths in the US 1973-1978
USArrests	Violent Crime Rates by US State
UScitiesD	Distances Between European Cities and Between US Cities
USJudgeRatings	Lawyers' Ratings of State Judges in the US Superior Court
USPersonalExpenditure	Personal Expenditure Data
uspop	Populations Recorded by the US Census
VADeaths	Death Rates in Virginia (1940)
volcano	Topographic Information on Auckland's Maunga Whau Volcano
warbreaks	The Number of Breaks in Yarn during Weaving
women	Average Heights and Weights for American Women
WorldPhones	The World's Telephones
WWWusage	Internet Usage per Minute

# UC Irvine Machine Learning Repository<sup>2</sup>

- Aproximadamente 300 datasets disponíveis para a comunidade de aprendizado de máquina (Jun/2014)
- Classificação, Regressão, Clusterização
- Áreas: Ciências da Vida, Ciências Físicas, Ciência da Computação / Engenharia, Ciências Sociais, Administração e Negócios, Jogos e Outras
- Atributos categóricos e numéricos (inteiros e reais)
- Cada dataset dá acesso à pasta com os dados, e informações sobre a estrutura do dataset
- Exemplos:
  - Iris: `http://archive.ics.uci.edu/ml/datasets/Iris`
  - Wine: `https://archive.ics.uci.edu/ml/datasets/Wine`
  - Seeds: `http://archive.ics.uci.edu/ml/datasets/seeds#`

---

<sup>2</sup>`http://archive.ics.uci.edu/ml/`



# Manipulando Dataframes

# Manipulando um dataframe

## ■ Selecionando variáveis de um dataframe pelo nome

```
> <novo> <- <dados>[c(<VAR1>, ..., <VARN>)]
```

## ■ Selecionando variáveis de um dataframe por índice e intervalo.

```
> <novo> <- <dados>[c(<lista de índices e/ou intervalos>)]
```

## ■ Exemplos: <https://www.statmethods.net/management/subset.html>

```
# select variables v1, v2, v3
myvars <- c("v1", "v2", "v3")
newdata <- mydata[myvars]

# another method
myvars <- paste("v", 1:3, sep="")
newdata <- mydata[myvars]

# select 1st and 5th thru 10th variables
newdata <- mydata[c(1,5:10)]
```

# Manipulando um dataframe

## ■ Removendo variáveis de um dataframe usando vetor de booleans

```
> <novo> <- <dados>[c(<Bool1>, ..., <BoolN>)]
```

## ■ Removendo variáveis de um dataframe indicando o índice da coluna

```
> <novo> <- <dados>[c(-<Idx1>, ..., -<IdxN>)]
```

## ■ Removendo variáveis de um dataframe atribuindo NULL à coluna

```
> <novo>$<LABEL.COL> <- NULL
```

## ■ Exemplos: <https://www.statmethods.net/management/subset.html>

```
# exclude variable v2 in v1, v2, v3
newdata <- mydata[c(TRUE, FALSE, TRUE)]

# exclude 3rd and 5th variable
newdata <- mydata[c(-3, -5)]

# delete variables v3 and v5
mydata$v3 <- mydata$v5 <- NULL
```

# Manipulando um dataframe

- **Selecionando observações** de um dataframe segundo as linhas do dataset

```
> <dados> <- <dados>[<Intervalo de Linhas> ,]
```

- **Selecionando observações** de um dataframe segundo o valor dos campos

```
> <dados> <- <dados>[ which(expressão lógica ) ,]
```

- Exemplos: <https://www.statmethods.net/management/subset.html>

```
# first 5 observations  
newdata <- mydata[1:5,]
```

```
# based on variable values  
newdata <- mydata[ which(mydata$gender=='F' & mydata$age > 65),]
```

# Manipulando um dataframe

## ■ Selecionando observações de um dataframe usando **subset**

```
> <dados> <- subset(<dados>, <expressão>, select = c(Labels))
```

## ■ Exemplos: <https://www.statmethods.net/management/subset.html>

```
# using subset function
newdata <- subset(mydata, age >= 20 | age < 10,
                  select=c(ID, Weight) )

# using subset function (part 2)
newdata <- subset(mydata, sex=="m" & age > 25,
                  select=weight:income)
```

# Manipulando um dataframe

## ■ Ordenando um dataframe

```
> <dados> <- <dados>[ order(<dados>$<var>), ]
```

## ■ Exemplos: <https://www.statmethods.net/management/sorting.html>

```
# sort by mpg and cyl  
newdata <- mtcars[order(mtcars$mpg, mtcars$cyl),]  
  
#sort by mpg (ascending) and cyl (descending)  
newdata <- mtcars[order(mtcars$mpg, -mtcars$cyl),]
```

# Manipulando um dataframe

- **Adicionando colunas** um dataframe. É feito juntando-se dois dataframes por meio de uma ou mais variáveis em comum (semelhante à um *inner join*)

```
> <novo.frame> <- merge(<frameA>, <frameB>, by="VAR" )
```

- **Adicionando linhas** à um dataframe. Junta-se dois dataframes verticalmente desde que possuam as mesmas variáveis

```
> <novo.frame> <- rbind(<frameA>, <frameB>)
```

- Exemplos: <https://www.statmethods.net/management/merging.html>

```
# merge two data frames by ID and Country
total <- merge(data frameA,data frameB,by=c("ID","Country"))
```

# Manipulando um dataframe

- **Agregando dados** um dataframe e retorna sumário estatístico da agregação

```
> <resultado> <- aggregate(<dataframe>, by=list(<VARS>),  
fun=<função> )
```

- Exemplos:

<https://www.statmethods.net/management/aggregate.html>

```
# aggregate data frame mtcars by cyl and vs, returning means  
# for numeric variables  
attach(mtcars)  
aggdata <- aggregate(mtcars, by=list(cyl,vs),  
FUN=mean, na.rm=TRUE)  
print(aggdata)  
detach(mtcars)
```



# Importando Datasets como Dataframes

# Importando Arquivo .CSV

- **Importando** um arquivo no formato *comma-separated values* (.csv) e salvando o resultado em um dataframe

```
> <frame> <- read.csv("<file>", header=<bool>), sep="<char>")
```

- Em algumas regiões, a vírgula é usada como separador de decimais. Neste caso alguns arquivos .csv podem utilizar o caracter ';' como separador de campo
- Exemplo:

```
# Le o conteudo do arquivo auto-mpg.csv em dataset
auto <- read.csv("auto-mpg.csv", header=TRUE, sep = ",", ")

# Verifica o nome das variáveis contidas nele
names(auto)

# Outro exemplo, usando ';' como separador de campos
frame <- read.csv("test.csv", sep=";", dec="," )
```

# Importando Arquivo ASCII

## ■ Importando um arquivo no formato ASCII

```
> <frame> <- read.table("<file>", header=<bool>, sep="<char>")
```

## ■ Exemplo:

```
# Carrega o conteúdo do arquivo de dados  
frame <- read.table("dados.txt", header=TRUE)
```

# Gráficos

# Gráficos no R

- Funções Alto Nível: **produzem ou inicializam** um plot. Exemplo:
  - `plot`: gráfico de dispersão XY
  - `hist`: histograma de frequências da amostra
  - `boxplot`: gráfico de caixas mostrando a dispersão da amostra
  - `pairs`: matriz com gráfico de dispersão XY entre pares de variáveis
- Funções Baixo Nível: **adicionam algo** à um plot já criado. Exemplo:
  - `points`: acrescenta pontos ao plot
  - `lines`: acrescenta linhas ao plot
  - `text`: acrescenta texto ao plot
  - `polygon`: acrescenta polígonos ao plot

# Dot plot

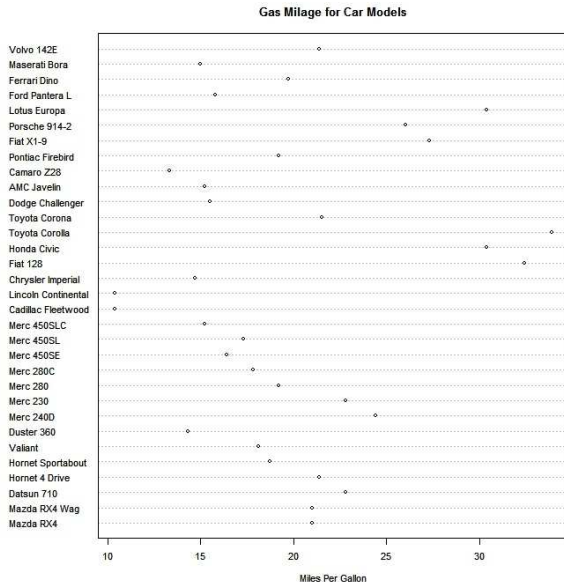
## ■ Plot de pontos

```
> dotchart(<frame>$<VAR>, labels=<vetor>)
```

## ■ Exemplos: <https://www.statmethods.net/graphs/dot.html>

```
# Simple Dotplot  
dotchart(mtcars$mpg, labels=row.names(mtcars), cex=.7,  
main="Gas Milage for Car Models",  
xlab="Miles Per Gallon")
```

# Dot plot



# Bar Plot

## ■ Plot de Gráfico de Barras

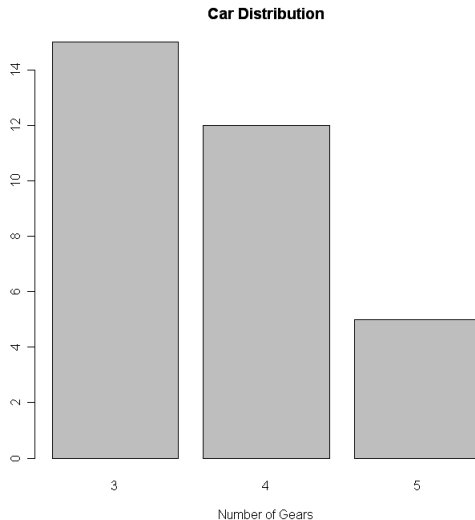
```
> barplot(<vetor>)
```

■ Exemplos: <https://www.statmethods.net/graphs/bar.html>

```
# Simple Bar Plot
counts <- table(mtcars$gear)
barplot(counts, main="Car Distribution",
        xlab="Number of Gears")
```



# Bar plot



# Scatterplot

## ■ Plot de Dispersão XY

```
> plot(<sequencial>, <sequencia2>)
```

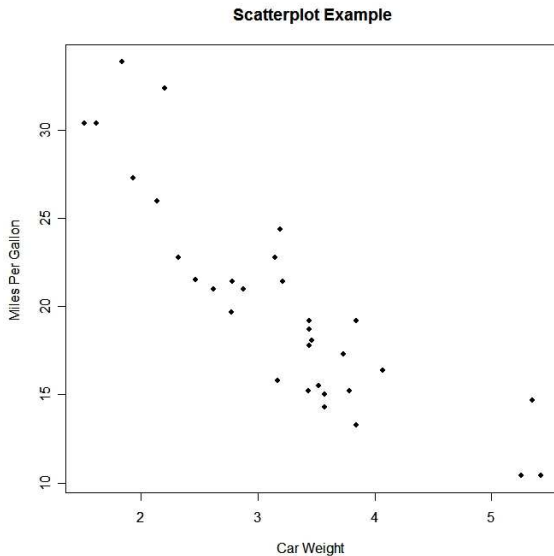
## ■ Parâmetros gráficos:

<https://www.statmethods.net/advgraphs/parameters.html>

## ■ Exemplos: <https://www.statmethods.net/graphs/scatterplot.html>

```
# Simple Scatterplot
attach(mtcars)
plot(wt, mpg, main="Scatterplot Example",
      xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
```

# Scatter plot



# Scatter Plot

## ■ Argumentos típicos do comando `plot`

- `xlab`: rótulo do eixo X
- `ylab`: rótulo do eixo Y
- `main`: título principal do plot
- `col`: cor dos pontos plotados
- `sub`: sub-título do gráfico

# Line Plot

## ■ Plot de Linhas

```
> lines(<sequencial>, <sequencia2>)
```

## ■ Parâmetros gráficos:

<https://www.statmethods.net/advgraphs/parameters.html>

## ■ Exemplo:

```
# Plota a funcao f(x) = x^2 + sin(x)
X <- seq(-10, 10, 0.5)
Y <- X^2 + sin(X)
plot(X,Y,type='n')
lines(X,Y,col='red',type='o')
```

## ■ Parâmetros para type

- p Pontos
- l Linhas
- o Linhas e pontos sobrepostos
- n Não produz pontos ou linhas
- h Linha vertical tipo histograma

# Line Plot

- Exemplos: <https://www.statmethods.net/graphs/line.html>

```
# Convert factor to numeric for convenience
Orange$Tree <- as.numeric(Orange$Tree)
ntrees <- max(Orange$Tree)

# get the range for the x and y axis
xrange <- range(Orange$age)
yrange <- range(Orange$circumference)

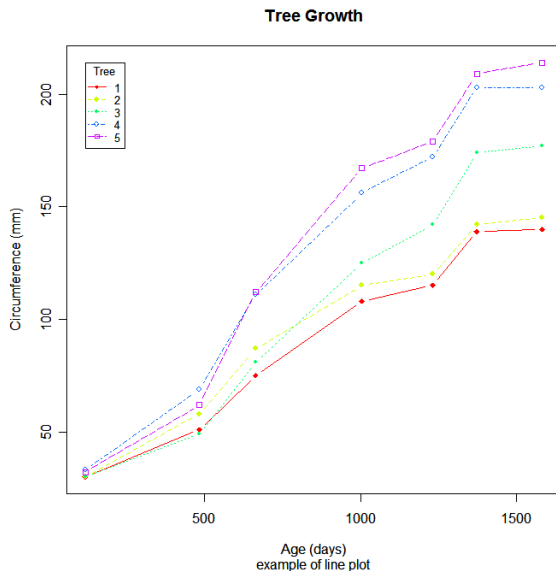
# set up the plot
plot(xrange, yrange, type="n", xlab="Age (days)",
     ylab="Circumference (mm)" )
colors <- rainbow(ntrees)
linetype <- c(1:ntrees)
plotchar <- seq(18,18+ntrees,1)

# add lines
for (i in 1:ntrees) {
  tree <- subset(Orange, Tree==i)
  lines(tree$age, tree$circumference, type="b", lwd=1.5,
        lty=linetype[i], col=colors[i], pch=plotchar[i])
}

# add a title and subtitle
title("Tree Growth", "example of line plot")

# add a legend
legend(xrange[1], yrange[2], 1:ntrees, cex=0.8, col=colors,
      pch=plotchar, lty=linetype, title="Tree")
```

# Line Plot



# Pie Chart

## ■ Pie Chart

```
> pie(<vetor>, labels=<vetLab>)
```

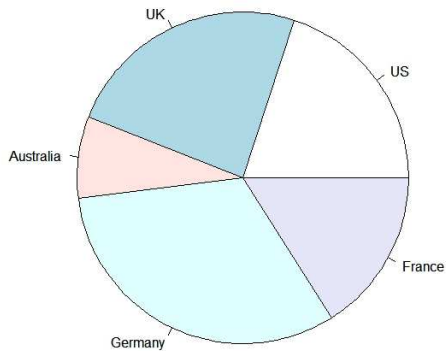
■ Exemplos: <https://www.statmethods.net/graphs/pie.html>

```
# Simple Pie Chart  
slices <- c(10, 12, 4, 16, 8)  
lbls <- c("US", "UK", "Australia", "Germany", "France")  
pie(slices, labels = lbls, main="Pie Chart of Countries")
```



# Pie Chart

**Pie Chart of Countries**



# Análise Descrita

# Estatística Descritiva

- Obtendo sumário estatístico das variáveis mediante função

```
> sapply(<frame>, fun=<função>)
```

```
# get means for variables in data frame mydata  
# excluding missing values  
sapply(mydata, mean, na.rm=TRUE)
```

- Funções que podem ser usadas:

- mean
- sd
- var
- min
- max
- median
- range
- quantile

# Estatística Descritiva

- Obtendo sumário estatístico das variáveis com algumas medidas de dispersão

```
> summar (<frame>)
```

```
# mean, median, 25th and 75th quartiles, min, max
```

```
summary (mydata)
```

```
# Tukey min, lower-hinge, median, upper-hinge, max
```

```
fivenum (x)
```

# Histogramas e Boxplots

# Boxplot

- Gráfico de caixas com min, max, mediana, Q1, Q3 e *outliers*

```
> boxplot(<sequencia>)
```

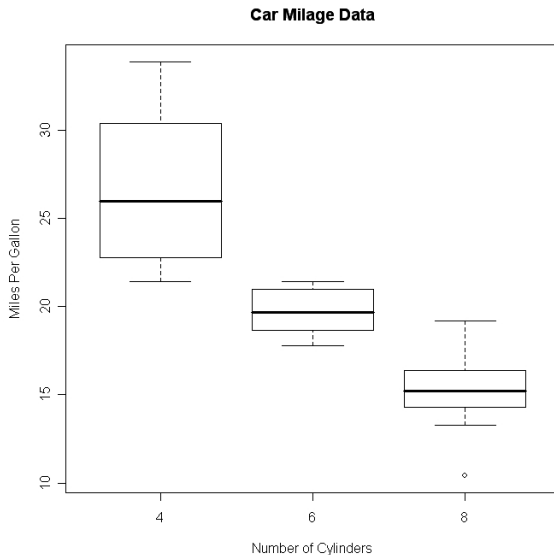
- Example

```
# Boxplot of MPG by Car Cylinders
boxplot(mpg~cyl,data=mtcars, main="Car Milage Data",
        xlab="Number of Cylinders", ylab="Miles Per Gallon")

# Another example
dados <- rexp(1000, 0.25)
boxplot(dados)

# Plotando a largura da pétala de iris por espécie
boxplot(iris$Petal.Width ~ iris$Species)
```

# Boxplot Chart



# Histogramas

- Constrói um histograma de frequências ou densidade dos dados

```
> hist(<sequencia>, freq=<Bool>)
```

- Example

```
# Histograma de uma amostra normal, com media 15 e desvio 2
hist(rnorm(1000,15,2))

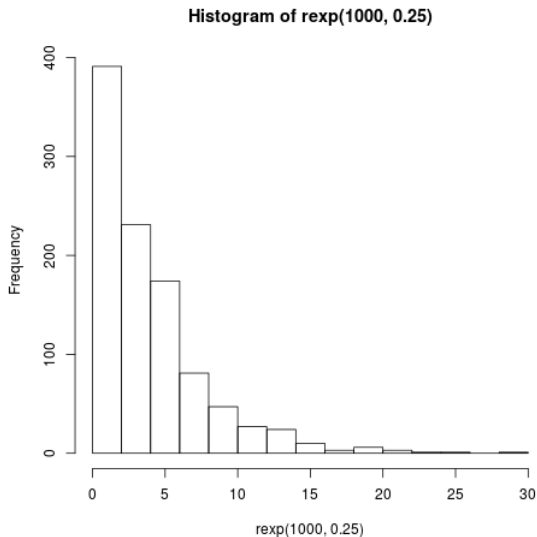
# Histograma de uma amostra exponencial, com lambda=0.25
hist(rexp(1000,0.25))

# Histograma da variável Speed do dataset cars
hist(cars$speed)

# Histograma da variável Speed do dataset cars, usando densidade
hist(cars$speed, freq=FALSE)
```



# Histograma



# Regressão Linear

# Modelo de Regressão Linear Simples

- Modela dados na forma  $y = ax + b + \text{resíduo}$

- Utiliza comando `lm` do R

- Exemplo com vetores:

<https://www.theanalysisfactor.com/linear-models-r-plotting-re>

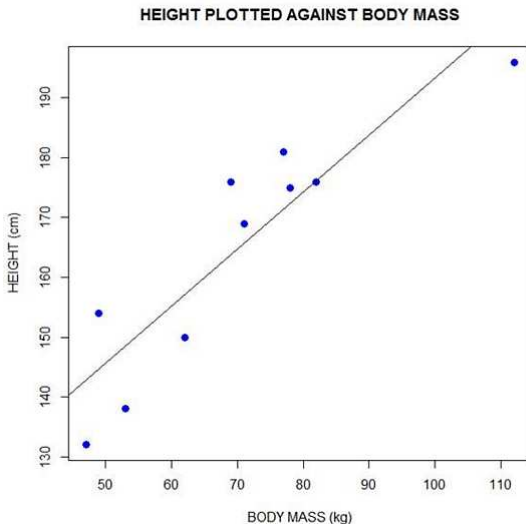
```
# Dados
height <- c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
bodymass <- c(82, 49, 53, 112, 47, 69, 77, 71, 62, 78)

# Plot
plot(bodymass, height, pch = 16, cex = 1.3, col = "blue",
      main = "HEIGHT PLOTTED AGAINST BODY MASS",
      xlab = "BODY MASS (kg)", ylab = "HEIGHT (cm)")

# Ajuste linear
fit <- lm(height ~ bodymass)
fit

# Acrescenta a reta de regressão com 'a' e 'b' calculados por lm
abline(lm(height ~ bodymass))
```

# Modelo de Regressão Linear Simples



# Modelo de Regressão Linear

- Podemos usar uma fórmula para associar uma variável independente com uma variável dependente
- Fórmula:  $Y \sim X$

```
> lm(<formula>, data=<dataframe>)
```

- Exemplo usando fórmula

```
# Ajustando a velocidade em função da distância no dataset cars
fit <- lm(speed ~ dist, cars)

# Apresenta o sumário dos dados ajustados
summary(fit)

# Multiple Linear Regression Example
fit <- lm(y ~ x1 + x2 + x3, data=mydata)
summary(fit) # show results
```

# Modelo de Regressão Linear

## ■ Exemplo sobre o dataset `trees`

```
# Inicia link com o dataset
attach(trees)

# Gera modelo de regressão para explicar Circunferencia em função de Volume
fit <- lm(Girth ~ Volume)

# Gera dataframe com valor a estimar, com apenas 1 observação
newdata <- data.frame(Volume=64)

# Faz a predição
predict(fit, newdata, interval="predict")

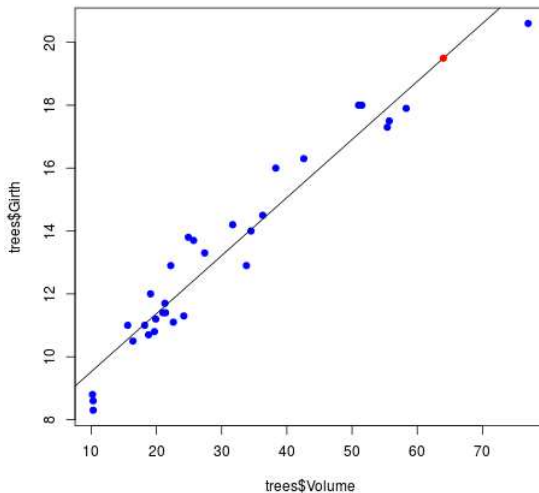
# Guarda resultado para plotar
ptoX <- 64
ptoY <- predict(fit, newdata, interval="predict")[1]

# Retira o link com o dataset
detach(trees)

# Plotando a reta de regressão juntamente com os dados do modelo
plot(trees$Volume, trees$Girth, col='blue', pch=19)
abline(fit)

# Plota ponto estimado de vermelho
points(ptoX, ptoY, col='red', pch=19)
```

# Plot com Reta de Regressão



# Anexo - Revisão Teórica



# Estatística Descritiva

## Definição (População)

Uma **população** consiste na totalidade das observações com o qual estamos lidando. O número de observações de uma população é chamado de **tamanho** da população. Frequentemente usamos distribuições de probabilidade como um modelo de população.

## Definição (Amostra)

Uma **amostra** é um subconjunto de observações selecionadas da população

- As variáveis aleatórias  $X_1, X_2, \dots, X_n$  são amostra aleatória de tamanho  $n$  se
  - (a) Os  $X_i$ 's são variáveis aleatórias independentes;
  - (b) Cada  $X_i$  tem a mesma distribuição de probabilidade.

## Definição (Estatística)

Uma **estatística** é qualquer função das observações de uma amostra aleatória

# Estatística Descritiva

## Definição (Média Populacional)

Seja uma população finita com  $N$  observações. A **média populacional** é dada por:

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i$$

## Definição (Média Amostral)

Seja  $x_1, x_2, \dots, x_n$  observações de uma amostra de tamanho  $n$ . A **média amostral** é dada por:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

# Estatística Descritiva

## Definição (Variância Populacional)

Seja uma população finita com  $N$  observações. A **variância populacional** é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

O desvio padrão populacional  $\sigma$  é a raiz quadrada positiva da variância populacional.

## Definição (Variância Amostral)

Seja  $x_1, x_2, \dots, x_n$  observações de uma amostra de tamanho  $n$ . A **variância amostral** é dada por:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

O desvio padrão amostral  $s$  é a raiz quadrada positiva da variância amostral.

# Estatística Descritiva

## Definição (Mediana)

A **mediana** de uma amostra é uma medida de tendência central que divide os dados em duas partes iguais, metade abaixo da mediana, e metade acima da mediana.

## Definição (Moda)

A **moda** é a observação de ocorrência mais frequente da amostra.

## Definição (Quartis)

Seja um conjunto de observações ordenada e dividida em 4 partes iguais. A divisão destes pontos são denominadas **quartis**. São eles:

$Q_1$  : valor que possui aproximadamente 25% das observações abaixo e 75% das observações acima deste valor (**quartil inferior**). Posição calculada por  $\frac{(n+1)}{4}$

$Q_2$  : valor que possui aproximadamente 50% das observações abaixo e 50% das observações acima deste valor (corresponde à **mediana**)

$Q_3$  : valor que possui aproximadamente 75% das observações abaixo e 25% das observações acima deste valor (**quartil superior**). Posição calculada por  $\frac{3(n+1)}{4}$

# Estatística Descritiva

## Definição (Percentis)

O  **$k$ -ésimo percentil** é o valor de dados tal que aproximadamente  $100k\%$  das observações estão abaixo ou neste valor, e aproximadamente  $100(1 - k)\%$  das observações estão acima dele.

## Definição (Intervalo Interquartis)

O **intervalo interquartis** é uma medida de variabilidade da amostra construído a partir da diferença  $IQR = Q_3 - Q_1$  (também denominado **interquartile range**)

# Estatística Descritiva

## Definição (Distribuição de Frequências)

Uma **distribuição de frequências** é uma tabela que exhibe a frequência dos vários resultados de uma amostra. São construídas a partir da divisão da amplitude da amostra em intervalos, denominados **intervalos de classe**, **celulas** ou **bins**:

$$k = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil$$

Existem diversas heurísticas para escolher o número  $k$  de bins:

■ Regra da Raíz Quadrada:

$$k = \sqrt{n}$$

■ Regra de Sturges:

$$k = \lceil \log_2(n) + 1 \rceil$$

■ Regra de Rice:

$$k = \lceil 2n^{1/3} \rceil$$

■ Regra de Scott:

$$h = \frac{3.5s}{n^{1/3}}$$

■ Regra de Freedman-Diaconis:

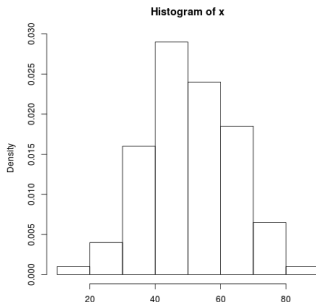
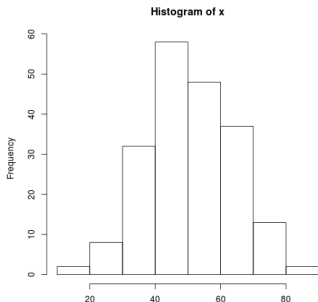
$$h = 2 \frac{IRQ}{n^{1/3}}$$

# Estatística Descritiva

## Definição (Histograma)

Um **histograma** é uma exibição visual de uma distribuição de frequências, construído pelos seguintes passos:

- (1) Rotule o limite de cada bin em um eixo vertical
- (2) Marque e rotule na escala vertical com as frequências (ou frequências relativas)
- (3) Acima de cada bin, desenhe um retângulo onde a altura é igual à frequência (ou frequência relativa) correspondente ao bin



# Estatística Descritiva

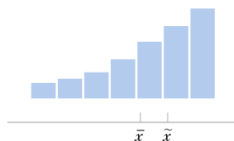
## Definição (Assimetria)

A **assimetria** mede a falta de simetria de uma distribuição de dados. Uma distribuição é simétrica se ela se parece a mesma do lado direito e esquerdo de seu ponto central.

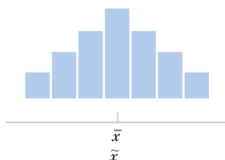
$$g_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3 / n}{s^3}$$

Quando o conjunto de dados é uma amostra, a assimetria é dada por

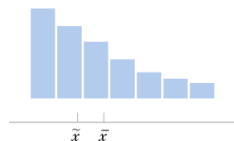
$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$



Negative or left skew  
(a)



Symmetric  
(b)



Positive or right skew  
(c)



## Definição (Curtose)

A **curtose** mede a forma de uma distribuição de probabilidade ou conjunto de dados.

$$Kurtosis = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 / n}{s^2}$$

A medida de **excesso de curtose** é dada por

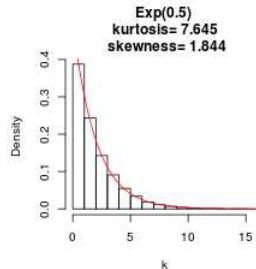
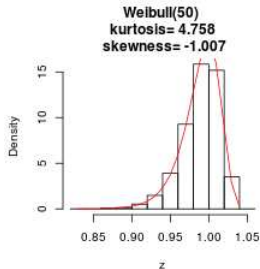
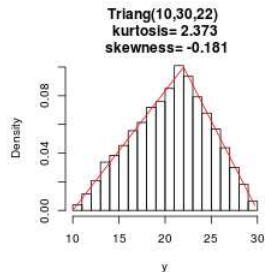
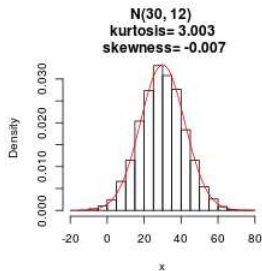
$$g_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 / n}{s^2} - 3$$

Classificação:

- $g_2 = 0$  (ou  $Kurtosis = 3$ ): mesocúrticas
- $g_2 > 0$  (ou  $Kurtosis > 3$ ): leptocúrticas
- $g_2 < 0$  (ou  $Kurtosis < 3$ ): platicúrticas (caudas pesadas)

# Estatística Descritiva

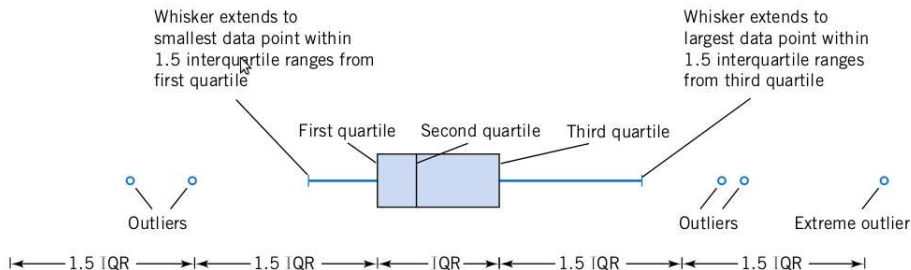
- Exemplos de cálculo de assimetria e curtose para 4 distribuições contínuas



# Estatística Descritiva

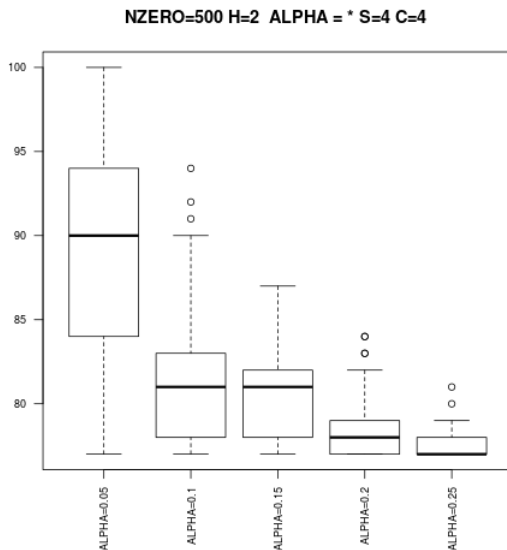
## Definição (Boxplot)

Um **boxplot** é uma exibição gráfica que descreve simultaneamente diversas características importantes sobre o conjunto de dados, tais como medida central, espalhamento, simetria e identificação de observações atípicas (outliers)



# Estatística Descritiva

- Boxplots são úteis em comparação de diferentes conjuntos de dados. Exemplo:



# Tratamento dos Dados

- Posição: média, mediana, moda, mínimo, máximo
- Dispersão: variância, amplitude, coeficiente de variação, coeficiente de assimetria
- Outliers: observações atípicas
  - Erro na coleta de dados
  - Eventos raros
- Tempos (seg) de chegadas sucessivas de clientes em supermercado

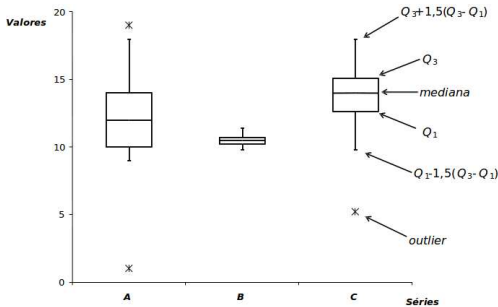
11	5	2	0	9	9	1	5	5	1
1	3	3	3	7	4	12	8	7	5
5	2	6	1	11	1	2	4	4	2
2	1	3	9	0	10	3	3	4	5
1	5	18	4	22	8	3	0	4	4
8	9	2	3	12	1	3	1	11	9
7	5	14	7	7	28	1	3	3	4
2	11	13	2	0	1	6	12	8	12
15	0	6	7	19	1	1	9	12	4
1	5	3	17	10	15	43	2	9	11
6	1	13	13	19	10	9	20	17	24
19	2	27	5	20	5	10	8	728	8
2	3	1	1	4	3	6	13	12	12
10	9	1	1	3	9	9	4	6	3
0	3	6	3	27	3	18	4	4	7
6	0	2	2	8	4	5	1	3	1
4	18	1	0	16	20	2	2	9	3
2	12	28	0	7	3	18	12	2	1
3	2	8	3	19	12	5	4	0	3
6	0	5	0	3	7	0	8	5	8

# Outliers

- Outliers afetam os resultados, distorcem as estimativas e nível de significância dos testes, e conduzem a conclusões errôneas sobre o processo
- Tempos (seg) de chegada sucessivas clientes em supermercado

	c/ outlier	s/ outlier
Média	10.44	6.83
Mediana	5	5
Variância	2643.81	43.60

- Podem ser detectados a partir de um intervalo construído sobre os quartis:



# Referências Bibliográficas

# Materiais Consultados



## R Tutorial

Quick-R: Acessing the power of R

<https://www.statmethods.net/r-tutorial/index.html>



## Lauretto, M. S.

Introdução à Análise de Dados Utilizando o Ambiente R

<http://each.uspnet.usp.br/lauretto/cursoR2015/cursoR2015.pdf>



## Jelihovschi, E.

Análise Exploratória de Dados usando R - UESC. ISBN: 9788574553702

<http://www.uesc.br/editora/livrosdigitais2/analiseexploratoria>



## Mourthé, I.

Análise Exploratória de Dados usando o R

<http://sbprimatologia.org.br/wp-content/uploads/2017/09/Italo->



## Carvalho, P.S., e Tandel, M.C.F.F.

Introdução ao Ambiente Computacional R

<http://www.ufscar.br/~des/docente/josemar/152056/05082008/tuto>



# Materiais Consultados



## R Graph Gallery

Site: The R Graph Gallery

<https://www.r-graph-gallery.com/>



## J H Maindonald, J. H.

Using R for Data Analysis and Graphics: Introduction, Code and Commentary

<https://cran.r-project.org/doc/contrib/usingR.pdf>



## R Development Core Group

Graphics with R

<https://csg.sph.umich.edu/docs/R/graphics-1.pdf>