

Probabilidades e Estatística

Capítulo 1. Análise exploratória de dados: Introdução ao R

Conceição Amado e Isabel M Rodrigues

Instituto Superior Técnico

Lisboa, 2023

Análise Exploratória de Dados (AED)

A finalidade da AED é:

- examinar os dados previamente à aplicação de qualquer técnica estatística
- deste modo o analista consegue um entendimento básico dos seus dados e das relações existentes entre as variáveis analisadas.

Após a coleta e a digitação de dados em uma base de dados apropriada, o próximo passo é a **análise descritiva**. Esta etapa é fundamental, pois uma **análise descritiva** detalhada permite ao pesquisador familiarizar-se com os dados, organizá-los e sintetizá-los de forma a obter as informações necessárias do conjunto de dados para responder as questões que estão sendo estudadas - **Estatística Descritiva**

Estatística Descritiva

Conceitos básicos em Estatística:

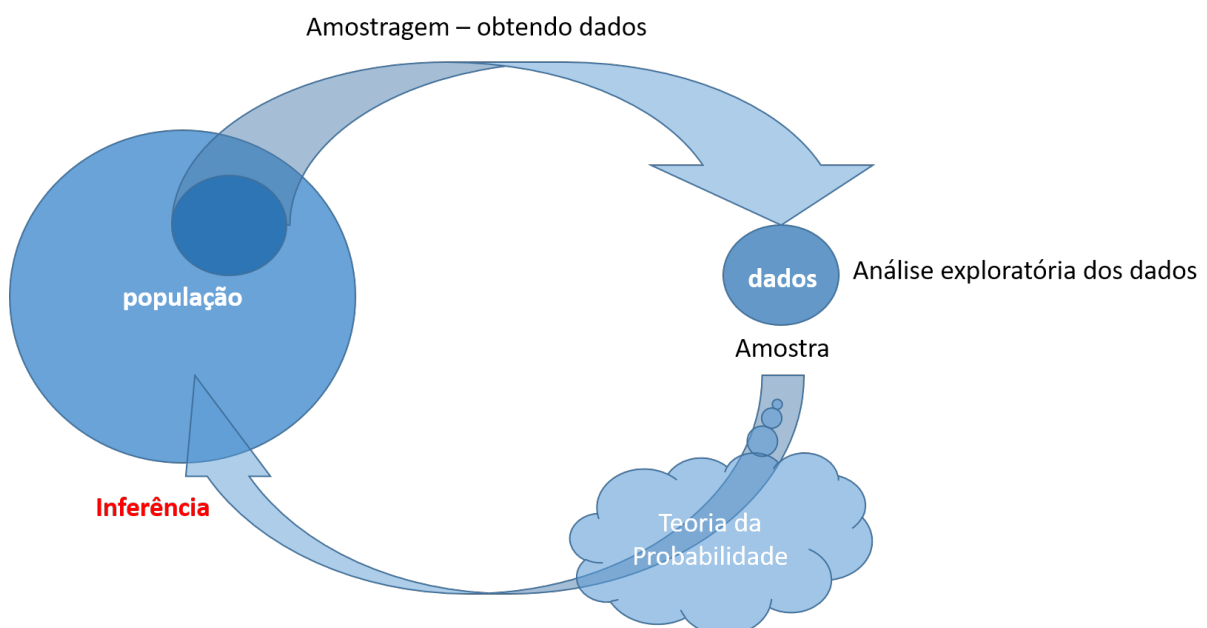
- **População ou universo:** conjunto de todos os elementos que têm alguma característica em comum (ex: todos os alunos do IST)
- **Unidades estatística:** elemento da população (aluno do IST)
- **Variável:** característica de interesse em estudo (ex: X - altura de os alunos do IST e x - altura observada de um aluno do IST (**dado**)).
- **Amostra:** subconjunto da população observado, $\mathbf{x} = (x_1, \dots, x_n)$

Objetivos da Estatística Descritiva:

- condensar os dados observados, $\mathbf{x} = (x_1, \dots, x_n)$, em tabelas;
- calcular indicadores de localização e de dispersão;
- fazer representações gráficas.

9 / 44

Estatística



Algumas Etapas da AED

- preparar os dados para serem acessíveis a qualquer técnica estatística
- realizar um exame gráfico da natureza das variáveis individuais a analisar e uma análise descritiva que permita quantificar alguns aspectos gráficos dos dados
- realizar um exame gráfico das relações entre as variáveis analisadas e uma análise descritiva que quantifique o grau de inter-relação entre elas
- identificar os possíveis casos atípicos (*outliers*)
- avaliar, se for necessário, a presença de dados ausentes (*missing*)

11 / 44

Tipos de variáveis

Variável: Qualquer característica associada a uma população

Classificação da variável:

- **Qualitativa:** são aquelas que apresentam como possíveis realizações uma qualidade ou atributo do indivíduo pesquisado
 - **Nominal:** sexo, cor dos olhos
 - **Ordinal:** classe social, grau de instrução
- **Quantitativa:** são aquelas que apresentam como possíveis realizações números resultantes de uma contagem ou mensuração
 - **Contínua:** peso, altura
 - **Discreta:** número de filhos, número de carros

12 / 44

Instalação do e do Rstudio

Instalação

- Instalar o programa R (<http://www.r-project.org>). Na secção download escolher qual o seu sistema operativo (Mac, Linux ou Windows) e qual o seu processador 32- ou 64-bits.
- Na página r-project.org, no lado esquerdo debaixo do texto download, aparece a opção CRAN e na nova página deve escolher um dos servidores internacionais.
- O programa R dispõe de uma interface gráfica própria. Porém, utilizaremos um interface gráfico avançado (IDE-Integrated Development Environment) que se chama RStudio (<https://www.rstudio.com/products/rstudio/download/>).

Ver também o documento *Instructions for installing R and RStudio.pdf* em https://web.tecnico.ulisboa.pt/~ist13493/PE_aulas2023/R_Material_exerciciosR/

13 / 44

Algumas estruturas de dados em

Estruturas de dados

- Um vector `c()`
- As matrizes correspondem a um conjunto de elementos do mesmo tipo definida através de linhas e colunas, `matrix`
- As `arrays` apresentam as mesmas características que as matrizes, mas apresentam a possibilidade de terem mais de duas dimensões.
- As `listas` são conjuntos de dados que podem ser de qualquer tipo.
- Uma `data frame` corresponde a um conjunto de vetores de igual tamanho. Esses vetores não têm de ser necessariamente do mesmo tipo de dados. As data frames são utilizadas para armazenar tabelas de dados.
- Um `factor` é um vector especializado para dados categóricos.

14 / 44

No RStudio e alguns comandos

- Linha de comandos (Console)
- Subjanela de script \Rightarrow mais adequado
- Se o resultado for uma variável ou um gráfico, estes vão aparecer nas subjanelas: Environment ou Plot.
- Concatenar e selecionar conj. de valores
Exemplo 1: `x <-c(1:10) x <-c(1,10)`
Exemplo 2: `y <- c(65.2, 73.2, 66.3, 56.7), y[1:3] y[y>60]`
- Construir matriz
Exemplo 1: `y <- matrix(1:6, nrow=3, ncol=2)`
Exemplo 2: `mat <- c(10, -3, 42, -10)`
`namesL <- c("l1", "l2") namesC <- c("C1", "C2")`
`matfinal <- matrix(mat, nrow=2, ncol=2, byrow=TRUE, dimnames=list(namesL, namesC))`

15 / 44

Algumas estruturas de dados em

Estruturas de dados: exemplo data frame

- `Alunos = as.factor(c("Pedro", "Maria", "João", "Ana"))`
- `Idade = c(15, 18, 22, 17)`
- `Estudos = as.factor(c("FIS", "MAT", "AMB", "INF"))`
- `dados.my = data.frame(Alunos, Idade, Estudos)`

	Alunos	Idade	Estudos
1	Pedro	15	FIS
2	Maria	18	MAT
3	João	22	AMB
4	Ana	17	INF

16 / 44

Variáveis Quantitativas - Indicadores amostrais

Indicadores numéricos (ou amostrais) de $\mathbf{x} = (x_1, \dots, x_n)$

- Medidas de **localização amostrais**: média, moda, quantis, mediana e quartis
- Medidas de **dispersão amostrais**: variância, desvio padrão e coeficiente de variação

Medidas de localização

- **Média**: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \Rightarrow \mathbb{R}$ `mean(variavelnome)`
- **Moda (mo)**: é o valor (ou atributo) que ocorre com maior frequência $\Rightarrow \mathbb{R}$ (ver no Exercício 1)
- **Quantil- α (q_α)**: chama-se quantil de ordem α , ($0 < \alpha < 1$), o valor q_α tal que há uma proporção α de observações inferiores ou iguais a q_α e uma proporção $(1 - \alpha)$ de observações maiores ou iguais a esse valor.

17 / 44

Variáveis Quantitativas - Indicadores amostrais

Seja $\mathbf{x} = (x_1, \dots, x_n)$, denota-se por $x_{(1)} \leq x_{(2)}, \dots, \leq x_{(n)}$ a amostra ordenada, i.e., $x_{(1)} = \min \mathbf{x}, \dots, x_{(n)} = \max \mathbf{x}$

$$q_\alpha = \begin{cases} \frac{x_{(n\alpha)} + x_{(n\alpha+1)}}{2}, & n\alpha \text{ inteiro} \\ x_{([\!n\alpha\!] + 1)}, & n\alpha \text{ não inteiro} \end{cases},$$

onde $[.]$ é a parte inteira do número.

$\Rightarrow \mathbb{R}$ `quantile(variavelnome, type=2)`

- **Mediana**: é o valor que divide a amostra ordenada em duas partes iguais (com o mesmo número de observações cada), i.e. a mediana é o $q_{\frac{1}{2}}$ $\Rightarrow \mathbb{R}$ `median(variavelnome)`
- **Quartis**: considerando a amostra ordenada dividida em quatro partes, cada uma com o mesmo número de observações, os pontos da divisão chamam-se quartis
 - $q_{\frac{1}{4}}$ é o 1º quartil
 - $q_{\frac{2}{4}}$ é a mediana (2º quartil)
 - $q_{\frac{3}{4}}$ é o 3º quartil

18 / 44

Média

```
# Criar um vector - valores de uma variável univariada
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Calcular a média
out.mean <- mean(x)
print(out.mean)
[1] 8.22
```

Quantis

```
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Calcular quantis
out.quantis <- quantile(x,type=2) # type=2 usa regra de acordo
                                     com as expressões acima
print(out.quantis)
      0%   25%   50%   75%  100%
-21.0   2.0   5.6  12.0  54.0
# Apenas a mediana
median(x)
[1] 5.6
```

19 / 44

Variáveis Quantitativas - Indicadores amostrais

Medidas de dispersão

- **Amplitude (R)**: diferença entre o valor máximo e o valor mínimo, $R = x_{(n)} - x_{(1)}$
- **Amplitude inter-quartis (IQR)**: $IQR = q_{\frac{3}{4}} - q_{\frac{1}{4}}$
- **Variância (s_x^2 ou s^2)**: média dos quadrados dos desvios em relação à média aritmética, $s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
- **Desvio padrão (s_x ou s)**: $s_x = +\sqrt{s_x^2}$ (mesma unidade de medida da média)
- **Coeficiente de variação (cv)**: dispersão relativa numa escala independente da unidade de medida ou da ordem de grandeza da variável e que só se calcula quando a variável toma valores de um só sinal; todos positivos ou todos negativos, $cv = \frac{s}{\bar{x}}$

20 / 44

Amplitude e IQR

```
# Criar um vector - valores de uma variável univariada
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Calcular a amplitude
out.amplitude <- range(x)
print(out.amplitude)
[1] -21 54
```

Variância, desvio padrão e coeficiente de variação

```
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Calcular a variância
out.var <- var(x)
print(out.var)
[1] 368.6618
# Desvio padrão
out.sd<-sd(x)
print(out.sd)
[1] 19.20057
# Coeficiente de variação
cv<-sd(x)/mean(x)
cv
[1] 2.335835
```

21 / 44

Variáveis Quantitativas - Indicadores amostrais

Duas variáveis observadas na unidade estatística

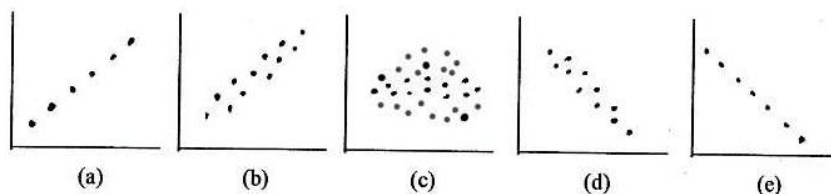
amostra bivariada: $(x_1, y_1), \dots, (x_n, y_n)$

Objetivo: Estudo simultâneo de duas séries de observações, pondo em evidência “relações” (associação linear) existentes entre elas

- **Covariância de x e y (s_{xy}):** $s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
 - $s_{xy} = 0 \implies$ x e y não associados linearmente
 - $s_{xy} > 0 \implies$ associação linear positiva
 - $s_{xy} < 0 \implies$ associação linear negativa
- **Coeficiente de correlação linear x e y (r_{xy}):** não é afetado, em valor absoluto, por transformações lineares e é adimensional
 $r_{xy} = \frac{s_{xy}}{s_x s_y}$, com $-1 \leq r_{xy} \leq 1$

Variáveis Quantitativas - Indicadores amostrais

- (a) $r = 1 \implies$ pontos observados sobre uma recta de declive positivo
- (b) $r \approx 1 \implies$ pontos observados próximos de uma recta de declive positivo
- (c) $r \approx 0 \implies$ a nuvem apresenta um aspecto arredondado ou alongado segundo um dos eixos.
- (d) $r \approx -1 \implies$ pontos observados próximos de uma recta de declive negativo
- (e) $r = -1 \implies$ pontos observados sobre uma recta de declive negativo.



23 / 44

Em 

Covariância e Coeficiente de Correlação

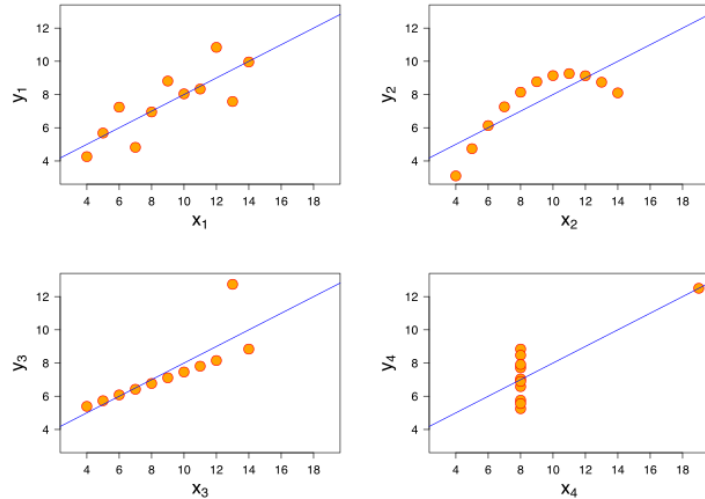
```
tempos <- c(16,24,32,40,48,48,48,48,56,64,72,80)
resistencias <- c(199,214,230,248,255,262,279,267,305,298,323,359)
# Calcular a covariância amostral
cov(tempos,resistencias)
[1] 843.6364
# Calcular o coeficiente de correlação
cor(tempos,resistencias)
[1] 0.9794227
```

24 / 44

Visualização dos dados

A importância de visualizar dados antes de aplicar testes estatísticos inferenciais e a limitação das estatísticas descritivas: Quarteto de Anscombe, para 4 conjuntos de dados temos praticamente as mesmas estatísticas descritivas (Tabela) mas ...

Propriedade	Valor
Média de x	9
Variância de x	11
Média de y	7,50
Variância de y	4,125
Correlação entre x e y	0,816
Reta de regressão linear	$y = 3,00 + 0,500x$
Coefficiente de determinação da regressão linear: R^2	0,67



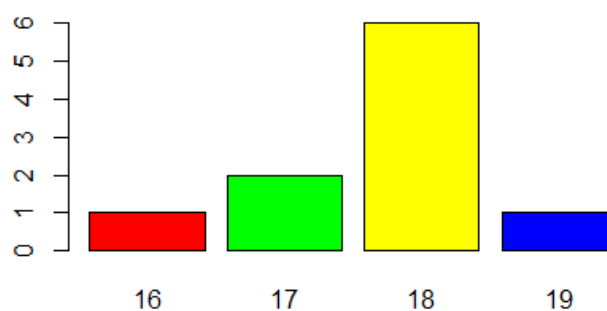
25 / 44

Alguns Métodos Gráficos: diagrama de barras (*barplot*)

Diagrama de barras → para dados univariados de natureza discreta ou categórica

Diagrama de barras

```
age <- c(17,18,18,17,18,19,18,16,18,18)
table(age)
barplot(table(age), col = c("red", "green", "yellow", "blue"))
```



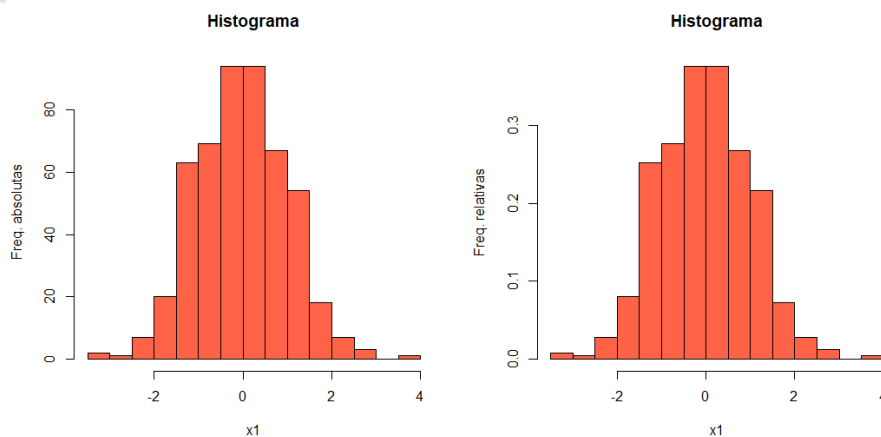
26 / 44

Alguns Métodos Gráficos: histograma

Histograma → para dados de natureza contínua (dados agrupados)

Histograma

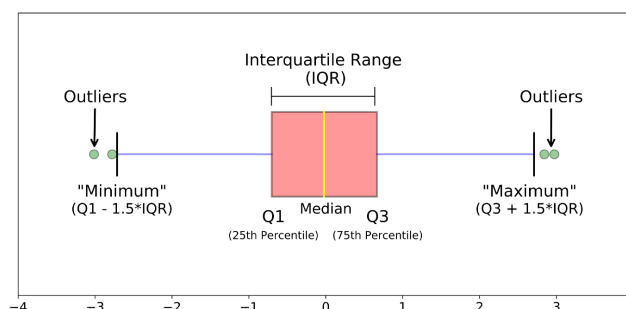
```
set.seed(2022)
x1 <- rnorm(500)
par(mfrow = c(1, 2))
hist(x1,col="tomato",pch=20,cex=4,breaks=15,ylab="Freq. absolutas"
     main="Histograma")
hist(x1,freq=FALSE,col="tomato",pch=20,cex=4,breaks=15,
     ylab="Freq.relativas",main="Histograma")
```



27 / 44

Alguns Métodos Gráficos. Caixa de bigodes (*Boxplot*)

- Fornece informações sobre posição, dispersão, assimetria, caudas e valores discrepantes (*outliers*)
- Para construí-la, desenha-se uma caixa com comprimento $IQR = q_{3/4} - q_{1/4}$. A mediana ($q_{1/2}$) é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até os valores máximo e mínimo, isto se não existirem observações discrepantes
- A observação x é classificada como discrepante (*outlier*) se:
 $x < q_{1/4} - 1.5 \times IQR$ ou $x > q_{3/4} + 1.5 \times IQR$



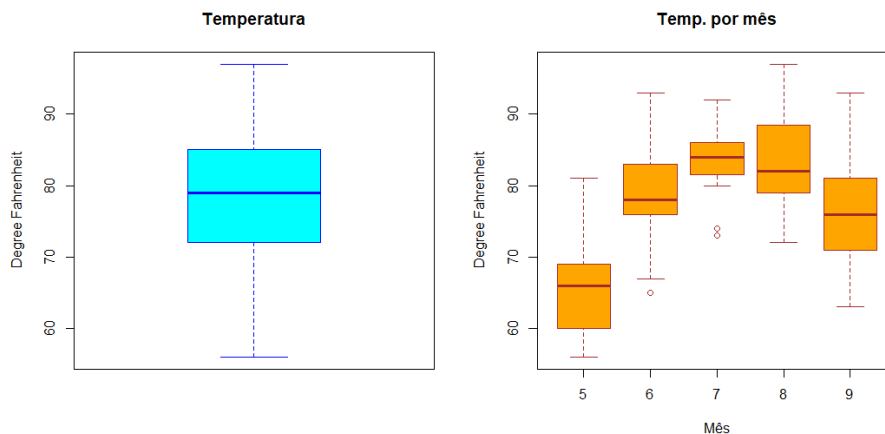
28 / 44

Alguns Métodos Gráficos. Caixa de bigodes (*Boxplot*)

Boxplot

```
par(mfrow = c(1, 2))  
boxplot(airquality$Temp,  
data=airquality,  
main="Temperatura",  
ylab="Degree Fahrenheit",  
col="cyan",  
border="blue")
```

```
boxplot(Temp~Month,  
data=airquality,  
main="Temp. por mês",  
xlab="Mês",  
ylab="Degree Fahrenheit",  
col="orange",  
border="brown")
```



29 / 44

Alguns Métodos Gráficos: Função distribuição empírica ($F_n(x)$)

- É uma função definida para todo o número x real e que para cada x dá a proporção de elementos da amostra menores ou iguais a x :

$$F_n(x) = \frac{\# \text{ observações } \leq x}{n}$$

- Para construí-la, seguem-se as seguintes etapas:
 - 1) Ordenar os n elementos da amostra: $(x_{(1)}, \dots, x_{(n)})$;
 - 2) Considerar um sistema de eixos coordenados e marcar no eixo do xx os valores da amostra ordenada;
 - 3) Começar a desenhar a função da esquerda para a direita, atribuindo o valor 0 à esquerda de $x_{(1)}$, o valor $1/n$ entre $x_{(1)}$ e $x_{(2)}$, o valor $2/n$ entre $x_{(2)}$ e $x_{(3)}$, e assim sucessivamente até esgotarmos todos os valores da amostra. Para um valor igual ou superior a $x_{(n)}$, a função toma o valor 1. Se na amostra um valor se repete d vezes, então o salto da função nesse ponto será d/n , em vez de $1/n$.

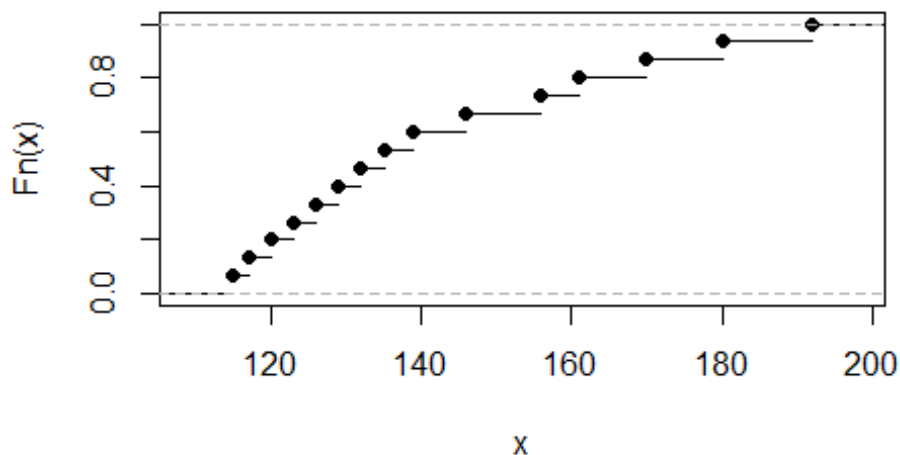
30 / 44

Alguns Métodos Gráficos: Função distribuição empírica ($F_n(x)$)

Função distribuição empírica

```
x<-c(115, 117, 120, 123, 126, 129, 132, 135, 139, 146, 156,  
161, 170, 180, 192)  
plot(ecdf(x), main="Função distribuição empírica")
```

Função distribuição empírica



31 / 44

Gráficos com a livreria do ggplot2

Links

- <https://www.r-graph-gallery.com/index.html>
- <https://ggplot2-book.org/index.html>
- <https://www.gapminder.org/>

Gráficos com a função GGPLOT - estrutura

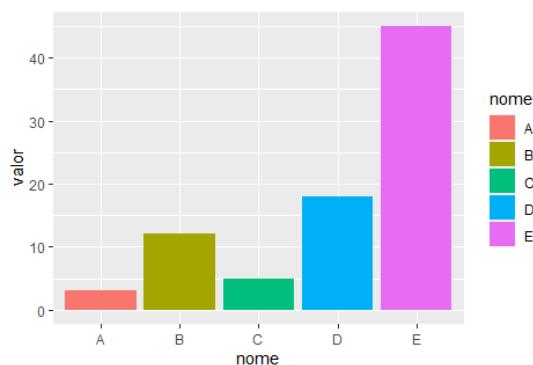
- A função `ggplot` permite definir os parâmetros iniciais do gráfico.
`ggplot(data=dados, aes(x= ... , y=...))`
A função `geom_xxx()` define o tipo de gráfico
- Exemplos: `xxx = density, dotplot, point, histogram, bar, dotplot, violin, line, freqpoly, ...`

32 / 44

Barplot - ggplot2

Barplot

```
library(ggplot2) # ativa as funções desta package
data <- data.frame(
  nome<-as.factor(c("A","B","C","D","E")),
  valor<-c(3,12,5,18,45))
# Barplot
ggplot(data, aes(x=nome, y=valor, fill=nome)) +
  geom_bar(stat = "identity")
```



33 / 44

Gráfico de dispersão - 2 variáveis

Gráfico de dispersão (*scatter plot*)

```
tempos <- c(16,24,32,40,48,48,48,48,56,64,72,80)
resistencias <- c(199,214,230,248,255,262,279,267,305,298,323,359)
plot(tempos,resistencias, main="Gráfico dispersao",
  xlab="Tempos", ylab="Resistencias", pch=10)
abline(lm(resistencias ~ tempos)) #inclue linha de regressao no gráfico
```

ou mais elaborado com a *package* do [Rggplot2](#)

Gráfico de dispersão (*scatter plot*) - ggplot2

```
tempos <- c(16,24,32,40,48,48,48,48,56,64,72,80)
resistencias <- c(199,214,230,248,255,262,279,267,305,298,323,359)
d<-data.frame(x=tempos, y=resistencias)
library(ggplot2) # ativa as funções desta package
ggplot(d, aes(x, y, color =x )) +
  geom_point(shape = 16, size = 5, show.legend = FALSE) +
  ggtitle("Gráfico dispersão -ggplot2")
```

34 / 44

Gráfico de dispersão

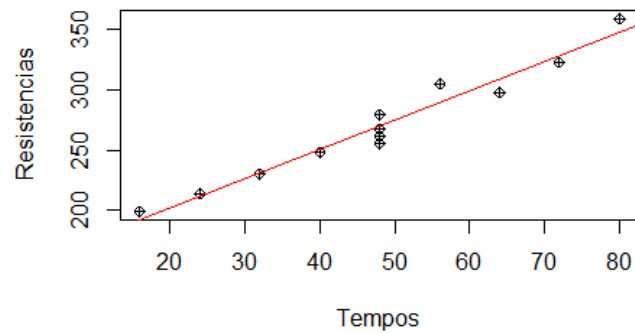
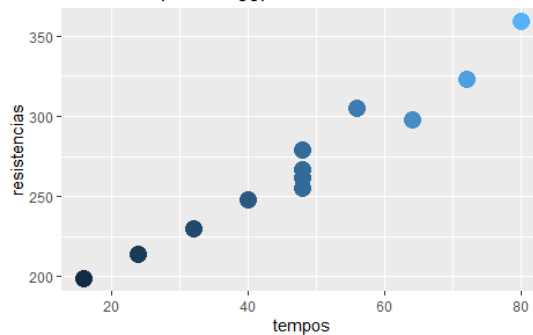


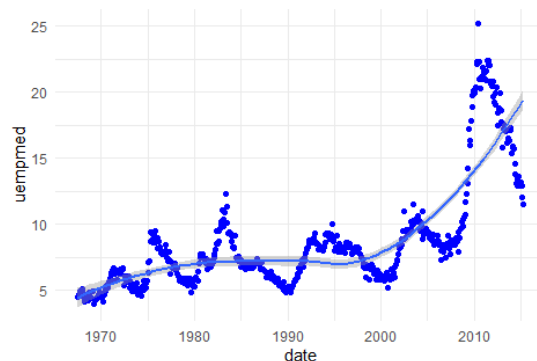
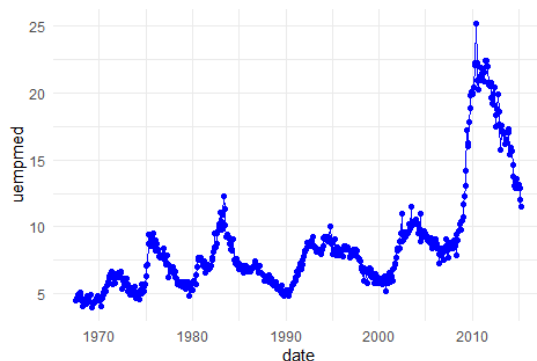
Gráfico dispersão -ggplot2



Outros Gráficos com ggplot

Gráficos de "linhas"

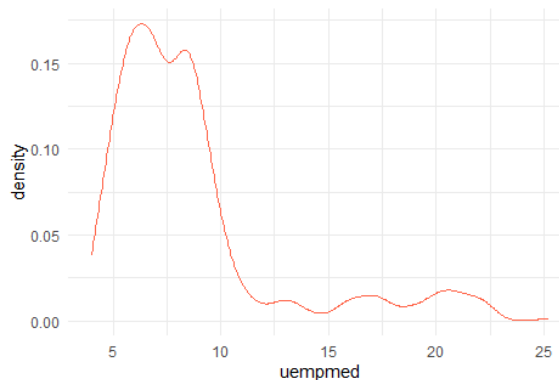
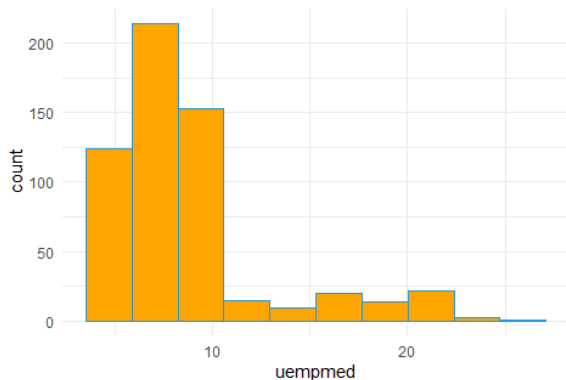
```
library(ggplot2)
theme_set(theme_minimal())
# detalhes sobre os dados fazer: >?economics
head(economics)
p <- ggplot(data= economics, aes(x = date, y = uempmed))
p + geom_line(colour="blue") + geom_point(colour="blue")
# Escolhendo outra cor e tamanho dos pontos
# p + geom_line(color = "#00AFBB", size = 2)
p + geom_point(colour="blue") + geom_smooth() # com ajuste de curva
```



Outros Gráficos com ggplot: cont.

Histograma e densidade estimada

```
p2 <- ggplot(data= economics, aes(x =uempmed ))  
p2 + geom_histogram(colour = 4, fill = "orange",bins = 10) # histograma  
p2 + geom_density(colour="tomato") # densidade estimada
```



37 / 44

Outros Gráficos com ggplot: cont.

Gráficos com subconjunto de dados

```
# Gráficos com subconjunto dos dados  
ss <- subset(economics, date > as.Date("2000-1-1"))  
ggplot(data = ss, aes(x = date, y = uempmed)) +  
geom_line(color = "#FC4E07", size = 2)  
  
ss2 <- subset(economics, date > as.Date("2006-1-1"))  
ggplot(data = ss2, aes(x = date, y = uempmed)) +  
geom_line(color = "#FC4E07", size = 2)
```

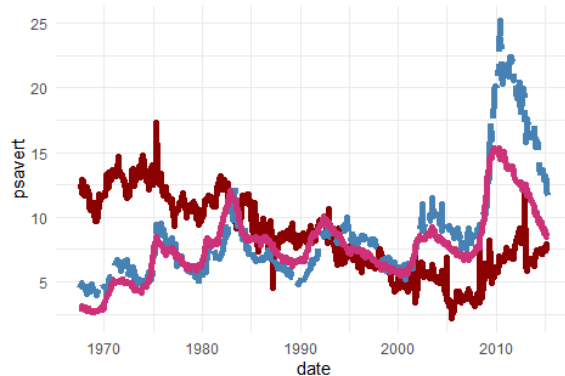
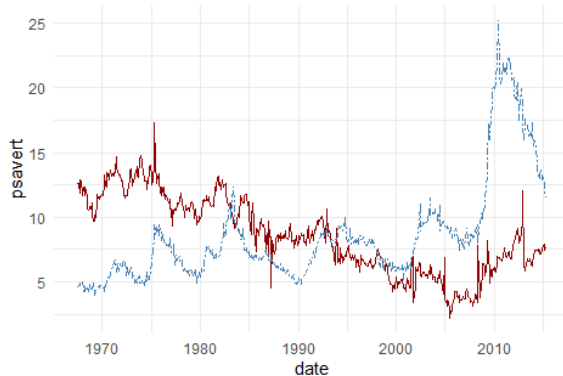


38 / 44

Outros Gráficos com ggplot: cont.

Gráfico com duas e três "linhas"/variáveis

```
p3<-ggplot(economics, aes(x=date))  
p3 + geom_line(aes(y = psavert), color = "darkred") +  
geom_line(aes(y = uempmed), color="steelblue", linetype="twodash")  
p3 + geom_line(aes(y = psavert), color = "darkred",size = 2)  
+ geom_line(aes(y = uempmed), color="steelblue",linetype="twodash",  
size = 2) + geom_line(aes(y = unemploy/1000), color="violetred3",size = 2)
```



39 / 44

Escrever funções em R

- Uma das mais importantes características do R é a capacidade de escrever e modificar funções para efetuar cálculos. A forma geral de uma função em R é: `function (arguments) expression`
- Considere-se a função para obter o quadrado de um número: A nova função, a qual será denominada por `quad()`, é definida do seguinte modo (usar a sub janela Script):

```
quad <-function(x) x * x
```

onde `x` é o argumento da função. Esta nova função pode ser usada agora como mais uma função do R. Na consola ou run no Script fazer:

```
> quad(2)  
[1] 4  
> quad(c(2,3,10))  
[1] 4 9 100
```

- Note-se que a função `quad` é "vetorizável", no sentido em que pode ser aplicada a elementos de um vetor. Isso deve-se à utilização da operação aritmética `*` que já é uma operação "vetorizável".

40 / 44

Escrever funções em

- A função abaixo, denominada `grande()`, toma dois vetores como argumento, compara-os, e origina um vetor constituído por elementos que correspondem ao máximo dos elementos das respetivas componentes dos dois vetores iniciais.

```
grande <-function(x,y) ifelse (y > x, y, x ) # criar a função
a <- c(1:10)
b<- rep(5,10)
grande(a,b) # aplicar a função
[1] 5 5 5 5 5 6 7 8 9 10
```

- Conversao de graus Celsius em Fahrenheit

```
toFahrenheit<-function(celsius){
  f = (9/5) * celsius + 32
  return(f)}
toFahrenheit(30) #aplicar a função
[1] 86
```

41 / 44

Escrever funções em

- Veja-se agora um exemplo de uma função que tem como resultado uma lista. A função `polin()` calcula as raízes reais de uma equação de segundo grau.

```
polin <- function(a, b, c) {
  x1 <- x2 <- NA
  d <- b^2 - 4 * a * c
  if(d < 0)
    stop("Sem raizes reais\n")
  else {
    x1 <- (-b + sqrt(d)) / (2 * a)
    x2 <-(-b - sqrt(d)) / (2 * a)
  }
  list(raiz1 = x1, raiz2 = x2)
}
```

42 / 44

Escrever funções em

- Aplique-se esta função ao cálculo das raízes da equação $6x^2 - 5x + 1 = 0$.

```
> polin(6,-5,1)
$raiz1:
[1] 0.5
$raiz2:
[1] 0.3333333
```

- Aplique-se esta função ao cálculo das raízes da equação $6x^2 + x + 1 = 0$.

```
> polin(6,1,1)
Error in polin(6, 1, 1) : Sem raizes reais
```

43 / 44

AED com /RStudio: outros recursos

- Iara Denise Endruweit Battisti, Felipe Micaíl da Silva Smolski (2019). Software R: Análise estatística de dados utilizando um programa livre. <http://www.editorafaith.com.br/ebooks/grat/978-85-68221-44-0.pdf>
- Tutoriais para R (Vanderlei Júlio Debastiani) <https://vanderleidebastiani.github.io/tutoriais/>
- The R Graph Gallery em <https://www.r-graph-gallery.com/> e DataCamp Course: Data Visualization in R.
- Quick-R - <https://www.statmethods.net/>

44 / 44