Statistical thinking will one day be as necessary for efficient citizenship as the ability to read or write.

Samuel S. Wilks (1906–1964) parafraseando H.G. Wells (1866–1946)

1. Introdução ao R

Nestas notas soltas, iremos rever brevemente conceitos de estatística (descritiva) com a maioria dos quais estão certamente familiarizadas(os), tirando partido do *software* estatístico R.

Estatística — *conjunto de conceitos e métodos utilizados na recolha e interpretação de dados [...], permitindo ainda descrever e predizer situações em que a variabilidade e a incerteza estão presentes* ([8, p. 2]). □

R — *O R é uma linguagem de programação de código aberto focada principalmente em processamento estatístico. As suas características colocam-no na elite dos softwares estatísticos. É fácil de usar, flexível, multiplataforma e [...] é um software gratuito [...]* ([11, p. 6]). □

Conceitos básicos em Estatística

Caraterística de interesse — Não passa de um atributo crucial para o conhecimento do fenónemo em estudo ([5, p. 279]). □

População; unidade estatística — Conjunto de todos os objetos que têm em comum pelo menos uma caraterística de interesse; a cada elemento da população damos o nome de unidade estatística ([5, p. 279]). □

Amostra; dado estatístico — Dada a impossibilidade de observar toda uma população[1] é essencial recolher um subconjunto que se pretende representativo da população e denominado amostra; a cada resultado observado — relativo à v.a. de interesse e respeitante a cada unidade estatística pertencente à amostra — damos o nome de dado estatístico ([5, p. 279]). □

Amostragem — O processo de seleção de uma amostra da população de modo a estimar algum aspeto de interesse da mesma é designado amostragem ([3, p. 332], [5, p. 280]).[2] □

Tipos de dados ([7, p. 4]) — Os dados podem ser:

- **discretos** (reportam-se a contagens ou números inteiros);
- **contínuos** (tomam qualquer valor num intervalo real);

[1] Ou devido ao facto de ser infinita, ou por implicar a sua destruição, ou por razões de economia, comodidade, ou tempo.

[2] A amostragem pode ser também entendida como um vasto conjunto de procedimentos estatísticos que encontra motivação na necessidade de obtenção de amostras, i. e., de *imagens à escala da população* ([5, p. 280]).

- **qualitativos** (referem-se a categorias/classes);
- qualitativos **nominais** (etiquetas/valores que designam uma categoria/classe, sem que haja relação de ordem entre as classes/categorias; e. g., 1=preto, 2=castanho, 3=azul, 4=verde, 5=cinzento);
- qualitativos **ordinais** (há relação de ordem entre as classes/categorias; e. g., classificação: *Mau*, *Suficiente*, *Bom*, *Muito Bom*, *Excelente*);
- **quantitativos** (dizem respeito a característica de interesse intrinsecamente numérica);[3]
- quantitativos com **escala intervalar** (e. g., temperatura; não podemos afirmar que a temperatura de 20^oC é duas vezes mais quente que a de 10^oC);
- quantitativos com **escala absoluta** (e. g., área; a área de 4 hectares é o dobro de outro com 2 hectares).
- **dados menos tradicionais** — Imagens, vídeos, áudio, *Here is your 2021 internet minute infographic!*. □

Exercício 1.1 (a) — Considere os dados disponíveis abaixo, referentes à massa (em kg) de 40 bicicletas.

4.3	6.8	9.2	7.2	8.7	8.6	6.6	5.2	8.1	10.9
7.4	4.5	3.8	7.6	6.8	7.8	8.4	7.5	10.5	6.0
7.7	8.1	7.0	8.2	8.4	8.8	6.7	8.2	9.4	7.7
6.3	7.7	9.1	7.9	7.9	9.4	8.2	6.7	8.2	6.5

(a) Identifique a população e classifique a variável em estudo.

Com a recolha da amostra obtém-se um conjunto de dados, com um aspeto caótico, cuja mera leitura depressa se reconhece nada contribuir para a compreensão do fenómeno aleatório em estudo; a estatística descritiva resolve (parcialmente) esta dificuldade ([5, p. 280]).

Estatística descritiva — Com efeito, a estatística descritiva tem por fim *descrever, resumir e representar a informação contida num conjunto de dados, através da construção de tabelas e gráficos ou através da determinação de medidas numéricas que adequadamente sintetizem os dados* ([7, p. 4]).

Manual

- *Texto sobre Estatística Descritiva* da autoria da Profa. Manuela Neves (ISA).

From data to knowledge and from knowledge to value!

[3] Por forma a identificar a que escala dizem respeito os dados quantitativos, pergunte-se se o dobro do valor da característica de interesse corresponde ao dobro de intensidade).

1.1 Aquisição e recolha de dados

Instalação do software estatístico R

Passa por:

- visitar *http://www.r-project.org*;
- clicar em CRAN (https://cran.r-project.org/mirrors.html), do lado esquerdo debaixo da palavra *Download*;
- seleccionar um dos servidores internacionais (e. g., https://cloud.r-project.org/);
- escolher o sistema operativo adequado (e. g. Download R for macOS);
- descarregar a versão do R compatível com o processador do computador pessoal.

O programa R dispõe de um interface gráfico próprio. Porém, utilizaremos um interface gráfico avançado (IDE-*Integrated Development Environment*) denominado *RStudio*:

- descarregar o *RStudio Desktop* (gratuito).

Após instalar o R e o RStudio, abra este último e no menu *Session* do RStudio escolha a *workspace* (directoria de trabalho) conveniente para a sua sessão, por exemplo:

- *Session* → *Set Working Directory* → *To Source File Location*

Manuais

- *Help* do RStudio ou *R Tutorial*
- *An Introduction to R: Notes on R – A Programming Environment for Data Analysis and Graphics, Version 4.1.1 (2021-08-10)*
- *https://cran.r-project.org/manuals.html*
- *Introdução à Programação em R*

Introdução de dados em R

- Manualmente[4]

```
bicicletas <- c(4.3,6.8,9.2,7.2,8.7,8.6,6.6,5.2,8.1,10.9,7.4,4.5,
3.8,7.6,6.8,7.8,8.4,7.5,10.5,6.0,7.7,8.1,7.0,8.2,8.4,8.8,6.7,8.2,
9.4,7.7,6.3,7.7,9.1,7.9,7.9,9.4,8.2,6.7,8.2,6.5)
```

[4] Utilize a *Console*, introduza uma linha de comandos e carregue na tecla *Return*. Em alternativa, crie um ficheiro *File* → *R Script* → *Save As...*, introduza as linhas de comandos e seleccione os comandos que deseja executar e carregue em *Run*.

```
bicicletas
length(bicicletas)  #dim. amostra
bicicletas[bicicletas>10]  #obs. superiores a 10
```

- A partir de ficheiro[5]

  ```
  Bicicletas = read.csv("Bicicletas.txt", header=TRUE)
  dim(Bicicletas)  #dim. data frame
  ```

  Os valores desta *data frame* aparecem na sub-janela *Environment*...

- Recorrendo a *File* → *Import Data Set* (e. g., From Text Base...)

- Tirando partido de conjuntos de dados (multivariados) famosos e existentes no R, por exemplo,

  ```
  iris
  dim(iris)
  colnames(iris)
  ```

Medidas descritivas

Não nos alongaremos na definição destas medidas, pois já foi dado bastante ênfase à Estatística Descritiva no ensino pré-universitário. De todo o modo, acrescentamos alguns comentários em notas de rodapé.

Estudaremos medidas descritivas referentes à localização, dispersão e forma (de distribuição) dos dados. Para o efeito, considere-se que:

- $(x_1, x_2, \ldots, x_n)$, representa a amostra (não agrupada);
- n, a dimensão da amostra;
- $(x_{(1)}, \ldots, x_{(n)})$, a amostra ordenada;
- $x_{(1)} = \min_{i=1,\ldots,n} x_i$, o mínimo da amostra;
- $x_{(n)} = \max_{i=1,\ldots,n} x_i$, o máximo da amostra.

Medidas de localização

Quantidades numéricas que nos dão uma ideia da posição/localização de um conjunto de dados ([8, p. 9]).

[5] Em *https://fenix.tecnico.ulisboa.pt/disciplinas/PEstatisticad4/2021-2022/1-semestre/dados* encontrará vários conjuntos de dados.

- **Média**[6]

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

  (Ver propriedades da média da amostra na *pag. 9* de [8].)

  Moda[7]

  mo

- **Mediana**[8]

$$me = \begin{cases} x_{((n+1)/2)}, & \text{para } n \text{ ímpar} \\ & \text{(observação central da amostra ordenada)} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2}, & \text{para } n \text{ par} \\ & \text{(média das 2 observações centrais da amostra ordenada)} \end{cases}$$

- **Quartis**[9]

  Primeiro quartil[10]

$$q_{1/4} = \begin{cases} x_{([n/4]+1)}, & \text{para } n/4 \text{ não inteiro} \\ \frac{x_{(n/4)}+x_{(n/4+1)}}{2}, & \text{para } n/4 \text{ inteiro} \end{cases}$$

  Segundo quartil

  $q_{1/2} = me$

  Terceiro quartil

$$q_{3/4} = \begin{cases} x_{([3n/4]+1)}, & \text{para } 3n/4 \text{ não inteiro} \\ \frac{x_{(3n/4)}+x_{(3n/4+1)}}{2}, & \text{para } 3n/4 \text{ inteiro} \end{cases}$$

---

[6] Ou média aritmética, média empírica. Trata-se do *ponto de equilíbrio* dos dados ([8, p. 9]) ou centro de massa/gravidade da amostra. É geralmente um valor que não pertence à amostra e pode não ter existência real, em particular quando estão a lidar com dados discretos.

[7] Medida de localização menos usual; mo é *definida, no caso discreto, como o valor que ocorre com mais frequência, ou como o intervalo de classe com maior frequência se os dados são de natureza contínua*; uma amostra pode não possuir *moda ou apresentar mais do que uma moda*; um conjunto de observações com uma única moda diz-se unimodal; a moda *é particularmente útil quando temos dados de natureza qualitativa, para os quais não é possível calcular a média ou [...] a mediana* ([8, p. 11]).

[8] A média da amostra, por basear-se na totalidade dos dados, é sensível a valores muito pequenos ou muito grandes no conjunto dos dados (candidatos a *outliers*). Uma medida *robusta* relativamente a este tipo observações, no sentido de não ser afectada por tais observações, é a mediana, o valor que divide um conjunto ordenado de dados em duas partes iguais; é o *valor do meio* ([8, p. 10]).

[9] Ao considerarmos a amostra ordenada dividida em quatro partes *iguais*, obtemos os pontos da divisão denominados quartis: $q_{1/4}, q_{1/2} = me, p_{3/4}$, o primeiro, segundo e terceiro quartis, respectivamente ([8, p. 10]).

[10] $q_{1/4}$ é tal que: pelo menos 25% das observações são menores ou iguais a $q_{1/4}$; e pelo menos 75% das observações são maiores ou iguais a $q_{1/4}$ ([8, p. 10]).

- **Quantil de ordem** α $(0 \leq \alpha \leq 1)$[11]

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

Medidas de dispersão[12]

- **Amplitude total**[13]

  $x_{(n)} - x_{(1)}$

  Amplitude inter-quartil (*interquartile range, IQR*)[14]

  $IQR = q_{3/4} - q_{1/4}$

  Variância[15]

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

  (Ver propriedades da variância da amostra na *pag. 13* de [8].)

  Desvio padrão[16]

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

  Coeficiente de variação[17]

  $cv = \frac{s}{\bar{x}}$

[11] Podemos definir de forma análoga: os decis, valores que dividem o conjunto das observações em 10 partes iguais quantis; os percentis, valores que resultam da divisão da amostra ordenada em 100 partes iguais. São, à semelhança dos quartis, o que designamos de quantis. q_α é tal que: pelo menos $\alpha \times 100\%$ das observações são menores ou iguais a q_α; e pelo menos $(1-\alpha) \times 100\%$ das observações são maiores ou iguais a q_α ([8, pp. 10–11]). *Aqui* poderá encontrar esta e outras formas de calcular quantis no R.

[12] As medidas de localização *não são suficientes para dar uma ideia clara da distribuição das observações*; os dois conjuntos de dados $(1, 2, 5, 8)$ $(-2, 3, 4, 1)$ possuem a mesma média $(\bar{x} = 4)$ e mediana $(me = 3.5)$, porém, o primeiro apresenta maior concentração dos dados que o segundo, pelo que são necessárias medidas que nos dêem informação sobre a dispersão das observações ([8, p. 12]).

[13] É a amplitude do intervalo de variação dos dados ([8, p. 12]).

[14] Sensivelmente 50% das observações encontram-se no intervalo $[q_{3/4}, q_{1/4}]$ ([8, p. 12]).

[15] No cálculo da IQR ignoramos as observações concretas *na zona central e nas zonas extremas* do conjunto de observações; faz sentido ter em conta a posição de todas as observações, relativamente a um ponto de referência como a média da amostra; usar como medida de dispersão $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$ não é razoável, pois $\sum_{i=1}^{n}(x_i - \bar{x})$=0. ([8, p. 12]). Em alternativa, podemos usar $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ou $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$; o uso de $(n-1)$ ao invés de n será justificado mais tarde; quanto menos dispersas estiverem as observações relativamente à média, menor será o valor de s^2 ([8, p. 13]).

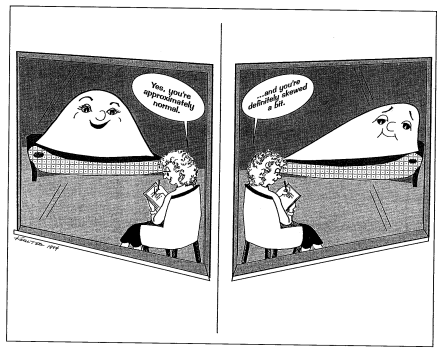
[16] Esta medida de dispersão tem uma vantagem sobre s^2: possui as mesmas unidades que as observações.

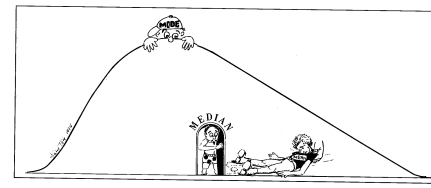
[17] É uma medida de dispersão relativa e não absoluta como a variância ou o desvio padrão; é adimensional e permite comparar a dispersão de distribuições com localizações (ou ordens de grandeza) distintas (ver [2, p. 65], [5, p. 90]). Atentemos que deve ser usada apenas quando as observações ou são todas positivas ou todas negativas ([8, p. 14]). O coeficiente de variação pode ser interpretado como a fracção da dispersão (desvio padrãoo) por que a localização (valor esperado) é responsável ([, p. 113], [5, p. 90]).

Medidas de forma

- **Coeficiente de assimetria** (*skewness coefficient*)[18]

$$sc = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}}$$





simetria assimetria negativa assimetria positiva

Coeficiente de achatamento ou curtose (*kurtosis coefficient*)[19]

$$kc = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}$$

Exercício 1.1 — Considere os dados referentes à massa (em kg) de 40 bicicletas.

(b) Obtenha algumas medidas descritivas dos dados e comente.[20]

R Tutorial — Numerical Measures

```
mean(bicicletas)
N <- length(bicicletas)
sum(bicicletas)/N

median(bicicletas)
sortbicicletas = sort(bicicletas)
sortbicicletas
(sortbicicletas[N/2]+sortbicicletas[N/2+1])/2
```

[18] Importa averiguar se o conjunto de dados é simétrico ($\bar{x} = me = mo$). As observações dir-se-ão com: assimetria negativa se $\bar{x} < me < mo$; assimetria positiva se $mo < me < \bar{x}$.

[19] Um coeficiente de curtose maior que 3 / igual a 3 / menor que 3 a sugere que um conjunto de dados leptocúrtico (histograma alongado) / mesocúrtico / platicúrtico (histograma achatado).

[20] Em *https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile* encontramos descritas nove possibilidades de cálculo dos quantis; o *default method* é o 7, um de seis tipos de *continuous sample quantiles*; os quantis que definimos correspondem ao método 2.

```
quantile(bicicletas, c(0.25, 0.5, 0.75), type=2)
(sortbicicletas[N/4]+sortbicicletas[N/4+1])/2
(sortbicicletas[N/2]+sortbicicletas[N/2+1])/2
(sortbicicletas[3*N/4]+sortbicicletas[3*N/4+1])/2
quantile(bicicletas, c(0.25, 0.5, 0.75), type=7) #default method
quantile(bicicletas, c(0.25, 0.5, 0.75))
quantile(bicicletas, c(0.25, 0.5, 0.75), type=1)
sortbicicletas[ceiling(N/4)]
sortbicicletas[ceiling(N/2)]
sortbicicletas[ceiling(3*N/4)]

max(bicicletas)-min(bicicletas)
sortbicicletas[N]-sortbicicletas[1]
range(bicicletas) #gama valores

IQR(bicicletas, type=2)
(sortbicicletas[3*N/4]+sortbicicletas[3*N/4+1])/2
  -(sortbicicletas[N/4]+sortbicicletas[N/4+1])/2
IQR(bicicletas, type=7)
IQR(bicicletas)

var(bicicletas)
1/(N-1)*(sum(bicicletas*bicicletas)-N*mean(bicicletas)^2)

sd(bicicletas)
sqrt(1/(N-1)*(sum(bicicletas*bicicletas)-N*mean(bicicletas)^2))

sd(bicicletas)/mean(bicicletas)*100

summary(bicicletas)

install.packages("moments")
library(moments)
skewness(bicicletas)
mmean <- rep(mean(bicicletas),N)
1/N*sum((bicicletas-mmean)^3)/ ((N-1)/N*var(bicicletas))^1.5

kurtosis(bicicletas)
mmean <- rep(mean(bicicletas),N)
1/N*sum((bicicletas-mmean)^4)/ ((N-1)/N*var(bicicletas))^2
```

(c) Determine o intervalo das 25% menores massas e o intervalo das 25% maiores massas da amostra, bem como a amplitude inter-quantil.

```
quantile(bicicletas,c(0,0.25))
quantile(bicicletas,c(0.75,1))
IQR(bicicletas)
```

(d) Indique o quantil amostral de ordem 0.68.

```
quantile(bicicletas,0.68)
```

1.2 Visualização de dados estáticos e dinâmicos

Diagrama de caule-e-folhas

Quando os dados não são muito numerosos, a interpretação e visualização dos mesmos pode ser facilitada usando o que se designa por *diagrama de caule-e-folhas*. Este dispositivo engenhoso, denominado *stem-and-leaf* na literatura anglo-saxónica, permite uma observação do aspecto global dos dados, sem (grande ou qualquer) perda da informação contida na colecção de dados inicial ([6, p. 18]).

Exercício 1.1

(e) Construa o diagrama de caule-e-folhas.

```
sortbicicletas
stem(bicicletas)
```

Agrupamento de dados e a regra de Sturges

Quando o número de observações distintas é elevado ou os dados são de natureza contínua, deverá agrupar-se os dados, tomando *as observações próximas por forma a evidenciar as características subjacentes aos dados* ([8, p. 6]).

É, pois, necessário decidir o número de classes, tendo sempre a consciência de que:

- um número exagerado de classes pode não mostrar a regularidade do fenómeno;
- um número muito pequeno de classes mascara a variabilidade do fenómeno.

É frequente recorrer à regra de Sturges ([13])[21] para obter o número de classes (k) à custa da dimensão da amostra (n):[22]

$$k \approx 1 + \log_2(n) = 1 + \frac{\ln(n)}{\ln(2)}.$$

Alguns autores aconselham o maior inteiro inferior ou igual a $1 + \frac{\ln(n)}{\ln(2)}$ ([8, p. 6]); outros há que recomendam que se tome $k = 1 + \left\lceil \frac{\ln(n)}{\ln(2)} \right\rceil$.[23]

[21] Hyndman *claims that* Herbert A. Sturges (1882–1958) *considered an idealised frequency histogram with k bins where the ith bin count is the binomial coefficient* $\binom{k-1}{i}$, $i = 0, 1, \ldots, k-1$. *As k increases, this ideal frequency histogram approaches the shape of a normal density. The total sample size is* $n = \sum_{i=0}^{k-1} \binom{k-1}{i} = (1+1)^{k-1} = 2^{k-1}$ *by the binomial expansion. So the number of classes to choose when constructing a histogram from normal data is* $k = 1 + \log_2(n)$.

[22] O k vem da palavra *klassen* em alemão.

[23] Há alternativas (mais sofisticadas) à regra de Sturges.

A amplitude h das classes obtém-se do seguinte modo:

$$\Delta = \frac{x_{(n)} - x_{(1)}}{k}$$

Esta sumarização inicia-se construindo uma tabela de frequências, cuja representação gráfica é feita por meio de um **histograma** ([8, p. 6]), proposto inicialmente pelo matemátivo e bio-estatístico inglês *Karl Pearson* (1857–1936), num seu trabalho datado de 1895 ([9]).

Um procedimento possível consiste em começar a construção das classes aberta à esquerda e fechadas à direita; contudo, a primeira classe deverá conter $x_{(1)}$ e coincide com o intervalo $[x_{(1)}, x_{(1)} + \Delta]$; a segunda será o intervalo $(x_{(1)} + \Delta, x_{(1)} + 2\Delta]$; ...; a última classe a formar-se deverá ser $(x_{(n)} - \Delta, x_{(n)}]$.[24]

Diagrama de extremos e quartis (*Box plot*)

A combinação dos extremos e dos quartis permite obter um gráfico muito sugestivo designado de diagrama de extremos-e-quartis ([6, p. 90]),[25] que permite que fiquemos com uma ideia quer da localização, quer da dispersão dos dados.

É à especialista em visualização de dados americana *Mary Eleanor Hunt Spear* (1897–1986) que devemos a primeira formulação do *box plot*, em 1952 em [12]. O matemático e estatístico americano *John W. Tukey* (1915 –2000) desenvolveu-o dezassete anos depois e popularizou-o.

O desenho do diagrama de extremos e quartis pressupõe, por exemplo, que marquemos num eixo os valores das seguintes medidas descritivas:

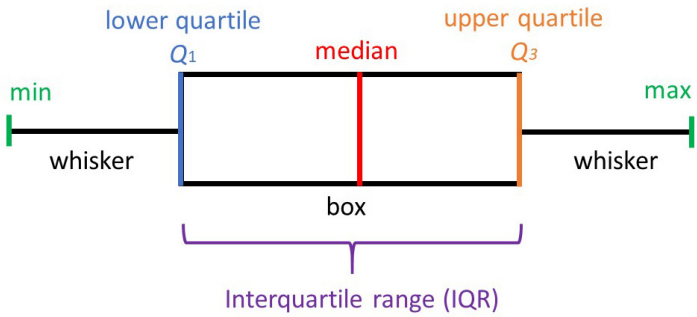
- $x_{(1)}$, o mínimo da amostra;
- $x_{(n)}$, o máximo;
- me, a mediana;
- $q_{1/4}$, o primeiro quartil;
- $q_{3/4}$, o terceiro quartil.

Tirando partidos destes valores desenhamos um rectângulo cujo lados inferior e superior são dados pelo primeiro e terceiro quartis, respectivamente; o rectângulo é dividido em duas partes pelo valor da mediana; um dos bigodes parte do centro do lado inferior do

[24] Há a possibilidade de considerar classes com amplitudes distintas.

[25] *Box plot*, *diagrama de caixa*, *Box-and-whiskers plot*, *diagrama de caixa-e-bigodes* ou *caixa-de-bigodes* são algumas das designações frequentes deste diagrama. Para uma descrição detalhada da construção de diagramas de extremos-e-quartis, sugere-se a leitura de [1, pp. 161–162].

rectângulo até ao mínimo da amostra; o outro do do centro do lado superior do rectângulo até ao máximo da amostra ([7, p. 12]), como ilustra a figura abaixo.



Trata-se, no entanto, de um dispositivo gráfico menos informativo que um histograma em que sejam marcados os quartis e a mediana, afinal tal histograma diz-nos como se distribuem as frequências nos intervalos $[x_{(1)}, q_1)$, $[q_1, me)$, $[me, q_3)$ e $[q_3, x_{(n)}]$.

Detecção de *outliers*

Toda e qualquer observação que não pertença ao intervalo

$$I = [q_{1/4} - 1.5 \times IQR,\ q_{3/4} + 1.5 \times IQR]$$

é considerada uma observação discordante ou *outlier*, de acordo com o *critério IQR* ([4, p. 61]).

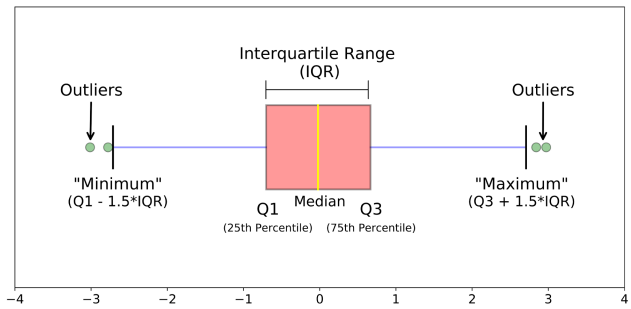
Caso os dados fossem provenientes de uma população normal com valor esperado nulo e variância igual a 1, a probabilidade de pertença ao intervalo I seria dada por

$$\Phi(\Phi^{-1}(0.75) + 1.5 \times [\Phi^{-1}(0.75) - \Phi^{-1}(0.25)])$$
$$-\Phi(\Phi^{-1}(0.25) - 1.5 \times [\Phi^{-1}(0.75) - \Phi^{-1}(0.25)]) \simeq 0.993023$$

e a de não pertença seria muitíssimo baixa, aproximadamente igual a 0.006977.

Ao adoptarmos uma pequena variante do *box plot*, em que na sua construção $x_{(1)}$ (resp. $x_{(n)}$) é substituído pelo "mínimo" da amostra, $\max\{x_{(1)},\ q_{1/4} - 1.5 \times IQR\}$ (resp. pelo "máximo" da amostra, $\min\{x_{(n)},\ q_{3/4} + 1.5 \times IQR\}$), podemos identificar observacões discordantes ([8, p. 16]).[26]

[26] É possível identificar estes valores discordantes no R. Há outros critérios para identificar potenciais *observações discordantes*.



Exercício 1.1

(f) Obtenha o histograma e a caixa de bigodes identificando possíveis *outliers*.

```
?hist
h <- hist(bicicletas,main="Histograma",xlab="Massa",ylab="Freq. absoluta")
h$breaks
sortbicicletas
h$counts

h <- hist(bicicletas,freq=FALSE,main="Histograma",xlab="Massa",
          ylab="Freq. relativa")

h <- hist(bicicletas, right=TRUE, include.lowest = TRUE)
% #intervals of the form (a, b], except the first [a,b]
h$breaks
sortbicicletas
h$counts

h <- hist(bicicletas, right=FALSE, include.lowest = TRUE)
% #intervals of the form [a, b), except the last [a,b]
h$breaks
sortbicicletas
h$counts

k <- nclass.Sturges(bicicletas)
print(k)
amp <- (max(bicicletas)-min(bicicletas))/k
breaaaks <- NULL
for (i in rep(0:k)) {
  breaaaks <- c(breaaaks,min(bicicletas)+i*amp)
}
print(breaaaks)

h <- hist(bicicletas, breaks=breaaaks, ylim=c(0,15))
text(h$mids,h$counts,labels=h$counts)
h$breaks
sortbicicletas
h$counts
```

```
?boxplot
boxplot(bicicletas, col="red", main="")
boxplot(bicicletas, col="green", horizontal=TRUE, main="Bicicletas")

boxplot.stats(bicicletas)$out
out <- boxplot.stats(bicicletas)$out
out_ind <- which(bicicletas %in% c(out))
print(out_ind)
print(Bicicletas)

#probabilidade de pertencer ao intervalo I, caso normal(0,1)
pnorm(qnorm(0.75)+1.5*(qnorm(0.75)-qnorm(0.25)), mean=0, sd=1)
     - pnorm(qnorm(0.25)-1.5*(qnorm(0.75)-qnorm(0.25)), mean=0, sd=1)
```

Medidas descritivas para dados agrupados

Muitas das medidas descritivas podem ser calculadas, caso não disponhamos dos dados originais mas sim dos dados agrupados em classes e de uma tabela de frequências (absolutas ou relativas).

O cálculo destas medidas assenta na hipótese básica de tabulagem ([6, p. 35]) que consiste em admitir que todos os valores de uma classe são iguais ao respectivo ponto médio. O erro associado designa-se por erro de tabulagem.

Consideremos que as n observações foram agrupadas k classes e, no que se refere à classe i ($i = 1, 2, \ldots, k$), sejam:

- c_i, o seu ponto médio;
- n_i, a respectiva frequência absoluta;
- $f_i = \frac{n_i}{n}$, a sua frequência relativa;
- $F_i = \sum_{l=1}^{i} f_i$, a frequência relativa acumulada da classe i.

É possível obtermos as medidas descritivas abaixo (ver [8, pp. 16–18] e [7, pp. 9–10]).[27]

- **Média**

  $\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \times c_i$

- **Classe modal; moda**

  A classe modal corresponde à classe com maior frequência absoluta.

[27] Preferimos utilizar as mesmas letras para representar estas medidas descritivas quando calculadas quer com base nos dados originais, quer nos dados agrupados (A), ao invés de substituir, por exemplo, $\bar{x}$ por $\bar{x}_A$.

De acordo com [8, p. 17], existem várias fórmulas empíricas para determinar a moda. A mais simples consiste em considerar o ponto médio da classe modal. A mais conhecida é a *fórmula de King*,

$$mo = x_\star^{inf} + (x_\star^{sup} - x_\star^{inf}) \times \frac{f_{\star+1}}{f_{\star-1}+f_{\star+1}},$$

onde: $\star$ representa o índice da classe modal; $x_\star^{inf}$ e $x_\star^{sup}$ são os limites inferior e superior da classe modal; e $f_{\star-1}$ e $f_{\star+1}$ são, respectivamente, as frequências relativas das classes anterior e posterior à classe modal.

- **Quantil de ordem** α $(0 \leq \alpha \leq 1)$

  A obtenção desta medida de localização para dados agrupados passa por identificar a primeira classe cuja frequência relativa acumulada seja superior ou igual a α, seja ela a classe com índice $\star$, extremos inferior e superior iguais a $x_\star^{inf}$ e $x_\star^{sup}$, frequência relativa $f_\star$ e frequência relativa acumulada $F_\star$. Então

  $$q_\alpha = x_\star^{inf} + (x_\star^{sup} - x_\star^{inf}) \times \frac{\alpha - F_{\star-1}}{f_\star},$$

  onde: $F_{\star-1}$ representa a frequência relativa acumulada da classe anterior à classe com índice $\star$; e $F_{\star-1} = 0$ se $\star = 1$.

  Escusado será dizer que o primeiro quartil, a mediana e o terceiro quartil se obtêm tomando $\alpha = 1/4, 1/2, 3/4$, respectivamente.

- **Amplitude total**[28]

  $x_k^{sup} - x_1^{inf}$

  Amplitude inter-quartil[29]

  $IQR = q_{3/4} - q_{1/4}$

- **Variância**

  $s^2 = \frac{1}{n} \sum_{i=1}^{k} n_i \times (c_i - \bar{x})^2$

  Desvio padrão

  $s = \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i \times (c_i - \bar{x})^2}$

  Coeficiente de variação

  $cv = \frac{s}{\bar{x}}$

[28] Parece razoável tomar a diferença entre o extremo superior da última classe e o extremo inferior da primeira classe considerada, como sugere [8, p. 18].

[29] Tiramos partido do terceiro e primeiro quartis obtidos com base nos dados agrupados.

- **Coeficiente de assimetria**

  $sc = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i \times (c_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{k} n_i \times (c_i - \bar{x})^2\right]^{3/2}}$

  Coeficiente de achatamento

  $kc = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i \times (c_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{k} n_i \times (c_i - \bar{x})^2\right]^{2}}$

Exercício 1.1

(g) Após agrupar os dados em classes, calcule a média e o desvio padrão para os dados agrupados e compare estas medidas descritivas com as dos dados originais.

Começamos por responder à questão tomando as classes propostas pelo R e tirando partido do *package gds* do R.

```
table(bicicletas)
table(cut(bicicletas,breaks=c(3,4,5,6,7,8,9,10,11),right=TRUE,
          include.lowest = TRUE))
ll <- c(3,4,5,6,7,8,9,10)
ul <- c(4,5,6,7,8,9,10,11)
freqabs <- c(1,2,1,8,11,11,4,2)
midpoints <- (ll+ul)/2
print(midpoints)

meang <- sum(freqabs*midpoints)/N
print(meang)
print(mean(bicicletas))

k <- length(freqabs)
mmeang <- rep(meang,k)
1/N*sum(freqabs*(midpoints-mmeang)^2)
print(var(bicicletas))

sqrt(1/N*sum(freqabs*(midpoints-mmeang)^2))
print(sd(bicicletas))

install.packages("gds")
library(gds)
?gds
gds(ll,ul,freqabs)
```

Desta feita respondemos à questão tomando as 7 classes resultantes da aplicação da regra de Sturges e usando o *package gds*.

```
k <- nclass.Sturges(bicicletas)
print(k)
amp <- (max(bicicletas)-min(bicicletas))/k
breaaaks <- NULL
for (i in rep(0:k)) {
  breaaaks <- c(breaaaks,min(bicicletas)+i*amp)
}
print(breaaaks)
table(cut(bicicletas,breaks=breaaaks,right=TRUE,
          include.lowest = TRUE))
ll <- breaaaks[-(k+1)]
print(ll)
ul <- breaaaks[-1]
print(ul)
freqabs <- c(3,1,8,9,13,4,2)
midpoints <- (ll+ul)/2
print(midpoints)

meang <- sum(freqabs*midpoints)/N
print(meang)
print(mean(bicicletas))

k <- length(freqabs)
mmeang <- rep(meang,k)
1/N*sum(freqabs*(midpoints-mmeang)^2)
print(var(bicicletas))

sqrt(1/N*sum(freqabs*(midpoints-mmeang)^2))
print(sd(bicicletas))

gds(ll,ul,freqabs)
```

Estatística descritiva a duas dimensões

É frequente lidarmos com observações de duas (ou mais) variáveis em cada unidade estatística, como nota [8, p. 19]. Importa averiguar, por exemplo, se existe uma relação estatística entre o par de varáveis em estudo; em caso afirmativo, dizemos haver correlação entre elas ou, equivalentemente, que são variáveis correlacionadas.

No estudo de séries de pares de observações, (x_i, y_i), $i = 1, \ldots, n$, discutiremos somente:

- a representação gráfica das observações;
- o cálculo do coeficiente de correlação amostral.[30]

[30] Para mais detalhes, sugerimos a leitura de [8, pp. 24-31].

Diagrama de dispersão

Admitamos que: x_i representa o *estímulo* a que é submetido o *indivíduo* i; y_i representa a resposta do indivíduo i ao estímulo x_i.

É crucial representar graficamente os pontos (x_i, y_i), $i = 1, \ldots, n$, para averiguar se a relação entre a variável (explicativa) x e a variável (resposta) y é, por exemplo, do tipo linear.

Este gráfico é usualmente designado por *diagrama de dispersão*, *scatter diagram* ou *scatter plot*.[31]

```
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="Diagrama de dispersao",
     xlab="Tempos", ylab="Resistencias", pch=10)
abline(lm(tempos ~ resistencias))
```

Coeficiente de correlação amostral (de Pearson)

Ao dispormos de um conjunto de n observações bivariadas, (x_i, y_i), $i = 1, \ldots, n$, o coeficiente de correlação amostral (de Pearson) é dado por

$$\begin{aligned} r &= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s'_x} \frac{y_i - \bar{y}}{s'_y} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})^2}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) \times (\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}} \end{aligned}$$

quantifica a associação linear entre as variáveis x e y.

Além de $-1 \leq r \leq 1$, para qualquer par de variáveis, importa notar o seguinte, como refere [8, pp. 26–27]:

(1) $r = 1$ (resp. $r = -1$) se só se todos os pontos se encontram sobre uma recta de declive positivo (resp. declive negativo);

(2) $r \simeq 1$ (resp. $r \simeq -1$) significa que todos os pontos se dispersam ligeira e aleatoriamente em torno de uma recta de declive positivo (resp. declive negativo).

[31] É possível recorrer ao *package car* e recorrer à função *scatterplot*.

(3) $r \simeq 0$ sugere a ausência de associação linear entre as variáveis x e y, os pontos no diagrama de dispersão fazem lembrar uma nuvem com aspecto arredondado ou alongado segundo um dos eixos.

O sinal da correlação entre x e y deve ser interpretado do seguinte modo: caso $r > 0$ (resp. $r < 0$), podemos afirmar que as variáveis x e y tenderão a variar no mesmo sentido (resp. em sentidos opostos) relativamente às respectivas médias ([5, p. 229]).

Como menciona [8, p. 27], o coeficiente de correlação amostral *mede a nitidez da ligação* [linear] *existente entre as variáveis*.

```
cor(tempos, resistencias)
```

Gráficos (mais sofisticados) com a função ggplot

Consultar

- *https://www.r-graph-gallery.com/index.html*
- *https://ggplot2-book.org/index.html*
- *Noções Básicas de Análise Exploratória de Dados — Prof. Manuel Scotto*

BIBLIOGRAFIA

[1] Abell, M.L., Braselton, J.P. e Rafter, J.A. (1999). *Statistics with Mathematica*. Academic Press, Inc.

[2] Devore, J. L. (1991). *Probability and Statistics for Engineering and the Sciences* (3rd. ed.). Pacific Grove, CA: Brooks/Cole.

[3] Everitt, B. S. (2002). *The Cambridge Dictionary of Statistics (Second Edition)*. Cambridge, UK: Cambridge University Press.

[4] Martins, M. E. G. (2005). Introdução à Probabilidade e à Estatística com complementos de Excel. Lisboa: Sociedade Portuguesa de Estatística.

[5] Morais, M. C. (2020). *Probabilidades e Estatística: Teoria, Exemplos & Exercícios*. Lisboa: IST Press (Coleção Ensino da Ciência e da Tecnologia).

[6] Murteira, B. J. F. e Black, G. H. J. (1983). *Estatística Descritiva*. Lisboa: Editora McGraw-Hill de Portugal, Lda.

[7] Natário, I. (2006). Probabilidades e Estaística D. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

[8] Neves, M. (2014). Introdução à Estatística e à Probabilidade. Instituto Superior de Agronomia, Universidade Técnica de Lisboa. *https://fenix.isa.ulisboa.pt/qubEdu/conteudos-publicos/ficheiros?oid=3972844786329*, acedido a 25/09/2021.

[9] Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew distribution in homogeneous material. *Philosophical Transactions of the Royal Society of London Ser. A* **186**, 343–414.

[10] Pestana, D. D. e Velosa, S. F. (2002). *Introdução à Probabilidade e à Estatística*. Lisboa: Fundação Calouste Gulbenkian.

[11] Ralha, T. I. C. (2014). Sinais válidos em esquemas conjuntos para a localização e a dispersão com exemplos em R. Tese de mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2014.

[12] Spear, M. E. (1952). *Charting Statistics*. New York: McGraw Hill.

[13] Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association* **21**, 65–66.

[14] Torgo, L. (2006). Introdução à Programação em R. Faculdade de Economia, Universidade do Porto. *https://cran.r-project.org/doc/contrib/Torgo-ProgrammingIntro.pdf*, acedido a 25/09/2021.