TECNOLOGIA SETÚBAL
ESCOLA SUPERIOR POLITÉCNICO SETÚBAL

MÉTODOS ESTATÍSTICOS

Estatística Descritiva

Licenciatura em Engenharia Informática

Departamento de Matemática
Escola Superior de Tecnologia de Setúbal
Instituto Politécnico de Setúbal
2023-2024

Estatística Descritiva

- Conceitos básicos.
- Dados qualitativos e quantitativos.
- Organização e interpretação de dados através de tabelas.
- Organização e interpretação de dados através de gráficos.
- Medidas de localização e dispersão.

Conceitos Básicos

População (ou universo)

Conjunto de objetos (pessoas, resultados experimentais, ...) com uma ou mais características comuns, que se pretendem estudar. A população poderá ser finita ou infinita. Aos elementos da população chamam-se **Unidades Estatísticas**.

Amostra

Subconjunto de dados que pertencem à população. Parte da população que é observada com o objetivo de obter informação para estudar a característica pretendida. As amostras são sempre finitas. Estudam-se amostras para tirar conclusões para a população.

Conceitos Básicos

População (ou universo)

Conjunto de objetos (pessoas, resultados experimentais, ...) com uma ou mais características comuns, que se pretendem estudar. A população poderá ser finita ou infinita. Aos elementos da população chamam-se **Unidades Estatísticas**.

Amostra

Subconjunto de dados que pertencem à população. Parte da população que é observada com o objetivo de obter informação para estudar a característica pretendida. As amostras são sempre finitas. Estudam-se amostras para tirar conclusões para a população.

Variável Estatística

Propriedade ou característica que se pretende estudar numa população.

Dado Estatístico

É cada um dos valores observados da variável em estudo.

Exemplo 1

- **População:** Os alunos do IPS
- **Unidade Estatística:** Alunos
- **Uma possível Amostra:** Os alunos do 2º ano
- **Possível Variável Estatística de interesse:** Número de unidades curriculares com aprovação
- **Dados Estatísticos:** 0, 1, 2,...
- **Possível Variável Estatística de interesse:** Se teve aprovação na unidade curricular Matemática I
- **Dados Estatísticos:** Sim, Não

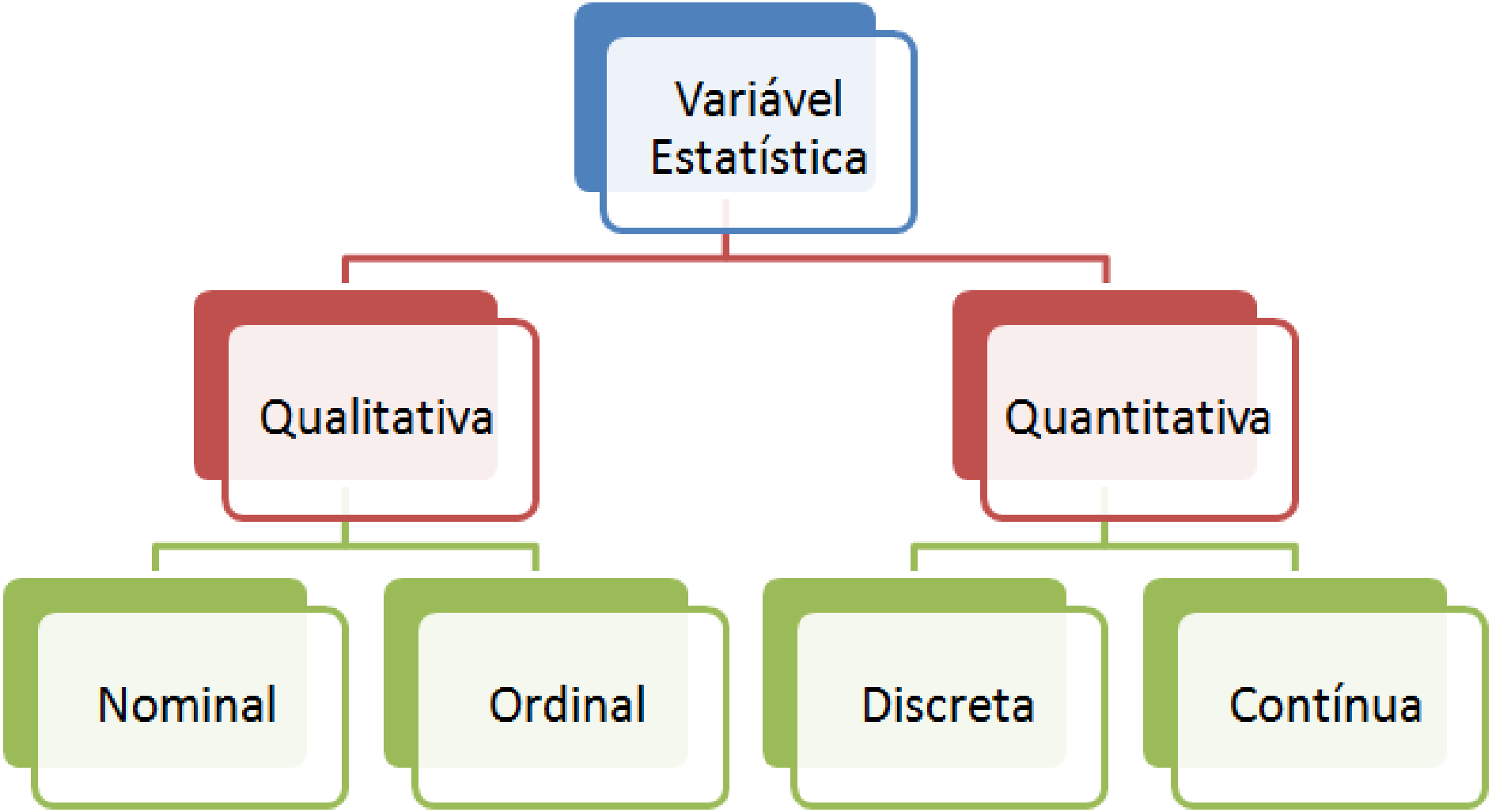
Exemplo 2

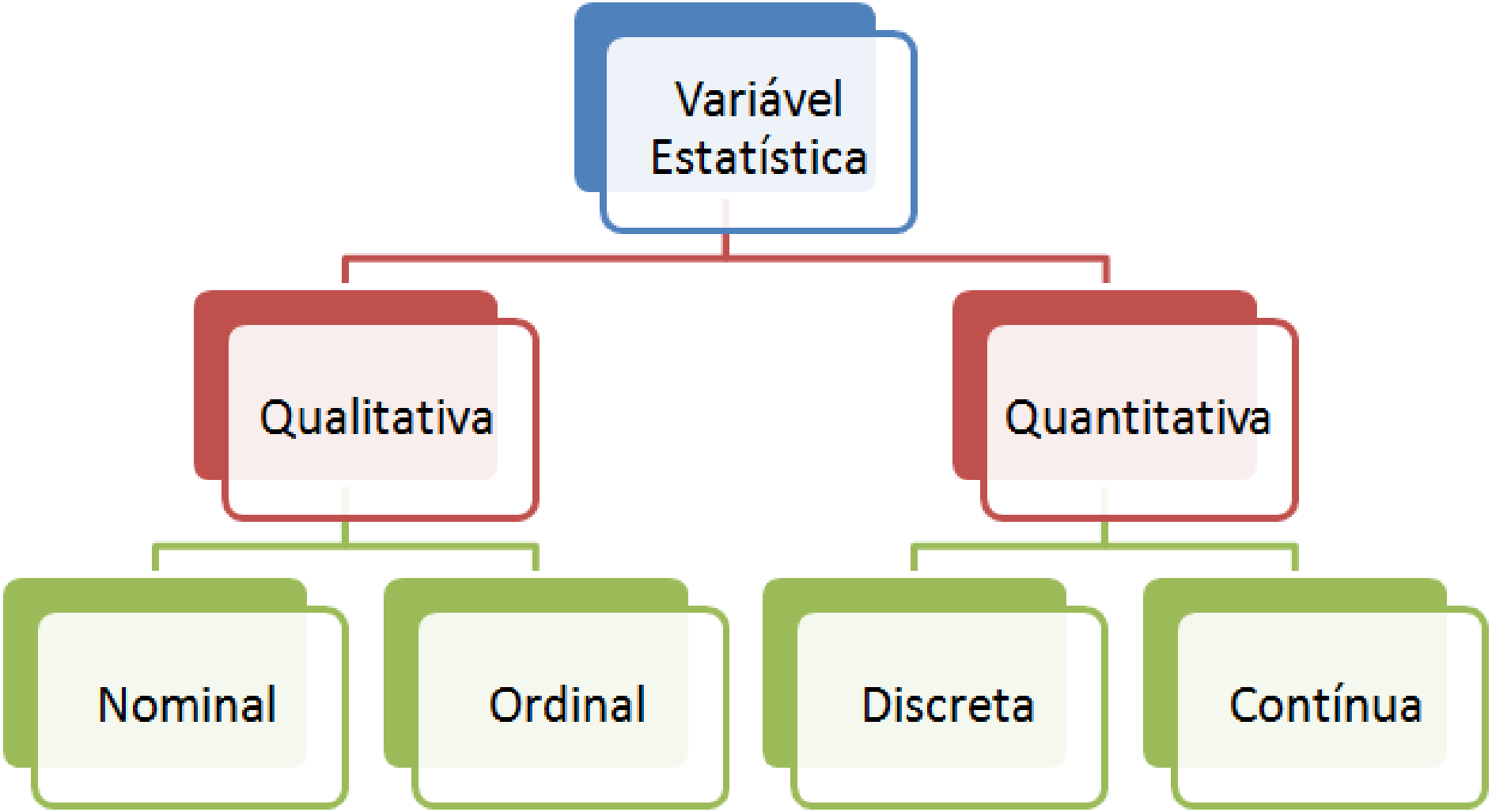
- **População:** O conjunto de todas as empresas portuguesas
- **Unidade Estatística:** Empresas portuguesas

- **Uma possível Amostra:** o conjunto das empresas do norte do país

- **Possível Variável estatística de interesse:** Ramo de atividade
- **Dados Estatísticos:** alimentação, brinquedos, moda, informática, lazer,...

- **Possível Variável estatística de interesse:** Volume de vendas mensal
- **Dados Estatísticos:** valores em euros no intervalo $[0, +\infty[$


Variável Estatística
Qualitativa
Quantitativa
Nominal
Ordinal
Discreta
Contínua



Como os dados estatísticos correspondem aos valores observados da variável em estudo, então a sua classificação é idêntica à das variáveis: **Dados Qualitativos e Quantitativos**

Dados Qualitativos e Quantitativos

Dados Qualitativos

Representam a informação que identifica alguma qualidade, categoria ou característica, não suscetível de medida, mas de classificação. Registam-se numa escala:

- **Nominal** - a ordem das categorias não tem significado
  - **Exemplos:**
    - Género: feminino, masculino
    - Cor dos olhos: pretos, castanhos, azuis,...
    - Grupo sanguíneo: O-, O+, A-, A+, B-, B+, AB-, AB+

- **Ordinal** - há uma ordem natural das categorias
  - **Exemplos:**
    - Nível de escolaridade: 1º ciclo, 2º ciclo, 3º ciclo,...
    - Classe social: baixa, média, alta
    - Fases de uma doença: inicial, intermédio, terminal

Dados Qualitativos e Quantitativos

Dados Quantitativos

Representam a informação resultante de características suscetíveis de serem medidas, apresentando-se com diferentes intensidades. Registam-se numa escala:

- **Discreta** - os valores podem ordenar-se, mas entre dois valores consecutivos não pode existir um valor intermédio (ou seja, o domínio da variável é um conjunto finito ou infinito numerável) - associado a contagens
  - **Exemplos:**
    - número de letras no nome,
    - número de assoalhadas numa casa,
    - número de cigarros fumados por dia.

- **Contínua** - pode tomar qualquer valor num certo intervalo (ou seja, o seu domínio é um conjunto de números reais) - associado a medições.
  - **Exemplos:**
    - tempos efetuados por um atleta para correr os 100 metros,
    - altura das pessoas,
    - peso de objetos.

Dados Qualitativos e Quantitativos

Observações:

- Muitas vezes os dados qualitativos (nominais ou ordinais) podem ser representados numericamente, isto é, são associados valores numéricos às diferentes categorias. Por exemplo, é possível associar os valores 1 e 2 às categorias masculino e feminino da variável género. Ou os valores 1, 2 e 3 às categorias baixo, médio e alto da variável classe social. Mas estes números são apenas símbolos para representar as categorias (nos dados qualitativos ordinais a numeração é feita de forma a respeitar a ordem).

  Estes valores numéricos não têm qualquer significado quantitativo, é apenas uma codificação.

- Os dados quantitativos são valores numéricos e estes números têm significado.
- As escalas de atitude também chamadas escalas de likert (por exemplo, escalas do tipo 1 a 5, onde 1 significa nada satisfeito e 5 significa muito satisfeito) são variáveis qualitativas ordinais, no entanto na prática são muitas vezes consideradas escalas de intervalos e são analisadas como variáveis quantitativas.
- Os dados originalmente podem ser quantitativos, mas podem ser recolhidos de forma qualitativa. Por exemplo, a variável idade, medida em anos é quantitativa (contínua), mas, se for obtida apenas a faixa etária (0 a 5 anos, 6 a 10 anos, ...), é qualitativa (ordinal).

Exemplo 3

Identifique a População e a sua dimensão, Amostra e a sua dimensão, Unidade Estatística, Variável Estatística e Dados estatísticos, classificando-os.

Numa fábrica produziram-se 1000 queijos durante um dia. Para classificar a qualidade do queijo, produzido nesse dia, em "mau", "razoável" ou "bom" foram retirados 10 queijos que foram testados.

Exemplo 3

Identifique a População e a sua dimensão, Amostra e a sua dimensão, Unidade Estatística, Variável Estatística e Dados estatísticos, classificando-os.

Numa fábrica produziram-se 1000 queijos durante um dia. Para classificar a qualidade do queijo, produzido nesse dia, em "mau", "razoável" ou "bom" foram retirados 10 queijos que foram testados.

- População:1000 queijos produzidos na fábrica durante um dia
- Dimensão da População: 1000 queijos
- Amostra: 10 queijos produzidos na fábrica durante um dia
- Dimensão da amostra: 10 queijos
- Unidade estatística: Cada queijo produzido na fábrica durante um dia
- Variável Estatística em estudo: Qualidade do queijo produzido na fábrica
- Dados Estatísticos: "mau", "razoável" ou "bom"
- Classificação da variável (ou dos dados) em estudo: Qualitativa Ordinal

Exemplo 4

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

Número de ordem	Dieta equilibrada	Intensidade dos treinos	Número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	14.3
2	Sim	Elevada	2	7.8
3	Não	Baixa	5	27.0
4	Sim	Moderada	6	11.0
5	Sim	Elevada	6	9.9
6	Não	Baixa	3	14.5
7	Sim	Baixa	4	15.4
8	Não	Baixa	4	20.8
9	Não	Elevada	7	10.5
10	Sim	Baixa	3	15.9

1. Identifique a População e a sua dimensão, Amostra e a sua dimensão, Unidade Estatística, Variável Estatística e Dados estatísticos, classificando-os.

Exemplo 4

- População: atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia
- a dimensão da população é desconhecida

- Amostra: 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia
- dimensão da amostra: $n = 10$

- Unidade estatística: atletas

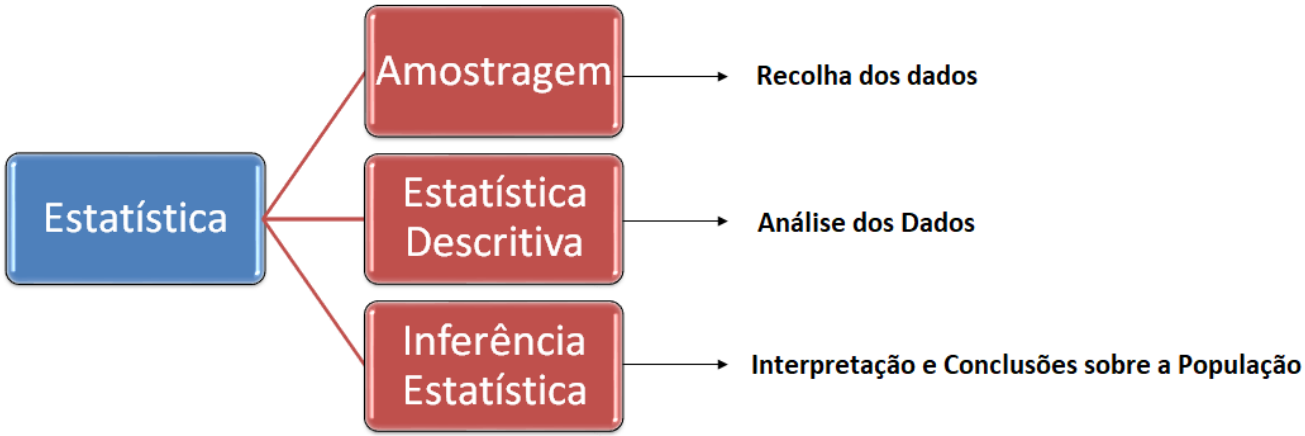
- Variável estatística: Dieta equilibrada
- Dados estatísticos: Sim, Sim, Não, Sim, Sim, Não, Sim, Não, Não, Sim
- Classificação: Qualitativa Nominal

Exemplo 4

- Variável estatística: Intensidade dos treinos
- Dados estatísticos: Moderada, Elevada, Baixa, Moderada, Elevada, Baixa, Baixa, Baixa, Elevada, Baixa
- Classificação: Qualitativa Ordinal

- Variável estatística: Número de suplementos alimentares (por semana)
- Dados estatísticos: 3, 2, 5, 6, 6, 3, 4, 4, 7, 3
- Classificação: Quantitativa Discreta

- Variável estatística: Nível de ferro (mg)
- Dados estatísticos: 14.3, 7.8, 27.0, 11.0, 9.9, 14.5, 15.4, 20.8, 10.5, 15.9
- Classificação: Quantitativa Contínua



O objetivo é a Estatística Descritiva, logo vamos supor que a primeira fase da análise estatística já foi efetuada:

- Estabelecer o objetivo de análise e definir a População e as Variáveis de interesse.
- Definir qual o método mais adequado para a recolha dos dados
- Recolher os dados e assim obter a Amostra.

Supondo que a Amostra já foi recolhida, passamos à segunda fase: a Estatística Descritiva. Primeiro é necessário organizar os dados recolhidos:

- Colocar toda a informação recolhida numa tabela de modo a respeitar a **regra**: 1 observação por linha e 1 coluna por variável
- Analisar a Amostra recolhida com recurso a meios computacionais

Meios Computacionais - Programa Estatístico

- O *R* é uma linguagem de programação "open source" para análise de dados que fornece uma grande variedade de ferramentas estatísticas e gráficas que inclui:
  - uma linguagem de programação com as mais comuns estruturas de programação
  - ligação a outras linguagens de programação ou outro software
  - uma coleção integrada de ferramentas para análise de dados

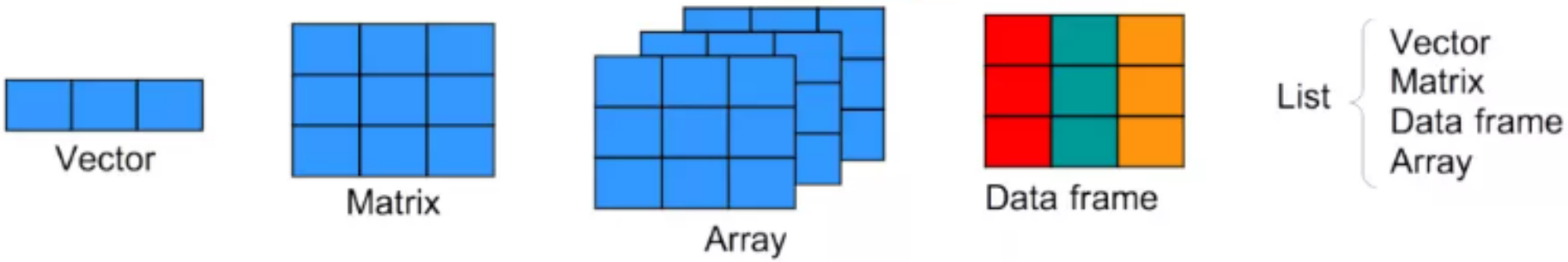
- O *RStudio* é um ambiente de desenvolvimento que permite utilizar diversas ferramentas que facilitam a utilização do *R*.

- Vamos utilizar o *R* a partir do *RStudio*:
  - instalar primeiro o *R* (*link* no Moodle)
  - instalar o *RStudio* depois de instalar o *R* (*link* no Moodle)

- O *RStudio* apresenta 4 janelas: editor, consola do R, *environment* e *output*.
- Os principais são:
  - editor: *script* onde se escreve o código
  - consola: onde corre o código e onde se recebe os resultados
- os restantes são auxiliares, por exemplo:
  - Environment: painel com todos os objetos criados na sessão
  - History: painel com um histórico dos comandos usados
  - Files: mostra os arquivos no diretório de trabalho. É possível navegar entre diretórios
  - Plots: painel onde os gráficos serão apresentados
  - Packages: apresenta todos os pacotes instalados e carregados
  - Help: janela onde a documentação das funções serão apresentadas

Estrutura básica de dados do R:

- Vector: coleção ordenada de elementos do mesmo tipo $\mapsto$ c()
- Matriz: generalização bidimensional de vector, com elementos do mesmo tipo $\mapsto$ matrix()
- Array: generalização multidimensional de vector, com elementos do mesmo tipo $\mapsto$ array()
- Data frame: como a matriz, mas com colunas de diferentes tipos $\mapsto$ data.frame()
- Lista: conjuntos de dados de diferentes tipos e dimensões $\mapsto$ list()



Vamos trabalhar principalmente com "**Vectores**" e "**Data frames**".

Escrever os dados:

- o símbolo # é utilizado para inserir comentários no script
- os símbolos $<-$ e $=$ são utilizados para atribuir valores ou nomes
- faz a distinção entre maiúsculas e minúsculas
- para visualizar todos os argumentos de uma função ou explicação sobre eles basta colocar o símbolo ? ou usar a função *help()*
- para escrever texto deve colocar os caracteres entre aspas

```
# escrever um vetor (feminino, masculino) com nome "genero"
genero <- c("feminino", "masculino")


# escrever uma tabela com o vetor "genero" e o vetor idade (20, 18)
tabela <- data.frame(genero=c("feminino", "masculino"), idade=c(20, 18))

# ver informação sobre "data.frame"
?data.frame()
help("data.frame")
```

Leitura dos dados:

- ficheiros de texto (.txt, .csv):

```
dados <- read.table(file, header = FALSE, sep = "", dec = ".", . . .)


         read.csv(file, . . .)
```

Através do RStudio basta fazer: "File $\mapsto$ Import Dataset $\mapsto$ From Text..."

- folhas de cálculo (.xls, .xlsx)

```
library(readxl)
dados <- read_excel(path, sheet = NULL, range = NULL,
                    col_names = TRUE, . . .)
```

Através do RStudio basta fazer: "File $\mapsto$ Import Dataset $\mapsto$ From Excel..."

Neste caso é necessário ter o package "readxl" instalado: no RStudio basta ir à janela "Packages", carregar em "Install" e escrever o nome do package pretendido e carregar em "Install".

Observação: Quando se instala packages que não pertencem à base do R, como é o caso do "readxl", só é necessário instalar uma vez, depois basta fazer "library(nome do package)" para abrir.

Exemplo 4

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

Número de ordem	Dieta equilibrada	Intensidade dos treinos	Número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	14, 3
2	Sim	Elevada	2	7, 8
3	Não	Baixa	5	27, 0
4	Sim	Moderada	6	11, 0
5	Sim	Elevada	6	9, 9
6	Não	Baixa	3	14, 5
7	Sim	Baixa	4	15, 4
8	Não	Baixa	4	20, 8
9	Não	Elevada	7	10, 5
10	Sim	Baixa	3	15, 9

3. Introduza a tabela no R.

Exemplo 4

Introdução da tabela no R:
a tabela já está na forma adequada para fazer uma análise exploratória de dados no R: 1 observação por linha e 1 coluna por variável.

- introduzir diretamente os dados:

```
dados <- data.frame(i=c(1:10),
                    dieta=c("Sim","Sim","Não", "Sim", "Sim", "Não", "Sim", "Não", "Não", "Sim"),
                    intensidade=c("Moderada", "Elevada", "Baixa", "Moderada", "Elevada", "Baixa",
                                  "Baixa", "Baixa", "Elevada", "Baixa"),
                    suplementos=c(3, 2, 5, 6, 6, 3, 4, 4, 7, 3),
                    ferro=c(14.3, 7.8, 27.0, 11.0, 9.9, 14.5, 15.4, 20.8, 10.5, 15.9))
```

- guardar os dados num ficheiro:

```
# escrever um ficheiro .txt
write.table(dados, file="dados.txt", quote=FALSE, row.names=FALSE)

# escrever um ficheiro .csv
write.csv(dados, file = "dados.csv")
```

- Importar os dados a partir do ficheiro:

  RStudio: "File $\mapsto$ Import Dataset $\mapsto$ From Text..."

data.frame - duas dimensões

- ver a informação da tabela

```
View(dados)  # ver a tabela toda
head(dados)  # ver as primeiras linhas da tabela
tail(dados)  # ver as últimas linhas da tabela

names(dados)  # ver os nomes das variáveis
str(dados)    # ver a estrutura da tabela

dim(dados)    # número de linhas e colunas da tabela
nrow(dados)   # número de linhas da tabela
ncol(dados)   # número de colunas da tabela
```

- acesso à informação da tabela por posição

```
dados[1,2]        # elemento da linha 1 e coluna 2
dados[1,]         # linha 1
dados[,2]         # coluna 2
dados[, c(2,4)]   # coluna 2 e coluna 4
dados[, 2:4]      # coluna 2, coluna 3 e coluna 4
dados[ c(2,4),]   # linha 2 e linha 4
dados[2:4,]       # linha 2, linha 3 e linha 4
```

data.frame - duas dimensões

- acesso à informação da tabela por nome das variáveis

```
dados$dieta                          # ver a variável "dieta"
dados[,c("dieta", "ferro")]  # ver as variáveis "dieta" e "ferro"
```

- acesso condicional à informação da tabela

```
dados[dados$dieta=="Sim",]          # tabela com as linhas onde a dieta é "Sim"
dados[dados$suplementos>3,]         # tabela com as linhas onde suplementos são >3
dados[dados$intensidade!="Baixa",] # tabela com as linhas onde a intensidade não é "Baixa"
```

  - operadores lógicos no R:

    - ★ igual (comparar): $==$
    - ★ diferente: $!=$
    - ★ maior: $>$
    - ★ maior ou igual: $>=$
    - ★ menor: $<$
    - ★ menor ou igual: $<=$
    - ★ pertencer: $\%in\%$
    - ★ ou: $|$
    - ★ e: $\&$

vector - uma dimensão

- vector

```
dados$dieta              # cada variável da tabela é um vector
v <- c(15,14,7,8)        # escrever um vector
```

- comprimento de um vector (número de elementos)

```
length(dados$dieta)
length(v)
```

- aceder à informação do vector por posição

```
dados$dieta[3]           # dado na posição 3
dados$dieta[c(3,7)]      # dados nas posições 3 e 7
dados$dieta[3:7]         # dados nas posições 3 a 7
```

Análise de uma base de dados

Antes de efetuar qualquer análise estatística é necessário fazer uma "limpeza" inicial aos dados, de modo a garantir que as conclusões são consistentes. Essa "limpeza" consiste em:

- verificar se existem valores absurdos que só podem ser erros e eliminar esses dados (ou corrigir, caso seja possível);
- verificar se existem dados omissos (em geral representados por NA), registar esse facto e, caso não seja adequado trabalhar com essa falta de informação, retirar os indivíduos nessa situação.

Atenção: Observações discordantes (também chamadas de "outliers") podem não ser erros, mas apenas valores que são possíveis de observar em situações raras.

Dados em falta ou Dados omissos

- no R os valores em falta são indicados com NA
  - usa-se a função "is.na()" para testar a presença de NA

```
dados2 <- data.frame(var1=2:5, var2=c(1,NA,6,12))
dados2

# ver se tem NA
is.na(dados2)
any(is.na(dados2))
```

  - muitas funções têm o argumento "na.rm" para remover valores em falta antes dos cálculos

```
# usar a função soma
sum(dados2$var2)
sum(dados2$var2, na.rm=TRUE)
```

  - a função "na.omit()" permite criar uma nova tabela sem as linhas que têm pelo menos um NA

```
# retirar as linhas que têm pelo menos 1 NA
na.omit(dados2)

dados3 <- na.omit(dados2)
dados3
any(is.na(dados3))
sum(dados3$var2)
```

Organização e interpretação de dados

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados por variável. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados por variável;
- a representações gráficas dos dados por variável;
- ao cálculo de indicadores numéricos por variável, indicadores de localização e dispersão.

Organização e interpretação de dados

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados por variável. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados por variável;
- a representações gráficas dos dados por variável;
- ao cálculo de indicadores numéricos por variável, indicadores de localização e dispersão.

Tabela de frequências

Tabelas de Frequências

- Numa tabela de frequências a informação é organizada, de um modo geral, em 3 colunas:
  1. Coluna dos valores ou modalidades que as variáveis podem assumir, caso sejam variáreis quantitativas ou qualitativas, respetivamente;
  2. Coluna das frequências absolutas;
  3. Coluna das frequências relativas.

- Podem, ainda, ser acrescentadas mais duas colunas, com as frequências acumuladas (não tem interesse nem significado no caso das variáveis qualitativas nominais):
  4. Frequência absoluta acumulada;
  5. Frequência relativa acumulada.

Tabelas de Frequências

- **Frequência absoluta** de um valor x_i da variável é o número de vezes que esse valor foi observado. Representa-se habitualmente por n_i.
  - A soma das frequências absolutas é igual à dimensão da amostra (ou à dimensão da população, caso tenham sido recolhidos dados relativos a todos os indivíduos da população).

- **Frequência relativa** de um valor da variável é o quociente entre a frequência absoluta desse valor e o número n de elementos da população (ou da amostra). Representa-se habitualmente por f_i.
  - É sempre um número entre 0 e 1.
  - Pode ser expressa em percentagem desde que se multiplique o número obtido por 100.
  - A soma das frequências relativas é igual a 1.

Tabelas de Frequências

As Tabelas de Frequências constroem-se de maneira diferente, consoante o tipo de variável.

Assim temos Tabelas de Frequências para:

- Variáveis Qualitativas Nominais - não incluem as frequências acumuladas
- Variáveis Qualitativas Ordinais ou Variáveis Quantitativas Discretas (com número pequeno de valores distintos) - incluem as frequências acumuladas
- Variáveis Quantitativas Contínuas ou Variáveis Quantitativas Discretas (com número elevado de valores distintos) - Neste caso há a necessidade de agrupar os dados em classes e incluem as frequências acumuladas

Tabela de Frequências

Variáveis qualitativas nominais

Valor da variável	Frequências Absolutas	Frequências Relativas
x_i	n_i	f_i
x_1	n_1	$f_1 = \frac{n_1}{n}$
x_2	n_2	$f_2 = \frac{n_2}{n}$
$\ldots$	$\ldots$	$\ldots$
x_k	n_k	$f_k = \frac{n_k}{n}$
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$

- Frequência absoluta (n_i) — número de observações iguais a x_i
- Frequência relativa (f_i) — proporção de observações iguais a x_i

Tabela de Frequências

Variáveis qualitativas ordinais ou quantitativas discretas

Valor da variável	Frequências Absolutas	Frequências Relativas	Frequências Absolutas Acumuladas	Frequências Relativas Acumuladas
x_i	n_i	f_i	N_i	F_i
x_1	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	$f_2 = \frac{n_2}{n}$	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_k	n_k	$f_k = \frac{n_k}{n}$	$N_k = \sum_{i=1}^{k} n_i = n$	$F_k = \sum_{i=1}^{k} f_i = 1$
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$		

- Frequência absoluta (n_i) — número de observações iguais a x_i
- Frequência relativa (f_i) — proporção de observações iguais a x_i
- Frequência absoluta acumulada (N_i) — número de observações menores ou iguais a x_i
- Frequência relativa acumulada (F_i) — proporção de observações menores ou iguais a x_i

Tabela de Frequências

Exemplo 4

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

Número de ordem	Dieta equilibrada	Intensidade dos treinos	número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	$14,3$
2	Sim	Elevada	2	$7,8$
3	Não	Baixa	5	$27,0$
4	Sim	Moderada	6	$11,0$
5	Sim	Elevada	6	$9,9$
6	Não	Baixa	3	$14,5$
7	Sim	Baixa	4	$15,4$
8	Não	Baixa	4	$20,8$
9	Não	Elevada	7	$10,5$
10	Sim	Baixa	3	$15,9$

3. Construa as tabelas de frequências das variáveis qualitativas e quantitativas discretas.

- Variável estatística: Dieta equilibrada
- Classificação: qualitativa nominal

número da linha	Níveis da variável	Frequências Absolutas	Frequências Relativas
i	x_i	n_i	f_i
1	Sim	6	0.6
2	Não	4	0.4
		$n = 10$	1

- Variável estatística: Intensidade dos treinos
- Classificação: qualitativa ordinal

número da linha	Níveis da variável	Frequências Absolutas	Frequências Relativas	Frequências Absolutas Acumuladas	Frequências Relativas Acumuladas
i	x_i	n_i	f_i	N_i	F_i
1	Baixa	5	0.5	5	0.5
2	Moderada	2	0.2	7	0.7
3	Elevada	3	0.3	10	1
		$n = 10$	1		

- Variável estatística: Número de suplementos alimentares
- Classificação: quantitativa discreta

número da linha	Níveis da variável	Frequências Absolutas	Frequências Relativas	Frequências Absolutas Acumuladas	Frequências Relativas Acumuladas
i	x_i	n_i	f_i	N_i	F_i
1	2	1	0.1	1	0.1
2	3	3	0.3	4	0.4
3	4	2	0.2	6	0.6
4	5	1	0.1	7	0.7
5	6	2	0.2	9	0.9
6	7	1	0.1	10	1
		$n = 10$	1		

Tabela de Frequências

Variáveis quantitativas contínuas

(ou variáveis quantitativas discretas com um número elevado de valores distintos)

- Neste caso há a necessidade de agrupar os dados em **classes**.
- Existem regras para construir classes.
- Podem ser usadas classes predefinidas.
- Habitualmente constroem-se **classes de igual amplitude**.

Uma possível regra para definir classes com a mesma amplitude:

Regra de Sturges

Para organizar uma amostra, de dados contínuos, de dimensão n, pode considerar-se para número de classes o valor k, onde k é o menor inteiro tal que $2^k > n$.
Da inequação anterior pode deduzir-se o seguinte resultado:

$$k = \lfloor 1 + \log_2 n \rfloor = \left\lfloor 1 + \frac{\ln n}{\ln 2} \right\rfloor$$

onde n é o número de dados e $\lfloor a \rfloor$ representa a parte inteira de a.

Formação das classes

- determinar o máximo $(max(x_i))$ e o mínimo $(min(x_i))$ dos dados
- a amplitude de cada classe é

$$h = \frac{max(x_i) - min(x_i)}{k}$$

  Se for necessário arredondar, deve ser sempre arredondado por excesso.
- Formar as classes:

- Classes como intervalos semiabertos, abertos à esquerda e fechados à direita:
  - $c_1 =]b_0; b_1]$ com $b_1 = b_0 + h$
  - $c_2 =]b_1; b_2]$ com $b_2 = b_1 + h$
  - $c_3 =]b_2; b_3]$ com $b_3 = b_2 + h$
  - $c_4 =]b_3; b_4]$ com $b_4 = b_3 + h$
  - ...
  - $c_k =]b_{k-1}; b_k]$ com $b_k = b_{k-1} + h$

  Se o extremo esquerdo do primeiro intervalo for o mínimo dos dados então o primeiro intervalo da tabela de frequências é fechado à esquerda e à direita:

$$c_1 = [b_0; b_1] \text{ com } b_0 = min(x_i)$$

- Classes como intervalos semiabertos, fechados à esquerda e abertos à direita:
  - $c_1 = [b_0; b_1[$ com $b_1 = b_0 + h$
  - $c_2 = [b_1; b_2[$ com $b_2 = b_1 + h$
  - $c_3 = [b_2; b_3[$ com $b_3 = b_2 + h$
  - $c_4 = [b_3; b_4[$ com $b_4 = b_3 + h$
  - ...
  - $c_k = [b_{k-1}; b_k[$ com $b_k = b_{k-1} + h$

  Se o extremo direito do último intervalo for o máximo dos dados então o último intervalo da tabela de frequências é fechado à esquerda e à direita:

$$c_k = [b_{k-1}; b_k] \text{ com } b_k = max(x_i)$$

Tabela de Frequências

Variáveis quantitativas contínuas

(ou variáveis quantitativas discretas com um número elevado de valores distintos)

- Uma vez escolhidas as classes, a construção da tabela de frequências é idêntica à considerada para dados discretos.
  1. Coluna das classes onde se indicam todas as classes definidas.
  2. Coluna das frequências absolutas.
  3. Coluna das frequências relativas.

- Podem, ainda, existir mais duas colunas:
  4. Coluna das frequências absolutas acumuladas.
  5. Coluna das frequências relativas acumuladas.

Tabela de Frequências

Variáveis quantitativas contínuas

(ou variáveis quantitativas discretas com um número elevado de valores distintos)

Classe c_i	Frequências Absolutas n_i	Frequências Relativas f_i	Frequências Absolutas Acumuladas N_i	Frequências Relativas Acumuladas F_i
$c_1 =]b_0; b_1]$	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = f_1$
$c_2 =]b_1; b_2]$	n_2	$f_2 = \frac{n_2}{n}$	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$c_k =]b_{k-1}; b_k]$	n_k	$f_k = \frac{n_k}{n}$	$N_k = \sum_{i=1}^{k} n_i = n$	$F_k = \sum_{i=1}^{k} f_i = 1$
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$		

- Classes (c_i) — intervalos semiabertos, abertos à esquerda e fechados à direita
- Frequência absoluta (n_i) — número de observações que pertencem à classe c_i
- Frequência relativa (f_i) — proporção de observações que pertencem à classe c_i
- Frequência absoluta acumulada (N_i) — número de observações menores ou iguais que o extremo superior da classe c_i
- Frequência relativa acumulada (F_i) — proporção de observações menores ou iguais que o extremo superior da classe c_i

Tabela de Frequências

Variáveis quantitativas contínuas

(ou variáveis quantitativas discretas com um número elevado de valores distintos)

Classe c_i	Frequências Absolutas n_i	Frequências Relativas f_i	Frequências Absolutas Acumuladas N_i	Frequências Relativas Acumuladas F_i
$c_1 = [b_0; b_1[$	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = f_1$
$c_2 = [b_1; b_2[$	n_2	$f_2 = \frac{n_2}{n}$	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$c_k = [b_{k-1}; b_k[$	n_k	$f_k = \frac{n_k}{n}$	$N_k = \sum_{i=1}^{k} n_i = n$	$F_k = \sum_{i=1}^{k} f_i = 1$
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$		

- Classes (c_i) — intervalos semiabertos, fechados à esquerda e abertos à direita
- Frequência absoluta (n_i) — número de observações que pertencem à classe c_i
- Frequência relativa (f_i) — proporção de observações que pertencem à classe c_i
- Frequência absoluta acumulada (N_i) — número de observações menores que o extremo superior da classe c_i
- Frequência relativa acumulada (F_i) — proporção de observações menores que o extremo superior da classe c_i

Tabela de Frequências

Exemplo 4

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

Número de ordem	Dieta equilibrada	Intensidade dos treinos	número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	$14,3$
2	Sim	Elevada	2	$7,8$
3	Não	Baixa	5	$27,0$
4	Sim	Moderada	6	$11,0$
5	Sim	Elevada	6	$9,9$
6	Não	Baixa	3	$14,5$
7	Sim	Baixa	4	$15,4$
8	Não	Baixa	4	$20,8$
9	Não	Elevada	7	$10,5$
10	Sim	Baixa	3	$15,9$

4. Suponha que os níveis de ferro são considerados muito baixos quando são inferiores a 15 mg, baixos se estiverem no intervalo $[15, 20[$ e adequados se forem no mínimo de 20 mg e no máximo 30 mg. Com base nestas classes construa a tabela de frequências da variável quantitativa.

- Variável estatística: Nível de ferro (mg)
- Classificação: Quantitativa Contínua
- número de classes $= k = 3$ classes
- classes: $c_1 = [0, 15[$, $c_2 = [15, 20[$, $c_3 = [20, 30]$

- mínimo da primeira classe $= 0$ mg
- mínimo dos dados $= 7.8$ mg
- máximo da última classe $= 30$ mg
- máximo dos dados $= 27.0$ mg

i	Nível de ferro c_i	Frequências Absolutas n_i	Frequências Relativas f_i	Frequências Absolutas Acumuladas N_i	Frequências Relativas Acumuladas F_i
1	$[0, 15[$	6	$\frac{6}{10} = 0.6$	6	0.6
2	$[15, 20[$	2	$\frac{2}{10} = 0.2$	$6 + 2 = 8$	$0.6 + 0.2 = 0.8$
3	$[20, 30]$	2	$\frac{2}{10} = 0.2$	$8 + 2 = 10$	$0.8 + 0.2 = 1$
		$n = 10$	1		

Tabela de Frequências

Exemplo 4

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

Número de ordem	Dieta equilibrada	Intensidade dos treinos	número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	14, 3
2	Sim	Elevada	2	7, 8
3	Não	Baixa	5	27, 0
4	Sim	Moderada	6	11, 0
5	Sim	Elevada	6	9, 9
6	Não	Baixa	3	14, 5
7	Sim	Baixa	4	15, 4
8	Não	Baixa	4	20, 8
9	Não	Elevada	7	10, 5
10	Sim	Baixa	3	15, 9

5. Construa a tabela de frequências da variável quantitativa considerando 3 classes, a primeira classe começa no valor 5 mg e cada classe tem amplitude 10 mg (classes abertas à esquerda e fechadas à direita).

- Variável estatística: Nível de ferro (mg)
- Classificação: quantitativa contínua
- Número de classes $= k = 3$ classes
- Amplitude das classes $= h = 10$ mg

- mínimo da primeira classe $= 5$ mg
- mínimo dos dados $= 7.8$ mg
- máximo da última classe $= 35$ mg
- máximo dos dados $= 27.0$ mg

número da linha	Classes da variável	Frequências Absolutas	Frequências Relativas	Frequências Absolutas Acumuladas	Frequências Relativas Acumuladas
i	classe$_i$	n_i	f_i	N_i	F_i
1	$]5, 15]$	6	0.6	6	0.6
2	$]15, 25]$	3	0.3	9	0.9
3	$]25, 35]$	1	0.1	10	1
		$n = 10$	1		

Tabela de Frequências

Exemplo 4

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

Número de ordem	Dieta equilibrada	Intensidade dos treinos	número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	$14,3$
2	Sim	Elevada	2	$7,8$
3	Não	Baixa	5	$27,0$
4	Sim	Moderada	6	$11,0$
5	Sim	Elevada	6	$9,9$
6	Não	Baixa	3	$14,5$
7	Sim	Baixa	4	$15,4$
8	Não	Baixa	4	$20,8$
9	Não	Elevada	7	$10,5$
10	Sim	Baixa	3	$15,9$

6. Construa a tabela de frequências da variável quantitativa contínua recorrendo à Regra de Sturges.

- Variável estatística: Nível de ferro (mg)
- Classificação: quantitativa contínua

- Número de classes $= k = \left\lfloor 1 + \frac{\ln 10}{\ln 2} \right\rfloor = \lfloor 4.32 \rfloor = 4$ classes

- Amplitude das classes $= h = \frac{27.0 - 7.8}{4} = 4.8$ mg

- mínimo da primeira classe = mínimo dos dados $= 7.8$ mg
- máximo da última classe = máximo dos dados $= 27.0$ mg

número da linha	Classes da variável	Frequências Absolutas	Frequências Relativas	Frequências Absolutas Acumuladas	Frequências Relativas Acumuladas
i	classe$_i$	n_i	f_i	N_i	F_i
1	$[7.8, 12.6]$	4	0.4	4	0.4
2	$]12.6, 17.4]$	4	0.4	8	0.8
3	$]17.4, 22.2]$	1	0.1	9	0.9
4	$]22.2, 27.0]$	1	0.1	10	1
		$n = 10$	1		

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados por variável. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados por variável;
- a representações gráficas dos dados por variável;
- ao cálculo de indicadores numéricos por variável, indicadores de localização e dispersão.

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados por variável. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados por variável;
- a representações gráficas dos dados por variável;
- ao cálculo de indicadores numéricos por variável, indicadores de localização e dispersão.

Gráficos

A principal vantagem dos gráficos, relativamente às tabelas, está na rapidez de leitura, pois permitem ter uma perceção imediata de quais as categorias de maior e menor frequência, assim como a ordem de grandeza de cada categoria relativamente às restantes.

Gráficos

Os gráficos mais usuais são:

- **Gráficos de Barras** – para representar graficamente dados qualitativos ou quantitativos discretos
- **Diagramas Circulares** – muito usados para representar graficamente dados qualitativos, mas também podem ser usados para representar dados quantitativos discretos
- **Histograma** – para representar graficamente dados quantitativos agrupados em classes, principalmente os dados quantitativos contínuos.

Gráficos de Barras

- Usados para representar graficamente dados qualitativos ou quantitativos discretos.
- No eixo horizontal colocam-se as modalidades ou categorias da variável em estudo e no eixo vertical colocam-se as frequências absolutas ou relativas.
- Constrói-se uma barra para cada modalidade ou categoria da variável em estudo, sendo a altura de cada barra igual à respetiva frequência absoluta ou relativa.
- Ao contrário das alturas das barras, a largura das barras não transmite qualquer informação. As barras devem ter todas a mesma largura (pois barras mais largas podem chamar mais a atenção, induzindo em erro) e a distância entre as barras deve ser a mesma.
- Quando não existe espaçamento entre as barras, as barras devem ter obrigatoriamente de cores diferentes.
- A metodologia apresentada refere-se a gráficos de barras verticais. Se trocar os eixos, então tem-se um gráfico de barras horizontal.

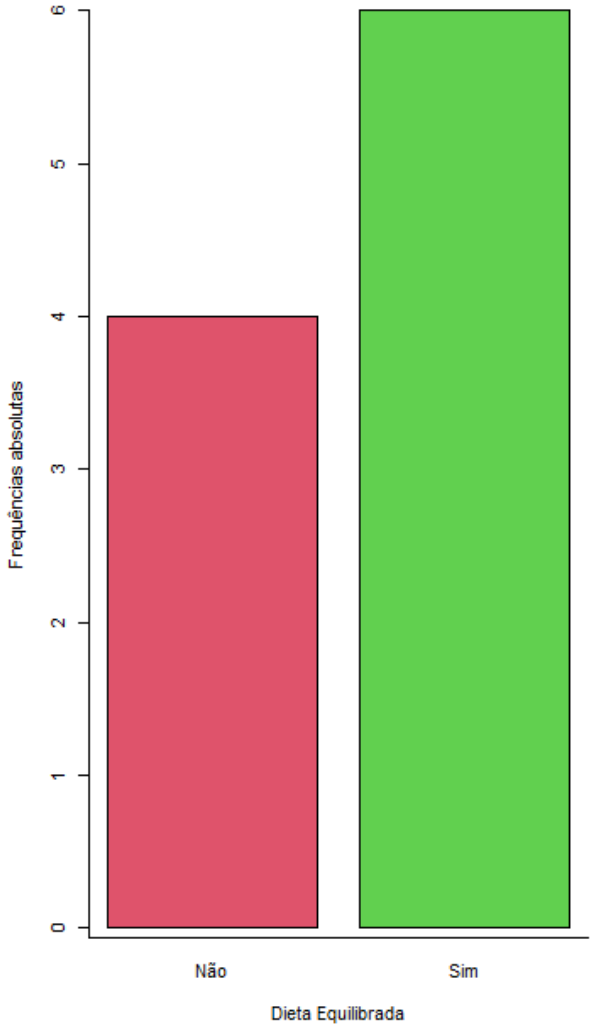
Exemplo 4

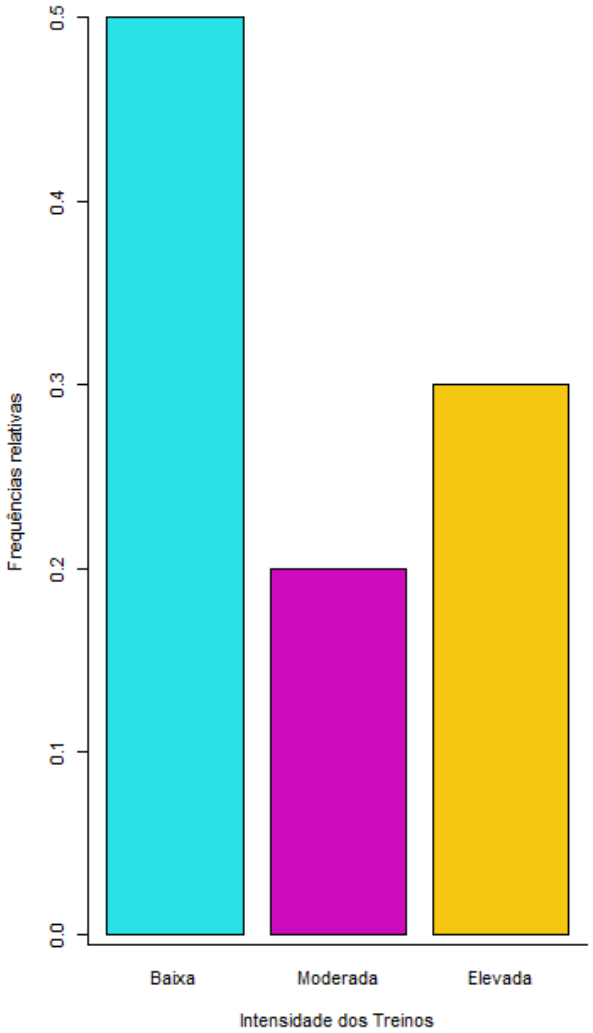
Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

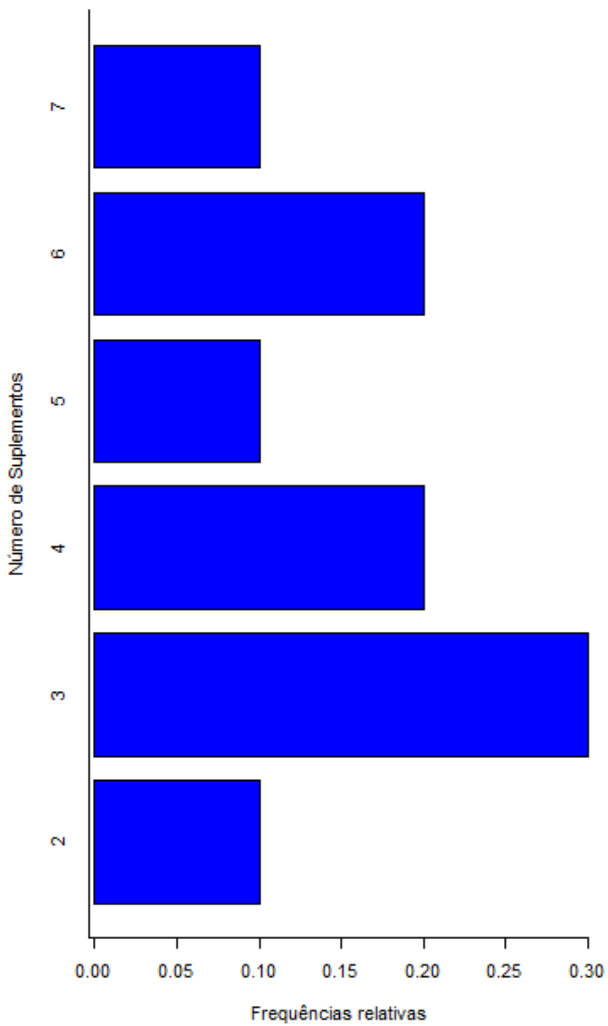
Número de ordem	Dieta equilibrada	Intensidade dos treinos	Número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	14, 3
2	Sim	Elevada	2	7, 8
3	Não	Baixa	5	27, 0
4	Sim	Moderada	6	11, 0
5	Sim	Elevada	6	9, 9
6	Não	Baixa	3	14, 5
7	Sim	Baixa	4	15, 4
8	Não	Baixa	4	20, 8
9	Não	Elevada	7	10, 5
10	Sim	Baixa	3	15, 9

7. Represente graficamente as variáveis qualitativas e quantitativas discretas.

Os gráficos de barras são feitos com base nas tabelas de frequências.







Gráficos de Barras

Exemplo 5

Os seguintes dados correspondem a respostas dadas por 30 pessoas de Lisboa e 50 pessoas do Porto sobre o desporto que praticam com mais frequência nos tempos livres:

Lisboa		
Desporto (x_i)	Freq. absolutas (n_i)	Freq. relativas (f_i)
Andebol	2	0.067
Atletismo	6	0.200
Basquetebol	1	0.033
Futebol	9	0.300
Ténis	5	0.167
Voleibol	7	0.233

Porto		
Desporto (x_i)	Freq. absolutas (n_i)	Freq. relativas (f_i)
Andebol	5	0.10
Atletismo	10	0.20
Basquetebol	9	0.18
Futebol	14	0.28
Ténis	5	0.10
Voleibol	7	0.14

Construa um gráfico de barras que permita comparar os dois conjuntos de dados.

Gráficos de Barras

Exemplo 5

Os seguintes dados correspondem a respostas dadas por 30 pessoas de Lisboa e 50 pessoas do Porto sobre o desporto que praticam com mais frequência nos tempos livres:

Lisboa		
Desporto (x_i)	Freq. absolutas (n_i)	Freq. relativas (f_i)
Andebol	2	0.067
Atletismo	6	0.200
Basquetebol	1	0.033
Futebol	9	0.300
Ténis	5	0.167
Voleibol	7	0.233

Porto		
Desporto (x_i)	Freq. absolutas (n_i)	Freq. relativas (f_i)
Andebol	5	0.10
Atletismo	10	0.20
Basquetebol	9	0.18
Futebol	14	0.28
Ténis	5	0.10
Voleibol	7	0.14

Construa um gráfico de barras que permita comparar os dois conjuntos de dados.

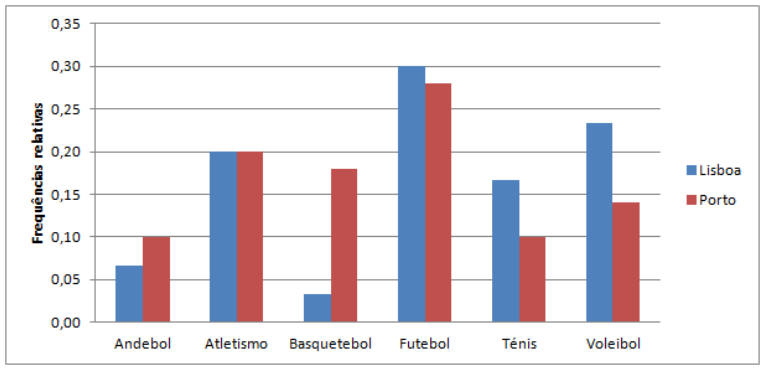


Diagrama Circular

- São mais usados para representar graficamente **dados qualitativos**.
- Esta representação é constituída por um círculo dividido em sectores.
- Tem tantos sectores circulares quantas as categorias ou classes consideradas na tabela de frequências.
- Podem mostrar as frequências absolutas, mas, em geral, apresentam as frequências relativas sob a forma de percentagens.
- O ângulo de cada sector circular é proporcional à frequência observada na modalidade que lhe corresponde, isto é, o ângulo do sector i é $f_i \times 360^o$.

Zonas (x_i)	Número de casas (n_i)	Frequências relativas (f_i)
A	19	0.475
B	16	0.400
C	5	0.125

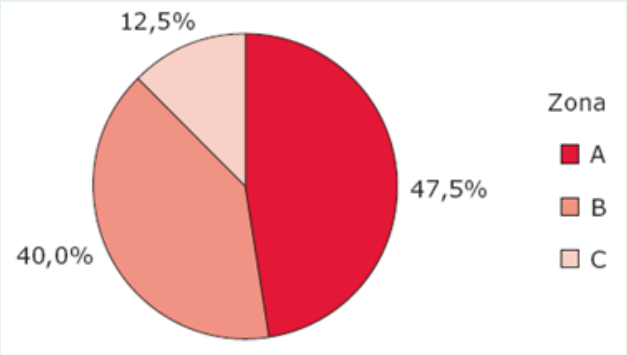


Diagrama Circular

Exemplo 6

Os seguintes dados correspondem ao número de vitórias, empates e derrotas de uma equipa desportiva durante um campeonato. Represente os dados recorrendo a um diagrama circular.

Resultados (x_i)	Frequências absolutas (n_i)	Frequências relativas (f_i)
vitória	10	0.40
empate	7	0.28
derrota	8	0.32
Total	25	1

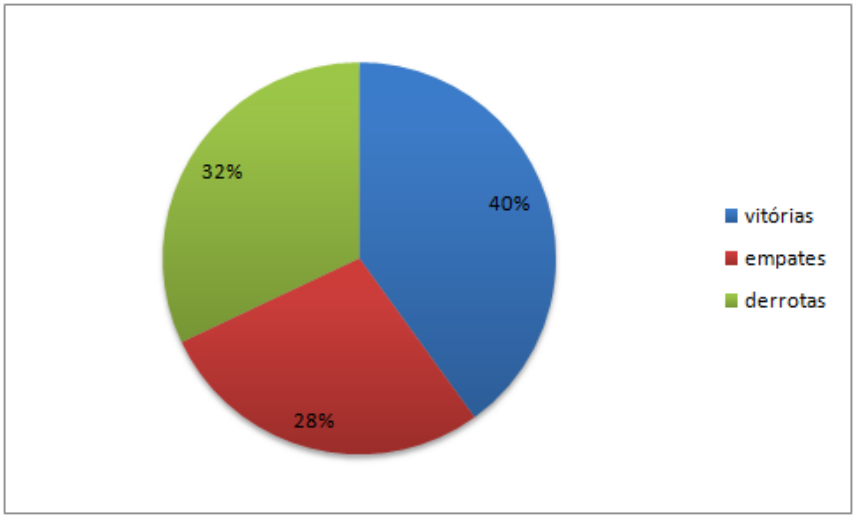
Diagrama Circular

Exemplo 6

Os seguintes dados correspondem ao número de vitórias, empates e derrotas de uma equipa desportiva durante um campeonato. Represente os dados recorrendo a um diagrama circular.

Resultados (x_i)	Frequências absolutas (n_i)	Frequências relativas (f_i)
vitória	10	0.40
empate	7	0.28
derrota	8	0.32
Total	25	1

Resultados (x_i)	Amplitude do ângulo $(f_i \times 360^o)$
vitória	144
empate	100.8
derrota	115.2
Total	360



Histogramas

- São usados para representar graficamente dados quantitativos agrupados em classes.
- É um gráfico formado por uma sucessão de retângulos adjacentes:
  - as barras são obrigatoriamente todas da mesma cor,
  - a base de cada retângulo representa uma classe,
  - Se as **classes têm todas a mesma amplitude**, então a **altura** de cada retângulo representa a frequência (relativa ou absoluta) com que os valores dessa classe ocorreram no conjunto de dados,
  - Se as **classes têm amplitudes diferentes**, então a **área** de cada retângulo representa a frequência (relativa ou absoluta) com que os valores dessa classe ocorreram no conjunto de dados,

Exemplo 4

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

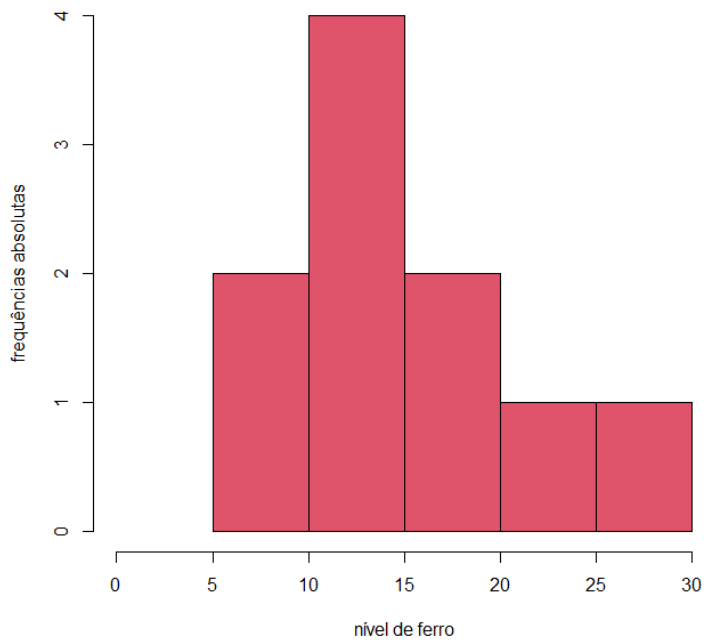
Número de ordem	Dieta equilibrada	Intensidade dos treinos	Número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	$14,3$
2	Sim	Elevada	2	$7,8$
3	Não	Baixa	5	$27,0$
4	Sim	Moderada	6	$11,0$
5	Sim	Elevada	6	$9,9$
6	Não	Baixa	3	$14,5$
7	Sim	Baixa	4	$15,4$
8	Não	Baixa	4	$20,8$
9	Não	Elevada	7	$10,5$
10	Sim	Baixa	3	$15,9$

8. Construa o histograma da variável Nível de ferro considerando 5 classes, a primeira classe começa no valor 5 mg e cada classe tem amplitude 5 mg.

Tabela de frequências:

i	Nível de ferro c_i	Frequências Absolutas n_i	Frequências Relativas f_i
1	$]5, 10]$	2	0.2
2	$]10, 15]$	4	0.4
3	$]15, 20]$	2	0.2
4	$]20, 25]$	1	0.1
5	$]25, 30]$	1	0.1

Como as **classes têm todas a mesma amplitude**, então a **altura** de cada retângulo pode ser representada pela frequência absoluta (ou pela frequência relativa).



Exemplo 4

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

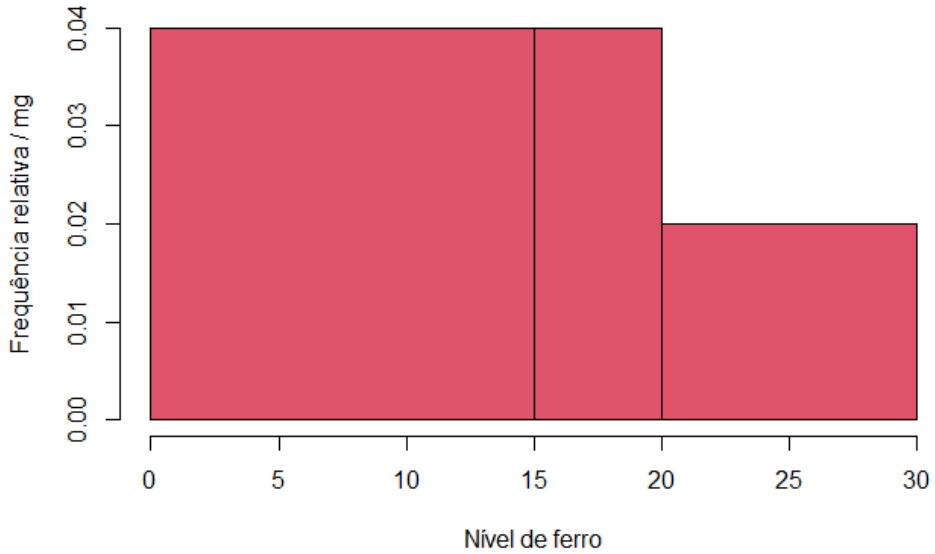
Número de ordem	Dieta equilibrada	Intensidade dos treinos	Número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	$14,3$
2	Sim	Elevada	2	$7,8$
3	Não	Baixa	5	$27,0$
4	Sim	Moderada	6	$11,0$
5	Sim	Elevada	6	$9,9$
6	Não	Baixa	3	$14,5$
7	Sim	Baixa	4	$15,4$
8	Não	Baixa	4	$20,8$
9	Não	Elevada	7	$10,5$
10	Sim	Baixa	3	$15,9$

9. Suponha que os níveis de ferro são considerados muito baixos quando são inferiores a 15 mg, baixos se estiverem no intervalo $[15, 20[$ e adequados se forem no mínimo de 20 mg e no máximo 30 mg. Com base nestas classes represente graficamente a variável Nível de ferro.

Considerando as classes predefinidas a tabela de frequências é

i	**Nível de ferro** c_i	**Frequências Absolutas** n_i	**Frequências Relativas** f_i	**Amplitude** h_i	**Altura** $\frac{f_i}{h_i}$
1	$[0, 15[$	6	0.6	$15 - 0 = 15$	$\frac{0.6}{15} = 0.04$
2	$[15, 20[$	2	0.2	$20 - 15 = 5$	$\frac{0.2}{5} = 0.04$
3	$[20, 30]$	2	0.2	$30 - 20 = 10$	$\frac{0.2}{10} = 0.02$
		$n = 10$	1		

Como as **classes têm amplitudes diferentes**, então é a **área** de cada retângulo que pode ser representada pela frequência relativa.



Quando as **classes têm amplitudes diferentes**, então é a **área** de cada retângulo que pode ser representada pela frequência relativa.

Como área de um retângulo é:

$$\textit{área} = \textit{base} \times \textit{altura}$$

tem-se para as **frequências relativas**:

$$\textit{frequência relativa} = \textit{amplitude da classe} \times \textit{altura}$$

ou seja

$$\boxed{\textit{altura} = \frac{\textit{frequência relativa}}{\textit{amplitude da classe}}}$$

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados por variável. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados por variável;
- a representações gráficas dos dados por variável;
- ao cálculo de indicadores numéricos por variável, indicadores de localização e dispersão.

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados por variável. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados por variável;
- a representações gráficas dos dados por variável;
- ao cálculo de indicadores numéricos por variável, indicadores de localização e dispersão.

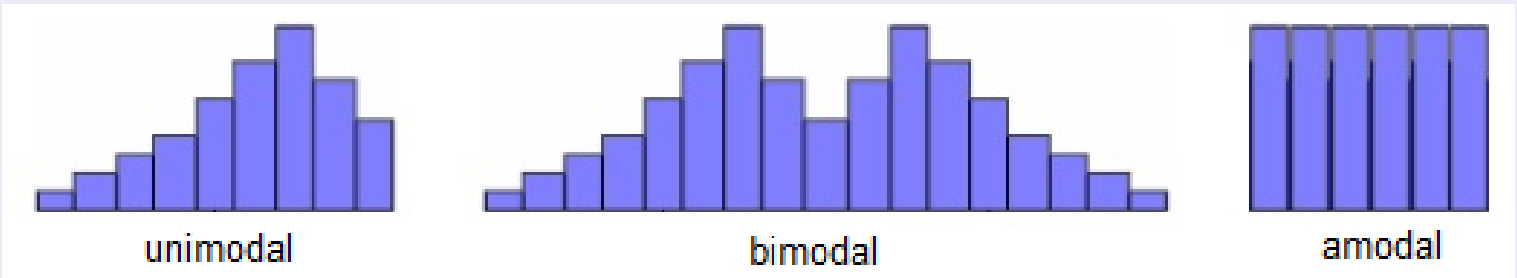
Medidas de Localização

Permitem resumir os dados calculando algumas características numéricas de modo a ter informação sobre a sua localização:

- Medidas de **localização central**:
  - moda,
  - média
  - mediana.
- Medidas de **localização não central**:
  - quantis.

Moda

- Habitualmente representa-se por mo.
- Para dados não agrupados, a moda define-se como o valor mais frequente.
- Para dados agrupados em classes (todas as classes com a mesma amplitude), a classe com maior frequência diz-se a **classe modal**.
- Um conjunto de dados pode não ter moda e diz-se **amodal**.
- Um conjunto de dados pode ter mais que uma moda. Isto acontece quando há dois ou mais valores que têm a maior frequência e diz-se
  - **bimodal** se tem duas modas;
  - **multimodal** ou **plurimodal** se tem mais do que duas modas.



Moda: Exemplos

Determine a moda.

1.

Naturalidade (x_i)	**Número de pessoas** (n_i)
Lisboa	72
Setúbal	42
Coimbra	31
Porto	50

Moda: Exemplos

Determine a moda.

1.

Naturalidade (x_i)	**Número de pessoas** (n_i)
Lisboa	72
Setúbal	42
Coimbra	31
Porto	50

A moda é Lisboa.

2.

14	14	12	13	15	15	16	11	10	9

Moda: Exemplos

Determine a moda.

1.

Naturalidade (x_i)	**Número de pessoas** (n_i)
Lisboa	72
Setúbal	42
Coimbra	31
Porto	50

A moda é Lisboa.

2.

14	14	12	13	15	15	16	11	10	9

Valores observados (x_i)	Frequência absoluta (n_i)
9	1
10	1
11	1
12	1
13	1
14	2
15	2
16	1

A moda é o 14 e o 15 $\rightarrow$ é bimodal

Moda: Exemplos

3. | 14 | 12 | 13 | 15 | 16 | 11 | 10 | 9 |
|---|---|---|---|---|---|---|---|

Moda: Exemplos

3. | 14 | 12 | 13 | 15 | 16 | 11 | 10 | 9 |
|---|---|---|---|---|---|---|---|

Valores observados (x_i)	Frequência absoluta (n_i)
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1

Não tem moda.
Não há nenhum valor que seja mais frequente.
↓
é amodal

4. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 13 |
|---|---|---|---|---|---|---|---|---|---|

Moda: Exemplos

3. | 14 | 12 | 13 | 15 | 16 | 11 | 10 | 9 |

Valores observados (x_i)	Frequência absoluta (n_i)
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1

Não tem moda.
Não há nenhum valor que seja mais frequente.
↓
é amodal

4. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 13 |

Valores observados (x_i)	Frequência absoluta (n_i)
10	1
11	1
12	1
13	2
14	2
15	2
16	1

A moda é o 13, o 14 e o 15
↓
é multimodal

Moda: Exemplos

5. Suponha que tem os seguintes dados

130.5;	130.5;	131.1;	132.0;	133.1;	134.0;	135.0;	136.2;	136.5;	137.0;
137.5	138.0;	138.5;	139.1;	139.9	140.0;	140.5	141.1;	141.5;	142.1;
142.5;	143.1;	143.5;	144.0;	144.5;	144.7;	145.0;	145.3;	145.5;	146.1;
146.3;	146.5;	146.7;	147.0;	147.5;	148.3;	148.5;	148.7;	149.0;	149.3;
150.0;	151.1;	152.3;	153.4;	154.5;	155.0;	156.1;	157.2;	158.4;	160.0

e que foram organizados na seguinte tabela de frequências:

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

Calcule a moda e a classe modal.

Moda: Exemplos

5. Suponha que tem os seguintes dados

130.5;	130.5;	131.1;	132.0;	133.1;	134.0;	135.0;	136.2;	136.5;	137.0;
137.5	138.0;	138.5;	139.1;	139.9	140.0;	140.5	141.1;	141.5;	142.1;
142.5;	143.1;	143.5;	144.0;	144.5;	144.7;	145.0;	145.3;	145.5;	146.1;
146.3;	146.5;	146.7;	147.0;	147.5;	148.3;	148.5;	148.7;	149.0;	149.3;
150.0;	151.1;	152.3;	153.4;	154.5;	155.0;	156.1;	157.2;	158.4;	160.0

e que foram organizados na seguinte tabela de frequências:

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

Calcule a moda e a classe modal.

A moda é 130.5 e a classe modal é $]145, 150]$.

Observação: A moda dos dados (sem estarem agrupados em classes) pode não pertencer à classe modal. É muito usual os dados serem amodais mas existir uma classe modal.

Média

- Representa-se por $\overline{x}$ (quando os dados correspondem a uma amostra) ou por μ (quando os dados correspondem à população).
- A média é a medida de localização central mais utilizada, sendo muitas vezes usada como valor "representativo" de um conjunto de dados.
- A **média** define-se como o quociente entre a soma de todos os valores observados e o número de elementos da amostra.

  Isto é, seja $\{x_1, x_2, \ldots, x_n\}$ um conjunto de dados com n observações, define-se média aritmética, ou simplesmente **média**, como

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

- Não existe média quando a variável é qualitativa.

Média

- A média de uma amostra apenas dá uma ideia da ordem de grandeza dos elementos da população, pois apenas é calculada com base nos elementos que foram incluídos na amostra.
- A média é muito sensível a valores extremos (muito grandes ou muito pequenos) dizendo-se por isso que é uma medida pouco resistente. Em alguns casos, a média pode não ser "representativa" de um conjunto de dados.

Média: Exemplos

Determine a média.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

Média: Exemplos

Determine a média.

1.

14	14	12	13	15	15	16	11	10	9

$$\overline{x} = \frac{14 + 14 + 12 + 13 + 15 + 15 + 16 + 11 + 10 + 9}{10} = 12.9$$

2.

Valores observados (x_i)	Frequência absoluta (n_i)	Frequência relativa (f_i)
45	3	0.091
47	10	0.303
50	7	0.212
53	10	0.303
54	3	0.091
Total	33	1

Média: Exemplos

Determine a média.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

$$\overline{x} = \frac{14 + 14 + 12 + 13 + 15 + 15 + 16 + 11 + 10 + 9}{10} = 12.9$$

2.

Valores observados (x_i)	Frequência absoluta (n_i)	Frequência relativa (f_i)
45	3	0.091
47	10	0.303
50	7	0.212
53	10	0.303
54	3	0.091
Total	33	1

$$\overline{x} = \frac{3 \times 45 + 10 \times 47 + 7 \times 50 + 10 \times 53 + 3 \times 54}{33} = 49.909$$

Quantis

- Representa-se por Q_p.
- Há vários métodos para calcular os quantis, nem todos conducentes aos mesmos valores, mas a valores próximos.
- Dado um número $0 \leq p \leq 1$, define-se **quantil de ordem** p, Q_p, como o valor contido no intervalo de variação das observações tal que, pelo menos $p \times 100\%$ das observações são inferiores ou iguais a esse valor e pelo menos $(1-p) \times 100\%$ das observações são maiores ou iguais a esse valor.

Quantis

- Alguns quantis são muito usados e têm nomes específicos:
  - **Quartis** - dividem a amostra em 4 partes iguais
    - 1^oquartil $= \mathrm{Q}_1 = Q_{0.25}$
    - 2^oquartil $= \mathrm{Q}_2 = Q_{0.50} =$ mediana
    - 3^oquartil $= \mathrm{Q}_3 = Q_{0.75}$
  - **Decis** - dividem a amostra em 10 partes iguais
    - 1^odecil $= D_1 = Q_{0.10}$
    - 2^odecil $= D_2 = Q_{0.20}$
    - $\ldots$
    - 8^odecil $= D_8 = Q_{0.80}$
    - 9^odecil $= D_9 = Q_{0.90}$
  - **Percentis** - dividem a amostra em 100 partes iguais
    - 1^opercentil $= P_1 = Q_{0.01}$
    - 2^opercentil $= P_2 = Q_{0.02}$
    - $\ldots$
    - 98^opercentil $= P_{98} = Q_{0.98}$
    - 99^opercentil $= P_{99} = Q_{0.99}$

Quantis

Mediana

- Um dos quantis mais importantes e mais utilizado em estatística é

$$2^o\text{quartil} = \text{Q}_2 = Q_{0.50} = \text{mediana}$$

- Habitualmente representa-se por $\widetilde{x}$ ou me.
- **A mediana** é o valor que ocupa a posição central quando se ordenam os dados estatísticos. Isto é, a mediana é o valor que separa as 50% das observações inferiores das 50% superiores. Por este motivo a mediana é considerada uma **medida de localização central**.
- A mediana é determinada pelo número de observações e não pelos seus valores, não sendo afetada por valores extremos. Diz-se, por isso, que é mais resistente do que a média.

Quando nos referimos aos quantis no geral, diz-se que são medidas de **localização não central** (a única exceção é a mediana que é uma medida de localização central).

Quantis

- Para determinar os quantis é necessário ordenar por ordem crescente as observações, pelo que não existem quantis quando a variável é qualitativa. No entanto há quem considere que é possível calcular quantis no caso da variável ser qualitativa ordinal. Aqui só vamos calcular os quantis para dados quantitativos.

- Os quantis são determinados pelo número de observações e não pelos seus valores, não sendo afetados por valores extremos.

- No caso dos **dados organizados numa tabela de frequências**, os quantis podem ser determinados a partir dos valores da **frequência relativa acumulada**.

Quantis

Dados em Tabelas de Frequências

No caso dos **dados organizados numa tabela de frequências**, os quantis podem ser determinados a partir dos valores da frequência relativa acumulada.

Dados que não estão agrupados em classes

- Se existir um valor com frequência relativa acumulada igual a p, o quantil é a média aritmética entre esse valor e o seguinte.
- Se não existir nenhum valor com frequência relativa acumulada igual a p, o quantil é o primeiro valor cuja frequência relativa acumulada ultrapassa p.

$$Q_p = \begin{cases} \frac{x_i + x_{i+1}}{2} & , \text{ para o valor } i, \text{ tal que } F_i = p \\ \\ x_i & , \text{ para o menor valor } i, \text{ tal que } F_i > p \end{cases}$$

Dados que estão agrupados em classes

- A primeira classe cuja a frequência relativa acumulada seja maior ou igual a p diz-se a **classe do quantil de ordem** p.

Quantis

Mediana: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} =$$

Quantis

Mediana: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} = 13$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.50
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} =$$

Quantis

Mediana: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} = 13$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.50
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} = \frac{12+13}{2} = 12.5$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} =$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$Q_1 = Q_{0.25} = 12$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$Q_1 = Q_{0.25} =$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} = 12$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_1 = Q_{0.25} = \frac{11+12}{2} = 11.5$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_3 = Q_{0.75} =$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} = 12$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_1 = Q_{0.25} = \frac{11+12}{2} = 11.5$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_3 = Q_{0.75} = 13$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_3 = Q_{0.75} =$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} = 12$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_1 = Q_{0.25} = \frac{11+12}{2} = 11.5$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_3 = Q_{0.75} = 13$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_3 = Q_{0.75} = \frac{12+13}{2} = 12.5$$

Quantis

Exemplo

Calcule as classes do 1^o Quartil, da mediana e do 3^o Quartil.

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)	Frequência Relativa Acumulada (F_i)
$[130, 135]$	7	0.14	0.14
$]135, 140]$	9	0.18	0.32
$]140, 145]$	11	0.22	0.54
$]145, 150]$	14	0.28	0.82
$]150, 155]$	5	0.10	0.92
$]155, 160]$	4	0.08	1
Total	50	1	

Quantis

Exemplo (1^o quartil)

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)	Frequência Relativa Acumulada (F_i)	Representante da classe $\left(x_i'\right)$
$[130, 135]$	7	0.14	0.14	$\frac{130+135}{2} = 132.5$
$]135, 140]$	9	0.18	0.32	$\frac{135+140}{2} = 137.5$
$]140, 145]$	11	0.22	0.54	$\frac{140+145}{2} = 142.5$
$]145, 150]$	14	0.28	0.82	$\frac{145+150}{2} = 147.5$
$]150, 155]$	5	0.10	0.92	$\frac{150+155}{2} = 152.5$
$]155, 160]$	4	0.08	1	$\frac{155+160}{2} = 157.5$
Total	50	1		

- $\mathrm{Q}_1 = Q_{0.25}$
- A **classe do 1º Quartil** é $]135, 140]$

  pois é a primeira classe cuja frequência relativa acumulada ultrapassa o valor de $p = 0.25$ $(0.32 > 0.25)$.

Quantis

Exemplo (mediana)

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)	Frequência Relativa Acumulada (F_i)	Representante da classe $\left(x_i'\right)$
$[130, 135]$	7	0.14	0.14	$\frac{130+135}{2} = 132.5$
$]135, 140]$	9	0.18	0.32	$\frac{135+140}{2} = 137.5$
$]140, 145]$	11	0.22	0.54	$\frac{140+145}{2} = 142.5$
$]145, 150]$	14	0.28	0.82	$\frac{145+150}{2} = 147.5$
$]150, 155]$	5	0.10	0.92	$\frac{150+155}{2} = 152.5$
$]155, 160]$	4	0.08	1	$\frac{155+160}{2} = 157.5$
Total	50	1		

- $\widetilde{x} = \mathrm{Q}_2 = Q_{0.50}$
- A **classe da mediana** é $]140, 145]$

  pois é a primeira classe cuja frequência relativa acumulada ultrapassa o valor de $p = 0.50$ $(0.54 > 0.50)$.

Quantis

Exemplo (3^o quartil)

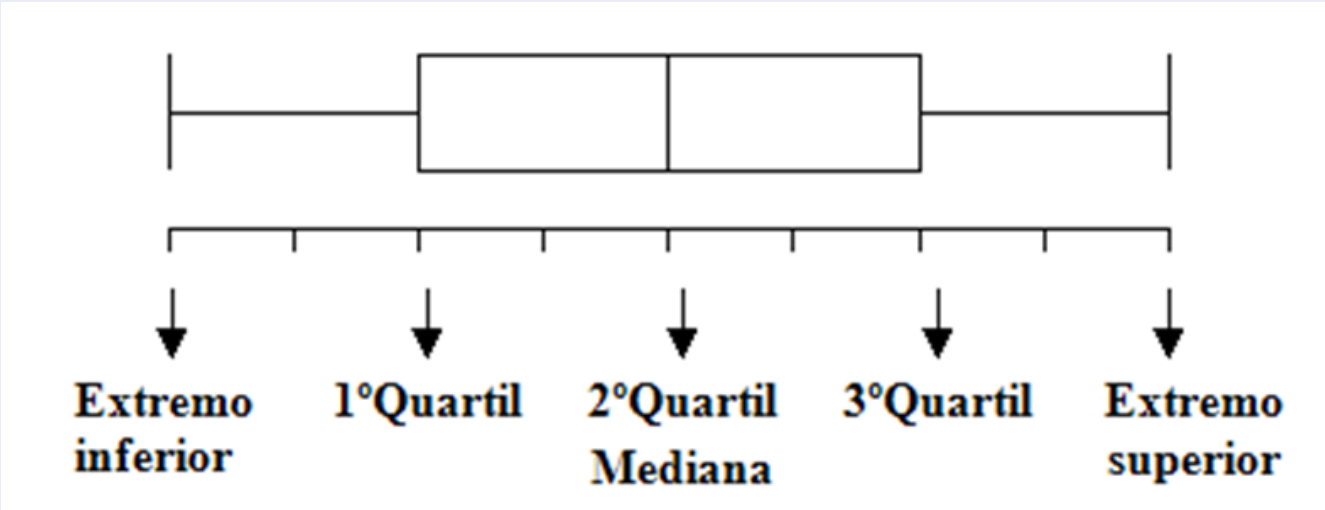
Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)	Frequência Relativa Acumulada (F_i)	Representante da classe $\left(x_i'\right)$
$[130, 135]$	7	0.14	0.14	$\frac{130+135}{2} = 132.5$
$]135, 140]$	9	0.18	0.32	$\frac{135+140}{2} = 137.5$
$]140, 145]$	11	0.22	0.54	$\frac{140+145}{2} = 142.5$
$]145, 150]$	14	0.28	0.82	$\frac{145+150}{2} = 147.5$
$]150, 155]$	5	0.10	0.92	$\frac{150+155}{2} = 152.5$
$]155, 160]$	4	0.08	1	$\frac{155+160}{2} = 157.5$
Total	50	1		

- $\mathrm{Q}_3 = Q_{0.75}$
- A **classe do 3º Quartil** é $]145, 150]$

  pois é a primeira classe cuja frequência relativa acumulada ultrapassa o valor de $p = 0.75$ $(0.82 > 0.75)$.

Diagrama de extremos e quartis

- O diagrama de extremos e quartis (BoxPlot) é uma forma esquemática de representar uma distribuição por cinco dos seus valores estatísticos: extremo inferior (mínimo), 1^o quartil, mediana ou 2^o quartil, 3^o quartil e extremo superior (máximo).



- Ficam definidas quatro zonas: duas centrais representadas por retângulos e duas caudas. Em cada uma destas zonas está 25% dos dados.
- Quanto mais estreita for uma zona, maior é a concentração de dados aí existente. Por isso, este diagrama dá algumas indicações gerais sobre o tipo de distribuição.

Diagrama de extremos e quartis: Exemplo

Construa o diagrama de extremos e quartis:

40	53	60	72	65	54	60	92	48	87

Diagrama de extremos e quartis: Exemplo

Construa o diagrama de extremos e quartis:

40	53	60	72	65	54	60	92	48	87

- Ordenar os dados:

40	48	53	54	60	60	65	72	87	92

- extremo inferior = mínimo dos dados $= 40$
- extremo superior = máximo dos dados $= 92$
- $Q_1 = Q_{0.25} = 53$
- $Q_2 = Q_{0.50} = \widetilde{x} = 60$
- $Q_3 = Q_{0.75} = 72$

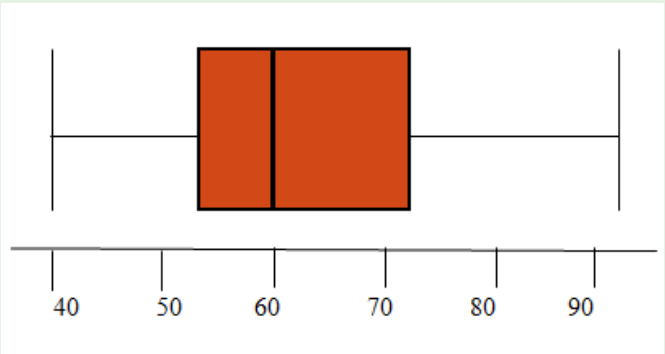
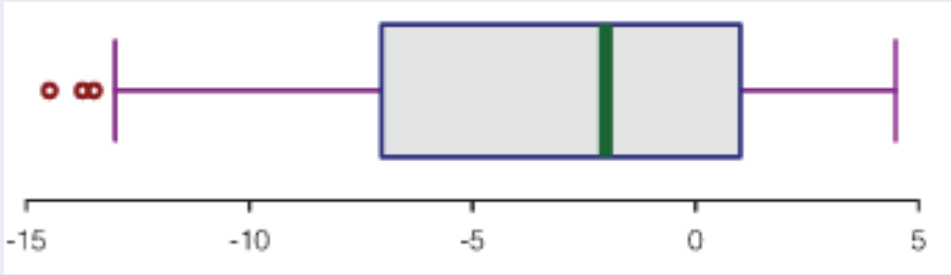


Diagrama de extremos e quartis

- No diagrama de extremos e quartis podemos identificar as observações que se afastam do padrão geral dos dados, os chamados "outliers" (observações discordantes) e representam-se por $*$ ou $\circ$.



- Existem vários critérios para classificar uma observação como um "outlier" :
  - Um valor x_i é um candidato a **"outlier" moderado** (habitualmente representa-se por $\circ$) se estiver
    - entre $Q_1 - 1.5 \times (Q_3 - Q_1)$ e $Q_1 - 3 \times (Q_3 - Q_1)$
      ou
    - entre $Q_3 + 1.5 \times (Q_3 - Q_1)$ e $Q_3 + 3 \times (Q_3 - Q_1)$.
  - Um valor x_i é um candidato a **"outlier" severo** (habitualmente representa-se por $*$) se
    - for maior que $Q_3 + 3 \times (Q_3 - Q_1)$
      ou
    - menor que $Q_1 - 3 \times (Q_3 - Q_1)$.

Diagrama de extremos e quartis: Exemplo

Construa o diagrama de extremos e quartis representando os "outliers" (caso existam).

22	14	23	6	20	21	55	22	25

Diagrama de extremos e quartis: Exemplo

Construa o diagrama de extremos e quartis representando os "outliers" (caso existam).

22	14	23	6	20	21	55	22	25

- Ordenar os dados:

6	14	20	21	22	22	23	25	55

- extremo inferior = mínimo dos dados $= 6$
- extremo superior = máximo dos dados $= 55$
- $Q_1 = Q_{0.25} = 20$
- $Q_2 = Q_{0.50} = \widetilde{x} = 22$
- $Q_3 = Q_{0.75} = 23$

Diagrama de extremos e quartis: Exemplo

Dados ordenados:

6	14	20	21	22	22	23	25	55

- $Q_1 = 20$; $\widetilde{x} = 22$; $Q_3 = 23$
- limites dos "outliers" moderados:
  - $Q_1 - 1.5 \times (Q_3 - Q_1) = 20 - 1.5 \times (23 - 20) = 15.5$
  - $Q_3 + 1.5 \times (Q_3 - Q_1) = 23 + 1.5 \times (23 - 20) = 27.5$
- limites dos "outliers" severos:
  - $Q_1 - 3 \times (Q_3 - Q_1) = 20 - 3 \times (23 - 20) = 11$
  - $Q_3 + 3 \times (Q_3 - Q_1) = 23 + 3 \times (23 - 20) = 32$
- "outliers" moderados: 14
- "outliers" severos: 6 e 55
- a caixa só é construída com os dados: 20, 21, 22, 22, 23, 25
  - extremo inferior $= 20$
  - extremo superior $= 25$

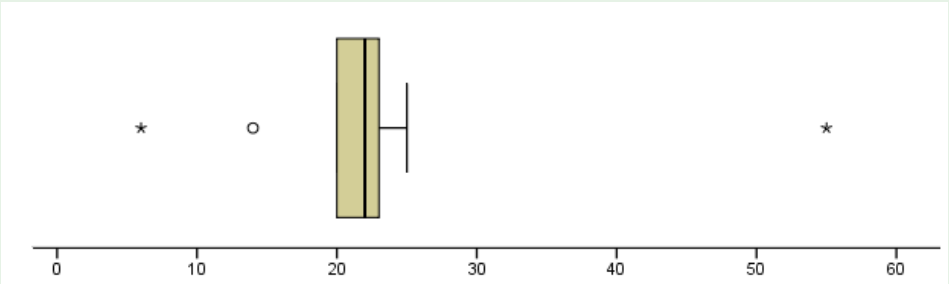
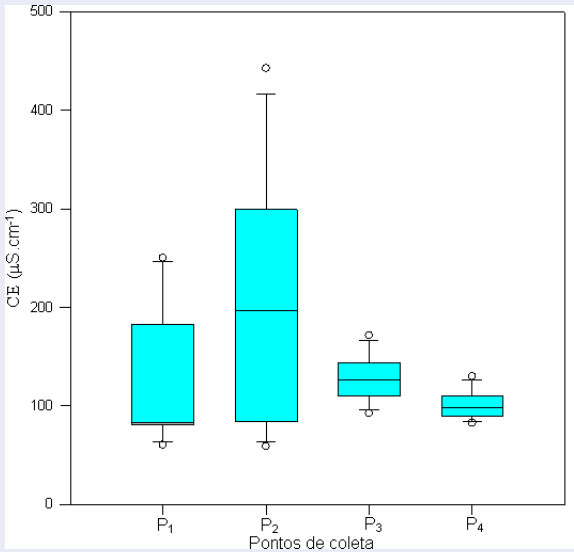
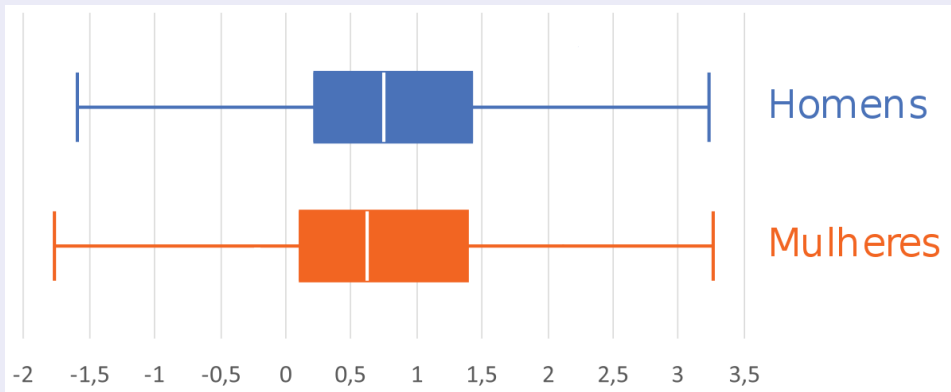


Diagrama de extremos e quartis

- Quando se pretende comparar várias amostras, o recurso a este tipo de diagramas, dispostos paralelamente, é uma ferramenta que permite, de forma fácil, obter uma primeira interpretação e comparação dos conjuntos de dados.





Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados por variável. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados por variável;
- a representações gráficas dos dados por variável;
- ao cálculo de indicadores numéricos por variável, indicadores de localização e dispersão.

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados por variável. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados por variável;
- a representações gráficas dos dados por variável;
- ao cálculo de indicadores numéricos por variável, indicadores de localização e dispersão.

Medidas de Dispersão

Permitem resumir os dados calculando algumas características numéricas de modo a ter informação sobre a sua a variabilidade ou dispersão:

- Medidas de **dispersão absoluta** (depende da unidade em que se exprime a variável):
  - amplitude: amplitude total e amplitude interquartis,
  - variância e desvio padrão.
- Medidas de **dispersão relativa** (não depende da unidade em que se exprime a variável):
  - coeficiente de variação.

Amplitude Total

- Habitualmente representa-se por **A**.
- A **Amplitude Total** é a medida mais simples para medir a variabilidade dos dados.

- Para dados não agrupados, a amplitude total define-se como a diferença entre o maior e o menor valor do conjunto de dados (diferença entre os extremos). Isto é, seja $\{x_1, x_2, \ldots, x_n\}$ um conjunto de dados com n observações,

$$\mathrm{A} = max(x_i) - min(x_i).$$

- Para dados agrupados em classes, a amplitude total é a diferença entre o limite superior da última classe e o limite inferior da primeira classe.

- É uma medida não negativa e será tanto maior quanto maior for a variabilidade dos dados.

Amplitude Total

Exemplos

Determine a amplitude.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

Amplitude Total

Exemplos

Determine a amplitude.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

   - Ordenar os dados:

     | 9 | 10 | 11 | 12 | 13 | 14 | 14 | 15 | 15 | 16 |

   - mínimo $= 9$
     máximo $= 16$
   - A $= 16 - 9 = 7$

2.

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

Amplitude Total

Exemplos

Determine a amplitude.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

   - Ordenar os dados:

   | 9 | 10 | 11 | 12 | 13 | 14 | 14 | 15 | 15 | 16 |

   - mínimo $= 9$
     máximo $= 16$
   - A $= 16 - 9 = 7$

2.

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

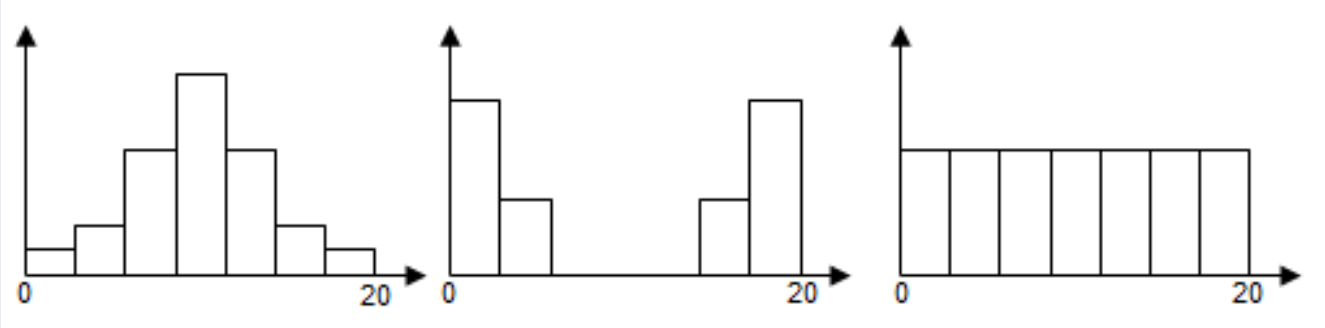
   - limite inferior da primeira classe $= 130$
     limite superior da última classe $= 160$
   - A $= 160 - 130 = 30$

Amplitude Total

Observações

- A amplitude é uma fraca medida de dispersão.

  Exemplo de distribuições com a mesma amplitude, mas com uma dispersão muito diferente:



- Desvantagem da amplitude enquanto medida de dispersão:
  - É insensível às alterações dos valores intermédios (nela só intervêm os extremos).
  - Não diz nada sobre o que se passa no intervalo entre os extremos. Em certas distribuições os valores extremos correspondem a casos excecionais e portanto pouco significativos.

Amplitude interquartis

- Habitualmente representa-se por AIQ.
- A amplitude interquartis define-se como a diferença entre o 3^o quartil e o 1^o quartil:

$$AIQ = \mathrm{Q}_3 - \mathrm{Q}_1 = Q_{0.75} - Q_{0.25}$$

- É uma medida não negativa e será tanto maior quanto maior for a variabilidade dos dados.
- Amplitude interquartis indica a amplitude do intervalo onde se situa a metade central dos dados, sendo pouco sensível aos valores extremos.
- Uma Amplitude Interquartis nula não significa que os dados não apresentem variabilidade.
- Desvantagem desta medida de dispersão:
  - É insensível às alterações dos valores que se encontram antes do 1^o quartil e depois do 3^o quartil.

Amplitude interquartis

Exemplo

Determine a amplitude interquartis.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

Amplitude interquartis

Exemplo

Determine a amplitude interquartis.

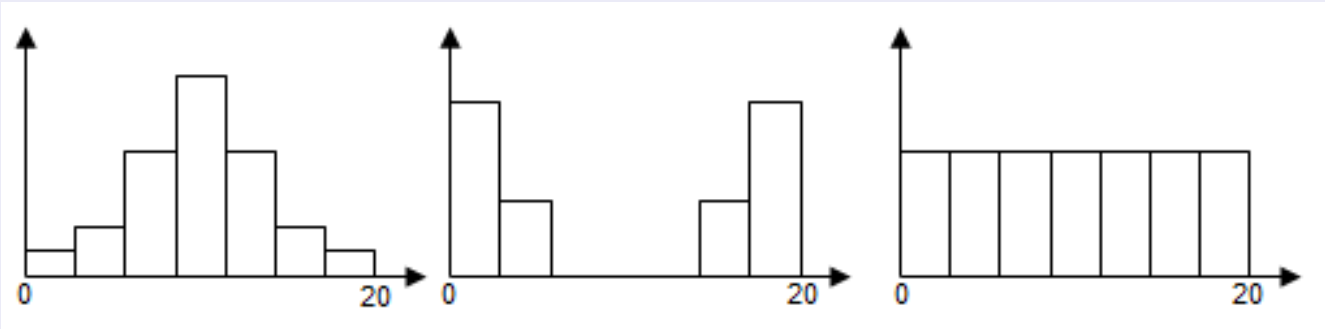
1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

   - $\mathrm{Q}_1 = Q_{0.25} = 11$
   - $\mathrm{Q}_3 = Q_{0.75} = 15$

2. $AIQ = 15 - 11 = 4$

Medidas de Dispersão

- Como vimos, embora a amplitude (total ou interquartil) seja uma possibilidade importante para analisar a variabilidade dos dados, tem limitações.
- Outra possibilidade para analisar a variabilidade dos dados consiste em comparar os dados com uma medida de localização central: a média.
- A dispersão dos dados em torno da sua média permite caracterizar um conjunto de dados, pois dados com a mesma média podem ter uma dispersão muito diferente:



- No entanto não é possível caracterizar a variabilidade somando os desvios em relação à média. A soma dos desvios é sempre zero.
- Deve-se considerar uma medida que não leve em conta o sinal dos desvios (o que importa é a magnitude do desvio). Assim, se considerarmos valor absoluto (módulo) dos desvios temos o **Desvio absoluto médio**, mas se considerarmos o quadrado dos desvios temos a **Variância**.

Variância

- Representa-se por s^2 (quando os dados correspondem a uma amostra) ou por σ^2 (quando os dados correspondem à população).
- A variância mede o afastamento dos dados em relação à média.
- A variância é a média dos quadrados dos desvios relativamente à média. Isto é, seja $\{x_1, x_2, \ldots, x_n\}$ um conjunto de dados com n observações, define-se variância como

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} =$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$$

Variância: Exemplos

Determine a variância.

(1)

14	14	12	13	15

Variância: Exemplos

Determine a variância.

(1)

14	14	12	13	15

- calcular a média:

$$\overline{x} = \frac{14 + 14 + 12 + 13 + 15}{5} = 13.6$$

- a variância é

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{(14 - 13.6)^2 + (14 - 13.6)^2 + (12 - 13.6)^2 + (13 - 13.6)^2 + (15 - 13.6)^2}{5 - 1} = 1.3$$

ou

- a variância é

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1} = \frac{\left(14^2 + 14^2 + 12^2 + 13^2 + 15^2\right) - 5 \times 13.6^2}{5 - 1} = 1.3$$

Variância: Exemplos

Determine a variância.

(2)

Valores observados (x_i)	Frequência absoluta (n_i)	Frequência relativa (f_i)
45	3	0.091
47	10	0.303
50	7	0.212
53	10	0.303
54	3	0.091
Total	33	1

Variância: Exemplos

Determine a variância.

(2)

Valores observados (x_i)	Frequência absoluta (n_i)	Frequência relativa (f_i)
45	3	0.091
47	10	0.303
50	7	0.212
53	10	0.303
54	3	0.091
Total	33	1

Calcular a variância considerando as frequências absolutas.

- calcular a média:

$$\overline{x} = \frac{45 \times 3 + 47 \times 10 + 50 \times 7 + 53 \times 10 + 54 \times 3}{33} = \frac{1647}{33} = 49.91$$

- a variância é

$$s^2 = \frac{(45 - 49.91)^2 \times 3 + (47 - 49.91)^2 \times 10 + (50 - 49.91)^2 \times 7 + (53 - 49.91)^2 \times 10 + (54 - 49.91)^2 \times 3}{33 - 1} =$$

$$= \frac{302.79}{32} = 9.46$$

Variância

Observações

- Note-se que a **Variância** envolve a soma de quadrados, e por isso a unidade medida em que se exprime não é a mesma que a dos dados, a **unidade de medida** fica **ao quadrado**.

- Vantagem da variância como medida de dispersão:
  - no seu cálculo entram todas as observações.

- Desvantagem da variância como medida de dispersão:
  - não é fácil de interpretar, uma vez que é expressa em unidades da variável ao quadrado;
  - facilmente assume valores muito elevados;
  - é uma medida pouco resistente a valores extremos (muito grandes ou muito pequenos).

Desvio Padrão

- Representa-se por s (quando os dados correspondem a uma amostra) ou por σ (quando os dados correspondem à população).
- O desvio padrão é a raiz quadrada da variância

$$s = \sqrt{s^2}$$

- O desvio padrão é sempre maior ou igual a zero.
- É a medida de dispersão mais utilizada uma vez que vem expressa na mesma unidade em que estão expressos os dados da amostra.
- O desvio padrão informa sobre o afastamento dos dados em relação à média. Quanto maior for o desvio padrão, maior é o afastamento dos dados em relação à média.
- O Desvio Padrão, assim como a média, é muito sensível a valores extremos, portanto é uma medida pouco resistente.

Desvio Padrão

Exemplo

Determine o desvio padrão (suponha que a unidade de medida dos dados é metros).

14	14	12	13	15

Desvio Padrão

Exemplo

Determine o desvio padrão (suponha que a unidade de medida dos dados é metros).

14	14	12	13	15

média:

$$\overline{x} = \frac{14 + 14 + 12 + 13 + 15}{5} = 13.6 \text{ metros}$$

variância:

$$s^2 = \frac{(14 - 13.6)^2 + (14 - 13.6)^2 + (12 - 13.6)^2 + (13 - 13.6)^2 + (15 - 13.6)^2}{5 - 1} = 1.3 \text{ metros}^2$$

desvio padrão:

$$s = \sqrt{s^2} = \sqrt{1.3} = 1.14 \text{ metros}$$

Coeficiente de Variação

- O desvio padrão por si só não traz muita informação. Ou seja, um desvio padrão de 2 unidades pode ser considerado pequeno para um conjunto de valores cuja média é 200, mas já pode ser considerado grande se a média for de 20.
- Como o desvio padrão vem na mesma unidade de medida dos dados, não se deve usar esta medida de dispersão para comparar conjuntos de dados com **unidades de medida diferentes** ou que **diferem consideravelmente em grandeza**. Neste caso deve-se recorrer ao **Coeficiente de Variação**.

Coeficiente de Variação

- O coeficiente de variação representa-se por CV.
- O coeficiente de variação é uma medida de dispersão relativa e corresponde ao quociente entre o desvio padrão (medida de dispersão) e a média (medida de localização):
  - quando os dados correspondem a uma **amostra**:
$$CV = \frac{s}{\overline{x}} \times 100\%$$
  - quando os dados correspondem à **população**:
$$CV = \frac{\sigma}{\mu} \times 100\%$$

Coeficiente de Variação

- O coeficiente de variação pode ser interpretado como a fração da dispersão pela qual a localização é responsável. Isto é, o coeficiente de variação indica a magnitude relativa do desvio padrão quando comparado com a média do conjunto de valores.
- Quanto maior for o coeficiente de variação, maior é a dispersão dos dados.
- O coeficiente de variação é independente da unidade de medida utilizada, sendo útil para comparar conjuntos de dados.
- Esta medida só deve ser usada quando a variável toma valores de um só sinal, isto é, todos os dados são positivos ou todos os dados são negativos.

Coeficiente de Variação: Exemplo 1

Na tabela seguinte são apresentados os resultados da altura e peso de um grupo de indivíduos:

	Altura	Peso
média $(\overline{x})$:	175 cm	68 kg
desvio padrão (s):	5 cm	2 kg

Qual dos conjuntos de dados apresenta maior dispersão, a altura ou o peso dos indivíduos?

Coeficiente de Variação: Exemplo 1

Na tabela seguinte são apresentados os resultados da altura e peso de um grupo de indivíduos:

	Altura	Peso
média $(\overline{x})$:	175 cm	68 kg
desvio padrão (s):	5 cm	2 kg

Qual dos conjuntos de dados apresenta maior dispersão, a altura ou o peso dos indivíduos?

As **unidades de medidas são diferentes**: a altura está em **centímetros** e o peso está em **quilos**. É necessário calcular o **coeficiente de variação**:

altura

$$CV = \frac{5}{175} \times 100\% = 2,86\%$$

peso

$$CV = \frac{2}{68} \times 100\% = 2,94\%$$

Conclui-se que neste grupo de indivíduos, os pesos apresentam maior grau de dispersão que as alturas.

Coeficiente de Variação: Exemplo 2

Considere os seguintes conjuntos de dados referentes aos preços (em euros) de frigoríficos e batedeiras em 7 lojas distintas:

Frigoríficos	750	800	790	810	820	760	780
Batedeiras	50	45	55	43	52	45	54

$\overline{x} = 787,14$	$s = 25,63$
$\overline{x} = 49.14$	$s = 4.81$

Qual dos produtos tem uma maior variabilidade de preços?

Coeficiente de Variação: Exemplo 2

Considere os seguintes conjuntos de dados referentes aos preços (em euros) de frigoríficos e batedeiras em 7 lojas distintas:

Frigoríficos	750	800	790	810	820	760	780
Batedeiras	50	45	55	43	52	45	54

$\overline{x} = 787,14$	$s = 25,63$
$\overline{x} = 49.14$	$s = 4.81$

Qual dos produtos tem uma maior variabilidade de preços?
As unidades de medidas são iguais mas **diferem consideravelmente em grandeza**.
É necessário calcular o **coeficiente de variação**:

Frigoríficos

$$CV = \frac{25.63}{787.14} \times 100\% = 3.3\%$$

Batedeiras

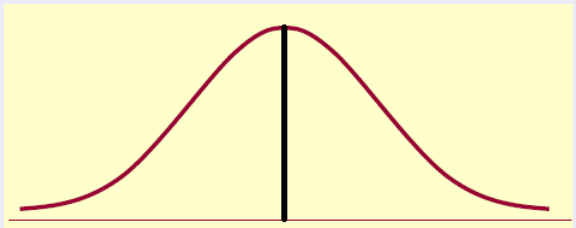
$$CV = \frac{4.81}{49.14} \times 100\% = 9.8\%$$

Conclui-se que neste conjunto de dados, os preços das batedeiras têm uma maior variabilidade do que os preços dos frigoríficos.

Caracterização da Distribuição de Frequências

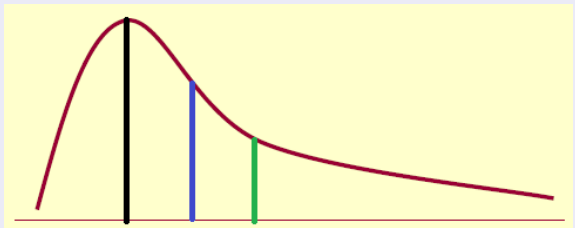
A posição relativa das medidas de localização média, mediana e moda possibilitam classificar as distribuições dos dados como: **Simétricas** ou **Assimétricas**.

- Se a distribuição dos dados for aproximadamente simétrica, a **média** aproxima-se da **mediana** e da **moda**.



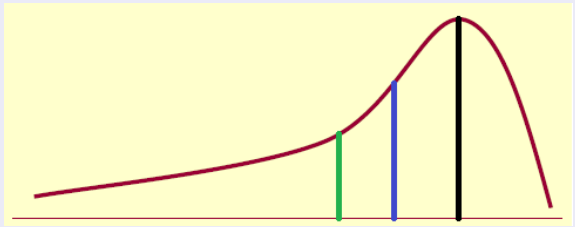
média = mediana = moda

- Se a distribuição dos dados for assimétrica positiva (ou enviesada para a direita), a média tende a ser maior que a mediana e que a **moda**.



moda < mediana < média

- Se a distribuição dos dados for assimétrica negativa (ou enviesada para a esquerda), a média tende a ser inferior à mediana e à **moda**.

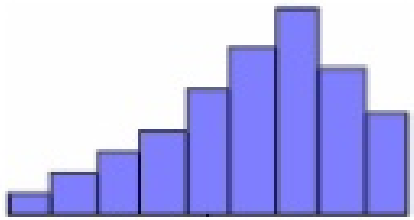


média < mediana < moda

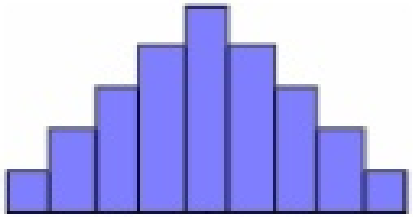
Caracterização da Distribuição de Frequências

Esta caracterização da distribuição de frequências em **Simétrica** ou **Assimétrica** também pode ser observada graficamente:

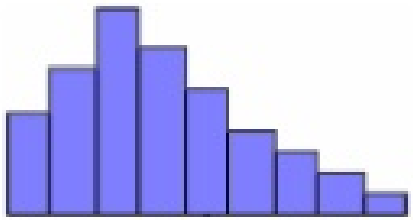
- através do Histograma:



Assimétrica negativa



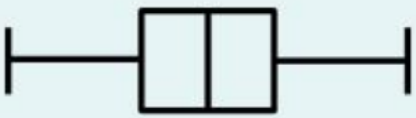
Simétrica



Assimétrica positiva

- através do Diagrama de extremos e quartis:

Assimétrico negativo



Simétrico

Assimétrico positivo

Caracterização da Distribuição de Frequências

- As medidas de localização e os gráficos, embora forneçam informação importante, são insuficientes para uma boa caracterização da distribuição de frequências em termos de assimetria.

- Para caracterizar adequadamente a distribuição de frequências é preciso estudar a sua forma, analisando o seu grau de assimetria com recurso às **Medidas de Assimetria**

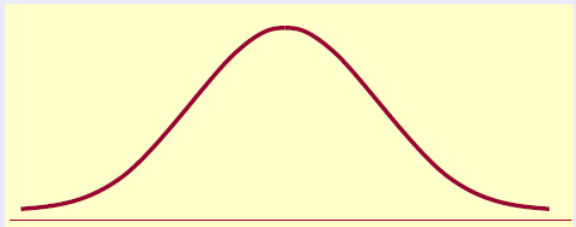
Medidas de Assimetria

Existem diversas medidas de assimetria, o coeficiente b_1 é um dos mais utilizados para avaliar a assimetria:

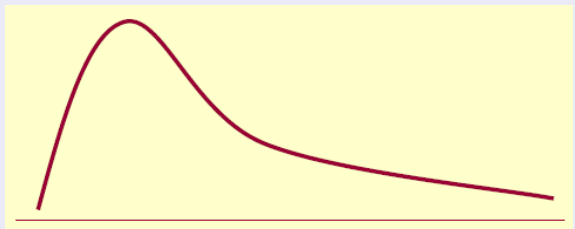
$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

onde

- $b_1 = 0 \rightarrow$ Simétrica



- $b_1 > 0 \rightarrow$ Assimétrica positiva (ou enviesada para a direita)



- $b_1 < 0 \rightarrow$ Assimétrica negativa (ou enviesadas para a esquerda)