

Preparado para:



REFORM/SC2022/126

DELIVERABLE 4

MÓDULO 3

ESTATÍSTICA BÁSICA

EM R

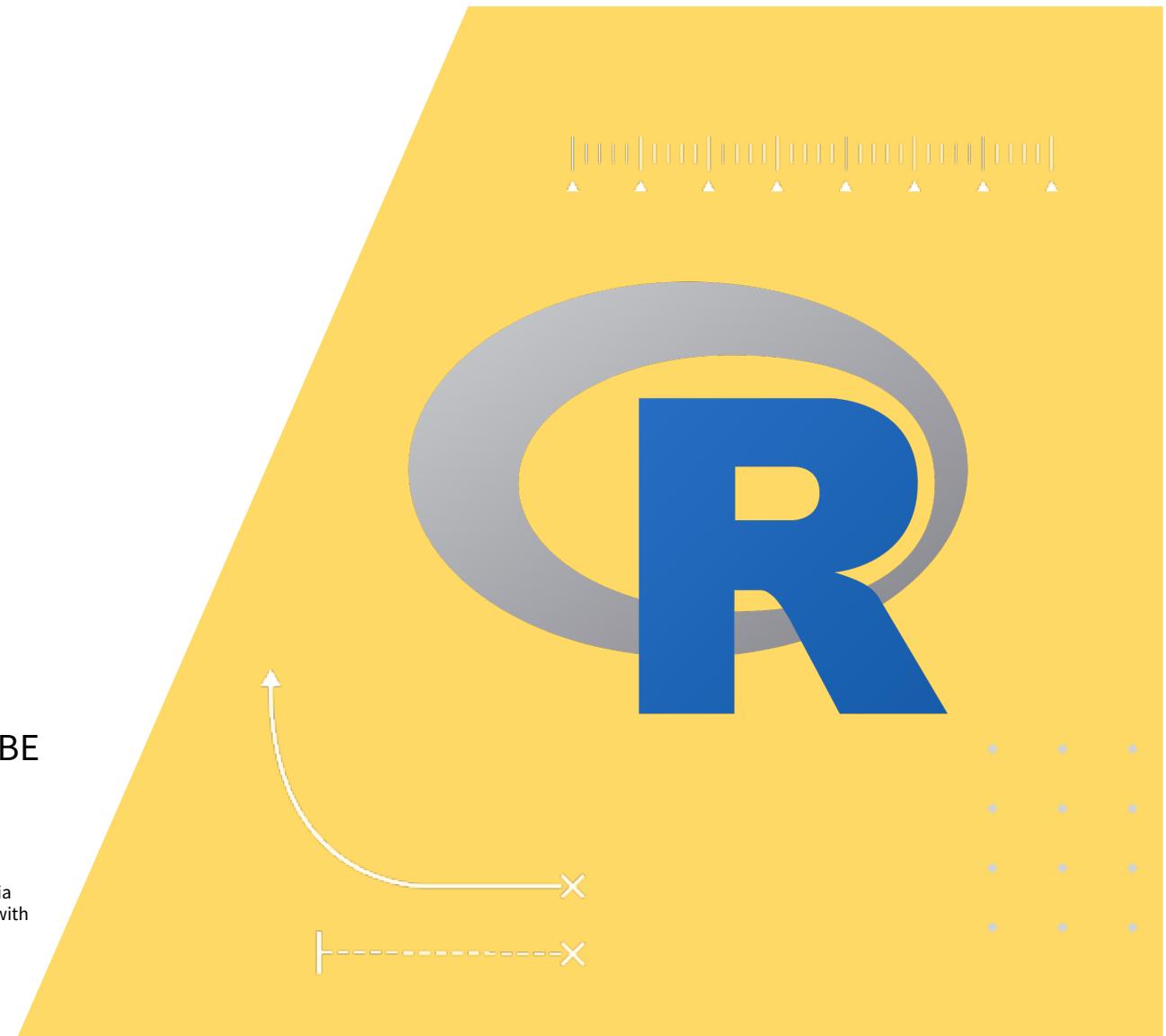
DESIGNING A NEW VALUATION MODEL
FOR RURAL PROPERTIES IN PORTUGAL

Parte I

Formador: Luís Teles Morais | Nova SBE
Lisboa, 15 junho 2023



This project is carried out with funding by the European Union via the Structural Reform Support Programme and in cooperation with the Directorate General for Structural Reform Support of the European Commission



Programa

MÓDULOS	DURAÇÃO
Módulo 1 – Introdução ao R: <ul style="list-style-type: none">- O que é o R?- Como instalar e configurar o R.- Sintaxe básica e comandos.- Tipos de dados, objetos e classes.	4 Horas
Módulo 2 – Gestão e tratamento de dados em R: <ul style="list-style-type: none">- Carregar dados no R.- Perceber as estruturas de dados e <i>subsetting</i>.- Limpeza de dados: <i>missing values</i>, <i>outliers</i> e transformações- Juntar bases de dados	8 Horas
Módulo 3 – Estatística básica em R: <ul style="list-style-type: none">- Estatísticas descritivas: medidas de dispersão central e variação.- Distribuições probabilísticas: variáveis discretas e contínuas.- Testes de hipóteses.	8 Horas

MÓDULOS	DURAÇÃO
Módulo 4 – Regressão Linear: <ul style="list-style-type: none">- O modelo classico linear.- Estimação de parametros segundo o MMQ.- Testes de hipóteses: significância estatística e ajuste do modelo.- Modelo de regressão múltipla.- Testar as premissas: multicolinearidade, heteroscedasticidade e normalidade dos resíduos.- Critérios de seleção dos modelos.	12 Horas
Módulo 5 – O modelo: <ul style="list-style-type: none">- Estrutura do modelo e premissas – Perceber o modelo (4 Hours).- Uso e tratamento dos dados (4 Hours).- Descrição do modelo (4 Hours).- Aplicação do modelo a cada piloto (12 Hours).- Aplicação autónoma do modelo a uma região (8 Hours).	32 Horas

Vamos a isso

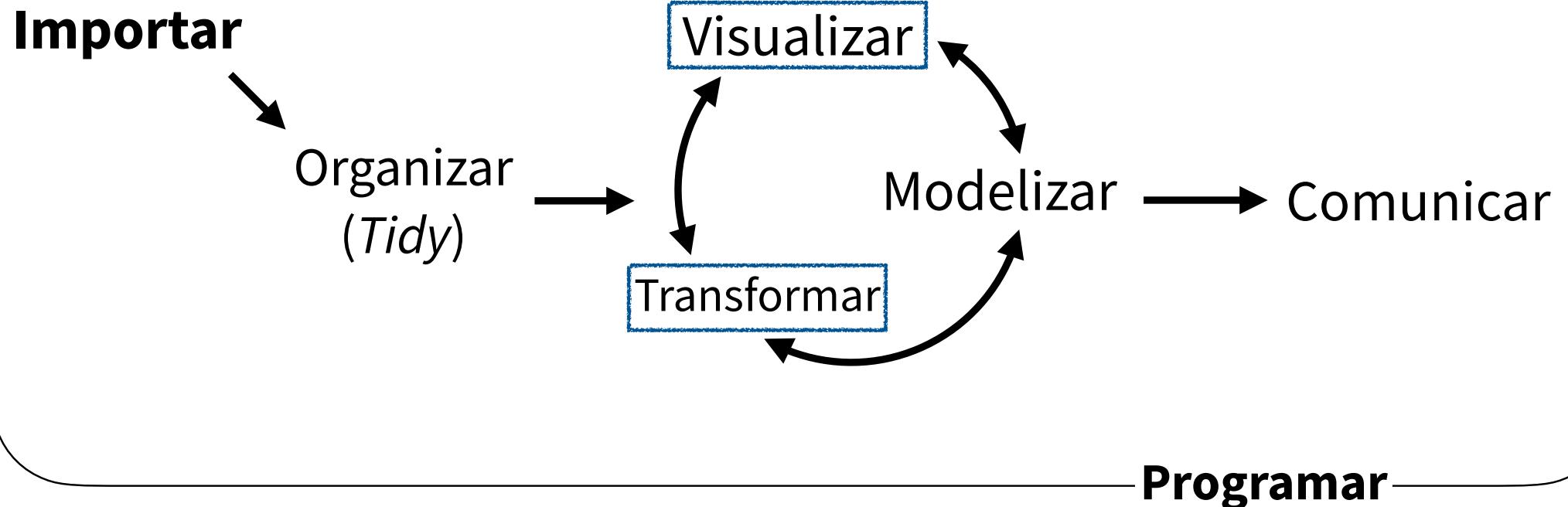
Aceda a este link para começar já

<https://posit.cloud/content/5906356>

Estatística básica em R (I)

Estatística descritiva

Ciência de dados



Alguns conceitos

Estatística **descritiva**

- Analisar um conjunto de dados, reduzindo-o a medidas sumárias simples

Visualizar

Inferência estatística

- Calcular ou estimar algo que não podemos observar diretamente, a partir dos dados existentes

Modelizar

Alguns conceitos

População

- O objeto de estudo

Amostra

- O conjunto de dados disponíveis sobre o objeto de estudo

As propriedades rurais em Portugal

Propriedades transacionadas entre 1990 e 2015
 $\{P_i\}$
 $i = 1, 2, 3, \dots, N$

Alguns conceitos

Descrever a amostra...

- Estatística descritiva
- e.g.: P médio = 25.000€

As propriedades rurais em Portugal

Propriedades transacionadas entre 1990 e 2015
 $\{P_i\}$
 $i = 1, 2, 3, \dots, N$

Alguns conceitos

Para **inferir** algo sobre a população

- Inferência estatística
- Hip.: medidas sumárias (momentos) da amostra são boas estimativas para a população [+ outras]
- e.g.: o P médio em Portugal é cerca de 25.000
 - e com 95% confiança está entre 24.000 e 26.000



As propriedades rurais em Portugal

Propriedades transacionadas entre 1990 e 2015
 $\{P_i\}$
 $i = 1, 2, 3, \dots, N$

Alguns conceitos

Parâmetro

- Quantidade teórica
- Valor exacto
- e.g. **Pm** preço médio das propriedades em Portugal

Medida

- Medida amostral, e.g. de tendência central
- Pode ser calculada a partir dos dados
- e.g. média amostral

As propriedades rurais em Portugal

Propriedades transacionadas entre 1990 e 2015
 $\{P_i\}$
 $i = 1, 2, 3, \dots, N$

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

Tipos de dados e implementação em R

Dados categóricos ou qualitativos: sem relação numérica

- **Escala nominal:** sem nenhuma ordem em particular
 - e.g. tipo de cultura: trigo, milho, batata

character ou “string” (*cultura*: “trigo”, “milho”,...)

- Quando só há 2 categorias: **variável binária**
 - e.g. prédio rústico ou urbano

logical (*urbano*: TRUE ou FALSE)

Tipos de dados e implementação em R

Dados numéricos ou quantitativos: contagem ou métrica

- **Dados discretos:** toma apenas valores dentro de um certo conjunto
 - normalmente, **inteiros**, e.g. contagem de algo
 - e.g. n.^o de árvores

integer (1L, 55L, 100L) - não tem grande vantagem

- **Dados contínuos:** quantidades que podem assumir qualquer valor numérico (mesmo que dentro de um certo intervalo)
 - e.g. área total

numeric (10.5, 55, 787)

Medidas de tendência central/posição

Média

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

- Serve para quaisquer dados numéricos (discretos ou contínuos)
- Medida de posição, mas sensível a *outliers*

`mean(<...>, na.rm = FALSE)`

Medidas de tendência central/posição

Mediana

Valor a que 50% das observações são inferiores

- Serve para quaisquer dados numéricos (discretos ou contínuos)
- Ordena-se os dados e seleciona-se o valor “do meio” (ou o ponto médio)
- Medida de posição, insensível a outliers

`median(<...>, na.rm = FALSE)`

Medidas de tendência central/posição

Moda

Valor ou categoria observada com maior frequência

- Útil especialmente em dados qualitativos, onde não se pode calcular a média ou mediana
- Medida de posição, insensível a outliers

?

slice() e slice_max()

babynames %>%

year	sex	name	n	prop
1880	M	John	9655	0.0815
1880	M	William	9532	0.0805
1880	M	James	5927	0.0501
1880	M	Charles	5348	0.0451
1880	M	Garrett	13	0.0001
1881	M	John	8769	0.081
1881	M	William	8524	0.0787
1881	M	James	5442	0.0503
1881	M	Charles	4664	0.0431



slice(1:3)

year	sex	name	n	prop
1880	M	John	9655	0.0815
1880	M	William	9532	0.0805
1880	M	James	5927	0.0501



slice_max(n)

year	sex	name	n	prop
1880	M	John	9655	0.0815

Ex.: qual a moda dos nomes, considerando todos os anos?

```
babynames %>% group_by(name) %>%
  summarize(total = sum(n)) %>% slice_max(total)

# A tibble: 1 × 2
  name    total
  <chr>   <dbl>
1 James  5173828
```

```
##   name    height  mass hair_color skin_color eye_color birth_year
##   <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl>
## 1 Luke S...     172     77 blond      fair       blue           19
## 2 C-3PO        167     75 <NA>       gold       yellow         112
## 3 R2-D2         96     32 <NA>      white, bl... red            33
## 4 Darth ...     202    136 none       white       yellow        41.9
## 5 Leia O...     150     49 brown      light       brown           19
## 6 Owen L...     178    120 brown, gr... light       blue           52
## # ... with 81 more rows, and 7 more variables: sex <chr>,
## #   gender <chr>, homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

starwars



Experimente

Considere apenas personagens das espécies (**species**) “Human” e “Droid”.

1. Prepare uma tabela com base nos dados starwars, que apresente em função da espécie (**species**):

- Média da altura (**height**)
- Mediana da altura
- N.º de observações

2. Prepare outra tabela que apresente em função da espécie:

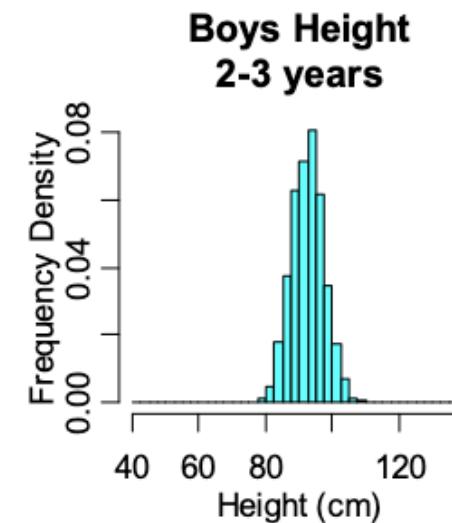
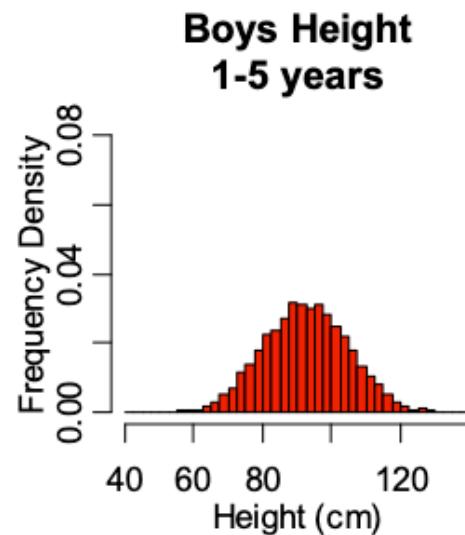
- Moda da cor dos olhos (**eye_color**)

```
starwars %>% filter(species %in% c('Human', 'Droid')) %>%  
  group_by(species) %>% summarize(media_alt = mean(height, na.rm = T),  
                                     mediana_alt = median(height, na.rm = T),  
                                     obs = n() )  
  
# A tibble: 2 × 4  
species media_alt mediana_alt   obs  
<chr>      <dbl>        <int> <int>  
1 Droid       131.         97     6  
2 Human       177.        180    35  
  
starwars %>% filter(species %in% c('Human', 'Droid')) %>%  
  group_by(species, eye_color) %>% summarize(total = n()) %>% slice_max(total)  
  
# A tibble: 2 × 3  
# Groups:   species [2]  
species eye_color total  
<chr>  <chr>     <int>  
1 Droid   red        3  
2 Human  brown      17
```

Medidas de dispersão

- Indicam a variabilidade de um determinado conjunto de dados
- Note que dois conjuntos de dados podem ter a mesma media e mediana mas um aspecto muito diferente (e significado):

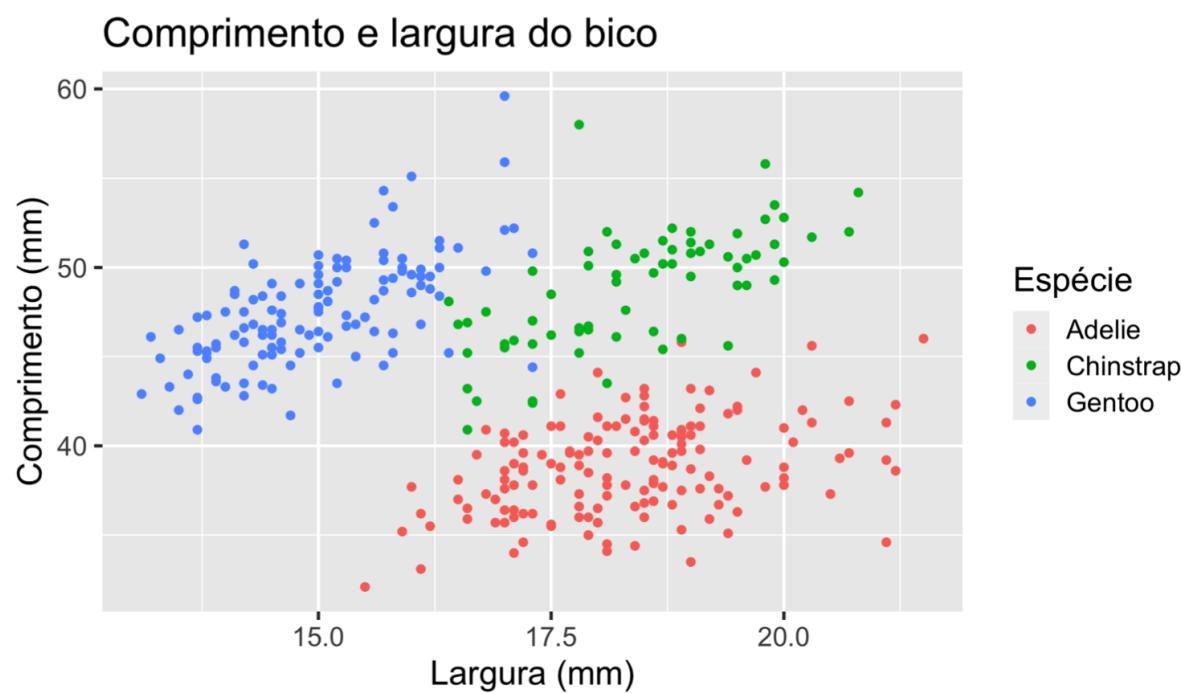
densidade
ou frequência
relativa f_i :
 $n.$ casos do tipo i
 $/ total$



- Visualmente **histograma**

Medidas de dispersão

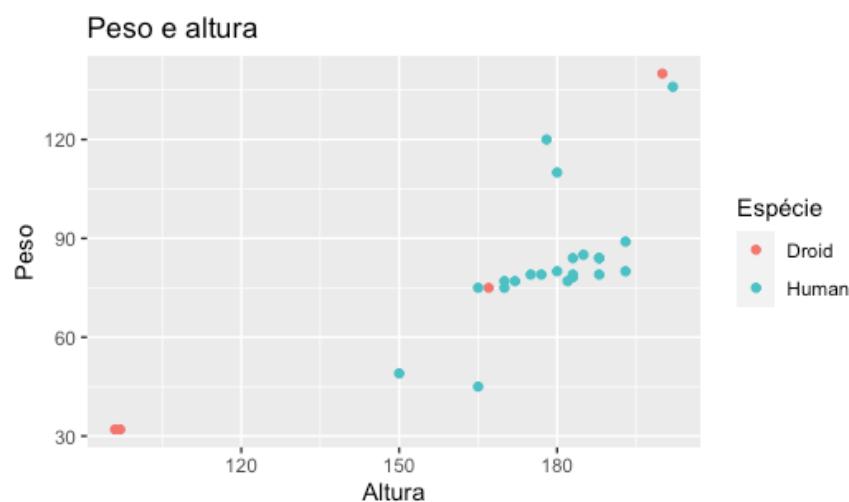
- Ou **scatter plot** (nuvem de pontos)



Experimente

Replique o gráfico anterior para as personagens de **starwars** pertencentes às espécies “Human” e “Droid”, com o peso (**mass**) no eixo dos xx e a altura (**height**) no eixo dos yy.

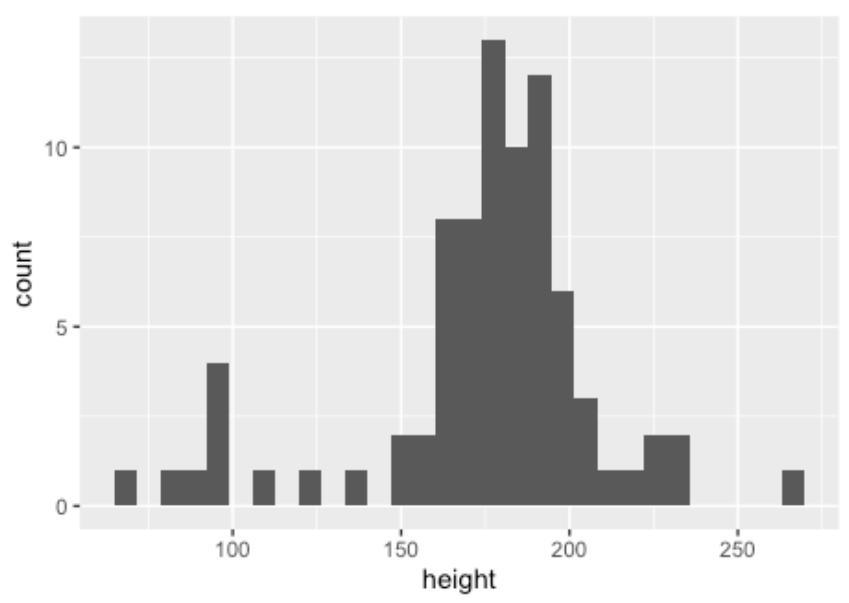
```
> ggplot(data = starwars %>% filter(species %in% c('Human', 'Droid'))),  
+         mapping = aes(x = height,  
+                           y = mass,  
+                           colour = species)) +  
+         geom_point() +  
+         labs(title = "Peso e altura",  
+               x = "Altura", y = "Peso",   colour = "Espécie")
```



Caracterização de distribuições

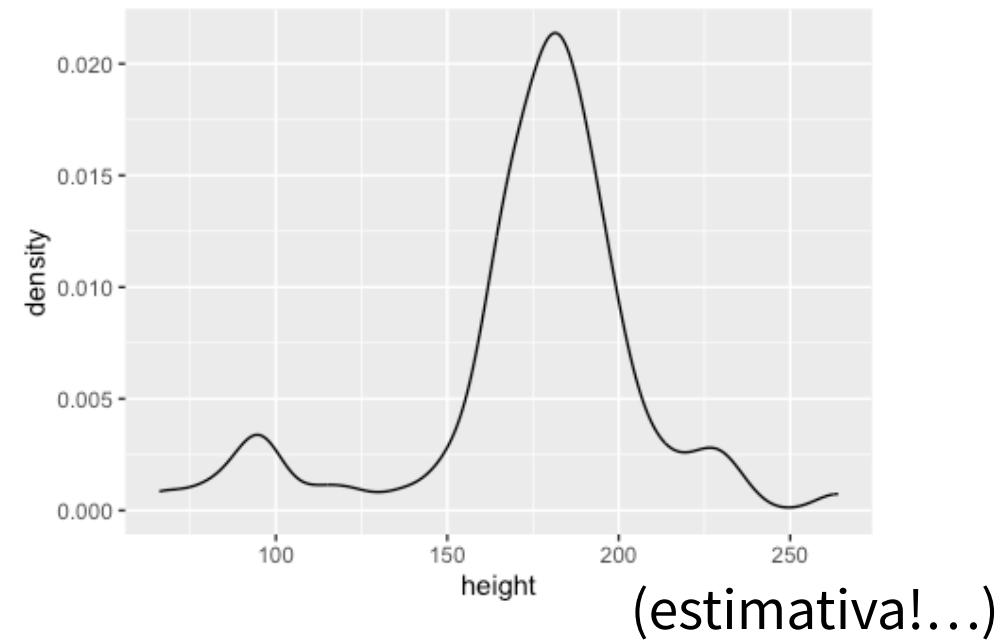
frequência **absoluta**

```
ggplot(starwars, aes(x=height))  
+ geom_histogram()
```



frequência relativa ou **densidade**:
n. casos do tipo i / total

+ geom_density()



Mais conceitos

Variável aleatória

- Quantidade empírica, cujo valor depende de fatores aleatórios
- ≠ realização ou resultado
- e.g. resultado do dado ou preço médio -> v.a.;
 - 6 ou 50.000€ -> realização

Distribuição de probabilidade

- Modelo matemático ou função que relaciona:
 - os diferentes resultados possíveis da v.a.
 - a probabilidade de os obter
- e.g. 1/6 para cada resultado do dado

As propriedades rurais em Portugal

Propriedades transacionadas entre 1990 e 2015

$\{P_i\}$

$i = 1, 2, 3, \dots, N$

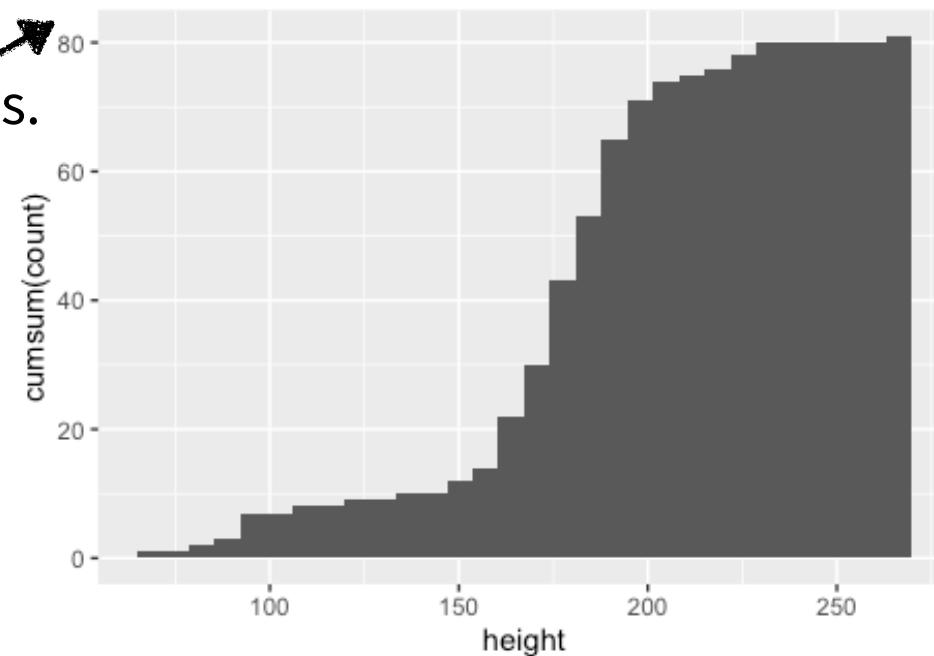
Frequênciacumulada

frequênciacumulada:

```
ggplot(starwars, aes(x=height))
```

```
+ stat_bin(aes(y=cumsum(after_stat(count)), geom="step"))
```

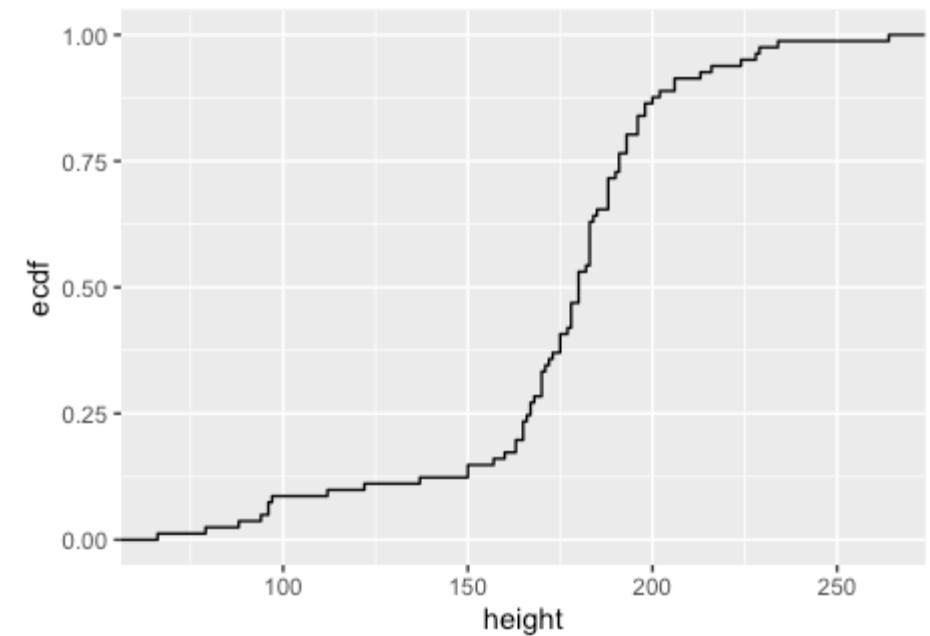
Total obs.



Dados ordenados!

frequênciacumulada:

```
+ stat_ecdf()
```

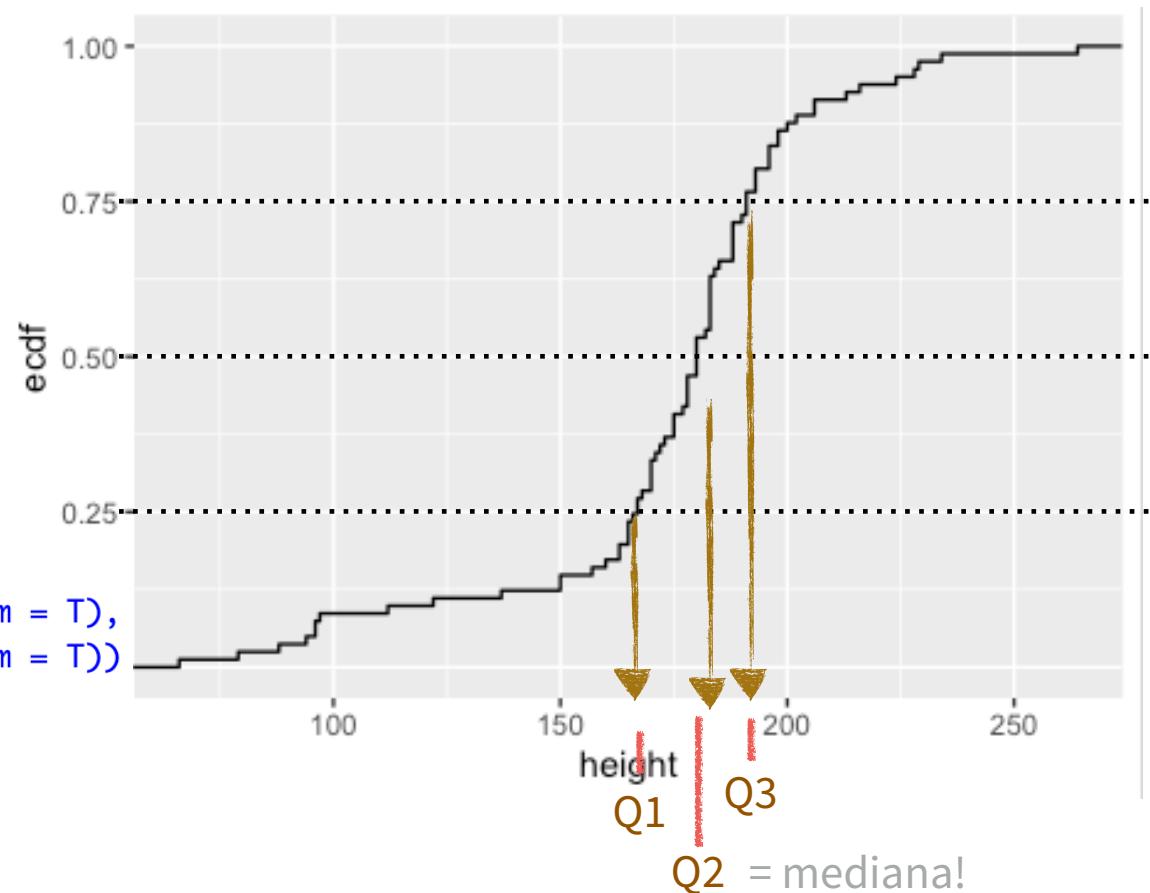


Quantis

- Dividir a distribuição em Q partes iguais e encontrar os valores associados
- Exemplo clássico: quartis 4 ($q = 1/4 = 0.25$)

```
quantile(x, probs = 0.5)
```

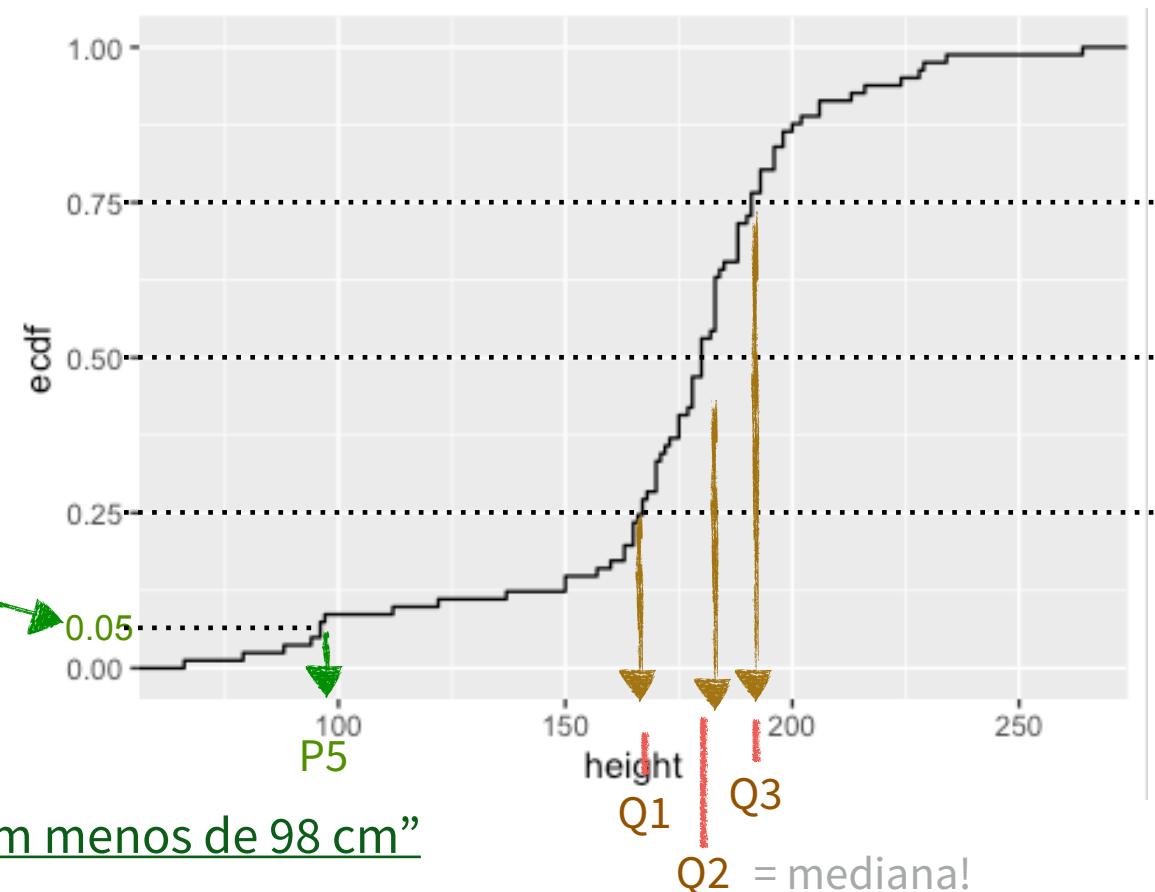
```
> starwars %>%  
+   summarize(Q1=quantile(height, probs=1/4, na.rm = T),  
+             Q3=quantile(height, probs=3/4, na.rm = T))  
# A tibble: 1 × 2  
  Q1     Q3  
  <dbl> <dbl>  
1 167    191
```



Quantis

- Dividir a distribuição em Q partes iguais e encontrar os valores associados
- Exemplo clássico: quartis
 4 ($q = 1/4 = 0.25$)
- Muito comum: percentil

“5% das personagens têm menos de 98 cm”



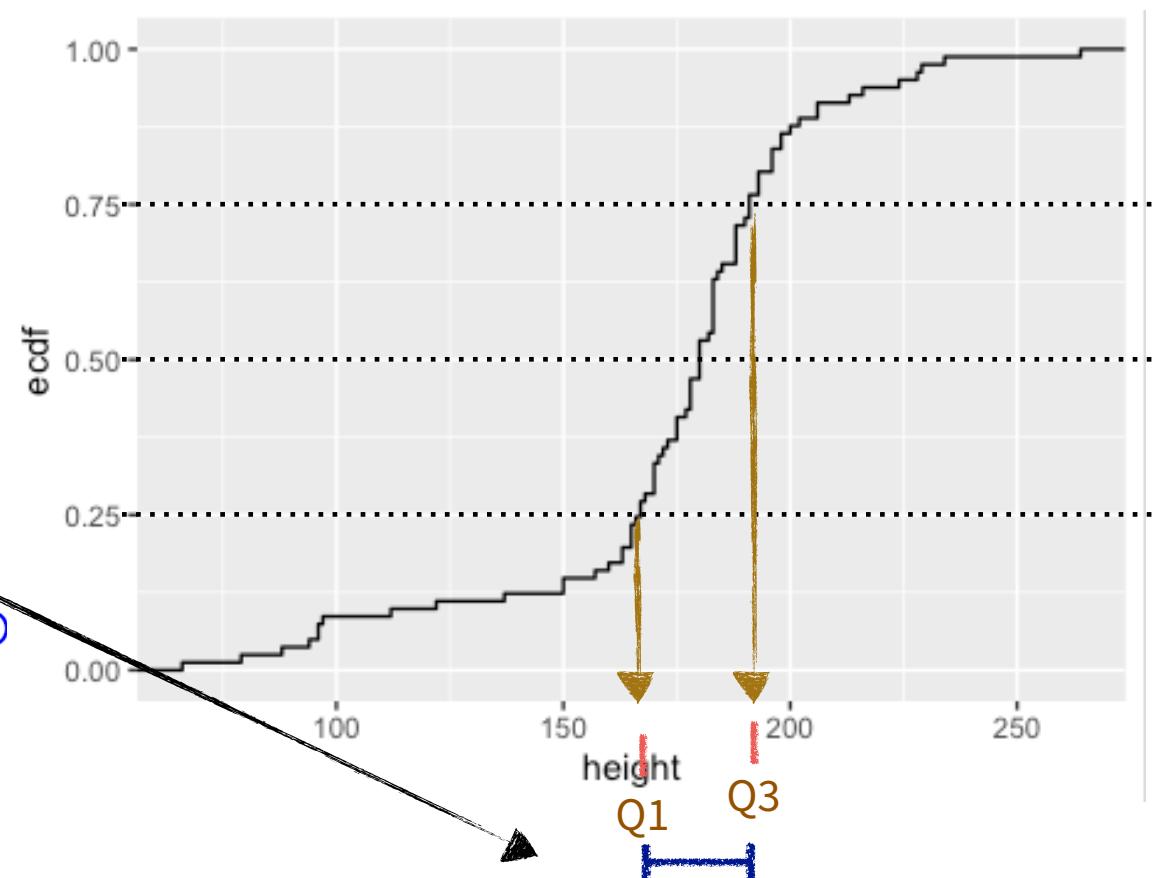
Amplitude interquartis

- Medida de dispersão simples
- Não é sensível a outliers
- Usa apenas os 50% “do meio”

IQR (*interquartile range*)
= $Q_3 - Q_1$

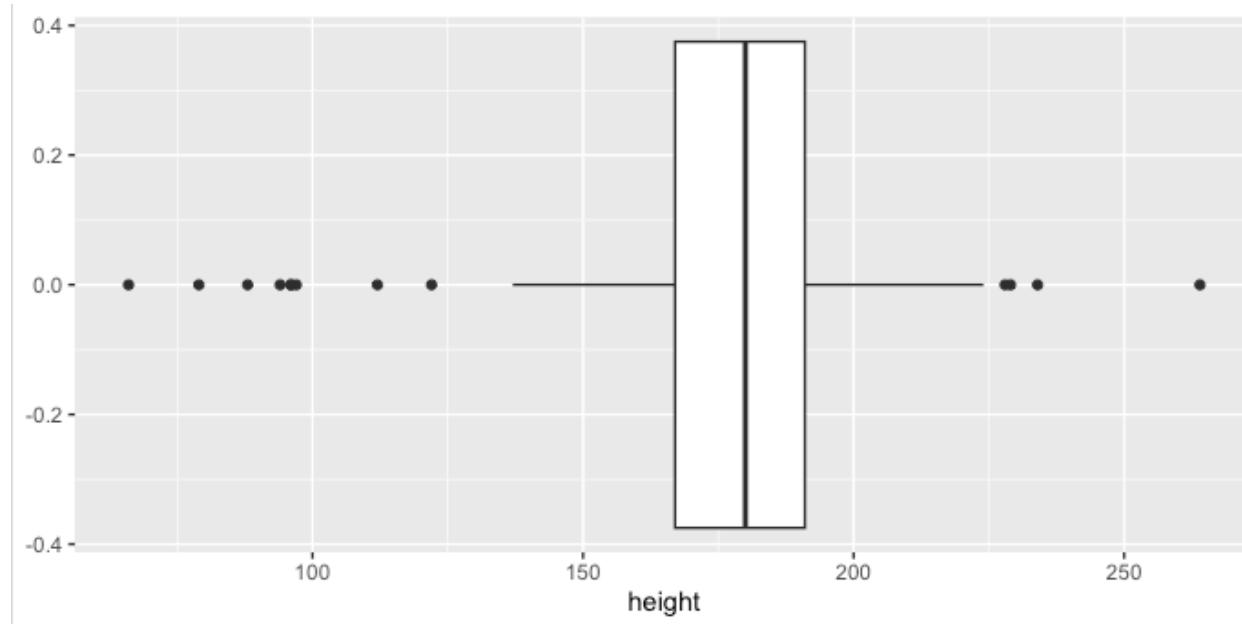
IQR()

```
> starwars %>% summarize(IQR(height, na.rm = T))  
# A tibble: 1 × 1  
`IQR(height, na.rm = T)`  
<dbl>  
1 24
```



Box plot - diagrama de extremos e quartis

- Sumário gráfico das principais características de uma distribuição



Experimente

Produza o diagrama de extremos e quartis para a variável altura (**height**) das personagens de **starwars**.

O que significam os “bigodes” (*whiskers*), e a que valores correspondem? Investigue utilizando a ajuda da função utilizada para o gráfico.

geom_boxplot()

```
> ggplot(starwars, aes(x=height)) + geom_boxplot()
```

Files Plots Packages Help Viewer Presentation

R: A box and whiskers plot (in the style of Tukey) ▾ Find in T

Summary statistics

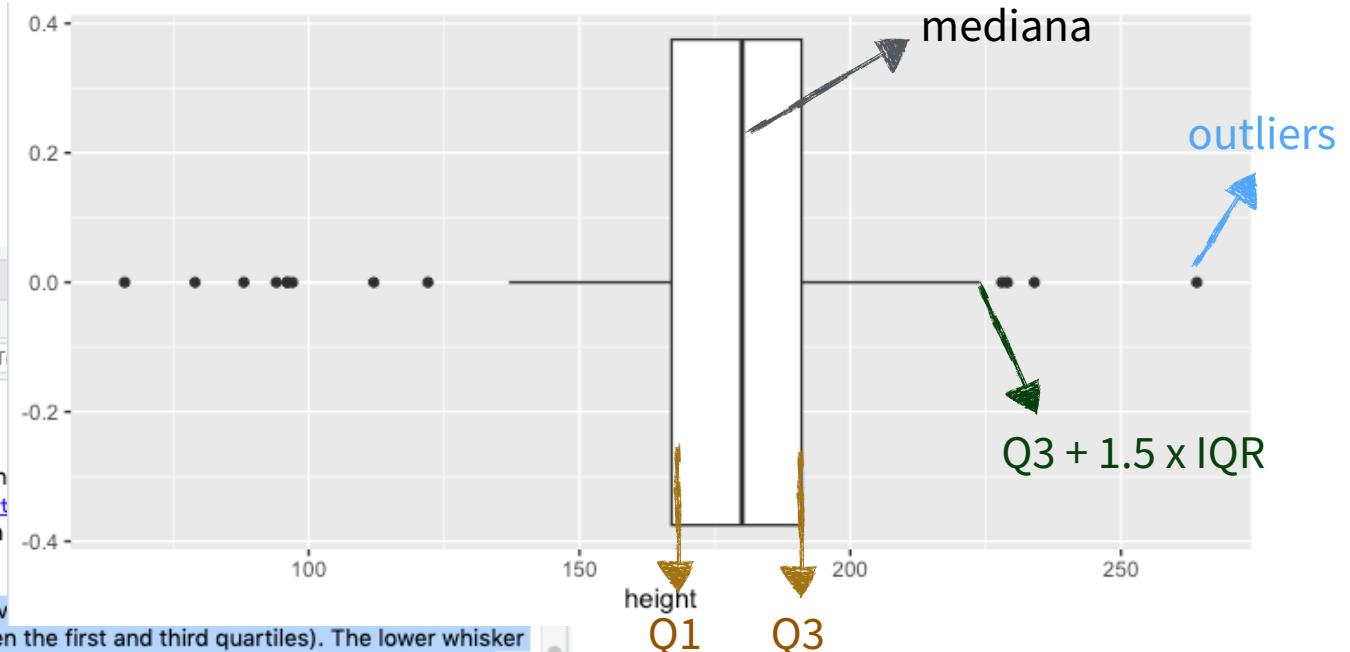
The lower and upper hinges correspond to the first and third quartiles. This differs slightly from the method used by the [boxplot](#) function in R's base samples. See [boxplot.stats\(\)](#) for more information on [boxplot\(\)](#).

The upper whisker extends from the hinge to the largest value within 1.5 times the IQR (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. Data beyond the end of the whiskers are called "outlying" points and are plotted individually.

In a notched box plot, the notches extend $1.58 * \text{IQR} / \sqrt{n}$. This gives a roughly 95% confidence interval for comparing medians. See McGill et al. (1978) for more details.

Aesthetics

`geom_boxplot()` understands the following aesthetics (required aesthetics are in bold):



**Obrigado
e bom fim-de-semana!**

luis.morais@novasbe.pt