

Preparado para:



# REFORM/SC2022/126 DELIVERABLE 4 **MÓDULO 3** **ESTATÍSTICA BÁSICA** **EM R**

DESIGNING A NEW VALUATION MODEL  
FOR RURAL PROPERTIES IN PORTUGAL

## Parte II

Formador: Luís Teles Morais | Nova SBE  
*Lisboa, 22 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

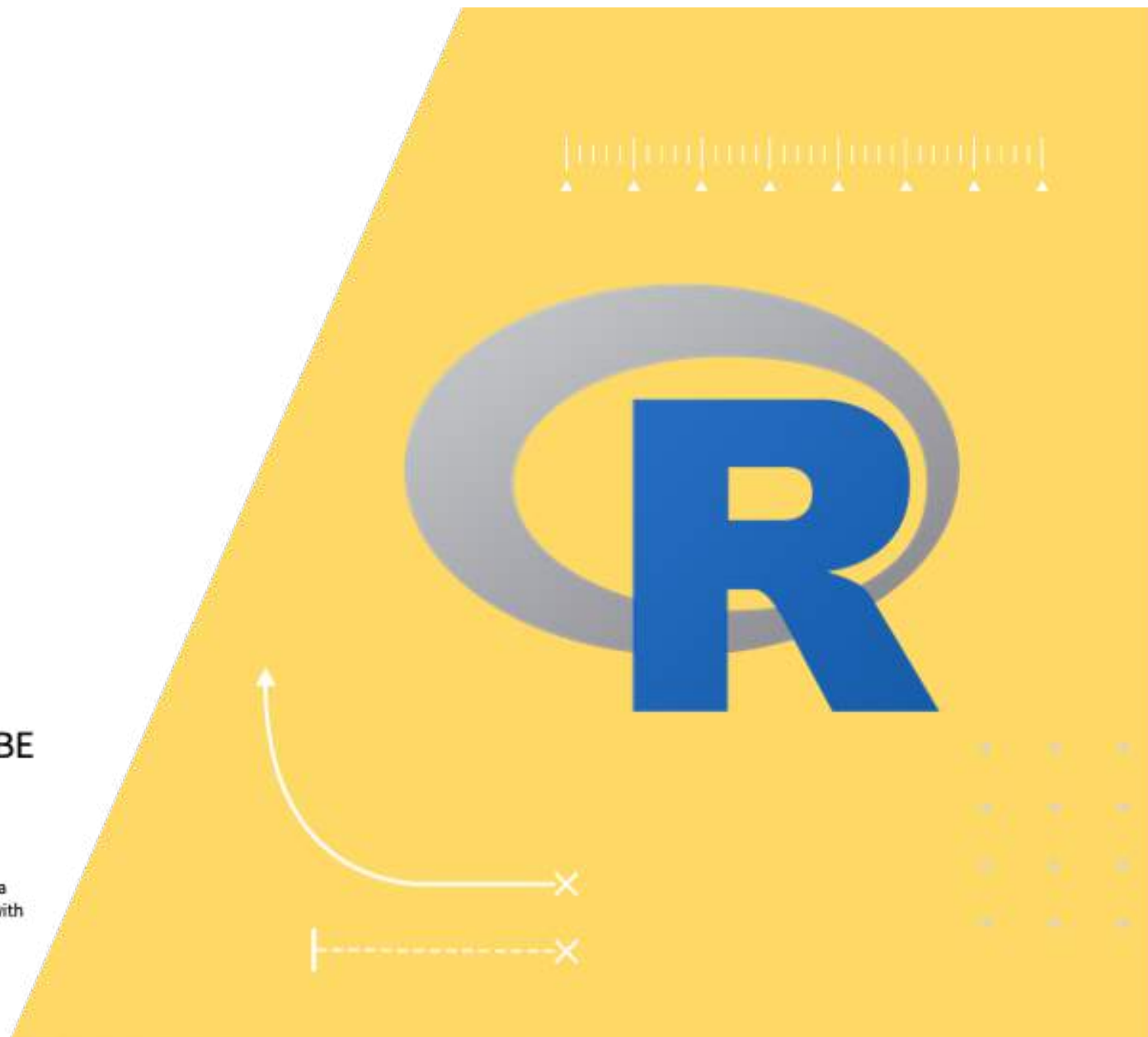
AARC

NOVA

LOBO VASQUES

esri

WINE LANDS



# Programa

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

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

# Vamos a isso

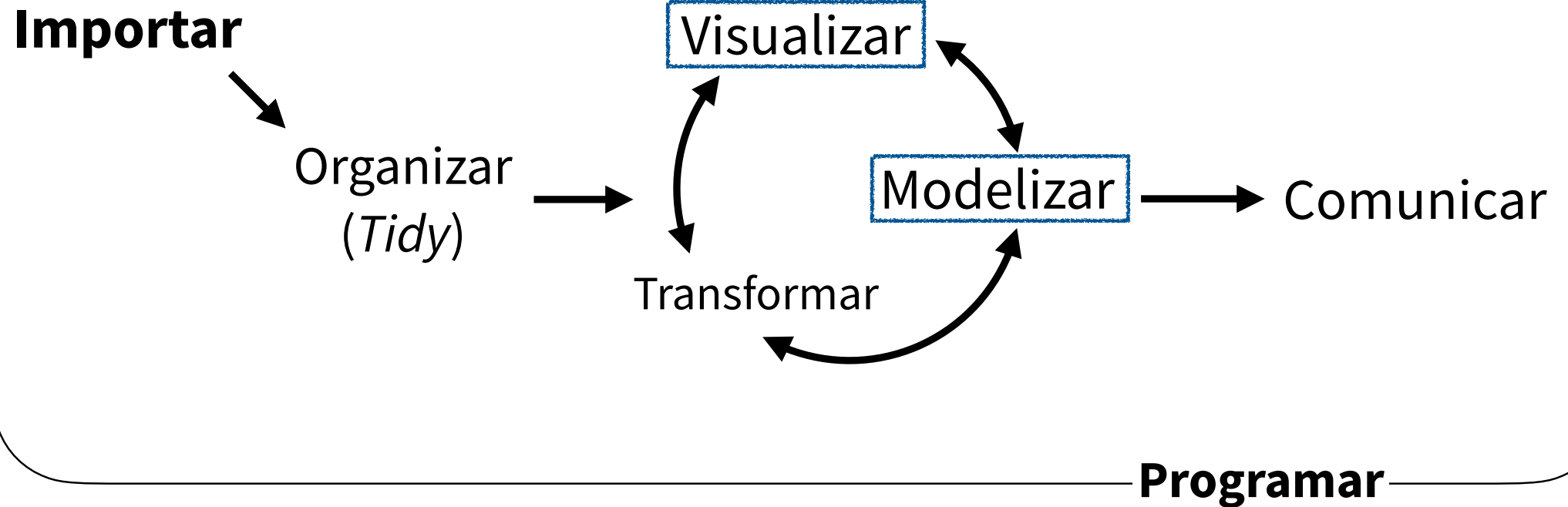
Aceda a este link para começar já

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

# **Estatística básica em R (II)**

## Distribuições probabilísticas

# Ciência de dados



# A semana passada

## Variável aleatória

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

$\{P_i\}$   
 $i = 1, 2, 3, \dots, N$

# Medidas de tendência central/posição

## Média

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

---

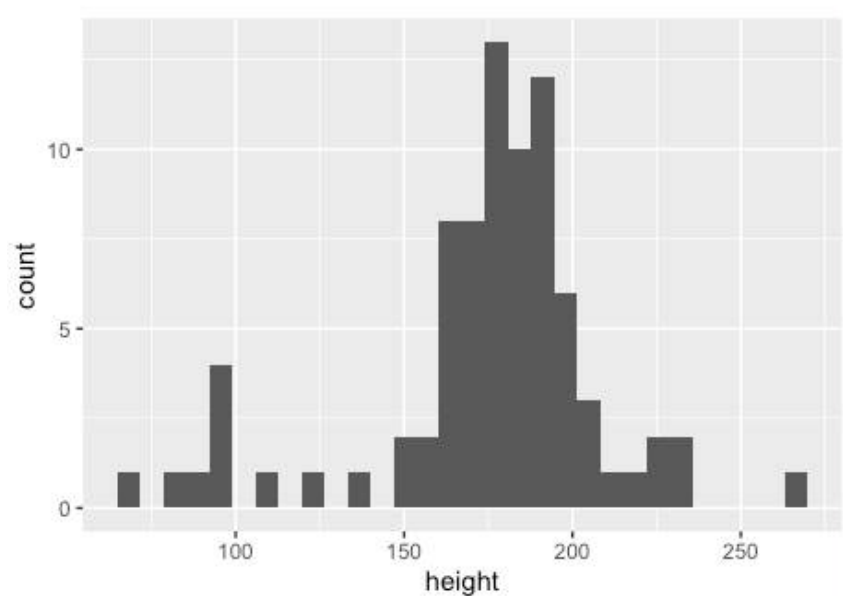
- Serve para quaisquer dados numéricos (discretos ou contínuos)
- Medida de posição, mas sensível a *outliers*

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

# Caracterização de distribuições

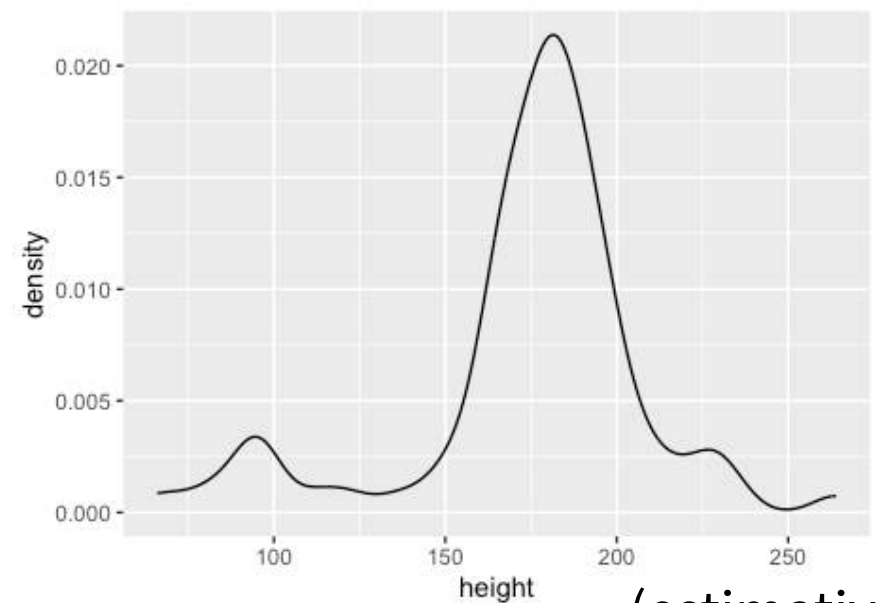
frequência **absoluta**

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



frequência relativa ou **densidade**:  
 $\text{n. casos do tipo } i / \text{total}$

```
+ geom_density()
```



(estimativa!...)

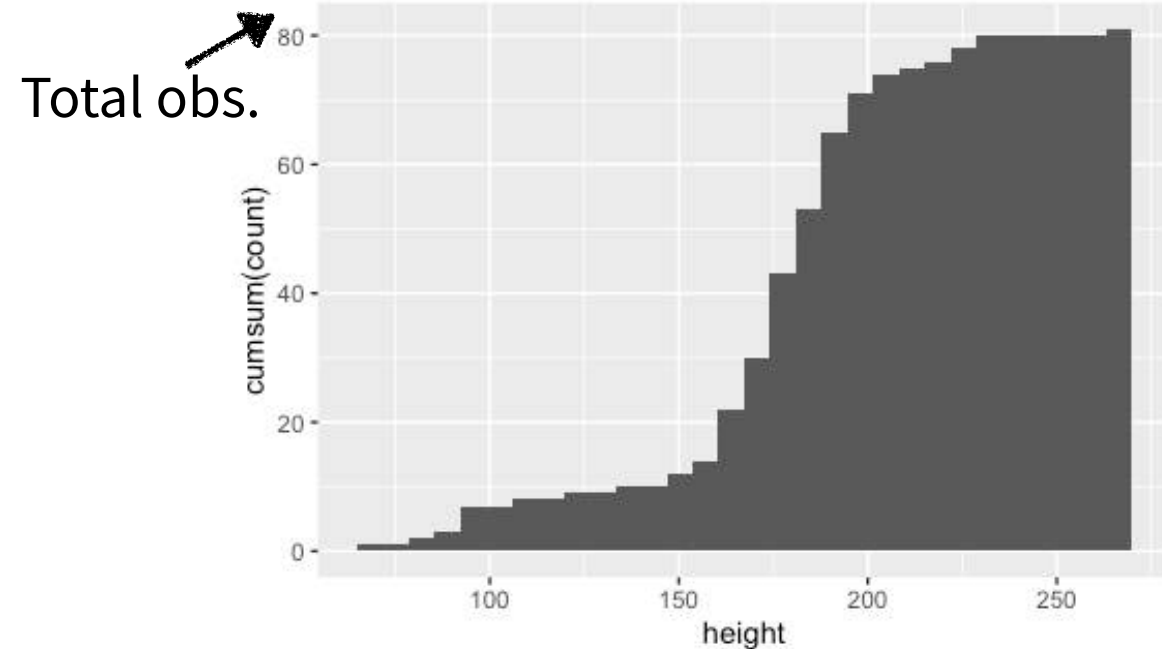


# Frequência acumulada

frequência absoluta acumulada:

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

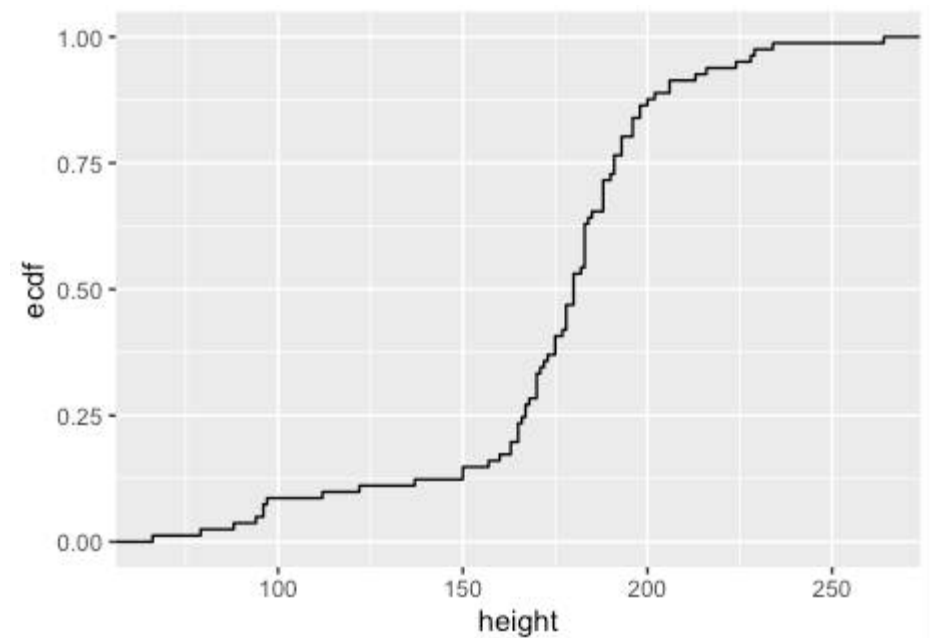
```
+ stat_bin(aes(y=cumsum(after_stat(count)), geom="step"))
```



Dados ordenados! →

frequência relativa acumulada:

```
+ stat_ecdf()
```

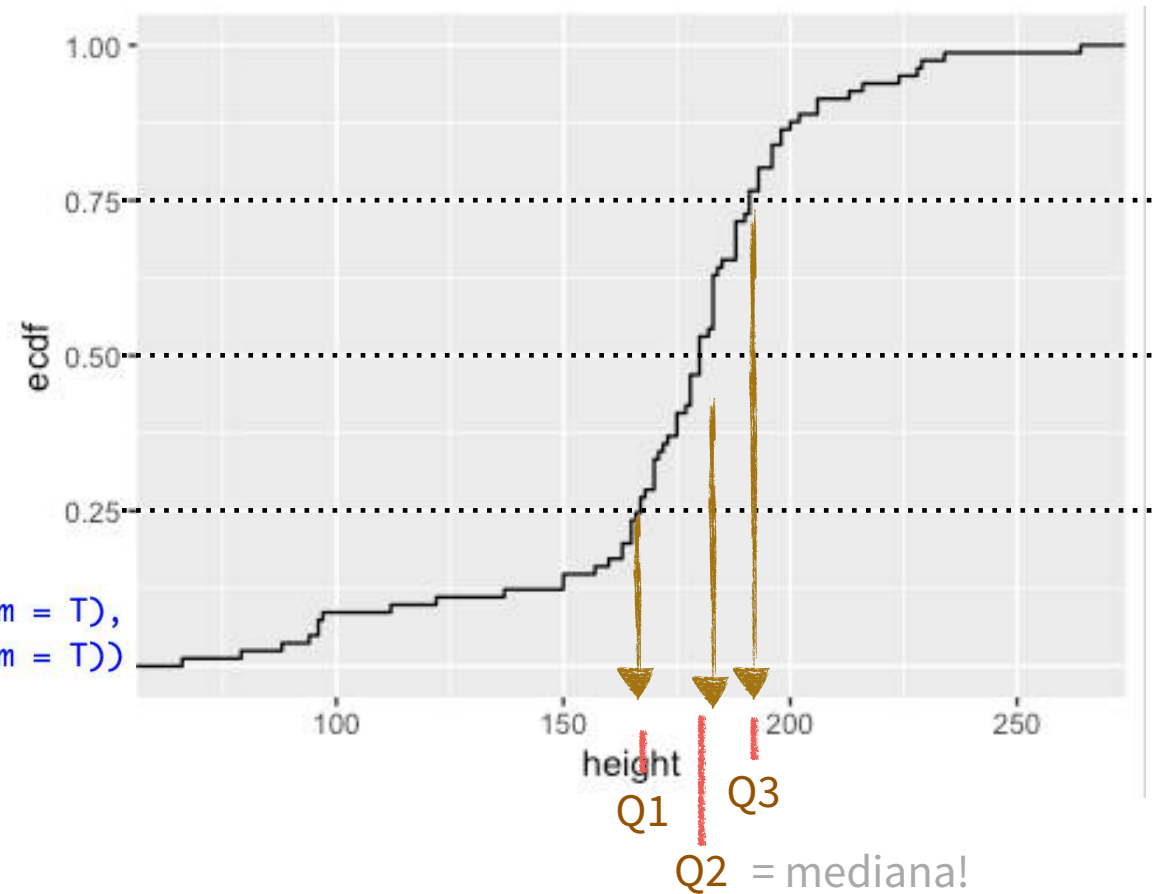


# Quantis

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

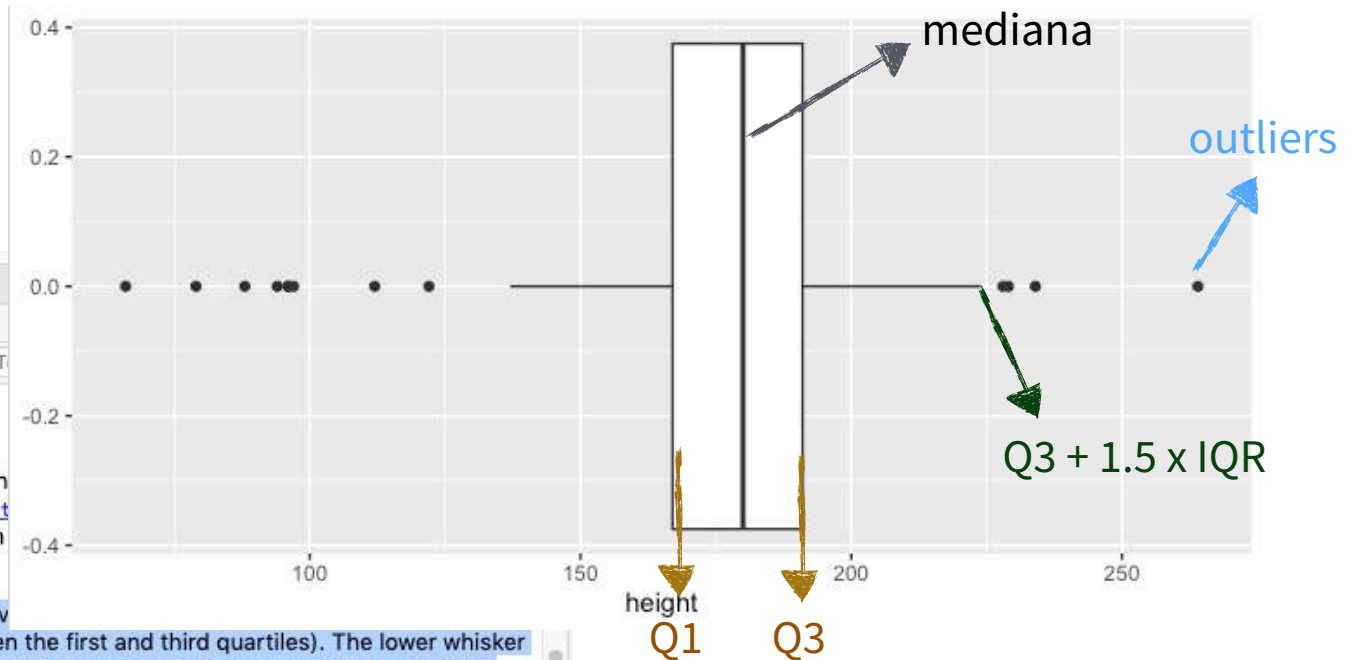
```
quantile(x, probs = 0.5)
```

```
> 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
```



# 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.stats()` function. See `boxplot.stats()` for more information on `boxplot()`.

The upper whisker extends from the hinge to the largest value not more than 1.5 \* IQR of the hinge. 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  $1.58 * IQR / \sqrt{n}$ . 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):

# Variância e desvio-padrão

- Variabilidade dos dados em torno da média

- Variância: média da distância quadrática à média

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

`var(x)`

- Desvio-padrão: raiz quadrada da variância
- Grande vantagem: **mesma unidade dos dados**

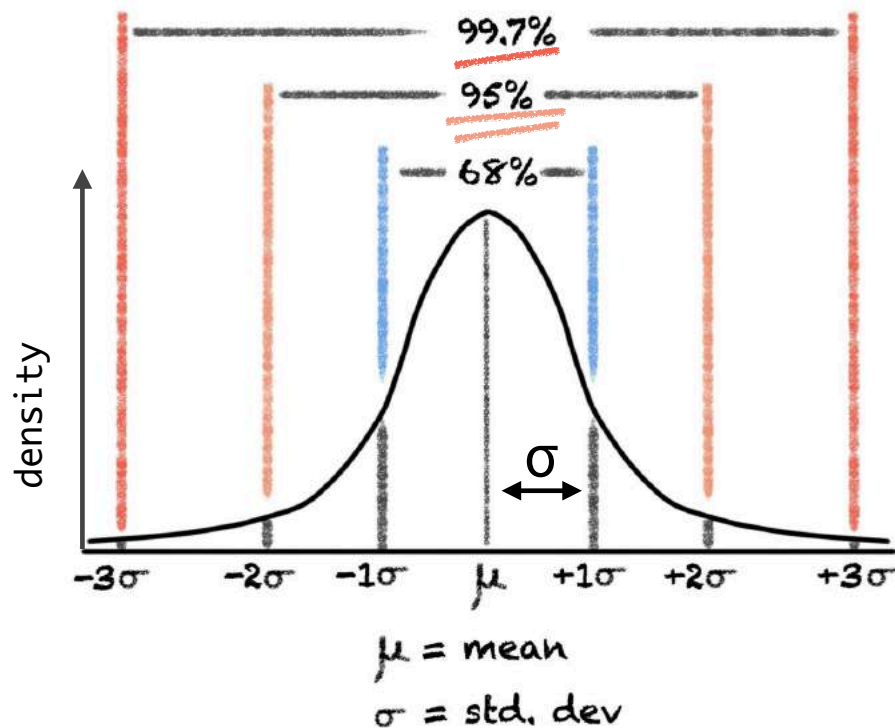
$$sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

`sd(x)`

```
> starwars %>% summarize(variancia = var(height, na.rm = T),
+                         desvpad = sd(height, na.rm = T),
+                         media = mean(height, na.rm = T))
# A tibble: 1 x 3
  variancia desvpad media
  <dbl>    <dbl> <dbl>
1    1209.    34.8  174.
```

# Distribuição normal

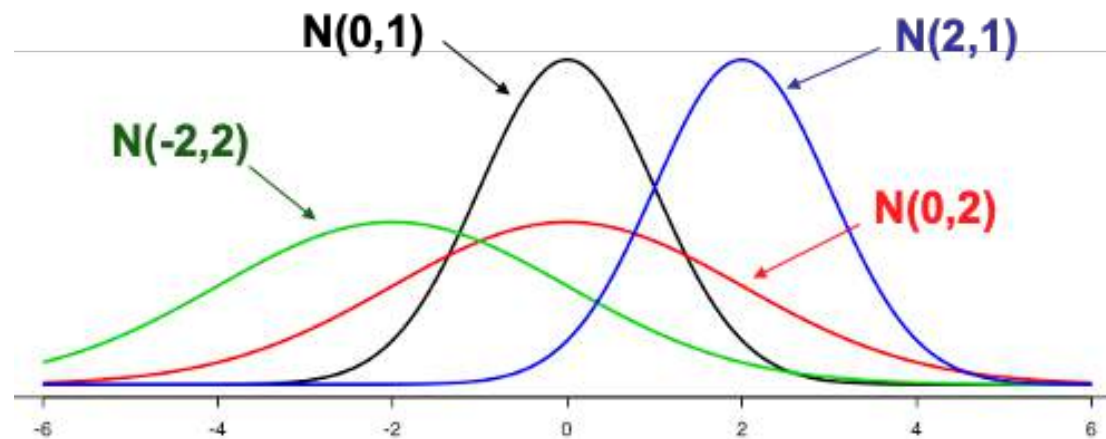
- Muitas variáveis na realidade são aproximadamente “normais”



- Também curva de Gauss
- $\sim N(\mu, \sigma)$ :
  - média  $\mu$  e desv. pad.  $\sigma$
  - ditam posição e dispersão

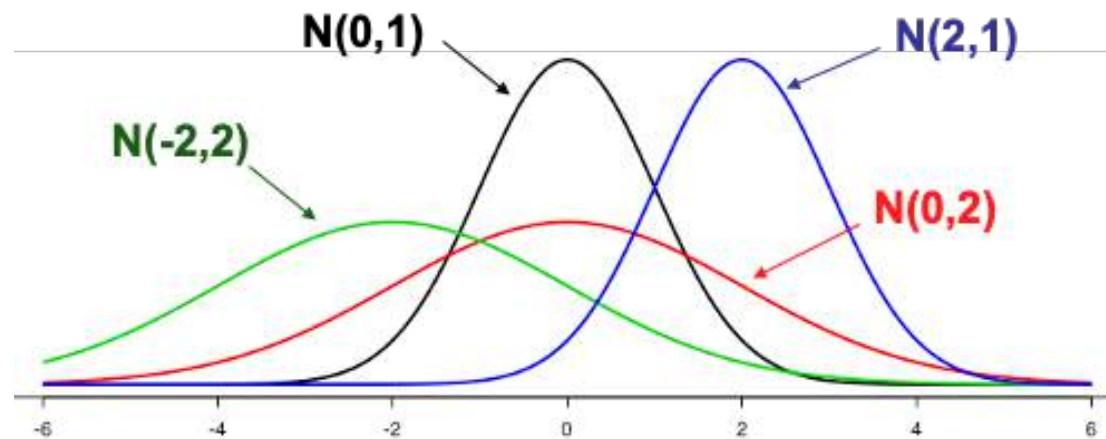
Moda?

# Distribuição normal



- $\sim N(\mu, \sigma)$ : qualquer normal tem esta forma, média  $\mu$  e desv. pad.  $\sigma$  ditam posição e dispersão
- Enviesamento = 0

# Distribuição normal



- $Z = N(0,1)$  — distribuição normal estandardizada
- Qualquer distribuição normal  $X$  pode ser transformada para ficar igual à estandardizada:  $Z = (X - \mu) / \sigma$

# Experimente

1. Adicione à tabela **starwars** uma nova variável (**I\_Human** por ex.) que indique se a observação é humana (**species == “Human”**).
2. Calcule a média e o desvio padrão (**sd()**) da altura (**height**) para os humanos e para os restantes, usando a variável de 1. para agrupar os dados.
3. Volte à tabela original com os dados completos e construa um gráfico com a densidade estimada para a distribuição da altura, para os humanos e para os restantes, com cores diferentes para cada grupo.

! Os casos sem valor de *species* não devem aparecer no gráfico !



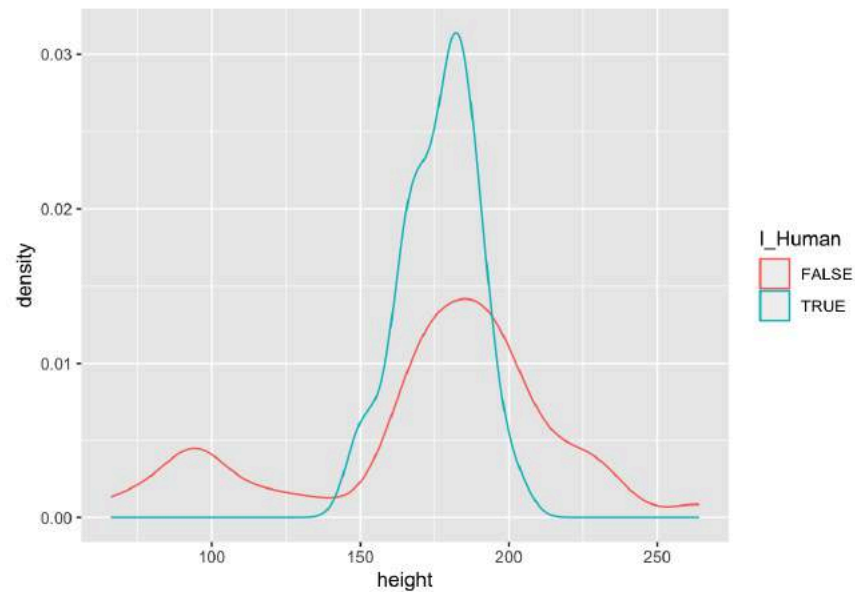
```
starwars <- starwars %>% mutate(I_Human = (species == "Human"))
```

```
starwars %>% group_by(I_Human) %>%  
  summarize(med = mean(height, na.rm = T),  
            sd = sd(height, na.rm = T))
```

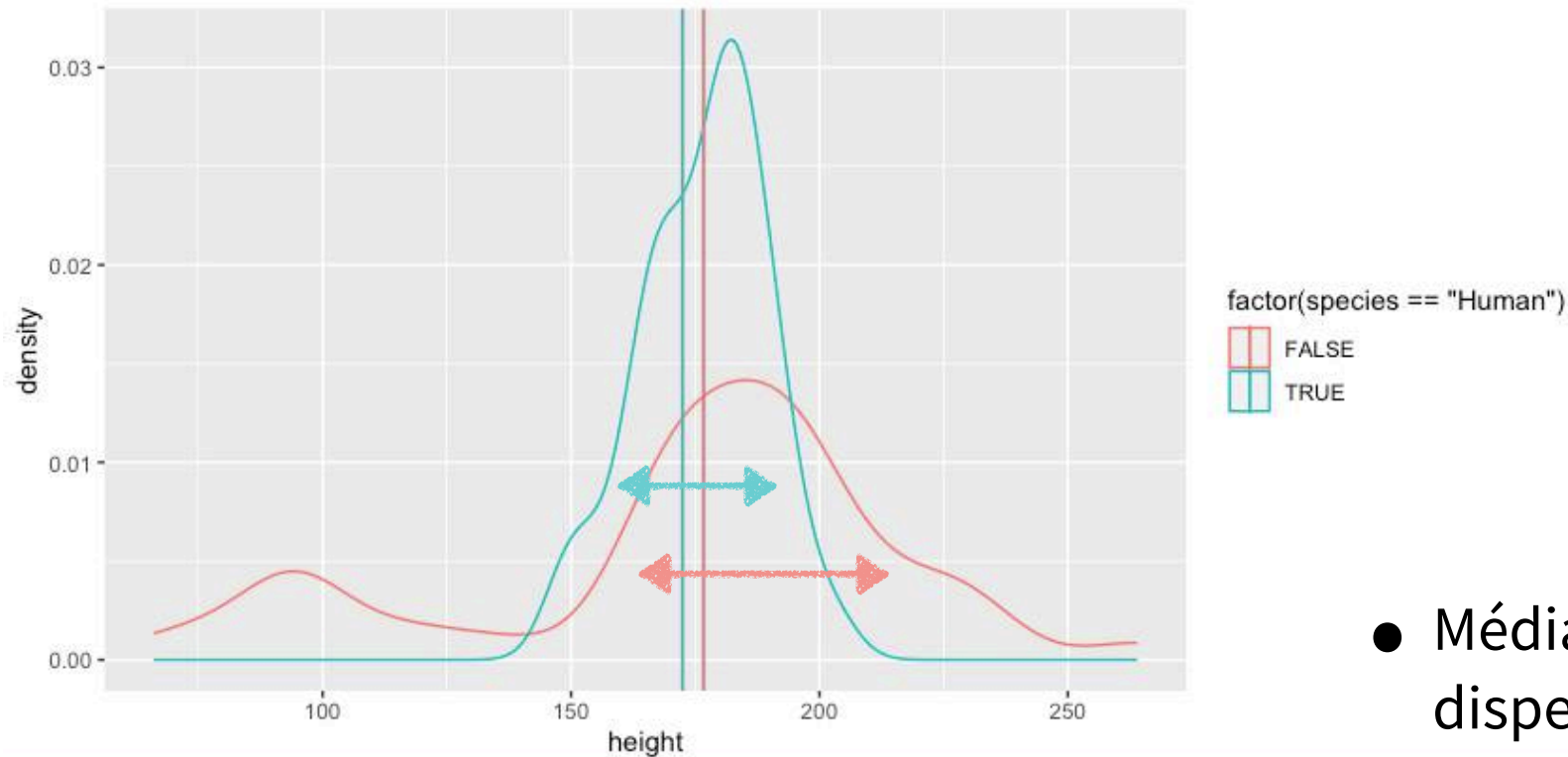
```
# A tibble: 3 × 3
```

	I_Human	med	sd
	<lgl>	<dbl>	<dbl>
1	FALSE	172.	44.6
2	TRUE	177.	12.5
3	NA	181.	2.89

```
ggplot(starwars %>% filter(!is.na(species)),  
       aes(x=height, color = factor(species == 'Human'))  
       ) +  
geom_density()
```



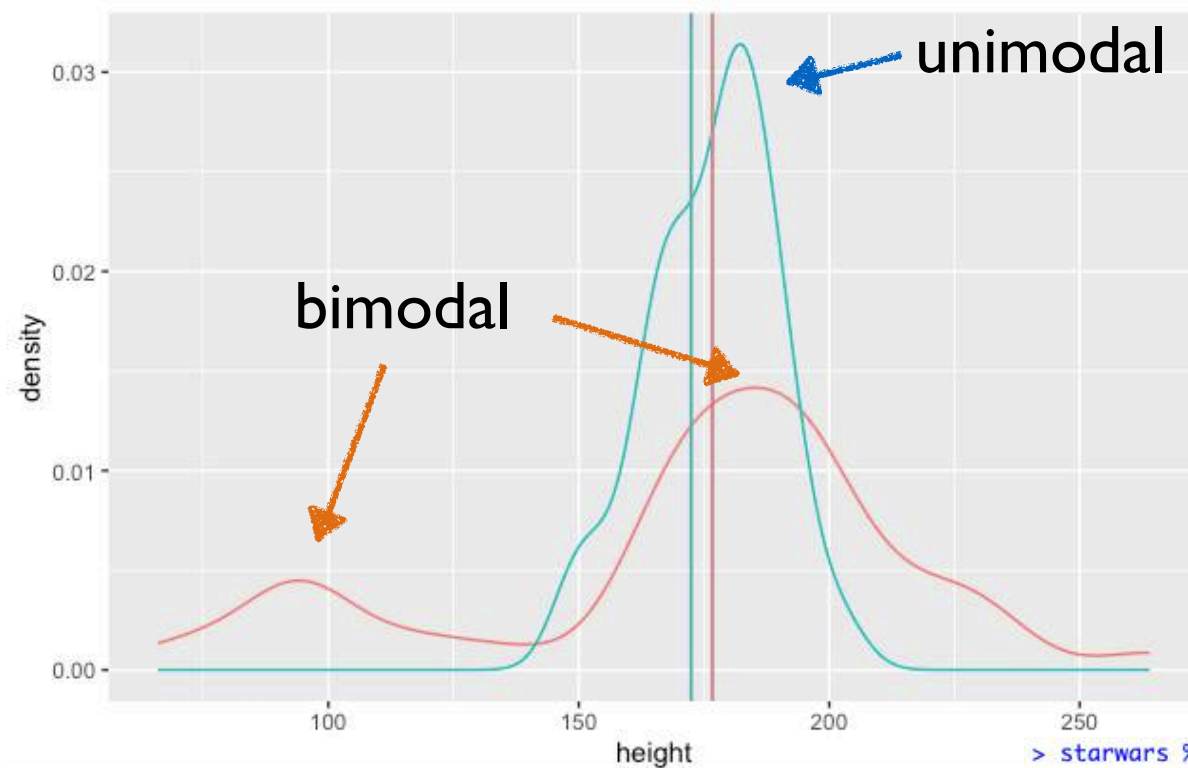
# Variância e desvio-padrão



- Média semelhante, dispersão diferente!

	<code>factor(species == "Human")</code>	med	sd
	<code>&lt;fct&gt;</code>	<code>&lt;dbl&gt;</code>	<code>&lt;dbl&gt;</code>
1	FALSE	172.	44.6
2	TRUE	177.	12.5
3	NA	181.	2.89

# Medidas de forma

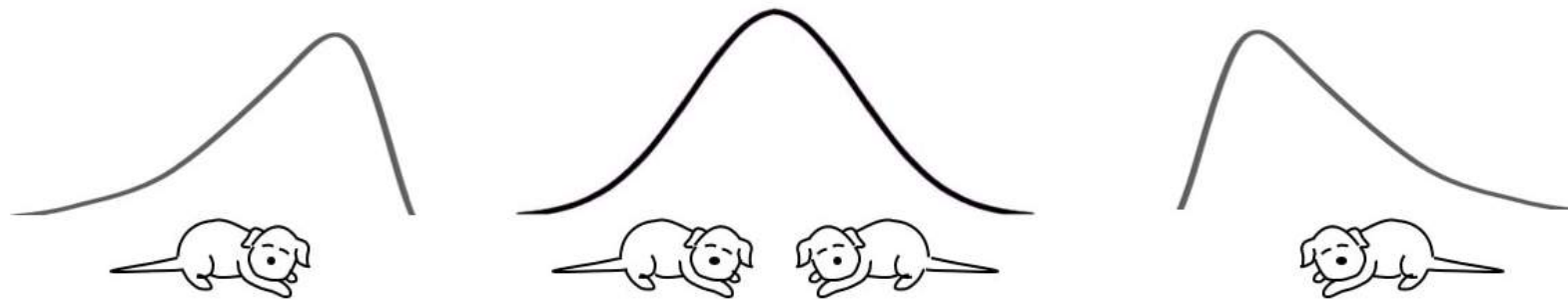


- Unimodalidade

```
> starwars %>% group_by(factor(species == 'Human')) %>%  
+   summarize(med = mean(height, na.rm = T), sd = sd(height, na.rm = T))  
# A tibble: 3 x 3  
  `factor(species == "Human")`   med    sd  
  <fct>                        <dbl> <dbl>  
1 FALSE                        172.  44.6  
2 TRUE                         177.  12.5  
3 NA                           181.   2.89
```

# Enviesamento ou assimetria

- Assimetria dos dados em torno da média

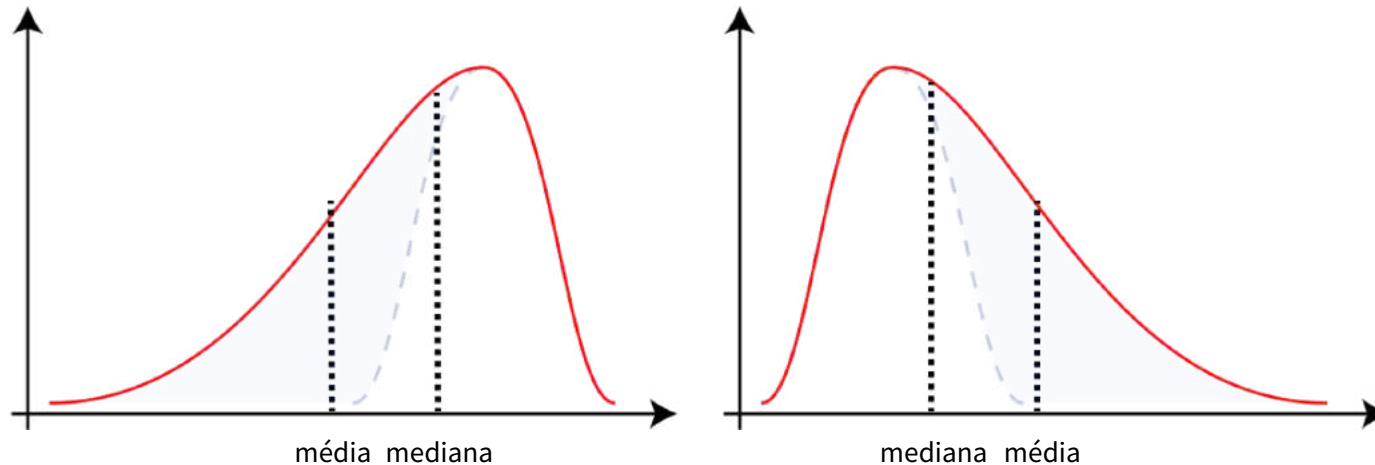


Reproduced by permission of John Wiley and Sons  
from the book, Statistics from A to Z – Confusing Concepts Clarified

- Enviesamento positivo: cauda longa à direita
- Enviesamento negativo: cauda longa à esquerda
- Simétrica: caudas iguais
- *Em variáveis assimétricas, especialmente importante olhar para a mediana e não só para a média*

# Enviesamento ou assimetria

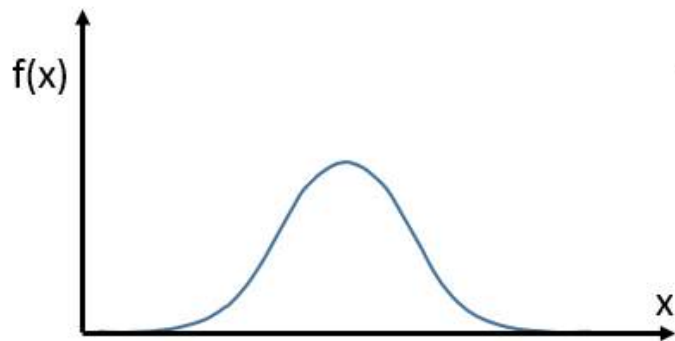
- Assimetria dos dados em torno da média



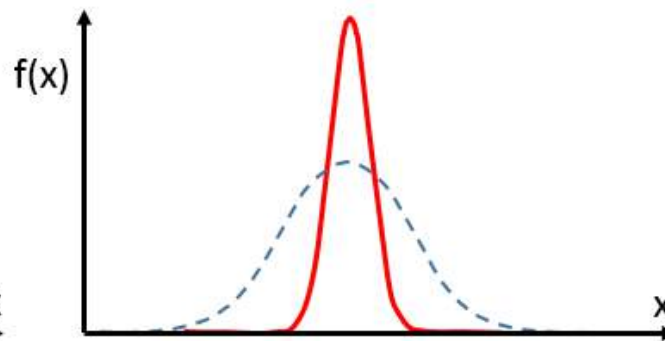
- Negativo:  $\text{mediana} > \text{média}$ 
  - e.g. idade da reforma
- Positivo:  $\text{mediana} < \text{média}$ 
  - e.g. rendimento

# Curtose

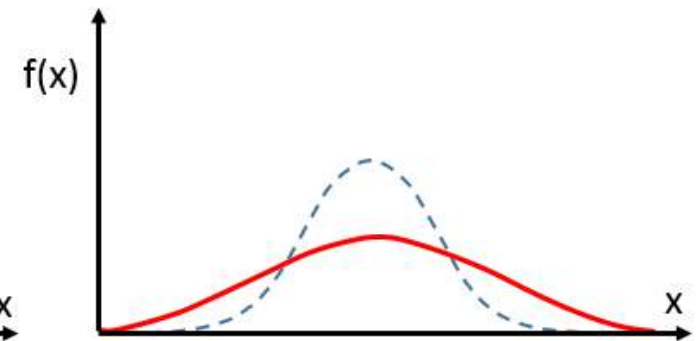
- Assimetria dos dados em torno da média



- Curtose = 3
- Dist. normal



- Curtose positiva ( $> 3$ )



- Curtose negativa ( $< 3$ )

- Largura das caudas
- Importância dos outliers

# Um momento de resumo

- Estivemos a analisar diferentes **momentos** das distribuições:
  - Momento  $k$  de qualquer variável:

$$\pi_k = \frac{1}{N} \sum_{i=1}^N \Psi_i^k$$

- 

- Primeiro momento: média
- Segundo momento: variância
- Terceiro momento: enviesamento
- Quarto momento: curtose



# **Estatística básica em R (II)**

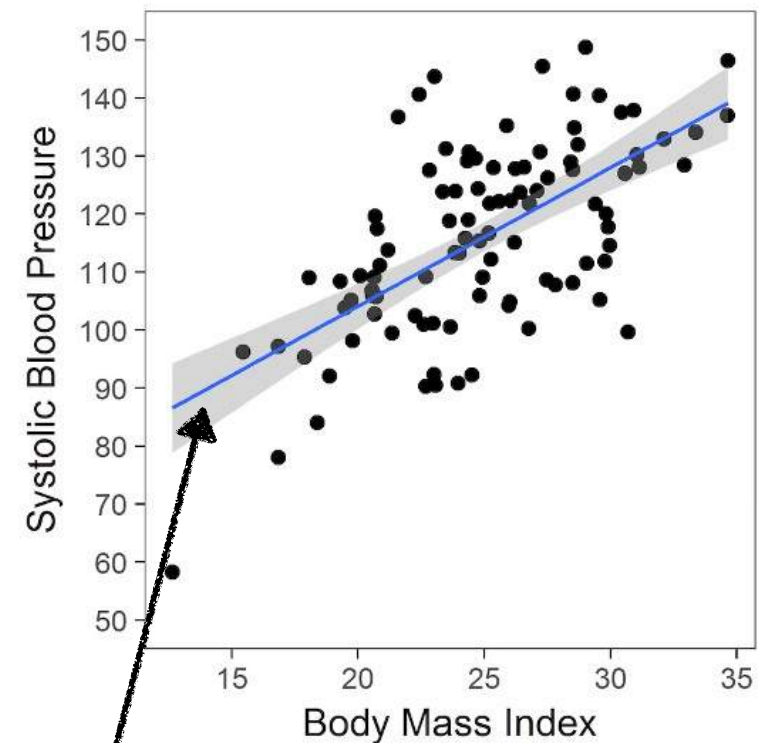
Distribuições bivariadas

# Correlação

- Relação de interdependência entre duas variáveis
- Variabilidade *conjunta*: quando uma mexe, a outra também?

## Também dependência linear

- Geralmente temos em mente uma relação linear
- Quão bem descrita por uma linha de tendência linear é a relação entre as duas variáveis?



```
geom_smooth(method = 'lm')
```

# Covariância e correlação

$\text{cov}(x, y)$

$$\text{COV}_{x,y} = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x}) \cdot (y_i - \bar{y})]$$

*O que significa a covariância de uma variável consigo própria?*

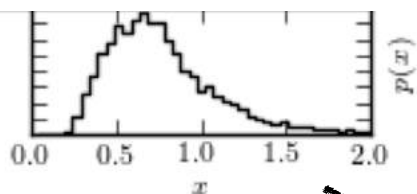
- Produto dos desvios em relação à respetiva média (variância conjunta)

## Coeficiente de correlação pearson

$\text{cor}(x, y)$

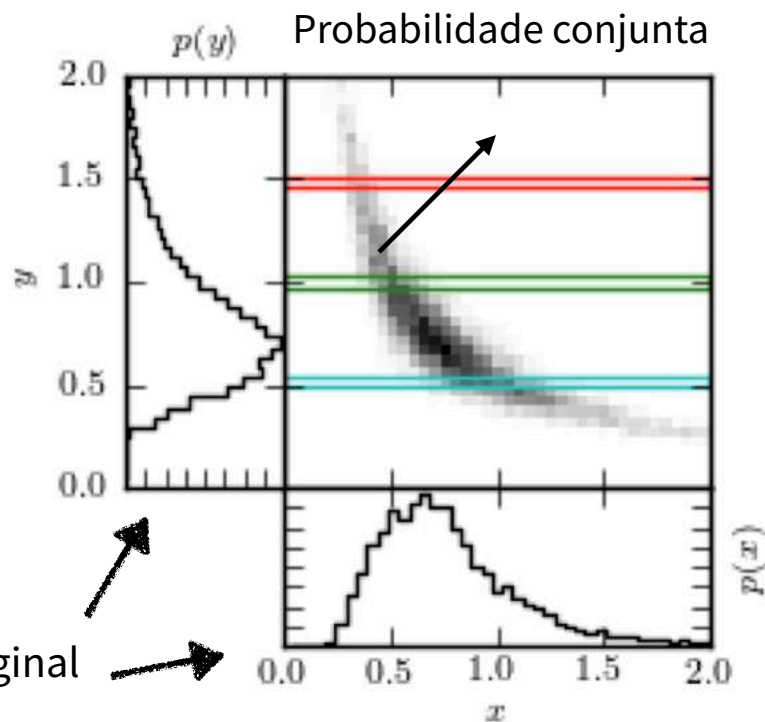
$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- 1 = correlação positiva perfeita
- -1 = correlação negativa perfeita



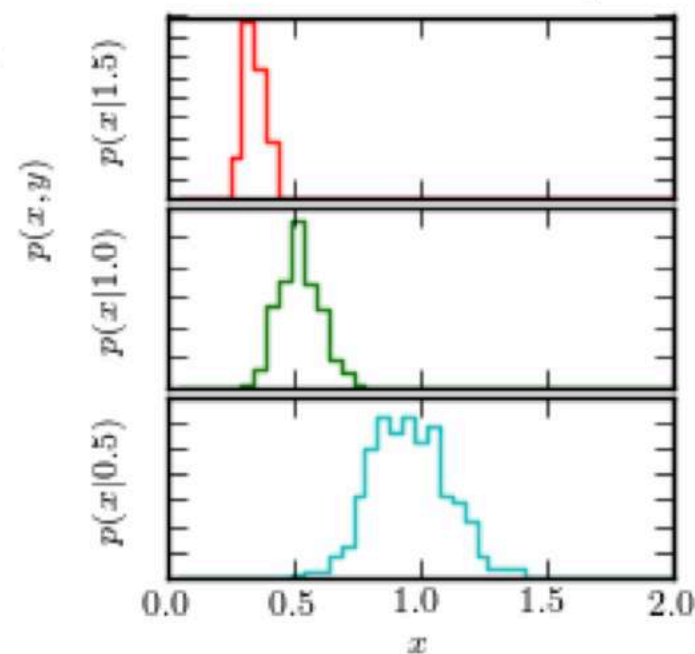
Distribuição univariada

Distribuição bivariada



Probabilidade(s) marginal

Probabilidade condicional



# Mini-teste I

1. Crie uma nova tabela apenas com observações que contenham um valor para o ano de nascimento (`birth_year`) e a altura (`height`).
2. Crie uma tabela sumária com a média e a variância de cada uma das variáveis referidas, a covariância entre elas e o coeficiente de correlação.
3. Desenhe um gráfico de pontos com `birth_year` e a linha de tendência linear.

```
starwars_filter <- starwars %>% filter(!is.na(birth_year) & !is.na(height))
```

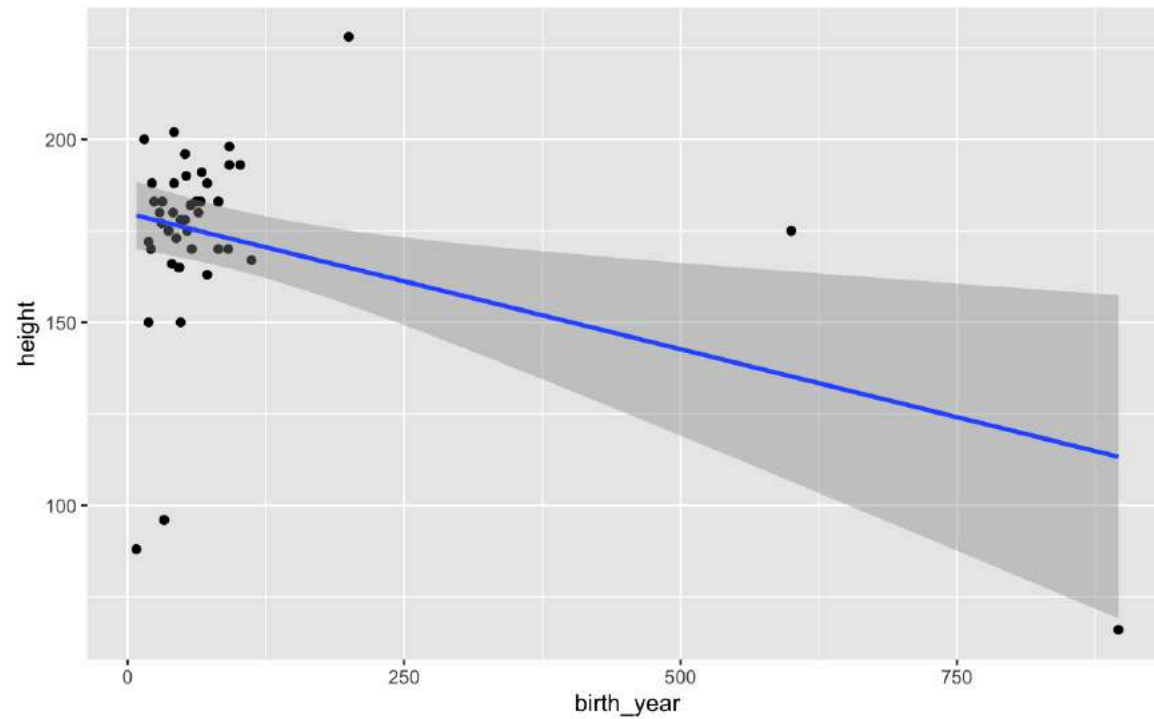
```
starwars_tbl <- starwars_filter %>%  
  summarize(  
    med_height = mean(height, na.rm = T),  
    med_birth_year = mean(birth_year, na.rm = T),  
    var_height = var(height, na.rm = T),  
    var_birth_year = var(birth_year, na.rm = T),  
    corr = cor(birth_year, height),  
    cov = cov(birth_year, height))
```

```
> starwars_tbl
```

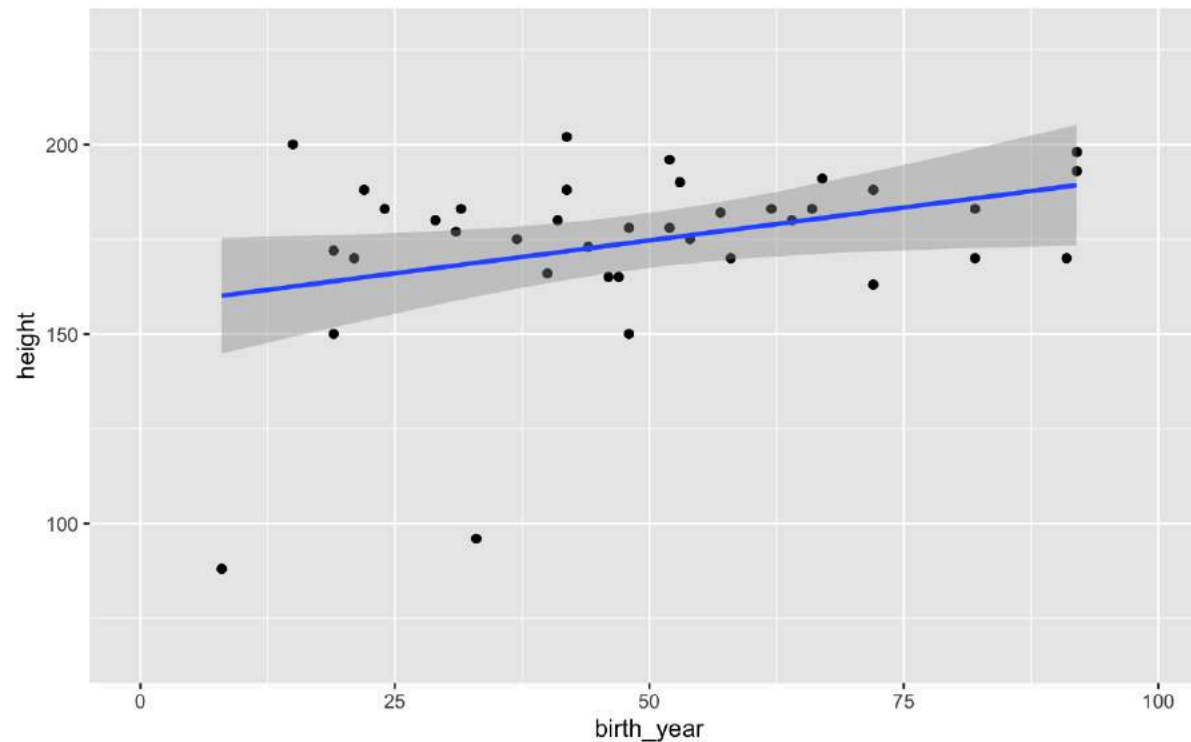
```
# A tibble: 1 × 6
```

	med_height	med_birth_year	var_height	var_birth_year	corr	cov
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	173.	87.6	825.	23929.	-0.400	-1777.

```
starwars_filter %>% ggplot(aes(x = birth_year, y = height)) +  
  geom_point() + geom_smooth(method = 'lm')
```

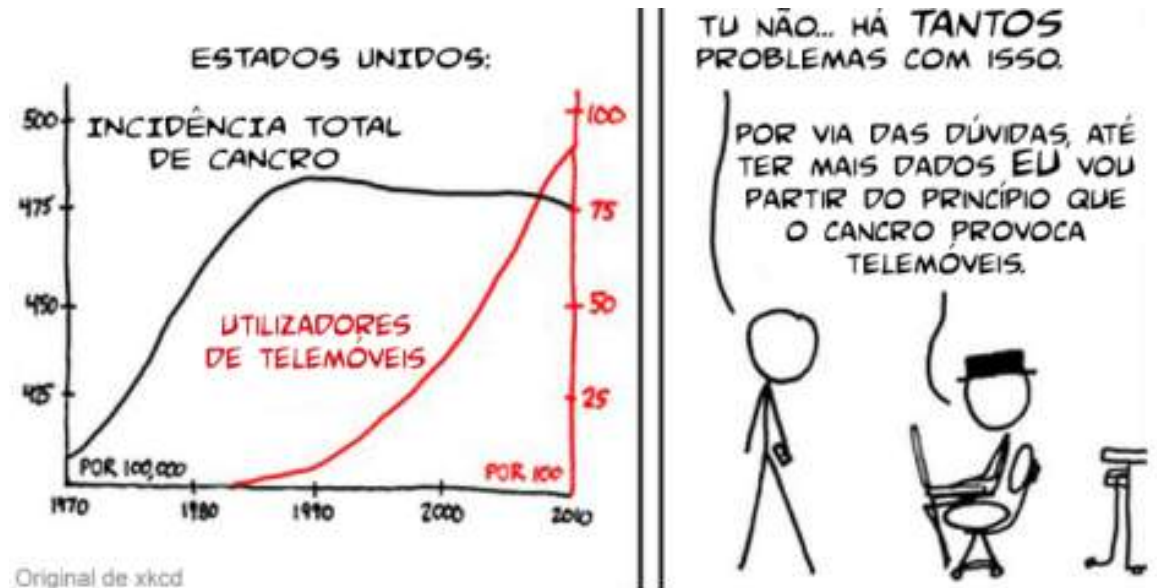


```
starwars_filter %>% ggplot(aes(x = birth_year, y = height)) +  
  geom_point() + geom_smooth(method = 'lm') + xlim(0,100)
```





# Correlação vs. causalidade



# Correlação vs. causalidade



- Relação causal? **Não!**
- Causalidade difícil de estabelecer fora de um ambiente de **experiência controlada**
- Ex.: dar o mesmo telemóvel a várias pessoas idênticas e verificar a incidência de cancro a posteriori

# Correlação vs. causalidade



**Obrigado  
e até amanhã!**

luis.morais@novasbe.pt