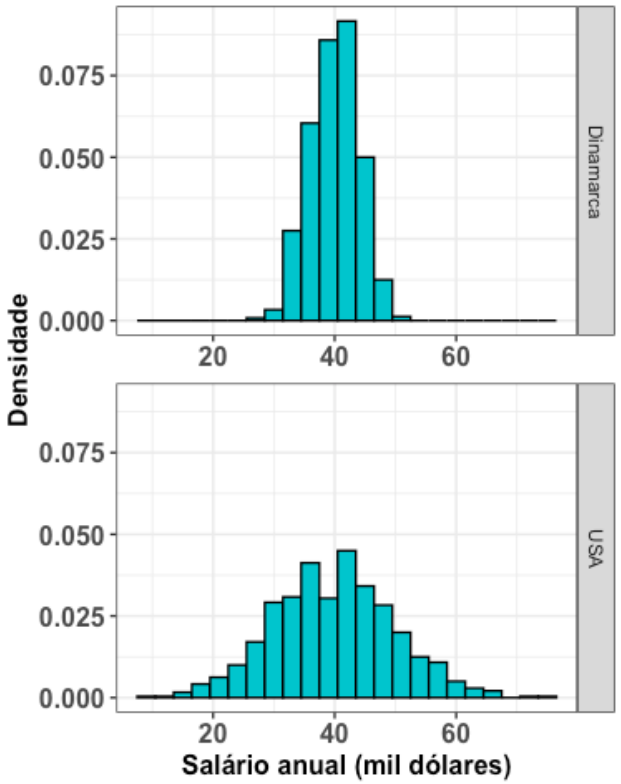
# ME414 - Estatística para Experimentalistas

Parte 3

# Medidas de Dispersão

# Exemplo: Salários de professores de música

Lembram do exemplo dos salários de professores de música na Dinamarca e EUA?




A média dos salários são equivalentes.
Então, para comparar, usamos uma medida de dispersão como, por exemplo, o **desvio padrão**:

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

**Dinamarca**: $\bar{x}$ = 40.02 e $s$ = 3.97
**EUA**: $\bar{x}$ = 39.87 e $s$ = 9.98

# Dispersão dos Dados

Considere dois conjuntos de dados:

$$A = \{1, 2, 3\} \quad \Longrightarrow \quad \overline{x}_A = 2, \quad s_A = 1$$
$$B = \{101, 102, 103\} \quad \Longrightarrow \quad \overline{x}_B = 102, \quad s_B = 1$$

Ambos têm o mesmo desvio padrão.

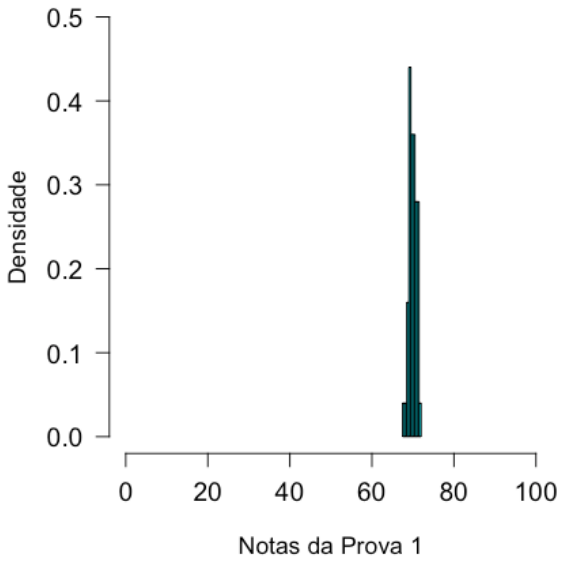
Se compararmos as escalas de cada conjunto de dados, poderíamos dizer que o segundo conjunto tem menor dispersão.

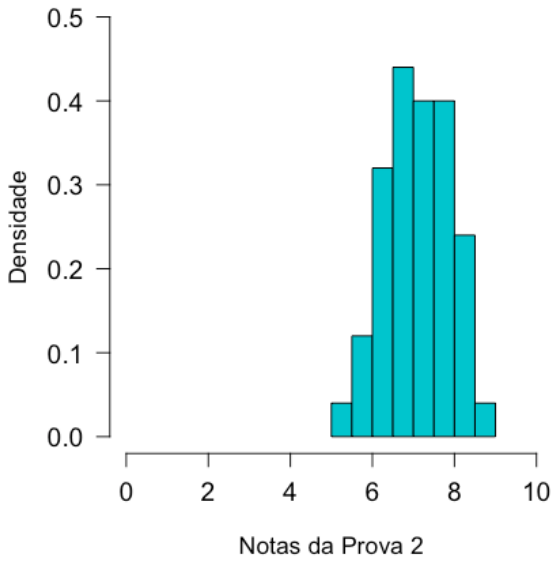
Veja que:

- A maior observação do conjunto $A$, 3, é 3 vezes maior do que a menor observação, 1.
- Já a maior observação do conjunto $B$, 103, é 2% maior do que a menor observação, 101.

# Exemplo: Notas

Considere as notas de 2 provas:




**Prova 1**: Notas de 0 a 100
**Média da turma**: $\bar{x}_1 = 70$
**Desvio padrão**: $s_1 = 1$

**Prova 2**: Notas 0 a 10
**Média da turma**: $\bar{x}_2 = 7$
**Desvio padrão**: $s_2 = 1$

Neste caso, como as escalas são diferentes, não podemos tirar conclusões usando apenas o desvio padrão.

# Coeficiente de Variação

**Coeficiente de variação (CV)**: razão do desvio padrão $s$ pela média $\bar{x}$, isto é

$$CV = \frac{s}{\bar{x}}.$$

**Exemplo:**

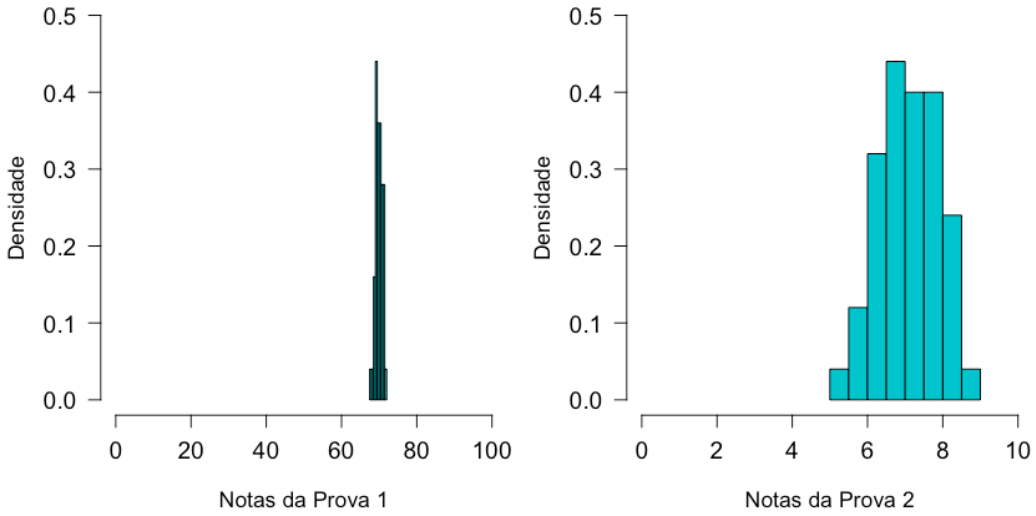
$$\begin{aligned} A = \{1, 2, 3\} &\implies \bar{x}_A = 2, \quad s_A = 1 \\ B = \{101, 102, 103\} &\implies \bar{x}_B = 102, \quad s_B = 1 \end{aligned}$$

Nesse caso,

$$CV_A = \frac{s_A}{\bar{x}_A} = 0.5 \qquad \text{e} \qquad CV_B = \frac{s_B}{\bar{x}_B} = 0.0098.$$

# Coeficiente de Variação

Exemplos das notas de duas provas:




**Prova 1:** $\bar{x}_1 = 70$ e $s_1 = 1$

**Prova 2:** $\bar{x}_2 = 7$ e $s_2 = 1$

**Coeficiente de Variação:** é o desvio padrão escalonado pela média dos dados.

Vamos calcular os CVs para esses dois casos:

$$CV_1 = \frac{s_1}{\bar{x}_1} = 0.014 \qquad \text{e} \qquad CV_2 = \frac{s_2}{\bar{x}_2} \approx 0.14.$$

# Medidas de posição para descrever dispersão

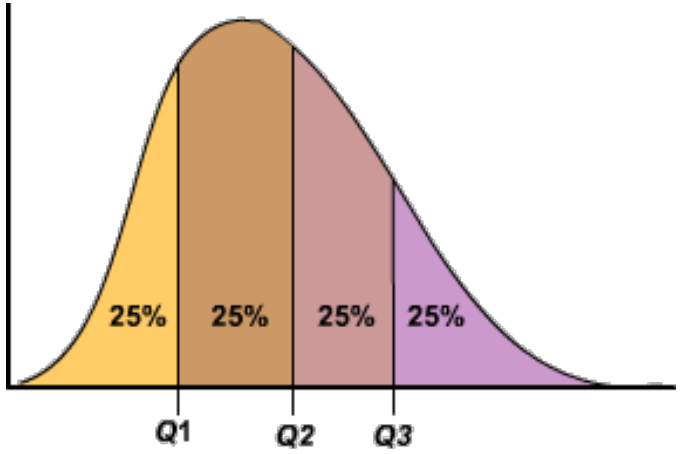
**Média e mediana**: medidas de posição **central**.

**Amplitude e desvio padrão**: medidas de dispersão.

Há outros tipos de medida de posição para descrever a distribuição dos dados: **quartis** e **percentis**.

**Quartis** dividem os dados em 4 partes iguais: primeiro quartil ($Q_1$), segundo quartil ($Q_2$) e o terceiro quartil ($Q_3$).

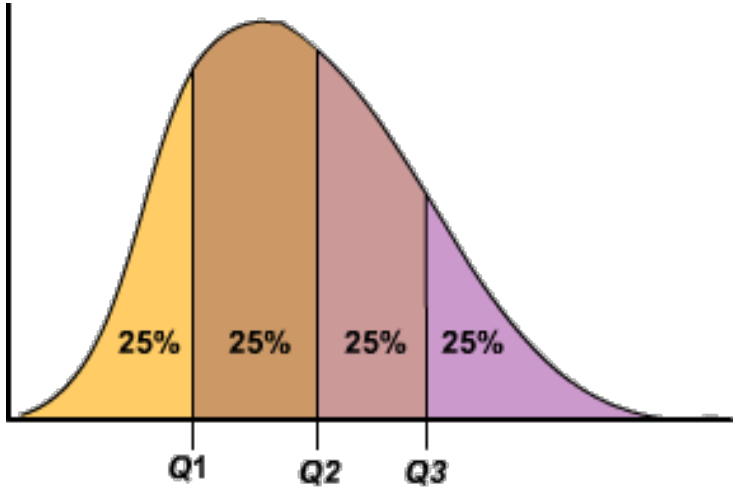
O **p-ésimo percentil** é o valor tal que uma porcentagem **p** dos dados ficam abaixo dele.



# Quartis

Para obter os quartis:

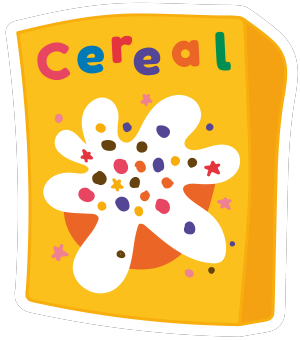
1. Ordene os dados em ordem crescente.
2. Encontre a mediana $Q_2$.
3. Considere o subconjunto de dados abaixo da mediana. $Q_1$ é a mediana deste subconjunto de dados.




4. Considere o subconjunto de dados acima da mediana. $Q_3$ é a mediana deste subconjunto de dados.

# Exemplo: Sódio em cereais matinais

Considere as quantidades de sódio (mg) em 20 cereais matinais:




0, 70, 125, 125, 140, 150, 170, 170, 180, **200**

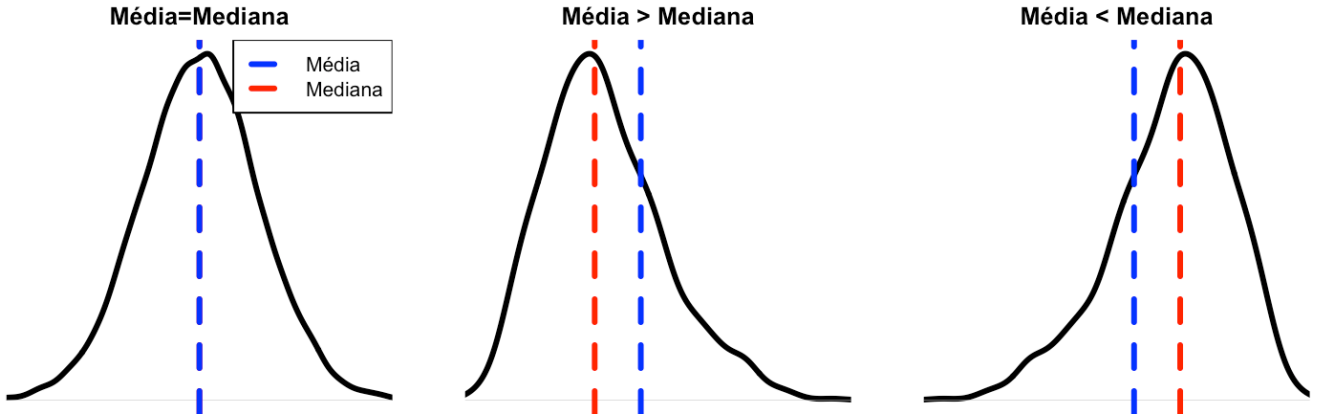
**200**, 210, 210, 220, 220, 230, 250, 260, 290, 290

Para obter $Q1$, calcula-se a mediana considerando apenas as 10 primeiras observações ordenadas: 0, 70, 125, 125, $\underbrace{140, 150}_{Q_1=145}$, 170, 170, 180, **200**

Para obter $Q3$, calcula-se a mediana considerando apenas as 10 últimas observações ordenadas: **200**, 210, 210, 220, $\underbrace{220, 230}_{Q_3=225}$, 250, 260, 290, 290

# Simetria e Assimetria da Distribuição

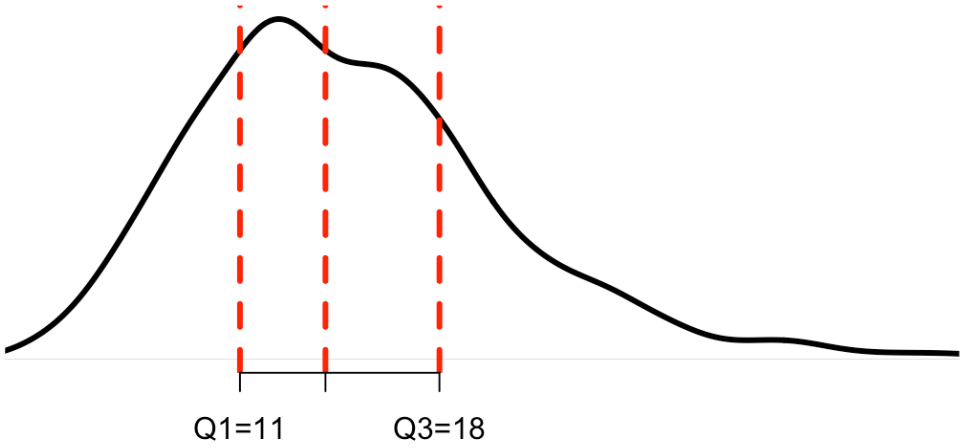
Vimos na aula passada que as posições da média e mediana fornecem informação sobre o formato da distribuição.




Em geral, se a distribuição é:
- **Perfeitamente simétrica**: média = mediana.
- **Assimétrica à direita**: média $>$ mediana.
- **Assimétrico à esquerda**: média $<$ mediana.

# Quartis e Assimetria

Os quartis também fornecem informação sobre o formato da distribuição.



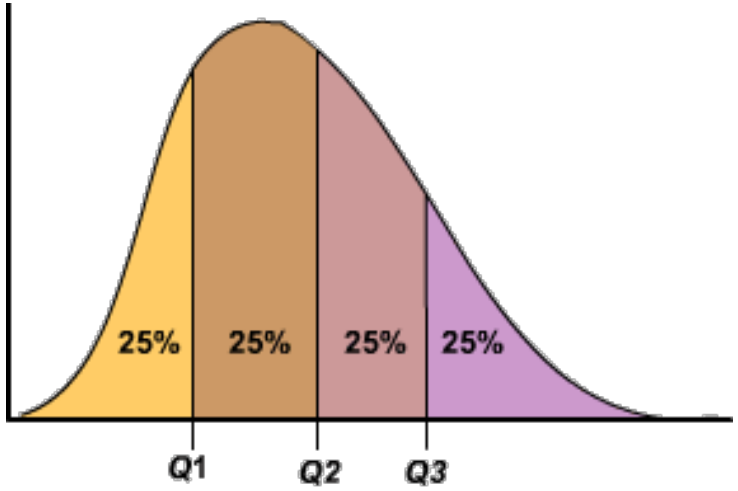

A mediana $Q_2$ é 14.

A distância entre $Q_1$ e $Q_2$ é 3, enquanto que a distância entre $Q_2$ e $Q_3$ é 4, indicando que a distribuição é assimétrica à direita.

# Quartis e simetria da distribuição

Para uma distribuição simétrica ou aproximadamente simétrica:

- $Q_2 - x_{(1)} \approx x_{(n)} - Q_2$
- $Q_2 - Q_1 \approx Q_3 - Q_2$
- $Q_1 - x_{(1)} \approx x_{(n)} - Q_3$




- distâncias entre a mediana e $Q_1, Q_3$ menores do que as distâncias entre os extremos e $Q_1, Q_3$.

# Exemplo: Pesos de alunas de Educação Física

Veja as medidas resumo dos pesos (em libras) de 64 alunas de Educação Física: $\bar{x} = 133$, $Q_1 = 119$, $Q_2 = 131.5$, e $Q_3 = 144$.

Como interpretar os quartis?

- 25% das alunas pesa até 119 libras.
- 25% das alunas pesa mais do que 144 libras.
- 75% das alunas pesa até 144 libras.

Você acredita que a distribuição seja simétrica?

$$Q_2 - Q_1 \approx Q_3 - Q_2 \quad (?)$$

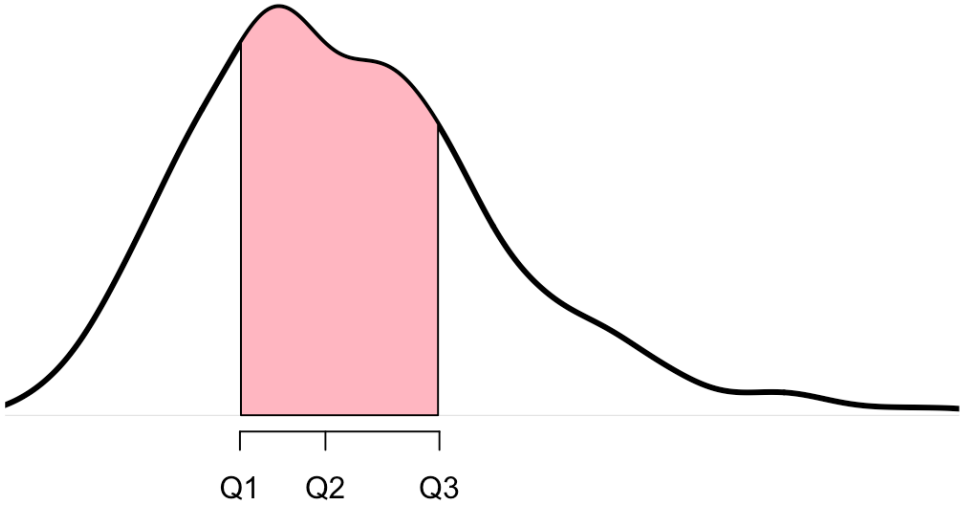
$$\underbrace{Q_2 - Q_1}_{131.5-119=12.5} = \underbrace{Q_3 - Q_2}_{144-131.5=12.5}$$

# Intervalo Interquartílico

A vantagem do uso de quartis sobre o desvio padrão ou a amplitude, é que os quartis são mais resistentes a dados extremos, ou seja, são mais **robustos**.

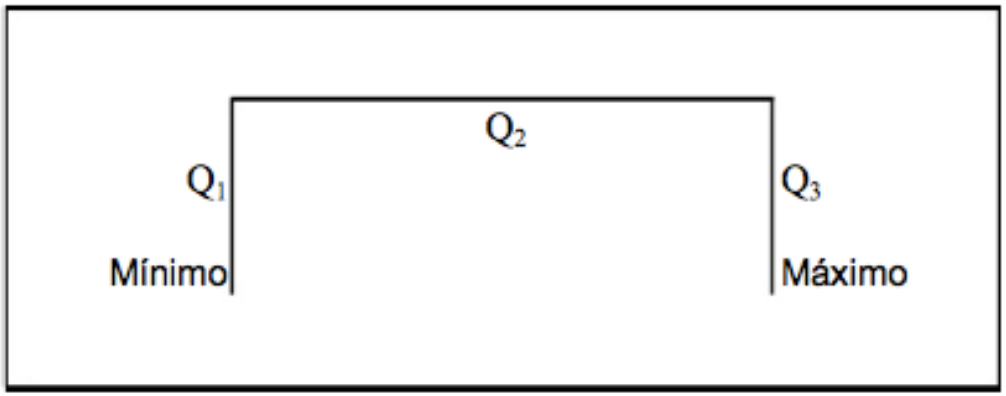
**Intervalo interquartílico (IQ)** = $Q_3 - Q_1$

Representa 50% dos dados localizados na parte central da distribuição.



# Esquema dos 5 números e Boxplot

# Esquema dos 5 números




**Notação:**
$x_{(1)}$: mínimo
$x_{(k)}$: $k$-ésima observação depois de ordenar os dados
$x_{(n)}$: máximo

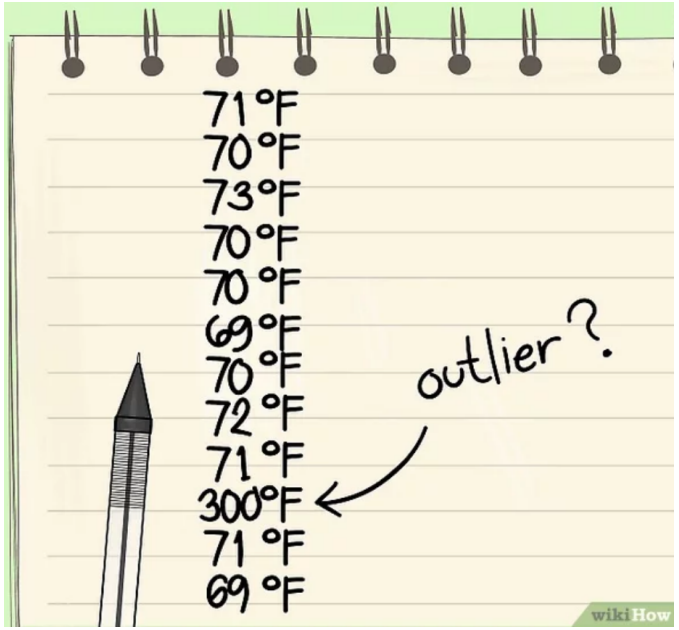
Lembrando que a fórmula da mediana ($Q_2$) é dada por:

$$Q_2 = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

# Dados discrepantes (*Outliers*)

**Importante:** examinar os dados para verificar se há observações discrepantes.

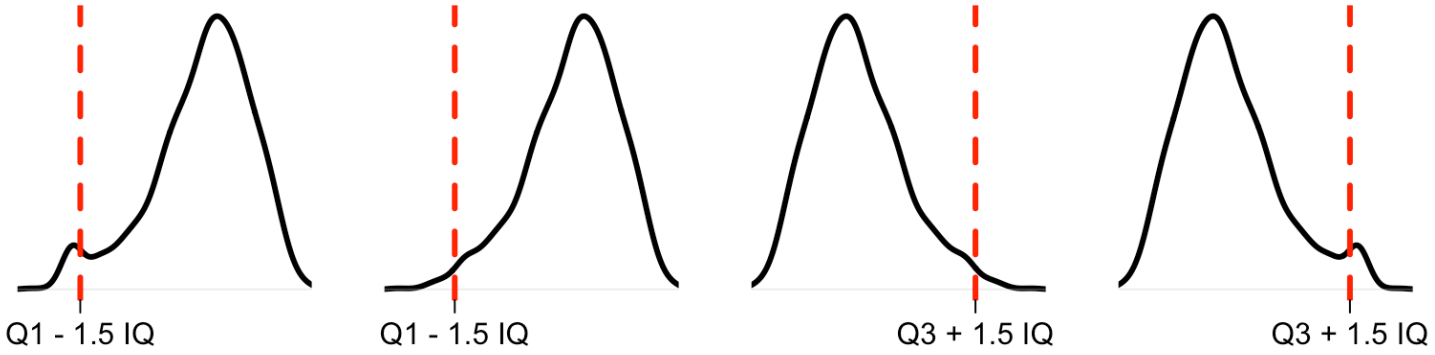
- Média e desvio padrão são muito afetados por observações discrepantes.
- Após detectar a observação discrepante, verificar se não é um erro de digitação ou um caso especial da sua amostra.
- Com poucos dados, podemos detectar um dados discrepante facilmente, apenas observando a sequência ordenada.
- Podemos usar o IQ como um critério mais geral de detecção de dados discrepantes.



# Dados discrepantes (*Outliers*)

Como regra geral, dizemos que uma observação é um potencial *outlier* se está:

- abaixo de $Q_1 - 1.5 \times IQ$ ou
- acima de $Q_3 + 1.5 \times IQ$.

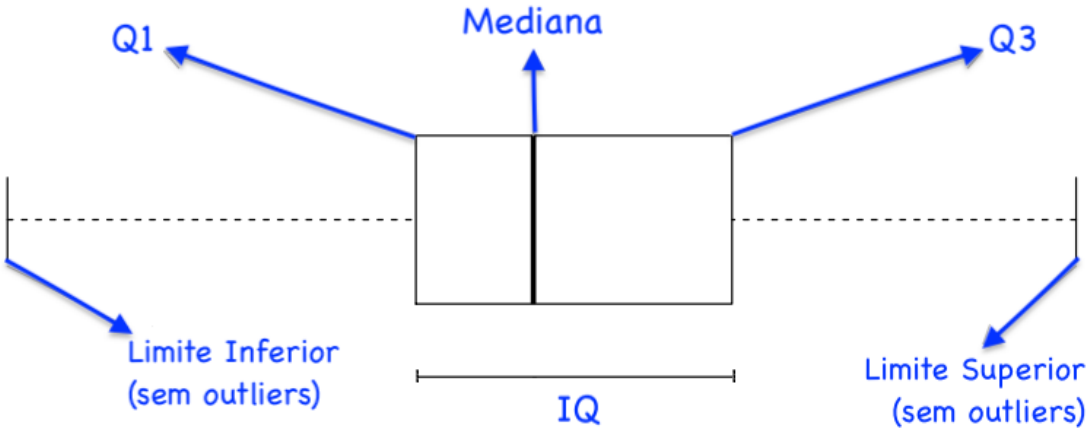



Dizemos *potencial outlier*, pois se a distribuição tem cauda longa, algumas observações irão cair no critério, apesar de não serem *outliers*.

# *Boxplot*

*Boxplot* : representação gráfica do esquema dos 5 números.

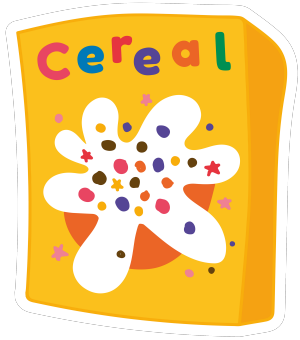
Esse gráfico permite resumir visualmente importante características dos dados (posição, dispersão, assimetria) e identificar a presença de *outliers*.




**ATENÇÃO:** Prestem atenção no que são os limites inferior e superior!!!

# *Boxplot*

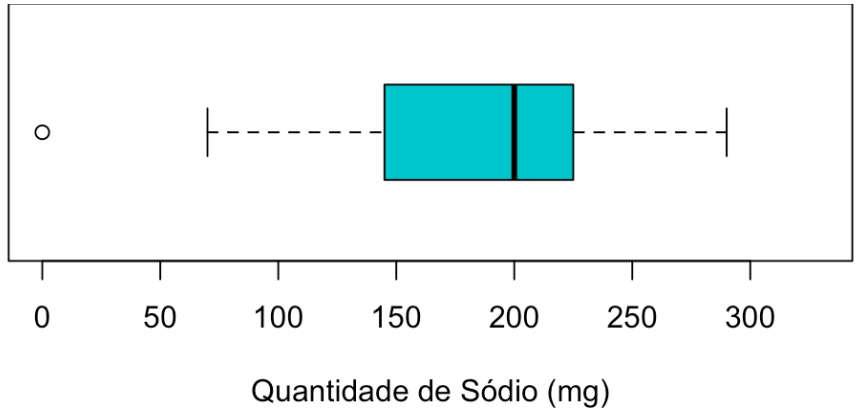
Voltando no exemplo das quantidades de sódio (mg) em 20 cereais matinais:



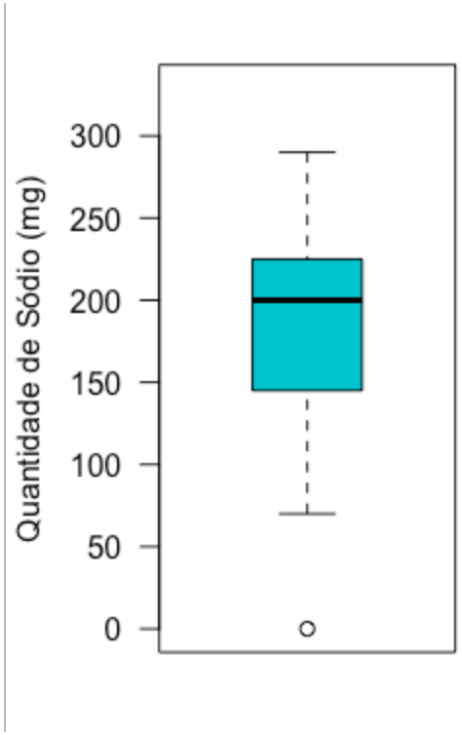

$$0, 70, 125, 125, 140, 150, 170, 170, 180, 200,$$

$$200, 210, 210, 220, 220, 230, 250, 260, 290, 290$$

Já calculamos anteriormente: $Q_2 = 200, Q_1 = 145$ e $Q_3 = 225$.
Esses valores podem ser representados pelo boxplot a seguir:



# Exemplo: Sódio em cereais matinais




Regra para detectar *outliers*:

$$IQ = Q_3 - Q_1 = 225 - 145 = 80$$

$$Q_1 - 1.5 \times IQ = 25 \quad \text{e} \quad Q_3 + 1.5 \times IQ = 345$$

Então, possíveis *outliers* são observações menores que 25 ou maiores que 345.

**Limites Superior e Inferior:** as linhas pontilhadas denotam o mínimo/máximo dos dados que estão na região entre 25 e 345.

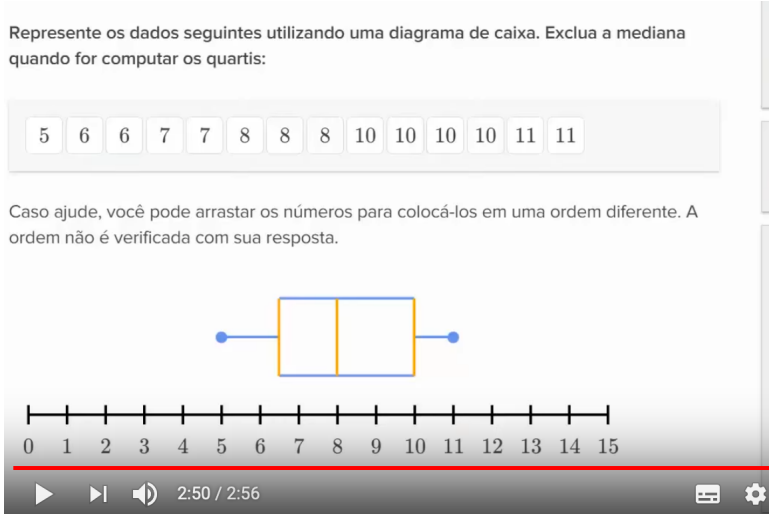
**Limite superior:** a observação máxima dos dados, 290, está no intervalo, então a linha superior vai até 290.

**Limite inferior:** a observação mínima dos dados, 0, está fora do intervalo (outlier=0). Desconsiderando o outlier, o valor mínimo dos dados é 70, que está no intervalo. Portanto, a linha inferior vai até 70.

# Construção de um *Boxplot*

Assista ao vídeo da Khan Academy sobre como criar um boxplot:

https://youtu.be/OanEVzmBD8Y




Vejam e pratiquem com o tutorial:

**Como resumir dados quantitativos**

https://pt.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/stats-box-whisker-plots

# Exemplo: Taxa de desemprego na UE em 2003

| País | Taxa | País | Taxa |
|---|---|---|---|
| **Bélgica** | 8.3 | **Luxemburgo** | 3.9 |
| **Dinamarca** | 6.0 | **Irlanda** | 4.6 |
| **Alemanha** | 9.2 | **Itália** | 8.5 |
| **Grécia** | 9.3 | **Finlândia** | 8.9 |
| **Espanha** | 11.2 | **Áustria** | 4.5 |
| **França** | 9.5 | **Suécia** | 6 |
| **Portugal** | 6.7 | **Reino Unido** | 4.8 |
| **Holanda** | 4.4 | | |

Responda:

1. Qual a amplitude dos dados?
2. Encontre os valores da mediana e de $Q_1$ e $Q_3$?
3. Desenhe um boxplot.

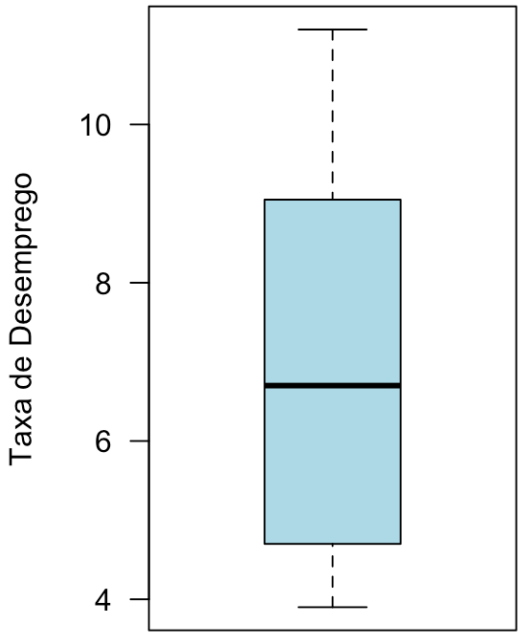
# Exemplo: Taxa de desemprego na UE em 2003

Ordenando os dados:

$$3.9,\ 4.4,\ 4.5,\ \mathbf{4.6},\ 4.8,\ 6.0,\ 6.0,\ \mathbf{6.7},\ 8.3,\ 8.5,\ 8.9,\ \mathbf{9.2},\ 9.3,\ 9.5,\ 11.2$$

Amplitude: 11.2 - 3.9 = 7.3
Mediana = 6.7
$Q_1$ = 4.6 e $Q_3$ = 9.2
$IQ = Q_3 - Q_1$ =4.6
$Q_1 - 1.5 \times IQ$ = -2.3
$Q_3 + 1.5 \times IQ$ =16.1

O mínimo e o máximo pertencem ao intervalo (-2.3, 16.1), portanto as linhas pontilhadas terminam no máximo (11.2) e no mínimo (3.9).



# Exemplo: População dos estados brasileiros

A tabela abaixo apresenta a população (em 1000 habitantes) dos 26 estados brasileiros e o Distrito Federal.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| RR | 325 | MS | 2079 | PB | 3444 | PR | 9564 |
| AP | 478 | MT | 2505 | GO | 5004 | RS | 10188 |
| AC | 558 | RN | 2777 | SC | 5357 | BA | 13071 |
| TO | 1158 | AM | 2813 | MA | 5652 | RJ | 14392 |
| RO | 1380 | AL | 2823 | PA | 6193 | MG | 17892 |
| SE | 1785 | PI | 2844 | CE | 7431 | SP | 37033 |
| DF | 2052 | ES | 3098 | PE | 7919 | | |

Temos 27 estados ($n$ é ímpar).

Portanto, a mediana é
$x_{(\frac{n+1}{2})} = x_{(\frac{27+1}{2})} = x_{(14)} = 3098$ (ES).

A metade inferior dos dados: 13 observações.
A mediana deste subconjunto é $Q_1 = x_{(7)} = 2052$ (DF).

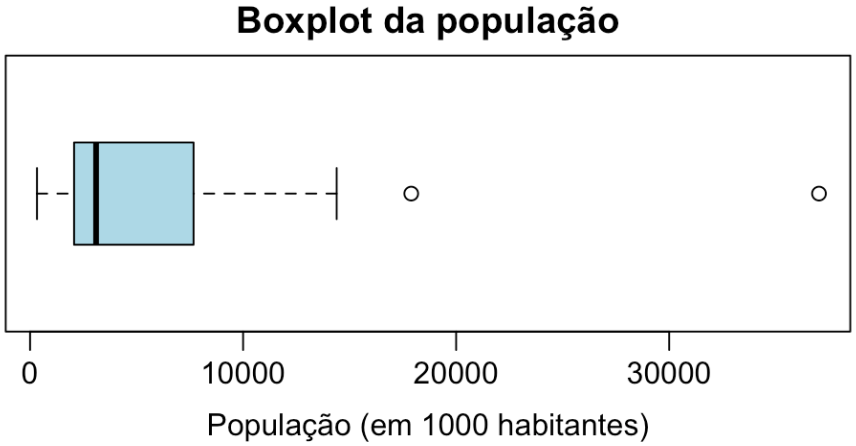
A metade superior dos dados: 13 observações.
A mediana deste subconjunto é $Q_3 = x_{(21)} = 7919$ (PE).

$$IQ = Q_3 - Q_1 = 7919 - 2052 = 5867$$

# Exemplo: População dos estados brasileiros

População (em 1000 habitantes):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| RR | 325 | MS | 2079 | PB | 3444 | PR | 9564 |
| AP | 478 | MT | 2505 | GO | 5004 | RS | 10188 |
| AC | 558 | RN | 2777 | SC | 5357 | BA | 13071 |
| TO | 1158 | AM | 2813 | MA | 5652 | RJ | 14392 |
| RO | 1380 | AL | 2823 | PA | 6193 | MG | 17892 |
| SE | 1785 | PI | 2844 | CE | 7431 | SP | 37033 |
| DF | 2052 | ES | 3098 | PE | 7919 | | |

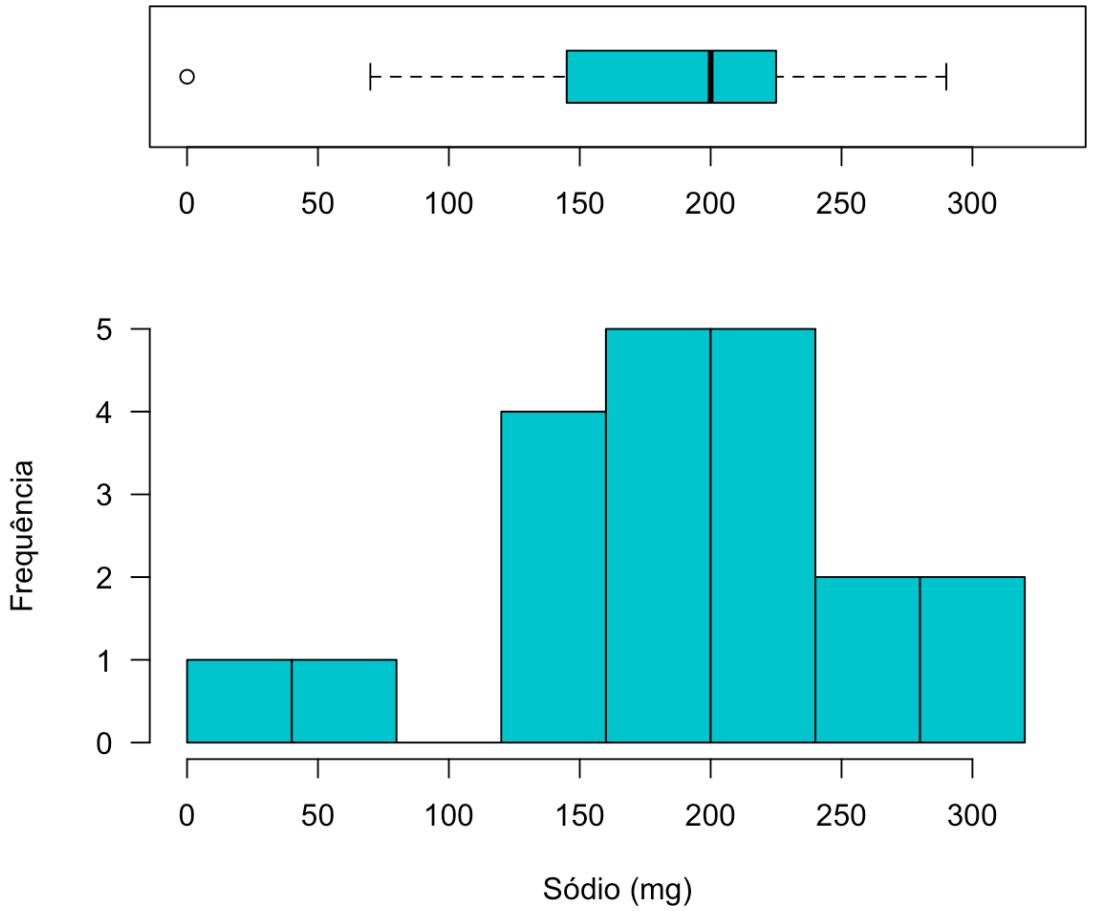



$$Q_1 - 1.5 \times IQ = -6748.5$$

$$Q_3 + 1.5 \times IQ = 16720$$

Temos outliers?

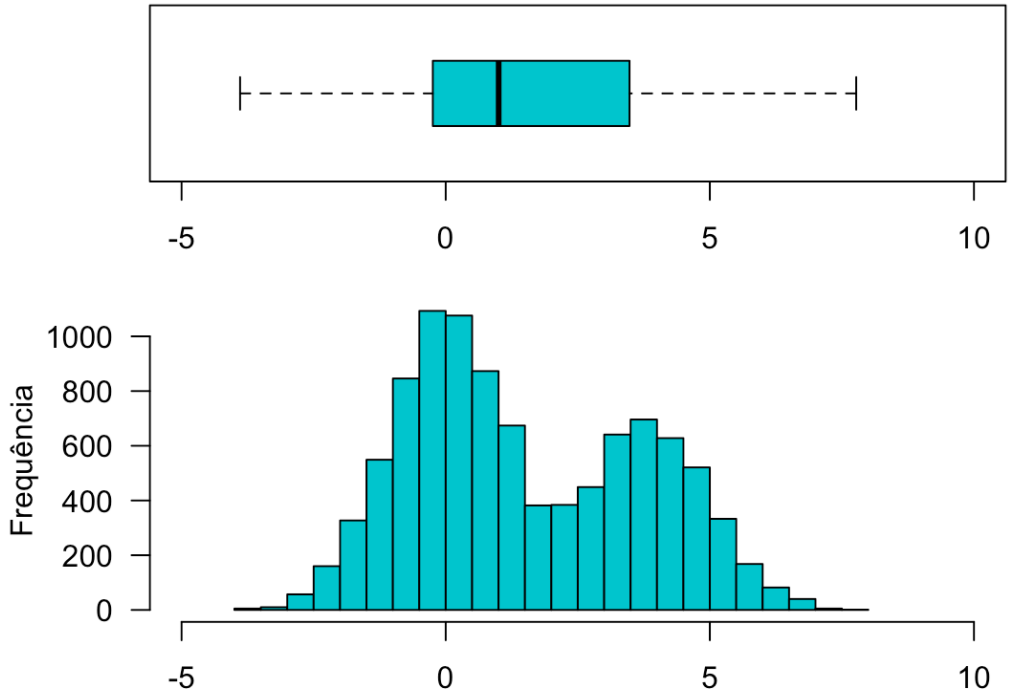
# Exemplo: Sódio em cereais matinais



# *Boxplot* x Histograma

*Boxplot* não substitui o histograma e vice-versa.

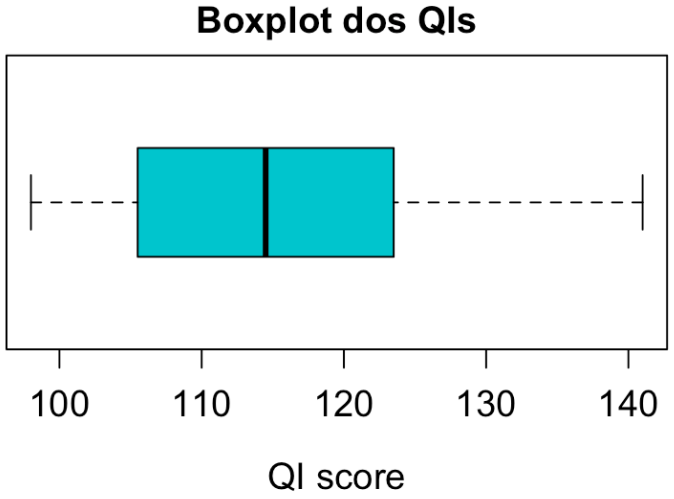
Por exemplo, se a distribuição é bimodal, não observamos isso pelo *boxplot*.

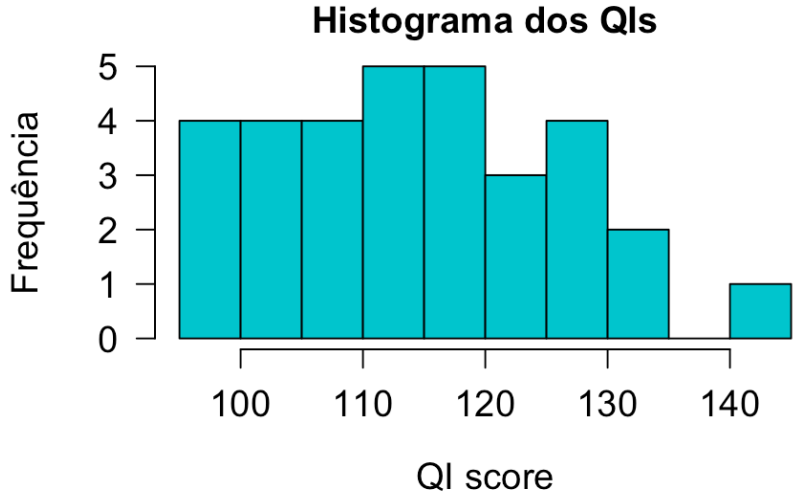


# Exemplo: QI

Para os dados dos QI's das 32 crianças, vamos calcular as medidas resumo, fazer o *boxplot* e histograma.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    98.0   105.5   114.5   115.2   123.5   141.0
```

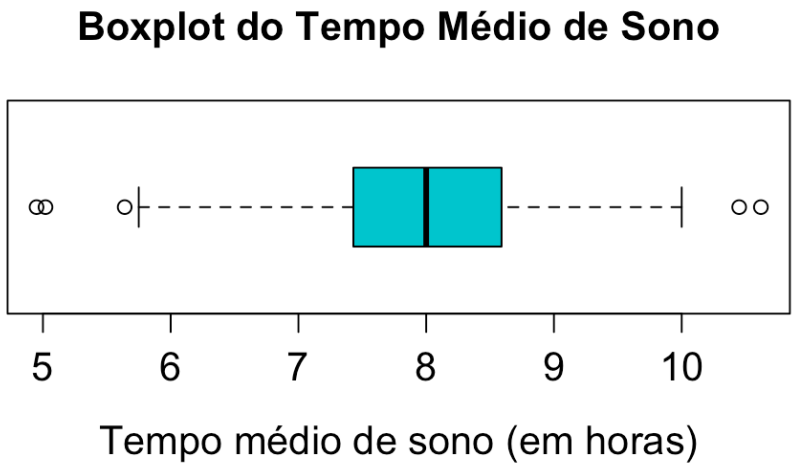


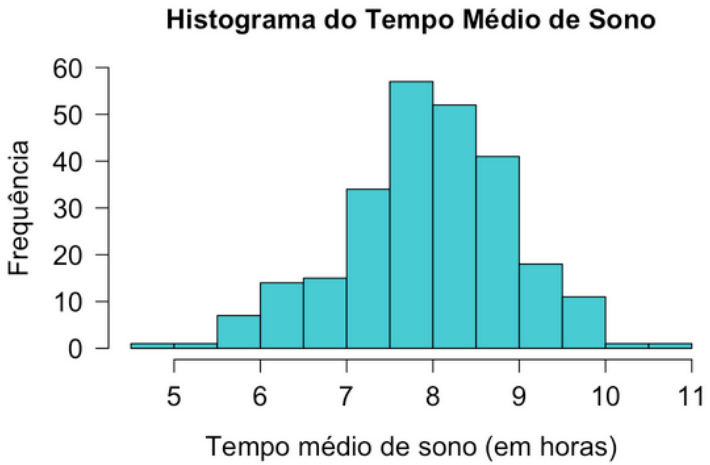



# Exemplo: `SleepStudy`

Vamos obter as medidas resumo e fazer o boxplot da variável `AverageSleep` do `SleepStudy`.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.950   7.430   8.000   7.966   8.590  10.620
```

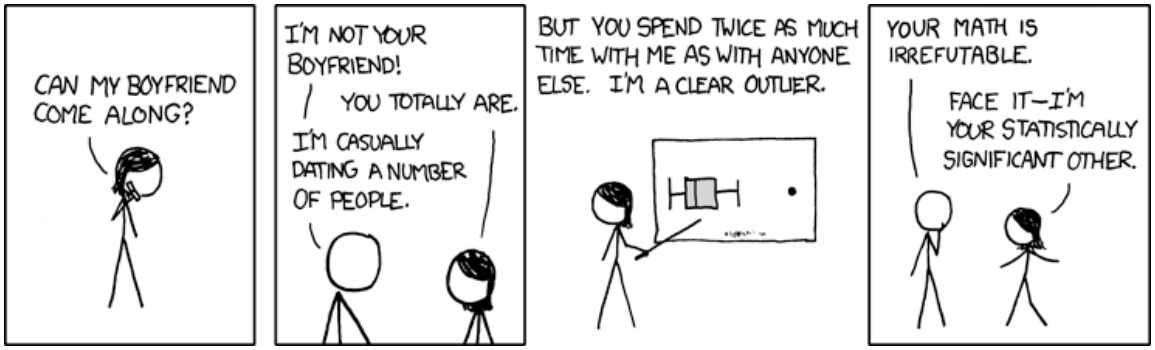


# Agradecimentos

Slides produzidos pelos professores:

- Benilton Carvalho
- Larissa Matos
- Samara Kiihl
- Tatiana Benaglia




https://xkcd.com/539/