# Estatística aplicada à Biologia

## Aula 3 – Medidas de tendência central e Dispersão

# O Dilema da Nutricionista de Capivaras

Vocês são pesquisadores responsáveis pela nutrição de dois grupos de capivaras em cativeiro. O Grupo A recebeu uma nova ração enriquecida e o Grupo B, a ração padrão.

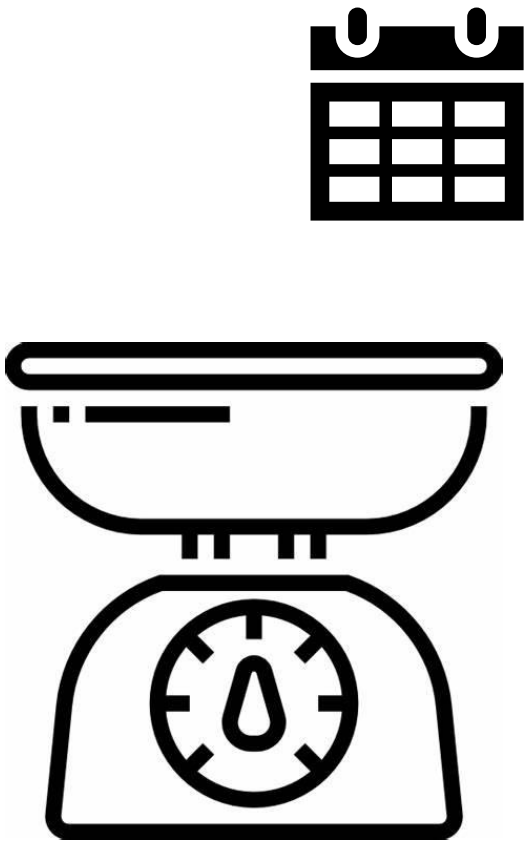
**Grupo A**

**Grupo B**

# Após um mês, vocês pesaram os animais. Os dados (em kg) foram:

**Grupo A**

50, 51, 52, 58, 59



**Grupo B**

35, 52, 53, 65, 65

# Se você pudesse usar uma só palavra ou número para descrever o peso 'típico' de cada grupo, qual seria?

**Observando os dados, qual grupo parece mais 'confiável' ou 'previsível' em seus pesos? Ou seja, em qual grupo os pesos dos animais são mais parecidos entre si?**

# Medidas resumo

- Servem para resumir/sumarizar mais ainda um conjunto de dados, **usando um único número**
- Apontam característica específicas dos dados
- Permitem entender o comportamento geral dos dados
- Medidas de posição (tendência central)
- Medidas de variabilidade (dispersão)

# Sequência numérica de dados

$$x_1, x_2, x_3 \ldots x_i \ldots x_n$$

**Exemplo 3.1 Representação de dados**

Os pesos, em quilogramas, de cinco recém-nascidos são:

| | | | | |
|---|---|---|---|---|
| 3,500 | 2,750 | 3,250 | 2,250 | 3,750 |

Em termos de símbolos, podemos escrever:

$$x_1 = 3,500; \quad x_2 = 2,750; \quad x_3 = 3,250; \quad x_4 = 2,250; \quad x_5 = 3,750.$$

Vieira 2015

# Somatória

$$x_1 + x_2 + x_3 + \ldots + x_n \qquad \sum_{i=1}^{n} x_i.$$

**Exemplo 3.2 Notação de somatório**

No Exemplo 3.1, são dados os pesos de cinco bebês:
$x_1 = 3{,}500$; $x_2 = 2{,}750$; $x_3 = 3{,}250$; $x_4 = 2{,}250$; $x_5 = 3,\ 750$

A soma desses pesos, usando a notação de somatório, fica como segue:

$$\sum_{i=1}^{5} x_i = 3,500 + 2,750 + 3,250 + 2,250 + 3,750 = 15,500$$

Vieira 2015

# Medidas de tendência central (ou de posição)

- **Média aritmética** => centro de equilíbrio do conjunto de dados

$$\overline{x} = \frac{\sum x}{n}$$

- **Mediana** => valor central do conjunto dos dados ordenados (quantil de 50%)
  - Seu cálculo exige ordenar o conjunto de dados de maneira crescente
  - Divide o conjunto de dados em 2 grupos: metade dos dados tem valor maior e metade valor menor do que a mediana
  - Para conjunto de **dados com números ímpar** de objetos, **a mediana é o próprio valor médio**, para **dados pares**, a **mediana é a média dos valores na posição central**
  - Pode ser mais **útil** que a média **para descrever conjuntos de dados com valores discrepantes** (outliers)

# Mediana

## Exemplo 3.6 Cálculo da mediana

Para obter a mediana do peso dos cinco bebês do Exemplo 3.1, coloque os dados em ordem crescente, como segue:

2,250; 2,850; 3,250; 3,500; 3,970

A mediana está no centro dos dados ordenados. Corresponde a 3,250 kg, mostrado na Figura 3.2.

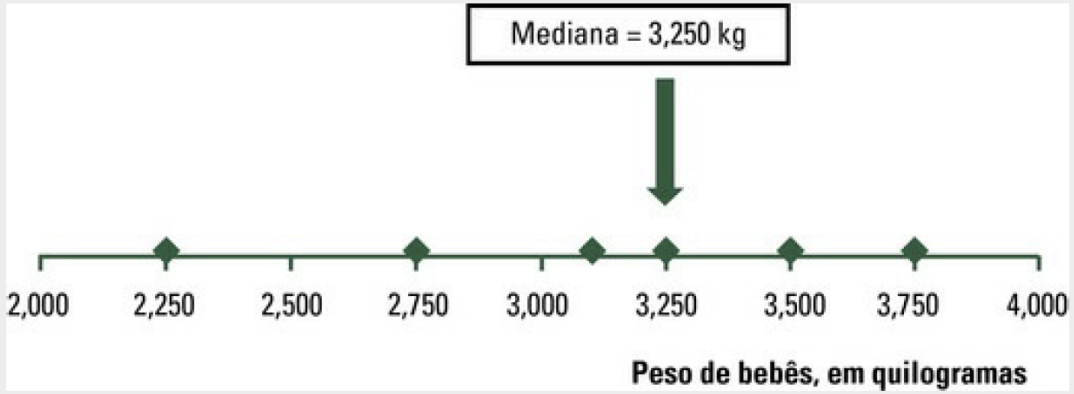



**FIGURA 3.2** Distribuição dos pesos de bebês em quilogramas, sobre um eixo, e a respectiva mediana

Vieira 2015

# Moda

- Valor que ocorre com maior frequência
- Única medida de tendência central que também pode ser usada para descrever dados *qualitativos*.
  - Nesse caso, a moda é a categoria da variável que ocorre com maior frequência.

**Exemplo 3.8 Determinando a moda**

A moda dos dados 0, 0, 2, 5, 3, 7, 4, 7, 8, 7, 9, 6 é 7, porque é o valor que ocorre maior número de vezes.

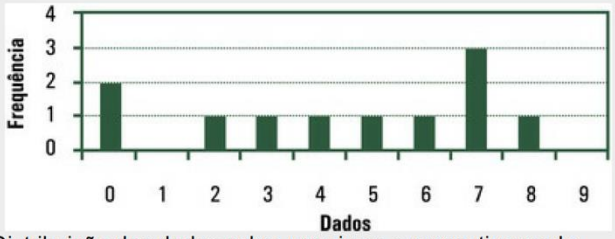



FIGURA 3.3 Distribuição dos dados sobre um eixo e a respectiva moda

**Exemplo 3.10 Determinação da moda**

Veja os dados apresentados na Tabela 3.8. O grupo sanguíneo O ocorreu com maior frequência, então é a moda.

**Tabela 3.8**

**Distribuição de indivíduos segundo o grupo sanguíneo**

| Grupo sanguíneo | Frequência |
|---|---|
| O | 550 |
| A | 456 |
| B | 132 |
| AB | 29 |
| Total | 1.167 |

Vieira 2015

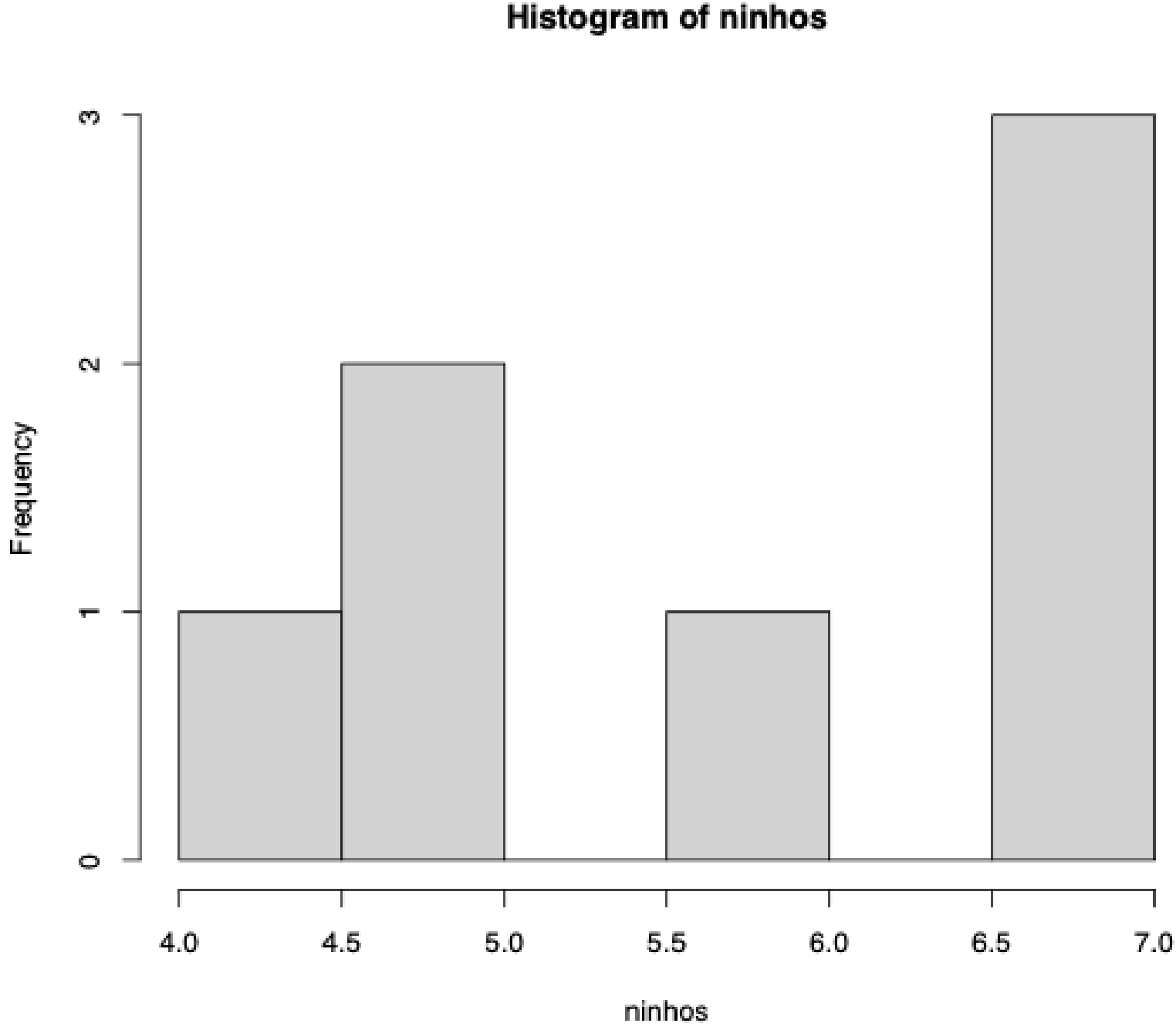
| Tipo da medida | Medida | Simbologia | |
|---|---|---|---|
| | | Amostra (n) | População (N) |
| Posição | Média | $\overline{x}$ | $\mu$ |
| | Mediana | md | Md |
| | Moda | mo | Mo |
| Variabilidade | Variância | $s^2$ | $\sigma^2$ |
| | Desvio Padrão | s | $\sigma$ |
| | Coeficiente de Variação | cv | CV |

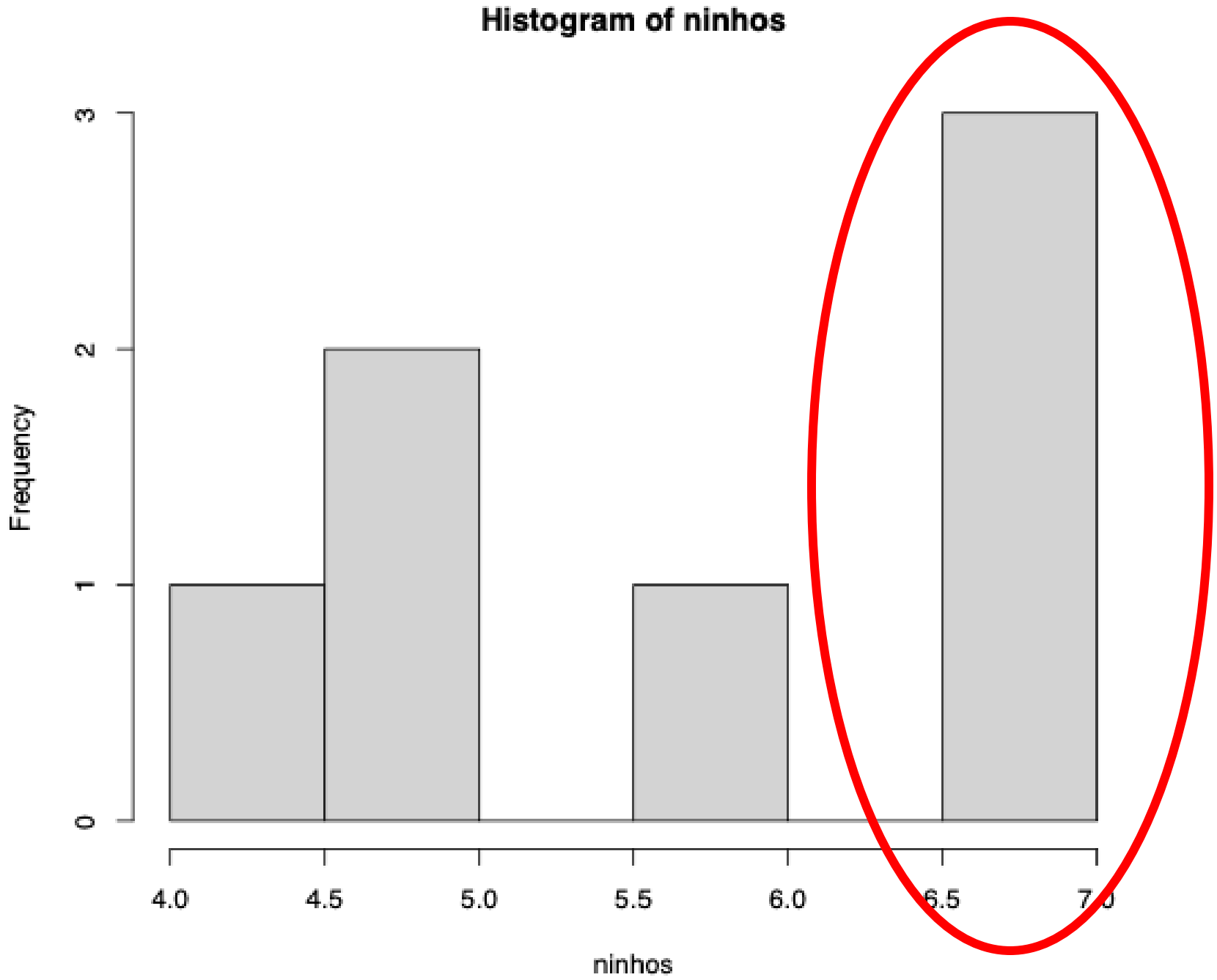
# Atividade: o efeito do mutante

Número de ovos em 7 ninhos de uma espécie de ave: **[4, 5, 5, 6, 7, 7, 7]**


Façam duplas, calculem a:

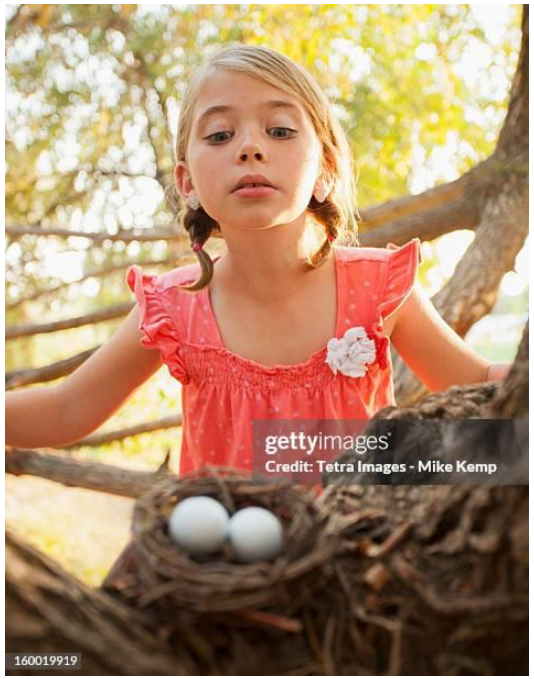
1) média,
2) mediana e
3) moda


Histogram of ninhos
Frequency
0
1
2
3
4.0
4.5
5.0
5.5
6.0
6.5
7.0
ninhos


Histogram of ninhos
Frequency
0
1
2
3
4.0
4.5
5.0
5.5
6.0
6.5
7.0
ninhos

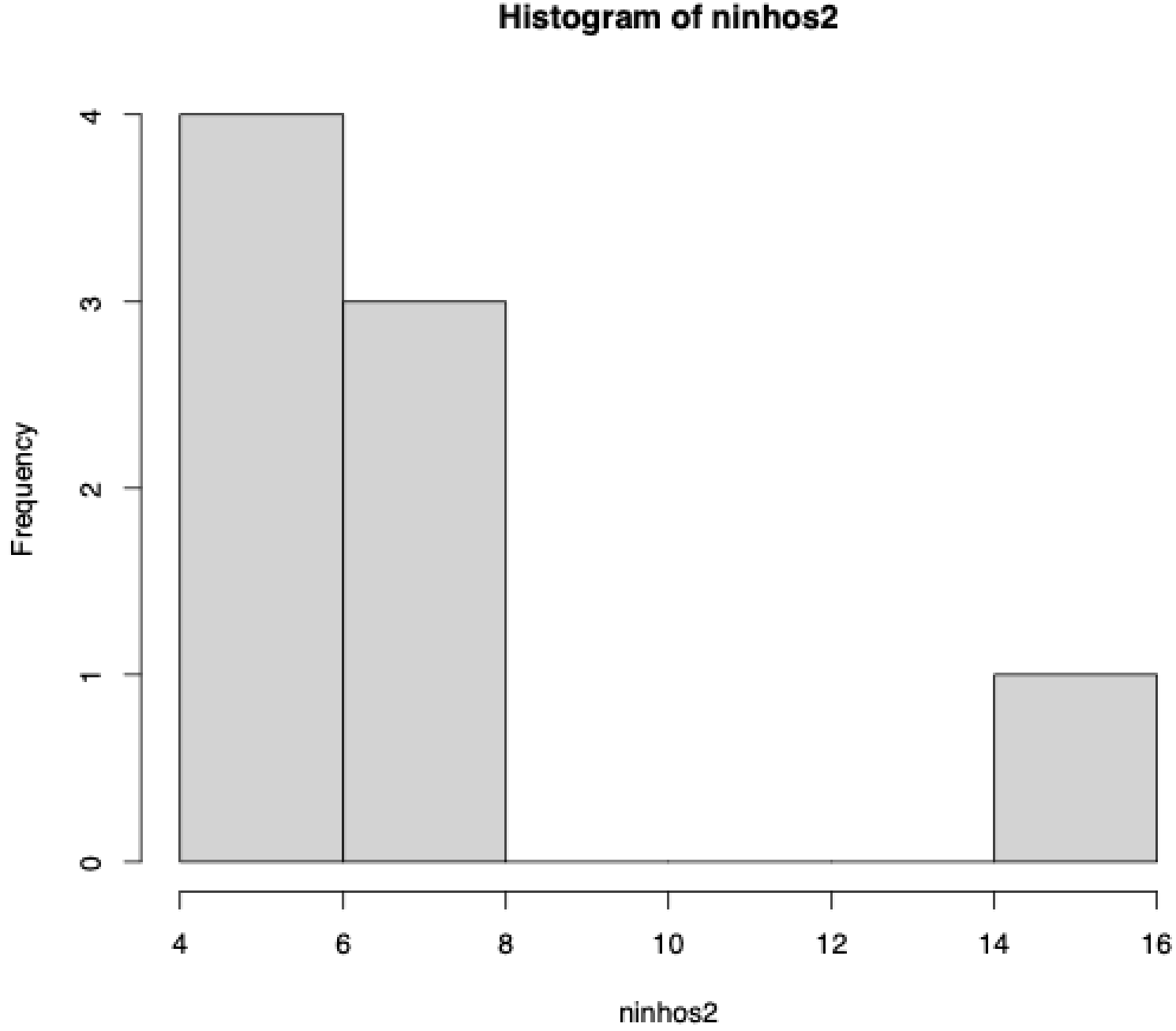
# Atividade : o efeito do mutante

Agora, imagine que encontramos um oitavo ninho, de uma ave que botou **15 ovos**! Um valor extremo, um 'mutante'.
O novo conjunto de dados é: **[4, 5, 5, 6, 7, 7, 7, 15]**




Refaçam as duplas e calculem a:
1) Média e
2) mediana

# Qual das duas medidas (média ou mediana) foi mais drasticamente afetada pelo ninho 'mutante'?


Histogram of ninhos2
Frequency
0
1
2
3
4
4
6
8
10
12
14
16
ninhos2

*A média é mais sensível a valores extremos (outliers) do que a mediana*

Quando um biólogo poderia preferir usar uma ou outra?

# Medidas de variabilidade (ou de dispersão)

- As medidas de posição não informam nada sobre a **variabilidade** dos dados
  - Podem esconder/mascarar informação
  - São mais úteis quando há pouca variação nos dados

- Analisar desvios/dispersão em relação à média

Grupo A (variável $X$): 3, 4, 5, 6, 7.
Grupo B (variável $Y$): 1, 3, 5, 7, 9.
Grupo C (variável $Z$): 5, 5, 5, 5, 5.
Grupo D (variável W): 3, 5, 5, 7.
Grupo E (variável V): 3, 5, 5, 6, 6.

Vemos que $\overline{x} = \overline{y} = \overline{z} = \overline{w} = \overline{v} = 5,0$ .

# Tipos de medidas de variabilidade

- Mínimo => menor valor do conjunto de dados
- Máximo => maior valor do conjunto de dados
- Amplitude => diferença entre o máximo e o mínimo
  - Muito sensível a valores discrepantes
- Quartil => divide os dados em quatro partes (25%)
  - noção da assimetria da distribuição.
  - Distância interquartílica: distância entre o primeiro e o terceiro quartis

$q(0,25) = q_1$ : 1º Quartil = 25º Percentil
$q(0,50) = q_2$ : Mediana = 2º Quartil = 50º Percentil
$q(0,75) = q_3$ : 3º Quartil = 75º Percentil
$q(0,40)$ : 4º Decil
$q(0,95)$ : 95º Percentil

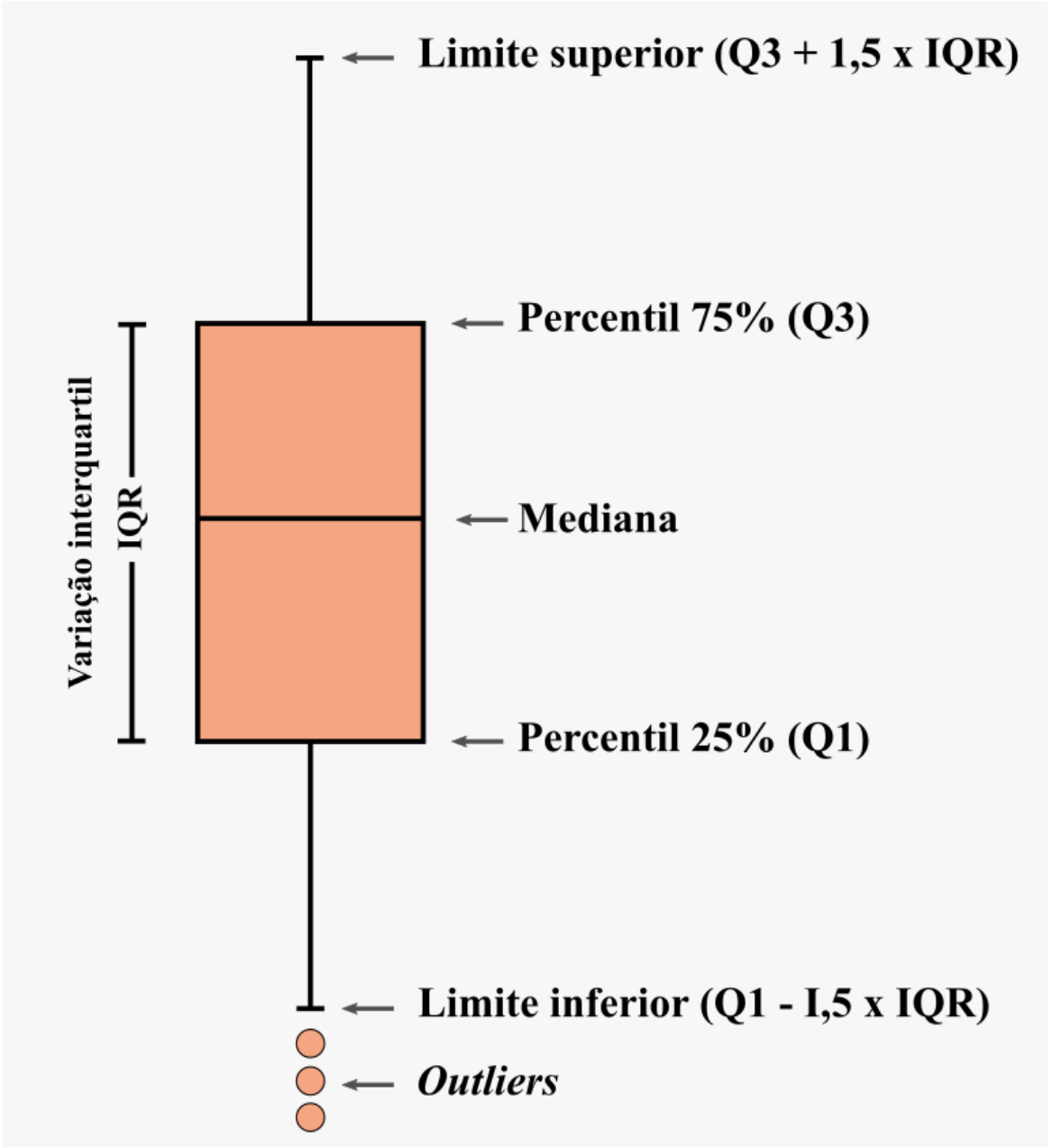
Morettin & Bussab 2024

# Aprendendo a desenhar um boxplot

1. Ordene os dados do maior para o menor
2. Calcule a mediana
3. Separe os dados cujos valores estão **abaixo** da mediana
4. Calcule a mediana destes dados, este é o Q1
5. Separe os dados cujos valores estão **acima** da mediana
6. Calcule a mediana destes dados, este é o Q3
7. Calcule a Variação Interquartil = Q3-Q1
8. Desenhe uma caixa ligando o Q1 e Q3.
9. Coloque uma linha dentro da caixa indicando a mediana
10. Desenhe uma reta da extremidade da caixa até 1,5 * Q1. Repita o mesmo procedimento para Q3.

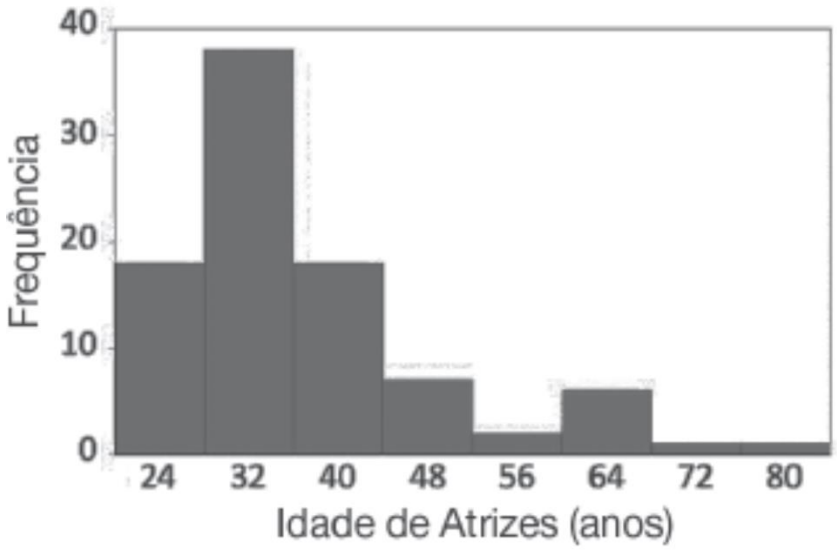
# Gráfico boxplot

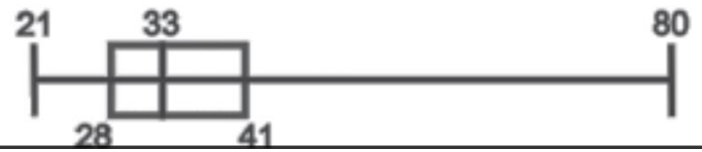
**O box plot dá uma ideia da posição, dispersão, assimetria, caudas e dados discrepantes.**



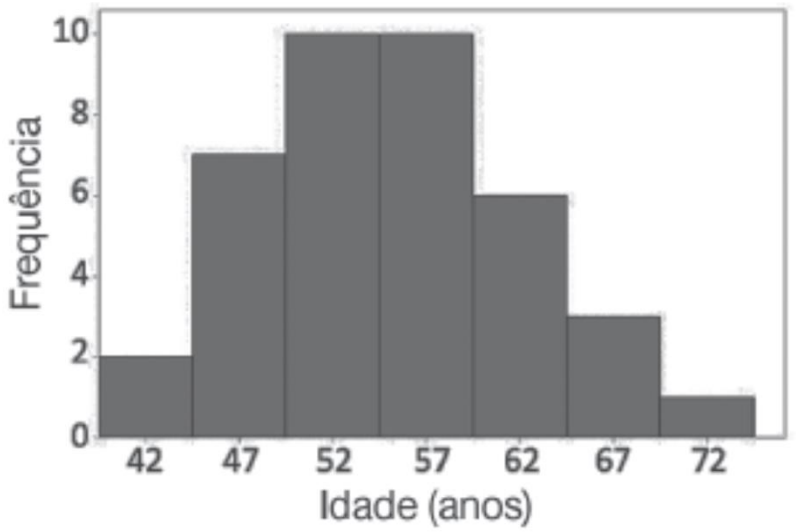

Da Silva et al.
Análises ecológicas

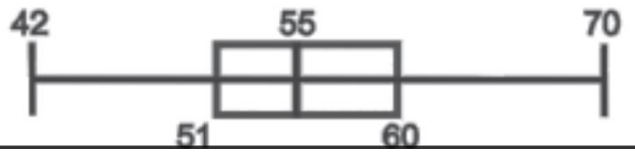
**Distribuição Assimétrica**
(Idades de Atrizes ao Ganharem o Oscar)



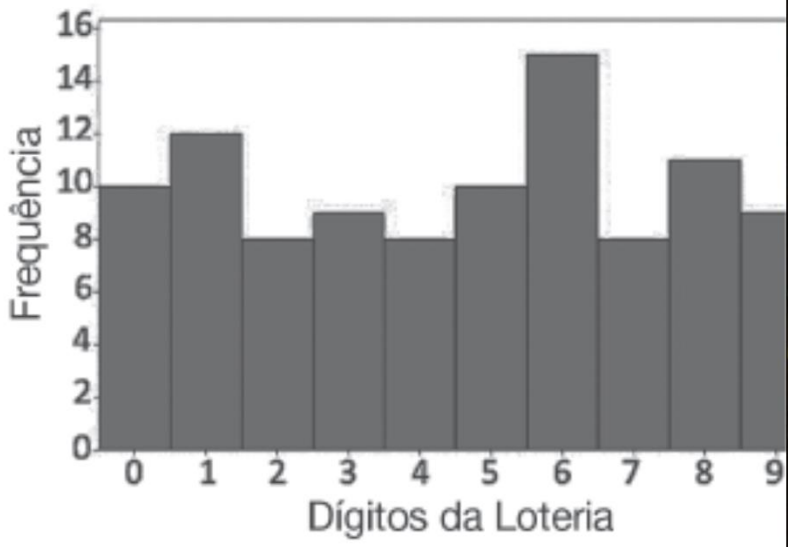



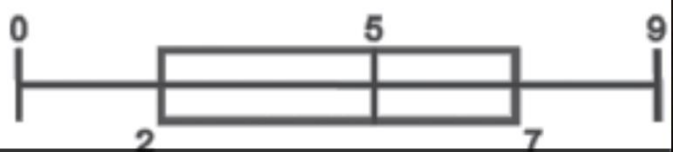

**Distribuição Normal**
(Idades de Presidentes na Primeira Posse)



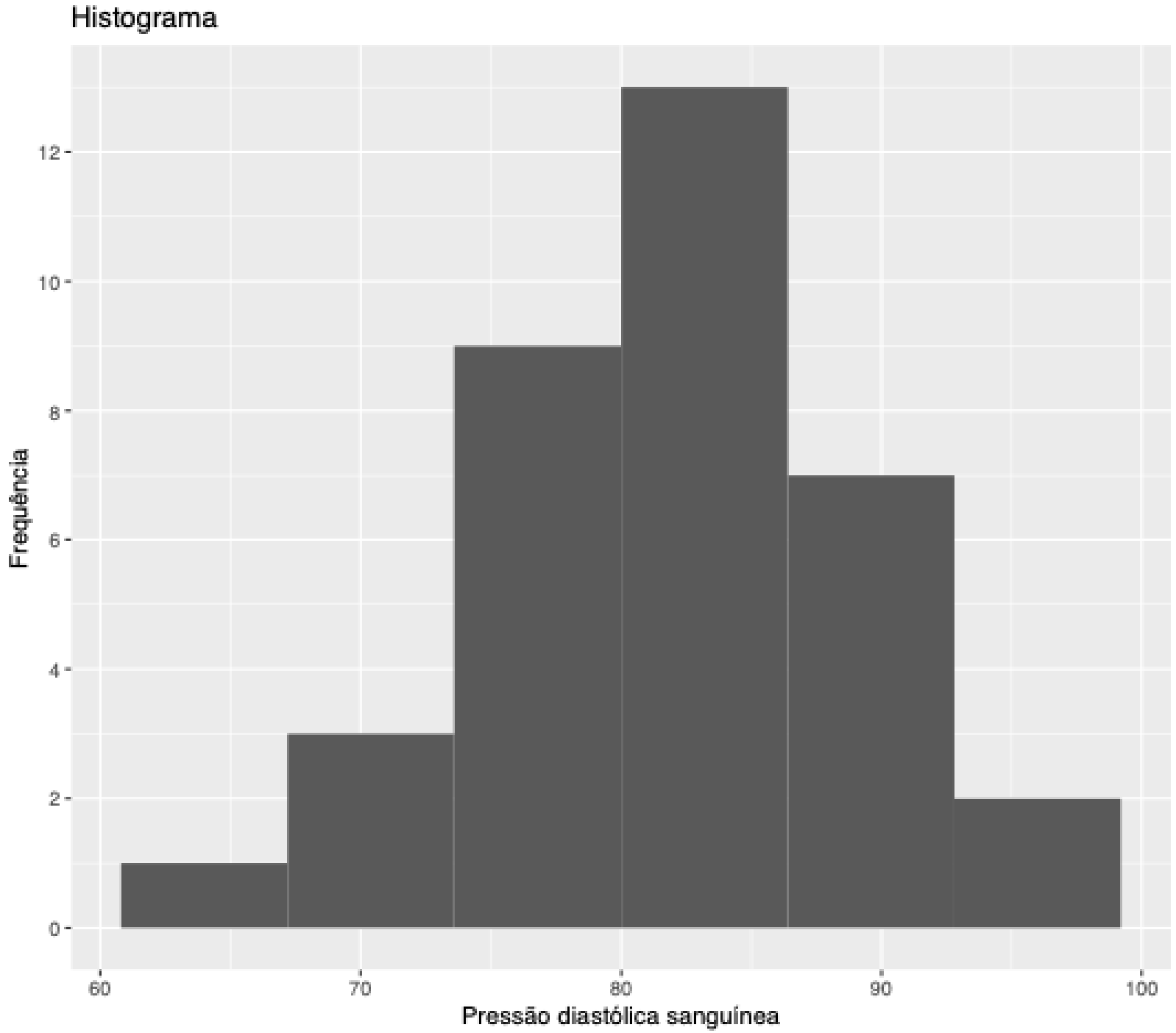

**Distribuição Uniforme**
(Dígitos nos Sorteios da Loteria)

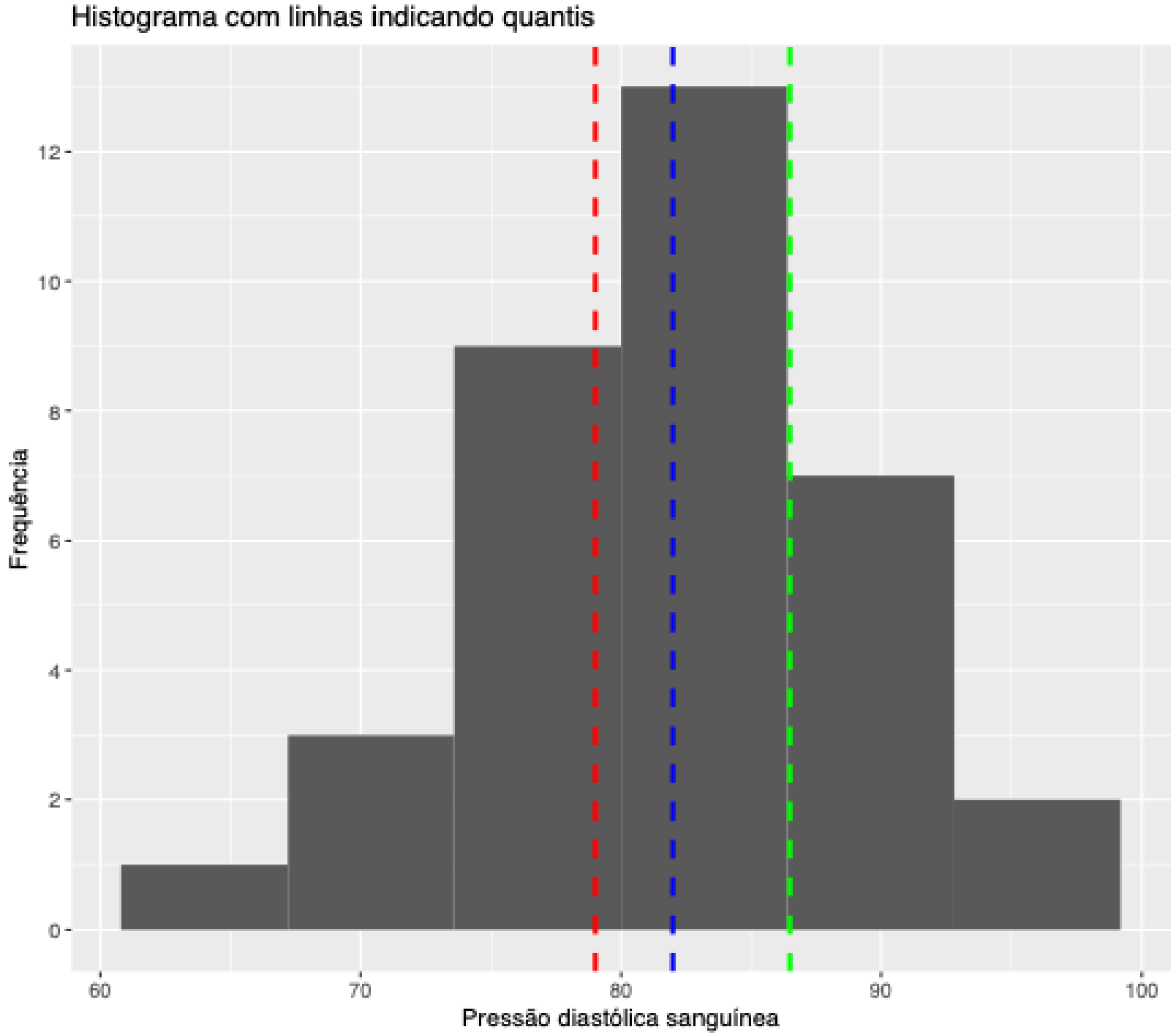


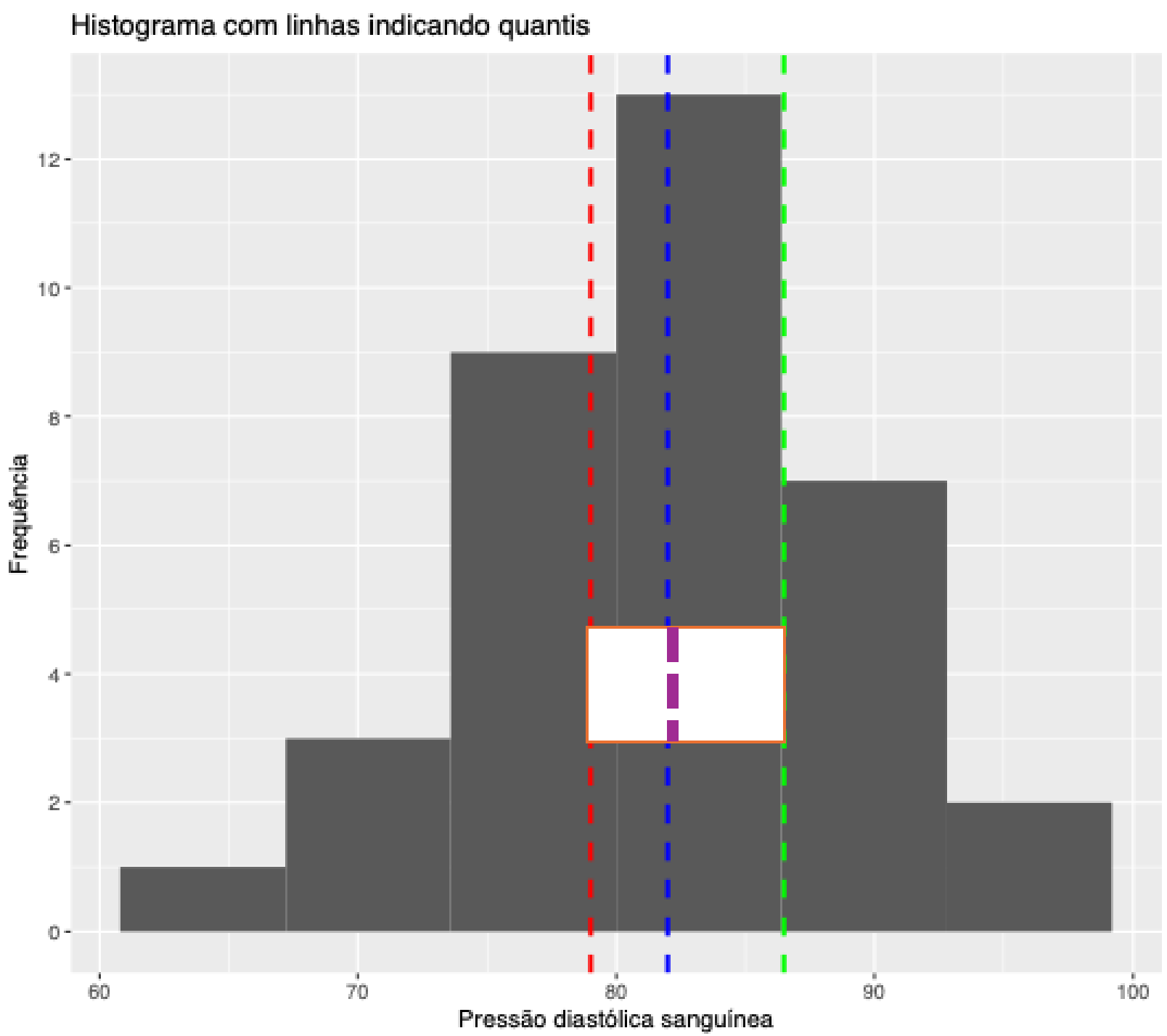

Triola 2024

Exemplo da aula passada






```
> quantiles
  25%   50%   75%
79.0  82.0  86.5
```




```
> quantiles
  25%   50%   75%
79.0  82.0  86.5
```

# Medidas de variabilidade

- Desvios da média

$$\text{desvio} = \text{dado} - \text{m é dia}$$

$$d_i = x_i - \overline{x}$$

- Desvios pequenos significam dados aglomerados, enquanto desvios grandes significam dados dispersos em torno da média

# Desvios da média

## Exemplo 4.7 Desvios em relação à média

No Exemplo 4.1, são dadas as idades de cinco crianças: 3, 6, 5, 7 e 9 anos. Para calcular os desvios em relação à média, subtraímos a média de cada observação. Como a média é 6, os desvios são os valores apresentados na Tabela 4.2.

**Tabela 4.2**

**Cálculo dos desvios**

| Observação | Desvio |
|---|---|
| x | 6 |
| 3 | 3 – 6 = –3 |
| 6 | 6 – 6 = 0 |
| 5 | 5 – 6 = –1 |
| 7 | 7 – 6 = 1 |
| 9 | 9 – 6 = 3 |

Vieira 2015

# Variância e soma dos quadrados

- O cálculo da variância nos fornece um conjunto de números (vetor). Mas precisamos de um único número que resuma a variabilidade dos dados
- Como fazer isso? Somando os desvios? Vejamos:

$$-3 + 0 - 1 + 1 + 3 = 0$$

- Não funciona porque os desvios positivos anulam os negativos
- Mas ao elevarmos os valores ao quadrado os sinais “desaparecem”

# Variância e soma dos quadrados

Variância da amostra é a soma dos quadrados dos desvios de cada observação em relação à média, dividida por $(n - 1)$.

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

Soma dos quadrados

Graus de liberdade

Para calcular a variância:

- Calcule a média
- Calcule o desvio de cada observação em relação à média
- Eleve cada desvio ao quadrado
- Some os quadrados dos desvios
- Divida o resultado por $n - 1$ ($n$ é o número de observações).

Vieira 2015

## Cálculo da variância

| Observação $x$ | Desvio | Desvio ao quadrado |
|---|---|---|
| 3 | $3-6=-3$ | $(-3)^2=9$ |
| 6 | $6-6=0$ | $0^2=0$ |
| 5 | $5-6=-1$ | $(-1)^2=1$ |
| 7 | $7-6=1$ | $1^2=1$ |
| 9 | $9-6=3$ | $3^2=9$ |
| $\Sigma x=30$ | $\Sigma(x-x)=0$ | |

A variância é

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

$$s^2=\frac{20}{4}=5$$

Vieira 2015

# Desvio padrão da amostra

- Medida de variabilidade **na mesma escala** (unidade de medida) **dos dados**
- O desvio-padrão é uma medida de **quanto os valores** de dados **se afastam da média**
- Não é um bom estimador para o DP da população

Desvio padrão é a raiz quadrada da variância, com sinal positivo.

$$s = \sqrt{\text{variância}} = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

# Por que dividir por *n-1*?

- Porque há apenas *n* – 1 valores independentes.

- Dada uma **média** de *n* elementos, podemos usar quaisquer valores para os *n* – 1 valores, mas o último valor será automaticamente determinado

- Regra: perdemos um grau de liberdade para cada parâmetro estimado

# Passo a passo para calcular o Desvio padrão

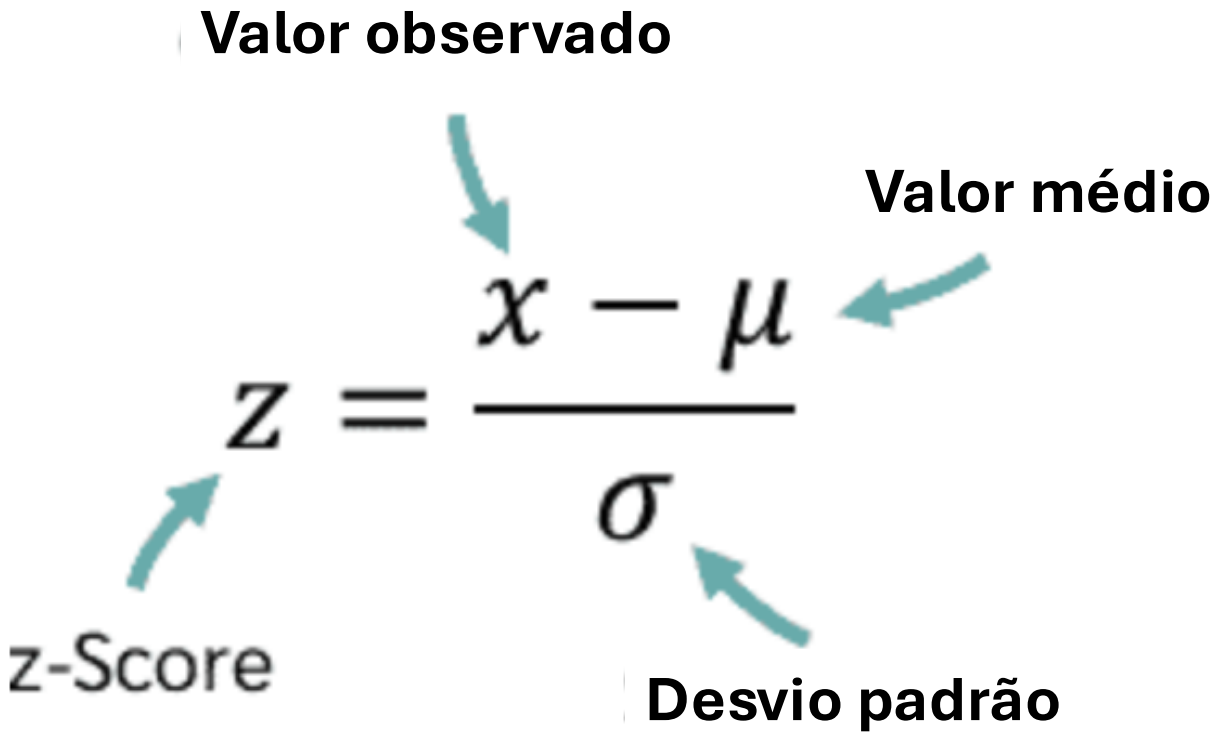
TABELA 3.4

| **Procedimento Geral para Encontrar o Desvio-padrão com a Fórmula 3.4** | **Exemplo Específico Usando Estes Valores Amostrais: 50, 25, 75, 35, 50, 25, 30, 50, 45, 25, 20** |
|---|---|
| **Passo 1:** calcule a média $\bar{x}$. | A soma de 50, 25, 75, 35, 50, 25, 30, 50, 45, 25, 20 é 430; portanto, $\bar{x} = \frac{\Sigma x}{n} = \frac{50+25+75+35+50+25+30+50+45+25+20}{11} = \frac{430}{11} = 39{,}1$ |
| **Passo 2:** subtraia a média de cada valor individual de dado. [O resultado é uma lista de desvios da forma $(x - \bar{x}$.] | Subtraia a média de 39,1 de cada valor amostral para obter esses desvios com relação à média: 10,9, –14,1, 35,9, –4,1, 10,9, –14,1, –9,1, 10,9, 5,9, –14,1, –19,1. |
| **Passo 3:** eleve ao quadrado cada um dos desvios obtidos no Passo 2. [Isso produz números da forma $(x - \bar{x})^2$.] | Os quadrados dos desvios do passo 2 são: 118,81, 198,81, 1.288,81, 16,81, 118,81, 198,81, 82,81, 118,81, 34,81, 198,81, 364,81. |
| **Passo 4:** adicione todos os quadrados obtidos no Passo 3. O resultado é $\Sigma(x - \bar{x})^2$. | A soma dos quadrados do Passo 3 é 2.740,91. |
| **Passo 5:** divida o total do Passo 4 pelo número $n - 1$, que é 1 a menos do que o número total dos valores amostrais presentes. | Com $n = 11$ valores de dados, $n - 1 = 10$, de modo que divida 2.740,91 por 10 para obter esse resultado: $\frac{2.740{,}91}{10} = 274{,}091$. |
| **Passo 6:** Ache a raiz quadrada do resultado do Passo 5. O resultado é o desvio-padrão, denotado por $s$. | O desvio-padrão é $\sqrt{274{,}091} = 16{,}556$. Expressando o resultado com uma casa decimal a mais do que nos dados originais, obtemos $s = 16{,}6$ minutos. |

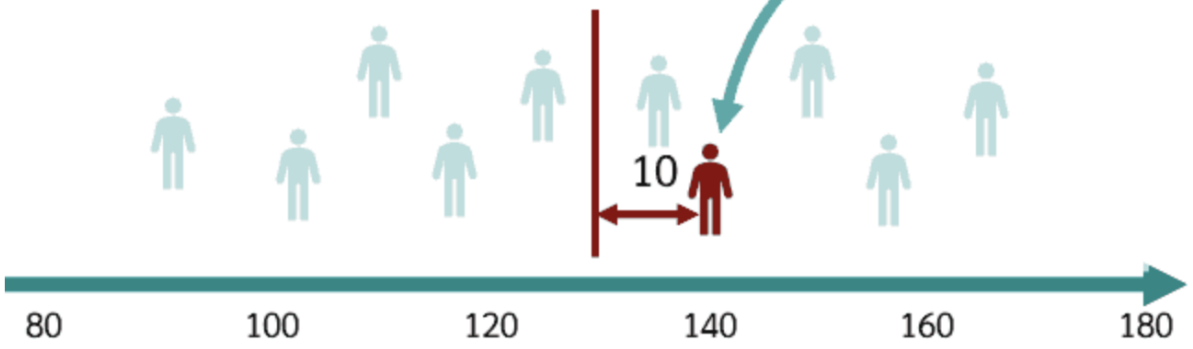
# Z-score

- Conceito:

Distância que um valor observado está da média em unidades de desvio padrão

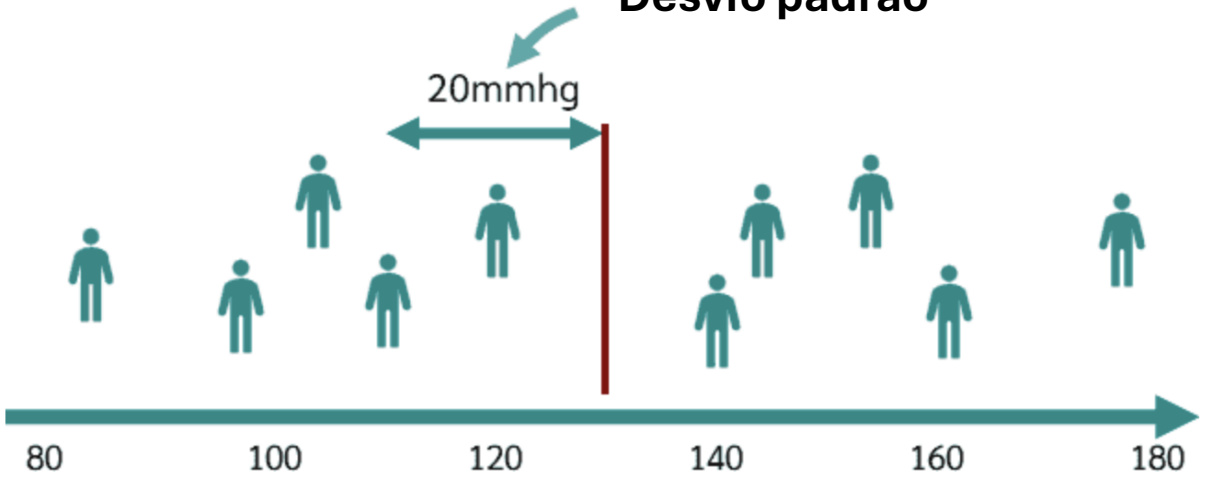


10 mmHg é pouco ou é muito?

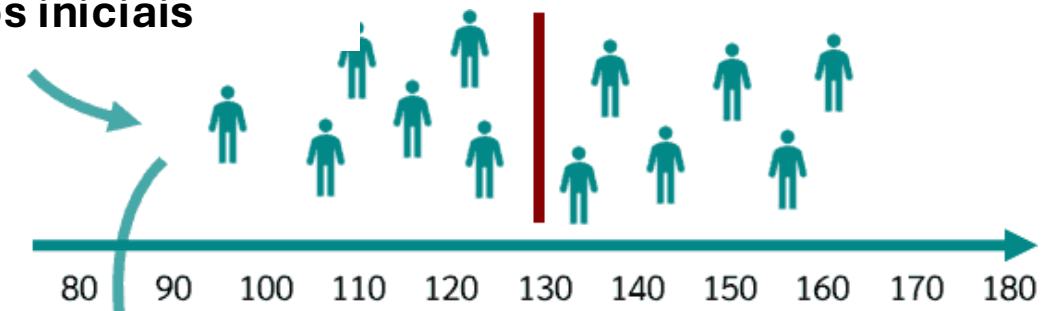



**Pressão sanguínea**

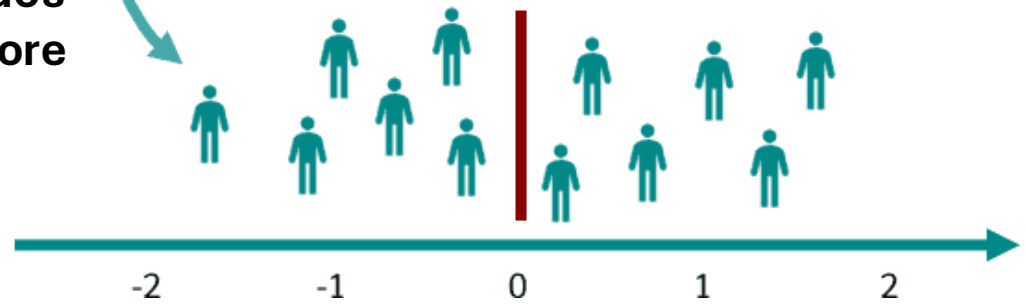
**Desvio padrão**

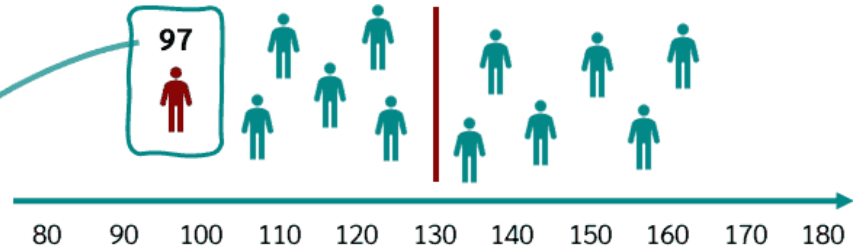



**Dados iniciais**



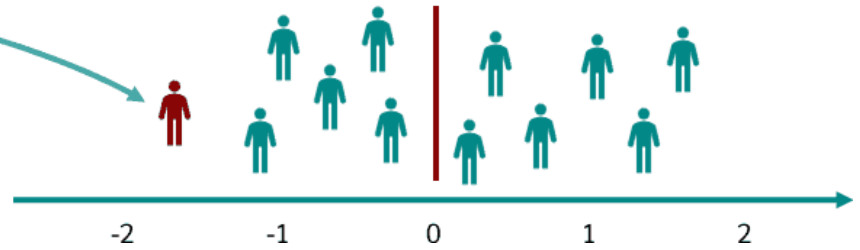

**Convertidos para z-score**



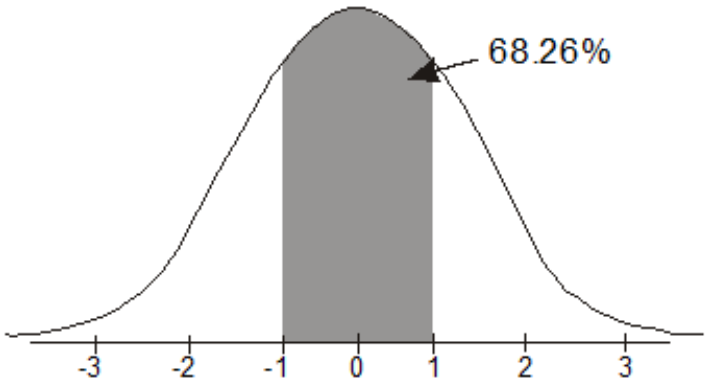




$$z = \frac{97 - 130}{20}$$

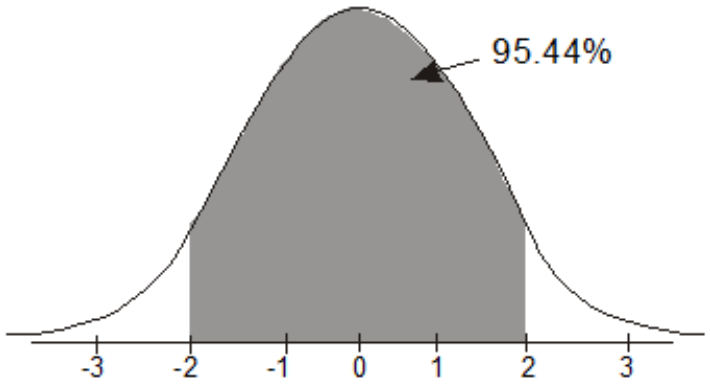
$$= -1.65$$




https://datatab.net/tutorial/z-score




The area under the curve is 0.6826




The area under the curve is 0.9544

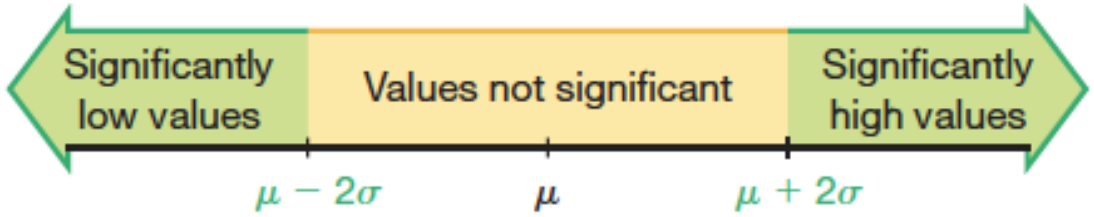



**FIGURE 3-3 Range Rule of Thumb for Identifying Significant Values**

Triola, M. 2024

**Por que isso é importante?**

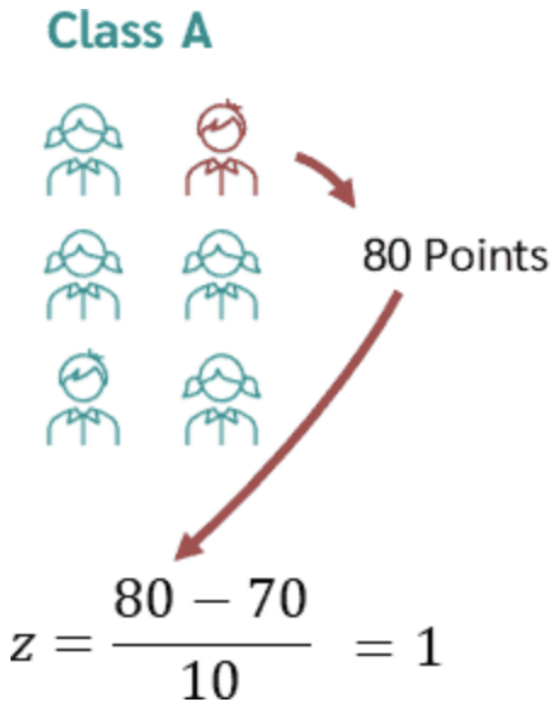
Porque para distribuições simétricas, sabemos que 95% dos dados estarão a uma distância de 2 desvios padrões da média! 😉

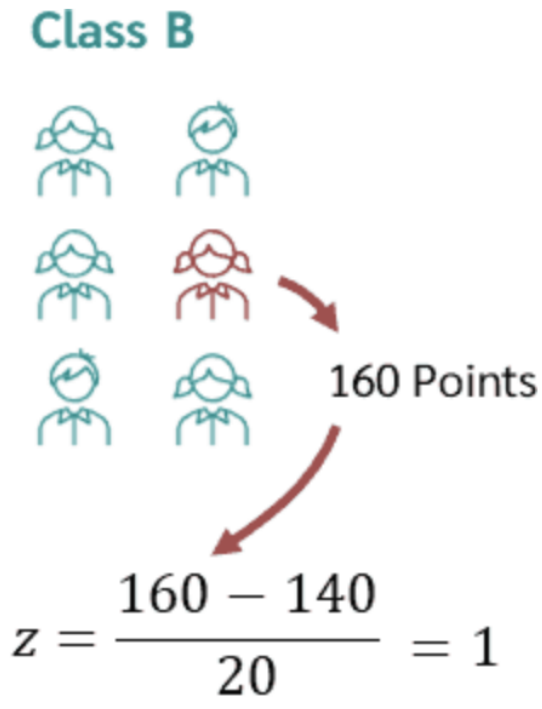
https://opentextbc.ca/businesstechnicalmath/chapter/8-4-z-scores-and-the-normal-curve/

# Por que isso é importante?

**Permite comparar amostras vindas de populações diferentes**

We now want to compare the performance of Max from class A, who scored 80 points, with the performance of Emma from class B, who scored 160 points.




https://datatab.net/tutorial/z-score

# Coeficiente de variação

- Razão entre o desvio padrão e a média

$$CV = \frac{s}{\bar{x}} \times 100$$

- Dois conjuntos de dados
  - 1, 3, 5
  - 55, 57, 53
  - Ambos têm variância $s^2 = 4$

$$CV = \frac{2}{3} \times 100 = 66,67\,\%$$

$$CV = \frac{2}{55} \times 100 = 3,64\,\%$$

- Não tem unidade, logo pode ser usado para **comparar conjuntos de dados de diferentes unidade de medida**

# Atividade: "Qual Remédio é Melhor?"

- (35 minutos)

- Formem 10-12 grupos (5-6 pessoas cada)

# Atividade: "Qual Remédio é Melhor?"

- Vocês trabalham para a ANVISA e precisam avaliar a eficácia e a segurança de dois novos anti-inflamatórios (Droga X e Droga Y) em comparação com um Placebo. **Os dados abaixo representam o tempo (em horas) que o efeito analgésico durou em diferentes pacientes.**
- **Placebo:** [1, 2, 2, 3, 3, 3, 4]
- **Droga X:** [4, 5, 5, 6, 6, 7, 15]
- **Droga Y:** [6, 6, 7, 7, 7, 8, 8]

ANVISA
Agência Nacional de Vigilância Sanitária


# Atividade: "Qual Remédio é Melhor?"

- Para cada grupo (Placebo, Droga X, Droga Y), calculem: **Média, Mediana, Amplitude e Desvio Padrão** (podem usar a calculadora do celular).

- Escrevam os resultados numa folha

# Atividade: "Qual Remédio é Melhor?"

- Após os cálculos, discutam e respondam:
  - **Eficácia:** Olhando para a tendência central (média e mediana), qual droga parece ser mais eficaz em prolongar a analgesia?
  - **Consistência:** Olhando para a dispersão (desvio padrão), qual droga tem o efeito mais consistente e previsível? A Droga X tem algum problema? (Eles devem notar o outlier e como ele afeta a média e o desvio padrão).
  - **Tomem uma Decisão Final:** Se vocês tivessem que aprovar apenas uma droga, qual seria e por quê? Justifiquem sua resposta usando os conceitos de tendência central E variabilidade.
  - **Visualização**. esboce rapidamente um boxplot para cada um dos três grupos (Placebo, Droga X, Droga Y) no verso da folha. Usem a mediana e a ideia de dispersão que vocês calcularam para guiar o desenho.

# Relatório final

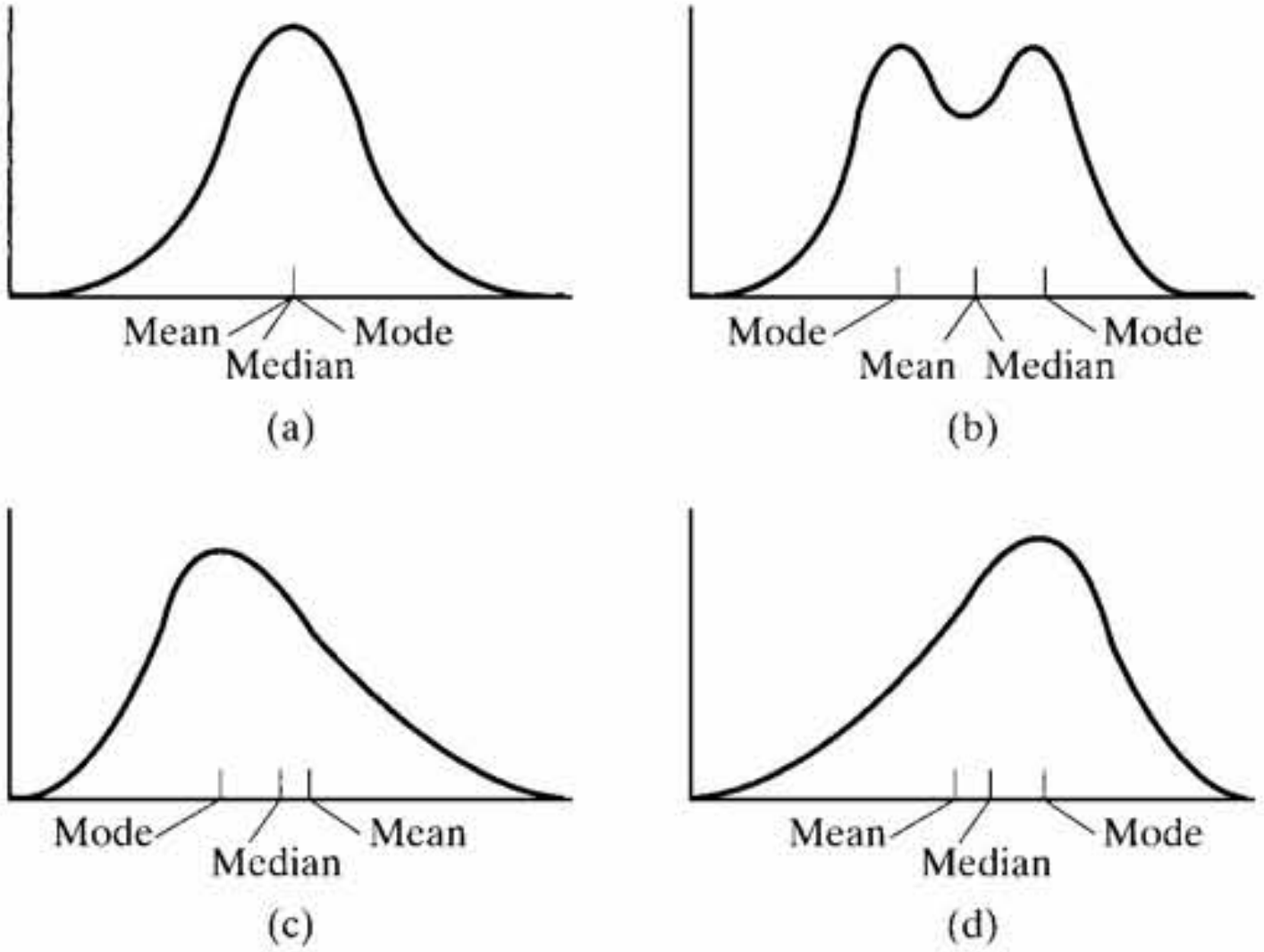
- 2 ou 3 grupos compartilhem suas conclusões com a turma
- Qual o comportamento da Droga X e Y, em termos de variância em relação à média?
- Qual das duas teria o feito mais confiável?

# Os 4 momentos de uma distribuição de probabilidade

- 1. Tendência central (e.g., média, mediana, moda)
- 2. Dispersão (e.g., variância)
- 3. Assimetria
- 4. Curtose

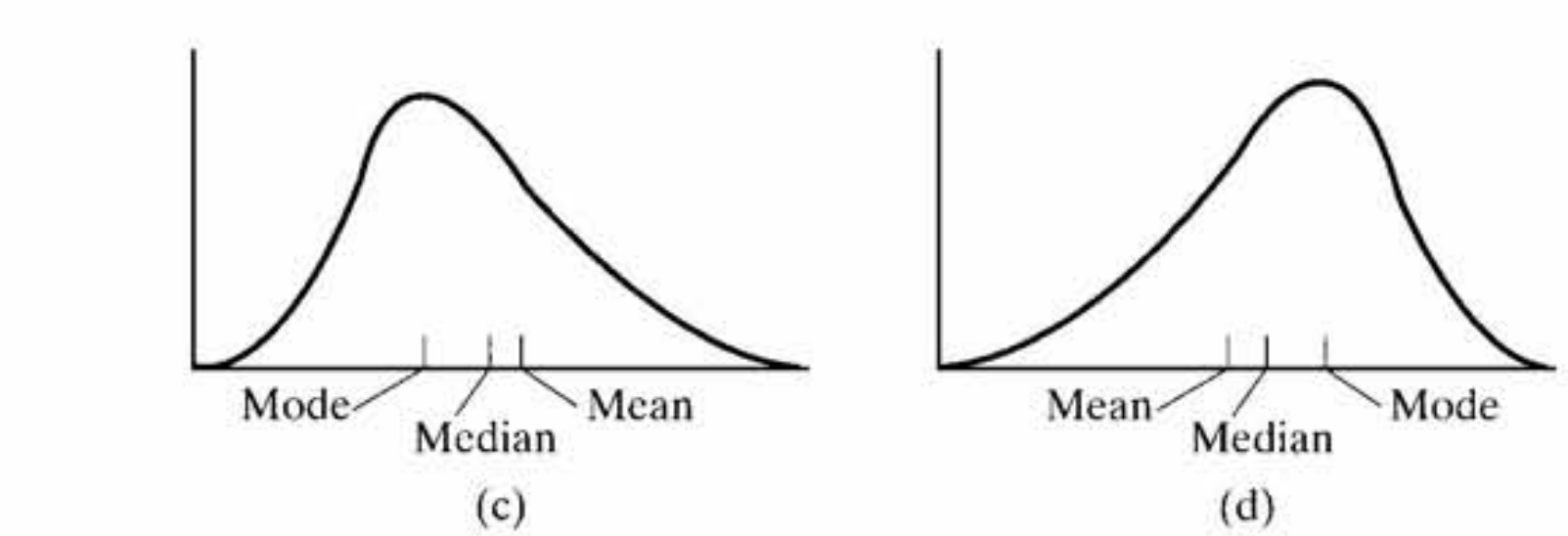
Simétricas

Assimétricas




Zar 2010

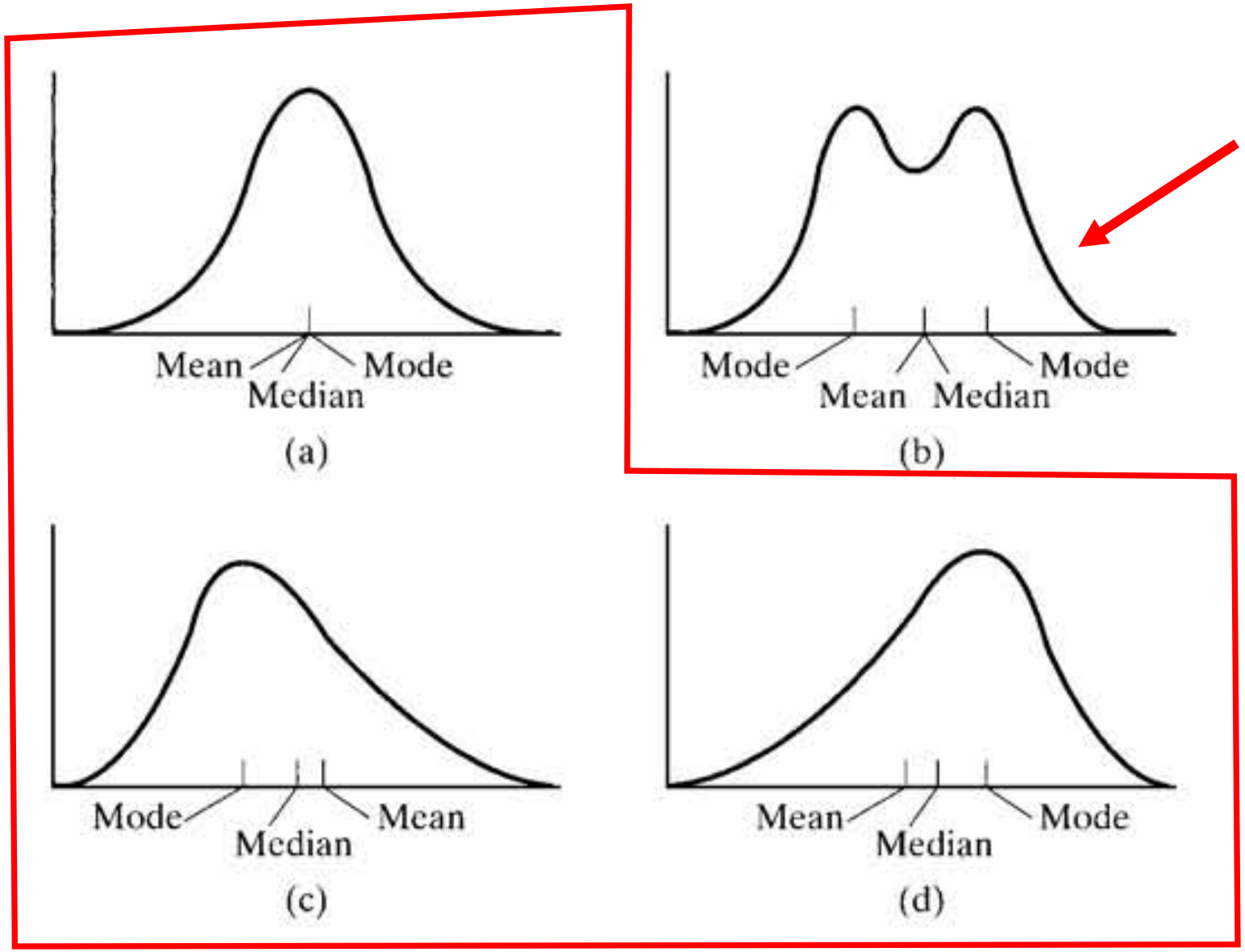
Assimétrica positiva




Assimétrica negativa

Zar 2010

Bimodal




Unimodais

Zar 2010

Kahoot!

https://play.kahoot.it/v2/lobby?quizId=44148d47-cfae-4aaf-af68-81710e641a21

15 minutos

# Como descreveríamos os dois grupos de capivaras agora com o que aprendemos ?

**Grupo A**

50, 51, 52, 58, 59

**Grupo B**

35, 52, 53, 65, 65

Os dois grupos têm uma **média idêntica, de 54 kg**, mas **o desvio padrão do B** é maior que o do **A**, indicando maior **variabilidade (=incerteza da medida central)**

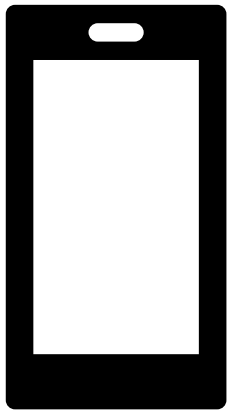
**Grupo A**

50, 51, 52, 58, 59

**Grupo B**

35, 52, 53, 65, 65

# Quais as suas principais dúvidas da aula de hoje?



5 minutos

https://app.sli.do/event/9EAtGB3oLhcpCV6HJqSJPX

# O que aprendemos hoje?

- A calcular um único valor, como a média, para representar o 'coração' dos nossos dados.
- Vimos que a **média** da Droga Y foi 7 h. Mas pensem nisso: isso foi só na nossa pequena **amostra** de pacientes. Será que podemos afirmar que a **média** de ***toda a população*** que usar a Droga Y será exatamente 7 h? Provavelmente não.
- Mas quão confiantes podemos estar de que a verdadeira **média da população** está *próxima* de 7? É possível criar uma faixa de valores, uma 'margem de segurança', para a nossa estimativa?
- É exatamente isso que vamos aprender na nossa próxima aula sobre **estimativa pontual e intervalos de confiança**.