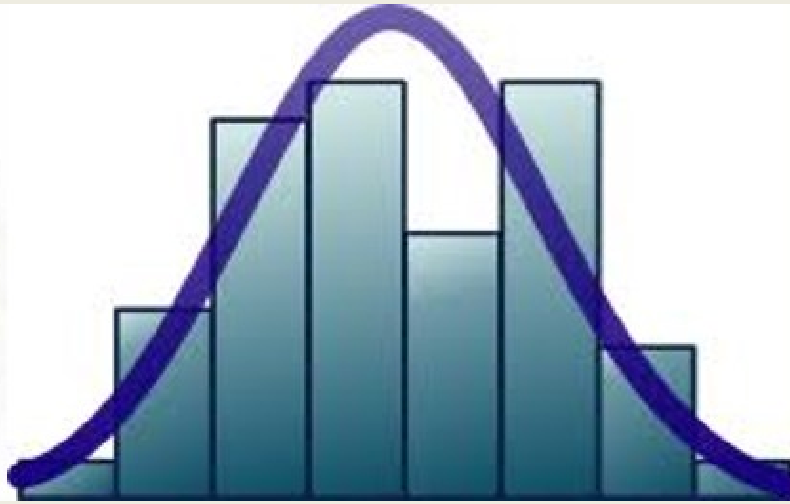
Probabilidade e Estatística





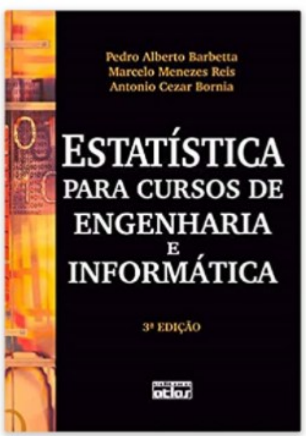
INSTITUTO FEDERAL
Catarinense
Campus Blumenau

Professor Jeovani Schmitt

Probabilidade e Estatística

Aula 5

Análise Exploratória de Dados (AED)



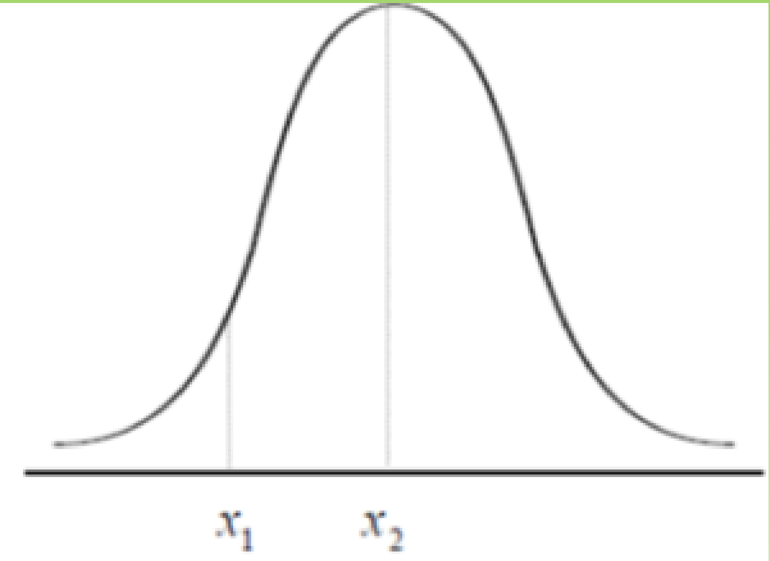
Capítulo 3

Medidas descritivas:

- Medidas de posição (Média, Mediana, Percentis, Moda)
- Medidas de dispersão (Variância, Desvio Padrão, Coeficiente de Variação)

Medidas de posição

As “Estatísticas” são índices numéricos que representam propriedades específicas das variáveis.



Qual o significado dos valores x_1 e x_2 de na distribuição?

Que locais da distribuição eles representam?

Medidas de posição – Média aritmética simples

Seja $x_1, x_2, x_3, \ldots, x_n$ uma amostra de n observações. A média aritmética simples dessas observações é definida por:

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

Medidas de posição – Média aritmética simples

Exemplo 1: Número de faltas dos alunos da turma de P&E do IFC no 1º. semestre de 2025 considerando 14 alunos:

0, 2, 3, 1, 0, 1, 2, 2, 3, 1, 2, 3, 2, 1

Calcule a média de faltas por aluno no 1º. semestre de 2025.

Medidas de posição – Média aritmética simples

Exemplo 1: Resolução

0, 2, 3, 1, 0, 1, 2, 2, 3, 1, 2, 3, 2, 1

$$\bar{x} = \frac{0+2+3+\dots+1}{14}$$

$$\bar{x} = \frac{23}{14} = \underline{1{,}64 \text{ faltas}}$$

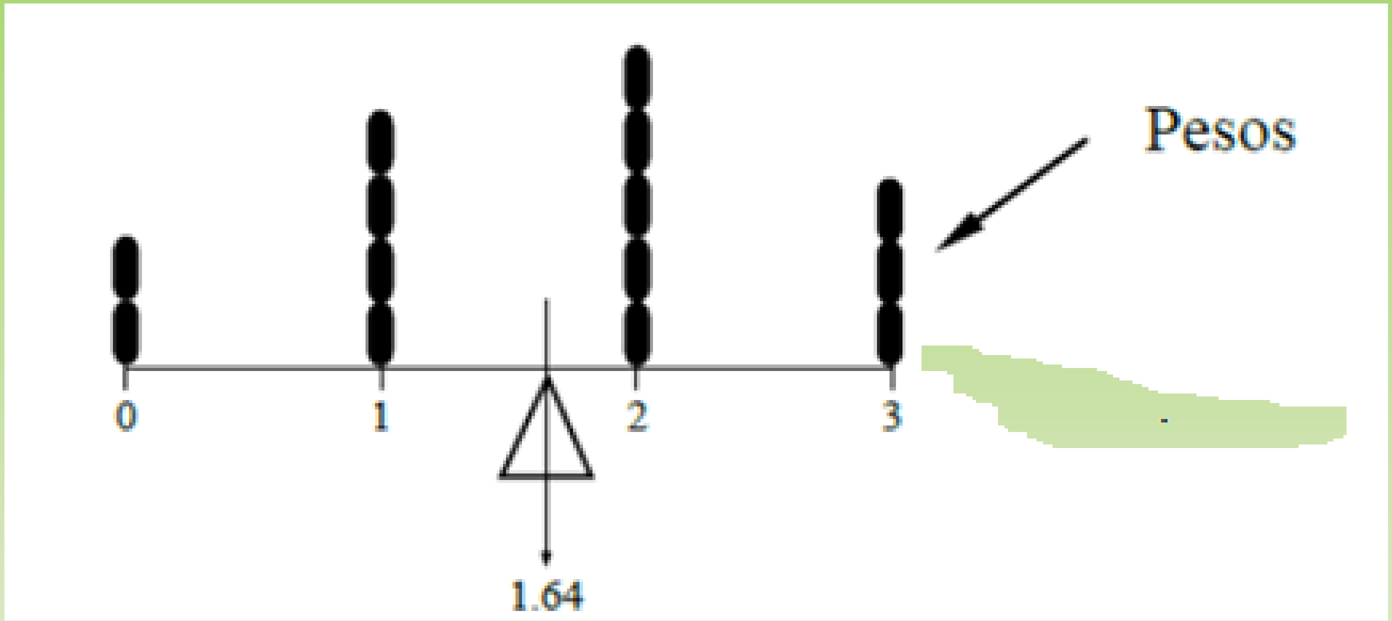
Medidas de posição – Média aritmética simples

O valor $\bar{x} = 1{,}64$ é o "número médio" de faltas por aluno no semestre.

É impossível um aluno faltar 1,64 vezes no semestre.

Medidas de posição – Média aritmética simples

O que significa o número médio de 1,64 vezes no semestre?



Medidas de posição – Média aritmética simples

A média aritmética nem sempre está no "centro"

Medidas de posição – Média aritmética simples

Exemplo 2: Número de faltas dos alunos da turma de P&E do IFC no 1º. semestre de 2025 considerando 14 alunos:

0, 2, 3, 1, 0, 1, 2, 2, 3, 1, 2, 3, 2, 25

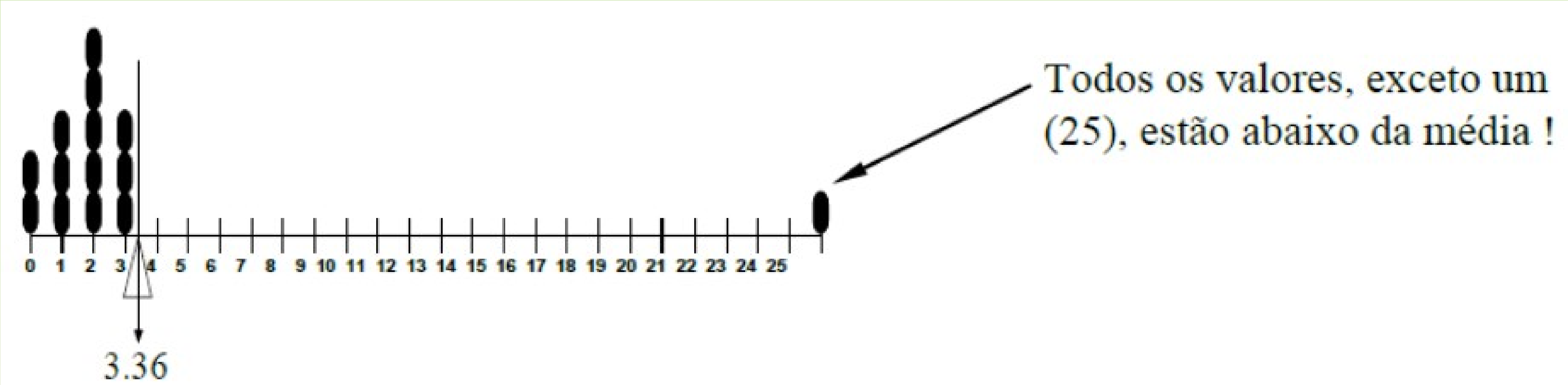
Calcule a média de faltas por aluno no 1º. semestre de 2025.

Medidas de posição – Média aritmética simples

Exemplo 2: Resolução

$$\bar{x} = \frac{0+2+3+1+0+1+2+2+3+1+2+3+2+25}{14}$$

$$\bar{x} = 3{,}36$$



Medidas de posição – Média aritmética simples

Sintaxe no R para o exemplo 1

```
faltas <- c(0,2,3,1,0,1,2,2,3,1,2,3,2,1)
mean(faltas)
```

```
faltas <- c(rep(0,2),rep(1,4),rep(2,5),rep(3,3))
```

Medidas de posição – Média aritmética simples

```
# Sintaxe no R para o exemplo 2

faltas <- c(0,2,3,1,0,1,2,2,3,1,2,3,2,25)
mean(faltas)

faltas <- c(rep(0,2),rep(1,3),rep(2,5),rep(3,3),25)
mean(faltas)
```

Medidas de posição – Média aritmética ponderada

A média ponderada dos números $x_1, x_2, x_3, \dots, x_n$ com pesos $p_1, p_2, p_3, \dots, p_n$, representada por $\bar{x}_p$, é definida como:

$$\bar{x}_P = \frac{p_1 x_1 + p_2 x_2 + \cdots + p_n x_n}{p_1 + p_2 + \cdots + p_n} = \frac{\sum_{i=1}^{n} p_i x_i}{\sum_{i=1}^{n} p_i}$$

Medidas de posição – Média aritmética ponderada

Exemplo 3: Em uma sala há 26 alunos no total, sendo que 16 estudantes têm 15 anos, 8 estudantes têm 16 anos e apenas 2 alunos têm 17 anos. Qual a idade média da turma?

Idade	nº alunos	
15	16	240
16	8	128
17	2	34
	26	402

$$\bar{x}_p = \frac{402}{26} \approx 15,46 \text{ anos}$$

Sintaxe no R para o exemplo 3

```
freq <- c(16,  8,  2)
peso <- freq/sum(freq)
x <- c(15,16,17)
xp <- weighted.mean(x, peso)
xp
```

Medidas de posição – Mediana

A mediana de um conjunto de n observações $x_1, x_2, x_3, \ldots, x_n$, é o valor “do meio” do conjunto, quando os dados estão dispostos em ordem crescente.

Medidas de posição – Mediana

Exemplo 4: Venda de pacotes turísticos em 9 agências de Blumenau no mês de fevereiro de 2025:

Dados brutos: 40, 52, 48, 54, 60, 58, 45, 54, 42

Dados em ordem crescente: 40, 42, 45, 48, **52**, 54, 54, 58, 60.

MEDIANA = 52

```r
# Sintaxe no R para o exemplo 4

turismo <- c(40,52,48,54,60,58,45,54,42)
median(turismo)
```

Medidas de posição – Percentil

O percentil é uma generalização do conceito de mediana. Enquanto a mediana divide um conjunto de valores dispostos em ordem crescente em duas partes iguais os percentis dividem em 100 partes iguais.

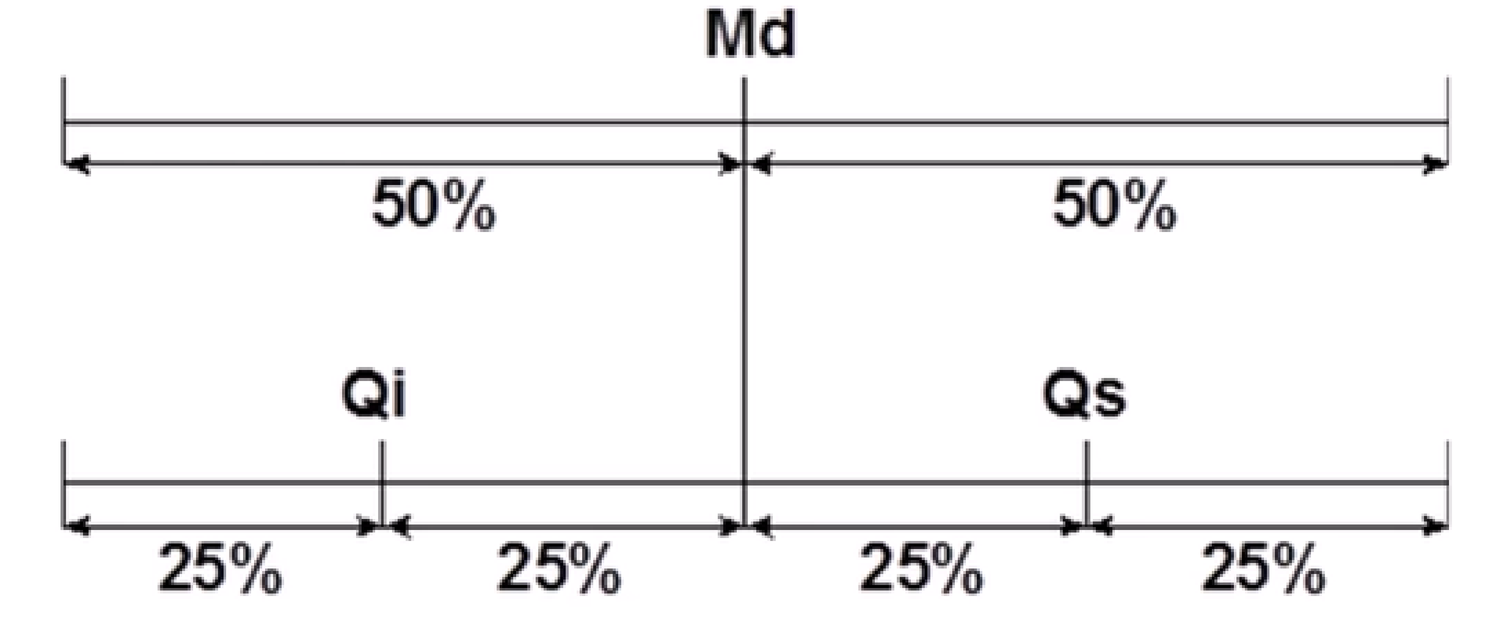
Medidas de posição – Percentil

★ PERCENTIL DE ORDEM 100p (P_{100p})

- O **Percentil de ordem 100p** de um conjunto de valores dispostos em ordem crescente é um valor tal que (100p)% das observações estão nele ou abaixo dele e 100(1-p)% estão nele ou acima dele (0 < p < 1).
- O percentil de ordem 50 (P_{50}) é a **mediana**
- Os percentis de ordens 25, 50 e 75, representados por Q_1, Q_2 e Q_3 são chamados **quartis (inferior, mediano e superior)**.

P_{75} = 75% dos valores estão nele ou abaixo dele

Medidas de posição – Separatrizes Q_i, Md, Q_S



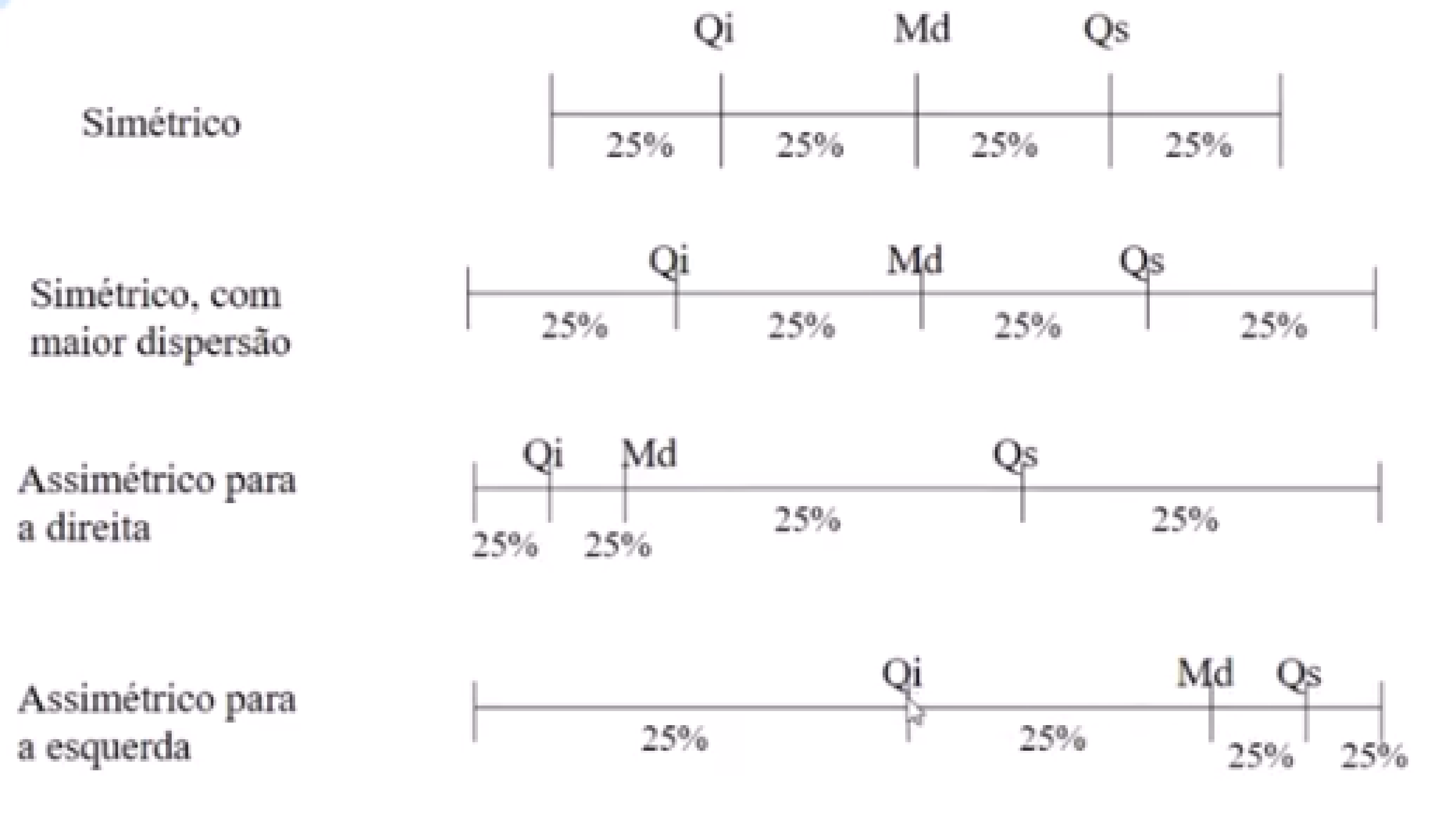
$Q_2 =$ Md = mediana

$Q_1 = Q_i =$ 1º. quartil ou quartil inferior

$Q_3 = Q_s =$ 3º. quartil ou quartil superior

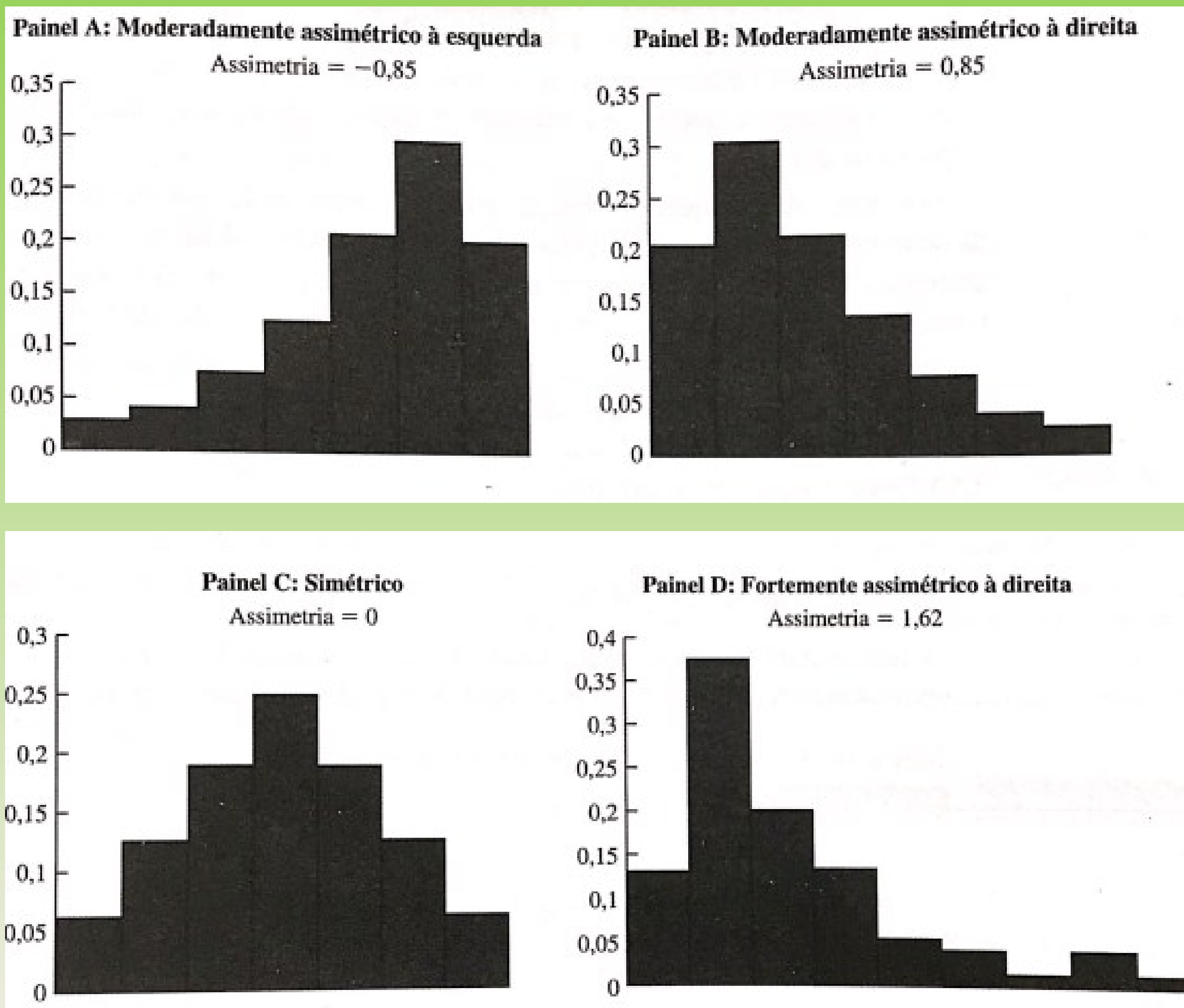
Medidas de posição – Separatrizes Q_i, Md, Q_S

- Avaliar a assimetria e dispersão dos valores



Medidas de posição – Separatrizes Q_i, Md, Q_S

- Avaliar a assimetria e dispersão dos valores



Medidas de posição

Exemplo 5: Calcular Q_i, Md, Q_S para o seguinte conjunto de dados:

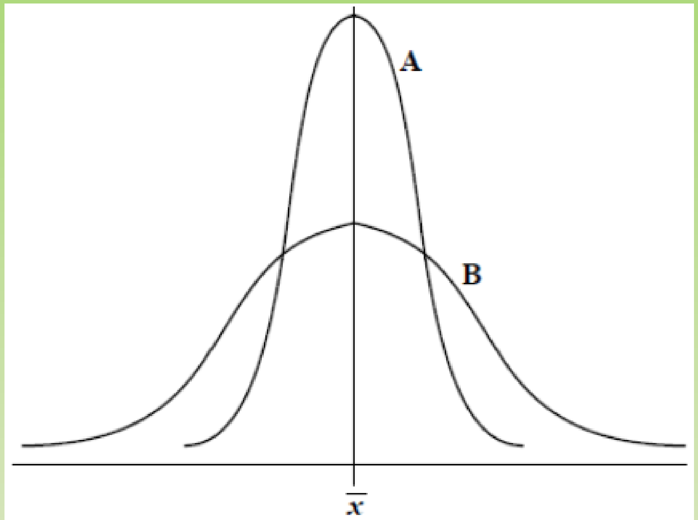
Dados brutos: **15**, **18**, **5**, **7**, **9**, **11**, **3**, **5**, **6**, **8**, **12**

Sintaxe no R para o exemplo 5

```
dados <- c(15,18,5,7,9,11,3,5,6,8,12)
dados
quantile(dados, probs=c(0.25,0.50,0.75))
```

Medidas de dispersão

Uma "Estatística" de dispersão refere-se à variabilidade ou heterogeneidade dos dados.



Nas duas distribuições (A e B), qual tem maior dispersão?

Medidas de dispersão

Exemplo 6: Peso do papel em gramas produzido em diferentes máquinas

Amostra	Máquinas		
	A	B	C
1	200	152	205
2	210	248	203
3	190	260	195
4	215	200	197
5	185	140	200
Média	200	200	200

Medidas de dispersão: variância

Exemplo 6: Calcular a variância dos dados da máquina A.

200, 210, 190, 215, 185

VARIÂNCIA

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

Medidas de dispersão: variância

PESO DO PAPEL EM GRAMAS
PRODUZIDO PELA MÁQUINA A

Observações	Desvio	Quadrado do Desvio
1	$(200 - 200) = 0$	$(200 - 200)^2 = 0$
2	$(210 - 200) = 10$	$(210 - 200)^2 = 100$
3	$(190 - 200) = -10$	$(190 - 200)^2 = 100$
4	$(215 - 200) = 15$	$(215 - 200)^2 = 225$
5	$(185 - 200) = -15$	$(185 - 200)^2 = 225$
SOMA	**0,00**	**650**

Soma dos Desvios é sempre **zero** (exceto por problemas de arredondamento).

Melhor utilizar a SOMA DE QUADRADOS DOS DESVIOS que será sempre **positiva**.

VARIÂNCIA

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

Medidas de dispersão: desvio padrão

DESVIO PADRÃO

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

mede a variabilidade independentemente do número de observações (n) e com a mesma unidade de medida da média.

Como as variáveis com que trabalhamos possuem **UNIDADES DE MEDIDA**, é importante considerá-las quando medimos a heterogeneidade dos dados.

A variância sempre eleva ao quadrado as unidades de medida, gerando escalas sem sentido prático.

Se utilizarmos a raiz quadrada da variância, recuperaremos as unidades originais.

Variável	Unidade de Medida	Unidade da Variância
altura de um parafuso	cm	**cm^2**
peso de um saco de arroz	kg	**kg^2**

Medidas de dispersão: desvio padrão

Exemplo 7: Calcular o desvio padrão dos dados da máquina A.

200, 210, 190, 215, 185

DESVIO PADRÃO

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

mede a variabilidade independentemente do número de observações (n) e com a mesma unidade de medida da média.

Medidas de dispersão: coeficiente de variação

Coeficiente de Variação

O Coeficiente de variação é uma forma de se medir a variabilidade de uma variável de modo independente da UNIDADE DE MEDIDA utilizada ou da ORDEM DE GRANDEZA dos dados.

Razão entre desvio padrão e média torna o CV um número puro.

COEFICIENTE DE VARIAÇÃO

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

mede a variabilidade numa escala percentual, independe da unidade de medida ou da ordem de grandeza da variável.

Medidas de dispersão: coeficiente de variação

Exemplo 8: Calcular o coeficiente de variação dos dados da máquina A.

200, 210, 190, 215, 185

$$coeficiente\ de\ variação = \frac{s}{\bar{x}} \times 100$$

Sintaxe no R para dp e cv dos exemplos 6, 7 e 8

```
mA <- c(200,210,190,215,185)
mB <- c(152,248,260,200,140)
mC <- c(205,203,195,197,200)

sd(mA)
sd(mB)
sd(mC)

cv_mA <- (sd(mA)/mean(mA))*100
cv_mB <- (sd(mB)/mean(mB))*100
cv_mC <- (sd(mC)/mean(mC))*100

round(cv_mA, digits = 2)
round(cv_mB, digits = 2)
round(cv_mC, digits = 2)
```

Exercício no R

Notas de uma turma de 40 alunos:

0 10 10 20 30 40 40 40 40 50
50 50 50 50 60 60 60 60 70 70
70 70 70 70 80 80 80 80 80 80
80 80 90 90 90 90 100 100 100 100

a) Calcular as seguintes estatísticas:

Média, Q_i, Md, Q_S , Percentil 95, Desvio padrão e

Coeficiente de variação

b) Construir o histograma e avaliar a simetria dos dados.

Exercício no R

Sintaxe

notas <-
c(0,rep(10,2),20,30,rep(40,4),rep(50,5),rep(60,4),rep(70,6),rep(80,8),rep(90,4),rep(100,4))

notas

mean(notas) # calcula a média

quantile(dados, probs=c(0.25,0.50,0.75)) # Calcula Q_i, Md, Q_S

quantile(notas,.95) # Calcule o percentil 95

sd(notas) # desvio padrão
cv_notas <- (sd(notas)/mean(notas))*100 # coeficiente de variação
cv_notas

hist(notas, ylab="número de alunos", main="Notas dos alunos")
hist(notas, scale="frequency", breaks="Sturges", col="blue", xlab="notas",
ylab="número de alunos", main="Notas dos alunos")