

Estatística para Cursos de Engenharia e Informática

Pedro Alberto Barbetta / Marcelo Menezes Reis / Antonio Cezar Bornia
São Paulo: Atlas, 2004

Cap. 3 – Análise exploratória de dados

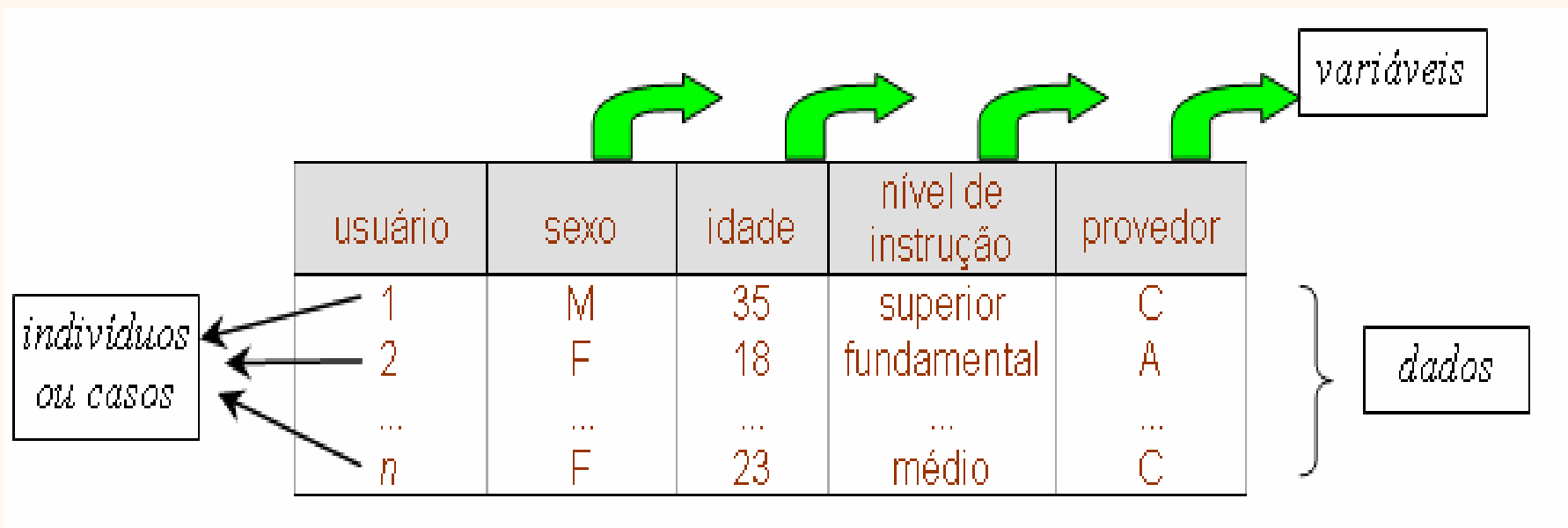
APOIO:

Fundação de Apoio à Pesquisa Científica e Tecnológica do Estado de Santa Catarina (FAPESC)

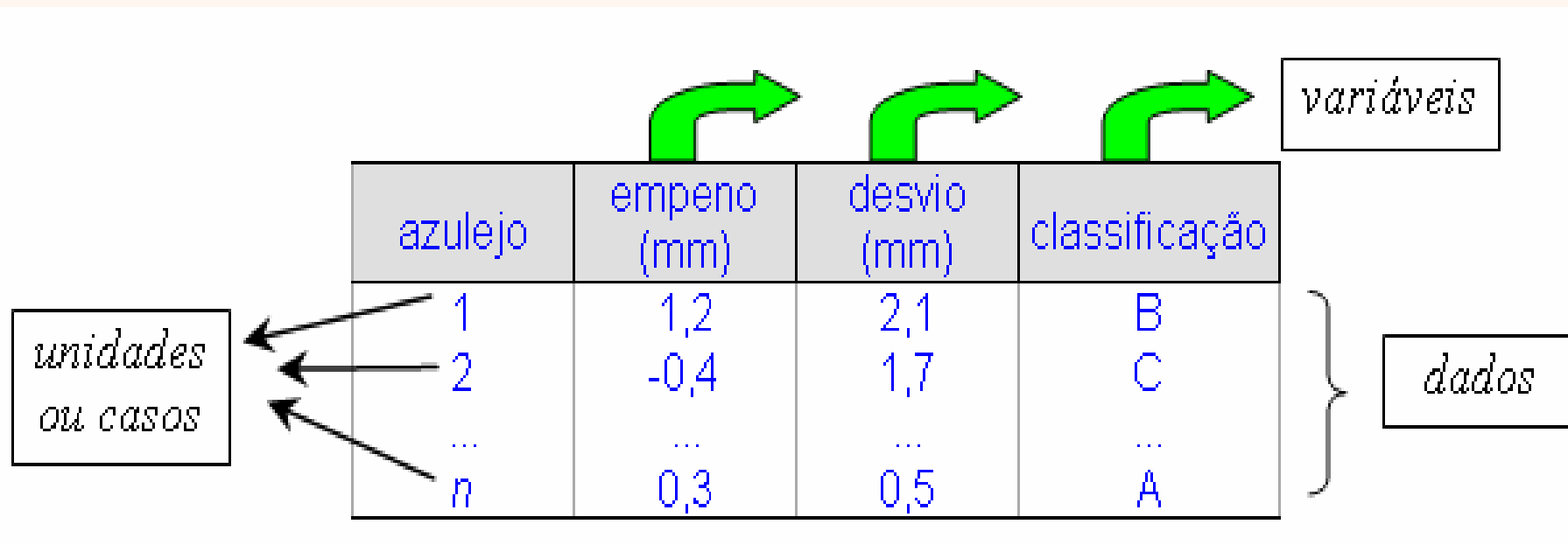
Departamento de Informática e Estatística – UFSC (INE/CTC/UFSC)

BARBETTA, REIS e BORNIA – Estatística para Cursos de Engenharia e Informática. Atlas, 2004

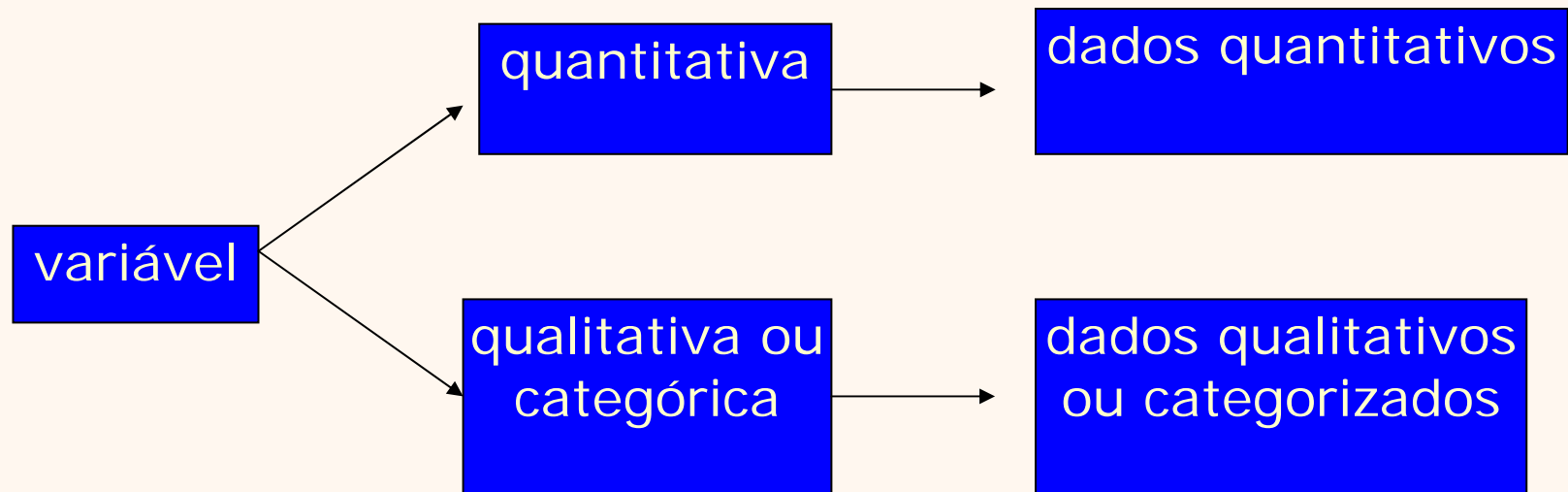
Dados e variáveis



Dados e variáveis



Dados e variáveis



Distribuição de freqüências

- A **distribuição de freqüências** consiste na organização dos dados de acordo com as ocorrências dos diferentes resultados observados.
- Pode ser apresentada em **tabela** ou **gráfico**.

Dados

Provedor usado por cada usuário

indivíduo	provedor	indivíduo	provedor	indivíduo	provedor	indivíduo	provedor
1	C	11	C	21	B	31	A
2	A	12	A	22	A	32	A
3	B	13	B	23	A	33	B
4	B	14	D	24	B	34	C
5	C	15	A	25	A	35	B
6	B	16	B	26	A	36	D
7	D	17	B	27	B	37	B
8	B	18	C	28	D	38	B
9	B	19	D	29	D	39	B
10	A	20	B	30	C	40	C

Distribuição de frequências para variáveis qualitativas

Tabela. Distribuição de frequências do provedor usado pelo visitante do *site*.

Provedor	Frequência	Percentagem
A	10	25,0
B	17	42,5
C	7	17,5
D	6	15,0
Total	40	100,0

Distribuição de frequências para variáveis qualitativas

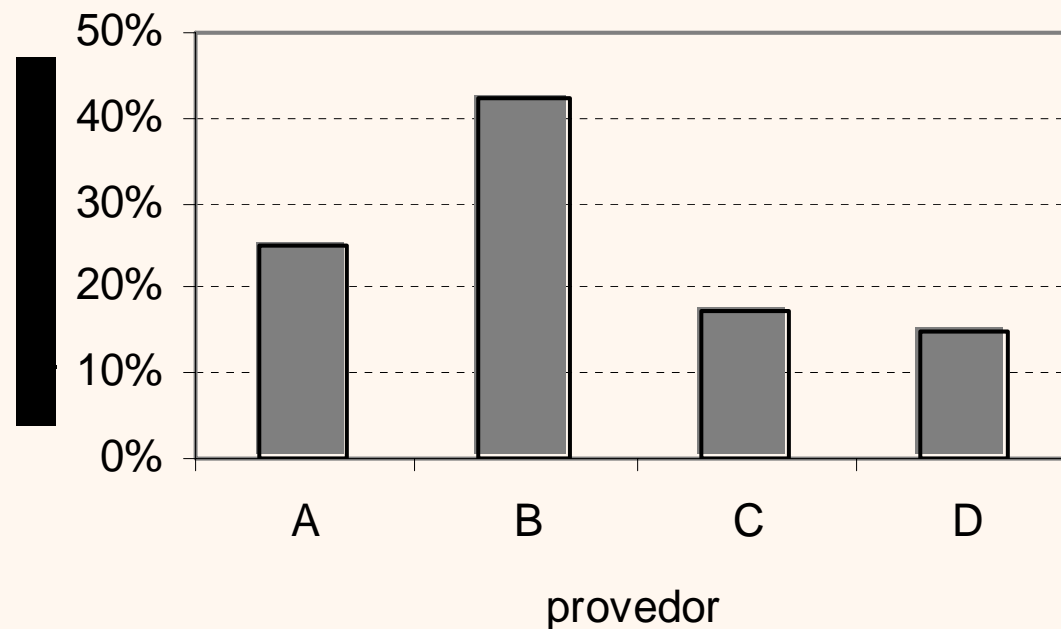


Gráfico de colunas para a apresentação da distribuição de frequências do provedor usado pelo visitante do *site*.

Distribuição de frequências para variáveis qualitativas

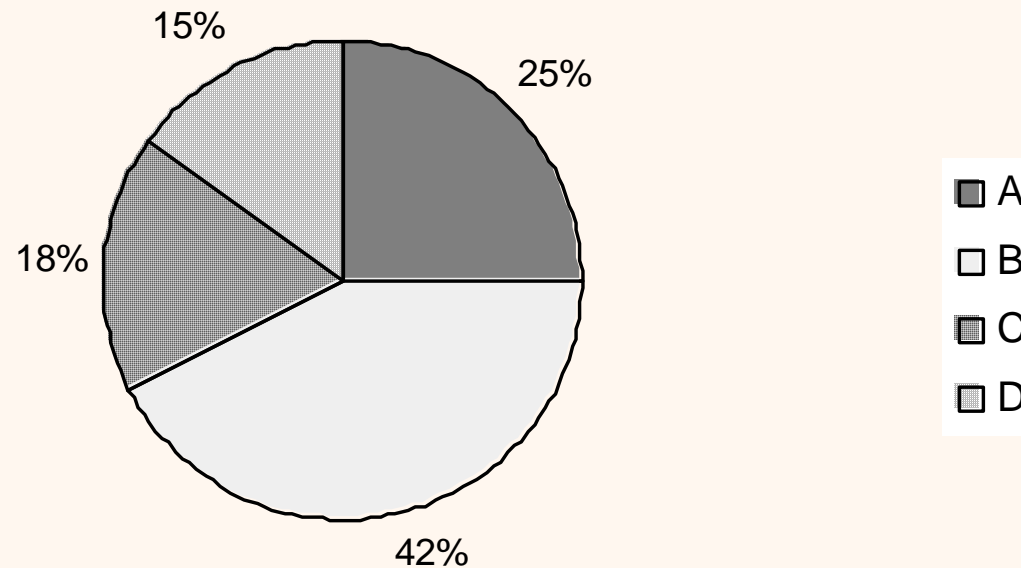
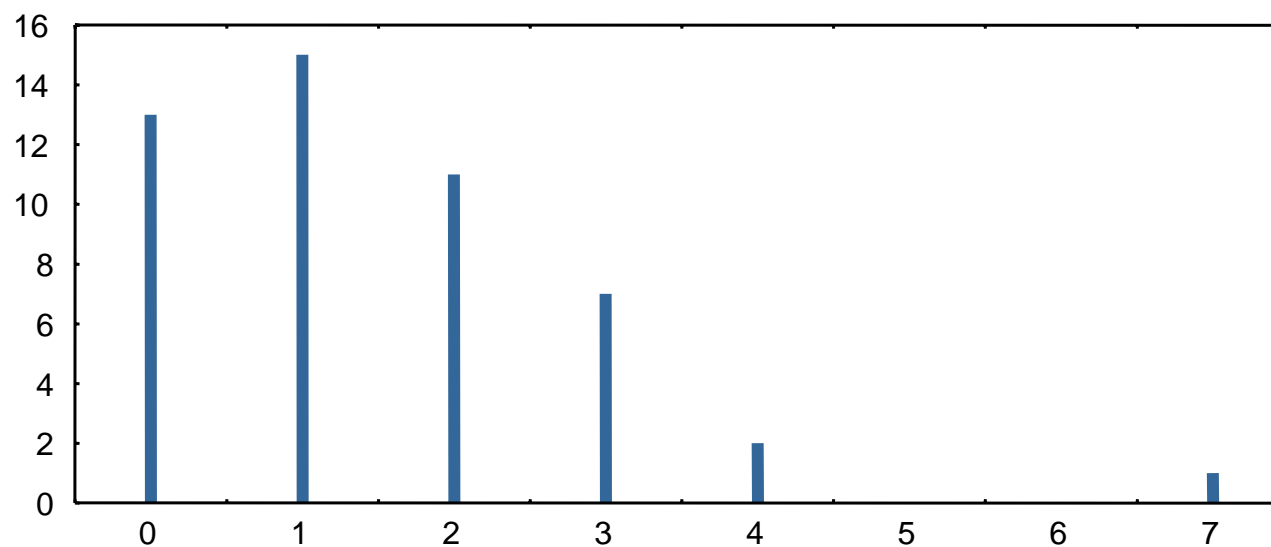


Gráfico de setores para a apresentação da distribuição de frequências do provedor usado pelo visitante do *site*.

Distribuição de freqüências para variáveis quantitativas discretas

% de itens



Número de defeitos

Variáveis contínuas

Construção da distribuição de frequências

5,2	6,4	5,7	8,3	7,0	5,4	4,8	9,1
5,5	6,2	4,9	5,7	6,3	5,1	8,4	6,2
8,9	7,3	5,4	4,8	5,6	6,8	5,0	6,7
8,2	7,1	4,9	5,0	8,2	9,9	5,4	5,6
5,7	6,2	4,9	5,1	6,0	4,7	14,1	5,3
4,9	5,0	5,7	6,3	6,0	6,8	7,3	6,9
6,5	5,9						

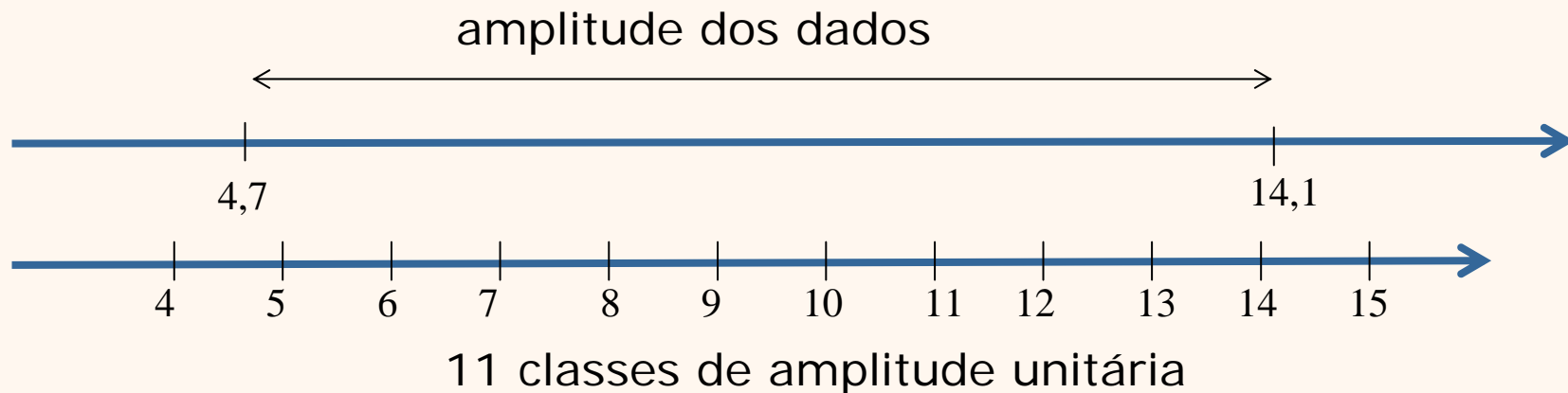
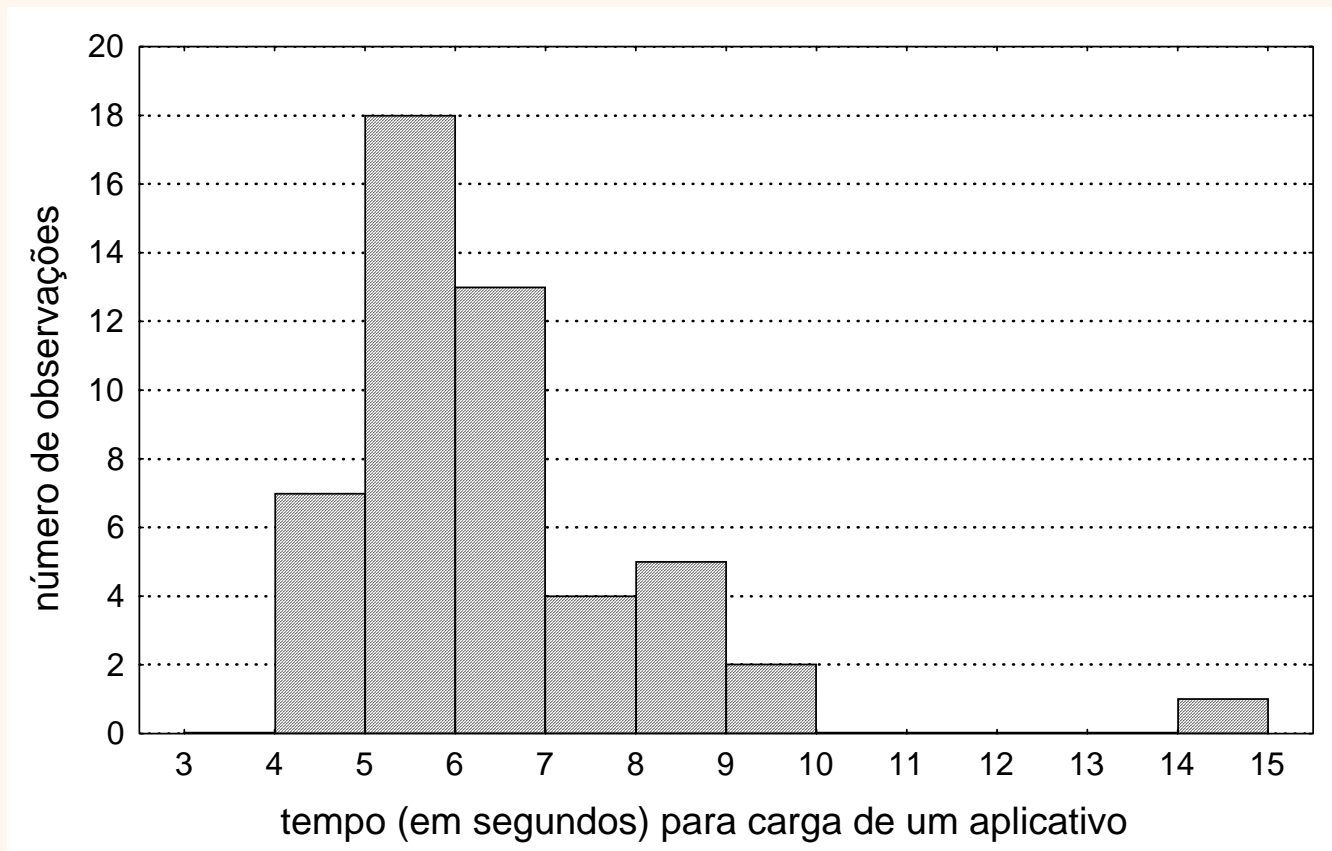


Tabela de frequências: variável contínua

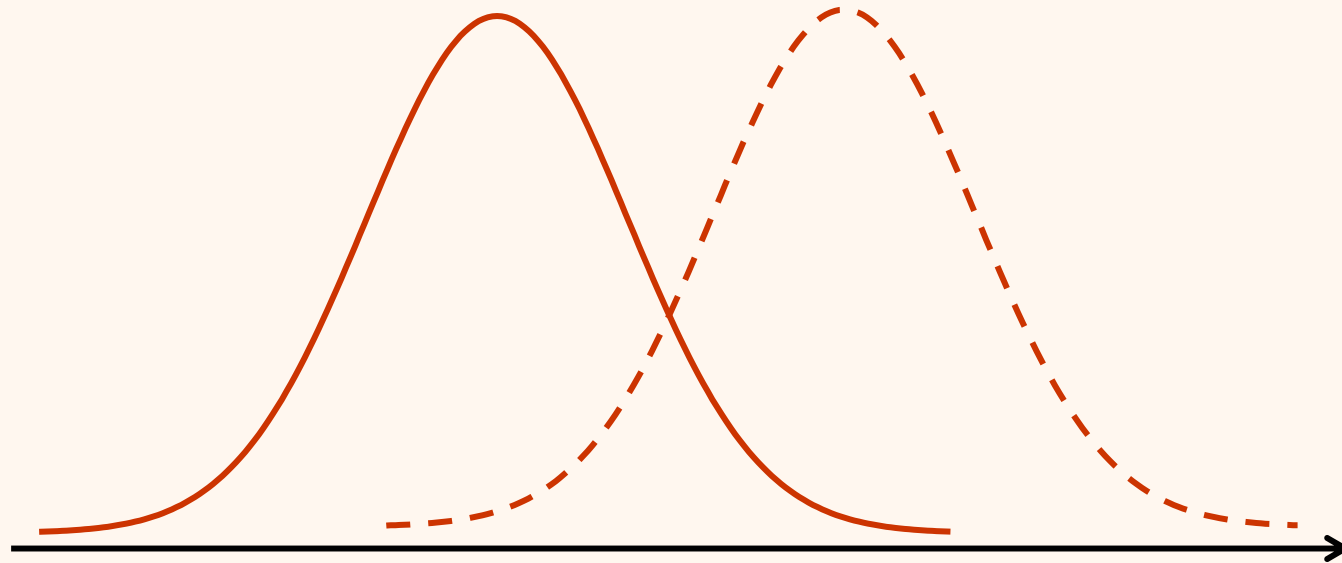
Classes de tempo	Ponto médio	Número de observações n_j	Percentagem de observações $100f_j$	Percentagem acumulada $100F_j$
4 — 5	4,5	7	14	14
5 — 6	5,5	18	36	50
6 — 7	6,5	13	26	76
7 — 8	7,5	4	8	84
8 — 9	8,5	5	10	94
9 — 10	9,5	2	4	98
10 — 11	10,5	0	0	98
11 — 12	11,5	0	0	98
12 — 13	12,5	0	0	98
13 — 14	13,5	0	0	98
14 — 15	14,5	1	2	100
Total	-	50	100	-

Histograma



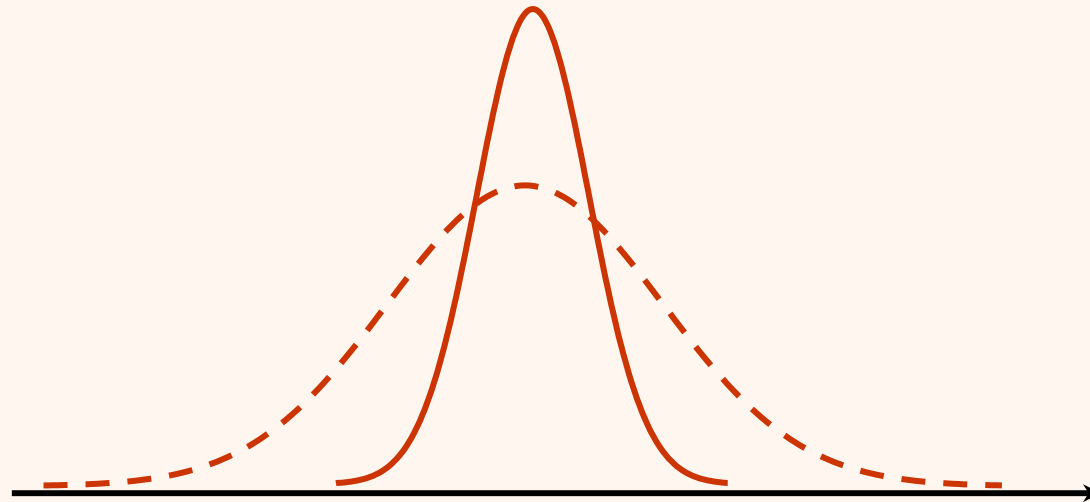
Formas de uma distribuição de frequências

- Distribuições diferentes em termos da posição central



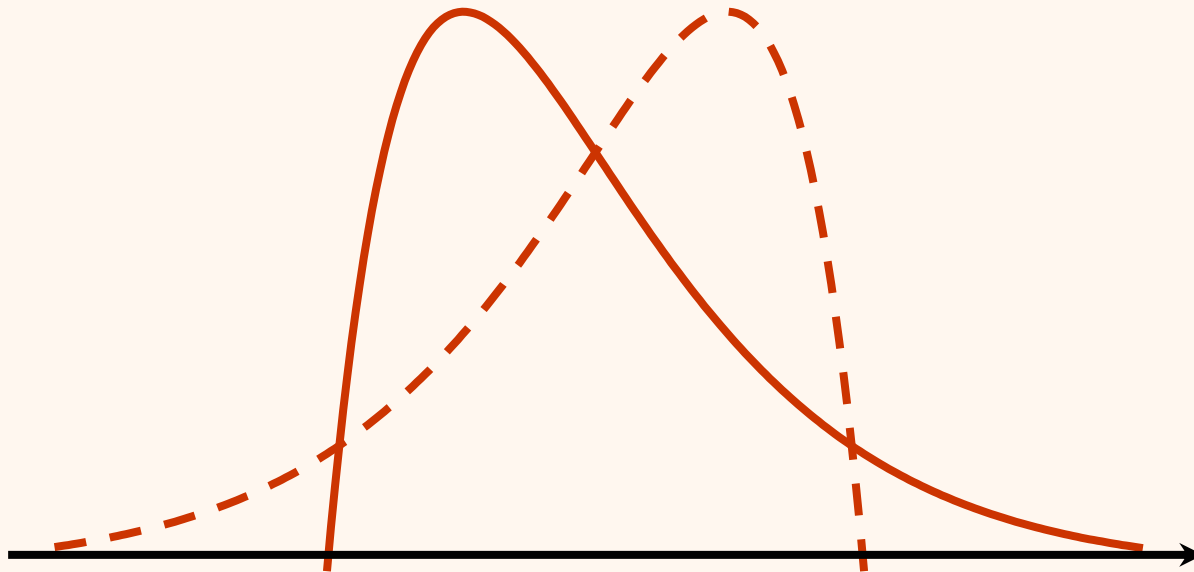
Formas de uma distribuição de frequências

- Distribuições diferentes quanto à dispersão



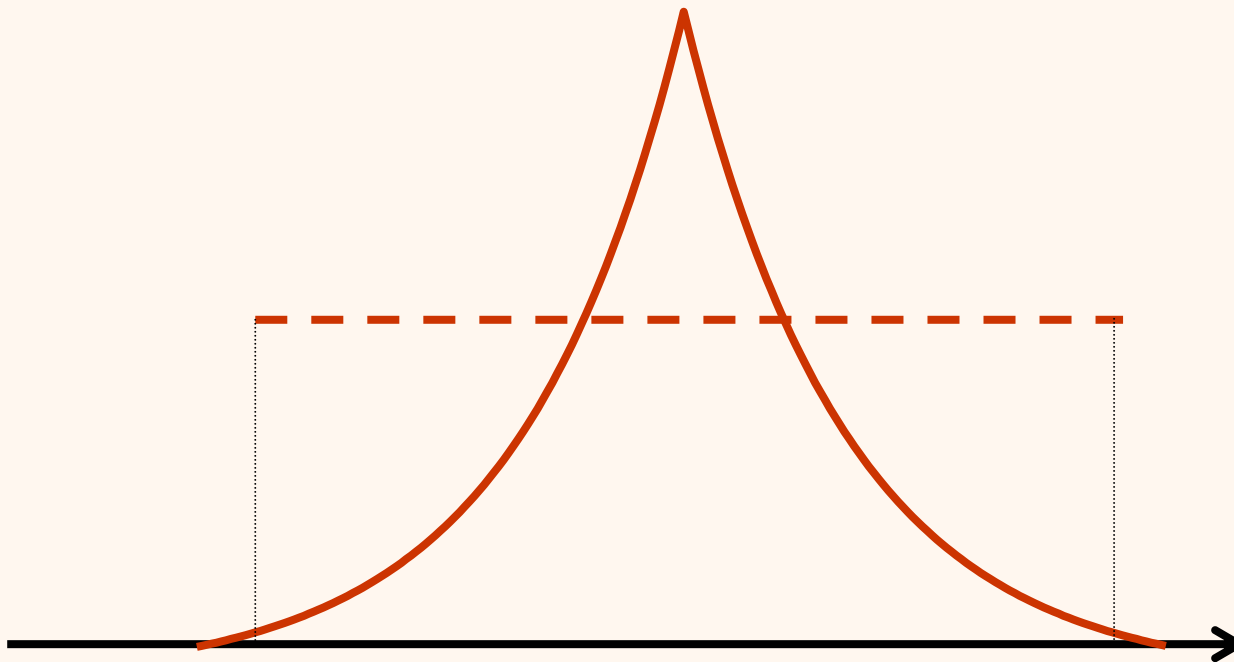
Formas de uma distribuição de frequências

- Distribuições diferentes quanto à assimetria



Formas de uma distribuição de frequências

- Distribuições diferentes quanto à curtose



Medidas descritivas

- A média aritmética: uma medida de posição central.

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

Exemplo

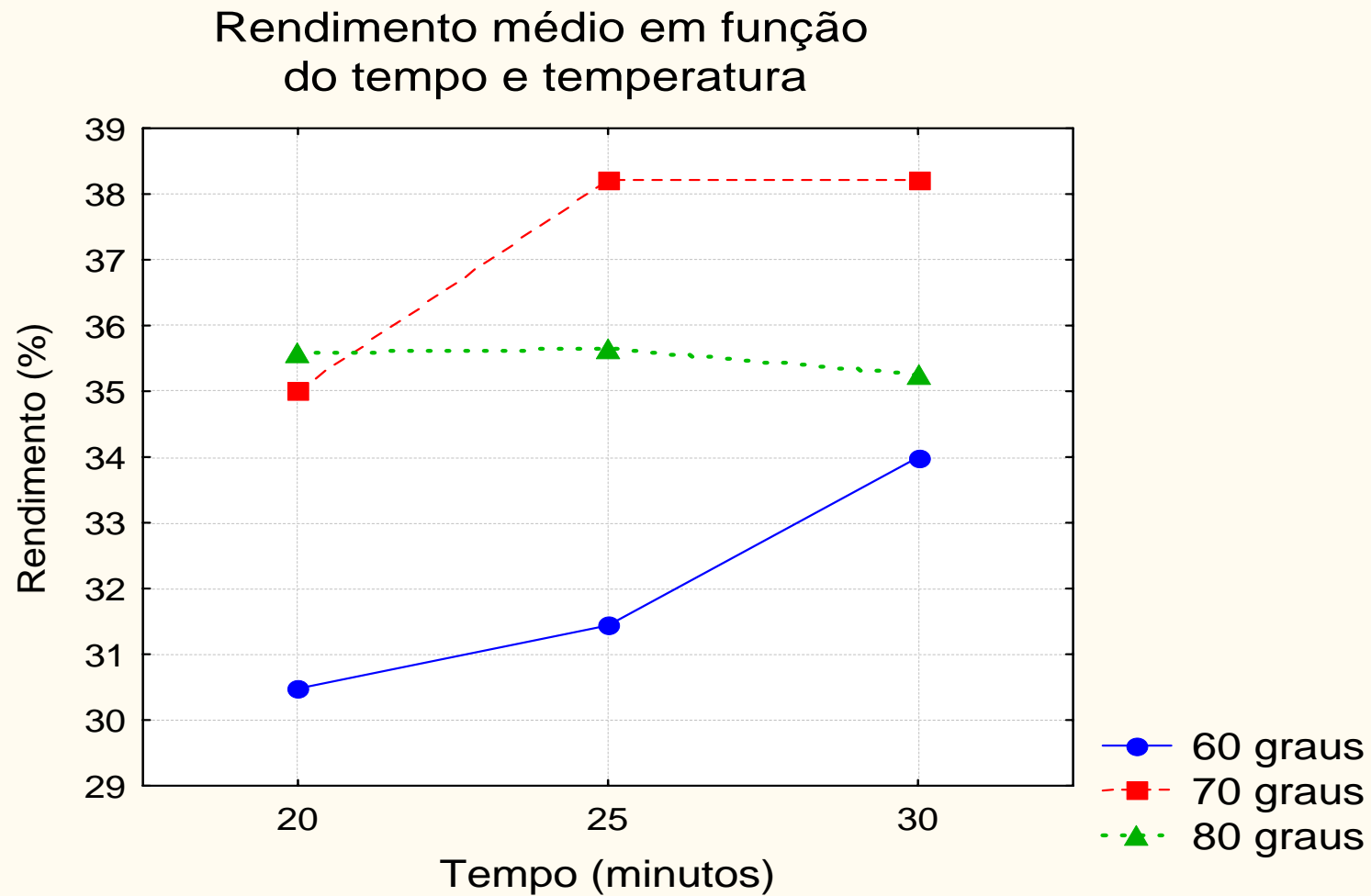
Temperatura (°C)	Tempo (minutos)								
	20			25			30		
60	29,7	28,7	30,2	31,0	30,6	32,8	32,9	32,7	34,8
	31,3	31,2	31,7	31,9	31,2	31,2	34,9	33,8	34,9
70	36,6	35,7	35,3	35,7	40,4	41,7	34,8	36,8	37,4
	35,1	30,2	37,2	36,9	34,5	40,0	38,9	38,7	42,5
80	40,2	33,6	33,4	37,0	34,4	29,8	36,0	31,3	36,6
	35,2	38,1	33,0	33,9	43,2	35,5	32,5	39,2	35,9

Exemplo

Médias aritméticas do rendimento, para diferentes níveis de temperatura e tempo de reação, num processo químico.

Temperatura (°C)	Tempo (minutos)		
	20	25	30
60	30,5	31,4	34,0
70	35,0	38,2	38,2
80	35,6	35,6	35,3

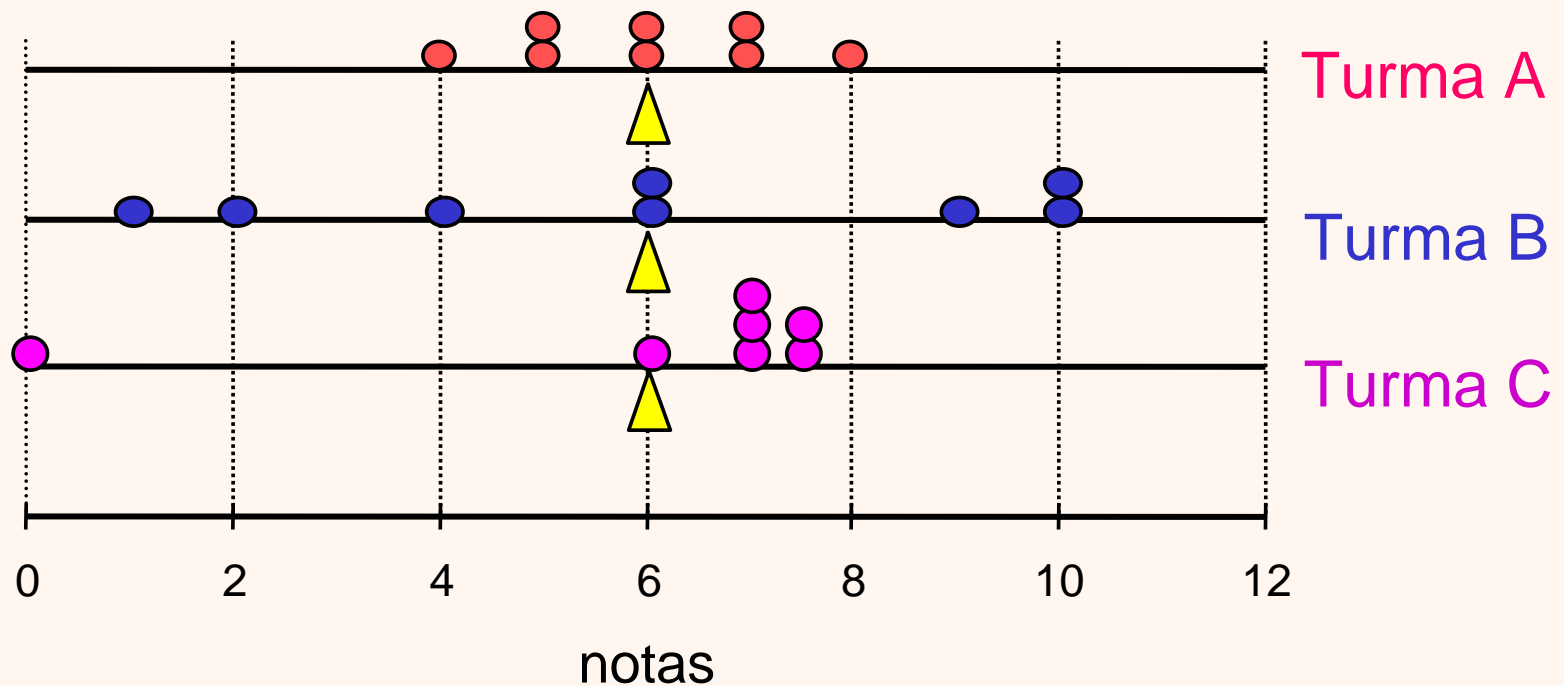
Exemplo



Exemplo: notas dos alunos de três turmas

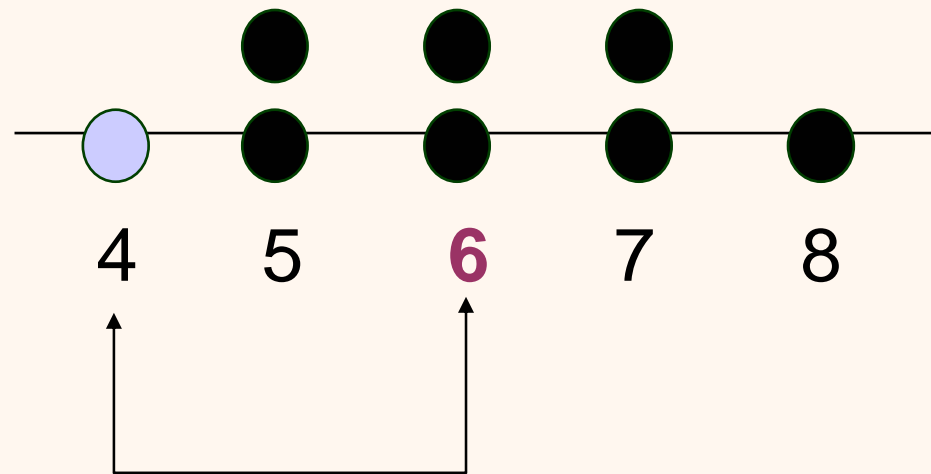
Turma	Notas dos alunos								Média da turma
A	4	5	5	6	6	7	7	8	6,00
B	1	2	4	6	6	9	10	10	6,00
C	0	6	7	7	7	7,5	7,5		6,00

Exemplo: notas dos alunos de três turmas



Como medir a dispersão?

Exemplo: Turma A (4 5 5 6 6 7 7 8)



distância (desvio) em relação à média

Como medir a dispersão?

Descrição	notação	resultados numéricos
Valores (notas dos alunos)	x_i	4 5 5 6 6 7 7 8
Média	\bar{x}	6
Desvios em relação à média	$x_i - \bar{x}$	-2 -1 -1 0 0 1 1 2
Desvios quadráticos	$(x_i - \bar{x})^2$	4 1 1 0 0 1 1 4

Variância (da amostra):

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

$$s^2 = \frac{4+1+1+0+0+1+1+4}{8-1} = 1,71$$

Como medir a dispersão?

Descrição	notação	resultados numéricos
Valores (notas dos alunos)	x_i	4 5 5 6 6 7 7 8
Média	\bar{x}	6
Desvios em relação à média	$x_i - \bar{x}$	-2 -1 -1 0 0 1 1 2
Desvios quadráticos	$(x_i - \bar{x})^2$	4 1 1 0 0 1 1 4

Desvio padrão (da amostra):

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

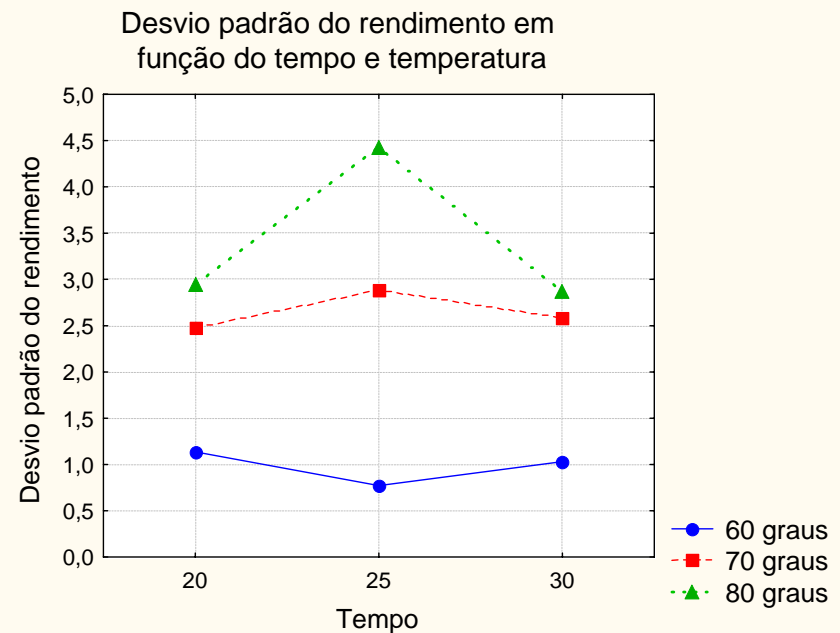
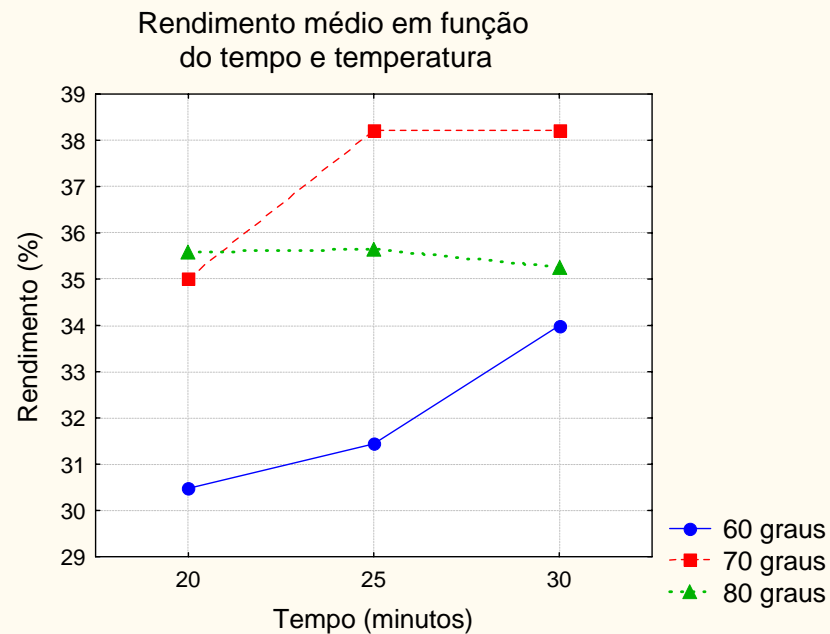
$$s = \sqrt{\frac{4+1+1+0+0+1+1+4}{8-1}} = \sqrt{1,71} = 1,31$$

Medidas descritivas das notas finais dos alunos de três turmas.

Turma	Número de alunos	Média	Desvio padrão
A	8	6,00	1,31
B	8	6,00	3,51
C	7	6,00	2,69

Interprete.

Ex: Rendimento de um processo químico



Interprete.

Outra forma de calcular o desvio padrão

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

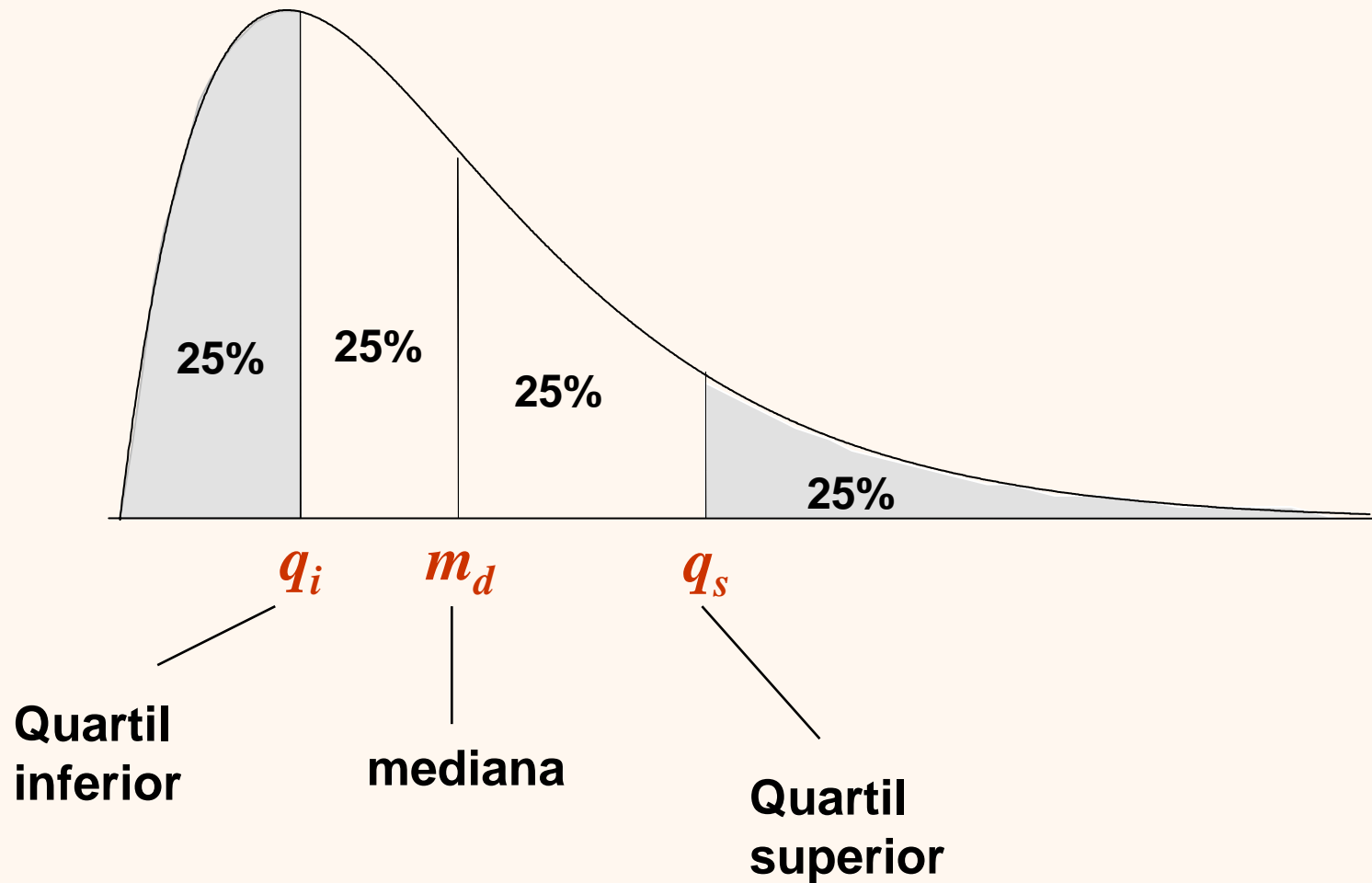
Valores x_i : 4 5 5 6 6 7 7 8

Valores ao quadrado x_i^2 : 16 25 25 36 36 49 49 64

$$\sum_{i=1}^n x_i = 48 \quad \bar{x} = 6 \quad \sum_{i=1}^n x_i^2 = 300$$

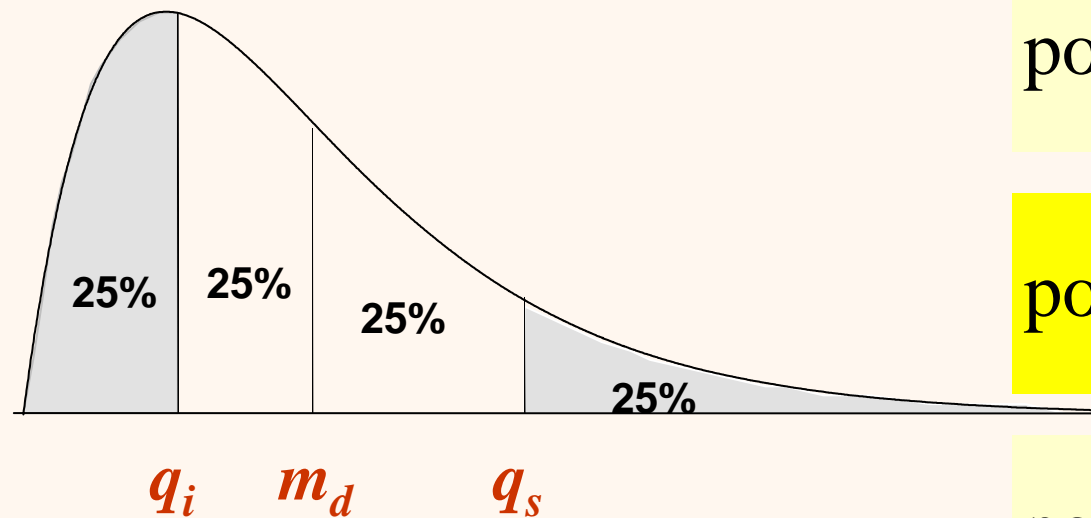
$$s = \sqrt{\frac{300 - 8 \cdot (6)^2}{7}} = \sqrt{\frac{300 - 288}{7}} = \sqrt{\frac{12}{7}} = 1,31$$

Medidas baseadas na ordenação dos dados



Medidas baseadas na ordenação dos dados

Dados ordenados:



$$\text{posição de } q_i : \frac{n+1}{4}$$

$$\text{posição de } m_d : \frac{n+1}{2}$$

$$\text{posição de } q_s : \frac{3(n+1)}{4}$$

Se fracionário → interpolação linear

Exemplo

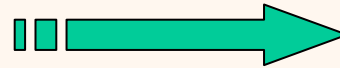
Observações: 15, 18, 5, 7, 9, 11, 3, 5, 6, 8, 12.

Ordenando:

3 5 5 6 7 8 9 11 12 15 18

$n = 11$

posição de q_i : $\frac{n+1}{4} = 3$



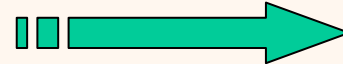
$q_i = 5$

posição de m_d : $\frac{n+1}{2} = 6$



$m_d = 8$

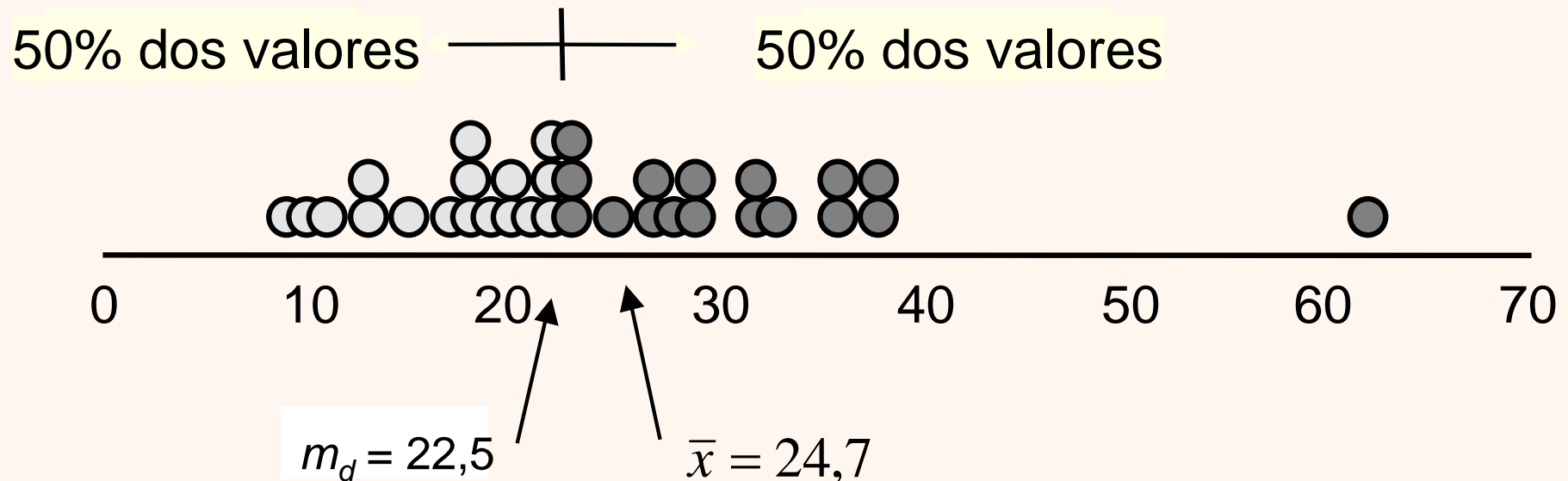
posição de q_s : $\frac{3(n+1)}{4} = 9$



$q_s = 12$

Comparação entre média e mediana

- A média é mais influenciada por valores discrepantes.



Comparação entre média e mediana

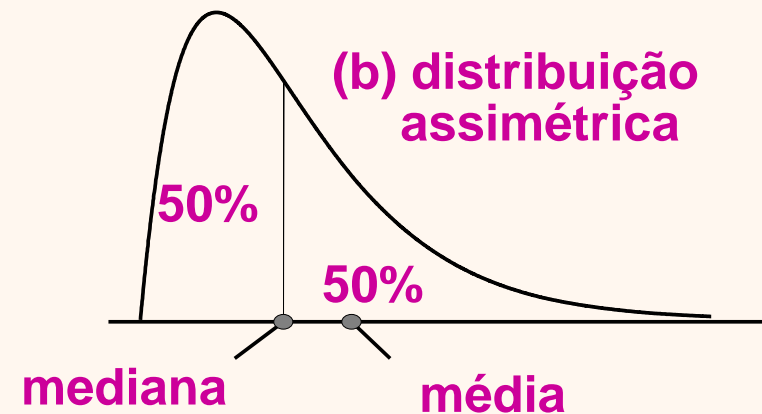
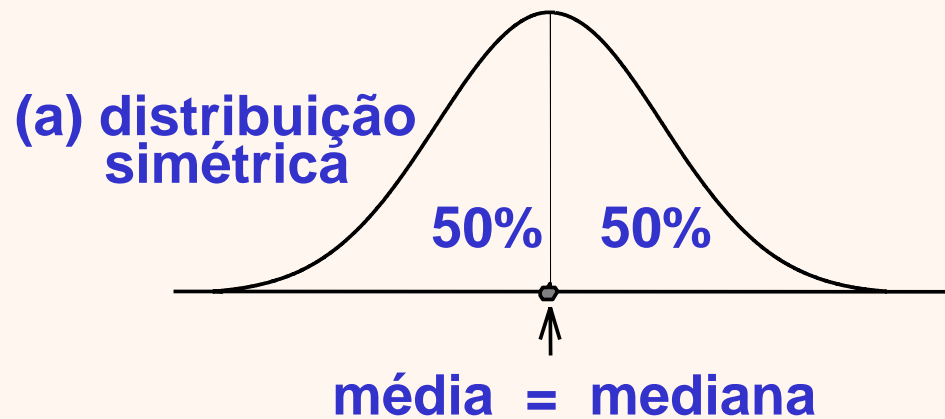


Diagrama em caixas

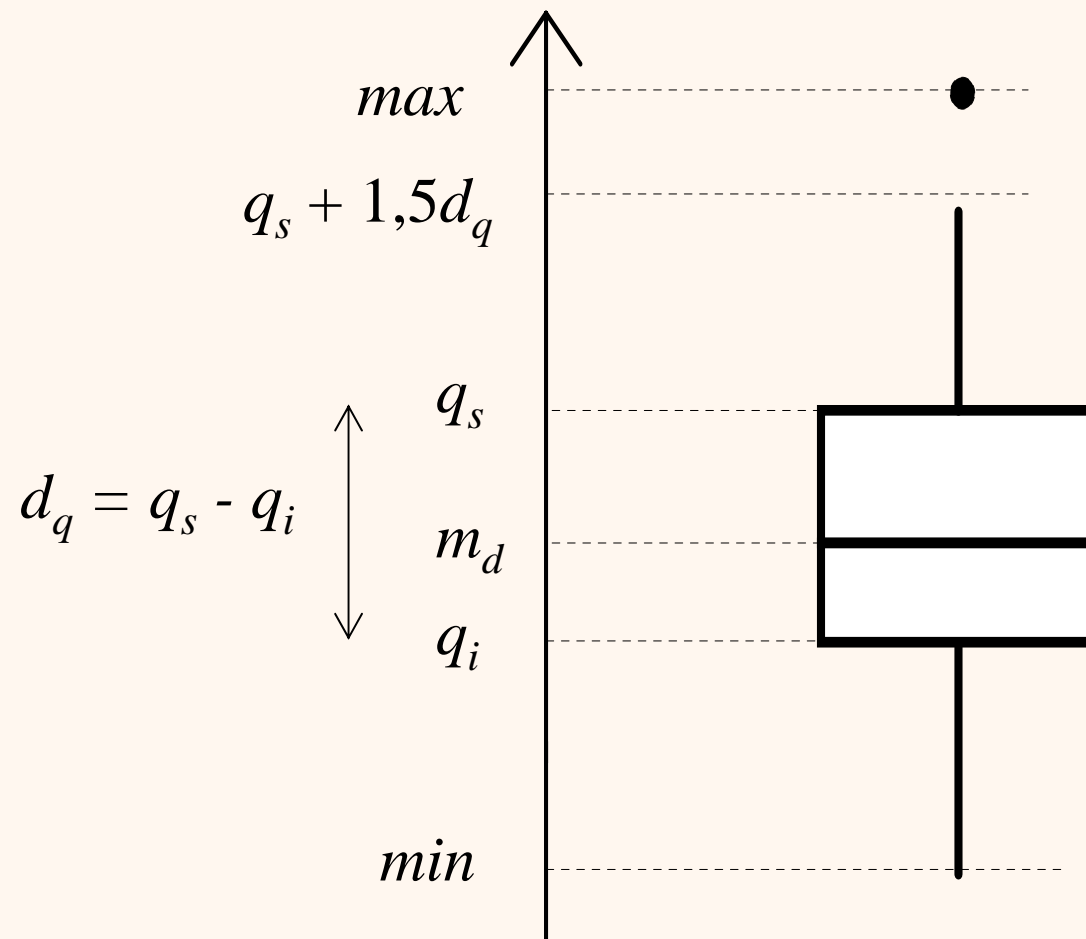
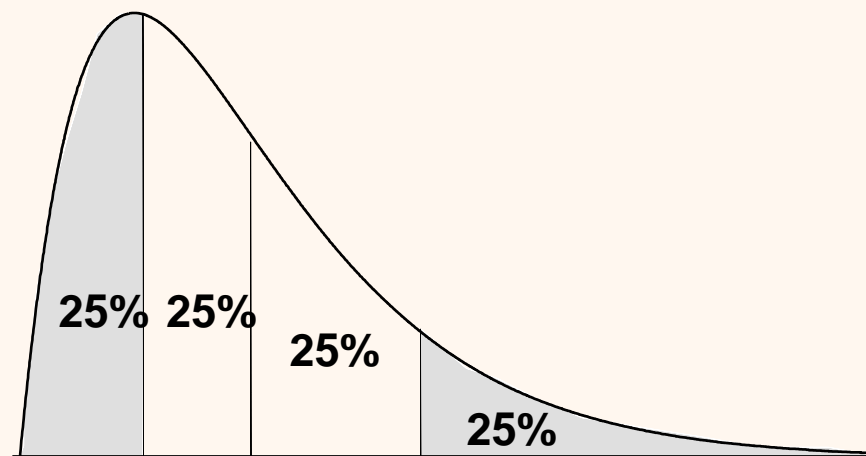
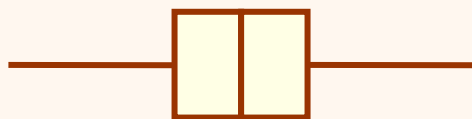
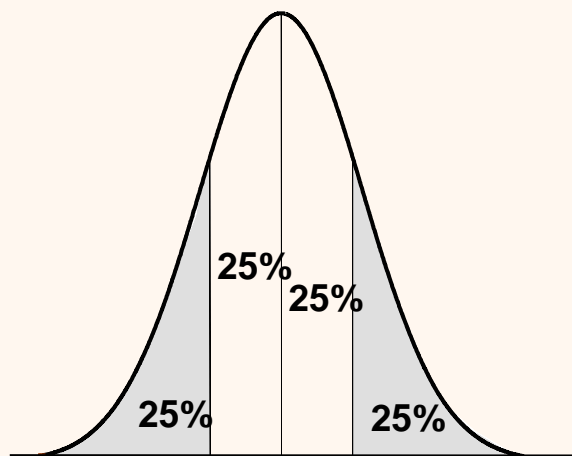
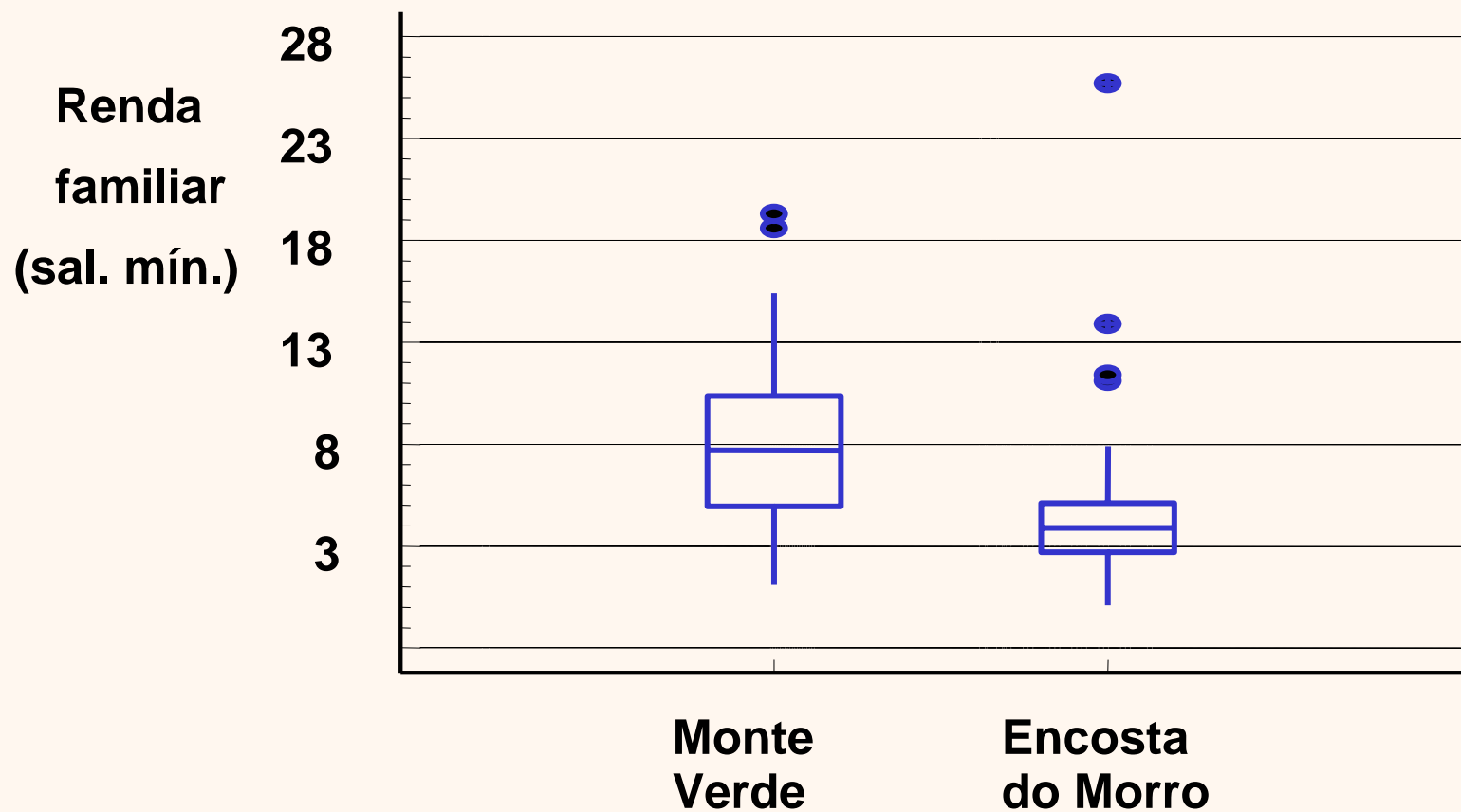


Diagrama em caixas e forma da distribuição



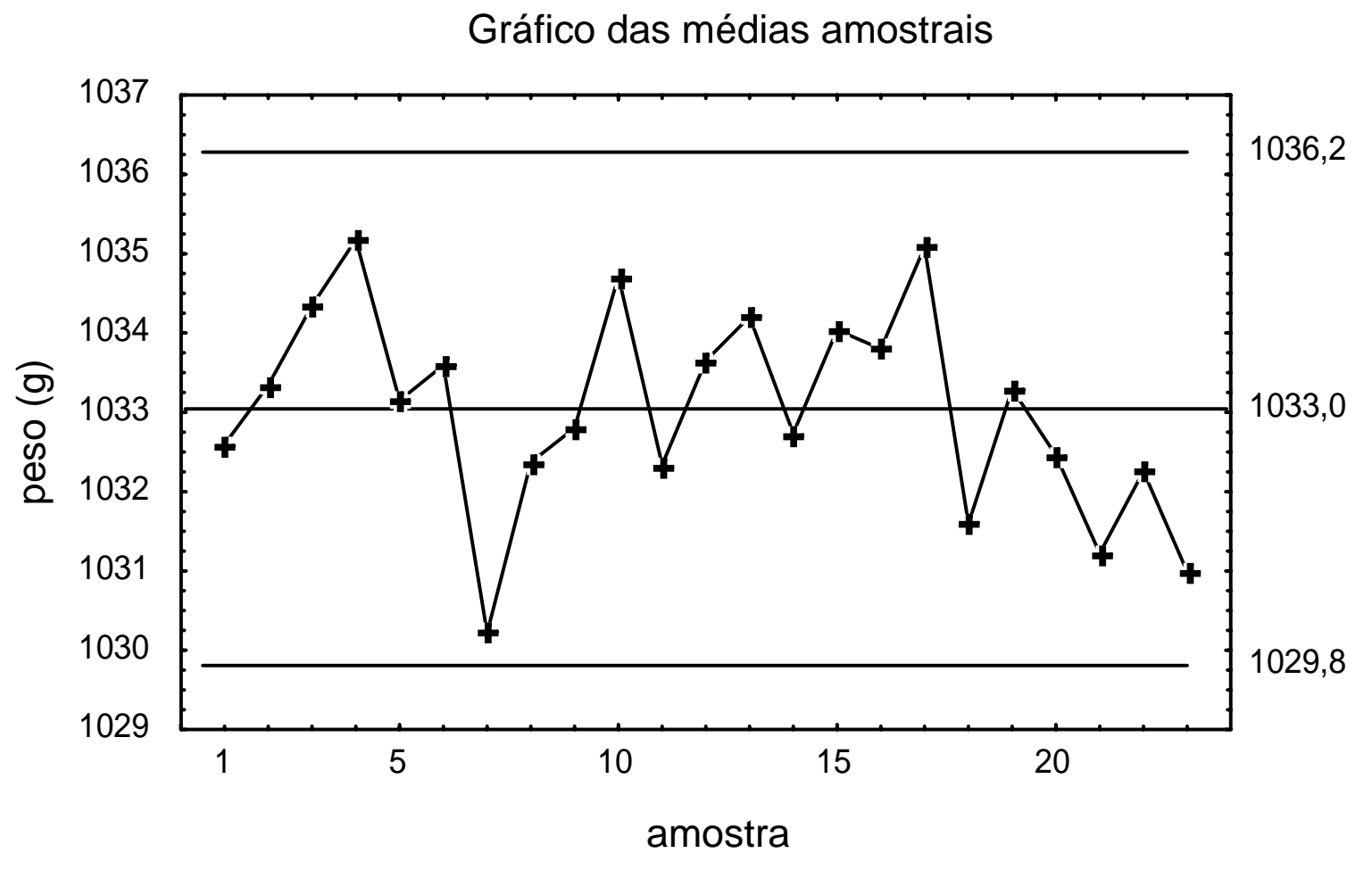
Interprete o gráfico



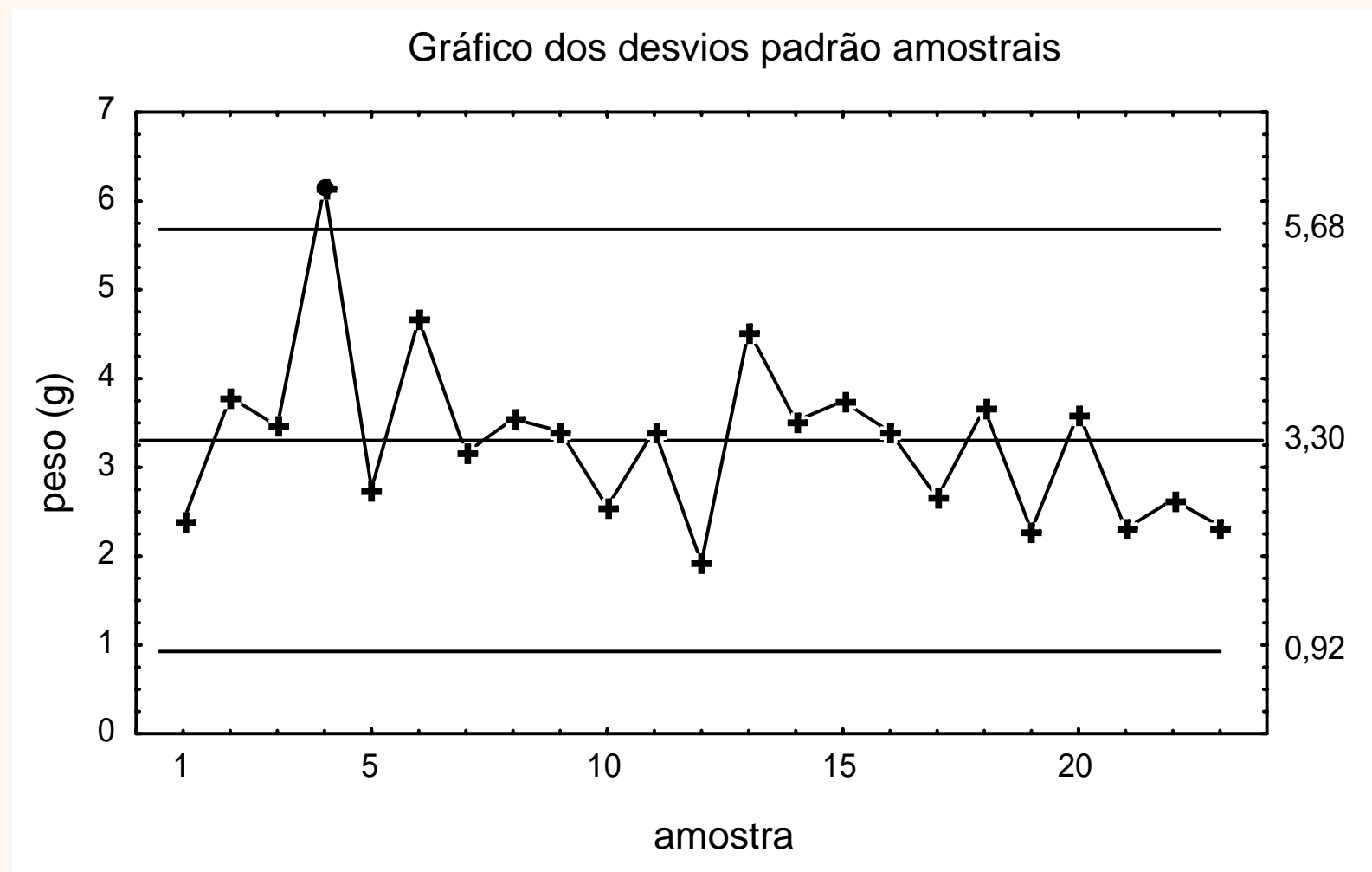
Observações ao longo do tempo

- EXEMPLO: todos os dias é retirada uma amostra de dez sacos de leite de um laticínio, durante 23 dias.
- Quer-se acompanhar o nível e a variabilidade do peso.

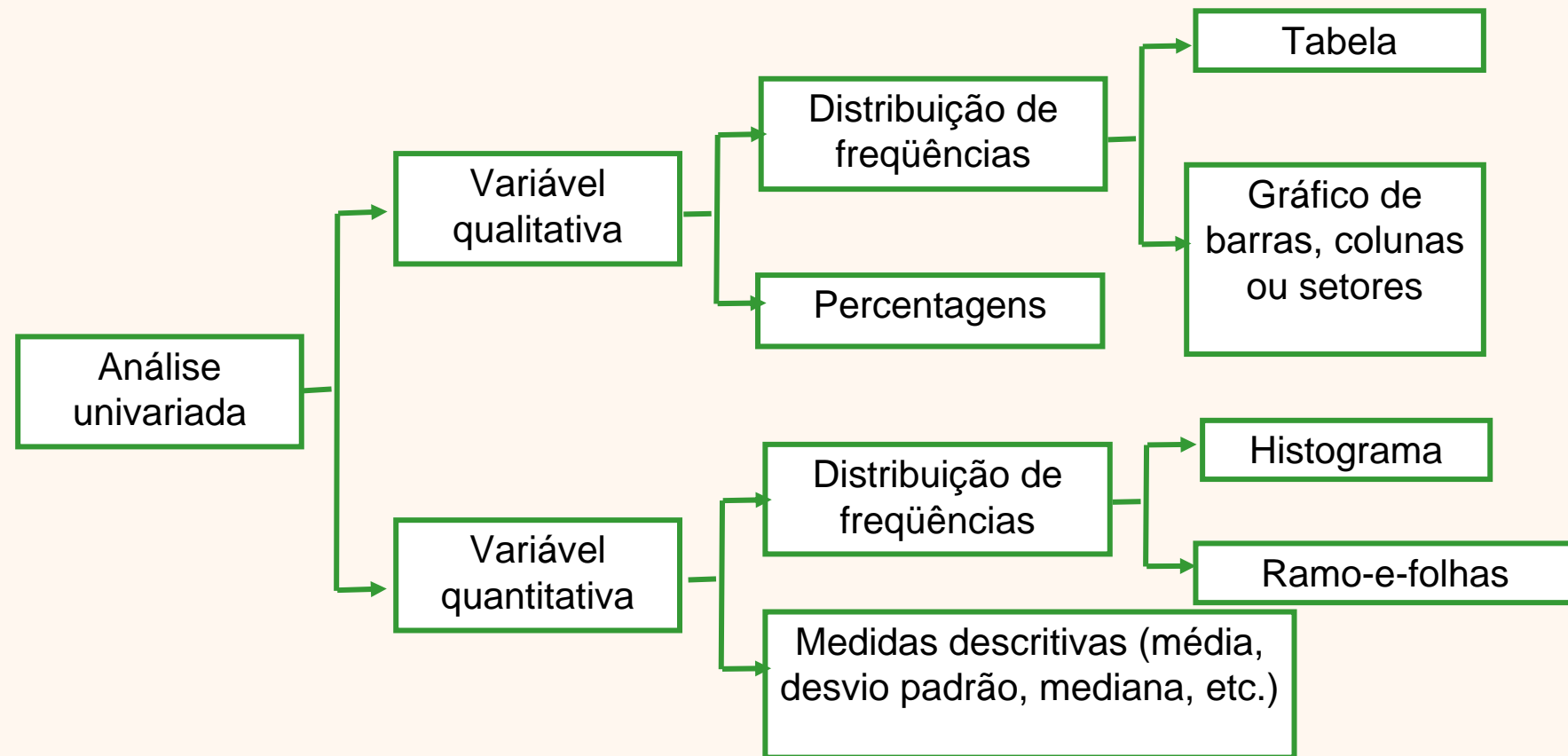
Observações ao longo do tempo



Observações ao longo do tempo



Orientação geral para análise exploratória de dados não temporais



Orientação geral para análise exploratória de dados não temporais

