# Análise Exploratória de Dados

Prof. Me. Ricardo Ávila
ricardo.avila@outlook.com.br

# Conhecendo os dados

O objetivo da análise exploratória de dados é examinar a estrutura subjacente dos dados e aprender sobre os relacionamentos sistemáticos entre muitas variáveis.

A análise exploratória de dados inclui um conjunto de ferramentas gráficas e descritivas, para explorar os dados, como pré-requisito para uma análise de dados mais formal (Predição, Previsão, Estimação, Classificação e Testes de Hipóteses), e como parte integral formal da construção de modelos.

# Análise Exploratória de Dados

A AED facilita a descoberta de conhecimento não esperado, como também ajuda a confirmar o esperado.

Como uma importante etapa em Data Mining, a AED emprega técnicas estatísticas descritivas e gráficas para estudar o conjunto de dados, detectando *outliers* e anomalias, e testando as suposições do modelo.

A AED é um importante pré-requisito para se alcançar o sucesso em qualquer projeto de data mining.

# Distribuições de Frequências

- organização dos dados de acordo com as ocorrências dos diferentes resultados observados.
  - Pode ser apresentada: em tabela ou em gráfico;
  - com freqüências absolutas, relativas ou porcentagens.

# Exemplo (com variável qualitativa)

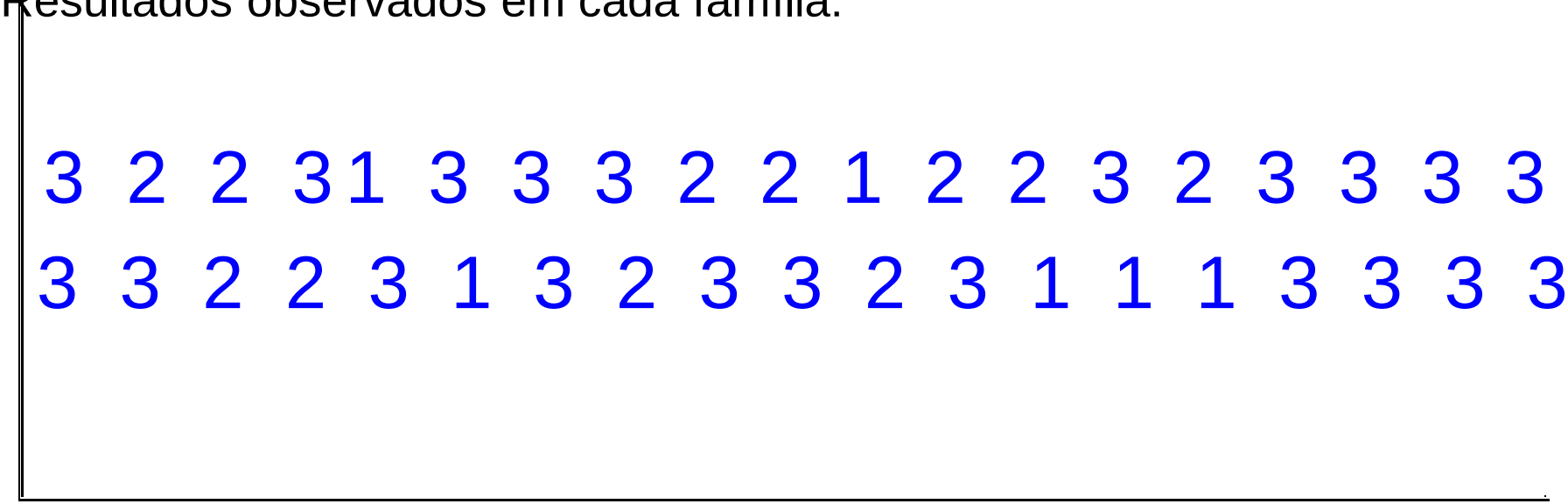
Grau de instrução do chefe da casa, numa amostra de 40 famílias do Conjunto Residencial Monte Verde, Florianópolis, SC, 1988.

Códigos: 1 – Nenhum grau de instrução completo;

2 – Primeiro grau completo;

3 – Segundo grau completo.

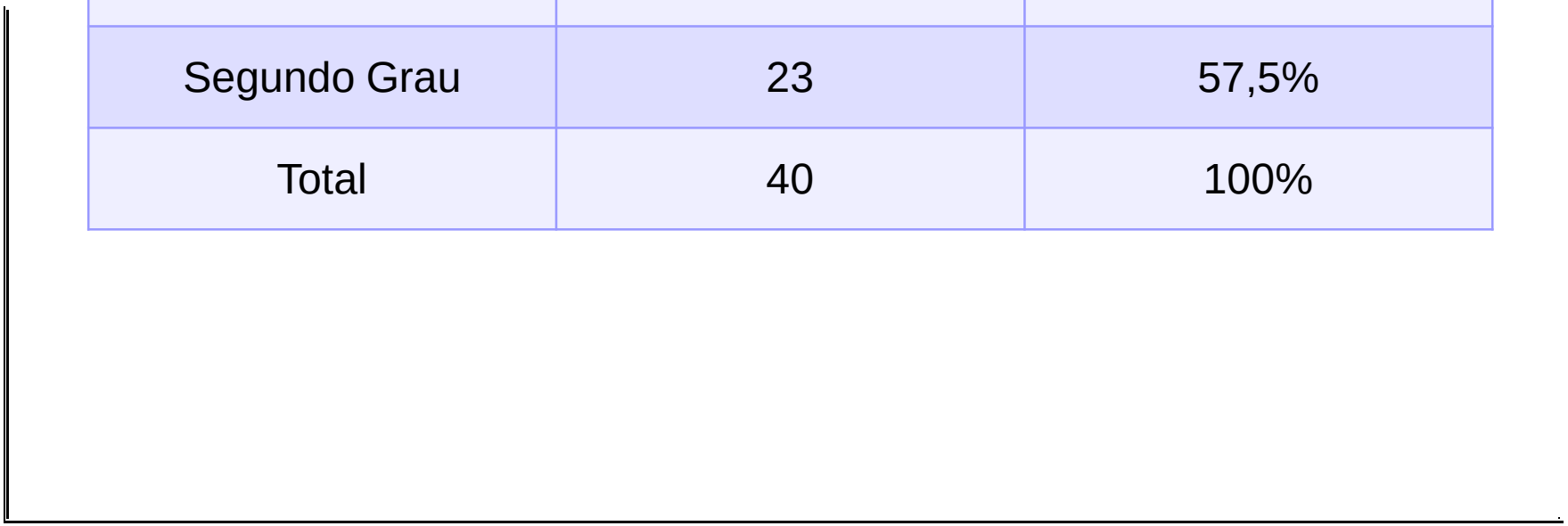
Resultados observados em cada família:



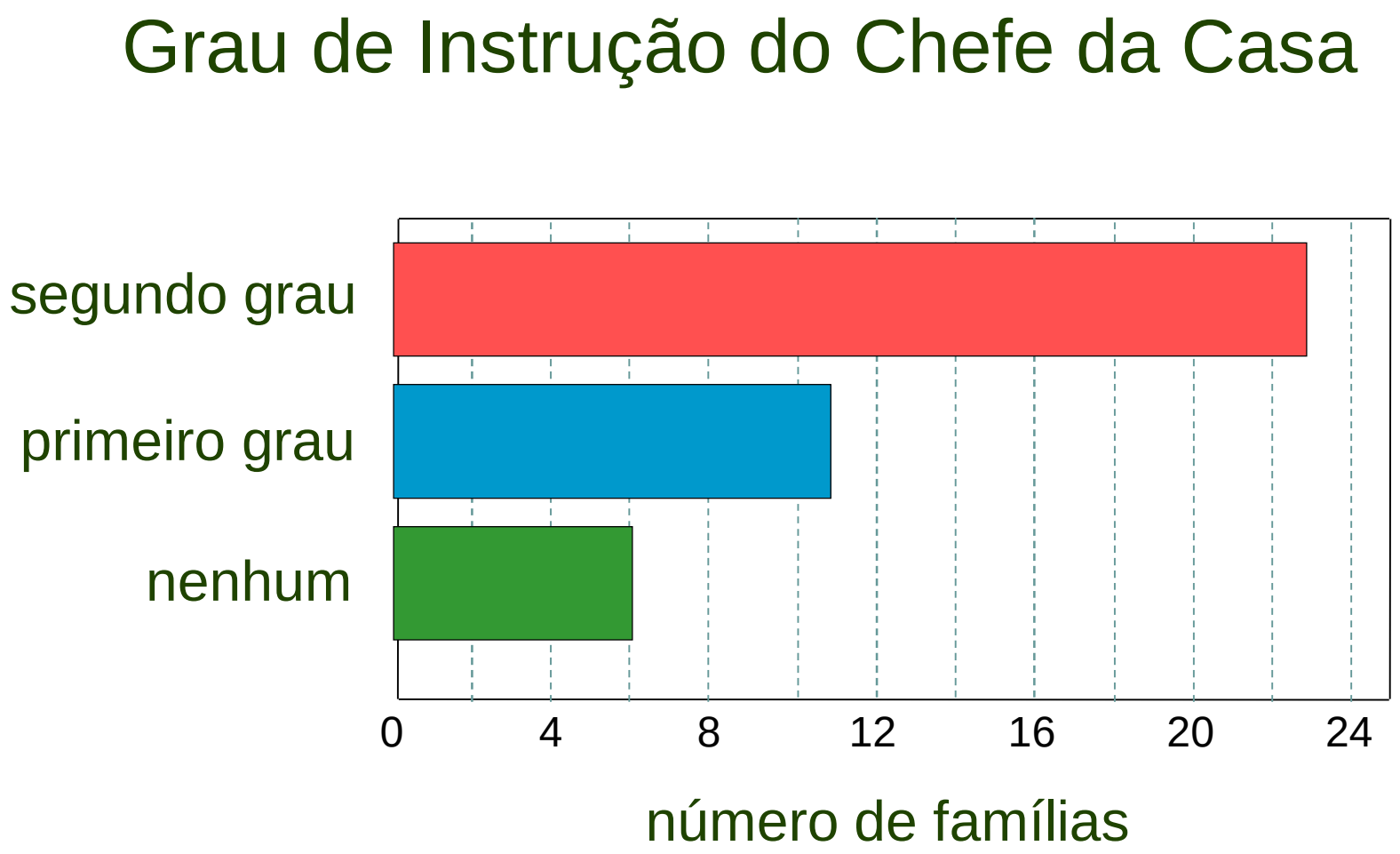
# Exemplo (com variável qualitativa)

Grau de instrução (Conjunto Residêncial Monte Verde)

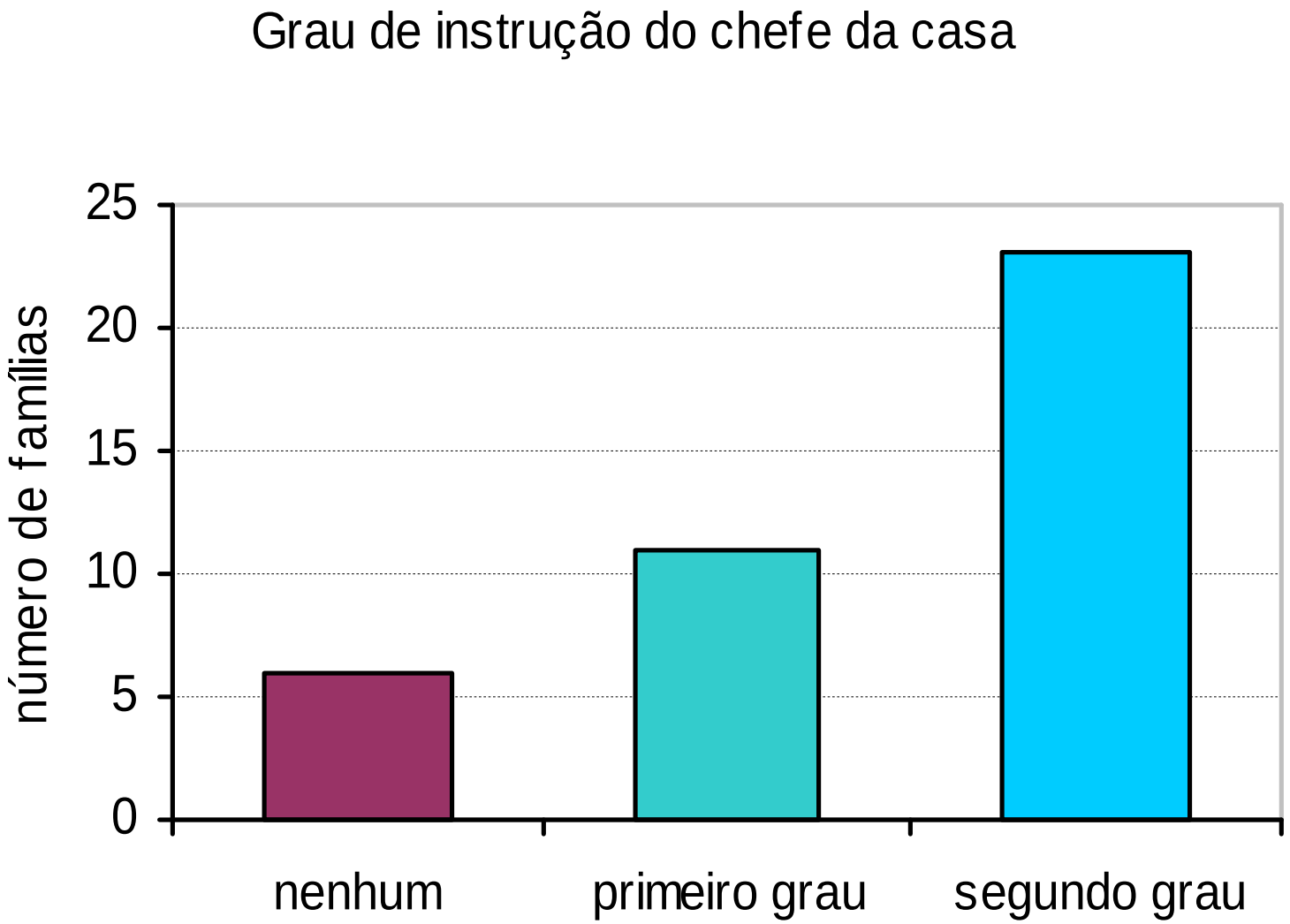
| Grau de instrução | Freqüência | Percentagem |
|---|---|---|
| Nenhum | 6 | 15% |
| Primeiro Grau | 11 | 27,5% |
| Segundo Grau | 23 | 57,5% |
| Total | 40 | 100% |



# Gráfico de Barras



# Gráfico em Colunas



# Gráfico de Setores (Proporções)

# Exemplo (com variável discreta)

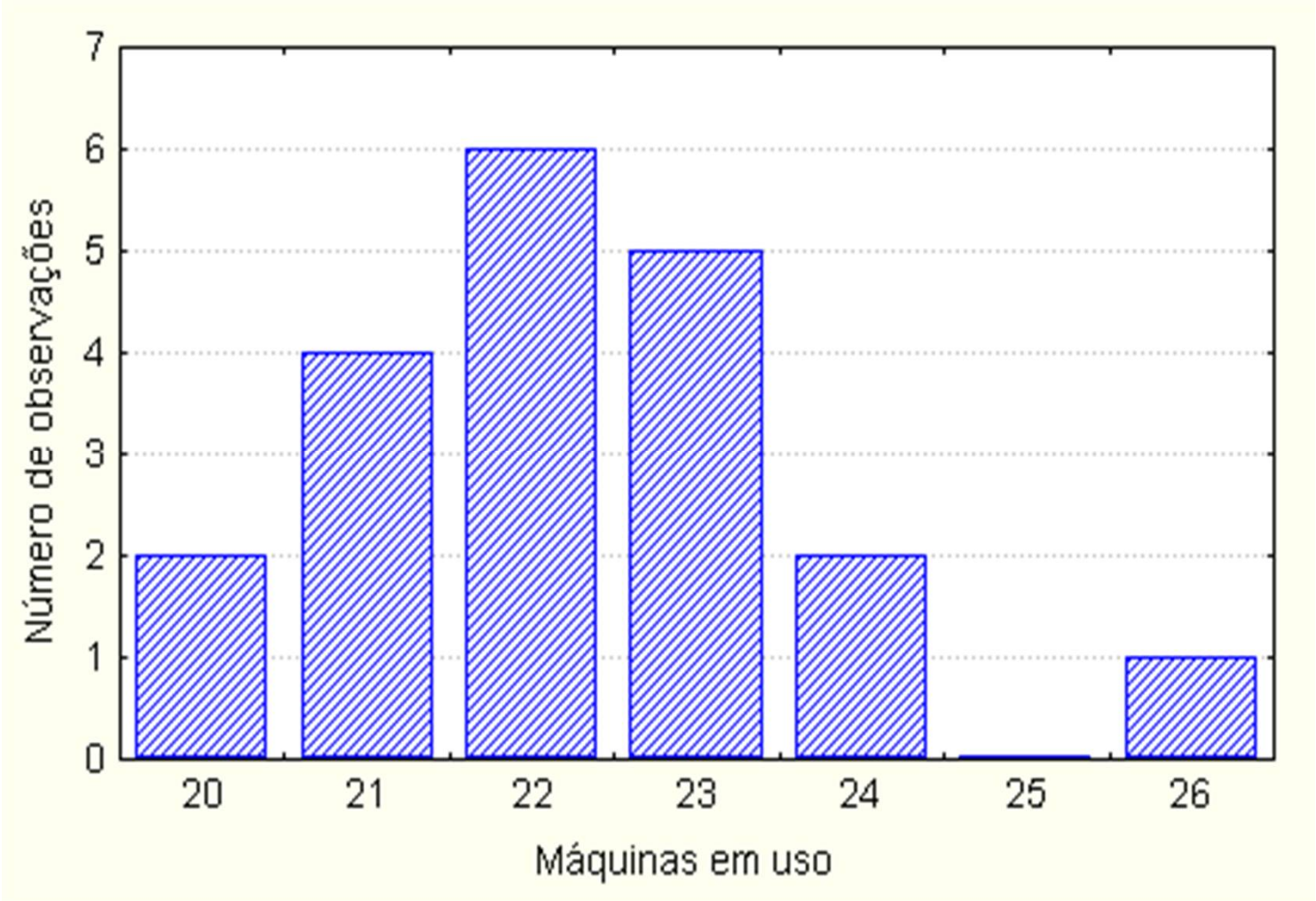
- Numa rede de computadores, a quantidade de máquinas ligadas, por dia

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | 26 | 21 | 21 | 20 | 21 | 23 | 22 | 24 | 22 |
| 22 | 22 | 23 | 23 | 23 | 22 | 23 | 22 | 24 | 21 |

# Distribuição de Frequências????

| Máquinas em uso | Freqüência (absoluta) | Proporção (%) |
|---|---|---|
| 20 | 2 | 0,10 (10%) |
| 21 | 4 | 0,20 (20%) |
| 22 | 6 | 0,30 (30%) |
| 23 | 5 | 0,25 (25%) |
| 24 | 2 | 0,10 (10%) |
| 25 | 0 | 0,00 (0 %) |
| 26 | 1 | 0,05 ( 5%) |
| Total | 20 | 1,00 (100%) |

# Gráfico de Colunas



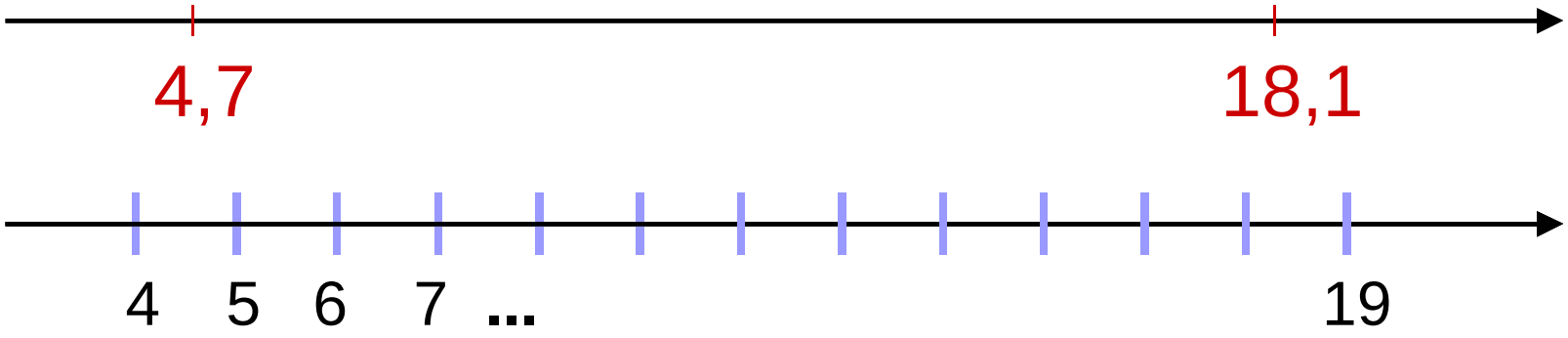
# Exemplo (com variável contínua)

Tempo (em segundos) para carga de um aplicativo num sistema compartilhado (50 observações):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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 | 18,1 | 5,3 |
| 4,9 | 5,0 | 5,7 | 6,3 | 6,0 | 6,8 | 7,3 | 6,9 |
| 6,5 | 5,9 | | | | | | |

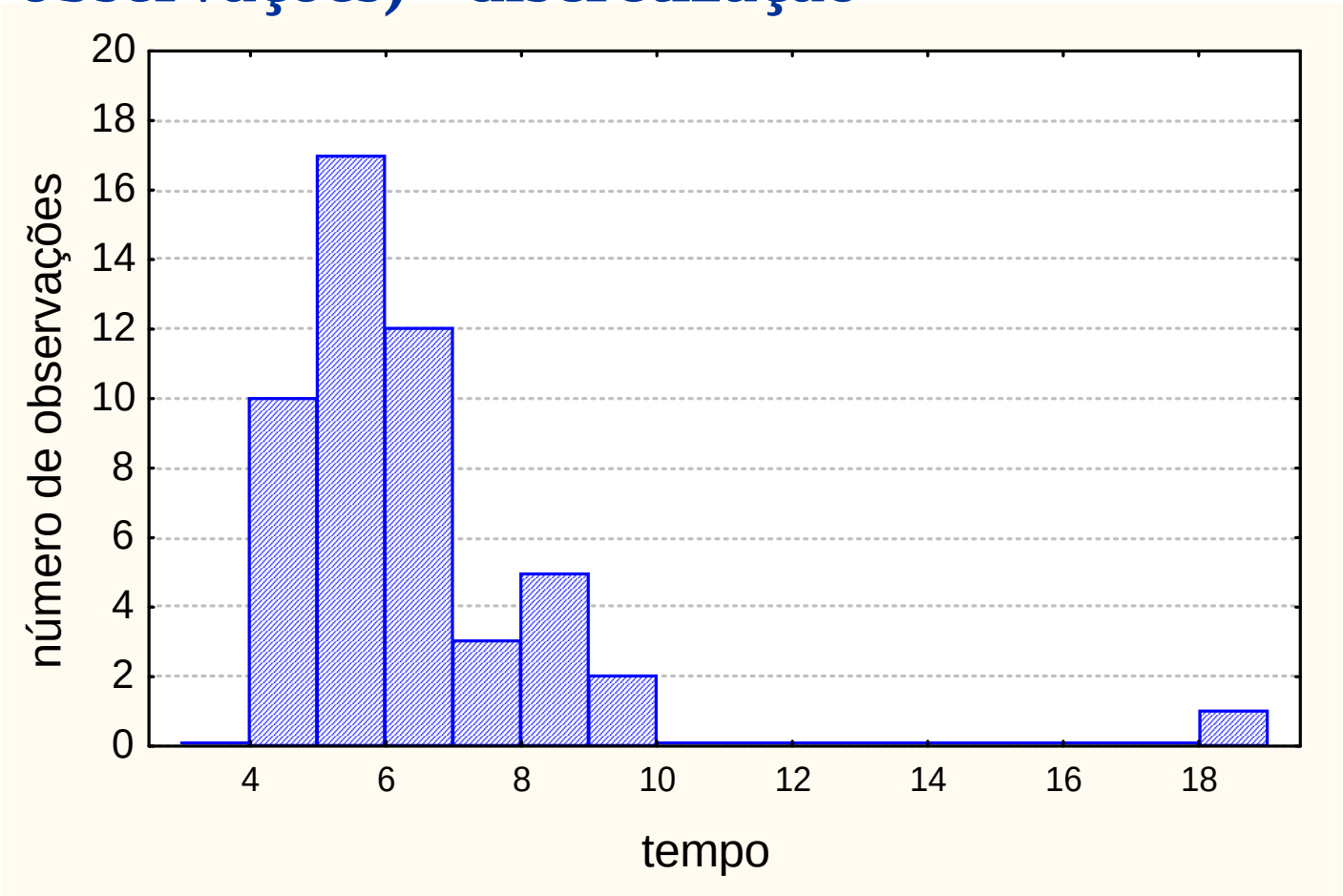
# DADOS

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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 | 18,1 | 5,3 |
| 4,9 | 5,0 | 5,7 | 6,3 | 6,0 | 6,8 | 7,3 | 6,9 |
| 6,5 | 5,9 | | | | | | |

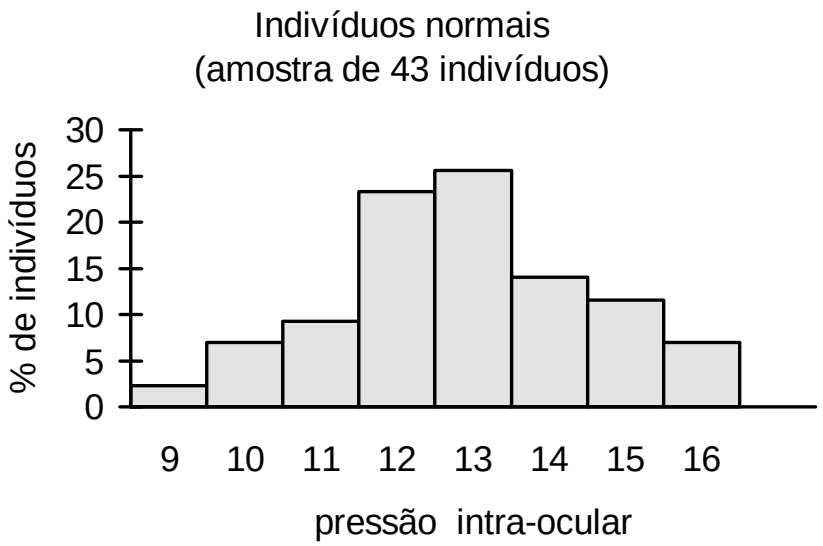


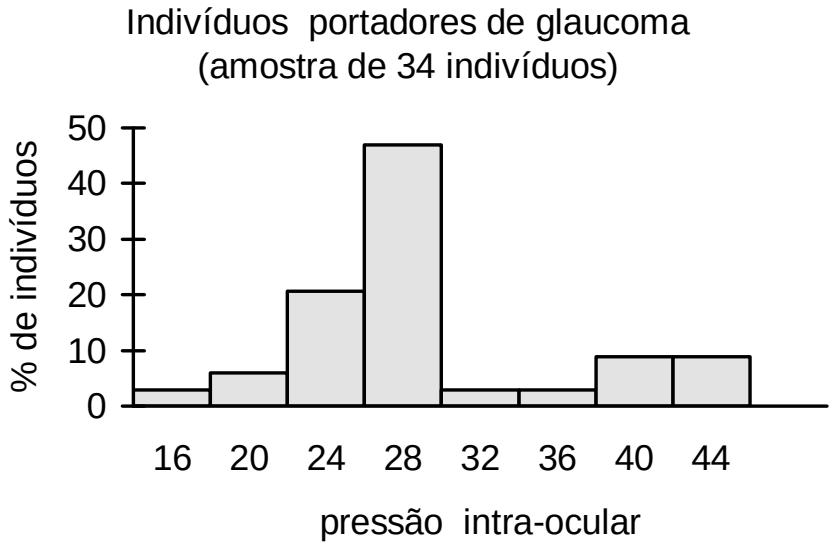
# Histograma

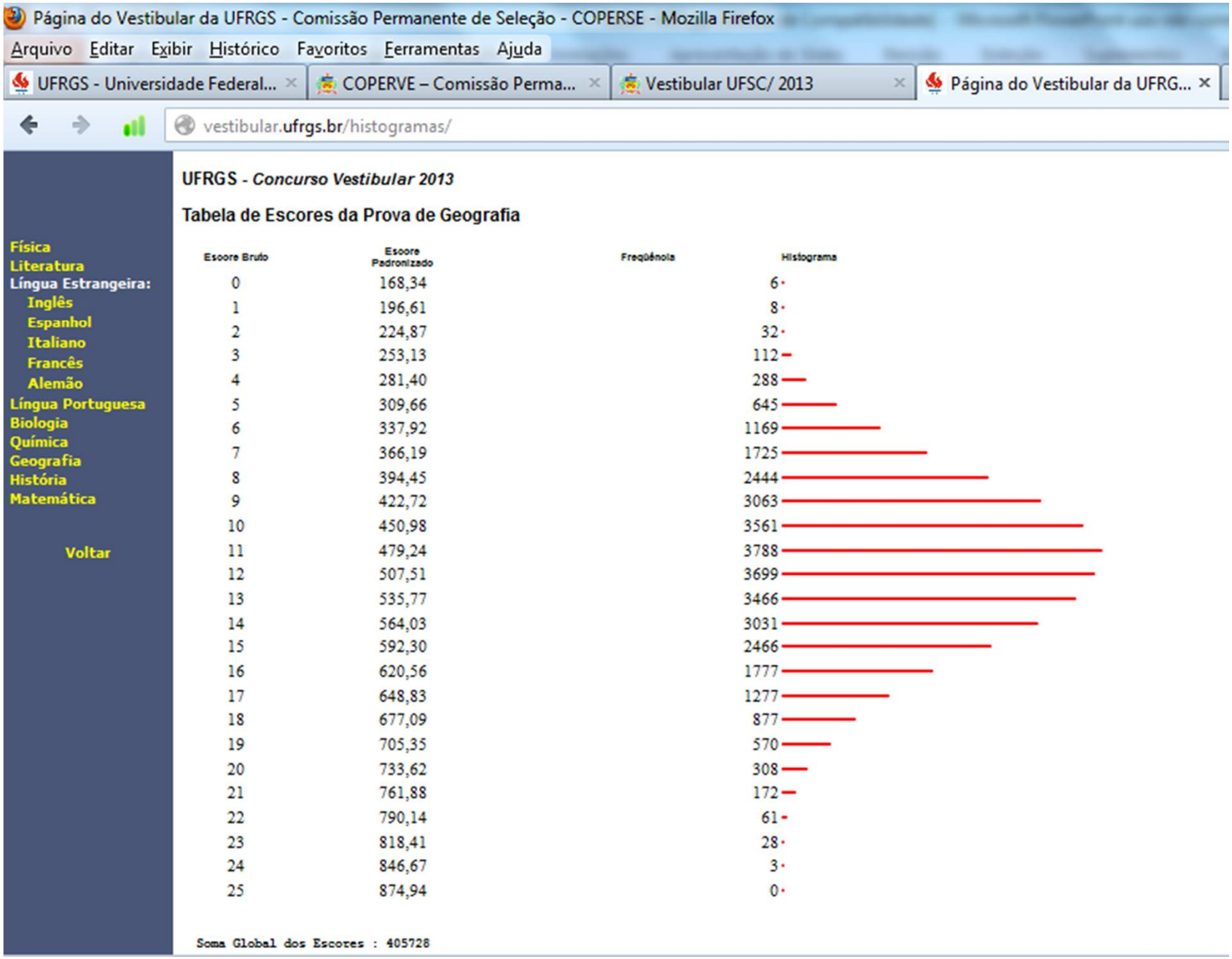
**Histograma do tempo (em segundos) para carga de um aplicativo num sistema compartilhado (50 observações) - discretização**

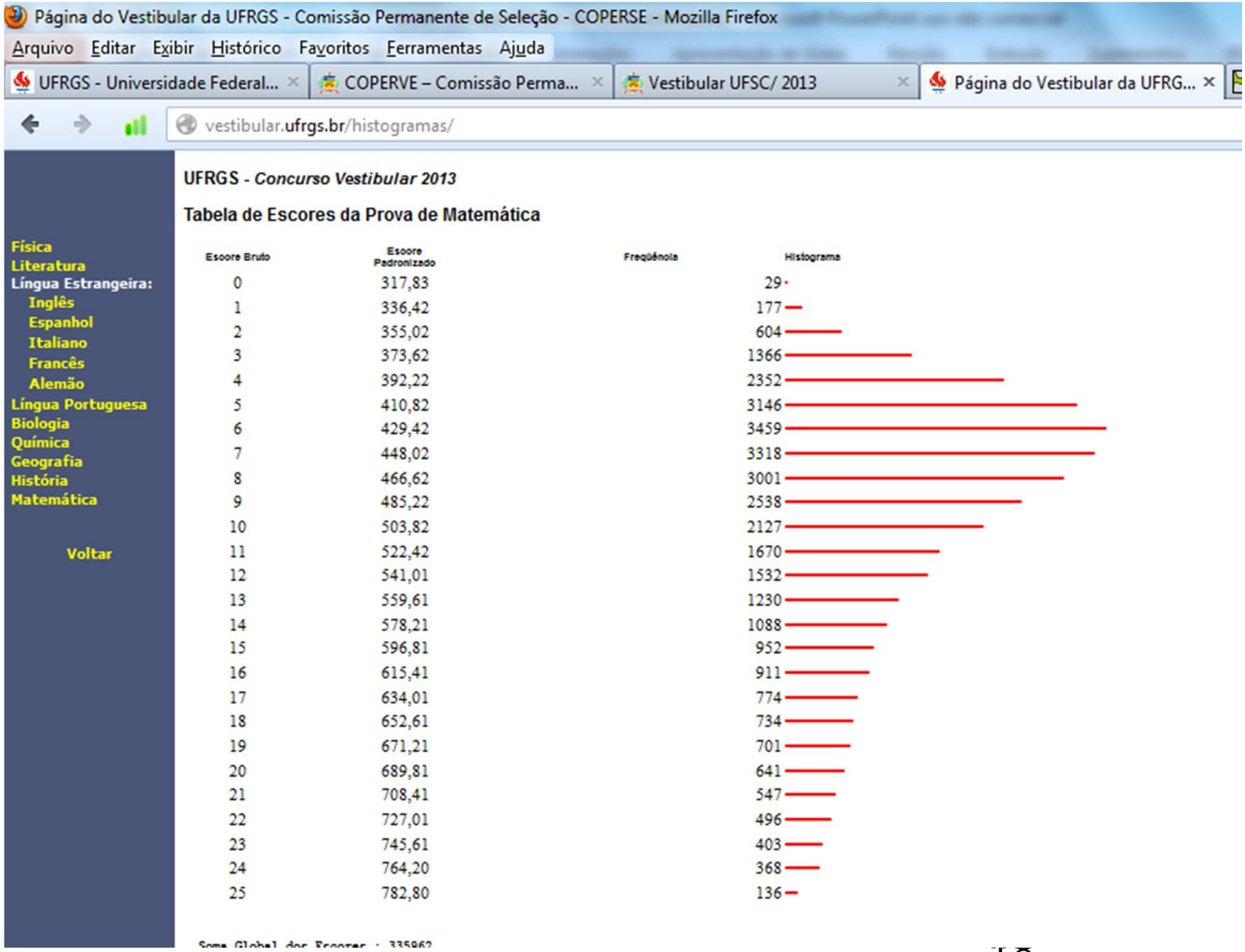


# Histograma


Indivíduos normais
(amostra de 43 indivíduos)
% de indivíduos
30
25
20
15
10
5
0
9
10
11
12
13
14
15
16
pressão intra-ocular


Indivíduos portadores de glaucoma
(amostra de 34 indivíduos)
% de indivíduos
50
40
30
20
10
0
16
20
24
28
32
36
40
44
pressão intra-ocular


Página do Vestibular da UFRGS - Comissão Permanente de Seleção - COPERSE - Mozilla Firefox
Arquivo Editar Exibir Histórico Favoritos Ferramentas Ajuda
UFRGS - Universidade Federal...
COPERVE – Comissão Perma...
Vestibular UFSC/ 2013
Página do Vestibular da UFRG...
vestibular.ufrgs.br/histogramas/
Física
Literatura
Língua Estrangeira:
Inglês
Espanhol
Italiano
Francês
Alemão
Língua Portuguesa
Biologia
Química
Geografia
História
Matemática
Voltar
UFRGS - Concurso Vestibular 2013
Tabela de Escores da Prova de Geografia
Escore Bruto
Escore Padronizado
Freqüência
Histograma
0 168,34 6
1 196,61 8
2 224,87 32
3 253,13 112
4 281,40 288
5 309,66 645
6 337,92 1169
7 366,19 1725
8 394,45 2444
9 422,72 3063
10 450,98 3561
11 479,24 3788
12 507,51 3699
13 535,77 3466
14 564,03 3031
15 592,30 2466
16 620,56 1777
17 648,83 1277
18 677,09 877
19 705,35 570
20 733,62 308
21 761,88 172
22 790,14 61
23 818,41 28
24 846,67 3
25 874,94 0
Soma Global dos Escores : 405728


Página do Vestibular da UFRGS - Comissão Permanente de Seleção - COPERSE - Mozilla Firefox
Arquivo Editar Exibir Histórico Favoritos Ferramentas Ajuda
UFRGS - Universidade Federal...
COPERVE – Comissão Perma...
Vestibular UFSC/ 2013
Página do Vestibular da UFRG...
vestibular.ufrgs.br/histogramas/
Física
Literatura
Língua Estrangeira:
Inglês
Espanhol
Italiano
Francês
Alemão
Língua Portuguesa
Biologia
Química
Geografia
História
Matemática
Voltar
UFRGS - Concurso Vestibular 2013
Tabela de Escores da Prova de Matemática
Escore Bruto
Escore Padronizado
Freqüência
Histograma
0 317,83 29
1 336,42 177
2 355,02 604
3 373,62 1366
4 392,22 2352
5 410,82 3146
6 429,42 3459
7 448,02 3318
8 466,62 3001
9 485,22 2538
10 503,82 2127
11 522,42 1670
12 541,01 1532
13 559,61 1230
14 578,21 1088
15 596,81 952
16 615,41 911
17 634,01 774
18 652,61 734
19 671,21 701
20 689,81 641
21 708,41 547
22 727,01 496
23 745,61 403
24 764,20 368
25 782,80 136

# Medidas Descritivas

- Existem medidas quantitativas que servem para descrever, resumidamente, características das distribuições.
- As mais utilizadas são a **média** e o **desvio padrão**.

# Média (X)

- A média aritmética simples ( $\overline{X}$ ) é a soma dos valores dividida pelo número de observações.

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

# Exemplo

- Deseja-se estudar o número de falhas no envio de mensagens, considerando três algoritmos diferentes para o envio dos pacotes:

| | |
|---|---|
| Algoritmo A | (8 observações) |
| Algoritmo B | (8 observações) |
| Algoritmo C | (7 observações) |

# Exemplo

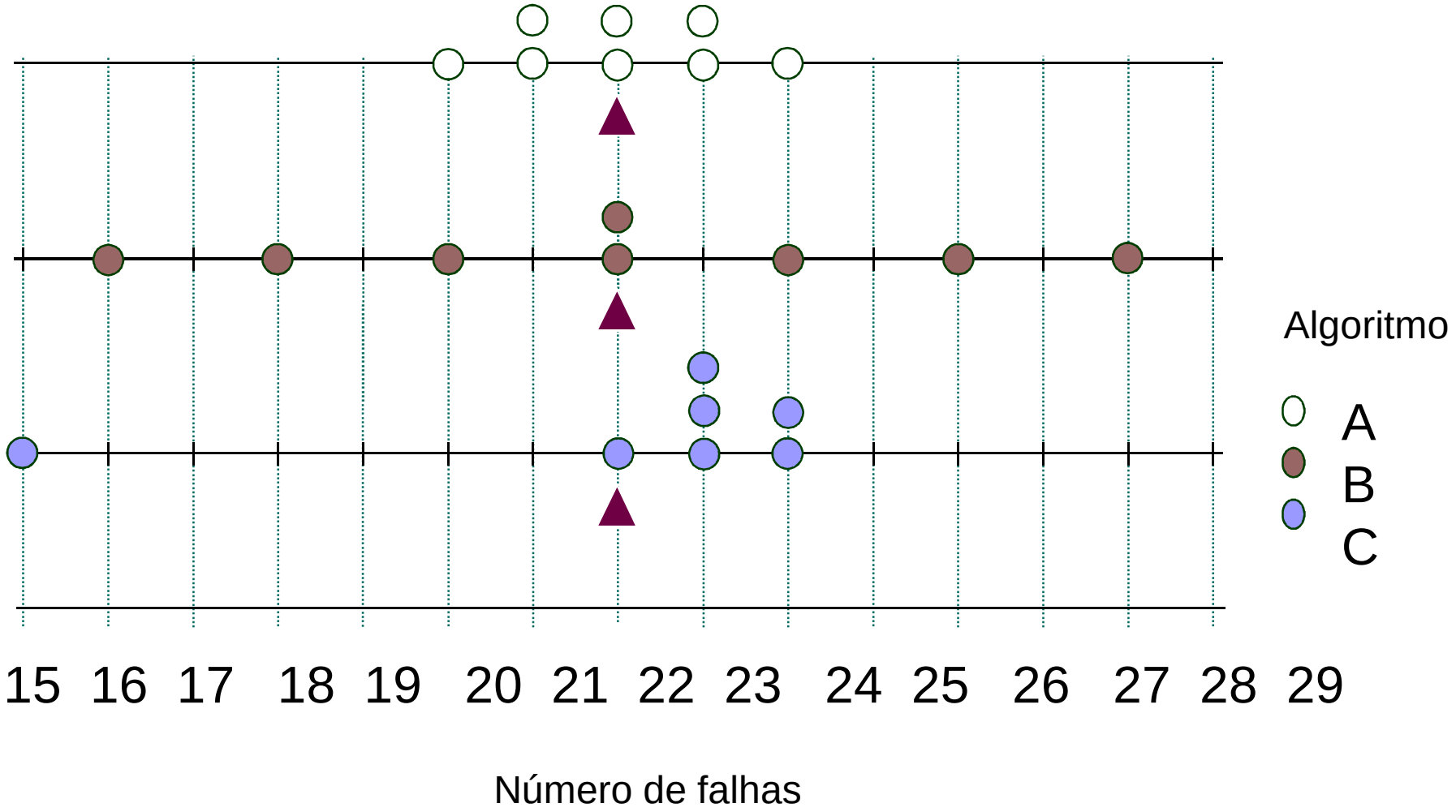
- Número de falhas a cada 10.000 mensagens enviadas.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A: | 20 | 21 | 21 | 22 | 22 | 23 | 23 | 24 |
| B: | 16 | 18 | 20 | 22 | 22 | 24 | 26 | 28 |
| C: | 15 | 22 | 23 | 23 | 23 | 24 | 24 | |

# Comparação dos três algoritmos pela média

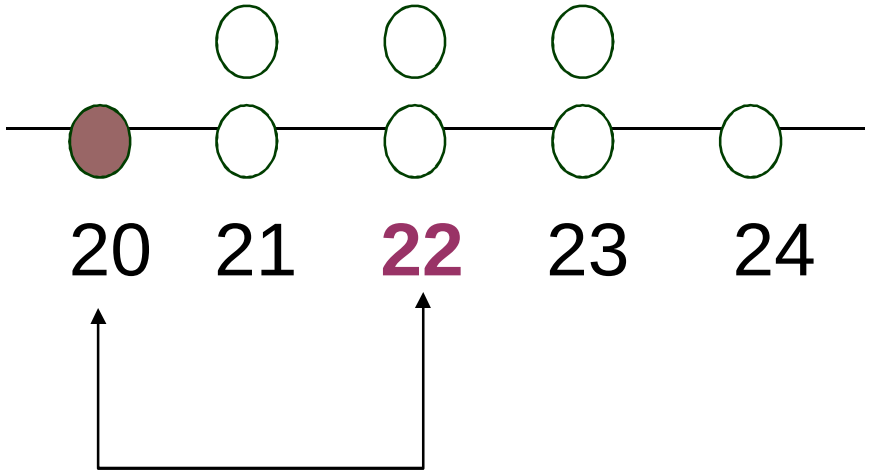
| algoritmo | falhas | | | | | | | | média |
|---|---|---|---|---|---|---|---|---|---|
| A | 20 | 21 | 21 | 22 | 22 | 23 | 23 | 24 | 22 |
| B | 16 | 18 | 20 | 22 | 22 | 24 | 26 | 28 | 22 |
| C | 15 | 22 | 23 | 23 | 23 | 24 | 24 | | 22 |

# Diagramas de Pontos



# Como medir a dispersão?

Exemplo: A ( 20 21 21 22 22 23 23 24 )

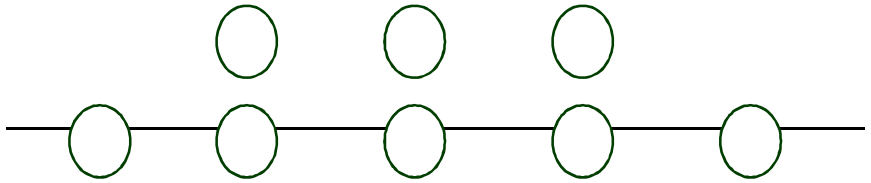



distância (desvio) em relação à média

# Desvios

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Valores | X | 20 | 21 | 21 | 22 | 22 | 23 | 23 | 24 |
| Média | $\bar{X}$ | 22 | | | | | | | |
| Desvios | $(X - \bar{X})$ | -2 | -1 | -1 | 0 | 0 | 1 | 1 | 2 |

# Desvios



20 21 **22** 23 24

| | | | | | | |
|---|---|---|---|---|---|---|
| Desvios: | -2 | -1 | 0 | 1 | 2 | Soma = 0 |

# Desvios Quadráticos

| | | | Soma |
|---|---|---|---|
| Valores | X | 20 21 21 22 22 23 23 24 | 176 |
| Média | $\bar{X}$ | 22 | - |
| Desvios | $X - \bar{X}$ | -2 -1 -1 0 0 1 1 2 | 0 |
| Desvios quadráticos | $(X-\bar{X})^2$ | 4 1 1 0 0 1 1 4 | 12 |

# Variância ($S^2$)

- A variância ($S^2$) é uma média dos desvios quadráticos. Usa-se no denominador n-1 ao invés de n quando trabalhamos com amostras e não a população completa.

$$S^2 = \frac{\sum \left(X - \overline{X}\right)^2}{n-1}$$

No exemplo apresentado (algoritmo A), a variância é:

$$S^2 = \frac{12}{7} = 1{,}71$$

# Desvio Padrão (S)

- O desvio padrão (S) é a raiz quadrada da variância.

$$S = \sqrt{S^2}$$

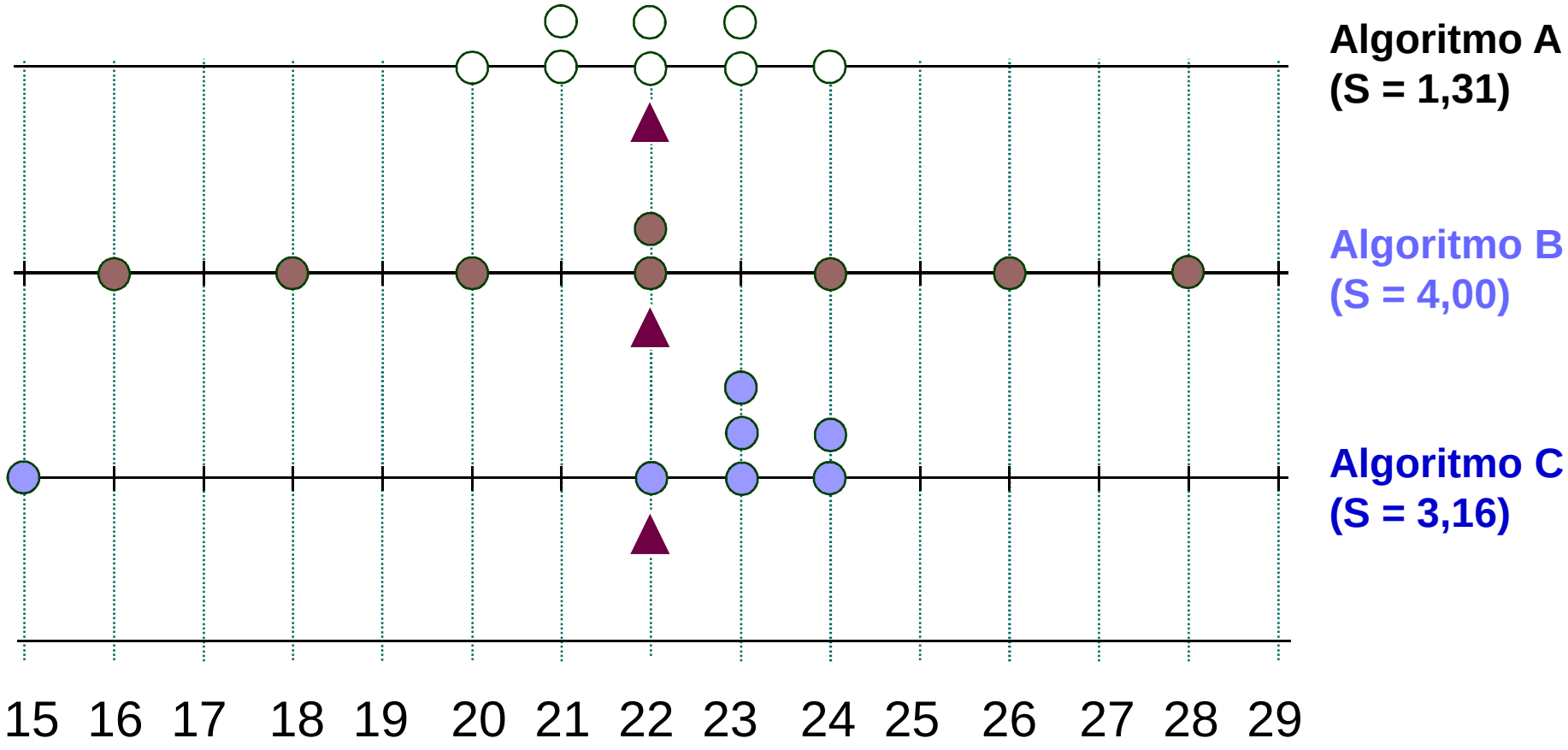
No exemplo apresentado (algoritmo A), o desvio padrão é:

$$S = \sqrt{1{,}71} = 1{,}31$$

# Comparação dos três algoritmos pela média e desvio padrão

| Algoritmo | Falhas | $\bar{X}$ | S |
|---|---|---|---|
| A | 20 21 21 22 22 23 23 24 | 22 | 1,31 |
| B | 16 18 20 22 22 24 26 28 | 22 | 4,00 |
| C | 15 22 23 23 23 24 24 | 22 | 3,16 |

# Diagramas de pontos e valores de S



# Exemplo

**TABELA** Medidas descritivas das notas finais dos alunos de três turmas

| Turma | Número de alunos | Média | Desvio padrão |
|---|---|---|---|
| A | 20 | 6,0 | 3,3 |
| B | 40 | 8,0 | 1,5 |
| C | 30 | 9,0 | 2,6 |

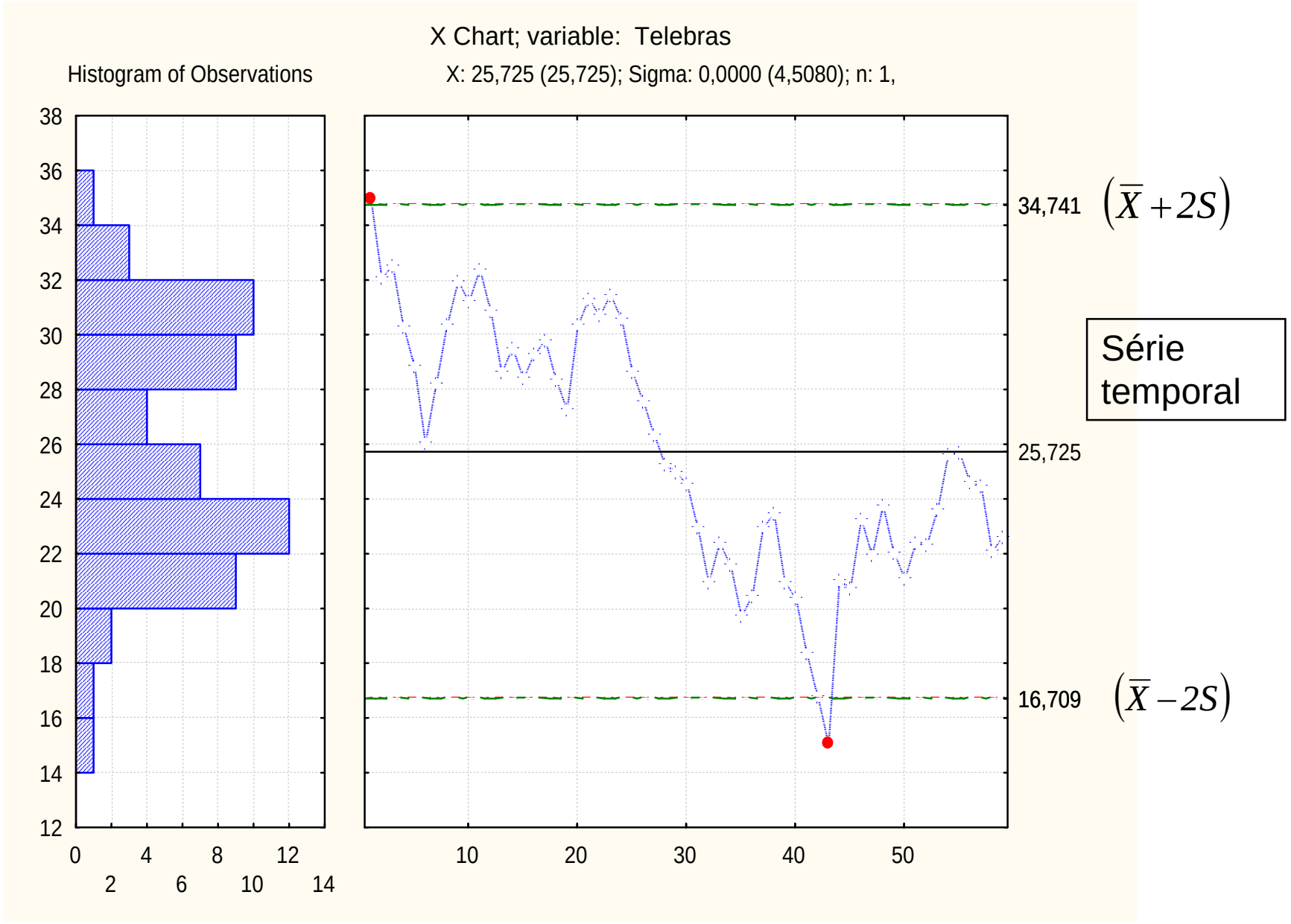
# Medida relativa de dispersão - Exemplo

**Coeficiente de variação = desvio padrão / média**

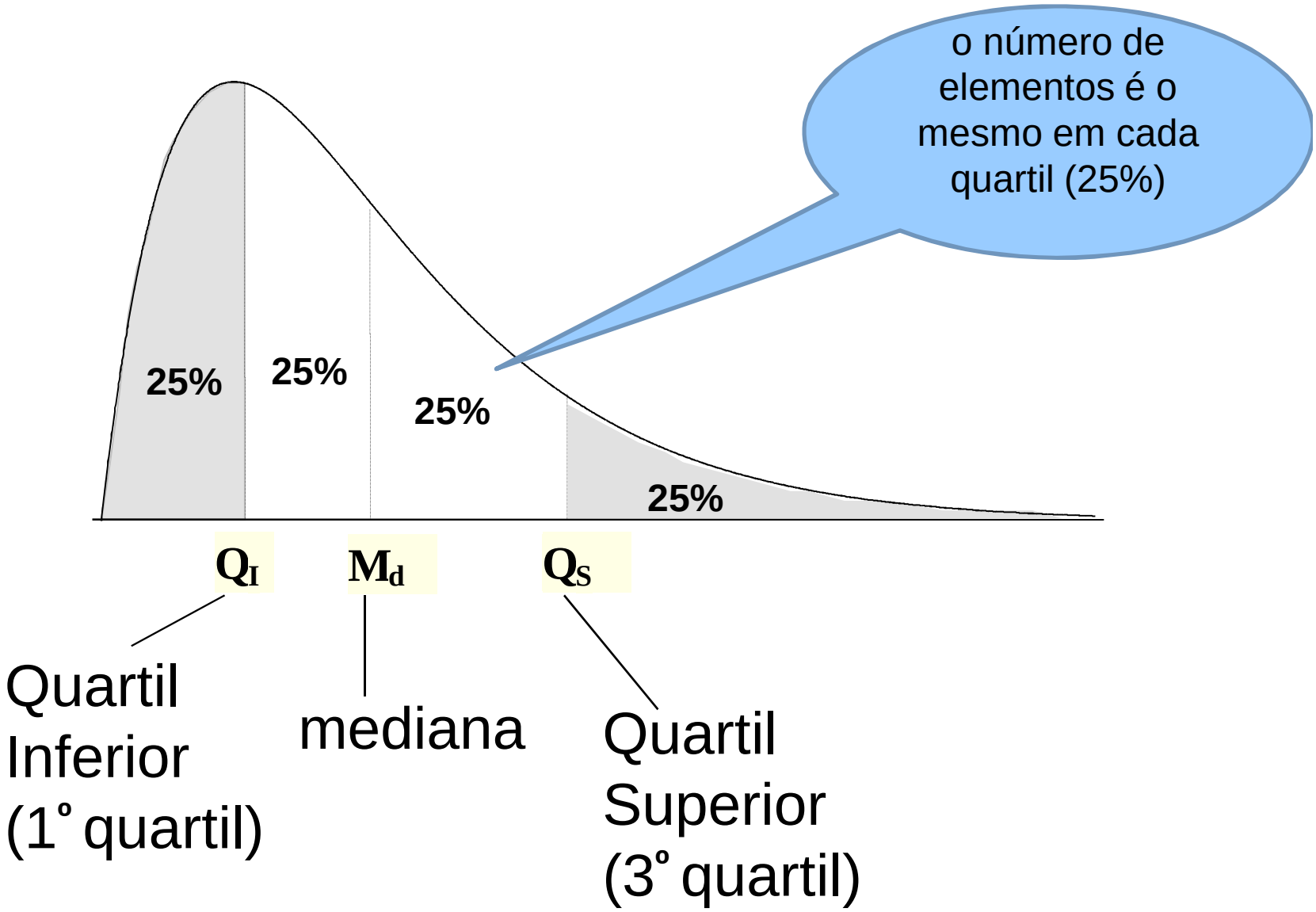
| | | | | |
|---|---|---|---|---|
| $X_1$: | 1 | 2 | 3 | média = 2<br>desvio padrão = 1<br>coeficiente de variação = 0,5 |
| $X_2$: | 100 | 101 | 102 | média = 101<br>desvio padrão = 1<br>coeficiente de variação = 0,01 |
| $X_3$: | 100 | 200 | 300 | média = 200<br>desvio padrão = 100<br>coeficiente de variação = 0,5 |

# Conjunto de dados: preços de fechamento de ações da telebrás

| 1<br>Mês | 2<br>dia | 3<br>id | 4<br>Telebras |
|---|---|---|---|
| jan | 2 | 1 | 34,99 |
| jan | 4 | 2 | 32,09 |
| jan | 5 | 3 | 32,56 |
| jan | 6 | 4 | 30,31 |
| jan | 9 | 5 | 28,91 |
| jan | 10 | 6 | 26,1 |
| jan | 11 | 7 | 28,25 |
| jan | 12 | 8 | 30,41 |
| jan | 13 | 9 | 32 |
| jan | 16 | 10 | 31,25 |
| jan | 17 | 11 | 32,37 |
| jan | 18 | 12 | 30,87 |
| jan | 19 | 13 | 28,63 |
| jan | 20 | 14 | 29,56 |
| jan | 23 | 15 | 28,44 |
| jan | 24 | 16 | 29,28 |
| jan | 26 | 17 | 29,84 |
| jan | 27 | 18 | 28,35 |
| jan | 30 | 19 | 27,32 |
| jan | 31 | 20 | 30,41 |
| fev | 1 | 21 | 31,34 |
| fev | 2 | 22 | 30,78 |
| fev | 3 | 23 | 31,44 |
| fev | 6 | 24 | 30,59 |
| fev | 7 | 25 | 28,63 |
| fev | 8 | 26 | 27,6 |
| fev | 9 | 27 | 26,38 |
| fev | 10 | 28 | 25,26 |
| fev | 13 | 29 | 24,98 |
| fev | 14 | 30 | 24,56 |
| fev | 1 | 31 | 23,02 |
| fev | 16 | 32 | 20,96 |
| fev | 17 | 33 | 22,45 |
| fev | 20 | 34 | 21,61 |
| fev | 21 | 35 | 19,74 |
| fev | 22 | 36 | 20,49 |
| fev | 23 | 37 | 23,02 |
| fev | 24 | 38 | 23,48 |
| mar | 2 | 39 | 20,96 |
| mar | 6 | 40 | 20,4 |
| mar | 7 | 41 | 18,43 |
| mar | 8 | 42 | 16,84 |
| mar | 8 | 42 | 16,84 |
| mar | 9 | 43 | 15,06 |
| mar | 10 | 44 | 21,05 |
| mar | 13 | 45 | 20,77 |
| mar | 14 | 46 | 23,3 |
| mar | 15 | 47 | 21,99 |
| mar | 16 | 48 | 23,75 |
| mar | 17 | 49 | 22,08 |
| mar | 20 | 50 | 21,14 |
| mar | 21 | 51 | 22,45 |
| mar | 22 | 52 | 22,36 |
| mar | 23 | 53 | 23,67 |
| mar | 24 | 54 | 25,63 |
| mar | 27 | 55 | 25,73 |
| mar | 28 | 56 | 24,61 |
| mar | 29 | 57 | 24,51 |
| mar | 30 | 58 | 22,13 |
| mar | 31 | 59 | 22,64 |


X Chart; variable: Telebras
Histogram of Observations
X: 25,725 (25,725); Sigma: 0,0000 (4,5080); n: 1,
38
36
34
32
30
28
26
24
22
20
18
16
14
12
0
2
4
6
8
10
12
14
10
20
30
40
50
34,741
$\left(\overline{X}+2S\right)$
25,725
16,709
$\left(\overline{X}-2S\right)$
Série temporal

# Medidas baseadas na ordenação dos dados



# Cálculo da mediana

Dados:

{2, 0, 5, 7, 9, 1, 3, 4, 6, 8}

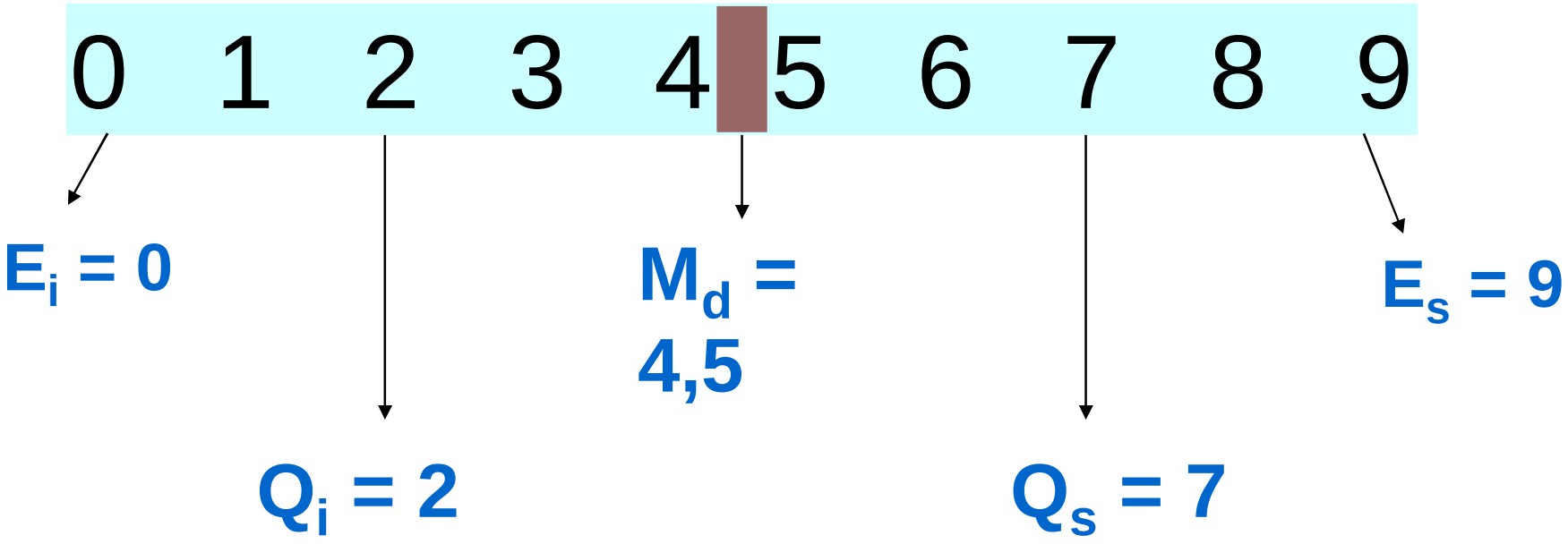
n = 10; (n + 1) / 2 = 5,5

0 1 2 3 4 5 6 7 8 9

↓

**$M_d$ = 4,5**

# Cálculo dos Quartis



# Exercício:

Cálculo da mediana

Dados:

{2, 0, 5, 7, 9, 1, 3, 4, 6, 8, 100}

n = 11; (n + 1) / 2 = 6

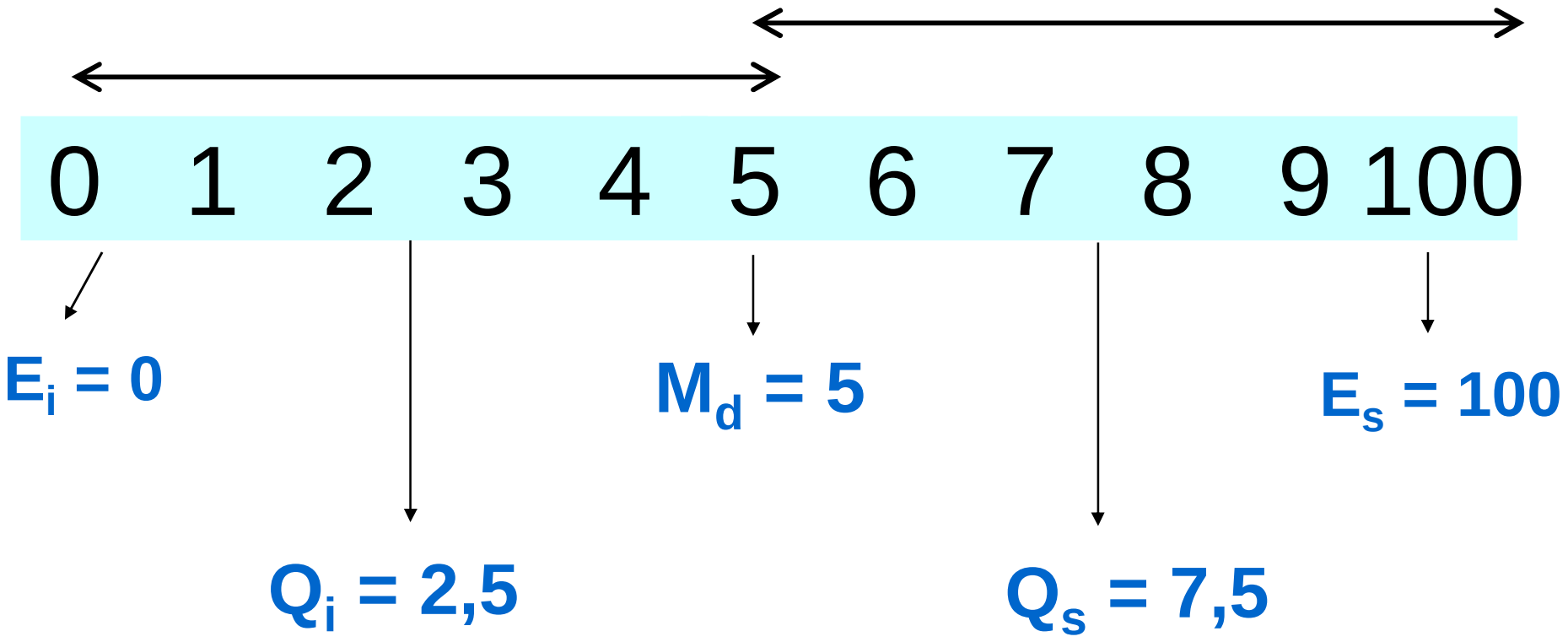
0 1 2 3 4 5 6 7 8 9 100



$M_d = 5$

# Exercício:

Cálculo dos quartis



# Medida de dispersão: Distância interquartílica

O desvio inter-quartílico é uma medida robusta de dispersão. Ele é calculado por:

$$Q_3 - Q_1$$

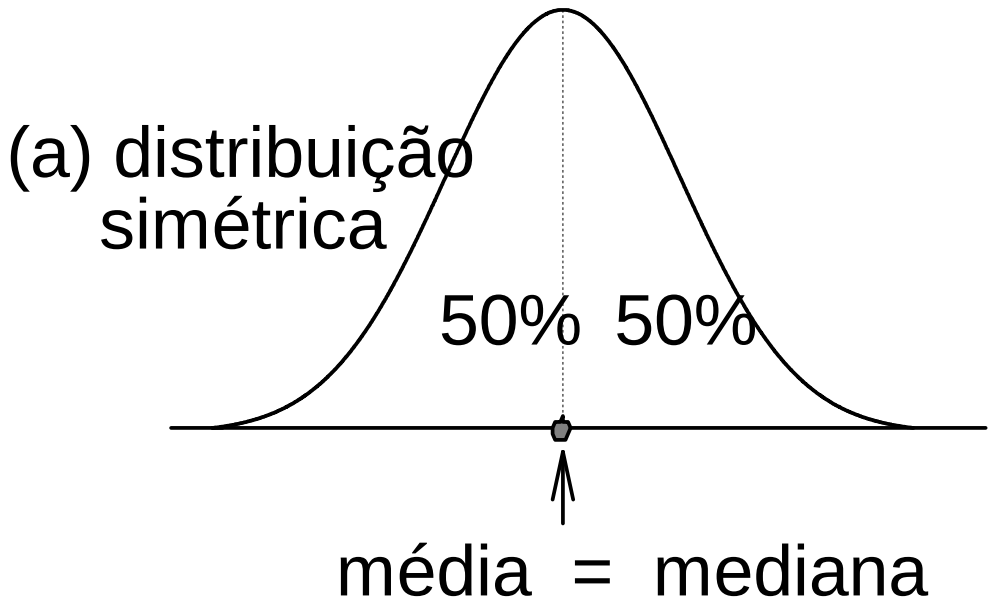
Onde $Q_3$ é o percentil 75, também chamado de quartil superior, e o $Q_1$ é o percentil 25, também chamado de quartil inferior. Ele é uma boa medida de dispersão para distribuições assimétricas. Para dados normalmente distribuídos, o desvio inter-quartílico é aproximadamente igual a 1,35 vezes o desvio padrão.

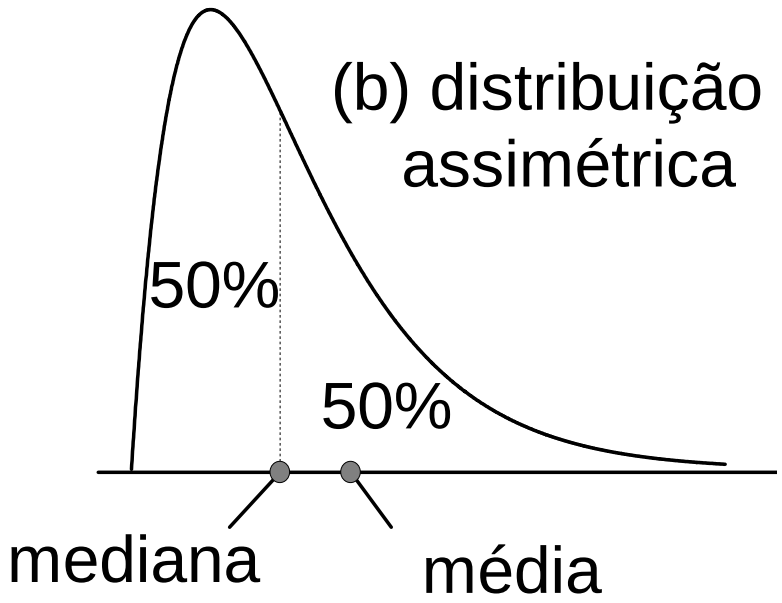
Medidas da variável IDADE de funcionários de uma empresa, do setor de tecidos:

Descriptive Statistics (Planilha_funcionarios_AED_Statistica)

| Valid N | Mean | Median | Lower Quartile | Upper Quartile | Quartile Range | Std.Dev. |
|---|---|---|---|---|---|---|
| 45 | 32,68889 | 32,00000 | 26,00000 | 38,00000 | 12,00000 | 8,920921 |

# Média e Mediana



# Cálculo dos Outliers

$$Q_I - 1{,}5(Q_S - Q_I)$$
$$Q_S + 1{,}5(Q_S - Q_I)$$

Onde $Q_I$ é o quartil inferior ou primeiro quartil da distribuição; $Q_S$ é o quartil superior ou terceiro quartil da distribuição. O valor 1,5 pode ser alterado.

# TRANSFORMAÇÃO DE DADOS

Objetivo: obter os dados em uma forma mais apropriada para os algoritmos de mineração

- Alisamento
- Generalização
- Normalização
- Transformação numérico para categórico
- Transformação categórico para numérico

# Alisamento

Eliminação de ruídos, exceções, ouliers, que são prejudiciais a muitos algoritmos de mineração

# Generalização

Utilizado quando os dados são muito esparsos e não se consegue bons resultados .

Então, dados primitivos são substituídos por conceitos de ordem superior via uma hierarquia de conceitos.

Exemplo:

- *calça*, *blusa*, *saia*, etc. são substituídos por *roupa*

- nomes de cidades são substituídas pelo nome do estado ao qual pertencem

# Normalização

O propósito da normalização é minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis.

Alguns algoritmos de mineração são beneficiados com a normalização (redes neurais, kNN, k-medias, ...)

# Normalização

**Objetivo**: ajustar as escalas de valores dos atributos para o mesmo intervalo : [-1 a 1] , [0 a 1],...

- Evita maior influência, em determinados métodos, de atributos com grande intervalo de valores

- Normalização linear
- Normalização por desvio padrão
- Normalização pelo valor máximo dos elementos
- Normalização por escala decimal

# Normalização

## Normalização linear no intervalo [0,1]

$$f(X) = \frac{X - Min}{Max - Min}$$

| CPF | Despesa | Despesa_normalizada |
|---|---|---|
| 99999999999 | 1000 | 0,14 |
| 11111111111 | 2000 | 0,43 |
| 33333333333 | 3000 | 0,71 |
| 55555555555 | 1500 | 0,29 |
| 22222222222 | 1500 | 0,29 |
| 00000000000 | 1000 | 0,14 |
| 88888888888 | 3000 | 0,71 |
| 77777777777 | 500 | 0 |
| 66666666666 | 4000 | 1 |
| 44444444444 | 1000 | 0,14 |

# Normalização

## Normalização por desvio padrão

- Objetivo: considera a posição média dos valores e os graus de dispersão em relação à posição média
- Útil quando mínimo e máximo são desconhecidos

$f(X) = (X - \text{média}) / \sigma$

onde $\sigma$ = desvio padrão

média = 1850
$\sigma$ = 1131,62

| CPF | Despesa | Despesa_normalizada |
|---|---|---|
| 99999999999 | 1000 | -0,75 |
| 11111111111 | 2000 | 0,13 |
| 33333333333 | 3000 | 1,02 |
| 55555555555 | 1500 | -0,31 |
| 22222222222 | 1500 | -0,31 |
| 00000000000 | 1000 | -0,75 |
| 88888888888 | 3000 | 1,02 |
| 77777777777 | 500 | -1,19 |
| 66666666666 | 4000 | 1,90 |
| 44444444444 | 1000 | -0,75 |

# Normalização

## Normalização pelo valor máximo dos elementos

- Dividir cada valor pelo maior valor
- Resultado similar à normalização linear
  - Igual se mínimo = 0 (zero)

$$f(X) = X / \text{máximo}$$

| CPF | Despesa | Despesa_normalizada |
|---|---|---|
| 99999999999 | 1000 | 0,25 |
| 11111111111 | 2000 | 0,50 |
| 33333333333 | 3000 | 0,75 |
| 55555555555 | 1500 | 0,38 |
| 22222222222 | 1500 | 0,38 |
| 00000000000 | 1000 | 0,25 |
| 88888888888 | 3000 | 0,75 |
| 77777777777 | 500 | 0,13 |
| 66666666666 | 4000 | 1 |
| 44444444444 | 1000 | 0,25 |

# Transformação numérico → categórico

Objetivo: transformação de valores numéricos para categóricos ou discretos

- Mapeamento direto
- Mapeamento em intervalos (discretização)

# Transformação numérico → categórico

## Mapeamento direto

- Objetivo: substituição de valores numéricos por valores categóricos

Exemplo: sexo

1 → M

0 → F

# Transformação numérico → categórico

## Mapeamento em intervalos (discretização)

- Objetivo: substituição de valores dentro de um intervalo por um identificador
- Identificador de intervalo:
  - Categórico: nome (sugestão: mneumônico)
  - Numérico
- Exemplo: número de dependentes

| Num_Dep: | 0 a 1 | 2 a 5 | 6 a 99 |
|---|---|---|---|
| **categórico** | poucos_dep | media_dep | muitos_dep |
| **numérico** | 0 | 1 | 2 |

# Transformação numérico → categórico

## Mapeamento em intervalos (discretização): formas

- **Intervalos com tamanho pré-definidos** (domínio da aplic.)

  0 a 1 → **0** , 2 a 5 → **1** , 6 a 99 → **2**

- **Intervalos de igual tamanho** (conhecimento dos limites do intervalo)

  2 intervalos / 10 valores: 0 a 4 → **0** , 5 a 9 → **1**

- **Intervalos com o mesmo número de elementos**

- **Intervalos por meio de clusterização**

  Utiliza algum algoritmo de agrupamento de dados para descobrir automaticamente a distribuição dos dados

# Transformação  categórico → numérico

Objetivo: transformação de valores categóricos em numéricos

- Mapeamento direto
- Representação binária 1-de-N

# Transformação categórico → numérico

## Mapeamento direto

Mapeamento em valores de 1 a N

| Est_Civil | mapeamento |
|---|---|
| Casado | 1 |
| Solteiro | 2 |
| Viúvo | 3 |
| Divorciado | 4 |
| Outro | 5 |

# Transformação  categórico → numérico

## Mapeamento direto

Quando o atributo categórico for **ordinal**, é importante  que os valores numéricos sigam a mesma ordem

| conceito | mapeamento |
|---|---|
| Ruim | 1 |
| Regular | 2 |
| Bom | 3 |
| Ótimo | 4 |

# Transformação categórico → numérico

## Representação binária 1-de-N

- Mapeamento em número cuja representação binária tenha N dígitos
  - Somente um dígito é “1”

| Est_Civil | Representação binária 1-de-N |
|---|---|
| Casado | 00001 |
| Solteiro | 00010 |
| Viúvo | 00100 |
| Divorciado | 01000 |
| Outro | 10000 |

# Outros tipos de dados: outras transformações

- Texto (ex: categorização de textos; “exame” de e-mails, ...)
- internet
  - conteúdo
  - estrutura
  - uso
- imagens
- seqüências de genes
- séries temporais
- dados de trajetórias
- dados de redes sociais
- .....

# Exercícios

- Dado o conjunto {1, 2, 3, 4, 5, 80}, calcular:
  - Média
  - Mediana

- Dados os números abaixo, calcular a mediana, o quartil inferior e o quartil superior

  23, 7, 12, 6, 10, 23, 7, 12, 6, 10, 7

# Exercícios

- Converter os dados abaixo para valores numéricos e normalizá-los em [0, 1]

| Febre | Enjôo | Mancha | Dor | Diagnóstico |
|---|---|---|---|---|
| baixa | sim | pequena | A | doente |
| média | não | média | C | saudável |
| alta | sim | grande | B | saudável |
| alta | não | pequena | A | doente |
| baixa | não | grande | D | saudável |
| média | não | ausente | C | doente |

# Exercícios

- Discretizar o atributo que possui os valores abaixo em 3 intervalos

0, 1, 1, 1, 2, 2, 2, 3, 4, 6, 6, 9, 10, 13, 20, 20, 21, 21, 22, 23, 23

Usar:

- Tamanhos iguais
- Freqüências iguais