Unidade II - Estatística descritiva

2.1. Apresentação de dados
- 2.1.1 Séries estatísticas
- 2.1.2 Tabelas
- 2.1.3 Gráficos

2.2. Distribuições de freqüências e gráficos
- 2.2.1 Tabelas de classificação simples
- 2.2.2 Tabelas de classificação cruzada

2.3. Medidas descritivas
- 2.3.1 Medidas de localização ou tendência central
- 2.3.2 Medidas separatrizes
- 2.3.3 Medidas de variação ou dispersão
- 2.3.4 Medidas de formato

2.4. Análise exploratória de dados

Estatística clássica

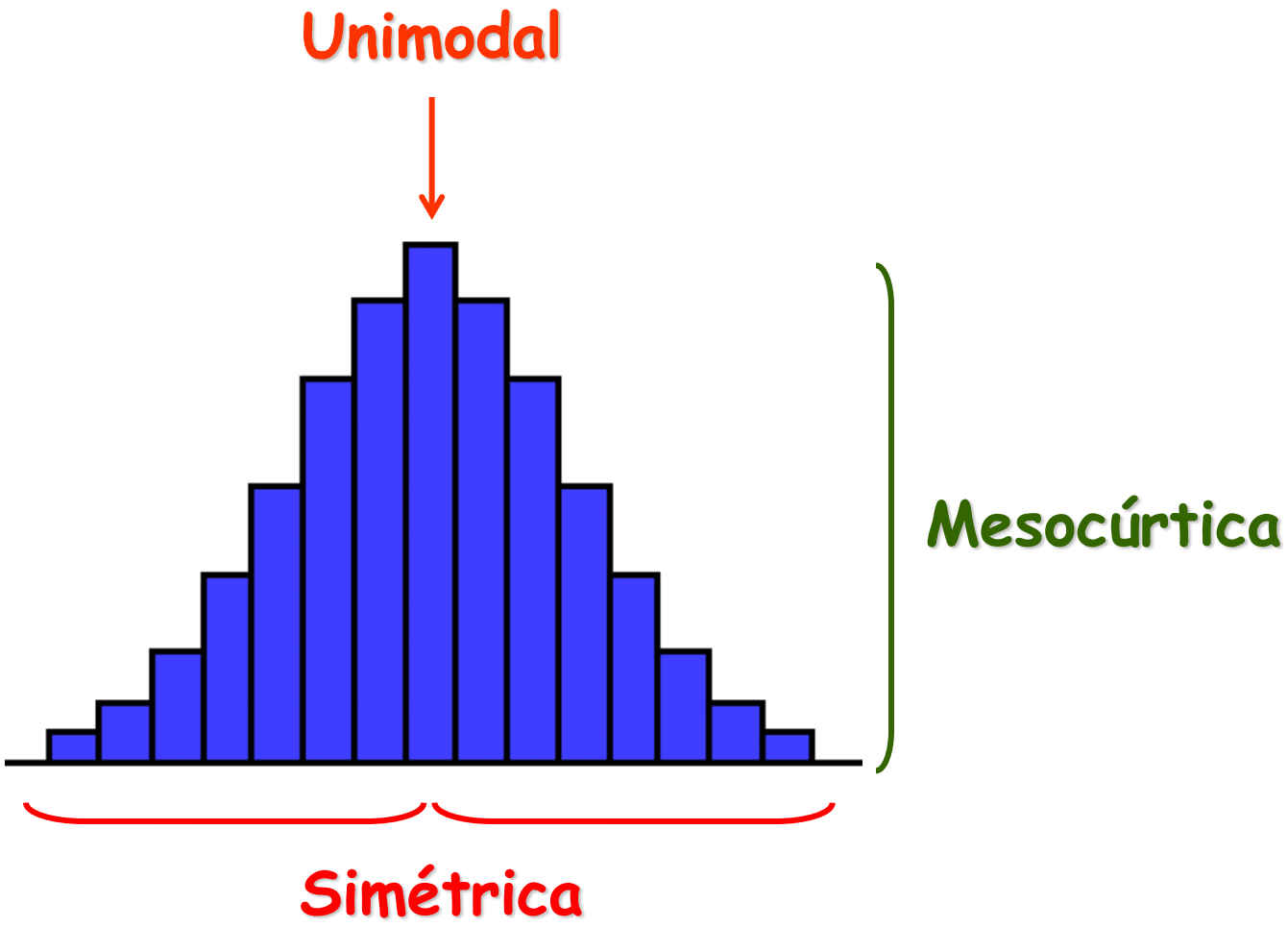
⇨ A **média aritmética** e a **variância** são medidas muito utilizadas por duas razões:

- são de fácil compreensão
- apresentam boas propriedades algébricas e estatísticas

Limitações:

⇨ Essas medidas **descrevem de forma ótima** distribuições de frequências **unimodais**, **simétricas** e **mesocúrticas**

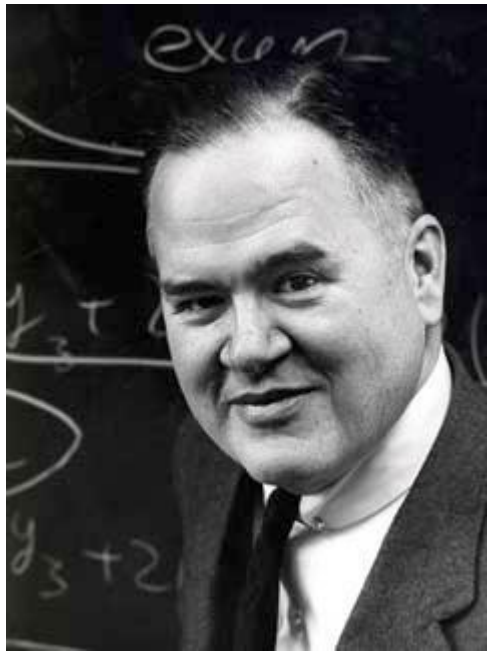
⇨ Numa distribuição **assimétrica** seus valores são bastante afetados pelos valores discrepantes


Unimodal
Mesocúrtica
Simétrica

1970 → John Tukey propôs algumas técnicas que contornavam esses problemas da média e da variância na descrição de distribuições assimétricas

John Wilder Tukey
(1915 - 2000)

O conjunto dessas técnicas recebeu a denominação de **Análise exploratória de dados.**



Formado em Química e Matemática, Tukey teve papel fundamental no desenvolvimento da Estatística durante a segunda metade do século XX. Trabalhou tanto na Universidade de Princeton quanto na Bell Laboratórios. Diferente da maioria de seus colegas, interessava-se pelos aspectos práticos, como questões do tipo: **"O que os dados querem dizer?"**.

Tukey inventou uma grande variedade de métodos, gráficos e numéricos, para aplicações estatísticas. Desenvolveu várias técnicas voltadas à análise exploratória de dados, tendo como objetivo examiná-los, descrevendo suas principais características, como um explorador numa terra desconhecida vai descrevendo o que vai vendo.

A ***Bell Labs*** desenvolveu uma série de tecnologias consideradas revolucionárias desde comutadores telefônicos, cabos de telefone, transístores, LEDs, lasers, a linguagem de programação C e o sistema operativo Unix. Conhecido por seu gosto por cunhar palavras e frases apropriadas, Tukey é creditado com a invenção das palavras **"bit"** (binary digit), em 1946, e **"software"**, em 1958. Também foi responsável pelo primeiro uso de muitos termos na Matemática Estatística.

- A **Análise exploratória de dados** não só constituiu um **complemento às técnicas estatísticas clássicas**, como foi também uma **valiosa alternativa** para descrever dados que não seguem o modelo unimodal, simétrico e mesocúrtico.

- O enfoque proposto pela **Análise exploratória de dados** pretende obter medidas **resistentes** e **robustas**.

- As medidas **resistentes** mostram-se **pouco sensíveis à presença de valores atípicos** (discrepantes do núcleo central da distribuição).

- Medidas **robustas** apresentam pouca sensibilidade diante da **violação dos pressupostos básicos** inerentes aos modelos probabilísticos, como, por exemplo, com relação à forma da distribuição.

⇨ As técnicas exploratórias, além de **descrever um conjunto de valores**, também ajudam a **comprovar as condições de aplicação dos testes de hipóteses** (Inferência Estatística).

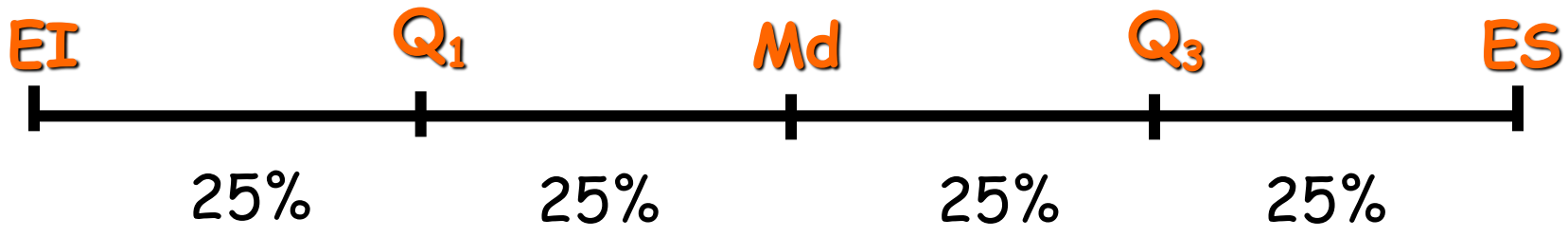
Principais técnicas exploratórias:

- **Resumo de cinco números**
- **Gráfico de caixa ("box plot")**
- **Diagrama de ramo e folhas**

Resumo de cinco números

Descreve o conjunto de dados através de cinco valores:

- mediana (Md)
- primeiro (Q_1) e terceiro (Q_3) quartis
- extremos inferior (EI) e superior (ES)



O resumo de cinco números fornece uma idéia da simetria (formato) da distribuição porque o percentual de valores dentro de cada intervalo é conhecido (25%).

Profª. Clause Piana

A partir dos cinco números podemos obter outras medidas:

- **Amplitude interquartílica (a_q):** diferença entre os **quartis**

$$a_q = Q_3 - Q_1$$

- **Dispersão inferior (DI):** diferença entre **Md** e **EI**
- **Dispersão superior (DS):** diferença entre **ES** e **Md**

Dispersão inferior

$$DI = Md - EI$$

Dispersão superior

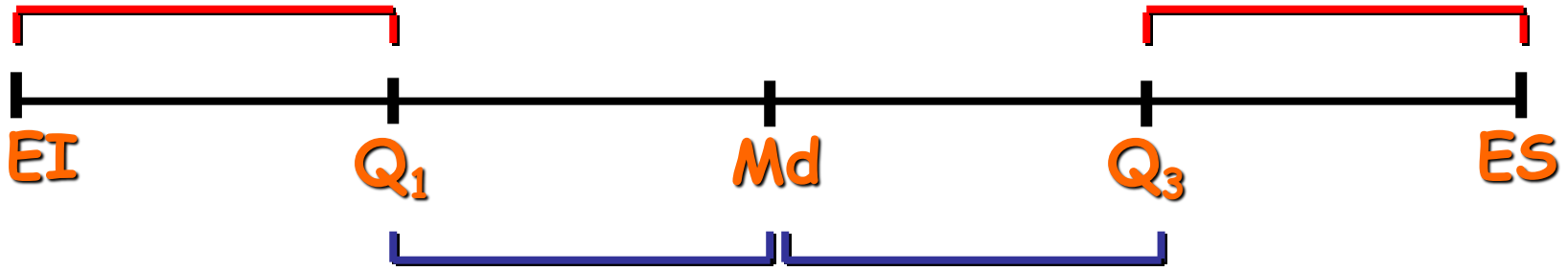
$$DS = ES - Md$$

Profª. Clause Piana

Simetria

A distribuição é considerada **simétrica** se:

1. A diferença entre o **primeiro quartil** e **extremo inferior** é aproximadamente igual à diferença entre o **extremo superior** e o **terceiro quartil** ($Q_1 - EI \cong ES - Q_3$)



2. A diferença entre a **mediana** e o **primeiro quartil** é aproximadamente igual à diferença entre o **terceiro quartil** e a **mediana** ($Md - Q_1 \cong Q_3 - Md$)

Profª. Clause Piana

Condições para a simetria $\begin{cases} Q1 - EI \cong ES - Q3 \\ Md - Q1 \cong Q3 - Md \end{cases}$

Casos simétricos

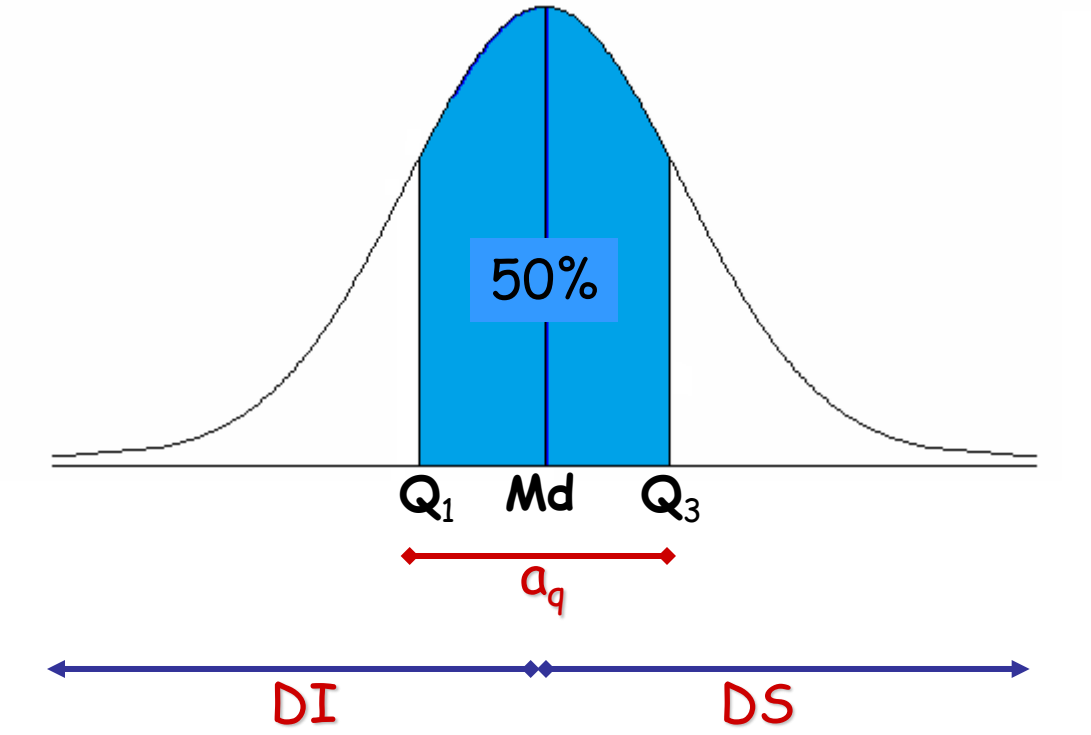
$a_q \cong DI \cong DS$

$a_q < DI \cong DS$ ← curva normal

$a_q > DI \cong DS$

Profª. Clause Piana

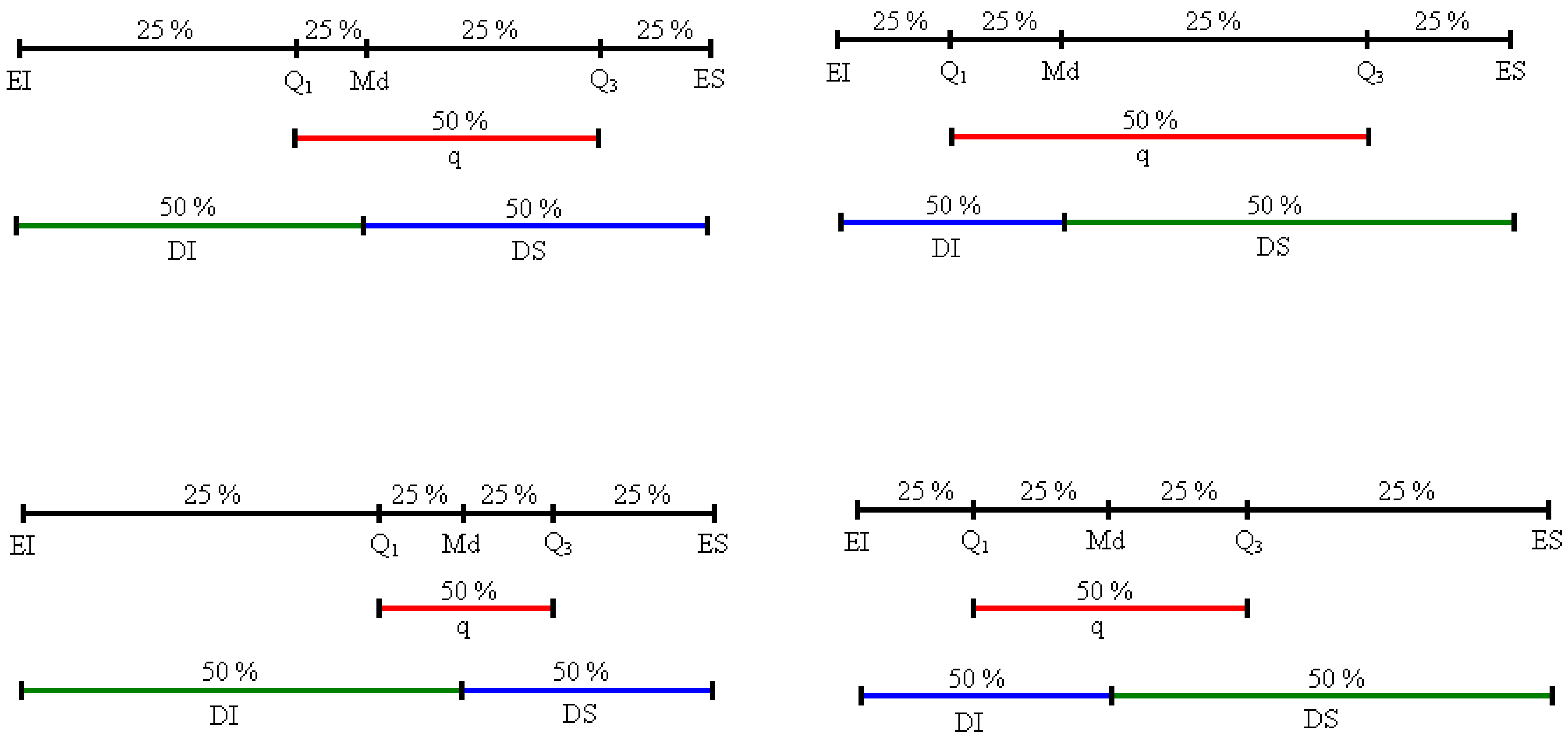
Distribuição normal



$$a_q < DI = DS$$

Profª. Clause Piana

Casos assimétricos

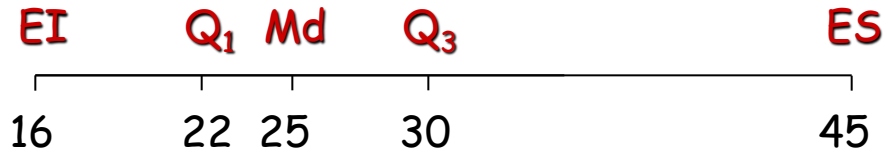


Exemplo resolvido:

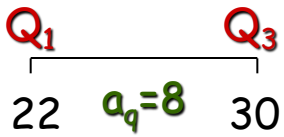
Os dados abaixo se referem aos pesos ao nascer (em kg) de 61 bovinos machos da raça Ibagé.

16 17 17 18 18 18 19 20 20 20 20 20 21 21 22 22 23 23
23 23 23 23 23 23 23 25 25 25 25 25 25 26 26 27 27 27
27 28 28 28 29 29 29 30 30 30 30 30 30 30 31 32 33 33
33 34 34 35 36 39 45

Resumo de cinco números ➡



Amplitude interquartílica ➡



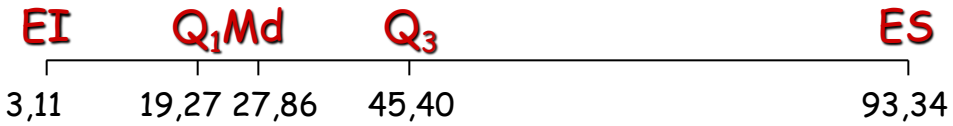
Dispersão inferior e Dispersão superior ➡

Exercício proposto:

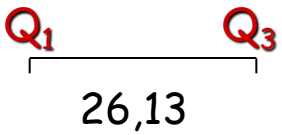
Os dados abaixo valores gastos (em reais) pelas primeiras 50 pessoas que entraram num determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

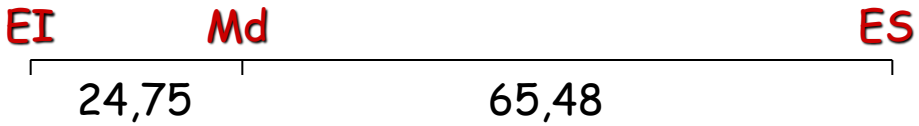
Resumo de cinco números ➡



Amplitude interquartílica ➡



Dispersão inferior e Dispersão superior ➡



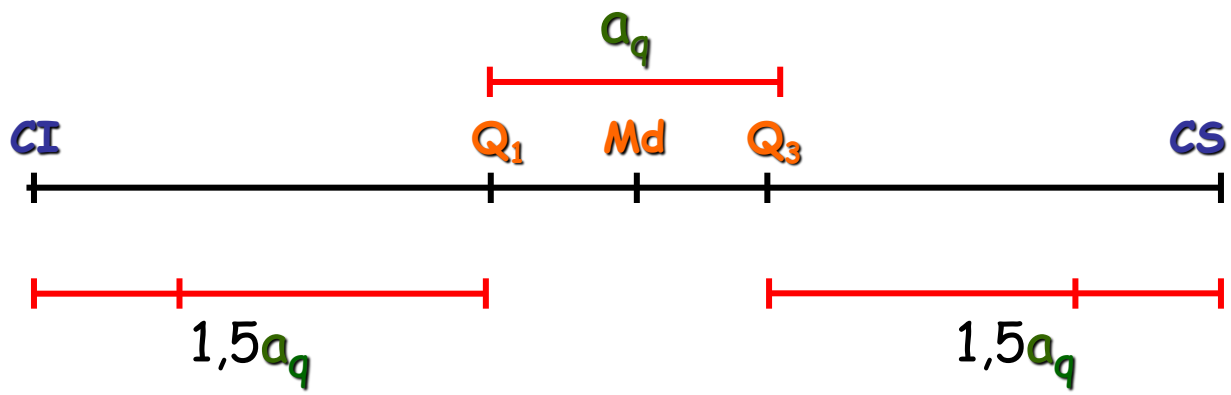
Profª. Clause Piana

Identificação de valores discrepantes ("outliers")

O critério usado para identificar valores discrepantes num conjunto de dados é baseado em duas medidas:

Cerca inferior → $CI = Q_1 - 1{,}5a_q$

Cerca superior → $CS = Q_3 + 1{,}5a_q$



Profª. Clause Piana

Identificação de valores discrepantes ("outliers")

O critério usado para identificar valores discrepantes num conjunto de dados é baseado em duas medidas:

Cerca inferior → $CI = Q_1 - 1{,}5a_q$

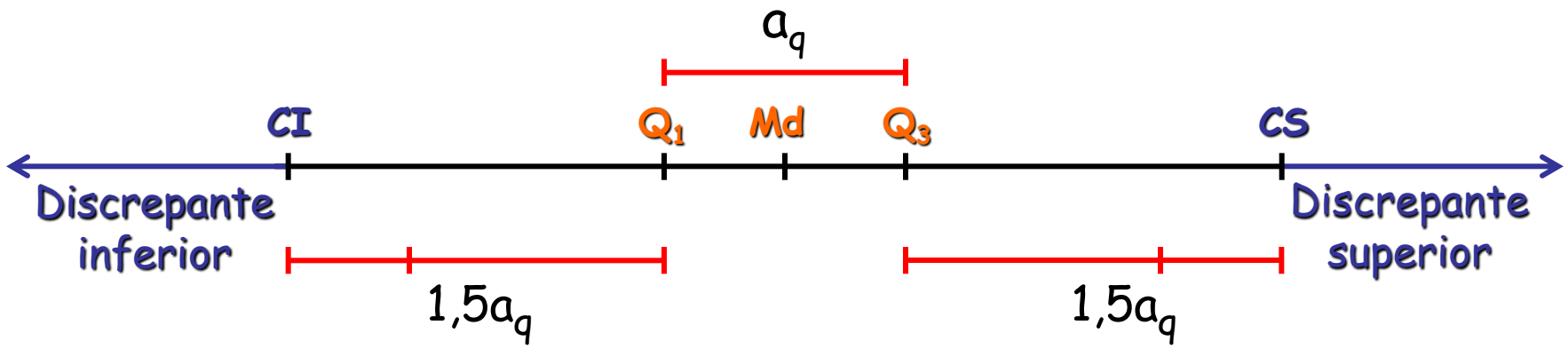
Cerca superior → $CS = Q_3 + 1{,}5a_q$

São considerados **discrepantes** os valores que estivem fora do seguinte intervalo:

$$[CI\ ;\ CS]$$

Valores **menores** que a **CI** são **discrepantes inferiores**
Valores **maiores** que a **CS** são **discrepantes superiores**

Profª. Clause Piana



Exemplo resolvido:

Verifique se existem valores discrepantes no conjunto abaixo:

16 17 17 18 18 18 19 20 20 20 20 20 21 21 22 22 23 23
23 23 23 23 23 23 23 25 25 25 25 25 25 26 26 27 27 27
27 28 28 28 29 29 29 30 30 30 30 30 30 30 31 32 33 33
33 34 34 35 36 39 45

$CI = Q_1 - 1{,}5\ a_q = 22 - 1{,}5 \times 8 = 10$

$CS = Q_3 + 1{,}5\ a_q = 30 + 1{,}5 \times 8 = 42$

Verificamos que o valor **45** ultrapassa a cerca superior, portanto, é classificado como **discrepante superior**.

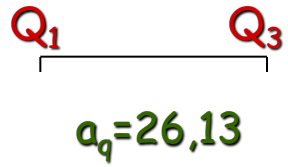
Exercício proposto:

Verifique se existem valores discrepantes no conjunto de valores abaixo:

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

$$CI = Q_1 - 1{,}5\, a_q = 19{,}27 - 1{,}5 \times 26{,}13 = -19{,}93$$

$$CS = Q_3 + 1{,}5\, a_q = 45{,}4 + 1{,}5 \times 26{,}13 = 84{,}60$$



Verificamos que os valores **85,76**, **86,37** e **93,34** ultrapassam a cerca superior, portanto, é classificado como discrepantes superiores.

Outras medidas resistentes

Grupo	Medidas	
Localização	Média de quartis	$\bar{Q} = \frac{Q_1 + Q_3}{2}$
	Trimédia	$TRI = \frac{Md + \bar{Q}}{2}$
	Média interquartílica	Média aritmética dos valores localizados entre o Q_1 e o Q_3
Dispersão	Amplitude interquartílica	$a_q = Q_3 - Q_1$
	Mediana dos desvios absolutos	$MAD = Md\|x_i - Md\|$
	Coeficiente de variação quartílico	$CVQ = \frac{a_q / 2}{\bar{Q}}$
Formato	Coeficiente de assimetria de Yule	$H_1 = \frac{Q_1 + Q_3 - 2Md}{2Md}$
	Coeficiente de assimetria de Kelly	$H_3 = \frac{P_{10} + P_{90} - 2Md}{2Md}$
	Coeficiente de curtose	$K_2 = \frac{P_{90} - P_{10}}{1,9(P_{75} - P_{25})}$ $K_1 = \frac{P_{87,5} - P_{12,5}}{1,7(P_{75} - P_{25})}$

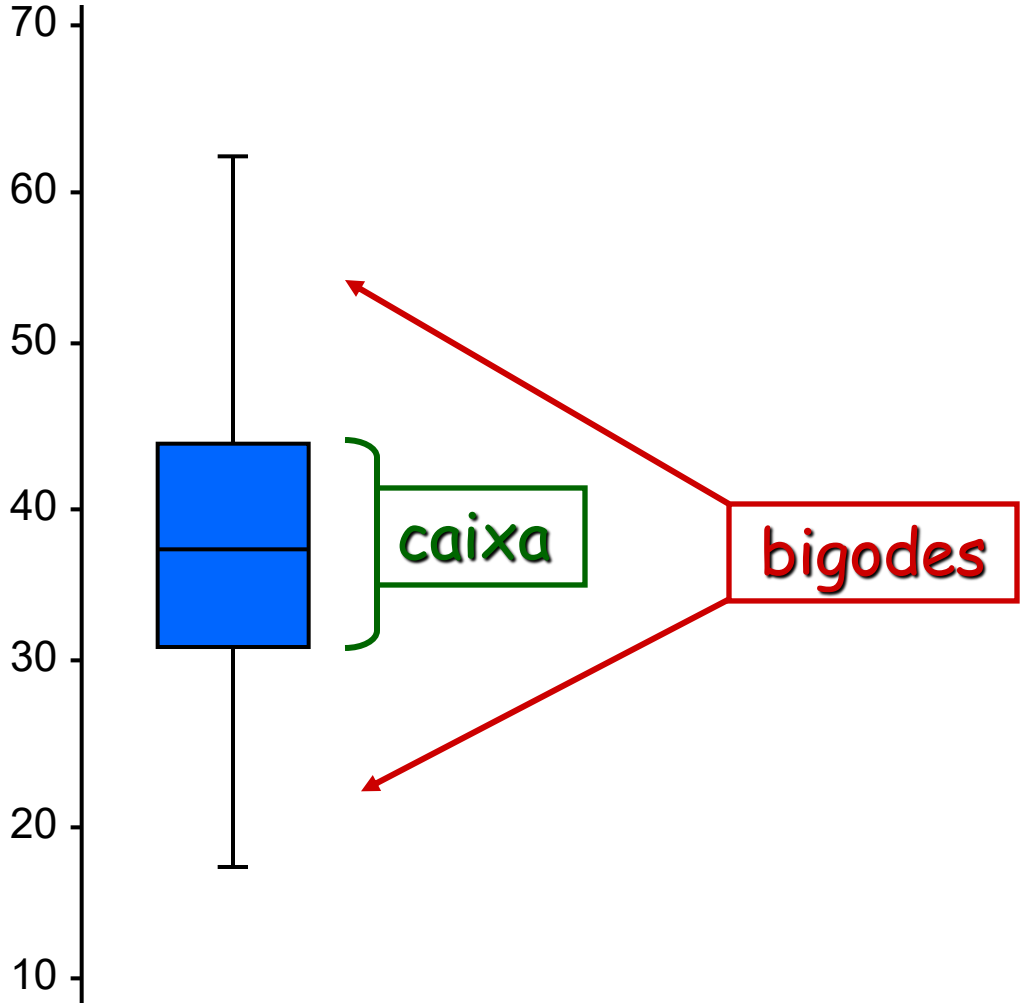
Profª. Clause Piana

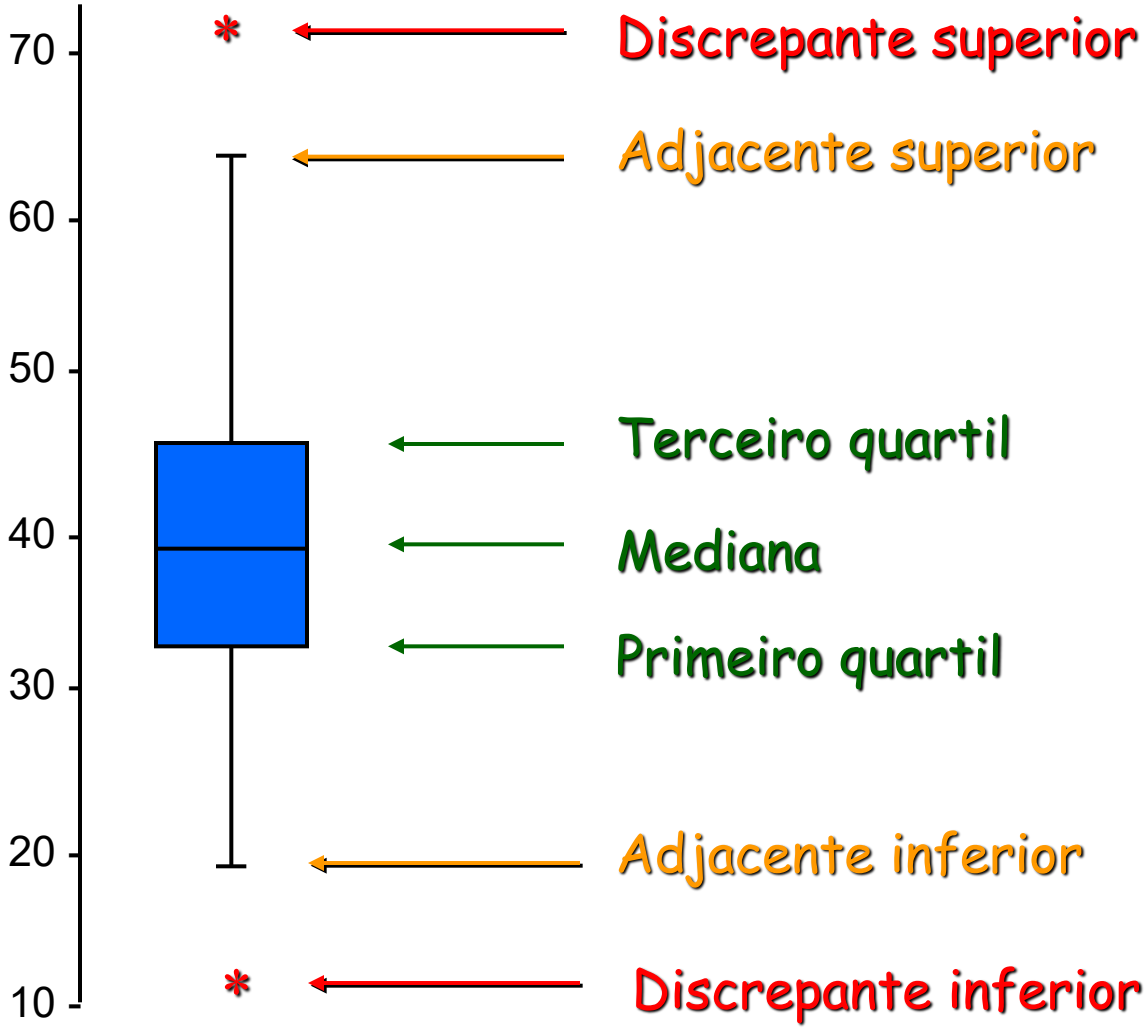
Gráfico de caixa (box plot)

A informação dada pelo resumo de cinco números pode ser apresentada na forma de um gráfico de caixa que agrega uma série de informações sobre a distribuição:

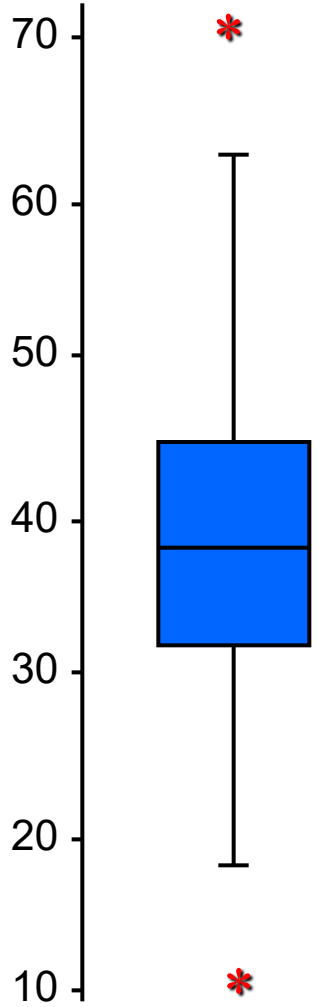
- localização (centro)
- dispersão
- assimetria
- caudas
- dados discrepantes

Profª. Clause Piana


70
60
50
40
30
20
10
caixa
bigodes



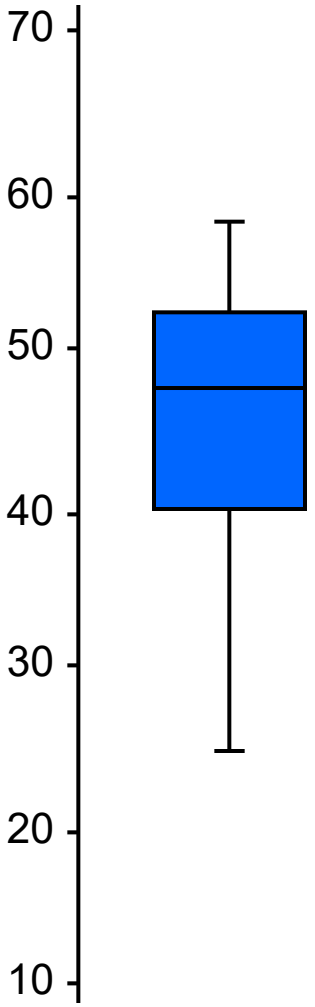
Profª. Clause Piana



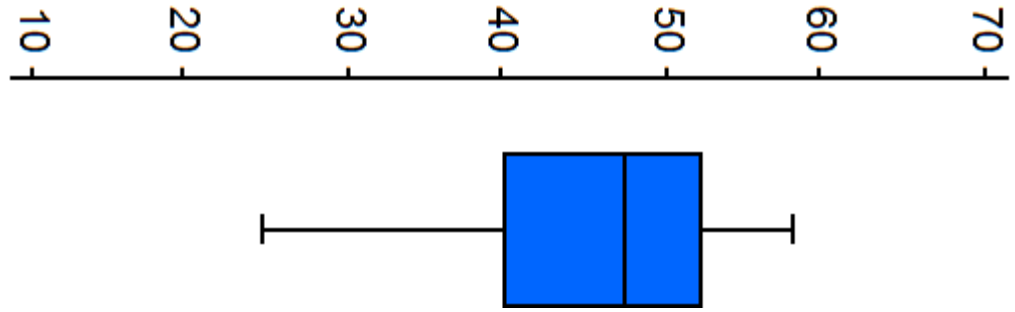
- A **posição central** dos valores é dada pela mediana e a **dispersão** pela amplitude interquartílica.
- As posições relativas da mediana e dos quartis e o formato dos bigodes dão uma noção da **simetria** e do **tamanho das caudas** da distribuição.

Profª. Clause Piana

Exemplos:



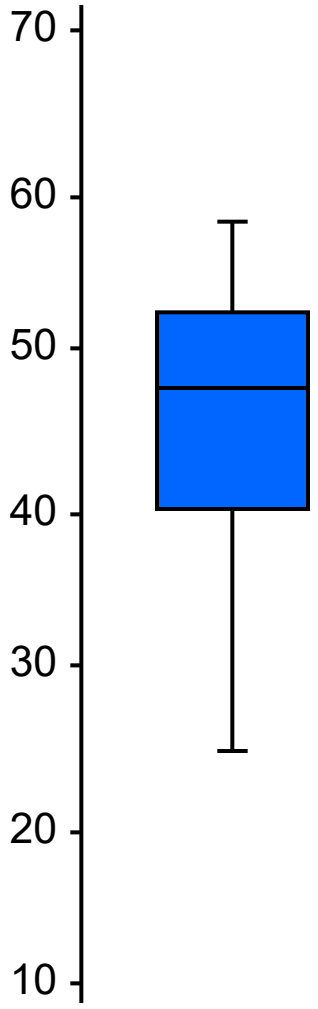
Interpretação



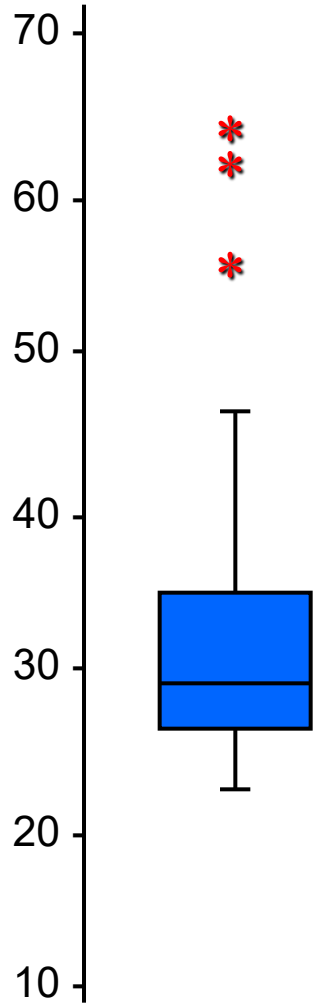
Assimétrica negativa

Profª. Clause Piana

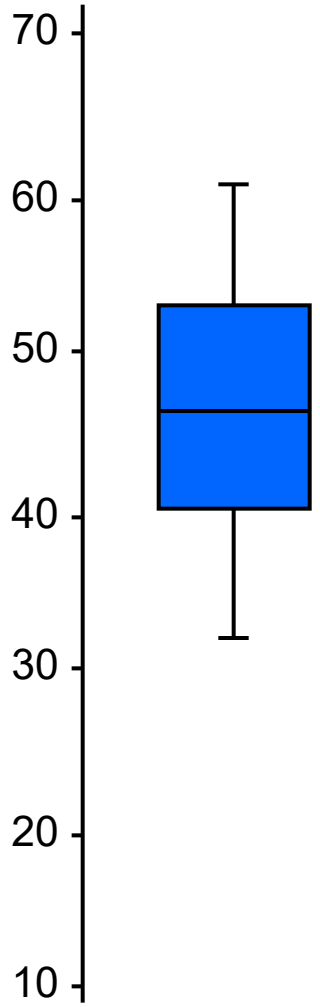
Exemplos:



Assimétrica negativa



Assimétrica positiva



Simétrica

Profª. Clause Piana

O que fazer quando identificamos valores atípicos?

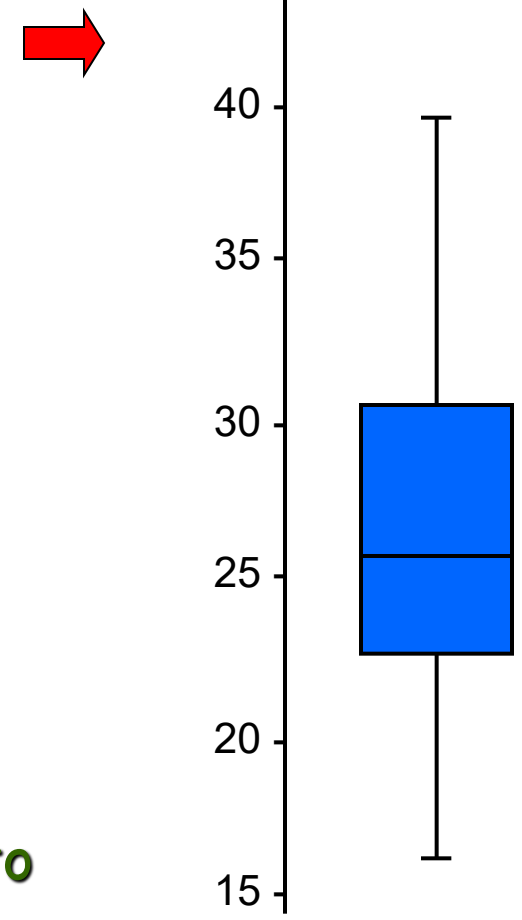
Investigar a sua origem.

- Valores atípicos podem, de fato, fazer parte do conjunto de dados, reforçando a característica assimétrica da distribuição.
- Distribuições com caudas longas têm uma tendência maior de produzir valores atípicos.
- Entretanto, eventualmente, esses valores podem ser oriundos de **erros** na aferição ou no registro dos dados.

Uma inspeção cuidadosa nos dados e nas eventuais causas da ocorrência de valores atípicos é sempre uma providência necessária antes que qualquer atitude seja tomada em relação a esses dados.

Consideremos o conjunto de dados referentes ao peso ao nascer (kg) de bovinos machos da raça Ibagé:

22	30	22	23	30	17	30	30	29	20	20
31	20	18	20	21	33	21	39	23	23	23
30	23	33	23	27	23	18	23	25	25	25
25	16	25	26	26	45	27	30	27	27	28
23	28	28	29	25	29	30	18	19	30	32
20	33	34	34	35	17	36				



O que os dados querem dizer?

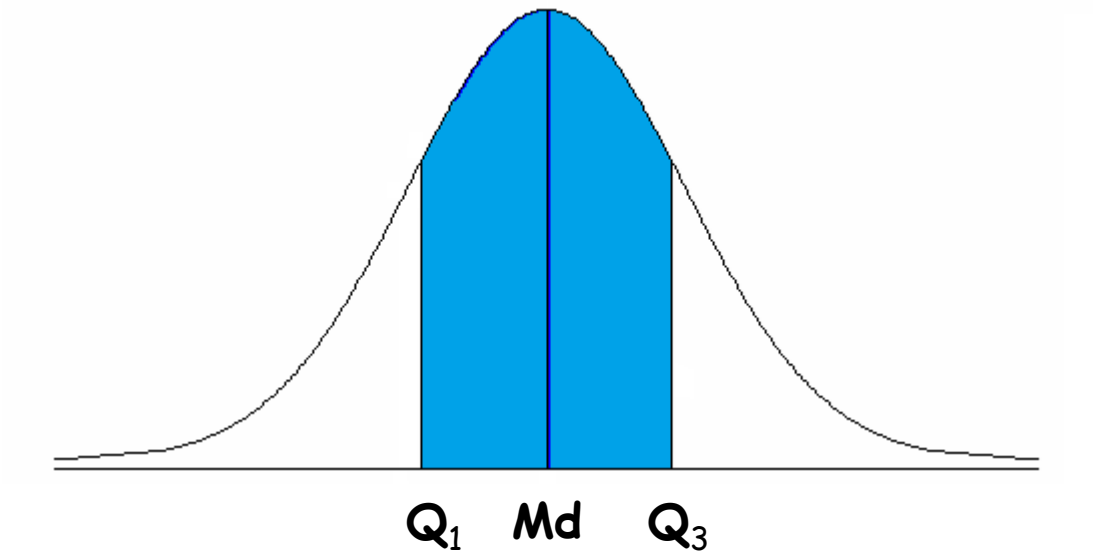
É difícil desvendar que padrões eles escondem.

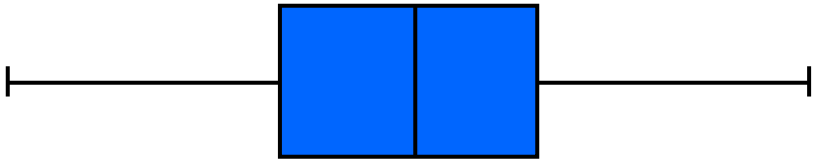
Vários gráficos foram desenvolvidos para revelar estes "padrões escondidos", transformando dados em informação.

O gráfico em caixa tem-se mostrado muito útil pelas análises que permite.

Profª. Clause Piana

Distribuição normal





Exercício proposto:

Construa o gráfico de caixa para os dados abaixo:

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

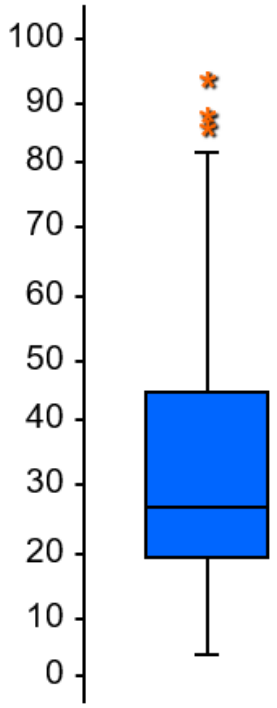
AI = 3,11

Q_1 = 19,27

Md = 27,86

Q_3 = 45,40

AS = 82,70



Profª. Clause Piana

Exercício proposto:

Os dados abaixo são os percentuais de retenção de enxofre em 42 vertentes do nordeste dos EUA, reportados em 1989. Faça o resumo de cinco números e construa o gráfico em caixa.

Ordenação vertical

17,8	25,9	33,2	39,8	43,0	47,8	54,7
18,3	28,2	34,3	41,7	43,2	48,8	56,2
19,0	31,7	37,0	41,9	43,5	49,7	57,4
19,3	32,4	37,6	42,0	44,2	51,0	59,6
21,8	33,0	38,0	42,1	45,3	53,7	60,6
24,3	33,2	39,0	42,2	45,7	53,9	66,2

Diagrama de ramo e folhas

- Trata-se de uma ferramenta exploratória útil para descrever pequenos conjuntos de dados.
- É um procedimento alternativo para resumir um conjunto de valores, com o objetivo de se obter uma ideia da forma de sua distribuição, semelhante a um histograma.
- O método fornece uma boa visão geral dos dados sem que haja perda de informação.
- O diagrama de ramo e folhas é uma boa opção quando temos em mãos somente os dados, caneta e papel.

Exemplo:

Consideremos os seguintes dados relativos às notas de 40 alunos em uma prova de Estatística.

78	59	86	94	43	56	78	84
57	49	96	68	67	65	75	73
67	87	84	45	56	94	87	56
85	76	86	79	78	77	59	76
68	49	86	87	83	94	85	96

Construção do diagrama

1º passo: separação dos dados, alocando todos os valores que pertencem à mesma dezena na mesma linha.

78	59	86	94	43	56	78	84
57	49	96	68	67	65	75	73
67	87	84	45	56	94	87	56
85	76	86	79	78	77	59	76
68	49	86	87	83	94	85	96

43 49 45 49

59 56 57 56 56 59

Construção do diagrama

1º passo: separação dos dados, alocando todos os valores que pertencem à mesma dezena na mesma linha.

43 49 45 49

59 56 57 56 56 59

68 67 65 67 68

78 78 75 73 76 79 78 77 76

86 84 89 87 84 87 85 86 86 87 83 85

94 96 94 94 96

Construção do diagrama

2º passo: apresentar o primeiro dígito que corresponde à dezena apenas uma vez em cada linha, à esquerda, separando-o dos demais dígitos por meio de uma linha vertical.

4| 3 9 5 9
59 56 57 56 56 59
68 67 65 67 68
78 78 75 73 76 79 78 77 76
86 84 89 87 84 87 85 86 86 87 83 85
94 96 94 94 96

Construção do diagrama

2º passo: apresentar o primeiro dígito que corresponde à dezena apenas uma vez em cada linha, à esquerda, separando-o dos demais dígitos por meio de uma linha vertical.

4	3 9 5 9
5	9 6 7 6 6 9
6	8 7 5 7 8
7	8 8 5 3 6 9 8 7 6
8	6 4 9 7 4 7 5 6 6 7 3 5
9	4 6 4 4 6

Cada linha é denominada **ramo**, cada número no ramo à esquerda da linha vertical é chamado **rótulo do ramo** e cada número à direita da linha vertical é denominado **folha**.

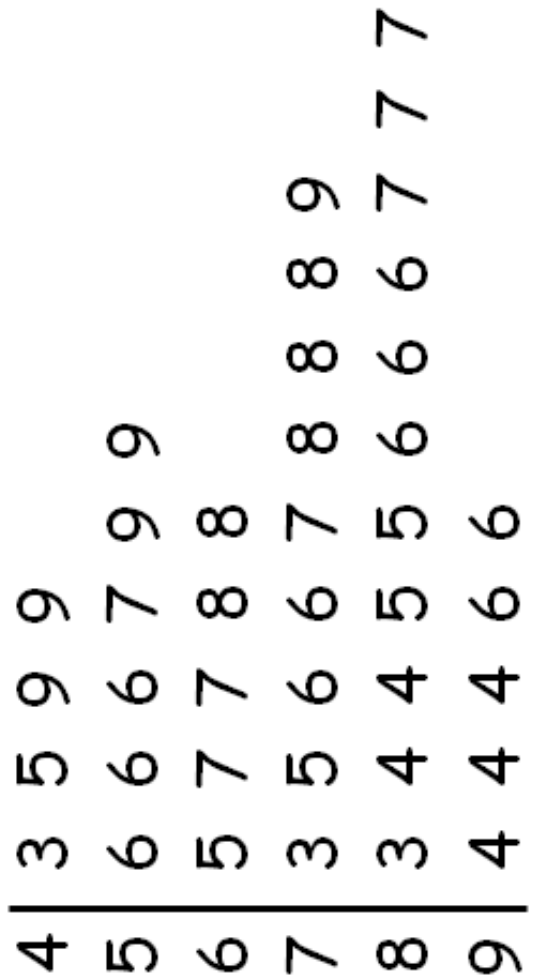
Construção do diagrama

3º passo: Ordenar as folhas nos ramos do diagrama.

4	3 5 9 9
5	6 6 6 7 9 9
6	5 7 7 8 8
7	3 5 6 6 7 8 8 8 9
8	3 4 4 5 5 6 6 6 7 7 7
9	4 4 4 6 6

Chave: 4|3 = 43

Interpretação do diagrama



Assimétrica negativa

Profª. Clause Piana

Outras maneiras de organizar os dados

- Os rótulos dos ramos ou as folhas podem ser de dois dígitos.

  Exemplo: 240, 242, 245, 248, 249

  24| 0 2 5 8 9 ou 2| 40 42 45 48 49

- Em casos de muitos valores, pode ser necessário obter mais ramos, repetindo cada rótulo de ramo, por exemplo, duas vezes, sendo o primeiro com as folhas de 0 a 4 e o segundo com as folhas de 5 a 9. Esse tipo de diagrama é chamado **diagrama de ramos duplos**.

- Um diagrama de ramo e folhas pode ainda ser complementando com informações adicionais, como o número de observações em cada ramo.

Exercício proposto:

Os dados abaixo se referem aos tempos de resposta (em picos por segundo) de 30 circuitos integrados:

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

Construa o diagrama de ramo e folhas e classifique distribuição quanto à simetria.

Ramos simples

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

Chave: 3|3 = 3,3

Ramos duplo

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

Distribuição assimétrica positiva

Histograma

Há perda de informação

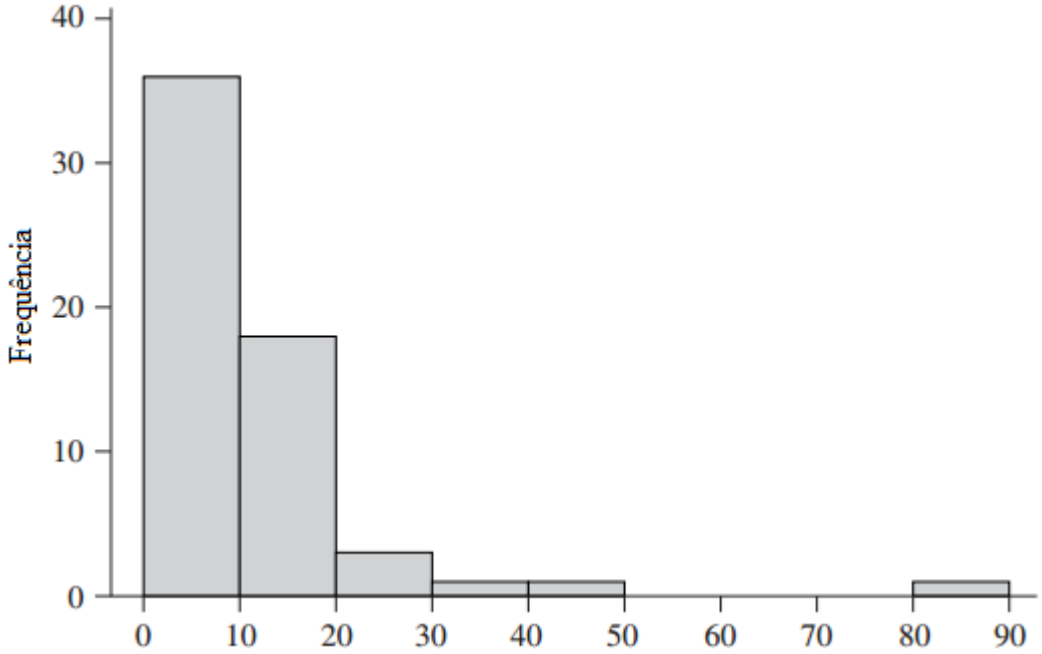
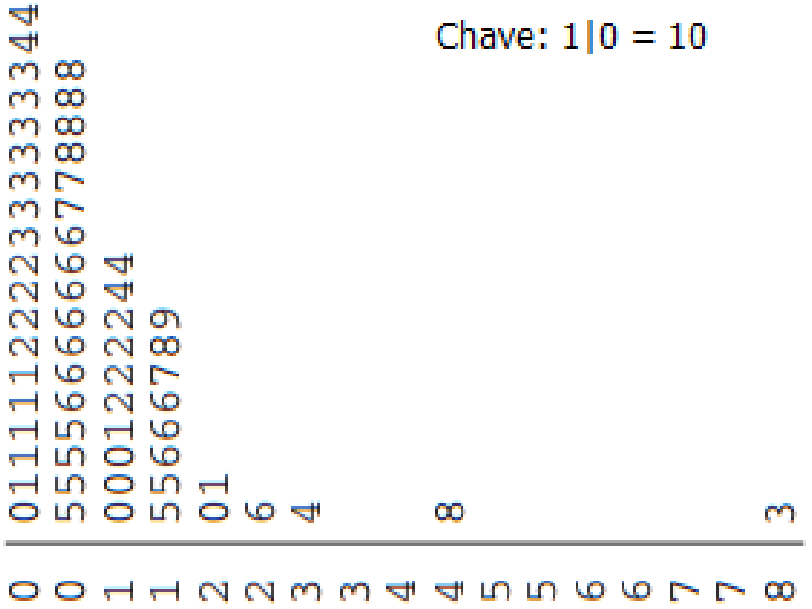


Diagrama de ramos e folhas

Não há perda de informação



Profª. Clause Piana

Exemplo resolvido:

Os dados abaixo são rendimentos de uma batelada de um processo químico.

61	62	64	65	65
66	70	71	71	73
75	77	78	78	79
81	83	84	84	87
88	88	92	93	95

Faça o diagrama de ramo e folhas para representar esses dados.

1) Ramos simples

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

Chave: 6|1 = 61

2) Ramos duplos

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

Chave: 6|1 = 61

3) Ramos divididos em cinco partes

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

Chave: 6|1 = 61

Na figura 1 o gráfico tem muito poucos ramos não provendo muita informação sobre os dados.

Na figura 2 cada ramo foi dividido em duas partes resultando em uma apresentação mais adequada dos dados.

Na figura 3 há um número excessivo de ramos que não diz muito sobre a forma da distribuição.

Profª. Clause Piana

Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

Construa um diagrama de ramo e folhas para esses dados.

Para facilitar a construção do diagrama podemos arredondar os números:
- para uma casa decimal
- para inteiros

Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

Construa um diagrama de ramo e folhas para esses dados.

0	31 89 93
1	08 27 38 52 56 70 74 84 84 93 95 95
2	02 06 22 30 45 46 51 62 63 77 81 81 84
3	20 64 86 90 92
4	10 30 41 47 54 67 87
5	04 28 48 91
6	12
7	03
8	27 58 64
9	33

Chave: 0|31 = 3,1

Profª. Clause Piana

Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

Construa um diagrama de ramo e folhas para esses dados.

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

Chave: 0|3 = 3

Profª. Clause Piana

Bibliografia

FERREIRA, D.F. **Estatística básica**. Lavras: Editora UFLA, 2005.

HOAGLIN, D.C.; MOSTELLER, F.; TUKEY, J.W. **Understanding robust and exploratory data analysis**. New York: John Wiley, 1983.

MONTGOMERY, D.C.; RUNGER, G.C.; HUBELE. N.F. **Estatística Aplicada à Engenharia**. 2 ed. Rio de Janeiro: Editora LTC. 2004. 335p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br/topico.html

VELLEMAN, P.F.; HOAGLIN, D.C. **Applications, basics and computing of exploratory data analysis**. Boston: Duxbury, 1981.