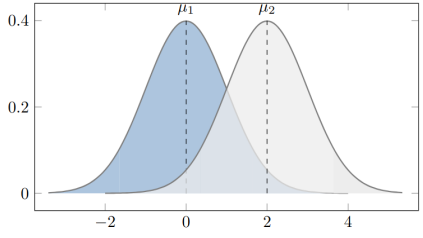
Análise exploratória de dados

Parte 6

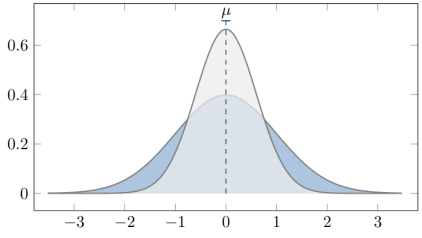
Prof.: Eduardo Vargas Ferreira

Medidas de posição e dispersão

Medidas de posição



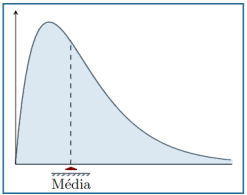
Medidas de dispersão



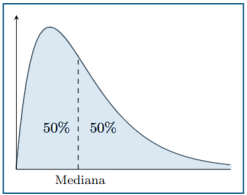
Medidas de Posição

- As medidas de posição (ou localização) são assim denominadas por indicarem um ponto em torno do qual se concentram os dados, p. ex.;

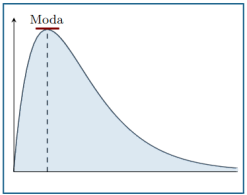
Média aritmética (Me)



Mediana (Md)



Moda (Mo)



Quando usamos um só valor, obtemos uma redução drástica dos dados!

Média Aritmética

Média Aritmética

- A **média aritmética** é a soma de todos os valores observados, dividido pelo total de observações. Considere as quantidades de sódio (mg) em 20 cereais matinais. Encontre a média dos dados.

0	70	125	125
140	150	170	170
180	195	205	210
210	220	220	230
250	260	290	290

$$\bar{x} = \frac{0 + 70 + 125 + \ldots + 290 + 290}{20}$$

$$= 185.5$$

Exemplo: acidentes no mês de janeiro

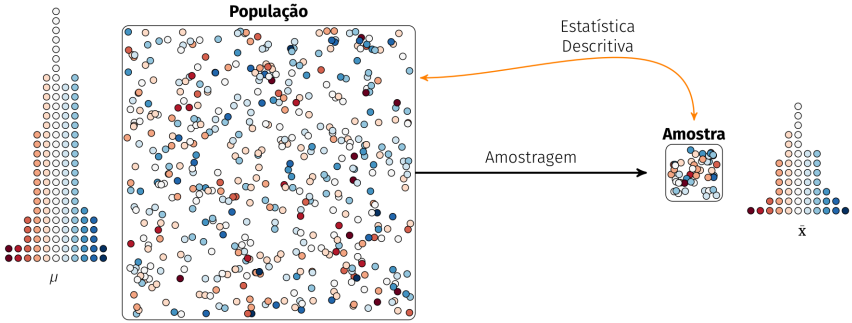
- A tabela abaixo refere-se ao número de acidentes no mês de janeiro. Encontre a média.

Número de acidentes	Frequência em dias
0	18
1	5
2	2
3	2
4	3
5	1
Total	31

$$\bar{x} = \frac{18 \cdot 0 + 5 \cdot 1 + \ldots + 3 \cdot 4 + 1 \cdot 5}{31}$$

$$= 1.03$$

Média Aritmética



Média populacional

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

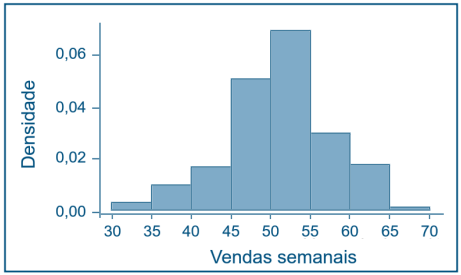
Média amostral

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Exemplo: vendas semanais

- Os dados representam as vendas semanais de vendedores de gêneros alimentícios:

Vendas semanais	Nº de vendedores
$30 \vdash 35$	2
$35 \vdash 40$	10
$40 \vdash 45$	18
$45 \vdash 50$	50
$50 \vdash 55$	70
$55 \vdash 60$	30
$60 \vdash 65$	18
$65 \vdash 70$	2



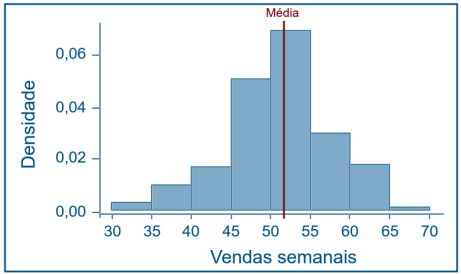
1. Calcule a média da amostra, $\bar{x}$.

$$\bar{x} = \frac{(32.5) \cdot 2 + (37.5) \cdot 10 + \ldots + (62.5) \cdot 18 + (67.5) \cdot 2}{200} = 51.2.$$

Exemplo: vendas semanais

- Os dados representam as vendas semanais de vendedores de gêneros alimentícios:

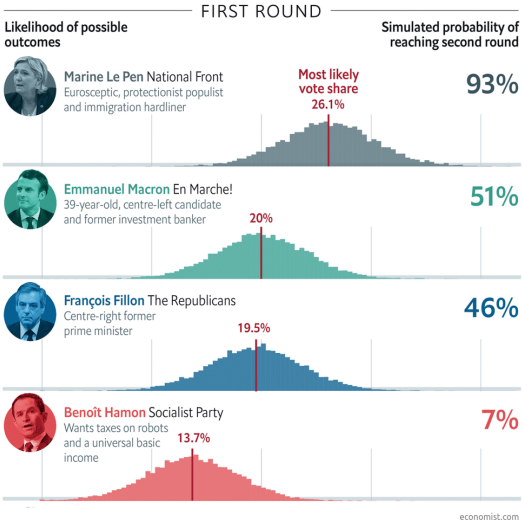
Vendas semanais	Nº de vendedores
$30 \vdash 35$	2
$35 \vdash 40$	10
$40 \vdash 45$	18
$45 \vdash 50$	50
$50 \vdash 55$	70
$55 \vdash 60$	30
$60 \vdash 65$	18
$65 \vdash 70$	2



1. Calcule a média da amostra, $\bar{x}$.

$$\bar{x} = \frac{(32.5) \cdot 2 + (37.5) \cdot 10 + \ldots + (62.5) \cdot 18 + (67.5) \cdot 2}{200} = 51.2.$$

Exemplo: eleição presidencial da França



Exemplo: número de filhos

- Numa pesquisa realizada com 100 famílias, levantaram-se as seguintes informações:

Número de filhos	0	1	2	3	4	5	mais que 5
Frequência de famílias	17	20	28	19	7	4	5

1. Que problemas você enfrentaria para calcular a média? Faça alguma suposição e encontre-a.

   Não temos um valor para "mais que 5".Supondo o valor 6:

$$\bar{x} = \frac{0 \cdot 17 + 1 \cdot 20 + \ldots + 5 \cdot 4 + 6 \cdot 5}{100} = 2,11$$

Exemplo: número de erros de impressão

- Deseja-se estudar o número de erros de impressão em um livro. Para tanto, escolheu-se uma amostra de 50 páginas, obtendo-se o resultado:

Erros	Frequência
0	25
1	20
2	3
3	1
4	1

1. Se o livro tem 500 páginas, qual o número de erros esperado no livro?

$$Me = \frac{(0 \cdot 25 + 1 \cdot 20 + 2 \cdot 3 + 3 \cdot 1 + 4 \cdot 1)}{50}$$

$$= 0,66 \text{ erros/página}$$

Assim, espera-se $500 \times 0,66 = 330$ erros no livro.

Mediana

Mediana

- A mediana (M_d) é o valor que ocupa a posição central da **série ordenada** de observações. Considere as observações ordenadas de forma crescente denotada por:

$$\underbrace{x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq}_{1^{\underline{a}} \text{ metade}} x_{(4)} \underbrace{\leq x_{(5)} \leq x_{(6)} \leq x_{(7)}}_{2^{\underline{a}} \text{ metade}}$$

Exemplo: determine a mediana no conjunto de observações abaixo:

$$5 \quad 130 \quad 8 \quad 7 \quad 1 \quad 8 \quad 2 \quad 8 \quad 11$$

$$\downarrow$$

$$\underbrace{1 \quad 2 \quad 5 \quad 7}_{1^{\underline{a}} \text{ metade}} \quad 8 \quad \underbrace{8 \quad 8 \quad 11 \quad 130}_{2^{\underline{a}} \text{ metade}}$$

Exemplo: cereais matinais

- Considere as quantidades de sódio (mg) em 20 cereais matinais. Encontre a mediana dos dados.

$$\underbrace{0 \;\; 70 \;\; 125 \;\; 125 \;\; 140 \;\; 150 \;\; 170 \;\; 170 \;\; 180 \;\; 195}_{\text{10 observações}} \;\; \underbrace{205 \;\; 210 \;\; 210 \;\; 220 \;\; 220 \;\; 230 \;\; 250 \;\; 260 \;\; 290 \;\; 290}_{\text{10 observações}}$$

$$md = \frac{195 + 205}{2} = 200$$

Mediana

- A mediana (M_d) é o valor que ocupa a posição central da **série ordenada** de observações. Considere as observações ordenadas de forma crescente denotada por:

$$\underbrace{x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)}}_{1^{\underline{a}} \text{ metade}} \leq \underbrace{x_{(5)} \leq x_{(6)} \leq x_{(7)} \leq x_{(8)}}_{2^{\underline{a}} \text{ metade}}$$

$$\downarrow$$

$$md = \frac{x_{(4)} + x_{(5)}}{2}$$

- **Note que a mediana pode ou não fazer parte dos dados observados!**

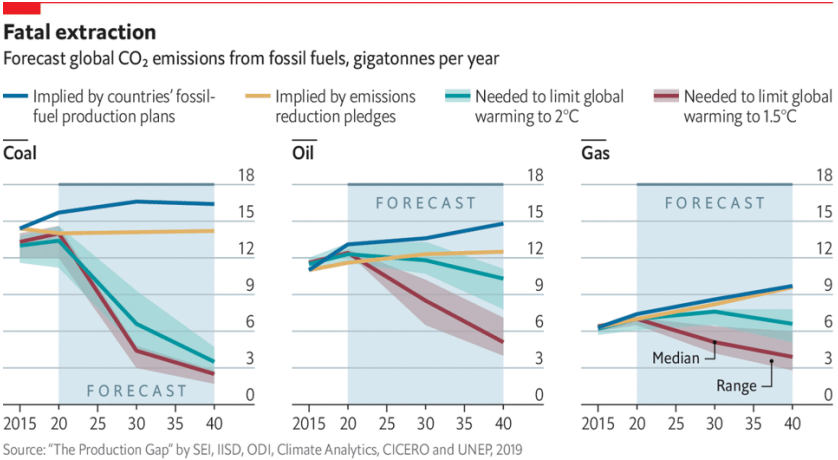
Exemplo: acidentes no mês de janeiro

- A tabela abaixo refere-se ao número de acidentes no mês de janeiro. Encontre a mediana.

Número de acidentes	Frequência em dias
0	18
1	5
2	2
3	2
4	3
5	1
Total	31

$$md = x_{(16)} = 0$$

Exemplo: meta de redução na emissão de gases



Exemplo: número de filhos

- Numa pesquisa realizada com 100 famílias, levantaram-se as seguintes informações:

Número de filhos	0	1	2	3	4	5	mais que 5
Frequência de famílias	17	20	28	19	7	4	5

- Qual a mediana do número de filhos?

Como temos $\underbrace{0\,0\,\ldots\,0}_{17\times}\ \underbrace{1\,1\,\ldots\,1}_{20\times}\ \underbrace{2\,2\,\ldots\,2}_{28\times}$.

Então, md = 2.

Moda

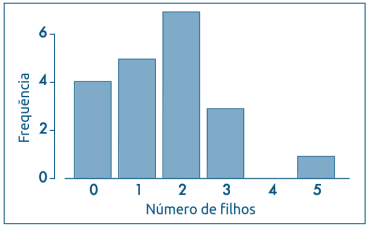
O que seria “estar na moda”?

Moda

- A moda (Mo) é o valor que apresenta a maior freqüência da variável entre os valores observados;

Nº de filhos x_i	**Frequência** n_i
0	4
1	5
2	7
3	3
5	1
Total	20

Figura: Gráfico de barras para a variável: número de filhos.



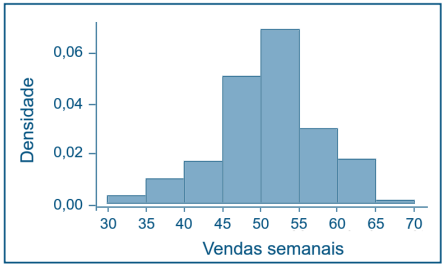
Fonte: Estatística Básica (Bussab e Morettin, 2017)

- Para o caso de valores individuais, a moda pode ser determinada imediatamente observando-se a freqüência absoluta dos dados.

Exemplo: vendas semanais

- Os dados representam as vendas semanais de vendedores de gêneros alimentícios:

Vendas semanais	Nº de vendedores
$30 \vdash 35$	2
$35 \vdash 40$	10
$40 \vdash 45$	18
$45 \vdash 50$	50
$50 \vdash 55$	70
$55 \vdash 60$	30
$60 \vdash 65$	18
$65 \vdash 70$	2

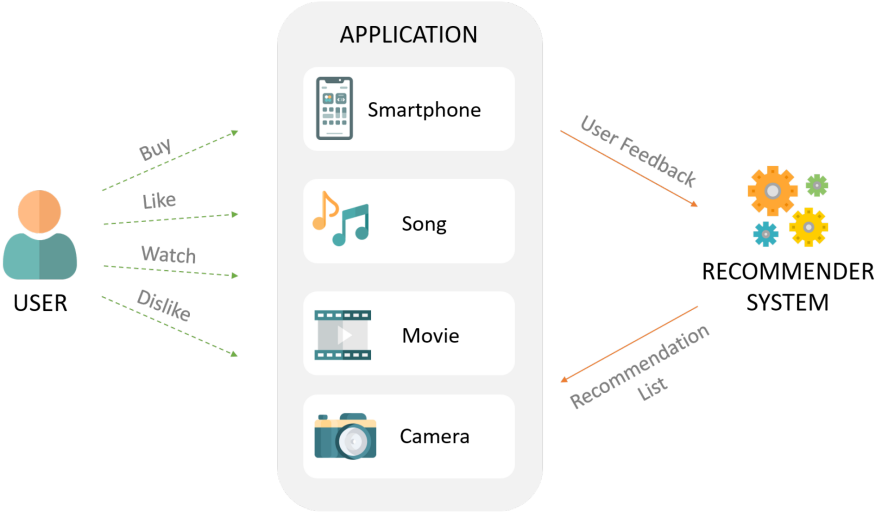


Exemplo: acidentes no mês de janeiro

- A tabela abaixo refere-se ao número de acidentes no mês de janeiro.

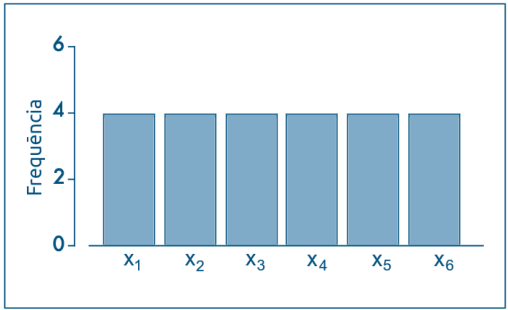
Número de acidentes	Frequência em dias
0	18
1	5
2	2
3	2
4	3
5	1
Total	31

Cold start problem



Distribuição amodal

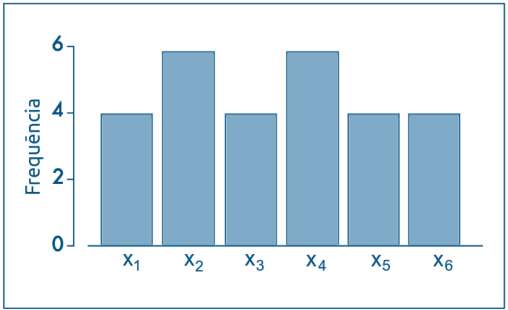
- Um conjunto de dados pode apresentar todos seus elementos com a mesma freqüência absoluta;



- Neste caso não existirá um valor modal, e a distribuição será classificada como **amodal**;

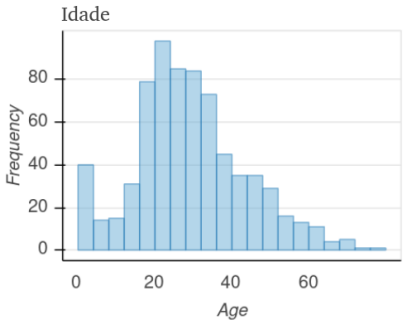
Distribuição plurimodal

- Há casos em que a seqüência de observações apresenta vários elementos com freqüência iguais, implicando numa **distribuição plurimodal**.



Exemplo: investigando o naufrágio do Titanic

Distinct Count	88
Unique (%)	12.3%
Missing	177
Missing (%)	19.9%
Infinite	0
Infinite (%)	0.0%
Memory Size	11.2 KB
Mean	29.6991
Minimum	0.42
Maximum	80
Zeros	0
Zeros (%)	0.0%



Referências

- Bussab, WO; Morettin, PA. Estatística Básica. São Paulo: Editora Saraiva, 2006 (5ª Edição).
- Magalhães, MN; Lima, ACP. Noções de Probabilidade e Estatística. São Paulo: EDUSP, 2008.