

Análise exploratória de dados

Parte 2

Prof.: Eduardo Vargas Ferreira

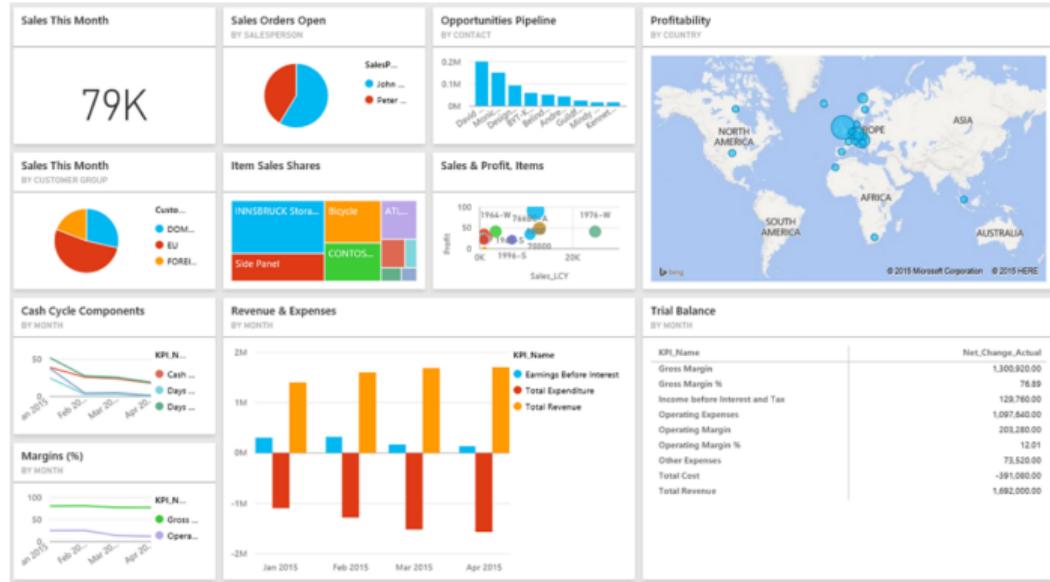


Precisamos aprender como apresentar os dados



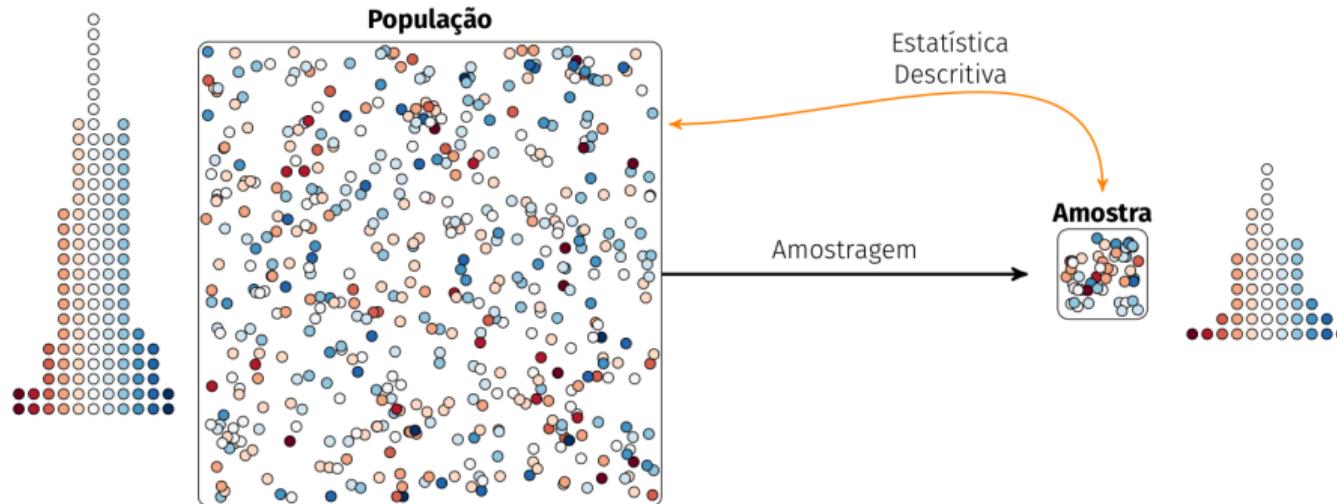
“Os números tem uma importante história para contar. Eles dependem de você dar-lhes um clara e convincente voz.”

-Stephen Few



Papel da estatística descritiva

- A **estatística descritiva** emprega métodos numéricos e gráficos para investigar padrões em um conjunto de dados apresentando-os de uma forma simplificada.



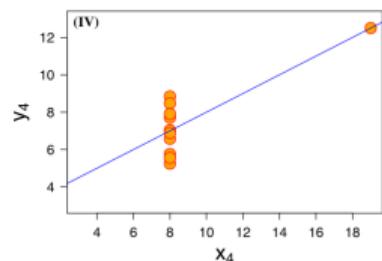
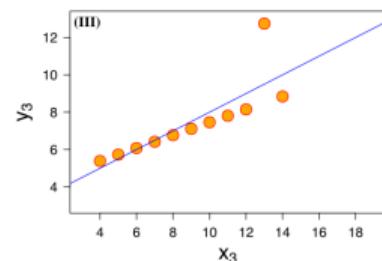
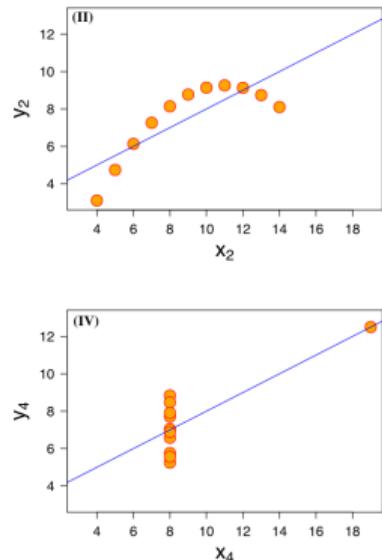
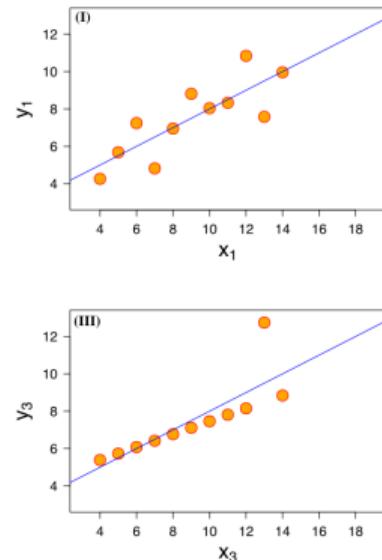
Objetivos com a descrição dos dados



1. Investigar o **comportamento** de uma variável.
2. Examinar a **relação** entre variáveis.
3. Enfatizar a **ordenação/classificação** de elementos/categorias.
4. Compreender a estrutura **organização** dos elementos/categorias.
5. Explorar a **evolução** cronológica de uma variável.
6. Revelar padrões **espaciais** nos dados.
7. Descrever a **conexão** entre elementos/categorias.

Qual o comportamento destes dados?

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Formas para descrição dos dados

Numérica:

- ▶ Tabelas de frequência;
- ▶ Medidas resumo (média, variância, etc).

Gráfica:

- ▶ Gráficos de uma variável;
- ▶ Duas ou mais variáveis.



Organização de dados em tabelas

Exemplo: Doctor Who

- Qual ator atuou no maior número de episódios da série Doctor Who?



Tabela de Frequências Absolutas e Relativas

Ator	Freq. Absoluta	Freq. Relativa
William Hartnell	136	0.157
Patrick Troughton	127	0.147
Jon Pertwee	129	0.149
Tom Baker	173	0.200
Peter Davison	70	0.081
Colin Baker	35	0.041
Sylvester McCoy	42	0.049
Christopher Ecclestone	20	0.023
David Tennant	52	0.060
Matt Smith	51	0.059
Peter Capaldi	29	0.034

Exemplo: cereais matinais

- Qual cereal apresenta menor quantidade de calorias?



Calorias e Carboidratos (Porções de 30g)

Cereal	Calorias	Carboidratos
Sucrilhos	109	26.0
All Bran	81	13.5
Nesfit	102	21.0
Nescau	115	23.0
Snow	113	25.0
Crunch	119	23.0
Moça	113	25.0
Fibra Mais	84	15.0
Froot Loops	113	25.0

Organização de dados em tabelas

Tabela: informações socioeconómicas de 20 funcionários da empresa.

ID	Estado civil	Grau de instrução	Nº de filhos	Salário*	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	-	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	interior
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	-	5,73	20	10	capital
5	solteiro	superior	3	6,26	40	07	interior
6	casado	ensino fundamental	-	6,66	28	00	outra
7	solteiro	ensino fundamental	-	6,86	41	00	interior
8	solteiro	ensino fundamental	2	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	outra
10	solteiro	ensino médio	1	7,44	23	06	interior
11	casado	ensino médio	2	8,12	33	10	interior
12	solteiro	ensino fundamental	5	8,46	27	06	interior
13	solteiro	ensino médio	3	8,74	37	11	interior
14	casado	ensino médio	4	8,95	44	10	interior
15	casado	ensino médio	-	9,13	30	11	capital
16	solteiro	superior	4	9,35	38	06	capital
17	casado	superior	4	9,77	31	10	interior
18	casado	superior	-	9,80	39	08	outra
19	solteiro	ensino fundamental	3	10,53	25	10	interior
20	solteiro	ensino médio	2	10,76	37	08	capital

* número de salários mínimos

Fonte: dados hipotéticos

Título: texto conciso, indicador do conteúdo de uma tabela

Tabela: informações socioeconómicas de 20 funcionários da empresa.

ID	Estado civil	Grau de instrução	Nº de filhos	Salário*	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	-	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	interior
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	-	5,73	20	10	capital
5	solteiro	superior	3	6,26	40	07	interior
6	casado	ensino fundamental	-	6,66	28	00	outra
7	solteiro	ensino fundamental	-	6,86	41	00	interior
8	solteiro	ensino fundamental	2	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	outra
10	solteiro	ensino médio	1	7,44	23	06	interior
11	casado	ensino médio	2	8,12	33	10	interior
12	solteiro	ensino fundamental	5	8,46	27	06	interior
13	solteiro	ensino médio	3	8,74	37	11	interior
14	casado	ensino médio	4	8,95	44	10	interior
15	casado	ensino médio	-	9,13	30	11	capital
16	solteiro	superior	4	9,35	38	06	capital
17	casado	superior	4	9,77	31	10	interior
18	casado	superior	-	9,80	39	08	outra
19	solteiro	ensino fundamental	3	10,53	25	10	interior
20	solteiro	ensino médio	2	10,76	37	08	capital

* número de salários mínimos

Fonte: dados hipotéticos

Cabeçalho: especifica o conteúdo das colunas

Tabela: informações socioeconómicas de 20 funcionários da empresa.

ID	Estado civil	Grau de instrução	Nº de filhos	Salário*	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	-	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	interior
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	-	5,73	20	10	capital
5	solteiro	superior	3	6,26	40	07	interior
6	casado	ensino fundamental	-	6,66	28	00	outra
7	solteiro	ensino fundamental	-	6,86	41	00	interior
8	solteiro	ensino fundamental	2	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	outra
10	solteiro	ensino médio	1	7,44	23	06	interior
11	casado	ensino médio	2	8,12	33	10	interior
12	solteiro	ensino fundamental	5	8,46	27	06	interior
13	solteiro	ensino médio	3	8,74	37	11	interior
14	casado	ensino médio	4	8,95	44	10	interior
15	casado	ensino médio	-	9,13	30	11	capital
16	solteiro	superior	4	9,35	38	06	capital
17	casado	superior	4	9,77	31	10	interior
18	casado	superior	-	9,80	39	08	outra
19	solteiro	ensino fundamental	3	10,53	25	10	interior
20	solteiro	ensino médio	2	10,76	37	08	capital

* número de salários mínimos

Fonte: dados hipotéticos

Corpo: linhas e colunas com informações sobre as variáveis

Tabela: informações socioeconómicas de 20 funcionários da empresa.

ID	Estado civil	Grau de instrução	Nº de filhos	Salário*	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	-	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	interior
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	-	5,73	20	10	capital
5	solteiro	superior	3	6,26	40	07	interior
6	casado	ensino fundamental	-	6,66	28	00	outra
7	solteiro	ensino fundamental	-	6,86	41	00	interior
8	solteiro	ensino fundamental	2	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	outra
10	solteiro	ensino médio	1	7,44	23	06	interior
11	casado	ensino médio	2	8,12	33	10	interior
12	solteiro	ensino fundamental	5	8,46	27	06	interior
13	solteiro	ensino médio	3	8,74	37	11	interior
14	casado	ensino médio	4	8,95	44	10	interior
15	casado	ensino médio	-	9,13	30	11	capital
16	solteiro	superior	4	9,35	38	06	capital
17	casado	superior	4	9,77	31	10	interior
18	casado	superior	-	9,80	39	08	outra
19	solteiro	ensino fundamental	3	10,53	25	10	interior
20	solteiro	ensino médio	2	10,76	37	08	capital

* número de salários mínimos

Fonte: dados hipotéticos

Notas: servem para esclarecer o conteúdo das tabelas

Tabela: informações socioeconómicas de 20 funcionários da empresa.

ID	Estado civil	Grau de instrução	Nº de filhos	Salário*	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	-	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	interior
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	-	5,73	20	10	capital
5	solteiro	superior	3	6,26	40	07	interior
6	casado	ensino fundamental	-	6,66	28	00	outra
7	solteiro	ensino fundamental	-	6,86	41	00	interior
8	solteiro	ensino fundamental	2	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	outra
10	solteiro	ensino médio	1	7,44	23	06	interior
11	casado	ensino médio	2	8,12	33	10	interior
12	solteiro	ensino fundamental	5	8,46	27	06	interior
13	solteiro	ensino médio	3	8,74	37	11	interior
14	casado	ensino médio	4	8,95	44	10	interior
15	casado	ensino médio	-	9,13	30	11	capital
16	solteiro	superior	4	9,35	38	06	capital
17	casado	superior	4	9,77	31	10	interior
18	casado	superior	-	9,80	39	08	outra
19	solteiro	ensino fundamental	3	10,53	25	10	interior
20	solteiro	ensino médio	2	10,76	37	08	capital

* número de salários mínimos

Fonte: dados hipotéticos

Fonte: identificador do responsável pelo fornecimento dos dados

Tabela: informações socioeconómicas de 20 funcionários da empresa.

ID	Estado civil	Grau de instrução	Nº de filhos	Salário*	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	-	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	interior
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	-	5,73	20	10	capital
5	solteiro	superior	3	6,26	40	07	interior
6	casado	ensino fundamental	-	6,66	28	00	outra
7	solteiro	ensino fundamental	-	6,86	41	00	interior
8	solteiro	ensino fundamental	2	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	outra
10	solteiro	ensino médio	1	7,44	23	06	interior
11	casado	ensino médio	2	8,12	33	10	interior
12	solteiro	ensino fundamental	5	8,46	27	06	interior
13	solteiro	ensino médio	3	8,74	37	11	interior
14	casado	ensino médio	4	8,95	44	10	interior
15	casado	ensino médio	-	9,13	30	11	capital
16	solteiro	superior	4	9,35	38	06	capital
17	casado	superior	4	9,77	31	10	interior
18	casado	superior	-	9,80	39	08	outra
19	solteiro	ensino fundamental	3	10,53	25	10	interior
20	solteiro	ensino médio	2	10,76	37	08	capital

* número de salários mínimos

Fonte: dados hipotéticos

Tabelas e gráficos devem ter significado próprio



Martin LeBlanc

@martinleblanc

Seguir

data visualization
A ~~user interface~~ is like a joke. If you have to
explain it, it's not that good.

10:56 - 14 de mai de 2014

2.811 Retweets 1.939 Curtidas



61

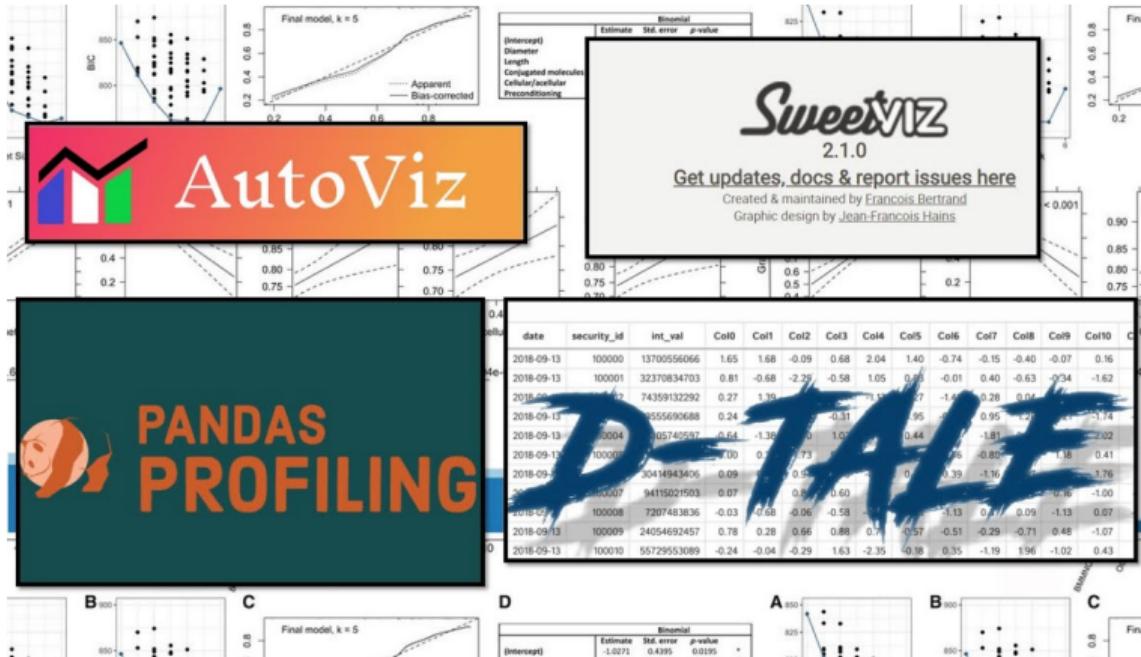
2,8 mil

1,9 mil

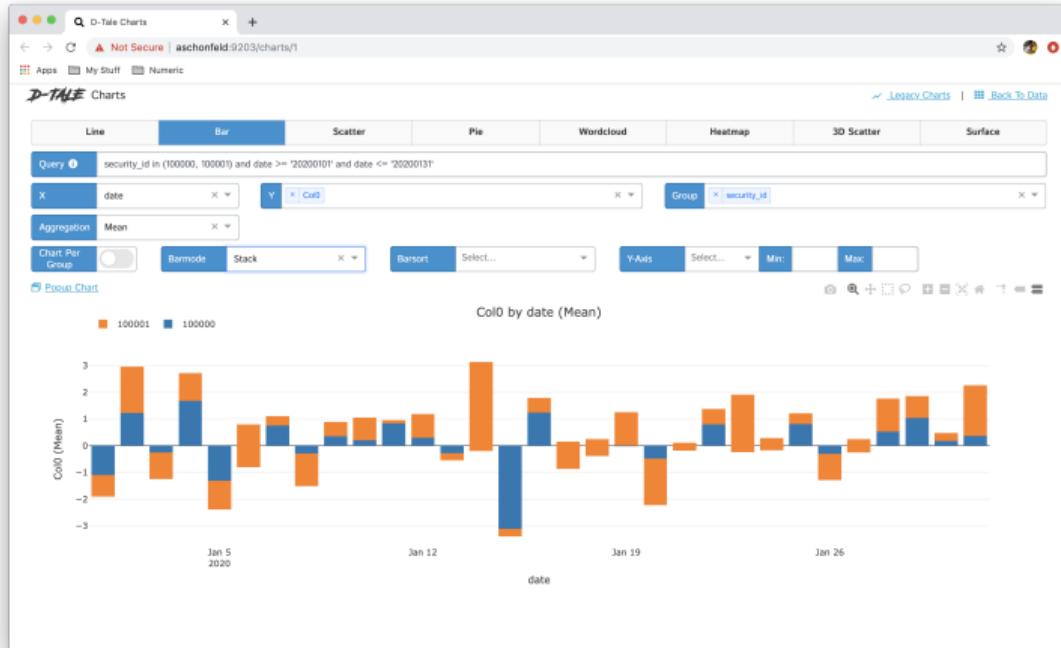
Como apresentar dados realmente grandes?



Já existem abordagens automáticas



Já existem abordagens automáticas



Já existem abordagens automáticas



Distribuição de frequências

Distribuição de frequências

ID	Estado civil	Grau de instrução
1	solteiro	ensino fundamental
2	casado	ensino fundamental
3	casado	ensino fundamental
4	solteiro	ensino médio
5	solteiro	superior
6	casado	ensino fundamental
7	solteiro	ensino fundamental
8	solteiro	ensino fundamental
9	casado	ensino médio
10	solteiro	ensino médio
11	casado	ensino médio
12	solteiro	ensino fundamental
13	solteiro	ensino médio
14	casado	ensino médio
:	:	:
32	solteiro	superior
33	casado	superior
34	casado	superior
35	solteiro	ensino fundamental
36	solteiro	ensino médio

- Quando se estuda uma variável, o maior interesse está em conhecer o **comportamento** dela.

Grau de instrução	Frequência n_i	Proporção f_i	Porcentagem $100f_i$
Fundamental	12	0.33	33%
Médio	18	0.50	50%
Superior	6	0.17	17%
Total	36	1.00	100%

Tipos de distribuição de frequências

- **Distribuição de frequências simples:** é uma tabela em que os valores da variável aparecem individualmente relacionados as suas frequências.

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

Interpretação: nos 31 dias do mês de janeiro,

- 18 dias não houve acidentes;
- 5 dias ocorreu 1 acidente;
- 2 dias ocorreram 2 acidentes.

:

Exemplo: número de irmãos

- Os dados referem-se ao nº de irmãos dos alunos da turma de Estatística:

1 - 3 - 0 - 5 - 2 - 1 - 1 - 0 - 0 - 1 - 4 - 3 - 1 - 0 - 1 - 2 - 2 - 1 - 3 - 1

- Para construir as frequências simples, primeiro ordenamos os dados:

0 - 0 - 0 - 0 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 2 - 2 - 2 - 3 - 3 - 3 - 4 - 5

Índice	Número de irmãos	Frequência (n_i)
1	0	4
2	1	8
3	2	3
4	3	3
5	4	1
6	5	1
Total		$\sum n_i = 20$

Frequênciá simples *vs* Frequênciá relativa

Frequênciá simples:

n_i = número de observações no nível i

Frequênciá relativa:

$f_i = \frac{\text{número de observações no nível } i}{\text{total das observações}}$

Índice	Número de irmãos	Frequênciá (n_i)	Frequênciá relativa (f_i)
1	0	4	0,2
2	1	8	0,4
3	2	3	0,15
4	3	3	0,15
5	4	1	0,05
6	5	1	0,05
Total		$\sum n_i = 20$	$\sum f_i = 1$

$$f_1 = \frac{n_1}{\sum n_i} = \frac{4}{20} = 0.2$$

$$f_2 = \frac{n_2}{\sum n_i} = \frac{8}{20} = 0.4$$

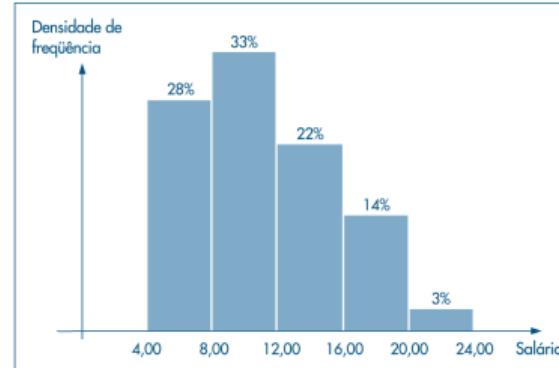
Tipos de distribuição de frequências

- **Distribuição de frequências por classes:** É uma tabela em que os valores observados são agrupados em intervalos (classes) de variação.

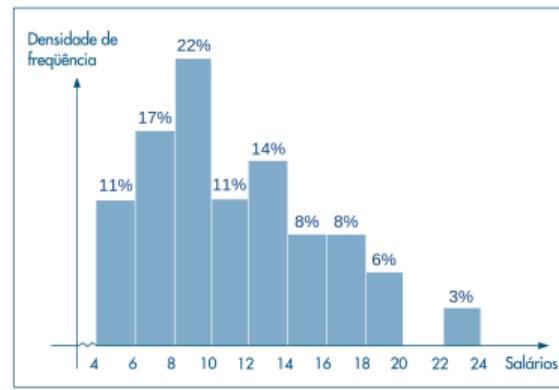
Classe de salários	Ponto médio	Frequência	Porcentagem
4,00 ← 8,00	6	10	27.78
8,00 ← 12,00	10	12	33.33
12,00 ← 16,00	14	8	22.22
16,00 ← 20,00	18	5	13.89
20,00 ← 24,00	22	1	2.78
Total	-	36	100

Distribuição de frequência variando o número de classes

Classe de salários	Frequência
4 ⊢ 8	10
8 ⊢ 12	12
12 ⊢ 16	8
16 ⊢ 20	5
20 ⊢ 24	1
Total	36



Classe de salários	Frequência
4 ⊢ 6	4
6 ⊢ 8	6
8 ⊢ 10	8
10 ⊢ 12	4
12 ⊢ 14	5
14 ⊢ 16	3
16 ⊢ 18	3
18 ⊢ 20	2
20 ⊢ 22	0
22 ⊢ 24	1
Total	36



Fonte: Estatística Básica (Bussab e Morettin, 2017)

Distribuição de frequências por classes

(a) Determinar a **amplitude total** dos dados (A_T);

(b) Determinar o **número de classes** (k):

Regra da raiz quadrada: $\begin{cases} k = 5, & \text{se } n \leq 25 \\ k = \sqrt{n}, & \text{c.c.} \end{cases}$

Regra de Sturges: $k = 1 + \log_{10} n$.

Classe de salários
4,00 ⌄ 8,00
8,00 ⌄ 12,00
12,00 ⌄ 16,00
16,00 ⌄ 20,00
20,00 ⌄ 24,00

(c) Determinar a **amplitude de classe** (h) que pode ser dada por $h = \frac{A_T}{k}$;

(d) Determinar os **limites de classes**. O limite da primeira classe e o limite da última, não precisam, necessariamente, pertencer ao conjunto.

Exemplo: distribuição das notas

- Os dados a seguir representam as notas de 50 alunos.

33	35	35	39	41	41	42	45	47	48
50	52	53	54	55	55	57	59	60	60
61	64	65	65	65	66	66	66	67	68
69	71	73	73	74	74	76	77	77	78
80	81	84	85	85	88	89	91	94	97

Classe de notas	Frequência n_i	Porcentagem $100f_i$
30 ⊢ 40		
40 ⊢ 50		
50 ⊢ 60		
⋮		
90 ⊢ 100		
Total		

- (a) **Amplitude total:** $A_T = 97 - 33 = 64$;
- (b) **Número de classes:** $k = \sqrt{50} \approx 7$;
- (c) **Amplitude das classes:** $h = \frac{A_T}{k} = \frac{64}{7} = 9,14 \approx 10$.
- (d) **Limites de classes:** $30 ⊢ 40, 40 ⊢ 50, \dots$

Exemplo: distribuição das notas

- Os dados a seguir representam as notas de 50 alunos.

33	35	35	39	41	41	42	45	47	48
50	52	53	54	55	55	57	59	60	60
61	64	65	65	65	66	66	66	67	68
69	71	73	73	74	74	76	77	77	78
80	81	84	85	85	88	89	91	94	97



Classe de notas	Frequência n_i	Porcentagem $100f_i$
30 \leftarrow 40	4	8%
40 \leftarrow 50	6	12%
50 \leftarrow 60	8	16%
\vdots	\vdots	\vdots
90 \leftarrow 100	3	6%
Total	50	100%

- Existem diversas maneiras de expressar as classes:

$$1. \ a \leftarrow b;$$

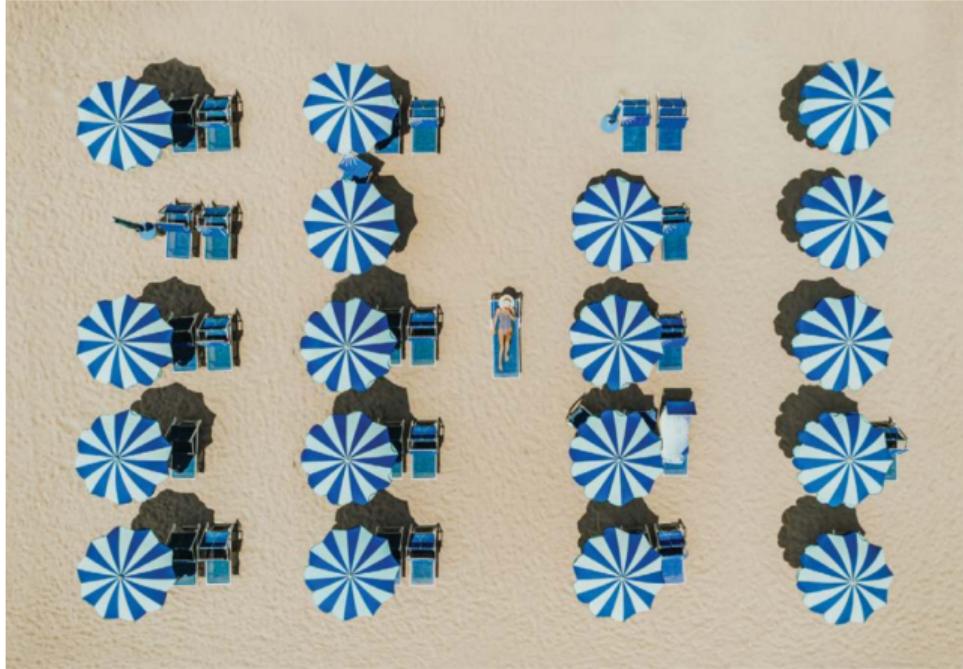
$$2. \ a \dashv b;$$

$$3. \ a - b;$$

Devemos atrair a atenção da audiência para o que queremos



Devemos atrair a atenção da audiência para o que queremos



Exemplo: clientes separados por classes

Classe	# de contas	% de contas	Receita (\$M)	% Receita
A+	19	2%	\$3,9	21%
A	77	7%	\$4,7	25%
B	338	31%	\$6,0	32%
C	425	39%	\$2,8	15%
D	24	2%	\$0,4	2%
Outros	205	19%	\$0,9	5%
TOTAL	1088	100%	\$18,7	100%

Classe	Contas		Receita	
	#	%	\$(M)	%
A+	19	2%	\$3,9	21%
A	77	7%	\$4,7	25%
B	338	31%	\$6,0	32%
C	425	39%	\$2,8	15%
D	24	2%	\$0,4	2%
Outros	205	19%	\$0,9	5%
TOTAL	1088	100%	\$18,7	100%

Exemplo: refeições servidas ao longo do tempo

Ano	Refeições servidas	Ano	Refeições servidas
2010	40139	2010	40139
2011	127020	2011	127020
2012	168193	2012	168193
2013	153115	2013	153115
2014	202102	2014	202102
2015	232897	2015	232897
2016	277912	2016	277912
2017	205350	2017	205350
2018	233389	2018	233389
2019	232797	2019	232797

Exemplo: compartilhamento de carteira por categoria



Referências

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

