SME0803 Visualização e Exploração de Dados

Organização e natureza dos dados

Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele

Baseado em

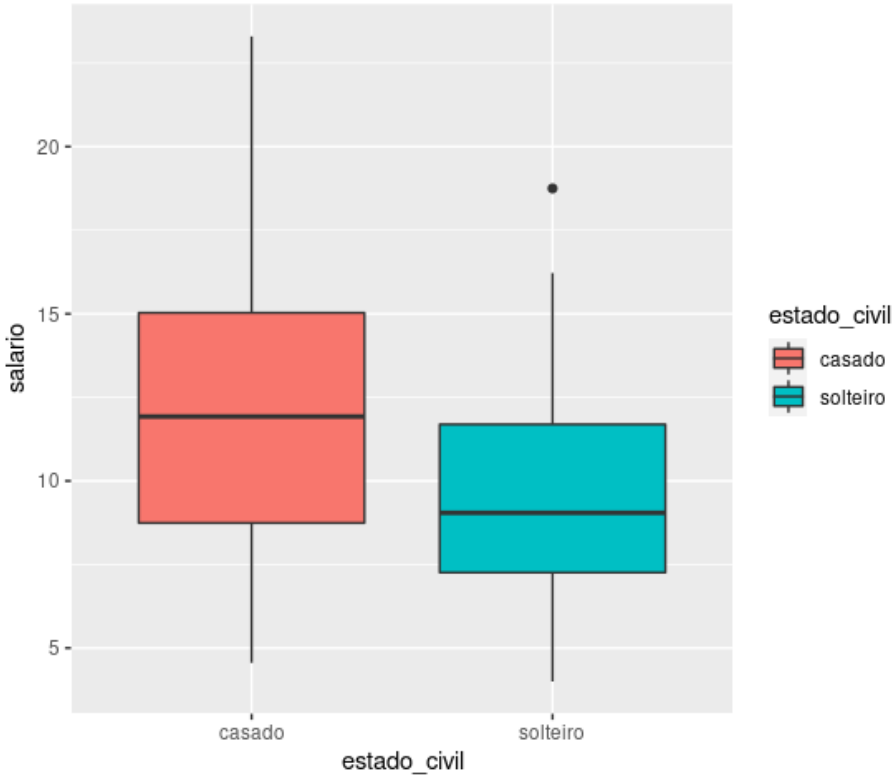
Bussab e Morettin, Estatística Básica. São Paulo: Saraiva 2010. 6a Edição.

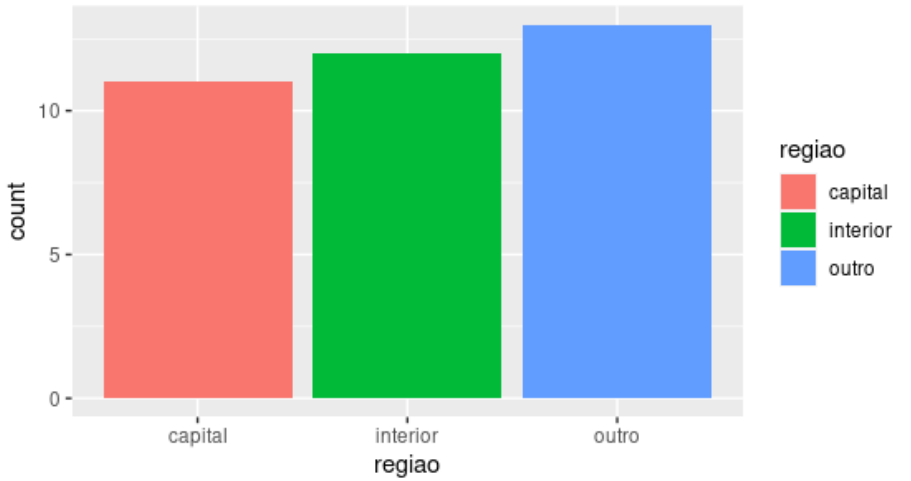
Notas de aula do Prof. Mário de Castro.

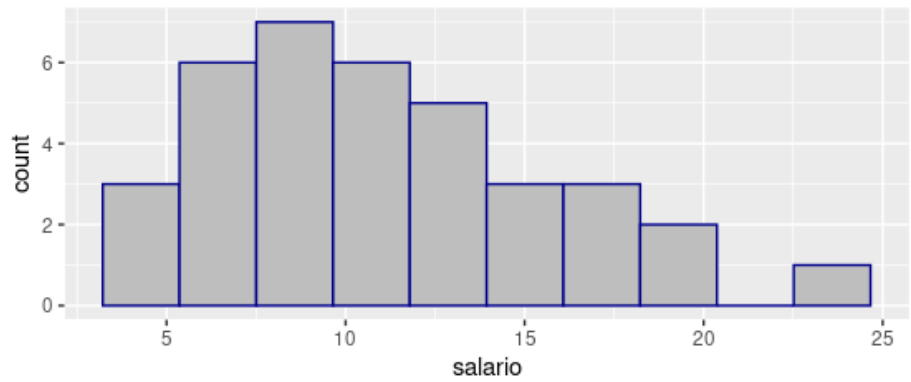
Resumo da aula

- Tipos de dados.
- Organização de banco de dados.
- Precisão e arredondamento de dados quantitativos

A natureza dos dados







A natureza dos dados

Trataremos aqui de **dados retangulares**, que tem nas linhas as **unidades amostrais** e nas colunas as **variáveis**.

A natureza dos dados

Trataremos aqui de **dados retangulares**, que tem nas linhas as **unidades amostrais** e nas colunas as **variáveis**.

Variável

Qualquer característica dos elementos em estudo e que temos interesse em medir de alguma forma.

A natureza dos dados

Os principais tipos de variáveis são:

- **Qualitativas** (não-numéricas)
  - **Nominais**: sexo, cor da pele, status de fumante (fumante ou não-fumante), classificação de inadimplente (sim ou não)

A natureza dos dados

Os principais tipos de variáveis são:

- **Qualitativas** (não-numéricas)
  - **Nominais**: sexo, cor da pele, status de fumante (fumante ou não-fumante), classificação de inadimplente (sim ou não)
  - **Ordinais**: escolaridade (em categorias), grau de satisfação (muito satisfeito, satisfeito, insatisfeito, muito insatisfeito), idade (em faixas)

A natureza dos dados

Os principais tipos de variáveis são:

- **Qualitativas** (não-numéricas)
  - **Nominais**: sexo, cor da pele, status de fumante (fumante ou não-fumante), classificação de inadimplente (sim ou não)
  - **Ordinais**: escolaridade (em categorias), grau de satisfação (muito satisfeito, satisfeito, insatisfeito, muito insatisfeito), idade (em faixas)
- **Quantitativas** (numéricas)
  - **Discretas**: número de defeitos em uma peça, número de produtos contratados

A natureza dos dados

Os principais tipos de variáveis são:

- **Qualitativas** (não-numéricas)
  - **Nominais**: sexo, cor da pele, status de fumante (fumante ou não-fumante), classificação de inadimplente (sim ou não)
  - **Ordinais**: escolaridade (em categorias), grau de satisfação (muito satisfeito, satisfeito, insatisfeito, muito insatisfeito), idade (em faixas)
- **Quantitativas** (numéricas)
  - **Discretas**: número de defeitos em uma peça, número de produtos contratados
  - **Contínuas**: peso, idade, pressão sanguínea, valor contratado de um produto

A natureza dos dados

Uma segunda classificação (por escala de medição)

(Bussab e Morettin, 2010, pág 14.)

- **Nominal** (exemplo: sexo).
  Valores não-numéricos. Atributos. Não é possível realizar operações aritméticas.
  Comparação: $\neq$.

A natureza dos dados

Uma segunda classificação (por escala de medição)

(Bussab e Morettin, 2010, pág 14.)

- **Nominal** (exemplo: sexo).
  Valores não-numéricos. Atributos. Não é possível realizar operações aritméticas.
  Comparação: $\neq$.
- **Ordinal** (exemplo: categorias de escolaridade)
  Valores não-numéricos com relação de ordem. Não é possível realizar operações aritméticas.
  Comparação: $\neq, >, <$.

A natureza dos dados

Uma segunda classificação (por escala de medição)

- **Intervalar** (exemplo: temperatura)
  Valores numéricos com origem arbitrária. Valor nulo não significa inexistência e podem comportar valores negativos. Não é possível realizar a operação de divisão.
  Comparação: $\neq, >, <$, magnitude da diferença.

A natureza dos dados

Uma segunda classificação (por escala de medição)

- **Intervalar** (exemplo: temperatura)
  Valores numéricos com origem arbitrária. Valor nulo não significa inexistência e podem comportar valores negativos. Não é possível realizar a operação de divisão.
  Comparação: $\neq, >, <$, magnitude da diferença.
- **Razão** (exemplo: idade)
  Valores numéricos com origem bem definida. É possível realizar operações aritméticas.
  Comparação: $\neq, >, <$, magnitude da diferença, magnitude do quociente.

Organização dos dados

Os estudos em geral envolvem diversas variáveis de diferentes tipos.

Organização dos dados

Os estudos em geral envolvem diversas variáveis de diferentes tipos.

Um conjunto de dados de n observações relativas a p variáveis pode ser armazenado em uma matriz $n \times p$ heterogênea.

Observações $(1, ..., n)$: linhas.

Variáveis $(1, ..., p)$: colunas.

Organização dos dados

Em R: folha de dados (data frame).

	funcionario	estado_civil	instrucao	nfilhos	salario	idade_anos	idade_meses	regiao
1	1	solteiro	ensino_fundamental	*NA*	4.00	26	3	interior
2	2	casado	ensino_fundamental	1	4.56	32	10	capital
3	3	casado	ensino_fundamental	2	5.25	36	5	capital
4	4	solteiro	ensino_medio	*NA*	5.73	20	10	outro
5	5	solteiro	ensino_fundamental	*NA*	6.26	40	7	outro
6	6	casado	ensino_fundamental	0	6.66	28	0	interior
7	7	solteiro	ensino_fundamental	*NA*	6.86	41	0	interior
8	8	solteiro	ensino_fundamental	*NA*	7.39	43	4	capital
9	9	casado	ensino_medio	1	7.59	34	10	capital
10	10	solteiro	ensino_medio	*NA*	7.44	23	6	outro
11	11	casado	ensino_medio	2	8.12	33	6	interior
12	12	solteiro	ensino_fundamental	*NA*	8.46	27	11	capital

Fonte: Bussab e Morettin, 2010, pág 11. Dados da Companhia MB.

Organização dos dados

A folha de dados do slide anterior foi retirada de Bussab e Morettin (2010) e ilustram o seguinte

Exemplo: Companhia MB
Um pesquisador está interessado em fazer um levantamento sobre alguns aspectos socioeconômicos dos empregados da seção de orçamentos da Companhia MB. Usando informações obtidas do departamento pessoal, ele elaborou a tabela descrita no arquivo CompanhiaMB.csv.

Exercício: Classifique as variáveis estado civil, grau de instrução, número de filhos, salário, idade, região. Que valores elas podem assumir?

Organização dos dados

Baixe e instale o R de `www.r-project.org`.

Você pode usar um editor como o RStudio se desejar.

Faça a leitura dos dados da Companhia MB e classifique suas variáveis.

Lendo os dados em R

```
dados <- read.csv("CompanhiaMB.csv",header=TRUE)
attach(dados)
View(dados)
```

Exemplo 2

Considere os dados de 100 mil clientes de um banco no arquivo dados_banco.csv. Estão disponíveis as variáveis:

Cliente: Identificador do cliente.
Sexo: Feminino (F) ou Masculino (M).
Idade: Idade do cliente, em anos completos.
Empresa: Tipo da empresa em que trabalha: Pública, Privada ou Autônomo
Salário: Salário declarado pelo cliente na abertura da conta, em reais.
Saldo_cc: Saldo em conta corrente, em reais.
Saldo_poupança: Saldo em poupança, em reais.
Saldo_investimento: Saldo em investimentos, em reais.
Devedor_cartao: Valor em atraso no cartão de crédito, em reais.
Inadimplente: Se o cliente é considerado inadimplente atualmente (1) ou não (0), de acordo com critérios preestabelecidos.

Classifique as variáveis por tipo e justifique.

Exemplo 2

Classificação das variáveis por tipo

- **Sexo**: qualitativa nominal
- **Idade**: quantitativa contínua
- **Empresa**: qualitativa nominal
- **Salário**: quantitativa contínua
- **Saldo_cc**: quantitativa contínua
- **Saldo_poupança**: quantitativa contínua
- **Saldo_investimento**: quantitativa contínua
- **Devedor_cartão**: quantitativa contínua
- **Inadimplente**: qualitativa nominal (embora numérica)

Exemplo 2

	Cliente	Sexo	Idade	Empresa	Salario	Saldo_cc	Saldo_poupança	Saldo_investimento	Devedor_cartao	Inadimplente
1	1	M	33	Privada	6019	1084.98	0.00	0.00	1214.35	0
2	2	F	31	Pública	5134	532.09	0.00	0.00	1662.96	0
3	3	M	31	Pública	5286	719.91	0.00	0.00	2167.97	0
4	4	F	32	Privada	5534	547.47	0.00	0.00	639.13	1
5	5	F	28	Autônomo	4559	412.81	0.00	0.00	1466.96	1
...	...	...	...	...	...	...	...	...	...	...
99996	99996	F	31	Autônomo	5246	650.93	0.00	0.00	2083.84	1
99997	99997	M	32	Privada	5678	877.58	16881.08	9482.21	0.00	0
99998	99998	M	31	Pública	5430	627.46	0.00	0.00	1239.22	0
99999	99999	F	30	Pública	5070	573.75	0.00	0.00	3427.16	0
100000	100000	M	31	Autônomo	5323	456.93	0.00	0.00	3119.86	0

100000 rows × 10 columns

Precisão e arredondamento

Algarismo significativo

É qualquer algarismo sobre o qual temos certeza na sua determinação.

Zeros à esquerda em um número não são considerados algarismos significativos.

Em inglês: **significant digit** ou **significant figure**.

Precisão e arredondamento

Algarismo significativo

É qualquer algarismo sobre o qual temos certeza na sua determinação.

Zeros à esquerda em um número não são considerados algarismos significativos.

Em inglês: **significant digit** ou **significant figure**.

Representações com k algarismos significativos (a.s.).

- $0,0000a_1a_2\ldots a_ka_{k+1}a_{k+2}$.
  Os algarismos significativos são $a_1,\ a_2,\ \ldots a_k$.
- $a_1a_2, a_3\ldots a_ka_{k+1}a_{k+2}$.
  Os algarismos significativos são $a_1,\ a_2,\ \ldots a_k$.

Regras de arredondamento

Resolução n. 886/66 da Fundação IBGE.

Um número fracionário deve ser arredondado na posição p.

1. **Algarismo na posição p + 1 é < 5: posição p não se altera.**
   1 decimal: 7,429 $\rightarrow$ 7,4.
   2 decimais: 5,324 $\rightarrow$ 5,32.
   2 decimais: –3,4510 $\rightarrow$ –3,45.

Regras de arredondamento

Resolução n. 886/66 da Fundação IBGE.

Um número fracionário deve ser arredondado na posição p.

1. **Algarismo na posição p + 1 é < 5: posição p não se altera.**
   1 decimal: 7,429 $\rightarrow$ 7,4.
   2 decimais: 5,324 $\rightarrow$ 5,32.
   2 decimais: –3,4510 $\rightarrow$ –3,45.
2. **Algarismo na posição p + 1 é > 5: posição p aumenta de uma unidade.**
   1 decimal: 3,18 $\rightarrow$ 3,2.
   2 decimais: 11,2986 $\rightarrow$ 11,30.
   2 decimais: –2,559 $\rightarrow$ –2,56.

Regras de arredondamento

3. **Algarismo na posição p + 1 é = 5 e após a posição p + 1 pelo menos um algarismo é diferente de 0: posição p aumenta de uma unidade.**

   1 decimal: 19,1501 $\rightarrow$ 19,2.

   2 decimais: 6,4254 $\rightarrow$ 6,43.

Regras de arredondamento

3. **Algarismo na posição p + 1 é = 5 e após a posição p + 1 pelo menos um algarismo é diferente de 0: posição p aumenta de uma unidade.**

   1 decimal: 19,1501 $\rightarrow$ 19,2.

   2 decimais: 6,4254 $\rightarrow$ 6,43.

4. **Algarismo na posição p + 1 é = 5 e este é o último algarismo ou se após a posição p + 1 todos os algarismos forem iguais a 0: posição p aumenta de uma unidade somente se for um número ímpar.**

   1 decimal: 2,35 $\rightarrow$ 2,4.

   1 decimal:8,6500 $\rightarrow$ 8,6.

   2 decimais: 3,7350 $\rightarrow$ 3,74.

Exercício. Verificar as regras na calculadora, R, Python, Excel, etc.

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.
3. Em números com valor absoluto $\in (0, 1)$, zeros antecedendo outros algarismos não são a.s.

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.
3. Em números com valor absoluto $\in (0, 1)$, zeros antecedendo outros algarismos não são a.s.
   0,00056 (2 a.s.); -0,000009 (1 a.s.).

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.
3. Em números com valor absoluto $\in (0, 1)$, zeros antecedendo outros algarismos não são a.s.
   0,00056 (2 a.s.); -0,000009 (1 a.s.).
4. Algarismos 0 entre algarismos de 1 a 9 são significativos.

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.
3. Em números com valor absoluto $\in (0, 1)$, zeros antecedendo outros algarismos não são a.s.
   0,00056 (2 a.s.); -0,000009 (1 a.s.).
4. Algarismos 0 entre algarismos de 1 a 9 são significativos.
   207 (3 a.s.); 107,46 (5 a.s.); 0,08009 (4 a.s.).

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.
3. Em números com valor absoluto $\in (0, 1)$, zeros antecedendo outros algarismos não são a.s.
   0,00056 (2 a.s.); -0,000009 (1 a.s.).
4. Algarismos 0 entre algarismos de 1 a 9 são significativos.
   207 (3 a.s.); 107,46 (5 a.s.); 0,08009 (4 a.s.).
5. Algarismos 0 finais em números fracionários são significativos.

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.
3. Em números com valor absoluto $\in (0, 1)$, zeros antecedendo outros algarismos não são a.s.
   0,00056 (2 a.s.); -0,000009 (1 a.s.).
4. Algarismos 0 entre algarismos de 1 a 9 são significativos.
   207 (3 a.s.); 107,46 (5 a.s.); 0,08009 (4 a.s.).
5. Algarismos 0 finais em números fracionários são significativos.
   3,0 (2 a.s.); -45,9000 (6 a.s.); -45,900 (5 a.s.); 10,00 (4 a.s.); 0,00210 (3 a.s.).

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.
3. Em números com valor absoluto $\in (0, 1)$, zeros antecedendo outros algarismos não são a.s.
   0,00056 (2 a.s.); -0,000009 (1 a.s.).
4. Algarismos 0 entre algarismos de 1 a 9 são significativos.
   207 (3 a.s.); 107,46 (5 a.s.); 0,08009 (4 a.s.).
5. Algarismos 0 finais em números fracionários são significativos.
   3,0 (2 a.s.); -45,9000 (6 a.s.); -45,900 (5 a.s.); 10,00 (4 a.s.); 0,00210 (3 a.s.).
6. Algarismos 0 finais em múltiplos de potências de 10 são ambíguos.

Regras sobre algarismos significativos (a. s.)

1. Os algarismos 1, 2, ..., 9 sempre são significativos.
2. A posição da vírgula não altera o número de a.s.
3. Em números com valor absoluto $\in (0, 1)$, zeros antecedendo outros algarismos não são a.s.
   0,00056 (2 a.s.); -0,000009 (1 a.s.).
4. Algarismos 0 entre algarismos de 1 a 9 são significativos.
   207 (3 a.s.); 107,46 (5 a.s.); 0,08009 (4 a.s.).
5. Algarismos 0 finais em números fracionários são significativos.
   3,0 (2 a.s.); -45,9000 (6 a.s.); -45,900 (5 a.s.); 10,00 (4 a.s.); 0,00210 (3 a.s.).
6. Algarismos 0 finais em múltiplos de potências de 10 são ambíguos.
   14200 (pelo menos 3 a.s.); 30 (pelo menos 1 a.s.); -19000 (pelo menos 2 a.s.).

Regras sobre algarismos significativos

Notação científica: $m \times 10^k$.

m é a **mantissa** (um número real), sendo que $1 \leq |m| < 10$,

k é a **ordem de grandeza** (um número inteiro, podendo ser < 0).

$2,00 \times 10^2$ (3 a.s.)

4×10^3 (1 a.s.)

$4,00 \times 10^3$ (3 a.s.)

$8,009 \times 10^{-2}$ (4 a.s.)

Operações: algarismos significativos

- **Soma e subtração**

  O número de casas decimais do resultado é determinado pelo operando com menor número de casas decimais, seguindo as regras de arredondamento.

  3,141593 + 2,5 = 5,6 (2 a.s.).

  2,718 – 0,800 + 20 = 22 (2 a.s.).

  2,718 – 0,80 + 20,0 = 21,9 (3 a.s.).

Operações: algarismos significativos

- **Soma e subtração**

  O número de casas decimais do resultado é determinado pelo operando com menor número de casas decimais, seguindo as regras de arredondamento.

  3,141593 + 2,5 = 5,6 (2 a.s.).

  2,718 – 0,800 + 20 = 22 (2 a.s.).

  2,718 – 0,80 + 20,0 = 21,9 (3 a.s.).

- **Multiplicação, divisão, funções trigonométricas, logaritmos**:

  O número de algarismos significativos do resultado é determinado pelo operando com menor número de algarismos significativos.

  log(0,031 x 8,15) = -1,4 (2 a.s.)

  12,74 / 3,31 = 3,85 (3 a.s.)

  $e^2 = 7.$ (1 a.s.)

  $e^2, 000 = 7{,}389$ (4 a.s.)

Arredondamento

Necessidade de arredondamento

- 1 / 310 = 0,003226 com quatro algarismos significativos.
- 41 / 7 = 5,86 com três algarismos significativos.
- 2,999... = 3 com um algarismo significativo.
- 2,999... = 3,0 com dois algarismos significativos.