CIÊNCIA DE DADOS COM LINGUAGEM R

Richard Guilherme dos Santos

Contents

Chapter 1

Introdução

Este livro tem como objetivo servir como guia para as aulas do curso Ciência de Dados com R. Nele apresentaremos os conceitos de:

1. **Estatística Básica:** Nesta parte do curso abordaremos conceitos de estatística como variáveis, tipos de distribuições discretas e contínuas, medidas descritivas e distribuição normal.
2. **Manipulação de dados no R:** Neste tópico serão abordados as principais formas de manipulação de dados utilizando a linguagem R, com ênfase nas bibliotecas dplyr e tidyr. Além disso, abordaremos a criação de gráficos pelo pacote ggplot2.
3. **Modelos de Regressão Linear:** Parte final do curso, onde o aluno aprenderá sobre diagrama de dispersão, coeficiente de correlação linear, regressão linear simples, múltipla e regressão logística, ganhando a capacidade de começar a criar modelos utilizando a linguagem R.

Chapter 2

Introdução a Probabilidade

Chapter 3

Introdução ao R

Aqui introduziremos alguns comandos da linguagem R. A linguagem utiliza de funções para realizar operações que vão desde leitura e manipulação de dados a operações matemáticas.

Comecemos criando um vetor de números:

```
x <- c(1,3,2,5)
# x = c(1,3,2,5) # Também podemos utilizar "=" para atribuir variáveis
x
```

```
## [1] 1 3 2 5
```

O comando acima combina os números 1,3,2 e 5 em um vetor de números e os salva em um objeto denominado x. Escrevemos x para recebermos os atributos do vetor.

A partir disto podemos utilizar outras funções para calcularmos informações destes atributos, como o tamanho de um vetor:

```
length(x)
```

```
## [1] 4
```

ou sua média:

```
mean(x)
```

```
## [1] 2.75
```

Há outros tipos de objetos que podem ser criados quando trabalhamos com R. Os mais importantes para manipulação de dados são as matrizes:

```
mat = matrix(data = c(1,2,3,4), nrow = 2, ncol = 2,
             byrow = TRUE)
mat
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

Funções aceitam os mais diversos tipos de argumentos, para termos uma ideia de quais utilizarmos e seus atributos devemos consultar na biblioteca do R:

```
help(matrix)
```

E os data.frames, tabelas que aceitam dados de diversos tipos:

```
nomes = c('Carol', 'Alfredo', 'Godoberto')
idade = c(18, 23, 19)
peso = c(69, 75, 80)
altura = c(1.70, 1.80, 1.85)
ICM = peso/altura^2
df = data.frame(nomes, idade, peso, altura, ICM)
df
```

```
##       nomes idade peso altura      ICM
## 1     Carol    18   69   1.70 23.87543
## 2   Alfredo    23   75   1.80 23.14815
## 3 Godoberto    19   80   1.85 23.37473
```

Chapter 4

Medidas Descritivas

4.1 Tipos de Variáveis

Antes de analisarmos conjuntos de dados propriamente, é necessário termos um conhecimento sobre tipos de variáveis. Para isto, consideremos a seguinte tabela:

```
nome = c('Guilherme', 'Leon', 'Nilce')
est_civil = c('Solteiro', 'Casado', 'Casado')
escolaridade = c('Ensino médio completo',
                 'Pós-graduação',
                 'Superior completo')
n_filhos = c(1, 0, 0)
salario = c(1500, 3000, 3000)
idade = c(21, 39, 32)
df = data.frame(nome, est_civil, escolaridade, n_filhos, salario, idade)
kable(df, align = 'c') # Melhor visualização dos dados para este PDF
```

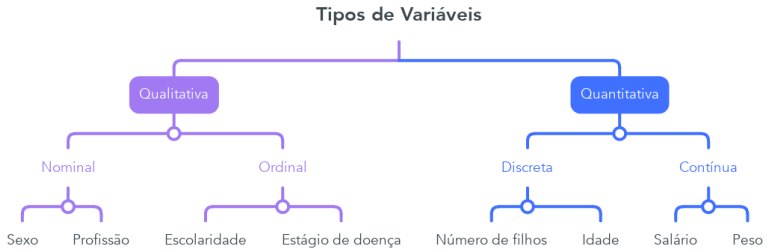
nome	est_civil	escolaridade	n_filhos	salario	idade
Guilherme	Solteiro	Ensino médio completo	1	1500	21
Leon	Casado	Pós-graduação	0	3000	39
Nilce	Casado	Superior completo	0	3000	32

Variáveis como sexo, escolaridade e estado civil apresentam realizações de uma qualidade ou atributo do indivíduo pesquisado, enquanto outras como número de filhos, salário e idade apresentam números como resultados de uma contagem ou mensuração. Chamamos as do primeiro tipo de **qualitativas** e as do segundo de **quantitativas**

Cada uma das duas ainda pode ser dividida em dois tipos:

- **Variável qualitativa nominal:** atributos não apresentam uma ordem lógica;

- **Variável qualitativa ordinal:** atributos apresentam uma ordem lógica bem estabelecida;
- **Variável quantitativa discreta:** dados de contagem, assumem apenas valores inteiros;
- **Variável quantitativa contínua:** dados que podem assumir qualquer tipo de valor.



Muitas vezes queremos resumir estes dados, apresentando um ou mais valores que sejam representativos da série toda. Neste contexto entram às **medidas de posição e dispersão**.

4.2 Medidas de Posição

Usualmente utilizamos uma das seguintes medidas de posição (ou localização): **média, mediana ou moda**. Vamos as suas definições:

- **Moda:** valor mais frequente do conjunto de valores observados.
- **Mediana:** valor que ocupa a posição central das observações quando estas estão ordenadas em ordem crescente.
  - Quando o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais.
- **Média:** soma de todos os elementos do conjunto dividida pela quantidade de elementos do conjunto

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

4.3 Medidas de Dispersão

O resumo de um conjunto de dados por uma única medida representativa de posição esconde toda a informação sobre a variabilidade de um conjunto de

observações. Consideremos que cinco alunos realizaram cinco provas, obtendo as seguintes notas:

```
nomes = c('alunoA', 'alunoB', 'alunoC',
          'alunoD', 'alunoE')
alunoA = c(3,4,5,6,7)
alunoB = c(1,3,5,7,9)
alunoC = c(5,5,5,5,5)
alunoD = c(3,5,5,5,7)
alunoE = c(3,5,5,6,6)
df = data.frame(alunoA, alunoB, alunoC, alunoD, alunoE)
row.names(df) = nomes
df
```

```
##        alunoA alunoB alunoC alunoD alunoE
## alunoA      3      1      5      3      3
## alunoB      4      3      5      5      5
## alunoC      5      5      5      5      5
## alunoD      6      7      5      5      6
## alunoE      7      9      5      7      6
```

4.4 Quantis Empíricos

4.5 Box Plot

4.6 Transformações

$$y = x^2$$

Chapter 5

Tipos de Distribuições Discretas

Chapter 6

Tipos de Distribuições Contínuas

Chapter 7

Introdução as bibliotecas do R

7.1 Dplyr

7.2 Tidyr

7.3 GGPlot2

Chapter 8

Regressão Linear