

*Jeidsan A. da C. Pereira*

---

# ***Estatística e Ciência de Dados***

***Notas e solução dos exercícios***



---

## Conteúdo

---

<b>Prefácio</b>	<b>ix</b>
<b>Prefácio</b>	<b>ix</b>
Pendências . . . . .	ix
<b>1 Estatística, Ciência de Dados e Megadados</b>	<b>1</b>
1.1 Introdução . . . . .	1
1.2 Aprendizado com estatística . . . . .	1
1.3 Aprendizado automático . . . . .	2
1.4 Uma cronologia do desenvolvimento da estatística . . . . .	3
1.5 Notação e tipos de dados . . . . .	3
1.6 Paradigmas para o aprendizado com estatística . . . . .	3
1.7 Este livro . . . . .	4
1.8 Conjuntos de dados . . . . .	4
1.9 Notas do capítulo . . . . .	4
<b>I Análise Exploratória de Dados</b>	<b>5</b>
<b>2 Preparação dos dados</b>	<b>7</b>
2.1 Considerações preliminares . . . . .	7
2.2 Planilhas de dados . . . . .	7
2.3 Construção de tabelas . . . . .	7
2.4 Construção de gráficos . . . . .	8
2.5 Notas de capítulo . . . . .	8
2.6 Exercícios . . . . .	8

<b>3</b>	<b>Análise de dados de uma variável</b>	<b>11</b>
3.1	Introdução . . . . .	13
3.2	Distribuição de frequências . . . . .	13
3.3	Medidas resumo . . . . .	13
3.4	<i>Boxplots</i> . . . . .	13
3.5	Modelos probabilísticos . . . . .	13
3.6	Dados amostrais . . . . .	13
3.7	Gráficos QQ . . . . .	13
3.8	Desvio padrão e erro padrão . . . . .	13
3.9	Intervalo de confiança e tamanho da amostra . . . . .	13
3.10	Transformação de variáveis . . . . .	13
3.11	Notas de capítulo . . . . .	13
3.12	Exercícios . . . . .	13
<b>4</b>	<b>Análise de dados de duas variáveis</b>	<b>15</b>
4.1	Introdução . . . . .	15
4.2	Duas variáveis qualitativas . . . . .	15
4.3	Duas variáveis quantitativas . . . . .	15
4.4	Uma variável qualitativa e outra quantitativa . . . . .	15
4.5	Notas de capítulo . . . . .	15
4.6	Exercícios . . . . .	15
<b>5</b>	<b>Análise de dados de várias variáveis</b>	<b>17</b>
5.1	Introdução . . . . .	17
5.2	Gráficos para três variáveis . . . . .	17
5.3	Gráficos para quatro ou mais variáveis . . . . .	17
5.4	Medidas resumo multivariadas . . . . .	17
5.5	Tabelas de contingência de múltiplas entradas . . . . .	17
5.6	Notas de capítulo . . . . .	17
5.7	Exercícios . . . . .	17

<b>6</b>	<b>Análise de Regressão</b>	<b>19</b>
6.1	Introdução . . . . .	19
6.2	Regressão linear simples . . . . .	19
6.3	Regressão linear múltipla . . . . .	19
6.4	Regressão para dados longitudinais . . . . .	19
6.5	Regressão logística . . . . .	19
6.6	Notas de capítulo . . . . .	19
6.7	Exercícios . . . . .	19
<b>7</b>	<b>Análise de Sobrevivência</b>	<b>21</b>
7.1	Introdução . . . . .	21
7.2	Estimação da função de sobrevivência . . . . .	21
7.3	Comparação de curvas de sobrevivência . . . . .	21
7.4	Regressão para dados de sobrevivência . . . . .	21
7.5	Notas de capítulo . . . . .	21
7.6	Exercícios . . . . .	21
<b>II</b>	<b>Aprendizado Supervisionado</b>	<b>23</b>
<b>8</b>	<b>Regularização e Modelos Aditivos Generalizados</b>	<b>25</b>
8.1	Introdução . . . . .	25
8.2	Regularização . . . . .	25
8.3	Modelos aditivos generalizados (GAM) . . . . .	25
8.4	Notas de capítulo . . . . .	25
8.5	Exercícios . . . . .	25
<b>9</b>	<b>Classificação por meio de técnicas clássicas</b>	<b>27</b>
9.1	Introdução . . . . .	27
9.2	Classificação por regressão logística . . . . .	27
9.3	Análise discriminante linear . . . . .	27
9.4	Classificador do vizinho mais próximo . . . . .	27
9.5	Algumas extensões . . . . .	27
9.6	Notas de capítulo . . . . .	27
9.7	Exercícios . . . . .	27

<b>10 Algoritmos de Suporte Vetorial</b>	<b>29</b>
10.1 Introdução . . . . .	29
10.2 Fundamentação dos algoritmos de suporte vetorial . . . . .	29
10.3 Classificador de margem máxima . . . . .	29
10.4 Classificador de margem flexível . . . . .	29
10.5 Classificador de margem não linear . . . . .	29
10.6 Regressão por algoritmos de suporte vetorial . . . . .	29
10.7 Notas de capítulo . . . . .	29
10.8 Exercícios . . . . .	29
<b>11 Árvores e Florestas</b>	<b>31</b>
11.1 Introdução . . . . .	31
11.2 Classificação por árvores . . . . .	31
11.3 <i>Bagging, boosting</i> e florestas . . . . .	31
11.4 Árvores para regressão . . . . .	31
11.5 Notas de capítulo . . . . .	31
11.6 Exercícios . . . . .	31
<b>12 Redes neurais</b>	<b>33</b>
12.1 Introdução . . . . .	33
12.2 <i>Perceptron</i> . . . . .	33
12.3 Redes com camadas ocultas . . . . .	33
12.4 O algoritmo de retropropagação ( <i>backpropagation</i> ) . . . . .	33
12.5 Aprendizado profundo ( <i>Deep learning</i> ) . . . . .	33
12.6 Notas de capítulo . . . . .	33
12.7 Exercícios . . . . .	33
<b>III Aprendizado não Supervisionado</b>	<b>35</b>

<i>Contents</i>	vii
<b>13 Análise de Agrupamentos</b>	<b>37</b>
13.1 Introdução . . . . .	37
13.2 Estratégias de agrupamento . . . . .	37
13.3 Algoritmos hierárquicos . . . . .	37
13.4 Algoritmos de partição: K-médias . . . . .	37
13.5 Notas de capítulo . . . . .	37
13.6 Exercícios . . . . .	37
<b>14 Redução de dimensionalidade</b>	<b>39</b>
14.1 Introdução . . . . .	39
14.2 Análise de Componentes Principais . . . . .	39
14.3 Análise fatorial . . . . .	39
14.4 Análise de componentes independentes . . . . .	39
14.5 Notas de capítulo . . . . .	39
14.6 Exercícios . . . . .	39
<b>Apêndice</b>	<b>39</b>
<b>A Otimização numérica</b>	<b>41</b>
A.1 Introdução . . . . .	41
A.2 O método de Newton-Raphson . . . . .	41
A.3 O método scoring . . . . .	41
A.4 O método de Gauss-Newton . . . . .	41
A.5 Métodos Quase-Newton . . . . .	41
A.6 Aspectos computacionais . . . . .	41
A.7 Notas de capítulo . . . . .	41
A.8 Exercícios . . . . .	41
<b>B Noções de simulação</b>	<b>43</b>
B.1 Introdução . . . . .	43
B.2 Método Monte Carlo . . . . .	43
B.3 Simulação de variáveis discretas . . . . .	43

B.4	Simulação de variáveis contínuas . . . . .	43
B.5	Simulação de vetores aleatórios . . . . .	43
B.6	Métodos de reamostragem . . . . .	43
B.7	Notas de capítulo . . . . .	43
B.8	Exercícios . . . . .	43
<b>C</b>	<b>Algoritmos para dados aumentados</b>	<b>45</b>
C.1	Introdução . . . . .	45
C.2	O algoritmo EM . . . . .	45
C.3	O algoritmo EM Monte Carlo . . . . .	45
C.4	Cálculo de erros padrões . . . . .	45
C.5	O algoritmo para dados aumentados . . . . .	45
C.6	Exercícios . . . . .	45



---

## ***Prefácio***

---

Esta página contém notas e solução para os exercícios propostos no livro **Estatística e Ciência de Dados**, de autoria de Pedro Alberto Morettin e Júlio da Motta Singer, publicado pela LTC em 2022 [Morettin and Singer, 2022].

É importante destacar que trata-se de um produto não oficial, as anotações e soluções de exercícios aqui apresentadas são de cunho pessoal e não possuem qualquer revisão ou análise por parte dos autores da obra ou da editora. Dessa forma e por se tratar de um produto construído durante o processo de aprendizagem, o conteúdo pode conter erros, tanto no texto em si, como na lógica utilizada para solução dos exercícios.

Dúvidas ou sugestões de melhoria podem ser encaminhadas para o e-mail *jeidsan.pereira@gmail.com*<sup>1</sup>.

---

## **Pendências**

•

---

<sup>1</sup>mailto:jeidsan.pereira@gmail.com



# 1

---

## *Estatística, Ciência de Dados e Megadados*

---

### 1.1 Introdução

---

Atualmente, os termos *Data Science* (**Ciência de Dados**) e *Big Data* (**Megadados**) são utilizados em profusão como se envolvessem conceitos novos, distintos daqueles com que os estatísticos lidam há cerca de dois séculos [Morettin and Singer, 2022, p. 1].

---

### 1.2 Aprendizado com estatística

---

O **aprendizado supervisionado** está relacionado com metodologia desenvolvida essencialmente para **previsão** e **classificação**. No âmbito da previsão, o objetivo é utilizar **variáveis preditivas** (sexo, classe social, renda, por exemplo) observadas em várias **unidades** (clientes de um banco, por exemplo) para “advinhar” valores de uma **variável resposta numérica** (saldo médio, por exemplo) de novas unidades. O problema de classificação consiste em qual categoria de uma **variável resposta qualitativa** (bons e maus pagadores, por exemplo) as novas unidades são classificadas [Morettin and Singer, 2022, p. 3].

---

---

No **aprendizado não supervisionado**, dispomos apenas um conjunto de dados, sem distinção entre preditoras e respostas, e o objetivo é descrever **associações** e **padrões** entre essas variáveis e **agrupá-las** com o objetivo de identificar características comuns e conjuntos de unidades de investigação ou desenvolver métodos para combiná-las e assim **reduzir sua dimensionalidade** [Morettin and Singer, 2022, p. 3].

---

---

Além de aprendizado supervisionado e não supervisionado, podemos acrescentar um terceiro tipo, denominado **aprendizado com reforço** (*reinforcement learning*), segundo o qual um algoritmo “aprende” a realizar determinadas tarefas por meio de repetições com o fim de maximizar um prêmio sujeito a um valor máximo [Morettin and Singer, 2022, p. 3].

---

---

Embora tanto o aprendizado supervisionado quanto o aprendizado com reforço utilizem um mapeamento entre entradas (*inputs*) e saídas (*outputs*), no primeiro caso a retroalimentação (*feedback*) fornecida ao algoritmo é um conjunto de ações corretas necessárias para a realização de uma tarefa; no aprendizado com reforço, por outro lado, a retroalimentação é baseada num sistema com prêmios e punições como indicativos de ações corretas ou incorretas [Morettin and Singer, 2022, p. 3].

---

---

### 1.3 Aprendizado automático

Jordan [2019 *apud* Morettin and Singer, 2022, p. 4] distingue três tipos de inteligência artificial: i) inteligência artificial imitativa humana; ii) aumento de inteligência; e iii) infraestrutura inteligente.

---

De modo informal, a inteligência artificial está relacionada com um esforço para automatizar tarefas intelectuais usualmente realizadas por seres humanos (Chollet, 2018) e consequentemente, intimamente ligada ao desenvolvimento da computação (ou programação de computadores) [Morettin and Singer, 2022, p. 4].

---

---

Convém ressaltar que o objetivo do aprendizado automático não é o mesmo daquele considerado na análise de regressão usual, em que se pretende entender como cada variável preditora  $X_0$  está associada com a variável resposta. O objetivo do aprendizado automático é selecionar o modelo que produz melhores previsões, mesmo que as variáveis selecionadas com essa finalidade não sejam aquelas consideradas numa análise padrão [Morettin and Singer, 2022, p. 5].

---

---

## 1.4 Uma cronologia do desenvolvimento da estatística

Sem notas para esta seção.

---

## 1.5 Notação e tipos de dados

Sem notas para esta seção.

---

## 1.6 Paradigmas para o aprendizado com estatística

Sem notas para esta seção.

---

---

## 1.7 Este livro

---

Independentemente do volume de dados disponíveis para análise, Ciência de Dados é uma atividade multidisciplinar que envolve: i) um problema a ser resolvido com questões claramente especificadas; ii) um conjunto de dados (seja ele volumoso ou não); iii) os meios para sua obtenção; iv) sua organização; v) a especificação do problema original em termos das variáveis desse conjunto de dados; vi) a descrição e resumo dos dados à luz do problema a ser resolvido; vii) a escolha das técnicas estatísticas apropriadas para a resolução desse problema; viii) os algoritmos computacionais necessários para a implementação dessas técnicas; ix) a apresentação dos resultados [Moret-tin and Singer, 2022, p. 11].

---

---

## 1.8 Conjuntos de dados

Sem notas para esta seção.

---

## 1.9 Notas do capítulo

Sem notas para esta seção.

## **Parte I**

# **Análise Exploratória de Dados**





## 2

---

### *Preparação dos dados*

---

#### 2.1 Considerações preliminares

---

O ramo da Estatística conhecido como **Análise Exploratória de Dados** se ocupa da organização e resumo dos dados de uma amostra ou, eventualmente, de toda a população e o ramo conhecido como **Inferência Estatística** se refere ao processo de se tirar conclusões sobre uma população com base em uma amostra dela [Morettin and Singer, 2022, p. 21].

---

#### 2.2 Planilhas de dados

Sem notas para esta seção.

---

#### 2.3 Contrução de tabelas

Sem notas para esta seção.

---

## 2.4 Construção de gráficos

Sem notas para esta seção.

---

## 2.5 Notas de capítulo

Sem notas para esta seção.

---

## 2.6 Exercícios

### Exercício 2.1

O objetivo de um estudo da Faculdade de Medicina da USP foi avaliar a associação entre a quantidade de morfina administrada a pacientes com dores intensas provenientes de lesões medulares ou radiculares e a dosagem dessa substância em seus cabelos. Três medidas foram realizadas em cada paciente, a primeira logo após o início do tratamento e as demais após 30 e 60 dias. Detalhes podem ser obtidos no documento disponível no arquivo morfina.doc.

A planilha morfina.xls, disponível no arquivo morfina foi entregue ao estatístico para análise e contém resumos de características demográficas além dos dados do estudo.

- Com base nessa planilha, apresente um dicionário com a especificação das variáveis segundo as indicações da Seção 2.2 e construa a planilha correspondente.
- Com as informações disponíveis, construa tabelas para as variáveis sexo, raça, grau de instrução e tipo de lesão segundo as sugestões da Seção 2.3.

*Solução.* x

### Exercício 2.2

A Figura 2.6 foi extraída de um estudo sobre atitudes de profissionais de saúde com relação a cuidados com infecção hospitalar. Critique-a e reformule-a para facilitar

sua leitura, lembrando que a comparação de maior interesse é entre as diferentes categorias profissionais.

*Solução.* x

### Exercício 2.3

Utilize as sugestões para construção de planilhas apresentadas na Seção 2.2 com a finalidade de preparar os dados do arquivo empresa para análise estatística.

*Solução.* x

### Exercício 2.4

Num estudo planejado para avaliar o consumo médio de combustível de veículos em diferentes velocidades foram utilizados 4 automóveis da marca A e 3 automóveis da marca B selecionados ao acaso das respectivas linhas de produção. O consumo (em L/km) de cada um dos 7 automóveis foi observado em 3 velocidades diferentes (40 km/h, 80 km/h e 110 km/h) Delineie uma planilha apropriada para a coleta e análise estatística dos dados, rotulando-a adequadamente.

*Solução.* x

### Exercício 2.5

Utilizando os dados do arquivo enforco, prepare uma planilha Excel num formato conveniente para análise pelo R. Inclua apenas as variáveis Idade, Altura, Peso, Frequência cardíaca e VO<sub>2</sub> no repouso além do quociente VE/VCO<sub>2</sub>, as correspondentes porcentagens relativamente ao máximo, o quociente VO<sub>2</sub>/FC no pico do exercício e data do óbito. Importe a planilha Excel que você criou utilizando comandos R e obtenha as características do arquivo importado (número de casos, número de observações omissas etc.)

*Solução.* x

### Exercício 2.6

A Figura 2.7 contém uma planilha encaminhada pelos investigadores responsáveis por um estudo sobre AIDS para análise estatística. Organize-a de forma a permitir sua análise por meio de um pacote computacional como o R.

*Solução.* x

**Exercício 2.7**

A planilha apresentada na Figura 2.8 contém dados de um estudo em que o limiar auditivo foi avaliado nas orelhas direita (OD) e esquerda (OE) de 13 pacientes em 3 ocasiões (Limiar, Teste 1 e Teste 2). Reformate-a segundo as recomendações da Seção 2.2, indicando claramente

- a) a definição das variáveis.
- b) os rótulos para as colunas da planilha.

*Solução.* x

**Exercício 2.8**

A planilha disponível no arquivo *idades* contém informações demográficas de 35548 municípios brasileiros.

- a) Importe-a para permitir a análise por meio do software R, indicando os problemas encontrados nesse processo além de sua solução.
- b) Use o comando `summary` para obter um resumo das variáveis do arquivo.
- c) Classifique cada variável como numérica ou alfanumérica e indique o número de observações omissas de cada uma delas.

*Solução.* x

**Exercício 2.9**

Preencha a ficha de inscrição do Centro de Estatística Aplicada ([www.ime.usp.br/-cea](http://www.ime.usp.br/-cea)) com as informações de um estudo em que você está envolvido.

*Solução.* x

# 3

## *Análise de dados de uma variável*



---

### 3.1 Introdução

---

### 3.2 Distribuição de frequências

---

### 3.3 Medidas resumo

---

### 3.4 *Boxplots*

---

### 3.5 Modelos probabilísticos

---

### 3.6 Dados amostrais

---

### 3.7 Gráficos QQ

---

### 3.8 Desvio padrão e erro padrão

---

### 3.9 Intervalo de confiança e tamanho da amostra

---

### 3.10 Transformação de variáveis

---

### 3.11 Notas de capítulo

---

### 3.12 Exercícios





# 4

---

## *Análise de dados de duas variáveis*

---

---

### 4.1 Introdução

---

### 4.2 Duas variáveis qualitativas

---

### 4.3 Duas variáveis quantitativas

---

### 4.4 Uma variável qualitativa e outra quantitativa

---

### 4.5 Notas de capítulo

---

### 4.6 Exercícios



# 5

---

## *Análise de dados de várias variáveis*

---

---

### 5.1 Introdução

---

### 5.2 Gráficos para três variáveis

---

### 5.3 Gráficos para quatro ou mais variáveis

---

### 5.4 Medidas resumo multivariadas

---

### 5.5 Tabelas de contingência de múltiplas entradas

---

### 5.6 Notas de capítulo

---

### 5.7 Exercícios



# 6

## *Análise de Regressão*

### 6.1 Introdução

### 6.2 Regressão linear simples

### 6.3 Regressão linear múltipla

### 6.4 Regressão para dados longitudinais

### 6.5 Regressão logística

### 6.6 Notas de capítulo

### 6.7 Exercícios



# 7

---

## *Análise de Sobrevivência*

---

---

### 7.1 Introdução

---

### 7.2 Estimação da função de sobrevivência

---

### 7.3 Comparação de curvas de sobrevivência

---

### 7.4 Regressão para dados de sobrevivência

---

### 7.5 Notas de capítulo

---

### 7.6 Exercícios





## **Parte II**

# **Aprendizado Supervisionado**



# 8

---

## *Regularização e Modelos Aditivos Generalizados*

---

### 8.1 Introdução

---

### 8.2 Regularização

---

### 8.3 Modelos aditivos generalizados (GAM)

---

### 8.4 Notas de capítulo

---

### 8.5 Exercícios



# 9

---

## *Classificação por meio de técnicas clássicas*

---

### 9.1 Introdução

---

### 9.2 Classificação por regressão logística

---

### 9.3 Análise discriminante linear

---

### 9.4 Classificador do vizinho mais próximo

---

### 9.5 Algumas extensões

---

### 9.6 Notas de capítulo

---

### 9.7 Exercícios



# 10

---

## *Algoritmos de Suporte Vetorial*

---

---

### 10.1 Introdução

---

### 10.2 Fundamentação dos algoritmos de suporte vetorial

---

### 10.3 Classificador de margem máxima

---

### 10.4 Classificador de margem flexível

---

### 10.5 Classificador de margem não linear

---

### 10.6 Regressão por algoritmos de suporte vetorial

---

### 10.7 Notas de capítulo

---

### 10.8 Exercícios

---





# 11

---

## *Árvores e Florestas*

---

### 11.1 Introdução

---

### 11.2 Classificação por árvores

---

### 11.3 *Bagging, boosting* e florestas

---

### 11.4 Árvores para regressão

---

### 11.5 Notas de capítulo

---

### 11.6 Exercícios



# 12

---

## *Redes neurais*

---

---

### 12.1 Introdução

---

### 12.2 *Perceptron*

---

### 12.3 Redes com camadas ocultas

---

### 12.4 O algoritmo de retropropagação (*backpropagation*)

---

### 12.5 Aprendizado profundo (*Deep learning*)

---

### 12.6 Notas de capítulo

---

### 12.7 Exercícios



## **Parte III**

# **Aprendizado não Supervisionado**



# 13

---

## *Análise de Agrupamentos*

---

---

### 13.1 Introdução

---

### 13.2 Estratégias de agrupamento

---

### 13.3 Algoritmos hierárquicos

---

### 13.4 Algoritmos de partição: K-médias

---

### 13.5 Notas de capítulo

---

### 13.6 Exercícios





# 14

---

## *Redução de dimensionalidade*

---

---

### 14.1 Introdução

---

### 14.2 Análise de Componentes Principais

---

### 14.3 Análise fatorial

---

### 14.4 Análise de componentes independentes

---

### 14.5 Notas de capítulo

---

### 14.6 Exercícios



# A

## *Otimização numérica*

### A.1 Introdução

### A.2 O método de Newton-Raphson

### A.3 O método scoring

### A.4 O método de Gauss-Newton

### A.5 Métodos Quase-Newton

### A.6 Aspectos computacionais

### A.7 Notas de capítulo

### A.8 Exercícios



# B

---

## *Noções de simulação*

---

### B.1 Introdução

---

### B.2 Método Monte Carlo

---

### B.3 Simulação de variáveis discretas

---

### B.4 Simulação de variáveis contínuas

---

### B.5 Simulação de vetores aleatórios

---

### B.6 Métodos de reamostragem

---

### B.7 Notas de capítulo

---

### B.8 Exercícios



# C

---

## *Algoritmos para dados aumentados*

---

### C.1 Introdução

---

### C.2 O algoritmo EM

---

### C.3 O algoritmo EM Monte Carlo

---

### C.4 Cálculo de erros padrões

---

### C.5 O algoritmo para dados aumentados

---

### C.6 Exercícios





---

## ***Bibliografia***

---

Pedro Alberto Morettin and Julio da Motta Singer. *Estatística e Ciência de Dados*. LTC, Rio de Janeiro, 2022.