Segunda Edição

Estatística Prática para Cientistas de Dados

+50 Conceitos Essenciais Usando R e Python

Segunda Edição

Estatística Prática para Cientistas de Dados

+50 Conceitos Essenciais Usando R e Python

Peter Bruce, Andrew Bruce e Peter Gedeck

ALTA BOOKS
EDITORA
Rio de Janeiro, 2025

Estatística Prática para Cientistas de Dados - 2ª Edição

ISBN: 978-85-508-2651-6

Impresso no Brasil — 2ª Edição, 2025 — Edição revisada conforme o Acordo Ortográfico da Língua Portuguesa de 2009.

Dados Internacionais de Catalogação na Publicação (CIP)

B889e
G293e
2.ed. Bruce, Peter; Bruce, Andrew; Gedeck, Peter
Estatística Prática para Cientistas de Dados: +50 Conceitos Essenciais Usando R e Python/
Peter Bruce, Andrew Bruce, Peter Gedeck; tradução Eveline Machado. – 2.ed. –Rio de Janeiro: Alta Books, 2025.
352 p.; il.; 15,7 x 23 cm.

Título original: Practical statistics for data scientists.
ISBN 978-85-508-2651-6

1. Estatística. 2. Ciência de dados. 3. Aprendizado de máquina. 4. Análise de dados. I. Bruce, Andrew. II. Gedeck, Peter. III. Título.
CDD 519.5

Índice para catálogo sistemático:
1. Estatística : Ciência de dados 519.5

Grupo Editorial Alta Books

Diretor Editorial: Anderson Vieira.
Vendas ao Governo: Cristiane Mutüs.
Gerência Comercial: Claudio Lima.
Gerência Marketing: Viviane Paiva.

Produtora Editorial: Isabella Gibara.
Tradução e Copidesque: Eveline Machado.
Revisão Gramatical: Denise Himpel.
Diagramação: Natalia Curupana.

ALTA BOOKS
GRUPO EDITORIAL

Rua Viúva Cláudio, 291 — Bairro Industrial do Jacaré
CEP: 20.970-031 — Rio de Janeiro (RJ)
Tels.: (21) 3278-8069 / 3278-8419
www.altabooks.com.br — altabooks@altabooks.com.br
Ouvidoria: ouvidoria@altabooks.com.br

Editora **afiliada à:**

Peter Bruce e Andrew Bruce gostariam de dedicar este livro à memória dos pais Victor G. Bruce e Nancy C. Bruce, que nutriam uma grande paixão por matemática e ciência; aos nossos primeiros mentores John W. Tukey e Julian Simon, e ao querido amigo de longa data Geoff Watson, que nos inspirou a seguir carreira em estatística.

Peter Gedeck gostaria de dedicar este livro a Tim Clark e Christian Kramer, com agradecimentos profundos por sua colaboração científica e amizade.

Sumário

Prefácio xv

1. Análise Exploratória de Dados 1

Elementos dos Dados Estruturados 2

Leitura Adicional 4

Dados Retangulares 5

Data Frames e Índices 6

Estruturas de Dados Não Retangulares 7

Leitura Adicional 8

Estimativas de Localização 8

Média 9

Mediana e Estimativas Robustas 11

Exemplo: Estimativas de Localização de População e Taxas de Homicídio 12

Leitura Adicional 14

Estimativas de Variabilidade 14

Desvio-padrão e Estimativas Relacionadas 15

Estimativas Baseadas em Percentis 17

Exemplo: Estimativas de Variabilidade da População Estadual 18

Leitura Adicional 20

Explorando a Distribuição de Dados 20

Percentis e Gráficos de Caixa 20

Tabelas de Frequência e Histogramas 22

Gráficos de Densidade e Estimativas 25

Leitura Adicional 27

Explorando Dados Binários e Categóricos 27
Moda 30
Valor Esperado 30
Probabilidade 30
Leitura Adicional 31
Correlação 31
Gráficos de Dispersão 35
Leitura Adicional 36
Explorando Duas ou Mais Variáveis 36
Compartimentação Hexagonal e Contornos (Representação Numérica versus Dados Numéricos) 37
Duas Variáveis Categóricas 40
Dados Categóricos e Numéricos 41
Visualizando Variáveis Múltiplas 43
Leitura Adicional 46
Resumo 46

2. Distribuições de Dados e Amostras 47
Amostragem Aleatória e Viés de Amostra 48
Viés 50
Seleção Aleatória 51
Tamanho versus Qualidade: Quando o Tamanho Importa? 52
Média Amostral versus Média Populacional 53
Leitura Adicional 53
Viés de Seleção 53
Regressão à Média 55
Leitura Adicional 56
Distribuição de Amostragem de uma Estatística 57
Teorema de Limite Central 60
Erro-padrão 60
Leitura Adicional 61
Bootstrap 61
Reamostragem versus Bootstrap 65
Leitura Adicional 65

Intervalos de Confiança 66
Leitura Adicional 68
Distribuição Normal 69
Normal Padrão e Gráficos QQ 70
Distribuições de Cauda Longa 72
Leitura Adicional 74
Distribuição t de Student 74
Leitura Adicional 77
Distribuição Binomial 77
Leitura Adicional 79
Distribuição Qui-quadrado 79
Leitura Adicional 80
Distribuição F 81
Leitura Adicional 81
Poisson e Distribuições Relacionadas 81
Distribuições Poisson 82
Distribuição Exponencial 83
Estimando a Taxa de Falha 83
Distribuição Weibull 84
Leitura Adicional 85
Resumo 85

3. Experimentos Estatísticos e Teste de Significância..... 87
Teste A/B 88
Por que Ter um Grupo de Controle? 90
Por que Apenas A/B? Por que Não C, D…? 91
Leitura Adicional 92
Teste de Hipótese 93
Hipótese Nula 94
Hipótese Alternativa 95
Teste de Hipótese Unilateral Versus Bilateral 95
Leitura Adicional 96
Reamostragem 96
Teste de Permutação 97

Exemplo: Engajamento na Web 98
Testes de Permutação Exaustiva e Bootstrap 101
Testes de Permutação: A Conclusão para a Ciência de Dados 102
Leitura Adicional 103
Significância Estatística e Valores-p 103
Valor-p 106
Alfa 107
Erros Tipo 1 e Tipo 2 108
Ciência de Dados e Valores-p 109
Leitura Adicional 109
Testes t 109
Leitura Adicional 111
Testagem Múltipla 112
Leitura Adicional 115
Graus de Liberdade 115
Leitura Adicional 117
ANOVA 117
Estatística F 120
ANOVA Bidirecional 122
Leitura Adicional 122
Teste de Qui-quadrado 122
Teste de Qui-quadrado: Uma Abordagem da Reamostragem 123
Teste de Qui-quadrado: Teoria Estatística 125
Teste Exato de Fisher 126
Relevância para a Ciência de Dados 129
Leitura Adicional 130
Algoritmo do Bandido Multibraços 130
Leitura Adicional 133
Potência e Tamanho da Amostra 133
Tamanho da Amostra 135
Leitura Adicional 137
Resumo 137

4. Regressão e Predição 139
Regressão Linear Simples 139

Equação de Regressão 141
Valores Ajustados e Resíduos 143
Mínimos Quadrados 144
Predição versus Explicação (Perfilamento) 145
Leitura Adicional 146
Regressão Linear Múltipla 147
Exemplo: Dados Imobiliários de King County 148
Avaliando o Modelo 149
Validação Cruzada 151
Seleção do Modelo e Regressão Passo a Passo 152
Regressão Ponderada 156
Leitura Adicional 157
Predição Usando Regressão 157
Os Perigos da Extrapolação 158
Intervalos de Confiança e Predição 158
Variáveis Fatoriais na Regressão 160
Representação de Variáveis Fictícias 161
Variáveis Fatoriais com Muitos Níveis 164
Variáveis Fatoriais Ordenadas 166
Interpretando a Equação de Regressão 166
Variáveis Preditoras Correlacionadas 167
Multicolinearidade 169
Variáveis de Confusão 169
Interações e Efeitos Principais 171
Diagnósticos da Regressão 173
Valores Atípicos 174
Valores Influentes 176
Heteroscedasticidade, Não Normalidade e Erros Correlacionados 178
Gráficos Residuais Parciais e Não Linearidade 181
Regressão Polinomial e por Spline 184
Polinomial 184
Splines 186
Modelos Aditivos Generalizados 188
Leitura Adicional 190
Resumo 190

5. Classificação 191
Naive Bayes 192
Por que a Classificação Bayesiana Exata é Inviável 193
A Solução Naive 194
Variáveis Preditoras Numéricas 196
Leitura Adicional 197
Análise Discriminante 197
Matriz de Covariância 198
Discriminante Linear de Fisher 199
Um Exemplo Simples 199
Leitura Adicional 203
Regressão Logística 203
Função de Resposta Logística e Logito 204
Regressão Logística e GLM 206
Modelos Lineares Generalizados 207
Valores Previstos a Partir da Regressão Logística 208
Interpretando os Coeficientes e as Razões de Chances 209
Regressões Linear e Logística: Semelhanças e Diferenças 210
Avaliando o Modelo 212
Leitura Adicional 215
Avaliando Modelos de Classificação 216
Matriz de Confusão 217
O Problema da Classe Rara 219
Exatidão, Revocação e Especificidade 219
Curva ROC 220
AUC 222
Lift 224
Leitura Adicional 225
Estratégias para Dados Desequilibrados 226
Undersampling 226
Oversampling e Ponderação Acima/Abaixo 228
Geração de Dados 229
Classificação Baseada em Custos 230
Explorando as Predições 230

Leitura Adicional 232
Resumo 232

6. Aprendizado de Máquina Estatístico 233
K-vizinhos Mais Próximos 234
Um Pequeno Exemplo: Prevendo Inadimplência em Empréstimos 235
Métricas de Distância 238
Codificação One-hot 239
Padronização (Normalização, Escores Z) 239
Escolhendo K 243
KNN como um Motor de Característica 244
Modelos de Árvore 246
Um Exemplo Simples 247
Algoritmo de Repartição Recursiva 249
Medindo Homogeneidade ou Impureza 251
Fazendo a Árvore Parar de Crescer 252
Prevendo um Valor Contínuo 254
Como as Árvores São Usadas 255
Leitura Adicional 256
Bagging e a Floresta Aleatória 256
Bagging 257
Floresta Aleatória 258
Importância da Variável 262
Hiperparâmetros 265
Boosting 267
Algoritmo de Boosting 268
XGBoost 269
Regularização: Evitando o Sobreajuste 271
Hiperparâmetros e Validação Cruzada 276
Resumo 279

7. Aprendizado Não Supervisionado 281
Análise de Componentes Principais 282
Um Exemplo Simples 283
Calculando os Componentes Principais 286

Interpretando os Componentes Principais 286
Análise de Correspondência 290
Leitura Adicional 292
Clusterização por K-médias 292
Um Exemplo Simples 293
Algoritmo de K-médias 295
Interpretando os Clusters 296
Escolhendo o Número de Clusters 299
Clusterização Hierárquica 301
Um Exemplo Simples 301
Dendrograma 302
Algoritmo Aglomerativo 304
Medidas de Dissimilaridade 305
Clusterização Baseada em Modelos 307
Distribuição Normal Multivariada 307
Misturas de Normais 308
Selecionando o Número de Clusters 311
Leitura Adicional 314
Escalonamento e Variáveis Categóricas 314
Escalonando as Variáveis 315
Variáveis Dominantes 317
Dados Categóricos e Distância de Gower 318
Problemas na Clusterização de Dados Combinados 321
Resumo 323
Bibliografia 325
Índice 327

PREFÁCIO

Este livro se destina a cientistas de dados que já têm alguma familiaridade com as linguagens de programação *R* e/ou *Python*, além de uma exposição prévia (talvez pontual ou efêmera) à estatística. Nós dois viemos do mundo da estatística para o mundo da ciência de dados, então reconhecemos a contribuição que a estatística pode dar para a arte da ciência de dados. Ao mesmo tempo, sabemos bem das limitações do ensino tradicional de estatística: como disciplina, tem um século e meio de idade, e a maioria dos livros e cursos de estatística é repleta do dinamismo e da inércia de um transatlântico. Todos os métodos neste livro têm alguma conexão — histórica ou metodológica — com a disciplina da estatística. Os métodos que evoluíram principalmente a partir da ciência da computação, como as redes neurais, não estão incluídos.
Dois objetivos fundamentam este livro:

- Expor, de forma digerível, navegável e de fácil referência, conceitos-chave da estatística que são relevantes para a ciência de dados.
- Explicar quais conceitos são importantes e úteis, da perspectiva da ciência de dados, quais são menos importantes e o porquê.

Convenções Usadas Neste Livro

As seguintes convenções tipográficas são utilizadas neste livro:
Itálico

Indica termos novos, URLs, endereços de e-mail, nomes e extensões de arquivo.

`Fonte monoespaçada`

Usada para listagens de programas e também no texto referente a elementos dos programas como variáveis ou nomes de funções, bancos de dados, tipos de dados, variáveis de ambiente, declarações e palavras-chave.

`Fonte monoespaçada em negrito`

Mostra comandos ou outro texto que deve ser digitado literalmente pelo usuário.

Termos-chave

A ciência de dados é uma fusão de múltiplas disciplinas, incluindo estatística, ciências da computação, tecnologia da informação e campos de domínio específico. Consequentemente, podem-se utilizar muitos termos diferentes para se referir a um dado conceito. Os termos-chave e seus sinônimos serão destacados no livro em caixas como esta.

Este elemento simboliza uma dica ou uma sugestão.

Este elemento simboliza um comentário geral.

Este elemento simboliza um aviso ou uma advertência.

Utilizando Exemplos de Código

Em todos os casos, este livro dá exemplos de código primeiro em *R* e depois em *Python*. Para evitar repetições desnecessárias, geralmente mostramos apenas a saída e os gráficos criados pelo código em *R*. Também pulamos o código necessário para carregar os pacotes e os conjuntos de dados necessários. Os materiais complementares (exemplos de código, exercícios etc.) estão disponíveis para download em *https://github.com/gedeck/practical-statistics-for-data-scientists* (em inglês) ou no site da Editora Alta Books; busque pelo ISBN do livro.

Este livro existe para ajudá-lo a fazer o seu trabalho. Em geral, se o código de exemplo é oferecido no livro, você pode usá-lo em seus programas e documentação. Não é preciso entrar em contato conosco para ter permissão, a menos que você esteja reproduzindo uma parte significativa do código. Por exemplo, escrever um programa que usa várias partes do código deste livro não requer

permissão. Vender ou distribuir exemplos dos livros não requer permissão. Responder a uma pergunta citando este livro e o código de exemplo não requer permissão. Incorporar uma quantidade significativa do código de exemplo deste livro na documentação do seu produto requer permissão.
Apreciamos, mas não exigimos, a atribuição. Uma atribuição costuma incluir título, autor, editora e ISBN. Por exemplo: "*Estatística Prática para Cientistas de Dados* de Peter Bruce, Andrew Bruce e Peter Gedeck. Copyright 2020 Peter Bruce, Andrew Bruce e Peter Gedeck, 978-85-508-2651-6."
Se você sentir que o uso dos exemplos de código está fora do limite da utilização justa ou da permissão informada antes, entre em contato conosco.

Agradecimentos

Os autores agradecem às muitas pessoas que ajudaram a tornar este livro realidade.
Gerhard Pilcher, CEO da empresa de pesquisa de dados Elder Research, leu os primeiros rascunhos deste livro e fez correções e comentários úteis e detalhados. Da mesma forma, Anya McGuirk e Wei Xiao, estatísticos na SAS, e Jay Hilfiger, autor membro da O'Reilly, forneceram feedbacks úteis sobre os rascunhos iniciais do livro. Toshiaki Kurokawa, que traduziu a primeira edição para o japonês, fez um trabalho abrangente de revisão e correção no processo. Aaron Schumacher e Walter Paczkowski revisaram minuciosamente a segunda edição do livro e deram inúmeras sugestões úteis e valiosas pelas quais somos extremamente gratos. É óbvio que quaisquer erros que permaneceram são nossos apenas.
Na O'Reilly, Shannon Cutt nos conduziu no processo de publicação com bom ânimo e a quantidade certa de estímulos, já Kristen Brown guiou nosso livro suavemente na fase de produção. Rachel Monaghan e Eliahu Sussman corrigiram e melhoraram nossa escrita com cuidado e paciência, enquanto Ellen Troutman-Zaig preparou o índice. Nicole Tache assumiu as rédeas da segunda edição e guiou o processo de forma eficaz, fornecendo muitas boas sugestões editoriais para melhorar a legibilidade do livro para um público amplo. Agradecemos também a Marie Beaugureau, que iniciou nosso projeto na O'Reilly, bem como a Ben Bengfort, autor da O'Reilly e instrutor na Statistics.com, que nos apresentou à O'Reilly.
Nós, e este livro, nos beneficiamos também das muitas conversas que Peter teve ao longo dos anos com Galit Shmueli, coautora em outros livros.
Finalmente, gostaríamos de agradecer especialmente a Elizabeth Bruce e Deborah Donnell, cuja paciência e apoio tornaram esta empreitada possível.

CAPÍTULO 1

Análise Exploratória de Dados

Este capítulo foca o primeiro passo de qualquer projeto de ciência de dados: explorar os dados.

A estatística clássica se concentrava quase que exclusivamente em *inferência*, um conjunto muitas vezes complexo de procedimentos, para tirar conclusões sobre grandes populações com base em pequenas amostras. Em 1962, John W. Tukey (Figura 1-1) sugeriu uma reforma na estatística em seu inovador estudo "The Future of Data Analysis" [Tukey, 1962]. Ele propôs uma nova disciplina científica chamada *análise de dados*, que incluía a inferência estatística como apenas um de seus componentes. Tukey firmou laços com as comunidades de engenharia e ciências da computação (ele criou os termos *bit*, abreviação de binary digit, e *software*), e suas crenças originais são surpreendentemente duráveis e fazem parte dos fundamentos da ciência de dados. O campo da análise de dados exploratórios nasceu com o, agora clássico, livro de Tukey, *Exploratory Data Analysis* [Tukey, 1977]. Tukey apresentou gráficos simples (por exemplo, gráficos de caixa, gráficos de dispersão) que, juntamente com a síntese estatística (média, mediana, quantis etc.), ajudam a pintar o quadro de um conjunto de dados.

Figura 1-1. John Tukey, o ilustre estatístico cujas ideias, apresentadas há mais de 50 anos, fundamentam a ciência de dados.

Com a disponibilidade da capacidade computacional e expressivos softwares de análise de dados, a análise exploratória de dados evoluiu muito além de seu escopo original. As principais características dessa modalidade têm sido o rápido desenvolvimento de novas tecnologias, o acesso a dados maiores e em maior quantidade, e o maior uso de análises quantitativas em diversas modalidades. David Donoho, professor de estatística na Universidade de Stanford e ex-aluno de Tukey na graduação, escreveu um artigo excelente com base em sua palestra no workshop do centenário de Tukey em Princeton, New Jersey [Donoho, 2015]. Donoho traça os primórdios da ciência de dados até o pioneiro trabalho de Tukey em análise de dados.

Elementos dos Dados Estruturados

Os dados vêm de diversas fontes: medições por sensores, eventos, textos, imagens e vídeos. A *Internet das Coisas* (IoT) jorra rios de informação. Muitos desses dados não são estruturados: as imagens são um conjunto de pixels, com cada pixel contendo informações de cor RGB (vermelho, verde, azul); os textos são sequências de palavras e caracteres non-word, geralmente organizados em seções, subseções etc.; os fluxos de cliques são sequências de ações de um usuário interagindo com um aplicativo ou uma página da internet. Na verdade, um dos maiores desafios da ciência de dados é trabalhar nessa torrente de dados brutos e transformá-la em informação prática. Para aplicar os conceitos estatísticos deste livro, os dados brutos não estruturados devem ser processados e manipulados de uma forma estruturada, pois podem vir de um banco de dados relacional ou ser coletados de um estudo.

Termos-chave dos Tipos de Dados

Numéricos

Dados expressos em uma escala numérica.

Contínuos
Dados que podem assumir qualquer valor em um intervalo.
Sinônimos
intervalo, flutuante, numérico

Discretos
Dados que podem assumir apenas valores inteiros, como contagens.
Sinônimos
inteiro, contagem

Categóricos
Dados que podem assumir apenas um conjunto específico de valores representando um conjunto de possíveis categorias.
Sinônimos
enumeração, enumerado, fatores, nominal

> ***Binários***
> Um caso especial de dados categóricos com apenas duas categorias de valores, por exemplo, 0/1, true/false).
> *Sinônimos*
> dicotômico, lógico, indicador, booliano
>
> ***Ordinais***
> Dado categórico que tem uma ordem explícita.
> *Sinônimo*
> fator ordenado

Existem dois tipos básicos de dados estruturados: numéricos e categóricos. Os dados numéricos aparecem de duas formas: *contínuos*, como a velocidade do vento ou o tempo de duração, e *discretos*, como a contagem de ocorrências de um evento. Os dados *categóricos* requerem apenas um conjunto fixo de valores, como um tipo de tela de TV (plasma, LCD, LED etc.) ou o nome de um estado (Sergipe, Paraná etc.). Os dados *binários* são um importante caso especial de dados categóricos que requerem apenas um de dois valores, como 0/1, sim/não ou true/false. Outro tipo útil de dados categóricos é o dado *ordinal*, no qual as categorias são ordenadas. Um exemplo é a classificação numérica (1, 2, 3, 4 ou 5). Por que a taxonomia dos tipos de dados é importante? Acontece que, para fins de análise de dados e modelagem preditiva, o tipo de dados é importante para ajudar a determinar a exposição visual, a análise de dados ou o modelo estatístico. Inclusive, os softwares de ciência de dados, como *R* e *Python*, utilizam esses tipos de dados para melhorar seu desempenho computacional. Além disso, o tipo de dados para uma variável determina como o software processará os cálculos para tal variável.
Os engenheiros de software e os programadores de bancos de dados podem se perguntar por que precisamos da noção de dados *categóricos* e *ordinais* para a análise. Afinal, as categorias são meramente um conjunto de valores de texto (ou numéricos) e o banco de dados subjacente processa automaticamente a representação interna. No entanto, a identificação explícita dos dados como categóricos, diferentes de texto, tem algumas vantagens:

- Saber que os dados são categóricos, como um sinal informando ao software como os procedimentos estatísticos — como produzir um gráfico ou ajustar um modelo — devem se comportar. Em particular, os dados

ordinais podem ser representados como `ordered.factor` em *R*, preservando uma ordenação especificada pelo usuário em gráficos, tabelas e modelos. Em *Python*, `scikit-learn` suporta dados ordinais com `sklearn.preprocessing.OrdinalEncoder`.
- O armazenamento e a indexação podem ser otimizados (como em um banco de dados relacional).
- Os possíveis valores que uma variável categórica pode requerer são reforçados no software (como uma enumeração).

O terceiro "benefício" pode levar a um comportamento não intencional ou inesperado: o comportamento padrão das funções de importação de dados em *R* (por exemplo, `read.csv`) é converter automaticamente uma coluna de texto em `factor`. As operações subsequentes nessa coluna presumirão que os únicos valores permitidos nela são aqueles importados originalmente, e atribuir um novo valor de texto introduzirá um aviso e produzirá um `NA` (valor ausente). O pacote `pandas` em *Python* não fará tal conversão automaticamente. No entanto, você pode especificar uma coluna como categórica explicitamente na função `read_csv`.

Ideias-chave

- Os dados geralmente são classificados por tipo nos softwares.
- Os tipos de dados incluem numéricos (contínuos, discretos) e categóricos (binários, ordinais).
- A tipagem de dados em um software atua como um sinal para o software de como processar os dados.

Leitura Adicional

- A documentação `pandas` descreve os diferentes tipos de dados e como eles podem ser manipulados em *Python*.
- Os tipos de dados podem ser confusos, pois eles podem se sobrepor, e a taxonomia em um software pode diferir em outro. O site R Tutorial trata da taxonomia para *R*.
- Os bancos de dados são mais detalhados em sua classificação dos tipos de dados, incorporando considerações de níveis de precisão, campos de comprimento fixo ou variável e mais. Consulte o guia W3Schools para SQL.

Dados Retangulares

O quadro de referências típico para uma análise em ciência de dados é um objeto de *dados retangulares*, como uma planilha ou uma tabela do banco de dados.

Termos-chave para Dados Retangulares

Data frame

Os dados retangulares (como uma planilha) são a estrutura básica de dados para os modelos estatísticos e de aprendizado de máquina (machine learning).

Atributo

Uma coluna na tabela costuma ser chamado de *atributo*.

Sinônimos

entrada, preditor, variável

Resultado

Muitos projetos de ciência de dados envolvem a previsão de um *resultado* — geralmente, uma saída sim/não (na Tabela 1-1, é "o leilão foi competitivo ou não"). Os *atributos* por vezes são usados para prever o *resultado* em um experimento ou um estudo.

Sinônimos

variável dependente, resposta, alvo, saída

Registros

Uma linha na tabela costuma ser chamada de *registro*.

Sinônimos

caso, exemplo, instância, observação, padrão, amostra

Dado retangular é o termo geral para uma matriz bidimensional com as linhas indicando os registros (casos) e as colunas indicando os atributos (variáveis); *data frame* é o formato específico em *R* e *Python*. Os dados nem sempre começam dessa forma: os dados não estruturados (por exemplo, texto) devem ser processados e tratados de modo a serem representados como um conjunto de atributos nos dados retangulares (veja "Elementos dos Dados Estruturados", anteriormente neste capítulo). Na maioria das tarefas de análise e modelagem de dados, os dados nos bancos de dados relacionais devem ser extraídos e colocados em uma única tabela.

Na Tabela 1-1 existe um mix de dados medidos ou contados (ex., duração e preço) e dados categóricos (ex., categoria e moeda). Como mencionado antes, uma forma especial de variável categórica é uma variável binária (sim/não ou 0/1), vista na coluna mais à direita na Tabela 1-1 — uma variável indicadora

mostrando se um leilão foi competitivo (tinha vários licitantes) ou não. Essa variável indicadora também é uma variável de *saída*, quando o cenário é prever se um leilão é competitivo ou não.

Tabela 1-1. Um formato de dados típico

Categoria	Moeda	ClassVend	Duração	DiaFinal	PreçoFim	PreçoInício	Competitivo?
Música/Filme/Game	US	3249	5	Seg	0.01	0.01	0
Música/Filme/Game	US	3249	5	Seg	0.01	0.01	0
Automotivo	US	3115	7	Ter	0.01	0.01	0
Automotivo	US	3115	7	Ter	0.01	0.01	0
Automotivo	US	3115	7	Ter	0.01	0.01	0
Automotivo	US	3115	7	Ter	0.01	0.01	0
Automotivo	US	3115	7	Ter	0.01	0.01	1
Automotivo	US	3115	7	Ter	0.01	0.01	1

Data Frames e Índices

As tabelas de banco de dados tradicionais têm uma ou mais colunas designadas como índice. Isso pode melhorar muito a eficiência em certas consultas no banco de dados. Em *Python*, com a biblioteca `pandas`, a estrutura básica de dados retangulares é um objeto `DataFrame`. Por padrão, um índice de inteiros automático é criado para um `DataFrame` com base na ordem das linhas. Em `pandas` também é possível definir índices multiníveis/hierárquicos para melhorar a eficiência de certas operações.

Em *R*, a estrutura básica de dados retangulares é um objeto `data.frame`. Um `data.frame` tem também um índice de inteiros implícito baseado na ordem das linhas. Uma chave personalizada pode ser criada com o atributo `row.names`, mas o `data.frame` de *R* nativo não suporta índices especificados pelo usuário ou multiníveis. Para resolver essa deficiência, dois novos pacotes são usados: `data.table` e `dplyr`. Ambos suportam índices multiníveis e oferecem boa aceleração no trabalho com um `data.frame`.

Diferenças de Terminologia

A terminologia para os dados retangulares pode ser confusa. Estatísticos e cientistas de dados utilizam termos diferentes para a mesma coisa. Para um estatístico, *variáveis preditoras* são usadas em um modelo para prever uma *resposta* ou uma *variável dependente*. Para um cientista de dados, *atributos* são usados para prever um *alvo*. Um sinônimo é particularmente confuso: os cientistas da computação utilizam o termo *amostra* para uma única linha; para um estatístico, uma *amostra* significa uma coleção de linhas.

Estruturas de Dados Não Retangulares

Existem outras estruturas além dos dados retangulares.

Os dados de séries temporais registram medições sucessivas da mesma variável. São o material bruto para os métodos de previsão estatística, além de serem um componente-chave dos dados produzidos por dispositivos — a Internet das Coisas (IoT).

As estruturas de dados espaciais, usadas em análises de mapeamento e localização, são mais complexas e variadas do que as estruturas de dados retangulares. Na representação do *objeto*, o foco do dado é um objeto (por exemplo, uma casa) e suas coordenadas espaciais. A exibição do *campo*, por outro lado, foca pequenas unidades de espaço e o valor de uma métrica relevante (brilho do pixel, por exemplo).

As estruturas de dados gráficos (ou de rede) são usadas para representar relacionamentos físicos, sociais e abstratos. Por exemplo, um grafo/gráfico de uma rede social, como Facebook ou LinkedIn, pode representar as conexões entre as pessoas na rede. Centros de distribuição conectados por estradas são um exemplo de rede física. As estruturas gráficas são úteis para certos problemas, como a otimização de redes e os sistemas de recomendação.

Cada um desses dados tem sua metodologia especializada em ciência de dados. O foco deste livro são os dados retangulares, o bloco de construção fundamental para a modelagem preditiva.

Gráficos em Estatística

Em ciências da computação e tecnologia da informação, o termo *grafo* geralmente se refere a uma representação das conexões entre as entidades e as estruturas de dados subjacentes. Em estatística, *gráfico* é usado para se referir a uma variedade de diagramas e *visualizações*, não apenas de conexões entre as entidades, e o termo se aplica apenas à visualização, não à estrutura de dados.

Ideias-chave

- A estrutura básica de dados na ciência de dados é uma matriz retangular, em que as linhas são registros e as colunas são variáveis (atributos).
- A terminologia pode ser confusa. Existem diversos sinônimos resultantes das diferentes disciplinas que contribuem com a ciência de dados (estatística, ciências da computação e tecnologia da informação).

Leitura Adicional

- Documentação sobre data frames em *R*
- Documentação sobre data frames em *Python*

Estimativas de Localização

As variáveis com dados de medição ou contagem podem ter milhares de valores diferentes. Um passo fundamental na exploração dos dados é definir um "valor típico" para cada atributo (variável): uma estimativa de onde a maioria dos dados está localizada (ou seja, sua tendência central).

Termos-chave para Estimativas de Localização

Média
A soma de todos os valores, dividida pelo número de valores.

Sinônimo
média aritmética simples

Média ponderada
A soma de todos os valores, multiplicada por um peso e dividida pela soma dos pesos.

Sinônimo
média aritmética ponderada

Mediana
O valor que ocupa a posição central dos dados antes e depois.

Sinônimo
50° percentil

Percentil
O valor que ocupa a porcentagem *P* dos dados depois.

Sinônimo
quantil

Mediana ponderada
O valor cuja posição está no centro da soma dos pesos, estando metade da soma antes e metade depois desse dado classificado.

Média aparada
A média de todos os valores depois da exclusão de um número fixo de valores atípicos.

Sinônimo
média truncada

Robusto
Não sensível a valores atípicos.

Sinônimo
resistente

Valor atípico
Um valor de dados que é muito diferente da maioria dos dados.

Sinônimo
valor atípico

À primeira vista, resumir os dados pode parecer bem simples: basta tirar a *média* deles. Na verdade, apesar de a média ser fácil de calcular e conveniente de usar, nem sempre é a melhor medida para um valor central. Por isso, os estatísticos desenvolveram e promoveram diversas estimativas alternativas à média.

Métricas e Estimativas

Os estatísticos costumam usar o termo *estimativa* para os valores calculados a partir dos dados em mãos, para traçar uma diferença entre o que vemos dos dados e a verdade teórica ou a situação real. Os cientistas de dados e os analistas de negócios costumam se referir a esses valores como *métrica*. A diferença reflete a abordagem estatística *versus* a ciência de dados: a contabilização de incertezas está no centro da disciplina estatística, enquanto os objetivos concretos corporativos ou organizacionais são o foco da ciência de dados. Portanto, os estatísticos estimam e os cientistas de dados medem.

Média

A estimativa de localização mais básica é a média ou o valor *médio*. A média é a soma de todos os valores, dividida pelo número de valores. Considere o seguinte conjunto de números: {3 5 1 2}. A média é (3 + 5 + 1 + 2) / 4 = 11 / 4 = 2,75. Você encontrará o símbolo $\bar{x}$ (chamado "x barra") para representar a média de uma amostra de população. A fórmula para calcular a média de um conjunto de valores n x_1, x_2, …, x_n é:

$$\text{Média} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

N (ou *n*) se refere ao número total de registros ou observações. Em estatística, ele é maiúsculo se é referente a uma população e minúsculo se é referente a uma amostra de população. Em ciência de dados, essa distinção não é vital, então, pode ser visto das duas formas.

Uma variação da média é uma *média aparada*, a qual se calcula excluindo um número fixo de valores selecionados em cada extremidade, então tirando uma média dos valores restantes. Representando esses valores selecionados por $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$, em que $x_{(1)}$ é o menor valor e $x_{(n)}$ é o maior, a fórmula para calcular a média aparada com os maiores e os menores valores p omitidos é:

$$\text{Média aparada} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

Uma média aparada elimina a influência dos valores atípicos. Por exemplo, em uma competição internacional de mergulho, as notas máxima e mínima dos cinco juízes são descartadas, e a nota final é a média dos três juízes restantes. Isso dificulta a manipulação do placar por um único juiz, talvez em favor do competidor de seu país. As médias aparadas são muito usadas, e em muitos casos são preferíveis à média comum (veja "Mediana e Estimativas Robustas" a seguir para ter mais informações).

Outro tipo de média é a *média ponderada*, a qual se calcula pela multiplicação de cada valor de dado x_i por um peso w_i especificado pelo usuário, dividindo sua somatória pela soma de todos os pesos. A fórmula para a média ponderada é:

$$\text{Média ponderada} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Existem duas razões principais para o uso da média ponderada:

- Alguns valores são intrinsecamente mais variáveis que outros e as observações altamente variáveis recebem um peso menor. Por exemplo, se tiramos a média de diversos sensores e um deles é menos preciso, então devemos diminuir o peso dos dados desse sensor.
- Os dados coletados não representam igualmente os diferentes grupos que estamos interessados em medir. Por exemplo, por causa do modo como um experimento online foi conduzido, podemos não ter um conjunto de dados que reflete precisamente todos os grupos na base de usuários. Para corrigir isso, podemos conferir um peso maior aos valores dos grupos que foram sub-representados.

Mediana e Estimativas Robustas

A *mediana* é o número central em uma lista de dados classificada. Se existe um número par de valores de dados, o valor central é aquele que não está realmente no conjunto de dados, mas sim a média dos dois valores que dividem os valores classificados nas metades superior e inferior. Comparada à média, que usa todas as observações, a mediana depende apenas dos valores no centro dos dados classificados. Ainda que isso pareça uma desvantagem, já que a média é muito mais sensível aos dados, existem muitos casos nos quais a mediana é uma estimativa melhor para a localização. Digamos que queiramos observar as rendas familiares em bairros próximos a Lake Washington, em Seattle. Ao comparar o bairro Medina com o bairro Windermere, usando a média teríamos resultados muito diferentes, pois Bill Gates mora em Medina. Se usarmos a mediana, não importa a fortuna de Bill Gates — a posição da observação central permanecerá a mesma.

Pela mesma razão que se usa uma média ponderada, também é possível calcular uma *mediana ponderada*. Como na mediana, primeiro classificamos os dados, porém cada valor de dado tem um peso associado. Em vez do número central, a mediana ponderada é um valor cuja soma dos pesos é igual para as metades antes e depois na lista classificada. Como a mediana, a mediana ponderada é resistente aos valores atípicos.

Valores atípicos

A mediana é chamada de estimativa *robusta* de localização, pois não é influenciada por *valores atípicos* (casos extremos), que podem enviesar os resultados. Um valor atípico é qualquer valor muito distante dos outros valores em um conjunto de dados. A definição exata de um valor atípico é bastante subjetiva, apesar de algumas convenções serem utilizadas em diversos sumários e gráficos de dados (veja "Percentis e Gráficos de Caixa" adiante neste capítulo). Ser um valor atípico por si só não torna um valor de dado inválido ou errado (como no exemplo anterior com Bill Gates). Ainda assim, os valores atípicos costumam ser resultado dos erros de dados, como misturar dados de unidades diferentes (quilômetros e metros) ou leituras ruins de um sensor. Quando os valores atípicos forem resultado de dados ruins, a média resultará em uma estimativa de localização ruim, mas a mediana ainda será válida. Em qualquer caso, os valores atípicos devem ser identificados e costumam ser dignos de maior investigação.

Detecção de Anomalias

Ao contrário da análise de dados comum, em que os valores atípicos, às vezes, são informativos e outras vezes um empecilho, na *detecção de anomalias,* os pontos de

interesse são os valores atípicos e a maior massa de dados serve principalmente para definir o "normal" com o qual as anomalias são medidas.

A mediana não é a única estimativa de localização robusta. Na verdade, a média aparada é muito usada para evitar a influência de valores atípicos. Por exemplo, cortar os 10% iniciais e finais (uma escolha comum) dos dados oferecerá, quase sempre, uma proteção contra os valores atípicos, exceto nos conjuntos de dados menores. A média aparada pode ser vista como um meio-termo entre a mediana e a média: é robusta com valores atípicos nos dados, porém usa mais dados para calcular a estimativa de localização.

Outras Métricas Robustas para a Localização

Os estatísticos desenvolveram uma infinidade de outros estimadores para a localização, principalmente para desenvolver um estimador mais robusto que a média e também mais eficiente (ou seja, mais capaz de discernir pequenas diferenças de localização entre os conjuntos de dados). Esses métodos podem ser muito úteis para os pequenos conjuntos de dados, mas não costumam oferecer maiores benefícios para os conjuntos de dados de tamanho grande ou moderado.

Exemplo: Estimativas de Localização de População e Taxas de Homicídio

A Tabela 1-2 mostra as primeiras linhas do conjunto de dados contendo a população e as taxas de homicídio (em unidades de homicídios a cada 100 mil pessoas por ano) em cada estado nos EUA (censo de 2010).

Tabela 1-2. Algumas linhas de `data.frame` para o estado da população e a taxa de homicídio por estado

	Estado	População	Taxa de homicídios	Abreviação
1	Alabama	4.779.736	5,7	AL
2	Alasca	710.231	5,6	AK
3	Arizona	6.392.017	4,7	AZ
4	Arkansas	2.915.918	5,6	AR
5	Califórnia	37.253.956	4,4	CA
6	Colorado	5.029.196	2,8	CO
7	Connecticut	3.574.097	2,4	CT
8	Delaware	897.934	5,8	DE