

---

# **Estatística Prática para Cientistas de Dados**

*50 Conceitos Essenciais*

*Peter Bruce e Andrew Bruce*



**ALTA BOOKS**  
GRUPO EDITORIAL  
Rio de Janeiro, 2019

## Estatística Prática para Cientistas de Dados - 50 Conceitos Essenciais

Copyright © 2019 da Starlin Alta Editora e Consultoria Eireli. ISBN: 978-85-508-0603-7

*Translated from original Practical Statistics for Data Scientists © 2017 by Peter Bruce and Andrew Bruce. All rights reserved. ISBN 978-1-491-95296-2. This translation is published and sold by permission of O'Reilly Media, Inc the owner of all rights to publish and sell the same. PORTUGUESE language edition published by Starlin Alta Editora e Consultoria Eireli, Copyright © 2019 by Starlin Alta Editora e Consultoria Eireli.*

Todos os direitos estão reservados e protegidos por Lei. Nenhuma parte deste livro, sem autorização prévia por escrito da editora, poderá ser reproduzida ou transmitida. A violação dos Direitos Autorais é crime estabelecido na Lei nº 9.610/98 e com punição de acordo com o artigo 184 do Código Penal.

A editora não se responsabiliza pelo conteúdo da obra, formulada exclusivamente pelo(s) autor(es).

**Marcas Registradas:** Todos os termos mencionados e reconhecidos como Marca Registrada e/ou Comercial são de responsabilidade de seus proprietários. A editora informa não estar associada a nenhum produto e/ou fornecedor apresentado no livro.

Impresso no Brasil — 1ª Edição, 2019 — Edição revisada conforme o Acordo Ortográfico da Língua Portuguesa de 2009.

Publique seu livro com a Alta Books. Para mais informações envie um e-mail para [autoria@altabooks.com.br](mailto:autoria@altabooks.com.br)

Obra disponível para venda corporativa e/ou personalizada. Para mais informações, fale com [projetos@altabooks.com.br](mailto:projetos@altabooks.com.br)

<b>Produção Editorial</b> Editora Alta Books	<b>Produtor Editorial</b> Juliana de Oliveira Thiê Alves	<b>Marketing Editorial</b> <a href="mailto:marketing@altabooks.com.br">marketing@altabooks.com.br</a>	<b>Vendas Atacado e Varejo</b> Daniele Fonseca Viviane Paiva <a href="mailto:comercial@altabooks.com.br">comercial@altabooks.com.br</a>	<b>Ouvidoria</b> <a href="mailto:ouvidoria@altabooks.com.br">ouvidoria@altabooks.com.br</a>
<b>Gerência Editorial</b> Anderson Vieira	<b>Assistente Editorial</b> Illysbelle Trajano	<b>Editor de Aquisição</b> José Rugeri <a href="mailto:j.rugeri@altabooks.com.br">j.rugeri@altabooks.com.br</a>		
<b>Equipe Editorial</b>	Adriano Barros Bianca Teodoro Ian Verçosa	Kelry Oliveira Keyciane Botelho Mária de Lourdes Borges	Paulo Gomes Thales Silva Thauan Gomes	
<b>Tradução</b> Luciana Ferraz	<b>Copidesque</b> Alessandro Thomé	<b>Revisão Gramatical</b> Hellen Suzuki Rochelle Lassarot	<b>Revisão Técnica</b> José G. Lopes Estatístico pela Universidade de Brasília	<b>Diagramação</b> Lucia Quaresma

**Erratas e arquivos de apoio:** No site da editora relatamos, com a devida correção, qualquer erro encontrado em nossos livros, bem como disponibilizamos arquivos de apoio se aplicáveis à obra em questão.

Acesse o site [www.altabooks.com.br](http://www.altabooks.com.br) e procure pelo título do livro desejado para ter acesso às erratas, aos arquivos de apoio e/ou a outros conteúdos aplicáveis à obra.

**Suporte Técnico:** A obra é comercializada na forma em que está, sem direito a suporte técnico ou orientação pessoal/exclusiva ao leitor.

A editora não se responsabiliza pela manutenção, atualização e idioma dos sites referidos pelos autores nesta obra.

### Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD

B886e	Bruce, Peter
	Estadística Prática para Cientistas de Dados: 50 Conceitos Essenciais / Peter Bruce, Andrew Bruce ; traduzido por Luciana Ferraz. - Rio de Janeiro : Alta Books, 2019. 320 p. : il. ; 17cm x 24cm.
	Tradução de: Practical Statistics for Data Scientists Inclui bibliografia e índice. ISBN: 978-85-508-0603-7
	1. Ciência de dados. 2. Estatística. I. Bruce, Andrew. II. Ferraz, Luciana. III. Título.
2019-460	CDD 005.13 CDU 004.62

Elaborado por Wagner Rodolfo da Silva - CRB-8/9410



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](http://www.altabooks.com.br) — [altabooks@altabooks.com.br](mailto:altabooks@altabooks.com.br)  
[www.facebook.com/altabooks](http://www.facebook.com/altabooks) — [www.instagram.com/altabooks](http://www.instagram.com/altabooks)



---

# Sumário

<b>Prefácio .....</b>	<b>XV</b>
<b>1. Análise Exploratória de Dados .....</b>	<b>1</b>
Elementos de Dados Estruturados	2
Leitura Adicional	4
Dados Retangulares	5
Quadros de Dados e Índices	6
Estruturas de Dados Não Retangulares	7
Leitura Adicional	8
Estimativas de Localização	8
Média	9
Mediana e Estimativas Robustas	10
Exemplo: Estimativas de Localização de População e	
Taxas de Homicídio	12
Leitura Adicional	13
Estimativas de Variabilidade	13
Desvio-padrão e Estimativas Relacionadas	15
Estimativas Baseadas em Percentis	17
Exemplo: Estimativas de Variabilidade de População Estadual	18
Leitura Adicional	19
Explorando a Distribuição de Dados	19
Percentis e Boxplots	20
Tabela de Frequências e Histogramas	21

Estimativas de Densidade	24
Leitura Adicional	26
Explorando Dados Binários e Categóricos	26
Moda	28
Valor Esperado	28
Leitura Adicional	29
Correlação	29
Gráficos de Dispersão	32
Leitura Adicional	34
Explorando Duas ou Mais Variáveis	34
Compartimentação Hexagonal e Contornos (Representando Numéricos versus Dados Numéricos)	35
Duas Variáveis Categóricas	37
Dados Categóricos e Numéricos	38
Visualizando Variáveis Múltiplas	40
Leitura Adicional	42
Resumo	42
<b>2. Distribuições de Dados e Amostras.....</b>	<b>43</b>
Amostragem Aleatória e Viés de Amostra	44
Viés	46
Seleção Aleatória	47
Tamanho versus Qualidade: Quando o tamanho importa?	48
Média Amostral versus Média Populacional	49
Leitura Adicional	49
Viés de Seleção	50
Regressão à Média	51
Leitura Adicional	53
Distribuição de Amostragem de uma Estatística	53
Teorema de Limite Central	55
Erro-padrão	56
Leitura Adicional	57
O Bootstrap	57
Reamostragem versus Bootstrapping	60
Leitura Adicional	61

Intervalos de Confiança	61
Leitura Adicional	64
Distribuição Normal	64
Normal Padrão e Gráficos QQ	66
Distribuições de Cauda Longa	68
Leitura Adicional	70
Distribuição t de Student	70
Leitura Adicional	72
Distribuição Binomial	73
Leitura Adicional	75
Poisson e Distribuições Relacionadas	75
Distribuições Poisson	76
Distribuição Exponencial	76
Estimando a Taxa de Falha	77
Distribuição Weibull	77
Leitura Adicional	78
Resumo	78
<b>3. Experimentos Estatísticos e Teste de Significância .....</b>	<b>79</b>
Testagem A/B	80
Por que Ter um Grupo de Controle?	82
Por que apenas A/B? Por que Não C, D...?	83
Leitura Adicional	84
Testes de Hipótese	85
A Hipótese Nula	86
Hipótese Alternativa	87
Teste de Hipótese Unilateral, Bilateral	87
Leitura Adicional	88
Reamostragem	88
Teste de Permutação	89
Exemplo: Aderência Web	90
Testes de Permutação Exaustiva e Bootstrap	93
Testes de Permutação: A conclusão para a Ciência de Dados	93
Leitura Adicional	94

Significância Estatística e Valores P	94
Valor P	96
Alfa	97
Erros Tipo 1 e Tipo 2	98
Ciência de Dados e Valores P	99
Leitura Adicional	99
Testes t	100
Leitura Adicional	101
Testagem Múltipla	102
Leitura Adicional	105
Graus de Liberdade	105
Leitura Adicional	107
ANOVA	107
Estatística F	110
ANOVA Bidirecional	111
Leitura Adicional	112
Teste de Qui Quadrado	112
Teste de Qui Quadrado: Uma Abordagem à Reamostra	113
Teste de Qui Quadrado: Teoria Estatística	114
Teste Exato de Fisher	116
Relevância para a Ciência de Dados	118
Leitura Adicional	119
Algoritmo de Bandido Multibraços	119
Leitura Adicional	123
Potência e Tamanho de Amostra	123
Tamanho da Amostra	124
Leitura Adicional	126
Resumo	127
<b>4. Regressão e Previsão .....</b>	<b>129</b>
Regressão Linear Simples	129
A Equação de Regressão	131
Valores Ajustados e Resíduos	133
Mínimos Quadrados	134

Previsão versus Explicação (Profiling)	135
Leitura Adicional	136
Regressão Linear Múltipla	136
Exemplo: Dados Imobiliários de King County	137
Avaliando o Modelo	138
Validação Cruzada	140
Seleção de Modelo e Regressão Passo a Passo	141
Regressão Ponderada	144
Previsão Usando Regressão	145
Os Perigos da Extrapolação	145
Intervalos de Confiança e Previsão	146
Variáveis Fatoriais em Regressão	148
Representação de Variáveis Fictícias	148
Variáveis Fatoriais com Muitos Níveis	151
Variáveis de Fator Ordenado	152
Interpretando a Equação de Regressão	153
Preditoras Correlacionadas	154
Multicolinearidade	155
Variáveis de Confundimento	156
Interações e Efeitos Principais	157
Testando as Suposições: Diagnósticos de Regressão	159
Outliers	160
Valores Influentes	161
Heteroscedasticidade, Não Normalidade e Erros Correlacionados	164
Gráficos Residuais Parciais e Não Linearidade	167
Regressão Polinomial e Spline	169
Polinomial	170
Splines	171
Modelos Aditivos Generalizados	173
Leitura Adicional	175
Resumo	175
<b>5. Classificação.....</b>	<b>177</b>
Naive Bayes	178
Por que a Classificação Bayesiana Exata é Impraticável	179

A Solução Naive	180
Variáveis Preditoras Numéricas	182
Leitura Adicional	182
Análise Discriminante	183
Matriz de Covariância	184
Discriminante Linear de Fisher	184
Um Exemplo Simples	185
Leitura Adicional	187
Regressão Logística	188
Função de Resposta Logística e Logito	189
Regressão Logística e o GLM	190
Modelos Lineares Generalizados	191
Valores Previstos a Partir da Regressão Logística	192
Interpretando os Coeficientes e as Razões de Chances	193
Regressão Linear e Logística: Semelhanças e Diferenças	194
Avaliando o Modelo	196
Leitura Adicional	199
Avaliando Modelos de Classificação	199
Matriz de Confusão	200
O Problema da Classe Rara	202
Precisão, Revocação e Especificidade	202
Curva ROC	203
AUC	205
Lift	206
Leitura Adicional	208
Estratégias para Dados Desequilibrados	208
Undersampling	209
Oversampling e Ponderação Acima/Abaixo	210
Geração de Dados	211
Classificação Baseada em Custos	212
Explorando as Previsões	212
Leitura Adicional	214
Resumo	214

<b>6. Aprendizado de Máquina Estatístico .....</b>	<b>215</b>
K-Vizinhos Mais Próximos	216
Um Pequeno Exemplo: Prevendo Inadimplência em Empréstimos	217
Métricas de Distância	219
One Hot Encoder	220
Padronização (Normalização, Escores Z)	221
Escolhendo K	223
KNN como um Motor de Característica	224
Modelos de Árvore	226
Um Exemplo Simples	227
O Algoritmo Recursivo de Repartição	229
Medindo Homogeneidade ou Impureza	230
Fazendo a Árvore Parar de Crescer	232
Prevendo um Valor Contínuo	233
Como as Árvores São Usadas	234
Leitura Adicional	235
Bagging e a Floresta Aleatória	235
Bagging	236
Floresta Aleatória	237
Importância da Variável	240
Hiperparâmetros	242
Boosting	243
O Algoritmo de Boosting	245
XGBoost	245
Regularização: Evitando Sobreajuste	247
Hiperparâmetros e Validação Cruzada	251
Resumo	254
<b>7. Aprendizado Não Supervisionado .....</b>	<b>255</b>
Análise dos Componentes Principais	256
Um Exemplo Simples	257
Calculando os Componentes Principais	259
Interpretando os Componentes Principais	260
Leitura Adicional	262

Agrupamento por K-Médias	263
Um Exemplo Simples	263
Algoritmo de K-Médias	266
Interpretando os Agrupamentos	267
Escolhendo o Número de Grupos	269
Agrupamento Hierárquico	271
Um Exemplo Simples	272
O Dendrograma	272
O Algoritmo Aglomerativo	273
Medidas de Dissimilaridade	274
Agrupamento Baseado em Modelos	276
Distribuição Normal Multivariada	276
Misturas de Normais	278
Selecionando o Número de Grupos	280
Leitura Adicional	282
Escalonamento e Variáveis Categóricas	282
Escalonando as Variáveis	283
Variáveis Dominantes	285
Dados Categóricos e Distância de Gower	286
Problemas com Agrupamento de Dados Mistos	288
Resumo	290
<b>Bibliografia</b> .....	<b>291</b>
<b>Índice</b> .....	<b>293</b>