

O'REILLY®

TURING

图灵程序设计丛书

面向数据科学家的 实用统计学

Practical Statistics for Data Scientists

系统梳理数据科学中重要的统计学概念，演示
统计学方法在数据科学中的应用



[美] 彼得·布鲁斯 安德鲁·布鲁斯 著
盖磊 译



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS

TURING

图灵程序设计丛书

面向数据科学家的实用统计学

Practical Statistics for Data Scientists
50 Essential Concepts

[美] 彼得·布鲁斯 安德鲁·布鲁斯 著
盖磊 译

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

O'Reilly Media, Inc. 授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目 (C I P) 数据

面向数据科学家的实用统计学 / (美) 彼得·布鲁斯
(Peter Bruce), (美) 安德鲁·布鲁斯 (Andrew Bruce)
著; 盖磊译. — 北京: 人民邮电出版社, 2018. 10
(图灵程序设计丛书)
ISBN 978-7-115-49366-8

I. ①面… II. ①彼… ②安… ③盖… III. ①统计软
件 IV. ①C819

中国版本图书馆CIP数据核字(2018)第212556号

内 容 提 要

本书解释了数据科学中至关重要的统计学概念, 介绍如何将各种统计方法应用于数据科学。作者以易于理解、浏览和参考的方式, 引出统计学中与数据科学相关的关键概念; 解释各统计学概念在数据科学中的重要性及有用程度, 并给出原因。

本书适合数据科学从业人员阅读。

-
- ◆ 著 [美] 彼得·布鲁斯 安德鲁·布鲁斯
译 盖 磊
责任编辑 岳新欣
责任印制 周昇亮
 - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
北京市艺辉印刷有限公司印刷
 - ◆ 开本: 800×1000 1/16
印张: 14.75
字数: 349千字 2018年10月第1版
印数: 1—3 000册 2018年10月北京第1次印刷
著作权合同登记号 图字: 01-2018-3440号
-

定价: 89.00元

读者服务热线: (010)51095186转600 印装质量热线: (010)81055316

反盗版热线: (010)81055315

广告经营许可证: 京东工商广登字 20170147号

站在巨人的肩上
Standing on Shoulders of Giants



iTuring.cn

版权声明

© 2017 by Peter Bruce and Andrew Bruce.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and Posts & Telecom Press, 2018. Authorized translation of the English edition, 2017 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版，2017。

简体中文版由人民邮电出版社出版，2018。英文原版的翻译得到 O'Reilly Media, Inc. 的授权。此简体中文版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc. 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 *Make* 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过图书出版、在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——*Wired*

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——*Business 2.0*

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

本书献给我们的父母维克多·布鲁斯和南希·布鲁斯，以纪念他们。正是他们培养了我们对于数学和科学的热爱。也献给我们早年的导师约翰·图基和朱利安·西蒙，以及我们毕生的朋友杰夫·沃森。他们启发我们以统计学作为一生的职业。

前言

本书所面向的读者是那些在一定程度上熟悉 R 编程语言，并具有一些统计学知识（即便是碎片化的知识，或是短期接触过统计学）的数据科学家。作为本书的作者，我们都是从统计学领域迈入数据科学领域的，因此对统计学在数据科学中可做的贡献有所了解。同时，我们也十分清楚传统的统计学教学的局限所在，即统计学作为一门学科已经有 150 多年的历史了，大多数统计学课本和课程都表现出远洋轮船般的动量和惯性，很难有所改变。

本书有两大目标：

- 以易于理解、浏览和参考的方式，引出统计学中与数据科学相关的关键概念；
- 解释各个统计学概念在数据科学中的重要性 and 有用程度，并给出原因。

本书的独到之处

主要术语

数据科学融合了多门学科，包括统计学、计算机科学、信息技术和一些特定领域的研究。因此，同一个概念可能会使用多个不同的术语表述。本书将使用类似此处的格式，突出显示各个主要术语及其同义词。

排版约定

本书将使用如下排版约定。

- **黑体字**
用于标识新的术语。
- 等宽字体 (constant width)
用于标识程序清单，以及段落内引用的程序元素，例如变量、函数名称、数据库、数据类型、环境变量、程序语句和程序语言关键字等。

- 等宽粗体 (**constant width bold**)
表示应由用户逐字输入的命令或其他一些文本内容。
- 等宽斜体 (*constant width italic*)
表示文本应被替换，替换内容由用户提供，或取决于上下文。



此图标表示一个知识点或一条建议。



此图标表示一处通用注解。



此图标表示一条警告或警示。

使用代码示例

本书的补充材料（即示例代码、练习等）可从 <https://github.com/andrewgbruce/statistics-for-data-scientists> 下载。

本书旨在帮助你更好地完成工作。一般来说，只要是本书提供的示例代码，你都可以用于自己的程序和文档。除非你需要大规模地使用本书的代码，否则无须联系作者以获得许可。例如，你在编写代码时使用了书中的几处代码是不需要获得许可的，但销售或分发 O'Reilly 图书中的 CD-ROM 则需要获得许可。在回答问题时引用本书内容和示例代码不需要获得许可，但在产品文档中整合本书中的大量示例代码需要获得许可。

我们很希望但并不强制要求你在引用本书内容时加上引用说明。引用说明通常包括书名、作者、出版商和 ISBN。例如：“*Practical Statistics for Data Scientists* by Peter Bruce and Andrew Bruce (O'Reilly). Copyright 2017 Peter Bruce and Andrew Bruce, 978-1-491-95296-2.”

如果你认为自己对示例代码的使用超出了合理使用的范围或是上面介绍的许可范围，可随时通过电子邮件 permissions@oreilly.com 联系我们。

Safari® Books Online



Safari Books Online 是一个按需提供服务的数字图书馆，所提供的图书和视频来自于在技术和商业上处于世界领先地位的作者。

Safari Books Online 已被专业技术人员、软件开发人员、Web 设计师以及商业和专业创意人员使用，成为科学研究、解决问题、学习与认证培训的主要资源。

Safari Books Online 为企业、政府、教育机构和个人提供了一系列的计划和定价。

会员访问一个功能完备的数据库检索系统，就可以获得上百家出版商的上千种图书、培训视频和预发行手稿，其中包括 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 等。有关 Safari Books Online 的更多信息，可在线访问。

联系我们

请把对本书的评价和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

本书页面提供了本书的勘误表、例子及其他信息，网址为 <http://shop.oreilly.com/product/0636920048992.do>。¹

对 O'Reilly 图书的评论和技术问题，可以发送电子邮件到 bookquestions@oreilly.com。

关于 O'Reilly 图书、课程、会议和新闻更多内容，参见网站 <http://www.oreilly.com>。

在 Facebook 上关注我们：<http://facebook.com/oreilly>。

在 Twitter 上关注我们：<http://twitter.com/oreillymedia>。

在 YouTube 上关注我们：<http://www.youtube.com/oreillymedia>。

致谢

作为本书的作者，我们希望能在此感谢对本书出版提供过帮助的许多人。

数据挖掘公司 Elder Research 的首席执行官 Gerhard Pilcher 审读了本书的初稿，并做出了详细而有用的修正和评论。同样，SAS 的统计学家 Anya McGuirk 和 Wei Xiao，以及同是 O'Reilly 作者的 Jay Hilfiger，也对初稿提出了有益的反馈。

注 1：本书中文版的勘误请到 <http://www.ituring.com.cn/book/2066> 查看和提交。——编者注

在 O'Reilly 出版社方面，Shannon Cutt 给予我们鼓励并适当地敦促我们，还在出版流程上提供了指导。Kristen Brown 使本书得以顺利地推进到生产制作阶段。Rachel Monaghan 和 Eliahu Sussman 耐心而又细致地修改了本书的书稿。Ellen Troutman-Zaig 为本书做了索引。我们还要感谢 O'Reilly 发起本书项目的 Marie Beaugureau，以及 statistics.com 讲师兼 O'Reilly 作者 Ben Bengfort，正是他将我们介绍给了 O'Reilly。

Galit Shmueli 曾与 Peter 合著过其他图书，并且多年来一直与 Peter 保持交流。这种交流使我们乃至本书都受益匪浅。

最后，我们要特别感谢 Elizabeth Bruce 和 Deborah Donnell，没有她们的耐心和支持，就不会有这本书。

电子书

扫描如下二维码，即可购买本书电子版。



目录

前言	xiii
第 1 章 探索性数据分析	1
1.1 结构化数据的组成	2
1.2 矩形数据	4
1.2.1 数据框和索引	5
1.2.2 非矩形数据结构	5
1.2.3 拓展阅读	6
1.3 位置估计	6
1.3.1 均值	7
1.3.2 中位数和稳健估计量	8
1.3.3 位置估计的例子：人口和谋杀率	9
1.3.4 拓展阅读	10
1.4 变异性估计	10
1.4.1 标准偏差及相关估计值	11
1.4.2 基于百分位数的估计量	13
1.4.3 例子：美国各州人口的变异性估计量	14
1.4.4 拓展阅读	14
1.5 探索数据分布	14
1.5.1 百分位数和箱线图	15
1.5.2 频数表和直方图	16
1.5.3 密度估计	18
1.5.4 拓展阅读	20
1.6 探索二元数据和分类数据	20
1.6.1 众数	21
1.6.2 期望值	22
1.6.3 拓展阅读	22

1.7	相关性	22
1.7.1	散点图	25
1.7.2	拓展阅读	26
1.8	探索两个及以上变量	26
1.8.1	六边形图和等势线 (适用于两个数值型变量)	26
1.8.2	两个分类变量	28
1.8.3	分类数据和数值型数据	29
1.8.4	多个变量的可视化	31
1.8.5	拓展阅读	33
1.9	小结	33
第 2 章	数据和抽样分布	34
2.1	随机抽样和样本偏差	35
2.1.1	偏差	36
2.1.2	随机选择	37
2.1.3	数据规模与数据质量: 何时规模更重要	38
2.1.4	样本均值与总体均值	38
2.1.5	拓展阅读	39
2.2	选择偏差	39
2.2.1	趋均值回归	40
2.2.2	拓展阅读	41
2.3	统计量的抽样分布	42
2.3.1	中心极限定理	44
2.3.2	标准误差	44
2.3.3	拓展阅读	45
2.4	自助法	45
2.4.1	重抽样与自助法	47
2.4.2	拓展阅读	48
2.5	置信区间	48
2.6	正态分布	50
2.7	长尾分布	53
2.8	学生 t 分布	55
2.9	二项分布	57
2.10	泊松分布及其相关分布	58
2.10.1	泊松分布	59
2.10.2	指数分布	59
2.10.3	故障率估计	60
2.10.4	韦伯分布	60
2.10.5	拓展阅读	61
2.11	小结	61
第 3 章	统计实验与显著性检验	62
3.1	A/B 测试	62

3.1.1	为什么要有对照组	64
3.1.2	为什么只有处理 A 和 B, 没有 C、D	65
3.1.3	拓展阅读	66
3.2	假设检验	66
3.2.1	零假设	67
3.2.2	备择假设	67
3.2.3	单向假设检验和双向假设检验	68
3.2.4	拓展阅读	68
3.3	重抽样	68
3.3.1	置换检验	69
3.3.2	例子: Web 黏性	69
3.3.3	穷尽置换检验和自助置换检验	72
3.3.4	置换检验: 数据科学的底线	72
3.3.5	拓展阅读	72
3.4	统计显著性和 p 值	72
3.4.1	p 值	74
3.4.2	α 值	75
3.4.3	第一类错误和第二类错误	76
3.4.4	数据科学与 p 值	76
3.4.5	拓展阅读	77
3.5	t 检验	77
3.6	多重检验	78
3.7	自由度	81
3.8	方差分析	82
3.8.1	F 统计量	84
3.8.2	双向方差分析	85
3.8.3	拓展阅读	86
3.9	卡方检验	86
3.9.1	卡方检验: 一种重抽样方法	86
3.9.2	卡方检验: 统计理论	88
3.9.3	费舍尔精确检验	88
3.9.4	与数据科学的关联	90
3.9.5	拓展阅读	91
3.10	多臂老虎机算法	91
3.11	检验效能和样本规模	93
3.11.1	样本规模	95
3.11.2	拓展阅读	96
3.12	小结	96
第 4 章	回归与预测	97
4.1	简单线性回归	97
4.1.1	回归方程	98
4.1.2	拟合值与残差	100

4.1.3	最小二乘法	101
4.1.4	预测与解释 (剖析)	102
4.1.5	拓展阅读	103
4.2	多元线性回归	103
4.2.1	美国金县房屋数据案例	103
4.2.2	评估模型	104
4.2.3	交叉验证	106
4.2.4	模型选择和逐步回归法	107
4.2.5	加权回归	108
4.3	使用回归做预测	109
4.3.1	外推法的风险	109
4.3.2	置信区间和预测区间	110
4.4	回归中的因子变量	111
4.4.1	虚拟变量的表示	112
4.4.2	多层因子变量	113
4.4.3	有序因子变量	114
4.5	解释回归方程	115
4.5.1	相关的预测变量	116
4.5.2	多重共线性	117
4.5.3	混淆变量	117
4.5.4	交互作用和主效应	118
4.6	检验假设: 回归诊断	119
4.6.1	离群值	120
4.6.2	强影响值	121
4.6.3	异方差性、非正态分布和相关误差	123
4.6.4	偏残差图和非线性	126
4.7	多项式回归和样条回归	127
4.7.1	多项式回归	128
4.7.2	样条回归	129
4.7.3	广义加性模型	131
4.7.4	拓展阅读	132
4.8	小结	133
第5章	分类	134
5.1	朴素贝叶斯算法	135
5.1.1	准确的贝叶斯分类是不切实际的	136
5.1.2	朴素解决方案	136
5.1.3	数值型预测变量	138
5.1.4	拓展阅读	138
5.2	判别分析	138
5.2.1	协方差矩阵	139
5.2.2	费希尔线性判别分析	139
5.2.3	一个简单的例子	140

5.2.4	拓展阅读	142
5.3	逻辑回归	142
5.3.1	逻辑响应函数和 Logit 函数	143
5.3.2	逻辑回归和广义线性模型	144
5.3.3	广义线性模型	145
5.3.4	逻辑回归的预测值	145
5.3.5	解释系数和优势比	146
5.3.6	线性回归与逻辑回归：相似之处和不同之处	147
5.3.7	模型评估	148
5.3.8	拓展阅读	150
5.4	评估分类模型	150
5.4.1	混淆矩阵	151
5.4.2	稀有类问题	152
5.4.3	准确率、召回率和特异性	153
5.4.4	ROC 曲线	153
5.4.5	AUC	155
5.4.6	提升	156
5.4.7	拓展阅读	157
5.5	不平衡数据的处理策略	157
5.5.1	欠采样	158
5.5.2	过采样以及上权重和下权重	158
5.5.3	数据生成	159
5.5.4	基于代价的分类	160
5.5.5	探索预测值	160
5.5.6	拓展阅读	161
5.6	小结	161
第 6 章	统计机器学习	162
6.1	K 最近邻算法	163
6.1.1	预测贷款拖欠的示例	164
6.1.2	距离度量	165
6.1.3	独热编码	166
6.1.4	标准化	166
6.1.5	K 值的选取	168
6.1.6	KNN 作为特征引擎	169
6.2	树模型	170
6.2.1	一个简单的例子	171
6.2.2	递归分区算法	172
6.2.3	测量同质性或不纯度	174
6.2.4	阻止树模型继续生长	175
6.2.5	预测连续值	176
6.2.6	如何使用树模型	176
6.2.7	拓展阅读	177

6.3	Bagging 和随机森林	177
6.3.1	Bagging 方法	178
6.3.2	随机森林	178
6.3.3	变量的重要性	181
6.3.4	超参数	183
6.4	Boosting	184
6.4.1	Boosting 算法	184
6.4.2	XGBoost 软件	185
6.4.3	正则化: 避免过拟合	186
6.4.4	超参数和交叉验证	189
6.5	小结	191
第 7 章 无监督学习		192
7.1	主成分分析	193
7.1.1	一个简单的例子	194
7.1.2	计算主成分	195
7.1.3	解释主成分	196
7.1.4	拓展阅读	198
7.2	K-Means 聚类	198
7.2.1	一个简单的例子	199
7.2.2	K-Means 算法	201
7.2.3	解释类	201
7.2.4	选择类的个数	203
7.3	层次聚类	204
7.3.1	一个简单的例子	205
7.3.2	树状图	205
7.3.3	凝聚算法	206
7.3.4	测量相异性	207
7.4	基于模型的聚类	208
7.4.1	多元正态分布	209
7.4.2	混合正态分布	210
7.4.3	类数的选取	212
7.4.4	拓展阅读	213
7.5	变量的缩放和分类变量	213
7.5.1	变量的缩放	214
7.5.2	控制变量	215
7.5.3	分类数据和高氏距离	216
7.5.4	混合数据的聚类问题	218
7.6	小结	219
作者简介		220
封面说明		220