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（CIP）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

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和有用程度，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、会议和新闻等更多内容，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