“本书从务实的角度出发，清晰阐释了R的基本知识及统计数据分析，为我提供了很大帮助。”

——亚马逊读者评论

R in Action
Data Analysis and Graphics with R

R语言实战

大数据时代已经到来，但数据分析、数据挖掘人才却十分短缺。由于“大数据”对每个领域的决定性影响，相对于经验和直觉，在商业、经济及其他领域中基于数据和分析去发现问题并作出科学、客观的决策越来越重要。开源软件R是世界上最流行的数据分析、统计计算及制图语言，几乎能够完成任何数据处理任务，可安装并运行于所有主流平台，为我们提供了成千上万的专业模块和实用工具，是从大数据中获取有用信息的绝佳工具。

本书从解决实际问题入手，尽量跳脱统计学的理论阐述来讨论R语言及其应用，讲解清晰透澈，极具实用性。作者不仅高度概括了R语言的强大功能、展示了各种实用的统计示例，而且对于难以用传统方法分析的凌乱、不完整和非正态的数据也给出了完备的处理方法。通读本书，你将全面掌握使用R语言进行数据分析、数据挖掘的技巧，并领略大量探索和展示数据的图形功能，从而更加高效地进行分析与沟通。

想要成为倍受高科技企业追捧的、炙手可热的数据分析师吗？想要科学分析数据并正确决策吗？不妨从本书开始，挑战大数据，用R开始炫酷的数据统计与分析吧！

MANNING

图灵社区：www.ituring.com.cn
新浪微博：@图灵教育　@图灵社区
反馈/投稿/推荐信箱：contact@turingbook.com
热线：(010)51095186转604

分类建议 计算机/程序设计/R

人民邮电出版社网址：www.ptpress.com.cn

本书内容

- R安装与操作
- 数据导入/导出及格式化
- 双变量关系的描述性分析
- 回归分析
- 模型适用性的评价方法以及结果的可视化
- 用图形实现变量关系的可视化
- 在给定置信度的前提下确定样本量
- 高级统计分析方法和高级绘图

Robert I. Kabacoff

R语言社区著名学习网站Quick-R（http://www.statmethods.net/）的幕后维护者，现为全球化开发与咨询公司Management研究集团研发副总裁。此前，Kabacoff博士是佛罗里达诺瓦东南大学的教授，讲授定量方法和统计编程的研究生课程。Kabacoff还是临床心理学博士、统计顾问，擅长数据分析，在健康、金融服务、制造业、行为科学、政府和学术界有20余年的研究和统计咨询经验。

ISBN 978-7-115-29990-1

ISBN 978-7-115-29990-1

定价：79.00元

版权声明

那些年，我们一起学过的R语言

公元前

2007年上半年的一天，一堆做基因组序列分析的代码让我认识了R语言。那是一份高质量的R代码，各种标准的向量化操作、匿名函数、优美的代码格式让我欣喜，也让我茫然。同年暑假，有幸到清华大学学习，刘军老师布置下来的作业是用HMM模型预测蛋白质二级结构。我壮着胆子，硬着头皮，以C语言的风格用R语言完成了作业（各种循环，各种丑陋的标量）。

那些年，R语言所有的参考资料就是官方的几本手册，但庆幸的是，还有丁国徽博士翻译的中文版。

公元纪年开始

2008年的初冬，北京市海淀区中国人民大学的一间阶梯教室内，举办了一场小众、既不太学术技术档次也不高的会议。教室外的墙上挂着一条横幅，上书“第一届中国R语言会议”。这算是R语言在国内发展历程中的一个里程碑。100多人参加了为期一天多的会议。参加那次会议的人不少都成为了现在中国R语言社区最活跃的人，比如谢益辉、刘思喆、李舰、张翔、魏太云、陈堰平等人。当然，其中也有当时就已经算是R语言社区元老的吴喜之老师和丁国徽博士。

没记错的话，当时出席会议的还有机械工业出版社的编辑。因为作为会议的承办机构，统计之都社区（http://cos.name）的骨干们已经想通过出版一本高质量R语言书来推动R语言在中国的发展，而出版界也已经开始注意到这个小众市场。那时候，大部分R语言书籍来自学术界。水平高深莫测的专家教授们，乃至R语言的发明人 Robert C. Gentleman 大叔写出来的书，都让我这种初窥门径的人越看越糊涂。同时，中文的R语言书籍也开始出现，但都只是将R语言作为某个特定领域（比如生物信息学）的数据分析工具。

文艺复兴

会议举办了，人也都混了个脸熟，但R语言在国内的发展依旧不温不火。直到2011年，大数据突然火了起来，R语言一举杀进编程语言排行榜前20名。刘思喆同学在“码农”界主流媒体《程序员》上的文章，让R语言一下子走到了很多人面前。大家发现，这个经常被描述成统计编程语

言的东西并不是仅仅擅长统计，其底层融合C/C++/Fortran等各种语言的优势、层出不穷的新模型、日趋成熟的开发设施，再加上它跟Hadoop、多核计算、MPI等高性能计算技术的迅速结合，让人们看到了它在大数据时代的潜力。

O'Reilly、Manning等技术图书出版商迅速跟进。与以往的学术出版社不同，它们的加入让R语言书籍更容易被普通读者接受，也迅速降低了R语言的门槛。

你现在翻开的这本书，就是由业内大名鼎鼎的Quick-R网站（http://www.statmethods.net）的创始人Robert I. Kabacoff撰写并由Manning出版的。全书分为四部分，由浅入深地介绍了R语言本身，以及如何用R语言实现或简单或复杂的数据分析和绘图。而书后8个附录中关于大数据分析、自定义启动环境、图形界面等方面的内容，有一些早已被志愿者翻译成中文在互联网上广泛流传。本书内容质量之高，权威性之强，由此可见一斑，足以帮读者快速地走过我曾经历过的迷茫。

结束公元纪年

这本书的翻译工作开始的时候我还在学校读书，实验室里有两三个人在使用R语言做数据分析，为发表论文而努力。现在，我坐在公司的工位上，周围的每个人都或多或少在使用R。整个研发部门一千多人，多半每天都会用到R语言。R语言已经是新员工技术培训的必修内容。

若干天前，同事们在讨论中描绘出一个宏大的愿景：用高效的数据分析手段，建立起海量生物实验数据到所有生物表型的预测模型。如果成功，这将结束公元纪年。这个愿景的核心不是新的生物实验技术，也不是如何采集样本，而是数据分析。

有点画大饼的意思了，就此打住。让我们翻开这本书，或许，公元纪年真的能在我们手中结束。

——陈钢

2012年9月4日夜于深圳华大基因研究院

译者致谢

感谢我的家人和女朋友敏敏，在翻译本书的过程中，他们给了我无限的支持与鼓励；感谢好友肖楠和师兄陈钢，他们细心的校正和耐心的解惑让翻译进行得更加顺畅；感谢统计之都的谢益辉博士和魏太云师兄，他们引领我走上了R之路，让我在统计的世界里获益匪浅；感谢所有R包贡献者无私的分享，他们让统计变得更加多姿多彩！

——高涛

http://www.gaotao.name

感谢我的家人，他们的支持和宽容让我在艰难时刻获得了内心的平静；感谢我的导师许青松教授在学术上的悉心指导和严格要求；感谢中国人民大学的魏太云师兄引导我走上技术的通途；感谢无名的R黑客们，是他们的无私贡献，铸就了这把统计计算的“屠龙刀”。

——肖楠

http://www.road2stat.com

首先要感谢两位合作者：肖楠和高涛。虽然整本书的翻译工作是我发起和协调的，但其实是他们高效地完成了这本书绝大部分内容的翻译，质量也大大超出了我的预期。我非常高兴能有机会跟他们合作。

感谢明尼苏达大学的龚午鸣博士和现在不知在何处高就的Tongbin Li博士带领我进入R语言的世界。感谢统计之都和历届R语言会议的主办者为R语言在中国的发展付出的不懈努力。感谢参与此书出版工作的傅志红、岳新欣和毛倩倩等编辑。另外要感谢Dirk Eddelbuettel同意我们将其在*Journal of Statistical Software*上对此书的评论翻译成中文出版。

谨以此书献给我的夫人王倩和儿子陈涤菲，他们是我所有努力的动力源泉。

——陈钢

http://www.gossipcoder.com

前　言

要是一本书里没有图画和对话，那还有什么意思呢？

——爱丽丝，《爱丽丝梦游仙境》

它太神奇了，满载珍宝，可以让那些聪明狡猾和粗野胆大的人得到充分满足；但并不适合胆小者。

——Q，“Q Who？”，《星际迷航：下一代》

在开始写这本书时，我花了很多时间搜索适合于开始本书的名言警句。最后，我找到了这两句话。R是一个非常灵活的平台，是专用于探索、展示和理解数据的语言，因此我引用了《爱丽丝梦游仙境》的句子来表示当今统计分析的潮流——一个探索、展示和理解的交互式过程。

第二句话反映了大部分人对R的看法：难学。但你完全没必要这样想。虽然R很强大，应用广泛，不论你是新手还是略有经验的用户，众多的分析和绘图函数（超过50 000个）都很容易让你望而却步，但实际上并非无规律可循。只要有合适的指导，你就可以畅游其中，选择所需的工具，用最优雅、最简洁、最高效的方式来完成工作——那真的很酷！

多年前，我在申请一个统计咨询职位时，第一次遇到了R。雇主在正式面试前发来的材料中问我是否熟悉R。根据猎头的建议，我立马回答“是的，我很熟悉”，然后开始恶补R。在统计和研究方面我有丰富的经验，作为SAS和SPSS程序员也有25年的工作经验，而且对各种编程语言也颇为精通。学习R能有多难？但事与愿违。

在学习这门语言的过程中（因为要面试，我要尽可能地快），我发现这门语言无论是底层的结构还是各种高级的统计方法，都是由各具体领域的专家为同行专家编写的。看在线帮助简直就是折磨，那不是教程，都是参考手册。每当我觉得自己已经对R的结构和功能有足够把握时，就会发现一些闻所未闻的新东西，它们让我感觉自己很渺小。

为了解决这些问题，我开始以数据科学家的角度学习R。我开始思考如何才能成功地处理、分析和理解数据，包括：

- 获取数据（从各种数据源将数据导入程序）；
- 整理数据（编码缺失值、修复或删除错误数据、将变量转换成更方便的格式）；
- 注释数据（以记住每段数据的含义）；

- 总结数据（通过描述性统计量了解数据的概况）；
- 数据可视化（一图胜千言）；
- 数据建模（解释数据间的关系，检验假设）；
- 整理结果（创建具有出版水平的表格和图形）。

然后，我试图用R来完成这些任务。通过教授别人来学习是最好的方式，所以我创建了一个网站（www.statmethods.net），不断把我学到的东西放在上面。

大概一年后，Marjan Bace（Manning的出版人）打电话给我，问我是不是能写一本关于R的书。那时我已经写了50篇期刊文章、4份技术手册，以及大量章节的内容，还写了一本关于研究方法的书，所以，写一本关于R的书能有多难？结果依然是事与愿违。

你现在捧着的这本书是我多年来梦寐以求的。我试图提供一份R的指南，让你能尽快感受到R的强大以及开源的魅力，不再感到沮丧和忧虑。我希望你能喜欢本书。

另外，虽然当年我成功地申请到了那个职位，但并未入职。不过，学习R的经历改变了我的职业方向，这是我未曾想到的。真可谓人生如戏。

致谢

很多人都对本书精益求精并付出了辛勤的劳动，在此让我对他们一一表示感谢。

- Marjan Bace，Manning出版人，最初劝说我撰写本书的人。
- Sebastian Stirling，进度编辑，花了大量时间与我电话沟通，帮我组织材料、理清概念，帮我润色文字，在整个出版过程中给了我很多帮助。
- Karen Tegtmeyer，评审编辑，帮助寻找审稿人并协调评审进度。
- Mary Piergies及其团队成员Liz Welch、Susan Harkins和Rachel Schroeder，他们指导了本书的出版过程。
- Pablo Domínguez Vaselli，技术审读人，帮我理清了很多易混淆的地方，从独立而专业的角度测试了代码。
- 所有花费时间审读本书内容，寻找书写错误和提供了宝贵建议的审稿人：Chris Williams、Charles Malpas、Angela Staples、Daniel Reis Pereira博士、D. H. van Rijn博士、Christian Marquardt博士、Amos Folarin、Stuart Jefferys、Dror Berel、Patrick Breen、Elizabeth Ostrowski、Atef Ouni博士、Carles Fenollosa、Ricardo Pietrobon、Samuel McQuillin、Landon Cox、Austin Ziegler、Rick Wagner、Ryan Cox、Sumit Pal、Philipp K. Janert、Deepak Vohra和Sophie Mormede。
- 在本书完成前参与MEAP（Manning早期试读计划）的同仁，他们提出了重要的问题、指出了书中的错误并提供了有益的建议。

他们每个人的贡献都让本书的质量更上一层楼。

我还想感谢为R成为如此强大的数据分析平台而做出卓越贡献的软件开发人员。这其中有R的核心开发者，还有那些开发R包和维护各种软件包的个人，他们极大地扩展了R的功能。附录F

罗列了本书中涉及的软件包的作者。其中，我要特别感谢John Fox、Hadley Wickham、Frank E. Harrell、Deepayan Sarkar和William Revelle。我会尽可能准确地介绍他们的贡献，并为本书中所有可能存在的错误或是误导性描述负责。

在本书开头，我还应该感谢我的妻子，同时她也是我的合作者：Carol Lynn。她对统计学和编程都没有太多兴趣，但却反复阅读了每一章的内容，帮助纠正了很多问题并提出了大量建议。为了他人而研读多元统计学实在是一件很有爱的事情。同样重要的是，她容忍我在深夜和周末编写此书，给予我无限的包容、支持和关怀。我真的感到非常幸运。

我还要感谢两个人。一位是我父亲，他对科学的热爱影响了我，还让我认识到了数据的价值。另一位是Gary K. Burger——我读研究生时的导师。我有段时间觉得自己想成为一名医生，是Gary引领我进入统计学和教育领域，这一切都是他赐予的。

关于本书

如果你翻开了本书，那么很有可能是因为要做一些数据的收集、总结、转换、探索、建模、可视化或呈现方面的工作。如果确实如此，那么R完全能够满足你的需求！R已经成了统计、预测分析和数据可视化的全球通用语言。它提供各种用于分析和理解数据的方法，从最基础的到最前沿的，无所不包。

R是一个开源项目，在很多操作系统上都可以免费得到，包括Windows、Mac OS X和Linux。R还在持续发展中，每天都在纳入新的功能。此外，R还得到了社区的广泛支持，这个社区里既有数据科学家也有程序员，他们很乐于为R的用户提供帮助或建议。

R以能创建漂亮优雅的图形而闻名，但实际上它可以处理各种统计问题。基本的安装就提供了数以百计的数据管理、统计和图形函数。不过，R很多强大的功能都来自社区开发的数以千计的扩展（包）。

但这些好处都是有代价的。对于新手来说，经常遇到的两个基本难题就是：R到底是什么以及R究竟能做什么？甚至是经验丰富的R用户也常常发现一些他们之前闻所未闻的新功能。

本书是一本R指南，高度概括了该软件和它的强大功能。本书会介绍基本安装中最重要的函数，以及90多个重要扩展包中的函数。整本书都是围绕实际应用展开的，你将学会理解数据并能够与他人交流这种对数据的理解。通读本书，你应该会对R的原理和功能有基本的了解，并知道从什么地方学习更多的相关知识。你将能用各种技术实现数据的可视化，还能解决各种难度的数据分析问题。

读者对象

每一个要处理数据的人都应该读读本书，他们不需要任何统计编程或R语言知识背景。R语言新手完全能够读懂本书，而有经验的R老手也能在本书中发现很多实用的新东西。

没有统计背景，但需要用R操作数据、总结数据、绘制图形的读者会觉得第1章～第6章、第11章和第16章比较容易理解。第7章和第10章则需要读者学过一学期的统计学课程；第8章、第9章和第12章～第15章则需要读者学过一学年的统计学课程。不过，我尽可能地让每一章都能同时迎合数据分析新手和专家的需求，让所有人都能从中获益。

本书结构

本书的目的是让读者熟悉R平台，重点关注那些能马上应用到数据操作、可视化和理解的方法。全书共16章，分为4部分："入门"、"基础方法"、"中级方法"和"高级方法"。在8个附录中还有更多的相关内容。

第1章首先简要介绍了R，以及它作为数据分析平台的诸多特性。这一章主要介绍了R的获取，以及如何用网上的扩展包增强R基本安装的功能。另外，它还介绍了用户界面，以及如何以交互方式和批处理方式运行程序。

第2章介绍了向R中导入数据的诸多方法。这一章的前半部分介绍了R用来存储数据的数据结构，以及如何用键盘输入数据。后半部分介绍了怎样从文本文件、网页、电子表格、统计软件和数据库向R导入数据。

很多用户最初接触R都是为了绘制图形，我们在第3章会对此作介绍。这一章介绍了创建、修改图形的方法，以及如何将图形保存为各种格式的文件。

第4章探讨了基本的数据管理，包括数据集的排序、合并、取子集，以及变量的转换、重编码和删除。

在第4章的基础上，第5章涵盖了数据管理中函数（数学函数、统计函数、字符函数）和控制结构（循环、条件执行）的用法。然后我们介绍如何编写自己的R函数，以及如何用不同的方法整合数据。

第6章演示了创建常见单变量图形的方法，例如柱状图、饼图、直方图、密度图、箱线图和点图。这些图形对于理解单变量的分布都很有用。

第7章首先演示了如何总结数据，包括使用描述统计量和交叉表。然后，这一章介绍了用于分析两变量间关系的基本方法，包括相关性、t检验、卡方检验和非参数方法。

第8章介绍了针对一个数值型结果变量与一系列数值型预测变量间的关系进行建模的回归方法，详细给出了拟合模型的方法、适用性评价和含义解释。

第9章介绍了基于方差及其变体对基本实验设计的分析。此处，我们通常感兴趣的是处理方式的组合或条件对数值结果变量的影响。这一章还介绍了如何评价分析的适用性，以及如何可视化地展示分析结果。

第10章详细介绍了功效分析。这一章首先讨论了假设检验，重点是如何判断在给定置信度的前提下需要多少样本才能判断处理的效果。这可以帮助我们安排实验和准实验研究来获得有用的结果。

第11章扩展了第5章的内容，介绍了创建表现两个或多个变量间关系的图形。这包括各种2D和3D的散点图、散点图矩阵、折线图、相关图和马赛克图。

第12章介绍了一些稳健的数据分析方法，它们能处理比较复杂的情况，比如数据来源于未知或混合分布、有小样本问题、有恼人的异常值，或者依据理论分布设计假设检验非常复杂且在数学上难以处理的情况。这一章介绍的方法包括重抽样和自助法——很容易在R中实现的需要大量计算机资源的方法。

第13章扩展了第8章中介绍的回归方法，分析非正态分布的数据。这一章首先介绍了广义线性模型，然后重点介绍了如何预测类别型变量（Logistic回归）或计数变量（泊松回归）。

多元数据分析的一个难点是简化数据。第14章介绍了如何将大量的相关变量转换成较少的不相关变量（主成分分析），以及如何发现一系列变量中的潜在结构（因子分析）。这些方法涉及许多步骤，每一步都有详细的介绍。

实际工作中面临的一个普遍问题是数据值缺失，第15章介绍了一个应对此问题的现代方法。R中有很多简捷的方法可以用来分析因各种原因导致缺失而生成的不完整数据。这一章对一些好的方法都有介绍，还具体说明了在什么情况下应该用哪一种以及应该避免使用哪些方法。

第16章介绍了R中最先进、最有用的数据可视化方法，包括用lattice图形表现非常复杂的数据，简要介绍新的ggplot2包，并对各种跟图形实时交互的方法做了综述。

后记中介绍了一些优秀的网站，有助于读者进一步学习R、加入R社区、获得帮助，并及时获得R这个快速发展的软件的最新信息。

最后的内容也很重要，8个附录（从A到H）扩展了正文的一些内容，包括R中的图形用户界面、自定义和升级R、导出数据到其他软件、创建出版级质量的输出、（像MATLAB一样）用R做矩阵计算，以及处理大型数据集。

例子

为了让本书内容尽可能接近各个领域的实际情况，我从心理学、社会学、医学、生物、商业和工程等诸多领域选取了一些例子。所有的这些例子都不需要读者具备这些领域的专业知识。

这些例子中所使用的数据集是经过精心挑选的，因为它们不仅提出了有趣的问题，而且比较小。这样能让读者专注于技术，快速地理解所涉及的过程。在学习新方法时，数据集小是有好处的。

这些数据集有些是R基本安装中就有的，有些则可以通过网上下载软件包来获得。每个例子的代码都可以从www.manning.com/RinAction[①]下载。为了更好地理解本书中的内容，我建议读者在阅读本书时试试这些例子。

经常听人引用这么一句话：如果你问两个统计学家该如何分析一个数据集，你会得到三个答案。反过来说，每个答案都能让你更好地理解数据集。对于一个问题，我不会说某种分析方式是最好的，或者是唯一的。读者应该用本书中学到的技术动手分析数据，看看都能得到什么。R是交互式的，最好的学习方法就是自己尝试。

排版约定

下面是本书的排版约定。

- 等宽字体用于代码清单。

① 也可在图灵社区（www.ituring.com.cn）本书网页免费注册下载。——编者注

- 等宽字体还用于在一般的正文中表示代码或之前定义的对象。
- 代码清单中的斜体表示占位符。你应该用自己问题中的文本和值来替换它们。例如，*path_to_my_file*就应该用该文件在你自己电脑上的实际路径来替换。
- R是一种交互式语言，用提示符（默认是>）表示已经准备好读取用户的下一行输入。本书中的很多代码清单都是从交互式会话中截取的。当你看到代码是以>开头时，不要输入这个提示符。
- 用行内注释作为代码注释（这是Manning图书的传统做法）。此外，有些注释会以有序项目符号的形式出现（如❶），它们对应稍后正文中对代码作出的解释。
- 为了节约版面，让正文更紧凑，我们会在交互式会话的输出中加入一些空白，同时也会删除一些与当前讨论问题无关的文字。

作者在线

在购买本书英文版的同时，你便获得了访问Manning出版社运营的私密Web论坛的权限，在这里你可以发表图书评论、询问技术问题，还可以从作者或其他读者那里获得帮助。用浏览器访问www.manning.com/RinAction就可以访问和订阅这个论坛。这个网页说明了注册后如何访问论坛、能获得何种帮助以及论坛上的行为规范等信息。

Manning致力于为读者之间以及读者和作者之间提供一个良好的交流空间。作者对论坛的参与完全是自愿的，他们对AO论坛的贡献都是（无偿的）志愿行为。我们建议读者向作者提一些有挑战性的问题，作者对这样的问题会更有兴趣。

在本书英文版的整个销售期中，大家都可以从出版商的网站上访问AO论坛，阅读以前的讨论。

关于封面图片

本书的封面图片标题是“来自扎达尔的男人”。这张图片取自19世纪中期Nikola Arsenovic的一本克罗地亚传统服饰图集的复刻版，由克罗地亚斯普利特的Ethnographic博物馆在2003年时出版。图片由Ethnographic博物馆一位热心的图书管理员提供。斯普利特在中世纪时是罗马帝国的核心，从大概公元304年起，卸任的帝国国王戴克里安（Diocletian）所居住的皇宫就在这里。这本书中涵盖了克罗地亚各个地区色彩斑斓的图片，并对服饰和日常生活做了介绍。

扎达尔（Zadar）是克罗地亚达尔马提亚（Dalmatian）海岸北方的一个古罗马时期的城镇，有着两千年的历史，曾在数百年的时间里是康斯坦丁堡和西方的贸易通道上的重要港口。它坐落于一个伸向亚得里亚海的半岛上，周围被各种大大小小的岛屿环绕，如画般的风景，加上罗马帝国时代的遗迹、护城河和古老的石头城墙，让这里成为了旅行者的圣地。封面图片上的人穿着蓝色的羊毛裤子和白色的麻质衬衫，外披点缀着当地特色刺绣的蓝色马甲和夹克，再加上红色羊毛腰带和帽子，就构成了一套完整的服饰。

在这过去的二百年里，服饰和生活方式都发生了巨大的变化，各地当时的特色已随时间流逝。现如今，来自不同大陆的人都已难以区分，更不用说相隔仅数英里的村子和城镇居民了。或许，文化多样性也是我们为获得丰富多彩的个人生活而付出的代价——现在生活无疑是更多姿多彩的快节奏的高科技生活。

Manning出版社用两个世纪前各地独具特色的生活方式来赞美计算机行业的诞生和发展，用古老书籍和图册中的图片让我们领略那个时代的风土人情。

Part 1 第一部分

入　门

欢迎阅读本书！R 是现今最受欢迎的数据分析和可视化平台之一。它是自由的开源软件，并同时提供 Windows、Mac OS X 和 Linux 系统的版本。通读本书，你将掌握精通这个功能全面的软件所需的技能，有效地使用它分析自己的数据。

本书共分四部分。第一部分涵盖了软件的安装、软件界面的操作、数据的导入，以及如何将数据修改成可供进一步分析的格式等基本知识。

第一章将带你熟悉 R 环境。这一章首先是 R 的概览，介绍使其成为强大的现代数据分析平台的独有特性。在简要介绍了如何获取和安装 R 之后，我们通过一系列的简单示例探索了 R 的用户界面。接着，你将学习如何通过可从在线仓库中免费下载的扩展（称为用户贡献包）来增强基本安装的功能。最后，本章以一个示例结尾，让你自测学到的新技术。

熟悉了 R 的界面之后，下一个挑战是将数据导入到程序中。在当今这个信息丰富的世界中，数据的来源和格式多种多样。第 2 章全面介绍向 R 中导入数据的多种方式。此章的前半部分介绍了 R 用以存储数据的各种数据结构，并描述了如何手工输入数据。后半部分讨论了从文本文件、网页、电子表格、统计软件和数据库导入数据的方法。

从工作流程的观点考虑，下一步理应讨论数据管理和数据清理问题。然而，许多第一次接触 R 的用户都对其强大的图形功能表现出了浓厚的兴趣。为了不扫你的兴，第 3 章我们直接开始探索图形的绘制问题。这一章对创建图形、自定义图形、以各种格式保存图形的方法进行了综述，描述了如何设定图形中使用的颜色、符号、线条类型、字体、坐标轴、标题、标签以及图例，最后还介绍了将多个图形组合为单个图形的方法。

尝试过 R 的图形功能之后，我们再重返数据分析的正题。由于数据很少以直接可用的格式出现，因此在开始解决感兴趣的问题之前，我们经常不得不将大量时间花在从不同的数据源组合数据、清理脏数据（误编码的数据、不匹配的数据、含缺失值的数据），以及新变量（组合后的变量、变换后的变量、重编码的变量）的创建上。第 4 章讲述了 R 中基本的数据管理任务，包括数据集的排序、合并、取子集，以及变量的变换、重编码和删除。

第 5 章在第 4 章的基础上，进一步讲解了数据管理中数值（算术运算、三角运算和统计运算）函数和字符处理（字符串取子集、连接和替换）函数的使用。为了阐明许多相关函数的用法，整章使用了一个综合示例进行讲解。接下来是关于控制结构（循环、条件执行）的讨论，你将学到如何编写 R 函数。编写自定义函数能够让你将许多程序执行步骤封装在单个的函数中进行灵活调用，这大大拓展了 R 的功能。因为数据的重塑和整合对于为进一步分析而准备数据的阶段通常很有用，所以最后将讨论一些重组（重塑）数据和整合数据的强大方法。

学习完第一部分之后，你将完全熟悉 R 环境的编程，并可掌握输入和访问数据、清理数据，以及为进一步分析做数据准备所需的技术。另外，你还会获得创建、自定义和保存多种图形的经验。

目　录

第一部分　入　门

第 1 章　R 语言介绍……3
1.1　为何要使用 R?……4
1.2　R 的获取和安装……6
1.3　R 的使用……7
1.3.1　新手上路……7
1.3.2　获取帮助……10
1.3.3　工作空间……10
1.3.4　输入和输出……12
1.4　包……14
1.4.1　什么是包……14
1.4.2　包的安装……14
1.4.3　包的载入……14
1.4.4　包的使用方法……15
1.5　批处理……15
1.6　将输出用为输入——结果的重用……16
1.7　处理大数据集……16
1.8　示例实践……17
1.9　小结……18
第 2 章　创建数据集……19
2.1　数据集的概念……19
2.2　数据结构……20
2.2.1　向量……21
2.2.2　矩阵……22
2.2.3　数组……23
2.2.4　数据框……24
2.2.5　因子……27
2.2.6　列表……29
2.3　数据的输入……30
2.3.1　使用键盘输入数据……31
2.3.2　从带分隔符的文本文件导入数据……32
2.3.3　导入 Excel 数据……33
2.3.4　导入 XML 数据……34
2.3.5　从网页抓取数据……34
2.3.6　导入 SPSS 数据……34
2.3.7　导入 SAS 数据……34
2.3.8　导入 Stata 数据……35
2.3.9　导入 netCDF 数据……35
2.3.10　导入 HDF5 数据……35
2.3.11　访问数据库管理系统……36
2.3.12　通过 Stat/Transfer 导入数据……37
2.4　数据集的标注……37
2.4.1　变量标签……38
2.4.2　值标签……38
2.5　处理数据对象的实用函数……38
2.6　小结……39
第 3 章　图形初阶……40
3.1　使用图形……40
3.2　一个简单的例子……42
3.3　图形参数……43
3.3.1　符号和线条……45
3.3.2　颜色……46
3.3.3　文本属性……47
3.3.4　图形尺寸与边界尺寸……49
3.4　添加文本、自定义坐标轴和图例……50
3.4.1　标题……51
3.4.2　坐标轴……52
3.4.3　参考线……54

3.4.4 图例……54
3.4.5 文本标注……56
3.5 图形的组合……58
3.6 小结……64
第 4 章 基本数据管理……65
4.1 一个示例……65
4.2 创建新变量……67
4.3 变量的重编码……68
4.4 变量的重命名……69
4.5 缺失值……70
4.5.1 重编码某些值为缺失值……71
4.5.2 在分析中排除缺失值……72
4.6 日期值……73
4.6.1 将日期转换为字符型变量……74
4.6.2 更进一步……74
4.7 类型转换……74
4.8 数据排序……75
4.9 数据集的合并……76
4.9.1 添加列……76
4.9.2 添加行……76
4.10 数据集取子集……77
4.10.1 选入（保留）变量……77
4.10.2 剔除（丢弃）变量……77
4.10.3 选入观测……78
4.10.4 subset()函数……79
4.10.5 随机抽样……79
4.11 使用 SQL 语句操作数据框……80
4.12 小结……81
第 5 章 高级数据管理……82
5.1 一个数据处理难题……82
5.2 数值和字符处理函数……83
5.2.1 数学函数……83
5.2.2 统计函数……84
5.2.3 概率函数……86
5.2.4 字符处理函数……89
5.2.5 其他实用函数……90
5.2.6 将函数应用于矩阵和数据框……91
5.3 数据处理难题的一套解决方案……93
5.4 控制流……96
5.4.1 重复和循环……97
5.4.2 条件执行……97
5.5 用户自编函数……99
5.6 整合与重构……101
5.6.1 转置……101
5.6.2 整合数据……101
5.6.3 reshape 包……102
5.7 小结……105
第二部分 基本方法
第 6 章 基本图形……108
6.1 条形图……108
6.1.1 简单的条形图……109
6.1.2 堆砌条形图和分组条形图……110
6.1.3 均值条形图……111
6.1.4 条形图的微调……112
6.1.5 棘状图……113
6.2 饼图……114
6.3 直方图……116
6.4 核密度图……118
6.5 箱线图……120
6.5.1 使用并列箱线图进行跨组比较……121
6.5.2 小提琴图……124
6.6 点图……125
6.7 小结……128
第 7 章 基本统计分析……129
7.1 描述性统计分析……130
7.1.1 方法云集……130
7.1.2 分组计算描述性统计量……133
7.1.3 结果的可视化……136
7.2 频数表和列联表……136
7.2.1 生成频数表……137
7.2.2 独立性检验……142
7.2.3 相关性的度量……144
7.2.4 结果的可视化……144