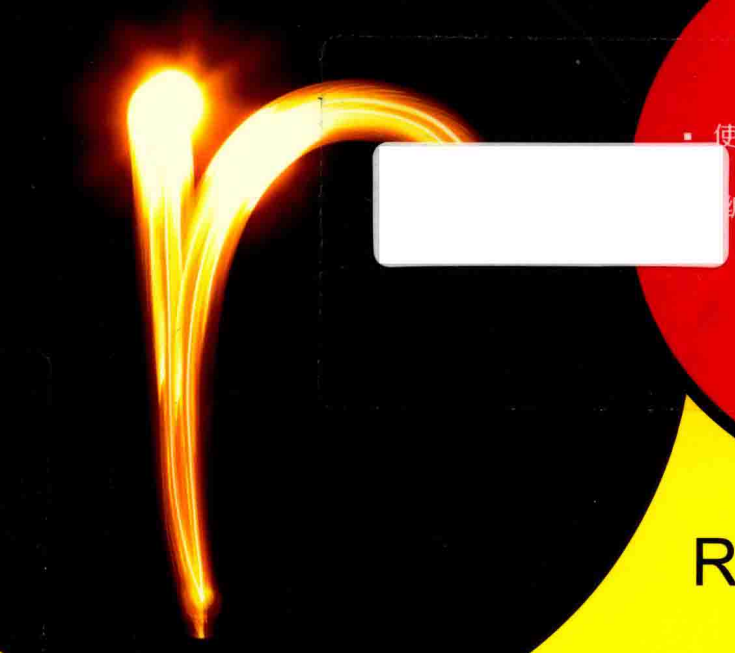
for dummies
A Wiley Brand

达人速

R语言（第2版）可以很简单

[法] Andrie de Vries [比] Joris Meys 著　李毅 译



让学习变得更简单

- 使用R语言对数据进行分析和操作
- 编写函数与脚本来解决重复分析
- 绘制高质量的图表
- 建立统计模型进行分析

R For Dummies
2nd Edition

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

达人迷

R语言可以很简单

for R dummies®

A Wiley Brand

第2版

[法]Andrie de Vries
[比]Joris Meys 著

李毅 译

for dummies®
A Wiley Brand

人民邮电出版社
北京

图书在版编目（CIP）数据

R语言可以很简单 : 第2版 / (法) 安德里·德弗里斯 (Andrie de Vries) , (比) 乔里斯·梅斯 (Joris Meys) 著 ; 李毅译. — 北京 : 人民邮电出版社, 2017.7
(达人迷)
ISBN 978-7-115-45539-0

Ⅰ. ①R… Ⅱ. ①安… ②乔… ③李… Ⅲ. ①程序语言—程序设计 Ⅳ. ①TP312

中国版本图书馆CIP数据核字(2017)第114297号

版权声明

◆ 著 [法] Andrie de Vries [比]Joris Meys
译 李 毅
责任编辑 王峰松
责任印制 焦志炜

◆ 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：24
字数：537 千字 2017 年 7 月第 1 版
印数：1 – 2 400 册 2017 年 7 月北京第 1 次印刷

著作权合同登记号 图字：01-2016-5849 号

定价：69.00 元

读者服务热线：(010) 81055410 印装质量热线：(010) 81055316
反盗版热线：(010) 81055315

内容提要

R是一个开源、跨平台的编程语言，用于统计计算和绘图，特别是其社区开发的数以千计的扩展包为R增加了强大的和前瞻性的功能。数据分析与挖掘已经成为大数据时代重要的技能之一，而R已经成为数据分析领域炙手可热的标志性语言。

本书作为业内外一致好评的Dummies系列书籍之一，是供R语言初学者学习的经典力作。本书通俗易懂地讲解了如何利用R语言基本知识，包括创建、运行以及调试R脚本，用户自定义R函数，用R绘制基本图形，R的循环语句和逻辑控制语句等，逐步引导读者迈入R语言高手行列。

本书适用于数据分析人员以及对R语言感兴趣的读者。

作者简介

Andrie de Vries：他从2009年开始使用R语言来分析调研数据，之后开始活跃在一些开源社区，开发引人注目的软件。同时，Andrie就职于PentaLibra公司，从事时尚品市场的调研和统计分析。随着Andrie对R疯狂地着迷，他加入Revolution Analytics公司使用R来服务于企业客户，帮助客户处理大数据和数据科学带来的挑战。为了陶冶生活，Andrie正在学习瑜伽。

Joris Meys：理学硕士，比利时根特大学讲师，同时也是统计咨询师和R程序员。在获得生物学硕士学位后，他从事了6年环境研究与管理相关的工作，之后获得了统计数据分析的高级硕士学位。Joris为他们系的项目开发了大量代码包，另外他也是R-Forge上某些包的维护者，还与其他合作者共同发表了若干统计专业方面的学术论文。在闲暇时刻，Joris是当地乐队的萨克斯手。

译者简介

李毅，韩国岭南大学理学博士，现为山西财经大学统计学院副教授、硕士生导师，研究方向为应用统计，主持国家自然基金、国家统计局多项重点课题等，发表学术论文20余篇，其中被SCI收录10余篇，其电子邮箱：liyi@sxufe.edu.cn。

献词

这本书献给我的妻子 Annemarie，谢谢她的鼓励、支持和耐心。同时还有我 9 岁的外甥女 Tanya，她数学非常棒，一直很认真地提醒我这本书的截稿日期！最后感谢我的父母，一生的鼓励。

——Andrie de Vries

献给我这一生最重要的女人——我的母亲，因为她成就了现在的我；献给 Eva，深爱着我的人；献给 Amelie，每次她的微笑感染着我；献给 Granny，因为她太酷了！

——Joris Meys

致谢

本书能够面世，完全得益于 Wiley 出版社编辑团队的大力支持。我们尤其感谢本书第一版的编辑 Elizabeth Kuball 和第二版的编辑 Katie Mohr。

感谢我们的技术编辑 Gavin Simpson，他非常仔细地审阅了本书并给出了很好的建议。

感谢 R 语言核心开发团队开发 R 语言、维护 CRAN，并通过各类邮件列表、文档和专题讲座为 R 社区贡献力量。感谢 R 语言社区，感谢大家开发了数以千计的代码包、博客文章，并回答各种疑难问题。

本书使用了若干个由 Hadley Wickham 编写的代码包，他对 ggplot 和 dplyr 等代码包的贡献至今引人注目。

在编写本书的过程中，我们收到大量 R tag 和 Stack Overflow 贡献者的帮助和支持。在这里要感谢：James (JD) Long、David Winsemius、Ben Bolker、Joshua Ulrich、Barry Rowlingson、Roman Luštrik、Joran Elias、Dirk Eddelbuettel、Richie Cotton、Colin Gillespie、Simon Urbanek、Gabor Grotendieck，还有一直努力让 Stack Overflow 成为更优秀的 R 资源社区的爱好者们。

Andrie：毫不夸张地说这本书改变了我的人生轨迹。我学习 R 归功于开源代码社区，写这本书归功于 Revolution Analytics 公司。同时我要感谢我的同事，特别是 Derek McCrae Norton、David Smith 和 Joseph Rickert。

Joris：感谢比利时根特大学数据建模、统计与生物信息系的各位教授和同事，他们在本书撰写过程中所进行了深入而细致的讨论，感谢他们提供的帮助。

前言

欢迎阅读《R 语言可以很简单》。这本书可以帮助读者快速轻松地学习统计编程语言 R。

读过这本书后，虽然不能保证读者会成为 R 语言专家，但读者应该能够掌握以下几点。

- 利用各种强大的工具进行数据分析。
- 使用 R 语言完成统计分析和数据处理任务。
- 领会到向量运算（而不是循环）进行高速计算的魅力。
- 欣赏下面代码的优美：

```
knowledge<- apply(theory, 1, sum)
```

- 知道如何查找、下载和使用 R 语言社区里活跃的开发者贡献的代码。
- 知道在哪里获取额外的帮助和资源，从而进阶 R 语言的编程水平。
- 能够绘制漂亮的图形，实现数据可视化。

关于这本书

《R 语言可以很简单》主要介绍 R 统计编程语言。从 R 语言的 IDE 界面介绍和最基本的语法入手，由浅入深，逐步到更复杂的数据处理和分析。

本书用浅显易懂的实例来解释 R 语言的要点，不仅包含了大量的代码片段，还有很多完整的数据分析脚本，给读者留下很大的动手尝试发挥空间。

本书不对 R 语言的内部实现技术进行介绍，而是更关注 R 语言为什么这样操作。R 语言有很多看起来可能很奇怪的属性，因此，本书将介绍如何对 R 发出命令，以及 R 怎样处理命令。读完本书后，读者不仅能够以其想要的数据格式来管理数据，还会知道如何使用本书没有提及的功能（也包括提到的）。

这是一本参考书。读者不必从头读到尾，而是可以通过目录和索引快速找到想要的内容，书中还会交叉引用其他章节，以提供更多的信息。

第2版的变化

第 1 版书出版以来，R 语言一直不断地发展和改进。为了让本书准确无误，作者更新了代码来反映最新 R 版本（3.2.0 版本）的变化。根据读者、学生和同事的反馈，作者重新修订了部分章节，更正了一些错误，例如，在使用文本字符串时，修改了代码用法——用双引号取代单引号；将列表的基本单元称为组件，而不是元素。

本书中有全新 R 包“rfordummies”案例代码。详情见附录 B。

R 和 R Studio

《R 语言可以很简单》适用于任何操作系统，无论是 Mac、Linux 还是 Windows，这本书都会让读者学会在自己的操作系统中使用 R。

与其说 R 是应用程序，不如说 R 是编程语言。当下载 R 时，适用于读者的操作系统的控制台程序也会一同被下载。然而，这些程序只具备基本的功能，在不同的操作系统中会略有差异。

R Studio 是跨平台应用程序，被认为是非常简洁的 R 语言集成开发环境（IDE）。本书不强制读者使用某种特定的 R 语言控制台。但是，R Studio 在不同的操作系统中提供通用的操作界面，本书将使用 R Studio 而非某个特定操作系统的 R 语言控制台来演示 R 语言的概念和程序。

本书的一些默认规则

下面的一段代码模拟出 2 个六面骰子掷 100 万次的例子。

```
> set.seed(42)
> throws <- 1e6
> dice<- replicate(2,
+                  sample(1:6, throws, replace = TRUE)
+ )
> table(rowSums(dice))

    2     3     4     5     6     7     8
28007 55443 83382 110359 138801 167130 138808
    9    10    11    12
110920 83389 55816 27945
```

这里每一行 R 代码之前都带有一个提示符，有以下两种情况。

- >: 提示符“>”不属于代码的一部分，当读者练习输入代码的时候不需要输入该符号。
- +: 连续符“+”表示该行代码仍属于上一行。读者不必将一行代码分成两行输入，本书中这么写通常是考虑到书中代码排版印刷后的可读性。

R 控制台不会输出提示符或连续符开头的行。在这个例子中，所投掷的 2 个骰子点数之和为 2 到 12。例如，在投掷的 100 万次中，有 28007 次投掷结果的和为 2。

读者可以试着复制这段代码在 R 中运行，但是在确保正确地输入这段代码的同时，需要注意以下三点。

- 不要输入提示符“>”。
- 不要输入连续符“+”。
- 除了关键字，可以在任何位置输入空格或制表符。另外，需要注意换行。

在 R 控制台输入代码时，代码前会出现提示符“>”，如下：

```
>print("Hello world!")
```

在控制台输入这行代码，然后敲击回车键，R 会运行出下面的结果：

```
[1] "Hello world!"
```

为了简便，将上面两行输入代码和输出结果合并成单一代码框来表示，如下：

```
>print("Hello world!")
[1] "Hello world!"
```

在本书中，R 语言的函数、参数等关键词用普通字体来表示。例如，可以使用 plot() 函数来绘制一张数据图。函数名称后面都会有一对小括号——例如 plot()。除非特别必要，一般不会在正文的函数名后面添加参数等额外信息。

本书也会介绍菜单命令，例如 File⇨Save，这表示打开 File 菜单，然后选择 Save 选项。

可以省略阅读的内容

如果这本书适合于读者的学习，但又时间紧迫（或者对某些技术细节不感兴

趣），则读者可以放心地跳过带有“Technical Stuff”图标的内容，也可以略过侧边栏（灰色框中内容），虽然其中包含了不少有趣的信息，但是这对把握整体知识没有任何影响。

读者要求

本书对读者及其计算机有如下要求。

- **要熟悉计算机的基本操作。**读者要知道如何下载和安装软件，可以连接互联网，并且知道如何在互联网上搜索信息。
- **可以不是一个程序员。**如果读者是程序员，并且使用过其他编程语言编写代码，那么推荐关注“Technical Stuff”中的内容，这里会告诉读者 R 语言和其他常见的编程语言之间的异同点。
- **可以不是来自统计学专业，但需要具备基本统计学知识。**这本书不是一本统计学著作，但展示了如何用 R 语言来完成一些基础的统计工作。如果读者想要更深入地了解统计学知识，推荐学习 Deborah J. Rumsey 博士编写的《Statistics For Dummies, 2nd Edition》（Wiley 出版）。
- **想要学习新的知识。**读者喜欢解决问题，并且不惧怕在 R 控制台界面上进行尝试。

本书结构安排

本书由 6 部分组成，每部分内容如下所示。

第一部分：R语言编程入门

在这部分，读者将要编写第一个脚本。使用强大的向量概念来对多个变量进行一次性计算。在 R 工作空间上工作（即如何创建、修改和删除变量），读者会学会保存工作，加载并修改之前编写的 R 脚本文件。另外，我们还会介绍 R 的一些基础知识（如如何安装程序包）。

第二部分：开始使用R

这部分介绍 R 语言三会：读、写和算，即处理文本和数字（包括日期）。另外，读者也可以学到列表和数据框这两种重要的数据结构。

第三部分：编写R代码

R是一种编程语言，因此需要了解如何编写和理解函数。这部分将介绍如何完成相关工作，如何使用if语句控制脚本的逻辑流，以及如何使用循环代码来实现重复操作。此部分还介绍了如何处理代码中产生的警告和错误提示。最后，还会介绍一些工具，帮助读者应对调试中可能遇到的问题。

第四部分：让数据说话

在这部分，首先介绍R语言中使用的不同类型的数据结构，如列表和数据框。读者将学会如何在R语言中将数据导入或者输出（例如，从文件夹或剪贴板中读取数据），并且还会掌握如何将R语言与其他应用程序（如微软Excel软件）进行交互使用。

其次，读者将会发现在R语言中做数据的高级变形和管理是非常容易的。此部分介绍如何对数据的子集进行选择、分类和排序，如何基于相同数据列合并不同的数据集。同时，针对数据子集应用函数来实现分离和合并数据，还介绍了一种非常有效的策略。当读者理解了这个策略后，可以反复地使用，只需要很少的几步就可以进行复杂数据的专业分析。

阅读完这部分后，读者会搞清楚如何使用R语言来描述、总结变量和数据，将能够做一些经典的检验（如t检验），并且学会如何使用随机数来模拟一些分布。

最后，介绍了一些线性模型的基本原理（如线性回归和方差分析）。此外，还展示了如何在R语言上利用数据拟合模型，从而预测新数据。

第五部分：绘制图形

有人说，一张图胜过千言万语。尤其是当读者想和其他人分享自己的研究成果时这一点是肯定的。在这部分中，读者会学习如何绘制基本的甚至更复杂的图形，以可视化形式查看数据。我们将从柱形图和折线图开始，介绍如何使用小平面来展示数据的不同方面。

第六部分：20条有用建议

这部分介绍了如何用R语言来做本可以使用微软Excel来做的10件事。例如，如何做出相同效果的透视图和查询表。另外，本书也给出了10条关于使用非基础R程序包的建议。

本书图标

当读这本书的时候，读者会在页面边缘发现一些小图片。这些图片，或者说是图标，标注了下面的类型。

当看到这个“TIP”图标的时候，一定可以找到一种更简单或者更快速的方法来替代某项操作。

不需要完全记住这本书，但是“REMEMBER”图标标出的内容都是很有用的东西，是必须要记住的。通常这些标示表明会在更多的章节中遇到它们。

当看到“WARNING”图标的时候，一定要提高警惕。它告诉读者哪些事情是千万不能做的。虽然使用 R 并不会导致一些灾难性的错误，但是需要用这个图标提醒读者可能会增添的麻烦。

“TECHNICAL STUFF”图标表示一些技术细节，可以放心地跳过。本书尽可能使这些内容变得有趣，但是如果读者没有时间，或者只需要获取想要的知识，可以跳过这些内容，继续下面的学习。

本书之外的内容

本书还包含了以下可以在互联网下载的相关信息及对应网址：

- **备忘录：** www.dummies. com/cheatsheet/r
- **额外的内容：** www.dummies. com/extras/r
- **源代码：** www.dummies.com/extras/r

下一步

学习 R 语言的唯一途径是：使用它！本书尽力让读者熟悉 R 的用法。但读者最好坐在电脑前亲身体验和尝试，翻开书，开始敲击键盘吧！

目录