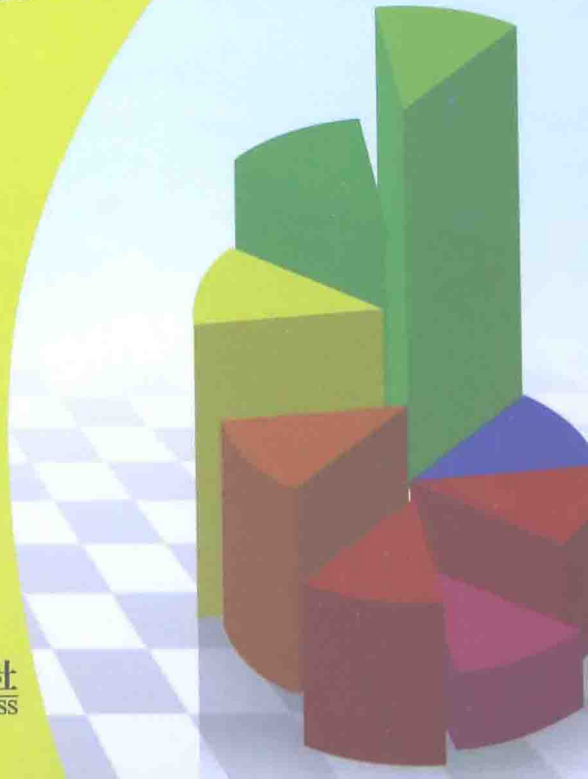
让一切变得更简单！

R FOR DUMMIES 达人迷

R语言轻松入门与提高

通过本书您将学到：

- 用R语言进行数据分析和处理
- 可复用的分析编写函数和脚本
- 制高品质的表格和图表
- 建数据模型，进行数据分析



◎【法】Andrie de Vries
【比利时】Joris Meys 著
◎ 麦秆创智 译

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

R FOR DUMMIES 达人速

R语言轻松入门与提高

◎【法】Andrie de Vries
【比利时】Joris Meys
◎ 麦秆创智 译

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

R语言轻松入门与提高 /（法）德弗里斯(Vries,A.D.)，（比）梅斯（Meys,J.）著 ；麦秆创智译. -- 北京 ：人民邮电出版社，2015.5
（达人迷）
ISBN 978-7-115-38764-6

Ⅰ. ①R… Ⅱ. ①德… ②梅… ③麦… Ⅲ. ①程序语言－程序设计 Ⅳ. ①TP312

中国版本图书馆CIP数据核字(2015)第055759号

版 权 声 明

商 标 声 明

◆ 著　［法］Andrie de Vries　［比利时］Joris Meys
译　麦秆创智
责任编辑　刘　洋
责任印制　彭志环

◆ 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
三河市潮河印业有限公司印刷

◆ 开本：800×1000　1/16
印张：24　2015 年 5 月第 1 版
字数：404 千字　2015 年 5 月河北第 1 次印刷
著作权合同登记号　图字：01-2014-1131 号

定价：69.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

本书首先从基本概念入手，介绍向量的计算与R语言向量化函数的强大之处，然后逐步引导你迈入R编程的世界，从一个统计分析师的角度，细致而深入地讲解了R语言中的数据提取与数据处理，并在科学而系统的统计分析中应用和实践。除此之外，本书还系统地介绍了如何使用R语言来绘制各类数据图表，使你可以方便地将数据转化成可视化元素，丰富数据报表与演示文档。

本书既适用于众多领域的数据分析师和数据处理人员，也适用于对R语言编程感兴趣的广大学生和科研工作者。

作者简介

安德里·德弗里斯（Andrie de Vries）：安德里从2009年开始使用R语言来进行调研数据的分析，同时还活跃在一些开源社区，贡献自己富有创造力的软件作品。安德里是PentaLibra Limited的总监，这是一家从事时尚品市场调研的公司，专注于数据的统计与分析。安德里已经为CRAN贡献了两个R包，并一直努力使人们今后的调研数据统计、分析与报表的生成工作变得更容易。另外，他还积极投入到LimeSurvey的研发工作中，这是一个开源的调研管理系统。

乔里斯·梅斯（JorisMeys）：乔里斯是Ghent大学（比利时）数学建模、统计与生物信息学院的统计咨询师，同时也是一名R程序员。在获得生物学硕士学位后，他从事了6年与环境研究和管理相关的工作，之后获得了统计数据分析的高级硕士学位。乔里斯为学院的具体项目编写了若干代码包，还完成了一些统计方法的通用实现。另外，他也是R-Forge上几个包的维护者，并且还与人一起发表了若干篇统计专业方面的科学论文。

献词

这本书送给我的妻子，Annemarie，谢谢她的鼓励、支持和耐心；同时送给我9岁的外甥女Tanya，她数学学得非常棒，一直很负责地提醒我本书的截稿日快要到了！

——安德里·德弗里斯

献给我的母亲，因为是她成就了现在的我；同时送给我的外婆，因为她太酷了！

——乔里斯·梅斯

致谢

本书能够得以面世，完全得益于Wiley出版社编辑团队的大力支持。我们尤其要感谢Elizabeth Kuball，感谢她细致而耐心的编辑和温柔的催稿。还有Sara Shlaer，她总是假装不知道我们已经错过最后的交稿期限了。谢谢Kathy Simpson教我们如何用for Dummies系列的风格进行写作，谢谢Chris Katsaropoulos帮助我们迈出第一步。

感谢我们的技术编辑Gavin Simpson，他非常仔细地审阅了本书的技术细节并给出了很多建议。

谢谢R语言核心开发团队，谢谢你们创造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，谢谢你们创建并编写了数以千计的代码包、博客文章，并帮助大家解决各种疑难问题。

在本书中，我们使用了若干个由Hadley Wickham编写的代码包，他对ggplot图形和plyr等代码包的贡献至今仍为人所称道。

在编写本书的过程中，我们收获了大量来自R tag和Stack Overflow的贡献者的帮助和支持。在这里要感谢James(JD) Long、David Winsemius、Ben Bolker、Joshua Ulrich、Kohske Takahashi、Barry Rowlingson、Roman Luštrik、David Purdy、Nick Sabbe、Joran Elias、Brandon Bertelsen、 Michael Sumner、Ian Fellows、Dirk Eddelbuettel、Simon Urbanek、Gabor Grotendieck，以及所有一直努力让Stack Overflow成为更优秀的R资源社区的各位。

安德里说：感谢所有促成我做出现在这个选择的人们。来自伦敦商学院的Bruce Hardie教授在2004年的一场讲座上对R语言做了非常细致的点评，这是我第一次注意到R语言，他这种通过建立数据相关性用以辅助市场决策的程序性方法非常令人鼓舞。来自Logit研究院的Gray Bennett一直以来都是很棒的合作伙伴，他也给了我大量的建议。来自Kindle研究院的Paul Hutchings在我创业早期为我提供了很多帮助，在此一并感谢。

乔里斯说：感谢比利时根特大学数学建模、数据和生物信息学院的各位教授和我的同事们，感谢他们在本书撰写过程中所进行的深入而细致的讨论，以及对我个人的许多其他帮助。

译者序

对这个时代做这样两个定义应该是没有人反对的：这是一个大数据的时代，这是一个互联网的时代。

我们正在接触越来越多的数据，这些数据的多样性、即时性正在变得越来越强，而我们想要从数据中探索的现象背后的本质正在变得越来越深刻，种种因素的综合作用，使得数据统计、数据分析、数据呈现的工作挑战变得越来越大。

R语言是一门可以在数据领域对人们产生帮助的语言，这也是它近年来备受追捧的原因之一。R语言的发源可以追溯到S语言时代。20世纪80年代，S语言就开始应用于统计领域。R语言在很多语法上借鉴了S语言，但与S语言截然不同的一点在于，R语言是开源且免费的，这就是互联网时代最大的特征。开源社区拥有无限的可挖掘潜力，在享受他人成果的同时，你也能够做出自己的贡献。至于R语言可以在任何环境下运行、支持丰富的扩展、有活跃的社区可以获得帮助、能够与其他语言紧密连接、无需编辑直接运行等优势，都已经成了开源这个根本之上的锦上添花了。

本书从最基础的代码编辑器和每一个程序员最熟悉的“Hello World!”开始说起，介绍了学习R语言你必须要具备的基本知识，如何进行简单的算术，如何在R语言中操纵文本，如何使用因子进行分类，如何使用日期，如何处理以矩阵为代表的高维数据，R语言中的函数是怎样运作的，逻辑流又该如何控制。进阶部分的知识点开始直击R语言的最佳实践领域——统计分析。R语言中提供大量实用且高效的数据操作和数据处理功能，能够涵盖绝大部分数据处理需求，并且每分钟覆盖范围都在变得更为广泛和全面。最后一个部分介绍了R语言中的图形功能，这对于数据呈现有着很重要的指导意义。当然，还少不了实用性极佳的Tips。

作为一门大数据领域的互联网气质语言，R语言凭借自身独特的优势脱颖而出了，对于拥有同样特质的你，想必也会遵循同样的规律。

译　者

2015年3月

前　言

欢迎你翻开这本《R语言轻松入门与提高》，它将帮助你轻松地掌握R语言。虽然我们不能保证你在学完这本书后成为一名专家，但至少完成下面这几件事情是没有任何困难的。

- 使用各种强大的工具进行数据分析。
- 使用R完成统计分析和数据处理任务。
- 掌握基于向量的运算操作，用它来代替冗长的循环，从而加速计算。
- 理解下面这行代码的含义，并欣赏它的优美：

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

- 学会如何查找、下载并使用R开发人员社区中其他人贡献的代码。
- 了解获取其他帮助和资源的渠道，从而使你的R编程水平迈上新的台阶。
- 实现数据的可视化，绘制漂亮的数据图。

关于本书

《R语言轻松入门与提高》是关于R这门统计编程语言的入门书籍，我们会从最基本的概念入手，由浅入深，逐步讲解如何使用这门语言实现复杂而专业的数据处理和分析。

在这本书中，几乎每个知识环节都包含浅显易懂的示例，其中不仅有大量代码片段，还有许多完整的数据分析脚本，同时还留出了供你自由发挥的空间。

尽管本书不会对R语言的内部实现原理进行介绍，但对任何问题，不仅会讲解“如何实现”，还会说明“为什么如此”。如果你熟悉其他脚本语言的话，刚接触R时，可能会感到陌生而诧异，因为它的确提供了很多独一无二的功能和特性。所以我们不仅会教会你如何对R发布指令，还会告诉你R将怎样理解你对它说的“话”。在读完这本书后，你不仅可以用它来完成数据处理的任务，

还将具备继续深入学习R语言的能力。

R和RStudio

不管你用来运行R的操作系统是什么，Mac、Linux或者Windows，《R语言轻松入门与提高》这本书都适合你。

R不仅是一个应用程序，还是一门编程语言。在你下载R的时候，一个适用于你的操作系统的控制台程序也会一同被下载。不过，这个程序只具备一些基本的功能，并且在不同的操作系统中会略有差异。

RStudio是一个跨平台的应用程序，提供了许多支持R语言的特性。在这本书中，我们不建议你使用前面提到的控制台程序，由于RStudio在不同平台下提供了统一的操作界面，并且能够方便而快速地运行，因此，我们将使用RStudio而非操作系统相关的各种编辑器来介绍各种R语言的概念、演示各类示例程序。

本书也是一本参考手册，所以你并不需要从头读到尾，而是可以通过目录快速地找到感兴趣的话题。除此之外，在本书的各个章节中，如果有其他章节涉及的内容，我们还会给出交叉引用供你深入阅读和参考。

本书约定

书中的代码片段都会呈现为下面这段“投骰子”代码的样式。

```
> set.seed(42)
> throws <- 1e6
> dice <- sapply(1:2,
+     function(x)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
```

上面这个片段中的每行代码前面都带有一个提示符，总共有以下两种。

- >：提示符“>”并不是代码的一部分，所以在你输入代码时不需要输入它。
- +：“+”是接续符，表示该行与上一行是同一行代码。事实上，你自己在编写代码的时候并不需要这么做，本书中这样写是因为考虑到本书在排版印刷后代码的可读性。

前面不包含提示符或接续符的代码为R产生的输出。例如，上面这段代码显示的结果就是在投100万次骰子后，两个骰子点数之和为2～12的情况分别出现的次数统计，比如和为2的情况出现了28 007次。

你可以把这段代码原模原样地输入到R中运行，不过要注意以下3点：

- 不要输入提示符“>”；
- 不要输入接续符“+”；
- 只要不在关键字内部，你可以在任何地方输入空格或制表符，但要注意换行。

在这本书的几乎每个章节中，我们都会编写R代码；在大部分情况下，都会以交互式的方式执行R程序。所以，你需要学会分辨哪些是代码、哪些是程序运行的结果。当需要你在R控制台中输入命令时，前面会出现命令提示符“>”，如下所示：

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

如果你把上面这行代码输入R控制台然后按下回车键，R就会显示下面这行文字：

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

为了简便，一般会将上面两行输入和输出代码合在一起，用一个代码框来表示，如下所示：

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

所以，对这段代码的完整表述为：在控制台输入命令（`print("Hello world!")`），然后R给出输出结果（`[1] "Hello world!"`）。

最后一点，由于R语言的许多关键字都借用英文来表达，因此为了避免混淆，我们将R的函数、参数等关键字都用等宽字体来表现。例如，要创建一张数据图，可以使用R语言的`plot()`函数。在讨论函数时，函数的名称后面始终都会加上一对小括号——比如这里的`plot()`。除非特别有必要，在正文中，我们一般不会在函数名称后面的小括号内添加参数等额外信息。

有时，书中还会介绍菜单命令，如File⇨Save，它表示打开File菜单，然后选择Save命令。

可以略过的内容

你可以以最符合你自身需求的方式阅读本书，如果时间仓促（或者你不关心太多技术细节）的话，可以放心跳过文中标注了“Technical Stuff”的内容。你还可以略过补充内容（放在灰色背景中的文字），虽然其中介绍了不少有趣的东西，但对你把握整体知识没有任何影响。

读者水平要求

本书对于你和你的计算机有如下要求。

- 熟悉一般的计算机操作。比如知道如何下载和安装软件，如何利用英特网检索需要了解的信息。当然，你的计算机必须连接网络。
- 你可以不是一名程序员。不过，如果你的确是一名程序员，而且使用过其他语言编写代码，那么推荐你关注“Technical Stuff”中的内容，其中包含许多R语言与其他常见语言之间的异同点。
- 你可以不是一名统计学家，但需要了解基本的统计学知识。《R语言轻松入门与提高》不是一本统计学著作，但仍然涉及如何使用R来完成基本的统计操作。如果你对更深入的统计学感兴趣的话，不妨阅读Deborah J. Rumsey博士编写的《Statistics For Dummies, 2nd Edition》一书（Wiley出版）。
- 拥有学习新知识的动力。你喜欢解决问题，并且不惧怕看到枯燥的R控制台界面。

本书导览

本书由6个部分构成，下面是每个部分所包含的内容。

第一部分：R You Ready?

在这部分中，我们将向你介绍R，并引导你编写第一个脚本。在此期间，你将接触向量这个强大的概念，它能一次性实现对多个变量的计算。你还将使用R Workspace（换句话说，学会如何创建、修改和删除变量），并掌握如何保存你的工作，供以后载入并继续编写。另外，我们还会介绍R的一些基础知识（例如如何通过安装扩展包来增强R的功能）。

第二部分：开始使用R

在这部分中，我们介绍了有关数据读取、写入以及算术的问题——换句话说，使用文本和数字（包括日期）。

另外，这部分还会介绍列表和数据框架这两种重要的数据结构。

第三部分：在R中编程

R是一门编程语言，所以你需要了解如何编写并理解函数。在这部分中，我们将介绍如何完成相关工作，包括使用if语句控制脚本的逻辑流，以及通过循环实现重复动作的执行。另外，这部分还会介绍如何处理代码中产生的错误和警告。最后，我们会为你介绍几个工具，它们可以帮助你完成调试工作。

第四部分：让数据说话

在这部分中，我们将介绍可以在R中使用的不同类型的数据结构，如列表和数据框架。你将学会如何将数据传入R以及如何从R中取出数据（例如，从文件或剪贴板读取数据）。另外，你还会掌握如何与其他应用程序进行交互，如Microsoft Excel。

然后，你将学习到一些高级的数据操作，并发现这是一件非常简单的事情。我们将向你介绍如何选择一个数据的子集，并对它们进行排序。我们还会介绍如何基于相似列进行数据集合并操作。最后，你还会学习到如何对数据子集应用函数操作，以实现数据的分离与合并。当你理解了这一方法后，可以不断地重复使用，只需要几步就可以完成非常专业的数据分析操作。

我们很想向你展现使用R进行统计分析的方法，毕竟这是R的精髓，不过我保证会让内容尽量简单。在读完这部分之后，你会学习到如何使用R来描述并汇总变量和数据。另外，你还会掌握一些经典的测试（例如，计算t-test），并学会如何使用随机数来模拟一些分布。

最后，我们将向你展示使用线性模型的基础（例如，线性回归和方差分析）。更进一步，在取得了数据模型之后，你还会学习到如何使用R来进行数据预测。

第五部分：绘制数据图

一图胜千言，所以，在这部分中，我们将教你如何用图来向别人展现数据结

果。你会学习如何创建简单和复杂的图形（Plot），以可视化地查看数据。我们将从柱状图和折线图开始，向你展示如何使用小平面（Facet）来呈现数据。

第六部分：20条有用的建议

在这部分中，我们将告诉你10件可以在R中完成，但无法通过Microsoft Excel来实现的事情。另外，我们还为你提供了使用非基础R包的10条建议。

本书使用的图标

在阅读这本书的过程中，你一定会注意到页面边缘处的图标，它们标记了具有某些特殊含义的文字：

当你看到这样的“TIP”图标时，可以肯定，它会告诉你完成某项操作更简单、更快捷的方法。

显然，你不可能完全记住整本书的内容，但是，当你看到“REMEMBER”图标时，表示一定要将相关内容记在脑海中。通常它所指示的都是我们会反复遇见和使用的设计模式或者原则。

当你看到“WARNING”图标时，一定要提高警惕，它告诉你哪些事情是千万不能做的。虽然使用R并不会导致真正的灾难发生，但“WARNING”图标往往意味着一不小心就会给自己增添很多麻烦。

“TECHNICAL STUFF”图标指示的是一些技术细节，如果你不感兴趣的话，可以放心跳过。虽然我们尽力让这些内容显得不那么枯燥无趣，但如果你的时间有限，或者只想学习重要而关键的知识，也可以略去不看。

接下来该怎么做

学习R语言只有一条路径：用它！尽管在这本书中，我们会尽力让你熟悉R的使用，但你最好还是坐在计算机旁亲身体验和尝试。把书放到一边扣在桌上，然后开始敲键盘吧！

目　录