大数据人才培养规划教材

以解决实际问题为**学习目标**
以实战案例贯穿为**学习手段**

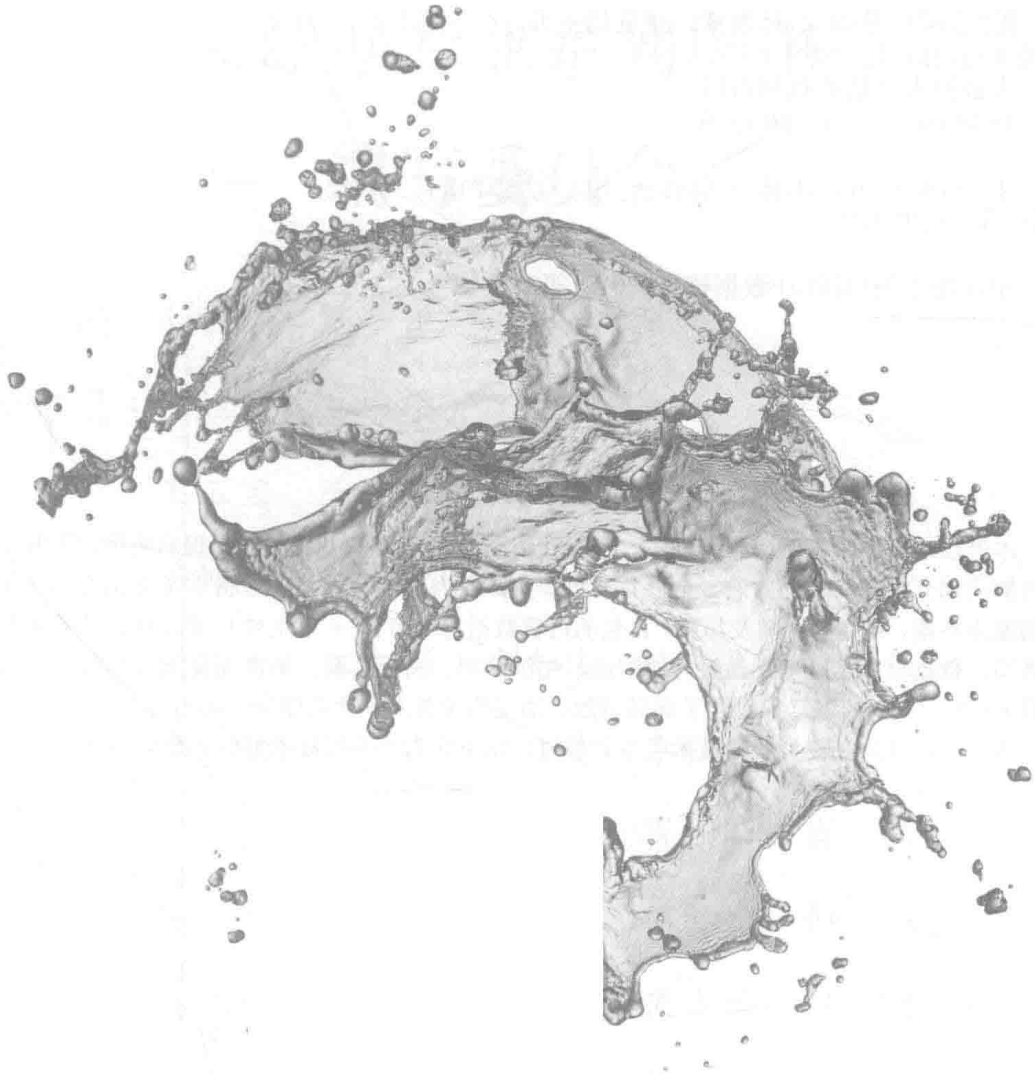
R语言编程基础

R Programming

林智章 张良均 ◉ 主编
李博文 杨惠 麦国炫 ◉ 副主编

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

大数据人才培养规划教材



R语言编程基础

R Programming

林智章 张良均◉主编
李博文 杨惠 麦国炫◉副主编

人民邮电出版社
北京

图书在版编目（CIP）数据

R语言编程基础 / 林智章，张良均主编. -- 北京 : 人民邮电出版社，2019.1
大数据人才培养规划教材
ISBN 978-7-115-49611-9

Ⅰ. ①R… Ⅱ. ①林… ②张… Ⅲ. ①程序语言－程序设计 Ⅳ. ①TP312

中国版本图书馆CIP数据核字(2018)第273148号

内 容 提 要

本书以理论结合示例操作的方式，全面介绍了R语言编程基础及其知识的应用，讲解了利用R语言解决部分实际问题的方法。全书共7章：第1章为R语言概述，包括学习R语言的优势、R语言的编译环境、R包的获取及加载、R包的内置数据等；第2～6章主要介绍R语言的数据对象与数据读写、数据集基本处理、函数与控制流、初级绘图、高级绘图；第7章主要介绍可视化数据挖掘工具Rattle。本书的每章都包含了课后习题，通过练习帮助读者巩固所学的内容。

本书可以作为高校大数据技术类专业教材，也可作为大数据技术爱好者自学用书。

◆ 主　　编　林智章　张良均
副 主 编　李博文　杨　惠　麦国炫
责任编辑　左仲海
责任印制　马振武

◆ 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
三河市君旺印务有限公司印刷

◆ 开本：787×1092　1/16
印张：16.25　　2019年1月第1版
字数：368千字　　2019年1月河北第1次印刷

定价：49.80元

读者服务热线：(010)81055256　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

大数据专业系列图书
编写委员会

序 PREFACE

随着大数据时代的到来，移动互联网络和智能手机迅速普及，多种形态的移动互联应用蓬勃发展，电子商务、云计算、互联网金融、物联网等不断渗透并重塑传统产业，大数据当之无愧地成了新的产业革命核心。

未来 5～10 年，我国大数据产业将会是一个飞速发展时期，社会对大数据相关专业人才有着巨大的需求。目前，国内各大高校都在争相设立或准备设立大数据相关专业，以适应地方产业发展对战略性新兴产业的人才需求。

人才培养离不开教材，大数据专业是 2016 年才获批的新专业，目前还没有成套的系列教材，已有教材也存在企业案例缺失等亟须解决的问题。由广州泰迪智能科技有限公司和人民邮电出版社策划，校企联合编写的这套图书，犹如大旱中的甘露，可以有效解决高校大数据相关专业教材紧缺的困难。

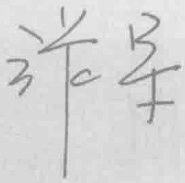
实践教学是在一定的理论指导下，通过引导学习者的实践活动，从而传承实践知识、形成技能、发展实践能力、提高综合素质的教学活动。目前，高校教学体系的设置有诸多限制因素，过多地偏向理论教学，课程设置与企业实际应用契合度不高，学生无法把理论转化为实践应用技能。课程内容设置方面看似繁多又各自为“政”，课程冗余、缺漏，体系不健全。本套图书的第一大特点就是注重学生的实践能力培养，根据高校实践教学中的痛点，首次提出“鱼骨教学法”的概念。以企业真实需求为导向，学生所学技能紧紧围绕企业实际应用需求，将学生需掌握的理论知识，通过企业案例的形式进行衔接，达到知行合一、以用促学的目的。

大数据专业应该以大数据技术应用为核心，紧紧围绕大数据应用闭环的流程进行教学，才能够使学生从宏观上理解大数据技术在行业中的具体应用场景及应用方法。高校现有的大数据课程集中在如何进行数据处理、建模分析、参数调整，使得模型的结果更加准确。但是，完整的大数据应用却是一个容易被忽视的部分。本套图书的第二大特点就是围绕大数据应用的整个流程，从数据采集、数据迁移、数据存储、数据

分析与挖掘，最终到数据可视化，覆盖完整的大数据应用流程，涵盖企业大数据应用中的各个环节，符合企业大数据应用真实场景。

我很高兴看到这套书的出版，也希望这套书能给更多的高校师生带来教学上的便利，帮助读者尽快掌握本领，成为有用之才！

教育部长江学者特聘教授
国家杰出青年基金获得者
IEEE Fellow
华南理工大学计算机与工程学院院长



2017 年 12 月

前 言 FOREWORD

随着云时代的来临，数据分析技术将帮助企业在合理时间内获取、管理海量数据，为企业经营决策提供积极的帮助。数据分析作为一门前沿技术，广泛应用于物联网、云计算、移动互联网等领域。虽然大数据目前在国内还处于初级阶段，但是其商业价值已经显现出来，特别是有实践经验的数据分析人才，更是各企业争夺的热门。为了满足日益增长的数据分析人才需求，很多高校开始尝试开设数据分析课程。“数据分析”作为大数据时代的核心技术，必将成为高校大数据相关专业的重要课程之一。

本书特色

本书定位于R大数据基础教材，深入浅出地介绍R语言编程基础的相关知识，包括R语言概述、数据对象与数据读写、数据集基本处理、函数与控制流、初级绘图、高级绘图。本书涉及的知识点简要精到，实践操作性强，能对R语言编程基础的学习、理解及应用提供有效的指导。

本书采用了理论结合示例操作的模式，按照解决实际问题的思路，逐步展开相关的理论知识点。全书大部分章节紧扣示例操作，不堆积知识点。通过从理论到示例操作的一系列体验，读者真正理解、掌握R语言的编程基础。

本书适用对象

（1）开设“数据分析”课程的高校教师和学生

目前，国内不少高校将数据分析引入教学中，在计算机、数学、自动化、电子信息、金融等专业开设了数据分析相关的课程，但目前这一课程教学的相关教材没有统一，有些高校使用SPSS、SAS等传统统计工具，并没有使用R语言作为数据分析工具。本书提供了R语言相关技术的介绍、原理、实践等，能有效指导高校教师和学生使用R语言解决企业实际问题，为以后的工作打下良好基础。

（2）数据分析开发人员

这类人员可以在理解数据分析、应用需求和设计方案的基础上，结合书中提供的R语言的使用方法快速实现数据分析应用编程。

（3）进行数据分析应用研究的科研人员

许多科研院所为了更好地对科研工作进行管理，纷纷开发了适应自身特点的科研业务管理系统，并在使用过程中积累了大量的科研数据。R语言可以提供一个优异的

环境对这些数据进行分析应用。

（4）关注高级数据分析的人员

R 语言作为一款专业的数据分析软件，能为数据分析人员提供可靠的依据。

代码下载及问题反馈

读者可登录人民邮电出版社教育社区（www.ryjiaoyu.com）或“泰迪杯”全国数据挖掘挑战赛网站（http://www.tipdm.org/tj/1309.jhtml）下载书中全部示例的数据文件及源代码。另外，为方便教师授课，我们还提供了 PPT 课件，读者可以从“泰迪杯”数据挖掘挑战赛网站（http://www.tipdm.org/tj/840.jhtml）下载并填写申请表，填写后发送至指定邮箱；其他图书资源，读者可通过热线电话（40068-40020）或以下微信公众号咨询。

我们已经尽最大努力避免在文本和代码中出现错误，但是由于水平有限，编写时间仓促，书中难免会有一些不足和疏漏之处。如果您有更多的宝贵意见，欢迎发送邮件至邮箱 13560356095@qq.com，期待能够得到您真挚的反馈。同时，本书更新内容会及时在“泰迪杯”全国数据挖掘挑战赛网站上发布，读者可以登录网站或关注泰迪大数据挖掘微信公众号查阅相关信息。

编　者

2018 年 7 月

目录 CONTENTS

第1章 R语言概述

R语言是一个体系庞大的应用软件，主要包括核心的R标准包和各专业领域的其他包。本书采用原理加示例的方式来对R语言相关函数进行介绍。本章主要对R语言的基本信息、R软件和RStudio的安装及升级、常用包的安装与加载，以及R包内置数据集进行简单介绍。

学习目标

（1）认识并安装R语言。
（2）认识R的编译环境。
（3）认识R包，并掌握R包的安装与加载方法。
（4）了解R语言的内置数据集。

1.1 认识R语言

本节主要介绍R语言的基本信息，如何下载R语言，以及如何在自己的计算机上实现安装。安装成功后，将介绍R语言的编辑窗口。

1.1.1 R语言的基本信息

R语言是一种为统计计算和图形显示而设计的语言环境，是贝尔实验室（Bell Laboratories）的Rick Becker、John Chambers和Allan Wilks开发的S语言的一种实现，提供了一系列统计和图形显示工具。R语言是面向对象的一种编程语言，也是一套开源的数据分析解决方案，由一个庞大且活跃的全球性研究型社区维护。它具有下列优势。

（1）R语言是完全免费的统计分析软件，可以在不同的平台上运行，包括Windows、UNIX、Mac OS和Linux。

（2）R语言可以轻松地从各种类型的数据源读写数据，包括带分隔符的文件、统计软件、数据库管理系统，以及专门的数据仓库。几乎所有类型的数据都可以用R语言进行分析统计。

（3）R语言的优势主要体现在其软件包生态系统具有较高的开放性（即免费开源）。R语言不仅提供功能丰富的内置函数供用户调用，也允许用户编写自定义函数来扩充功能。读者无须申请权限即可直接查看软件包或程序包的源码，并且对其进行拓展。如果某项统计技术已经存在，那么必然存在着一款R软件包与之对应。

（4）R 语言具有顶尖水准的制图功能。R 语言的拓展包 dplyr 与 ggplot2 可分别用于数据处理与绘图，且能够非常直观地提升用户对数据的理解。

图 1-1 所示是信用卡客户经济情况分布的直方图，展示了 R 语言的绘图能力。该图用来分析信用卡客户的个人月开销、月刷卡额、个人月收入和家庭月收入等变量。由图 1-1 可知，信用卡客户的个人月开销主要集中在 1 万元以下和 1 万元至 2 万元之间；多数客户的月刷卡额在 2 万元至 8 万元之间；个人月收入中有 1/3 左右的客户无收入，其余客户个人月收入主要集中在 2 万元至 4 万元之间，4 万元以上的占少数；家庭月收入为 2 万元至 4 万元的客户尤为突出，说明大部分客户的家庭经济水平中等。

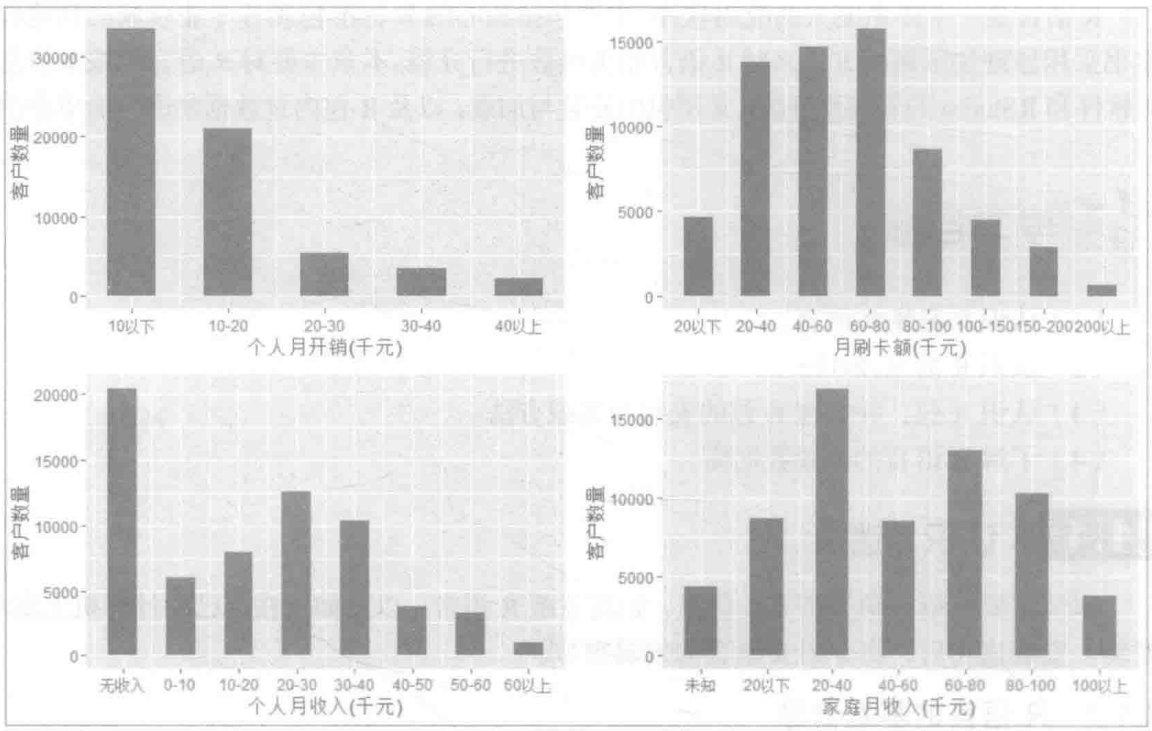


图 1-1　信用卡客户经济情况分布直方图

第 5 章及第 6 章将继续讨论这些图形，介绍更多的 R 语言在图形展示方面的强大功能，让用户以简单方便的方式创建优雅、信息丰富、高度定制的专业图形。

1.1.2　获取与安装 R 语言

本书使用的 R 版本为 R 3.4.2。根据操作系统不同，读者可选择安装 64 位或 32 位版本。读者安装时直接运行下载的 R-3.4.2-win.exe。Linux、Mac OS X 和 Windows 都有相应的编译好的二进制版本，读者根据所选择平台的安装说明进行安装即可。

这里以在 Windows 操作系统下安装 R 为例，操作步骤如下。

（1）打开浏览器访问 R 的官网 http://www.r-project.org/，如图 1-2 所示。

（2）单击“Download”栏目下的“CRAN”，即跳转到 R 综合资料网（Comprehensive R Archive Network，CRAN）的路径上，如图 1-3 所示。

从镜像路径中选择 China 栏目（如图 1-4 所示）下的任意一个链接，单击进入 R 的下载界面，如图 1-5 所示。

[name]

Download
CRAN

R Project
About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

R Foundation
Foundation
Board
Members
Donors
Donate

The R Project for Statistical Computing

Getting Started

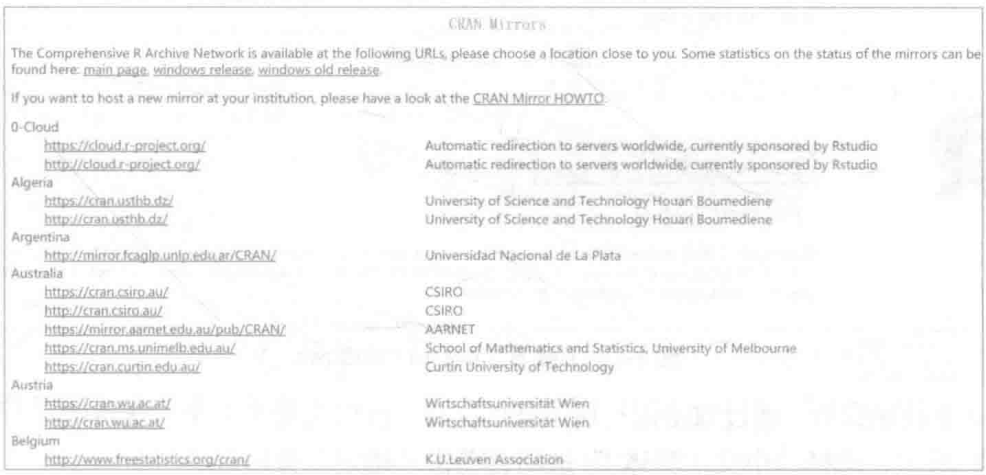
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 3.4.2 (Short Summer)** has been released on Thursday 2017-09-28.
- **The R Journal Volume 9/1** is available.
- **R version 3.3.3 (Another Canoe)** has been released on Monday 2017-03-06.
- **The R Journal Volume 8/2** is available.
- **useR! 2017** (July 4 - 7 in Brussels) has opened registration and more at http://user2017.brussels/
- Tomas Kalibera has joined the R core team.
- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kalibera, and Balasubramanian Narasimhan.
- **The R Journal Volume 8/1** is available.
- The **useR! 2017** conference will take place in Brussels, July 4 - 7, 2017.

图 1-2 R的官网



CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

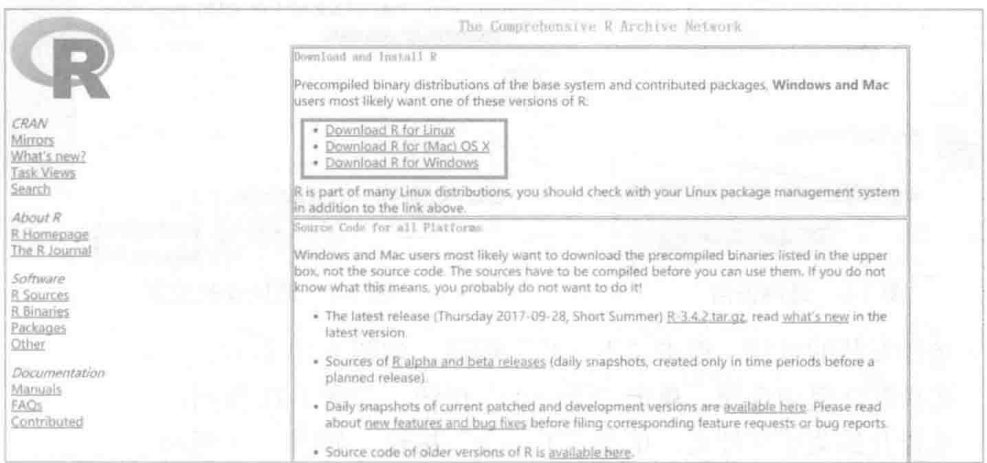
If you want to host a new mirror at your institution, please have a look at the CRAN Mirror HOWTO.

0-Cloud	
https://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
http://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
Algeria	
https://cran.usthb.dz/	University of Science and Technology Houari Boumediene
http://cran.usthb.dz/	University of Science and Technology Houari Boumediene
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
https://cran.csiro.au/	CSIRO
http://cran.csiro.au/	CSIRO
https://mirror.aarnet.edu.au/pub/CRAN/	AARNET
https://cran.ms.unimelb.edu.au/	School of Mathematics and Statistics, University of Melbourne
https://cran.curtin.edu.au/	Curtin University of Technology
Austria	
https://cran.wu.ac.at/	Wirtschaftsuniversität Wien
http://cran.wu.ac.at/	Wirtschaftsuniversität Wien
Belgium	
http://www.freestatistics.org/cran/	K.U.Leuven Association

图 1-3 R的下载镜像路径

China	
https://mirrors.tuna.tsinghua.edu.cn/CRAN/	TUNA Team, Tsinghua University
http://mirrors.tuna.tsinghua.edu.cn/CRAN/	TUNA Team, Tsinghua University
https://mirrors.ustc.edu.cn/CRAN/	University of Science and Technology of China
http://mirrors.ustc.edu.cn/CRAN/	University of Science and Technology of China
https://mirrors.eliteu.cn/CRAN/	Elite Education
https://mirror.lzu.edu.cn/CRAN/	Lanzhou University Open Source Society
http://mirror.lzu.edu.cn/CRAN/	Lanzhou University Open Source Society

图 1-4 R的China下载镜像路径



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Thursday 2017-09-28, Short Summer) R-3.4.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.

图 1-5 R的下载界面

（3）如果是第一次安装 R 语言，单击“base”项目，如图 1-6 所示。进入 R 的下载页面，单击“Download R 3.4.2 for Windows”链接（如图 1-7 所示），即可下载相应版本的 R 语言。

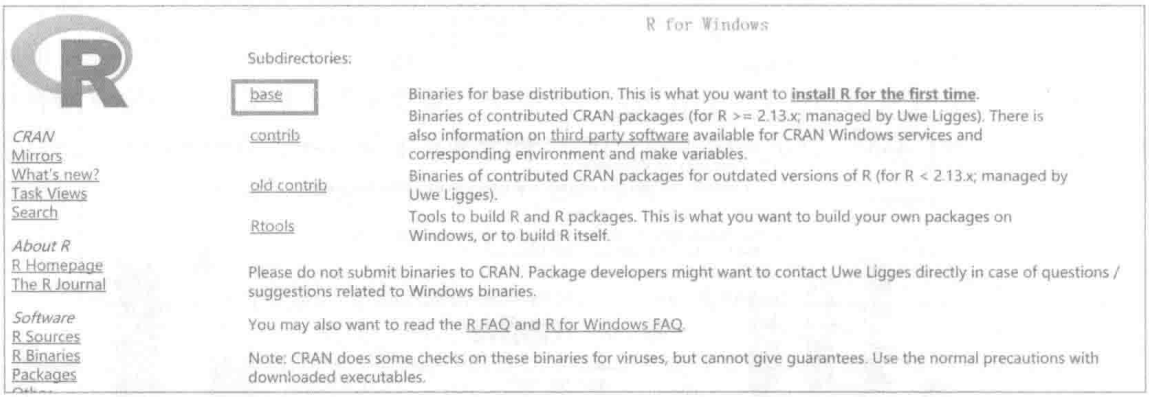


图 1-6　base 项目

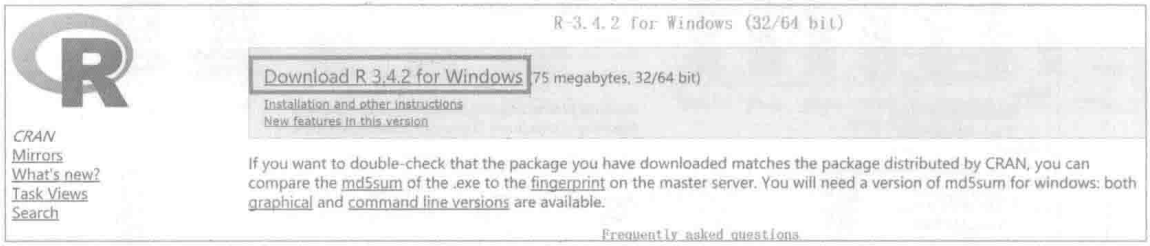


图 1-7　下载 R 3.4.2 for Windows

（4）下载完成后，通过双击运行所下载的文件，此时会弹出一个“选择语言”对话框，如图 1-8 所示，选择“中文（简体）”选项，单击“确定”按钮。

（5）弹出安装向导后，根据指示不断单击“下一步”按钮，直到出现图 1-9 所示的界面，选择软件的安装位置，单击“下一步”按钮。

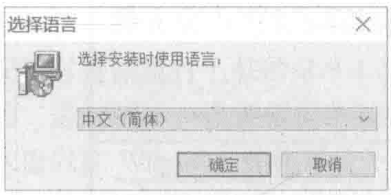


图 1-8　选择语言

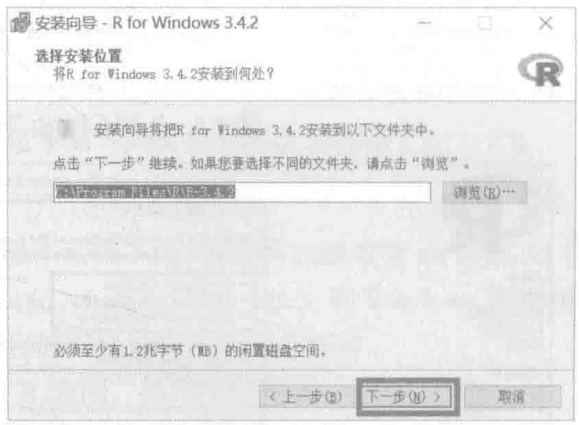


图 1-9　选择安装位置

（6）选择安装的组件，单击“下一步”按钮，如图 1-10 所示。

（7）选择默认启动选项，单击“下一步”按钮，如图 1-11 所示。

（8）选择开始菜单文件夹，单击“下一步”按钮，如图 1-12 所示。

（9）选择附加任务，如添加快捷方式等，单击“下一步”按钮，如图 1-13 所示。

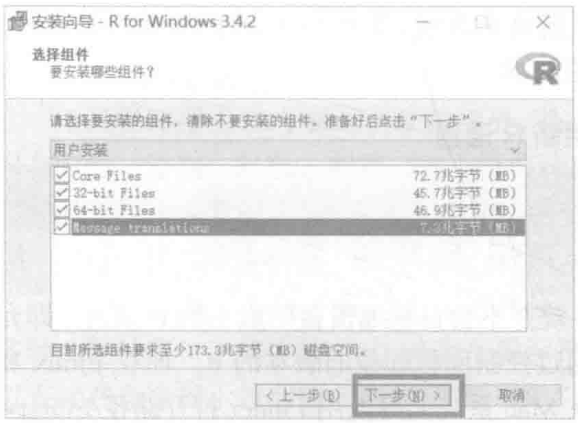


图 1-10 选择安装组件

图 1-11 选择启动选项

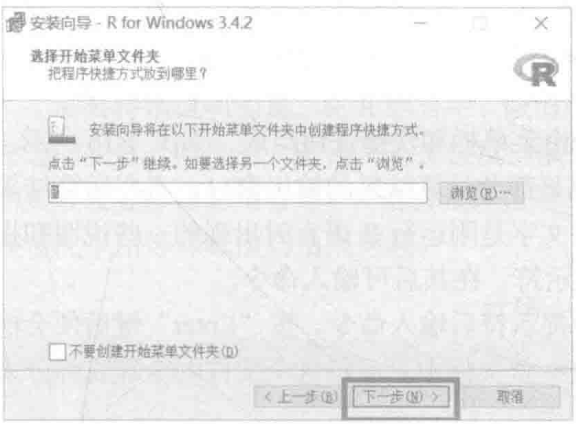


图 1-12 选择开始菜单文件夹

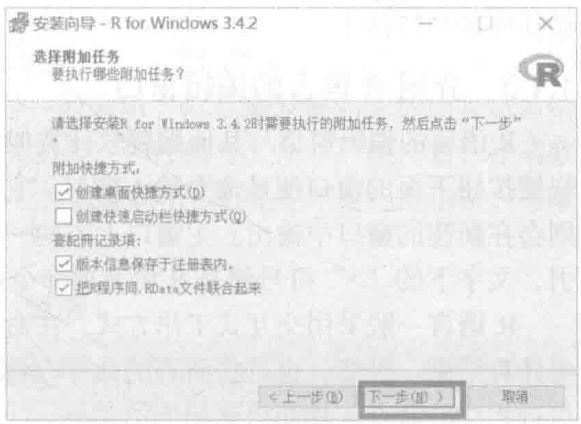


图 1-13 选择附加任务

（10）安装完成，单击“结束”按钮，如图 1-14 所示。此时安装完成。

（11）安装好 R 语言后，单击安装目录中 bin 目录下的 Rgui.exe 文件启动 R 语言（或者双击桌面快捷方式打开），打开的界面如图 1-15 所示。

图 1-14 安装完成

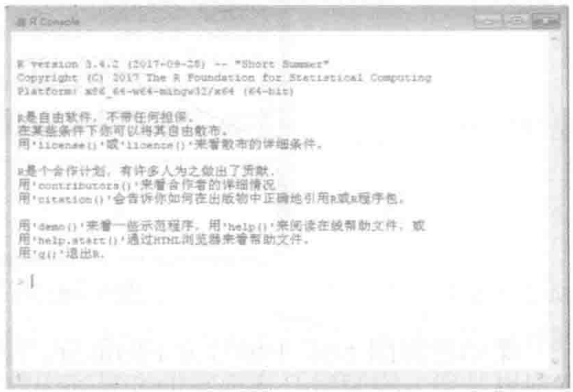


图 1-15 R 3.4.2 的初始界面

R 语言的升级通常是通过从 CRAN（http://cran.r-project.org/bin/）上下载和安装最新版的 R 语言来实现。这种方式需要重新设置各种自定义选项，包括之前安装的扩展包。可以将 R 安装目录下 etc 文件夹中的 Rprofile.site 文件及 R 安装目录下的 library 文件夹保存到其他地方，待安装新版本的 R 语言后再移动到相应的位置进行覆盖。

在 Windows 系统上，有一种更加方便的更新 R 的方式，如代码 1-1 所示。输入代码后，按照提示即可很方便地将 R 语言升级至最新的版本。

代码 1-1　更新 R 语言

```
> install.packages("installr")
> require(installr)          #load / install+load installr
> updateR()
```

安装了最新版本的 R 语言后，Windows 系统并不会自动地覆盖旧版本的 R 语言，即允许系统中存在多种版本的 R 语言，此时可以通过控制面板卸载旧版本的 R。而在 Linux 和 Mac 系统上，新版的 R 语言会覆盖老版本。在 Mac 系统上可以用 Finder 打开路径/Library/Frameworks/R.frameworks/versions/，删除其中旧版本的文件夹。在 Linux 系统上，不需要做任何额外的操作。

1.1.3　介绍 R 语言的编辑窗口

R 语言的编辑窗口与其他编程软件类似，由菜单栏和快捷按钮组成，如图 1-16 所示。快捷按钮下面的窗口便是命令输入窗口，它也是部分运算结果的输出窗口，有些运算结果则会在新建的窗口中输出。主窗口上方的一些文字是刚运行 R 语言时出现的一些说明和指引，文字下的“>”符号便是 R 语言的命令提示符，在其后可输入命令。

R 语言一般采用交互式工作方式，在命令提示符后输入命令，按“Enter”键后便会输出计算结果。当然，也可将所有的命令存储在一个文件中，运行这个文件的全部或部分来执行相应的命令，从而得到相应的结果。

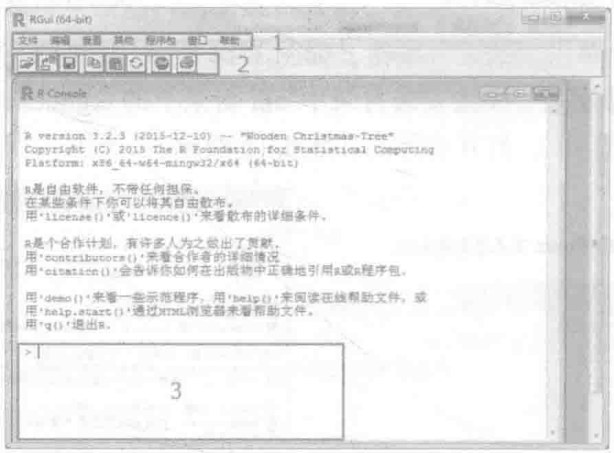


图 1-16　R 3.4.2 操作界面

菜单栏即图 1-16 中标号为 1 的部分，位于工作环境的最上方。文件菜单可以实现的功能有输入 R 语言代码、建立新的程序脚本、打开程序脚本、显示文件、载入工作空间、保存工作空间、载入历史、保存历史、改变当前目录、打印、保存到文件及退出。编辑菜单可以实现复制、粘贴、清除控制台和数据编辑等功能。查看菜单可以选择是否显示工具栏。其他菜单可以实现中断目前计算、缓冲输出及列出目标对象等功能。程序包菜单可以实现载入程序包、设定 CRAN 镜像、安装及更新程序包等功能。窗口菜单可以将所有窗口层叠或者平铺。帮助菜单提供 R 语言的常见问答和帮助途径。当执行不同的窗口操作时，菜单