数据科学与大数据技术系列

R语言：大数据分析中的统计方法及应用

薛　薇◎著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

数据科学与大数据技术系列

本成果受到中国人民大学2018年度“中央高校建设世界一流大学（学科）和特色发展引导专项资金”支持

R语言：大数据分析中的统计方法及应用

薛 薇◎著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

大数据分析，其学习起点应是大数据的统计分析；大数据分析，其学习特点应是案例化、工具化和业务导向化。本书面向大数据分析实践，基于大数据案例，以问题为线索，以解决问题为导向讲解统计方法及 R 语言实现；突出大数据应用特色，兼顾统计方法的经典性和普适性、理论讲解的通俗性和严谨性、R 语言代码的实操性和示范性。本书提供配套全部案例数据及各章节 R 语言程序代码，可登录华信教育资源网 www.hxedu.com.cn 免费下载。

本书可作为大数据相关专业、统计学专业及其他有关专业的本科生或硕士研究生数据分析课程的教材，也可作为从事大数据分析工作人员的参考用书。

图书在版编目(CIP)数据

R 语言：大数据分析中的统计方法及应用 / 薛薇著. —北京：电子工业出版社，2018.7

ISBN 978-7-121-33915-8

I. ①R… II. ①薛… 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 060949 号

策划编辑：秦淑灵
责任编辑：秦淑灵
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1092　1/16　印张：15　字数：384 千字
版　　次：2018 年 7 月第 1 版
印　　次：2018 年 7 月第 1 次印刷
定　　价：48.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：qinshl@phei.com.cn。

前　言

大数据时代，数据是生产资料，计算是生产力，互联网是生产关系，而数据分析就是串联各个生产要素的基本生产方式。

目前比较有代表性的大数据定义，来自麦肯锡全球研究院(McKinsey Global Institute)、高德纳公司(Gartner)和IBM公司等先行研究机构的综合观点。从狭义角度来讲，大数据是一个具有5V特征的大规模数据集合。5V即海量的数据规模(Volume)、快速流转且动态激增的数据体系(Velocity)、多样异构的数据类型(Variety)、潜力大但密度低的数据价值(Value)，以及受噪声影响的数据质量(Veracity)。而从广义角度来讲，大数据的概念还应包含大数据的理论、技术、应用和产业生态这四个基本范畴。

近年来，我国大数据事业迅猛发展，大数据人才的需求与培养也日趋紧迫。全国高校“大数据技术与应用”和“数据科学与大数据技术”专业建设不断升温。一般我们可将大数据技术概括为两大方向：一是大数据工程，二是大数据分析，并分别对应着大数据工程师和大数据分析师这两个角色。总体而言，随着大数据系统架构和基础设施的不断完善和普及，以大数据工程为核心的相关项目终究是有限的。而随着移动互联网和物联网的广泛应用，以及各方对精细化管理、个性化营销和智能化决策的渴望，大数据分析将不断深入到各行各业，大数据分析人才的需求也必将呈现出长期性、有规模的增长态势。

数据分析的理论发展和实践经验都证明，掌握大数据分析，其学习起点应是大数据的统计分析。进一步，我们认为，学习大数据的统计分析应面向市场需求、面向实际应用，所以应具有以下三个特点。

第一，要结合大数据分析的实际案例。

面对“5V俱全”的大数据体系，许多经典的统计分析方法仍然有效，是我们分析问题、解决问题的可靠手段，但需要突破那种“小样本、习题式”的传统学习模式，要精挑有针对性的大数据集合，细选有说明性的大数据案例，以这些数据和案例为引导，有条理地形成分析思路，并贯穿整个学习过程，从而真正实现由表及里、深入浅出的学习体验。

第二，要结合大数据分析的应用工具。

大数据的统计分析应进一步突破“重理论讲解，重公式推导，轻技能培养，轻工具实现”的传统学习模式，要将各个知识点言简意赅地阐述透彻，同时也要同步掌握一个有效的软件工具，进而可对相应的数据与案例进行实操破解。

第三，要结合大数据分析的目标导向。

大数据的统计分析应进一步突破“方法导向”的传统学习模式，应围绕大数据案例，确定分析目标，细化研究问题，明确分析思路，并以业务问题为出发点，形成以目标为导向的学习模式，努力培养大数据分析人才的数据敏感性，以及发现问题和运用恰当统计分析方法解决问题的能力。最终针对整个知识体系建立“问题→概念→方法→工具→结果→分析解释”一条龙式的学习模式。

本书正是结合上述三个特点而筹划推出的，具体表现在以下三个方面。

第一，选择典型的大数据分析案例。

选用三个典型的大数据案例贯穿全书，并提供数据集和分析程序的下载，主要内容为手机APP

美食餐馆食客点评数据、北京市空气质量监测数据、超市顾客购买行为数据等。这些案例具有大数据分析应用的代表性，而且业务问题直观明了，数据含义通俗易懂。一方面使读者能够直接感知大数据处理规模，另一方面也可有效避免由于专业领域不同而带来的数据理解问题。

第二，选择开源的大数据分析工具 R 语言。

选用 R 语言作为大数据分析工具。从分析工具的方法覆盖全面性、学习难易程度、使用流行性、未来发展潜力和开源性等多方面考虑，R 语言都是进行大数据统计分析的最恰当工具。

第三，设计并提出研究问题和分析思路。

本书在每章开篇，均首先围绕大数据案例提出若干分析需求的问题，同时提炼总结出这些问题的共性特征，进而提出可行的统计分析思路，建立学习途径；然后讨论方法原理，给出解决案例问题的 R 语言程序代码和详细的结果说明。

为确保内容的完整性和实用性，本书在大数据分析案例的选择、分析工具讲解的详略程度、以目标为导向的主流统计方法覆盖的全面性等方面，都进行了精心安排和综合设计。本书共 12 章。第 1 章在大数据基本定义的基础上，明确给出了本书的学习目标和定位。然后，对 R 语言的基本概念和入门知识进行了较为详尽的讲解。之后，提出了大数据的统计分析整体框架和思路，并基于大数据分析案例，对相关统计概念和内容进行了说明，旨在方便读者尽快明晰统计分析路线。数据组织是数据分析的基础，数据整理是数据分析不可或缺的必要环节。因此第 2 章和第 3 章直入主题，讨论了 R 语言的数据组织、整理以及编程基础，引入三个大数据分析案例并贯穿全书。大数据的统计分析起步于数据的基本分析，包括从单个变量分布特征到两个变量相关性的基本描述等，因此第 4 章和第 5 章首先基于大数据分析案例，提出了若干个基本数据分析问题，然后逐一讲解问题、阐述解决方法并给出 R 代码实现。第 6 章和第 7 章，继续针对大数据分析案例中更广泛的应用问题，细致地讨论了解决应用问题的诸多统计方法，包括单个总体的均值检验方法、两个及多个总体的均值对比方法和相应的 R 代码设计。第 8 章、第 9 章和第 11 章分别涉及线性回归分析、Logistic 回归分析和线性判别分析。这些分析方法均是当前大数据分析中应用极为广泛的主流核心方法，旨在探究影响因素，解决分类预测等问题。第 10 章的聚类分析关注数据分组，不仅普遍存在于大数据的一般统计分析中，也广泛拓展到了数据挖掘、机器学习等诸多领域。同时第 12 章的因子分析更是大数据特征工程中的最常用方法。

总之，作者希望为致力于大数据分析和 R 语言实践的初学者，奉献一本具有大数据统计分析应用特色、R 语言代码可操作性和示范性、统计方法经典性和普适性的优秀作品。本书提供配套的全部案例数据以及各章节 R 语言程序代码，可登录华信教育资源网 www.hxedu.com.cn 免费下载。本书可作为大数据相关专业、统计学专业及其他有关专业的本科生或硕士研究生数据分析的教材，也可作为从事大数据分析实际工作人员的参考用书。

本书写作过程中，杨志峰和陈笑语同学对北京市空气质量监测案例数据进行了整理并利用 R 完成了初步分析。林家嗣和王琪等同学收集整理了超市顾客购买行为案例数据和美食餐馆食客点评案例数据，完成了部分 R 代码的初步编写和调试工作。这里对他们为本书做出的贡献，一并表示诚挚的感谢！本成果受到中国人民大学 2018 年度“中央高校建设世界一流大学（学科）和特色发展引导专项资金”支持。书中不妥和错误之处，诚望读者不吝指正。

薛　薇

于中国人民大学应用统计科学研究中心

中国人民大学统计学院

目　录

第1章　R语言与统计分析概述

1.1　写在前面的话

1.1.1　大数据的广义概念

当前，大数据技术正以全面、快速、渗透式的发展态势席卷整个世界。其发展势头迅猛，发展潜力巨大，发展前景辉煌。如果说，大数据时代数据是生产资料，计算是生产力，互联网是生产关系，那么大数据分析就是串联各个要素的最重要的生产方式。

比较有代表性和权威性的大数据定义是来自麦肯锡全球研究院 McKinsey Global Institute、高德纳公司 Gartner 和 IBM 公司等先行研究机构和企业的综合观点。认为大数据首先是一个大规模的数据集合或信息资源。同时，这个数据集合具有典型的 5V 特征，即海量的数据规模(Volume)、快速流转且动态激增的数据体系(Velocity)、多样异构的数据类型(Variety)和潜力大但密度低的数据价值(Value)，以及噪声影响的数据质量(Veracity)。

事实上，大数据还有更为宽泛的广义概念，即大数据是围绕大数据集的一个包括大数据理论、技术、应用和生态四个方面的组合架构概念，如图 1-1 所示。

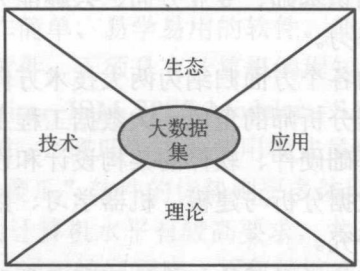

图 1-1　大数据的广义概念示意图

1. 大数据理论

目前，大数据理论层面主要是从计算机科学、统计学、数学以及实践等方面汲取营养，旨在最终构建数据科学理论基础，建立数据空间的科学认知体系。这个数据空间是独立且关联于自然世界和人类社会之外的新维度。

2. 大数据技术

大数据技术是推动大数据发展最活跃的因素。其关键技术可分为大数据采集技术，大数据集成与预处理技术，大数据存储技术(包括云计算技术、数据库与数据仓库技术、分布式数据处理技术、数据湖技术等)，大数据分析技术(包括统计分析、数据挖掘、机器学习

以及深度学习和增强学习等），大数据可视化技术，大数据平台技术，以及大数据隐私与安全技术等。

3．大数据应用

大数据应用在我国呈现出大力推进和积极拓展的局面。多领域应用场景的有效开发成为带动大数据发展的重要引擎。大数据应用一般可分为个人、企业与行业、政府以及时空综合应用等若干方面。其中，个人大数据应用的典型代表，如根据医疗健康大数据建立个人数据中心，指导个人就医、运动和饮食等；企业与行业大数据应用的典型案例，如商业银行通过征信和交易大数据系统，对各种贷款业务进行风险控制等；政府大数据应用的典型场景，如管理部门利用公共大数据系统，进行交通管理、旅游规划和环境保护等；时空综合大数据是指，在一定区域的时序化基础地理信息之上，与其他专题大数据构成一体化数据资源。其典型案例有智慧城市时空大数据应用、百度公司时空大数据城市计算等。

4．大数据生态

大数据生态是指大数据事业与其相关环境所形成的相互作用、相互影响的共生系统。主要包括大数据市场需求、政策法规、人才培养、产业配套与行业协调、区域协同与国际合作等要素。

1.1.2 目标定位

本书聚焦于大数据分析技术中的统计分析，将大数据集定位在各种事物的客观动态记录上，并从以下三个方面确定定位。

1．明确目标

初学者首先应结合自己的知识基础、专业方向、兴趣能力以及职业发展规划，搞清楚自己将向大数据的哪个技术方向努力。

一般可将大数据技术涉及的各个方面归纳为两大技术方向：大数据工程和大数据分析，其对应着大数据工程师和大数据分析师的角色。大数据工程主要涉及物联网、云计算、分布式存储和计算、大数据平台等基础硬件、软件的架构设计和设施搭建；大数据分析主要涉及数据采集管理、数据可视化、数据分析与建模、机器学习、数据安全、数据应用服务等相关的模型、算法、工具的探讨和研究。

总体而言，随着大数据物理基础和系统架构的不断成熟和持续普及，以大数据工程为核心的项目通常是有限的，而随着精细化管理和智能化技术的蓬勃发展，大数据分析则会进入各行各业，形成长期且持续的需求，从而使更多的大数据研究项目进入这个领域。本书正是为有志从事**大数据分析**的初学者准备的。

2．明确途径

初学者学习大数据分析，可以从统计方法、数据挖掘、机器学习、深度学习这个途径循序渐进，逐步提升。大数据的统计分析是一个出发点，也是后续分析的基础。本书正是为起步于统计方法的大数据分析初学者准备的。

3．明确工具

大数据的统计分析需要功能强大、灵活易用的实现工具。作为面向统计分析的计算机语

言，R 无疑是一个最好的选择。R 的前身是 1976 年美国贝尔实验室开发的 S 语言。20 世纪 90 年代 R 语言正式问世，因两名主要研发者 Ross 和 Robert 名字首字母均为 R 而得名。目前，R 已发展成为具有共享性，可运行于 Windows、Linux、Mac OS X 操作系统之上，支持交互式数据探索、可视化和分析实践，支撑统计理论研究和大数据统计分析的强大平台。本书正是为聚焦大数据分析中的统计方法及其 R 实现的初学者准备的。

1.1.3　初识 R

数据分析，一方面需要深厚的统计专业知识，另一方面也离不开有效的软件工具。目前，包括 R 在内的数据分析工具繁多，功能上各有侧重，操作上有简有繁，数据处理效率有高有低。可从不同角度对这些工具进行粗略分类。

第一种角度：商业软件和共享软件。

商业软件的字面含义是指可作为商品进行交易的计算机软件。通常商业软件的所有权属于相应的软件公司，软件公司负责整个软件的研发、升级和服务等，能够有效确保软件功能的完备性，软件开发的规范性，以及可推广性等。数据分析软件中常见的商业软件包括 IBM SPSS Statistics，IBM SPSS Modeler，SAS，Matlab 等。商业软件的销售价格一般较高，囊括的分析方法都是经典的和成熟的。

共享软件属于非商业软件，其最大特点之一是共享性，使用者可以到相应的网站上免费下载和使用。与商业软件相比，共享软件一般没有专门特定的软件公司主导研发，通常由爱好者们自由研发并上传到网上。共享软件具有较大的随意性，除囊括众多的经典分析方法外，还拥有大量前沿的新模型、新算法，且更新速度快。R 正是数据分析软件中最常见的一款非商业的共享软件。

第二种角度：“傻瓜”软件和“非傻瓜”软件。

所谓“傻瓜”软件是指操作简单、易学易用的软件。使用者只需通过窗口、菜单、对话框等的操作即可自如应用这些软件，无须具有计算机编程知识。为利于推广使用，商业数据分析软件，如 IBM SPSS Statistics，IBM SPSS Modeler，SAS，Matlab 等一般都支持窗口、菜单、对话框等“傻瓜”式操作。“傻瓜”式的使用模式虽便于操作，但无法最大限度地满足用户个性化的分析需要。“非傻瓜”软件的优势则更多体现于此。

“非傻瓜”软件对使用者的计算机水平有较高要求，尤其对计算机编程能力和技巧要求较高。但使用者一旦掌握了它，通过编写程序，不仅能够实现常见的数据分析目标，还能自行对现有的模型、算法进行改进、扩充，从而充分满足不同个性化分析的需求。R 正是这样一款“非傻瓜”的统计分析软件。

共享性使得 R 博大精深，但也会令初学者眼花缭乱，常常因无从下手而感觉软件体系杂乱无章。根据由浅入深的数据分析需求，依据统计分析的基本框架，分阶段、分步骤地学习 R，是一种快速有效掌握 R 的基本策略。本书后续章节将对统计分析的基本框架做详细说明。

1.2　R 语言入门

1.2.1　R 中的基本概念

利用 R 语言进行统计分析的核心内容，是在 R 的工作空间中创建和管理 R 对象，调用

已被加载的 R 包中的系统函数或编写用户自定义函数，以逐步完成数据的管理和分析。这里涉及几个基本概念：包、函数、工作空间、对象。掌握这些基本概念的含义，对快速入门R将有很好的帮助。

1. 包(Packages)

R 语言是一种面向统计分析的共享性和开源性软件平台，是一种面向统计分析的计算机高级语言。具体讲，R 是一个关于包的集合。包是关于函数、数据集、编译器等的集合。

包是R的核心，可划分为基础包(Base)和共享包(Contrib)两大类。

(1)基础包

基础包，顾名思义，为 R 的基本核心系统，是默认下载和安装的包，由 R 核心研发团队(Development Core Team，简称 R Core)维护和管理。基础包支持各类基本统计分析和基本绘图等功能，并包含一些共享数据集供用户使用。

(2)共享包

共享包是由 R 的全球性研究型社区和第三方提供的各种包的集合。迄今为止，共享包中的“小包”已多达 12000 多个，涵盖了各类现代统计和数据挖掘方法，涉及地理数据分析、蛋白质质谱处理、心理学测试分析等众多应用领域。使用者可根据自身的研究目的，有选择地自行指定下载、安装和加载。

2. 函数

R 函数是存在于 R 包中的实现某个计算或某种分析的程序段，每个函数都有一个函数名。可通过函数调用的方式，直接调用已有函数解决分析中的各类计算问题。函数名是函数调用的唯一标识。可通过以下两种格式实现函数调用。

格式一：函数名(形式参数列表)

函数名后括号中附带形式参数的函数调用，称为有形式参数的函数调用。函数调用时，用户须在函数名后的括号中，依顺序给出一个或多个参数值，各参数值间以英文逗号隔开。R 将这些参数值“喂给”函数，并根据参数值进行计算。

格式二：函数名()

函数名后括号中无任何内容的函数调用，称为无形式参数的函数调用。这类函数无须用户指定参数值，不必“喂给”函数任何具体值即可进行既定的计算。

3. 工作空间(Workspace)

工作空间，简单讲就是 R 的运行环境，也称工作内存。R 的所有计算都是基于工作内存的，即需要将外存中的 R 包、数据等，首先加载到工作内存中，然后才能够进行后续的计算。基于这种工作机制，R 成功启动后会首先自动将基础包加载到工作空间中，旨在提供最基本的程序运行环境。

说明：

① R 的基础包由很多“小包”组成。R 启动后仅自动将其中的部分“小包”加载到工作内存，还有些“小包”需用户根据分析需要适时选择性地手工加载。

② 用户只能调用已加载到工作空间的R包中的函数和数据集等。

4. R 对象

R 对象是存在于工作空间中的基本单元。分析的数据及相关分析结果等均将以 R 对象的形式组织。每个 R 对象都有一个对象名，是对象访问的唯一标识。直接给出对象名，即可访问 R 对象，查看其中存储的数据或分析结果或将新值赋值给 R 对象。

总之，R 的工作空间中加载了哪些包，包中包含了哪些函数，如何利用这些函数创建和管理 R 对象并进行数据分析，是 R 语言学习的主要线索。

1.2.2　R 的下载安装

可从 R 的网站 www.r-project.org 上免费下载并安装 R 软件。网站的主页如图 1-2 所示。

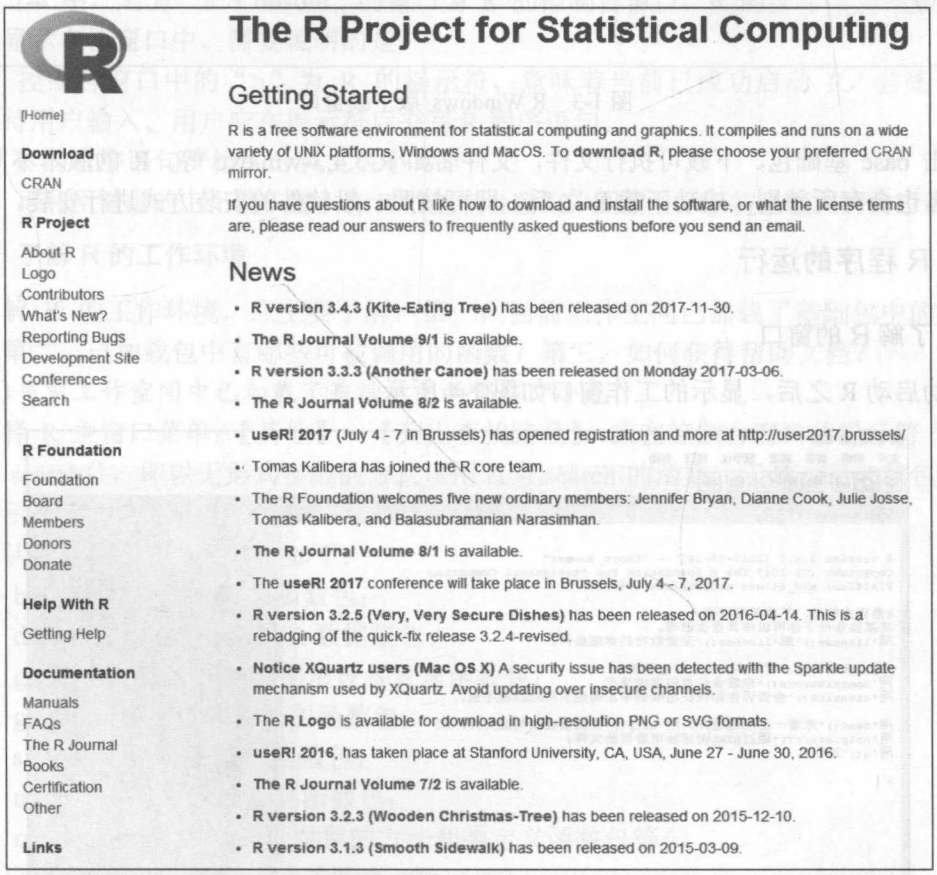

图 1-2　R 的网站主页

R 网站主页列出了与 R 有关的各类信息，包括 R 社区的主要成员情况、R 的相关帮助文档等。R 的基础包、相关文档和大多数共享包以 CRAN(Comprehensive R Archive Network，http://CRAN.R-project.org)的形式集成在一起。同时，为确保不同地区 R 用户的下载速度，在全球众多国家均设置了镜像链接地址。镜像可视为一种全球范围的缓存，每个镜像地址对应一个镜像站点(Mirror Sites)，它们有各自独立的域名和服务器，存放的 R 系统是主站点的备份，内容与主站点完全相同。用户下载 R 时，须首先单击 CRAN 链接，选择一个镜像链接地址。国内的 R 用户可以选择 R 在中国的镜像站点。

R 支持在 Windows、Linux、Mac OS X 操作系统上运行，用户可根据不同情况选择不同的链接。如选择 Download R for Windows，表示下载运行于 Windows 操作系统下的 R，显示窗口如图 1-3 所示。

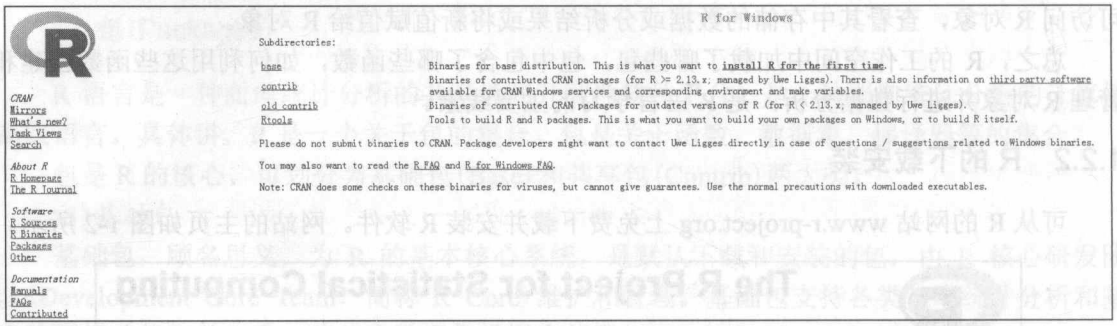

图 1-3　R Windows 版下载窗口

单击 base 基础包，下载可执行文件，文件名如 R-3.4.3-win.exe 等。R 的版本不同，可执行文件名也会有所差别。成功下载 R 之后，即可按照一般软件的安装方式进行安装。

1.2.3　R 程序的运行

1. 了解 R 的窗口

成功启动 R 之后，显示的工作窗口如图 1-4 所示。

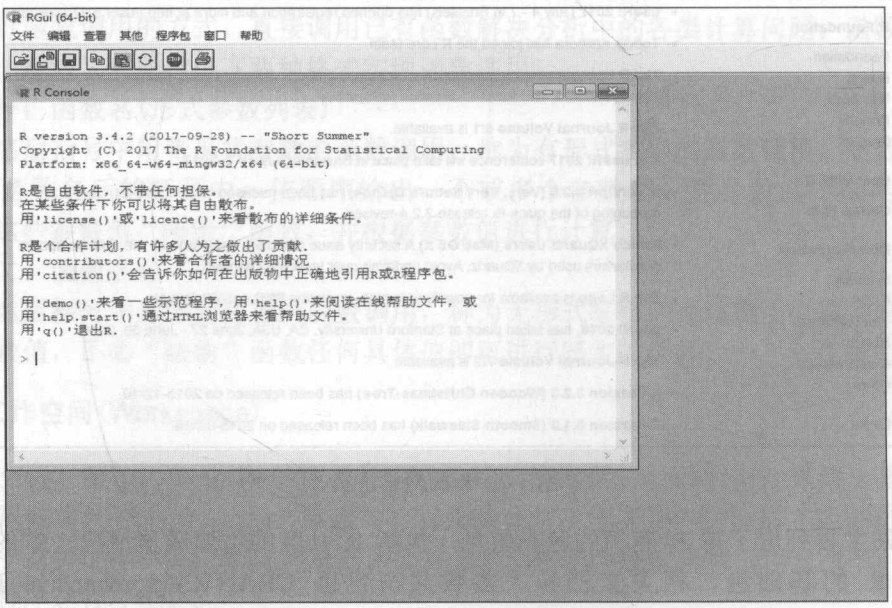

图 1-4　R 的工作窗口

图 1-4 中，名为“RGui(64-bit)”的窗口为 R 的主窗口，用于管理 R 的运行环境。可通过窗口菜单和工具栏完成以下工作。

①【文件】菜单：新建、打开、打印和保存 R 程序文件，管理 R 的工作空间。