统计数据分析与应用丛书

Statistics

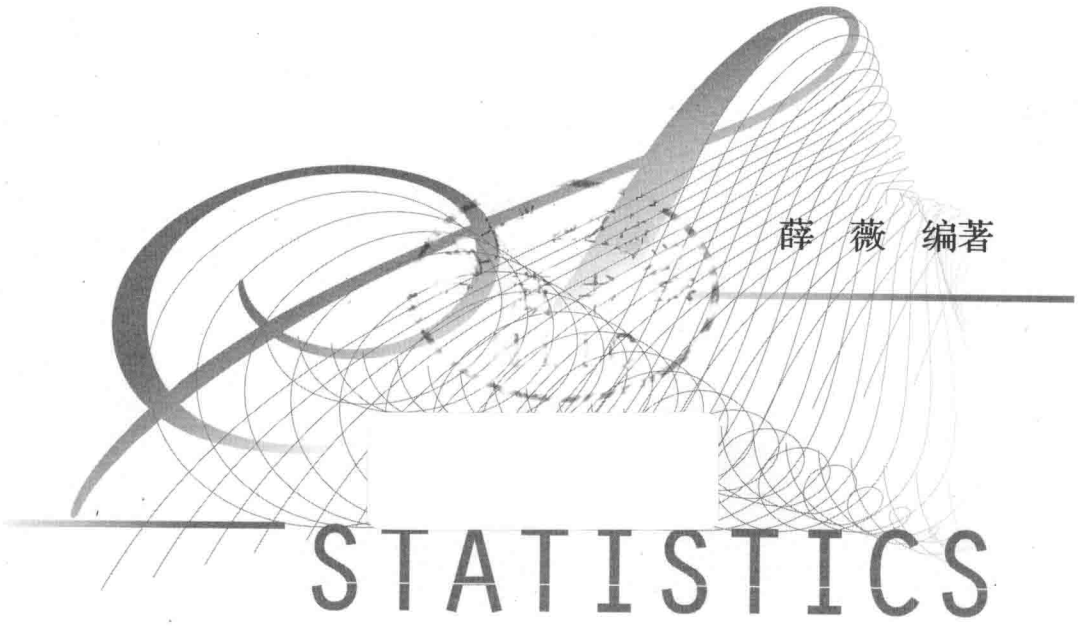
基于R的统计分析与数据挖掘

薛 薇 编著

中国人民大学出版社

统计数据分析与应用丛书 Statistics

基于R的统计分析与数据挖掘



薛 薇 编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

基于R的统计分析与数据挖掘/薛薇编著. —北京：中国人民大学出版社，2014.4
（统计数据分析与应用丛书）
ISBN 978-7-300-19074-7

Ⅰ.①基… Ⅱ.①薛… Ⅲ.①统计分析-软件工具 Ⅳ.①C819

中国版本图书馆CIP数据核字（2014）第062115号

统计数据分析与应用丛书

基于R的统计分析与数据挖掘

薛　薇　编著

Jiyu R de Tongji Fenxi yu Shuju Wajue

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511770（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	涿州中煤制图印刷厂北京分厂		
规　　格	185 mm×260 mm　16开本	**版　　次**	2014年5月第1版
印　　张	25.5插页1	**印　　次**	2014年5月第1次印刷
字　　数	596 000	**定　　价**	48.00元

前　言
Preface

R是一款应用前景广阔的数据分析工具，一个共享的开源软件平台。

R的不断发展受惠于这个方兴未艾的大数据时代，此谓天时。互联网、物联网和移动客户端的广泛应用，让人们置身于数据的汪洋大海和崇山峻岭之中。对于一个有进取心的探险者来说，商机与危机同在，机遇与挑战共存。利用R开展有效的数据分析工作，无疑是前进征途的指南针，同时也是挖掘职业生涯宝藏的利器。

R在国内的逐步普及得益于这个复杂激烈的市场竞争环境，此谓地利。目前以及未来相当长的一段时期，是我国改革开放的关键时期，也是产业结构转型和企业升级换代的重要战略机遇。面临竞争国际化、利益多元化、服务个性化、方式便利化的市场形势，以数据分析为重要管理手段，是促使国家科学决策和企业健康发展的战略选择。

R在全球广泛应用的根源在于其一贯的软件免费和程序代码开放策略，此谓人和。R像一个数据分析的生态群落，从封闭的实验室迁生到互联网的开放环境中，按照全新的生存法则迅速发展。任何人只需访问相应网站，便可免费下载获得R的系统、文档、数据集等全部相关资源。同时，R今天的学习者完全有可能成为明天的合作开发者。

《基于R的统计分析与数据挖掘》聚焦当今备受国内外数据分析师和数据应用者关注的R语言，企图借助R实现统计分析和数据挖掘。理由很简单：R不仅囊括了几乎所有的经典统计方法，而且拥有众多前沿的现代统计模型、数据挖掘算法以及顶尖的绘图功能；不仅可以解决数据分析的共性问题，而且能够服务于电商、金融、医学、生物、地理、环境、传媒等领域的特色数据应用；不仅适合统计分析的学习者、学术研究的探索者，而且适合致力数据应用开发的实践者和掘金者。

一方面，本书全面系统地介绍了R的数据对象、常用系统函数、用户自定义函数以及流程控制等与统计编程和数据模拟相关的基础知

识；另一方面，以由浅入深的数据分析过程为线索，详细讨论R的数据组织、数据加工和可视化图形处理、回归预测建模、变量降维处理以及分类和聚类研究等。内容上涵盖从基础编程到统计模拟，从单变量描述性分析到多变量相关性研究，从线性模型到非线性模型，从满足分布假设的经典统计模型到以随机化为基础的现代统计建模乃至数据挖掘等众多方面。

本书既不是仅侧重理论讲解的统计分析和数据挖掘教材，也不是仅侧重编程操作的R的使用手册，而是以数据分析贯穿全书，是两者的有机结合。在以数据模拟的直观方式论述方法原理的同时，通过案例强化R的操作实践性；在以解决应用问题为目标讨论R操作的同时，通过原理论述强化模型结果的理解解读。

本书定位于统计分析和数据挖掘的学习者、实践者和研究者，旨在使读者理解统计分析原理，熟练操控R软件，拓展数据应用，提升研究水平。

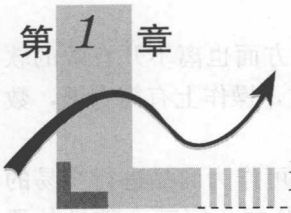
掌握基于R的主流数据分析技术，以洞察大数据的敏锐智慧，懂得大数据分析精髓的务实态度，掌控大数据分析的卓越能力，每个人都可能成为大数据分析师。数里淘金，R的大数据时代即将到来。

请读者到人大经管图书在线（http://www.rdjg.com.cn）下载本书案例数据和R程序代码。在此特别感谢中国人民大学出版社对本书出版的大力支持。北京城市学院经管学部2013级陈笑语等同学，为本书案例数据的收集和整理做了很多工作，这里也一并表示感谢。

书中不妥和错误之处，望读者不吝指正。

薛　薇

目　录



第1章 关于R

随着移动互联网技术的发展和普及，物联网建设进程的推进和应用，人们正迎接着一个“大数据”时代的到来。顾名思义，“大数据”时代的重要特点之一是数据量巨大。据估计，2007年全球所积累的各种各样的数字化数据容量高达到300EB。如果一部压缩的数字电影需要大约1GB的存储容量，那么，1EB相当于10亿部数字电影的存储容量，300EB则意味着全球已拥有了3 000亿部数字电影容量的数据。南加利福尼亚大学安嫩伯格通信学院的马丁·希尔伯特教授预测2013年年底全球积累数据量将达到1.2ZB。如果将这些数据以文字形式印制在普通纸质图书上，则这些图书平铺起来可覆盖52个美国国土面积。尽管该预测结果尚未得到权威验证，但仍可作为参考。如果将这些数据刻在普通VCD光盘中，则这些光盘堆积起来的高度是地球到月球距离的5倍。进一步，“大数据”给人类带来的不仅是日益丰富堪比下一个社会发展阶段的石油和金矿的数据资源，更深层的意义是“大数据”正促使人们逐步改变原有的思维模式、商业模式、管理模式，甚至国家层面的发展战略模式。所有变革背后的技术根本是以数据处理和分析为重要核心的科学管理和决策。进行数据处理和分析，需要深厚的统计专业知识，也离不开有效的软件工具。因此，如果说关于数据的知识应成为当代个人知识结构中的必备要素，那么熟练掌握一款数据分析软件，则是当代数据分析者的必备技能。这样的“数据科学家”必将是“大数据”时代的人才瑰宝。

1.1 为什么选择R

R语言是一种面向统计分析的计算机高级语言，属于数据分析软件的范畴。R语言的前身是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，它因两名主要研发者Ross和Robert的名字首字母均为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，支持交互式数据探索和分析实践，

支撑统计理论研究和探讨的强大平台。

进行数据处理和分析，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在内的数据分析软件繁多，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软件进行粗略分类。

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，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。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和成熟的。共享软件属于非商业软件，其最大的特点之一就是共享性，使用者可以到相应的网站上免费下载使用。与商业软件中的非免费软件相比，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之外，还拥有众多较为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，是指操作非常简单的软件。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、对话框等的“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要求较高，尤其对计算机编程能力和技巧要求较高。使用者一旦掌握，通过编写程序，不仅能够实现常见的数据分析目标，而且能够自行对现有的模型、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”型的统计软件。

了解不同软件的特点，根据自身的专业水平以及数据分析的目标，选用一款或几款恰当适用的分析软件，是数据分析者的明智之举。例如，对于本科低年级的统计专业学生，学习重点是深刻理解统计原理。统计模拟是帮助学生生动直观地理解经典统计理论的重要手段。由于“傻瓜”型的商业软件在灵活性方面存在不足，不易实现统计模拟，而“非傻瓜”型软件具有明显优势。因此，此时选择R是再恰当不过的了。又如，对于实际应用领域的数据工作者，其工作过程可细分为组织管理数据、展示数据分布特征、揭示数据背后规律等多个阶段。数据组织管理阶段，由于商业软件通常依托数据库技术组织和管理数据，相比共享软件有更高的处理效率，因此选择商业软件事半功倍。图形应用是数据分布特征描述展示的核心，由于R具有顶尖的图形绘制功能，因此该阶段R具有绝对优势。数据建模阶段，因经典分析方法通常可满足实际数据的分析需求，即便再前沿的模型和算法也没有更大的施展空间，所以选择操作简单的“傻瓜”型商业软件是非常务实的做法。再如，对于统计学科领域的学者，由于更多关注从理论角度探讨模型的优劣并改进模型，提出新模型新算法，因此更需要一种能够有效实现新算法、评价新模型优劣的软件工具。为此，采用支持高效编程的R软件应是最恰当的。

1.2　如何学习 R

共享性使得 R 博大精深，但会令初学者眼花缭乱，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，依据数据分析过程分阶段学习 R，是一种快速有效地掌握 R 的基本方法。

数据分析主要包括数据的组织、数据的基本描述和预处理、数据的分析建模、模型验证和模型应用等主要阶段（见图 1—1）。统计学对各阶段的主要任务、重点和难点都有完备和严谨的论述，在此仅就与 R 直接相关的内容作简要阐述，以明确快速学习 R 的知识要点和路径。

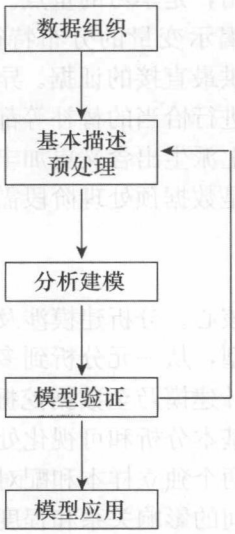


图 1—1　数据分析的主要阶段

1.2.1　数据的组织

利用统计软件实现数据分析首先要进行数据的组织。统计分析中的数据通常被组织成二维表的形式。其中，列变量用来描述研究对象的某种属性，例如商品销售额、人们的受教育程度、产品的质量等级等。各列都有一个用以标识相应属性的称谓，即变量名，是数据访问和分析的基本单位。根据变量所对应属性的类别，计量尺度变量可进一步细分为数值型变量（metric variable）和分类型变量（categorical variable）。数值型变量用于存储诸如收入、身高、年龄等连续或非连续的定距属性数据。分类型变量又分为顺序分类型变量和名义分类型变量。顺序分类型变量用于存储诸如职称、年龄段、教育水平等具有内在顺序的类别属性数据。名义分类型变量用于存储诸如性别、职业、籍贯等无内在顺序的类别属性数据。

二维表中的一行称为一个观测，是个体所有属性取值的集合。一组观测的集合称为一个样本。

学习 R 的首要任务是关注 R 以怎样的方式实现上述二维表数据的组织，以及 R 是否

还有其他的数据组织方式等。

1.2.2 数据的基本描述和预处理

数据的基本描述是数据分析的起点，也是后续数据建模的基础。基本描述的根本是选择恰当的统计图形和统计量，全面、直观、准确地展示变量的分布特征，以及与其他变量的相关性特征。对于分类型变量，通常选用条形图、饼图等图形工具和频数、频率等统计量来描述。对于数值型变量，通常选用直方图、折线图、茎叶图、散点图等图形工具和均值、标准差、偏态系数、峰度系数等统计量来描述。

R的图形制作功能强大，包括的图形种类繁多，不仅可绘制传统的统计图形，还可绘制其他更具特色的图形，如小提琴图、克利夫兰点图、相关系数图、马赛克图等。R在变量分布特征的图形展示方面表现突出，是学习的重点。

数据的基本描述，一方面用于揭示变量的分布特征；另一方面为数据是否包含异常值，以及包含多大比例的缺失值提供最直接的证据。异常值和缺失值会对后续的数据建模产生重要影响。发现它们，剔除或进行恰当的替补等都是必须的，也是进行数据预处理的重要方面。此外，在现有数据基础上派生出含义更加丰富的新变量，对现有变量进行恰当的变换处理为后续建模做准备，也是数据预处理阶段需要解决的问题。

1.2.3 数据的分析建模

数据的分析建模是数据分析的核心。分析建模涉及面非常广，包括从单变量描述到多变量相关，从线性模型到非线性模型，从一元分析到多元统计，从满足分布假设的传统统计建模到以随机化为基础的现代统计建模乃至数据挖掘等内容。后续将按照由浅入深的建模线索逐一展开讨论。对数据进行基本分析和可视化处理之后，通常需要对两个或者多个总体的均值做比较检验，这将涉及两个独立样本和配对样本的均值检验，以及方差分析等问题。进一步，为精确刻画变量之间的影响关系和程度，需要进行回归建模，包括建立一般线性回归模型和广义线性回归模型等。再进一步，多元统计分析是数据建模的重要方面，主要包括以剖析变量整体结构为目的的聚类分析、以变量降维为目的的因子分析、以分类预测为目的的判别分析等。经典统计方法对变量有诸多假定。对不能满足假定的数据建模时，需要采用更加宽松的数据挖掘方法，主要涉及决策树和神经网络等方法。

1.2.4 模型的验证

当数据建模的目的是预测时，模型的预测精度或误差是评价模型的主要依据。基于随机化和重抽样的统计推断和模型评价，是现代模型评价的基本策略。如何将数据建模和模型评价有机结合起来，如何通过组合预测模型提高预测精度，都是需要关注的问题。

1.3 R入门必备

1.3.1 R的包

R语言是一种面向统计分析的计算机高级语言。具体来讲，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。编写 R 程序的过程就是通过创建 R 对象组织数据，通过调用系统函数，或者创建并调用自定义函数逐步完成数据分析任务的过程。

包是 R 的核心，可划分为基础包（base）和共享包（contrib）两大类。

基础包，顾名思义即为 R 的底层核心，是下载 R 时默认下载和安装的包。其中包括很多“小包”。成功安装并启动 R 后，有些“小包”被默认加载到 R 的工作空间，用户可直接调用其中的函数。但还有些“小包”需要使用者手工加载，只有手工加载后才可以调用其中的函数。

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 4 000 多个，涵盖了各类现代统计方法，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。使用者可根据自己的研究目的，有选择地自行指定下载。

1.3.2　R 的下载安装

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

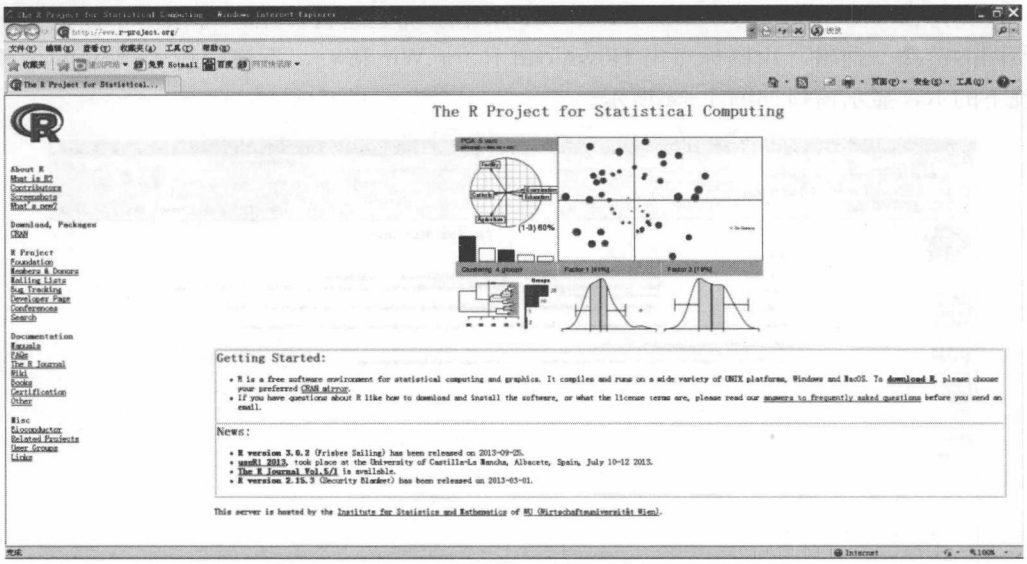


图 1—2　R 网站主页

R 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。用户下载 R 时，需首先用鼠标点击 CRAN 链接，选择一个镜像链接地址。镜像可视为一种全球范围的缓存，是为提高用户在不同地区的下载速度而专门设立的。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。国内的 R 用户可以选择 R 在中国的镜像站点，如厦门大学（Xiamen University）、中国科技大学（University of Science and Technology of China）、北京交通大学（Beijing Jiaotong University）等下载 R。指定镜像站点后，将显示如图 1—3 所示的窗口。

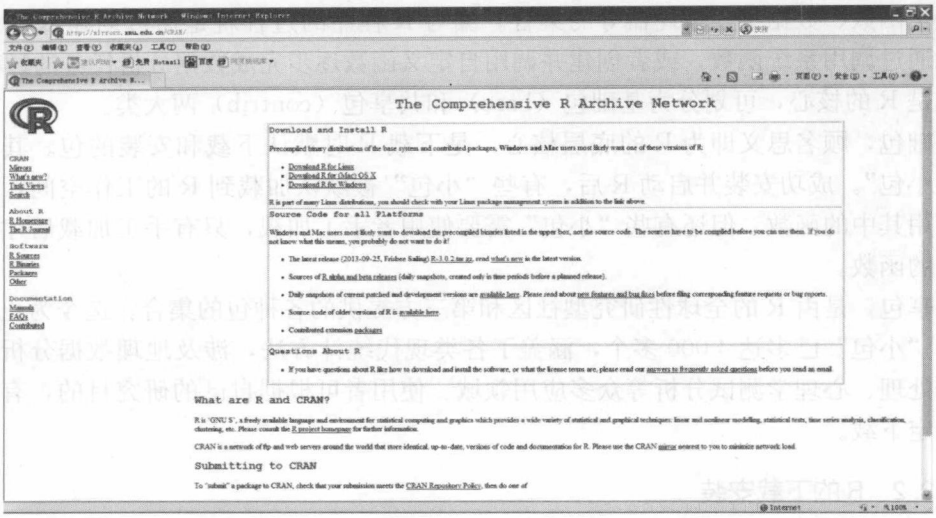


图 1—3　R 的下载窗口（1）

R 支持在 Windows，Linux，Mac OS X 操作系统上运行，用户可根据不同的情况选择不同的链接。通常，用鼠标点击 Download R for Windows，下载运行于 Windows 操作系统下的 R，显示窗口如图 1—4 所示。

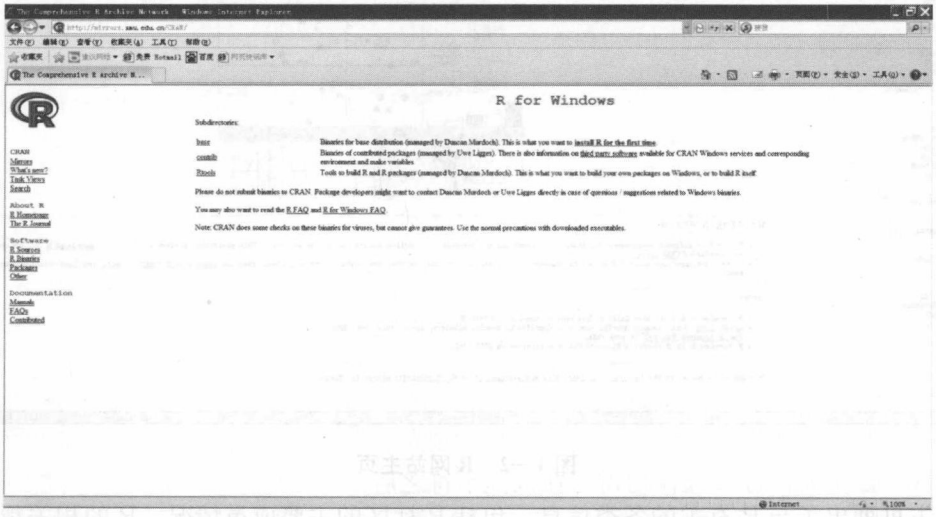


图 1—4　R 的下载窗口（2）

用鼠标点击基础包，下载可执行文件，如 R-3. 0. 1-win. exe 文件等。R 的版本不同，可执行文件的名称也会有所差别。成功下载 R 之后，即可按照 Windows 软件的一般安装方式进行安装。

1.3.3　启动 R

成功启动 R 之后显示的窗口如图 1—5 所示。

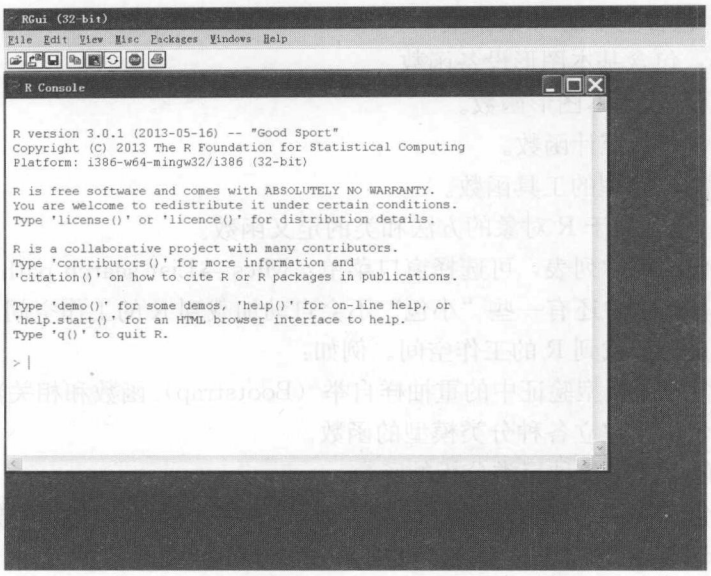


图 1—5 R 的工作窗口

图 1—5 中，名为 RGui(32-bit) 的窗口为 R 的主窗口，包括窗口菜单和工具栏。其中：

- File 菜单：主要用于 R 程序文件的新建、打开、打印和保存，R 工作空间的管理等。
- Edit 菜单：主要服务于 R 程序的编写以及 R 控制台清空管理。
- View 菜单：指定在主窗口中是否显示状态栏，是否显示工具栏。
- Misc 菜单：主要用于终止当前或所有运算，显示或删除工作空间中包含的 R 对象，显示当前已经加载的包名称列表等。
- Packages 菜单：主要用于加载已下载的包。联网条件下，指定镜像地址、下载安装其他包、对已下载安装的包进行更新等。
- Windows 菜单：主要用于指定 R 主窗口所包含的其他窗口（如控制台窗口、程序编辑窗口等）的排列形式。如左右排列（Tile horizontally）或上下排列（Tile vertically）等。
- Help 菜单：主要用于以各种方式浏览 R 的帮助文档。

图 1—5 中，名为 R Console 的窗口为 R 的控制台窗口，R 的操作以及计算结果均显示在该窗口中。控制台窗口中的“>”为 R 的提示符，意味着当前已成功启动 R，且处于就绪状态。后续所有的 R 操作均需写在该提示符之后。

需要注意的是：R 代码的编写是严格区分英文大小写的；利用键盘上的上下箭头键，可重复显示以往的书写内容。

成功启动后首先应关注三个问题：第一，当前已经加载了哪些包；第二，当前可以做哪些事情；第三，如何获得 R 的帮助文档。

一、当前已经加载了哪些包

成功启动 R 意味着基础包中的默认加载包已成功加载到 R 的工作空间，用户可以直接调用其中的函数。这些包主要有：

- base：包含基本的 R 函数。

- datasets：包含基本的R数据集。
- grDevices：包含基本图形设备函数。
- graphics：包含基本图形函数。
- stats：包含各类统计函数。
- utils：包含R管理的工具函数。
- methods：包含关于R对象的方法和类的定义函数。

为显示上述包的名称列表，可选择窗口菜单：Mics →List search path。

除此之外，基础包中还有一些“小包”不会自动加载到R的工作空间。同时，已下载的共享包也不会自动加载到R的工作空间。例如：

- boot：包含统计模型验证中的重抽样自举（Bootstrap）函数和相关数据集。
- class：包含用于建立各种分类模型的函数。
- cluster：包含用于各种聚类分析的函数。
- foreign：包含读取各种格式，如SPSS，SAS等格式数据文件的函数。
- KernSmooth：包含用于核密度估计的函数。
- lattice：包含各种格栅函数，用于高级图形的绘制。
- MASS：包含Venables和Ripey所著的*Modern Applied Statistics with S*一书的配套函数、工具和数据集。
- mgcv：包含用于带多平滑参数选择的广义域回归函数。
- nlme：包含用于线性和非线性混合效应的建模函数。
- nnet：包含建立带单个隐层的前馈式神经网络模型、多项式对数线性模型的函数。
- rpart：包含建立分类回归树的函数。
- spatial：包含空间分析的函数。
- surival：包含生存分析的函数。

为显示上述包的名称列表，可选择窗口菜单：Packages →Load Package。

二、当前可以做哪些事情

成功启动R意味着用户可在R工作空间中创建和管理R对象，调用已加载包中的函数，实现对对象的管理和对相关数据的分析等。其中：

R对象是R程序处理的基本单元，用于待分析数据的组织，以及分析结果的组织等。每个R对象均有一个对象名作为唯一的标识。一般可直接通过对象名访问对象中的数据或其他内容。

函数是实现计算或分析的程序段，可视为一种特殊的对象。每个函数均有一个函数名。用户可通过两种形式调用函数：

1. 函数名()

这是一种无形式参数的函数调用，即括号中不给出任何内容。R将以默认的参数值调用并运行函数，运行结果即函数值将自动显示在R的控制台窗口中。

例如，为显示默认加载包的名称列表，可在提示符“>”后书写：“search()”，即以无形式参数的方式调用名为search的函数。

2. 函数名(形式参数列表)

这是一种带形式参数的函数调用，即括号中依顺序给出一个或多个形式参数，各形式

参数之间以英文逗号隔开。R 将以所指定的参数值调用并运行函数。

例如，为了解各包中包含哪些函数、各函数的功能以及如何调用函数，可在提示符“>”后书写：“library(help="包名称")”。具体如 library(help="stats")，即以带形式参数（这里的形式参数为 help="stats"）的方式调用名为 library 的函数。

三、如何获得 R 的帮助文档

对于 R 的初学者，学会使用 R 的帮助文档是非常有必要的。获得 R 的帮助可通过在提示符“>”后调用以下主要函数实现。

1. help. start()

该函数以浏览器的形式打开 R 的帮助文档，如图 1—6 所示。

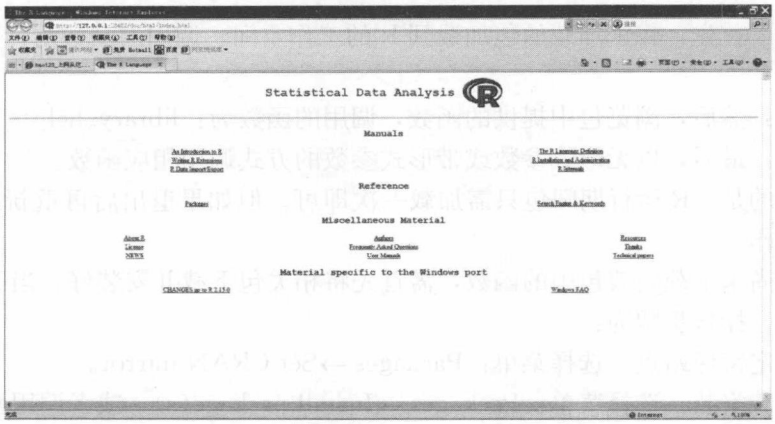


图 1—6　R 的帮助手册

图 1—6 以链接的形式显示了 R 的相关资源和整套帮助手册。用户只需用鼠标点击相关链接，即可浏览相应的内容。

2. help(函数名)

该函数用于查看指定函数的帮助文档。例如 help(boxplot)，显示内容如图 1—7 所示。

图 1—7　函数 boxplot 的帮助文档