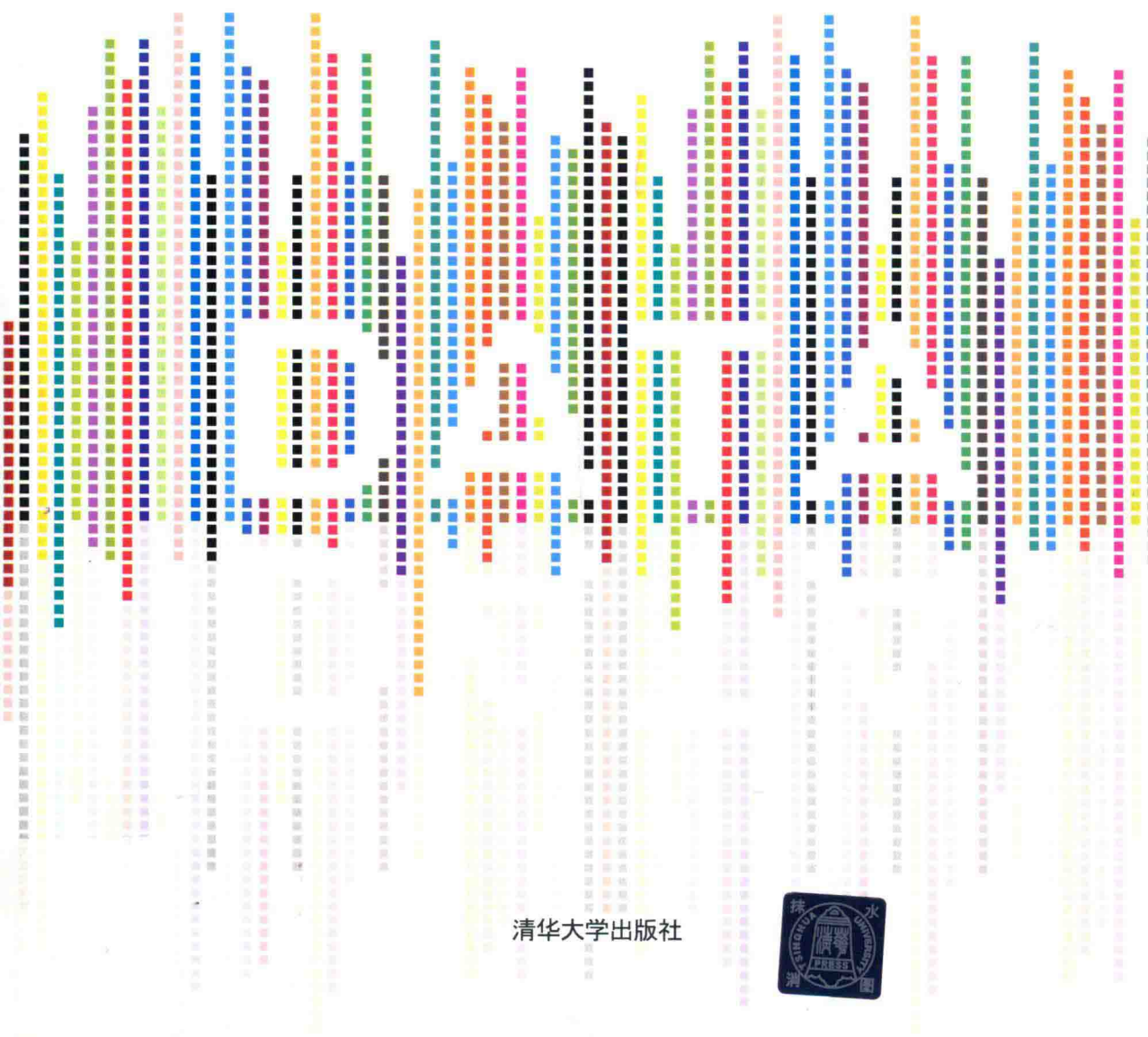


大数据时代经济与金融数据分析系列丛书



数据统计分析的R软件应用

朱顺泉 编著



清华大学出版社



● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

朱顺泉 编著

北京

内 容 简 介

本书内容包括 R 软件下载、安装与启动, R 软件数据结构, R 软件数据存储与读取, R 软件编程, R 软件绘图, 描述性统计的 R 软件应用, 参数估计的 R 软件应用, 参数假设检验的 R 软件应用, 相关分析与回归分析的 R 软件应用, 主成分分析与因子分析的 R 软件应用, 聚类分析与判别分析的 R 软件应用, 典型相关分析与对应分析的 R 软件应用。

本书紧跟大数据分析时代, 内容新颖、全面, 实用性强, 融理论、方法、应用于一体, 是一部供统计学、数量经济学、管理科学与工程、应用数学、计算数学、概率统计、金融工程、投资学、金融专业硕士、金融学、经济学、财务管理、会计学、工商管理等专业本科高年级学生与研究生使用的实验教材或参考书。

本书封面贴有清华大学出版社防伪标签, 无标签者不得销售。

版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目(CIP)数据

数据统计分析的 R 软件应用/朱顺泉编著. --北京: 清华大学出版社, 2015

大数据时代经济与金融数据分析系列丛书

ISBN 978-7-302-39970-4

I. ①数… II. ①朱… III. ①统计分析—应用软件 IV. ①C819

中国版本图书馆 CIP 数据核字(2015)第 087774 号

责任编辑: 刘向威 薛 阳

封面设计: 文 静

责任校对: 李建庄

责任印制: 王静怡

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-62770175 邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

课 件 下 载: <http://www.tup.com.cn>, 010-62795954

印 装 者: 北京国马印刷厂

经 销: 全国新华书店

开 本: 185mm×260mm

印 张: 12.5

字 数: 310 千字

版 次: 2015 年 6 月第 1 版

印 次: 2015 年 6 月第 1 次印刷

印 数: 1~2000

定 价: 25.00 元

产品编号: 063623-01



大数据时代,数据成为决策最为重要的参考之一,数据分析行业迈入了一个全新的阶段。《数据统计分析的 R 软件应用》侧重于 R 软件的数据存取、图形展示和统计分析,重点介绍数据基础统计与多元统计分析的 R 软件应用,同时结合大量实例,对 R 软件进行科学、准确和全面的介绍,以便使读者能深刻理解 R 软件的精髓和灵活、高效的使用技巧。

本书之所以采用 R 软件,是因为它具有强大的图形展示和统计分析功能、免费使用、更新及大量可随时加载的有针对性的程序包。MATLAB、SAS、SPSS、Eviews、Stata、S-PLUS 等都是收费软件。R 简洁的输出和强大的帮助系统为用户提供了极好的自学环境,因此它受到广大用户的欢迎和喜爱。

本书通过丰富的实例,详细介绍 R 3.1.1 在数据统计分析中的应用,侧重于理论方法与应用相结合,实例丰富且通俗易懂,对 R 软件的各种绘图方法、与数据表格的连接、基础统计分析和多元统计分析应用等方面描述很有自己的特色,详细介绍了各种统计方法在 R 软件中的实现过程。本书的特点是:以问题为导向,通过问题来介绍 R 软件的使用方法,因此,读者通过本书不仅能掌握使用 R 软件及相关的程序包的使用方法,而且能学会从实际问题分析入手,使用 R 软件解决实际数据的统计分析问题。

本书的内容这样安排:第 1 章介绍 R 软件下载、安装与启动;第 2 章介绍 R 软件数据结构;第 3 章介绍 R 软件数据存储与读取;第 4 章介绍 R 软件编程;第 5 章介绍 R 软件绘图;第 6 章介绍描述性统计的 R 软件应用;第 7 章介绍参数估计的 R 软件应用;第 8 章介绍假设检验的 R 软件应用;第 9 章介绍相关分析与回归分析的 R 软件应用;第 10 章介绍主成分分析与因子分析的 R 软件应用;第 11 章介绍聚类分析与判别分析的 R 软件应用;第 12 章介绍典型相关分析与对应分析的 R 软件应用。

本书实例与内容丰富,有很强的针对性,书中各章详细地介绍了实例的 R 软件具体操作过程,读者只需按照书中介绍的步骤一步一步地实际操作,就能掌握全书的内容。为了帮助读者更加直观地学习本书,特将书中实例的全部数据文件收录在本书的配套光盘中。读者在自己的计算机中建立一个 data 目录(其他目录名也可以),将所有数据文件复制到此目录,即可进行操作。

本书适合作为统计学、经济学、管理学、金融学等相关专业的本科生或研究生学习统计学数据分析等课程的实验参考用书,同时对从事数据分析的实际工作者也大有裨益。

本书的出版得到了清华大学出版社及编辑的大力支持、帮助,应该感谢他们为读者提供了一个这么好的数据统计分析工具。由于时间和水平的限制,书中难免出现一些纰漏,恳请读者谅解并提出宝贵意见。

作 者

2015 年 4 月于广州



第 1 篇 R 软件简介

第 1 章 R 软件下载、安装与启动	3
1.1 选择 R 软件的理由	3
1.2 R 软件的下载	4
1.3 R 软件及其程序包的安装	5
1.3.1 R 软件的安装	5
1.3.2 R 软件程序包的安装	6
1.4 R 软件的启动和退出	7
1.5 R 软件的在线帮助系统	8
练习题	8
第 2 章 R 软件的数据结构	9
2.1 R 的对象与属性	9
2.2 对象信息的浏览和删除	12
2.3 向量对象	12
2.3.1 数值型向量对象	12
2.3.2 字符型向量对象	13
2.3.3 逻辑型向量	14
2.3.4 因子型向量	14
2.3.5 数值型向量的运算	16
2.3.6 常用的统计函数	17
2.3.7 向量的下标与子集的提取	17
2.4 数组与矩阵对象	19
2.4.1 数组的建立	19
2.4.2 矩阵的建立	20
2.4.3 数组与矩阵的下标与子集的提取	22
2.4.4 矩阵的运算函数	23
2.5 数据框对象	25
2.5.1 数据框的直接建立	26
2.5.2 数据框的间接建立	26

2.5.3 适用于数据框的函数	27
2.5.4 数据框的下标与子集的提取	28
2.5.5 数据框中添加新变量	29
2.6 时间序列对象	30
2.7 列表对象	31
练习题	32
第 3 章 R 软件数据存储与读取	33
3.1 数据存储	33
3.2 数据读取	34
3.2.1 文本文件数据的读取	34
3.2.2 Excel 数据的读取	36
3.2.3 R 软件中数据集的读取	37
3.2.4 R 软件中的格式数据	38
练习题	38
第 4 章 R 软件编程	40
4.1 R 函数基础	40
4.2 循环和向量化	41
4.2.1 控制结构	41
4.2.2 向量化	42
4.3 用 R 编写程序	42
4.4 用 R 编写函数	43
4.5 用 R 编写标准函数的实例	43
练习题	46
第 5 章 R 软件图形的绘制	47
5.1 绘图基础知识	47
5.1.1 绘图函数	47
5.1.2 低级绘图命令	48
5.1.3 绘图参数	49
5.2 直方图的绘制	49
5.3 散点图的绘制	52
5.4 曲线标绘图的绘制	55
5.5 连线标绘图的绘制	57
5.6 箱图的绘制	59
5.7 饼图的绘制	60
5.8 条形图的绘制	61
5.9 点图的绘制	63

5.10 复杂图形的绘制	64
练习题	66

第2篇 基础统计分析的 R 软件应用

第6章 描述性统计的 R 软件应用	69
6.1 统计分布	69
6.1.1 正态分布	69
6.1.2 t 分布	71
6.1.3 卡方分布	72
6.1.4 F 分布	73
6.2 描述性统计量	74
6.2.1 总体和样本	74
6.2.2 量度尺度	74
6.2.3 频数分布	74
6.2.4 集中趋势的量度	76
6.2.5 中位数	77
6.2.6 众数	77
6.2.7 分位数	77
6.2.8 离散程度的量度	78
6.3 单组数据描述性统计的 R 软件应用	79
6.3.1 总体描述	80
6.3.2 五数及样本分位数描述	80
6.3.3 离差描述	81
6.3.4 偏度与峰度描述	81
6.4 多组数据描述性统计的 R 软件应用	82
6.4.1 多组数据的概括	82
6.4.2 方差和与协方差的计算	83
6.5 分类数据描述性统计的 R 软件应用	83
6.5.1 列联表的制作	84
6.5.2 获得边际列表	84
6.5.3 频数列联表	84
6.6 列联表图形描述的 R 软件应用	85
练习题	86
第7章 参数估计的 R 软件应用	87
7.1 参数估计概述	87
7.2 点估计的 R 软件应用	87
7.3 单正态总体均值区间估计的 R 软件应用	88

7.4	单正态总体方差区间估计的 R 软件应用	91
7.5	双正态总体均值差区间估计的 R 软件应用	92
7.6	双正态总体方差比区间估计的 R 软件应用	94
7.7	确定样本容量的 R 软件应用	95
7.7.1	估计正态总体均值时样本容量的确定	95
7.7.2	估计比例 p 时样本容量的确定	96
	练习题	97
第 8 章	参数假设检验的 R 软件应用	98
8.1	参数假设检验的基本理论	98
8.2	单个样本 t 检验的 R 软件应用	107
8.3	两个独立样本 t 检验的 R 软件应用	108
8.4	配对样本 t 检验的 R 软件应用	110
8.5	单样本方差假设检验的 R 软件应用	112
8.6	双样本方差假设检验的 R 软件应用	113
	练习题	115
第 9 章	相关分析与回归分析的 R 软件应用	116
9.1	相关分析基本理论	116
9.2	相关分析的 R 软件应用	117
9.3	一元线性回归分析基本理论	118
9.3.1	一元线性回归分析模型	118
9.3.2	一元线性回归的假设	119
9.3.3	方差分析	119
9.3.4	决定系数	120
9.3.5	估计的标准误	120
9.3.6	回归系数的假设检验	120
9.3.7	回归系数的置信区间	121
9.4	一元线性回归分析的 R 软件应用	121
9.5	多元线性回归分析基本理论	124
9.5.1	多元回归模型	124
9.5.2	方差分析	124
9.5.3	决定系数	125
9.5.4	估计的标准误	125
9.5.5	回归系数的 t 检验和置信区间	125
9.5.6	回归系数的 F 检验	126
9.5.7	虚拟变量	126
9.6	多元线性回归分析的 R 软件应用	127
9.7	多重共线性问题的 R 软件应用	131

9.8	自相关问题的 R 软件应用	134
9.8.1	自相关问题	134
9.8.2	自相关问题诊断	135
9.8.3	自相关问题的解决	135
9.8.4	自相关问题的 R 软件实现	135
9.9	异方差问题的 R 软件应用	137
9.9.1	异方差问题	137
9.9.2	异方差问题诊断	137
9.9.3	异方差问题的解决	137
9.9.4	异方差问题的 R 软件实现	138
9.10	Logistic 回归的 R 软件应用	140
9.10.1	Logistic 回归	140
9.10.2	广义线性回归模型	140
9.10.3	与广义线性回归模型有关的 R 函数: glm()	140
9.10.4	基于二项分布的广义线性模型应用实例	141
9.11	Huber 法和 bisquare 法回归的 R 软件应用	144
9.11.1	线性回归中的几个术语	144
9.11.2	数据描述	145
9.11.3	稳健回归的 R 软件实现	146
	练习题	150

第 3 篇 多元统计分析的 R 软件应用

第 10 章	主成分分析与因子分析的 R 软件应用	153
10.1	主成分分析基本理论	153
10.1.1	主成分分析	153
10.1.2	对主成分分析法进行综合评价特点的讨论	154
10.1.3	主成分分析法综合评价方法的改进——原始数据的无量纲化方法的改进	155
10.1.4	用主成分分析法对被评价对象进行综合评价的实施步骤	156
10.2	主成分分析的 R 软件应用	158
10.3	因子分析基本理论	159
10.3.1	因子分析的基本原理	160
10.3.2	因子分析与主成分分析的异同点	164
10.3.3	用因子分析方法进行综合评价分析的基本步骤	165
10.4	因子分析的 R 软件应用	167
10.4.1	R 软件在科技企业发展评价中的应用	167
10.4.2	R 软件在各省市社会经济发展水平评价中的应用	169
	练习题	172

第 11 章 判别分析与聚类分析的 R 软件应用	173
11.1 判别分析基本理论	173
11.1.1 距离判别分析法	173
11.1.2 Fisher 判别分析	174
11.1.3 数据预处理	175
11.2 判别分析的 R 软件应用	176
11.3 聚类分析基本理论	178
11.3.1 距离的统计量	178
11.3.2 系统聚类法	179
11.3.3 快速聚类法	179
11.4 聚类分析的 R 软件应用	180
练习题	181
第 12 章 R 软件的典型相关分析与对应分析	182
12.1 典型相关分析的基本理论	182
12.2 典型相关分析的 R 软件应用	183
12.3 对应分析的基本理论	185
12.4 对应分析的 R 软件应用	187
练习题	188

第 1 篇 R 软件简介



R 软件下载、安装与启动

1.1 选择 R 软件的理由

R 是统计领域广泛使用的诞生于 1980 年左右的 S 语言的一个分支,可以认为 R 是 S 语言的一种实现。而 S 语言是由 AT&T 贝尔实验室开发的一种用来进行数据探索、统计分析和作图的解释型语言。最初 S 语言的实现版本主要是 S-PLUS。S-PLUS 是一个商业软件,它基于 S 语言,并由 MathSoft 公司的统计科学部进一步完善。后来 Auckland 大学的 Robert Gentleman 和 Ross Ihaka 及其他志愿人员开发了一个 R 系统。由“R 开发核心团队”负责开发。

R 是基于 S 语言的一个 GNU 项目,所以也可以当作 S 语言的一种实现,通常用 S 语言编写的代码都可以不做修改地在 R 环境下运行。R 的语法是来自 Scheme。R 的使用与 S-PLUS 有很多类似之处,这两种语言有一定的兼容性。S-PLUS 的使用手册,只要稍加修改就可作为 R 的使用手册。所以有人说: R 是 S-PLUS 的一个“克隆”。

R 是一个有着统计数据分析功能及强大作图功能的软件系统,是一种新兴的统计学软件、语言、环境,而且它的源代码开放。由于 R 的强大功能和它在统计理论及分析上的优势,近年来在统计、经济、管理、金融等相关领域,受到有关人士的广泛欢迎和关注。虽然笔者使用过 MATLAB、SAS、SPSS 和 Stata 等统计计算方面的软件,但现在 R 是笔者的首选。因为

(1) R 是自由免费软件。它不收取任何费用,但其能力不会比任何同类型的商业软件差。从功能相似的角度来说,R 和 MATLAB 最像。

(2) 通过 R,用户可以和全球一流的统计计算方面的专家进行讨论,它是全世界统计学家思维的最大集中。

(3) R 是彻底的面向对象的统计编程语言。对于在面向对象编程模式年代里的人是非常容易理解和使用的。

(4) R 和其他编程语言/数据库之间有很好的接口。代码整合的时候感觉 R 为用户提供了一系列对象,用其他语言只要调用这些对象就可以了,这对数据整合工作非常有用。

(5) R 浮点运算功能强大。R 可以作为一台高级科学计算器,因为 R 与 MATLAB 一样不需要编译就可执行代码。

(6) R 不依赖于操作系统。R 可以运行在 Windows、UNIX、Linux 和 Macintosh 等操作系统上,它们的安装文件以及安装说明都可以在 CRAN 社区上下载。

(7) R 的帮助功能完善。R 嵌入了非常实用的帮助系统,这个帮助系统随软件所附的 pdf 或 html 帮助文件可以随时通过主菜单打开浏览或打印,通过 help 命令可随时了解 R 所

提供的各类函数的使用方法和例子。

(8) R 作图功能强大。R 内嵌的作图函数能将产生的图片展示在一个独立的窗口中,并能将之保存为各种形式的文件(如 jpg, png, bmp, ps, pdf, emf, pictex, xfig 等)。

(9) R 统计分析能力尤为突出。R 内嵌了许多实用的统计分析函数,统计分析的结果也能被直接显示出来,一些中间结果(如 p 值、回归系数、残差等)既可保存到专门的文件中,也可直接用于进一步分析。R 的部分统计功能整合在 R 语言的底层,但是大部分功能则以包的形式提供。大约有 25 个包和 R 同时发布(被称为“标准包”和“推荐包”),更多的包可以通过网上或其 CRAN 社区(<http://cran.r-project.org>)得到,它们都配有完整的 pdf 帮助文件,且其版本会随 R 新版本的发行得到更新,通过在线(或者下载后)安装加载后就可融入原来的 R 中,实现有针对性的分析。

(10) R 可移植性强。R 程序可以很容易地移植到 S-PLUS 程序中,反之 S-PLUS 的许多过程直接或稍做修改就可用于 R; R 与 MATLAB 有许多相似的地方,如都可作为高级计算器,都可不经过编译直接运行源代码,但是 R 侧重于统计分析,而 MATLAB 侧重于工程,例如信号处理。现在通过 R, MATLAB 程序包可实现两者之间许多功能的共享,具体见程序的说明;许多常用的统计分析软件(如 SAS, SPSS, Stata 等)的数据文件都可读入 R,这样其他软件的数据或分析的中间结果可用于 R,并做出进一步的分析。

(11) R 强大的拓展与开发能力。R 是开发新的交互式数据分析方法一个非常好的工具。可以编制自己的函数来扩展现有的 R 语言,或者制作相对独立的统计分析包。

(12) R 灵活而不死板。一般的软件往往会直接展示分析结果,而 R 则将这些结果都放在一个对象里,所以常常分析执行结束后并不显示任何结果,使用者特别是初学者或非专业人员可能会对此感到困惑,其实这样的特点是非常有用的,因为可以有选择地显示自己感兴趣的结果。而有的软件如 SAS 和 SPSS 会同时显示几个窗口,内容太多会使使用者无从选择和解释。

1.2 R 软件的下载

输入以下网址: <http://cran.r-project.org/> 即可下载 R 软件,界面如图 1-1 所示。

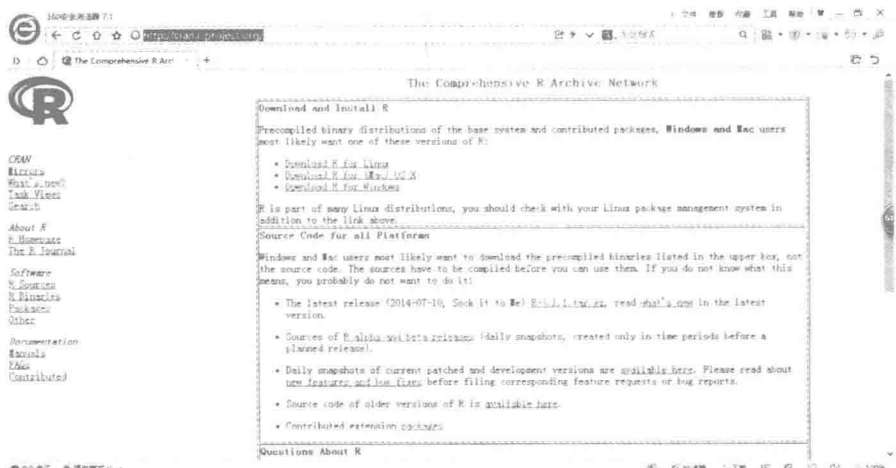


图 1-1 R 软件下载界面

在图 1-1 中,提供了丰富的 R 软件资源,包括 R 简介、R 更新、R 常用手册、R 图书、R 通信和会议信息等。

单击 Download R for Windows,选择要下载的盘和目录,如 E:\R,即可下载所需要的 R 软件。与 SAS 和 SPSS 相比,SAS 和 SPSS 为统计分析等提供了丰富的屏幕输出内容,但 R 给出的屏幕输出内容却很少。它将结果保存在一些合适的对象中以便于用 R 里面的函数做进一步分析。

1.3 R 软件及其程序包的安装

1.3.1 R 软件的安装

R 软件有以下三个版本:

① Linux 版本;② (Mac) OS X 版本;③ Windows 版本。本书使用的是 Windows 版本。R 软件大概每 3 个月更新一次版本。

双击 R 图标,即可得到如图 1-2 所示的界面。

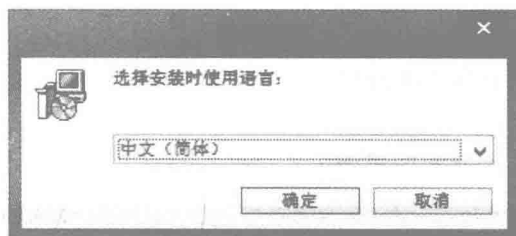


图 1-2 选择安装语言

建议读者选择 English,本书选择中文(简体)。

在图 1-2 中单击确定按钮,得到如图 1-3 所示的界面。



图 1-3 安装向导

单击图 1-3 中的下一步按钮,得到如图 1-4 所示的界面。

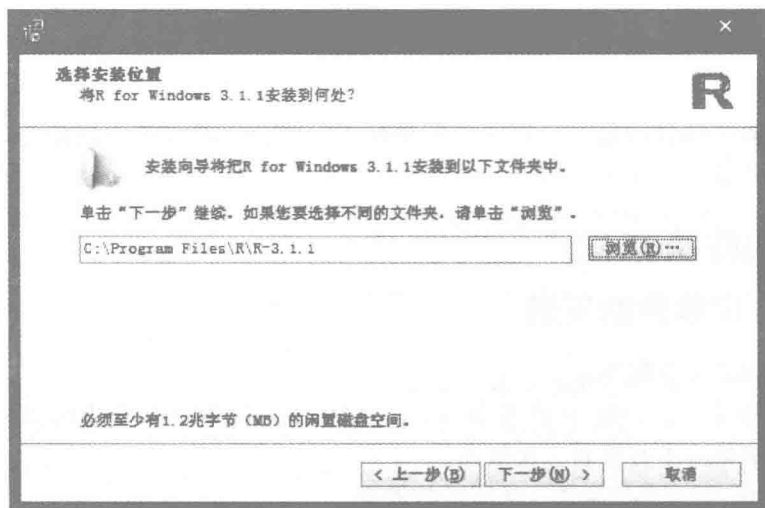


图 1-4 安装向导

选择图 1-4 中的默认文件夹或根据实际需要改变图 1-4 中的文件夹,例如 E:\R-3.1.1,即可安装 R 软件。

安装完后出现如图 1-5 所示的界面。



图 1-5 R 软件图标

1.3.2 R 软件程序包的安装

使用命令: `install.packages("package_name", "dir")` 可以安装所需要的程序包。

`package_name`: 指定要安装的程序包名,请注意大小写。

`dir`: 程序包安装的路径。默认情况下是安装在 `..\library` 文件夹中的。可以通过本参数进行修改,来选择安装的文件夹。

例如：

```
> install.packages("DAAG")
```

即安装了数据分析与图形程序包。

```
> install.packages("fBasics")
```

即安装了 fBasics 程序包,有了这个程序包,就可以求偏度、峰度等。

程序包安装后,如果要使用程序包的功能。必须先把程序包加载到内存中(默认情况下,R启动后默认加载基本包),加载程序包命令:

```
Library("程序包名")
```

```
Require("程序包名")
```

查看程序包帮助可以用:

```
library(help = package_name)
```

主要包括程序包名、作者、版本、更新时间、功能描述、开源协议、存储位置、主要的函数等。

例如:

```
> library(fBasics)
```

即加载 fBasics 加载包。

查看当前环境哪些程序包加载可以用:

```
find.package()或者.path.package()
```

移除程序包出内存可以用:

```
detach()
```

把其他程序包的数据加载到内存中可以用:

```
data(dsname, package = "pkgname")
```

查看这个程序包里的所有数据可以用:

```
data(package = "程序包名")
```

列出所有安装的程序包可以用:

```
library()
```

1.4 R 软件的启动和退出

1. R 软件的启动

双击图 1-5 中两个 R 图标中的任意一个,即可启动 R 软件的交互式用户界面(R-GUI),如图 1-6 所示。R 是按照问答的方式运行的,即在提示命令符>后输入命令并按 Enter 键,R 就完成了相应的操作。