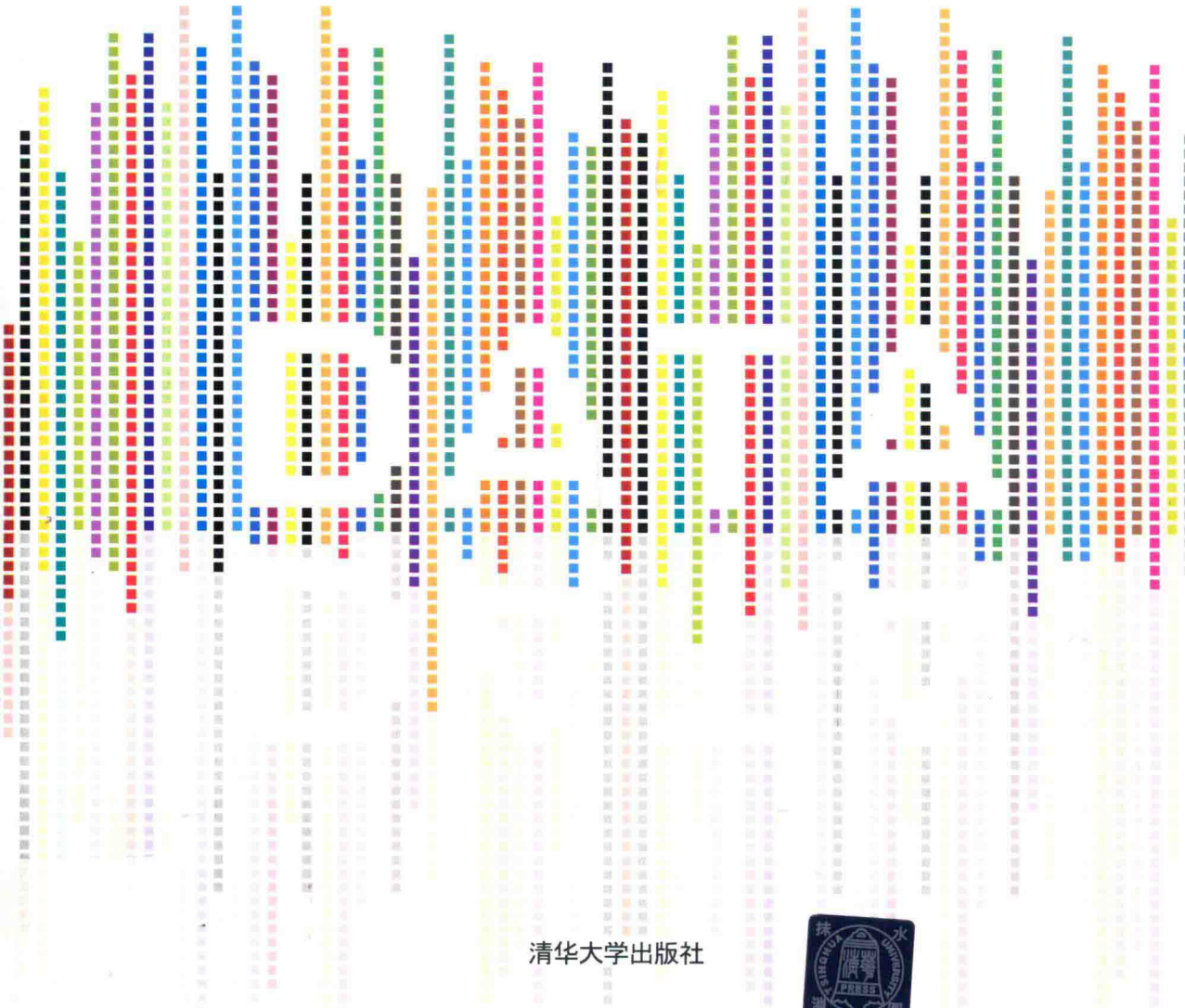
大数据时代经济与金融数据分析系列丛书

数据统计分析的R软件应用

朱顺泉 编著



清华大学出版社

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
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、工商管理等专业的本科高年级学生与研究生使用的实验教材或参考书。

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

前言
FOREWORD

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

目录

CONTENTS

第1篇 R软件简介

第2篇 基础统计分析的R软件应用

第3篇 多元统计分析的R软件应用

第1篇　R软件简介

第1章 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。

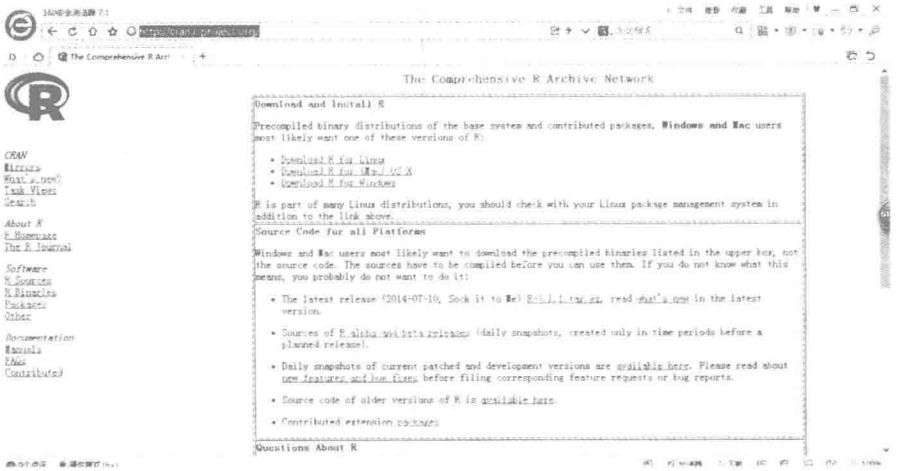


图1-1 R软件下载界面

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；③Windows 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。

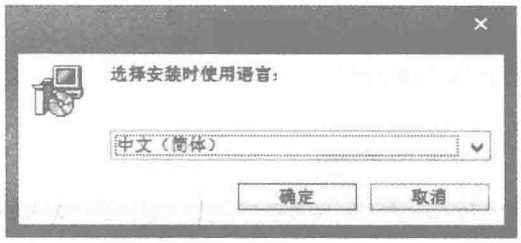


图 1-2　选择安装语言

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。

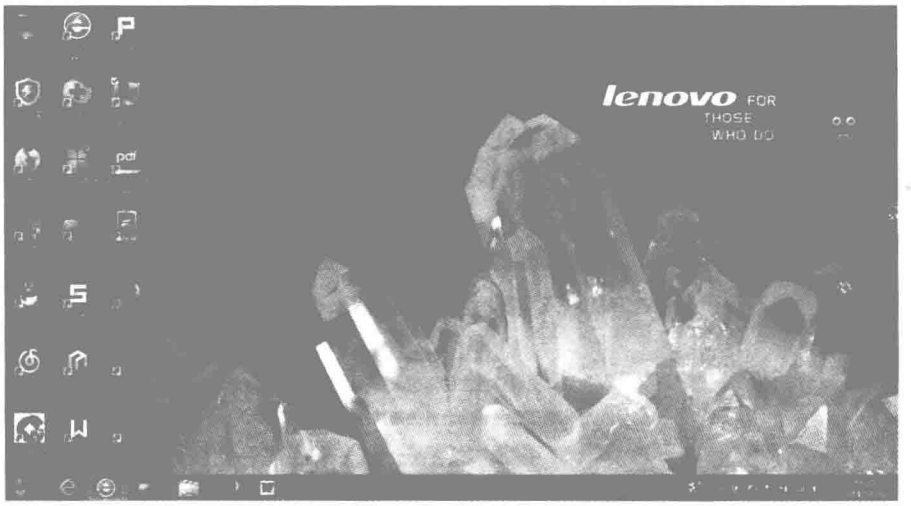


图 1-5　R 软件图标

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)
```

主要内容包括程序包名、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。