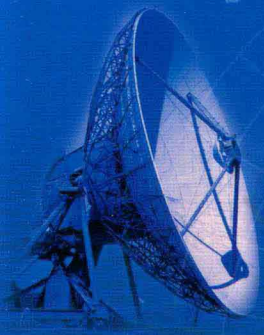
全国高校素质教育教材研究编审委员会审定

R语言统计分析软件教程

R YUYAN TONGJI FENXI RUANJIAN JIAOCHENG

国务院侨办教学改革项目

王斌会★主编



中国教育文化出版社

全国高校素质教育教材研究编审委员会审定
国务院侨办教学改革项目（JYQ0621）

R语言统计分析软件教程

主　编　王斌会
副主编　方匡南　谢佳斌

中国教育文化出版社

图书在版编目（CIP）数据

R 语言统计分析软件教程/王斌会 主编. —中国：中国教育文化出版社，2007.1

ISBN 988-98193-6-8

I. R… II. 王… III. 统计分析—软件教程—计算机人员—参考书 IV. TP31

中国版本图书馆CIP数据核字（2007）

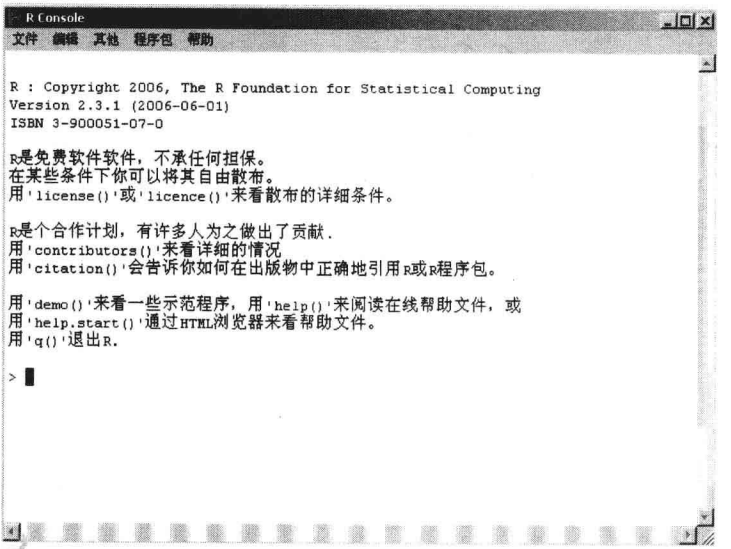
R 语言统计分析软件教程

王斌会 主编

责任编辑：王方玉
封面设计：张骐年
出版发行：中国教育文化出版社
排　　版：科士洁文印中心
印　　刷：新颖印务有限公司
开　　本：850mm×1168mm　1/32
印　　张：6.9375
字　　数：180 千字
版　　次：2007 年 1 月第 1 版
印　　次：2007 年 1 月第 1 次印刷
书　　号：ISBN 988-98193-6-8/G・413
定　　价：16.80 元

地址：北京市海淀区交大东路 62 号西楼 208 室　100044

版权声明



R 是属于 GNU 系统的一个自由、免费、开放源代码的软件，是一个用于统计计算、数据分析和统计制图的优秀工具。

作为一个免费的统计软件，它有 UNIX、LINUX、MacOS 和 WINDOWS 等版本，均可免费下载使用。R 的官方网站是 http://www.r-project.org。在官方网站可以下载到 R 的安装程序、各种外挂程序和文档。在 R 的安装程序中只包含了 8 个基础模块，其他外挂模块可以通过镜像（CRAN）获得（http://cran.r-project.org）。

本书所用的数据和程序可向作者索取 w_b_h@21cn.com

也可到作者网站 www.Qstat.net 下载。

内容提要

随着计算机技术的迅速发展，现代统计方法解决问题能力的深度和广度都有了很大的拓展。而统计软件正是我们应用统计方法不可或缺的工具。统计软件随着计算机技术和统计技术的发展不断推陈出新，名目繁多，各具特色，令人有无所适从之感。随着全球对知识产权保护要求的不断提高，而开放源代码逐渐开始形成另一种市场，R 语言正是在这个大背景下发展起来的，以 S 语言环境为基础的R语言由于其鲜明的特色一出现就受到了统计专业人士的青睐，成为国外大学里相当标准的统计软件。

本书是一本介绍R语言软件基础应用的统计教科书，要求读者有一定的统计知识，并准备应用R语言解决实际问题。本书内容详实、结构清楚、实例丰富、图文并茂，并第一次在国内统计教学中引入大量随机模拟技术。其突出的特点是实用性强，既可作为高校统计学各专业的本科及研究生的教学用书，又可作为研究人员及各类数据分析人员学习的参考书。

前　言

统计学是研究不确定性现象数量规律性的方法论科学，在众多的专业、学科领域中都起着重要的作用，具有很强的应用性，是进行科学研究的一项重要工具，在自然科学、社会科学和经济管理等各领域中得到越来越广泛的应用。随着计算机的普及和统计软件的广泛使用，了解和运用它的人迅速增加。作为数据处理的非常有用的方法，统计学在各个领域都取得了卓有成效的成果。

众所周知，数据的统计分析是以概率统计为基础、应用统计学的基本原理和方法并结合计算机对实际资料和信息进行收集、整理和分析的一门科学。因此，它的原理较为抽象，对学生的数学基础要求也较高，教学中存在着大量的数学公式、数学符号、矩阵运算和统计计算，而且计算量大，手工计算有的几乎无法进行，必须借助于现代化的计算工具。

R 语言是属于 GNU 系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。在目前保护知识产权的大环境下，开发和利用 R 语言对我国统计事业的发展有着非常重大的现实意义。

本书是关于 R 语言的一个入门教材。由于主要针对初学者，我将重点放在了对 R 语言的工作原理的解释上。R 语言涉及广泛，因此对于初学者来讲，了解和掌握一些基本概念及原理是很有必要的。在打下扎实的基础后，进行更深入的学习将会变得轻松许多。本着深入浅出的宗旨，本书将大量配合图表等形式，尽可能使用通俗的语言，使读者容易理解但又不失细节。

本书的特色是：

1．原理、方法、算法和实例分析相结合。鉴于目前计算机统计分析软件已是统计分析应用中不可缺少的工具，本书特别注意各种统计分析的算法实现，使得给出的计算方法更有实用价值。

2．解决统计软件用于统计学教学和科研中存在的问题。国内目前缺乏适合开展统计分析教学科研的统计分析软件，SAS、SPSS、Splus 等统计软件，由于没有版权，需要昂贵的购买费用，更新速度慢，并要大量的维护费用，许多内容与教科书设置不完全一致，使得文科类学生和部分研究人员使用较为困难。

3．提供了一些用于统计分析的 R 语言程序，特别是统计模拟方面的内容，并及时加入了一些现代统计的方法。本书中的所有结果、图形和算法都是由 R 语言软件给出的。

4．研究如何将统计软件的数据处理与统计教学相结合，形成一套完整的教学科研相结合的统计过程。在教学与科研一体化的功能上，在数据编辑、统计分析、统计设计、统计绘图和统计帮助上充分体现多媒体教学的特点。

本书的内容安排吸收了国内外有关统计分析教材的特点，在章节的安排上遵循由浅入深、由简到繁的原则，对统计量和分布进行了较为详细的介绍，增加了许多探索性统计分析的内容和一些统计推断的内容，同时附加了一些数据结构和矩阵运算的概念。书中的主要内容是根据我在暨南大学多年从事统计计算教学的研究基础上编写而成的，包括一些作者多年从事统计教学的心得体会。

本书由王斌会和方匡南、谢佳斌共同完成，其中第 1～5 章由王斌会和方匡南完成，第 6～10 章由王斌会和谢佳斌完成，全书由王斌会统稿。

今年年初，我在日本访问期间，同志社大学的金明哲教授告诉我，即使是在知识产权保护相当完善的发达国家，许多大学也在广泛采用 R 语言进行统计分析和教学，不仅因为它是免费的，而且它是及时更新的（大约每两个月更新一次），更重要的是，它不断吸收最先进的统计技术。所以他主张我在国内开展 R 语言方

面的研究，并积极鼓励我撰写 R 语言学习指导书，介绍 R 语言的特色和优势，于是才形成了这本教程。在此，我衷心地感谢金明哲教授！

感谢暨南大学统计系主任刘建平教授对编写本书的支持和鼓励！

本书是国内第一个用 R 语言软件编写的统计分析教程，由于作者知识和水平有限，书中难免有错误和不足之处，恳请读者批评指正！

王斌会

2006 年 10 月于暨南花园

目　录

第 1 章　关于R语言

1.1　R 语言简介

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。S 语言也是目前比较流行的统计软件 S-PLUS 的基础。

R 语言的创始人为 Ross Ihaka 和 Robert Gentleman，由于这两位“R 之父”的名字都是以 R 开头，所以就称之为 R 语言。

R 语言是一组数据操作、计算和图形显示工具的整合包。相对于其它同类软件，其特色在于：

1．有效的数据处理和保存机制。

2．拥有一整套数组和矩阵的操作运算符。

3．一系列连贯而又完整的数据分析中间工具。

4．图形统计可以对数据直接进行分析和显示，可用于多种图形设备。

5．一种相当完善、简洁和高效的程序设计语言。它包括条件语句、循环语句、用户自定义的递归函数以及输入输出接口。

6．R 语言是彻底面向对象的统计编程语言。

7．R 语言和其它编程语言、数据库之间有很好的接口。

8．R 语言是自由软件，可免费使用，但其功能却不比任何其它同类软件差。

9．R 语言具有丰富的网上资源，更为重要的一点是 R 提供

了非常丰富的程序包，除了推荐的标准包外还有很多志愿者贡献的贡献包，可以直接利用这些包，大大提高工作效率。R 语言的官方网站是 http://www.r-project.org，与 R 语言有关的网站还有 CRAN（镜像），其主站网址是：http://www.cran.r-project.org，相应的中国镜像网是：http://www.lmbe.seu.edu.cn/CRAN/，在这些网站可以下载到很多程序包以及有关 R 语言的资料。

1.2 R 语言和统计

R 语言具有丰富的统计方法，大多数人使用 R 语言是因为其强大的统计功能。不过对 R 语言比较准确的认识是一个内部包含了许多经典统计技术的环境。部分的统计功能整合在 R 环境的底层，但是大多数统计功能则以包的形式提供。大约有 25 个包和 R 同时发布，也被称为标准包，如果想得到更多的其它包，可以在 R 的中国镜像里找到（http://www.lmbe.seu.edu.cn/CRAN/），镜像里还提供了其它比如关于 R 使用的一些资料（详见 10.4）。大多数经典的统计方法和最新的技术都可以在 R 中直接得到，终端用户只要花点时间去寻找就可以了。

R 语言的统计分析过程常常被分解成一系列步骤，并且所有的中间结果都被保存在对象（Object）中，以便使用 R 里面的函数作进一步的分析。虽然 SAS、SPSS 和 Minitab 也提供了丰富的屏幕输出内容，但其中间结果很难在后续分析过程中直接使用。

R 是一套完整的数据处理、计算和绘图软件系统。其功能主要包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言：可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能等。

与其说 R 是一种统计软件，还不如说 R 是一种统计计算的环境，因为 R 语言提供了大量的统计程序，使用者只需指定数据库和若干参数便可进行统计分析。R 语言的思想是：它可以提供一些集成的统计工具，但更大量的是它提供的各种统计计算函数，从而使使用者能灵活地进行数据分析，甚至创造出符合需要的新的统计计算方法。

1.3　R 语言的启动和退出

R 语言的启动很简单，与其它程序相类似，只要双击桌面图标或在开始菜单里的所有程序下打开。R 语言启动后将会出现有关版权、如何获得 R 语言帮助以及如何退出 R 语言的一些说明性文字。图 1-1 是 R 语言的控制台（R Console），R 的主要命令都在这里执行。

> 是 R 语言的命令提示符，等待用户输入命令。

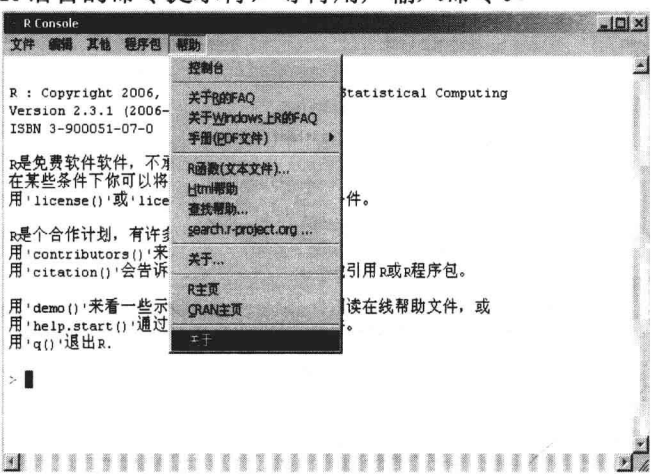


图 1-1　R 语言的控制台界面

R 语言的退出非常简单，可以直接关闭控制台窗口，或是用命令 q() 退出，当退出 R 语言时会提示是否要保存工作环境。如果保存当前的工作环境，下次启动 R 时就会恢复上次的工作环境。

1.4 R 语言的帮助系统

R 语言提供了强大的内置帮助系统，为了得到 R 语言里任何特定名字的函数的帮助，例如想要知道 mean 函数的用法，可以使用如下命令：

```
>help(mean)
```

也可以使用：

```
>?mean()
```

这两个命令是等价的，都会出现 R 帮助窗口，见图 1-2。

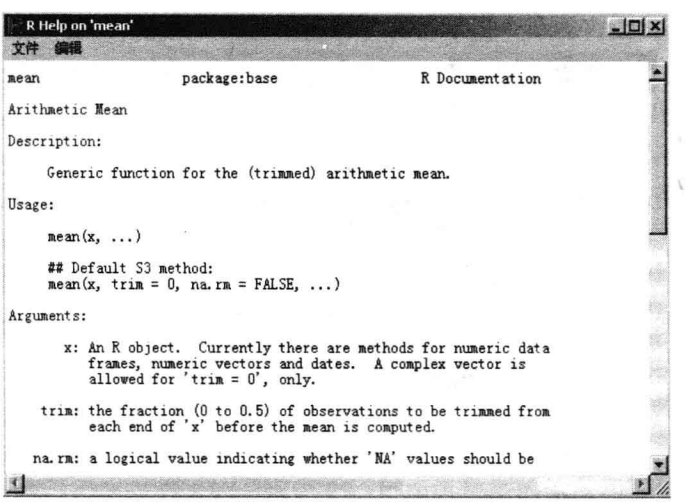


图 1-2 R 语言的帮助界面

除了可以获得 R 语言的内置系统帮助外，也可以利用命令help.start()启动一个网页浏览器，允许你通过超级链接访问帮助页，在网上获得相应帮助。R 的网上帮助主要提供了 R 指南（manuals）、R 参考书（Reference）以及 R 的其它各种资料（miscellaneous material）。R 指南主要提供了 An Introduction to R、Writing R Extensions、The R Language Definition、R Data Import/Export、R Installation and Administration 等的链接，其实这些指南在 R 的控制台帮助菜单下都可以获得（见图 1-3）。R 参考书主要是有关各个包的介绍说明（Packages）、搜索引擎和关键词（Search Engine & Keywords）的链接，其中搜索引擎和关键词的链接特别有用，因为通过它搜索可以使用的函数而提供一个高层次的概念列表。这会让你很快认清自己所处的位置和理解 R 所提供的函数能力范围。

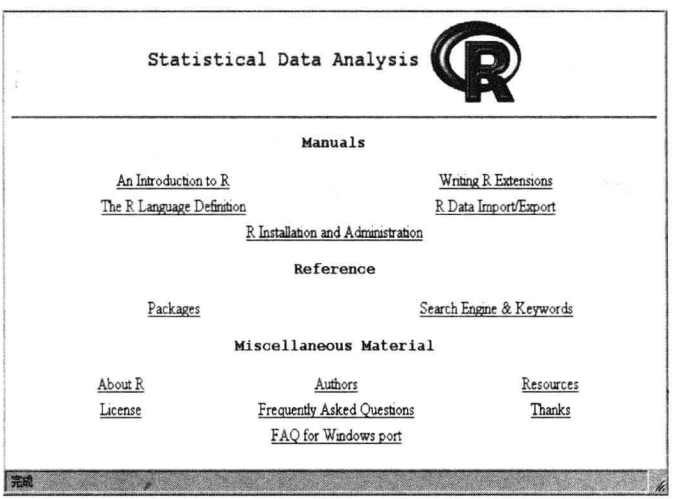


图 1-3　R 语言的在线帮助

有时我们需要知道关于某个函数的使用范例，例如我们想知道 mean()的范例，这时可以用命令：example(mean)

R 提供了如下范例：

```
mean> x <- c(0:10, 50)
mean> xm <- mean(x)
mean> c(xm, mean(x, trim = 0.1))
[1] 8.75 5.50
mean> mean(USArrests, trim = 0.2)
  Murder   Assault   UrbanPop     Rape
    7.42    167.60      66.20    20.16
```

1.5 本书使用的 R 语言版本

R 语言可以在 windows、UNIX、MacOS 等操作系统上使用，本书主要讨论在 windows95 以上的操作系统上使用。

R 语言的版本很多，每隔一段时间就会推出新的版本，本书所使用的软件是 R 2.3.1 的汉化版，是 2006 年 6 月 1 日发布的，ISBN 3-900051-07-0。2006 年 10 月已发布了 2.4.0 版，请读者自行下载使用。

第 2 章　数据对象与运算

2.1　数据对象及类型

2.1.1　数据对象

R 语言创建和控制的实体称为对象（object），它们可以是变量、数组、字符串、函数或者其它通过这些实体定义的更一般的结构（structures）。在 R 语言里，对象是通过名字创建和保存的。在 R 控制台（console）窗口里可以用 ls()命令来查看当前系统里的数据对象。如：

```
> ls()
  character(0)
```

输出结果表示当前系统里没有数据对象，

```
> x<-c(1,2,3,4,5,6)          # 创建变量 x          （# 在 R 中表示注释）
> ls()
  [1] "x"
```

上面列出的是新创建的数据对象 x 的名称。R 对象的名称必须以一个英文字母打头，并由一串大小写字母、数字或句点（.）组成。值得注意的是：R 语言区分大小写，比如 Orange 与 orange 数据对象是不同的。

此外，不要用 R 的内置函数名作为数据对象的名称，比如 c、length 等。

注：R 的内置函数见本章表 2-1，2-2，2-3 等。