



普通高校“十三五”规划教材·统计学系列

# R语言与现代统计方法

刘 强  
裴艳波 © 编著  
张贝贝



清华大学出版社



普通高校“十三五”规划教材·统计学系列

# R语言与现代统计方法

刘 强 裴艳波 张贝贝 © 编著



清华大学出版社  
北京

## 内 容 简 介

本书主要基于统计学、经济金融、生物医学等专业人才培养的基本要求,结合作者多年的教学、科研经验和研究案例,对 R 语言的浩瀚内容进行了有针对性的系统整理,较为全面地介绍了 R 软件的使用.在此基础上,对一些现代统计分析方法的理论及 R 软件的实现进行了系统阐述,力图反映 R 语言最新发展成果.内容涵盖了 R 基础内容、R 数据结构、数据的输入与输出、数据管理与 R 编程、图形管理、概率与抽样分布、探索性数据分析、参数估计、假设检验、方差分析、回归分析、生存分析、贝叶斯计算、时间序列分析、变量选择、非参数回归、缺失数据分析等.本书的主要特点是内容系统全面、深入浅出、详略得当,读者可以根据自己的需求学习相关的章节.

本书既可以作为数理统计、经济统计、经济金融、生物医学等专业本科生、研究生的数据分析教材,也可以作为从事高级统计分析的研究人员和工程技术人员的参考用书.

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

R 语言与现代统计方法/刘强,裴艳波,张贝贝编著. —北京:清华大学出版社,2016  
(普通高校“十三五”规划教材·经济学系列)  
ISBN 978-7-302-45260-7

I. ①R… II. ①刘… ②裴… ③张… III. ①程序语言—程序设计—应用—统计方法—高等学校—教材  
IV. ①C819

中国版本图书馆 CIP 数据核字(2016)第 248321 号

责任编辑:彭欣  
封面设计:汉风唐韵  
责任校对:宋玉莲  
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](mailto:c-service@tup.tsinghua.edu.cn)

质 量 反 馈:010-62772015, [zhiliang@tup.tsinghua.edu.cn](mailto:zhiliang@tup.tsinghua.edu.cn)

课 件 下 载: <http://www.tup.com.cn>, 010-62770175-4506

印 装 者:北京密云胶印厂

经 销:全国新华书店

开 本:185mm×260mm 印 张:23.5 字 数:511千字

版 次:2016年11月第1版 印 次:2016年11月第1次印刷

印 数:1~4000

定 价:49.00元

产品编号:063598-01

# 前言

R 语言作为一种免费的开源统计软件，已经在统计学、运筹学、生物信息学、经济学、金融学以及工程技术等诸多领域得到广泛应用。翻开任何一本关于 R 语言的书籍，都会看到 R 语言有着众多优势：完全免费、完备的统计分析功能、交互式数据分析平台、强大的制图功能、无与伦比的帮助系统以及无私奉献的 R 语言开发核心团队有力支撑等，这使得 R 语言成为当前最流行的统计分析软件之一。事实上，R 语言体系庞大，应用领域广阔，程序包种类繁多，完全掌握 R 语言几乎是一项无法完成的挑战。

本教程主要基于统计学、经济金融、生物医学等专业人才培养的基本要求，有针对性地对 R 语言的浩瀚内容进行了系统整理，全面介绍 R 软件的使用。在此基础上，我们对一些现代统计方法的理论与 R 软件的实现进行了系统阐述，尝试编写一本符合高级统计分析人才培养目标的 R 软件学习教程。

R 语言的发展可以称得上是日新月异：一方面体现在使用 R 软件的人越来越多；另一方面也体现在 R 语言的程序包不断地更新换代，R 软件版本的更新速度也越来越快。我们在编写过程中，尽量追随 R 的发展步伐，力图反映 R 语言最新发展成果。

全书内容共分 15 章，其中第 1~5 章由刘强编写，第 6~10 章、12 章由裴艳波编写，第 10~11 章、13~14 章由张贝贝编写，第 15 章由编写组共同编写，最后由刘强负责统一定稿。

在本书的撰写过程中，中国科学院数学与系统科学研究院的陈敏研究员、北京工商大学的曹显兵教授、北京工业大学的薛留根教授、程维虎教授、李高荣教授，昆明理工大学的吴刘仓教授，首都经济贸易大学统计学院的张宝学教授、马立平教授等都给予了极大的支持和热心的帮助。首都经济贸易大学的王琳老师也为本书的编写做了大量工作。本书的编写与出版也得到了清华大学出版社刘志彬主任和彭欣女士的鼓励和关心，在此一并表示衷心的感谢。本书的编写也得到了北京市青年拔尖人才培养计划项目（CIT&TCD201404133）、国家自然科学基金项目（11201315）的资助，在此表示感谢。

由于作者水平有限，书稿中一定存在许多不足甚至错谬之处，欢迎读者和同行批评指正，电子邮件地址：[cuebliuqiang@163.com](mailto:cuebliuqiang@163.com)。

作者



# 教学支持说明



扫描二维码在线填写  
更快捷获取教学支持

尊敬的老师：

您好！为方便教学，我们为采用本书作为教材的老师提供教学辅助资源。鉴于部分资源仅提供给授课教师使用，请您填写如下信息，发电子邮件给我们，或直接手机扫描上方二维码在线填写提交给我们，我们将会及时提供给您教学资源或使用说明。

（本表电子版下载地址：<http://www.tup.com.cn/subpress/3/jsfk.doc>）

## 课程信息

书 名			
作 者		书号 (ISBN)	
开设课程1		开设课程2	
学生类型	<input type="checkbox"/> 本科 <input type="checkbox"/> 研究生 <input type="checkbox"/> MBA/EMBA <input type="checkbox"/> 在职培训		
本书作为	<input type="checkbox"/> 主要教材 <input type="checkbox"/> 参考教材	学生人数	
对本教材建议			
有何出版计划			

## 您的信息

学 校			
学 院		系/专业	
姓 名		职称/职务	
电 话		电子邮件	
通信地址			

清华大学出版社客户服务：

E-mail: [tupfuwu@163.com](mailto:tupfuwu@163.com)

电话：010-62770175-4506/4903

地址：北京市海淀区双清路学研大厦 B 座 506 室

网址：<http://www.tup.com.cn/>

传真：010-62775511

邮编：100084

# 目 录

<b>第 1 章 R 语言简介</b> .....	1
1.1 R 语言的特点 .....	1
1.2 R 语言运行平台 .....	1
1.2.1 工作目录 .....	2
1.2.2 工作空间 .....	3
1.2.3 历史命令 .....	4
1.2.4 帮助系统 .....	4
1.3 R 程序包的安装使用 .....	7
1.3.1 R 软件的下载与安装 .....	7
1.3.2 程序包的安装与加载 .....	7
1.3.3 与程序包有关的一些函数 .....	8
1.4 初识 R 语言 .....	9
1.4.1 对象的命名 .....	9
1.4.2 对象的赋值 .....	9
1.4.3 一个实例 .....	10
<b>第 2 章 R 数据结构</b> .....	12
2.1 向量 .....	12
2.1.1 向量的建立与赋值 .....	12
2.1.2 向量的运算 .....	13
2.1.3 向量的下标运算 .....	14
2.1.4 与向量有关的一些函数 .....	15
2.2 R 的数据结构 .....	16
2.2.1 对象及属性 .....	16
2.2.2 数据对象的类别 .....	17
2.2.3 与对象有关的一些函数 .....	17
2.3 向量问题的扩展 .....	18
2.3.1 字符型向量 .....	18
2.3.2 逻辑型向量 .....	20

2.3.3	复数型向量	20
2.4	因子及其运算	21
2.4.1	因子的建立与赋值	21
2.4.2	与因子运算有关的一些函数	23
2.5	数组与矩阵	23
2.5.1	数组的建立	23
2.5.2	矩阵的建立	24
2.5.3	数组(矩阵)的下标运算	25
2.5.4	数组(矩阵)的运算	25
2.5.5	矩阵的代数运算	26
2.5.6	与矩阵运算有关的一些函数	30
2.6	数据框	31
2.6.1	数据框的建立	31
2.6.2	数据框元素的引用	32
2.6.3	数据框的编辑	33
2.6.4	与数据框有关的一些函数	35
2.7	列表	35
2.7.1	列表的建立	36
2.7.2	列表元素的引用	36
2.7.3	列表元素的相关运算	37
2.7.4	不同数据结构间的转换	38
<b>第 3 章</b>	<b>数据的输入与输出</b>	<b>40</b>
3.1	数据的输入	40
3.1.1	利用键盘录入数据	40
3.1.2	读取固定格式的文本文件	40
3.1.3	读取复杂格式的文本文件	43
3.1.4	读取其他格式的数据文件	44
3.1.5	R 中数据集的读取	47
3.2	数据的输出	48
3.2.1	运行结果的定向输出	48
3.2.2	数据的定向输出	50
3.2.3	图形的定向输出	52
<b>第 4 章</b>	<b>数据管理与 R 编程</b>	<b>53</b>
4.1	缺失值问题	53

4.1.1	缺失值的识别	53
4.1.2	缺失值的处理	54
4.2	日期值问题	55
4.3	变量的整理	56
4.3.1	变量的重新编码	56
4.3.2	变量的重新命名	59
4.4	数据的整理	59
4.4.1	数据的合并	59
4.4.2	数据的排序	60
4.4.3	数据的汇总与重构	61
4.5	控制结构	63
4.5.1	循环结构	63
4.5.2	条件结构	65
4.6	定义自己的函数	66
4.7	R 编程基础	68
<b>第 5 章</b>	<b>R 图形管理</b>	<b>69</b>
5.1	一个引例	69
5.2	graphics 图形系统简介	70
5.2.1	绘图函数概述	71
5.2.2	图形参数	72
5.2.3	图形设备	74
5.3	基本图形的绘制	76
5.3.1	散点图	76
5.3.2	折线图	77
5.3.3	直方图	80
5.3.4	箱线图	84
5.3.5	条形图	88
5.3.6	饼形图	89
5.3.7	透视图	91
5.4	自定义图形	92
5.4.1	添加拟合曲线	92
5.4.2	绘制多图	93
5.4.3	添加图例和标注	94
5.4.4	添加多边形及填充颜色	95



5.5	lattice 绘图系统	96
5.5.1	lattice 包简介	97
5.5.2	lattice 绘图函数	98
5.5.3	lattice 图形的定制	99
<b>第 6 章</b>	<b>概率与抽样分布</b>	<b>104</b>
6.1	常用的概率分布	104
6.1.1	离散分布的分布律	104
6.1.2	连续分布的密度函数	105
6.2	与分布相关函数的 R 实现	109
6.2.1	概率密度函数	109
6.2.2	分布函数	110
6.2.3	分位数函数	111
6.2.4	随机数产生函数	111
6.3	随机抽样	112
<b>第 7 章</b>	<b>探索性数据分析</b>	<b>114</b>
7.1	常用描述统计量	114
7.1.1	数据集中程度的描述	114
7.1.2	数据离散程度的描述	117
7.1.3	数据分布形状的描述	119
7.1.4	两组样本相关性分析	120
7.2	图形描述	122
7.2.1	茎叶图	123
7.2.2	Q-Q 图	124
7.2.3	经验分布函数图	124
<b>第 8 章</b>	<b>参数估计</b>	<b>126</b>
8.1	点估计	126
8.1.1	矩估计	126
8.1.2	极大似然估计	129
8.2	区间估计	133
8.2.1	区间估计的定义	134
8.2.2	正态总体参数的区间估计	134
8.2.3	比率 $p$ 的区间估计	143

<b>第 9 章 假设检验</b> .....	147
9.1 参数假设检验 .....	147
9.1.1 单个正态总体的参数检验 .....	148
9.1.2 两个正态总体参数的检验 .....	151
9.1.3 单总体比率的检验 .....	156
9.1.4 两个总体比率的检验 .....	159
9.2 非参数假设检验 .....	160
9.2.1 单个样本的非参数检验 .....	160
9.2.2 两样本的独立性检验 .....	165
9.2.3 两样本的非参数检验 .....	168
9.2.4 多样本的非参数检验 .....	174
<b>第 10 章 回归分析</b> .....	177
10.1 一元线性回归 .....	177
10.1.1 一元线性回归模型 .....	177
10.1.2 参数估计 .....	178
10.1.3 回归方程的显著性检验 .....	180
10.1.4 预测 .....	182
10.2 多元线性回归 .....	183
10.2.1 回归模型 .....	183
10.2.2 估计 .....	184
10.2.3 回归模型的显著性检验 .....	184
10.2.4 预测 .....	185
10.2.5 自变量的选择 .....	185
10.3 回归诊断 .....	191
10.3.1 高斯-马尔科夫假定的诊断 .....	191
10.3.2 多重共线性的诊断 .....	195
10.3.3 异常值和影响点的诊断 .....	198
10.4 二分类 Logistic 回归 .....	204
10.4.1 回归模型 .....	204
10.4.2 参数的估计 .....	205
10.4.3 模型的预测 .....	206
10.4.4 拟合优度的测度 .....	206
<b>第 11 章 方差分析</b> .....	208
11.1 单因素方差分析 .....	209

11.1.1	单因素方差分析模型	209
11.1.2	单因素方差分析的 R 实现	211
11.1.3	方差齐性检验	213
11.1.4	多重比较	213
11.2	双因素方差分析	214
11.2.1	不考虑交互作用	214
11.2.2	考虑交互作用	216
11.2.3	双因素方差分析的 R 函数和实例分析	218
11.3	方差分析在模型选择中的应用	219
<b>第 12 章</b>	<b>生存分析</b>	<b>221</b>
12.1	R 程序包的载入及生存对象的建立	222
12.2	非参数建模方法	223
12.2.1	KM 估计	223
12.2.2	生存曲线的比较	226
12.3	参数建模方法	227
12.4	半参数模型方法	229
<b>第 13 章</b>	<b>贝叶斯计算</b>	<b>233</b>
13.1	贝叶斯统计推断的基本概念	233
13.1.1	贝叶斯公式	233
13.1.2	参数估计	234
13.1.3	假设检验	234
13.1.4	预测	235
13.2	单参数模型	235
13.2.1	离散先验分布	236
13.2.2	贝塔先验分布	237
13.2.3	直方图先验	239
13.3	多参数模型	241
13.3.1	均值和方差均未知的正态模型	241
13.3.2	多项模型	242
13.4	蒙特卡洛抽样方法	244
13.4.1	拒绝抽样	245
13.4.2	重要性抽样	248
13.5	马尔可夫链-蒙特卡洛抽样方法	252
13.5.1	马尔可夫链	252

13.5.2	Metropolis-Hastings 算法	255
13.5.3	Gibbs 抽样方法	260
<b>第 14 章</b>	<b>时间序列分析</b>	<b>266</b>
14.1	时间序列的探索性分析	266
14.1.1	时间序列实例	266
14.1.2	传统的分解方法	272
14.2	时间序列的相关概念与简单时序模型	274
14.2.1	平稳性	274
14.2.2	可逆性	275
14.2.3	自协方差函数和自相关函数	275
14.2.4	白噪声模型	276
14.2.5	随机游走	277
14.3	自回归移动平均模型	279
14.3.1	AR 模型	279
14.3.2	MA 模型	283
14.3.3	ARMA 模型	285
14.4	非平稳时间序列模型	295
14.4.1	带漂移的随机游走过程	295
14.4.2	ARIMA 模型	296
14.4.3	单位根检验	296
14.5	季节模型	299
14.5.1	季节 ARMA 模型	299
14.5.2	航空模型	299
14.5.3	乘法季节模型	302
14.6	条件异方差模型	307
14.6.1	模型的结构与模型的建立	307
14.6.2	ARCH 模型	310
14.6.3	GARCH 模型	313
<b>第 15 章</b>	<b>统计方法进阶</b>	<b>317</b>
15.1	非参数密度估计	317
15.1.1	直方图法	317
15.1.2	核密度估计法	320
15.1.3	窗宽的选择方法	323
15.2	多重借补方法	326

15.3	Bootstrap 方法	330
15.4	EM 算法	333
15.5	变量选择	339
15.5.1	岭回归	339
15.5.2	Lasso 方法	342
15.5.3	自适应 Lasso 方法	346
15.5.3	SCAD 方法	347
<b>附录 A</b>	<b>图形用户界面</b>	<b>349</b>
A.1	R Commander	349
A.2	RStudio	353
A.2.1	RStudio 的安装	353
A.2.2	RStudio 界面介绍	353
A.2.3	尝试使用 RStudio	356
<b>附录 B</b>	<b>数据集</b>	<b>359</b>

# 第 1 章

## R 语言简介

R 软件最初是由奥克兰 (Auckland) 大学的 Robert Gentleman 和 Ross Ihaka 及其他志愿者在 1997 年前后开发的一个统计分析系统。其前身是贝尔实验室所开发的 S 语言。R 语言现在由 R 开发核心小组 (R Development Core Team) 维护, 他们的开发维护完全出于自愿, 将全球优秀的统计软件打包提供给大家共享。R 软件免费下载网址: <http://www.r-project.org/> 或 <http://CRAN.R-project.org>。

### 1.1 R 语言的特点

自诞生至今不到 20 年的时间, R 语言已经成为全球众多统计学者和统计工作者的首选统计分析软件。R 语言最大的特点或优势在于: 它是一款免费的统计计算软件, 并有着强大的软件维护和扩展团队。R 语言的主要特点还包括:

(1) 不受操作系统的限制。R 语言可以在 Windows, UNIX, Macintosh 操作系统上运行, 这就意味着 R 语言几乎可以在任何一台计算机上运行。本书主要基于 Windows 操作系统上 R 软件的使用进行介绍。

(2) R 语言是一种解释性的编程语言。它同 Matlab 一样, 不需要编译即可执行代码。

(3) 拥有完善的帮助系统。R 软件内嵌一个非常实用的帮助系统: 包括随软件所附的 pdf 帮助文件 (An Introduction to R) 和 Html 帮助文件。另外, 通过 help 命令可以随时了解 R 软件所提供的各类函数的使用方法。

(4) 具有强大的绘图系统。R 支持的主要图形系统有: 基础图形 (base)、网格图形 (grid)、lattice 图形和 ggplot2。这些系统使得数据可视化更为便捷。此外, R 软件生成的图形文件可以保存为各种形式的文件 (jpg, png, bmp, ps, pdf, emf, xfig, pictex 等), 有利于进一步分析与使用。

(5) 具有强大的统计分析功能。R 语言的部分功能 (大约 25 个程序包) 嵌入在 R 语言底层, 其他都能以 Package 形式下载, 几乎涵盖了现有的全部统计分析方法。

### 1.2 R 语言运行平台

R 软件的运行平台为: RGui (graphic user's interface)。启动 R 软件, 我们看到 RGui, 即图形用户界面的主窗口, 见图 1.1。



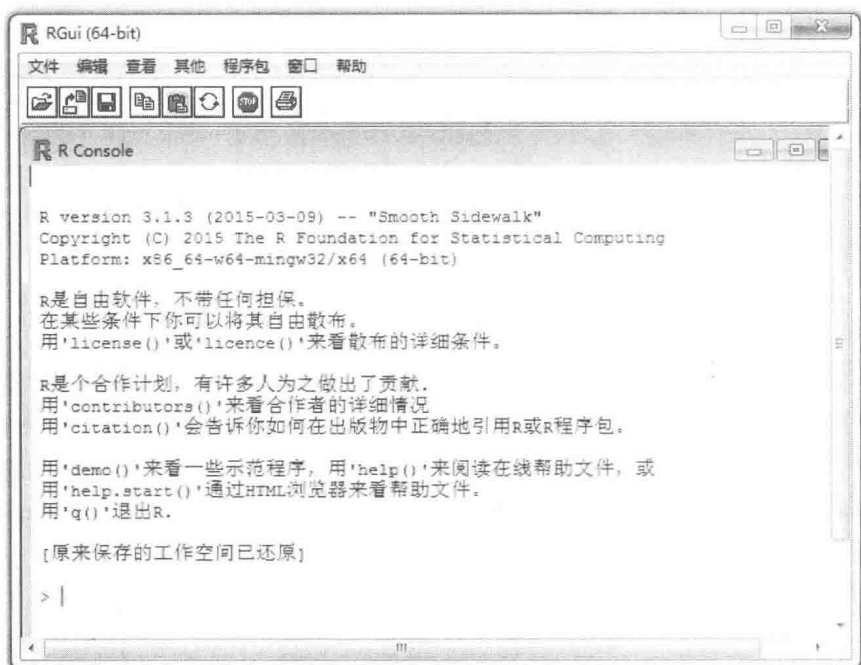


图 1.1 R 软件的运行平台：RGui

R 软件的运行平台 RGui 由三部分组成：主菜单、工具条、R Console（R 语言运行窗口）。

R Console：绝大部分工作都是通过在这里发布命令来完成的，包括数据集的建立、数据的读取、作图等，在这里也可以得到在线帮助。

### 1.2.1 工作目录

工作目录是 R 语言数据输入输出的默认位置，默认状态下是软件安装时的目录。R 的很多操作，包括读写数据，打开、保存脚本文件，读取保存工作空间的镜像等都是在工作目录中进行的。为管理方便，在首次运行 R 软件前，可以建立一个自己的目录，启动 R 后将工作目录改变到自己的目录上。在 Windows 版本中，更改工作目录可以利用菜单方式，选择“文件”|“改变工作目录”选项，选择自己的目录即可，见图 1.2。

另外，也可以利用 `getwd()` 命令获得当前工作目录，并直接利用 `setwd()` 命令改变当前工作目录。例如：

```
> getwd()
[1] "C:/Users/tongji/Documents"
> setwd("C:/Users")
> getwd()
[1] "C:/Users"
```

需要说明的是，在 Windows 操作系统中，以不同的方式打开 R 软件，如通过桌面快捷方式或双击文档中的.RData 文件运行 R 软件，其工作目录可能会不同。因此，每次运行 R 软件时，需要注意工作目录问题。

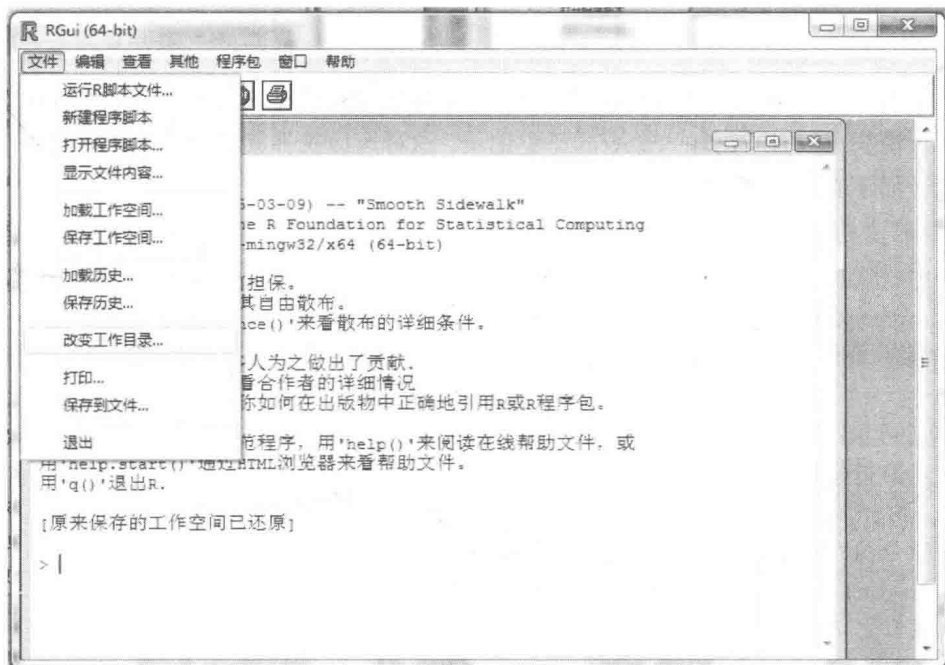


图 1.2 主窗口文件菜单

命令 `list.files()` 或者 `dir()` 可以用来显示当前工作目录中的所有文件和文件夹。例如：

```
> list.files()
[1] "360js Files"      "Adobe"           "desktop.ini"     "Downloads"
[5] "My eBooks"       "My Music"        "My Pictures"     "My Videos"
[9] "save_data.Rdata" "SPSS_data.sav"
```

可以看到，在当前工作目录下，共有 10 个文件或文件夹。

### 1.2.2 工作空间

对于初学者而言，工作空间（workspace）可以理解为 R 当前的工作环境或工作场所，它存储着在运行 R 时所定义的变量、向量、矩阵等所有的对象与函数。很多时候我们希望在下次运行 R 时能够继续以前的工作，这时只需将工作空间保存到一个镜像中，下次运行 R 时载入工作空间镜像即可。

工作空间存放在当前工作目录下的一个后缀名为 `.RData` 的文件中，当启动 R 时，工作空间将自动创建。当直接单击运行窗口 R Console 中的“关闭”按钮或利用命令

```
>q()
```

退出 R 时，系统将提示我们是否需要保存工作空间。

如果我们想在不退出 R 时保存工作空间，可以选择“文档”|“保存工作空间”选项（见图 1.2）或利用命令

```
>save.image()
```

来保存。以后运行 R 时可以通过选择“文件”|“加载工作空间”选项（见图 1.2）或命令

```
>load()
```

加载，进而可以继续前一次的工作。

### 1.2.3 历史命令

在运行 R 时，我们往往在运行窗口 R Console 中交互式输入很多条命令。使用上行箭头或下行箭头可以查看已输入命令的历史记录，这样可以选择某条命令进行适当修改后再次运行，而不必烦琐地重复录入。

选择“文档”|“保存历史”选项可以将运行窗口中的所有记录保存到后缀名为 .RHistory 的文件中；选择“文档”|“加载历史”选项（见图 1.2），可以载入历史命令。利用函数

```
>history()
```

也可以显示最近使用过的命令，默认值为最近的 25 条。也可以自由定制显示更多条，例如：

```
>history(50)
```

可以显示最近使用过的 50 条命令；利用命令

```
> savehistory("myhistory")
```

可以将命令保存在文件名为 myhistory.RHistory 的文件中；命令

```
> loadhistory("myhistory")
```

将载入文件名为 myhistory.RHistory 的命令历史。

### 1.2.4 帮助系统

学习并较好地掌握一门语言或软件，快捷方便的帮助系统是其关键。R 软件提供了十分强大的帮助系统，见图 1.3。

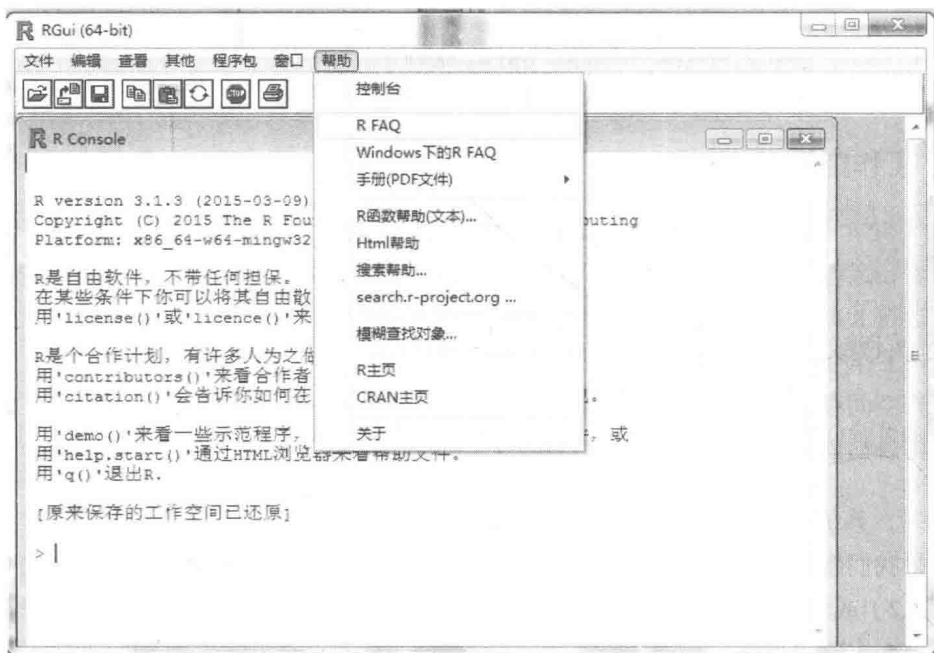


图 1.3 R 中的“帮助”菜单

(1) 在 R 用户界面中，“帮助”菜单中的“R FAQ”选项（见图 1.3）给出了关于 R 软