

□ 应用统计学丛书

Statistical Analysis with R

R 语言与统计分析

汤银才 主编



高等教育出版社
HIGHER EDUCATION PRESS

□ 应用统计学丛书

Statistical Analysis with R

R 语言与统计分析



高等教育出版社
HIGHER EDUCATION PRESS

图书在版编目(CIP)数据

R 语言与统计分析/汤银才主编. —北京:高等教育出版社,2008.11

ISBN 978-7-04-025062-6

I. R… II. 汤… III. ①程序语言—程序设计②统计分折 IV. TP312 C813

中国版本图书馆 CIP 数据核字(2008)第 147944 号

策划编辑 王丽萍
版式设计 王艳红

责任编辑 董达英
责任校对 俞声佳

封面设计 王凌波
责任印制 韩刚

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120
总 机 010-58581000

经 销 蓝色畅想图书发行有限公司
印 刷 北京外文印刷厂

开 本 787×1092 1/16
印 张 24.5
字 数 460 000

购书热线 010-58581118
免费咨询 800-810-0598
网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.landaco.com>
<http://www.landaco.com.cn>
畅想教育 <http://www.widedu.com>

版 次 2008 年 11 月第 1 版
印 次 2008 年 11 月第 1 次印刷
定 价 38.00 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换。

版权所有 侵权必究

物料号 25062-00

前 言

统计学的任务是研究有关收集、整理、分析数据,从而对所考察的问题作出统计推断。作为一门科学,统计学有其坚实的理论基础,研究统计学方法的理论基础问题的那一部分,构成了所谓数理统计学的内容。其次,统计学就其本质来讲,是一门实用性很强的科学,它在人类活动的各个领域有着广泛的应用。因此数理统计的理论与方法应该与实际相结合,解决社会、经济、工农业生产、生物制药、航空航天、质量管理、环境资源等领域中的各种问题。最后,统计学又是一门技术性很强的科学,由于所研究的问题越来越复杂、变量之间关联性越来越强、数据的规模越来越大,使得原有的计算方法无法顺利实现。现在,随着计算机的不断发展与普及,特别是近 20 年来统计计算的突破性进展及统计软件的不完善和成熟,使得解决这些问题不仅成为可能,而且越来越容易、快速。

目前许多大学中几乎所有理工科、甚至文科的许多专业都开设了《数理统计》或《应用统计》之类的课程,有的还编写了相应的教材,这是可喜的。这些课程与教材的共同特点是以较大的篇幅介绍数理统计的理论、方法与实际背景,并配有一定数量的例子和习题。部分学校还为有统计专业和应用数学专业的学生开设 SAS 或 MATLAB 统计软件课程,为经济统计专业的学生开设 SPSS 或 EViews 统计软件课程,但这还远远不够。

作者长期从事概率论与数理统计、统计计算及统计软件的教学工作,我们发现目前的统计教学普遍存在的问题有:一、关于教学内容:在有限的课时中,对于非统计专业的学生采用针对统计专业学生的教学方式,过多强调理论的重要性,从而忽视了统计思想和数据处理能力的培养;有的因为仅用一学期(54 课

时或更少) 讲授概率论与数理统计, 面面俱到的概率论教学使学生无法学到诸如回归分析与方差分析的重要内容。二、关于软件教学: 由于没有软件支持, 使用传统的教学方法和教材, 无论是老师讲解例题, 还是学生完成习题都要花费大量的时间进行手工计算, 且错误率高。使用软件可使数据分析更具直观性、灵活性和可重复性, 可起到举一反三的作用, 提高学生的学习兴趣 and 动手 (操作或编程) 能力。三、关于统计教学与软件教学是否分开: 统计教学与软件教学分开会产生一定的重复性, 从而浪费有限的教学课时, 降低学习的效率。分开的教学会使大部分非统计专业的学生不能得到统计软件操作和数据分析能力的培养。有了统计软件, 可大大增加教学的信息量, 将节省下来的时间用于培养学生统计软件的上机操作能力; 有了统计软件, 使得大规模或海量数据分析和精确计算成为可能, 也使教材中的许多附表 (如常用分布的分位数表) 失去其必要性。四、关于 R 软件: 本书之所以采用 R 软件, 主要原因是其具有强大的数据的图形展示和统计分析功能, 可以免费使用和更新, 同时又有大量可随时加载的有针对性的软件包。而 SAS、MATLAB、SPSS、EViews 却都是收费软件, 与 R 功能几乎相同的 S-PLUS 也是收费的。R 高效的代码、简洁的输出和强大的帮助系统使得在统计软件辅助下的统计教学成为可能。基于 R 开发的菜单式驱动的图形界面工具 R Commander 和 PMG (见附录 B) 使得基础统计分析像 SPSS 一样容易实现。

本书介绍了 R 的基本功能、常用的数据处理与分析方法及它们在 R 中的实现。全书共分十一章及三个附录: 第一章: R 介绍, 介绍了 R 软件的功能与安装。第二章: R 的基本原理与核心, 简明扼要地介绍了 R 软件的使用方法, 主要侧重于不同类型的数据的操作与函数的使用。第三章: 概率与分布, 介绍了常用的离散与连续型分布及 R 中有关的四类函数: 分布函数、概率函数、分位数函数和随机数生存函数。第四章: 探索性数据分析, 介绍了单组和多组数据中特征量的提取方法及数据的图形展示方法。第五章: 参数估计, 主要介绍了单总体与两总体正态及二项分布参数的点估计与区间估计。第六章: 参数的假设检验, 主要介绍了单总体与两总体正态及二项分布参数的假设检验。第七章: 非参数的假设检验, 主要介绍了常用的几种非参数检验方法。第八章: 方差分析, 主要介绍了多组数据比较的单因子与双因子方差分析及协方差分析方法。第九章: 回归分析与相关分析, 介绍了随机变量之间相关性的度量与回归分析及诊断方法。第十章: 多元统计分析介绍, 介绍了多元分析中常用的主成分分析、因子分析、判别分析、聚类分析、典型相关分析及对应分析方法。第十一章: 贝叶斯统计分析, 介绍了贝叶斯分析中单参数与多参数模型、分层模型及回归模型的分析方法。最后是附录, 附录 B 介绍了基于 R 开发的基础统计分析的菜单式工具 R Commander 和 PMG, 附录 C 介绍了 R 的 3 个编程环境: R WinEdt、Tinn-R 及 SciViews-R。全书的所有程序都在 R 的 2.6.0 版本上调试通过, 原则上在其他版本上也可以运行。

本书的特点是:注重统计思想、实用性和可操作性。我们在内容的设计上尽可能简化统计理论与方法的推导过程,对于主要的统计知识都通过一个具体例子展开、讲清要解决问题的思想、方法和具体的实现过程。所有方法的实现都给出了相应的 R 函数的调用格式,而例子讲解的 R 程序都全部嵌入在正文中,便于读者举一反三,解答习题或进行其他类似的数据分析。

本书可作为各专业本科生、研究生数理统计或应用统计课程的基础教材或实验教材,也可作为从事数据统计分析研究人员、工程技术人员的工具书或参考读物。本书整个教材的教学安排可考虑以 1:3 的比例安排上机时间。教学内容可根据需要进行取舍,具体可参考下表安排课时:

教学内容	选取章节	课时安排
R 语言入门	第一章、第二章、附录 B	12
探索性数据分析	第三章、第四章	12
数据统计分析	第五章、第六章、第八章、第九章	24
选讲内容	第七章	8
	第十章	8
	第十一章	8

本书编写过程中,参考了大量的资料文献,得到了华东师范大学金融与统计学院全体老师,特别是终身教授茆诗松老师的支持。我的学生魏晓玲参与了本书第四章和第五章初稿的编写工作,徐安察参与了本书第六章和第七章初稿的编写工作,于巧丽参与了本书第八章和第九章初稿的编写工作,岳映婕参与了本书第十一章初稿的编写工作,上海师范大学的朱杰老师参与了本书第十章的编写工作。在本书的编写过程中,得到了高等教育出版社领导和自然科学学术著作分社王丽萍女士的关心和帮助,在此一并表示感谢。

由于编者水平有限,书中一定存在不足甚至错误之处,欢迎读者不吝指正。

电子邮件地址: yctang@stat.ecnu.edu.cn。

作者

2008 年 5 月

目 录

第一章 R 介绍	1
§1.1 S 语言与 R	1
§1.2 R 的特点	2
§1.3 R 的资源	3
§1.4 R 的安装与运行	3
1.4.1 R 软件的安装、启动与关闭	3
1.4.2 R 程序包的安装与使用	4
第一章习题	5
第二章 R 的基本原理与核心	7
§2.1 R 的基本原理	7
§2.2 R 的在线帮助	9
§2.3 一个简短的 R 会话	11
§2.4 R 的数据结构	17
2.4.1 R 的对象与属性	17
2.4.2 浏览对象的信息	20
2.4.3 向量的建立	21
2.4.4 数组与矩阵的建立	29
2.4.5 数据框 (data frame) 的建立	37
2.4.6 列表 (list) 的建立	42

2.4.7 时间序列 (ts) 的建立	43
§2.5 数据的存储与读取	45
2.5.1 数据的存储	45
2.5.2 数据的读取	46
§2.6 R 的图形功能	50
2.6.1 绘图函数	51
2.6.2 低级绘图命令	52
2.6.3 绘图参数	54
2.6.4 一个实例	56
§2.7 R 编程	62
2.7.1 循环和向量化	63
2.7.2 用 R 写程序	64
2.7.3 编写你自己的函数	65
2.7.4 养成良好的编程习惯	67
第二章习题	68
第三章 概率与分布	70
§3.1 随机抽样	70
§3.2 排列组合与概率的计算	71
§3.3 概率分布	72
3.3.1 离散分布的分布律	72
3.3.2 连续分布的密度函数	74
§3.4 R 中内嵌的分布	79
§3.5 应用: 中心极限定理	81
3.5.1 中心极限定理	81
3.5.2 渐近正态性的图形检验	81
3.5.3 举例	83
第三章习题	87
第四章 探索性数据分析	89
§4.1 常用分布的概率函数图	89
§4.2 直方图与密度函数的估计	97
4.2.1 直方图	97
4.2.2 核密度估计	97

§4.3 单组数据的描述性统计分析	99
4.3.1 单组数据的图形描述	99
4.3.2 单组数据的描述性统计	104
§4.4 多组数据的描述性统计分析	107
4.4.1 两组数据的图形概括	107
4.4.2 多组数据的图形描述	112
4.4.3 多组数据的描述性统计	114
4.4.4 分组数据的图形概括	118
§4.5 分类数据的描述性统计分析	125
4.5.1 列联表的制作	125
4.5.2 列联表的图形描述	127
第四章习题	129
第五章 参数估计	132
§5.1 矩法估计和极大似然估计	132
5.1.1 矩法估计	132
5.1.2 极大似然估计	135
§5.2 单正态总体参数的区间估计	137
5.2.1 均值 μ 的区间估计	138
5.2.2 方差 σ^2 的区间估计	142
§5.3 两正态总体参数的区间估计	143
5.3.1 均值差 $\mu_1 - \mu_2$ 的置信区间	143
5.3.2 两方差比 σ_1^2/σ_2^2 的置信区间	146
§5.4 单总体比率 p 的区间估计	148
§5.5 两总体比率差 $p_1 - p_2$ 的区间估计	151
§5.6 样本容量的确定	153
5.6.1 估计正态总体均值时样本容量的确定	153
5.6.2 估计比例 p 时样本容量的确定	155
第五章习题	156
第六章 参数的假设检验	159
§6.1 假设检验与检验的 p 值	160
6.1.1 假设检验的概念与步骤	160
6.1.2 检验的 p 值	162

§6.2 单正态总体参数的检验	162
6.2.1 均值 μ 的假设检验	162
6.2.2 方差 σ^2 的检验: χ^2 检验	165
§6.3 两正态总体参数的检验	166
6.3.1 均值的比较: t 检验	166
6.3.2 方差的比较: F 检验	168
§6.4 成对数据的 t 检验	169
§6.5 单样本比率的检验	171
6.5.1 比率 p 的精确检验	171
6.5.2 比率 p 的近似检验	172
§6.6 两样本比率的检验	174
第六章习题	175
第七章 非参数的假设检验	177
§7.1 单总体位置参数的检验	177
7.1.1 中位数的符号检验	178
7.1.2 Wilcoxon 符号秩检验	180
§7.2 分布的一致性检验: χ^2 检验	182
§7.3 两总体的比较与检验	185
7.3.1 χ^2 独立性检验	185
7.3.2 Fisher 精确检验	187
7.3.3 Wilcoxon 秩和检验法和 Mann-Whitney U 检验	189
7.3.4 Mood 检验	191
§7.4 多总体的比较与检验	193
7.4.1 位置参数的 Kruskal-Wallis 秩和检验	193
7.4.2 尺度参数的 Ansari-Bradley 检验	195
7.4.3 尺度参数的 Fligner-Killeen 检验	196
第七章习题	197
第八章 方差分析	200
§8.1 单因子方差分析	200
8.1.1 数学模型	200
8.1.2 均值的多重比较	204
8.1.3 同时置信区间: Tukey 法	206
8.1.4 方差齐性检验	209

§8.2 双因子方差分析	211
8.2.1 无交互作用的方差分析	211
8.2.2 有交互作用的方差分析	215
§8.3 协方差分析	219
第八章习题	224
第九章 回归分析与相关分析	229
§9.1 相关性及其度量	229
9.1.1 相关性概念	229
9.1.2 相关分析	230
§9.2 一元线性回归分析	233
9.2.1 数学模型	233
9.2.2 估计与检验	235
9.2.3 预测与控制	238
9.2.4 计算例子	240
§9.3 多元线性回归分析	243
9.3.1 数学模型	243
9.3.2 估计与检验	243
9.3.3 预测与控制	246
9.3.4 计算例子	247
§9.4 回归诊断	251
9.4.1 残差分析	251
9.4.2 影响分析	258
9.4.3 共线性诊断	262
§9.5 Logistic 回归	265
第九章习题	271
第十章 多元统计分析介绍	277
§10.1 主成分分析与因子分析	278
10.1.1 主成分的简要定义与计算	278
10.1.2 主成分 R 通用程序	279
10.1.3 因子分析的简要定义与计算	282
10.1.4 因子分析 R 通用程序	283
§10.2 判别分析	286
10.2.1 距离判别	286
10.2.2 Fisher 判别法	288

10.2.3 R 通用程序	289
§10.3 聚类分析	292
10.3.1 基本思想	293
10.3.2 R 通用程序	294
§10.4 典型相关分析	297
10.4.1 基本思想	297
10.4.2 R 通用程序	299
§10.5 对应分析	301
10.5.1 基本思想	302
10.5.2 R 通用程序	304
第十章习题	305
第十一章 贝叶斯统计分析	316
§11.1 贝叶斯统计分析与经典统计分析的比较	316
11.1.1 经典统计分析中存在的问题	317
11.1.2 对贝叶斯统计分析的质疑及褒奖	317
§11.2 贝叶斯统计分析与先验分布的选取	318
11.2.1 贝叶斯公式	318
11.2.2 先验分布的选取	320
11.2.3 贝叶斯分析体现了科学探索过程	323
§11.3 单参数贝叶斯统计分析	323
11.3.1 两项分布下的贝叶斯推断	323
11.3.2 正态分布下的贝叶斯统计推断	334
§11.4 多参数贝叶斯统计分析	338
11.4.1 方法概述	338
11.4.2 正态分布参数中的贝叶斯分析	338
11.4.3 随机模拟方法	339
11.4.4 一个实例	340
§11.5 分层贝叶斯统计分析	346
11.5.1 分层模型的建立及其贝叶斯推断	346
11.5.2 $N-N$ 模型与应用	348
§11.6 贝叶斯线性回归分析	357
11.6.1 模型的表示	357
11.6.2 后验分布	358
11.6.3 回归拟合	358
11.6.4 后验预测	358

第十一章习题	362
附录 A 秩与结的介绍	364
附录 B R 的图形界面	366
§B.1 R Commander	366
B.1.1 功能	366
B.1.2 (网络) 安装	366
B.1.3 运行	367
B.1.4 结构与使用	367
§B.2 PMG	368
B.2.1 功能	368
B.2.2 安装	368
B.2.3 结构与使用	368
附录 C R 的编程环境	371
§C.1 R WinEdt	371
C.1.1 (网络) 安装	371
C.1.2 运行	371
C.1.3 R WinEdt 的特点	373
§C.2 Tinn-R	373
§C.3 SciViews R	374
参考文献	376

第一章 R 介绍

本章概要

- ◇ R 的功能与特点
- ◇ R 的安装与运行
- ◇ R 程序包的安装与运行

§1.1 S 语言与 R

R 是一个有着强大统计分析及作图功能的软件系统, 在 GNU 协议 General Public Licence 下免费发行, 最先是由 Ross Ihaka 和 Robert Gentleman 共同创立, 现在由 R 开发核心小组 (R Development Core Team) 维护, 他们的开发完全出于自愿、工作努力负责, 并将全球优秀的统计应用软件打包提供给我们共享.

R 可以看作是贝尔实验室 (Bell Laboratories) 的 Rick Becker, John Chambers 和 Allan Wilks 开发的 S 语言的一种实现或形式. 因此, R 是一种软件也可以说是一种语言. S 语言现在主要内含在由 Insightful 公司经营的 S-PLUS 软件中. 我们可以将 R 和 S-PLUS 视为 S 语言的两种形式, S/S-PLUS 方面的文档都可以直接用于 R, 不过 R 和 S 在设计理念上存在着许多不同, 关于这方面的

详细内容大家可以参考 Ihaka & Gentleman (1996) 或随 R 同时发布的 R-FAQ^[16]. 本书今后主要使用 R, 有时也使用 R 软件、R 语言或 R 系统来称呼这种形式的 S 语言.

§1.2 R 的特点

现在越来越多的人开始接触、学习和使用 R, 因为它有其显著的优点, 主要包括:

- 1) **免费:** 尽管 S-PLUS 是非常优秀的统计分析软件, 但你需要支付一笔费用, 而 R 是一个免费的统计分析软件 (环境);
- 2) **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 同 MATLAB 一样不需要编译就可执行代码;
- 3) **不依赖于操作系统:** R 可以运行于 UNIX, Linux, Windows 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 (Comprehensive R Archive Network) 社区上下载;
- 4) **帮助功能完善:** R 嵌入了一个非常实用的帮助系统——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;
- 5) **作图功能强大:** 其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 (例如 jpg, png, bmp, ps, pdf, emf, pictex, xfig);
- 6) **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, 也可以直接用于进一步的分析.

R 的部分统计功能整合在 R 语言的底层, 但是大多数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”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 (或下载后) 安装并加载后就可融入原来的 R 中, 实现有针对性的分析;

- 7) **可移植性强:**
 - R 程序容易地移植到 S-PLUS 程序中; 反之 S-PLUS 的许多过程直接或稍作修改可用于 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 SPSS, SAS, Stata 及 Excel) 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, 并作出进一步的分析.
- 8) **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. 例如附录 A 介绍的 R Commander 就是一个非常成功的例子. 我们可以编制自己的函数来扩展现有的 R 语言, 或制作相对独立的统计分析包.
- 9) **灵活而不死板:** 一般的软件往往会直接展示分析的结果, 而 R 则将这些结果都存放在一个对象 (object) 里, 所以常常在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, 因为我们可以有选择地显示我们感兴趣的结果. 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所适从.

§1.3 R 的资源

R 的核心开发与维护小组通过 R 的主页, 即 R 工程 (R Project) 网站 (<http://www.r-project.org>) 及时发布有关信息, 包括 R 的简介、R 的更新及宏包信息、R 常用手册、已经出版的关于 R 的图书、R 通讯和会议信息等. 你还可以通过该主页预订邮件, 通过电子邮件发出求助或对求助者提供帮助.

R 的 CRAN 社区是我们获得软件 (及源代码) 和资源的主要场所, 通过它或其镜像站点我们可以下载最新版本及大量的统计程序包 (packages).

本书将使用 Windows(95 及以后) 操作系统上的 R, 其他操作系统上 R 的使用方法请参考 R 相关说明. 除 R 自带的运行平台 R-GUI (R Graphic User's Interface) 外, 本书附录 B 和 C 还提供了 Windows 下几个 R 的运行平台: R-Commander, PMG, R-WinEdt, R-Sciview 和 R-Tinn.

§1.4 R 的安装与运行

1.4.1 R 软件的安装、启动与关闭

R 的安装: 从 CRAN 社区下载最新的封装好的 R 安装程序到本地计算机, 运行可执行的安装文件, 通常缺省的安装目录为 C:\Program Files\R\R-x.x.x, 其中 x.x.x 为版本号. 安装时可以改变目录, 从 2.2.0 以后还可以选择中文作为基本语言, 这样 R 的图形用户界面, 即 RGui 窗口的菜单都是中文的.

R 的启动: 安装完成后点击桌面上 R x.x.x 图标就可启动 R R-GUI. R 是按问答的方式运行的, 也即你在命令提示符 “>” 后键入命令并回车, R 就完成一些操作. 例如输入命令

```
> plot(rnorm(1000))
```

就可得到图 1.1, 此命令的具体含义我们将在后面第二章叙述.

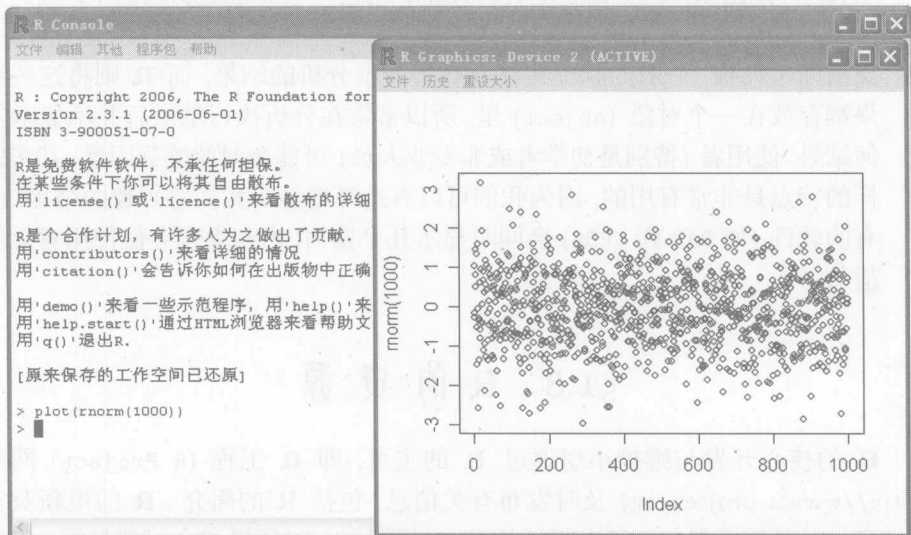


图 1.1 R 的启动

R 的退出: 在命令行键入 `q()` 或点击 RGui 右上角的叉号. 退出时可选择保存工作空间, 缺省文件名为 R 安装目录的 `bin` 子目录下的 `R.RData`. 以后可以通过命令 `load()` 或通过菜单“文件”下的“载入工作空间”加载, 进而继续你前一次的工作.

1.4.2 R 程序包的安装与使用

R 程序包的安装有三种方式:

- 1) **菜单方式:** 在已经联网的条件下, 按步骤“程序包 ⇒ 安装程序包 ⇒ … ⇒ 选择 CRAN 镜像服务器 ⇒ 选定程序包”进行实时安装;
- 2) **命令方式:** 在已经联网的条件下, 在命令提示符后键入


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

 完成程序包 `PKname` 的安装.
- 3) **本地安装:** 在无上网条件下, 先从 CRAN 社区下载需要的程序包及与之关联的程序包, 再按第一种方式通过“程序包”菜单中的“用本机的 zip 文件安