统计应用与实战系列

Yingyong Tongji Fenxi yu
R Yuyan Shizhan

应用统计分析与R语言实战

吕书龙　梁飞豹　/主　编
刘文丽　薛美玉　/副主编

北京大学出版社
PEKING UNIVERSITY PRESS

Yingyong Tongji Fenxi yu
R Yuyan Shizhan

应用统计分析与R语言实战

吕书龙　梁飞豹　/主　编
刘文丽　薛美玉　/副主编

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

应用统计分析与R语言实战/吕书龙，梁飞豹主编. —北京：北京大学出版社，2017. 11
ISBN 978-7-301-28590-9

Ⅰ.①应… Ⅱ.①吕… ②梁… Ⅲ.①统计分析—统计程序—高等学校—教材
Ⅳ. ①C819

中国版本图书馆CIP数据核字（2017）第195757号

书　　名：应用统计分析与R语言实战
著作责任者：吕书龙　梁飞豹　主编
责任编辑：潘丽娜
标准书号：ISBN 978-7-301-28590-9
出版发行：北京大学出版社
地　　址：北京市海淀区成府路205号　100871
网　　址：http://www.pup.cn　　新浪官方微博：@北京大学出版社
电子信箱：zpup@pup.cn
电　　话：邮购部62752015　发行部62750672　编辑部62752021
印 刷 者：三河市北燕印装有限公司
经 销 者：新华书店
787毫米×960毫米　16开本　31.25印张　760千字
2017年11月第1版　2017年11月第1次印刷
定　　价：69.00元

内 容 简 介

本书以应用概率统计方法为主线，引入自由开放的统计软件 R，并按照数理统计的逻辑顺序，逐步展开理论分析、方法介绍和软件实现. 主要内容有：统计软件与概率计算、数理统计初步与模拟计算、参数估计、假设检验、案例的直观分析、方差分析与正交试验设计、回归分析、多元统计初步、R 统计软件、R 软件的图形环境、R 软件中的数学运算. 本书着重介绍统计方法的基本思想、背景、应用条件、实际意义及软件计算，使学生能够对数理统计方法的应用有一个较为系统、全面的了解，同时确保本书提到的所有的统计方法都能在 R 软件中得到快速实现，使学生对统计方法的实现和应用有一个标准参照. 书中各章具有内容小结、知识点网络结构图和适量的习题，书末配有大部分习题的参考答案、附录中给出经典问题的索引页、部分源程序、常见软件包等. 本书编写得到福州大学研究生“应用概率统计”重点课程建设项目和福州大学教务处重点教材建设项目的资助.

本书可作为高等院校工学、经济、管理、农学、医学等非数学类硕士、博士研究生以及高年级本科生学习应用概率统计（或统计分析方法或应用统计分析）课程的教材，也可作为相关学科和工程技术人员的工具参考书. 为了方便教师教学和学生学习，我们将提供基本的课件、数据、程序脚本等，请通过电子邮件 wujispace@fzu.edu.cn 与作者联系.

作 者 简 介

吕书龙，福建闽侯人，理学硕士. 现为福州大学数学与计算机科学学院副教授. 近年参编《应用统计方法》等 3 部教材；在《实验技术与管理》等核心刊物发表论文十余篇. 主持或参与多项省级、校级课题以及横向课题. 目前主要从事概率统计、统计应用与软件研发等教学与研究工作.

梁飞豹，福建莆田人，理学学士. 现为福州大学数学与计算机科学学院副教授. 近年来主编教材《概率论与数理统计》《应用统计方法》等 4 部；在 SCI 等核心刊物上发表学术论文二十余篇. 目前主要从事概率统计的教学与研究工作.

刘文丽，湖北枝江人，理学硕士. 现为福州大学数学与计算机科学学院副教授. 近年参编《概率论与数理统计》等 4 部教材；在 *Expert Systems With Applications* 等核心刊物发表论文十余篇. 主持或参与多项省级、校级课题及横向课题. 目前主要从事概率统计、统计应用与系统工程等教学与研究工作.

薛美玉，福建福清人，理学硕士. 现为福州大学数学与计算机科学学院讲师. 近年参编《应用统计方法》等多部教材；在《福州大学学报》等核心刊物发表论文若干篇，主持或参与多项省级、校级教改课题. 目前主要从事概率统计、时间序列等教学与研究工作.

前　言

本书是根据编者多年来积累的研究生教学经验和经历, 结合各专业研究生的专业特点和科研需求, 并参照教育部 “工学硕士研究生应用概率统计课程教学基本要求” 编写. 本书编写融合概率统计理论及其应用、统计计算及软件模拟, 内容编排由浅入深, 直观易懂. 它可以作为高等院校工学、经济、管理、农学、医学、心理学等非数学类硕士、博士研究生以及高年级本科生学习应用概率统计（或统计分析方法或应用统计分析）课程的教材, 也可作为相关学科和工程技术人员的参考书.

应用统计是关于收集、整理、分析和解释受随机因素影响的数据的一门科学, 是一种认识世界的方法论, 是现代应用数学的一个重要组成部分, 也是统计学一级学科的重要分支. 它以概率论为基础, 以统计软件为实现途径, 旨在探索数据内在的数量规律性, 以达到对客观事物的科学认识. 随着计算机和互联网的发展, 大数据时代的悄然而至以及人们对数据处理的需求和要求的不断提升, 各种统计软件也应运而生, 发展迅猛. 不同学科与应用统计的交叉融合, 使得应用统计在各个领域得到了广泛的应用, 也推动了应用统计的纵深发展. 正因为如此, 应用统计课程成为高等院校非数学类硕士研究生最重要的公共基础课之一, 也是非数学类硕士研究生观测世界所必备的文化修养, 统计软件也成为硕士研究生进行学科深造和科学研究的必备工具之一. 由于这门课程自身的特点, 以及非数学类硕士研究生在本科阶段所学课程和统计软件应用水平的差异, 我们在教材的编写过程中, 力求体现以下几个特点：

(1) 以统计方法为主, 着重介绍各种统计方法的应用背景、统计思想、适用条件、实际意义和软件实现, 使学生能够对统计方法及其应用有一个系统、全面、直观的了解, 以实用为主, 尽量避免纯数学化的论证.

(2) 在内容编排上, 考虑到不同学科和不同程度学生的需要, 以概率论基础及直观模拟作为引入形成基础篇；着重介绍各种基本的统计方法, 同时加入一些应用广泛的统计方法, 如优化方法、非参数方法、回归专题、多元统计方法等形成方法篇；最后专门介绍 R 软件形成软件篇. 不同的安排适合不同的教学模式. 建议第一章和第二章作为数理统计的基础与第九章和第十章作为 R 软件的基础同步授课, 并且早于其他章讲授.

(3) 考虑到专业的差异性, 为了便于对照学习和应用, 各章配备了一定数量的专业性统计案例分析以及完整的 R 代码实现, 方便学生操作实践.

(4) 本书各章都配备适量针对性强的习题, 并给出大部分习题的解答和 R 程序实现, 便于教师教学和学生自学.

(5) 本书附录给出书中各种统计问题的索引页, 便于读者快速检索；附录还给出丰富的 R 命令集函数、网络资源及数据集、常见软件包的介绍等.

本书主要由吕书龙、梁飞豹、刘文丽、薛美玉编写, 并在集体讨论的基础上, 由吕书龙和梁飞豹对全书进行统编、定稿. 在此对所有参与本书编写、校对的人员表示衷心的感谢!

本书在编写、出版的过程中, 得到福州大学研究生院、福州大学教务处以及北京大学出

版社的大力支持；本书列入福州大学高水平建设规划教材，得到福州大学研究生处优质课程建设项目和福州大学教务处重点教材建设项目的资助.

限于编者的水平，书中难免存在不足之处甚至错误，恳请同行和广大读者批评指正.

编 者

2016 年 12 月于福州大学

目　录

基础篇　R 软件与概率统计

方法篇　应用统计分析

软件篇 R 软件

基础篇　R 软件与概率统计

第一章　统计软件与概率计算

本章重点阐述统计软件及其在概率论中的应用. 数理统计是统计学中一个应用极其广泛的学科分支, 它以随机现象中的问题为研究对象并确立为总体, 然后通过对总体进行试验或观察得到数据, 再采用合适的方法对数据进行分析, 最后对总体的客观规律性作出合理的估计或推断. 数理统计是以概率论为理论基础, 以统计软件为实现手段. 统计软件集成了各种统计方法和运算, 可实现数据的预处理与分析、提供报表和图形输出、集成编程和计算环境等, 是进行统计分析不可或缺的工具.

本章第 1 节介绍统计软件的基本使用, 第 2 节通过统计软件把抽象的理论和知识直观化, 以便读者对概率论的内容有个全面且直观的回顾.

1.1　R 软件简介

R 软件是由 Aucklang(奥克兰) 大学的 Robert Gentleman 和 Ross Ihaka 及广大志愿者开发的一套包含数学运算、统计计算与数据分析、图形制作、程序设计等功能的自由软件系统, 可在 Unix/Linux, Windows 和 Machintosh 操作系统上运行. R 软件提供一个开放的、可互动的、可编程的、解释型的统计计算环境. 它支持命令执行、脚本执行和远程执行, 并内嵌一个非常实用的帮助系统和用户手册, 还可外连互联网帮助系统. R 软件完全免费使用, 更新迅速, 使用方便, 其网络资源丰富, 特别是软件包数量庞大、功能丰富.

一、 R 软件的获取与安装

R 软件的官方网站提供最新的 R 软件和各种软件包的下载, 官方网站位于:

http://www.r-project.org

建议到最近的镜像站点下载 R 软件

http://mirrors.xmu.edu.cn/CRAN/

对于 Windows 操作系统上运行的 R 软件, 目前最新的版本是 R-3.3.2(32/64 bit), 于 2016 年 10 月 31 日发布, 大概 70MB 大小. 它支持 Windows 所有的 32 位和 64 位操作系统.

R 软件的安装非常简单, 双击下载的 R-3.3.2-win.exe, 按照提示 (或每次弹出对话框都单击下一步) 即可顺利安装. 安装成功后, 会在 Windows 系统的开始菜单和桌面上出现 R 的图标, 双击该图标即可运行 R 软件.

启动 R 软件后, 出现的是 R 软件的图形界面, 包括主菜单、工具栏和一个窗口, 其标题为 "R Console", 如图 1.1.1 所示. 该窗口可接受用户的命令输入, 同时也输出执行命令后的结果, 但有些结果会显示在新建的窗口中 (特别是图形). 界面中 ">" 表示命令提示符, 闪烁

的 "|" 符号表示光标, 出现它们意味着可以输入 R 命令了, 或者说上一条命令执行完成, 可继续下一条命令了. 该窗口是用户通过命令交互来使用 R 软件的主要场所. 用户输入命令时如果命令太长, 可按下回车键 (Enter 键), 在下一行续写命令, R 程序会自动识别命令的完整性, 直到命令结束, 用户再按下回车键 (Enter 键) 确认.

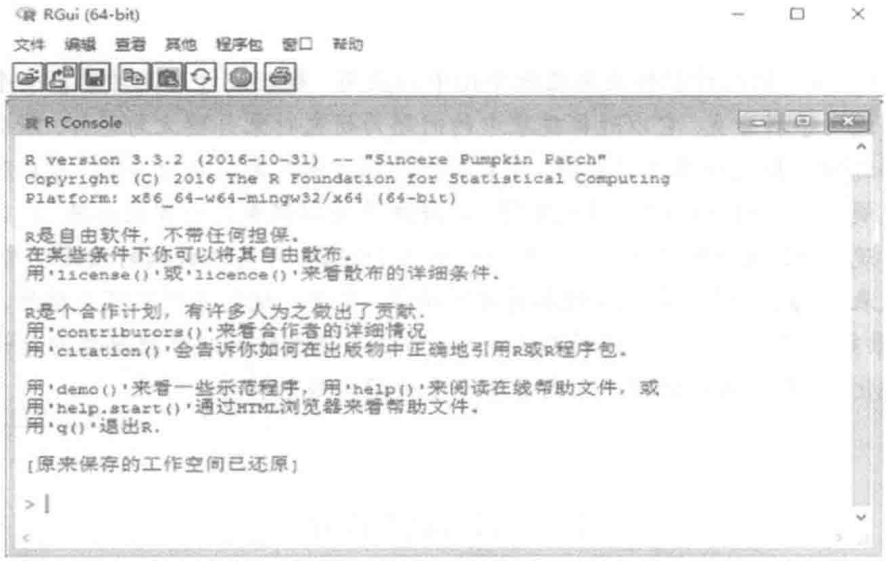

图 1.1.1 R 软件主界面

二、R 软件的基本使用

下面通过几个例子来阐述 R 软件的基本使用.

例 1.1.1 对于收集到的 (x, y) 数据, 见表 1.1.1, 绘制散点–折线图.

表 1.1.1 (x, y) 数据

x	1	2	3	4	5	6	7	8	9	10
y	5	7	6	9	8	10	11	4	6	8

解 R 软件执行过程如下:

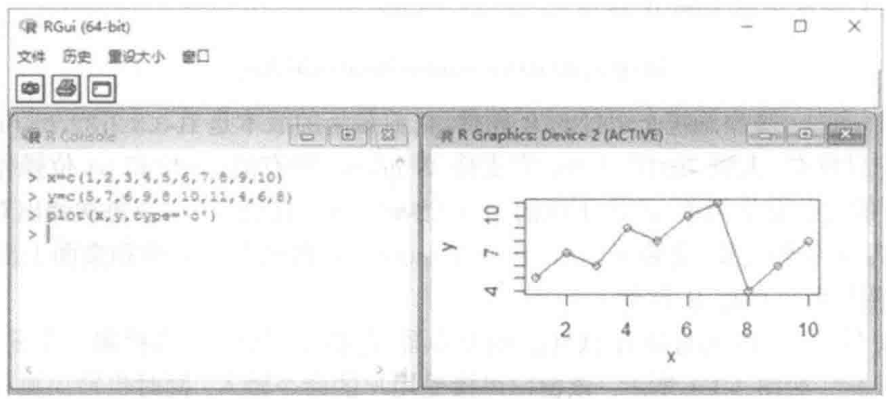

图 1.1.2 例 1.1.1 的命令和输出

对于第一条和第二条命令, 等号表示赋值语句, 其左侧是变量名, 右侧是一维数组, 由 “c( )” 表示; 而第三条命令采用 “plot( )” 将 x,y 当作坐标绘制在二维平面上, 并将图形显示在新建窗口中.

例 1.1.2　在 Excel 表格中有一批数据, 如图 1.1.3 所示, 分别计算两组数据的平均值.

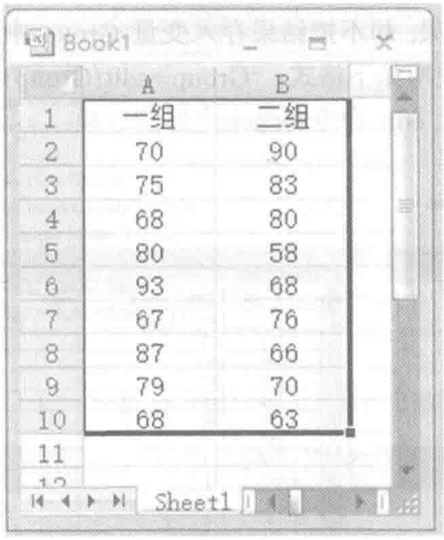

图 1.1.3　Excel 表格数据

解　在计算平均值之前, 必须将数据载入 R 程序中. 此处通过剪贴板实现数据的转移, 先在 Excel 中选中数据并复制到剪贴板 (如图 1.1.3 所示), 然后在 R 的命令窗口中输入如下命令:

图 1.1.4　Exceel 数据求均值

如图 1.1.4 所示, 第一条命令通过 “read.table( )” 函数, 从剪贴板对象 “clipboard” 中获取数据, 存入 “Group” 变量中, 其中 “header=TRUE” 表示该数据的各列变量有名字. 第二条命令 “colMeans(Group)” 将数据按列计算平均值, 并将结果输出在命令窗口中.

例 1.1.3　将例 1.1.2 中第一组和第二组数据分别增加 60, 80 和 70, 75.