统计数据分析与应用丛书

Statistics

统计分析与R

陈正昌　贾俊平　著

中国人民大学出版社

统计数据分析与应用丛书

Statistics

统计分析与R

陈正昌　贾俊平　著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

统计分析与R/陈正昌，贾俊平著. —北京：中国人民大学出版社，2016.11
统计数据分析与应用丛书
ISBN 978-7-300-23584-4

Ⅰ.①统… Ⅱ.①陈…②贾… Ⅲ.①统计分析-统计程序 Ⅳ.①C819

中国版本图书馆CIP数据核字（2016）第270365号

统计数据分析与应用丛书
统计分析与R
陈正昌　贾俊平　著
Tongji Fenxi yu R

出版发行　中国人民大学出版社
社　　址　北京中关村大街31号　　**邮政编码**　100080
电　　话　010－62511242（总编室）　010－62511770（质管部）
010－82501766（邮购部）　010－62514148（门市部）
010－62515195（发行公司）　010－62515275（盗版举报）
网　　址　http://www.crup.com.cn
http://www.ttrnet.com（人大教研网）
经　　销　新华书店
印　　刷　北京鑫丰华彩印有限公司
规　　格　185 mm×260 mm　16开本　　**版　　次**　2016年11月第1版
印　　张　25.5插页1　　**印　　次**　2016年11月第1次印刷
字　　数　598 000　　**定　　价**　52.00元

前　言
Preface

精熟应用统计软件，是研究人员与数据分析人员须具备的基本能力。目前有许多统计软件可供选择，然而，多数都需要付费购买，也无法免费升级，通常不是个别人员可以负担得起的。幸好，在商业软件之外，R提供了另一个选择。R统计软件不仅免费，而且功能强大，包含众多的分析包，提供了绝大多数统计软件的全部分析方法，此外，R还可以自行编写程序完成自建模型的求解和分析。甚至，像SPSS或Stata等统计软件也可以直接调用R，扩充其分析功能。现在，R正逐步成为统计分析的主流，国外也有大量的图书介绍R统计软件。

《统计分析与R》是在陈正昌所著《SPSS与统计分析》和《Minitab与统计分析》（均由台北五南图书出版公司出版，前者亦由北京教育科学出版社出版简体版）的基础上，搭配最新版之R统计软件改写而成。本书包罗了多数的单变量统计方法，以及常用的多变量分析技术，主要供基础统计学及进阶统计学教学之用，也配合研究生及学者进行量化研究分析与撰写论文之需。

全书共分为九大部分。第一部分（第1章）是R的安装、数据处理及初步分析之简介。第二部分（第2章及第3章）是统计图表及描述统计。第三部分（第4章）是各种概率分布简介，也是本书各章统计方法的基础。第四部分（第5章及第6章）说明均值之区间估计及统计检验的基本概念。第五部分（第7章至第15章）为均值差异检验，分别针对t检验及各种方差分析加以说明。第六部分（第16章至第18章）是变量间的相关分析，包含简单相关、偏相关及典型相关。第七部分（第19章及第20章）为回归分析，包含简单及多元回归。第八部分（第21章及第22章）是卡方检验及定性变量的分析。第九部分（第23章至第25章）是量表的信度及效度分析。

第7章至第25章都涵盖七个重点。一是，每章开头提醒该种统计方法适用的情境，叙述虽然简短，却相当重要。二是，简要说明基本统计概念，建议读者仔细阅读这一节的内容。三是，使用各学科领域的范例数据，并提出研究问题及统计假设。四是，配合R进行分

析，此部分说明如何使用程序套件及函数进行分析，并针对分析所得的结果逐一解读，对重要的统计量数说明计算方法。五是，针对目前各学术期刊都强调的效应量（effect size）加以介绍。六是，将分析发现以APA格式写成研究结果。七是，强调该种统计方法的基本假定，避免误用工具。

本书能够发行简体版，首先要感谢中国人民大学出版社慨允出版，贾俊平协助各项出版事宜。其次要感谢林素秋老师及黄俊维先生审阅初稿，并提出许多宝贵建议。教学过程中，学生对《SPSS与统计分析》的回馈及提问，也使《统计分析与R》更加完善，在此一并致谢。

本书主要的统计方法及分析结果解读由屏东大学陈正昌负责，第1、第4两章及R命令的撰写由中国人民大学贾俊平负责。书中所有的统计图形都由我们亲自绘制，虽然投入许多心力，但是难免会有疏漏之处，敬请读者不吝来信指教（陈正昌的电子邮箱：chencc@mail.nptu.edu.tw）。

需要书中所用的数据文件，请到中国人民大学出版社工商管理分社网站（www.rdjg.com.cn），或是陈正昌的教学网站（http://faculty.nptu.edu.tw/～chencc/）的“个人著作”区下载。

屏东大学　陈正昌
中国人民大学　贾俊平

目　录

第1章 R简介

R是基于R语言的一种优秀统计软件. R语言是一款统计计算语言，它根源于贝尔实验室开发的S语言. R语言有许多优点，比如：

1. 与多数统计软件相比，R语言是免费的.
2. 更新速度快，包含很多最新方法，而其他软件的更新需要比较长的时间.
3. R提供丰富的数据分析技术，功能十分强大.
4. R绘图功能强大，可以按照需求画出所需的图形对数据进行可视化分析.

1.1 R的初步使用

1.1.1 R的安装

在CRAN网站 http://www.r-project.org/上可以下载R的各种版本，包括Windows、Linux和MacOX三个版本，使用者可以根据自己的平台选择相对应的版本.

下载完成后，双击程序文件（.exe文件）即可完成安装. 安装R后，启动R（本书使用的是3.3.0版本）出现的开始界面如图1—1所示. 其中，显示了R的版本信息及一些简要的R软件说明.

R命令（指令）要在提示符号“>”后输入，每次输入一条命令，按“Enter”后再输入下一条命令. 也可以连续输入多条命令，命令之间可以用分号“;”隔开. R每次执行一条命令，也可以连续执行多条命令.

1.1.2 为对象赋值并运行

R运行的是一个对象，在运行前需要给对象赋值. R语言中标准的赋值符号是“<-”，也允许使用“=”进行赋值，但推荐使用更加标准的前者.

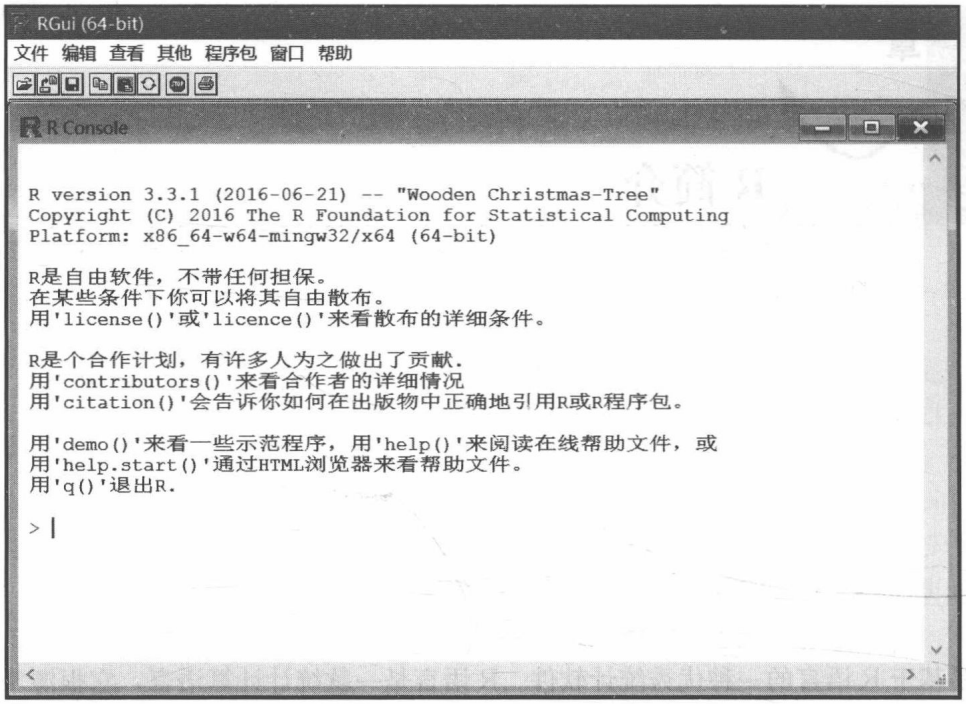


图 1—1　Windows 中的 R 界面

我们可以给对象赋一个值、一个向量、一个矩阵，或一个数据框. 比如，将数值 8 赋值给对象 x，将 5 个数据 75，82，93，68，95 赋值给对象 y，将数据框 table1 _ 1 赋值给对象 z，命令如下：

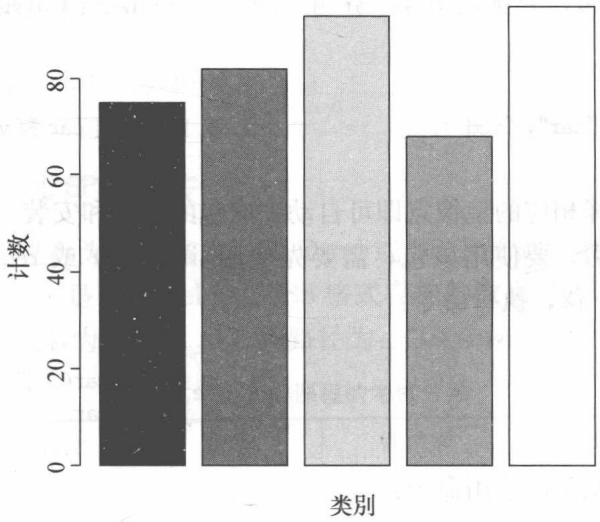
```
x<-8                                    # 将数值8赋值给对象x
y<-c(75, 82, 93, 68, 95)                # 将5个数据赋值给对象y
z<-table1_1                             # 将数据框table1_1赋值给对象z
```

为对象赋值后，就可以对对象进行各种计算和绘图. 比如，要计算对象 y 的总和、均值和方差，命令如下（命令下一行即为输出结果）：

```
> sum(y)                                # 计算对象y的总和
[1] 413
> mean(y)                               # 计算对象y的均值
[1] 82.6
> var(y)                                # 计算对象y的方差
[1] 133.3
```

要绘制对象 y 的条形图，命令为：

```
> barplot(y,xlab="类别",ylab="计数",col=c(1:5))   # 绘制对象y的条形图,x轴标签为"类
                                                  别",y轴标签为"计数",每个条的颜色
                                                  为R颜色系中的1~5种颜色
```



1.1.3　查看 R 的帮助文件

R 软件中的每个函数和包都有相对应的帮助说明，使用中遇到疑问时，可以随时查看帮助文件. 比如，要想了解 mean 函数的功能和使用方法，使用 help 或“?”命令查询该函数，命令为：

```
> help(mean)
> ? mean
```

R 软件就会输出 mean() 函数的具体说明，包括函数中参数设定、结果的结构、使用例子等内容. 当对一个函数不太清楚时，通过帮助文件，可以得到很大的协助.

1.1.4　包的安装和加载

R 中的包（package）是指包含 R 数据、函数等的集合. 目前，R 中有 9 000 多个包. 大部分统计分析都可以通过已有的 R 包来实现. 一个 R 包中可能包含多个函数，能做多种分析，而对于同类问题的分析也可以使用不同的包来实现，使用者可根据个人需要和偏好来选择所用的包.

在安装 R 时，附带了一系列默认包（包括 base，datasets，utils，grDevices，graphics，stats 以及 methods），它们提供了种类繁多的默认函数和数据集，分析时直接加载这些包，即可使用. 其他包可经由下载来进行安装. 使用函数 library() 则可以显示你的 R 中已经安装了哪些包.

在使用 R 时，可根据需要随时在线安装所需的包. 比如，要安装包 car（Companion to Applied Regression），命令为：

```
> install.packages("car")                    # 安装包 car
```

要同时安装多个包，比如要安装 car 和 vcd（Visualizing Categorical Data）两个包，命令为：

```
> install.packages(c("car", "vcd"))                    # 同时安装 car 和 vcd 两个包
```

输入命令后，选择相应的镜像点即可自动完成包的下载和安装.

在完成包的安装后，要使用该包，需要先使用 library() 或 require() 函数加载这个包. 比如，要使用 car 包，执行命令：

```
> library(car)                                         # 载入包 car(二择一)
> require(car)                                         # 载入包 car
```

要查看包的使用说明，使用命令：

```
> help(package = "package_name")                       # 查看加载包的帮助文件
```

R 可以输出某个包的简短描述以及包中的函数名称和数据集名称的列表，使用函数 help() 可以查看其中任意函数或数据集的更多细节.

本书各章将使用以下 R 包，建议读者先行安装，以方便随时调用，也避免不能连网无法分析的情况.

- agricolae
- biotools
- BSDA
- car
- CCP
- compute.es
- corpcor
- DescTools
- DTK
- effects
- ez
- foreign
- ggm
- GPArotation
- gplots
- heplots
- lavaan
- lm.beta
- lsr
- multcomp
- onewaytests
- openxlsx
- pastecs
- phia
- plyr
- psych
- reshape
- Rmisc
- semPlot
- subselect
- userfriendlyscience
- vcd
- xlsx
- yacca

1.2 数据的读入与保存

1.2.1 读取 R 格式数据

如果要在 R 中分析数据，可以在 R 中直接输入数据，也可以读取已有的数据文件. 如果是已有的数据，在运行程序时，首先需要将数据读入 R. R 可以读入很多类型的数据文件，包括 Excel，SPSS，SAS，Stata 数据等.

如果数据已存为 R 数据格式，读取 R 数据文件的命令为：

```
> load("mydata.RData")
```

在 mydata 中需要说明文件存放的路径和文件名，注意路径中分隔符是“\\”或者

“/”. 比如，要将存放在 C：/mydata/chap01/目录下的 R 文件 table1 _ 1 读入 R，命令为：

```
> load("C:/mydata/chap01/table1_1.RData")
```

1.2.2　读取 Excel 和 SPSS 文件

有时，已有的数据已存为其他格式，如 Excel 或 SPSS 格式，在分析时需要先将其读入 R. 假定下面的表 1—1 是以 Excel 或 SPSS 格式存储的数据文件，并已存放在 C:/mydata/chap01/目录下，取名为 table1 _ 1，即路径为：“C:/mydata/chap01/table1_1”.

表 1—1　　**10 名学生 5 门课程的考试分数**

学生姓名	统计学	数学	营销学	管理学	会计学
张青松	68	85	84	89	86
王奕翔	85	91	63	76	66
田新雨	74	74	61	80	69
徐丽娜	88	100	49	71	66
张志杰	63	82	89	78	80
赵颖睿	78	84	51	60	60
王智强	90	78	59	72	66
宋媛婷	80	100	53	73	70
袁四方	58	51	79	91	85
张建国	63	70	91	85	82

使用包“xlsx”或“openxlsx”可以读入 Excel 数据，使用包“foreign”则可以读入 SPSS 数据（注：如果数据中有中文，读取时会有错误）. 命令为：

```
# 读取 Excel 数据-1
> install.packages("xlsx")                          # 下载并安装 xlsx 包
> library(xlsx)                                     # 载入 xlsx 包
> table1_1<- read.xlsx("C:/mydata/chap01/table1_1.xlsx", as.data.frame = TRUE)
                                                    # 读取 Excel 数据,存入 table1_1 对象
# 读取 Excel 数据-2
> install.packages("openxlsx")                      # 下载并安装 openxlsx 包
> library(openxlsx)                                 # 载入 openxlsx 包
> table1_1<- read.xlsx("C:/mydata/chap01/table1_1.xlsx")
                                                    # 读取 Excel 数据,存入 table1_1 对象
# 读取 SPSS 数据
> install.packages("foreign")                       # 安装 foreign 包
> library(foreign)                                  # 载入 foreign 包
> table1_1<- read.spss(file = "C:/mydata/chap01/table1_1.sav", use.value.labels = TRUE, to.
data.frame = TRUE)                                  # 读取 SPSS 数据,存入 table1_1 对象
```

也可以先将 Excel 或 SPSS 格式数据存为 csv 格式数据（以逗号分隔的文本文件），并将其存放在指定路径的文件中. 比如，将表 1—1 存为 csv 格式，存放路径为：“C:/mydata/chap01/table1_1.csv”.

然后，在R中读取csv文件. 如果csv文件中包含标题（即变量名，如table1 _ 1的“学生姓名”和“统计学”“数学”等课程名称），使用以下命令：

```
> table1_1<- read.csv("C:/mydata/chap01/table1_1.csv")
```

如果csv文件中不包含标题，使用以下命令：

```
> table1_1<- read.csv("C:/mydata/chap01/table1_1.csv",header = FALSE)
```

这样，以csv格式存放的table1 _ 1就被读入到R中，可以进行各种分析了.

如果要将读入的数据存为R数据格式，使用save函数可将该数据存为一个R文件. 比如，将读入的table1 _ 1数据以R格式存放在指定的目录“C:/mydata/chap01”下，并命名为table1 _ 1. RData，命令为：

```
> save(table1_1,file = "C:/mydata/chap01/table1_1.RData")
```

其中，file=" " 指定文件的存放路径和名称. 后缀名必须是“. RData”. 这样表1—1就已经被存为一个名为table1 _ 1的R数据文件了.

1.2.3 在R中查看数据

将数据载入R时，R并不显示该数据. 如果要在R中看table1 _ 1的数据，使用命令：

```
> load("C:/mydata/chap01/table1_1.RData")          # 加载数据框table1_1
> table1_1                                         # 查看table1_1的全部数据
   学生姓名 统计学 数学 营销学 管理学 会计学
1    张青松     68   85     84     89     86
2    王奕翔     85   91     63     76     66
3    田新雨     74   74     61     80     69
4    徐丽娜     88  100     49     71     66
5    张志杰     63   82     89     78     80
6    赵颖睿     78   84     51     60     60
7    王智强     90   78     59     72     66
8    宋媛婷     80  100     53     73     70
9    袁四方     58   51     79     91     85
10   张建国     63   70     91     85     82
> head(table1_1, 5)                               # 查看table1_1的前5行数据
   学生姓名 统计学 数学 营销学 管理学 会计学
1    张青松     68   85     84     89     86
2    王奕翔     85   91     63     76     66
3    田新雨     74   74     61     80     69
4    徐丽娜     88  100     49     71     66
5    张志杰     63   82     89     78     80
```

```
> tail(table1_1)                                    # 查看table1_1的最后几行(默认值为后6行)数据
   学生姓名 统计学 数学 营销学 管理学 会计学
5    张志杰     63   82     89     78     80
6    赵颖睿     78   84     51     60     60
7    王智强     90   78     59     72     66
8    宋媛婷     80  100     53     73     70
9    袁四方     58   51     79     91     85
10   张建国     63   70     91     85     82
```

如果要对数据框或矩阵做转置处理（行列互换），使用命令：

```
> t(table1_1)                                       # 将数据框table1_1转置
```

1.3 数据的使用和编辑

有时，我们要对一个数据框（表1—1就是一个数据框）的某个或某些变量进行分析，就需要指定这些特定的分析变量.

1.3.1 选定数据框（或矩阵）的行或列进行分析

比如，要对数据框table1_1的某个特定的列或变量进行分析，例如计算均值，可以使用以下命令：

```
> load("C:/mydata/chap01/table1_1.RData")           # 加载数据框table1_1
> table1_1[,2]                                      # 选定数据框table1_1的第2列
或
> table1_1$统计学                                   # 选定数据框table1_1的“统计学”变量
> mean(table1_1[,2])                                # 对数据框table1_1的第2列求均值
[1] 74.7
或
> mean(table1_1$统计学)                             # 对数据框table1_1的“统计学”求均值
[1] 74.7
```

再比如，要选定矩阵matrix1_1的第5行，使用命令：

```
> matrix1_1[5,]                                     # 选定矩阵matrix1_1的第5行
```

1.3.2 编辑数据框

有时需要对数据框中的变量名或数据进行编辑，并用编辑后的数据覆盖原有的数据.

1.3.2.1 变量的重新命名

比如，将数据框table1_1中的“学生姓名”重新命名为“姓名”，将“统计学”重新命名为“统计”，命令为：