# 认识数据分析和R语言

# 本章内容

- 参考书目
- 数据分析简介
- 常用数据分析工具
- R简介
- R集成开发环境Rstudio

# 参考书目


R语言实战
[美] Robert I. Kabacoff 著
人民邮电出版社
数据分析：R语言实战
李诗羽 张飞 王正林 编著
電子工業出版社
数据挖掘：R语言实战
黄文 王正林 编著
電子工業出版社
ggplot2：数据分析与图形艺术
[美] 哈德利·威克姆 著
西安交通大学出版社
机器学习 实用案例解析
Conway & White 著
机械工业出版社
R语言数据分析与挖掘实战
张良均 云伟标 王路 刘晓勇◎著
机械工业出版社
数据挖掘与R语言
Data Mining with R
Learning with Case Studies
(葡) Luís Torgo 著
机械工业出版社

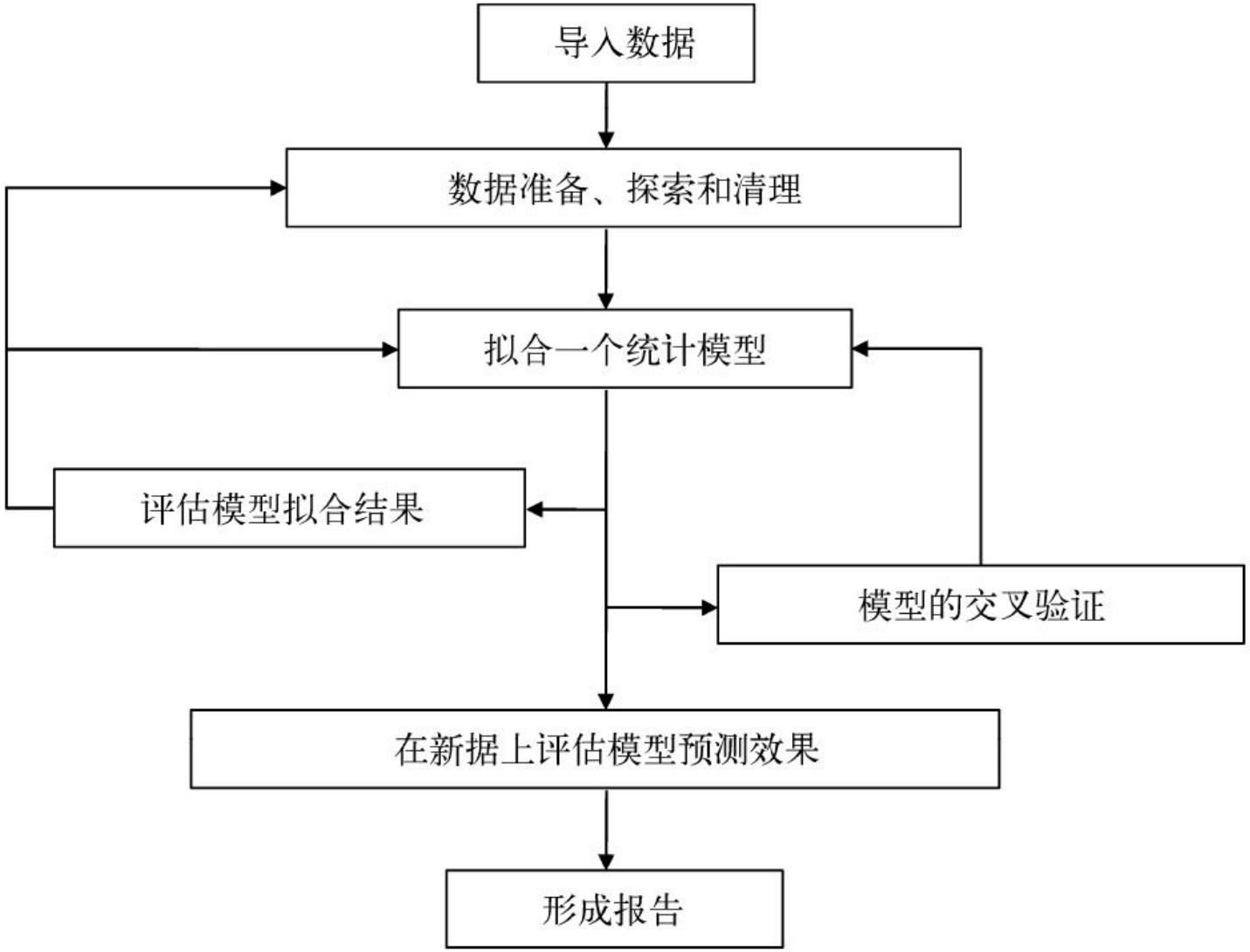
# 数据分析

- 数据分析是指用适当的**统计分析方法**对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以**详细研究**和**概括总结**的过程。

# 数据分析的原则

1. 数据分析是为了验证假设的问题，必须提供必要的数据验证。即构建完分析模型后，需要利用测试数据验证模型正确性。
2. 数据分析是为了发现更多问题，并找到深层次原因。
3. 不能为了做数据分析而分析。必须有明确的问题或目标。

# 数据分析步骤（1/2）


导入数据
数据准备、探索和清理
拟合一个统计模型
评估模型拟合结果
模型的交叉验证
在新据上评估模型预测效果
形成报告

# 数据分析步骤（2/2）

1. 探索性数据分析

   找到数据中隐含信息；探索规律性的可能的形式（探索方向和方式）；通常需要数据清洗和整合。

2. 模型选定分析

   通过定量分析，提出一类或几类可能的模型；再进一步分析，确定一类适合的模型。

3. 推断分析

   使用数理统计方法， 对所确定模型或估计的可靠程度和精度做出推断。

# 传统数据分析过程

1. 明确目标
2. 搜集数据
3. 加工整理

   缺失值处理、数据分组、数据取值转换……
4. 选择方法
5. 解释结果

# “大数据”分析过程

1. 数据采集
2. 预处理

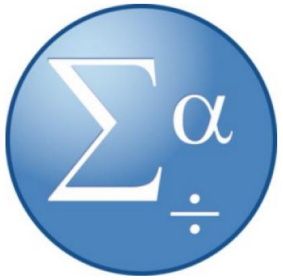
   数据清洗等。
3. 统计和分析

   可以满足大多数常见的分析需求
4. 数据挖掘

   与统计分析不同，数据挖掘一般没有什么预先设定好的主题

# 数据分析常用工具

- Excel：ToolPak（分析数据库）和Solver（规划求解加载项）
- SPSS：始于1968。世界上最早统计分析软件。已被IBM收购



- SAS：始于1976。该公司专做商业分析。
- Matlab
  - 开源替代：Octave

- R（本门课的内容）
- Python（后继课程涉及）

# R语言简介

- R是用于统计分析和绘图的编程语言和软件环境。
- R是GNU包，因此是自由软件。源码使用C、Fortran和R进行编写。

# R语言历史

- R语言是S语言的一种方言。
- 1976年，贝尔实验室的John Chambers开发S语言，以替代昂贵的SPSS和SAS工具。
- 1992年，新西兰奥克兰大学两位统计学教授**R**oss Ihaka和**R**obert Gentleman开发，Chambers也是开发成员。

# R的优势

- 统计学家发明的：全面的统计研究平台，提供格式各样的数据分析技术
- 开源：可以自己修改（包和函数）；免费。
- 交互式数据分析
- 可以从多中数据源导入数据
- 新算法（新的包）会迅速在R中实现
- 轻量级，安装文件小（不超过100M）
- 兼容不同的OS

# R的劣势

- 统计学家发明的：语法和一般程序设计语言差别很大，学习曲线陡峭，对于程序员来说“奇怪的”术语
- 开源导致package的质量、版本兼容性等问题
- 内存管理、速度与效率问题
- 不能直接利用R开发应用程序

# R 的获取和安装

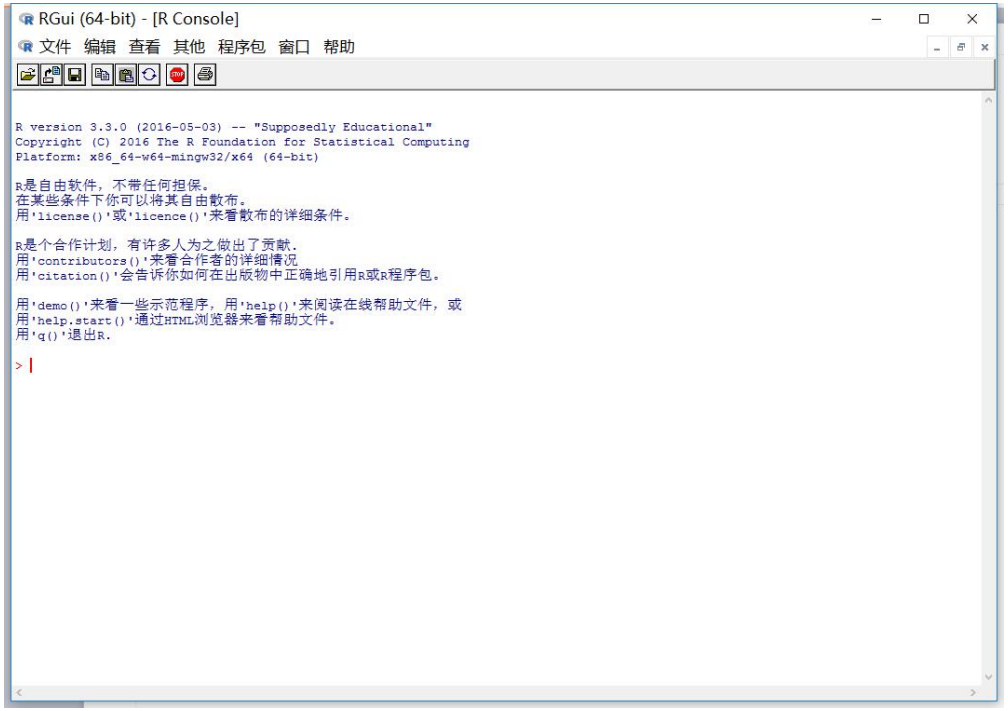
- R可以在CRAN（Comprehensive R Archive Network）http://cran.r-project.org上免费下载。Linux、Mac OS X和Windows都有相应编译好的二进制版本。

# R的使用（1/2）

- R是一种区分大小写的**解释型**语言。
- R中的多数功能是由程序内置函数和用户自编函数提供的，一次交互式会话期间的所有数据对象都被保存在内存中。一些基本函数是默认直接可用的，而其他高级函数则包含于按需加载的程序包中。
- **R语句由函数和赋值构成**。R使用 <-，而不是传统的 = 作为赋值符号。注释由符号 # 开头。

# R的使用（2/2）

- Windows从开始菜单中启动R。Mac需要双击应用程序文件夹中的R图标。Linux在终端命令提示符下敲入R并回车。



# 例子

```
# Listing 1.1 - A Sample R session
age <- c(1,3,5,2,11,9,3,9,12,3)
weight<-c(4.4,5.3,7.2,5.2,8.5,7.3,6.0,10.4,10.2,6.1)
mean(weight) #均值
sd(weight) #标准差
cor(age,weight) #相关度
plot(age,weight) #画图
```

# 获取帮助

- R的内置帮助系统提供了当前已安装包中所有函数的细节、参考文献以及使用示例。

**help.start()** 打开帮助文档首页

**help("foo")或?foo** 查看函数foo的帮助（引号可以省略）

**help.search("foo")或??foo** 以foo为关键词搜索本地帮助文档

**example("foo")** 函数foo的使用示例（引号可以省略）

**RSiteSearch("foo")** 以foo为关键词搜索在线文档和邮件列表存档

**apropos("foo", mode="function")** 列出名称中含有foo的所有可用函数

**data()** 列出当前已加载包中所含的所有可用示例数据集

**vignette()** 列出当前已安装包中所有可用的vignette文档(一般是PDF文章)

**vignette("foo")** 为主题foo显示指定的vignette文档

# 工作空间（workspace）

- **工作空间**就是当前R的工作环境，它储存着所有用户定义的对象（向量、矩阵、函数、数据框、列表。可以将当前工作空间保存到一个镜像中，并在下次启动R时自动载入它。
- 当前的**工作目录**（working directory）是R用来读取文件和保存结果的默认目录。

# 用于管理R工作空间的函数

| 函　数 | 功　能 |
| --- | --- |
| getwd() | 显示当前的工作目录 |
| setwd("*mydirectory*") | 修改当前的工作目录为*mydirectory* |
| ls() | 列出当前工作空间中的对象 |
| rm(*objectlist*) | 移除（删除）一个或多个对象 |
| help(options) | 显示可用选项的说明 |
| options() | 显示或设置当前选项 |
| history(#) | 显示最近使用过的#个命令（默认值为25） |
| savehistory("*myfile*") | 保存命令历史到文件*myfile*中（默认值为.Rhistory） |
| loadhistory("*myfile*") | 载入一个命令历史文件（默认值为.Rhistory） |
| save.image("*myfile*") | 保存工作空间到文件*myfile*中（默认值为.RData） |
| save(*objectlist*, file="*myfile*") | 保存指定对象到一个文件中 |
| load("*myfile*") | 读取一个工作空间到当前会话中（默认值为.RData） |
| q() | 退出R。将会询问你是否保存工作空间 |

# 输入和输出

1. 输入：函数source("filename")可在当前会话中执行一个脚本。
2. 文本输出：函数sink("filename")将输出重定向到文件filename中。
3. 图形输出

| 函　数 | 输　出 |
|---|---|
| pdf("filename.pdf") | PDF文件 |
| win.metafile("filename.wmf") | Windows图元文件 |
| png("filename.png") | PBG文件 |
| jpeg("filename.jpg") | JPEG文件 |
| bmp("filename.bmp") | BMP文件 |
| postscript("filename.ps") | PostScript文件 |

# 包（package）

- 包是R函数、数据、预编译代码以一种定义完善的格式组成的集合。计算机上存储包的目录称为库（library）。
- 函数library()则可以显示库中有哪些包。
- 目前有2500多个包可从http://cran.r-project.org/web/packages下载。这些包提供了横跨各种领域的新功能，包括分析地理数据、处理蛋白质质谱，甚至是心理测验分析的功能。

# 包的安装

- 执行install.packages()将显示一个CRAN镜像站点的列表，选择其中一个镜像站点之后，将看到所有可用包的列表，选择其中的一个包即可进行下载和安装。
- 例如，可以用可以使用命令install.packages("gclus")来下载和安装gclus包（注意有引号）。

# 包的载入

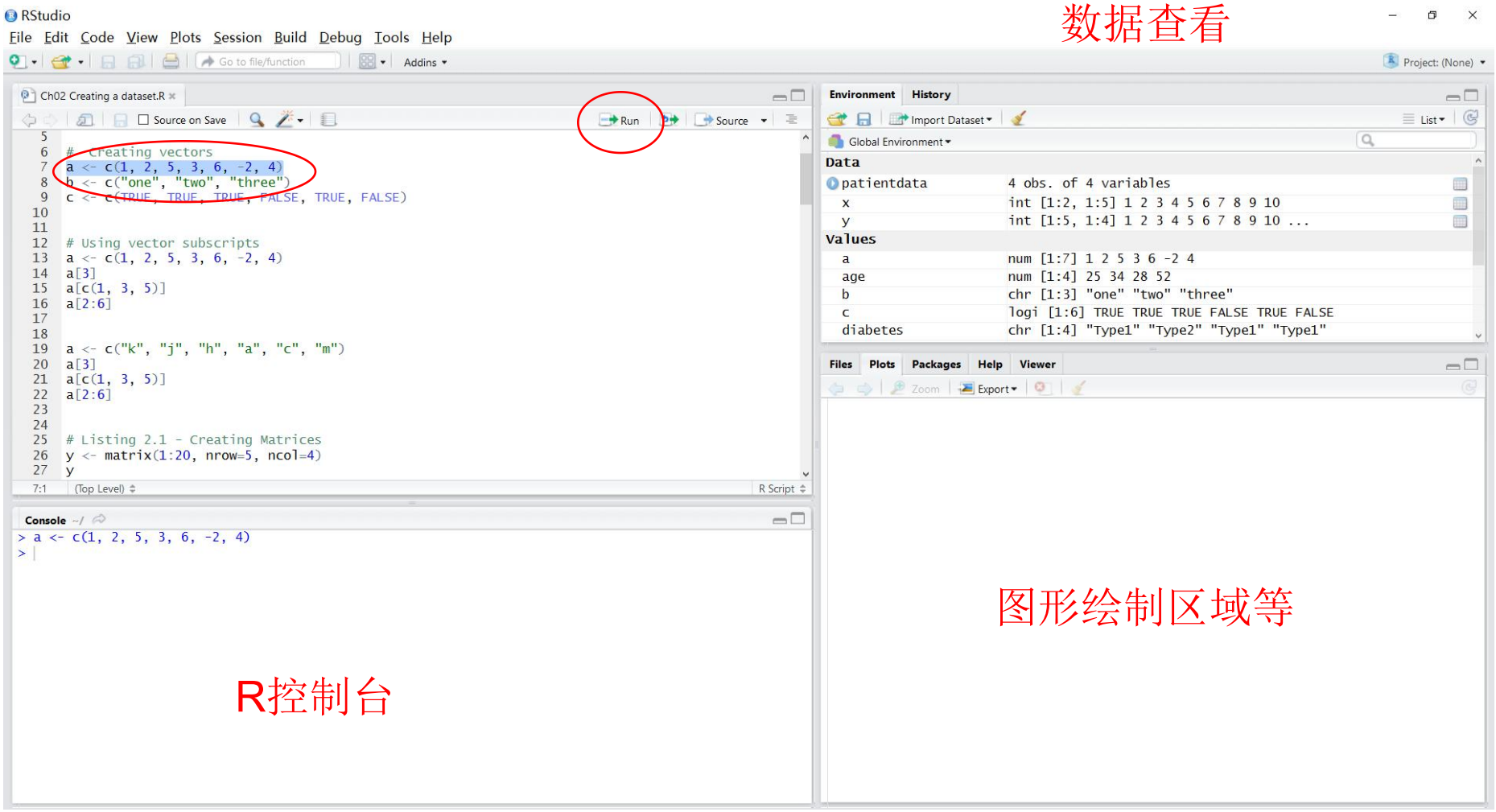
- 要在R会话中使用包，需要用library()命令载入这个包。
- 例如，要使用gclus包，执行命令library(gclus)即可（注意**没有**引号）。

# RStudio


- RStudio是一个自由开源的R的集成开发环境 (IDE)
- RStudio 有两个版本:
  - RStudio Desktop
  - RStudio Server：在Linux服务器上运行，客户端使用web浏览器访问。
- RStudio Desktop支持的系统包括Windows, OS X, and Linux

# RStudio的使用

选中R源文件中部分代码，单机Run即可执行




数据查看

R控制台

图形绘制区域等