R语言浅析

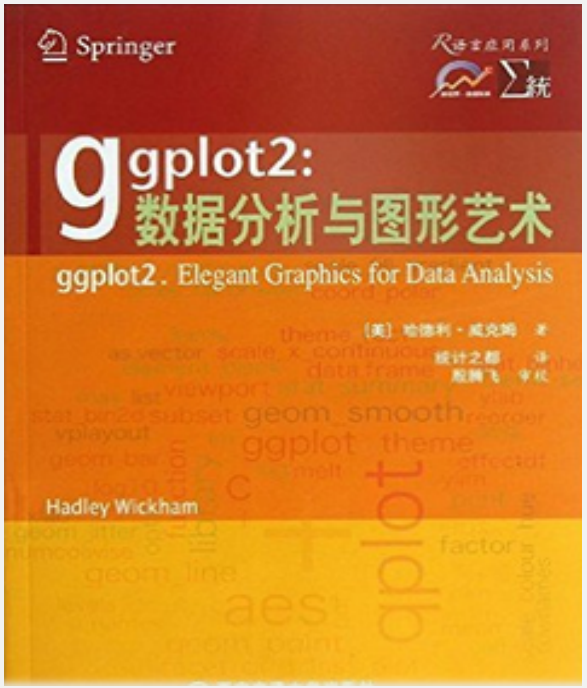
宋玉平

songyuping@shnu.edu.cn

办公室:徐汇六教504

R语言参考书目:

- R语言初学者指南
- R语言数据操作
- 统计建模与R软件
- ggplot2:数据分析与图形艺术



内容架构:

- R语言前世今生
- R语言简介: 软件安装，获取帮助，包的安装
- 数据结构: 向量、矩阵、数组、数据框（Dataframe）、因子、列表
- 数据导入
- 控制流/函数
- R语言统计分析
- ggplot2图形艺术

R语言前世今生:

- 奥克兰大学Ross Ihaka, Robert Gentleman 在20世纪
- 90年代初开发
- 前身是20世纪70年代的S语言（贝尔实验室开发）
- 免费, 开源, 开放等优点

R语言简介:

- 下载网址: **https://cran.r-project.org/**
- （百度R Cran第一个链接）
- 安装: 一般软件安装方式（适合电脑操作系统）
- 运行并简单操作

R语言优点:

- R免费, 永远正版
- R 资源公开(不是黑匣子), 优秀的内在帮助系统
- R可以在linux, Windows和Macos X上运行
- R有优秀的画图功能
- 学生能够轻松地转到商业支持的 S-Plus程序
- R语言有一个强大的, 容易学习的语法, 有许多内在的统计函数
- 对计算机初学者, 学习R语言使得学习下一步的其他编程不那么困难（譬如金融大数据分析软件Python等）

R语言数据结构:

- 向量
- 矩阵
- 数组
- 数据框
- 列表
- 因子

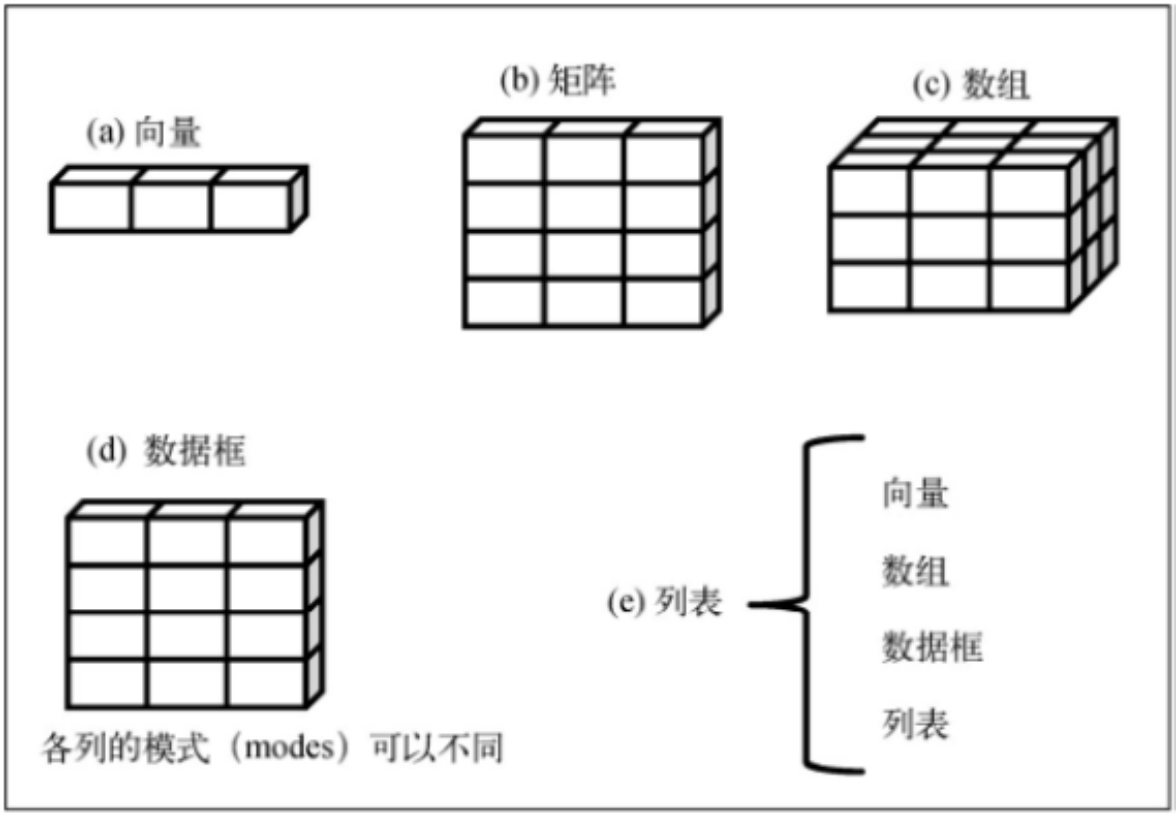


图2-1　R中的数据结构

R语言数据结构:

常量&变量

R 中变量名命名规则：

- 数字、字母、句点组成
- 首字符不能是数字，以句号开头时的第二个字符不允许是数字
- 区分大小写

三种常量：逻辑型、数值型、字符型。

- 逻辑型：TRUE，FALSE。
- 数值型：123，123.45，1.234e10，2.1-3.5i
- 字符型：“小明”，“Weight”
- 缺失值：NA

R语言数据结构:

变量赋值

- <- 最常用
- = 不建议使用, 和假设检验的等号容易混淆
- -> 不建议使用, 改变方向会降低代码可读性
- assign 函数, 很少使用

```
> x <- 3;y <- 2;z <- x+y
> z
[1] 5
```

R语言数据结构-向量

- 向量生成及相关运算函数
- 向量元素选取

R语言数据结构:

向量

使用 c() 操作来生成向量

```
x <- c(1,2,3); x
```

```
## [1] 1 2 3
```

```
y <- c(2*5, 3-4, 12/5, pi); y
```

```
## [1] 10.000000 -1.000000  2.400000  3.141593
```

基本统计函数:(并非全部)

- min, max, range, sum
- mean, median, quantile
- sd, var, cor, cov

R语言数据结构:

向量—序列生成

```
1:5
```

```
## [1] 1 2 3 4 5
```

```
seq(from = 4, to = 10, by = 2)
```

```
## [1]  4  6  8 10
```

```
seq(from = 0, to = 2, length.out = 5)
```

```
## [1] 0.0 0.5 1.0 1.5 2.0
```

```
seq(from = 0, to = 2, length.out = 6)
```

```
## [1] 0.0 0.4 0.8 1.2 1.6 2.0
```

```
rep(1, times = 5)
```

```
## [1] 1 1 1 1 1
```

```
rep(c(1,2), times = 3)
```

```
## [1] 1 2 1 2 1 2
```

```
rep(c(1,2), each = 3)
```

```
## [1] 1 1 1 2 2 2
```

R语言数据结构:

向量—选取元素

```
x <- 2:5
x[1]
```

```
## [1] 2
```

```
x[c(1,4)]
```

```
## [1] 2 5
```

```
x[2:3]
```

```
## [1] 3 4
```

```
x[-1]
```

```
## [1] 3 4 5
```

R语言数据导入:

- 读纯文本文件

read.table() 函数【非工作目录需采用绝对路径】

常用使用格式read.table(file,header=TRUE,sep="",skip=0）

例

```
rt<read.table("C:\\Desktop\\houses.txt",header=TRUE
)
```

OR rt2<-read.table(file.choose(),header=TRUE)【推荐】

scan()函数

scan(file,what=list(指定scan数据类型))

R语言数据导入:

- **从数据库中读取数据(**install.packages(“RODBC”))

```
> library(RODBC)
> connection <-
  odbcConnect(dsn=“servername”,uid=“userid”,pwd=
   “******”)
> query <-  “select * from lib.table where …”
> mydata <- sqlQuery(connection,query,errors=TRUE)
> odbcClose(conenction)
```

R语言数据结构-矩阵

- 矩阵创建
- 矩阵元素选取
- 矩阵计算

R语言数据结构-矩阵

矩阵创建--matrix

```
A <- matrix(1:12, nrow = 3); A
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

重要参数:

- nrow, ncol
- byrow = T(缺省值是 FALSE)
- dimnames, rownames, colnames

```
dim(A)
```

```
## [1] 3 4
```

```
nrow(A)
```

```
## [1] 3
```

```
ncol(A)
```

```
## [1] 4
```

```
length(A)
```

```
## [1] 12
```

R语言数据结构-矩阵

矩阵创建--rbind

```
rbind(c(1,2),c(3,4))
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

```
matrix(1:4,nrow = 2,ncol = 2,byrow = TRUE)
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

R语言数据结构-矩阵

矩阵创建--cbind

```
cbind(c(1,2),c(3,4))
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
```

```
matrix(1:4,nrow = 2,ncol = 2,byrow = FALSE)
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
```

R语言数据结构-矩阵

矩阵元素选取

索引的几种方法:

- 正整数
- 负整数
- 逻辑值
- 元素名称

注意:

- 不同维度可以使用不同引用方法
- 每个维度下标用逗号分隔

R语言数据结构-矩阵

矩阵元素选取

- 访问矩阵的第 i 行第 j 列的元素：A[i,j]
- 访问矩阵的第 i 行的元素：A[i,]
- 访问矩阵的第 j 列的元素：A[,j]
- 访问矩阵的第 a,b 行与第 x,y,z 列交叉的元素形成的子矩阵：A[c(a,b),c(x,y,z)]

R语言数据结构-矩阵

矩阵计算

- 矩阵转置：t(A)
- 矩阵相乘：A%*%B
- 矩阵求逆：solve(A)
- 特征值和特征向量：eigen(A)
- eigen(A)$values
- eigen(A)$vecotrs

R语言数据结构-矩阵

apply系列函数

```
> A<-matrix(1:12,nrow=2,byrow=FALSE)
> A
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    3    5    7    9   11
[2,]    2    4    6    8   10   12
```

- apply, mapply, tapply, lapply
- apply 函数处理矩阵和数组

```
apply(A, MARGIN = 1, FUN = sum)
```

```
## [1] 36 42
```

```
apply(A, MARGIN = 2, FUN = mean)
```

```
## [1]  1.5  3.5  5.5  7.5  9.5 11.5
```

R语言数据结构-数组

数组创建--array

```
z <- array(1:12,dim=c(2,6))
z
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    3    5    7    9   11
## [2,]    2    4    6    8   10   12
```

```
z <- array(1:12,dim=c(2,3,2))
z
```

```
## , , 1
##
##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
##
## , , 2
##
##      [,1] [,2] [,3]
## [1,]    7    9   11
## [2,]    8   10   12
```

R语言数据结构-数组

数组创建--dim

数组同矩阵一样，是向量加维数

```
z <- 1:12
dim(z) <- c(2,6)
z
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    3    5    7    9   11
## [2,]    2    4    6    8   10   12
```

R语言数据结构-数据框

- 数据框创建
- 数据框元素选取
- Attach、detach、with

R语言数据结构-数据框

R语言内置数据集

```
> library(help=datasets)
> head(CO2)
  Plant   Type  Treatment conc uptake
1   Qn1 Quebec nonchilled   95   16.0
2   Qn1 Quebec nonchilled  175   30.4
3   Qn1 Quebec nonchilled  250   34.8
4   Qn1 Quebec nonchilled  350   37.2
5   Qn1 Quebec nonchilled  500   35.3
6   Qn1 Quebec nonchilled  675   39.2
```

R语言数据结构-数据框

数据框（data.frame）创建

- 数据框是一个矩阵形式的数据结构
- 不同的列的数据类型可不同
- 但是每一列内的数据类型都相同
- 每列是一个变量，每行是一个观测
- 类似于 Excel 表格

R语言数据结构-数据框

数据框（data.frame）创建

- frame <- data.frame(x, y)
- x 和 y 应该是长度相同的向量
- frame 把 x 作为第一列, 把 y 作为第二列
- 可创建任意多列, 任意多行

```
x <- 1:3
gender <- c("M", "F", "M")
age <- c(25, 20, 30)
graduate <- c(T,F,T) #graduated or not
math <- c(80, 90, 85); physics <- c(70, 75, 80)
stu <- data.frame(x,gender,age,graduate,math,physics); stu
```

```
##   x gender age graduate math physics
## 1 1      M  25     TRUE   80      70
## 2 2      F  20    FALSE   90      75
## 3 3      M  30     TRUE   85      80
```

R语言数据结构-数据框

数据框（data.frame）引用

由于数据框是二维数据, 类似矩阵, 所以引用方法相似

```
stu[1,3]
```

```
## [1] 25
```

```
stu[2:3,-2]
```

```
##   x age graduate math physics
## 2 2  20    FALSE   90      75
## 3 3  30     TRUE   85      80
```

R语言数据结构-数据框

数据框（data.frame）特殊引用方式

[[]] 和 $ 是两个特殊引用列的方式

```
stu[[3]]
```

```
## [1] 25 20 30
```

```
stu$graduate
```

```
## [1]  TRUE FALSE  TRUE
```

R语言数据结构-数据框

数据框（data.frame）引用--subset

- ?subset 查看使用帮助
- select 参数选择列, subset 参数选择行 (使用逻辑表达式)

```
subset(mtcars, select = mpg, subset = (cyl == 6))
```

```
##                 mpg
## Mazda RX4      21.0
## Mazda RX4 Wag  21.0
## Hornet 4 Drive 21.4
## Valiant        18.1
## Merc 280       19.2
## Merc 280C      17.8
## Ferrari Dino   19.7
```