R 语言简介及 R 对象详解

谢益辉

中国人民大学统计学院

生物多样性数据分析和生物统计培训班

2008年11月17至20日于中国科学院植物研究所

目录

目录

贝尔实验室的 S 语言

- Fortran程序对于统计分析来说过于低层
- 统计数据分析过程复杂，模式化的程序难以适应分析需要
- 统计图形是（探索型）数据分析的重要输出
- S语言的主要作者John Chambers获得了ACM的软件系统奖

奥克兰大学的 R 语言

- 作者Ross Ihaka和Robert Gentlemen（首字母都是R）
- 基于Scheme语言，恰逢S语言的发布
- 改进 S 语言
- 作者都对统计计算感兴趣

R 语言现状

- 开源、免费、灵活、统计方法模型繁多
- 19 位核心成员
- 超过 1500 个程序附加包
- 论文引用次数呈指数增长
- 邮件列表中的邮件不计其数
- 跨地域、跨行业的协作

主页导航

http://www.r-project.org

- 关于
- 下载镜像（中国香港有一个镜像网站）
- R 组织
- 文档（官方文档、用户贡献文档、卡片）
- 其它

其它网络资源

- 入门示例: http://www.statmethods.net/
- 小技巧和小提示: http://onertipaday.blogspot.com/
- COS 论坛R版块: http://www.cos.name/bbs/thread.php?fid=15

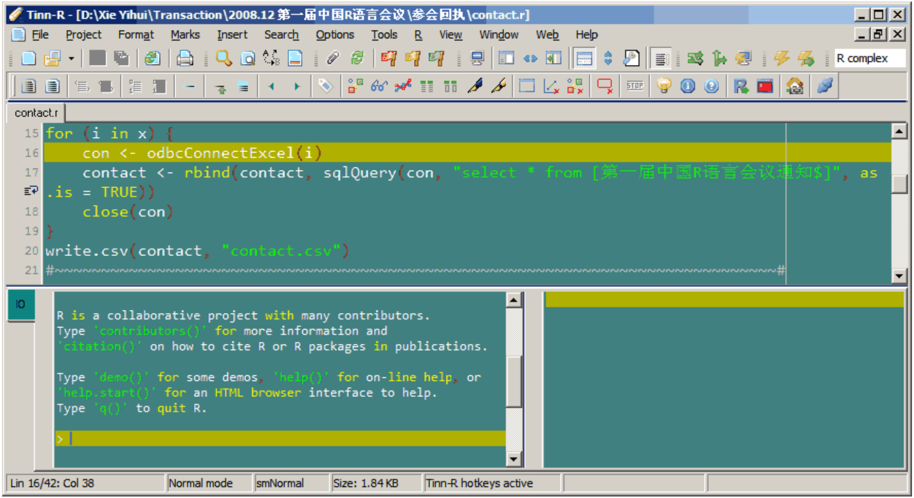
相关书籍

- Peter Dalgaard, *Introductory Statistics with R*（初等统计）
- Brian S. Everitt and Torsten Hothorn, *A Handbook of Statistical Analyses Using R*（涵盖较多统计模型，理论部分少，实例多）
- Venables and Ripley, *Modern Applied Statistics with S* (MASS)（经典，注重理论和统计计算细节）
- Paul Murrell, *R Graphics*（详细解释R图形）

我怎样学习R

- 一天看两次，一次看半天

Tinn-R 编辑器

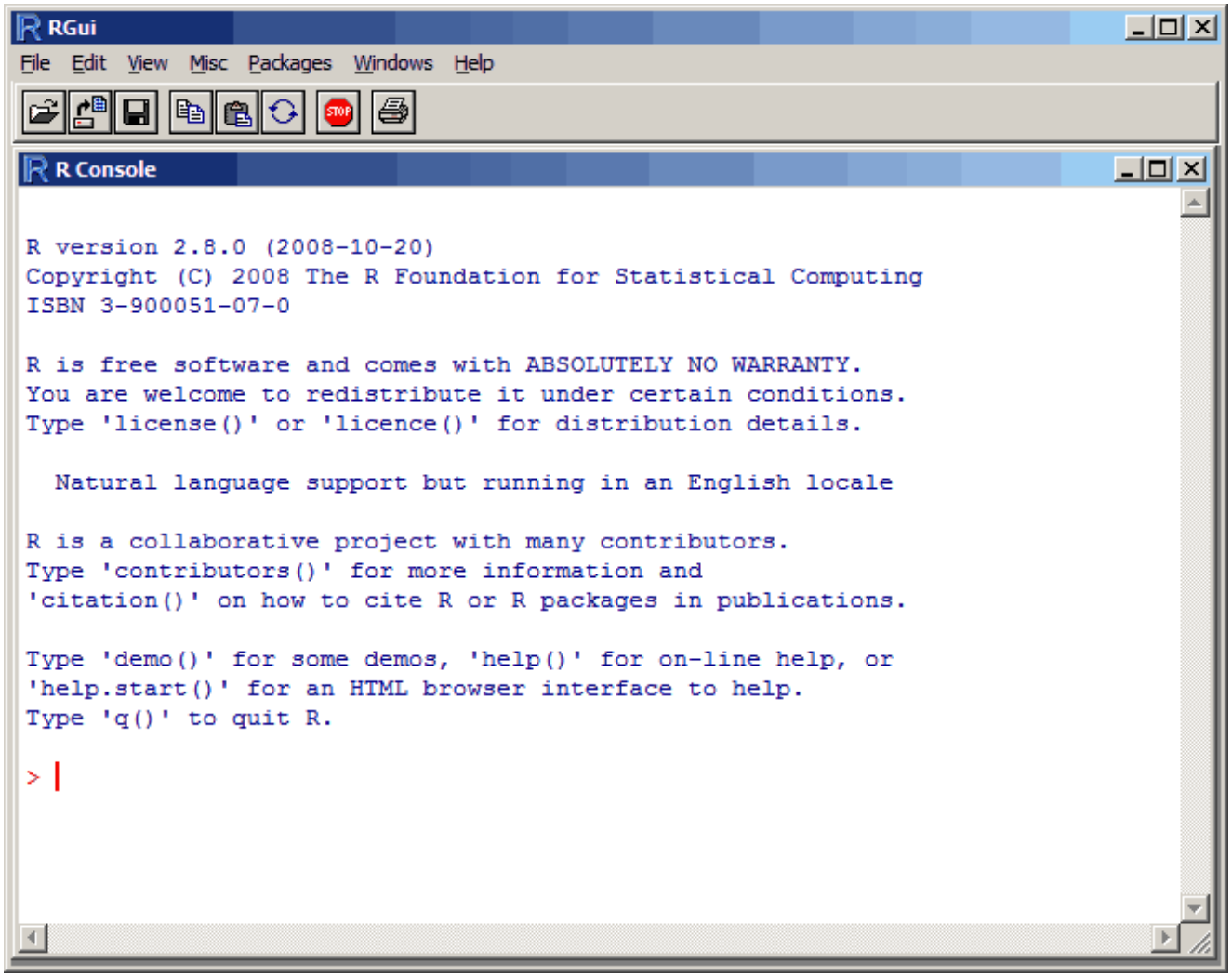

Tinn-R - [D:\Xie Yihui\Transaction\2008.12 第一届中国R语言会议\参会回执\contact.r]
File Project Format Marks Insert Search Options Tools R View Window Web Help
R complex
contact.r
for (i in x) {
con <- odbcConnectExcel(i)
contact <- rbind(contact, sqlQuery(con, "select * from [第一届中国R语言会议通知$]", as
.is = TRUE))
close(con)
}
write.csv(contact, "contact.csv")
#~~~~~~#
IO
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Lin 16/42: Col 38
Normal mode
smNormal
Size: 1.84 KB
Tinn-R hotkeys active

其它编辑器

任意文本编辑器都可以

- R 自身的编辑器
- 记事本
- Emacs/ESS
- WinEdt/R-WinEdt
- Kate

R 控制台


RGui
File Edit View Misc Packages Windows Help
R Console
R version 2.8.0 (2008-10-20)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>

工作目录

```
# 工作目录的好处在于可以使用相对路径
> getwd()
[1] "C:/Documents and Settings/Xie"
> setwd('E:/')
> getwd()
[1] "E:/"
> x = read.table("file.txt")
# 若不设定工作目录，则需要使用完整路径"E:/file.txt"
```

搜索路径

```
# 搜索路径让R知道从哪里获取对象
> search()
[1] ".GlobalEnv"        "package:stats"     "package:graphics"
[4] "package:grDevices" "package:utils"     "package:datasets"
[7] "package:methods"   "Autoloads"         "package:base"
> attach(iris)
> search()
 [1] ".GlobalEnv"        "iris"              "package:stats"
 [4] "package:graphics"  "package:grDevices" "package:utils"
 [7] "package:datasets"  "package:methods"   "Autoloads"
[10] "package:base"
> Species
  [1] setosa     setosa     setosa     setosa     setosa
  [6] setosa     setosa     setosa     setosa     setosa
....
```

目录

对象的属性

- 类型（数值、字符、复数、逻辑值）mode()
- 类（除了类型以外，还有矩阵、数组、因子和数据框等）class()
- 长度（向量、列表）length()
- 维数dim()
- 任意属性attr()
- 一个非常有用的查看函数：str()（对对象进行“庖丁解牛”）

特殊的值和对象

- 空对象：NULL （0或’’都不是空对象！）
- 缺失值：NA （Not Available）
- 非数值：NaN （Not a Number）
- 常数：pi、letters、LETTERS 等

向量

一串数字或字符

```
> c(1, 3, 7, 2, 9)
[1] 1 3 7 2 9
```

规则序列
重复序列

矩阵

矩形的二维表

```
> matrix(1:10, 2, 5)
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    5    7    9
[2,]    2    4    6    8   10
```

数组

多维度的数据（矩阵是其特例）

```
> array(1:24, c(3, 4, 2))
, , 1

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

, , 2

     [,1] [,2] [,3] [,4]
[1,]   13   16   19   22
[2,]   14   17   20   23
[3,]   15   18   21   24
```

因子

统计中的分类变量（或无序的离散变量）

```
> factor(c("甲", "乙", "丙")[sample(1:3, 10, TRUE)])
 [1] 乙 丙 乙 乙 乙 乙 乙 乙 乙 甲
Levels: 丙 甲 乙
```

有序因子

统计中的定序变量

```
> ordered(factor(c("甲", "乙", "丙")[sample(1:3, 10, TRUE)]))
 [1] 甲 丙 乙 乙 乙 丙 甲 丙 乙 丙
Levels: 丙 < 甲 < 乙
```

数据框

最常用的数据结构

```
> data.frame(a = 1:8, b = letters[1:8])
  a b
1 1 a
2 2 b
3 3 c
4 4 d
5 5 e
6 6 f
7 7 g
8 8 h
```

与矩阵的区别：各列的类型可以不同！

列表

最灵活的数据结构

```
> list(x = 1:3, y = factor(3:4), z = matrix(10:1, 2))
$x
[1] 1 2 3

$y
[1] 3 4
Levels: 3 4

$z
     [,1] [,2] [,3] [,4] [,5]
[1,]   10    8    6    4    2
[2,]    9    7    5    3    1
```

函数

自定义执行代码

```
> f = function(x) {
+     variance = sum((x - mean(x))^2)/(length(x) - 1)
+     return(variance)
+ }
> f(1:5)
[1] 2.5
> var(1:5)
[1] 2.5
```

转化

- 转数值as.numeric()、as.integer()
- 转字符as.character()
- 转逻辑值as.logical()
- 转因子as.factor()
- 转矩阵as.matrix()
- 转数据框as.data.frame()
- 等等系列as.*() 函数

下标

- 整数下标：中括号和整数表示将某位置上的元素提出来
- 逻辑下标：中括号和逻辑值表示将符合条件的元素提出来
- 名称下标：$符号和字符名称将某名称的子对象提出来
- 列表对象还可以用双中括号取子对象[[]]
- 负数下标：删除相应位置上的元素

计算

- 加减乘除、模、幂
- 数学和统计函数: log() log10() exp() sin() cos() tan() asin() acos() atan() min() max() range() pmin() pmax() sum() prod() cumsum() cumprod() mean() sd() var() median() quantile() cor() 等等

隐循环

- 向量化操作（R语言的一大便利）：允许我们针对某个对象整体操作，而不必对每一个元素循环操作（很多低层语言都必须显式循环）

```
> x = 1:10
> y = x + 1
> y
 [1]  2  3  4  5  6  7  8  9 10 11
# 不必 for(i in 1:10) y[i] = x[i] + 1
```

目录

注意事项

- R对大小写敏感！
- 把代码写整齐，注意缩进、空格（赋值符号<-、->），尽管它们大多数情况下不重要（animation 包中的tidy.source()函数）
- 随时查帮助！要经常使用问号'?'
- 出错了别想象，看看对象取值是什么

顺序

按照代码书写顺序一步步执行

```
> x = 1:10
> m.x = mean(x)
> s.x = 1/9 * sum((x - m.x)^2)
> s.x
[1] 9.166667
> var(x)
[1] 9.166667
```

循环

按照某个循环体依次执行

```
for(var in seq) expr
while(cond) expr
```

如果你发现你用的循环代码行数超过了总的代码行数的一半时，那么一般来说，这段程序说明了你的三种可能：

1. 你在糟蹋R
2. 你是故意的
3. 参见第1条

选择

满足条件则执行

```
if(cond) expr
if(cond) cons.expr  else  alt.expr
switch(EXPR, ...)
```

常见问题:

- 条件判断“等于”写错: 是==而不是=
- 条件cond不是TRUE或FALSE（而是NA或其它）