153 分钟学会 R

刘思喆

2008 年 12 月 13 日

文档结构

1. 前言
2. 基础知识
3. 输入输出
4. 数据处理
5. 数学运算
6. 字符操作
7. 日期时间
8. 绘图相关
9. 统计模型
10. 其他

Getting Started:

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

CRAN: *C*omprehensive *R* *A*rchive *N*etwork

CTAN: *C*omprehensive *TEX* *A*rchive *N*etwork

获取资料

- http://cran.r-project.org/other-docs.html
- 统计之都 *bbs*

推荐阅读

1. R for beginner
2. An Introduction to R
3. 153
4. Modern Applied Statistics with S

R 使用 cpu 不能超过 50%

这是 Windows 下任务管理器的误导，它将多个 CPU 看作是单个 CPU，同时计算使用比例。而 R 是单线程计算软件，它不能同时使用 2 个以上的 CPU。当你的计算机应用的是双核技术，你会发现 CPU 应用会定格在 50% 上。

获得在线帮助

- ?
- help.search()
- RSiteSearch()

在 2.5.0 版本中，R 引入了命令自动补全功能，使用 Tab 键能自动补全 R 命令；或使用第二次 Tab 后，返回所有可能的补全命令列表。

R 的内存使用

R 的工作内存大小的设定值为 32Mb 到 3Gb 间的任意数值。但需要提示的是：Windows 平台可用最大有效内存为 2Gb，也就是说，实际上 R 的工作内存区间为 32Mb 至 2Gb。

Windows 下升级 R，但不想重装 packages

update.packages()

R 初始加载的包

包	描述
stats	常用统计函数
graphics	基础绘图函数
grDevices	基础或 grid 图形设备
utils	R 工具函数
datasets	基础数据集
methods	用于 R 对象和编程工具的方法和类的定义
base	基础函数

获得 R 命令的源码

1.
2. methods(foo)
3. *.tar.gz

读取其他软件录入的文件

foreign 包，它可以读取 Minitab, S, SAS, SPSS, Stata, Systat, dBase 保存的数据

R 读取 Excel

- 另存为 csv 文件，read.csv() 读入；
- 加载 RODBC 包，使用 odbcConnectExcel() 函数；
- xlsReadWrite 包中的 read.xls 函数。

R 输出 TEX 文本

1. Hmisc 包中的 latex()
2. xtable 包中的 xtable()
3. quantreg 包中的 latex.table()

R 处理缺失值

- is.na()
- NA,TRUE,FALSE

两个数据框是否相等

如果每个元素都相同，那么这两个数据框也相同

```
a1 <- data.frame(num = 1:8,lib = letters[1:8])
a2 <- a1
a2[[3,1]] <- 2 -> a2[[8,2]]
any(a1!=a2) #     all(a1 == a2)
identical(a1,a2)
which(a1!=a2, arr.ind = TRUE)
```

去除相同的行

```r
x <- c(9:20, 1:5, 3:7, 0:8)
(xu <- x[!duplicated(x)])
unique(x)    # is more efficient
```

如何对不规则数组进行统计分析?

```
attach(warpbreaks)
tapply(breaks, list(wool, tension), mean)
aggregate(breaks, list(wool, tension), mean)
##   from the help
aggregate(state.x77,
          list(Region = state.region,
               Cold = state.x77[,"Frost"] > 130),
          mean)
```

随机抽取

sample(n)	随机组合 $1, \ldots, n$
sample(x)	随机组合向量 $x, length(x) > 1$
sample(x, replace = T)	bootstrap
sample(x,n)	非放回的从 x 中抽取 n 项
sample(x,n, replace = T)	放回的从 x 中抽取 n 项
sample(x,n, replace = T ,prob = p)	以概率 p, 放回的从 x 中抽取 n 项

如何进行复数计算?

```
x <- 1 + 1i # x <- complex(1,1)
Mod(x)  ; Conj(x)
```

求矩阵的特征值和特征向量的函数是什么?

已知 $A = \begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & -2 \\ 2 & -2 & -1 \end{bmatrix}$ 试求 $B = (\frac{1}{2}A^{-1}) + E$ 的特征值。

```
A <- matrix(c(-1,2,2,2,-1,-2,2,-2,-1),3,3)
m <- solve(0.5*A) + diag(c(1,1,1))
eigen(m)
```

这里还使用了函数 solve()，这个函数用于运算

```
a%*%x = b
```

而得到 x，当然也可以用来求矩阵的逆。

求立方根如何运算？

x^(1/3)。在 R 里面 sqrt() 函数可以计算开平方，故新手容易推测开立方也有函数。事实上 R 里面使用 ^ 来作幂函数运算。^ 不但是运算符号，还可以看作是函数：

```
"^"(x , 1/3)
```

在 R 中的运算符号包括：

R 中的运算符号

数学运算	+,-,*,/,^,%%,%/%	加、减、乘、除、乘方、余数、整除
逻辑运算	>, <, >=, <=, ==, !=	大于，小于，大于等于，小于等于，等于，不等于

如何求矩阵各行 (列) 的均值?

- apply()
- rowMeans(),colMeans()

如何计算组合数或得到所有组合？

choose() 用于计算组合数 $\binom{n}{k}$，函数 combn() 可以得到所有元素的组合。使用 factorial() 计算阶乘。

如何模拟高斯（正态）分布数据？

使用 rnorm(n，mean，sd) 来产生 n 个来自于均值为 mean，标准差为 sd 的高斯（正态）分布的数据。在 R 里面通过分布前增加字母 ‘d’ 表示概率密度函数，‘p’ 表示累积分布函数，‘q’ 表示分位数函数，‘r’ 表示产生该分布的随机数。这些分布具体可以参考第 26 页中 “*R 的分布函数*”，或 R-intro 中的 Probability distributions 章节

R 的分布函数

分布	R 函数	附加参数	默认参数
beta	beta	shape1(α),shape2(β)	
二项	binom	size(n),prob(p)	
χ^2	chisq	df	
均匀	unif	min(a),max(b)	$min = 0, max = 1$
指数	exp	rate	$rate = 1$
F	f	df1(r_1),df2(r_2)	
伽玛	gamma	shape(α),scale(θ)	$scale = 1$
超几何	hyper	$m = N_1, n = N_2, k = n$	
正态	norm	mean(μ),sd(σ)	$mean = 0, sd = 1$
泊松	pois	lamda(λ)	
t	t	df	
威布尔	weibull	shape(α),scale(θ)	$scale = 1$

对大小写敏感么?

R 中有很多基于 Unix 的包，故 R 对大小写是敏感的。可以使用 tolower() 、toupper() 、casefold() 这类的函数对字符进行转化。

```
x <- "MiXeD cAsE 123"
chartr("iXs", "why", x)
chartr("a-cX", "D-Fw", x)
tolower(x)
toupper(x)
```

日期可以做算术运算么?

一般我们需要使用 as.Date() ，as.POSIXct() 函数将读取的日期（字符串）转化为 “Date” 类型数据，“Date” 类型数据可以进行算术运算。

```
d1 <- c("06/29/07") ;    d2 <- c("07/02/07")
D1 <- as.Date(d1,"%m/%d/%y")
D2 <- as.Date(d2,"%m/%d/%y")
D1 + 2   ;  D1 - D2
difftime(D1,D2,units = "days")
```

如何将日期表示为 “星期日, 22 七月 2007”？

使用 format() 函数。

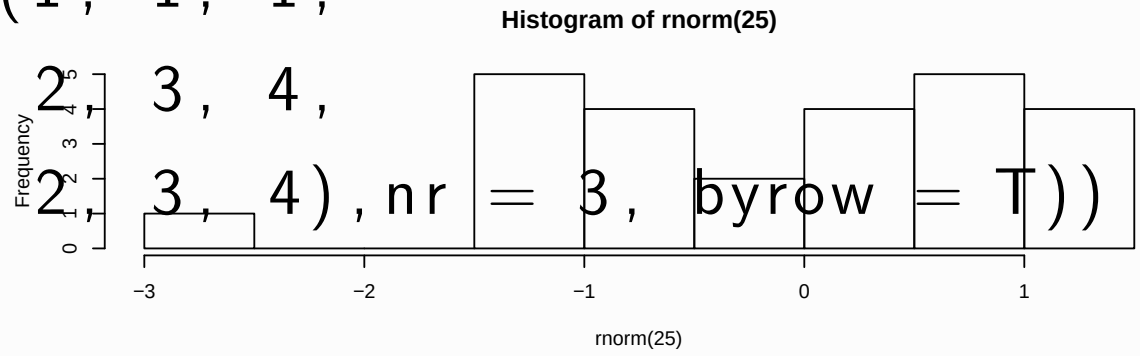
```
format((Sys.Date(), format="%A, %d %B %Y")
```

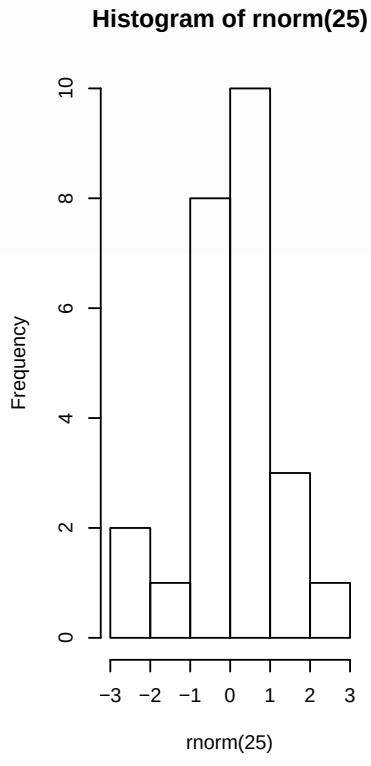
具体 format 参数可以参考 help(strptime) 的 details 部分。

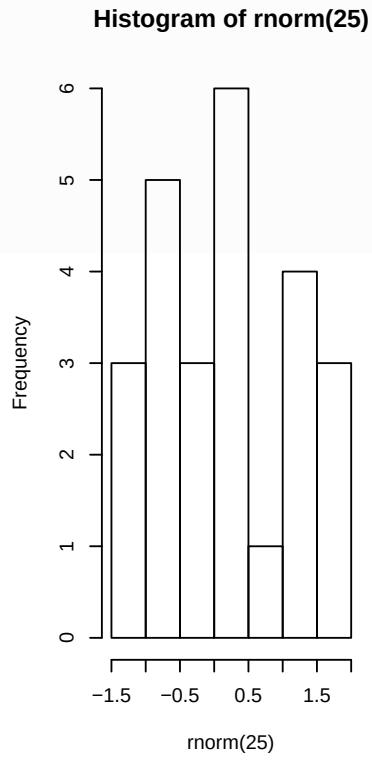
如何在同一画面画出多张图?

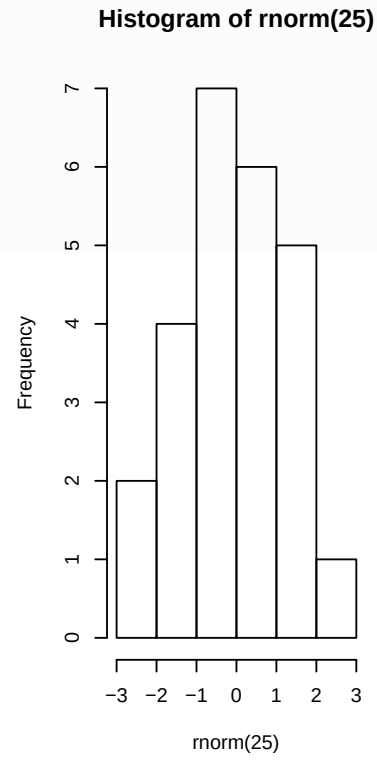
- 绘图参数：par(mfrow = c(2,2)) 或 par(mfcol = c(2,2))；
- 更为强大功能的 layout函数，它可以设置图形绘制顺序和图形大小；
- split.screen()函数。

```
layout(matrix(c(1, 1, 1,
                2, 3, 4,
                2, 3, 4),nr = 3, byrow = T))
hist(rnorm(25))
hist(rnorm(25))
hist(rnorm(25))
hist(rnorm(25))
```









如何在已有图形上加一条水平线

使用低水平绘图命令 abline()，它可以作出水平线（y 值 h=）、垂线（x 值 v=）和斜线（截距 a=，斜率 b=）。

R 中的绘图命令可以分为 “高水平”（High_level）、“低水平（Low_level）” 和 “交互式”（Interactive）三种绘图命令。

简要地说，“高水平” 绘图命令可以在图形设备上绘制新图；“低水平” 绘图命令将在已经存在图形上添加更多的绘图信息，如点、线、多边形等；使用 “交互式” 绘图命令创建的绘图，可以使用如鼠标这类的定点装置来添加或提取绘图信息。在已有图形上添加信息当然要使用 “低水平” 绘图命令。

常用的绘图设备都有哪些？

	名称	描述
屏幕	x11	X 窗口
显示	windows	Windows 窗口
文件设备	postscript	ps 格式文件
	pdf	pdf 格式文件
	pictex	供 LaTeX使用的文件
	png	png 格式文件
	jpeg	jpeg 格式文件
	bmp	bmp 格式文件
	xfig	供 XFIG 使用的图形格式
	win.metafile	emf 格式的文件

为什么 R 不能显示 8 种以上的颜色？

当绘图参数 col 使用数字来代替颜色名时会有这种情形，这是因为 R 内置调色板默认为 8 种颜色：

```
palette()
barplot(rnorm(15, 10 , 3) , col = 1:15)
palette(rainbow(15))
barplot(rnorm(15, 10 , 3) , col = 1:15)
palette("default")
```

如何用不同的颜色来代表数据？

高级绘图函数一般都有 col 参数可以设置。对于像 barplot() 这类图形，可以使用 “颜色组”(color sets) 来设置颜色，颜色组包括如下几类：

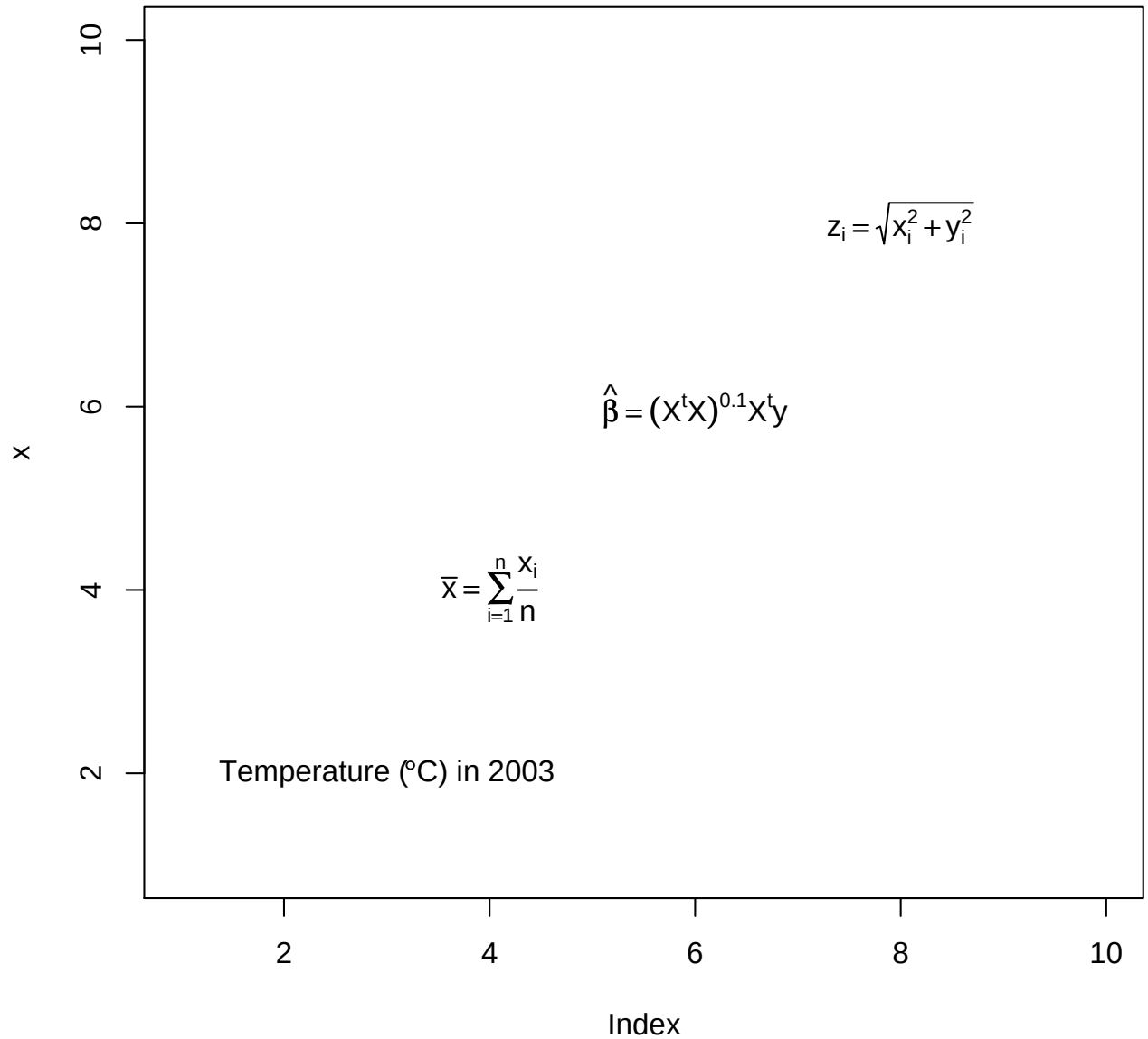
名称	描述
rainbow()	彩虹色 ()
heat.colors()	红色至黄色 ()
terrain.colors()	绿色、棕色至白色 ()
topo.colors()	深蓝色至浅棕色 ()
cm.colors()	浅蓝到白色，浅紫色 ()
gay()、grey()	灰色 ()

网格 (lattice) 绘图和普通绘图有什么区别?

网格（lattice）绘图实际上是 S-plus 中 Trellis 绘图在 R 中的实现，是多元数据可视化的方法。网格绘图相对于普通绘图来说，是一种拥有 “固定格式” 的绘图方式，当然它相对来说较难修改。适合对分属不同类数据绘图：

函数	说明
xyplot(y~x)	双变量散点图
dotplot(y~x)	Cleveland 点图 (逐行逐列累加图)
barchart(y~x)	y 对 x 的条形图
stripplot(y~x)	一维图，x 必须是数值型，y 可以是因子
bwplot(y~x)	箱线图
histogram(~x)	直方图

如何在 R 的绘图中加入数学公式或希腊字符？



在 word 里如何使用 R 生成的高质量绘图？

矢量绘图的效果是最好的，比如 eps、pdf，而不是位图（png、jpg、tiff 等）。在 word 里面，可以使用 eps，虽然在屏幕上显示不是很好，但打印效果却不错。

如何使用逐步回归？

在 R 里，可以使用计算逐步回归的 step()函数。它以计算 AIC 信息统计量为准则，选取最小的 AIC 信息统计量来达到逐步回归的目的。

如何做聚类分析?

- kmeans()
- hlust()
- cluster 包

如何做主成分分析?

1. princomp()
2. loadings()
3. screeplot()

如何对样本数据进行正态检验？

比较常见的方法：shapiro.test()，ks.test()(Kolmogorov-Smirnov 检验)，jarque.bera.test() (需要 tseries 包)。或者参考专门用作正态检验的 normtest 包，fBasics 包中的相关函数。这几个包（包括基础包）大概提供了十几种检验函数。

假设检验？

bartlett.test	方差齐次性检验	binom.test	二项检验
chisq.test	χ^2 检验	cor.test	相关性检验
fisher.test	Fisher 精确检验	friedman.test	Friedman 秩和检验
kruskal.test	Kruskal-Wallis 秩和检验	mcnemar.test	McNemar 检验
pairwise.t.test	均值的多重比较	PP.test	Phillips-Perron 检验
var.test	方差比检验	wilcox.test	Wilcoxon 秩和检验

logistic 回归相关函数是?

logistic 回归是关于响应变量为 0–1 定性变量的广义线性回归问题，这里需要使用广义线性模型 glm() 函数，且广义线性模型的分布族为二项分布。

广义线性模型中的常用分布族

分布	函数	模型
高斯（Gaussian）[1]	$E(y) = x^T\beta$	普通线性模型
二项（Binomial）	$E(y) = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}$	Logistic 模型和概率单位（probit）模型
泊松（Poission）	$E(y) = \exp(x^T\beta)$	对数线性模型

如何求 Spearman 等级（或 kendall）相关系数

cor() 函数默认为求出 Person 相关系数，修改其 method 参数即可求得 Kendall τ 和 Spearman 秩相关系数。

```
cor(longley, method = "spearman")
```

名称	方法	用途（条件）
Pearson	线性	正态总体假定
Kendall τ	协同	非参数检验
Spearman	样本秩	非参数检验

如何使用时间序列相关模型?

R 中使用 $arima(x, order = c(0,0,0), seasonal = list(order = c(0,0,0))$ 对模型进行拟合:

如何做判别分析?

参考 MASS 包中的 lda() 函数（Fisher *L*inear *D*iscriminant *A*nalysis）和 qda() 函数。

R 有类似于 SPSS 的界面么?

安装包 Rcmdr，加载包后，使用命令

```
Commander()
```

调出可供使用的图形使用界面。由于这个图形使用界面需要若干基础包外的其他函数，故还需要包 car 、effects 、abind、lmtest、multcomp、relimp、RODBC、rgl 的支持。

Sweave 是用来做什么的？

Sweave 提供了一种为 “混排 TEX 文本和 S 编码” 生成文档的机制。单个的 Sweave 文档中既包含 TEX 文本又包含 S 编码，通过编译最终形成的文档包含：

- TEX 文档的编译输出；
- S 编码和（或）；
- S 编码的代码输出（文本、图形）。

它的文档形成过程：

$$\text{Sweave 文档} \xrightarrow{\text{Sweave(in R)}} \text{TEX文档} \xrightarrow[\text{dvipdfmx}]{\text{LATEX}} \text{最终 pdf 文档}$$

如何释放 R 运行后占用的内存?

因为 R 是在内存中运算，所以当 R 读入了体积比较大的数据后，即使删除了相关对象，内存空间仍不能释放。gc() 函数虽然主要用来报告内存使用情况，但是一个重要的用途便是释放内存。

用什么文本编辑器比较好？

比较常用的是 Tinn–R ，RWinEdt [2] ，ESS(Emacs Speaks Statistics) ，甚至任意一款编辑器，如 UltraEdit[3]，这些都支持 R 语法的高亮显示。如果是 Windows 桌面环境下的用户，对这些不是很了解，记事本也不失为一种选择。

[2]下载、安装 WinEdt 后，在 R 中安装 RWinEdt 包即可使用

[3]需要下载、修改 wordfile

Thanks