R 语言入门

1. 预备工作

1.1 安装 R 本体和工具

适用于 Windows OS, 其余 OS 自行搜索

下载 R for Windows: https://www.r-project.org/

或 https://mirrors.tuna.tsinghua.edu.cn/CRAN/

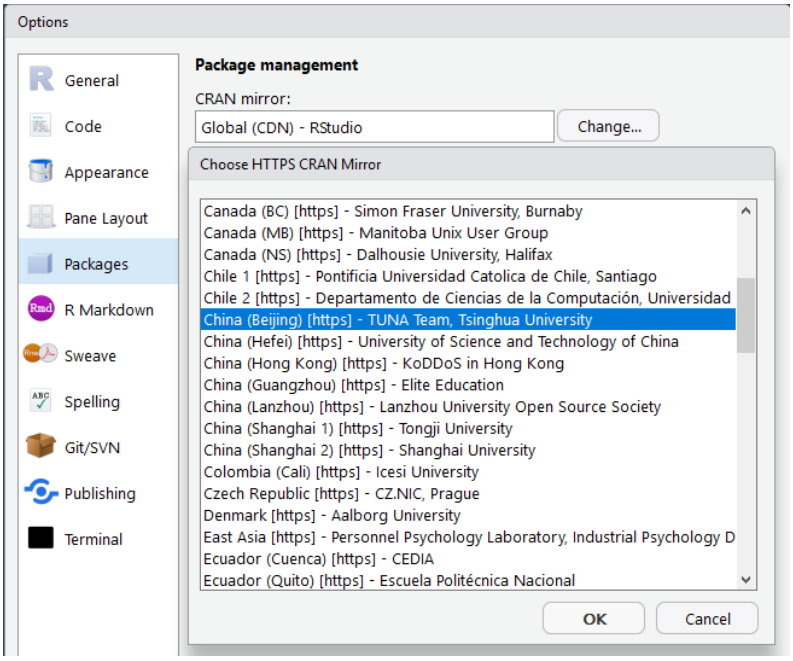
下载 Rstudio Desktop Free: https://www.rstudio.com/products/rstudio/download/

根据 OS 版本(32bit 或 64bit), 依次安装

1.2 安装扩展包

运行命令: `install.packages('扩展包名')`

如果下载速度太慢, 可以在 Tools—Options—Packages 中选择一个国内 CRAN 镜像



1.3 注意事项

- 工作空间默认为 current user\documents, 使用命令 `getwd()`查看, 更改可使用命令 `setwd('folderpath')`;

- 英文操作系统中文显示问题, 可用命令 `Sys.setlocale("LC_ALL","Chinese")`, 并设置 Tools --> global options --> code --> saving --> defalt text encoding: --> change, 选择 UTF-8
- 函数名、变量名等注意区分大小写;
- Rstudio 会自动补全引号、括号等; 单引号和双引号基本等价;
- 使用英文输入法以免误输入中文符号

1.4 入门教程推荐

- **中文在线教程: 李东风《R 语言教程》**
- 英文 RStudio 内置教程: Swirl

```
install.packages("swirl")
library(swirl)
swirl()
```

2. 基本操作

2.1 向量

R 语言以向量为最小单位, 用<-或=赋值 (有细微差别, 推荐用<-, 但初阶水平无所谓), 用 c( ) 生成向量, 向量[序数]取出元素, 序数从 1 开始

```
x1 <- 1:10
x2 = c(4, 2, 3, 5, 1, 6)
x3 = c(1,2,3,4:6)
x4 = c(x2,x3)
x5 = seq(3, 21, by=2)
x3[2]; x3[1:3]; x3[c(4:6,1,2)]; x[3]=22
ages <- c("李明"=30, "张聪"=25, "刘颖"=28)
ages <- c(30, 25, 28); names(ages) <- c("李明", "张聪", "刘颖")
ages[c("李明", "刘颖")]
```

2.2 函数 f()

- ?f 调出内置帮助

- 基本数学运算: +, -, *, /, ^, sqrt, exp, log, log10, round, floor, ceiling, ……
- 大部分函数可用向量作为自变量, 结果是自变量的每个元素各自的函数值
- 排序函数: sort(x)返回排序结果; rev(x)返回把各元素排列次序反转后的结果; order(x)返回排序用的下标

```
sort(x2); order(x2); rev(sort(x2)); x2[order(x2)]
```

- 统计函数: sum, mean, var, sd, min, max, range, prod, cor, etc.
- 概率函数
  - dxxx(x), 即 xxx 分布的分布密度函数(PDF)或概率函数(PMF) p(x)
  - pxxx(q), 即 xxx 分布的分布函数(CDF) F(q) = p(x≤q), 可加选项 lower.tail=FALSE 来计算 p(x>q)
  - qxxx(p), 即 xxx 分布的分位数函数 q(p), 对连续型分布, $q(p)=F^{-1}(p)$
  - rxxx(n), 即 xxx 的随机数函数, 可以生成个 xxx 的随机数
  - 离散分布有 dbinom 二项分布, dpois 泊松分布，dgeom 几何分布，dnbinom 负二项分布, dmultinom 多项分布，dhyper 超几何分布
  - 连续分布有 dunif 均匀分布，dnorm 正态分布，dchisq 卡方分布，dt t 分布(包括非中心 t)，df F 分布，dexp 指数分布，dweibull 威布尔分布，dgamma 伽马分布，dbeta 贝塔分布，dlnorm 对数正态分布，dcauchy 柯西分布, dlogis 逻辑斯谛分布

2.3 矩阵和数据框

矩阵用 matrix 函数定义，实际存储成一个向量，根据保存的行数和列数对应到矩阵的元素， 存储次序默认为按列存储

```
A <- matrix(11:16, nrow=3, ncol=2); print(A)
B <- matrix(c(1,-1, 1,1), nrow=2, ncol=2, byrow=TRUE); print(B)
```

用 nrow()和 ncol()函数可以访问矩阵的行数和列数

用 A[1,]取出 A 的第一行，变成一个普通向量。 用 A[,1]取出 A 的第一列，变成一个普通向量。 用 A[c(1,3),1:2]取出指定行、列对应的子矩阵。

x1, x2, x3 是等长的向量， cbind(x1, x2, x3)把它们看成列向量并在一起组成一个矩阵

```
cbind(c(1,2), c(3,4), c(5,6))
cbind(1, c(1,-1,10))
```

rbind()用法类似， 可以等长的向量看成行向量上下摞在一起

数据框是一个特殊的列表， 其每个列表元素都是一个长度相同的向量

```
d <- data.frame(name=c("李明", "张聪", "王建"), age=c(30, 35, 28),
height=c(180, 162, 175),stringsAsFactors=FALSE)

print(d); names(d)

d[[2]]; d[["age"]]; d$age

d[3,'age']; d[1:2, c('age', 'height')]
```

2.3 读入 csv 文件

```
tax.tab=read.csv('taxsamp.csv')
```

如果 csv 文件较大建议使用或者内含中文导致乱码, 建议使用 readr 扩展包的 read_csv 函数

```
install.packages('readr'); library('readr')

tax.tab=read_csv('taxsamp.csv')
```

如果内含中文的 csv 文件读入报错, 使用记事本程序将其打开, 点击<另存为>, 编码选择<UTF-8>, 保存覆盖原文件.

2.4 绘图

```
curve(sin(x), 0, 2*pi); abline(h=0)

barplot(c('男生'=10, '女生'=7), main='男女生人数')

plot(1:10, sqrt(1:10))


d.class <- read_csv("class.csv")

plot(d.class$height, d.class$weight)

with(d.class, plot(height, weight,main='关系', xlab='身高', ylab='体重'))


r=rnorm(1000,0,1); hist(r,breaks=30);

hist(r,breaks=30,freq=FALSE); lines(density(r))
```

2.5 查看和简单统计

```
head(tax.tab)

table(tax.tab[,'征收方式'])

summary(tax.tab[,'营业额']); mean(tax.tab[,'营业额']); sd(tax.tab[,'营业额'])
```