Beginning R

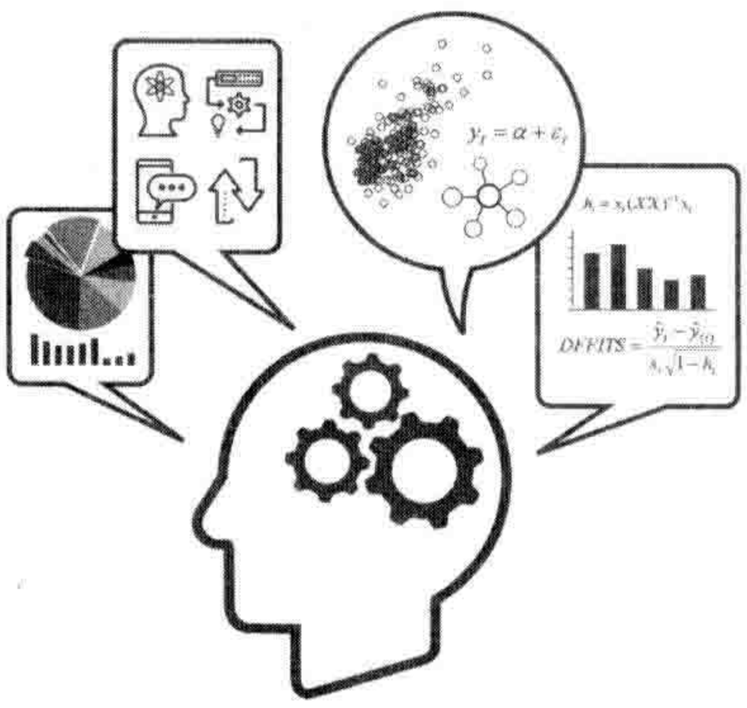
零基础学R语言

数学计算、统计模型与金融大数据分析

循序渐进地讲解 R 语言的基本语法，并通过三个热门领域的实用案例来讲解：数据的输入与输出，绘制统计数据图表，统计模型的处理与分析，金融工具的分析与获取，以及金融大数据的挖掘。

丰士昌 著

清華大學出版社



零基础学R语言

数学计算、统计模型与金融大数据分析

丰士昌 著

清华大学出版社
北京

内 容 简 介

R 具有高效的数据存储和数据处理功能，随着大数据技术的崛起，R 语言已成为大数据处理必备的工具之一。

R 语言并不是独立存在的程序设计语言，我们习惯说的 R 其实是指 R 系统。本书从建立 R 系统的基本环境入手，讲述 R 语言的基本函数及数据分析图形的绘制，用丰富的范例来讲解 R 语言的基础知识，并切入三个热门领域：金融分析、统计模型、数学计算。通过解析在这些领域的实用案例及数据处理分析的过程，让你在最短的时间内掌握 R 语言的核心知识，并可以用这些知识解决自己实际工作中遇到的问题。

若你是初学者，本书可以作为你学习 R 语言应用基础的快速入门教材。若你有一定基础，本书则可以进一步拓展你的视野，提升你使用 R 系统进行专业数据分析的能力。

图书在版编目（CIP）数据

零基础学 R 语言数学计算、统计模型与金融大数据分析 / 丰士昌著. —北京：清华大学出版社，2018
ISBN 978-7-302-50285-2

Ⅰ. ①零… Ⅱ. ①丰… Ⅲ. ①程序语言－程序设计Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字(2018)第 112035 号

责任编辑：夏毓彦
封面设计：王　翔
责任校对：闫秀华
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京国马印刷厂
经　　销：全国新华书店
开　　本：190mm×260mm　　**印　张**：17.75　　**字　　数**：454 千字
版　　次：2018 年 8 月第 1 版　　**印　　次**：2018 年 8 月第 1 次印刷
定　　价：59.00 元

产品编号：078580-01

前　言

R 语言并不是独立存在的程序设计语言，当我们单独称 R 而不是 R 语言时，其实是指 R 系统。R 是用于统计分析、绘图的语言和操作环境，或者说 R 是一个集成的环境，其中包含一整套数据操作、计算和图形绘制的软件包。R 定位于提供一个完善和统一的系统，所以 R 语言并不会脱离 R 环境而独立存在，因而不像其他数据分析语言那样成为一个附属工具。

作为 GNU 系统的一个自由、免费、源代码开放的软件环境，R 具有高效的数据存储和处理功能、一整套完整的数组和矩阵计算能力以及开放、完整的数据分析体系，同时为数据分析、统计及其结果的图形展示提供了强大的绘图功能。随着大数据技术的兴起，R 也成为大数据处理必备的工具之一。

R 语言在矩阵处理、统计分析、金融应用、图表绘制等方面都拥有十分便捷的函数与工具，操作方式十分类似 MATLAB 语言。将 R 应用于数学计算、统计模型，特别是股票和期货等金融交易数据的分析、回测，甚至是行情走势的研判，变得越来越热门。例如，只需要寥寥几条语句就可以绘制出专业的 K 线图、均线系统、布林线、MACD 等技术图形。

目前，每年都会举办 R 语言大会，届时邀请学界与产业界的人士发表最新的开发工具或产业应用。微软公司在 2015 年 1 月宣布收购了 R 的商业方案提供商 Revolution Analytics，可见 R 语言也是一个被看好的工具软件。

为了让初学者迅速步入 R 语言的殿堂，本书从 R 基本环境的建立开始介绍，而后切入 R 语言的基本函数和分析图形的绘制，在丰富的范例中迅速掌握 R 的核心知识，以便读者可以继续自学，为提升 R 的应用能力打下坚实的基础。本书还花了不少篇幅教授读者如何从公开的信息网站和财经网站获取实际的证券、期货交易的历史数据，并以此数据为基础在范例中加以运用，达到在实战中学习的效果。

本书从一般性的使用、函数介绍与图表绘制开始，让读者快速地对 R 具备基本的使用技能，接下来从三个热门的领域：数学计算、统计模型与金融分析介绍实用的案例。

如果你对这些领域之一感兴趣，并想试试 R 在这些领域的功力，即大数据分析和处理、数学计算、统计分析、财务数据分析、证券交易数据分析与研判等，那么本书就非常适合你用来打通自己潜力的“任督二脉”。

虽然本书在撰写与编排上力求尽善尽美，但是疏漏之处在所难免，恳请读者与专家不吝指正。

丰士昌

目　录

第 1 章　建立 R 语言的环境

R 语言是当今排名进入前 10 的程序设计语言，也是大数据处理的常用工具之一。在本章中，我们将从认识 R 语言开始逐步介绍在不同系统上的安装方式，分析解释型与编译型语言的差异，并介绍在服务器端的用法。

1.1　认识 R 语言

1.1.1　R 语言的诞生

R 语言是由新西兰奥克兰大学（The University of Auckland）的 Ross Ihaka 和 Robert Gentleman 所开发的，两人名字开头都为 R，因此就以 R 语言来命名。R 语言是一个 GNU 项目，源代码可自由地下载、修改、发布，并有编译好的软件可直接使用，可在多种平台下执行，如 Windows、UNIX、Linux、FreeBSD 与 MacOS 等，如今已交由“R 开发核心团队”负责后续的开发。

R 语言是一种高级[1]解释型[2]语言，本身也是一个系统，其中包含许多常用的科学工具，对于非信息相关背景的人士容易上手，因此在短短的数年中，R 已经在热门开发软件[3]中上升到前 15 名。

R 语言在矩阵处理、统计分析、金融应用、图表绘制等方面都拥有十分便捷的函数与工具，操作方式十分类似 MATLAB 语言。目前，每年都会举办 R 语言大会，邀请学界与产业界的人士发表新的开发工具或产业应用。微软公司在 2015 年 1 月宣布收购 R 的商业方案提供商 Revolution Analytics，可见 R 工具软件被市场看好。

R 的官方网站为 https://www.r-project.org/，其中包含一般性的介绍、软件的下载、帮助文件与参考资料等，如图 1-1 所示。

---

[1] 高低级程序设计语言是对计算机而言的名词，低级语言接近于机器语言，计算机易懂而人难学；高级语言则相反，人好学但计算机必须花更多时间理解和处理。因此，高级语言容易入门，但处理性能较差。

[2] 许多计算机语言（如 C 语言）需要编译后计算机才能执行，解释型的语言不需要编译，直接在运行环境中执行就可以得到结果。R 语言本身是解释型的语言，但提供了 Rcpp 的软件包供用户转换成语法编译，以提供执行性能。

[3] 在 TIOBE（https://www.tiobe.com/tiobe-index/）的资料中，R 语言从 2015 年的第 44 名上升到 2016 年 12 月的第 17 名，到 2017 年 12 月更是上升到第 8 名。

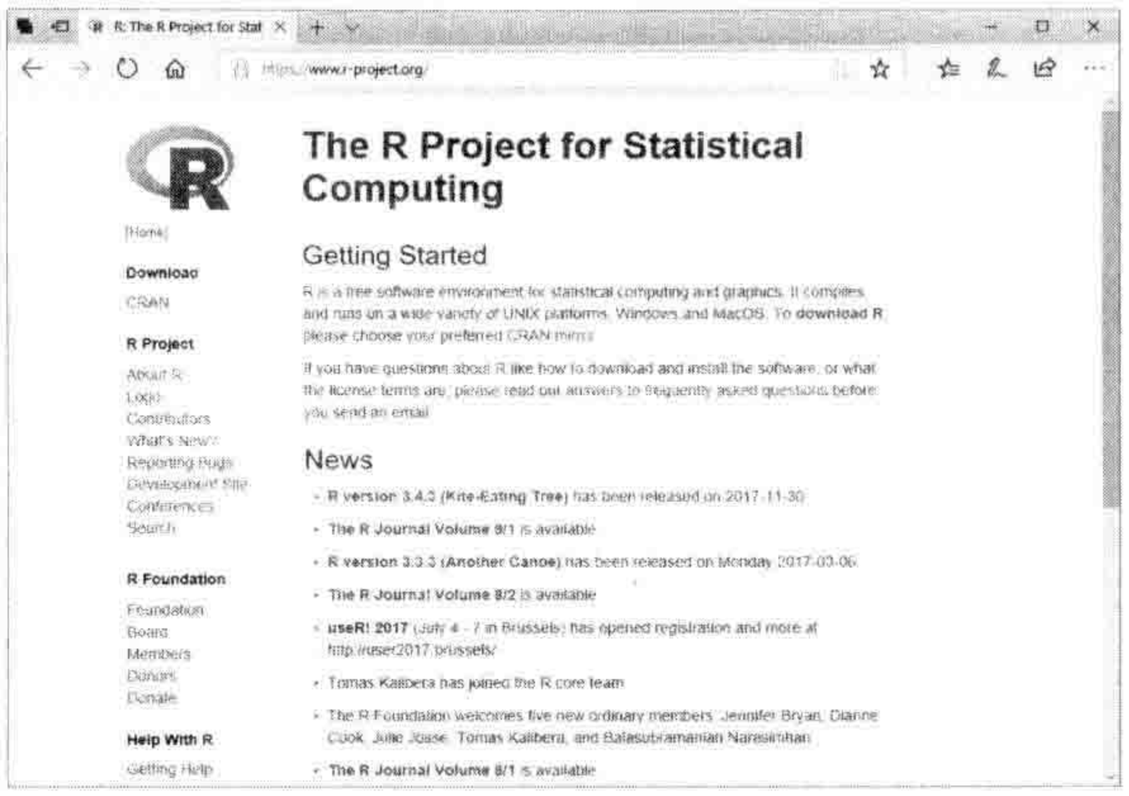


图 1-1

R 是一个开放源代码的自由软件，在国内外都拥有社区网站（网上论坛），提供使用的心得、讨论区以及更多的帮助文件。R 语言中文论坛网站如图 1-2 所示。

图 1-2

1.1.2 关于大数据

大数据是近年来兴起的一个名词，意指海量的数据分析与应用。和传统的数据分析相比，大数据更强调互动行为的数据分析，包括人与人的互动、人与物的互动。缺乏互动行为的庞大数据较难创造商业营利模式，将偏向理论性的研究。

约在 2000 年初期，在谷歌地图（Google Map）尚未出现之前，曾有一家地图服务公司（以下简称 A 公司）拥有最全的地图信息并制作了友好的网页界面，结合本地的项目与店家的联系信息，但最后的营收却不如预期，且在谷歌（Google）推出地图服务后被快速超越了，为何会如此呢？

地图信息拥有大量丰富且具有互动性的数据，很显然是大数据的一种类型，可是缺乏 3G 或 4G 网络以及方便、可携带式的移动设备（如手机或平板电脑）就无法产生互动的应用。我们坐在交通工具上拿着手机使用微信就可以和朋友互动，在路上搜索附近的店家也可以产生消费者与商店的互动，因此这一切互动应用在云计算（Cloud Computing）和云网络（Cloud Web）兴起且移

动设备普及之后才能真正发挥作用。

大数据是在云计算之后产生的名词。当网络的基础设施已经完备之后，用户通过移动设备产生的互动行为就可以累积成有价值的大数据。因此，回到刚刚所提到的A公司，很显然是推出服务时没有网络的基础设施，并且缺乏相关的应用作为支撑，导致实用性不如预期，后期也没有持续跟上市场发展的脚步。

一般把大数据的应用分为以下三类。

❖ 记录文件的应用

记录文件的工具是最常被使用的大数据分析工具，最常见的是从网页的操作行为中进行分析。例如，中国最大的购物网站之一——京东商城（https://www.jd.com/）就会分析客户购物的记录，并向客户推销他们可能感兴趣的商品，如图1-3所示。当我们曾经购买了日光灯管，下次登录该网站时，就会在该网站网页下方的广告中自动出现其他日光灯管的相关产品。

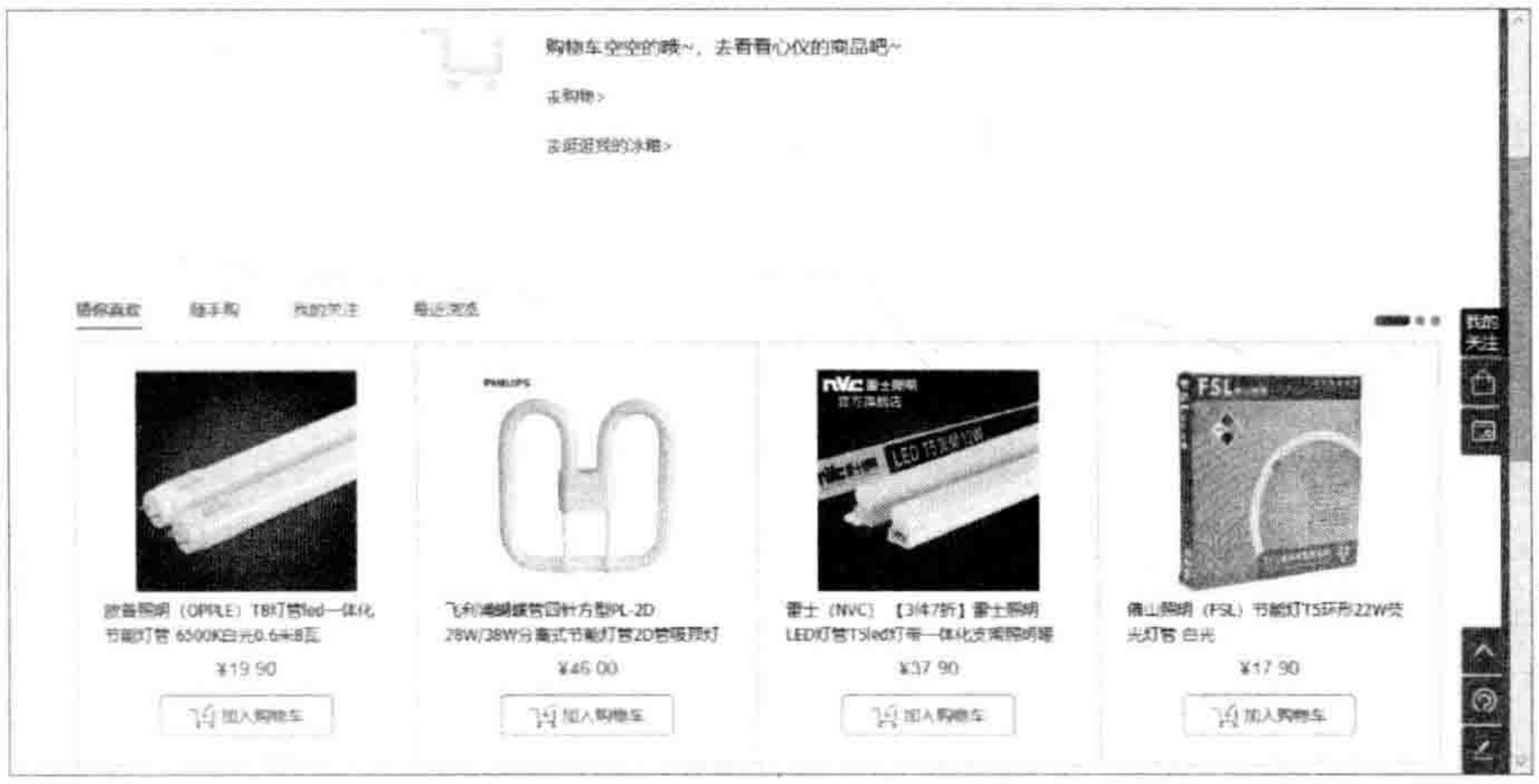


图1-3

❖ 社区用户行为

目前有许多大型的社交网站，如QQ、微博、百度贴吧、微信等，社区用户的行为记录与分析也是大数据的应用之一。举例来说，通过点赞的数量、关键词的分析等就能进行用户意向的分析。

另外，很多人会在社交网站上分享个人的身体状况，当一个人的好友们谈论到“生病”“感冒”等话题时，就可以对所在的区域进行分析，作为流行疾病开始发展的判断依据。

❖ 物联网

物联网是将人与人、人与物、物与物建立起关联关系而串接起来的大型网络。在物联网中，我们会寻找关联性高的人与物，例如一堆人出门吃饭时，很多人会表达“吃什么好？”（跟随者），会有一些人提供意见，称为“意见领袖”。团体中的意见领袖将主导（深深影响）其他人的想法与行为。以推销而言，通过意见领袖传达的效果会远高于一般人；以对象或软件而言，用户越多代表营销能力越强，广告效果越优异。

我们将人、软件、硬件当作一个个元件，只要有关联性，就通过线条连接起来，形成一个网状的图形。

如图 1-4 所示就是一个简单的物联网关系图，其中关联性[1]高的就是关键角色。

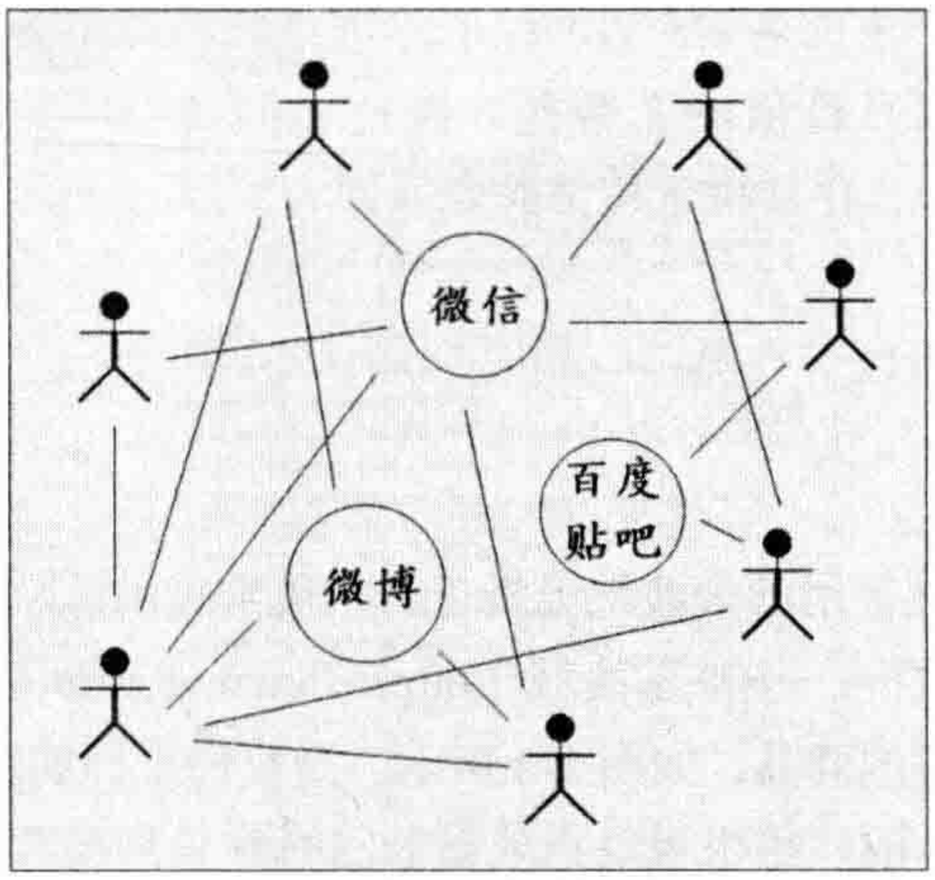


图 1-4

1.1.3 R 语言在大数据中的应用

R 语言是目前大数据应用的程序设计语言之一，使用 R 语言的理由包括下面 4 点：

❖ 简单的解释型语言

对于很多程序设计学习者而言，编译[2]是一个巨大的学习障碍。R 语言是一种不需要编译的程序设计语言，让程序编写者专心于功能性的正确，不需要花费太多时间在程序语法的调试上。下面举例说明 C 语言与 R 语言矩阵变量的定义与使用方式。

在 C 语言中，如果要定义一个矩阵，我们必须正确地给定需要加载的 header 头文件，并定义变量类型，代码如下：

```
#include <stdio.h>

void main() {
  int A[3][2];

  A[1][1]=1;
  A[2][1]=2;
  A[3][1]=3;
  A[1][2]=4;
  A[2][2]=5;
  A[3][2]=6;

}
```

[1] 一般学术的用语称为中心性（Centrality），目前有多种中心性的计算方式，包括 Degree Centrality、Closeness Centrality、Betweenness Centrality、Eigenvector Centrality、Katz Centrality 等，读者有兴趣可参考 https://en.wikipedia.org/wiki/Centrality。

[2] 编译，英文为 compile，程序设计者编写了程序源代码之后，通过编译程序（compiler）转换为可执行文件。编译程序在执行时会检查程序语法、变量声明、内存分配等语句，需完全正确才能编译完成。举例来说，一般常见的 C 语言就是由程序设计者编写一个或多个扩展文件名为 c 或 cpp 的文件，通过 GCC、VC、Borland C++等编译为可执行文件（在 Windows 上为 exe 文件）之后，就可以直接运行该可执行文件。

将该文件保存为 test.c，通过编译程序编译（在 VC 中选择 Compile 或 Build 命令）为 test.exe 之后，执行 test.exe 才会生效。

在 R 语言中，定义矩阵是一件很简单的工作，通过单行语句输入即可：

matrix(c(1,2,3,4,5,6), nrow=2)

其中，c(1,2,3,4,5,6)表示 6 个向量，nrow=2 表示 row（行）的数量为 2，执行后程序垂直按序排列，代码如下：

```
> matrix(c(1,2,3,4,5,6), nrow=2)
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

或者执行：

rbind(c(1,3,5),c(2,4,6))

其中，c(1,3,5)与 c(2,4,6)表示两个向量(1,3,5)与(2,4,6)，通过 rbind（意思为 row bind）将两行合并，执行过程如下：

```
> rbind(c(1,2,3),c(4,5,6))
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

或者执行：

cbind(c(1,2),c(3,4),c(5,6))

其中，c(1,2)、c(3,4)与 c(5,6)表示三个向量(1,2)、(3,4)与(5,6)，通过 cbind（意思为 column bind）将三列合并，执行过程如下：

```
> cbind(c(1,2),c(3,4),c(5,6))
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

在 R 语言中，可以通过行或列的合并（或者直接定义矩阵）直观地呈现一个我们需要的矩阵，非常简单。

提示

在现实世界的数据系统都属于与时间有关的动态系统，会记录每个时间点的数据，例如在物理上可记录每个时间点的位置、速度、能量等，在金融上可记录每个时间点的价格、数量、指标等，因此呈现的数据类型都为数组或矩阵。

❖ 大型的数据吞吐量

在 R 语言中，一个向量数量的理论值可达 2 的 31 次方，一个两列（例如第一列为时间，第二列为数值）的矩阵可以存储 10 亿项数据，因此一般而言，海量的数据可以存入一个大矩阵之中，

直接用于运算，便于用户直接操作或设计模型。

当然，矩阵的内容多就会导致内存的耗损并延迟执行的速度，因此数据的分类、分割与预处理是十分必要的。

❖ 多样的工具软件包

目前，在R语言上的软件包已经超过7000个，领域包含数学计算、数值分析、物理应用、金融相关等，内容包含公式计算、图表绘制、外部程序链接与数据库应用等，在各个领域中几乎都能找到对应的软件包与方便的工具。

由于R语言是标准的Open Source软件，因此任何人都可以上传自己做好的软件包到CRAN上（https://cran.r-project.org/submit.html），只要通过审核就能成为官方版的软件包。

❖ 免费且跨平台的软件

R语言是一款遵循GNU的自由软件，保证最终用户执行、学习、分享（复制）及编辑软件的自由，授予使用者以下权利：

- 以任何目的执行此程序。
- 将软件复制后再发行。
- 改良程序并公开发布。

目前，在常见的操作系统（如Windows、Linux、MAC OS）上都有对应的R语言版本可供安装。由于软件跨平台且随处可取得，增加了方便性与流通性，让更多人愿意使用并在上面进行开发，因此成为当下流行的程序设计语言之一。

1.2 单机版的R语言

对于个人而言，单机版的R软件是最容易安装上手的，下面介绍在Windows与Linux上的安装方式。

1.2.1 在Windows上安装R语言软件

步骤01 在Windows上安装时，R语言软件可到官方网站下载，如图1-5所示。

URL https://cran.r-project.org/bin/windows/base/

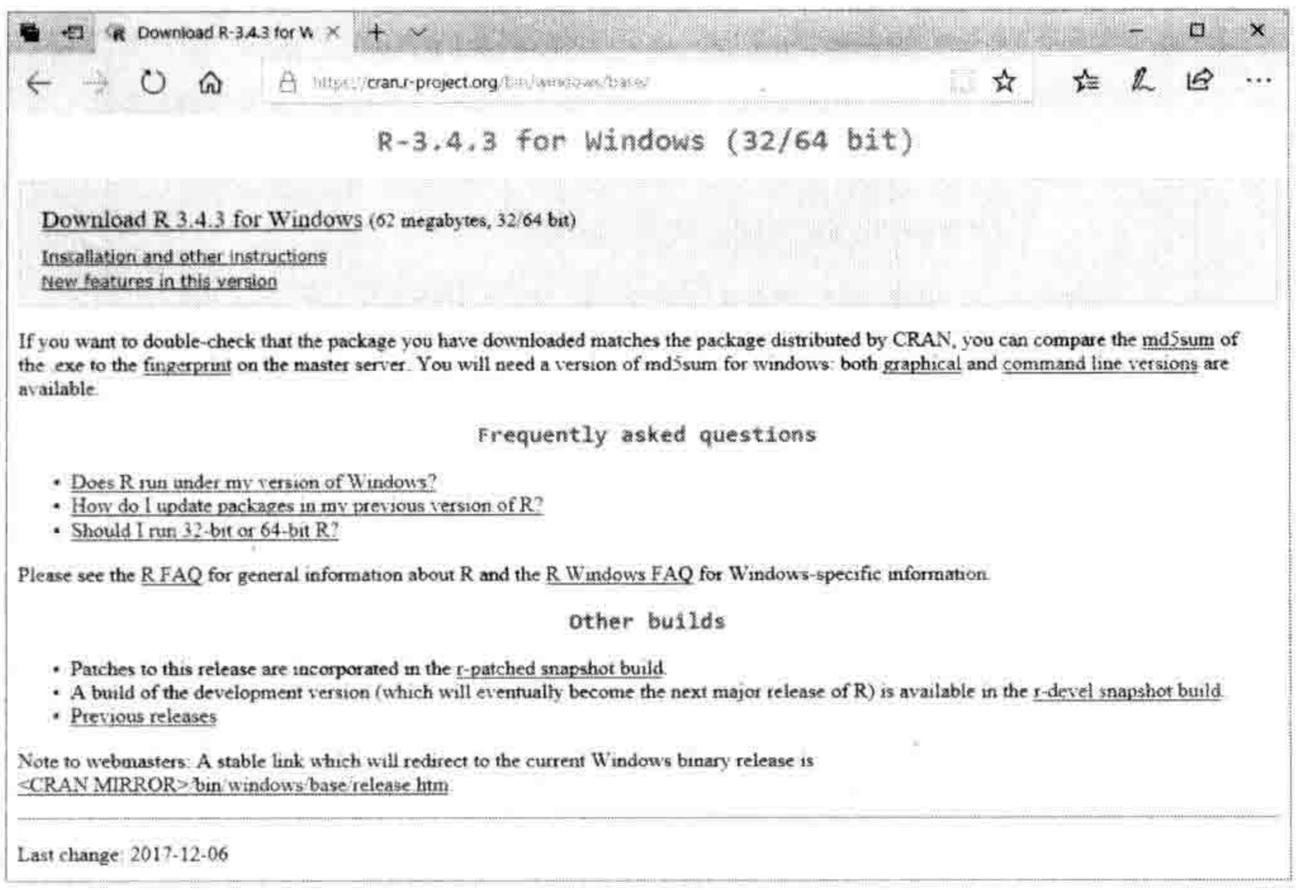


图 1-5

步骤02 下载后直接执行安装程序，可以选择安装过程的语言，如图 1-6 和图 1-7 所示。

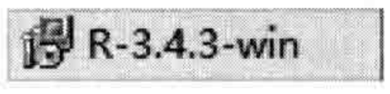


图 1-6

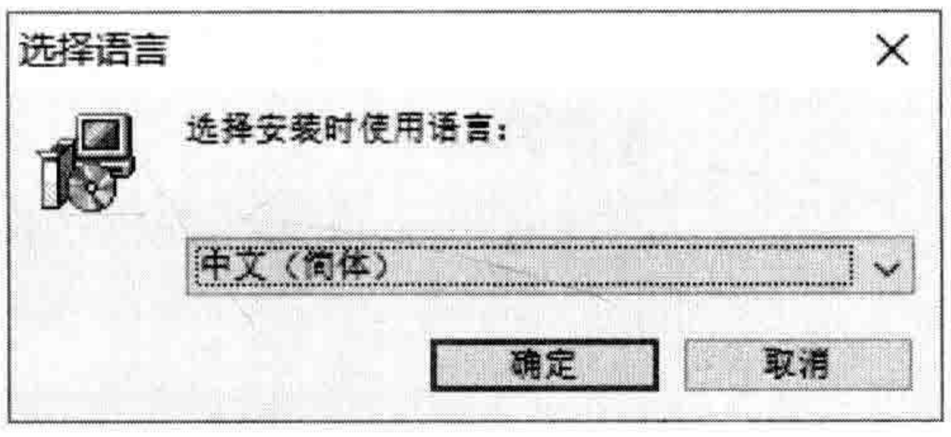


图 1-7

步骤03 出现版权说明，单击“下一步”按钮继续，如图 1-8 所示。

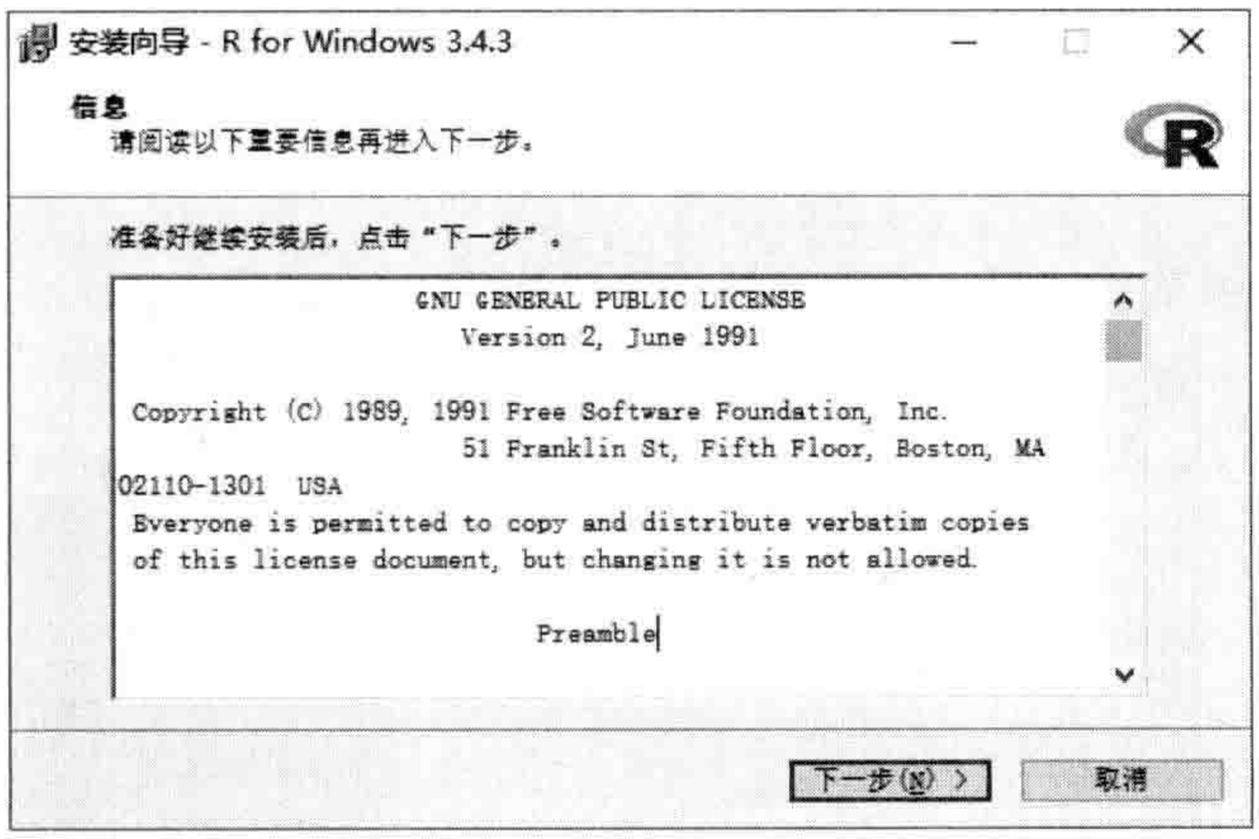


图 1-8

步骤04 显示要安装的路径，单击“下一步”按钮继续，如图 1-9 所示。

图 1-9

步骤05 选择要安装的内容，如果是 64 位的系统，可以取消勾选“32-bit Files”复选框（如果不取消勾选，在系统内会同时存在 32 位与 64 位的 R 语言执行程序），单击“下一步”按钮继续，如图 1-10 所示。

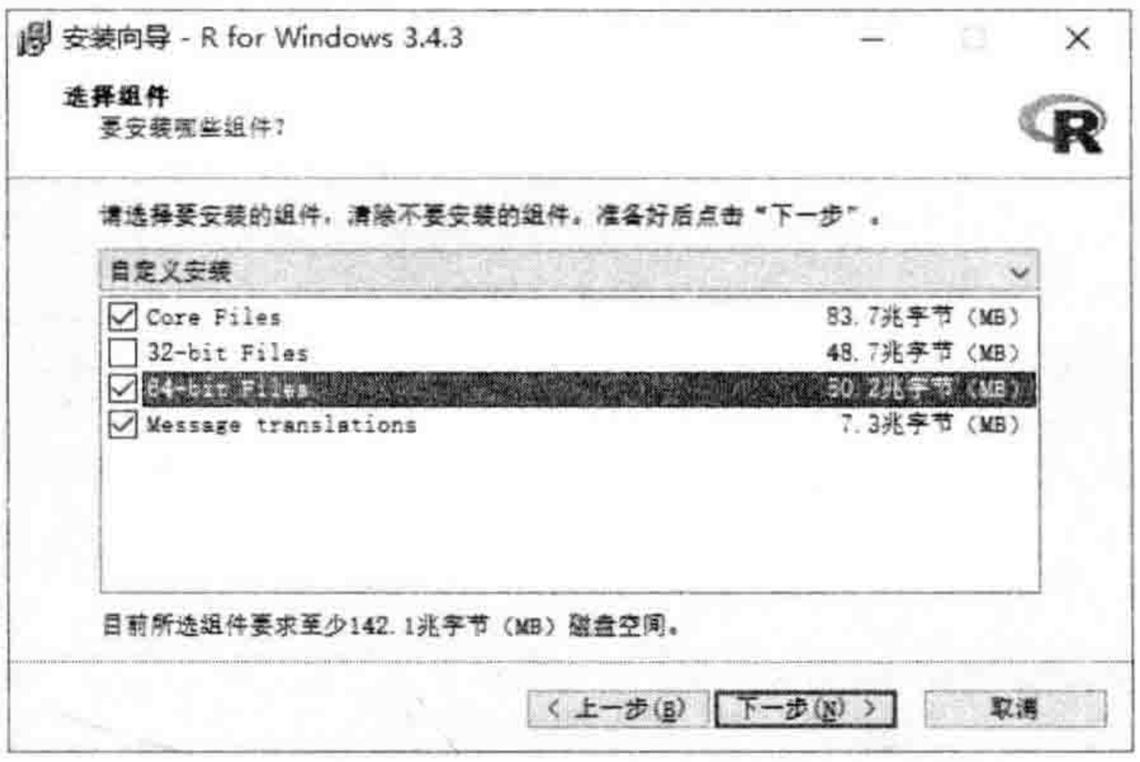


图 1-10

步骤06 使用默认值并单击“下一步”按钮继续，如图 1-11 所示。

图 1-11

步骤07 设置开始菜单内的程序名称，可使用默认值并单击“下一步”按钮继续，如图 1-12 所示。