R 语言编程 — 基于 tidyverse

作者：张敬信

单位：哈尔滨商业大学

时间：2020-08-28

版本：*1.0*

Email：zhjx_19@163.com

目录

作者序 1

开发环境 2

运行环境 2

测试数学公式 2

前言 3

0.1 怎么学习编程语言？ 3

例 1.1 计算并绘制 ROC 曲线 5

0.2 R 语言简介 9

0.2.1 什么是数据科学？ 9

0.2.2 什么是 R 语言？ 10

0.2.3 一个改变了 R 的人 12

0.3 R 语言编程思想 13

0.3.1 面向对象 13

0.3.2 面向函数 14

0.3.3 向量化编程 16

0.4 本书的特色与内容安排 18

0.4.1 本书的特色 18

0.4.2 本书的内容安排 19

1 基础语法 20

1.1 搭建 R 环境及常用操作 20

1.1.1 搭建 R 环境 20

1.1.2 常用操作 22

1.2 数据结构 I：向量、矩阵、多维数组 30

1.2.1 向量（一维数据） 30

1.2.2 矩阵（二维数据） 37

1.2.3 多维数组（多维数据） 40

1.3 数据结构 II：列表、数据框、因子 42

1.3.1 列表（list） 42

1.3.2 数据框（数据表） 45

1.3.3 因子（factor） 52

1.4 数据结构 III：字符串、日期时间 57

1.4.1 字符串 57

© 张敬信

1.4.2 日期时间 62
1.4.3 时间序列 68
1.5 正则表达式 70
1.6 控制结构 70
1.7 自定义函数 70

2 数据操作 **71**

3 可视化 **72**

4 应用统计 **73**

5 文档沟通 **74**

附录 **75**
5.1 R6 类面向对象编程简单实例 75

参考文献 **78**

© 张敬信

作者序

R 语言是专业的统计编程语言，具有顶尖水准的绘图功能，且开源免费有着丰富的扩展包和活跃的社区。R 语言这些优质的特性，使得它始终在数据统计分析领域的 SAS、STATA、SPSS、Python、Matlab 等同类软件中占据领先地位。

R 语言曾经最为人们津津乐道的是 Hadley 大神开发的 ggplot2 包，泛函式图层化语法赋予了绘图一种 ‘‘优雅’’ 美。近年来，R 语言在国外蓬勃发展，ggplot2 这个 ‘‘点’’ 在 2016 年以来，已被 Hadley 大神 ‘‘连成线、张成面、形成体（系）’’，这就是 tidyverse 包，集

数据导入 — 数据清洗 — 数据操作 — 数据可视化 — 数据建模 — 可重现与交互报告

整个数据科学流程于一身，而且是以 ‘‘现代的’’、‘‘优雅的’’ 方式，以管道式、泛函式编程技术实现。不夸张地说，tidyverse 操作数据比 pandas 好用、易用数倍！再加上可视化本来就是 R 所擅长，可以说 R 在数据科学领域强于 Python。这种整洁、优雅的 **tidy-流**，又带动了 R 语言在很多研究领域涌现出了一系列 **tidy-风格**的包。

在机器学习领域，曾经的 R 靠单打独斗的包，如今也正在从整合技术上迎头赶上 python，出现了 **tidy-风格**的 tidymodels 包，以及真正最新理念、最新技术、最新一代的机器学习 mlr3verse 包，它比 sklearn 还先进，基于 R6 类面向对象，data.table 神速数据底层，开创性的 **Graph-流**模式（图/网络流，区别于通常的线性流）。

在其它领域，如时间序列、金融、空间数据分析、大数据、生信等，R 语言也都涌现出一系列好用、易用的新包。

然而，我发现这些近几年出现的 R 语言新技术，在国内很少有人问津，绝大多数 R 语言的教师、教材、博客文章、R 学习者仍在沿用那些过时的、晦涩的 R 语法，对 R 语言的印象停留在 5 年前。

有感于此，我想写一本用最新 R 技术，方便新手真正快速入门 R 语言编程的书，来为 R 语言正名，以在国内推广已如此优秀好用的 R 语言。

我是一名大学数学教师，热爱编程、热爱 R 语言，奉行终生学习理念，一直喜欢跟踪和学习新知识、新技能。我对编程和 R 语言有一些独到的理解体会，因为我觉得数学语言与编程语言是相通的，都是用语法元素来表达和解决问题，我想把这些理解体会用符合国人的语言习惯表达出来。

希望我这本书，如果有幸进入了您的法眼，能让您学到正确的编程思想，学到最新的 R 语言编程知识，能真正让您完成 R 语言入门或汰旧换新。

— 张敬信，2020 年 8 月于哈尔滨

© 张敬信

开发环境

运行环境

本书是用黄湘云和叶飞的 ElegantBookdown 模板开发。

```
xfun::session_info(dependencies = FALSE)
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 16299)
##
## Locale:
##   LC_COLLATE=Chinese (Simplified)_China.936
##   LC_CTYPE=Chinese (Simplified)_China.936
##   LC_MONETARY=Chinese (Simplified)_China.936
##   LC_NUMERIC=C
##   LC_TIME=Chinese (Simplified)_China.936
##
## Package version:
##   compiler_4.0.2  magrittr_1.5    bookdown_0.20   htmltools_0.5.0
##   tools_4.0.2     yaml_2.2.1      stringi_1.4.6   rmarkdown_2.3
##   knitr_1.29      stringr_1.4.0   digest_0.6.25   xfun_0.16
##   rlang_0.4.7     evaluate_0.14
```

测试数学公式

引理 0.1 对任意两个随机变量 X_1, X_2, 它们具有相同的概率分布，当且仅当

$$\varphi_{X_1}(t) = \varphi_{X_2}(t)$$

♡

© 张敬信