

• 第二版

北京大学出版社，暨南大学出版社 2017.6（第二版）

• 数据收集与整理

• 1 引言

- 1.1 数据分析的未来
- 1.2 工欲善其事必先利其器
 - 1.2.1 四大分析利器简介
 - 一、数据管理工具
 - 二、报告撰写工具
 - 三、结果展示工具
 - 四、数据分析工具
 - 1.2.2 四大分析利器的比较
 - 一、办公软件比较
 - 二、统计分析软件比较
 - 1.2.3 数据分析工具的选择
 - 一、首选 WPS+R
 - 二、次选 Excel+R
 - 三、差钱 SQL+R
 - 四、专业 Oracle+R
 - 1.2.4 常用的数据分析软件
 - 一、专业的数据分析软件 SAS
 - 二、方便的数据分析软件 SPSS
 - 三、强大的数值分析软件 Matlab
 - 四、免费的数据分析软件 R语言
- 1.3 数据统计分析语言R简介
 - 1.3.1 什么是R语言
 - 1.3.2 为什么要用R语言
 - 1.3.3 R语言的优劣势
 - 1.3.4 如何发挥R的优势
- 练习题1

• 2 数据收集过程

- 2.1 统计数据
 - 2.1.1 基本概念
 - 定性数据
 - 定量数据
 - 2.1.2 分析思路
- 2.2 收集数据
 - 2.2.1 数据格式：结构化数据
 - 2.2.2 数据收集：表格或数据库
- 2.3 数据管理
 - 2.3.1 保存数据：电子表格
 - 2.3.2 输入数据

- 一、在R语言中输入数据 `c()`
- 二、从外部文件输入数据 `read.table()`
- 三、其他数据R语言读取 `read.csv()`
- 2.3.3 数据形式
 - 一、R语言数据类型
 - 数值型
 - 字符型
 - 逻辑型
 - 日期型
 - 缺省值
 - 二、R语言数据对象
 - 向量 `c()`
 - 数据框 `data.frame()`
- 练习题2

• 3 数据处理步骤

- 3.1 基本方法
 - 3.1.1 基本函数
 - 3.1.2 自定义函数 `function()`
 - 3.1.3 控制语句
 - 一、循环语句 `for()`
 - 二、分支语句 `if/else()`
- 3.2 数据选择
 - 3.2.1 选取观测
 - 一、下标法 `[],`
 - 二、记号法 `$`
 - 三、子集法 `subset`
 - 3.2.2 选取变量
 - 一、下标法 `[],`
 - 二、记号法 `$`
 - 三、数据框绑定法 `attach()`
 - 四、with函数法 `with()`
 - 3.2.3 选取观测与变量 `subset()`
 - 3.2.4 剔除观测与变量 -
- 3.3 数据转换
 - 3.3.1 修改变量名 `names()`
 - 3.3.2 创建变量
 - 3.3.3 变量变换 `as.`
 - 3.3.4 删除变量 `rm()`
 - 3.3.5 重新编码 `cut()`
- 3.4 数据整理
 - 3.4.1 数据集排序 `order()`
 - 3.4.2 数据集合并 `merge`
 - 3.4.3 缺失数据处理
 - 一、缺失值定义 `NA`

- 二、缺失值识别 `is.na()`
- 三、缺失值排除 `na.omit()`

- 练习题3

• 统计分析与建模

• 4 基本统计描述

- 4.1 基本图形函数
 - 4.1.1 高级绘图函数
 - 一、R中常用的高级函数
 - 二、高级函数的主要参数
 - 4.1.2 低级绘图函数
 - 4.1.3 绘图函数参数
 - 一、绘图参数的设置
 - 二、参数设置的用法
- 4.2 单变量(向量)数据分析
 - 4.2.1 计数数据分析
 - 一、分类频数表 `table()`
 - 二、分类条图 `barplot()`
 - 三、分类饼图 `pie()`
 - 4.2.2 计量数据分析
 - 一、集中趋势和离散程度 `summary()`
 - 二、茎叶图 `stem()`
 - 三、数值分类函数 `cut()`
 - 四、直方图 `hist()`
 - 五、正态概率图 `qqnorm()`
 - 七、箱式图 `boxplot()`
 - 4.2.3 分析函数构建
 - 一、基本统计量 `Stats()`
 - 二、探索性统计图 `EDA()`
 - 三、频数表构造函数 `Ftab()` , `Freq()`
- 4.3 多变量(数据框)数据分析
 - 4.3.1 计数类数据分析
 - 一、列联表 `table()`
 - 二、复式条图 `barplot()`
 - 4.3.2 计量类数据分析
 - 一、双变量散点图 `plot()`
 - 二、多变量散点图 `pairs()`
 - 4.3.3 计数计量数据分析
 - 一、分组函数 `by()`
 - 二、点带图函数 `stripchart()`
 - 三、分组散点图 `coplot()`
 - 4.3.4 应用类函数的应用
 - 一、矩阵应用函数 `apply()`
 - 二、列表应用函数 `lapply()`

- 三、分组应用函数 `tapply()`
- 四、聚集应用函数 `aggregate()`

- 练习题4

• 5 随机变量及其分布

- 5.1 随机变量及其分布
 - 5.1.1 离散型随机变量
 - 一、二项分布 `dbinom()`
 - 二、超几何分布 `dhyper()`
 - 三、泊松分布 `dexp()`
 - 5.1.2 连续型随机变量
 - 一、均匀分布 `dunif()`
 - 二、正态分布 `dnorm()`
 - 三、指数分布 `dexp()`
 - 5.1.3 R语言分布函数列表
- 5.2 随机抽样与随机数
 - 5.2.1 离散变量随机数
 - 一、二项分布随机数 `rbinom()`
 - 二、超几何分布随机数 `rhyper()`
 - 三、泊松分布随机数 `rpois()`
 - 5.2.2 连续变量随机数
 - 一、均匀分布随机数 `runif()`
 - 二、正态分布随机数 `rnorm()`
 - 三、指数分布随机数 `rexp()`
- 5.3 统计量及其抽样分布
 - 5.3.1 样本与统计量
 - 一、基本概念 `sample()`
 - 二、常用统计量
 - 5.3.2 常用的抽样分布
 - 一、几个常用的随机分布
 - 二、抽样分布的基本性质
 - 三、中心极限定理
 - 5.3.3 抽样分布的临界值
 - 一、标准正态分布概率表
 - 二、t分布临界值表

- 练习题5

• 6 常用统计推断方法

- 6.1 正态总体的参数估计
 - 6.1.1 参数估计的方法
 - 一、点估计
 - 二、区间估计 `z.conf.plot()`
 - 6.1.2 均值的区间估计 `t.conf.plot()`
- 6.2 正态总体的假设检验
 - 6.2.1 假设检验的概念

- 一、假设检验的基本思想
- 二、假设检验的基本步骤
- 6.2.2 单样本均值t检验
 - 一、正态性检验 `qqnorm()`
 - 二、t检验 `t.test()`
- 6.2.3 两样本均值t检验
 - 一、正态性检验 `shapiro.test()`
 - 二、方差齐性检验 `var.test()`
 - 三、均值的检验（方差不齐时）`t.test()`
 - 四、均值的检验（方差齐性时）`t.test()`
- 6.2.4 多样本均值方差分析
 - 一、方差分析简介
 - 二、单因素方差分析 `oneway.test()`
- 6.3 分布自由的非参数统计
 - 6.3.1 非参数统计简介
 - 一、非参数统计的用途
 - 二、秩的概念 `rank()`
 - 6.3.2 单样本非参数检验
 - 一、单样本分布K-S检验 `ks.test()`
 - 二、单样本中位数估计 `wilcox.test()`
 - 三、单样本符号秩检验 `wilcox.test()`
 - 6.3.3 两样本非参数检验
 - 一、两样本分布K-S检验 `ks.test()`
 - 二、两独立样本秩和检验 `wilcox.test()`
 - 6.3.4 多样本非参数检验 `kruskal.test()`
- 6.4 计数数据的统计推断
 - 6.4.1 单样本数据统计推断
 - 一、比例的区间估计 `prop.test()`
 - 二、单样本拟合优度检验 `chisq.test()`
 - 6.4.2 列联表数据卡方检验 `chisq.test()`

• 练习题6

• 7 基本统计分析模型

- 7.1 线性相关分析模型
 - 7.1.1 线性相关系数的计算 `cor()`
 - 7.1.2 相关系数的假设检验 `cor.test()`
 - 7.1.3 分组数据的相关分析
- 7.2 线性回归分析模型
 - 7.2.1 一元线性回归模型
 - 一、一元线性回归模型建立 `lm()`
 - 二、一元线性回归模型检验 `summary()`
 - 7.2.2 多元线性回归模型
 - 一、多元线性回归模型简介 `lm()`
 - 二、多元线性回归模型建立 `summary()`

- 7.2.3 多元回归模型诊断
 - 一、残差值 resid()
 - 二、杠杆值
 - 三、学生化残差 rstudent()
 - 四、强影响值
- 7.2.4 分组多元回归模型
- 7.3 数据分类与模型选择
 - 7.3.1 数据与模型
 - 一、变量的取值类型
 - 二、模型选择方式
 - 7.3.2 线性模型分析
 - 一、单因素方差分析模型 anova()
 - 二、两因素方差分析模型 anova()
- 练习题7

• 大数据分析入门

• 8 R语言的高级应用

- 8.1 R语言的编程概述
 - 8.1.1 R语言编程基础
 - 8.1.2 R语言编程对象
 - 一、向量 c()
 - 二、矩阵 matrix()
 - 三、数组 array()
 - 四、因子 factor()
 - 五、数据框 data.frame()
 - 六、列表 list()
 - 七、对象识别与转换 as.
 - 8.1.3 R程序的数学运算
 - 一、R语言的运算符
 - 二、常用的数学与字符函数
 - 三、R语言的数值计算
 - 8.1.4 R中字符与时间函数
- 8.2 R语言高级编程举例
 - 8.2.1 自定义函数技巧 function()
 - 8.2.2 自定义统计函数
 - 一、计算基本统计量 Stats()
 - 二、绘制探索性统计图 EDA()
 - 三、自定义计数频数表函数 Ftab()
 - 四、自定义计量频数表函数 Freq()
 - 8.2.3 自定义检验函数
 - 一、自定义置信区间函数
 - 二、自编单样本t检验程序 t.test1()
 - 三、自编两样本t检验程序 t.test2()
- 8.3 R语言高级绘图功能

- 8.3.1 绘制特殊统计图
- 8.3.2 lattice绘图系统
 - 一、lattice绘图和普通绘图的区别
 - 二、常用的 lattice 绘图函数
 - 三、模拟例子
- 8.3.3 ggplot2绘图系统
 - 一、ggplot2简介
 - 二、ggplot2绘图 ggplot()
 - 三、为什么要用ggplot2
 - 四、qplot绘图函数快速入门 qplot()
- 8.4 结果输出与报告生成
 - 8.4.1 脚本的输入和结果的输出
 - 一、脚本输入 source()
 - 二、文本输出 sink()
 - 三、图形输出 pdf()
 - 8.4.2 使用Rmarkdown统计分析
 - 8.4.3 使用Rmarkdown生成报告
 - 8.4.3 使用Markdown的好处
- 练习题8

• 9 R语言大数据分析入门

- 9.1 统计模拟实验
 - 9.1.1 随机模拟方法简介
 - 一、随机模拟 cumsum()
 - 二、模拟大数定律 Bernoulli()
 - 三、利用模拟方法求积分
 - 9.1.2 模拟函数的建立方法
 - 一、R的重复函数 replicate()
 - 二、模拟函数应用
 - 9.1.3 对模拟的进一步认识
 - 一、t统计量的稳健性验证
 - 二、各种分布的模拟研究
- 9.2 R语言中数据库的使用
 - 9.2.1 为何要使用数据库
 - 9.2.2 关系型数据库简介
 - 9.2.3 R语言数据库包
 - 一、从Excel中读取数据 odbcConnectExcel()
 - 二、从数据库中读取数据 odbcConnect()
- 9.3 调查数据的设计与分析
 - 9.3.1 调查表的设计
 - 9.3.2 调查数据的管理
 - 9.3.3 调查数据的分析
 - 一、单因素分析
 - 二、两因素分析
 - 三、多因素分析
- 练习题9

• 附录 Rstudio简介