电子信息类新技术丛书

数据分析与R

韩忠明 段大高◎著

SHUJU FENXI YU R

北京邮电大学出版社
www.buptpress.com

电子信息类新技术丛书

数据分析与R

韩忠明　段大高　著

北京邮电大学出版社
www. buptpress. com

内 容 简 介

信息系统、互联网、移动通信等的快速发展催生了海量的数据。从数据中分析、挖掘隐藏在其中的模式、规律等是发挥数据价值的根本途径。采用有效的工具、方法是分析挖掘数据的基础。R是一个开放、高效的数据分析平台，本书介绍了R的基本功能、数据管理功能、详细描述了R实现各种分析图形的方法。本书详细地介绍了数据分析的整体流程，涵盖了数据获取、数据预处理和常见的数据分析方法。采用R实现了主流的数据预处理方法，详细介绍了方差分析、Logistic回归、聚类和分类以及用于数据分析的EM算法和MCMC模拟，分析了这些技术的基本原理和实现算法，应用R实现了分析模型与应用过程。本书采用大量真实数据和案例作为驱动，分析了在实际问题中如何利用相关技术解决分析问题。

本书既可供从事数据分析、数据挖掘等的研究者、应用者参考，也可供在市场营销、金融、医疗等行业从事数据分析的人士参考。

图书在版编目(CIP)数据

数据分析与R / 韩忠明，段大高著. --北京 ：北京邮电大学出版社，2014.8
ISBN 978-7-5635-4064-8

Ⅰ. ①数… Ⅱ. ①韩… ②段… Ⅲ. ①统计分析 Ⅳ. ①O212.1
中国版本图书馆CIP数据核字(2014)第168625号

书　　名：数据分析与R
著作责任者：韩忠明　段大高　著
责任编辑：刘　颖
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路10号(邮编：100876)
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：北京鑫丰华彩印有限公司
开　　本：720 mm×1 000 mm　1/16
印　　张：16.75
字　　数：297千字
版　　次：2014年8月第1版　2014年8月第1次印刷

ISBN 978-7-5635-4064-8　　定　价：38.00元

前　言

在互联网快速发展之前，数据分析一直被认为是统计学家的数字游戏，数据分析只是服务于金融、市场等领域专家的辅助研究手段。随着信息时代的到来，数据无处不在，数据变成问题产生和问题解决的出发点，数据变成了经济增长、企业赢利、疾病预测的推动力。数据的形态、产生方式、速度、数据量都在快速变化，这些变化推动着数据分析的需求增长和技术发展。

信息时代的数据分析正以从未想象过的方式影响着人类的经济、社会以及日常生活。在知识经济与信息技术时代，每个人都面临着如何有效从大量公开的数据中探索能够激发创新机会的机遇和挑战。“工欲善其事，必先利其器”，有效的工具、方法是分析挖掘数据的基础，R 作为一个开放、高效、强大、简洁的数据分析与挖掘平台，正受到越来越多学术界、工业界用户的欢迎。无论是金融，还是市场、医疗、互联网、甚至食品等领域，采用 R 作为数据分析的平台正成为一个趋势。

数据分析理论、方法、工具的发展推动了数据的产生；从数据直接产生的数据分析需求反过来也促进了数据分析相关理论与算法的发展，数据分析领域正面临着许多新的挑战：

（1）数据来源多，数据的异构性强，所以数据预处理对分析结果非常重要，在数据预处理时常用的方法有哪些？如何在 R 中实现这些数据与处理方法？我们将在本书揭晓。

（2）数据复杂性高，高维复杂数据快速产生的同时，也带来大量噪声、异常、非规范的数据。高效的数据分析平台和算法是研究复杂数据的基础，本书将阐述如何利用 R 实现高效的分析算法。

（3）数据分析的展现，现代数据分析要求数据分析结果有直观的展

示。数据分析者是专业人员，而数据分析的用户是领域专家和决策者，充分利用各种直观的图表进行结果展示是数据分析的重要任务。本书将利用R实现数据分析的直观展示。

(4) 多学科的融合，现代数据分析融合了计算机科学、统计学、机器学习等不同领域。Logistic回归，EM算法等将会在数据分析中有广阔用武之地。本书将重点研究和实现数据分析中的核心技术以及R实现。

(5) 大数据问题。一方面，现代数据分析面临的数据量大、数据复杂性高等问题；另一方面，很多现实问题需要的大量真实数据又难以保证有效获取，所以仿真算法和一些新型模型是解决这个矛盾的一个有效方法。本书将研究和实现MCMC仿真的传统和新算法。

本书包含着我们最新的一些研究成果和实际项目经验，我们希望将数据分析的传统技术、现代方法以及实际问题中的应用结合，从应用出发、从用户出发，构建基于R的现代数据分析系统。

在本书的写作过程中，我们得到了很多热情的鼓励和支持，包含国家自然科学基金、中央财政专项人才建设项目等项目的支持，也受到了北京工商大学的很多支持，团队实验室的研究生谭旭升、张梦参与了部分章节的编写。很多人都对本书付出了辛勤的劳动，在此我们对他们一并表示感谢。

作　者

目录

第1章 数据分析基础

信息时代涌现了海量数据，海量数据中蕴含着揭示事物本质的知识与规律，所以需要对这些数据进行分析得出这些知识与规律。本章介绍数据分析所需数理统计的基础知识以及数据分析涉及的基本工具。

1.1 统计基础

1.1.1 概率与统计

1. 随机试验与随机事件

(1) 随机试验

根据特定研究目的，在一定条件下对自然现象所进行的观察或试验统称为试验。如果一个试验满足下述三个特性，则称其为一个随机试验：

① 试验可以在相同条件下多次重复进行；

② 每次试验的可能结果不止一个，并且事先知道会有哪些可能的结果；

③ 每次试验总是恰好出现这些可能结果中的一个，但在一次试验之前却不能肯定这次试验会出现哪一个结果。

如果没有特殊声明，后文中简称随机试验为试验。如在一定条件下，观察产品的合格情况；又如从52张扑克牌中随机抽取一张等，它们都具有随机试验的三个特征，因此都是随机试验。

(2) 随机事件

随机试验的每一种可能结果，在一定条件下可能发生，也可能不发生，称为随

机事件，简称事件，通常用大写斜体字母 A、B、C 等来表示。

① 基本事件。不能再分的事件称为基本事件，也称为样本点。

例 1 在编号为 1、2、3、…、10 的十个球中随机抽取 1 个，有 10 种不同的可能结果："取得一个编号是 1"、"取得一个编号是 2"、…、"取得一个编号是 10"，这 10 个事件都是不可能再分的事件，它们都是基本事件。由若干个基本事件组合而成的事件称为复合事件。如"取得一个编号是偶数"是一个复合事件，它由"取得一个编号是 2"、"是 4"、"是 6"、"是 8"、"是 10"5 个基本事件组合而成。

② 必然事件。在一定条件下必然会发生的事件称为必然事件，用 Ω 表示。例如，在例 1 中，取出编号小于 10 的球，就是一个必然事件。

③ 不可能事件。在一定条件下不可能发生的事件称为不可能事件，用 Φ 表示。例如，在例 1 中，取出编号大于 10 的球，就是一个不可能事件。

必然事件与不可能事件实际上是确定性现象，即它们不是随机事件，但是为了方便起见，我们把它们看作为两个特殊的随机事件。

2. 概率

概率是一个能够刻画事件发生可能性大小的数量指标，是事件本身所固有的，且不随人的主观意志而改变。事件 A 的概率记为 $P(A)$。下面介绍概率的统计定义。

在相同条件下进行 n 次重复试验，如果随机事件 A 发生的次数为 m，那么 m/n 称为随机事件 A 的频率；当试验重复数 n 逐渐增大时，随机事件 A 的频率越来越稳定地接近某一数值 p，那么就把 p 称为随机事件 A 的概率。这样定义的概率称为统计概率，或者称后验概率。

在一般情况下，随机事件的概率 p 是不可能准确得到的。通常以试验次数 n 充分大时随机事件 A 的频率作为该随机事件概率的统计定义的近似值，即

$$P(A)=p\approx m/n \quad (n \text{ 充分大})$$

对于某些随机事件，用不着进行多次重复试验来确定其概率，而是根据随机事件本身的特性直接计算其概率。

有很多随机试验具有以下特征：

① 试验的所有可能结果只有有限个，即样本空间中的基本事件只有有限个；

② 各个试验的可能结果出现的可能性相等，即所有基本事件的发生是等可能的；

③ 试验的所有可能结果两两互不相容。

具有上述特征的随机试验，称为古典概型。

对于古典概型，概率的定义如下：设样本空间由 n 个等可能的基本事件所构成，其中事件 A 包含有 m 个基本事件，则事件 A 的概率为 m/n，即

$$P(A)=m/n$$

这样定义的概率称为古典概率或先验概率。

根据概率的定义，概率有如下基本性质：

① 对于任何事件 A，有 $0\leqslant P(A)\leqslant 1$；

② 必然事件的概率为 1，即 $P(\Omega)=1$；

③ 不可能事件的概率为 0，即 $P(\Phi)=0$。

3. 常见概率分布

(1) 离散型随机变量及其分布

离散型随机变量 ξ 的分布满足下列性质：

①（非负性）$p_i\geqslant 0$；

②（规范性）$\sum_{i=1}^{+\infty} p_i = 1$。

常见的离散型分布如下：

① 退化分布（单点分布）

$$F(x)=\begin{cases}0, & x<a,\\ 1, & x\geqslant a,\end{cases}\qquad P\{\xi=a\}=1$$

② 贝努里分布（两点分布）

$$\begin{pmatrix}0 & 1\\ q & p\end{pmatrix}\text{或}P\{X=x\}=p^x(1-p)^{1-x},\quad x=0,1$$

③ 二项分布

$$B(k;n,p)=P\{\mu=k\}=\binom{n}{k}p^kq^{n-k},\quad k=0,1,2,\cdots,n$$

④ 泊松分布

$$P\{\xi=k\}=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},\quad k=0,1,2,\cdots,n,\quad \lambda>0$$

(2) 连续性随机变量及概率密度函数

设 ξ 是随机变量，$F(x)$ 是它的分布函数，若存在一个非负可积函数 $p(x)$ 使得对任意的 $x\in(-\infty,+\infty)$，有 $F(x)=P\{\xi\leqslant x\}=\int_{-\infty}^{x}p(t)\mathrm{d}t$，则称 ξ 为连续性随

机变量，称 $p(x)$ 为 ξ 的概率密度函数或分布密度函数。

密度函数具有如下性质：

① （非负性）$p(x)\geqslant 0, x\in \mathbf{R}$；

② （规范性）$\int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1$；

③ 若 $p(x)$ 在 x 处是连续的，则 $F'(x)=p(x)$；

④ 设 a,b 为任意实数，且 $a<b$，则 $p\{a<\xi\leqslant b\} = \int_a^b p(x)\mathrm{d}x$；

⑤ 若 ξ 是连续型随机变量，则 $\forall a\in\mathbf{R}, P\{\xi=a\}=0$。

常见的连续型分布如下：

① 均匀分布

$$U[a,b],\quad p(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{其他}\end{cases}$$

② 正态分布

$$N(a,\sigma^2),\quad p(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}},\quad -\infty<x<+\infty$$

③ 指数分布

$$P(\lambda),\quad p(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & x\geqslant 0,\\ 0, & x<0,\end{cases}\quad \lambda>0$$

下面我们重点介绍数据分析中常见的几种离散型与连续性随机变量的分布。

① 正态分布

若连续型随机变量 x 的概率分布密度函数为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1-1}$$

其中，μ 为平均数，σ^2 为方差，则称随机变量 x 服从正态分布，记为 $x\sim N(\mu,\sigma^2)$。相应的概率分布函数为

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \tag{1-2}$$

分布密度曲线如图 1.1 所示。

由式(1-1)和图 1.1 可以看出正态分布具有以下几个重要特征：

a. 正态分布密度曲线是单峰、对称的悬钟形曲线，对称轴为 $x=\mu$。

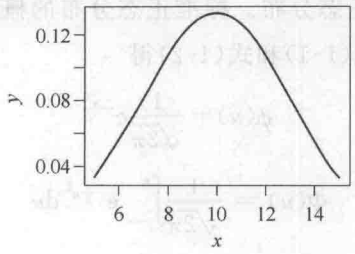

图 1.1　正态分布密度曲线图

b. $f(x)$在 $x=\mu$ 处达到极大，极大值 $f(\mu)=\dfrac{1}{\sigma\sqrt{2\pi}}$。

c. $f(x)$是非负函数，以 x 轴为渐近线，分布从$-\infty$至$+\infty$。

d. 曲线在 $x=\mu\pm\sigma$ 处各有一个拐点，即曲线在$(-\infty,\mu-\sigma)$和$(\mu+\sigma,+\infty)$区间上是下凸的，在$[\mu-\sigma,\mu+\sigma]$区间内是上凸的。

e. 正态分布有两个参数，即平均数 μ 和标准差 σ。μ 是位置参数，如图 1.2 所示。当 σ 恒定时，μ 愈大，则曲线沿 x 轴愈向右移动；反之，μ 愈小，曲线沿 x 轴愈向左移动。σ 是变异度参数，如图 1.2 所示。当 μ 恒定时，σ 愈大，表示 x 的取值愈分散，曲线愈“胖”；σ 愈小，x 的取值愈集中在 μ 附近，曲线愈“瘦”。

f. 分布密度曲线与横轴所夹的面积为 1，即

$$P(-\infty<x<+\infty)=\int_{-\infty}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=1$$

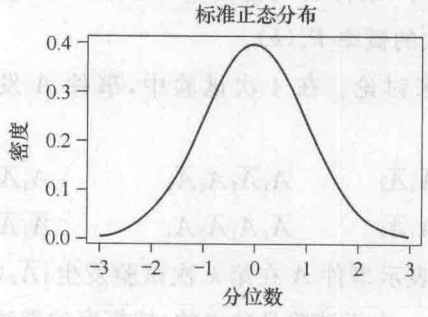

图 1.2　标准正态分布图

由上述正态分布的特征可知，正态分布是依赖于参数 μ 和 σ^2（或 σ）的一簇分布，正态曲线之位置及形态随 μ 和 σ^2 的不同而不同。这就给研究具体的正态总体带来困难，需将一般的 $N(\mu,\sigma^2)$ 转换为 $\mu=0,\sigma^2=1$ 的正态分布。我们称 $\mu=0$，

$\sigma^2=1$ 的正态分布为标准正态分布。标准正态分布的概率密度函数及分布函数分别记作 $\varphi(u)$ 和 $\Phi(u)$，由式(1-1)和式(1-2)得

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \tag{1-3}$$

$$\Phi(u)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-\frac{1}{2}u^2}\,du \tag{1-4}$$

随机变量 u 服从标准正态分布，记作 $u\sim N(0,1)$。

对于任何一个服从正态分布 $N(\mu,\sigma^2)$ 的随机变量 x，都可以通过标准化变换：

$$u=(x-\mu)/\sigma \tag{1-5}$$

将其变换为服从标准正态分布的随机变量 u。u 称为标准正态变量或标准正态离差。

② 二项分布

将某随机试验重复进行 n 次，若各次试验结果互不影响，即每次试验结果出现的概率都不依赖于其他各次试验的结果，则称这 n 次试验是独立的。

对于 n 次独立的试验，如果每次试验结果出现且只出现对立事件 A 与 $\overline{A}$ 之一，在每次试验中出现 A 的概率是常数 $p(0<p<1)$，因而出现对立事件 $\overline{A}$ 的概率是 $1-p=q$，则称这一串重复的独立试验为 n 重贝努利试验，简称贝努利试验。

在生物学研究中，我们经常碰到的一类离散型随机变量，例如孵 n 枚种蛋的出雏数、n 头病畜治疗后的治愈数、n 尾鱼苗的成活数等，可用贝努利试验来概括。

在 n 重贝努利试验中，事件 A 可能发生 $0,1,2,\cdots,n$ 次，现在我们来求事件 A 恰好发生 $k(0\leqslant k\leqslant n)$ 次的概率 $P_n(k)$。

先取 $n=4,k=2$ 来讨论。在 4 次试验中，事件 A 发生 2 次的方式有以下 C_4^2 种：

$$A_1A_2\overline{A}_3\overline{A}_4 \qquad A_1\overline{A}_2A_3\overline{A}_4 \qquad A_1\overline{A}_2\overline{A}_3A_4$$
$$\overline{A}_1A_2A_3\overline{A}_4 \qquad \overline{A}_1A_2\overline{A}_3A_4 \qquad \overline{A}_1\overline{A}_2A_3A_4$$

其中，$A_k(k=1,2,3,4)$ 表示事件 A 在第 k 次试验发生；$\overline{A}_k(k=1,2,3,4)$ 表示事件 A 在第 k 次试验不发生。由于试验是独立的，按概率的乘法法则，于是有

$$P(A_1A_2\overline{A}_3\overline{A}_4)=P(A_1\overline{A}_2A_3\overline{A}_4)=\cdots=P(\overline{A}_1\overline{A}_2A_3A_4)$$
$$=P(A_1)\cdot P(A_2)\cdot P(\overline{A}_3)\cdot P(\overline{A}_4)=p^2q^{4-2}$$

又由于以上各种方式中，任何两种方式都是互不相容的，按概率的加法法则，在 4 次试验中，事件 A 恰好发生 2 次的概率为

$$P_4(2)=P(A_1A_2\overline{A}_3\overline{A}_4)+P(A_1\overline{A}_2A_3\overline{A}_4)+\cdots+P(\overline{A}_1\overline{A}_2A_3A_4)=C_4^2p^2q^{4-2}$$

一般，在 n 重贝努利试验中，事件 A 恰好发生 $k(0\leqslant k\leqslant n)$ 次的概率为

$$P_n(k)=C_n^kp^kq^{n-k},\quad k=0,1,2,\cdots,n \tag{1-6}$$

若把式(1-6)与二项展开式

$$(q+p)^n=\sum_{k=0}^{n}C_n^kp^kq^{n-k}$$

相比较就可以发现，在 n 重贝努利试验中，事件 A 发生 k 次的概率恰好等于 $(q+p)^n$ 展开式中的第 $k+1$ 项，所以也把式(1-6)称作二项概率公式。

二项分布定义如下：

设随机变量 x 所有可能取的值为零和正整数：$0,1,2,\cdots,n$，且有

$$P_n(k)=C_n^kp^kq^{n-k},\quad k=0,1,2,\cdots,n$$

其中，$p>0$，$q>0$，$p+q=1$，则称随机变量 x 服从参数为 n 和 p 的二项分布，记为 $x\sim B(n,p)$。

显然，二项分布是一种离散型随机变量的概率分布。参数 n 称为离散参数，只能取正整数；p 是连续参数，它能取 0 与 1 之间的任何数值（q 由 p 确定，故不是另一个独立参数）。

容易验证，二项分布具有概率分布的一切性质，即

a. $P(x=k)=P_n(k)\geqslant 0,\quad k=0,1,\cdots,n$。

b. 二项分布的概率之和等于 1，即

$$\sum_{k=0}^{n}C_n^kp^kq^{n-k}=(q+p)^n=1$$

c. $P(x\leqslant m)=P_n(k\leqslant m)=\sum_{k=0}^{m}C_n^kp^kq^{n-k}$。

d. $P(x\geqslant m)=P_n(k\geqslant m)=\sum_{k=m}^{n}C_n^kp^kq^{n-k}$。

e. $P(m_1\leqslant x\leqslant m_2)=p_n(m_1\leqslant k\leqslant m_2)=\sum_{k=m_1}^{m_2}C_n^kp^kq^{n-k}(m_1<m_2)$。

二项分布由 n 和 p 两个参数决定：

a. 当 p 值较小且 n 不大时，分布是偏倚的。但随着 n 的增大，分布逐渐趋于对称，如图 1.3 所示。

b. 当 p 值趋于 0.5 时，分布趋于对称，如图 1.4 所示。

c. 对于固定的 n 及 p，当 k 增加时，$P_n(k)$ 先随之增加并达到其极大值，以后又下降。

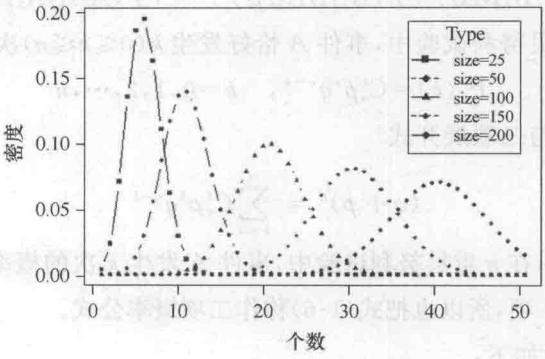

图 1.3　n 值不同的二项分布比较

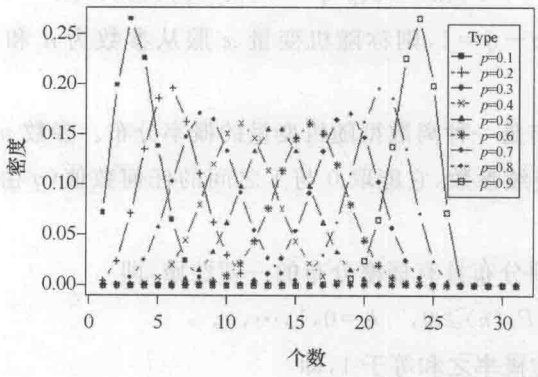

图 1.4　p 值不同的二项分布比较

此外，在 n 较大，np、nq 较接近时，二项分布接近于正态分布；当 $n\to\infty$ 时，二项分布的极限分布是正态分布。

前面已经指出二项分布由参数 n 和 p 决定。统计学证明，服从二项分布 $B(n,p)$ 的随机变量之平均数 μ、标准差 σ 与参数 n、p 有如下关系。

当试验结果以事件 A 发生次数 k 表示时

$$\mu=np$$

$$\sigma=\sqrt{npq}$$

③ 泊松分布

泊松分布是一种可以用来描述和分析随机地发生在单位空间或时间里的稀有事件的概率分布。要观察到这类事件，样本含量 n 必须很大。在生物、医学研究