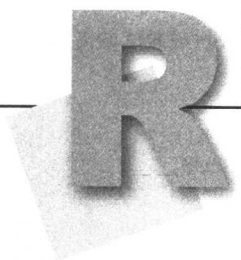
R统计建模与R软件

薛 毅 陈立萍 编著

http://www.tup.com.cn

清华大学出版社



统计建模与R软件

薛 毅 陈立萍 编著

清华大学出版社
北京

内 容 简 介

本书以统计理论为基础，按照数理统计教材的章节顺序，在讲明统计的基本概念的同时，以 R 软件为辅助计算手段，介绍统计计算的方法，从而有效地解决统计中的计算问题.

书中结合数理统计问题对 R 软件进行科学、准确和全面的介绍，以便使读者能深刻理解该软件的精髓和灵活、高效的使用技巧. 此外，还介绍了在工程技术、经济管理、社会生活等各方面的丰富的统计问题及其统计建模方法，通过该软件将所建模型进行求解，使读者获得从实际问题建模入手，到利用软件进行求解，以及对计算结果进行分析的全面训练.

本书可作为理工、经济、管理、生物等专业学生数理统计课程的辅导教材或教学参考书，也可作为统计计算课程的教材和数学建模竞赛的辅导教材.

图书在版编目（CIP）数据

统计建模与 R 软件/薛毅，陈立萍编著. —北京：清华大学出版社，2007.4

ISBN 978-7-302-14366-6

Ⅰ. 统…　Ⅱ. ① 薛… ② 陈…　Ⅲ. 统计分析—应用软件　Ⅳ. C819

中国版本图书馆 CIP 数据核字(2006)第 158986 号

责任编辑：刘　颖　赵从棉

责任校对：赵丽敏

责任印制：孟凡玉

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座

http://www.tup.com.cn　**邮　　编**：100084

c-service@tup.tsinghua.edu.cn

社 总 机：010-62770175　**邮购热线**：010-62786544

投稿咨询：010-62772015　**客户服务**：010-62776969

印 刷 者：北京市清华园胶印厂

装 订 者：三河市溧源装订厂

经　　销：全国新华书店

开　　本：185×230　**印　张**：33.75　**字　数**：716 千字

版　　次：2007 年 4 月第 1 版　　**印　次**：2007 年 4 月第1 次印刷

印　　数：1～4000

定　　价：49.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：(010)62770177 转 3103　　产品编号：020971-01

前　言

本书力求将实用统计方法的介绍与在计算机上 R 软件如何实现这些方法紧密地联系起来, 不仅介绍了各种数理统计方法的统计思想、实际背景、统计模型和计算方法, 并且结合 R 软件, 给出相应解决问题的步骤并对计算结果进行分析.

关于数理统计的教材或教科书已非常多. 这类教材主要以数理统计的理论为基础, 讲述其理论、方法与应用背景, 但对于计算, 讲的较少, 基本是以手工计算为主, 目的是为了帮助读者理解相应的统计方法, 可操作性不强.

关于统计计算的书也有不少. 目前, 统计计算的教材一般是讲算法 (这一点与数值分析或计算方法差不多), 而没有相应的软件做支撑, 有些内容是数值分析内容的重复, 统计味不足.

结合软件讲统计的书, 目前最多的是结合 SAS 软件、SPSS 软件. 这类书籍基本上相当于软件使用说明书, 虽然谈到一些统计概念, 但讲得很少.

本书既不是一本单纯的关于数理统计或统计计算的教科书, 也不是一本关于 R 软件的使用手册, 而是一本将两者相结合的教科书. 其特点是结合 R 软件讲述数理统计的基本概念与计算方法.

R 软件是一种共享的统计软件, 也是一种数学计算环境. 它提供了有弹性的、互动的环境来分析和处理数据; 它提供了若干统计程序包, 以及一些集成的统计工具和各种数学计算、统计计算的函数. 用户只需根据统计模型, 指定相应的数据库及相关的参数, 便可灵活机动地进行数据分析等工作, 甚至创造出符合需要的新的统计计算方法. 使用 R 软件可以简化数据分析过程, 从数据的存取, 到计算结果的分享, R 软件提供了非常方便的计算工具, 帮助用户更好地分析和解决问题. 通过 R 软件的许多内嵌统计函数, 可以很容易学习和掌握 R 软件的语法, 也可以编制自己的函数来扩展现有的 R 语言, 完成科研工作.

本教材的编写风格是: (1) 以目前常见的数理统计教材的内容为基准, 首先对数理统计的基本概念、基本方法作一个简单、清晰的介绍, 在注重基础的同时, 侧重统计思想和统计方法的介绍. (2) 以 R 语言为主, 编写相应的计算程序. 这部分内容的目的有两个: 第一是学习 R 软件的编程方法, 掌握 R 软件的基本技巧; 第二是通过编程加深对统计方法的了解与掌握, 同时, 还可以通过编程, 加深对 R 软件中相关函数的了解. (3) 介绍相关的计算函数. 针对许多统计方法, R 软件提供了大量的内嵌计算函数, 使用者只需输入数据并且调用相应的内嵌函数, 就可得到相应的结果. 本书这一部分的写作重点放在对计算结果的统计解释上, 即如何通过结果来分析已有数据中所包含的统计信息, 着重介绍相应的统计建模方法. 这些是本教材最主要的特色, 也是不同于其他与软件有关的教材之处. 本

书着重强调统计建模, 以及如何使用 R 软件得到其计算结果和相应的结果解释.

本书的主要内容: 第 1 章, 概率统计的基本知识. 主要是复习统计的基本知识, 便于对后面各章内容的理解. 第 2 章, R 软件的使用. 主要介绍 R 软件的基本使用方法. 第 3 章, 数据描述性分析. 从数据描述开始分析数据, 主要介绍数据的基本特征, 如均值、方差, 还有与数据有关的各种图形, 如直方图、散点图等. 第 4 章, 参数估计. 介绍参数估计的基本方法, 如点估计和区间估计. 着重介绍 R 软件中与估计有关的函数. 第 5 章, 假设检验. 介绍假设检验的基本方法, 一类是参数检验, 另一类是非参数检验. 重点是介绍 R 软件中与假设检验有关的 R 函数及相关的使用方法. 第 6 章, 回归分析. 介绍回归分析的基本方法, 着重介绍回归分析的过程与方法和如何使用 R 软件作回归分析. 除一般的回归方法外, 还谈到逐步回归、非线性回归等内容. 第 7 章, 方差分析. 介绍单因素方差分析、双因素方差分析, 以及正交试验设计与方差分析之间的关系. 第 8 章, 应用多元分析 (Ⅰ). 介绍判别分析和聚类分析, 这些内容与判别和分类有关. 第 9 章, 应用多元分析 (Ⅱ). 介绍主成分分析、主因子分析和典型相关分析, 它是应用多元分析中降维计算的内容. 第 10 章, 计算机模拟. 介绍计算机模拟的 Monte Carlo 方法, 以及系统模拟方法, 最后介绍模拟方法在排队论中的应用. 此外, 还包括两个附录, 内容分别是作者自编函数的索引和 R 软件中函数的索引.

在学习本书的内容之后, 可以发现, 尽管有些统计内容其计算相当复杂, 但在使用 R 软件之后, 这些问题可以很轻松地得到解决.

本书所编写的 R 函数, 以及所介绍的 R 函数均以 R-2.1.1 版为基础 (目前的版本是 R-2.3.1, 大约每 3 至 4 个月版本会更新一次), 而且全部程序均运行通过, 读者如果需要作者自编的 R 程序, 可以通过电子邮件向作者索取, 邮件地址: xueyi@bjut.edu.cn.

本书是为理工、经济、管理、生物等专业学生或专业人员解决统计计算问题编写的, 可以作为上述专业学生数理统计课程的辅导教材或教学参考书, 也可作为统计计算课程的教材和数学建模竞赛的辅导教材.

由于编者水平有限, 书中一定存在不足甚至错误之处, 欢迎读者不吝指正, 我们的电子邮件地址是: xueyi@bjut.edu.cn (薛毅); chenliping@bjut.edu.cn (陈立萍).

编 者

2006 年 7 月

于北京工业大学

目　录

第 1 章　概率统计的基本知识

本书是一本统计建模与软件应用相结合的教科书, 讲述重点是数理统计的基本方法和用 R 软件进行相应的计算. 众所周知, 数理统计是以概率论为基础, 应用非常广泛的数学学科分支, 是通过对试验或观察数据进行分析, 来研究随机现象以达到对研究对象的客观规律性做出合理的估计和推断的目的, 因此在介绍统计建模和 R 软件知识之前, 有必要先回顾一下相关的概率与数理统计的基本概念, 以及数理统计的各个应用分支.

本章用 4 节内容简单回顾概率论的基础知识, 用 1 节内容简单介绍数理统计的基本概念. 这样做的目的是使读者对已有概率论的知识有一个全面的了解与回顾, 对数理统计的概念有一个基本的认识.

1.1　随机事件与概率

1.1.1　随机事件

1. 随机事件

在一定条件下, 所得的结果不能预先完全确定, 而只能确定是多种可能结果中的一种, 称这种现象为随机现象. 例如, 抛掷一枚硬币, 其结果有可能是出现正面, 也有可能是出现反面; 电话交换台在 1 min 内接到的呼叫次数, 可能是 0 次、1 次、2 次、$\cdots\cdots$; 在同一工艺条件下生产出的灯泡, 其使用寿命有长有短; 测量同一物体的长度时, 由于仪器及观察受到环境的影响, 多次测量的结果往往有差异, 等等. 这些现象都是随机现象.

使随机现象得以实现和对它观察的全过程称为随机试验 (random experiment), 记为 E. 随机试验满足以下条件:

(1) 可以在相同条件下重复进行;

(2) 结果有多种可能性, 并且所有可能结果事先已知;

(3) 做一次试验究竟哪个结果出现, 事先不能确定.

称随机试验的所有可能结果组成的集合为样本空间 (sample space), 记为 Ω. 试验的每一个可能结果称为样本点 (sample point), 记为 ω.

称 Ω 中满足一定条件的子集为随机事件 (random event), 用大写字母 A, B, C, $\cdots$ 表示.

若一个随机事件只含一个不可再分的试验结果称为一个基本事件 (即一个样本点所组成的集合 $\{\omega\}$).

在试验中, 称一个事件发生是指构成该事件的一个样本点出现. 由于样本空间 Ω 包含了所有的样本点, 所以在每次试验中, 它总是发生, 因此称 Ω 为必然事件 (certain event). 空集 $\varnothing$ 不包含任何样本点, 且在每次试验中总不发生, 所以称 $\varnothing$ 为不可能事件 (impossible event).

2. 随机事件之间的关系

若事件 A 的发生必然导致事件 B 的发生, 则称事件 A 包含于事件 B, 或事件 B 包含事件 A, 记为 $A \subset B$, 亦称为事件的包含 (contain) 关系.

若 $A \subset B$, 且 $B \subset A$, 则称事件 A 与事件 B 等价 (equivalent), 记为 $A = B$.

若事件 A 与事件 B 至少有一个发生, 则称为事件的和 (union), 记为 $A \cup B$. 若 n 个事件 $A_1, A_2, \cdots, A_n$ 中至少有一个发生, 则称为 n 个事件的和, 记为 $A_1 \cup A_2 \cup \cdots \cup A_n$ 或 $\bigcup\limits_{i=1}^{n} A_i$.

同样, 可以定义可列无穷个事件的和 $A_1 \cup A_2 \cup \cdots \cup A_n \cup \cdots$ 或 $\bigcup\limits_{i=1}^{\infty} A_i$, 表示无穷个事件中至少有一个发生.

若事件 A 发生而事件 B 不发生, 则称为事件 A 与事件 B 的差, 记为 $A - B$.

若事件 A 与 B 同时发生, 则称事件 A 与事件 B 的积 (intersection), 记为 $A \cap B$ 或 AB. 若 n 个事件 $A_1, A_2, \cdots, A_n$ 同时发生, 则称为 n 个事件的积, 记为 $A_1 \cap A_2 \cap \cdots \cap A_n$ 或 $\bigcap\limits_{i=1}^{n} A_i$.

同样, 可以定义可列无穷个事件的积 $A_1 \cap A_2 \cap \cdots \cap A_n \cap \cdots$ 或 $\bigcap\limits_{i=1}^{\infty} A_i$, 表示无穷个事件同时发生.

若事件 A 与 B 不能同时发生, 则称事件 A 与事件 B 为互斥事件 (mutually exclusive event) 或不相容事件 (incompatiable event), 记为 $AB = \varnothing$.

在一次试验中, 基本事件之间是两两互斥的.

若 A 为随机事件, 称 "事件 A 不发生" 的事件为事件 A 的对立事件 (opposite event) 或逆事件 (complementary event), 记为 $\overline{A}$. 事件与对其立事件有如下关系:

$$A \cup \overline{A} = \Omega, \quad A\overline{A} = \varnothing.$$

由定义可知: 对立事件一定是互斥事件, 但互斥事件不一定是对立事件.

3. 随机事件的运算律

(1) 交换律

$$A \cup B = B \cup A, \quad AB = BA. \tag{1.1}$$

(2) 结合律

$$(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cap (B \cap C). \tag{1.2}$$

(3) 分配律

$$(A\cup B)C=(AC)\cup(BC),\quad A\cup(BC)=(A\cup B)(A\cup C). \tag{1.3}$$

(4) De Morgan(德·摩根) 律

$$\overline{A_1\cup A_2}=\overline{A}_1\cap\overline{A}_2,\quad \overline{A_1\cap A_2}=\overline{A}_1\cup\overline{A}_2. \tag{1.4}$$

对于 n 个或可列无穷个事件有

$$\overline{\bigcup_{k=1}^{n}A_k}=\bigcap_{k=1}^{n}\overline{A}_k,\quad \overline{\bigcap_{k=1}^{n}A_k}=\bigcup_{k=1}^{n}\overline{A}_k,\quad \overline{\bigcup_{k=1}^{\infty}A_k}=\bigcap_{k=1}^{\infty}\overline{A}_k,\quad \overline{\bigcap_{k=1}^{\infty}A_k}=\bigcup_{k=1}^{\infty}\overline{A}_k. \tag{1.5}$$

(5) 减法满足

$$A-B=A\overline{B}\quad 或\quad A-B=A\cap\overline{B}. \tag{1.6}$$

1.1.2 概率

1. 概率的公理化定义

在概率论中并非样本空间 Ω 的任何子集均可以看作事件, 所定义的事件之间应满足一定的代数结构.

定义 1.1 设随机试验 E 的样本空间为 Ω, $\mathcal{F}$ 是 Ω 的子集组成的集族, 满足

(1) $\Omega\in\mathcal{F}$;

(2) 若 $A\in\mathcal{F}$, 则 $\overline{A}\in\mathcal{F}$;(对逆运算封闭)

(3) 若 $A_i\in\mathcal{F}$, $i=1,2,\cdots$, 则 $\bigcup\limits_{i=1}^{\infty}A_i\in\mathcal{F}$. (对可列并运算封闭)

则称 $\mathcal{F}$ 为 Ω 的一个 σ 代数 (事件体), $\mathcal{F}$ 中的集合称为事件. 样本空间 Ω 和 σ 代数的二元体 $(\Omega,\mathcal{F})$ 称为可测空间.

定义 1.2 随机试验 E 的样本空间为 Ω, $(\Omega,\mathcal{F})$ 是可测空间, 对于每个事件 $A\in\mathcal{F}$, 定义一个实数 $P(A)$ 与之对应, 若函数 $P(\cdot)$ 满足条件:

(1) 对每个事件 A, 均有 $0\leqslant P(A)\leqslant 1$;

(2) $P(\Omega)=1$;

(3) 若事件 A_1, A_2, $\cdots$ 两两互斥, 即对于 $i,j=1,2,\cdots$, $i\neq j$, $A_iA_j=\varnothing$ 均有

$$P(A_1\cup A_2\cup\cdots)=P(A_1)+P(A_2)+\cdots.$$

则称 $P(A)$ 为事件 A 的概率 (probability), 称 $(\Omega,\mathcal{F},P)$ 为概率空间.

2. 概率的性质

性质 1 $P(\varnothing)=0$, 即不可能事件的概率为零.

但性质 1 反过来不成立, 即 $P(A)=0\not\Rightarrow A=\varnothing$.

性质 2 若事件 $A_1, A_2, \cdots, A_n$ 两两互斥, 则有

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n), \tag{1.7}$$

即互斥事件和的概率等于它们各自概率的和.

性质 3 对任一事件 A, 均有 $P(\overline{A}) = 1 - P(A)$.

性质 4 对两个事件 A 和 B, 若 $A \subset B$, 则有

$$P(B - A) = P(B) - P(A), \quad P(B) \geqslant P(A). \tag{1.8}$$

性质 5(加法公式) 对任意两个事件 A 和 B, 有

$$P(A \cup B) = P(A) + P(B) - P(AB). \tag{1.9}$$

性质 5 可以推广为

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) = & P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_1A_3) \\ & - P(A_2A_3) + P(A_1A_2A_3), \end{aligned} \tag{1.10}$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = S_1 - S_2 + S_3 - S_4 + \cdots + (-1)^{n-1} S_n, \tag{1.11}$$

其中 $S_1 = \sum\limits_{i=1}^{n} P(A_i)$, $S_2 = \sum\limits_{1 \leqslant i < j \leqslant n} P(A_iA_j)$, $S_3 = \sum\limits_{1 \leqslant i < j < k \leqslant n} P(A_iA_jA_k)$, $\cdots$, $S_n = P(A_1A_2 \cdots A_n)$.

1.1.3 古典概型

设随机事件 E 的样本空间中只有有限个样本点, 即 $\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$, 其中 n 为样本点总数. 每个样本点 $\omega_i (i = 1, 2, \cdots, n)$ 出现是等可能的, 并且每次试验有且仅有一个样本点发生, 则称这类现象为古典概型 (classical probability model). 若事件 A 包含 m 个样本点, 则事件 A 的概率定义为

$$P(A) = \frac{m}{n} = \frac{\text{事件 } A \text{ 包含的基本事件数}}{\text{基本事件总数}}. \tag{1.12}$$

例 1.1 设有 k 个不同的 (可分辨) 球, 每个球都能以同样的概率 $1/l$ 落到 l 个格子 ($l \geqslant k$) 的每一个中, 且每个格子可容纳任意多个球, 试分别求如下两事件 A 与 B 的概率.

A: 指定的 k 个格子中各有一个球;

B: 存在 k 个格子, 其中各有一个球.

解 由于每个球可以落入 l 个格子中的任一个, 并且每一个格子中可落入任意多个球, 所以 k 个球落入 l 个格子中的分布情况相当于从 l 个格子中选取 k 个的可重复排列, 故样本空间共有 l^k 种等可能的基本结果.