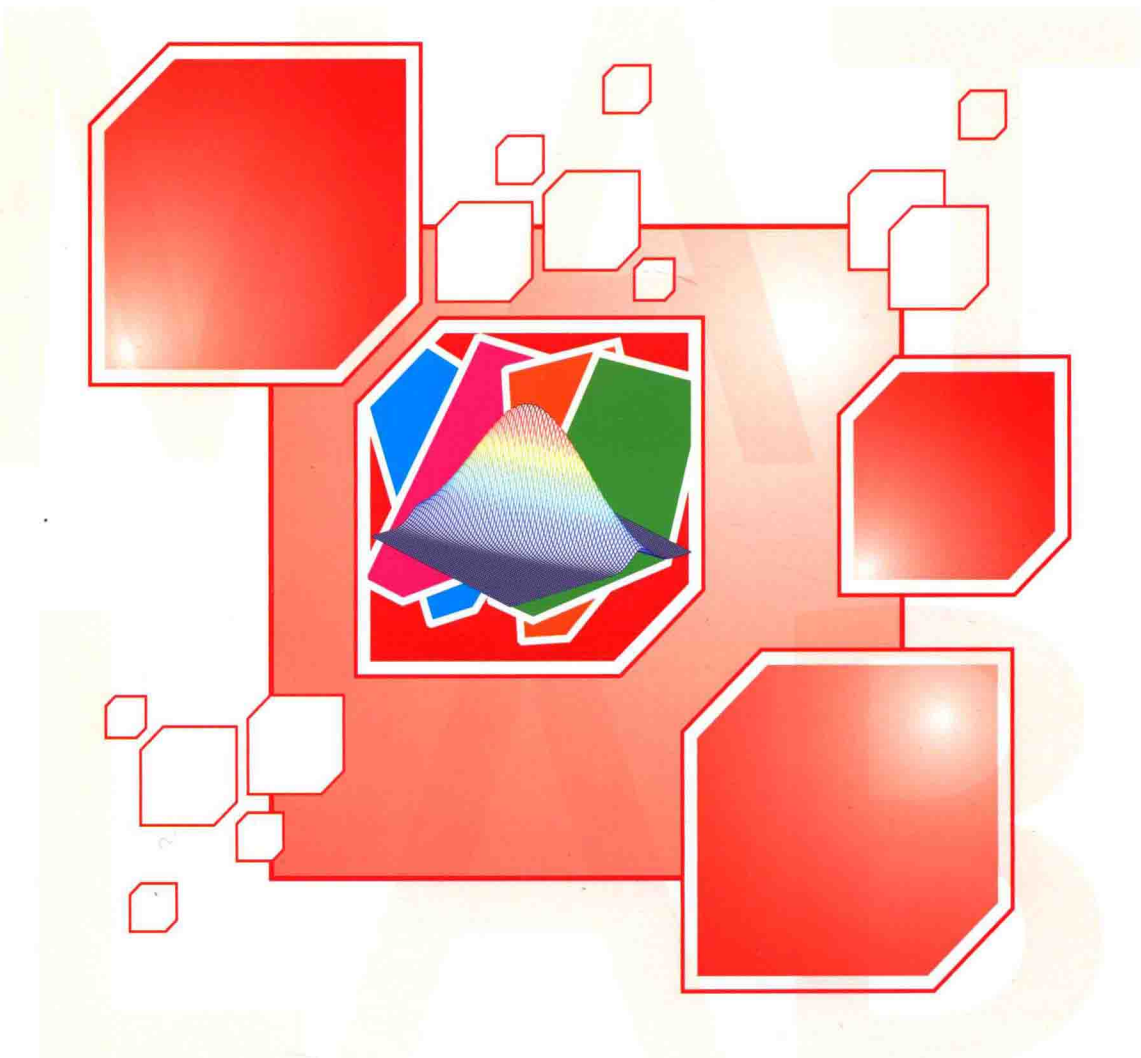
实用数据分析与MATLAB软件

薛 毅 陈立萍◎编著



北京工业大学出版社

实用数据分析与MATLAB软件

薛　毅　陈立萍　编著

北京工业大学出版社

内 容 简 介

本书以数据处理的过程为基准，介绍相关的基本概念、基本方法，在注重统计思想的同时，侧重于将统计方法与计算机软件相结合。淡化计算、注重方法是本书的宗旨。因此，书中数据处理的计算工作将由 MATLAB 软件来完成。本书着重介绍了 MATLAB 统计工具箱中与数据分析相关的函数。有了这些函数，读者就可以轻松地完成过去认为不可能完成的工作，体会到使用 MATLAB 软件的乐趣。

本书是为理工、经济、管理、生物等专业学生或专业人员编写的，目的是为了解决数据分析中可能遇到的问题。本书可作为上述专业学生相关课程的辅导教材或教学参考书，也可作为信息与计算科学专业学生数据分析或统计计算课程的教材和概率论、数理统计、多元分析、回归分析与实验设计等课程的辅助教材，还可作为数学建模竞赛的辅导教材，以及作为科技工作者和工程技术人员学习和参考用书。

图书在版编目(CIP)数据

实用数据分析与 MATLAB 软件/薛毅，陈立萍编著. —北京：北京工业大学出版社，2015.7

ISBN 978-7-5639-4375-3

Ⅰ.①实… Ⅱ.①薛… ②陈… Ⅲ.①Matlab 软件-应用-统计分析 Ⅳ.①C819

中国版本图书馆 CIP 数据核字(2015)第 144064 号

实用数据分析与 MATLAB 软件

编　　著：薛　毅　陈立萍
责任编辑：李周辉　贺　帆
封面设计：何　强
出版发行：北京工业大学出版社
（北京市朝阳区平乐园 100 号　邮编：100124）
010-67391722(传真) bgdcbs@sina.com
出 版 人：郝　勇
经销单位：全国各地新华书店
承印单位：徐水宏远印刷有限公司
开　　本：787 毫米×1092 毫米　1/16
印　　张：32.5
字　　数：687 千字
版　　次：2015 年 8 月第 1 版
印　　次：2015 年 8 月第 1 次印刷
标准书号：ISBN 978-7-5639-4375-3
定　　价：58.00 元

前　言

当今的时代是信息爆炸的时代，数据无处不在，从数据中得到有用的信息是分析和解决问题的重要环节。因此，如何从大量的、看似杂乱无章的数据中揭示隐藏在内部的规律、挖掘出有用的信息，以指导人们进行科学的推断和决策，这就需要用到数据分析。

顾名思义，数据分析就是分析和处理数据的方法，从这个层面上讲，数据分析不存在固定的模式与解决方法，一切与分析和处理数据有关的方法都包含在其中。但由于数据带有一定的随机性，因此狭义地讲，数据分析就是用数理统计的方法来分析和处理数据。本书正是在这个层面上开展工作的。

本书之所以命名为《实用数据分析与 MATLAB 软件》，其目的是使用概率或数理统计的方法来处理数据，借用 MATLAB 软件强大的计算功能，完成各种统计方法的计算，从而将研究工作从繁杂的计算中解放出来，将研究问题的重点放在对数据分析和处理的方法上。

MATLAB 是美国 MathWork 公司于 1984 年开发的主要用于工程计算的计算机高级语言，是当今国际上优秀的数学软件之一。MATLAB 以矩阵形式处理数据，具有强大的科学计算、图形处理、可视化、开放式和可扩展的功能，MATLAB 及其附带的几十种面对不同领域的工具箱(ToolBox)，能够广泛应用于数值分析、信号与图像处理、控制系统设计、通信仿真、工程优化、数学建模和统计分析等各个领域。

从教学层面来讲，概率论、数理统计、多元分析等统计课程，都有大量实例可供学生完成，但复杂的计算过程，会削弱学生的学习兴趣。使用 MATLAB 之后，计算工作交给计算机完成，学生学习的重点可放在问题的分析与理解方面。本书着重介绍了 MATLAB 数据统计工具箱中与数据分析相关的函数。有了这些函数，读者就可以轻松地完成过去认为不可能完成的工作，体会到学习的乐趣，通过对实例的计算，加深对所学理论知识的理解。

本书的主要内容共分 8 章来介绍。第 1 章，概率论与数理统计基础，简单回顾一下相关的概率论与数理统计的基本概念。第 2 章，数据描述性分析，介绍数据位置和分散程度的度量，数据的分布，以及数据的图形表示。第 3 章，参数估计，介绍分布参数的估计方法，第 4 章，假设检验，介绍常用的参数检验和非参数检验方法。第 5 章，回归分析。介绍回归分析和回归诊断的内容。第 6 章，方差分析，介绍单、双和

多因素分析方法。第7章，应用多元分析(I)，介绍与分类有关的判别分析和聚类分析。第8章，应用多元分析(II)，介绍主成分分析、因子分析、典型相关分析和非负矩阵分解的知识。本书的基本结构是：在介绍统计知识之后，重点介绍如何使用MATLAB数据统计工具箱中的函数完成统计方法的计算，同时介绍如何利用MATLAB提供的函数，编写一些有效地便于求解问题的程序。为帮助读者对文中MATLAB函数或程序的理解，本书给出了两个附录：附录A，MATLAB软件简介；附录B，MATLAB数理统计工具箱。

本书既不是纯粹的概率统计教材，也不是MATLAB软件的使用说明书，而是以数据处理的过程为基准，介绍相关的基本概念、基本方法，在注重统计思想的同时，侧重于将统计方法与计算机软件相结合，淡化计算、注重方法是本书的宗旨。因此，数据处理的计算工作将由MATLAB软件来完成。具体地说，绝大多数的计算都由MATLAB数据统计工具箱中的函数完成。在学习这些函数之后，可以达到以下三个目的：第一，学会使用MATLAB的相关函数求解问题，在科研与实际工作中会选择合适的函数进行计算。第二，读懂MATLAB的计算结果，会运用计算结果来解释所要处理的问题。第三，将MATLAB的函数作为基础函数，编写自己需要的函数程序，能够“站在巨人的肩膀上”工作。

本书的全部程序均通过计算检验，书中的程序已在MATLAB 7.9.0(R2009b)环境下运行通过。读者所持的软件的版本可能与编者的不一致，这基本不会影响到书中的程序的运行，但书中的某些函数不能在较低版本环境下运行，请读者见谅。

本书是为理工、经济、管理、生物等专业学生或专业人员编写的，目的是为了解决数据分析中可能遇到的问题。本书可作为上述专业学生相关课程的辅导教材或教学参考书，也可作为信息与计算科学专业学生数据分析或统计计算课程的教材和概率论、数理统计、多元分析、回归分析与实验设计等课程的辅助教材，还可作为数学建模竞赛的辅导教材，以及作为科技工作者和工程技术人员学习和参考用书。

由于受编者水平限制，可能在内容的取舍、结构的编排及课程的讲法上存在着不妥之处，希望使用本书的教师、学生和同行专家及其他读者提出宝贵的批评和建议。

在本书出版之际，谨向对本书提供过帮助的各位教师和专家表示感谢，感谢北京工业大学研究生院对数学建模课程的支持，同时感谢北京工业大学出版社为本书的出版做的大量工作。

目　录

第 1 章　概率论与数理统计基础

数据分析就是分析和处理数据的理论与方法。就广义而言，目前还没有固定的数据分析方法；但就狭义而言，数据分析需要用数理统计、多元分析的方法进行分析和处理。

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，在介绍用统计方法分析数据之前，有必要先回顾一下相关的概率论与数理统计的基本概念，以及数理统计的各个应用分支。

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 分钟内接到的呼叫次数，可能是 0 次、1 次、2 次……；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，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。

随机试验的所有可能结果组成的集合称为样本空间，记为 Ω。试验的每一个可能结果称为样本点，记为 ω。

称 Ω 中满足一定条件的子集为随机事件，用大写字母 A，B，C，…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。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。

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$A\subset B$，亦称为事件的包含关系。

若$A\subset B$且$B\subset A$，则称事件A与事件B等价，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，记为$A\cup B$。若n个事件A_1，A_2，…，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$\bigcup_{i=1}^{n}A_i$。

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$\bigcup_{i=1}^{\infty}A_i$，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，记为$A\cap B$或AB。若n个事件A_1，A_2，…，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$\bigcap_{i=1}^{n}A_i$。

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$\bigcap_{i=1}^{\infty}A_i$，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为互斥事件或互不相容事件，记为$AB=\varnothing$。

在一次试验中，基本事件之间是两两互斥的。

若事件A与事件B必有一个发生，且仅有一个发生，则称事件A与事件B互为对立事件或互逆事件，记为$\overline{A}$。事件A与事件B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)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

$$\overline{\cup_{k=1}^{n} A_k} = \cap_{k=1}^{n} \overline{A}_k, \quad \overline{\cap_{k=1}^{n} A_k} = \cup_{k=1}^{n} \overline{A}_k,$$
$$\overline{\cup_{k=1}^{\infty} A_k} = \cap_{k=1}^{\infty} \overline{A}_k, \quad \overline{\cap_{k=1}^{\infty} A_k} = \cup_{k=1}^{\infty} \overline{A}_k。 \tag{1.5}$$

(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关系。

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

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

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

(3)若 $A_i \in \mathscr{F}$，$i=1, 2, \cdots$，则 $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$。（对可列并运算封闭）

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

定义 1.2　设随机试验 E 的样本空间为 Ω，(Ω, $\mathscr{F}$)是可测空间，对于每个事件 $A \in \mathscr{F}$，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，…两两互斥，即对于 i，$j=1, 2, \cdots$，$i \neq j$，$A_i A_j = \varnothing$ 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，称(Ω, $\mathscr{F}$, 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。但性质反过来不成立，即 $P(A) = 0 \nRightarrow A = \varnothing$。

性质 2　若事件 A_1，A_2，…，A_n 两两互斥，则有

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

即互斥事件之和的概率等于它们各自的概率之和。

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

性质 4 对两个事件 A 和 B，若 $A \subset B$，则有

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

性质 5 对任意两个事件 A 和 B，

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

性质 5 可以推广为

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

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

式中，$S_1 = \sum_{i=1}^{n} P(A_i)$，$S_2 = \sum_{1 \leqslant i < j \leqslant n} P(A_iA_j)$，$S_3 = \sum_{1 \leqslant i < j < k \leqslant n} P(A_iA_jA_k)$，$\cdots$，$S_n = P(A_1A_2 \cdots A_n)$。

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。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，试分别求以下两事件 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。

事件 A 所含基本结果数应是 k 个球在指定的 k 个格子中的全排列数，即 $k!$，所以

$$P(A) = \frac{k!}{l^k}。$$

为了算出事件 B 所含的基本事件数，可设想分两步进行：因为 k 个格子可以是任意选取的，故可先从 l 个格子中任意选出 k 个来，选法共有 $\binom{l}{k}$ 种；对于每种选定的 k 个格子，依上述各有一个球的推理，则有 $k!$ 个基本结果，故 B 含有 $\binom{l}{k} k!$ 个基本结果。所以

$$P(B) = \binom{l}{k} \frac{k!}{l^k} = \frac{l!}{(l-k)!\, l^k}。$$

概率论的历史上有一个颇为著名的问题——生日问题：求 k 个同班同学没有两人生日相同的概率。

若把这 k 个同学看作例 1.1 中的 k 个球，而把一年 365 天看作格子，即 $l=365$，则上述的 $P(B)$ 就是所要求的概率。例如，$k=40$ 时，$P(B)=0.109$。或者换句话说，40 个同学中至少两个人同一天过生日的概率是：$P(\overline{B})=1-0.109=0.891$，其概率大得出乎意料。

1.1.4　几何概型

当随机试验的样本空间是某一可度量(长度、面积或体积)的区域，并且任意一点落在度量相同的子区域内是等可能的，则事件 A 的概率定义为

$$P(A)=\frac{S_A}{S}=\frac{\text{构成事件 A 的子区域的度量}}{\text{样本空间的度量}}。\tag{1.13}$$

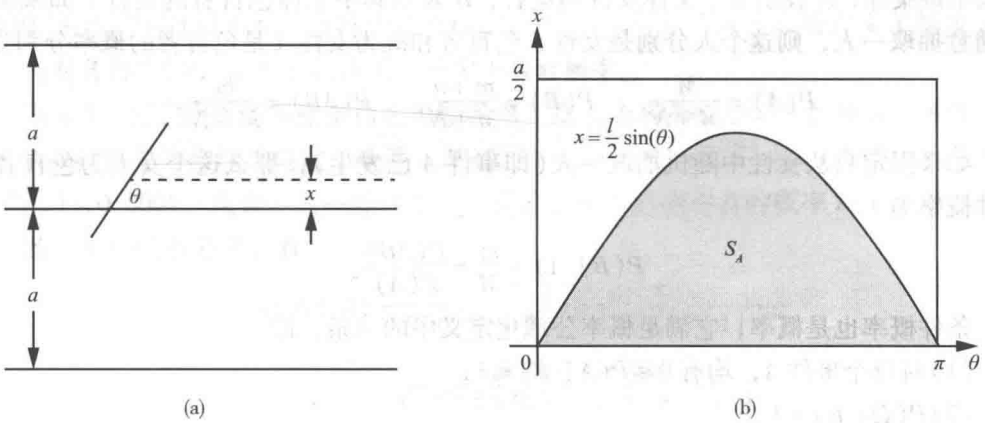
这种概率模型称为几何概型。

例 1.2　蒲丰投针问题。设平面上画有等距为 a 的一簇平行线。取一枚长为 $l(l<a)$ 的针随意扔到平面上，求针与平行线相交的概率。

解　设 x 表示针的中心到最近一条平行线的距离，θ 表示针与此直线间的交角，如图 1.1(a)所示，则(θ, x)完全决定针所落的位置。针的所有可能的位置为

$$\Omega=\left\{(\theta, x): 0\leqslant\theta\leqslant\pi, 0\leqslant x\leqslant\frac{a}{2}\right\},$$

它可用 $\theta-x$ 坐标平面上的一个矩形来表示，如图 1.1(b)所示。



(a)　　(b)

图 1.1　蒲丰投针的几何概率

针与平行线相交的充分必要条件是 $x\leqslant\frac{l}{2}\sin\theta$，即图 1.1(b)中阴影部分，它的面积为