数理统计与MATLAB工程数据分析

王　岩　隋思涟　王爱青　编著

清华大学出版社

ISBN 7-302-13999-7

ISBN 7-302-13999-7

定价：35.00元（含光盘）

数理统计与MATLAB工程数据分析

王 岩 隋思涟 王爱青 编著

清华大学出版社
北京

内容简介

本书介绍了数理统计的基本原理、典型应用，以及使用 MATLAB 进行实际工程数据分析的基本方法. 本书将统计分析方法与快速实现工程数据处理的应用软件工具融为一体，既从理论层面上介绍了假设检验、方差分析、回归分析、正交实验设计、判别分析等常用统计分析方法的基本原理和应用，同时又在配书盘中给出了快速实现工程数据处理的 MATLAB 应用程序库，包括书中所有例题的 MATLAB 实现程序、可执行文件等. 读者可以直接使用这些程序进行数据计算，还可以通过研究和学习相应的源程序代码来学会使用 MATLAB.

本书着重基础、强化应用、便于自学，可作为工科硕士研究生应用数理统计课程的基础教材、本科生相关专业的专业基础教材或实验教材，也可作为科研人员、工程技术人员的工具书或理论参考书.

图书在版编目(CIP)数据

数理统计与 MATLAB 工程数据分析/王岩，隋思涟，王爱青编著. —北京：清华大学出版社，2006.10
ISBN 7-302-13999-7

Ⅰ. 数… Ⅱ. ①王… ②隋… ③王… Ⅲ. ①数理统计 ②计算机辅助计算－软件包，MATLAB
Ⅳ. ①O212 ②TP391.75

中国版本图书馆 CIP 数据核字(2006)第 121809 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**客户服务**：010-62776969
组稿编辑：石　磊
文稿编辑：李　嫚
印 装 者：北京市清华园胶印厂
发 行 者：新华书店总店北京发行所
开　　本：185×230　**印张**：21.5　**字数**：453 千字
版　　次：2006 年 10 月第 1 版　2006 年 10 月第1 次印刷
书　　号：ISBN 7-302-13999-7/O·580
印　　数：1～3500
定　　价：35.00 元(含光盘)

前　言

随着计算机的发展与普及，数理统计已成为处理信息、进行决策的重要理论和方法. 在科学研究中，用数理统计方法从数据中获取信息和判别初步规律，往往成为重大科学发现的先导. 数理统计是数学方法与实际相结合，应用最为广泛、最为重要的方式之一. 因此，现代科研人员和工程技术人员都应该掌握数理统计的基础知识. MATLAB 是一套高性能的数值计算和可视化软件，它集矩阵运算、数值分析、信号处理和图形显示于一体，构成了一个界面友好、使用方便的用户环境，是实现数据分析与处理的有效工具.

本书介绍了数理统计的基本原理、典型应用，以及使用 MATLAB 进行实际工程数据处理与分析的基本方法. 全书共分 9 章，前 3 章为概率论与数理统计初步，简单概括了概率论与数理统计的基本内容和基本思想，属于概述内容. 第 4 章至第 8 章依次介绍了非参数假设检验、方差分析、回归分析、正交实验设计和判别分析的原理及其工程应用，以及使用作者开发的 MATLAB 应用程序进行实际数据处理的具体方法和步骤. 第 9 章 MATLAB 中的数理统计简单介绍了 MATLAB 中提供的主要数理统计函数及其使用方法，以供读者了解和学习 MATLAB.

本书是作者根据广大学生、科研人员、工程技术人员进行数据处理的需求而编写的，凝聚了作者近十年来从事工科研究生数理统计和本科生实验设计方法教学的经验，以及参与工程研究项目和指导数学建模竞赛过程中的体会；是作者进行工科数理统计课程教学改革的研究成果. 本教材具有以下特点：第一，注重数理统计的思想方法介绍. 在阐述某一统计概念方法时，一般是从具体实例开始引出相关内容的客观背景，让学生带着实际问题去学习和思考. 第二，注重应用性，数理统计是一门应用性很强的学科，其应用几乎遍及各个领域，成为解决实际问题的重要工具. 因此，本教材充实了许多应用性内容，以适应读者解决实际问题的需要. 第三，重视 MATLAB 应用于统计方法时的简单性、实用性和可操作性. 实际中，数据处理工作往往是庞大而繁琐的，使很多学生、科研人员、工程技术人员对此望而兴叹，感到无助. 本书对每一章节的例题都编制了 MATLAB 例题代码程序. 即便是对 MATLAB 知之甚少，或者对统计方法掌握得不够全面的读者，只要按照本书中的操作说明进行操作，就可得到计算与分析的结果. 操作方法简单、快捷. 对没有安装 MATLAB 的计算机，我们还提供了重要章节内容（方差分析、回归分析及正交实验设计）的可执行程序，读者可按照使用说明在任何 Windows 操作系统中进行计算. 因此，本书不仅为教师教学提供了方便，也为科研人员、工程技术人员从事科学研究和工程计算，以及研究生、大学生撰写论文时进行数据处理提供了可直接使用的平台. 本书的配书盘中包含

这些 MATLAB 例题代码程序、可执行文件及其使用说明. 读者可以直接使用这些程序进行数据计算，还可以通过研究和学习相应的源程序代码来学会使用 MATLAB.

本书不仅可作为工科硕士研究生应用数理统计课程的基础教材、本科生相关专业的专业基础教材或实验教材，也可作为科研人员、工程技术人员的工具书或参考读物.

本书在编写过程中，参考了大量的资料文献. 青岛理工大学吕平教授提供了相关的实验数据，王谦源教授、王军英教授和西安交通大学周家良教授给予了热情的关心与支持，在本书出版之际，谨向他们表示衷心的感谢.

由于作者水平有限，本教材虽然是在经多年使用和修改的讲稿基础上整理编写的，但书中一定还存在很多缺点和不足，恳请读者批评指正. 作者的联系方式是 wangyan1231@sohu.com.

作　者

2006 年 7 月

目　　录

第 1 章　概率论与数理统计基础

数理统计是研究随机现象规律性的一门学科. 它以概率论为基础，研究如何以有效的方式获得、整理和分析受到随机性影响的数据，并以这些数据为依据，建立有效的数学模型，去揭示所研究问题的统计规律性.

数理统计的理论和方法已广泛应用在自然科学、技术科学、社会科学和人文科学等各个领域. 随着计算机的发展和普及，数理统计已成为处理信息、进行决策的重要的理论和方法.

数理统计研究的内容概括起来可分为两大类：其一是研究如何对随机现象进行观察、实验，以便更合理和更有效地获取观察资料的方法，即实验的设计和研究；其二是研究如何对所获得的有限数据进行整理、加工，并对所讨论的问题做出尽可能可靠、精确的判断，这就是统计推断问题.

1.1　概率论基础

概率论是数理统计的基础，为此，我们先简要复习概率论的基本概念、性质与公式.

1.1.1　概率论基本概念

1. 随机事件与概率

自然界和人类社会中所发生的现象是多种多样的，但大致可分为两类：一类是确定性现象，即在一定条件下必然发生的现象. 比如在标准大气压下“水加热至 100℃时沸腾”等；另一类是随机事件，即在相同条件下重复进行某种实验，有多种可能的结果发生，而在实验或观察之前不能预知确切的结果. 例如：投掷一枚均匀硬币，结果可能出现“正面”，也可能出现“反面”，掷前无法确定哪个结果会出现；远射一个目标可能击中也可能击不中，射前无法确定哪个结果会出现；从一袋小麦种子中任取 10 粒做发芽实验，实验的结果可能是有 10，9，…，1 粒种子发芽或全部不发芽，而实验前无法知道有几粒小麦种子会发芽，等等. 这些都是随机现象. 随机现象有两个特点：(1)在一次观察中，现象可能发生也可能不发生，即结果呈现不确定性；(2)在重复观察中，其结果具有统计

规律性，例如，多次重复投掷硬币，出现“正”“反”面的次数大致相同. 概率论就是研究随机现象统计规律性的学科.

我们把具有以下几个特点的实验称为随机实验：(1)在相同的条件下可以重复进行；(2)实验的结果不止一个，所有结果事先明确知道；(3)进行一次实验前不能确定哪个结果会出现. 随机实验通常以字母 E 表示.

随机实验中，可能出现也可能不出现的事情称为随机事件，用 $A,B,C,\cdots$ 表示；每一个可能出现的结果，称为基本事件. 例如随机实验“掷一只骰子，观察出现的点数”中“点数大于 3”，“点数为偶数”等都是随机事件，而“点数为 1”、“点数为 2”、…、“点数为 6”，都是基本事件.

样本空间是概率论中的重要概念. 在随机实验中，每一个基本事件称为样本点；样本点的全体称为样本空间，即必然事件，记作 Ω. 由此，随机事件是样本空间的子集合.

在一个随机实验中，随机事件是否发生是很重要的，但更重要的是事件发生的可能性的大小，它是随机事件的客观属性，是可以度量的，于是，我们就把刻画随机事件 A 发生的可能性大小的量 p 叫做事件 A 的概率. 记作

$$P(A) = p.$$

在一个随机实验下，怎样确定事件 A 的概率呢？这就涉及频率的概念.

首先说明什么是频率. 设 A 为某一实验可能出现的随机事件，在同样的条件下，这种实验重复 n 次，在这 n 次实验中，事件 A 出现了 m 次($0\leqslant m\leqslant n$)，则称 m 为 A 在这 n 次实验中出现的频数. 称 m/n 为 A 在这 n 次实验中出现的频率. 例如，在食品抽样检查中，每次抽一件，共抽了 10 件，其中正品 7 件，那么“出现正品”这一事件 A 在这 10 次实验中的频数为 7，频率为 $7/10=0.7$. 进一步实验结果见表 1.1.1.

表 1.1.1 抽样实验结果

项　目	实验结果						
抽样件数 n	10	60	150	600	900	1200	1800
正品件数 m	7	53	131	548	820	1091	1631
正品频率	0.7	0.883	0.873	0.913	0.911	0.909	0.906

从表 1.1.1 可见，实验次数较少时，A 出现的频率差异会很大，但是随着实验次数的增多，事件 A 出现的频率虽然不是一个确定的数，但波动却减少，且稳定在 0.9 附近.

上例说明，频率 m/n 本身是不确定的，但随实验次数的增加，频率总是在某一常数附近摆动，而且 n 愈大，频率与这个常数的偏差往往愈小，这种性质称为频率的稳定性. 这个常数是客观存在的，与所做的若干次具体实验无关，它反映了事件本身所蕴含的规律性，反映了事件出现的可能性大小. 因此，这个常数就是事件 A 的概率，即事件 A 的概率就是事件 A 发生的频率的稳定值.

明确了概率与频率的关系后，就可以利用频率来估计概率了.

(1) 对任意事件 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 $P(\varnothing)=0$.

2. 事件的独立性

设 A,B 是两个随机事件，若 $P(AB)=P(A)P(B)$，则称事件 A,B 相互独立(简称独立). 设 $A_1,A_2,\cdots,A_n$ 是 n 个事件，若对任意的 $k(2 \leqslant k \leqslant n)$ 和任意一组 $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$ 都有

$$P(A_{i_1}A_{i_2}\cdots A_{i_k}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k})$$

成立，则称 n 个事件 $A_1,A_2,\cdots,A_n$ 相互独立.

1.1.2　随机变量及其分布

随机变量是概率论中另一个重要概念. 引进随机变量的概念后，可把对事件的研究转化为对随机变量的研究. 由于随机变量是以数量的形式来描述随机现象，因此它给理论研究和数学运算都带来极大方便.

设随机实验 E 的样本空间为 Ω，如果对于每一个样本点 e，都有一个实数 X 与之对应，则称 X 为随机变量.

随机变量分为离散型随机变量和连续型随机变量.

1) 离散型随机变量

若随机变量 X 的所有可能的取值只有有限多个或可列无限多个，则称为离散型随机变量. 离散型随机变量的取值规律称为分布律. 设离散型随机变量 X 所有可能的取值为 $x_k(k=1,2,3,\cdots)$，取这些值的概率为 $p_k(k=1,2,3,\cdots)$，称式

$$P(X = x_k) = p_k, \quad k = 1,2,\cdots$$

为随机变量 X 的分布列(或分布律或概率分布).

分布律也可用表 1.1.2 列出.

表 1.1.2　X 的分布律表

X	x_1	x_2	$\cdots$	x_k	$\cdots$
p	p_1	p_2	$\cdots$	p_k	$\cdots$

离散型随机变量的概率分布图如图 1.1.1. 图中横轴上点的横坐标表示随机变量所取的值，横坐标上各点对应的纵轴的平行线表示随机变量取该值的概率.