北京大学数学教学系列丛书

本科生
数学基础课教材

应用多元统计分析

高惠璇 编著

北京大学出版社

北京大学数学教学系列丛书

应用多元统计分析

高惠璇　编著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

应用多元统计分析/高惠璇编著. —北京：北京大学出版社，2005.1

（北京大学数学教学系列丛书）

ISBN 7-301-07858-7

Ⅰ.应… Ⅱ.高… Ⅲ.多元分析：统计分析-高等学校-教材 Ⅳ.O212.4

中国版本图书馆 CIP 数据核字(2004)第 124211 号

书　　名：应用多元统计分析

著作责任者：高惠璇　编著

责 任 编 辑：邱淑清

标 准 书 号：ISBN 7-301-07858-7/O・0613

出 版 发 行：北京大学出版社

地　　址：北京市海淀区中关村北京大学校内　100871

网　　址：http://cbs.pku.edu.cn　电子信箱：zpup@pup.pku.edu.cn

电　　话：邮购部 62752015　发行部 62750672　理科编辑部 62752021

印 刷 者：北京大学印刷厂

经 销 者：新华书店

890 mm×1240 mm　A5　13.625 印张　392 千字

2005 年 1 月第 1 版　　2005 年 1 月第 1 次印刷

印　　数：0001—4000 册

定　　价：21.00 元

《北京大学数学教学系列丛书》编委会

名誉主编：姜伯驹

主　　编：张继平

副 主 编：李　忠

编　　委：（按姓氏笔画为序）

王长平　刘张炬　陈大岳　何书元

张平文　郑志明

编委会秘书：方新贵

责任编辑：刘　勇

内容简介

本书为北京大学数学科学学院概率统计系“应用多元统计分析”课程使用多年的教材，它主要介绍一些实用的多元统计分析方法的理论及其应用，并列举了各方面的应用实例，同时还以国际上著名的统计分析软件SAS系统作为典型工具，通过实例介绍如何处理数据分析中的各种实际问题。

本书共分十一章。第一章为绪论；第二、第三章介绍多元统计分析的理论基础——多元正态分布及其参数的估计和检验问题；第四章重点介绍多因变量的多元线性回归的有关问题，包括模型、参数的估计及其性质、假设检验、变量筛选，以及双重筛选逐步回归问题；第五、第六章介绍分类问题（判别与聚类）；第七到第九章介绍降维的多变量方法（主成分分析、因子分析和对应分析方法）；第十章讨论两组相关变量的典型相关分析；第十一章介绍近年来发展的偏最小二乘回归分析方法；并且在每一章内都配有适量的习题。“附录”中介绍了本课程所需的矩阵代数的有关内容；书末附有“部分习题参考解答或提示”，这些都将更便于读者自学。

本书可作为综合大学、工科大学或高等师范学院数学系、应用数学系、经济学等相关专业的本科生或研究生教材或教学参考书；对于其他领域中从事应用统计的工作人员也是一本极好的学习参考书。

作者简介

高惠璇　北京大学数学科学学院教授。1965年毕业于北京大学数学力学系。长期从事概率论与数理统计的教学、科研工作，主要研究方向是统计计算、统计软件与应用多元统计方法，曾参加过国家教委《数学软件的研究与开发》项目和统计软件的开发及推广普及工作。

试读结束：需要全本请在线购买：　www.ertongbook.com

序　言

自 1995 年以来，在姜伯驹院士的主持下，北京大学数学科学学院根据国际数学发展的要求和北京大学数学教育的实际，创造性地贯彻教育部“加强基础，淡化专业，因材施教，分流培养”的办学方针，全面发挥我院学科门类齐全和师资力量雄厚的综合优势，在培养模式的转变、教学计划的修订、教学内容与方法的革新，以及教材建设等方面进行了全方位、大力度的改革，取得了显著的成效。2001 年，北京大学数学科学学院的这项改革成果荣获全国教学成果特等奖，在国内外产生很大反响。

在本科教育改革方面，我们按照加强基础、淡化专业的要求，对教学各主要环节进行了调整，使数学科学学院的全体学生在数学分析、高等代数、几何学、计算机等主干基础课程上，接受学时充分、强度足够的严格训练；在对学生分流培养阶段，我们在课程内容上坚决贯彻“少而精”的原则，大力压缩后续课程中多年逐步形成的过窄、过深和过繁的教学内容，为新的培养方向、实践性教学环节，以及为培养学生的创新能力所进行的基础科研训练争取到了必要的学时和空间。这样既使学生打下宽广、坚实的基础，又充分照顾到每个人的不同特长、爱好和发展取向。与上述改革相适应，积极而慎重地进行教学计划的修订，适当压缩常微、复变、偏微、实变、微分几何、抽象代数、泛函分析等后续课程的周学时。并增加了数学模型和计算机的相关课程，使学生有更大的选课余地。

在研究生教育中，在注重专题课程的同时，我们制定了 30 多门研究生普选基础课程（其中数学系 18 门），重点拓宽学生的专业基础和加强学生对数学整体发展及最新进展的了解。

教材建设是教学成果的一个重要体现。与修订的教学计划相

配合，我们进行了有组织的教材建设，计划自 1999 年起用 8 年的时间修订、编写和出版 40 余种教材，这就是将陆续呈现在大家面前的《北京大学数学教学系列丛书》。这套丛书凝聚了我们近十年在人才培养方面的思考，记录了我们教学实践的足迹，体现了我们教学改革的成果，反映了我们对新世纪人才培养的理念，代表了我们新时期的数学教学水平。

经过 20 世纪的空前发展，数学的基本理论更加深入和完善，而计算机技术的发展使得数学的应用更加直接和广泛，而且活跃于生产第一线，促进着技术和经济的发展，所有这些都正在改变着人们对数学的传统认识。同时也促使数学研究的方式发生巨大变化。作为整个科学技术基础的数学，正突破传统的范围而向人类一切知识领域渗透。作为一种文化，数学科学已成为推动人类文明进化、知识创新的重要因素，将更深刻地改变着客观现实的面貌和人们对世界的认识。数学素质已成为今天培养高层次创新人才的重要基础。数学的理论和应用的巨大发展必然引起数学教育的深刻变革。我们现在的改革还是初步的。教学改革无禁区，但要十分稳重和积极；人才培养无止境，既要遵循基本规律，更要不断创新。我们现在推出这套丛书，目的是向大家学习。让我们大家携起手来，为提高中国数学教育水平和建设世界一流数学强国而共同努力。

张 继 平

2002 年 5 月 18 日

于北京大学蓝旗营

前　言

多元统计分析是数理统计学30多年来迅速发展起来的一个分支。特别在计算机非常普及、各种统计分析软件不断推出的今天，多元统计分析方法已广泛地应用到社会科学和自然科学的许多领域中。北京大学概率统计系自1985年成立以来，一直开设“应用多元统计分析”课程。编者在近20年来教学和科研的基础上，编写了《应用多元统计分析》一书。本书的目的是介绍一些实用的多元统计分析方法的理论及其应用，并以国际上著名的标准统计分析软件SAS系统作为典型工具，通过实例介绍如何用统计软件处理数据分析中的各种实际问题。

本书共有十一章及附录。第一章“绪论”介绍多元统计分析研究的对象，应用领域及多元数据的图表示法；第二章介绍多元正态分布及其参数的估计和性质；第三章首先介绍三个重要分布，即威沙特(Wishart)分布、霍特林(Hotelling) T^2 分布、威尔克斯(Wilks)分布及它们的性质，然后讨论多元正态总体中参数的假设检验问题；第四章“回归分析”重点介绍多因变量的多元线性回归的有关问题，包括模型、参数的估计及其性质、假设检验、变量筛选，以及双重筛选逐步回归问题。从第五章至第十章介绍另一些常用的统计方法，如判别分析、聚类分析、主成分分析、因子分析、对应分析方法以及典型相关分析。第十一章介绍近年来发展的偏最小二乘回归分析方法。在“附录”中介绍了本课程所需的矩阵代数的有关内容。书末还给出书中部分习题参考解答或提示。

“应用多元统计分析”是一门应用性很强的课程。本书不仅介绍了各种常用的多元统计分析方法的统计背景和实际意义，说明该方法的统计思想、数学原理及解题步骤，还列举出各方面的应用实例。本书将多元统计方法的介绍与在计算机上实现这些方法的统计软件(SAS系统)结合起来，使读者不仅学到统计方法的理论知识，还知

道如何解决实际问题。书中全部实例都是用 SAS 系统完成分析计算，并且每一章都配有适量的习题，其中大部分习题都附有参考解答或提示，以便于读者自学。

本书是北京大学数学科学学院概率统计系为开设的限选专业课“应用多元统计分析”所编写的教材。国内目前虽有一些介绍多元统计方法的教材，因偏重的方面不相同，并不能很好地满足要求。国外这方面较好的教材目前虽已有中译本，但由于篇幅太大给学生增加了经济上的负担。为达到本课程所要求的目的，编者在已编写的讲义基础上，通过反复使用、多次修改后编写出版了此书。

本书的读者对象是理工科类、经济类，特别是统计学学科等各专业学习应用统计的本科生，以及其他各个领域中需要进行数据分析处理的实际工作者。本书适用于每周 3～4 学时、每学期约讲授 54～72 学时“应用多元统计分析”课程或相关课程的教材，其中有些内容可供任课教师酌情选用。

本书因篇幅关系，应用实例的 SAS 程序没有在正文中给出，正文中只列出主要计算结果。为方便读者学习与掌握本书内容，我们另准备了《应用多元统计分析》附盘(3 寸软盘)一张，其内容包括正文所有实例的 SAS 程序，各章所有练习题的原始数据及用编程方法解答的 SAS 程序，以供读者参考。需要此附盘的读者请从网站“ftp://162.105.69.120/gaohx”上下载附盘上的文件，或与北京大学数学科学学院(邮编：100871)作者联系。

高惠璇

2003 年 7 月于北京大学

目　录

试读结束：需要全本请在线购买：　www.ertongbook.com

第一章 绪 论

§1.1 引 言

多元统计分析(简称多元分析)是运用数理统计的方法来研究多变量(多指标)问题的理论和方法,它是一元统计学的推广.

在实际问题中,很多随机现象涉及到的变量不是一个,而经常是多个变量,并且这些变量间又存在一定的联系.我们常常需要处理多个变量的观测数据.例如考察学生的学习情况时,就需了解学生在几个主要科目的考试成绩.表 1.1 给出某年级随机抽取的 12 名学生 5 门主课期末考试的成绩.

表 1.1 12 名学生 5 门课程的考试成绩

序号	政治(X_1)	语文(X_2)	外语(X_3)	数学(X_4)	物理(X_5)
1	99	94	93	100	100
2	99	88	96	99	97
3	100	98	81	96	100
4	93	88	88	99	96
5	100	91	72	96	78
6	90	78	82	75	97
7	75	73	88	97	89
8	93	84	83	68	88
9	87	73	60	76	84
10	95	82	90	62	39
11	76	72	43	67	78
12	85	75	50	34	37

表 1.1 提供的数据,如果用一元统计方法,势必要对多门课程分别分析,每次分析处理一门课程的成绩.这样处理,由于忽视了课程之间可能存在的相关性,因此,一般说来,丢失信息太多,分析的结果

不能客观全面地反映某年级学生的学习情况. 本书将要讨论的多元统计方法,它同时对多门课程的成绩进行分析. 这样的分析对诸课程间的关系、相依性和相对重要性等都能提供有用的信息. 如果说一元统计分析是研究一个随机变量统计规律性的学科,那么多元统计分析则是研究多个随机变量之间相互依赖关系以及内在统计规律性的一门统计学科.

由于大量实际问题都涉及到多个变量,这些变量又是随机变量,如学生的学习成绩随着被抽取学生的不同,成绩也有变化(我们往往需要依据它们来推断全年级的学习情况). 所以要讨论多元随机变量的统计规律性. 多元统计分析就是讨论多元随机变量的理论和统计方法的总称. 其内容既包括一元统计学中某些方法的直接推广,也包括多元随机变量特有的一些问题. 多元统计分析是一类范围很广的理论和方法.

就以学生成绩为例,我们可以研究很多问题:用各科成绩的总和作为综合指标,来比较学生学习成绩的好坏;根据各科成绩相近程度对学生进行分类(如成绩好的与成绩差的,又如文科成绩好的与理科成绩好的);研究各科成绩之间的关系(如物理与数学成绩的关系,文科成绩与理科成绩的关系);等等. 所有这些都属于多元统计分析的研究内容.

综上所述,多元统计分析是以 p 个变量的 n 次观测数据所组成的数据矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

为依据的. 根据实际问题的需要,给出种种方法. 英国著名统计学家肯德尔(Kendall)在《多元分析》一书中把多元统计分析所研究的内容和方法概括为以下几个方面.

1. 简化数据结构(降维问题)

简化数据结构即是将某些较复杂的数据结构通过变量变换等方

法使相互依赖的变量变成互不相关的;或把高维空间的数据投影到低维空间,使问题得到简化而损失的信息又不太多的.例如主成分分析、因子分析,以及对应分析等多元统计方法就是这样的一类方法.

2. 分类与判别(归类问题)

归类问题即是对所考察的观测点(或变量)按相似程度进行分类(或归类).例如聚类分析和判别分析等方法就是解决这类问题的统计方法.

3. 变量间的相互联系

(1) 相互依赖关系:分析一个或几个变量的变化是否依赖于另一些变量的变化?如果是,建立变量间的定量关系式,并用于预测或控制——回归分析.

(2) 变量间的相互关系:分析两组变量间的相互关系——典型相关分析.

4. 多元数据的统计推断

这是关于参数估计和假设检验的问题.特别是多元正态分布的均值向量及协方差阵的估计和假设检验等问题.

5. 多元统计分析的理论基础

多元统计分析的理论基础包括多维随机向量及多维正态随机向量,以及由此定义的各种多元统计量,推导它们的分布并研究其性质,研究它们的抽样分布理论.这些不仅是统计估计和假设检验的基础,也是多元统计分析的理论基础.

多元统计分析起源于 20 世纪初,1928 年威沙特(Wishart)发表的论文《多元正态总体样本协方差阵的精确分布》,可以说是多元分析的开端.之后费希尔(Fisher)、霍特林(Hotelling)、罗伊(Roy)、许宝騄等人作了一系列奠基性的工作,使多元统计分析在理论上得到迅速的发展,在许多领域中也有了实际应用.由于用统计方法解决实际问题时需要的计算量很大,使其发展受到影响,甚至停滞了相当长的时间.20 世纪 50 年代中期,随着电子计算机的出现和发展,使得多元统计分析在地质、气象、医学、社会学等方面得到广泛的应用.60

年代通过应用和实践又完善和发展了理论，由于新理论、新方法的不断出现又促使它的应用范围更加扩大. 70 年代初期在我国才受到各个领域的极大关注，近 30 年来我国在多元统计分析的理论研究和应用上也取得了很多显著成绩，有些研究工作已达到国际水平，并已形成一支科技队伍，活跃在各条战线上.

§1.2　多元统计分析的应用

多元统计分析是解决实际问题的有效的数据处理方法. 随着电子计算机使用的日益普及，多元统计方法已广泛地应用于自然科学、社会科学的各个方面. 以下我们列举多元统计分析的一些应用领域.

一、教育学

n 个考生报考北京大学概率统计系. 每个考生参加 p 门课(语文、数学、政治、外语、物理、化学……)的考试，各门课的成绩记为 $y_{i1}, y_{i2}, \cdots, y_{ip}(i=1,2,\cdots,n)$. 又每个考生在高中学习期间，$m$ 门主要课程成绩为 $x_{i1}, x_{i2}, \cdots, x_{im}(i=1,2,\cdots,n)$. 经过对这些大量的资料作统计分析，我们能够得出：

(1) 高考成绩和高中学习期间成绩的关系，即给出两组变量线性组合间的关系，从而可由考生在高中学习期间的成绩来预测高考的综合成绩或某些科目的成绩.

(2) 给出考生成绩次序排队的最佳方案(最佳组合). 总分可以体现一个考生成绩好坏，但对报考概率统计系的学生，按总分从高到低的顺序录取并不是很合适的，如果按适当的权重加权求和，比如数学、物理、外语的权重相对高些，然后按加权和的顺序录取也许更合适些.

此外利用 n 个学生在高中学习期间 m 门主要课程的考试成绩，可对学生进行分类，如按文、理科成绩分类，按总成绩分类等. 若准备给优秀学生发奖，那么一等奖、二等奖的比例应该是多少？应用多元统计分析的方法可以给出公平合理地确定.

试读结束：需要全本请在线购买：　www.ertongbook.com