高等院校统计专业规划教材

多元统计分析

Multivariate Statistical Analysis

于秀林　任雪松　编著

中国统计出版社

高等院校统计专业规划教材

多元统计分析

于秀林　任雪松　编著

中国统计出版社

（京）新登字 041 号

图书在版编目（CIP）数据

多元统计分析/于秀林　任雪松编著．
—北京：中国统计出版社，1999.5
高等院校统计专业规划教材
ISBN 7-5037-2931-7

Ⅰ．多…
Ⅱ．于…
Ⅲ．多元分析：统计分析-高等学校-教材
Ⅳ．C812

中国版本图书馆 CIP 数据核字（1999）第 10217 号

责任编辑：徐　颖
封面设计：张建民
出版发行：中国统计出版社
通信地址：北京市三里河月坛南街 75 号　邮政编码：100826
办公地址：北京市丰台区西三环南路甲 6 号
电　　话：（010）63450984、63266600-22500（发行部）
印　　刷：科伦克三莱印务(北京)有限公司
经　　销：新华书店
开　　本：850×1168mm 1/32
字　　数：254 万字
印　　张：10
印　　数：1—5 000 册
版　　别：1999 年 8 月第 1 版
版　　次：1999 年 8 月第 1 次印刷
书　　号：ISBN 7-5037-2931-7/C・1611
定　　价：17.00 元

出版说明

“九五”期间是我国社会主义市场经济体制逐步完善和发展的重要时期。一方面，随着高等教育体制改革和统计改革的深入发展，对统计教育模式和统计人才培养目标都提出新的要求，另一方面，科学技术的飞速发展也促使统计技术发生了重大变革，新理论、新方法和新技术不断涌现并被应用于统计实践。为了适应这种新形势的需要，全国统计教材编审委员会制定了《1996—2000年全国统计教材建设规划》，根据《规划》的要求，编委会采取招标的方式组织全国有关院校的专家、学者编写了这批统计学专业“规划教材”。

这批教材力求以邓小平理论为指导，在总结“八五”期间规划统计教材建设经验的基础上，认真贯彻以下原则：①理论紧密联系实际的原则；②解放思想、转变观念、大胆探索、努力创新的原则；③正确处理继承与发展关系的原则。通过不懈努力，把这批教材建设成为质量高、适应性强、面向21世纪的新教材。

相信通过这批教材的出版、发行，对推动我国统计教育改革和加快更新、改造我国统计教材体系、教材内容的步伐将起到积极的促进作用，同时对我国统计教材建设也将起到较好的示范、导向作用。

限于水平和经验，这批教材的编审、出版工作还会有缺点和不足之处，诚恳欢迎教材的使用单位、广大教师和同学们提出批评和建议。

全国统计教材编审委员会

1999年3月

前　言

多元统计分析简称多元分析，是统计学的一个重要分支。随着电子计算机的普及和发展，了解和使用它的人迅速增加，它的作用也越来越大，几乎在国民经济许多领域中都有着广泛的应用，并已取得很多具有卓越成效的成果。实践证明，多元分析方法是处理多维数据不可缺少的重要工具，并日益显示出无比的魅力。

为了适应经济问题研究中定量分析的需要以及当前教学改革不断深入的需要，国内很多理工和财经院校相继给研究生和本科生开设了该课程。

作者在多年开设多元统计分析课程和科研工作的基础上，曾出版过多元统计分析的书，这次在原出版《多元统计分析及程序》一书的基础上，根据全国统计教材编委会专家评审组通过的《编写大纲》要求，对原书进行修改、充实；力争写出一本适合财经、统计、管理等专业用的教材，同时也想给对此方法感兴趣的科研人员、教师、经济方面的读者，提供一个较系统掌握这一方法的良好参考书。

本书特点：1. 概念清晰，方法明了，强调实际应用。2. 在一元统计分析的基础上深入浅出地介绍多元分析的内容，并着重介绍多元分析中常用的各种方法，讲清各种方法的实际背景和统计思想，同时每种方法都给出具体经济实例。3. 为了适合不同层次读者的需要和加深对各种方法的理解以及期望读者能灵活地运用这些方法，作者对多元分析中的一些理论也给出适当的论证和说明，但大多数理论，只是叙述结果，而有关理论证明，可查看本

书后面列出的参考书。4. 本书介绍的各种统计方法可使用国内外通用的SPSS和SAS软件去实现，不再附计算程序。5. 本书对主要章节给出附注，目的有两个：一是对本章节所介绍的内容进一步引伸。二是扩展，即进一步补充一下所介绍的内容，因此附注内容根据学生情况可选讲或不讲，对实际工作者可选读或不读，对掌握全书主要内容不受影响。

学习本书之前应具备以下三方面的知识：1. 由于向量和矩阵是研究多元数据的重要工具，所以要求读者具有一定的线性代数知识，本书附录中针对本书的需要复习了有关这方面的基本知识。2. 多元统计分析是建立在一元统计分析基础上的，因此要求读者具有初等数理统计知识。3. 多元统计分析是依赖于计算机的发展而发展的，如果不使用计算机，多元统计分析中许多计算几乎是不可能完成的，为了做到学以致用，要求读者会调用国内外通用的某一种统计软件包能上机操作即可，并不要求自编程序去实现各种方法的计算。

本书适用范围：可作为高等院校统计、财经、管理等专业的本科生教材，也可作为非数学专业的研究生和广大科技工作者的参考书，所需讲授学时约60左右。

本书在编写过程中得到全国统计教材编委会专家们的关心和帮助，特别是北京大学数学学院陈家鼎教授给以热心的指导。书稿写完后，又经过全国统计教材编委会召开的专家审稿会作了评审，他们对书稿提出许多很好的建议。最后由中国科学院系统科学研究所吴启光教授审稿。另外，本书的实例，一部分是作者科研课题，一部分是学生完成的作业，在此一并向他们表示衷心感谢。

书中全部例题，都经过任雪松老师上机核实验算，并做了补充和修改。

希望这本书的出版，为多元分析的普及和发展起到一定的促进作用，也为面向新世纪统计教材的改革，做些有意义的工作，使

这一有效的数学工具更好地为社会主义市场经济服务。

由于水平有限，书中难免有不足之处，欢迎读者批评指正。

作　　者

1999年3月于北京

目　录

第一章 绪 论

§1.1 什么是多元统计分析

在工业、农业、医学、气象、环境以及经济、管理等诸多领域中，常常需要同时观测多个指标。例如，要衡量一个地区的经济发展，需观测的指标有：总产值、利润、效益、劳动生产率、万元生产值能耗、固定资产、流动资金周转率、物价、信贷、税收等等；要了解一种岩石，需观测或化验的指标也很多，如：颜色、硬度、含碳量、含硫量等等；要了解一个国家经济发展的类型也需观测很多指标，如：人均国民收入，人均工农业产值、人均消费水平等等。在医学诊断中，要判断某人是有病还是无病，也需要做多项指标的体检，如：血压、心脏脉搏跳动的次数、白血球、体温等等。总之，在科研、生产和日常生活中，受多种指标共同作用和影响的现象是大量存在的，举不胜举。上述指标，在数学上通常称为变量，由于每次观测的指标值是不能预先确定的，因此每个指标可用随机变量来表示。

如何同时对多个随机变量的观测数据进行有效地分析和研究呢？一种做法是把多个随机变量分开分析，一次处理一个去分析研究；另一种做法是同时进行分析研究。显然前者做法有时是有效的，但一般来说，由于变量多，避免不了变量之间有相关性，如果分开处理不仅会丢失很多信息，往往也不容易取得好的研究结果。而后一种做法通常可以用多元统计分析方法来解决，通过对多个随机变量观测数据的分析，来研究变量之间的相互关系以及

揭示这些变量内在的变化规律，如果说一元统计分析是研究一个随机变量统计规律的学科，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。同时，利用多元分析中不同的方法还可以对研究对象进行分类（如指标分类或样品分类）和简化（如把相互依赖的变量变成独立的或降低复杂集合的维数等等）。在当前科技和经济迅速发展的今天，在国民经济许多领域中特别对社会经济现象的分析，只停留在定性分析上往往是不够的。为提高科学性、可靠性，通常需要定性与定量分析相结合。实践证明，多元分析是实现做定量分析的有效工具。

多元分析包括的主要内容：有多元正态总体的参数估计和假设检验以及常用的统计方法。这些方法是多元数据图表示法、聚类分析、判别分析、主成分分析、因子分析、对应分析、多重多元回归分析、典型相关分析、路径分析、多维标度法等。本书重点介绍多元分析中常用的各种方法。

多元分析起源于本世纪初，1928 年 Wishart 发表论文《多元正态总体样本协差阵的精确分布》，可以说是多元分析的开端。20 世纪 30 年代 R. A. Fisher、H. Hotelling、S. N. Roy、许宝騄等人作了一系列的奠基性工作，使多元分析在理论上得到了迅速的发展。40 年代在心理、教育、生物等方面有不少的应用，但由于计算量大，使其发展受到影响，甚至停滞了相当长的时间。50 年代中期，随着电子计算机的出现和发展，使多元分析方法在地质、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，由于新的理论、新的方法不断涌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，20 余年来我国在多元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 多元分析能解决哪些类型的实际问题

下面例举一些实际问题，从中不仅可以看到多元分析能解决哪些不同类型的问题，而且还可以看到多元分析应用的广度和深度，它将会引起学习者们的浓厚兴趣。

经济学：

1. 对我国30个省市自治区的社会情况进行分析，一般不是逐个省市自治区去分析，而较好地做法是选取能反映社会情况的代表性指标，如：人口密度、城市和农村的平均每人每月收入和支出情况、居住面积、城市绿化覆盖率等等，根据这些指标对30个省市自治区进行分类，然后根据分类结果对社会情况进行综合评价。又如要考察北京、天津等几所大城市的企业情况，首先要选取企业方面有代表性指标，如：企业个数、工业总产值、平均人数、固定资产净值、资金利税率、资金利润率、全员劳动生产率等等。由于要考察的指标多，通常先对指标进行分类，按分类结果对指标进行综合分析给出企业的评价。如何分类？可用Q型和R型聚类分析法。

2. 在经济学中，可根据人均国民收入、人均工农业产值、人均消费水平等多种指标判定一个国家的经济发展程度所属的类型。又如在市场预测中如何根据以往调查所得的种种指标判别下季度产品是畅销、平常或滞销，可用判别分析法。

3. 如何研究国民收入变量（工农业国民收入、运输业国民收入、建筑业国民收入等）与投资性变量（劳动者人数、货物周转量、生产建设投资等）之间的相关关系。如何研究全国所有制独立核算工业企业的经济效益指标与其资金、利税等主要财务指标之间的关系，可用典型相关分析法。

4. 对全国28个省市自治区经济效益作综合评价（未包括西藏、海南），显然要选取的指标很多。如固定资产投资完成额、年

末银行贷款余额、职工工资总额、工业全员劳动生产率、工业可比产品成本降低率、工业销售利税率、工业资金利税率、万元工业总产值能耗等等。如何将这些具有错综复杂关系的指标综合成几个较少的因子，既有利于对问题进行分析和解释，又能便于抓住主要矛盾做出科学的评价。可用主成分分析和因子分析法。

5. 如何考察某产品的质量指标（多个）与影响产品质量的因素（多个）之间的关系。在商品需求研究中，同时要考察某商品销售量与商品的价格、消费者的收入等等之间的相互关系，如何揭示它们之间的相互依赖关系，以及建立数学模型进行预测预报？可用多重多元回归分析法。

6. 某一产品是用两种不同原料生产的，试问此两种原料生产的产品寿命有无显著差异？又比如，若考察某商业行业今年和去年的经营状况，这时需要看这两年经营指标的平均水平是否有显著差异以及经营指标之间的波动是否有显著差异。可用多元正态总体均值向量和协差阵的假设检验。

在其它领域研究中也同样存在上述类似问题，为说明多元分析应用的广泛性，简单举例如下：

工业：

企业的经济效益是人力、财力、物力、信息、市场条件等等因素共同作用的结果，如何对企业经济效益作出评价？又如，某服装厂要生产一批新型服装，为了适应大多数顾客的需要如何确定服装的主要指标及分类的型号？

农业：

如何按照城乡居民消费水平，对我国30个省市自治区进行分类？如何根据全国各地区农民生活消费支出情况研究农民消费结构的趋势？

医学：

随机抽取200名患有抑郁症病人，按照测量到的指标，可以将他们分成几种类型？如何根据某病人的多种症状（体温、白血

球、恶心、呕吐、腹部压疼感等）判别此人患何种类型阑尾炎（急性、慢性、有无穿孔等）？

教育学：

如何对高考的考生成绩作因素分析？学生入学后的考试成绩和入学考试的各门课程成绩有何相关关系？

体育科学：

如何对运动员的多项心理、生理测试指标如简单反应、时间知觉、综合反应等作主要因素分析？如何研究体力测试指标（反复横向跳、立定体前屈、俯卧上体后仰等）与运动能力测试指标（耐力跑、跳远、投球等）之间相关关系？

生态学：

研究中国七星瓢虫在黄海、渤海的群聚与近期气象条件的关系。对 1 000 个类似的鱼类样本，如何根据测量的特征如体重、身长、鳍数、鳍长、头宽等，将这类鱼分成几个不同品种？

地质学：

在地质勘探中，如何根据岩石标本的多种特征来判别地层的地质年代，是有矿还是无矿，是铜矿还是铁矿等等？

社会学：

调查青年对婚姻家庭的态度如对文化和职业的要求、对经济收入的态度、对老人的责任、对相貌的重视等等作主要因素分析以便进行正确引导和思想教育。

考古学：

考古学家对挖掘出来的人头盖骨的高、宽等特征来判别是男或女，根据挖掘出的动物牙齿的有关测试指标，判别它是属于哪一类动物牙齿、是哪一个时代的。

环境保护：

研究多种污染气体（CO、CO_2、SO_2）的浓度与污染源的排放量和气象因子（风向、风速、温度、湿度）等之间的相互关系。

军事科学：

研究某飞机洞库可燃性气体变化的规律以及对气体浓度的预测。

文学：

我国古典小说的著名作品《红楼梦》一书的版权鉴定问题也用了多元统计分析方法，为使读者相信这一作法，并从中受到启发，这里不妨稍多做一点说明。众所周知，《红楼梦》一书共120回，一般认为前80回为曹雪芹所写，后40回为高鹗所续，长期以来对这个问题一直有争议。能否从数学上作出论证？1985、1986年复旦大学李贤平教授带领他的学生作了这项有意义的工作，他们创造性想法是将120回看成是120个样本，然后确定与情节无关的虚词作为变量（所以要抛开情节，是因为在一般情况下，同一情节大家描述的都差不多，但由于个人写作特点和习惯的不同，所用的虚词是不会一样的），让学生数出每一回里变量出现的次数，作为数据，用多元分析中的聚类分析法进行分类，果然将120回分成两类即前80回为一类，后40回为一类，很形象地证实了不是出自同一人的手笔。之后又进一步分析前80回是否为曹雪芹所写？这时又找了一本曹雪芹的其它著作，做了类似计算，结果证实了用词手法完全相同，断定为曹雪芹一人手笔，而后40回是否为高鹗写的呢？论证结果推翻了后40回是高鹗一个人所写。这个论证在红学界轰动很大，他们用多元统计分析方法支持了红学界观点，使红学界大为赞叹，之后他们还综合运用多元统计分析中其它方法作了一系列有意义的工作。

§1.3 主要内容安排

本书共分十二章。

第一章绪论，主要介绍多元分析研究对象及应用范围。第二章到第四章介绍多元分析的基本概念和基本理论。主要有四个重要的统计量分布即多元正态分布、Wishart分布、Hotelling T^2 分

布、Wilks 分布以及多元正态总体的参数估计和假设检验。其实，上述内容都是一元统计中相应内容的推广，因此这几章内容的介绍都是借助复习相应地一元统计内容而自然地引出新的知识，使读者不会感到抽象和困难。之后简要地介绍多元数据的图表示法。

第五章和第六章主要研究分类问题，介绍聚类分析法和判别分析法。实际应用时两种方法往往联合起来使用。因为判别分析要求对新样品进行判别分类之前，必先知道已有几类总体，然后建立判别式，对新样品进行判别归类。如果一批给出样品要划分几类事先不知道，这时可先做聚类分析然后再做判别分析。

第七章到第九章介绍主成分分析、因子分析和对应分析法。主要研究结构化简问题，将具有错综复杂关系的变量（或样品）综合成数量较少的因子尽可能简单地表示所研究的对象，又不致于损失很多有价值的信息。

第十章和第十一章研究两组变量之间的相关关系，介绍典型相关分析和多重多元回归法，前者用于简化两组变量为少数综合变量以再现原来两组变量之间的相关关系，后者侧重于建立数学表达式解决预测问题。

第十二章简介定性资料统计分析，对定性变量如：性别（男、女）、天气（阴、晴）、职业（工人、职员、教员等）如何进行统计分析，这里主要介绍列联表、对数线性模型和 Logistic 回归，本章不是详细介绍这方面的理论、方法和应用。而是初步反映一下这方面的内容。目的是展示进一步可学的知识，以便更好地解决实际问题。

本书除第二章给出习题之外，其余各章在统计方法介绍之后，都给出应用性课题的列举，供选作题参考，读者不妨就这些课题，收集有关数据，按每章所述方法去计算和分析，定有收获。

期望读者读完这本书能达到以下目的：

1. 清楚理解每种统计方法所要解决的问题、前题条件和局限