普通高等教育“十三五”规划教材

多元统计分析

DUOYUAN TONGJI FENXI

李亚杰　主编

北京邮电大学出版社
www.buptpress.com

普通高等教育“十三五”规划教材

多元统计分析

主编　李亚杰

北京邮电大学出版社
www.buptpress.com

内 容 简 介

本书主要介绍多元统计分析的理论和方法。全书共14章，内容包括绪论、多变量的可视化、多元分布的基本概念及数字特征、多元统计量及抽样分布、多元正态分布的参数估计和假设检验、相关性度量、主成分分析、因子分析、典型相关分析、对应分析、聚类分析、判别分析、定性数据的建模方法、多维标度分析。本书在介绍各种多元统计分析理论和方法时，由浅入深，注重理论联系实际，通过通俗易懂的案例进行从数据到结论的分析，以适合不同层次读者的需求。

本书可作为高等院校数学、计算机、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工作者的参考书。

图书在版编目(CIP)数据

多元统计分析 / 李亚杰主编. -- 北京：北京邮电大学出版社，2018.9

ISBN 978-7-5635-5399-0

Ⅰ. ①多… Ⅱ. ①李… Ⅲ. ①多元分析—统计分析—教材 Ⅳ. ①O212.4

中国版本图书馆 CIP 数据核字(2018)第 036971 号

书　　名：多元统计分析

著作责任者：李亚杰　主编

责 任 编 辑：徐振华　孙宏颖

出 版 发 行：北京邮电大学出版社

社　　址：北京市海淀区西土城路 10 号(邮编：100876)

发 行 部：电话：010-62282185　传真：010-62283578

E-mail：publish@bupt.edu.cn

经　　销：各地新华书店

印　　刷：北京玺诚印务有限公司

开　　本：787 mm×1 092 mm　1/16

印　　张：22.75

字　　数：596 千字

版　　次：2018 年 9 月第 1 版　2018 年 9 月第 1 次印刷

ISBN 978-7-5635-5399-0　　定价：54.00 元

前　言

作为教授多元统计分析课程的老师，笔者希望能够撰写一本书，帮助学生和普通公众去理解多元统计分析对生活带来的影响。多元统计分析主要对多维数据进行处理，包括简化多维数据和数据结构、对案例进行假设检验、分类和组合、综合评价、预测、控制等。多元统计分析的方法广泛应用于经济学、医学、教育学、心理学、社会学、考古学、环境科学、文学等领域，已经成为解决实际问题非常有效的方法。

统计理论研究为统计学的应用奠定了基础，理论研究和应用研究从总体上说是“源”和“流”的关系。如果理论不成熟，方法不完善，统计应用研究很难达到较高的水平，因此，笔者在编写本书时，注重了每个方法理论思想的介绍和理论知识的推导。虽然多元统计方法多种多样，但几乎所有方法都要求简化问题的复杂性，对那些重要的、后面仍会反复使用的理论知识，本书从多角度详细进行介绍。当然，本书也注重把理论知识与实际应用紧密结合，兼顾两类不同的读者：技术科技人员及一般性的对科学感兴趣的读者。

多元统计分析是利用统计学和数学方法，对多维复杂数据进行科学分析的理论和方法，主要内容包括多元正态总体的参数估计和假设检验，以及常用的多元统计方法。常用的多元统计方法有多元数据图表示法、多元回归分析、聚类分析、判别分析、主成分分析、因子分析、对应分析、典型相关分析、定性数据分析、多维标度分析等。由于很多院校把回归分析的内容单独授课，所以本书没有讨论多元回归分析这一方法。

本书的特色之一是章节的安排，经过教学实践发现“过山车”式的章节安排比较吸引学生：首先让学生总览课程，然后介绍有趣的多元图形让学生开拓思路，激发其学习兴趣；其次复习一元经典统计推断学的理论和方法，将之推广到多元，让学生体会到数学推理的逻辑性和继承性；最后进行案例方法的逐一介绍，在案例方法的章节安排上，我们考虑各个方法之间的联系和循序渐进。本书的特色之二是案例的软件操作，从统计学发展的历史可以看到，在统计数据的收集、整理、加工、分析过程中，起决定性作用的是高速计算工具——计算机，培养计算机统计软件的使用能力对统计教学很重要，我们在每个方法的理论介绍之后，都给出了应用案例的统计软件操作步骤，使学生学会每个方法后，可以自己动手进行类似的案例分析，让学生体会到从数据到结论的乐趣和成就感。本书的特色之三是案例的选择，本书挑选了经典案例或者通俗易懂的案例。例如，在 Fisher 判别分析中的鸢尾花案例是经典案例，能帮助我们了解一个方法最初产生时的实证分析情况；再如，在多元图形可视化章节中，我们对一个易懂的经济学例子进行多个方法的分析，展示各种不同的观点怎样导出各自的处理方法，其目的是帮助读者了解这些观点的特点，使其在讨论更复杂的问题时能把握基本的思想，不致被烦琐的推导和形式复杂的公式所迷惑。

本书可以作为高等院校多元统计分析课程的教材使用，建议采用授课与实际案例相结合的教学方式。学习本课程的先修课程有高等数学、线性代数、概率论和数理统计，这些先修课程可以帮助读者理解统计方法的原理。建议读者将各种方法与统计软件紧密结合，领悟各种方法的实际背景、基本思想、理论依据、应用场合，从而会用各种方法解决实际问题。

本书的出版要感谢北京邮电大学教务处的支持和各位数学系同人的帮助，感谢理学院数学系孙洪祥、闵祥伟、胡细宝、莫骄、刘吉佑、丁金扣、李鹤等老师提出的宝贵意见，感谢理学院数学系 2014212101 班的吕正东、邵亚男、孔祥钊、骆雪婷等同学在课堂上的积极反馈，感谢家人和朋友们的鼓励。

由于编者水平有限，书中难免有不足之处，请读者批评指正。

编　者

目　录

第1章　绪论:爱上多元统计学

"在终极的分析中,一切知识都是历史;在抽象的意义下,一切科学都是数学;在理性的基础上,所有的判断都是统计学。"

——C. R. Rao,《统计与真理——怎样运用偶然性》

很多现实问题需要同时研究多个指标。例如,经济学中研究企业划分,可以考虑指标:从业人数、销售额和资产总额等;医学上研究病情诊断,可以考虑指标:血压、脉搏、白细胞、体温等。上述案例都是对多个(多维)变量进行研究,寻找背后的规律,这就是多元统计分析要解决的问题。

1.1　什么是多元统计分析

一元统计分析是研究一个随机变量统计规律的学科,有其理论和现实的局限性。多元统计分析,顾名思义,是对多维随机变量进行分析和研究,研究它们之间的相互依赖关系以及内在统计规律性的统计学科。

如何同时对多个随机变量的观测数据进行有效的分析和研究?假如把多个随机变量分开分析,每个随机变量用一元统计分析方法研究,就不会清楚多个变量之间的相关性,会丢失信息,不易获得好的研究结果。科学的方法是对多个变量同时进行分析研究,采用多元统计分析方法,通过同时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。

法国著名数学家庞加来(J. H. Poincaré,1854—1912年)说过,"如果我们想预测数学的未来,那么正确的途径是研究其历史与现状"。史学研究是任何学科永恒的研究主题,多元统计学自然不能例外,统计学史上曾涌现多位杰出的多元统计学家。

首先涉足多元分析方法的是英国统计学家高尔顿(F. Galton),他于1889年把双变量的正态分布方法运用于传统的统计学,他于六年中测量了近万人的"身高、体重、阔度、呼吸力、拉力和压力、手击的速率、听力、视力、色觉及个人的其他资料",在探究这些数据内在联系的过程中提出了今天在自然科学和社会科学领域中广泛应用的"相关"思想,创立了线性回归,他的学生皮尔逊(K. Pearson)受其影响,给出积矩相关系数、复相关等研究多个变量之间关系的概念和方法。其后,斯皮尔曼(C. E. Spearman)提出对多维变量进行降维的因子分析法,费希尔(R. A. Fisher)提出方差分析和判别分析,美国的威尔克斯(S. S. Wilks)发展了多元方差分析,

美国的霍特林(H. Hotelling)确定了主成分分析和典型相关分析。到20世纪前半叶,多元分析理论基础基本确立,1928年英国的维希特(J. Wishart)发表论文《多元正态总体样本协方差阵的精确分布》,是学术界公认的多元统计分析理论研究的开端。R. A. Fisher、H. Hotelling、S. N. Roy、M. A. Girshick、许宝騄等人做了一系列奠基的工作,使多元统计分析在理论上得到迅速的发展,在许多领域中有了实际应用。21世纪初,人们获得的数据正以前所未有的速度急剧增加,产生了很多超大型数据库,遍及超级市场销售、银行存款、天文学、粒子物理、化学、医学以及政府统计等领域,多元统计与人工智能、数据库技术相结合,已在经济、商业、金融、天文等行业得到成功应用。

为了更清楚地了解多元统计分析史的发展脉络,我们给出图1.1.1,图的横轴表示时间。

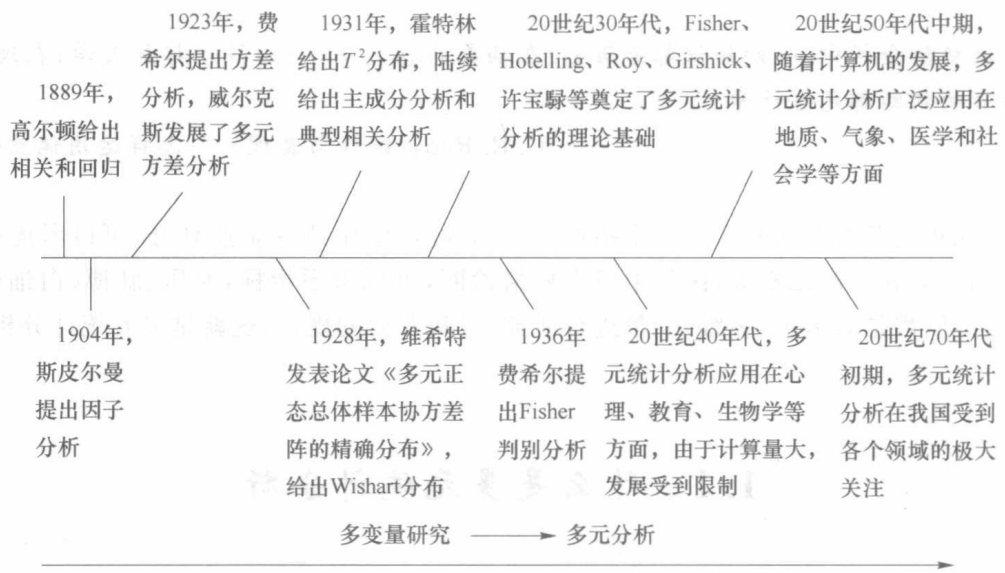

图1.1.1 多元统计分析史的发展脉络

1.2 多元统计分析的主要内容和方法

多元统计分析是应用数理统计学来研究多变量(多指标)问题的理论和方法,是统计学的一个重要分支。它是一元统计学的推广和发展,是一门具有很强应用性的课程,在自然科学和社会科学等领域中得到广泛的应用,包括了很多非常有用的数据处理方法。

英国著名统计学家肯德尔(M. G. Kendall)先后出版了《多元分析》《统计理论入门》《高等统计学理论》《等级相关方法》《时间序列》《几何概率》《统计学和概率史研究》等著作。

Kendall在*Multivariate Analysis*(1983年)一书中把多元分析所研究的内容和方法概括为以下几个方面。

1. 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。

2. 多元数据的统计推断

多元数据的统计推断主要研究多元正态分布的均值向量和协差阵的估计和假设检验等问题。

3. 简化数据结构

简化数据结构主要研究降维问题。例如，通过变量变换等方法使相互依赖的变量变成互不相关的变量；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。主成分分析、因子分析、对应分析等多元统计方法就是这样的一类方法。

4. 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——典型相关分析等。

③ 定性变量间的相互关系：对应分析等。

5. 分类与判别

对所考察的对象（样品点或变量）按相似程度进行分类（或归类）。聚类分析和判别分析等方法是解决这类问题的统计方法。

上述内容可由表 1.2.1 概括，要注意一种方法有时会解决多个问题。

表 1.2.1　多元统计分析的内容和方法

问　题	研究内容	可采用的方法名称
数据或结构性化简	尽可能简单地表示所研究的现象，但不损失很多有用的信息，并希望这种表示能够很容易地解释	多元回归分析、聚类分析、主成分分析、因子分析、对应分析、多维标度法、可视化分析
分类和组合	基于所测量的一些特征，给出好的分组方法，对相似的对象或变量进行分组	判别分析、聚类分析、主成分分析、可视化分析
变量之间的相关关系	变量之间是否存在相关关系，相关关系如何体现	多元回归、典型相关、主成分分析、因子分析、对应分析、多维标度法、可视化分析
预测与决策	通过统计模型或最优准则，对未来进行预见或判断	多元回归、判别分析、聚类分析、可视化分析
假设的提出及检验	检验由多元总体参数表示的某种统计假设，能够证实某种假设条件的合理性	多元总体参数估计、假设检验

例如，在考查大学生的学习情况时，需了解学生的几个主要课程的考试成绩。表 1.2.2 给出从某大学某学院随机抽取的 100 名学生中 6 门主要课程的期末考试成绩。

表 1.2.2　某大学学生的主要课程成绩

序　号	概率与统计	高等数学	大学英语	通信原理	线性代数	信号与系统
1	73	78	75	81	88	83
2	78	83	65	80	73	81
3	61	63	59	64	72	76
4	84	84	78	88	80	85
5	60	65	57	54	77	61
⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	66	85	59	47	75	63

如果使用一元统计方法，就要把多门课程分开分析，每次分析处理一门课的成绩，这样处理，忽视了课程之间可能存在的相关性，丢失信息太多，使得分析的结果不能客观全面地反映学生的学习情况。

如果使用多元分析方法，可以同时对多门课程成绩进行分析。例如：可以运用典型相关、对应分析、图形可视化了解这些课程之间的相互关系、相互依赖性等；可以运用主成分分析、因子分析研究影响成绩的主要因素，用主要因素（综合指标）来比较学生学习成绩的好坏；可以运用聚类分析对学生进行分类，从而对不同类别的学生分析成绩构成，制订相应学习计划；可以运用多元回归分析根据一些课程成绩预测其他课程成绩；可以运用判别分析根据一些课程成绩判别学生类别。上面提到的典型相关、对应分析、图形可视化、主成分分析、因子分析、聚类分析、判别分析等都属于多元统计分析的研究内容。

1.3 多元统计分析的主要应用

多元统计分析是解决实际问题有效的数据处理方法。随着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，如教育学、医学、气象学、环境科学、地质学、考古学、经济学、农业、社会科学、文学，从中可看到多元统计分析应用的广度和深度。

1. 教育学

若有 n 个高中考生高考成绩和高中学习期间的课程成绩数据，对其做多元统计分析，我们能够得出：

① 预测高考情况。由学生高考成绩和高中学习期间成绩的历史数据，研究高考成绩与学习过程成绩两组变量的关系，从而可由考生在高中期间的学习成绩来预测高考的综合成绩或某科目的成绩。

② 由考生成绩进行招生排队的最佳方案。虽然加和总分可以体现考生总的成绩好坏，但对报考数学系的学生，按加和的总分从高到低的顺序录取并不是最合适的，对于数学系的招生，数学、物理、外语的权数相对高些是比较合理的，应适当加权求和。

③ 奖学金的合理发放。利用 n 个学生在高中学习期间 m 门主科的考试成绩，可对学生进行分类，如按文科、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对其给出公平合理的解决方案。

感兴趣的读者可以参考书籍：《教育心理多元统计学与 SPSS 软件》，梁荣辉等；《多元描述统计方法》，李伟明。

2. 医学

运用多元统计的方法可以在医学研究中进行比较、判断关系、患病预测、病情分类、综合评价等。

通常医生对病人的诊断是通过观测病人的若干指标来综合评定的。例如，研究患糖尿病情况，根据医理确定研究指标，采集病人和正常人的年龄、家族史、工作性质、BMI、腰臀比等指标的样本数据，可以对正常人和糖尿病患者进行数据比较；利用历史病例资料，运用多元统计方法建立计算机辅助诊断系统（专家系统），进行分类及预测。对来就诊的病人，观测若干项指标后，可作出健康状况评价。可以对主要指标给出参考范围，预测哪些人更容易患糖尿病，并

对其进行生活健康指导。例如，比较不同地区儿童生长发育情况，对口腔牙病进行分类，预防牙病。再如，根据年龄、家族史、并发症、复发、化疗等预测乳腺癌患者手术后的生存时间等。以上问题都可以采用多元统计分析方法进行研究。

感兴趣的读者可以参考书籍：《医用多元统计分析方法》，陈峰；《医用多元统计方法》，张家放；《应用多变量统计分析》，孙尚拱。

3. 气象学

俗话说“天上钩钩云，地上雨淋淋。天有城堡云，地上雷雨临。天上扫帚云，三天雨降淋……”，看云识天气，就是通过历史资料做天气预测，要想更准确地预报天气情况，离不开多元数据分析。国内外各地建立了很多气象站，在不同时间各气象站都记录了降雨量、气温、气压、湿度、风速、风向等气象指标。对这些指标做多元统计分析，可以得出：

① 指标间的关系。如降雨与前一天的气温、气压、湿度等的关系，利用该关系可对降雨的可能性作出预报。

② 不同地点气象指标的关系。例如，计划建大型化工厂，我们关心化工厂区气象的精准情况。如果气象台站较远，可以先在厂区建个临时观测站，与气象台站同时测定气象指标。然后分析临时观测站和气象台站这两站气象指标的关系，以达到今后可由气象台站的气象资料来预报厂区的气象情况的目的。

感兴趣的读者可以参考书籍：《气象科研与预报中的多元分析方法》，施能；《气象统计分析与预报方法》，黄嘉佑。

4. 环境科学

① 研究大气环境污染的评估，以及环境污染与职工健康的关系。

② 国外学者研究了洛杉矶地区大气中污染物质的浓度。在较长的一段时间内，每天定时测定与污染有关的几个指标值。用多元统计检验的方法首先判断洛杉矶地区空气污染程度，在一周内是固定不变，还是周末与平时有显著差异。其次对这庞杂的观测数据用一种易解释的方法加以归纳化简。

③ 研究多种污染气体（CO，CO_2，SO_2）的浓度与污染源的排放量、气象因子（风向、风速、温度、湿度等）之间的相互关系。

感兴趣的读者可以参考书籍：《基于多元统计和 GIS 的环境质量评价研究》，王晓鹏和曹广超。

5. 地质学

随着地质科学向定量化发展，地质学和数学（主要是多元统计方法）结合起来产生了交叉学科——数学地质，多元地质分析是其主要内容之一。

例如，可以应用多元统计方法处理各种地质观测数据，对成矿规律进行评价，对矿产资源进行预测，对矿物构造进行推断和进行勘探工程部署等。

感兴趣的读者可以参考书籍：《地质数据的多变量统计分析》，王学仁；《实用地质统计学：空间信息统计学》，侯景儒；《数学地质的方法与应用：地质与化探工作中的多元分析》，於崇文。

6. 考古学

① 考古学家利用坟墓中的陪葬品（特别是陶瓷和珠宝）在式样和装饰上的差别，把它们按时间顺序排列起来。

② 考古学家对挖掘出来的人头盖骨可测得多种数据（如高、宽等），利用头盖骨的数据来

判断所属的种族，或判别性别，并研究最佳的测量法以及最少的测量数目。

③ 考古学家根据挖掘出的动物牙齿的有关测试指标，判别它属于哪类动物牙齿，是哪一个时代的。

感兴趣的读者可以参考书籍：《简明考古统计学》，陈铁梅和陈建立；*Statistics for Archaeologists: A Common Sense Approach*，Robert D. Drennan；《定量考古学》，陈铁梅。

7. 经济学

经济学是研究人类经济活动规律的学科。

① 城镇居民消费水平通常用以下指标来描述，如人均粮食支出、人均副食支出、人均烟酒茶支出、人均衣着商品支出、人均日用品支出、人均燃料支出、人均非商品支出。这些指标存在一定的关系，需要将相关的指标归并到一起，这实际上就是对指标进行聚类分析。根据分类结果还可进一步研究各类地区农民的生活水平、富裕程度，以便进一步研究城镇居民的消费结构和经济发展对策。

② 构造中国国民收入的生产、分配与最终使用的计量经济模型。例如，根据我国 1952—2017 年财政收入与国民收入、工农业总产值、人口、就业人口、固定投资等数据，用回归方法建立预测模型，对今后的财政收入作预测。

③ 在商业经济中，常常需要将很复杂的数据综合成商业指数形式，如物价指数、货币工资比、生活费用指数、商业活动指数等，用主成分分析可以从多个变量中构造出所需的商业指数。

④ 服装公司希望生产足够多的成衣以适应大多数顾客的要求，为此目的，首先在各地做抽样调查，对被调查人测量身体几十个部位的尺寸，然后对庞大的调查资料用多元统计方法进行分析和处理，确定一种服装究竟要有几种型号，每种型号服装的比例是多少，由身体的哪几个主要部位的尺寸决定。

感兴趣的读者可以参考书籍：《经济管理多元统计分析》，雷钦礼；《金融市场中的统计模型和方法》，黎子良和邢海鹏。

8. 农业

① 有 n 个不同地区，每个地区记录多种农作物的收获量，用多元统计方法对各个地区的总生产效率进行比较，并对不同的农业区域进行分类。

② 为了节省能源，对某地农用的手扶拖拉机的能源消耗进行抽样调查。调查的内容为拖拉机在田间进行运输、排灌、加工等作业时的燃油耗，再测月数、年平均更变零件数及平均燃油耗。通过对调查资料作多元统计分析，达到对拖拉机的平均燃油耗作预测，并对拖拉机进行分类（划分淘汰类、大修类、小修类和继续使用类）的目的。

感兴趣的读者可以参考书籍：《农业气象统计》，魏淑秋；《田间试验与统计分析》，明道绪；《土壤和环境研究中的数学方法与建模》，刘多森和曾志远。

9. 社会科学

青少年犯罪问题是一个很大的社会问题。对待青少年犯罪，我们采取“以防为主”的原则。可以研究目前犯罪的青少年的指标数据，进行聚类分析和判别分析，从而进行预测和防治。

感兴趣的读者可以参考书籍：《社会研究方法》，艾尔·巴比。

10. 文学

英国统计学家 Yule 把统计方法引入到文学词汇的研究，俄国著名数学家马尔可夫（1865—1922 年），在对俄语字母序列的研究中，提出了马尔可夫随机过程，语言结构中所蕴藏着的统计规律，成了思想的源泉。这种统计学的新分支称为文献计量学、数理语言学、计算风

格学、文字 DNA 等,即通过文献来搜寻信息。

例如,对美国立国三大历史文献之一的《联邦主义者》文集的研究。由于历史原因,文集 85 篇文章中,有 73 篇文章的作者身份较为明确,其余 12 篇署名都为 Federalist 的文章的真正作者身份曾引起长期争议。1955 年,哈佛大学统计学家 Fredrick Mosteller 和芝加哥大学的统计学家 David Wallance,通过研究"in""an""of""upon""while""whilst" "enough" "there" "on"等多个词汇的使用规律,花了十多年的时间,甄别了 12 篇文章的作者,引起了统计学界极大的轰动。

例如,中国古典名著《红楼梦》作者的研究。前 80 回的作者为曹雪芹,后 40 回的作者有争议。1985 年复旦大学统计运筹系的李贤平教授对著作权进行研究,选定数十个与情节无关的虚词(如了、吗、嘛、喱、呢、么……)作为变量,把全书中的 120 回作为 120 个样品,统计每一回这些虚词(即变量)出现的规律。在研究中主要使用聚类分析、主成分分析、典型相关分析等多元统计分析方法,结合历史考证,主要得出如下结果:

① 前 80 回和后 40 回不是出自同一个人的手笔;

② 前 80 回是否为曹雪芹所写?通过用曹雪芹的另一著作,做类似的分析,结果证实了用词手法完全相同,断定为曹雪芹一人手笔;

③ 后 40 回是否为高鹗所写?结论推翻了后 40 回是高鹗一人所写。后 40 回的成书比较复杂,既有残稿也有外人笔墨,不是高鹗一人所续。

感兴趣的读者可以参考文章:《〈红楼梦〉成书新说》,复旦学报(社会科学版),1987 年,第 5 期,李贤平。

本书主要介绍多元统计分析的理论及常用的方法,同时,利用 SPSS、S-Plus 等统计软件进行实证分析,做到在应用中体会理论。本书的章节结构如图 1.3.1 所示。

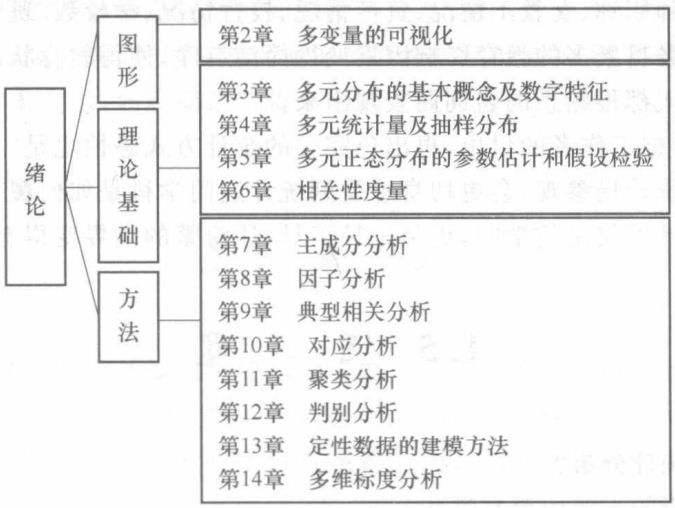

图 1.3.1　本书的章节结构

1.4　小贴士

俗话说"读万卷书,行万里路",读者有机会可以去中国统计资料馆参观一下。1952 年,为