Method and Demonstration Study of Multivariate Statistical Analysis

符想花◎著

多元统计分析方法与实证研究

METHOD AND
DEMONSTRATION STUDY OF MULTIVARIATE
STATISTICAL ANALYSIS

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

本书获得河南省教育统计数据分析和研究中心、河南省高等学校哲学社会科学创新团队、河南省高等学校哲学社会科学基础研究重大项目的资助

符想花◎著

多元统计分析方法与实证研究

METHOD AND
DEMONSTRATION STUDY OF MULTIVARIATE
STATISTICAL ANALYSIS

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

图书在版编目（CIP）数据

多元统计分析方法与实证研究/符想花著. —北京：经济管理出版社，2017.10
ISBN 978-7-5096-5426-2

Ⅰ.①多… Ⅱ.①符… Ⅲ.①多元统计—统计分析—研究 Ⅳ.①O212.4

中国版本图书馆 CIP 数据核字（2017）第 249162 号

组稿编辑：杨　雪
责任编辑：范美琴
责任印制：司东翔
责任校对：董杉珊

出版发行：经济管理出版社
（北京市海淀区北蜂窝 8 号中雅大厦 A 座 11 层　100038）
网　　址：www. E-mp. com. cn
电　　话：（010）51915602
印　　刷：玉田县昊达印刷有限公司
经　　销：新华书店
开　　本：720mm×1000mm/16
印　　张：19. 5
字　　数：310 千字
版　　次：2017 年 10 月第 1 版　　2017 年 10 月第 1 次印刷
书　　号：ISBN 978-7-5096-5426-2
定　　价：59.00 元

·版权所有　翻印必究·
凡购本社图书，如有印装错误，由本社读者服务部负责调换。
联系地址：北京阜外月坛北小街 2 号
电话：（010）68022974　　邮编：100836

前言

随着电子计算机的日益普及，各行各业都开始采用计算机及相应的信息技术进行管理和决策，数据量与日俱增，大数据时代扑面而来，对复杂数据进行处理的重要理论基础之一是多元统计分析，它起源于20世纪初，但由于在应用时需要大量的计算，使其发展受到影响。20世纪50年代中期开始，随着电子计算机的发展和普及，多元统计分析在生物、医学、经济分析等许多领域得到了广泛的应用，同时也促进了理论的发展。各种统计软件的开发和使用，使实际工作者利用多元统计分析方法解决实际问题更加简单、方便。20世纪80年代起在我国许多领域拉开了多元统计分析应用的序幕，并在其理论研究和应用上取得了显著成就。对实际领域中的研究者和高等院校的学生来说，要学习掌握多元统计分析的各种模型和方法，手头有一本既有多元统计分析方法又有实证分析的参考书是非常必要的，为此笔者独著了这本《多元统计分析方法与实证研究》。

本书的写作突出了以下四点：①对数学基础知识的要求较低，读者只需掌握线性代数、概率论与数理统计、微积分的基本知识。②在一元统计分析的基础上深入浅出地介绍多元统计分析的内容，这有助于读者对内容的理解和

掌握。③在介绍多元统计分析方法的同时，利用案例，介绍其在SPSS软件中的实现过程，并对部分输出结果做了解释和说明，将分析方法、案例及如何在SPSS中实现密切地结合在了一起。④注重实际应用，围绕多元统计分析方法选取实际案例，这些案例要么是笔者的科研成果，要么是笔者指导的本科毕业生论文。

本书共分两篇：第一篇为多元统计分析方法及在SPSS中实现，共十一章内容。第一章为绪论，介绍多元统计分析的作用及应用范围。第二章介绍多元分析的基本概念和基本理论，这些内容是在复习一元统计分析内容的基础上自然引申出来的。第三章介绍多元正态总体均值向量和协差阵的假设检验。第四章和第五章主要介绍分类问题，包括聚类分析和判别分析。第六章到第八章介绍数据简化问题，包括主成分分析、因子分析和对应分析。第九章和第十章阐述两组变量之间的相互关系，包括典型相关分析和多重多元回归分析。第十一章是定性资料的统计分析。第二篇为实证分析部分，共四个案例，分别是加快河南省高技术产业发展研究、滴滴打车营销模式分析——以郑州市为例、基于现代统计分析的区域高技术产业技术创新能力比较研究、劳动者报酬占比变化的模拟与分析。

由于笔者水平有限，书中难免有不足之处，恳请读者批评指正。

符想花

2017年7月

试读结束：需要全本请在线购买： www.ertongbook.com

目录

第一篇　多元统计分析方法及在SPSS中实现

第二篇 实证分析

第一篇
多元统计分析方法及在SPSS中实现

第一章　绪论

多变量统计分析又称多元统计分析，它起源于20世纪初，1928年J. 维希特（J. Wishart）发表论文《多元正态总体样本协差阵的精确分布》，被认为是多变量分析研究的开端。20世纪30年代，R. A. 费希尔（R. A. Fisher）、H. 霍特林（H. Hotelling）、许宝禄以及S. N. 罗伊（S. N. Roy）等人做了一系列奠基性的工作，使多元分析在理论上得到迅速发展。20世纪40年代，多变量分析方法在心理学、教育学、生物学等领域获得了一些应用，但由于在应用时需要大量的计算，加上第二次世界大战，使其发展受到影响，甚至停滞了相当长的时间。20世纪50年代中期开始，随着电子计算机的发展和普及，多元统计分析方法在地质、气象、生物、医学、经济分析等许多领域得到了广泛的应用。20世纪60年代，通过应用和实践又完善、发展了理论，由于新的理论、新的方法不断涌现，促使它的应用范围更加扩大。20世纪80年代起，我国许多领域都拉开了多元统计分析应用的序幕，并在其理论研究和应用上取得了显著成就。

第一节　多元统计分析的作用

一、多元统计分析的概念

在工业、农业、医学、气象、环境以及经济、管理等众多领域，常常需要同时观测多个指标来分析和研究问题。例如，在经济管理中，要对国有企业资本金效绩进行评价，需观测净资产收益率、总资产报酬率、总资产周转

试读结束：需要全本请在线购买：　www.ertongbook.com

率、流动资产周转率、资产负债率、已获利息倍数、销售增长率和资本积累率等多个指标。在统计学上，通常将指标称为变量，变量有确定性变量和随机性变量。由于受某些确定性因素的影响，现象的量沿着某一方向持续变化，这样的量就是确定性变量。如由于科学技术的不断提高和医疗卫生条件的不断改善，人类的平均寿命在不断延长，因此，从长期看，人的平均寿命是一个确定性变量。而有些变量的变动受许多因素变动的影响，变量值的大小没有明确的方向，这样的变量就是随机性变量。如净资产收益率、总资产报酬率、总资产周转率等在不同的企业表现为不同的数值，这些变量都是随机性变量。

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？传统的方法是把多个随机变量分开进行分析和研究，每次只处理一个变量逐次分析研究，但当变量较多时，变量之间不可避免地存在相关性，如果分开处理变量不仅会丢失很多信息，也不容易取得较好的研究结果。而现代的分析方法是对多个变量同时进行分析研究，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变量内在的变化规律。多元统计分析就是研究多个随机变量之间相互依赖关系及内在统计规律性的一门统计学科。

二、多元统计分析方法的作用

（一）能够简化数据和数据结构

要认识客观现象，往往需要从多角度、多方面进行系统的相互联系的考察，这必然要用到统计指标体系。构成统计指标体系的多个指标，各有侧重地解释、说明同一事物，多重共线性的出现不可避免。为此，在不损失数据蕴含信息量的情况下，通过变换数据和构造模型剔除指标间相互制约的成分，尽可能简单地将被研究现象描述出来，就可达到简化数据和数据结构之目的。主成分分析、因子分析等均可达到这样的目的。

（二）能够进行分类或分组

通过统计分组将性质相同的单位（或者变量）归为一组，将性质不同的单位（或者变量）归为不同的组，有利于对问题做深入细致的分析和研究。根据所测量到的某些特征数据对研究现象进行分类或分组，是多元统计分析的另一目的所在。聚类分析和判别分析等可达到这样的目的。

（三）能够研究变量间的相互依赖关系

人们对变量间关系的本质感兴趣。是否所有的变量都相互独立？还是一个变量或一些变量依赖于另一个或一些变量？如果是后者，这种依赖关系是怎样的？相关分析（包括简单相关分析、复相关分析、偏相关分析、典型相关分析）和回归分析（一元回归分析、多元回归分析、多重多元回归分析）就可研究此类问题，主成分分析、因子分析、对应分析等亦可研究此类问题。

（四）能够进行预测或控制

在生产与生活中，探索多变量系统内在的客观规律性及其与外部环境的关系，进行预测预报，以实现对系统的最优控制，是应用多元统计分析技术的主要目的，而多个变量之间关系的建立，为从一些变量观测值预测另一个或一些变量值提供了可能。

（五）能够进行假设检验

检验多元总体参数表示的某种统计假设，可以证实某种假设的合理性或支持事先树立的某种信念。

三、多元统计分析在现实生活中经常处理的问题

多元统计分析作为科学研究的重要工具，在自然科学、社会科学等诸多方面有着广泛的应用。

（一）在经济、管理中经常处理的问题

（1）要对我国大型工业企业的经济效益进行综合评价，其做法是选取能够反映经济效益的代表性指标，如百元固定资产实现利税、资金产值率、资金利税率、全员劳动生产率等，根据这些指标对大型工业企业分类，根据分类结果可给出对企业的评价。对大型工业企业分类应用的是聚类分析法。

（2）给出美国、日本、英国等经济发达国家和印度、罗马尼亚、南非等发展中国家反映其经济发展程度的若干指标，如人均国内生产总值、人均收入、人均消费支出等，要根据这些指标判断中国所属类型，应用的是判别分析法。

（3）要对我国大型工业企业经济效益进行综合评价，评选出前 500 强，其做法同样是选取能够反映经济效益的代表性指标，根据这些指标采用主成

分分析或因子分析法即可达到此目的。

（4）农村居民人均消费支出是用来反映和研究农村居民实际生活消费水平的重要指标，主要包括食品烟酒、衣着、居住、生活用品及服务、交通通信、教育文化娱乐、医疗保健、其他用品及服务。如果收集到某个年份我国各个省、直辖市、自治区该指标数据，可以用对应分析（相应分析）的方法，研究各个省、直辖市、自治区农村居民人均消费支出的分布规律。

（5）某一产品质量的好坏可用多个指标进行衡量，而影响产品质量好坏的因素亦有多个，要揭示影响产品质量的多个因素与衡量产品质量的多个指标之间的依赖关系，需要用典型相关分析法，如果它们之间存在相关关系，这时要用多重多元回归分析法建立回归模型，以便进行预测预报。

（6）不同原料生产的两种产品，其产品寿命有无显著性差异，可进行假设检验。

（二）在其他领域研究中的应用

（1）在一项癌症患者对放射性疗法反应的研究中，对一批患者测量多项反应指标，因为难以同时解释所有变量的观测值，于是人们需要寻找刻画患者的一个简单综合性指标，主成分分析或因子分析能够帮助研究者达到这一目标，且不损失数据中蕴含的许多信息。

（2）在地质学中，常常要研究矿石中所含化学成分之间的关系。设在某矿体中采集了 60 个标本，对每个标本测得 20 个化学成分的含量。我们希望通过对这 20 个化学成分的分析，了解矿体的性质和矿体形成的主要原因。

（3）在考古学中，考古学家对挖掘出来的人头盖骨的高、宽等特征来判断是男或女。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，是哪一个时代的动物。

（4）在植物育种学中，植物收获时，要选择种子，以保证在一些特征上，下一代优于上一代，这时往往要测量、评价很多特征，将这些特征通过变量变换变为综合指标，根据综合指标得分的高低来选择种子。

（5）在社会学中，对当代大学生择偶标准（对文化程度、职业、居住条件、收入、相貌等的要求）进行调查，在此基础上做主要因素分析，以便对他们的家庭婚姻观作正确引导。

（6）在文学研究中，同样可以应用多元统计分析方法。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，但长期以来对这一结论存在争议。1985~1986年，复旦大学的李贤平教授带领他的学生对《红楼梦》一书采用多元统计分析的方法进行了研究，他们创造性地将120回看成120个样品，并确定与情节无关的虚词作为变量，让学生数出每一回里各变量出现的次数，用聚类分析的方法进行了分类，分类的结果是前80回为一类，后40回为一类。为了判断前80回是否为曹雪芹所写，他们找到曹雪芹的其他著作，做了类似的分析计算，证实用词手法完全相同，结果断定前80回为曹雪芹一人所写。而后40回是否为高鹗所写？论证结果是并非如此。他们用多元统计分析的方法支持了红学界的观点，使红学界大为赞叹。

尽管上述问题不同、领域不同，但从对数据的处理上看，却有很多相似之处。正是多元统计分析方法能够应用于各个领域，才使得该学科能够伴随着计算技术的普及而得到广泛的应用。

第二节　主要内容安排

本书分为两篇。第一篇为多元统计分析方法及在SPSS中实现部分，具体包括：

第一章是绪论，介绍多元统计分析的作用及其应用范围。第二章和第三章介绍多元统计分析的基本概念和基本理论。包括多元正态分布、Wishart分布、HotellingT^2分布、Wilks分布、多元正态总体的参数估计和假设检验。而这些内容都是一元统计分析中相应内容的推广，因此这两章的内容是在复习一元统计分析内容的基础上自然引申出来的，这样读者在学习时就不会感觉抽象和困难。

第四章和第五章阐述分类问题。聚类分析是研究“物以类聚”的一种现代多元统计分析方法，判别分析是在研究对象用某种方法分好若干类的情况下，确定新样品属于已知类别中哪一类的多元统计分析方法。实际应用时往往将这两种方法结合起来使用，因为判别分析要求对新样品进行判别归类之前，应先知道已对总体分了几类，然后建立判别式，根据一定的判别规则对

新样品进行判别归类。如果在判别分析之前，并不知道事先划分了几类，这时可先做聚类分析然后再做判别分析。

第六章到第八章介绍主成分分析、因子分析和对应分析。主要阐述数据简化问题，将具有错综复杂关系的变量（或样品）综合成数量较少的综合指标，用较少的综合指标以尽可能简单地表示所研究的对象，又不至于损失很多有价值的信息。

第九章和第十章介绍两组变量之间的相互关系，包括典型相关分析和多重多元回归分析。典型相关分析能够反映两组变量之间相互线性依赖关系，但根据典型相关关系不能根据自变量的取值来求解因变量的可能取值，要解决这一问题，需要多重多元回归分析。这两章内容是在简单相关和一元回归分析的基础上引申而来的。

第十一章是定性资料的统计分析。在实际应用中，不可避免地要涉及定性变量，例如人的性别、人的职业、天气状况、产品等级等。对诸如这些定性变量给予相应的数量描述，从而进行有关的统计分析，是定性资料统计分析要解决的问题。但本章并不详细介绍这方面的理论、方法，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。

在第一篇的各章中，除介绍统计分析方法的理论外，还有统计分析方法在 SPSS 等统计软件上的实现及对输出结果的解释和说明。多元统计分析方法若要运用于实际，需要大量的计算，而这些计算靠手工几乎无法完成，如果不借助于计算机，许多问题根本无法解决。本书中各种统计方法均要借助案例，采用国际上流行的通用统计软件包 SPSS 来实现，这样不仅能体现多元统计分析方法的理论价值，而且能更好地显示出其应用价值。

第二篇为实证分析部分。在许多实际问题中，常常需要同时观测多个指标来分析和研究某一个问题，从哪个角度、哪个方面选取指标，选取的依据是什么，到底选取哪些指标分析问题，这是至关重要的，但这往往又是被学生所忽视的，我们的大学生在写毕业论文时，也会选用一些统计指标，但他们是不加任何说明地选用，为此，在第二篇中选用作者部分科研成果，给出指标的选取原则及选取的依据，以供读者参考使用。统计指标选取出来后，就需要采用一定的分析方法分析问题，分析问题的方法很多，每种方法都有

其自己的理论和实用价值，但也会有一定的局限性。这需要根据研究问题的目的、研究对象的特点及所掌握的资料加以选择。实证分析部分主要是围绕多元统计分析方法的使用选取实证案例，这些实证案例要么是作者的科研成果，要么是作者指导的本科毕业生论文。具体包括：加快河南省高技术产业发展研究（2011 年度河南省哲学社会科学规划项目，批准号：2011FJJ027，结项证书号：2013A037）、滴滴打车营销模式分析——以郑州市为例（河南财经政法大学 2017 年本科毕业论文）、基于现代统计分析的区域高技术产业技术创新能力比较研究（发表在《商业时代》2012 年第 30 期）、劳动者报酬占比变化的模拟与分析（发表在《商业经济研究》2015 年第 18 期）。通过这些实证案例，希望能为读者提供写作的些许帮助。

试读结束：需要全本请在线购买： www.ertongbook.com