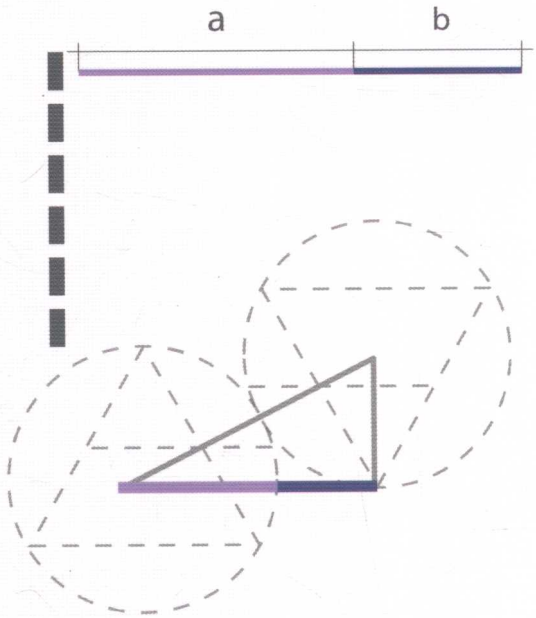
普通高等教育“十三五”规划教材

多元统计分析：从数据到结论

韩　明　编著



上海财经大学出版社
SHANGHAI UNIVERSITY OF FINANCE & ECONOMICS PRESS

普通高等教育“十三五”规划教材

多元统计分析：从数据到结论

韩　明　编著

上海财经大学出版社

内容提要

全书共由12章组成，在介绍多元统计分析的有关概念、相关背景的基础上，突出统计思想，着重讲解常用方法及其应用. 主要内容包括多元数据的表示及可视化、线性回归分析、逐步回归与回归诊断、广义线性模型与非线性模型、方差分析、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. 本书图文并茂，注重可读性，着重于多元统计分析方法在各个领域中的应用，将应用案例贯穿始终，并给出了R软件、MATLAB的相关程序.

本书可以作为高等院校有关专业本科生、研究生“多元统计分析”课程的教材或参考书，也可作为全国大学生(研究生)“数学建模竞赛”、全国大学生“统计建模大赛”的培训教材或参考书，还可以供有关专业的教师、研究人员和工程技术人员以及广大自学者参考.

图书在版编目(CIP)数据

多元统计分析：从数据到结论 / 韩明编著. —上海：上海财经大学出版社，2016.8

(普通高等教育“十三五”规划教材)

ISBN 978-7-5642-2519-3/F.2519

Ⅰ.①多… Ⅱ.①韩… Ⅲ.①多元分析—统计分析—高等学校—教材 Ⅳ.①O212.4

中国版本图书馆CIP数据核字(2016)第175052号

□ 责任编辑 温 涌

□ 封面设计 杨雪婷

DUOYUAN TONGJI FENXI CONG SHUJU DAO JIELUN

多元统计分析：从数据到结论

韩 明 编著

上海财经大学出版社出版发行

(上海市武东路321号乙 邮编200434)

网 址：http://www.sufep.com

电子邮箱：webmaster@sufep.com

全国新华书店经销

同济大学印刷厂印刷

上海叶大印务发展有限公司装订

2016年8月第1版 2016年8月第1次印刷

787 mm×1 092 mm 1/16 18.5印张 450千字

印数：0001—3000 定价：45.00元

前　言

“多元统计分析”课程已经被越来越多的将来需要与数据打交道的本科生和研究生的相关专业列为必修课或选修课.随着我国高等教育进一步“大众化”,特别是相关软件的普及,学习“多元统计分析”的人越来越多,人们不再只满足于学习一些理论知识,大家更希望将此作为工具,借助计算机和相关软件进行数据处理和分析.

作者结合多年来的教学实践,深感一本内容简练但又实用的“多元统计分析”教材的重要性.在已有的相关教材中,有的侧重理论的讲述,读者需要具备较深厚的数学基础;有的则注重模型的应用,理论和技术细节不是重点.本书在介绍多元统计分析的有关概念、背景的基础上,突出统计思想,着重讲解常用方法及其应用,并侧重于应用.本书书名为《多元统计分析:从数据到结论》(*Multivariate Statistical Analysis: From Data to Conclusions*),意在“应用”.书中将一些严格的数学推导过程略去而只列出结论(降低了对数学基础的要求),读者学习时关键是理解这些结果,清楚它们的意义和背景.对一些被略去的推理论证部分,感兴趣者可参考书后列出的有关文献.

本书汲取了国内外相关教材中流行的直观、灵活的教学方式,以及通过图表和应用案例进行教学这些长处.本书中的例题可以分为两类:一类是为了说明有关理论或方法的简单问题(这类问题一般不需要借助软件);另一类是为了应用有关理论或方法解决一些比较复杂的问题(应用案例),这类问题的解决一般需要借助软件才能实现.

考虑到作为一款免费软件,R 软件具有丰富的资源、良好的扩展性和完备的帮助系统,并且考虑到 MATLAB 在工程等领域中应用的广泛性、在国内外各高等院校中使用的普及性,本书的应用案例采用 R 软件和 MATLAB,并给出了相应的程序.

感谢王家宝教授在作者写作本书过程中给予的指导和鼓励.本书的编写得到宁波工程学院理学院的支持,在此表示感谢.

作者结合多年的教学实践,把一些教学经验、教学研究成果和教学心得体会等写进了本书,希望能和广大读者一起分享.虽然作者努力使本书成为一本既有特色又便于教学(或自学)的教材,但由于水平所限,书中难免还存在一些疏漏甚至是错误,恳请专家和读者批评和指正.

韩　明

2016 年 8 月

目　录

第1章 绪 论

多元统计分析(Multivariate Statistical Analysis)是应用统计方法来研究多变量(多指标)问题的理论和方法,它是一元统计学的推广.

有了一元统计学的理论和方法,为什么还要多元统计分析呢?在实际问题中,很多随机现象涉及的变量不是一个,而经常是多个变量,并且这些变量之间又存在一定的联系.我们常需要处理多个变量的观测数据,那么如何对多个变量的观测数据进行有效的分析和研究呢?一种做法是,把多个变量分开分析,一次处理一个地去分析和研究;另一种做法是,同时对多个变量进行分析和研究.显然前者的做法有时是有效的,但一般来说,由于变量多,避免不了变量之间有相关性,把多个变量分开处理不仅会丢失一些信息,往往也不容易取得很好的研究结果.而后一种做法通常可以用多元统计分析方法来解决,通过对多个变量的观测数据进行分析,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.

1.1 多元统计分析概述

如果说一元统计分析是研究一个变量统计规律的学科,那么多元统计分析则是研究多个变量之间的内在统计规律的统计学科.

早在19世纪就出现了处理二维正态总体的一些方法,但系统地处理多维概率分布总体的统计分析问题则开始于20世纪.多元统计分析起源于20世纪初,1928年Wishart发表的论文《多元正态总体样本协方差矩阵的精确分布》,可以说是多元统计分析的开端.之后Fisher、Hotelling、Roy、许宝騄等人做出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.

20世纪40年代,多元统计分析在心理学、教育学、生物学等方面有不少的应用,但由于计算量大,使其发展受到影响.20世纪50年代,随着计算机的出现和发展,使多元统计分析在地质学、医学、气象学、社会学等方面得到了广泛的应用.20世纪60年代,通过应用和实践又完善和发展了多元统计分析理论,由于新理论和新方法的不断出现,又促使它的应用范围更加扩大.20世纪七八十年代,多元统计分析在我国才受到各个领域的极大关注.近40年来,我国在多元统计分析的理论和应用上取得了许多显著的成绩.

进入21世纪后,人们获得的数据正以前所未有的速度迅速增加,产生了海量数据、超大型数据库等,遍及超级市场销售、银行存款、天文学、粒子物理学、化学、医学、生物学以及政

府统计等领域,多元统计分析与人工智能、数据库技术等相结合,已经在经济学、商业、金融、天文、地理、农业、工业等方面取得了成功的应用.

"多元统计分析"也称为"多元分析"(Multivariate Analysis). 例如,Mardia et al. (1979)的书名为 *Multivariate Analysis*. 英国著名的统计学家 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)

简化数据结构就是将某些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. 例如,主成分分析、因子分析、对应分析等就是这样的一类方法.

(2) 分类与判别(归类问题)

归类问题就是对所考察的观测点(或变量)按照相近程度进行分类(或归类). 例如,聚类分析、判别分析等就是解决这类问题的统计方法.

(3) 变量间的相互联系

相互依赖关系:分析一个或几个变量的变化是否依赖于另外一些变量的变化. 如果是,建立变量之间的定量关系式,并用于预测或控制——回归分析.

变量之间的相互关系:分析两组变量之间的相互关系——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和协方差矩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(特别是多维正态随机向量),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.

1.2 多元统计分析的应用

多元统计分析可以应用于几乎所有的领域,主要包括经济学、农业、地质学、医学、工业、气象学、金融、精算、物理学、地理学、军事科学、文学、法律、环境科学、考古学、体育科学、遗传学、教育学、生物学、管理科学、水文学等,还有一些交叉学科或方向等. 多元统计分析的应用实在是难以一一罗列,以下简要地介绍一下多元统计分析在文学、数据挖掘(作为交叉学科或方向的代表)领域的应用.

在**文学**方面,自从 20 世纪 30 年代末英国著名的统计学家 Yule 把统计方法引入文学词汇的研究以来,这个领域已经取得了不少进展,其中最有名的是 Mosteller 与 Wallace 在 20 世纪 60 年代初对美国立国三大文献之一的《联邦主义者》文集的研究.

在 1985～1986 年间,复旦大学李贤平教授对我国名著《红楼梦》的著作权进行了研究,使用的统计方法主要是多元统计分析. 先选定数十个与情节无关的虚词作为变量,把《红楼梦》一书中的 120 回作为 120 个样品,统计每一回(即每个样品)选定的这些虚词(即变量)出现的频数,由此得到的数据矩阵作为分析的依据.

在《红楼梦》著作权的研究中，使用较多的是聚类分析、主成分分析、典型相关分析等方法，由分析结果可以看出：

(1) 前80回和后40回截然地分为两类，证实了前80回和后40回不是出于同一个人的手笔.

(2) 前80回是否为曹雪芹所写？通过对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) 而后40回是否为高鹗所写？分析结果发现，后40回依回目的先后可分为几类，得出的结论否定了后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.

以上这些论证在红学界引起了轰动. 李贤平教授等人用多元统计分析方法提出了关于《红楼梦》作者和成书过程的新学说.

李贤平教授等人还把这类方法用于分析其他作家和作品. 结果证明，统计方法的分辨能力是很强的.

在**数据挖掘**方面，随着科学技术的发展，利用数据库技术来存储、管理数据，利用机器学习的方法来分析数据，从而挖掘出大量的隐藏在数据背后的知识，这种思想的结合形成了深受人们关注的非常热门的研究领域——数据库中的知识发现(knowledge discovery in databases)，数据挖掘(data mining)技术便是其中的一个最为关键的环节. 数据挖掘、机器学习(machine learning)等为统计学(包括“多元统计分析”)提供了一个新的应用领域，同时也提出了很多挑战. 多元统计分析中的聚类分析(cluster analysis)是按照某种相近程度，将用户数据分成一系列有意义的集合. 例如，在金融领域中，将贷款对象分为低风险和高风险等. 数据挖掘是一个交叉学科，它涉及数据库、人工智能、统计学、并行计算等不同学科和领域，近年来受到各界的广泛关注. 应该指出，Johnson 和 Wichern 在 *Applied Multivariate Statistical Analysis*(6th ed. 2007)中补充了“数据挖掘”部分，以及多元统计分析方法在数据挖掘中的应用. 数据挖掘与统计学有着密切的关系，那么统计学如何为数据挖掘服务呢？这是在“数据挖掘”飞速发展的今天，统计学必须回答的一个问题. (令人高兴的是，现在可以从统计学在数据挖掘领域中的研究与应用情况，看到对这个问题的各种回答.)数据挖掘对统计学带来的挑战，无疑将推动统计学的发展(韩明. 数据挖掘及其对统计学的挑战[J]. 统计研究，2001.). 关于统计分析与数据挖掘，感兴趣的读者可参考：张尧庭，谢邦昌，朱世武(2001)；朱建平(2005)；Johnson and Wichern(2007)；薛薇(2014)等.

1.3 有关软件介绍

相关软件的种类很多，有些功能齐全，有些价格便宜，有些容易操作，有些需要更多的实践才能掌握. 这里简要介绍最常见的以下几种.

(1) SAS：这是功能非常齐全的软件；尽管价格相当不菲，但许多公司，特别是美国制药公司都在使用，这多半因为其功能众多和某些美国政府机构一些人的偏爱. 尽管现在已经尽量“傻瓜化”，但仍然需要一定的训练才可以进入. 也可以对它编程，但对于基本统计课程

则不那么方便.

(2) SPSS:这是一个很受欢迎的统计软件;它容易操作,输出漂亮,功能齐全,价格合理.它也有自己的程序语言,但基本上已经"傻瓜化".它对于非专业统计工作者是很好的选择.

(3) Excel:它严格来说并不是统计软件,但作为数据表格软件,必然有一定的统计和计算功能.而且凡是有 Microsoft Office 的计算机,基本上都装有 Excel.但要注意,有时在装 Office 时,没有装数据分析的功能,那就必须装了才行.当然,画图功能是已经具备的了.对于简单分析,Excel 还算方便,但随着问题的深入,Excel 就不那么"傻瓜化",需要使用宏命令来编程;这时就没有相应的简单选项了.多数专门一些的统计推断问题,还需要其他专门的统计软件来处理.

(4) S-plus:这是统计学家喜爱的软件.不仅由于其功能齐全,而且由于其强大而又方便的编程功能,使得研究人员可以编制他们的程序来实现其自己创造的理论和方法.它也在进行"傻瓜化"以争取顾客,但仍然以编程方便为顾客所青睐.

(5) R 软件:完全免费,并且安装非常方便.(在网站 https://cran.r-project.org/bin/windows/base/上可下载到 R 软件的 Windows 版,点击 Download R 3.2.4 for Windows下载,按照提示安装即可,非常简便.作者在写作本书后期时的最新版本是R 3.2.4.)这是由志愿者管理的软件.其编程语言与 S-plus 所基于的 S 语言一样,很方便.还有不断加入的从事各个方向研究者编写的软件包和程序.从这个意义上说,其函数的数量和更新远远超过其他软件.它的所有计算过程和代码都是公开的,它的函数还可以被用户按需要改写.它的语言结构和 C++、Fortran、MATLAB、Pascal、Basic 等很相似,容易举一反三.对于一般非统计工作者来说,主要问题是它没有"傻瓜化".

(6) MATLAB:这也是应用于各个领域的以编程为主的软件,在工程上应用广泛.编程类似于 S 和 R.MATLAB(Matrix Laboratory)提供了一个人机交互的数学系统环境,并以矩阵作为基本的数据结构,可以大大节省编程时间.MATLAB 具有强大的符号演算、数值计算和图形分析功能.

当然,还有很多其他的软件,这里就不一一罗列了.其实,读者只要学会使用一种软件,使用其他的软件也不会困难,最多看看帮助和说明即可.学习软件的最好方式是需要时在使用中学.

本书的案例采用 R 软件和 MATLAB.(吴喜之教授用 SPSS、SAS 和 R 软件写的书《统计学:从数据到结论》影响很大,值得一读.)

1.4 本书的基本框架和内容安排

《多元统计分析》课程已经被越来越多的将来需要与大量数据打交道的本科生和研究生的相关专业列为必修课或选修课.《多元统计分析》的教材版本众多,其中有的教材侧重理论的讲述(传统的相关教材大多属于此类),读者需要具备较深厚的数学基础;有的教材则注重模型的应用,理论和技术细节不是重点.本书在介绍多元统计分析的有关概念、背景

和常用方法的基础上，侧重于应用，书名为《多元统计分析：从数据到结论》(*Multivariate Statistical Analysis: From Data to Conclusions*)，意在“应用”.

为了吸引广大读者学习多元统计分析，本书不得不在某种程度上牺牲内容难度的一致性；有些章节的难度可能会略大一些，因而初次阅读时会感到一些困难，希望教师们能在选择适合学生的章节时设法弥补这种不平衡，必要时可以降低一些要求. 本书将一些严格的数学推导略去而只列出结论(降低了数学基础的要求)，读者学习时关键是理解这些结果，清楚它们的意义和背景；对一些被略去的推理论证部分，可参考书后列出的有关文献.

作者结合多年来的教学实践，深感一本内容简练但又实用的《多元统计分析》教材的重要性. 随着我国高等教育进一步“大众化”，特别是相关软件的普及，学习《多元统计分析》的人越来越多，人们不再只满足于学习一些理论和方法，大家更希望将此作为工具，借助计算机和相关软件进行数据的处理和分析.

考虑到作为一款免费软件，R 软件具有丰富的资源(涵盖了多种行业数据分析中几乎所有的方法)、良好的扩展性(方便的编写函数和程序包，可以胜任复杂数据的分析、绘制精美的图形)、完备的帮助系统(每个函数都有统一格式的帮助)；另外，考虑到 MATLAB 在工程等领域中应用的广泛性、在国内外各高等院校中使用的普及性，本书的案例采用 R 软件和 MATLAB.

本书在介绍有关概念、背景的基础上，突出统计思想，重点介绍多元统计分析中的常用方法，主要包括多元数据的表示及可视化、线性回归分析、逐步回归与回归诊断、广义线性模型与非线性模型、方差分析、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. 本书注重体现多元统计分析在各个领域的应用，将应用案例贯穿于理论讲解的始终，并给出了 R 软件、MATLAB 的相关程序. 本书汲取了国内外教材中流行的直观、灵活的教学方法，以及通过图表和应用案例进行教学这些长处.

本书中的例题可以分为两类：一类是为了说明有关理论或方法的简单问题(这类问题一般不需要借助软件)；另一类是为了应用有关理论或方法解决一些比较复杂的问题(应用案例)，这类问题的解决一般需要借助软件才能实现.

说明：本书不再给出关于 R 软件、MATLAB 的“使用说明”，建议需要的读者可参考：

(1) 关于 R 软件的“使用说明”，可参考：薛毅，陈立萍(2007)；Cryer & Chan(2008)；汤银才(2008)；何春雄，朱锋峰，龙卫江(2012)；吴喜之(2013)；Kabacoff(2013)等.

(2) 关于 MATLAB 的“使用说明”，可参考：Freedman(2008)；包科研(2011)；韩明，王家宝，李林(2015)等.

1.5 思考与练习题

1. 根据你感兴趣的领域，查阅有关资料并说明多元统计分析在该领域中的应用情况.
2. 根据你感兴趣的软件(比如 R 软件、MATLAB 等)，请选择一种安装在你的个人计算机上(或已安装该软件)，为配合本课程的学习，请熟悉该软件的基本操作.

第 2 章　多元数据的表示及可视化

每天翻开报纸或打开电视，就可以看到各种数据，比如高速公路通车里程、物价指数、股票行情、外汇牌价、犯罪率、房价、流行病的有关数据；当然还有国家统计局定期发布的各种国家经济数据、海关发布的进出口贸易数据等. 从这些数据中，各有关方面可以提取对自己有用的信息.

某些企业每年都要花掉数目可观的经费来收集和分析数据. 它们调查其产品目前在市场中的状况和地位并确定其竞争对手的态势；它们调查不同地区、不同阶层的民众对其产品的认知程度和购买意愿，以改进产品或推出新品种以争取新顾客；它们还收集各地方的经济交通等信息，以决定如何保住现有市场和开发新市场. 市场信息数据对企业是至关重要的. 面对着一堆数据，我们该如何简洁明了地反映出其中规律性的东西或所谓的信息呢？一般首先对收集来的数据进行描述性分析，以发现其内在的规律性，然后再选择进一步分析的方法.

数据作为信息的载体，当然要分析数据中包含的主要信息，也就是分析数据的主要特征——数字特征. 对一元数据，即样本数据（或观测值）x_1，x_2，…，x_n 是从一元总体中抽取的. 一元数据的数字特征主要有：均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$、方差 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$、标准差 $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ 等. 对于多元数据，除分析各分量的取值特征外，还要分析各分量之间的相关关系.

由于多元统计分析中的符号多而杂，因此需要**说明**：在一元统计学中一般用大写和小写字母分别来区分随机变量及其观测值，在本书后面的章节里，由于其他复杂的符号，我们可能不再遵守此约定（Anderson 在其所著 *An Introduction to Multivariate Statistical Analysis*（3rd ed. 2003）中也采用了类似的做法），请读者注意一个符号在每一章中的意义.

对于多元数据，通常要研究其分量指标的相关性，图形表示（可视化）就显得尤其重要. 将数据显示在一个平面图上，可以非常直观地了解、认识数据，发现其中的可能分布规律. 多元数据的可视化（图形表示）方法主要有直方图、散点图（二维和三维）、Q-Q 散点图、散点图矩阵、条形图、饼图、尾箱图、小提琴图、星相图等.

2.1　多元数据的矩阵表示

2.1.1　多元数据的一般格式

当人们要研究一个社会现象或自然现象时，通常要选择一些变量的特征来进行记录，从而形成多元数据. 对于每个项目，这些变量的值被记录下来.

我们用 x_{ij} 表示第 j 个变量 $X_j(j=1, 2, \cdots, p)$ 在第 i 项或第 $i(i=1, 2, \cdots, n)$ 次试验中的观测值，因此 p 个变量的 n 个观测值如表 2-1 所示.

表 2-1　　**p 个变量的 n 个观测值**

	变量 X_1	变量 X_2	…	变量 X_p
记录 1	x_{11}	x_{12}	…	x_{1p}
记录 2	x_{21}	x_{12}	…	x_{2p}
…	…	…	…	…
记录 n	x_{n1}	x_{n2}	…	x_{np}

可以用一个有 n 行 p 列的矩阵来表示这些数据，称为**数据矩阵**，记为

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = (x_{ij})_{n\times p}.$$

于是以上数据矩阵包含了全部变量的所有观测值.

当这些变量处于同等地位时，就是聚类分析、主成分分析、因子分析、对应分析等模型的数据格式；当其中一个变量是因变量，而其他变量为自变量时，就是回归分析等模型的数据格式；若此时因变量还是分类变量，则为方差分析、判别分析等模型的数据格式.

例 2.1.1　从一个大学的书店收集到 4 张收据来了解书的销售情况. 每张收据提供了售书数量以及总金额. 用第一个变量来表示总销售金额，用第二个变量来表示售出书的数量. 然后我们可以把收据上的相关数据看作这 2 个变量的 4 个观测值，假定数据如表 2-2 所示.

表 2-2　　**2 个变量的 4 个观测值**

	变量 X_1	变量 X_2
记录 1	42	4
记录 2	52	5
记录 3	48	4
记录 4	58	3

而数据矩阵由 4 行 2 列组成，即