“十二五”国家重点图书出版规划项目

应用统计工程前沿丛书

多元统计分析

杜子芳 著

清华大学出版社

“十二五”国家重点图书出版规划项目

应用统计工程前沿丛书

多元统计分析

杜子芳 著

清华大学出版社
北京

内 容 简 介

本书内容广泛，通俗易懂，对数学和数理统计的要求很低，是一本极具特色的统计学教科书和工具书，既适合那些学习统计学课程的经济学、社会学、管理学和统计学专业的大学高年级本科生与研究生，也适合那些从事数据分析工作需正确理解各种多元统计方法的原理，掌握基本操作技巧的数据工程师，对于那些备考研究生的考生更不失为一本深入浅出、简明扼要的参考书。作者拥有多年授课经历和丰富的实际经验，力求说理透彻，应用地道，注意将复杂方法溯源至常理、常识，对一个方法要解决的问题与解决问题的逻辑思路、前提条件，存在的障碍进行全面介绍，引导读者进入每种方法实际应用时的情景设定：比较重视交代方法的适用场合、变量类型和量纲、数据基础，后续动作；尤其重视不同方法间以及同一类方法内部的子方法间的逻辑联系，以及在介绍经典方法的同时，自然、平滑地引入适合处理大数据分析的方法。

图书在版编目(CIP)数据

多元统计分析/杜子芳著. —北京：清华大学出版社，2016
(应用统计工程前沿丛书)
ISBN 978-7-302-44892-1

Ⅰ. ①多…　Ⅱ. ①杜…　Ⅲ. ①多元分析－统计分析　Ⅳ. ①O212.4

中国版本图书馆 CIP 数据核字(2016)第 201667 号

责任编辑：刘　颖
封面设计：傅瑞学
责任校对：王淑云
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：三河市君旺印务有限公司
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：170mm×230mm　　**印　　张**：20　　**字　　数**：380 千字
版　　次：2016 年 10 月第 1 版　　**印　　次**：2016 年 10 月第1次印刷
印　　数：1～2000
定　　价：49.00 元

产品编号：066381-01

“应用统计工程前沿丛书”
编　委　会

前言

公元2009年，受时任教育部应用统计科学研究中心主任赵彦云教授的鼓励，作者作为第一负责人申请了一个名为“多元统计分析及其应用的统计理论研究”的教育部人文社会科学重点研究基地重大项目，并幸运地获得了批准，项目编号为2009JJD910240，研究成果约定为一本专著。本书就是这一项目的主要成果之一，这其中还凝结了近20余年教学与应用等方面的经验：其中大部分内容在中国人民大学的本科生和研究生课堂讲授过，涉及的所有方法都在研究课题或咨询项目中有过实际应用。

经过几年的努力，期间几次延宕，现在这部专著终于要出版了。值此付梓之际，感慨良多。作者志大才疏，兼生性疏懒，倘若不是受到一些专家朋友的鼓励（如暨南大学的刘建平教授对多元统计框架给予了肯定；首都经贸大学的纪宏教授2000年前后与本人的讨论启发了本书聚类分析部分的研究；北方工业大学的李从珠教授则把他的判别分析的著作供我参考；中国科学院的冯士雍研究员和中国标准化研究院的肖惠总工程师在20世纪80年代中国人体尺寸数据案例上极具价值的慷慨相助；而北京航空航天大学的王惠文教授对本书部分内容的称许令我受宠若惊）和我的学生尤其是硕士、博士们的帮助（刘东硕士最早、杨进硕士继之帮我验证了方差分析与联合分析结论一致的想法；刘亚文博士、郑坤硕士验证了判别分析与logistic回归结果的一致；徐一丁博士验证了聚类距离计算的不一致；硕士生王维和于焕杰演算了大部分例题），成书恐怕遥遥无期，而那些已毕业的学生关于“何时见到书”的追问更是极其有力的鞭策，往往使我羞愧万分，不得不暗下决心，务必有个交代。学生之中，刘亚文和于焕杰两位出力最多，前者缜密细致，每每发现许多学理细节方面的意外错误；后者聪明勤奋，常常能够以令人吃惊的速度完成许多计算、绘图、编辑、排版等繁琐无比的工作。此外，我们项目组的主要成员广东商学院林海明教授，在项目研究过程中发表了许多很好的论文，但由于本书定位为专著，出于尊重知识产权的考虑，在此领域其诸多贡献并未体现于本书。借此机会，作者要向所有贡献者致以最诚挚的谢意！

本书的内容在招标申请书已经列明，除个别的如对应分析限于篇幅未予介绍外，令人欣慰的是其余的都完全兑现了，有些不在计划的如某些适用于大数据的分析与挖掘的内容也有涉猎，尽管作者认为，大数据的统计分析方法既不复杂，也不高级，但有关算法相对说实居于关键地位。敝帚自珍，高明不敢轻言，但学术上差不多毕生的心力融会于此，独到处是有些的，相信会对读者在透彻的理解与地道的应用方面有所裨益。现将项目申请书中关于内容的说明转录如下，兹以为序。

“由于现实问题往往比较复杂，并非一两个变量所能概括反映，多元统计分析本应是统计数据处理的最适合的手段，但以往因没有计算机或计算机不够普及，极大限制了多元统计分析的应用，以至于多元统计分析长时期内被束之高阁，虽然有些多元统计分析方法如因子分析早在1904年就已提出，而真正被广泛应用却是20世纪80年代以后的事情。在我国，多元统计分析的普及年代更晚。有记载许宝禄先生20世纪50年代中期曾说当时从事数理统计专业的连他本人在内不超过12人；从研究生课程里抽出一部分多元统计分析内容纳入本科生课程在中国人民大学统计学院其历史也仅有10年左右；时至今日，在国内的大多数高校里，作为三大多元统计之一的回归分析仍在多元统计分析课程之外独立地充当一门课程。

改革开放以后，伴随着我国整个教育的进步，统计教育的改善也堪称突飞猛进，大学里设置统计学院系的越来越多，开出统计课程的越来越多；中小学里统计知识介绍甚至超越概率论进入了抽样与推断统计的领域。一方面得益于这样的大环境和计算机与统计软件如SPSS、SAS和STATSTICA等的普及，多元统计分析中纳入教学内容的方法日益增加。另一方面，多元统计分析的应用领域，统计科学对科研经济社会建设的全面渗透而日趋扩展，从地质学、生物学、医学、心理学迅速扩展到经济学、社会学、营销学、管理学和教育学等诸多领域；应用频率也大幅地日渐增加，发表在期刊上的多元统计分析文章明显增多，具体信息见表1。

表1　部分多元方法在CPCI(原ISI proceedings)检索的文献数及学科分类

	判别	logistic	联合	方差	因子	主成分	聚类	对应
合计	26668	80659	1706	100000+	100000+	17379	85660	14469
数学	38.7	47.3	17.8	64.4	17.3	28.8	26.4	18.7
行为科学	21.2	22.9	22.7	32.1	10.8	18.4	7.6	9.7
心理学	19.7	20.7	24.2	28.6	8.7	17.3	7.2	8.5
神经科学	12.7	12.1	5.3	26.5	10.8	10.4	6.6	7.4
生物化学	14.3	15.5		26.2	47.8	12.9	35.5	15.5
生理学	6.8			15.4	10.1	6.6		
遗传学	9.9	8.2		14.3	33.2	9.4	35.2	11.2
老年病学	12.2	31.9	6.8	14.2	12.9	5.1		
儿科学	11.7	22.7	5.2	13.5	7.5	6.4		
心血管学	7.4	18.1		11.9	12.5			
免疫学	5.1	11.3	3.3	8.6	22.4		9.4	
健康护理	5.8	20.1	15.2	7.4				

续表

	判别	logistic	联合	方差	因子	主成分	聚类	对应
环境生态学	9.0			7.3		14.4	10.1	26.1
细胞生物学				6.6	21.5		9.1	
肿瘤学	7.2	11.8			16.2			
计算机科学	15.9		11.8		6.6	12.6	9.5	13.0
人口学		19.4						
商业与经济			43.0					
工程学	6.2		11.1			8.4	0.0	10.1
化学	6.4				7.3	11.9	13.0	6.4
农学	4.9		6.2		0.0	8.5	6.5	7.4
微生物学					6.4		17.4	4.4
传染病学		12.9			6.4		9.3	
生物多样性								11.5
海洋及淡水生物学								10.2
植物科学								10.1

然而，由于多元统计方法的出现与实际应用间隔太久，在我国其大规模应用也就是近几年的事情，对内容的掌握尚属生疏与实用场合的明显增多同时交汇，客观上难免造成一些生吞活剥与误用滥用现象的出现。除此之外，多元统计分析的很多方法都是其他学科而非统计学科的人士所提出，例如回归分析是遗传学家所发现，因子分析是心理学家所开创，联合分析拥有心理学和营销学的血缘，而分层分析则有教育学的基因，这些外来‘物种’极大丰富了统计学的内容，促进了统计学的应用。但众多原本起于其他学科的方法在融入统计学大冶炉的过程中，难免因带有浓厚的原来学科的色彩而有些水土不服，术语庞杂混乱，原理的统计学基础薄弱，因此当我们今天从统计学的视角重新审视多元统计分析的构成时，可以发现其中存在着一些明显的问题，以下是几例。

1. 聚类分析、回归分析和判别分析并称三大多元统计方法，其使用价值之大可见一斑，但迄今仍未解决其理论基础问题，致使这一方法是否应归到统计学科尚有疑问。

2. 距离判别、费歇判别和贝叶斯判别三种判别之间原理上存在怎样的联系？孰优孰劣？各自的使用场合是什么？

3. 联合分析与方差分析同属自变量为分类型变量而因变量为数值型变量的分析方法，同样使用 F 统计量作为判定依据，要达到的目标——衡量因素的重要性和优选

因素的水平，也是一样的。两者之间原理上存在怎样的联系？孰优孰劣？各自的使用场合是什么？

4. logistic 回归分析与判别分析同属自变量为数值型变量而因变量为分类型变量的分析方法，要达到的目标也是一致的，但 logistic 回归分析多被看成回归分析的推广，这一方法与判别分析更近的‘亲缘’却不被公认，两者之间原理上存在怎样的联系？孰优孰劣？各自的使用场合是什么？

5. 主成分分析被认为是求得因子的方法之一，与其他求取因子方法相比孰优孰劣？使用场合上有何区别？

6. 对应分析原本属于列联分析的复杂情形(因素水平较多)，又被看做因子分析的深入，但列联分析的‘自变量’和‘因变量’都是分类型变量，而因子限于处理数值型变量，一个数据如何既是数值型的又是分类型的？显然存在着明显的矛盾。

所有这些问题国内外文献均无完整明确的叙述，教科书里对此也不提及。但毫无疑问，这些问题的解决将有助于明确上述方法本身的统计学理论背景，廓清方法间的联系是包含的还是并列的、抑或是递进的，使学生和使用者从数学上的‘在这些条件下，方法甲与方法乙等价’的模糊叙述中解脱出来，以清晰的逻辑和语言阐明在特定条件下究竟何种方法更优，或者倒过来说各个方法的适用场合怎样，从而促进对多元统计分析的理论推广，防止对各种多元统计分析方法的误用滥用。鉴于多元统计分析是统计数据处理最重要的工具，同时是统计学应用最为广泛的一个分支，因此这项研究的理论意义与实际价值都是不难理解的，在降低多元统计分析的学习成本和误用概率方面尤其具有明显的、巨大的促进作用。

本项研究的目标定为完成一篇对上述问题有很好答案的、确有新意的专著，期待可以成为全国统计学科发展与研究生培养的核心参考文献之一。根据我们对有关课题的兴趣与经验的多年积淀，我们有信心使这项研究成功完成，也有信心这项研究结果可对多元统计分析的教学与科研有所助益。”

限于作者水平，本书难免存在一孔之见或错漏舛误，敬希同行不吝赐教，哪怕是严厉的理性批判，以使本书日后能够渐臻完善，以飨读者。

杜子芳

2016 年 5 月

目 录

第1章　多元统计描述

——反映多个随机变量的分布（联合、边缘、条件）及分布特征

现实中，出于特定的目的或目标，针对具体的任务或问题，在一些场合，人们只关注一个变量，如物理老师只关注学生们物理考试的成绩，数学老师则只关注学生们数学考试的成绩，高考只有总成绩（相当于一个复合变量）一条分数线，买衬衣的只要说出一个号型，个人所得税征税只与个人所得挂钩等。在另一些场合，人们则关注多个变量，如家长不仅关注子女的数学考试成绩，还关注子女其他各门考试成绩；而考研则除有总成绩的分数线外，各门课程还各有一条分数线；跳水、体操等缺乏客观标准的国际体育比赛，运动员的比赛成绩取决于几个裁判员而非一个裁判员的打分；服装制作需注意人体许多部位的尺寸；购买汽车需要考虑品牌、安全性、舒适性、颜色、排量、售后、价格、按揭等许多因素，购买住房则要考虑地段、房龄、朝向、层数、层高、居室数、卫生间数，客厅大小、厨房大小、房间布局，交通条件、学校、医院、商店等配套设施；经济学认为生产率决定于人力、土地、资源、技术、基础设施等生产要素的合理配置；一国的国家竞争力评价因素动辄几十个；美国的穆迪、欧洲的惠誉、中国的大公等信用评级机构所依据的指标虽然有所不同但其数目都很多。值得注意的是，需要关注许多变量的场合要远多于只需关注一个变量的场合。与此相类似，进行统计分析时，一些场合只关注一个变量的统计特性，更多的场合则要关注许多变量的统计特性，这便是多元统计分析产生与存在的实际背景。

1.1 多元统计分析的内容

多元统计分析中的“统计分析”，内容包含两个方面，阐述性分析与因果性分析，其中阐述性分析有描述性分析与推断性分析之分。

描述性分析反映或表达随机变量的分布及其分布特征，分为两条平行的线，第一条线**是关于总体分布或总体分布特征的描述**，第二条线**是关于样本分布或样本分布特征的描述**。从数据的来源说，第一条线**对应的是全面调查**[①]，第二条线**对应的是严格定义下的抽样调查（概率抽样）**。至于**推断性分析**，也有两条线，第一条线**是根据大数定理尤其是格列文科（Glivenko）定理，以样本分布估计总体分布**；第二条线则**是借由抽样分布**[②]**的中介作用，根据中心极限定理，解决如何用样本分布特征来估计总体分布特征的问题，部分或全部地实现由样本分布特征推断总体分布**

① 一些场合下，人们充分利用抽象的想象，得到理论上的各种常见分布，如正态分布、二项分布等，这些种类的随机变量分布很多与重复试验及所谓无限总体有关，实施完全调查是不可能的。

② 在抽样理论中，抽样分布是指总体规模为 N，样本规模为 n，由特定的抽样方式（如分层、系统、整群、简单随机抽样等）所决定的所有可能样本（如在简单随机抽样中的所有可能样本的数目为 C_N^n 个）的样本均值所形成的分布。

的目标。

因果性分析即**解释性分析**的内容，其实与数学中的函数概念比较近似，研究的是如何**使用一些随机变量或非随机变量去解释其他随机变量**，在这种解释过程中，往往要引用阐述性分析的结论或以阐述性分析为基础，所以阐述性分析与解释性分析的关系一如哲学中的本体论与关系论的关系，不同而相辅相成。**阐述性分析的基础是随机变量的分布**；而**解释性分析的基础是条件分布**，彼此明显不同，联系却也一目了然。

多元统计分析之“多元”是相对于“一元”而言的，尽管一元统计分析的说法似乎从来未见诸于任何文献中。多元统计分析之“多元”，乃多个变量之意。“多元”之“多”，不止一个也，可二可三可四……究竟多少，取决于客观允许与主观需要之交集。“多元”之“元”意为变量也。这里的变量，在阐述性分析中皆为随机变量，而在解释性分析中，因变量概为随机变量，自变量则可能为随机变量，也可能为非随机变量（所谓确定性变量），还可能为两者之混合，即自变量中同时存在随机变量与确定性变量两种不同类型的变量。在统计学的不同领域里，变量有着各种不同的别名，如“因素”、“因子”、“条件”、“解释变量”和“控制变量”以及“辅助变量”，其实都是自变量的别称；而“应变量”、“效用”、“主变量”等则是因变量的别称。以“元”来称变量，并不会作为名词独立使用。“元”与“多”或“一”的数词结合用作修饰词时才会被赋予变量简称的含义。然而在多元统计分析中，“多元”所表示的多个变量一般是同质的或同类的。在阐述性分析场合，不分自变量与因变量，所有变量自然地皆属同类；而在解释性分析场合，一般是指自变量，但某些学者对多元回归分析[①]的定义则是指多个因变量。不过无论如何，可以肯定的是，“元”从来不会同时混指因变量的个数和自变量的个数。多元与高维几乎同义。

前面所述，很有些琐碎的八股式题解的意味，但正如“论语”中的“名不正则言不顺”这句“名言”所陈述的“至理”[②]，在细述多元统计分析之前，我们只有先将“多元统计分析”正名，而后的具体分析内容的介绍才会“言顺”。事实上，当下统计学界普遍存在着学科界限不清，术语混乱，重视方法论而忽视本体论，常以某个领域覆盖其他领域等诸多弊病。这种混乱而糟糕的景况，既反映了统计方法来自于五湖四海、各行各业的历史，又是统计学界迄今仍然漠视统计教育日趋普及的现实，

---

① Rechard A. Johnson 在《实用多元统计分析》中将多个自变量的回归模型称为多重回归，多个因变量的回归模型称为多元回归。

② 其中一个明显的例子，数理统计的基础是抽样，然而数理统计中的抽样都是无限总体抽样，或重复抽样，但这种抽样，实际上少有场合采用，且与抽样论中所讲的抽样完全不同。

依旧囿于自我狭窄的领域坐井观天、夜郎自大、党同伐异而又故步自封、墨守成规的真实写照。其结果是在一级学科或大数据的虚假繁荣气氛里，既未获得应有的尊重，也未得到足够的学科收益。

大致而论，在整个统计学的内容框架里，多元统计分析的位置大抵如图 1.1 所示。

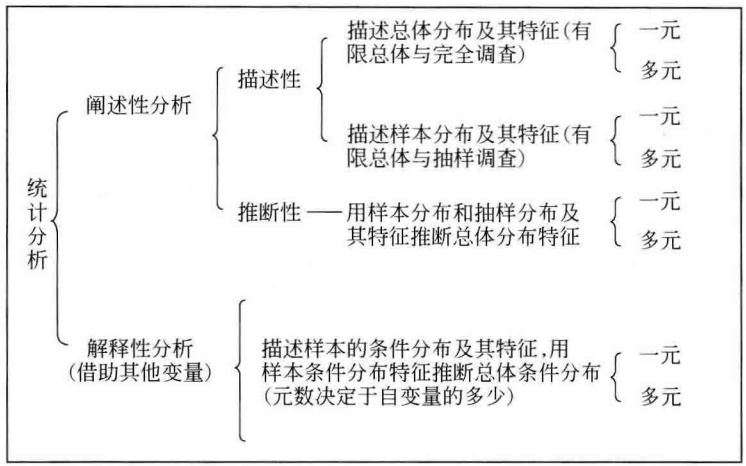

图 1.1　统计分析内容框架图

上述框架同时决定了本书的基本框架安排，先是描述性分析，再是推断性分析，然后是解释性分析(解释性分析乃本书之重点)。在描述性分析中，首先依多个变量间关系的疏密程度进行平行描述与整体描述，其次在具体描述时依变量尺度将变量区分为分类变量和数值变量进行展开，并考虑多个变量的尺度结构：纯粹的单一尺度，二尺度的混合等。在推断性分析中，首先依多个变量间关系的疏密程度进行平行推断与整体推断(依据单一分布还是联合分布)，其次在具体推断时充分考虑分类变量、顺序变量和数值变量的特点。在解释性分析中，内容分两条线展开，一是将样品之间的关系与变量之间的关系分开讨论；二是将变量区分为因变量与自变量，并将三种尺度的变量归并为分类变量与数值变量两种尺度，然后按一定顺序对各类因变量与自变量的可能组合分别进行介绍。

1.2　数据及其来源

从应用的角度看，多元统计分析属于数据分析的范畴，其分析对象是形如表 1.1 的数据阵。

表 1.1　多元统计分析数据阵

样品 \ 变量	$\boldsymbol{X}_1$	…	$\boldsymbol{X}_j$	…	$\boldsymbol{X}_p$
$X_{(1)}$	x_{11}	…	x_{1j}	…	x_{1p}
⋮	⋮		⋮		⋮
$X_{(i)}$	x_{i1}	…	x_{ij}	…	x_{ip}
⋮	⋮		⋮		⋮
$X_{(n)}$	x_{n1}	…	x_{nj}	…	x_{np}

这样的数据阵是调查或观测的终点和工作结果，又是多元统计分析的起点和工作基础。表 1.1 的最上面一行内容是变量(字段)的名称，其余的行，名叫样品，如第一行是第一个样品，第二行是第二个样品，以此类推。样品是样本的单元，又称个体或实验单元，其内容是该样品中所有变量的变量值，故又称记录；表 1.1 中最左一列内容是记录的号码，其余的列，名叫变量(计算机有关学科称字段)。行与列的交叉处称为单元格或表项，表项里所填内容为相应样品与相应变量的变量值(变量值，有时称观测值，也经常被称作数据)。最左边一列之外的表的总列数等于变量数，记为 p；最上面一行之外的表的总行数等于样品数，记为 n；表项的总数即变量值的个数为 $n\times p$。数据量一般用 $n\times p$ 与每个变量值的平均位数或字节数的乘积来表示。

在多元统计分析中，为了行文方便，通常将表 1.1 行号和列名省略，内化为以 X_{ij} 为元素的数据矩阵 $\boldsymbol{X}$，$\boldsymbol{X}$ 一般称为数据阵，有时也被称为数据集，纯由表项及其填充值构成。

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

其中 X_{ij} 表示数据表第 i 行第 j 列的数据，即变量 X_j 的第 i 个样品对应的变量值(字段值、观测值或记录)。

例 1.1(某大学 13 个学生的学分统计)　研究者调查了某大学一个班级的学分记录，其中的 13 个学生的学分记录结果列于表 1.2。

表 1.2　大学一个班级部分学生的学分记录

编号	性别	民族	学院	专业	已获得学分	本学期修读学分	预计毕业总学分
1	男	汉族	信息工程学院	计算机技术	144	18	162
2	男	汉族	信息工程学院	计算机技术	162	16	178