教育部高等学校统计学类专业
教学指导委员会推荐用书

博雅·21世纪统计学规划教材

Applied Multivariate Statistical Analysis

应用多元统计分析

朱建平　主编

北京大学出版社
PEKING UNIVERSITY PRESS

教育部高等学校统计学类专业
教学指导委员会推荐用书

博雅·21世纪统计学规划教材

Applied Multivariate Statistical Analysis

应用多元统计分析

朱建平 主编

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目 (CIP) 数据

应用多元统计分析 / 朱建平主编. — 北京 : 北京大学出版社, 2017.8
(21 世纪统计学规划教材)
ISBN 978-7-301-28505-3

Ⅰ. ①应… Ⅱ. ①朱… Ⅲ. ①多元分析—统计分析—高等学校—教材 Ⅳ. ① O212.4

中国版本图书馆 CIP 数据核字 (2017) 第 166793 号

书　　名	应用多元统计分析 Yingyong Duoyuan Tongji Fenxi
著作责任者	朱建平　主编
责任编辑	潘丽娜
标准书号	ISBN 978-7-301-28505-3
出版发行	北京大学出版社
地　　址	北京市海淀区成府路 205 号　100871
网　　址	http://www.pup.cn　新浪微博 : @ 北京大学出版社
电子信箱	zpup@pup.cn
电　　话	邮购部 62752015　发行部 62750672　编辑部 62752021
印 刷 者	北京大学印刷厂
经 销 者	新华书店
	787 毫米 × 980 毫米　16 开本　15.75 印张　330 千字 2017 年 8 月第 1 版　2017 年 8 月第 1 次印刷
定　　价	39.00 元

“21世纪统计学规划教材”
编委会

前　言

随着互联网 (Internet) 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, 这使得各企事业单位生成、收集、存储和处理数据的能力大大提高, 数据量与日俱增, 大量复杂信息层出不穷. 大数据时代已经到来, 数据产生的经济效益愈发凸显. 然而, 大量信息在给人们带来方便的同时也带来一系列问题. 比如: 信息量过大, 超过了人们掌握、消化的能力; 一些信息真伪难辨, 从而给信息的正确应用带来困难; 信息组织形式的不一致性导致难以对信息进行有效的统一处理; 在公共的网络环境之中, 用户隐私的保护, 不仅需要法律支持, 更需要社会公认的数据标准和规范; 等等. 因此, 我们将面临着复杂数据的处理问题, 特别是研究客观事物中多个变量 (或多个因素) 之间相互依赖的统计规律性, 它的重要理论基础之一是多元统计分析. 多元统计分析是统计学中一个非常重要的分支, 具有很广泛的应用性, 它在自然科学、社会科学和经济学等各领域中得到了越来越广泛的应用, 是一套非常有用的数据处理方法. 为了能更好地从统计学的角度解决这些问题, 我们组织编写了《应用多元统计分析》这本书, 并且作为 "教育部统计学类专业教学指导委员会" 推荐系列教材之一.

按照国家级教材规划要求, 本书的编写力求以统计思想为主线, 以 SPSS 软件为工具, 深入浅出地介绍各种多元统计方法的应用. 其基本框架是: 第一章为概述, 第二章和第三章介绍多元正态总体的参数估计和假设检验、多元线性回归模型, 第四章至第九章介绍常用的多元统计方法, 这些方法包括聚类分析、判别分析、主成分分析、因子分析、相应分析、典型相关分析, 第十章介绍多变量的可视化分析等.

在本书的编写过程中, 我们根据统计专业的要求, 突出以下特点:

第一, 把握统计实质, 贯穿统计思想. 注重统计思想的讲述, 在多元统计方法的应用上把握实质, 从实际问题入手, 在不失严谨的前提下, 淡化统计方法本身的数学推导, 体现统计学的实用性.

第二, 应用 SPSS 软件, 实现统计计算. 根据多元统计固有的特点, 我们选用在我国广泛流行的 SPSS 软件作为计算工具. 在每一章的最后, 都要讲述所介绍的多元统计方法在 SPSS 软件中的实现. 这样将 SPSS 软件的学习和案例分析有机结合, 不仅使得学生在实践运用中学习了 SPSS 软件的操作方法, 而且还使学生对多元统计分析的意义有深入的体会.

第三, 加强统计理论, 完成统计实践. 根据实际介绍的统计方法, 我们将编写的习题分为两类: 一类是继续巩固和加强统计理论和方法, 包括基本概念和基本思路训练的习题; 另一类是针对实际问题, 培养学生结合统计方法独立解决实际问题的能力和素质的习题.

为了提高学生的学习兴趣和学习的效率, 考虑到不同的使用对象和教学特点, 对部分内容可根据实际情况进行选讲.

本书第一、四、五、九章由厦门大学朱建平教授编写, 第六、七章由广东财经大学林海明教授编写, 第八章由厦门大学刘云霞副教授编写, 第二、三章由东北石油大学辛华副教授编写, 第十章由东北石油大学任晓萍助理教授编写. 本书由朱建平教授担任主编并进行统稿和总纂.

本书在编写和出版过程中, 得到了厦门大学数据挖掘研究中心、厦门大学管理学院 MBA 中心、广东财经大学经济学院、东北石油大学数学与统计学院、浙江工商大学现代商贸流通体系协同创新中心和北京大学出版社的支持, 潘丽娜编辑为本书的组稿、编辑做了大量的工作, 在此表示衷心感谢! 编写一本好的教材并不容易, 尽管我们努力想奉献给读者一本满意的书, 但仍有达不到读者各方面要求的地方. 书中难免有疏漏或错误之处, 恳请读者多提宝贵意见, 以便今后进一步修改与完善.

本书的编写得到了国家社会科学基金重大项目 “大数据与统计学理论的发展研究” (13&2D148) 的资助.

朱建平

2017 年 7 月于厦门珍珠湾花园

目　　录

第一章　多元统计分析概述

第一节　引　　言

多元统计分析是运用数理统计方法来研究解决多指标问题的理论和方法. 近 30 年来, 随着计算机应用技术的发展和科研生产的迫切需要, 多元统计分析技术被广泛地应用于地质、气象、水文、医学、工业、农业和经济等许多领域, 已经成为解决实际问题的有效方法. 由于计算机处理技术发生着日新月异的变化, 人们处理大规模复杂数据的能力日益增强, 从大规模复杂数据中提取有价值的信息能力也日益提高, 人们将会迅速进入大数据时代. 大数据时代, 不仅会带来人类自然科学技术和人文社会科学的发展变革, 还会给人们的生活和工作方式带来焕然一新的变化. 大数据时代的到来, 给多元统计分析理论的发展和方法的应用带来了发展壮大机会的同时, 也使其面临着重大的挑战.

多元统计分析起源于 20 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, 使多元统计分析在理论上得到了迅速的发展. 20 世纪 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. 20 世纪 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进一步扩大. 20 世纪 70 年代初期在我国才受到各个领域的极大关注, 并在多元统计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. 20 世纪 80 年代初期数据在不同信息管理系统之间的共享使数据接口的标准化越来越得到强调, 为数据的共享和交流提供了捷径; 80 年代后期, 互联网的概念兴起、"普适计算" (ubiquitous computing) 理论的实现以及传感器对信息自动采集、传递和计算成为现实, 为数据爆炸式增长提供了平台, 为多元统计理论和方法的应用开辟了新的领域. 20 世纪 90 年代, 由于数据驱动, 数据呈指数增长, 企业界和学术界也不断对此现象及其意义进行探讨, 为大数据概念的广泛传播提供了途径. 进入 21 世纪以来, 世界上许多国家开始关注大数据的发展和应用, 一些学者和专家发起了关于大数据研究和应用的深入探讨, 例如 Vikor Mayer-Schǒnberger and Kenneth Cukier 所著的《大数据时代》等, 对大数据促进人们生活、工作与思维的变革奠定了基础. 在此期间, 多元统计与人工智能和数据库技术相结

合, 将通过互联网和物联网在经济、商业、金融、天文等行业得到更广泛的应用.

为了让人们更好地、较为系统地掌握多元统计分析的理论与方法, 本书重点介绍多元正态总体的参数估计和假设检验、多元线性回归模型以及常用的统计方法. 这些方法包括判别分析、聚类分析、主成分分析、因子分析、对应分析、典型相关分析及多变量的可视化分析等. 与此同时, 我们将利用在我国广泛流行的 SPSS 统计软件来实现实证分析, 做到 "在理论的学习中体会应用, 在应用的分析中加深理论".

第二节 大数据时代与大数据

大数据是信息科技高速发展的产物, 如果要全面深入理解大数据的概念, 必须理解大数据产生的时代背景, 然后根据大数据时代背景理解大数据概念.

一、"大数据时代" 背景介绍

格雷布林克 (Grobelink. M) 在《纽约时报》2012 年 2 月的一篇专栏中称, "大数据时代" 已经降临, 在商业、经济及其他领域中, 管理者决策越来越依靠数据分析, 而不是依靠经验和直觉. "大数据" 概念之所以被炒得如火如荼, 是因为大数据时代已经到来.

如果说 19 世纪以蒸汽机为主导的产业革命时代终结了传统的手工劳动为主的生产方式, 并从而推动了人类社会生产力的变革; 那么 20 世纪以计算机为主导的技术革命则方便了人们的生活, 并推动人类生活方式发生翻天覆地的变化. 我们认为, 随着计算机互联网、移动互联网、物联网、车联网的大众化和博客、论坛、微信等网络交流方式的日益普及, 数据资料的增长正发生着 "秒新分异" 的变化, 大数据时代已经到来毋庸置疑. 据不完全, 一天之中, 互联网产生的全部数据可以刻满 1.68 亿张 DVD. 国际数据公司 (IDC) 的研究结果表明, 2008 年全球产生的数据量为 0.49ZB (1024EB=1ZB, 1024PB=1EB, 1024TB=1PB, 1024GB=1TB), 2009 年的数据量为 0.8ZB, 2010 年增长为 1.2ZB, 2011 年的数量量高达 1.82ZB, 相当于全球每人产生 200GB 以上的数据, 而到 2012 年为止, 人类生产的所有印刷材料的数据量是 200PB, 全人类历史上所有语言资料积累的数据量大约是 5EB. 哈佛大学社会学教授加里·金说: "大数据这是一场革命, 庞大的数据资源使得各个领域开始了量化进程, 无论学术界、商界还是政府, 所有领域都将开始这种进程." 在大数据时代, 因为等同于数据的知识随处可寻, 对数据的处理和分析才显得难能可贵. 因此, 在大数据时代, 能从纷繁芜杂的数据中提取有价值的知识才是创造价值的源泉.

我们可以这样来定义大数据时代, 大数据时代是建立在通过互联网、物联网等现代网络渠道广泛大量数据资源收集基础上的数据存储、价值提炼、智能处理和展示的信息时代. 在这个时代, 人们几乎能够从任何数据中获得可转换为推动人们生活方式变化的有价值的知识. 大数据时代的基本特征主要体现在以下几个方面:

(1) 社会性. 在大数据时代, 从社会角度看, 世界范围的计算机联网使越来越多的领域以数据流通取代产品流通, 将生产演变成服务, 将工业劳动演变成信息劳动. 信息劳动的产品不需要离开它的原始占有者就能够被买卖和交换, 这类产品能够通过计算机网络大量复制和分配而不需要额外增加费用, 其价值增加是通过知识而不是手工劳动来实现的, 而实现这一价值的主要工具就是计算机软件.

(2) 广泛性. 在大数据时代, 随着互联网技术的迅速崛起与普及, 计算机技术不仅促进自然科学和人文社会科学各个领域的发展, 而且全面融入了人们的社会生活中, 人们在不同领域采集到的数据量之大, 达到了前所未有的程度. 同时, 数据的产生、存储和处理方式发生了革命性的变化, 人们的工作和生活基本上都可以用数字化表示, 在一定程度上改变了人们的工作和生活方式.

(3) 公开性. 大数据时代展示了从信息公开运动到数据技术演化的多维画卷. 在大数据时代会有越来越多的数据被开放, 被交叉使用. 在这个过程中, 虽然考虑对于用户隐私的保护, 但是大数据必然产生于一个开放的、公共的网络环境之中. 这种公开性和公共性的实现取决于若干个网络开放平台或云计算服务以及一系列受到法律支持或社会公认的数据标准和规范.

(4) 动态性. 人们借助计算机通过互联网进入大数据时代, 充分体现了大数据是基于互联网的及时动态数据, 而不是历史的或严格控制环境下产生的内容. 由于数据资料可以随时随地产生, 因此, 不仅数据资料的收集具有动态性, 而且数据存储技术、数据处理技术也随时更新, 即处理数据的工具也具有动态性.

二、“大数据”的定义

在大数据时代, 数据引领人们生活, 引导商业变革和技术创新. 从大数据的时代背景来看, 我们可以把大数据作为研究对象, 从数据本身和处理数据的技术两个思路理解大数据, 这样理解大数据就有狭义和广义之分: 狭义的大数据是指数据的结构形式和规模, 是从数据的字面意义理解; 广义的大数据不仅包括数据的结构形式和数据的规模, 还包括处理数据的技术.

狭义角度的大数据, 是指计量起始单位至少是 PB, EB 或 ZB 的数据规模, 其不仅包括结构化数据, 还包括半结构化数据和非结构化数据. 我们应该从横向和纵向两个维度解读大数据: 横向是指数据规模, 从这个角度来讲, 大数据等同于海量数据, 指大数据包含的数据规模巨大; 纵向是指数据结构形式, 从这个角度来说, 大数据不仅包含结构化数据, 更多的是指半结构化数据和非结构化数据, 大数据包含的数据形式多样. 大数据时代, 由于有 90% 的信息和知识在 “结构化” 数据世界之外, 因此, 人们通常认为大数据的分析对象为半结构化数据和非结构化数据.

此外, 大数据时代的战略意义不仅在于掌握庞大的数据信息, 而且在于如何处理数据. 这

就需要从数据处理技术的角度理解大数据.

广义角度的大数据, 不仅包含大数据结构形式和规模, 还泛指大数据的处理技术. 大数据的处理技术是指能够从不断更新增长、有价值信息转瞬即逝的大数据中抓取有价值信息的能力. 在大数据时代, 针对小数据处理的传统技术可能不再适用. 这样, 就产生了专门针对大数据的处理技术, 大数据的处理技术也衍生为大数据的代名词. 这就意味着, 广义的大数据不仅包括数据的结构形式和规模, 还包括处理数据的技术. 此时, 大数据不仅是指数据本身, 还指处理数据的能力.

不管从广义的角度, 还是从狭义的角度来看, 大数据的核心是数据, 而数据是统计研究的对象, 从大数据中寻找有价值的信息关键在于对数据进行正确的统计分析. 因此, 鉴定 "大数据" 应该在现有数据处理技术水平的基础上引入统计学的思想.

从统计学科与计算机科学性质出发, 我们可以这样来定义 "大数据": 大数据指那些超过传统数据系统处理能力、超越经典统计思想研究范围、不借用网络无法用主流软件工具及技术进行单机分析的复杂数据的集合, 对于这一数据集合, 在一定的条件下和合理的时间内, 可以通过现代计算机技术和创新统计方法, 有目的地进行设计、获取、管理、分析, 揭示隐藏在其中的有价值的模式和知识.

根据大数据的概念和其时代属性, 我们认为大数据的基本特征主要体现在以下四个方面:

(1) 大量性. 这是指大数据的数据量巨大. 在大数据时代, 高度发达的网络技术和承载数据资料的个人电脑、手机、平板电脑等网络工具的普及, 数据资料的来源范围在不断拓展, 人类获得数据资料在不断更改数据的计量单位. 数据的计量单位从 PB 到 EB 到 ZB, 反映了数据量增长质的飞跃. 据统计, 截止 2012 年底, 全球智能手机用户 13 亿, 仅智能手机每月产生的数据量就有 500MB, 每个月移动数据流量有 1.3EB 之巨.

(2) 多样性. 这是指数据类型繁多, 大数据不仅包括以文本资料为主的结构化数据, 还包括网络日志、音频、视频、图片、地理位置等半结构或非结构化的数据资料. 多样化的数据产生的原因主要有两个方面: 一是由于非结构化数据资料的广泛存在; 二是挖掘价值信息的需要, 传统的数据处理对象是结构式的, 我们从数据的大小多少来感受对象的特征, 但这远远不够具体. 很多时候, 我们希望了解得更多, 除了了解对象的数量特征外, 我们还希望了解对象的颜色、形状、位置, 甚至是人物心理活动, 等等, 这些是传统数据很难描述的. 为了满足人们对数据分析深层次的需要, 同时由于大数据时代对音频、视频或图片等数据资料处理技术不再是难题, 于是半结构化数据和非结构化数据也成为数据处理的对象.

(3) 价值性. 这指大数据价值巨大, 但价值密度低: 大数据中存在反映人们生产活动、商业活动和心理活动各方面极具价值的信息, 但由于大数据规模巨大, 数据在不断更新变化, 这些有价值的信息可能转瞬即逝. 一般来讲, 价值密度的高低与数据规模的大小成反比. 因此, 在大数据时代, 对数据的接收和处理思想都需要转变, 如何通过强大的机器算法更迅速地完

成数据的价值 “提纯” 成为目前大数据背景下亟待解决的难题.

(4) 高速性. 这指数据处理时效性高, 因为大数据有价值信息存在时间短, 要求能迅速有效地提取大量复杂数据中的有价值信息. 根据 IDC 的 “数字宇宙” 的报告, 预计到 2020 年, 全球数据使用量将达到 35.2ZB. 在如此海量的数据面前, 处理数据的效率关乎智能型企业的生死存亡.

毫无疑问, 由于计算机处理技术发生着日新月异的变化, 人们能处理大规模复杂数据的能力日益增强, 从大规模数据中提取有价值的信息能力也日益提高, 人们将会迅速进入大数据时代. 大数据时代, 不仅会带来人类自然科学技术和人文社会科学的发展变革, 还会给人们的生活和工作方式带来焕然一新的变化.

统计学是一门古老的学科, 已经有三百多年的历史, 在自然科学和人文社会科学的发展中起到了举足轻重的作用; 统计学又是一门生命力及其旺盛的学科, 她海纳百川又博采众长, 她随着各门具体学科的发展不断壮大自己. 毫不例外, 大数据时代的到来, 给统计学科带来了发展壮大机会的同时, 也使得统计学科面临着重大的挑战. 怎样深刻地认识和把握这一发展契机, 怎样更好地理解和应对这一重大挑战, 这就迫使我们对多变量统计分析的理论和方法进行学习和研究的基础上, 重新审视并提出适合现代数据分析的思想、理念与方法.

第三节 应 用 背 景

统计方法是科学研究的一种重要工具, 其应用颇为广泛. 特别地, 多元统计分析方法常常被应用于自然科学、社会科学等领域的问题中. 为了进一步体现多元统计分析方法的应用, 我们首先从宏观的角度认识统计学应用的背景, 然后从微观的角度显示多元统计分析应用的广泛性.

一、统计学的生命力在于应用

(一) 统计学产生于应用

从统计学的发展过程中可以看出统计学产生于应用, 在应用过程中发展, 它的生命力在于应用.

300 年前, 威廉 · 配第 (1623–1687) 写的《政治算术》, 从其研究方法看, 被认为是一本统计学著作. 政治算术学派的统计学家将统计方法应用于各自熟悉和感兴趣的研究领域, 都还是把其应用对象当作肯定性事物之间的联系来进行研究的. 他们确信, 事物现象存在着简单明了的数量关系, 需要用定性与定量的方法将这种关系 (规律) 揭示或描述. 使人们能够更具体、更真切地认识世界.

数理统计学派的奠基人凯特勒在统计学中引入了概率论, 把它应用于自然界和社会的许多方面, 从而为人们认识和说明不确定现象及其相互之间的联系开辟出了一条道路. 在自然

科学和社会科学的许多领域, 都留下凯特勒应用统计学研究的烙印. 自从凯特勒把概率论引入了应用中的统计学, 人们对客观世界的认识及描述更全面、更接近于实际了. 他在广泛应用拉普拉斯等人概率论中的正态曲线、误差法则、大数法则等成果的过程中, 为统计学增添了数理统计方法, 进而又扩展了统计学的应用范围.

在应用中对发展统计方法贡献显著的当推生物统计学派的戈尔登 (1822–1921)、皮尔逊 (1857–1936) 和农业实验学派的孟德尔 (1822–1884)、戈塞特 (1876–1937) 等. 戈尔登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. 将大量数据加以综合描述和比较, 从而能使他的遗传理论建立在比较精确的基础上, 为统计学引入了中位数、四分位数、分布、回归等极为重要的概念和方法. 皮尔逊在检验他老师戈尔登的 "祖先遗传法则" 和自然选择中 "淘汰" 对器官的相关及变异的影响中, 导入了复相关的概念和方法. 在讨论生物退化、反祖、遗传、随机交配等问题中, 展开了回归与相关的研究, 并提出以 χ^2 检验作为曲线配合适合度的一种量度的思想.

农业实验学派的孟德尔和戈塞特同样是在实验回答各自应用领域中出现的新要求、新课题, 发展了统计思想和统计分析方法. 孟德尔及其后继者贝特森等人创建的遗传试验手段, 比通过记录生命外部联系曲折反映事物内在本质的描述统计更加深刻. 他们运用推断的理论与实验的方法, 通常只用小样本来处理. 戈塞特的 t 分布与小样本思想更是在由于 "有些实验不能多次地进行", 从而 "必须根据极少数的事例 (小样本) 来判断实验结果的正确性" 的情况下产生的. 今天, 这些统计思想和分析推断方法已经成为了科学家们不可缺少的基本研究工具了.

近现代, 统计学已经空前广泛应用于最高级的运动形式 —— 社会. 其结果便是出现了一系列与其应用对象指导理论和其他相关学科交织在一起的边缘学科. 如在社会经济方面的投入产出经济学、经济计量学、统计预测学、统计决策学等. 在这些边缘学科中, 统计学与其应用对象结合更紧密、更自然. 这些学科的专家学者至少在两个或两个以上的专业领域里有比较深厚的学术造诣. 统计学的应用帮助他们在各自的应用领域中取得辉煌的成就.

可见, 统计学的发展一刻也离不开应用. 它在应用中诞生, 在应用中成熟、独立, 在应用中扩充自身的方法内容, 同时扩展了应用领域, 又在应用中与其他学科紧密结合形成新的边缘学科. 一部统计理论发展史同时又是一部应用统计发展史, 正因如此, 统计学的生命力在于应用.

(二) 理论研究为统计学的应用奠定了基础

统计理论问题的研究和应用研究从总体上说应该属于 "源" 和 "流" 的关系. 如果理论不成熟, 方法不完善, 统计应用研究也很难达到较高的水平. 因此, 充分发挥统计学的生命力,

必须建立在统计理论研究的基础之上.

从国际上看, 近十几年来, 统计分析技术的研究有了新的发展. 这些研究的总体特征是, 广泛吸收和融合相关学科的新理论, 不断开发应用新技术和新方法, 深化和丰富了统计学传统领域的理论与方法研究, 并拓展了统计研究的新领域. 这些都充分地体现了统计学强有利的生命力, 其具体表现在下面的三个方面:

第一, 统计学为计算机科学的发展发挥作用. 在计算机协助的电子通讯、网络创新、资源及信息统计中的统计软件等方面, 对统计信息搜集、存贮和传递中利用计算机提高工作效能, 建立统计信息时空结构有了新的发展. 在网络推断、统计软件包、统计建模中的计算机诊断方面, 提出了统计思想直接转化为计算机软件, 通过软件对统计过程实行控制的作用, 以及利用计算机程序识别模型、改善估计量性质的新方法. 这些研究成果使人们兴奋地看到计算机技术正在促使统计科研工作发生革命性变化. 在软件的质量评估上及统计程序和方法在软件可靠性检验等方面也有了新的发展.

第二, 统计理论与分析方法的新发展. 近年来, 统计方法成果丰硕, 反映了统计理论与分析方法在不断的发展中趋于成熟和完善. 在贝叶斯方法、非线性时间序列、多元分析、统计计算、线性模型、稳健估计、极值统计、混沌理论及统计检验等方面, 内容广泛而翔实, 可以归纳为三个方面:

(1) 理论上有新的开拓. 如应用混沌理论提出混沌动态系统、混沌似然分析; 引入数学中象分析、谱分析的方法, 探讨象分析中同步模型化的方法, 建立经验谱类函数的假设检验方法等.

(2) 不同的分析方法相互渗透、交叉结合运用, 衍生新的分析方法. 如马尔可夫链, 蒙特卡罗方法在叶贝斯似然计算中的应用, 参数估计方法的非参数校正, 状态空间模型与月份时间序列的结合运用.

(3) 借助现代计算机技术活跃新的研究领域. 在计算机技术迅速发展的带动下, 模拟计算理论和方法有了长足的发展, 这给非线性模型等因计算繁琐而沉闷多时的研究领域注入了新的活力, 提出了非线性结构方程模型的特征向量估计方法, 非线性回归中的截面有效性逼近, 带噪声的非线性时间序列的识别等富有见地的新思路. Logistic 模型、向量时间序列模型的研究也因计算技术的解决而不乏新成果.

第三, 统计调查方法与记述的创新. 调查方法是统计方法论的重要组成部分, 近年来, 在抽样理论与方法、抽样调查、实验设计方面十分关心如何改进调查技术、减少抽样误差等问题. 调查过程的综合管理、不等概率抽样设计、分层总体的样本分配、抽样比例的回归分析和实验设计正交数组的构造方法等方面有了新见解. 再抽样及随机加权方法、随机模型及连续调查报告的趋势计量、辅助信息和抽样方法, 则涉及多种统计分析和计算方法的应用, 在转换样本调查设计等方面也取得一定成果. 计算机辅助调查有了新的发展.

众所周知, 理论来源于实践, 反过来又服务于实践. 统计理论的研究和分析技术的发展,

无疑对统计的实践起到了一定的指导作用. 从另一角度也显示出了, 统计理论和分析技术的不断完善, 为统计学的应用奠定了基础, 确保了统计学强大的生命力.

二、多元统计分析方法的应用

这里我们要通过一些实际的问题, 解释选择统计方法和研究目的之间的关系, 这些问题以及本书中的大量案例能够使得读者对多元统计分析方法在各个领域中的广泛应用有一定的了解. 多元统计分析方法从研究问题的角度可以分为不同的类, 相应有具体解决问题的方法, 参看表 1.1.

表 1.1 统计方法和研究目的之间的关系

问题	内容	方法
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、相应分析、多维标度法、可视化分析
分类和组合	基于所测量到的一些特征, 给出好的分组方法, 对相似的对象或变量分组.	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, 相关关系又是怎样体现.	多元回归、典型相关、主成分分析、因子分析、相应分析、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

多元统计分析方法在经济管理、农业、医学、教育学、体育科学、生态学、地质学、社会学、考古学、环境保护、军事科学、文学等方面都有广泛的应用, 这里我们例举一些实际问题, 进一步了解多元统计分析的应用领域, 让读者从感性上加深对多元统计分析的认识.

例 1.1 银行希望根据客户过去的贷款数据, 来预测新的贷款者核贷后逾期的机率, 以做为银行是否核贷的依据, 或者提供给客户其他类型的贷款产品. 在此可以利用聚类分析和因子分析方法.

例 1.2 城镇居民消费水平通常用八项指标来描述, 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. 这八项指标存在一定的线性关系. 为了研究城镇居民的消费结构, 需要将相关强的指标归并到一起, 这实际就是对指标进行聚类分析.

例 1.3 频发的网络服务侵权纠纷已经成为制约我国信息网络产业有序发展的主要障碍, 利用统计方法进行定性分析、对比分析、主成分分析以完善相关法律法规, 厘清各类型