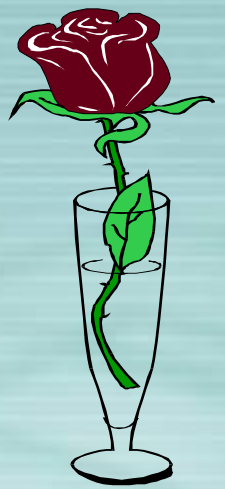
第一章　多元分析概述

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。然而，随着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。在信息爆炸的今天，人们已经意识到数据最值钱的时代已经到来。
- 显然，大量信息在给人们带来方便的同时也带来一系列问题。

多元统计

- 比如：信息量过大，超过了人们掌握、消化的能力；一些信息真伪难辩，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统一处理等等，这种变化使传统的数据库技术和数据处理手段已经不能满足要求.Internet的迅猛发展也使得网络上的各种资源信息异常丰富，在其中进行信息的查找真如大海捞针。这样又给多元统计分析理论的发展和方法的应用提出了新的挑战。



多元统计

- 多元统计分析起源于上世纪初，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得奠基性工作，使多元分析在理论上得到了迅速得发展。20世纪40年代在心理、教育、生物等方面有不少得应用，但由于计算量大，使其发展受到影响，甚至停滞了相当长得时间。20世纪50年代中期，随着电子计算机得出现和发展，使多元分析方法在地质、气象、医学、社会学等方面得到广泛得应用。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更加扩大。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。

多元统计

- 在20世纪末与本世纪初，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和数据库技术相结合，已在经济、商业、金融、天文等行业得到了成功的应用。
- 为了让人们更好的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、多维标度法以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。

第一节 应用背景

一 统计学的生命力在于应用

二 多元统计分析方法的应用

多元统计

- 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

- 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、真切地认识世界。

多元统计

- 数理统计学派的奠基人凯特勒在统计学中引入了概率论，把它应用与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。

多元统计

- 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，并提出以检验作为曲线配合适合度的一种量度的思想。

多元统计

- 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。

多元统计

- 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它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。
- 这一些都充分地体现了统计学强有利的生命力，其具体表现在：

多元统计

- 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。

多元统计

- 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烦而沉闷多时的研究领域住入了新的活力，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。

多元统计

- 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。
- 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。

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的解释。	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

1、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。

2、在企业经济效益的评价中，涉及到的指标往往很多，如百元固定资产原值实现产值、百元固定资产原值实现利税、百元资金实现利税、百元工业总产值实现利税、百元销售收入实现利税、每吨标准煤实现工业产值、每千瓦时电力实现工业产值、全员劳动生产率、百元流动资金实现产值。如何将这些具有错综复杂关系的指标综合成几个较少的因子，既有利于对问题进行分析和解释，又能便于抓住主要矛盾做出科学的评价。可用主成分分析和因子分析法。

多元统计

3、某一产品是用两种不同原料生产的，试问此两种原料生产的产品寿命有无显著差异？又比如，若考察某商业行业今年和去年的经营状况，这时需要看这两年经营指标的平均水平是否有显著差异以及经营指标之间的波动是否有显著差异。可用多元正态总体均值向量和协差阵的假设检验。

4、按现行统计报表制度，农村家庭纯收入是指农村常住居民家庭总收入中扣除从事生产和非生产经营用支出、税款和上交承包集体任务金额以后剩余的、可直接用于进行生产的、非生产性建设投资、生产性消费的那一部分收入。如果我们收集某年各个省、自治区、直辖市农民家庭人均纯收入的数据，可以用相应分析，揭示全国农民人均纯收入的特征以及各省、自治区、直辖市与各收入指标的关系。

多元统计

5、某医院已有100个分别患有胃炎、肝炎、冠心病、糖尿病等的病人资料，记录了他们每个人若干项症状指标数据。如果对于一个新的病人，当也测得这若干项症状指标时，可以利用判别分析方法判定他患的是哪种病。

6、有100种酒，品尝家可以对每两种酒进行品尝对比，给出一种相近程度的得分（越相近得分越高，相差越远得分越低），希望用这些得分数据来了解这100种酒之间的结构关系。这样的问题就可以用多维标度法来解决。

7、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。

多元统计

8、研究中国七星瓢虫在黄海、渤海的群聚与近期气象条件的关系。对1000个类似的鱼类样本，如何根据测量的特征如体重、身长、鳍数、鳍长、头宽等，我们可以利用聚类分析方法将这类鱼分成几个不同品种。

9、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。

10、在高考招生工作中，我们知道每个考生的基本情况，通过分析我们不仅可以了解到学生喜欢学习的科目，还可以进一步从考生每门课程的成绩，分析出学生的逻辑思维能力、形象思维能力和记忆力等等对学习成绩的影响。

第三节 计算机在统计分析中的应用

一 加强计算机统计应用教学

二 计算机统计分析的基本步骤

一、加强计算机统计应用教学

- 从统计学产生和发展的历史我们可以看到，统计数据的收集、整理、加工、分析的过程中，对统计学的昌盛发展起决定性作用的工具就是高速的计算工具——计算机。同样，它对统计教学也是相当重要的。首先，应在统计教学中大力加强通用统计应用软件的教学。在国外比较流行的统计应用软件如SAS、SPSS、S-PLUS、MINITAB、EXCEL等，都不仅仅是一个统计分析软件，它们都可用于统计工作的全过程，如统计调查方案设计、统计整理、数据库的建立与管理等等。
- 因此，加强通用统计应用软件的教学十分重要。

多元统计

- 其次，应把掌握一种算法语言和一定的数据库知识或网络知识作为对统计专业学生计算机应用知识的基本要求。应注重于应用，根据统计课程的特点，处理好通用统计应用软件课程教学与应用统计方法课程教学间的关系，尽可能把它们有机地结合起来。这样不仅能突出有关统计方法课程的应用特色，更好地理解其原理、基本思想及适用条件，而且能使学生通过课程的反复学习，熟练掌握通用统计软件的使用。
- 这里我们应该清楚的认识到，多元统计分析的数学计算比较复杂，如果不借助于计算机，许多问题根本无法解决。在多元统计分析的教学中，加强计算机的应用教学就显得尤为重要。因此，本书在案例分析中，大部分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，而且能更好的显示出其应用价值。

二、计算机统计分析的基本步骤

- 计算机统计分析的基本过程为：

1．数据的组织。数据的组织实际上就是数据库的建立。数据组织有两步。第一步是编码，即用数字代表分类数据（有时也可以是区间数据或比率数据）。第二步是给变量赋值，即设置变量并根据研究结果给予其数字代码。

2．数据的录入。数据的录入就是将编码数据输入计算机、即输入已经建立的数据库结构，形成数据库。数据录入关键的是保证录入的正确性。录入错误主要有认读错误和按键错误。在数据录入后还应进行检验，检验可采取计算机核对和人工核对两种方法。

3．统计分析。首先根据研究目的和需要确定统计方法，然后确定与选定的统计方法相应的运行程序，既可以用计算机存储的统计分析程序，也可以用其他的统计软件包中的程序。

4．结果输出。经过统计分析，计算结果可用计算机打印出来，输出的形式有列表、图形等。

本章结束