21世纪高等教育计算机规划教材

COMPUTER

数据分析方法及应用
——基于SPSS和EXCEL环境

Application & Method of Data Analysis
(Based on SPSS and EXCEL)

马秀麟 姚自明 邬彤 王敏 主编

- 强调分析方法的适用范畴
- 准确解读分析方法的输出表格
- 基于案例学习数据分析技术

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

21世纪高等教育计算机规划教材

COMPUTER

数据分析方法及应用
——基于SPSS和EXCEL环境

Application & Method of Data Analysis
(Based on SPSS and EXCEL)

马秀麟 姚自明 邬彤 王敏 主编

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

数据分析方法及应用 : 基于SPSS和EXCEL环境 / 马秀麟等主编. -- 北京 : 人民邮电出版社, 2015.5
21世纪高等教育计算机规划教材
ISBN 978-7-115-39086-8

Ⅰ. ①数… Ⅱ. ①马… Ⅲ. ①统计分析－应用软件－高等学校－教材 Ⅳ. ①C812

中国版本图书馆CIP数据核字(2015)第094414号

内 容 提 要

本书是在教育部高等学校大学计算机课程教学指导委员会提出的“加强在校大学生计算思维能力培养”的指导思想下，基于大数据时代对人才培养的要求而编写的。本书从信息处理与应用的视角入手，探索了基于 SPSS 和 EXCEL 环境的数据预处理和数据分析技术。本书由 6 章组成：数据统计分析的概念、数据梳理与统计描述、数据的差异显著性检验、数据的关联性分析、数据的降维与聚类分析、信度与效度的检验内容。

与同类教材相比，本书比较注重对各种统计分析方法适应范畴的讲解，以保证读者在面对具体研究项目时，能够正确地选择有效方法；与此同时，本书还非常注重对各统计分析方法的输出结果进行讲解，对输出表格内相关数据项之间的关系及其边界值进行了重点说明，从而保证读者在获得了数据的分析结果后能够准确地总结出有价值的研究结论；另外，本书主要面向非统计类专业学生，注意了语言和术语的通俗化和易于理解性。

本书深入浅出，注重系统性和理论性，涵盖知识面较广，既可以作为高等院校数据处理类课程的教材，也可作为有志青年的自学参考资料。

◆ 主　　编　马秀麟　姚自明　邬彤　王　敏
责任编辑　邹文波
执行编辑　吴　婷
责任印制　沈　蓉　彭志环
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京圣夫亚美印刷有限公司印刷
◆ 开本：787×1092　1/16
印张：19　　2015 年 5 月第 1 版
字数：496 千字　　2015 年 5 月北京第 1 次印刷

定价：42.00 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前言

计算机科学与技术与其他学科最大的不同就是突破了学科范式的限制，渗透到各个学科，形成了一套有效的思维模式——计算思维，促使学科的研究范式走向多元化。学习者的计算思维能力水平对他们未来从事科研的能力、适应社会的能力具有重要影响。在大数据时代，数据处理与分析的方法和策略是计算思维的重要组成部分，对学习者科研能力的提升具有重要意义。

随着大数据时代的来临，基于社会调查和项目评价的定量研究日益增加，在教育学领域、经济学领域、社会学领域、中文信息处理领域都有着非常重要的地位。然而，不可否认的是：在众多研究项目中，经常存在着误用或者滥用数据分析方法的问题。笔者作为评委评价北京师范大学的学生科研项目时，每年都会发现多份基于定量分析的科研项目误用了不恰当的数据分析方法。研究方法的错误，直接导致研究结论的可信度不高，严重影响了研究的质量。分析学生和研究者在数据定量分析中出现的这些问题，编者认为导致这一现象的原因主要包括以下 3 个方面：首先，部分研究者并没有完全掌握每个数据分析方法的约束条件，不知道各分析方法对原始数据有哪些要求，数据分析的输入数据应该满足哪些规范；其次，部分研究者并不了解各数据分析方法的基本原理，没有掌握该分析方法是以什么样的算法来评价数据的；第三，部分研究者对各个分析方法的输出结果并不了解，只是简单地知道“检验概率”值以 0.05 为界，对输出表格中的其他信息知之甚少。正是因为存在这些问题，所以在教学类的定量研究中，就出现了对实验班和对照班的数据采用“配对样本 T 检验”的错误，也出现了对低测度的定序数据实施“Pearson 相关”分析的错误；在关于社会调查的数据分析中，更是出现了把无效的线性回归模型作为最终研究成果的研究报告。

根据编者接触到的学生和研究者在数据分析过程中存在的各种问题，编者认为：由于当前已经有很多现成的数据分析软件，所以对于非统计学专业的学生来讲，了解每个数据分析方法中算法的准确数学公式并不重要，更重要的是要求他们把精力聚焦于各个分析方法的输入与输出，即要求学习者准确地掌握每个数据分析方法对原始数据的要求，并能够正确地解读数据分析结果表格。在此过程中，不仅要求他们掌握结果表格中的关键数据，还应掌握结果表格中各个数据项之间的关系，从而更加全面地掌握分析工具的使用。基于这一思路，我们决定组织教师撰写《数据分析方法及应用——基于 SPSS 和 EXCEL 环境》教材，通过此教材把数据分析处理的基本方法介绍给希望从事定量研究的学生和研究者，以促使他们在数据分析过程中少犯研究方法方面的错误，同时也把我们的教学理念渗透到新教材之中。

全书共分 6 章，第 1 章数据统计分析的概念由邬彤副教授负责；第 2 章数据梳理与统计描述由姚自明老师负责；第 3 章数据的差异显著性检验、第 4 章数据的相关性与回归分析、第 5 章数据的降维和聚类分析、第 6 章信度与效度的检验由马秀麟副教授负责。另外，湖北文理学院的王敏老师负责了全书的文字校对、案例验证。最后，全书由马秀麟负责统稿和审定。

本书主要面向需要借助数据分析手段开展定量研究的本科生和硕士研究生，通过对本书 6 章的学习和训练，读者不仅能够准确地掌握数据分析的常见方法，而且能够规范读者的定量研究流程和定量研究方法，提升其定量研究水平。本书的参考学时为 48 ~ 64 学时，建议采用理论与实践相结合的教学模式，以讲授课时与学生上机实践课时等额分配的方式组织教学活动，并在期末预留时间组织学生开展综合性定量研究项目的交流与分析。各模块的学时可参考下面的学时分配表。

学时分配表

章节	课程内容	学时
第 1 章	数据统计分析的概念	3～4
第 2 章	数据梳理与统计描述	8～10
第 3 章	数据的差异显著性检验	10～12
第 4 章	数据的相关性与回归分析	10～12
第 5 章	数据的降维与聚类分析	8～12
第 6 章	信度与效度的检验	6～10
综合应用	综合作业交流（定量研究项目汇报）	3～4
课时总计		48～64

本书的出版得益于多方面的帮助。首先，从事北京师范大学计算机基础课教学的全体同事的长期积累和经验是本书的坚实基础。其次，教育学、经济学、哲学、社会学和法学等专业的本科生和硕士生在开展实证性研究项目过程中对数据分析方法的热切需求，是我们撰写本书的重要动力。另外，在本书成书的过程中，得到了北京师范大学教育技术学院衷克定教授和计算机基础课教学指导委员会的大力支持，并听取了他们许多中肯的建议和批评。同时，人民邮电出版社的编辑对本书的出版给予了自始至终的关心和指导，并给出了很多中肯的意见，保证了图书的质量。在此，对他们表示衷心的感谢！

对于本书，虽然编者尽了很大的努力，尽量避免出现问题，然而由于诸多因素的制约，书中难免有疏漏错误之处，诚恳地请各位老师和同学批评指正。编者的 E-mail：maxl@bnu.edu.cn。

马秀麟
于北师大科技楼
2015 年 2 月

目　录

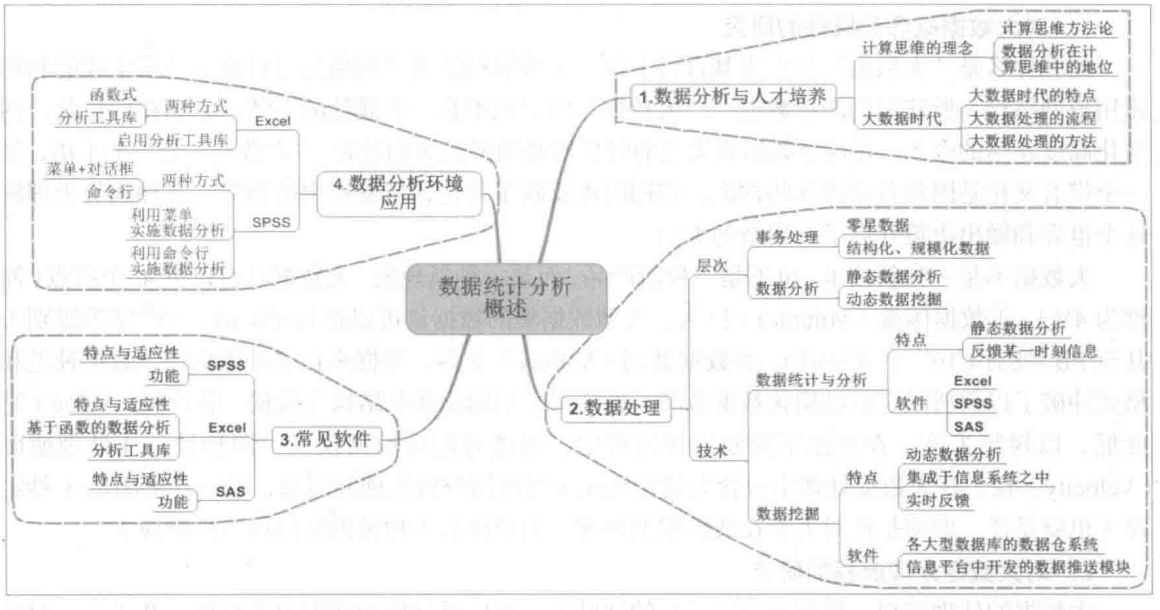
第 1 章
数据统计分析的概念

学习指导

涵盖内容：

本章简要地阐述了大数据时代数据处理的特点、关键方法以及在大数据时代开展实证性研究所必须具备的数据分析处理技能，并从实用性的视角，对主流的数据分析软件进行了简要介绍。

知识地图：



重点与难点：

重点了解大数据时代数据分析的主要技术及其适应性，掌握主流的数据分析软件及其特点。难点在于掌握各数据分析技术的适应范围，能够在恰当的场合选用恰当的技术。

1.1 数据分析能力培养的背景及其意义

人类已经迈入大数据时代，具备较好的数据管理与应用能力已经成为社会对当代大学生的重要要求。从人类使用数据的层次看，人们对数据的管理与应用主要集中在两个层次：针对日常事务处理的数据管理，服务于决策的数据分析。因此，在大学计算机教育中，进行数据分析能力的培养是非常有意义的。

1.1.1 数据分析能力培养的背景

随着信息化的普及和各类信息系统的应用，各类信息系统中都积累了大量的原始数据，分析这些数据内部所蕴含的规律、预测相关系统的运行趋势，已经成为当代信息处理的主要任务。大数据处理就是应时代的需求出现并发展起来的，大数据知识服务是为适应信息服务业智慧化、协作化、绿色化、先觉化和泛在化的发展趋势而衍生的一种基于网络的信息服务新模式，用以对结构化、半结构化及非结构化数据进行多维度处理；是嵌入式协作化知识服务模式的一种新发展，是现代信息服务理念的具体体现。

对数据分析能力培养的研究就是在大数据研究的背景下进行的。对大数据的研究，主要包括了以下几个领域：

1. 对大数据概念和属性的研究

到底什么是“大数据”呢？维基百科上说：大数据指的是“网络公司日常运营所生成的和积累用户网络行为所获得的海量数据”。“大数据”的出现不是一个偶然的事情，它是在信息化、网络化高度发达的今天，在这个数据满天飞的时代所必须要经历的过程。“大数据”是一个术语，是一个带有文化基因和营销理念的词汇，但同时也反映了科技领域发展中的趋势，这种趋势为理解这个世界和做出决策开启了一扇新的大门。

大数据不是一种新技术，也不是一种新产品，而是一种新现象。大数据具有以下 4 个特点（简称为 4V）。①数据体量（Volumes）巨大，大型数据集的数据量可以达到 TB 级（10^{12} 字节级别），甚至 PB 级别（10^{15} 字节级别）。②数据类别（Variety）繁多，数据来自多种数据源，数据种类和格式冲破了以前所限定的结构化数据范畴，囊括了半结构化和非结构化数据。③价值（Value）密度低，以视频为例，在连续不间断监控过程中，可能有用的数据仅仅一两秒钟。④处理速度（Velocity）快，在大数据处理中包含大量在线或实时数据分析处理的需求，通常要求满足 1 秒定律（也就是说，即使是针对大量在线数据的处理，也应该在 1 秒钟的时间内给出响应）。

2. 对大数据处理流程的研究

大数据的处理流程，是指在合适工具的辅助下，对广泛异构的数据源进行抽取和集成，对结果按照一定的标准统一存储。然后，利用合适的数据分析技术对存储的数据进行分析，从中提取有益的知识并利用恰当的方式将结果展现给终端用户。具体来说，可以分为数据获取与集成、数据分析以及数据解释。

大数据的一个重要特点就是数据类型多样性，这就意味着数据来源极其广泛，数据类型极为繁杂，这种复杂的数据环境给大数据的处理带来极大的挑战。要想处理大数据，首先必须对数据源提供的数据进行筛选和集成，从中提取出关系和实体，经过关联和聚合之后采用统一定义的结构来存储这些数据。在此过程中，应该注意对数据进行必要的清洗，清理掉垃圾数据和无效信息，

保证数据质量及可信性，同时还要兼顾大数据的模式和数据内在的关系。

数据分析是整个大数据处理流程的核心，因为大数据的价值产生于分析过程。从异构数据源抽取和集成的数据构成了数据分析的原始数据，然后根据不同应用的需求可以从这些数据中选择全部或部分进行分析。鉴于大数据的特点，传统的分析技术如数据挖掘、机器学习、统计分析可以应用在大数据处理中，在特定情况下则需要根据大数据的时代需要做出调整。

尽管数据分析是大数据处理的核心，但普通用户往往更关心结果的展示。比较传统的就是以文本形式简要地陈述分析结论，也可以直接在电脑终端上显示结果。这种方法在面对小数据量时是一种很好的选择。但是，大数据时代的数据分析结果往往也是海量的，同时结果之间的关联关系极其复杂，因此借助于可视化的技术或者使用特定的数字指标来呈现数据分析结果是必要的。

3. 对大数据挖掘、分析技术的研究

随着大数据热的兴起，数据分析与数据挖掘的算法日益成熟，统计学、数据分析的技术手段被引入到大数据处理过程中，起到了重要的作用。

首先，统计学的理论被引入到大数据处理领域，数据统计分析的手段已经变成了数据分析的常规手段。由于大数据的规模比较大，经过数据清洗的有效数据通常符合统计规律，因此信度系数检验、关联性分析、数据的离散度分析（方差、标准差）、聚类分析、主成分分析等被广泛地应用到大数据处理的过程中。目前，这些技术已经被集成到多种计算机信息系统中，发挥着越来越重要的作用。

其次，除了传统的数据分析技术之外，遗传算法、神经网络、语义网络、分布式数据库管理等面向大数据的处理技术已经成熟。

第三，专业的数据挖掘软件、数据推送技术快速发展。应大数据处理的要求，IBM 公司、微软公司、Oracle 公司都在自己的大型数据库处理系统（即 DBMS）中集成了数据挖掘技术，强化时间序列特点、支持数据挖掘技术的数据仓已经成为主流数据库系统的重要组件，为基于大数据的数据挖掘提供了强大的技术支撑。

1.1.2　数据分析能力培养的意义

1. 数据分析是大数据时代各行业和学科发展的迫切要求

管理信息化、教育信息化、企业现代化的快速发展，促使各行各业在近十年都出现了极大的、极快的数据积累。不论在商业贸易领域，还是在经济建设领域、教育领域，都积累了海量数据。如何充分地利用这些数据，从中总结出规律，为下一步的决策提供依据，或者依据数据分析实现智能化的数据推送，已经成为社会科学的重要研究领域。基于这一需求而快速发展的数据建模技术、数据挖掘技术已经成长为计算机科学的重要应用领域，也是管理与决策的重要依据。

不同学科所形成的科学研究方法在学科自身发展的推动下迅速发展，在计算机科学与技术的支持下，自然科学研究的主流研究方法范式——面向实证研究的量化数据处理，发展出了一整套形式语言理论、编译理论、检验理论以及优化理论。而人文社会科学研究的主流研究方法范式——思辨研究质性分析，则受到了计算机科学和数据处理理论的冲击，从基本文本分析到语义分析、语料分析处理，也都能借助计算机将原本只有人工才能分析的复杂内容机器化、形式化和程序化，并借助数据处理的理论和方法获得可信度更高的研究结论。由于研究者群体开展研究活动时所遵从的一系列规范的结构性组合是针对“问题域”本身的，当数据分析方法作为工具和技术所承载的方法论属性渗透进来后，它将超越学科疆域的研究“规则和框架”，成为跨学科的研究范式。数

据分析的理念和模式必将对相关学科的研究方法体系产生重要影响，甚至从根本上改变其原有的研究范式。

尽管在计算机类公共课程中直接讨论数据挖掘和数据分析技术的原理和算法会存在困难，但是，如果只是把数据建模和数据挖掘技术的概念、方法和工具介绍给学生，允许学生在借用数据建模和数据挖掘的现有工具时不必详细掌握其内部的算法结构，只需了解每个工具的输入、输出及其参数规范，让学生逐步具备准确地使用数据分析工具并解读数据分析结果的能力，还是完全可行的。如果做到了这一点，我们的学生们在参与社科类的科研活动时就能借助这些工具开展数据分析并能根据分析结果获得比较准确的量化结论。与此同时，也一定能够拓展学生的解题方法，使研究的科学性、严谨性都能得到很大的提升，从而优化其思维方式，促进其科研能力的发展。

从另一个视角看，计算机科学的发展和大数据时代对数据分析的迫切需求，促生了许多数据统计分析软件，诸如 SPSS、SAS 等。正是诸多专业的数据分析软件的出现，使数据分析和数据挖掘技术的门槛进一步降低。诸如相关性分析、差异显著性检验（T-检验与方差分析）、归因分析、聚类分析（降维分析）、信度效度检验等算法已经成为人文科学研究中的基本方法。目前，专业化的数据分析不再是统计学专业人士的专利，教育学、经济学、心理学、社会学等人文学科专业的研究人员都应该能通过数据分析软件实现专业水准的定量分析。事实上，随着定量研究法的普及，许多定量分析算法已经被集成到了常规的办公软件中（例如 Excel 中就集成了大量的数据分析模块），使定量研究中所需的数据分析算法也不再神秘。

2. 数据分析能力发展已经成为当代社会人才培养的重要战略目标

随着数据分析与数据挖掘技术的日益普及，建立在数据分析和计算科学基础上的研究方法也逐步向诸多学科渗透，已经有越来越多的学者认识到计算科学在基础理论研究、社会发展和人才培养中的价值，于是计算思维的概念与理论应运而生。

（1）计算思维的概念及其价值

计算机和网络技术日益普及，计算机中的思维方式、解决问题的方法已经逐渐向其他领域渗透，并影响了其他学科，促进了相关学科的发展，甚至形成了一些交叉学科。因此，计算机技术已经不仅仅是一种工具，而是逐步演化成为了一种思维习惯。也就是说，人们在学习和应用计算机的过程中，已经自觉或者不自觉地使用着计算机科学中的思维方式、技术手段，在以计算机处理问题的过程中蕴含着方法论。并在此基础上，逐步拓展了其他学科的研究方法和内容体系，丰富和深化了其他学科的研究范畴。李廉教授指出：自然问题和社会问题内部就蕴含着丰富的属于计算的演化规律，这些演化规律伴随着物质的变化、能量的变化和信息的变换而发展。因此正确地提取这些信息变换，并通过恰当的方式表达出来，使之成为可以被计算机处理的形式，就是基于计算思维概念的基本原理和方法论。

与其他学科相比，计算科学、数据处理科学的最大不同就是突破了学科范式的限制，渗透到了各个学科乃至于推向其前沿，形成了一套有效的思维模式——计算思维，促使学科走向范式多元化，因此，没有哪个学科有如此广泛的研究领域和实践范畴。未来人才的计算思维能力、数据处理水平将对他们从事科研的能力、适应社会的能力具有重要影响。

在大数据时代，数据处理与分析的方法和策略是计算思维的重要组成部分，对学习者科研能力的提升具有重要意义。

（2）计算思维能力培养引起了各方面的重视

计算机技术、网络技术、数据处理技术三者的快速发展，促成了计算思维概念。当前，计算

思维的重要作用已引起了中国学者与美国学者的共同注意。2010 年 7 月 19 日～20 日，国内著名的 9 所高校在西安交通大学举办了“九校联盟计算机基础课课程研讨会”，由陈国良院士做了“计算思维能力研究培养”的报告，强调了“计算思维”能力培养在当前大学计算机基础课教学中的重要意义，强调了计算科学中的思维方式、操作方法对现代化人才培养的重要价值，为新时期大学计算机基础课教学指明了方向。

在教育部高等学校计算机课程教学指导委员会的推动下，教育部高教司于 2012 年启动了“以计算思维为导向的大学计算机基础课程”教改立项工作，共有 22 个项目同时获得教育部立项，标志着以计算思维为导向的大学计算机培养模式正式启动。至此，计算思维能力的培养已经正式列入国内高校计算机基础课教学计划，成为新世纪人才培养的核心内容。

（3）数据分析与处理技术计算思维的核心内容

数据分析与处理技术是计算思维的核心内容，在人才培养中具备开展普遍教育的必要性和可能性。

在计算思维的概念中，建立在计算机和网络技术基础上的“计算”是其核心，在其中起着引导作用的计算方法则是计算思维的灵魂，而数据分析与数据挖掘的相关技术则在其中起着骨架与脊梁的作用，在诸多领域都发挥着重要影响，进而对相关领域的后备人才培养方案都提出了相应的新要求。

从计算思维能力培养的内含看，计算思维的内容博大精深，针对不同层次、不同专业的人才，应该有不同的培养目标和培养模式。因此，在大学计算机教育中，应该分层、分类开展计算思维能力的培养。然而，作为计算思维核心内容的数据分析与处理技术，则随着大数据时代的来临而面临着普遍性的需求。这是由于随着大数据时代的来临，每一个科研工作者都不能回避大数据的冲击，在他们开展研究活动的过程中，都或多或少地需要借助数据分析与数据挖掘的相关技术。

3. 基于数据分析的定量研究方法能改变科研人员的思维方式，促进学科融合

在人文科学的研究中，传统的研究以质性研究法为主。如果想基于数据开展量化研究，则需要以统计学、数据分析的理论为基础，通过大量的数据计算分析数据之间的相关性、差异性，甚至包括归因分析、聚类分析（降维分析）等，才能获得研究结论。在计算机科学和数据分析软件真正地普及以前，基于大量的调查数据开展统计与分析是一项计算量很大的工作，而且要求研究者精确地了解统计学的基本理论、掌握每个数据分析算法的机理和规范。因此，彼时对定量研究者的要求非常高。然而随着专业化的数据分析软件的普及，借助数据分析工具开展定量研究已经成为很多文科科研人员的常规研究方法。目前，对多数从事人文科学研究的科研人员来讲，SPSS 和 SAS 中的各类数据分析工具就像一个只有“输入”和“输出”的“黑匣子”，在开展定量研究的过程中，不需要了解黑匣子的内部结构，只需要能精确地掌握其输入数据和各项参数，并解读其各类输出结果所代表的具体含义，就能够很好地使用它们。

尽管基于数据分析软件的定量研究过程并没有专门要求其用户在计算机操作和数据分析原理方面具备有多深的水平，然而研究发现：很多从事人文科学研究的人员在多次利用定量分析工具开展实证性研究后，其思维习惯和解决问题的方法都有了很大变化，在论证的严谨性、对数据的应用方法等层面，都比以前有了很大的提高，反映了数据分析工具对人们思维方式所产生的重要影响。与此同时，在基于数据分析工具开展研究的过程中，不同学科的科研人员由于使用了相同的研究工具，使他们在一定程度上有了共同的研究语境，促进了学科间的融合和研究成果的分享。

1.2 数据处理的层次与数据分析

1.2.1 数据管理与数据采集的三个层次

1. 日常生活中的琐细数据及其管理

在人们的日常生活中，不可避免地与各种各样的数据打交道，如个人的工资收入数据、个人的日常支出数据、单位内部的固定资产和设备数据。这些数据有的需要长期存储，有的可能仅仅是临时数据。由于这种数据形态多样、比较分散、结构化程度较差，人们通常采用比较随意的记录方式。

在传统社会中，人们通常使用纸和笔、便条等手段记录和管理这些数据。然而，随着计算机技术的普及与发展，越来越多的人借助电子表格等应用软件（例如 Excel）来管理个人的日常琐细数据。

2. 社会调查与数据采集

社会调查方法已经成为 21 世纪科学工作者了解社会现象、掌握社会实际、进而辅助决策的一种基本手段。在这种研究方法的指导下，科学工作者通过设置调查评价指标系统、选择抽样样本，然后针对样本展开调查并获取第一手资料。最后，在调查数据的基础上，开展数据分析，基于“用数据说话”的理念拿出研究结论。这一过程就是实证性研究的基本流程。

在基于社会调查的科学研究中，数据的采集与规范化是整个研究工作的核心环节。人们通常借助电子表格软件手段收集数据，然后利用专业的数据分析软件开展研究。

3. 基于信息系统的数据管理

信息系统也叫信息管理系统（Information Management System，IMS），它是一个进行信息处理的系统，是一个由人、计算机软硬件和数据资源组成的系统。信息系统应该建立在数据库技术的基础上，其目的是及时正确地收集、加工、传递决策所需的数据，实现组织内部活动中对信息的管理、调节和控制。当前社会中，所有的信息系统都是人机系统，都是利用计算机技术、网络技术和数据库技术对信息进行管理、应用的系统。

信息系统建立在数据库技术的基础上，以数据库中的有效数据为核心。当前信息系统的基本结构为：由一台功能强大的计算机充当服务器，在服务器上安装大型数据库，多个终端上的用户可以通过 Internet 或局域网访问服务器上的数据库，实现数据的多用户共享。信息系统的数据库中可以存储各种功能的数据，既可以有管理信息，也可以有其他信息资源，例如教学课件、课堂教学资源数据等。其终端用户也可能是各种人员，既可以是教师、学生，也可以是普通游戏玩家。

目前的信息系统，既有学校管理部门所使用的管理信息系统，也有企业和商业部门为了提高办公效率和经营效益而开发的各种信息系统。例如，淘宝网平台、新浪微博平台、北京师范大学教务系统等。随着信息化水平的提升，信息系统中积累的数据规模日益增加，如何利用这些数据，已经成为信息系统建设者必须考虑的问题。

1.2.2 数据分析与数据挖掘技术的出现

1. 大数据时代对数据分析与挖掘技术的需求

随着大数据热的兴起，数据分析与数据挖掘的算法日益成熟，统计学、数据分析的技术手段

被引入到大数据处理过程中，起到了重要的作用。

首先，统计学的理论被引入到大数据处理领域，数据统计分析的手段已经变成了数据分析的常规手段。由于大数据的规模比较大，经过数据清洗的有效数据通常符合统计规律，因此信度系数检验、关联性分析、数据的离散度分析（方差、标准差）、聚类分析、主成分分析等被广泛地应用到大数据处理的过程中。目前，这些技术已经被集成到多种计算机信息系统中，发挥着越来越重要的作用。

其次，除了传统的数据分析技术之外，遗传算法、神经网络、语义网络、分布式数据库管理等面向大数据的处理技术已经成熟。

第三，专业的数据挖掘软件、数据推送技术快速发展。应大数据处理的要求，IBM 公司、微软公司、Oracle 公司都在自己的大型 DBMS 中集成了数据挖掘技术，强化时间序列特点、支持数据挖掘技术的数据仓已经成为主流数据库系统的重要组件，为基于大数据的数据挖掘提供了强大的技术支撑。

2. 数据分析的两种思路

在数据分析的发展过程中，始终伴随着两条思路。其一，面向静态数据的数据分析；其二是面向动态数据的实时数据挖掘。

所谓面向静态数据的数据分析，是指把通过社会调查、科学实验获得的数据，或者从信息系统导出的针对某一时间段的数据，借助数据分析专业软件，对已有数据进行分析。这种分析没有考虑数据的动态性、变化性，往往是针对某一时间段状态的数据分析。

所谓面向动态数据的实时数据挖掘，是指在信息系统中集成数据挖掘算法，以便信息系统能够随时针对动态数据开展分析。这种技术强化时间序列特点、依托支持动态数据采集和集成的数据仓技术，开展实时的数据分析。它对用户具有很高的要求，已经成为主流数据库系统的重要组件。

1.3　数据描述与数据分析简介

1.3.1　常见的数据描述方法

在数据分析中，人们获得的通常是来自一组样本或者多组样本的调查数据，或者一个数据序列，也有可能是多个数据序列。在对数据序列进行复杂的数据分析前，掌握每个数据序列的基本特征是非常必要的。

1. 对数据序列的集中性描述

在数据分析过程中，人们通常需要了解数据序列集中于哪一个数据点周围。常见的描述量主要有均值、众数和中位数。

均值（Mean）即平均值，是对整个序列求和后再除以数据个数所得到的结果。

众数（Mode）即个数最多的数，它指在整个序列中，那个出现次数最多的数值。简单的说，就是一组数据中占比例最多的那个数。它是在统计分布上具有明显集中趋势点的数值，代表数据的一般水平（众数可以不存在或多于一个）。

中位数（Median）即对数据序列排序后位于正中间的那个数值，它可将数值集合划分为相等的上下两部分。需要注意的是：如果原序列中数据的个数为偶数，则中位数为正中间两个数值的

平均值。

2. 对数据序列的离散性描述

在数据分析中，人们通常需要了解数据序列的波动情况，即数据的离散性。对于数据序列来讲，数据在均值附近的波动性大小是序列的重要属性之一，对于未来的统计分析有重要价值。衡量数据序列离散性的描述量主要有方差、标准差。

方差（Variance），即数据序列中 n 个离差（当前数值与均值的差）的平方和与数据个数 n 的比值。在概率论和数理统计中，方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。

标准差（Standard Deviation）是方差的平方根，也是描述数据离散性的量，中文环境中又常称均方差。

3. 对数据序列分布形态的描述

对于待分析的数据序列，数据的分布形态对分析方法的选择具有重要影响。因此，在数据的描述中，了解数据序列的分布形态也非常重要。在统计学中，数据的分布形态主要有正态分布、均匀分布、指数分布、泊松分布等。另外，偏度和峰度是描述数据分布形态的重要指标。

1.3.2 常见的数据分析技术

计算机科学的发展，促成了许多数据统计分析软件，诸如 SPSS、SAS，就连最简单的 Excel，也提供了数据分析功能等。而且，随着定量研究法的普及，许多定量分析算法已经被集成到了常规的办公软件中，使定量研究中所需的数据分析算法不再神秘，诸如相关性分析、差异显著性检验（T 检验与方差分析）、归因分析、聚类分析（降维分析）、信度效度检验等算法已经成为人文科学研究中的基本方法。

1. 相关性分析

相关性分析是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。相关性的元素之间需要存在一定的联系或者概率才可以进行相关性分析。

在统计分析学中，对两个数据序列相关性的分析主要通过相关系数和相关性检验概率两个指标来体现。相关性系数的绝对值在 0～1 之间，反映两列数据的关联性程度；而检验概率用于反映两列数据不存在相关性的概率值。

2. 差异显著性检验

差异显著性检验也叫差异显著性检验，用于判断两个数据序列是否存在显著的差异。对于数据序列的差异性检验，分为均值差异性和分布差异性两种形式。对于具有正态分布形态的两列连续型数据，通常可检验其均值差异性，而对不明形态或非正态分布的数据，则常常检查其分布差异性。

差异显著性检验是一种推断检验。通常首先假设两列数据没有显著性差异，通过计算相应的统计量判断无显著性差异的概率值（即检验概率）。在统计学中，通常以 0.05（即百分之五为标准），若两列数据的差异显著性检验概率大于 0.05，则认为两列数据没有显著性的差异；反之，若两列数据的差异显著性检验概率小于 0.05，则认为它们具有显著性差异。

3. 降维分析

在数据统计分析过程中，常常从多个视角制作调查或评价指标，从而能够全面地反映调查对象的属性和特点。然而，在调查完成后，常常发现以下问题：多个指标项的语义有重叠，需要获得凝练的分析结论。

为此，需要对调研指标进行凝练，减少评价指标的维数，使结论变得更加易于表述和理解。

这就是降维分析。

4. 聚类分析

在数据统计与分析过程中，常常需要把成千上万的个案分成若干类，以便于操作。例如，可以把学生分为男生、女生，还可以把学生按照综合表现分为优等生、良好生、普通生和差生。这种依据某些因素，对个案分类的过程就是聚类分析，也叫分类。所以，分类分析就是对收集到的数据分析其内在规律和特点，把相似的数据归结为一类的过程。

在数据统计分析过程中，聚类分析可以分为针对个案（记录）的分类和针对变量（字段）的分类。针对变量的聚类过程实际上也是一种降维过程。

1.4　数据分析与挖掘软件

大数据时代的数据分析技术可分为两种不同的类型，一种是针对某一时间点的静态数据的数据分析，另一种是面向动态变化的数据的实时数据挖掘技术。

1.4.1　数据统计与分析软件

1. 专业化的数据统计分析软件

（1）SPSS

SPSS 是 IBM 公司推出的一系列用于统计学分析运算、数据挖掘、预测分析和决策支持任务的软件产品及相关服务的总称，被广泛地应用于教育、心理、经济以及生物、地理、医学等学科领域，是世界上著名的统计分析软件之一。

SPSS 软件最初全称为“Statistical Package for Social Science”，即“社会科学统计软件包”，但是随着 SPSS 产品服务领域的扩大和服务深度的增加，SPSS 公司已于 2000 年正式将英文全称更改为“Solutions Statistical Package for the Social Sciences”，即统计产品与服务解决方案，标志着 SPSS 的战略方向正在做出重大调整。

SPSS for Windows 是一个组合式软件包，它集数据录入、整理、分析功能于一身。用户可以根据实际需要和计算机的功能选择模块。SPSS 的基本功能包括数据管理、统计分析、图表分析、输出管理等。SPSS 统计分析过程包括描述性统计、均值比较、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、数据简化、生存分析、时间序列分析、多重响应等几大类。SPSS 也有专门的绘图系统，可以根据数据绘制各种图形。

（2）SAS

SAS 是一款广泛地应用于化学、生物、心理、农业、医学等领域的统计分析软件。SAS 系统全称为“Statistics Analysis System”，即“数据统计分析系统”，它最早是由北卡罗来纳州立大学的两位生物统计学研究生编制并研发的，并于 1976 年正式推出了 SAS 软件。SAS 是用于决策支持的大型集成信息系统，但该软件系统最早的功能限于统计分析，至今，统计分析功能也仍是它的重要组成部分和核心功能。

SAS 是由大型机系统发展而来，其核心操作方式就是程序驱动，经过多年的发展，现在已经成为商业分析软件与服务领域的领跑者。

（3）Systat

Systat 的含义是 System Statistical，这是一款强大的统计分析软件，拥有高效的数据分析和各