国家自然科学基金出版支持

数据分析与数据挖掘

姜　维　编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内容简介

本书讲解了数据分析与数据挖掘的理论和方法，包括描述性统计、假设检验、方差分析、回归分析、关联规则、决策树、贝叶斯模型、判别分析、支持向量机、神经网络、聚类分析、离群点分析等，同时配有应用举例。大数据分析、人工智能与互联网的发展为该领域的研究提出了新的需求，本书在阐述理论方法的同时，也注重实践，更注重知识体系结构。书中的理论和技术既能作为科研的基础，也能直接用来解决实际问题。

本书可作为相关专业高年级本科生和研究生的教学用书，也可作为数据分析与数据挖掘研究人员的参考用书。各种编程语言均可实现本书中的理论方法，如 Python、C++和 R 等，还有许多软件工具可用，如 SPSS 等。本书配套的编程软件工具有利于将理论和技术应用于实践。

图书在版编目（CIP）数据

数据分析与数据挖掘 / 姜维编著. —北京：电子工业出版社，2023.1
ISBN 978-7-121-44743-3

Ⅰ. ①数… Ⅱ. ①姜… Ⅲ. ①数据处理－高等学校－教材②数据采集－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2022)第 243582 号

责任编辑：石会敏　　文字编辑：苏颖杰　　特约编辑：侯学明
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×1092　1/16　印张：27.75　　字数：763 千字
版　　次：2023 年 1 月第 1 版
印　　次：2023 年 1 月第 1 次印刷
定　　价：89.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：shhm@phei.com.cn。

前　言

随着大数据分析、人工智能、互联网等的快速发展，数据分析与数据挖掘理论方法的研究与应用已成为当前的热点，在多个领域已经取得有价值的研究成果。

受国家自然科学基金(No.71671052，No.71271066，No.70801022)的支持，我基于多年在数据分析和数据挖掘方面的成果积累，结合多年为本科生和研究生讲授数据分析、数据挖掘、文本分析与文本挖掘、专家系统、人工智能方法等课程的经验，撰写了本书。

本书的定位和特点如下：

(1)可作为大学高年级本科生和研究生的教材。兼顾知识深度、系统性和学习者的知识掌握规律，循序渐进地展开知识讲述，专注科学技术，引导创新。

(2)知识体系结构清晰。注重知识的内在逻辑性，系统阐述统计分析、机器学习等理论方法及其在数据分析和数据挖掘中的应用。

(3)理论与实践相结合。注重理论方法的工作原理、工作过程，借助流程、公式、算法、程序，清晰地给出应用的具体细节。阐述典型应用并举例。书中讲解的方法可借助相关软件工具或编程语言直接使用。

(4)配以计算举例、应用举例、图表等，既严谨又通俗易懂。书中的大多数内容已经在课堂中讲授，撰写本书的目标是“在科学的框架下将知识讲述清楚”。

(5)多方位配套相关书籍。《数据分析与数据挖掘建模及工具》侧重案例和工具，《文本分析与文本挖掘》是数据分析与数据挖掘的理论方法在文本领域的应用，《数据分析与数据挖掘 C++建模工具》讲述本书配套软件建模，《高级数据分析与数据挖掘》面向研究生和科研工作者阐述前沿深层知识。

(6)配套软件工具。目前有多种软件工具和编程语言可供数据分析和数据挖掘使用。本书的配套资源有 C++软件研发包，全面支持本书内容，同时也支持《文本分析与文本挖掘》等相关书籍中的理论方法。该软件研发包还提供了向量、矩阵、分布式编程库，可直接编程使用，也支持理论方法的教学、科研和应用。特别说明，考虑目前很多软件工具都是英文版的，故书中有些图表保留了英文，从而与软件更加紧密地结合。

(7)可分块学习。第 1、2、3、4、5、6、7、11 章可作为数据分析的学习内容；第 1、5、7、8、9、10、11、12、13、14、15 章可作为数据挖掘的学习内容。各章之间有内在的逻辑性，可酌情选择；还可参考相关图书《数据分析与数据挖掘建模及工具》中的案例。

感谢课题组成员的支持，感谢姜绍航对本书出版的支持。撰写本书时参考了国内外同行的研究成果，特别是一些基础理论方法，在此表示感谢。数据分析与数据挖掘的需求在不断变化，前沿问

题、新的理论方法、新技术也层出不穷，书中难免存在错误和不足之处，敬请各位专家与学者批评指正，以进一步完善。

本书在配套网站（网址：http://www.jiangw.cn 和 http://www.orsci.com）上提供共享技术资料、在线研讨群、书籍勘误表、最新研究文档、常见问题、联系方式等，读者可根据需要下载使用。

姜　维
哈尔滨工业大学
jiangw@hit.edu.cn
2022 年 10 月 20 日

目　录

第 1 章　数据分析与数据挖掘基础

数据分析(Data analysis)是指用统计分析方法对收集来的大量数据进行分析、研究和概括总结，提取数据中隐含的有用信息和知识。常说的数据分析通常是指统计数据分析。数据挖掘(Data mining)一般是指从大量的数据中利用挖掘方法和算法提取隐藏在数据中的信息和知识。数据挖掘所用的方法一般来自统计、在线分析处理、情报检索、机器学习、模式识别、专家系统等。相比常用的统计数据分析方法，数据挖掘更强调对那些自身分布规律不明显的数据，甚至很难找到分布规律的数据进行分析，典型的处理方法是使用机器学习和模式分类中的关联规则、分类、聚类等技术。

数据分析和数据挖掘两个概念并没有严格的区分边界，有些学者认为数据挖掘是数据分析研究中的一个前沿分支。数据分析和数据挖掘的作用一般包括：描述作用，即描述数据的一般性质，提出信息和知识；预测作用，即利用数据进行推断，能对新的数据对象做预测。[①]

1.1　数据分析与数据挖掘需求

1.1.1　数据分析与数据挖掘

数据分析与数据挖掘有着非常广泛的应用，特别是计算机提高了数据处理的能力，甚至可以通过多机协同运算、云计算等新的计算形式，提供更强的数据处理能力。在这种情况下，以往的经典数据分析方法和现代的数据挖掘方法，都可以借助强大的现代计算工具进行大规模数据处理。现在人们都强调处于大数据时代，有许多值得分析和挖掘的数据，想充分利用数据就需要做好三件事：掌握数据分析和挖掘方法、探索新的分析和挖掘方法，以及耐心细致地动手付诸实践。

数据分析与数据挖掘方面的应用案例不胜枚举，这里列举几个常见的例子：①大多数超市都有用户购买记录，也有会员卡，忠实记录了消费者的购买行为，深入分析这些数据能够挖掘用户的行为特点，提高超市的收入。②在电子商务领域，除了商务自身数据，商品的评论、人们的购物习惯、市场政策、人们的生活数据等都影响着电子商务的发展。如何能够进行有效分析是电子商务运营的一项重要工作。③在企业客户关系管理中，强调对客户分级管理，有针对性地运用打折、奖励、优惠服务等营销手段，那么根据企业中的一些客户信息，如何对客户进行分级就是客户关系管理中的一项重要任务。④在专车、出租车运营公司中，利用大量用户出行数据，通过针对不同季节、不同节假日、每天的不同时段进行车辆部署优化，提高专业的服务水平和运营收益。⑤银行信贷机构，通过企业或者个人的基础信息、财务信息等多个因素，可以自动地为客户的信用进行评级打分，分析出优质客户、普通客户或可能坏账的客户。⑥一些大型游乐场，为用户提供游园时使用的智能佩戴设备，通过实时数据分析，辅助用户制订游园的方案，并根据用户的年龄、性别、位置等信息实时推荐一些餐饮、休息区等服务项目，既优化了用户游园体验，又增加了游乐场的收入。⑦Google 的 Flu Trends(流感趋势)使用特殊的搜索项作为流感活动的指示器。它发现了搜索流感相关信息的人数与实际具有流感症状的人数

① 本书侧重理论知识，《数据分析与数据挖掘建模与工具》侧重案例、工具与实践。

之间的紧密联系。当与流感相关的所有搜索都聚集在一起时，一个模型就出现了。使用聚集的搜索数据，Google 的 Flu Trends 可以比传统的系统早两周对流感活动做出评估。⑧随着人工智能技术的发展，自动问答系统、智能机器人的技术都在日益发展，可以把互联网看作一个巨大的半结构化数据库，有些问题的答案可以通过自动挖掘技术从互联网中获得，自动问答系统就是一个研究方向。智能机器人是智能化发展的一个代表，也可以将领域知识集成，甚至可以根据应用场合自动学习一些新的知识提供智能化服务。

科研机构、企业、政府和数据分析机构是一些典型的数据分析和数据挖掘研究和应用场所。现在已经有一些专门从事数据分析的岗位，称作数据分析岗位。数据分析师掌握较多的数据分析方法，拥有丰富的数据分析经验，当解决应用领域问题时，其作用包括：①快速挖掘数据背后的实用价值；②提供科学、合理的决策依据。例如，许多大型企业都会设置专门的数据分析部门，有专门一些人进行数据分析，为企业服务，而有些中小型企业可能选择第三方数据分析机构来帮助企业进行数据分析。企业应用数据分析和数据挖掘技术，能够帮助企业掌握消费者的需求、发现市场趋势、了解竞争对手的动向、提高员工工作效率等。

1.1.2 大数据处理需求

随着计算机技术和互联网的快速发展，许多行业都进入了大数据时代。人们想充分挖掘数据中的有用信息和知识，以期为人们提供更好的服务。大数据(Big data)是指数据量大并且通过现有常规软件工具很难直接处理的数据。现有的“大数据”概念一般是指面临着大数据分析这一问题，既需要采集、收集或整理大数据，又需要对大量数据集进行存储、处理和分析。大数据给人的直观印象首先是数据量大，而“数据量多少才算是大”是一个与时俱进的问题。在互联网刚出现时，人们感觉互联网上的共享信息数据量很大，而 21 世纪初出现更大规模的数据，如互联网上的海量数据、飞机飞行数据、人脸图像采集、视频数据等。2012 年，美国开始启动大数据研究与发展项目。目前，通常数据量达到 1PB=1024TB=1024×1024GB 级别的数据，才能称为大数据。大数据问题在某些大型企业、政府和科研机构等部门都面临着如何存储和如何处理的问题，如飞机制造公司采集的飞机发动机飞行数据、图像研究机构收集的海量图像数据等。如今，在磁盘阵列、分布式存储，特别是互联网的“云存储”和“云计算”的支持下，大数据的收集和处理变得更加可行和便利，一些新的大数据挖掘方法逐渐被提出来。人们已经感受到大数据分析带来的重要益处。

大数据的处理和分析技术可看作传统数据分析与数据挖掘的一个重要分支，它为人工智能技术提供新的动力，促进了“数据驱动”的人工智能研究发展。大数据将会是一个长期的研究领域，目前在教育、营销、金融、医疗、企业、政务、互联网等许多领域都有着较为广泛的研究，人们也将在探索、研究和实践中获得益处。大数据的基本特征可以使用四个 V 来概括，即 Volume(体量大)、Variety(多样性)、Value(价值密度低)和 Velocity(处理速度快)。Volume 特性是指数据量比较大，通常可达到 PB 级。Variety 特性是指数据形式除数字外，还可能包括文字、声音、图片、视频、地理位置等。Value 特性是指相比传统数据来说，数据的价值密度一般比较低，但是大量的数据的价值通常就有很大的价值了。Velocity 特性包括两方面：①数据增长快；②要求高效地挖掘其中有价值信息与知识。常规数据和大数据的数据分析对比如表 1.1 所示。

表 1.1 常规数据和大数据的数据分析对比

比 较 内 容	常 规 数 据	大 数 据
数据量	数据规模小或者适中	数据海量，规模庞大
价值密度	价值密度高	价值密度低

续表

比 较 内 容	常 规 数 据	大 数 据
存储方式	单机或若干台计算机组网存储	阵列存储、分布式存储或云存储
分析方法	现有数据分析模型，一般单台计算机就能处理	一般采用三种方式：①经过数据抽取和整理方式进行一定数据整理，提炼出分析数据集，然后使用单机进行分析；②构造近似计算方法，利用单机或若干组网计算机进行不完全精确的但有一定价值的数据分析；③采用大规模分布式计算或者云计算方式，研究专门的大数据分析方法
分析目的	信息获取、知识发现，数据分类、目标预测等	信息获取、知识发现，数据分类、目标预测等。目前在大数据研究上，更多关注其预测作用

数据分析和数据挖掘是期望从数据中获取有用的信息和知识，无论是常规数据还是大数据，都强调把蕴含在数据中的有用信息和知识挖掘出来。有一类方法是利用已知的领域知识去解决问题，这类解决方法可以称之为“知识驱动方法”。例如，利用营销知识进行产品销售推广宣传、利用专家经验进行客户分类等。还有一类方法是从已有的数据中获取信息和知识，然后利用该信息和知识去解决领域问题，称之为“数据驱动方法”。例如，电子商务推荐系统根据大量用户的购买记录进行个性化推荐，企业利用互联网数据进行市场预测，Google 搜索引擎根据大量用户的搜索行为进行精准广告推送等。

1.1.3 数据分析误区与隐私问题

进行数据分析和数据挖掘的目的是从数据中提炼出信息和知识，然而仅仅在给定数据集上建模并生成模型结果并非数据分析的终点，还应该结合实际应用问题来判断数据分析本身是否具备有效的意义。数据分析应用中常存在一些误区，其中包括虚假性和误导性分析、无意义的分析等。例如，在有些电子商务网站上，商家通过对购买者评论进行奖励等策略，诱使大量用户提供好评，甚至通过一些手段创造一些评论，而基于此的统计数据可能会误导他人。某些影评网站，通过一些手段为一些关照的电影增加好评，误导观众。某些调查故意将问卷有选择性地发放，获得的调查结果有失偏颇，只为得到所期望的调查结论。某些私立医院宣传治愈率高，实质上，许多重症患者都转向了公立医院。某部门统计一段时期内交通事故与星座的关系，这种无意义的数据分析有可能误导他人。上述例子说明，应该结合实际应用来判别数据分析本身是否具有有效意义。再举例，有学者研究发现刑事犯罪的数量与所在城市的厕所数量呈正相关。该现象背后可能存在其他内在影响，如人口数量等，而不能简单地得出结论：消除厕所将减少犯罪。数据分析并非终点，其内在意义仍需结合实际问题辨析。

数据分析中也应该注意数据隐私问题。在精准营销中，为了更有针对性地营销，有可能会过度收集用户个人数据，故需要控制数据收集界限，避免出现隐私数据收集问题。在客户关系管理中，需要注意收集客户信息时不要侵犯客户隐私。有些软件跟踪用户的运动位置，收集用户使用行为，甚至用户通话、相册等内容，若未经授权，则可能已侵犯用户隐私。有些类似的不良数据收集行为，如某些禁止的互联网数据抓取，有可能违背商业规则。

1.2 数据分析与数据挖掘的工作过程

1.2.1 数据分析的主要工作过程

数据分析与数据挖掘可看作“数据驱动方法”，并且是数据驱动方法的核心模块。进行数据分析与数据挖掘有两种常见工作方式：一种强调信息和知识的获取；另一种强调信息和知识的

获取及运用。针对第一种方式，进行数据分析与数据挖掘通常包括五个主要步骤：①明确数据分析与数据挖掘的目的；②进行数据收集和整理；③数据分析与数据挖掘；④数据分析结果展示；⑤形成数据分析报告。针对第二种方式，整个处理过程通常包括六个步骤：①明确数据分析与数据挖掘的目的；②进行数据收集和整理；③数据分析与数据挖掘；④运用数据分析和数据挖掘的结果，可能根据应用情况反馈重新调整模型，直到达到满意结果；⑤数据分析和应用情况结果展示；⑥形成数据分析与应用报告。

明确数据分析与数据挖掘的目的，能够保证数据处理整体方案的逻辑合理性，对各项工作的边界加以约束，保证其不会偏离主题。虽然科研中也存在意外发现，但本书将意外发现引发的新研究视作一个新的研究目的，需要开展新的研究工作。

数据的收集和整理是数据预处理的重要组成部分，这部分的工作是为数据分析和数据挖掘提供数据准备的。数据的来源可能是直接的也可能是间接的，企业内部数据，可看作直接数据来源，并且是权威准确的。有些数据可能来自其他途径，如在互联网上抓取的数据，这属于间接数据来源，此时需要仔细分析数据的可靠性。在数据的收集上，要尽量保证数据具有代表性，能够涵盖所要研究的问题的各方面规律。在数据的整理上，要尽量保证所需要的信息在整理时不丢失，并且能够形成数据分析和数据挖掘所需要的格式。特别值得注意的是，数据的质量和数量对数据分析和数据挖掘有着重要的影响。数据的收集和整理工作通常比较耗费时间，但在实际应用中却是极具价值的，需要认真仔细地开展。

数据分析与数据挖掘中有许多方法，可以根据具体目的加以选择，也可以进行各种尝试以期获得意外发现，还可以根据研究目的，尝试多种方法组合、改进算法甚至用新的方法探索研究。有些研究工作只需进行初步的数据分析和数据挖掘就足够了，如研究衡量客户价值的因素、研究房地产价格的影响因素、研究销售商品的关联关系、研究网络新闻的聚类关系；而有些研究则需要将数据分析和数据挖掘结果进一步应用，如用于分类、预测，并进行进一步的评价。例如，研究衡量客户价值的因素之后，建立客户价值评价模型，并根据该模型对新的客户进行分类；研究房地产价格影响要素之后，建立房地产价格预测模型对新的房地产价格进行预测；研究大型游乐场内用户的游览习惯，向用户推荐游园路线、餐馆等。

数据分析和应用情况结果一般需要以较为直观的方式展示。观看结果的人可能具备专业的知识，也可能不完全熟悉专业的知识，因此展示的方式需要根据目标人群来选择。在通常情况下，图表展示属于广泛采用的形式，如直方图、趋势图、对比图等。数据分析和应用情况展示需要注意：①考虑目标人群关注的信息和知识，如对于决策者需要提供其决策时需要参考的各项信息，对于专业人士要提供更专业的数据以便其查看和分析专业数据；②展示一些数据分析和数据挖掘的关键概念和关键数据，可以令观察者更加信服，同时又有助于观察者进行本次数据分析和数据挖掘研究价值的判断，但需要注意所提供的概念要明确统一，既有专业的定义，又有较为直观的形象解释，对所提供的关键数据应说明其作用并注重前后的逻辑性。

数据分析和数据挖掘报告大致包括三个部分：前序部分、正文部分和结论部分。前序部分主要包括封面、标题、目录、前言；正文部分包括数据分析和数据挖掘的目的，数据来源，数据分析、挖掘和应用过程概述，数据的描述分析，关键分析过程和关键应用过程，应用效果；结论部分包括数据分析的若干细节结论、应用的若干细节结论和总结论。数据分析报告主要给同事、上级领导或下级员工查看，需要根据目标人群来确定要展示的内容；报告中要适当用图表展示相关内容，提供关键数据分析和数据挖掘的内容，提供数据分析和数据挖掘应用情况分析；报告中可以适当使用专业术语，这样会增强阅读者的信任感，但如果专业术语过多，可能会影响阅读者的理解。撰写报告时要严谨、认真、前后逻辑合理并突出有意义的结论。

1.2.2 数据收集

数据分析与数据挖掘的数据来源广泛，按照数据的获取方式来分，一般可分为直接数据来源和间接数据来源。直接数据来源是指所需要的数据直接来自产生原始数据的企业、政府等部门，如某生产制造企业的信息管理系统中的数据、某政府单位内部的数据、某大型超市的销售数据。在许多情况下，数据直接来自对数据分析有应用需求的部门，它们常以数据库或者数据仓库的形式存储数据，数据存储比较规范，数据相对可靠。间接数据来源是指获取的数据并非直接来自产生原始数据的企业、政府等部门，而是以间接途径获取的数据，如通过互联网抓取的某电子商务网站的销售数据、从互联网上收集的评论某产品的数据。间接数据的可靠性较低，需要根据具体问题进行可靠性分析。按照数据获取途径来分，一般可分为内部途径和外部途径。内部途径一般是指来自需要数据分析的部门自身，通常直接数据来源就是来自内部途径。外部途径是通过协作单位、问卷调查、互联网等方式获取的。一般来说，间接数据来源是通过外部途径获取的。需要注意，直接数据来源和间接数据来源的区分要点是看数据是否直接由需要数据分析的部门直接生成。例如，协作单位提供了竞争对手企业的内部数据，那么这属于外部途径，但是直接数据来源。

按照数据是否直接显式地表示，可以划分为结构化数据、非结构化数据和半结构化数据。结构化数据，是指数据以所需要的形式存储和表示，典型的就是数据库或者数据仓库等规范化的存储形式，其中所需要的各项数据细节被直接描述出来，如数据库以“字段”来显式地指明某“记录”的对应属性值。例如，企业 ERP、财务系统、医疗数据库、学生成绩库、出租专车的 GPS 运行数据、政府行政审批、超市的销售数据库等的数据就是结构化数据。非结构化数据是指数据结构不规则或不完整，没有预定义的数据模型，没有显式地标示出数据分析所需要的各项数据值。例如，办公文档、文本、图片、声音、视频等，通常是非结构化数据。半结构化数据是介于结构化数据和非结构化数据之间的一种类型数据，数据中有一部分指示表示，但又没有完全显式地标示出所需要的数据。例如，HTML、XML 等类数据中有些标记，但又不完全规范，通常称这类数据为半结构化数据。

一般来说，为了保证数据收集的质量，需要遵守以下原则：

(1)可靠性原则。是指无论是直接数据来源还是间接数据来源，都应该保证数据满足可靠性要求、由真实环境或对象产生。

(2)时效性原则。是指所获取的数据要满足数据分析的时效性要求。通常，近期数据的描述能力要好于过于陈旧的数据。因此，数据收集时既要分析即将获取的数据本身的时效性，也要分析数据收集时间的时效性，以保证所获取数据的综合时效价值。

(3)完整性原则。是指收集的每个数据样本在内容上都是完整无缺的，或者满足数据分析的完整性要求。

(4)准确性原则。是指所收集的数据样本与具体应用目标和工作需求的关联程度的高低。关联程度越高，准确性越高，越能作为总体的一个样本。

(5)代表性原则。获取的数据样本应该具有代表性，能够反映数据分析所要研究问题的全貌。对于数据量小、数据容易获取的情况，收集所有数据进行数据分析是可行的，并能获取准确的分析数据；但对于数据量较大，很难获取全部数据的情况，或者数据量过大只能抽样(抽取其中部分数据)进行分析时，要求所获取的样本数据具有代表性。

(6)预测性原则。数据分析和数据挖掘的研究一般不只是为了根据存在的数据去分析已有的情况，往往还要根据现有的分析结果去预测未来的情况，因此数据收集既要着眼于现实的需求，又要具有一定的超前性，使其能够用于预测应用。

数据收集的方法与数据来源有密切关系，直接数据来源的收集方法简单直接，一般是通过

硬件设备直接采集，或者经过审批手续从部门内部获取，或者在协作单位的帮助下直接获取数据。间接来源的数据则需要通过一些方法来收集，其收集方法通常包括以下几种：

(1)当面询问法。当面询问法是依据事先拟定的调研提纲，以询问的方式，针对各个问题咨询相关专家或者了解该问题的人员，以获取问题的答案。所有的问题可能由某个人或某个部门来回答，或者由若干不同人员或不同部门来回答。

(2)问卷调查法。事先拟定问卷，发放给相关人群，对回收的问卷进行整理分析后获得数据集。具体的调查方式包括直接问卷调查法、互联网问卷调查法、E-mail 问卷调查法、电话问卷调查法等。例如，在某大型超市门口做直接问卷调查、通过互联网做互联网问卷调查、向相关人群发送 E-mail 做问卷调查，以及指定或随机选择调查对象通过电话向被调查者进行问询。

(3)会议调查法。事先拟定会议的主题和需要讨论的若干问题，邀请相关专家开会集中讨论问题。这种方法能集思广益，促进专家深入思考所要讨论的问题。

(4)收集互联网数据。从互联网获取所需要的数据，常见三种方式：①利用现有搜索引擎，通过设置相关关键字，多次检索，检索出所需要的数据，并进行适当整理；②从指定的网站下载相关数据，这适用于数据相对集中并容易下载的场合；③利用下载工具或自己编制爬虫软件，进行相关数据的抓取。

(5)收集政府、文献发布的数据。收集政府通过权威渠道，如统计年鉴、内部资料、官方网站等方式发布的政府认可的数据。此外，一些权威的书籍、论文中的数据通常可靠性也较高，也可用作数据分析。

间接获取数据的方法很多，前面提到的只是其中的几种，在实际应用中可以采用混合方法。无论采用哪些方法，一般都按照数据获取的原则来制订具体的获取方案。

1.2.3 数据展示

在数据分析和数据挖掘的研究阶段、数据分析结果展示阶段，常需要利用图表进行数据展示。数据可视化(Data visualization)可直观、形象地展示数据，其目的是以准确、直观、全面、高效(易与理解和易于分析)的方式展示数据。目前常用的数据展示方式包括表和图。科研人员在发表论文时广泛使用这两种方式，但这两种方式并不是唯一的，还可以用声音、动画、视频等方式进行展示。但声音、动画、视频等方式难以在纸制书籍、文章中展现，目前在与纸制文档类似的电子文档中也无法展现，例如，现有的 PDF(Portable Document Format)格式的电子文档、DOCX 格式的电子文档都无法展示。[①]互联网有更多可选择的展示方式，如听声音、看动画和视频等。

数据可视化可以帮助用户理解和运用数据。准确、直观、全面、高效是数据可视化的四个重要原则。“准确”既指数据来自真实数据或实际产生的数据，准确地展现了实际情况，又指能向观察者传递准确的信息和知识。“直观”是指数据的展示要突出所要传递的信息和知识的要点，简单直观地突出所要展示的主要内容。“全面”是相对的，是指尽可能向观察者提供利用图表等进行分析的数据。“高效”是指有利于观察者的理解和深入分析。可视化能将不可见的数据现象转化为可见的图形符号，能将错综复杂、看起来难以解释和关联的数据，建立起联系和关联，发现其规律和特征，获得更有商业价值的洞见和价值。观察者借助可视化的图表可以快速、准确地理解数据，还能够高效地运用数据，如寻找数据规律、分析推理、预测未来趋势。

① 本书出版时，这类电子版文档不直接支持声音、动画、视频，也许未来可能提供支持。

在研究阶段进行数据的可视化有助于研究者对数据进行了解、对数据的处理过程和处理方法进行分析、对研究的初步结果的性能进行有效性评价。在结果展示阶段，数据的可视化有助于向决策者、同事、下属提供直观的、易于解释的、易于决策分析的数据。图表是长期流行的数据可视化展示方式，在许多时候能够体现出“准确、直观、全面、高效”四个原则。表可以使用简单的(行列)二维表格、分组表格，或者构造更为复杂形式的表格。图的常见展示方式包括直方图、柱状图、饼图、盒图、条形图、折线图、曲面图、散点图、雷达图、气泡图、平行坐标图、树状图、透视地形图等。

1.3 数据的组织和数据的类型

1.3.1 数据的一般组织形式

统计学中学过几个重要概念：总体、个体、样本、样本容量、属性。总体(Population)是指所有考查对象的集合。个体(Individual)是总体中的每个考查对象。样本(Sample)是指由总体中取出的部分个体所组成的集合。样例(Sample case)即样本容量，是指样本中个体的数目。属性是对一个对象的抽象刻画，一个对象的属性就是该对象所具有的性质与关系。例如，想统计学校学生的某次英语考试成绩，那么学校内的所有学生构成总体，每个学生就是个体。如果只是抽查一部分同学的英语成绩，那么抽查的这些同学构成的集合就称为样本，样本容量就是所抽查的学生总数，属性就是学生的信息，如学号、姓名、性别、年龄、地区、英语分数。

数学集合概念中的元素具有确定性、互异性和无序性。元素的互异性要求集合中任意两个元素不能相同，而在数据分析的数据集(集合)中，可能存在一些个体在所考查的属性集中具有相同的信息。数据分析中的数据集代表着数据规范地组织在一起，允许重复对象存在。数据分析中的数据集通常只满足确定性和无序性，但这两个条件也并非绝对的，如果需要进行数据的不确定性分析，常采用概率论、模糊理论或粗糙集理论、不确定性推理等方法，所以“确定性”这一条件在某些条件下也允许不满足。如果做时间序列分析研究，数据一般按照时间顺序组织，那么此时“无序性”这一条件也不满足。综合来说，数据分析中的数据集只是强调数据按照对象排列组织在一起，常以二维表格形式组织，行代表个体(考查对象样例)，列代表对象的属性(特征)。

数据分析与数据挖掘的知识同以下多门课程有着密切的关系：①数据库和数据仓库，它们为数据分析提供基本的数据准备。②信息检索，这门课为数据的间接获取及一些数据分析方法提供方法参考和借鉴。例如，PageRank 方法虽然起源于信息检索的研究，但目前仍可看作一种经典的数据挖掘方法。③统计学，提供统计分析方法。它利用随机变量及其概率分布刻画目标类对象的行为。统计模型也广泛用于数据分析的建模中，如回归模型仍属于经典的数据分析方法。④机器学习，是研究如何利用计算机自动地从一组实例集合中学习知识，并加以运用。这门课与数据挖掘关系密切，本书中的关联分析、分类、聚类技术也直接应用于数据挖掘中。与之相似的还有一门课程——模式识别，它们有许多重复的内容。

由于数据分析与数据挖掘是从多门课程逐步发展而来的，所以某些概念来自多门课程，目前也尚未完全统一，仍旧保留多个学科中的原有概念。本书采用典型的若干概念的定义，先给出最基本的三个概念：①对象，有时也称个体、记录、实例、行；②属性，有时也称特征、字段、变量、维、列；③属性值，有时也称特征值、变量值、字段值/单元。表 1.2 给出了常见数据的行列组织形式。

表 1.2　常见数据的行列组织形式

	属性/特征/字段/变量/维/列–1	……	属性/特征/字段/变量/维/列–p
对象/个体/记录/实例/行–1	属性值/特征值/变量值/字段值/单元–1, 1	……	属性值/特征值/变量值/字段值/单元–1, p
……	……	……	……
对象/个体/记录/实例/行–n	属性值/特征值/变量值/字段值/单元–n, 1	……	属性值/特征值/变量值/字段值/单元–n, p

表 1.2 描述了 n 个个体/对象和 p 个属性/特征的数据记录组织形式。这种形式也可以使用矩阵的形式进行描述，称为样本矩阵[①]或对象属性矩阵，如图 1.1 所示。

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

图 1.1　样本矩阵的表示形式

图 1.1 中的行代表对应的个体/对象，列代表对应的属性/特征，而元素 x_{if} 代表个体/对象 i 的属性/特征 f 的取值。

1.3.2　数据类型

数据类型(属性值)可划分为两大类：分类数据和数值数据。在利用各个属性来表示个体对象的时候，一个属性列具有一样的数据类型。在数据库中进行表设计时需要指定各个字段的类型，结构化数据表示成行列数据时，各个列中也应是相同的数据类型，如表 1.2 中各列的数据类型相同。分类数据(Categorical data)用文本中的词、分类符号或者分类意义的数字串等形式进行描述，代表个体对象在该数据属性上的类别。例如，性别可用男和女来描述，头发颜色可用黑色、棕色、白色等来表述，教师的职称可用教授、副教授、讲师、助教来描述，一个学生所在地区可以使用省(自治区)或直辖市名称来描述。再比如，身份证号采用数字串来表示，性别也可以使用 0 和 1 来表示，但此处的数值具有分类的含义，一般不应使用四则运算(加减乘除)。常用的分类数据是离散数据(Discrete data)，它具有有限个不同值。数值数据(Numeric data)用可度量的量来描述，常用实数、整数形式，某些应用中也可以使用包括虚数等复数形式，或者构造向量或矩阵等复杂数据表示形式。

数据属性一般都用规范的数据类型来描述，在分类数据和数值数据的划分基础上，属性类型划分为四种：标称属性、二元属性、序数属性和数值属性，如图 1.2 所示。依照使用习惯，数值属性可分为两种子类型，但是大多数应用中只用到“数值属性”这一级别。

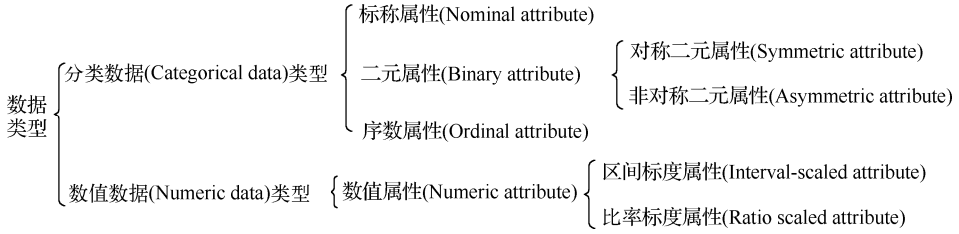


图 1.2　数据类型的分类

属性是一个数据字段，也是个体对象的一个特征，如表 1.2 中的各列。从数据的角度对属性类型进行划分，有助于有针对性地设计相应的操作(计算/度量)方法。

标称属性是分类数据中的一种，它的值是文本形式的类别名称，分类意义的符号或者分类

① 虽然有时数据集是总体，但仍习惯上称为样本，使用时需注意区分数据是总体还是样本。

意义的数字串，代表某种类别、编码或状态，如职业类型(教师、学生、医生等)、所属地区、头发颜色、身份证号码。标称属性的值是无序的，它们可看作枚举的值，对其进行数学四则运算是没有意义的。

二元属性是一种特殊的标称属性，只有两种类别，分别代表两种状态。这两种类别有多种描述形式，如“1”和“0”、“true”和“false”、“是”和“否”等，如是否是大学生可用 1 和 0 描述，也可用“是”和“否”描述；性别可用“男”和“女”来描述，也可用 1 和 0 等其他二值方式描述。本书中一般用 1 和 0 表示。如果二元属性的两种类别具有同等的重要程度，则称为对称的二元属性(Symmetric binary attribute)；如果两种类别的重要程度不同，则称为非对称的二元属性(Asymmetric binary attribute)。例如，对于是否是大学生同等对待，这时的两种状态(1 和 0)作用相同，可看作对称的二元属性；在疾病诊断中，是否吸烟采用 1 表示吸烟、0 表示不吸烟，假设更专注于吸烟的情况，此时 1 和 0 的作用程度不同，则可视作非对称的二元属性。

序数属性也是一种分类数据，但各类别存在顺序含义。例如，职称分为教授、副教授、讲师、助教，各个类别存在顺序关系；研究生、大学生、高中生、初中生、小学生、无学习经历也存在顺序关系；非常满意、满意、一般、不满意、非常不满意也存在顺序关系。序数属性在使用时不仅用于类别区分，还可用于类别之间的顺序关系。

数值属性是定量的，一般使用实数或整数。在某些特殊应用中，也可以使用虚数、复数等形式，甚至向量或矩阵等复杂数据表示形式。除使用实数或整数进行分类外，还可以划分为区间标度属性和比率标度属性。区间标度属性用相等的单位尺度度量，不具有倍数关系，加减法有意义而乘除法通常没有意义。区间标度属性中的 0 通常只代表一个参照点。例如，常说 2050 年比 2010 年多 40 年，但不会说 2050 年是 2010 年的 1.02 倍。当用摄氏度描述温度时，在 1 标准大气压下，纯净的冰水混合物的温度为 0℃，而沸腾的水温度为 100℃。虽然在数值上 100℃水温是 50℃水温的 2 倍，但在含义上一般只使用其长度标定，而不考虑倍数关系。比率标度属性具有倍数关系含义，数值的加减乘除都有意义。比率标度属性中的 0 常代表“空”“没有”的含义。例如，针对一个参照物，距离为 10m 是距离为 5m 的 2 倍，类似的还有速度、数量、容量、能量等属性。

数值属性是定量描述，而标称属性、二元属性和序数属性都是定性描述。除上述分类标准外，数据还可以按照数值是否连续，分为离散属性(Discrete attribute)和连续属性(Continuous attribute)。离散属性的各个值不连续，例如，职业、职称、颜色、地区等都属于离散属性；分类属性都是离散属性，如数值中的整数类型属性就是离散属性。连续属性一般都使用实数来表示，实数可能在整个实数空间或某段实数空间取值，如距离、温度等都属于连续属性。一般来说，实数类型通常都是连续属性，但若只利用实数中的某些离散点，则仍然属于离散属性。例如，满意度按照{0.2-非常不满意，0.4-不满意，0.6-一般，0.8-满意，1.0-非常满意}来度量，此时只是使用实数中的若干离散点，因此该属性仍然为离散属性。离散属性根据所有可能取值状态是否有限，还可进一步划分为有限状态离散属性和无限状态离散属性。

属性还可分为无量度属性和有量度属性两大类。无量度属性是离散属性，它指属性各可能取值之间没有大小关系，也没有先后序关系。标称属性、不考虑大小的整数编号、不考虑大小的若干离散实数取值等都可作为无量度属性，如头发颜色、性别、人员编号等。有量度属性可以是连续属性，也可以是离散属性，它是指属性各取值之间存在大小关系或者先后序关系。有大小关系的实数属性、有大小关系的整数、序数属性都是有量度属性。

1.3.3 分类数据的编码

为了便于数据处理，常对分类数据进行编码。标称属性的值本身没有顺序，但可以通过编

码形式转换为一组数据，便于计算机处理。例如，对职业进行编码{1-教师，2-学生，3-医生，4-护士，…}，对颜色进行编码{1-红色，2-橙色，3-黄色，4-绿色，…}。所谓标称属性编码，就是为标称属性中的各个类别进行编码，编码一般采用数字串或者字母、数字组合的形式。编码广泛应用在管理中，如学生号、身份证号等都使用编码形式。

标称属性编码的优点包括：①编码可以规范表示格式，如职业类型名称的描述长短不一，而编码之后长度统一，表示规范。规范的格式还有利于在数据库中存储。②编码便于计算机处理，如有助于用数组、向量、矩阵等形式来表示数据。例如，在班级所有学生的各门课成绩表示中，如果将学号按顺序排列后作为行，将各门课程按顺序排列后作为列，则可用矩阵形式来描述，这样便于应用矩阵上的统计分析操作。③编码可以形成唯一编号，如学号和身份证号都能解决人名重名的问题。标称属性编码中有两点需要注意：一是在设计编码时赋予一定的含义，如学号中常含有入学年份和班级编号，我国的身份证号前两位代表所在地区；二是在设计编码时需要考虑后续编码的使用，如前面的课程编码例子中，采用整数编码，便于形成矩阵的列索引。

二元属性常用 0 和 1 编码表示。一般 0 代表“false”“否”“不存在”等含义，1 代表“true”“是”“存在”等含义。例如，是否是大学生{0-否，1-是}，吸烟{0-不吸烟，1-吸烟}，性别{0-女，1-男}。

序数属性的各个类别之间存在序关系(Ranking)。一种是简单顺序编码，它只是给出序关系，如{0-无学习经历，1-小学生，2-中学生，3-大学生、4-研究生}；另一种方法不仅给出序关系，还关注各值的度量，即给出一种量化度量，包括等间隔量化序编码和复杂量化序编码，如{0.2-非常不满意，0.4-不满意，0.6-一般，0.8-满意，1.0-非常满意}。选用哪种方法需要根据数值分析的需要。如果只是利用序关系，则可以用简单顺序编码；如果还需要用程度量化值，就需要用量化序编码。序数属性只给出序关系，没有指定相继序数值之间的差值，而量化序编码是在序关系基础上额外指定量化值。等间隔量化序编码是将量化值等间隔映射到 0～1 区间的编码，设有 n 个序数，第 $k(1\leqslant k\leqslant n)$ 个序数量化值如式(1.1)所示。

$$\text{value}(k)=\begin{cases}k/n & \text{不允许存在 0}\\ (k-1)/(n-1) & \text{允许存在 0}\end{cases} \tag{1.1}$$

式(1.1)中，量化时如果允许存在 0，则采用$(k-1)/(n-1)$；如果不允许存在 0，则采用 k/n。

等间隔量化序编码只是一种编码方法，是在只给定序关系而没有其他信息的情况下，提供的一种度量。复杂量化序编码可以根据专家的经验或者基于数据分布统计等技术来提供。例如，假设更强调“一般”和“满意”，可以依经验赋予{0.2-非常不满意，0.4-不满意，0.8-一般，0.9-满意，1.0-非常满意}。如果使用专家经验，则可以借助“调查问卷”中的研究技术，如专家直接赋值、专家打分法等量化方法。如果使用数据分布统计技术，则可以根据各类别在数据中的分布比例，建立量化模型，实现各序数值的量化。

有一种基于数据分布统计的复杂量化序编码方法，具体过程如下：①统计数据集中 n 个序数属性值的出现频数(次数)，如表 1.3 中频数 $\text{Freq}=(5,8,15,13,9)$。②总频数为 $\text{sum}(\text{Freq})=50$，相对频率 $\text{RFreq}_i=\text{Freq}_i/\text{sum}(\text{Freq})$，累积频数 $\text{SFreq}_i=\text{Freq}_1+\cdots+\text{Freq}_i$。③概率值对照点 $\text{CP}_1=\text{RFreq}_1/2$，$\text{CP}_i=\text{RFreq}_i/2+\text{SFreq}_{i-1}/\text{sum}(\text{Freq}),(i>2)$。④假设数据满足正态分布，如图 1.3 所示，按正态分布概率查表获得 $P(x<x_i)=\text{CP}_i$ 中的每个 x_i 值。⑤如果需要量化值归一化到[0，1]区间，则可利用 $x_i^*=[x_i-\min(x)]/[\max(x)-\min(x)]$ 计算最终量化值 x^*。图 1.3 中，$q_i(i<n)$ 对应概率 CP_i 的分位点，用作参照，以便于理解。

表 1.3　复杂序数量化编码过程举例

序　数　值	非常不满意	不　满　意	一　　般	满　　意	非 常 满 意
频数(Freq)	5	8	15	13	9
相对频数(RFreq)	0.1	0.16	0.3	0.26	0.18
累积频数(SFreq)	0.1	0.26	0.56	0.82	1
概率值对照点(CP)	0.05	0.18	0.41	0.69	0.91
正态分布分位点(x)	–1.64485	–0.91537	–0.22754	0.49585	1.340755
最终归一化(x^*)	0	0.244335	0.474713	0.717008	1

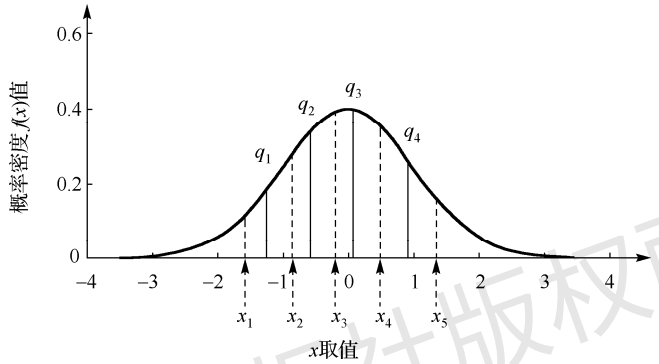


图 1.3　基于正态分布假设的复杂量化序编码示意图

1.4　数据的常用描述性统计量

1.4.1　数据的中心趋势

数据的中心趋势(Central tendency)是利用某一中心指标值来描述一组数据的集中位置，反映了一组数据中心点的位置。常用指标包括均值、中位数、众数、中列数和四分位中点。令一组数据采用 $X=\{x_1,x_2,x_3,\cdots,x_k,\cdots,x_n\}$ 来描述。举例，有四组数据用于后文的示例，A：{1, 2, 4, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 11, 15}，B：{15, 20, 22, 26, 28, 31, 32, 32, 33, 35, 38, 39}，C：{2, 9, 1, 0, 4, 6, 1, 2}，D：{29, 10, 4, 1, 6, 0, 2, 6, 2, 9}。常见的平均数包括三种：算术平均数(Arithmetic mean)、几何平均数(Geometric mean)和调和平均数(Harmonic mean)。常用的概念“均值”(Mean)和“平均数”是指算术平均数，如式(1.2)所示。

$$\text{Mean}(X)=E(X)=\frac{\text{sum}(X)}{\text{count}(X)}=\bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k=\frac{x_1+x_2+\cdots+x_n}{n} \tag{1.2}$$

式(1.2)中，$\bar{x}$ 代表 X 这组数的平均数，而如果将 X 看作一个随机变量，各数据代表各个取值，则平均数代表 X 的数学期望值 $E(X)$。几何平均数如式(1.3)所示，调和平均数如式(1.4)所示。

$$\text{Mean\_geom}(X)=\sqrt[n]{x_1x_2\cdots x_n} \tag{1.3}$$

$$\text{Mean\_harm}(X)=\frac{1}{\frac{1}{n}\times\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)} \tag{1.4}$$

按照式(1.2)～式(1.4)，A 组数据的算术平均数为 6.667、几何平均数为 5.670、调和平均数为 4.342。B 组、C 组和 D 组数据的算术平均数分别为 29.25、3.125 和 7.9。在 C 组数据中，由于存在 0，所以几何平均数为 0，而没有调和平均数。调和平均数易受极端值的影响，且受极小值的影响比受极大值的影响更大；只要有一个变量值为零，就不能计算调和平均数。

若一组数 X 中各个数在计算均值时的重要程度不同，则可为每个值 x_k 赋予一个权重 w_k，权重 w_k 反映 x_k 的重要性有计算加权均值(加权平均数)。加权均值的计算如式(1.5)所示。

$$\text{Mean\_weight}(X,W)=\frac{1}{n}\sum_{k=1}^{n}(w_k x_k)=\frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{n} \tag{1.5}$$

若一组数中的一些极端值对均值的计算影响较大，应用中还可以使用“截尾均值”，即去掉数值较大的若干个值和数值较小的若干个值后，再计算均值。

中位数(Median)也是一个中心趋势度量指标。如果 X 有序，则数值个数为 n：①若 n 为奇数，则中位数就是中间那个值，即第(n+1)/2 个数；②如果 n 为偶数，中位数就是中间那两个数的算术平均值[①]，即第 n/2 个数和第 n/2+1 个数的平均值。中位数的直观含义是：不考虑中位数本身时，有一半的数小于中位数，另一半的数大于中位数。例如，A 组中数据是有序的，数据的个数为 15 个(奇数)，因此选择位置居中的那个数，即第 7 个数作为中位数，计算依据为 7=(15+1)/2，中位数值为 6。

如果给定的数据 X 是无序的，常采用两种方法计算：①排序后取中间值。首先将这组数据从小到大排序或者从大到小排序，如果 n 是奇数时，取中间值，即第(n+1)/2 个数；如果 n 是偶数时，取中间两个值(第 n/2 个数和第 n/2+1 个数)的算术平均数。②如果存在计数统计，则可根据中位数左右两侧数值的个数相同这一含义，当 n 是奇数时，选择这样一个数，其左侧的数据个数占(n−1)/2；当 n 是偶数时，选择数值大小相同或邻近的两个数，这两个数的左侧占 n/2−1，然后取这两个数的中值作为中位数。例如，B 组数据中的 15, 20, 22, 26, 28, 31, 32, 32, 33, 35, 38, 39 已经排序，共 12 个数据，则中位数为中间位置的两个数的均值，即(31+32)/2=31.5。

众数(Mode)也是一个衡量中心趋势的指标，有时也称“模”。一组数据中，出现次数最多的数就叫这组数据的众数。只有 1 个峰值(1 个数值出现的次数最多)称为单模，有 2 个峰值称为 2 模，有多个峰值称为多模。即对应有两种可选方案：方案一，多众数法。众数可以是一个、多个或没有。如果有 2 个或 2 个以上数值出现次数都是最多的，则它们都是众数；而如果所有的数值出现次数都相同，则没有众数。例如，A 组数据的众数是 6，B 组数据的众数是 32，C 组数据的众数是 1 和 2，D 组数据的众数是 2，而{2, 9, 1, 0, 3, 5, 8, 7}中所有数值出现的次数都相同，所以没有众数。方案二，唯一众数法。如果数据中只有 1 个元素出现次数最多，则该元素是众数；如果多个元素出现次数都是最多，则取这些出现次数最多的元素的均值作为众数；如果所有元素出现次数都相同，则取全部元素的均值作为众数。

中列数(Midrange)也是一个衡量中心趋势的指标。中列数是数据集的最大值和最小值的平均值。例如，A 组数据的中列数是(1+15)/2=8，B 组数据的中列数是(15+39)/2=27。

1.4.2 数据的离散程度

离散程度(Measures of dispersion)是指一组数据中各取值之间的差异程度，有时也称分散程度。常用的离散程度度量指标包括方差、标准差、离差、平均绝对离差、极差、四分位极差。

① 有些软件在整数数组的中位数计算上，考虑输出为整数，所以当数组中的数为偶数个数时，只取中间两个数中较小那个数为中位数。

方差(Variance)广泛用于衡量数据的离散程度。概率论中的方差用来度量随机变量和其数学期望值(均值)之间的偏离程度。在统计学中，方差分为两种：①对于样本，方差的度量如式(1.6)所示，称为样本方差[①]；②对于总体，方差的度量如式(1.7)所示，称为总体方差。

$$\mathrm{Var}(X)=D(X)=S^2=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x}) \tag{1.6}$$

$$\mathrm{Var\_pop}(X)=\sigma^2=\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x}) \tag{1.7}$$

当数据分布比较分散(数据在平均数附近波动较大)时，各个数据与平均数的差的平方和较大，方差就较大；当数据分布比较集中时，各个数据与平均数的差的平方和较小。统计中由于许多数据集是抽样得到的，所以常用样本方差。

标准差(Standard deviation)也称标准偏差和均方差。计算上，标准差是对方差进行开平方运算。由样本方差开平方之后得到样本标准差，如式(1.8)所示，而总体方差开平方之后得到总体标准差，如式(1.9)所示。

$$\mathrm{stdev}(X)=S=\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x})} \tag{1.8}$$

$$\mathrm{stdev\_pop}(X)=\sigma=\sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x})} \tag{1.9}$$

如果一组数据来自一个随机变量，按照大数定律切比雪夫不等式，一个观测值一般不会远离均值超过标准差的数倍，最少 $(1-1/k^2)\times 100\%$ 的观测值不会超过 k 个标准差。这也解释了统计中一组数据常呈现集中分布的现象。当然，如果明确知道数据的分布，则估算通常会更准确些。例如，在正态分布中，以均值为中心处于正负标准差之内的数据达到 68%，处于正、负 2 倍标准差之内的数据达到 95%，处于正、负 3 倍标准差之内的数据达到 97.7%。

离差(Deviation)指各个数值到特定的参照点(通常是均值，特定情况下也可能是中位数等)之间的距离之和，如式(1.10)所示。

$$\mathrm{Deviation}(X)=\sum_{k=1}^{n}\left|x_k-\bar{x}\right| \tag{1.10}$$

平均绝对离差(Mean absolute deviation)是对离差取算术平均数，又称平均离差或标准离差，如式(1.11)所示。

$$\mathrm{Deviation\_Mean}(X)=\frac{1}{n}\sum_{k=1}^{n}\left|x_k-\bar{x}\right| \tag{1.11}$$

离差平方和(Sum of square deviation)，又称和方差，是指各个数值到特定的均值之间的距离的平方和，如式(1.12)所示。与离差中使用绝对值计算相比较，使用平方将放大偏离中心数值的作用。

$$\mathrm{SSD}(X)=\sum_{k=1}^{n}(x_k-\bar{x})^2 \tag{1.12}$$

① 样本方差中分母处以 $n-1$ 是为了获得总体方差的无偏估计，详见概率论或统计学。

极差(Range)是一组数据中的最大值减去最小值，也称全距，是数据分散程度的一个度量指标。相比方差、标准方差、离差等度量指标，该指标更容易受极值的噪声影响。

分位数(Quantile)是利用若干个数据点将有序数组 X 划分为数值个数相等的若干份，而这些点称为分位数。常用的分位数为四分位数(Quartiles)、十分位数(Deciles)和百分位数(Percentiles)。中位数也是二分位数的分位点。图 1.4 展示了四分位数的位置，它利用 3 个点 Q_1 ~ Q_3 将有序数划分为四个部分，每个部分占 25%。Q_1、Q_2、Q_3 左侧分别为 25%、50%、75%。其中，Q_2 为中位数。因此，四分位点包括 Q_1、Q_2(中位数)和 Q_3。

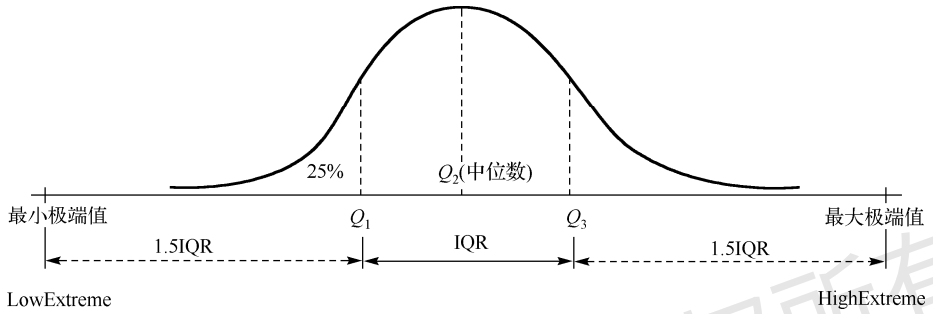


图 1.4　四分位数的分位点示意图

与四分位数类似，十分位数是利用 9 个分位点(数据点)将有序数组 X 划分为 10 份，每份占 10%。百分位数是利用 99 个分位点将有序数组 X 划分为 100 份，每份占 1%。由于 X 的数值个数 n 不一定恰好可以按照整数位置进行等份划分，所以对于分位数的分位点计算，如果不能够等比例划分，就采用最可能接近等比例的划分方法。分位数一般有三种计算方法：加一划分法、减一划分法和近似划分法。

(1)加一划分法：基于 $n+1$ 确定四分位数的位置，即

$$Q_1\text{的位置}=(n+1)\times 1/4=(n+1)\times 25\%$$
$$Q_2\text{的位置}=(n+1)\times 2/4=(n+1)\times 50\%$$
$$Q_3\text{的位置}=(n+1)\times 3/4=(n+1)\times 75\%$$

(2)减一划分法：基于 $n-1$ 确定四分位数的位置，即

$$Q_1\text{的位置}=1+(n-1)\times 1/4=1+(n-1)\times 25\%$$
$$Q_2\text{的位置}=1+(n-1)\times 2/4=1+(n-1)\times 50\%$$
$$Q_3\text{的位置}=1+(n-1)\times 3/4=1+(n-1)\times 75\%$$

(3)近似划分法：基于 $n+1$ 计算，再“下取整数”，令 floor() 为下取整函数，即

$$Q_1\text{的位置}=\text{floor}((n+1)\times 1/4)=\text{floor}((n+1)\times 25\%)$$
$$Q_2\text{的位置}=\text{floor}((n+1)\times 2/4)=\text{floor}((n+1)\times 50\%)$$
$$Q_3\text{的位置}=\text{floor}((n+1)\times 3/4)=\text{floor}((n+1)\times 75\%)$$

本书默认采用“加一划分法”。例如，对于 A：{1, 2, 4, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 11, 15}，该组数据有序，共 15 个数据，计算四分位数的各分位点的位置：

$$Q_1\text{的位置}=(n+1)\times 1/4=(15+1)\times 25\%=4$$
$$Q_2\text{的位置}=(n+1)\times 2/4=(15+1)\times 50\%=8$$
$$Q_3\text{的位置}=(n+1)\times 3/4=(15+1)\times 75\%=12$$

计算 X 中对应位置的值，得到 Q_1 值为 $x_4=5$，Q_2 值为 $x_8=6$，Q_3 值为 $x_{12}=8$。再看 B 组有序数据{15, 20, 22, 26, 28, 31, 32, 32, 33, 35, 38, 39}，共有 12 个数，计算分位点的位置：

$$Q_1\text{的位置}=(n+1)\times 1/4=(12+1)\times 25\%=3.25=3+0.25$$
$$Q_2\text{的位置}=(n+1)\times 2/4=(12+1)\times 50\%=6.5=6+0.5$$

Q_3 的位置$=(n+1)\times 3/4=(13+1)\times 75\%=9.75=9+0.75$

可见，分位点的位置不是整数，可采用线性插值法计算，Q_1 的位置是 3+0.25，代表分位点处于第 3 个数 $x_3=22$ 到第 4 个数 $x_4=26$ 的连线上，并且处在以 x_3 出发到 x_4 的 0.25 倍的长度位置，因此利用线性插值法 $Q_1=x_{3.25}=x_3+(x_4-x_3)\times 0.25=22+(26-22)\times 0.25=23$。同理，$Q_2=x_{6.5}=31+(32-31)\times 0.5=31.5$，由于 Q_2 就是中位数，所以也可以按照中位数计算方法计算。$Q_3=x_{9.5}=33+(35-33)\times 0.75=34.5$。

与四分位数的分位点计算方式类似，可计算十分位数、百分位数或其他分位数的分位点。

四分位中点(Inter-Quartile Midpoint，IQM)，也称四分位中心，$\mathrm{IQM}=(Q_1+Q_3)/2$。IQM 代表了到 Q_1 和 Q_3 距离相等的点，即两侧四分位点的中点。例如，对于 A 组数据，IQM=$(Q_1+Q_3)/2$=(5+8)/2=6.5；对于 B 组数据 IQM=$(Q_1+Q_3)/2$=(23+34.5)/2=28.75。

四分位极差(Inter-Quartile Range，IQR)，也称四分位间距，还称内距，$\mathrm{IQR}=Q_3-Q_1$。IQR 反映了中间 50%数据的离散程度，如图 1.4 所示。相比于极差(全距)使用最大值减去最小值易于受极值的不稳定性干扰，四分位极差表示的离散程度相对稳定，相当于去掉较小的 25%数据和较大的 25%数据后的离散程度度量，属于一种截尾方法。对于 A 组数据，IQR=Q_3-Q_1=8−5=3；对于 B 组数据，IQR=Q_3-Q_1=34.5−23=11.5。

IQR 还可以用于离群点的计算，离群点(Outlier)是一个数据值，它显著地偏离于数据群体。如图 1.4 所示，通常将1.5IQR 作为一个阈值，将低于 $\mathrm{LowExtreme}=Q_1-1.5\mathrm{IQR}$ 的数值视作离群点，将高于 $\mathrm{HighExtreme}=Q_3+1.5\mathrm{IQR}$ 的数值也视作离群点。以 A 组数据为例，$Q_1-1.5\mathrm{IQR}=5-1.5\times 3=0.5$，$Q_3+1.5\mathrm{IQR}=8+1.5\times 3=12.5$，所以超过[0.5, 12.5]区间的数都可视作离群点，于是 A 组数据中的 15 就是离群点。再以 D 组数据为例，排序后为{0, 1, 2, 2, 4, 6, 6, 9, 20, 29}，Q_1=1.75，Q_3=11.75，因此 IQR=Q_3-Q_1=10，$Q_1-1.5\mathrm{IQR}=-13.25$，$Q_3+1.5\mathrm{IQR}=26.75$，因此低于−13.25 和超过 26.75 的数值都视作离群点，于是 29 视作离群点。

1.4.3 数据的形态统计量

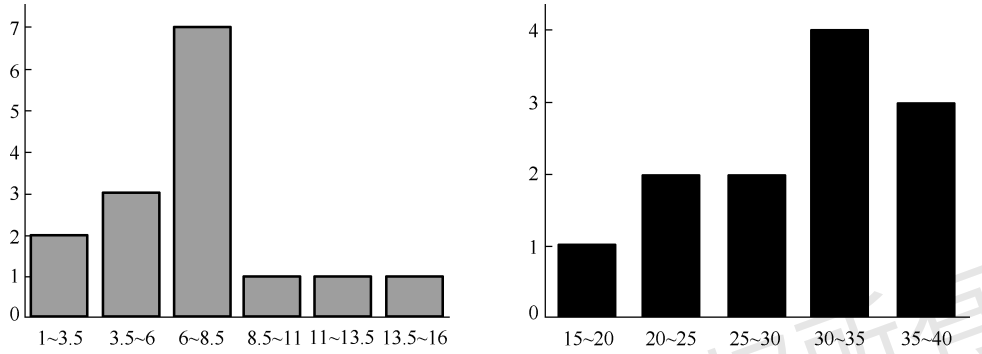
频数(Frequency)是指针对某一个属性在各个属性值上的对象个数，用于展现在该属性上各属性取值的分布情况。相对频数(Relative frequency)，又称频率(Rated frequency)，等于频数除以对象总数，即等于每个属性值上对象个数的百分比。

例如，对表 1.4 中的“性别”属性进行频数统计，存在两个取值“男”“女”，频数依次为 4 和 2；对“年龄”进行频数统计，所有取值包括 8、9、10，对应的频数依次为 1、3、2；对“数学”进行频数统计，所有取值包括 89、93、95、97，对应的频数依次为 1、2、2、1。在频数统计上，有时还划分区间统计区间频数(Interval frequency)，特别是对于实数属性，进行合适大小的区间划分有助于利用频数观察总体情况。相对区间频数(Relative interval frequency)等于区间频数除以对象总数。例如，对表 1.4 中的“英语”按照 85～89、90～94、95～100 进行区间频数统计，频数依次为 2、2、2，相对区间频数依次为 33.3%、33.3%、33.3%。

表 1.4 学生成绩单

姓　名	性　别	年　龄	语　文	数　学	英　语	总 成 绩	评　价
Jiang	男	10	98	97	97	292	优
Tom	男	9	96	95	97	288	优
Jerry	女	9	88	95	85	268	良
Samb	男	9	92	93	92	277	良
Susam	女	10	86	89	88	263	良
Lucas	男	8	95	93	91	279	优

条形图(Bar diagram，Bar chart)可用于直观地展示频数分析结果。条形图可以横向，也可以纵向。纵向的条形图又称柱状图(Column diagram，Column chart)。例如，将 A 组数据{1, 2, 4, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 11, 15}按照 2.5 间隔从 0 开始统计区间频数，则各区间[1, 3.5)、[3.5, 6)、[6, 8.5)、[8.5, 11)、[11, 13.5)、[13.5, 16)的频数依次为 2、3、7、1、1、1。B 组数据{15, 20, 22, 26, 28, 31, 32, 32, 33, 35, 38, 39}这里按照[15, 20)、[20, 25)、[25, 30)、[30, 35)、[35, 40)分段进行统计区间频数，依次为 1、2、2、4、3。A 组和 B 组数据的频数图如图 1.5 所示。



(a) A组数据　　(b) B组数据

图 1.5　A 组和 B 组数据的频数图

峰度(Kurtosis)又称峰度系数，是一个用于描述数据分布形态的统计量。它描述某属性上各属性值分布的形态，即陡缓程度，其计算如式(1.13)所示。该统计量与正态分布相比较来度量峰度的尖锐度或平坦度，峰度为 0 表示该数据总体分布与正态分布的陡缓程度相同；峰度大于 0 表示该数据总体分布与正态分布相比较为陡峭(比正态分布更尖锐)，为尖顶峰；峰度小于 0 表示该数据总体分布与正态分布相比较为平坦(比正态分布更平坦)，为平顶峰。峰度的绝对值数值越大，表示其分布形态的陡缓程度与正态分布的差异程度越大。

$$\text{Kurtosis}(X)=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{k=1}^{n}((x_k-\bar{x})/\text{stdev}(X))^4-\frac{3(n-1)(n-1)}{(n-2)(n-3)} \tag{1.13}$$

式(1.13)需满足两个条件：数据点个数 n 大于或等于 4，样本标准偏差 $\text{stdev}(X)$ 大于 0。如果 $n<4$ 或 $\text{stdev}(X)=0$，则式(1.13)无法计算峰度值。还有些统计学教材中采用了其他计算峰度的公式，如式(1.14)所示。

$$\text{Kurtosis\_simplify}(X)=\frac{1}{n-1}\sum_{k=1}^{n}((x_k-\bar{x})/\text{stdev}(X))^4-3 \tag{1.14}$$

例如，A 组数据的 Kurtosis 系数按式(1.13)计算为 1.575462259，而按照式(1.14)计算的结果为 0.474050468，由于计算结果为正数，表明这组数据的形态相比于正态分布更为尖锐。B 组数据的 Kurtosis 系数按式(1.13)计算为–0.325816472，而按照式(1.14)计算的结果为–0.861047964，由于计算结果为负数，表明这组数据的形态，相比正态分布更为平坦。图 1.5 也直观地展示了 A 组和 B 组数据的峰度情况。本书默认按照式(1.13)进行计算，与 SPSS 软件和 Excel 软件一致。

偏度(Skewness)又称偏度系数、偏斜度或偏态，是一个用于描述样本分布形态的统计量。它描述某属性取值分布的对称性，表明分布相对于平均值的不对称程度，计算如式(1.15)所示。该统计量与正态分布相比较，偏度为 0 表示其数据分布形态与正态分布偏度相同，可视作左右对称(Symmetric)；偏度大于 0 表示正偏差数值较大，为正偏或右偏，正偏斜度表明分布的不对称尾部趋向于更多正值，有一条长尾巴拖在右边，如图 1.6(a)所示；偏度小于 0 表示负偏差数

值大，为负偏或左偏，负偏斜度表明分布的不对称尾部趋向于更多负值，有一条长尾巴拖在左边，如图 1.6(b)所示。偏度的绝对值越大，表示分布形体的偏斜程度越大。

$$\text{Skewness}(X)=\frac{n}{(n-1)(n-2)}\sum_{k=1}^{n}((x_k-\bar{x})/\text{stdev}(X))^3 \tag{1.15}$$

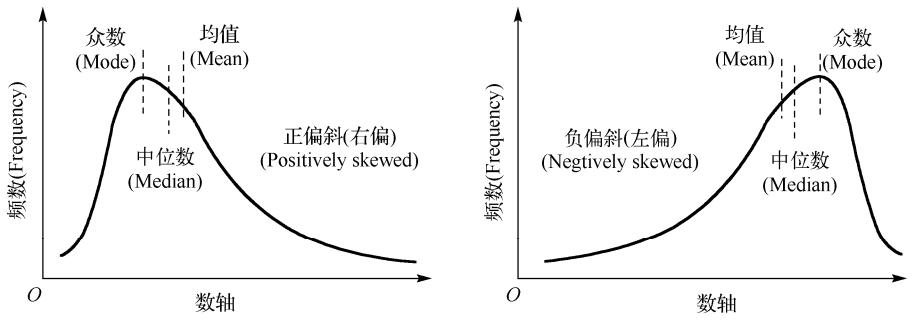


图 1.6　偏度的右偏与左偏情况示意图

式(1.15)代表样本的偏度计算，它需要满足两个条件：数据点个数 n 大于或等于 3，样本标准偏差 stdev(X)大于零。如果不满足条件，则无法计算。在总体上进行偏度计算与式(1.15)略有不同，总体的偏度计算如式(1.16)所示。

$$\text{Skewness\_pop}(X)=\frac{1}{n}\sum_{k=1}^{n}((x_k-\bar{x})/\text{stdev\_pop}(X))^3 \tag{1.16}$$

有些统计教材对于样本的偏度采用如式(1.17)的计算。

$$\text{Skewness\_simplify}(X)=\frac{1}{(n-1)}\sum_{k=1}^{n}((x_k-\bar{x})/\text{stdev}(X))^3 \tag{1.17}$$

本书采用式(1.15)计算样本的偏度，采用式(1.16)计算总体的偏度，这与 SPSS 和 Excel 软件的计算方式一致。例如，A 组数据，式(1.15)计算的样本偏度为 0.741023415，式(1.16)计算的总体偏度为 0.664761158，式(1.17)计算的样本偏度为 0.642220293，数值为正数，表明是正偏或右偏，右侧有一条长尾。图 1.5(a)也直观地展现了该情况，A 组数据的右侧尾部较长。B 组数据，式(1.15)计算的样本偏度为–0.614991131，式(1.16)计算的总体偏度为–0.535281073，式(1.17)计算的样本偏度为–0.512492609，可见该组数据属于负偏或左偏，即左侧有一条相对长的尾部。图 1.5(b)也直观地展示出左侧的长尾现象较明显。

除使用偏度系数进行偏度描述外，还可以使用中位数(Median)、均值(Mean)的大小关系进行偏度估计，可参考图 1.6。例如，通过前面的计算得到，对于 A 组数据，Median=6，Mean=6.667，Mode = 6，Median<Mean，说明峰部数据(中间一些集中数据)在均值中心的左侧，如图 1.5(a)所示，较尖锐的峰部在均值中心的左侧，这时可以估计出右侧有相对较长的尾部，数据呈现右偏，图 1.6(a)给出了相应示意图。对于 B 组数据，Median=31.5, Mean=29.25, Mode=32，Median> Mean 说明峰部数据在均值中心的右侧，如图 1.5(b)所示，所以通常左侧有一个相对较长的尾部，数据呈现左偏，图 1.6(b)给出了相应示意图。有三点需要强调：①使用中位数和均值之间的关系进行数据偏斜度度量只是一种估计方法，数据量充足的情况下通常比较准确，可作为数据偏斜程度的初步判别。需要注意该方法也可能受到极端数据的噪声干扰。②使用偏度统计量进行偏斜度度量更准确些，并且还能给出偏斜程度的度量。③中位数、均值和众数在数据量足够大

的统计数据中，通常满足一定的序关系，参看图 1.6。如果数据偏斜度接近 0，说明接近对称图形(Symmetric)，则此时中位数、均值和众数值相同或接近；如果数据偏斜度大于 0，数据右偏，参看图 1.6(a)，此时一般有 Mode<Median<Mean；如果数据偏斜度小于 0，数据左偏，参看图 1.6(b)，此时一般有 Mean<Median<Mode。有学者的研究中给出一些统计数据上满足的经验关系： $\text{Mode} - \text{Mean} \approx 3(\text{Median} - \text{Mean})$ 。

1.5 数据的基本描述性统计分析

1.5.1 数据的描述性统计

描述性统计(Descriptive statistics)分析是指使用一些统计量来描述一组数据的总体概况，通常包括如下几方面信息：①基本信息；②中心趋势；③离散程度；④形态描述。具体来说，统计量常包括数据个数、最小值、最大值、均值、中位数、方差、标准差、极差、四分位极差、偏度、峰度。例如，A 组数据{1, 2, 4, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 11, 15}的描述性统计如表 1.5 所示。

表 1.5 A 组数据的描述性统计

样本数	均值	中位数	方差	标准差	最小值	最大值	极差	四分位极差	偏度	峰度
15	6.6667	6.0000	11.810	3.43650	1.00	15.00	14.00	3.00	0.741	1.575

B 组数据{10, 13, 20, 30, 40, 45, 50, 38, 30, 20, 14, 10}的描述性统计如表 1.6 所示。

表 1.6 B 组数据的描述性统计

样本数	均值	中位数	方差	标准差	最小值	最大值	极差	四分位极差	偏度	峰度
12	29.2500	31.5000	53.659	7.32524	15.00	39.00	24.00	11.50	–0.615	–0.326

描述性统计只是对数据总体情况进行概括，因此只包括一些常用的宏观统计量。具体应用还有可能关心一些其他的宏观统计量，如众数、离差等。

频数分析也是描述性统计中常用的工作，表 1.7 给出了 A 组数据的频数统计。累积频率(Cumulativerate frequency)又称累积百分比(Cumulative percent)，是对频率的累积计算，常用于计算数据的经验分布。

表 1.7 A 组数据的频数统计

数 值 (Data)	频 数 (Frequency)	相对频数(%) (Percent)	累积百分比(%) (Cumulative percent)
1	1	6.7	6.7
2	1	6.7	13.3
4	1	6.7	20.0
5	2	13.3	33.3
6	3	20.0	53.3
7	2	13.3	66.7
8	2	13.3	80.0
9	1	6.7	86.7
11	1	6.7	93.3
15	1	6.7	100.0

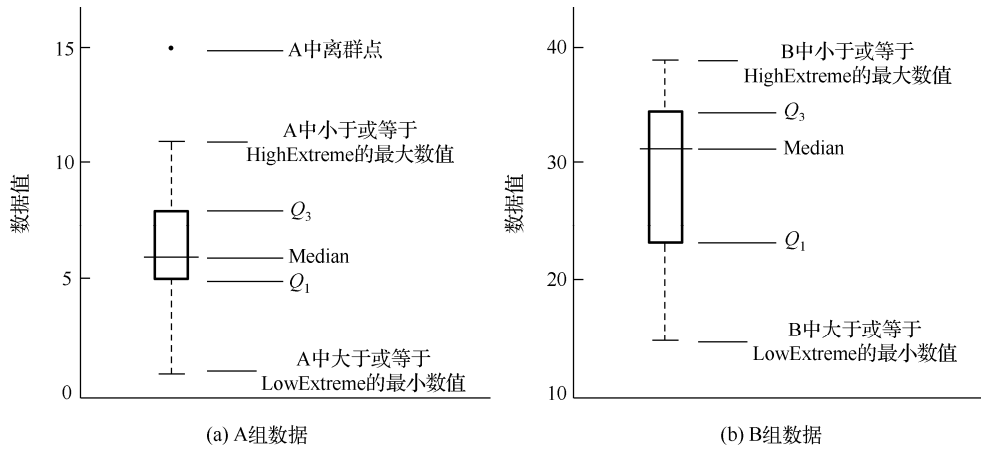
为了解数据分布特性，在获得数据后，进行描述性统计是很重要的工作。通过数据的频数分析、数据的集中趋势分析、数据离散程度分析、数据的分布分析，以及一些基本的统计图形，来揭示数据分布的特征。

1.5.2 五数概括与盒图

有时候，希望能够简洁地描述出一组数据的总体情况。五数概括(Five-number summary)由最小值(Min)、四分位数的第一分位点 Q_1、中位数(Median)(也是四分位数第二分位点 Q_2)、四分位数第三分位点 Q_3 和最大值(Max)组成。

尽管五数概括只使用五个数来大致描述数据的总体情况(Min、Q_1、Median、Q_3、Max)，却描述了多种统计信息：①描述了中心趋势。中位数衡量了中心趋势，还可以计算中位数 (Min + Max)/2，计算四分位中心 $\mathrm{IQM}=(Q_1+Q_3)/2$ 用作中心趋势的度量指标。②描述了离散程度。极差(全距)可由 Max − Min 计算得到，四分位极差 $\mathrm{IQR}=Q_3-Q_1$。③描述了基本形态。虽然五数概括没有提供均值，但可以计算四分位中心 IQM=(Q_1+Q_3)/2，然后通过中位数和 IQM 之间的关系估计数据的偏斜情况。例如，前面已计算出 A 组数据的 IQM=6.5，Median=6，可见峰值位置在四分位中心的左侧，估计右侧有一个长尾，因此可以估计 A 组数据为右偏。B 组数据的 IQM= 28.75，Median=31.5，可见峰值位置在四分位中心的右侧，估计左侧有一个长尾，因此可以估计 A 组数据为左偏。关于峰度情况也可以适当估计，根据极差 (Max − Min) 和四分位极差 (Q_3-Q_1) 的比例关系来估计，但该峰度情况的估计只供参考，这是因为数据中的极小值和极大值有时受噪声干扰较大。④判别是否存在离群点。五数概括可以计算[$Q_1-1.5\mathrm{IQR}$，$Q_3+1.5\mathrm{IQR}$]的两个区间端点，计算最小值和最大值是否处在区间之内，来判别是否存在离群点。

盒图(Boxplot)，又称箱线图，是基于五数概括的可视化展示方式，也是一种常用的数据可视化图形。例如，A 组数据{1, 2, 4, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 11, 15}的五数概括为：Min=1，Q_1=5，Median=6，Q_3=8，Max=15，盒图如图 1.7(a)所示；B 组数据{15, 20, 22, 26, 28, 31, 32, 32, 33, 35, 38, 39}的五数概括为：Min=15，Q_1=23，Median=31.5，Q_3=34.5，Max=39，盒图如图 1.7(b)所示。



(a) A组数据　　(b) B组数据

图 1.7　A 组数据与 B 组数据的盒图

盒图的绘制过程如下：

(1)绘制一个矩形框(盒)，下边框对应 Q_1 的值，上边框对应 Q_3 的值。如图 1.7(a)中，因为 A 组数据的 Q_1=5，Q_3=8，所以矩形的下边框对应刻度 5，上边框对应 8；图 1.7(b)中，因为 B 组数据的 Q_1=23，Q_3=34.5，所以矩形的下边框对应 23，上边框对应 34.5。

(2)在盒中绘制一条横线，对应中位数(Median)值。例如，A 组数据的 Median=6，B 组数据的 Median=31.5。

(3)依据四分位极差 IQR，计算最小极端值(LowExtreme)和最大极端值(HighExtreme)。例如，A 组数据 $\mathrm{IQR}=Q_3-Q_1=8-5=3$，$\mathrm{LowExtreme}=Q_1-1.5\mathrm{IQR}=5-1.5\times3=0.5$，$\mathrm{HighExtreme}=Q_3+1.5\mathrm{IQR}=8+1.5\times3=12.5$；B 组数据 $\mathrm{IQR}=Q_3-Q_1=34.5-23=11.5$，$\mathrm{LowExtreme}=Q_1-1.5\mathrm{IQR}=23-1.5\times11.5=5.75$，$\mathrm{HighExtreme}=Q_3+1.5\mathrm{IQR}=34.5+1.5\times11.5=51.75$。

(4)计算数组中不小于 LowExtreme 的最小值，将该值作为最下边界线(绘制一段横线)，再从矩形框向下边框引出虚线至最下边界线，而数组中小于 LowExtreme 的各个数将作为离群点单独以小圆点的形式在对应的位置分别单独绘制出来。例如，A 组数据和 B 组数据中没有低于 LowExtreme 的离群点，因此不必单独绘制。

(5)计算数组中不大于 HighExtreme 的最大值，将该值作为最上边界线(绘制一段横线)，再从矩形框向上边框引出虚线至最上边界线，而数组中大于 HighExtreme 的各个数将作为离群点单独以小圆点的形式在对应的位置分别单独绘制出来。例如，A 组数据中的 15 超过 HighExtreme 值 12.5，因此图 1.7(a)单独将离群点 15 以小圆点绘制出来；B 组数据没有离群点，因此图 1.7(b)中没有需要单独绘制的小圆点。

1.5.3 数据的描述性统计图

描述性统计中以图的形式展示一些统计信息，属于数据可视化的一种方法，对数据分析和数据挖掘有重要作用。例如，盒图可以展示比五数概括更多的统计信息，并且易于直观感受中心趋势、离散程度、数据形态、离群点等信息。

直方图(Histogram)，由一系列高度不等的纵向柱状条组成，其高度代表着相应数值的大小。图 1.8 所示是 A 组数据和 B 组数据的频数直方图，与图 1.5 使用的条形图比较，频数分布直方图的特点如下：

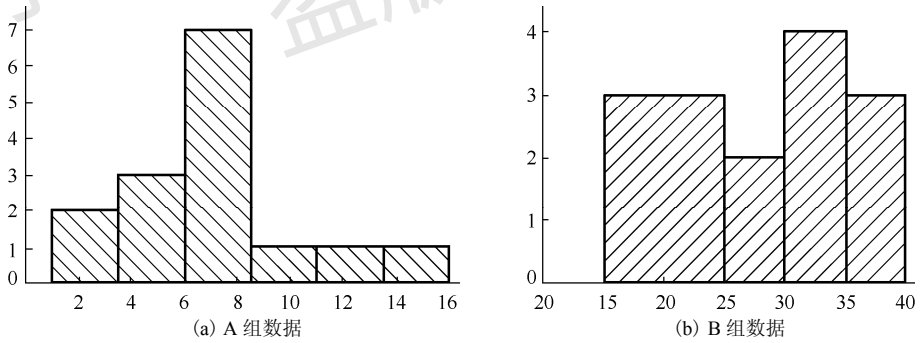


图 1.8　A 组和 B 组数据的频数直方图

(1)直方图中条状之间无间隔，而条形图的同一组数据的条状之间存在间隔。

(2)直方图中横轴是连续的实数，各条状的宽度对应着统计区间的跨度。例如，图 1.8(a)来自图 1.5 的数据，利用每个条状跨度 2.5 来表示各条状所对应的区间。在条形图中，横轴上的数据是离散的，如图 1.5(a)所示，所以各条状之间不连接。

(3)直方图用条状高度表示数据的值，而条形图通过条状的面积表示值。只有各条状的宽度相同时，才相当于利用高度表示值，因此通常条形图的宽度都相同，以易于理解。例如，图 1.8(a)的与图 1.5(a)相同，而图 1.8(b)特意将[15, 20)与[20, 25)区间合并为[15, 25)一个区间，使用直方图的一个条状来表示，此时，该条状的跨度从 15 到 25，而高度 3 代表[15, 25)区间内的频数(等于原来合并的两个区间对应的两个数值之和)，这也表明直方图是用高度描述值的。

直方图可以描述比五数概括更多的信息，常用于直观地展示划分的各个区间频数。直方图的跨度大小设置较为重要，如果数据量足够大，则跨度小些可以展示更多的数据分布细节；而跨度大些则条状数少，故展示的数据分布细节少一些。

分位数图(Quantile plot)是一种常用的展示属性数据的图形。横轴通常取值 0～1，代表分位点左侧的百分比(Percent value)，纵坐标对应着相应分位点的取值。例如，四分位图的三个点的横坐标对应着 0.25、0.5 和 0.75，分别代表 Q_1 左侧有 25%的数据量、Q_2 左侧有 50%的数据量、Q_3 左侧有 75%的数据量，而四分位图上标记的三个点分别为$(0.25, Q_1)$、$(0.5, Q_2)$和$(0.75, Q_3)$。q-分位数图上有 $q-1$ 个点，分别是$(k/q, Q_k)$。为了观察更多的点，常使用软件绘制百分位数图。分位数图常用于两种情况：①当数据量特别大的时候，利用分位数只抽取其中若干重要点的数据，按照顺序(分位点从前到后顺序)进行展示，有助于观察数据总体情况。例如，利用百分位数图绘制出上百万个数的数据总体分布情况。②在原始数据图中描述若干个重要数据点，以便进行关键点趋势分析。例如，在上千个数据构成的图形中，增加十分位数点，可增加图形的关键点描述。图 1.9 所示为某家超市的商品价格数据的十分位数图。

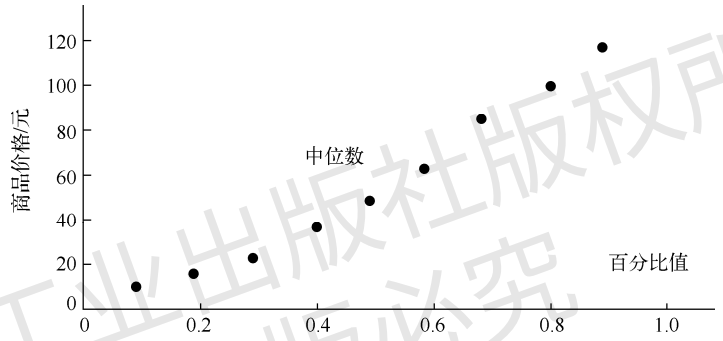


图 1.9　某家超市的商品价格数据的十分位数图

分位数-分位数图(Q-Q 图)常用于对比两组数据的分布情况，横轴描述一组分位数，纵轴描述另一组分位数。Q-Q 图的绘制方法是：①对两组数据计算 q-分位数，如百分位数。假设有两组数据 X 和 Y，它们各自所包括的数据个数可能相同也可能不同。令 XA 表示 X 的 q-分位数，令 YB 表示 Y 的 q-分位数，这样 XA 代表着 X 的主要数据分布情况，YB 代表着 Y 的主要数据分布情况。更重要的是，XA 和 YB 都按照从小到大排序，并且二者的数据个数相同，都是 q-1 个。②将两组分位数形成数据对，绘制到图形中。将 XA 和 YB 中的数从前到后依次配对$(\mathrm{XA}_k, \mathrm{YB}_k)$，其中 $k \in [1, q-1]$，将数据点$(\mathrm{XA}_k, \mathrm{YB}_k)$绘制到 Q-Q 图中，这就完成了 Q-Q 图的绘制。例如，有两家大型超市，所销售的商品总体上相似，以人们日常生活需求的商品为主。现在想对比一下这两家超市的销售价格高低情况。假设能够获取两家商品的全部价格，就可以对每家超市的价格数据计算十分位数[①]，这样获得 XA 和 YB 两组十分位数，分别代表着第一家价格的十分位数和第二家价格的十分位数，然后绘制 Q-Q 图，如图 1.10 所示。

在图 1.10(a)中，以中位数(Median)为例，X 超市的销售单价大约为 50 元，而 Y 超市的销售单价大约为 60 元，因此 Y 超市相对较贵。类似地，可以分析其他分位点。在图 1.10(b)中，X 超市销售价格相对较高些。值得注意的是：①这里没有针对每个相同商品进行一一比较，而是采用分位数进行对比，对比的前提是必须两家超市所销售的商品大致相同，而且商品数量非常多，比分位点的个数要多得多。②Q-Q 图相当于将对两组数据进行分布上的对比。

Q-Q 图常用于两种情况：①将一组数据与另一组已知数据进行对比，判断是否有相似的数

① 本例中使用十分位数是为了便于绘图讲述，实践中可以采用百分位数。

据分布。②观察一种数据是否满足一种已知的分布。如果将两组数据 X 和 Y 看作两个随机变量，Q-Q 图就是利用两组分位数，将两个随机变量的分布组合绘制在一张图上，是一条以分位数间隔为参数的曲线。如果两个分布相似，则该 Q-Q 图趋近于落在 $y=x$ 线上；如果两分布线性相关，则点在 Q-Q 图上趋近于落在一条直线上，但不一定在 $y=x$ 线上。例如，在图 1.10 中，如果两个超市的销售单价接近，则曲线上的点应该近似落在 $y=x$ 线上。有时还利用 Q-Q 图观察样本数据 X 是否近似于正态分布，这时只需将 X 和一组正态分布采样数据 Y 绘制成 Q-Q 图，然后看 Q-Q 图上的点是否近似地在一条直线附近。如果在直线附近，则初步判别符合正态分布，否则可能不符合正态分布，更进一步判别的可能需要做统计检验。

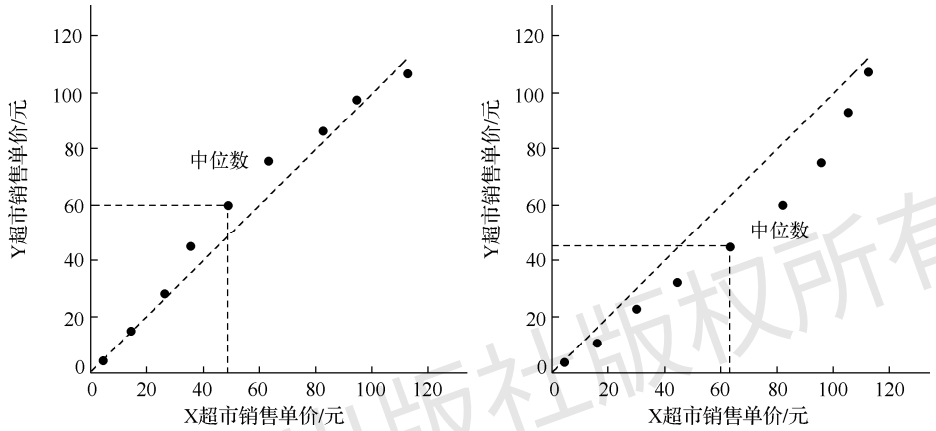


图 1.10　两个超市的销售单价对比

1.6 本 章 小 结

本章讲述了数据分析和数据挖掘的需求。随着计算机和互联网的广泛应用，人类已进入数据时代，互联网+、大数据、人工智能等热点和重点研究方向层出不穷。各行各业广泛进行数据分析和数据挖掘已经成为一项基本要求。本章还介绍了一般数据分析和数据挖掘的主要工作过程，列出了数据的收集原则，阐述了数据的常见分类。数据常分为四种类型：标称属性、二元属性、序数属性和数值属性，有时也按照离散属性和连续属性进行划分。

对于给定的一组数据进行基本的描述性统计是认识和了解数据的基础，中心趋势的相关统计量包括均值、中位数、众数、中列数、四分位中心。我们常说的均值是指算术平均值，此外还有几何平均值和调和平均值。离散程度度量的统计量包括样本方差、样本标准差、总体方差、总体标准差、极差(全距)、四分位极差(内距)、离差、平均绝对离差。描述数据形态的常用统计量包括峰度系数、偏度系数、频数。

五数概括和盒图也常用于数据分析和数据挖掘，用于对数据的总体进行描述。五数概括虽然只使用五个数，但可以描述数据的中心趋势、离散程度、基本形态和是否有离群点，盒图在五数概括的基础上提供直观的可视化展示。直方图、分位数图、Q-Q 图也常用于数据的可视化展示。

本章概念与关键词

数据分析　　数据挖掘　　大数据　　知识驱动方法　　数据驱动方法

总体	样本	对象	属性	特征
属性值	特征值	样本矩阵	标称属性	二元属性
序数属性	数值属性	离散属性	连续属性	对称二元属性
非对称二元属性	区间标度属性	比率标度属性	标称属性编码	二元属性编码
序数属性编码	均值	中位数	众数	中列数
四分位中点	方差	标准差	离差	平均绝对离差
极差(全距)	四分位极差(内距)	离群点	频数	区间频数
峰度系数	偏度系数	四分位数	描述性统计	五数概括
盒图	直方图	条形图	分位数图	
分位数-分位数图(Q-Q 图)				

练习与思考

1．举例说明数据分析和数据挖掘的重要性。

2．思考你是否遇到过数据分析对你造成误导的情况。你关心个人数据隐私吗？

3．简述数据分析和数据挖掘的主要工作过程。

4．数据收集的常用原则包括哪些？为什么有这些原则要求？

5．请举例说明数据的四种类型：标称属性、二元属性、序数属性和数值属性。

6．试比较离散属性和连续属性。

7．数据的中心趋势常用度量指标有哪些？

8．数据的离散程度常用度量指标有哪些？

9．数据描述中的五数概括包括哪五个指标？思考从五数概括中能分析出哪些信息。

10．对一组数据{2, 9, 1, 0, 3, 5, 8, 7, 5, 1, 8, 1}，手工计算均值、中位数、众数、最小值、最大值、第一四分位数、第三四分位数、四分位极差，并绘制盒图。

11．思考计算截尾平均数的作用是什么。如何计算 5%截尾平均数？

12．请说明峰度系数近似 0、大于 0 和小于 0 的含义，偏度系数近似 0、大于 0 和小于 0 的含义。

13．思考直方图、分位数图和 Q-Q 图常用于哪些分析的直观展示。

14．简述利用加一划分法计算四分位点的计算过程。

15．四分位极差(内距)如何计算？四分位极差能克服边缘数据不稳定问题，可较好衡量数据的离散程度。简述如何利用四分位极差计算离群点数据。

16．直方图的一般作用是什么？直方图中的分组数会影响图形的形状，请查阅资料说明一般如何选择分组数。

17．有人说直方图与盒图的作用有一定的相关性。请从生活中或互联网上收集至少 30 个数据，利用软件绘制直方图与盒图，并简要分析二者在展示分布上的异同。

18．统计学中属性的类型常划分为定性属性、定量属性。分析第 5 题中四种属性类型如何归到这两种类型中。

19．借助编程或软件计算第 10 题，并统计频数，绘制频数条形图。

20．借助编程或软件，绘制一组数据的盒图、直方图、分位数图、频数条形图。