



计算机类专业“互联网+”创新型精品教材

数据分析方法

# 数据分析方法

主 编 沈 琳 张节群 彭严松

主  
编  
沈  
琳  
张  
节  
群  
彭  
严  
松

北京出版集团  
北京出版社

北京出版集团  
北京出版社

图书在版编目 (CIP) 数据

数据分析方法 / 沈琳, 张节群, 彭严松主编.  
北京: 北京出版社, 2024.8. -- ISBN 978-7-200-18801-1

I . 017

中国国家版本馆 CIP 数据核字第 202406J002 号

数据分析方法

SHUJU FENXI FANGFA

主 编: 沈 琳 张节群 彭严松

出 版: 北京出版集团

北 京 出 版 社

地 址: 北京北三环中路 6 号

邮 编: 100120

网 址: www.bph.com.cn

总 发 行: 北京出版集团

经 销: 新华书店

印 刷: 定州启航印刷有限公司

版 印 次: 2024 年 8 月第 1 版 2024 年 8 月第 1 次印刷

成品尺寸: 185 毫米 × 260 毫米

印 张: 12

字 数: 270 千字

书 号: ISBN 978-7-200-18801-1

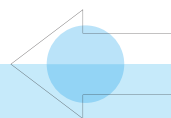
定 价: 39.00 元

教材意见建议接收方式: 010-58572341 邮箱: jiaocai@bphg.com.cn

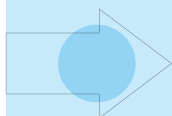
如有印装质量问题, 由本社负责调换

质量监督电话: 010-82685218 010-58572341 010-58572393

|                    |            |
|--------------------|------------|
| <b>项目一 初识数据分析</b>  | <b>1</b>   |
| 任务一 数据分析的认知        | 2          |
| 任务二 数据分析流程         | 5          |
| <b>项目二 数据描述性分析</b> | <b>17</b>  |
| 任务一 一维数据的数字特征      | 18         |
| 任务二 数据的分布          | 25         |
| 任务三 多维数据的数字特征及相关分析 | 36         |
| <b>项目三 线性回归分析</b>  | <b>45</b>  |
| 任务一 一元线性回归         | 46         |
| 任务二 多元线性回归分析       | 50         |
| <b>项目四 对比分析</b>    | <b>80</b>  |
| 任务一 对比分析的图表工具      | 81         |
| 任务二 内部对比分析和外部对比分析  | 89         |
| 任务三 综合评价           | 95         |
| 任务四 数据报告的制作        | 102        |
| <b>项目五 聚类分析</b>    | <b>110</b> |
| 任务一 相似程度的度量        | 111        |
| 任务二 快速聚类法          | 129        |
| <b>项目六 主成分分析</b>   | <b>136</b> |
| 任务一 主成分分析的认知       | 137        |
| 任务二 主成分分析的基本原理     | 139        |



|                           |            |
|---------------------------|------------|
| 任务三 主成分分析的计算步骤及应用举例 ..... | 142        |
| <b>项目七 相关分析.....</b>      | <b>149</b> |
| 任务一 相关分析相关基础知识 .....      | 150        |
| 任务二 皮尔逊相关系数与秩相关系数 .....   | 155        |
| 任务三 对应分析 .....            | 160        |
| <b>项目八 时间序列分析.....</b>    | <b>164</b> |
| 任务一 时间序列的认知 .....         | 165        |
| 任务二 向量自回归模型 .....         | 168        |
| 任务三 因素分解模型 .....          | 172        |
| 任务四 基于循环神经网络的时间序列分析 ..... | 175        |
| <b>参考文献.....</b>          | <b>184</b> |





## 项目一

# 初识数据分析

当今社会，处处充满数据，“没有调查，就没有发言权”“用数据说话”已成为人们的口头禅。人们也常说“凡事要做到心中有数”，这里的“数”就是“数据”。数据分析是指从数据中提取有用的信息，并指导实践。本项目主要学习各种常见的数据形式和分析方法以及数据分析的一般流程。

### 知识目标

- » 了解数据的定义与分类。
- » 掌握数据分析的方法。
- » 熟悉数据分析流程。

### 能力目标

- » 能进行基本的操作和数据处理。
- » 能够识别数据中的关键信息和潜在规律，为后续的深入分析打下基础。
- » 能够将所学理论知识运用于实际的数据分析项目中。
- » 能够根据实际需求和项目特点，合理规划数据分析的流程。

### 素养目标

- » 让学生认识到在数据处理和传递过程中应遵守的伦理规范和法律法规，承担起相应的社会责任。
- » 在数据处理和分析过程中，学生需要遵循客观、公正、真实的原则，不得篡改或捏造数据，树立正确的价值观和道德观，培养诚实守信、尊重事实的良好品质。
- » 让学生认识到数据在国家治理、经济发展、科技创新等方面的重要作用，强化学生的国家意识和民族自豪感，激发他们为国家 and 民族的发展贡献力量的热情。

#### 任务描述

数据无处不在，同时又形式多样。合理的数据分析，能够帮助人们提取出数据中蕴含的有价值的信息，并辅助人们做出合理的决策。本任务学习数据的定义、分类以及数据分析技术与方法。

#### 任务目标

1. 能够识别并区分不同类型的数据。
2. 能够更高效地处理和分析数据。
3. 能根据不同的数据类型和特性，选择准确的数据分析方法。

#### 任务实施

### 一、数据的定义与分类

#### （一）数据的定义

数据泛指描述事物的符号记录，是构成信息和知识的原材料。实际应用中数据的形式多种多样，除了数字，还包括文字、图像、声音、影像等。随着信息技术的高速发展和大数据时代的到来，数据不再是社会生产的“副产物”，而是已经成为可以被二次乃至多次加工利用的生产资料。

实际生活、学习和工作中，数据已无处不在。例如，用户想买某电子产品时，在京东或淘宝上搜索了相关的关键词，这些行为数据会被系统记录下来。系统后台会对这些行为数据进行分析，并结合用户个人和相关人群的历史行为进行用户画像，进而产生个性化的推荐结果，帮助用户快速地从海量候选商品中找到自己需要且喜欢的商品。数据的收集和使用正在各个方面和领域不断地改善着人们的日常生活。例如，智慧城市平台会利用城市的实时交通信息数据、社交网络信息和天气预报数据来优化交通管理和调度。

数据既是构成信息和知识的原材料，也是信息和知识的表现形式和载体。原始的数据通常是杂乱无章、毫无规律的。为了从数据（data）中挖掘出有价值的信息（information），进而形成结论或知识（knowledge），并帮助人们在特定的场景下做出合理的决策（智慧，wisdom），需要进行相应的数据分析。这四者的关系可以用 DIKW 金字塔结构进行表示，如图 1-1 所示。为了传递和利用数据分析所得到的结果（信息和知识），还需要通过数据对其进行表达和记录。

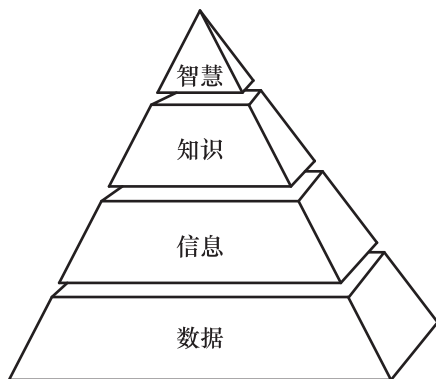


图 1-1 DIKW 金字塔示意图

## （二）数据的测量尺度及类型

对于不同的研究对象，人们往往针对其某些性质展开研究，因此能够用来计量或测度的标准和程度也是不同的。有些事物只能对其属性进行分类和研究，如人的性别、婚姻状况、受教育程度，产品的型号和质量等级等；有些事物可以用比较精确的数字进行计量，如产品的重量、长度、价值，收入的多少等。根据计量学的分类方法，按照研究对象的不同性质和计量的精确程度，可以采取不同的方式将变量分为四个层次：定类变量、定序变量、定距变量和定比变量。它们所采用的计量尺度由低级到高级，由粗略到精确。不同层次的统计数据适用于不同的统计分析方法。

### 1. 定类变量

定类变量是按照事物的某种属性对其进行平行的分类或分组。定类变量的取值只是测度事物之间的类别差，无法反映各类之间的其他差别。例如，根据性别将人口分为男、女两类；按照经济性质将企业分为国有企业、集体企业、私营企业、合资企业、独资企业等。使用定类尺度对事物进行分类时，它们必须符合穷尽和互斥的要求。类别穷尽是指在所做的全部分类中，必须保证每一个元素都能归属于某一类别，不能有所遗漏；类别互斥是指每一个元素或个体只能归属于一个类别，而不能在其他类别中重复出现。为了分析上的方便，对于定类变量的取值也可以用数字来表示，但这些数字不能区分大小，也不能进行数学运算。

### 2. 定序变量

定序变量的取值用于描述事物之间的等级差或顺序差别，该类变量不仅可以将事物分成不同的类别，而且还可以确定这些类别的优劣或顺序，但不能测量出类别之间的准确差值。例如，在收购苹果时会将苹果分为直径 65 cm、70 cm、75 cm、80 cm、85 cm 以上等，产品等级就是对产品质量优劣的一种次序测度。定序变量测量的结果只能比较顺序，不能进行数学运算，尽管它们的取值是用数字表示的。

由定类变量和定序变量形成的数据说明的是事物的品质特征，不能用数值表示，其结果均表现为类别，也称为定性数据或品质数据。

### 3. 定距变量

定距变量不仅能将事物区分为不同类型并进行排序,而且可以准确地指出类别之间的差距是多少。例如,收入用元度量(如月收入8 000元、5 000元等),质量用克度量,考试成绩用百分制度量等。定距变量的每一间隔都是相等的,可以进行加减运算。

### 4. 定比变量

定比变量除了具有上述三种变量尺度的全部特征,还具有一个特性,即可以计算出两个测度值之间的比值。这要求定比尺度中必须有一个绝对固定的“零点”,这也是它与定距变量的唯一差别。

由定距变量和定比变量形成的数据说明的是现象的数量特征,能够用数值来表示,也称为定量数据或数量数据。

区分测量的层次和数据的类型是非常重要的,因为针对不同的数据类型可以采取不同的统计方法进行处理和分析。例如,对于定类变量的数据,通常可以计算出各组的频数或频率,计算众数和变异系数,进行列联表分析和 $\chi^2$ 检验等;对于定序变量形成的数据,可以计算中位数和四分位数,进行等级相关系数等非参数分析;对于定距变量和定比变量形成的数据,可以用更多的统计方法如参数估计和检验等进行处理。需要特别指出的是,人们所处理的数据大多为数量数据,适用于低层次测量数据的统计方法同样也适用于较高层次的测量数据,因为后者具有前者的数字特征。例如,在描述数据的集中趋势时,对定类数据通常是计算众数,对定序数据通常是计算中位数,同样,对定距数据和定比数据也可以计算众数和中位数。但是,适用于较高层次测量数据的统计方法则不能用于较低层次的测量数据,因为较低层次的数据不具备较高层次测量数据的数字特征。

## 二、数据分析技术与方法

数据分析是指采用适当的方法和技术对收集得到的数据进行探索和分析,以提取有价值的信息并形成结论或知识的一系列过程。其目的是把隐藏在一大批看起来杂乱无章的数据中的信息萃取和提炼出来,以找出所研究对象的内在特性或规律。在实践中,数据分析可帮助人们定量地做出判断,以便采取适当的决策和行动。

数据分析这一概念涵盖多种技术,并已广泛应用于商业、科学、社会学等不同领域。例如,数据挖掘(data mining)关注数据建模与知识发现的过程。商务智能(business intelligence)聚焦商业信息的提取与利用,通过将企业中现有的数据转化为信息和知识,帮助企业做出明智的业务经营决策,以实现商业价值。文本分析(text analytics)关注非结构化的文本数据,综合应用统计学、语言学和结构化分析等技术对文本数据进行分析。概括而言,数据分析就是系统地应用统计、逻辑等各种技术来描述、说明、浓缩、概括、评价和利用数据的过程。不同类型的数据分析技术具有不同的侧重点。

在实际应用中,数据分析可以帮助人们做出正确的判断和合理的决策。例如,约翰



揭秘日常数据  
背后的故事

尼斯·开普勒(Johannes Kepler)通过分析天文学家第谷·布拉赫(Tycho Brahe)20多年观测与收集到的天文数据,不仅证实了哥白尼的日心说,而且还将行星的运动轨迹从圆形修正为椭圆形,进而得出了行星运动的三大定律。在产品研发、设计、推广和售后的整个生命周期中,从市场调研到售后服务等各个环节都需要进行数据分析,以实现对用户群体和市场的全面刻画。企业的管理者需要通过市场调查和相应的数据分析来掌握市场动向,从而制定合适的生产和销售计划。例如,全球零售业巨头沃尔玛在对消费者购物行为数据进行分析时发现,男性顾客在购买婴儿尿布时,常常会顺便购买几瓶啤酒。于是推出了将尿布和啤酒摆在一起的促销手段,使得两种商品的销量都大幅增加。

数据分析有多种不同的形式和方法。从不同角度,可以对数据分析方法进行不同的划分。常用的数据分析方法大多是以数理统计为基础,可以分为描述性分析、探索性分析和推断性分析三大类。描述性分析是借助各种描述指标对数据进行概要分析,以获得反映客观现象整体的各种定量特征,主要包括集中趋势分析、离中趋势分析和分布形状分析。描述性分析的目的是提供一套简单有效的概括和表征数据的方法。探索性分析是一种数据驱动的分析方法,不依赖于任何分布假设和复杂理论,完全从客观数据出发,通过图、表等可视化方式探索数据内在的规律和模式,主要包括直方图、箱形图、时序图等。探索性分析能够帮助人们检查数据预处理的结果,理解数据,并提出合理的数据分布和模型假设。推断性分析是指根据带有随机性的(样本)观测数据,基于合理的(模型)假设,对未知事物(总体)做出推断,主要包括参数估计、假设检验、方差分析等。推断性分析的基本思想是根据(少量)样本数据推断(大量)总体的特征。

根据分析目的的不同,数据分析方法可以分为回顾性分析和预测性分析。回顾性分析是以现在为结果,回溯过去的分析方法。回顾性分析是医学和心理学常用的数据分析方法。预测性分析的本质是根据事物的过去和现在预测未来,即根据已知预测未知,从而减少对未来事物认识的不确定性,以指导我们的决策行动,减少决策的盲目性。常用的预测性分析方法有回归分析、分类分析、时间序列预测等。

根据数据类型和特性的不同,数据分析方法可以分为数值型数据分析、分类型数据分析、顺序型数据分析;横截面数据分析、时间序列分析、面板数据分析;文本分析、语音分析、图像分析、视频分析等。

## 任务二

## 数据分析流程

### 任务描述

一个完整的数据分析流程包含多个不同的环节,每个环节的作用各不相同,但整体却密不可分。为了确保数据分析工作的高效、有序进行,制订一个明确的数据分析流程是必不可少的。本任务学习数据分析的基本流程以及各步骤之间的关系与作用。

## 任务目标

1. 能够发现数据中隐藏的模式、趋势和关联，从而提炼出有价值的信息。
2. 能够培养对问题进行整体把握和分析的能力。
3. 能够识别关键变量、建立假设，并设计出合适的数据分析策略来解决问题。
4. 能提升编写数据分析报告、展示分析结果的能力。

## 任务实施

一般而言，数据分析的基本流程主要包括以下步骤：问题定义与数据收集、数据预处理、描述性与探索性分析、推断性分析、基于模型与算法的分析、数据分析报告与应用等，如图 1-2 所示。其中，问题定义与数据收集是前提，数据预处理和描述性与探索性分析是基础，推断性分析和基于模型与算法的分析是核心，数据分析报告与应用是结果。

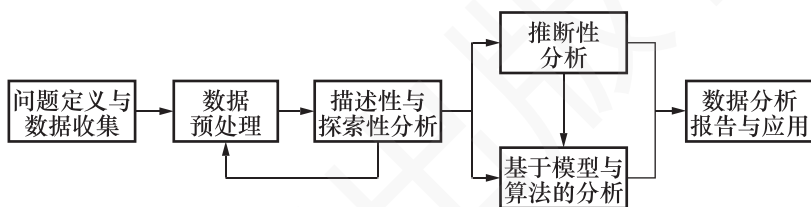


图 1-2 数据分析流程

### 一、问题定义与数据收集

问题定义是确保数据分析过程有效性的一个基础环节，主要包含两部分内容：确定目标和划定边界。数据分析师和实际决策者应根据具体应用和实际决策的需求，确定数据分析的目标，这是数据分析流程顺利进行的先决条件。若分析目标不明确，陷入“为了分析而分析”的误区，后续的分析流程将无法进行或有效进行。

错误的问题定义将导致毫无意义或错误的分析结果。例如，一家汽车制造商决定开展一个情感分析项目，为期 6 个月，耗资一千万美元。项目结束后，该厂商将分析结果分享给经销商并试图改变销售模式。但是，所得出的结果最终被证明是错误的。后来发现，项目失败的主要原因是项目团队没有花足够的时间去了解经销商所面临的真正问题和业务需求，从而导致其分析结果毫无价值。

实际中的问题往往是由目标与现实之间存在的差距或矛盾所产生的。在定义问题时，首先需要结合现实情况，明确差距或矛盾所在，从而挖掘出有意义、有价值的问题。但矛盾是普遍存在的，世间万物也是普遍联系的，在解决一个具体问题时，无法考虑到所有可能的影响因素。因此，在定义问题时需要做出取舍，通过边界划定确定需要考虑的主要相关因素，而忽略那些我们认为（或假设）不重要的因素。其次要对问题进行明确的、可量化的描述，需要将非量化的描述词汇转化为具有确定标准的可量化指



标,例如,将“汽车销量高”转化为“汽车销售量每年超过 100 万台”。只有将问题定义清楚,才能有针对性地开展后续的数据分析流程,从而得出合理的结论。

数据收集能够帮助数据分析师和实际决策者更好地定义问题,并为后续的数据分析提供数据资料。实际应用中,数据可以从各种不同的来源进行获取。可以从指定的内部数据库中进行采集,采用爬虫技术从互联网上进行爬取,利用环境中的传感器(如摄像机、GPS 等)进行收集,或通过访谈和问卷调查进行获取等。除了获取所需的数据,还需要了解数据的原始面貌,包括数据产生的时间、条件、格式、内容、限制条件等。这些信息能够帮助人们更有针对性地控制数据生产和采集的过程,避免由于违反数据采集规则导致的数据问题。

对数据采集逻辑的认识能够增加人们对数据的理解程度,尤其是数据中的异常变化。在数据采集阶段,数据分析师需要更多地了解数据生产和采集过程中的异常情况,如此才能更好地追本溯源。另外,这也能避免“无用数据进、无用数据出”(garbage in, garbage out)的无效劳动。

问题定义和数据收集二者密不可分,互相依赖。实际应用中的问题都来源于物理世界。通过问题定义能够指导人们进行数据收集,采用实验与观测等方法从物理世界中获取问题分析与求解所需的数据。同时,通过数据收集可以帮助人们理解业务、验证问题定义的合理性,并帮助人们修正和完善问题的定义。问题定义、数据收集与物理世界之间的关系,如图 1-3 所示。

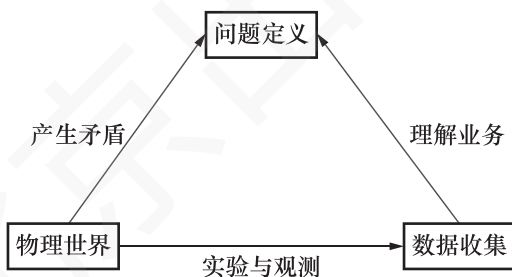


图 1-3 问题定义、数据收集与物理世界之间的关系示意图

## 二、数据预处理

数据预处理是指综合运用数据清理、数据集成、数据归约、数据变换等多种处理方法,将各种原始数据加工成人们需要的、标准的、干净的数据的过程。现实世界中的数据有结构化的数值表格,也有非结构化的文本和图像,而且大多是不完整、不一致的脏数据,包含噪声和异常数据,且存在数据缺失、量纲不统一等问题。

面对类型不同、质量各异的数据,数据分析师难以或是无法直接进行数据分析,需要通过数据预处理检测并纠正数据中损坏的或不准确的数据,并对数据进行变换以达到适应或匹配模型需求的目的。数据预处理能大大提高数据的质量,进而提高后续数据分析的效率。

数据预处理主要包括数据清理、数据集成、数据归约和数据变换等处理方法，如图 1-4 所示。

## （一）数据清理

数据清理是指对数据进行重新审查和校验的过程，其目的是检测并纠正损坏的或不准确的数据，为后续的分析提供干净的、一致的数据。如果决策者认为数据是脏乱的，则他们很难相信基于这些数据所分析或挖掘出来的结果，即会认为输出的结果是不可靠的。数据清理主要包括缺失值处理、异常值处理、不一致性纠正等。

数据缺失是指在信息和数据的获取过程中，由于各种原因产生的数据空缺或丢失。数据缺失是一种常见现象。针对数据的缺失值，通常会根据变量的分布特性和变量的重要性（信息量和预测能力）采用不同的处理方法。常用的缺失值处理方法包括变量删除、样本删除、定值填充、统计量填充、插值法填充、模型填充、哑变量填充等。

异常值是指数据的值处于特定分布区域或范围之外。异常值分为两种：伪异常和真异常。伪异常是由特定的业务运营动作而产生的，是正常反映业务状态的数据，而不是数据本身的异常。真异常是指由于操作、测量、记录、处理等环节中出现错误而导致的数据本身的异常。异常值处理的目的是对异常数据进行检测并去除或修正。常用的异常值检测方法有简单统计法（如箱形图法）、基于距离的方法、基于密度的方法和基于聚类的方法等。常用的异常值处理方法有简单删除、对数变换、平均值替换、中位数替换等。

不一致性是指原始数据中包含矛盾的或不相容的数据或信息。造成数据不一致性的主要原因有数据冗余、并发控制不当、各种故障或错误等。可以通过定义一致性约束来检测数据中的不一致性，也可以根据每个变量的合理取值范围和相互关系来检查数据是否符合要求，进而找出那些超出正常范围、逻辑不合理或者相互矛盾的数据。例如，用 1~7 级量表测量的变量值中出现了 0 值、体重数据中出现了负数等，都应被视为超出正常值域范围。逻辑不一致的数据可能以多种形式出现。例如，调查对象说自己开车上班，又报告自己没有汽车；或者调查对象称自己是某品牌的重度购买者和使用者，但同时又在该品牌的熟悉度上给了很低的分值。针对不一致数据，可以根据关联数据结合领域知识进行纠正，也可以把不一致数据当作缺失数据，并采用缺失值处理方法进行处理。

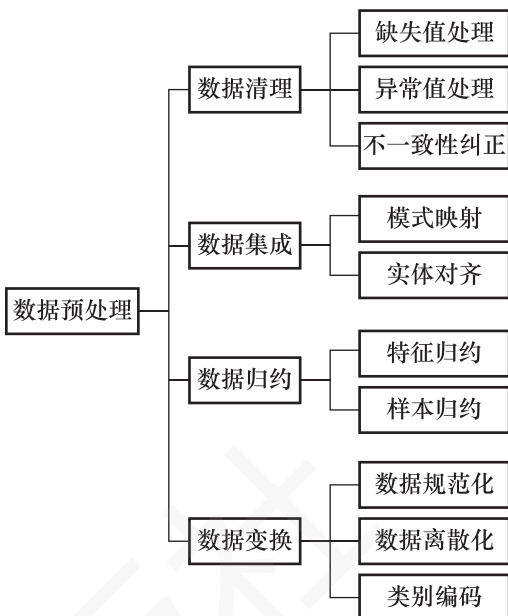


图 1-4 数据预处理方法



## （二）数据集成

数据集成是指将不同来源的数据进行合并，以得到一个完整的、一致的数据集。数据集成是后续数据分析和数据挖掘的基础。数据集成过程中可能会出现冗余数据或不一致数据，需要采用数据清理中的相应方法进行处理，否则可能会降低后续数据分析的效率或误导数据分析的进程。数据集成主要包括模式映射和实体对齐。

模式映射是指将来自不同数据源的表示相同含义的字段或属性映射为同一字段。在数据库中，模式（schema）是描述数据库结构的一组元数据，包括表、字段（列）、关系等。描述同一个概念的字段或属性在不同数据库中的取名可能不同，例如针对学生编号，有的数据库用“学号”，而有的则用“学生 ID”。直接通过字段拼接对这些数据进行集成会导致数据冗余，进而影响后续的数据分析。模式映射可以采用基于专家知识的手工映射，也可以采用基于内容匹配的自动映射。

实体对齐的目标是把不同数据源中同一实体的不同记录进行匹配和对齐。当存在统一的唯一标识时，例如用户身份证号，可以通过标识匹配进行实体对齐；否则需要借助其他可用信息（例如实体属性值）进行实体对齐。在实体对齐过程中可能会出现数据冲突或不一致，例如针对同一用户的学历属性，一个数据源中的记录为本科，而另一个数据源中的记录则为研究生，此时需要采用数据清理中相应的方法对不一致数据进行修正。

## （三）数据归约

数据归约是指在尽可能保留数据所包含的有用信息的前提下，最大限度地精简数据量。数据归约主要包括特征归约和样本归约，分别针对原始数据集中的属性和记录进行归约。

特征归约，也称维度归约，是指从原有的特征集中删除冗余的或不相关的特征，或是通过对特征进行重组来减少特征的个数，进而降低数据维度。常用的特征归约方法可以分为两大类：特征选择和特征构造。前者是直接从原始特征集中筛选出一个子集，而后者则是根据原始特征集信息构造出一组新的特征。

样本归约是指从数据集中选出一个具有代表性的样本子集。子集大小的确定需要综合考虑计算成本、存储要求、空间覆盖以及其他一些与算法和数据特性有关的因素。常用的样本归约方法包括随机抽样、分层抽样、系统抽样等。

## （四）数据变换

数据变换是指对数据进行变形和转换，使其适用于后续的数据分析和数据挖掘。数据变换主要包括数据规范化、数据离散化和类别编码。

数据规范化是指将数据平移并按比例进行缩放，以使其落入一个特定的范围或区域。数据中各个特征的量纲不一致会导致其数值间的差异较大，进而影响数据分析的结果。通过数据规范化可以去除数据的量纲限制，将其转化为无量纲的纯数值，以实现对不同量纲的特征或指标进行综合比较和分析。常用的数据规范化方法包括最小 - 最大（min-max）归一化和 Z-score 标准化，前者根据特征数据的极值将其归一化到 [0, 1]

范围，后者则根据特征数据的均值和标准差将其分布标准化成均值为 0 和标准差为 1 的分布。

数据离散化是指通过数据分段将数值型数据转化为类别型或顺序型数据的过程。其基本思想是通过对连续的数值区间进行分段，以得到少量的子区间，并将每个子区间内的数值映射到一个离散符号，进而将连续数值型特征的取值离散化。例如，根据年龄将用户分为儿童、少年、青年、中年和老年。常用的数据离散化方法有等宽法、等频法、聚类法和有监督学习法。有效的数据离散化不仅能够降低后续分析算法的计算时间和空间开销，还能提高模型对样本的区分能力和抗噪能力。

类别编码是指通过编码变换将离散的类别型数据转化为数值表示。常用的类别编码方法有虚拟变量 (dummy variable) 编码和独热 (one-hot) 编码。虚拟变量又称哑变量，是用来反映类别属性的人工变量，通常为二值变量，取值为 0 或 1。例如，针对类别型变量“性别”，可以构造一个虚拟变量“is\_男性”进行表示，相应的“男性”取值为 1，“女性”取值为 0。独热编码又称一位有效编码，其基本思想是使用  $n$  位状态寄存器对  $n$  个可能状态进行编码，每个状态都有独立的寄存器位，并且在任意时刻其中只有一位有效。例如，针对类别型变量“性别”，可以构造一个二维的二进制向量进行表示，第一位表示“is\_男性”，第二位表示“is\_女性”，相应的“男性”取值为 (1, 0)，而“女性”取值则为 (0, 1)。可以看出，哑变量编码和独热编码的思想类似，唯一的差别在于针对有  $n$  个不同可能取值的类别型变量，哑变量编码采用  $n-1$  维的二进制向量进行表示，而独热变量则是采用  $n$  维的二进制向量进行表示。若是类别型变量的可能取值数量较多，则可根据频数将出现次数较少的取值统一归为一类“rare”。这种稀疏化处理既有利于后续分析模型的快速收敛，又能提升模型的抗噪能力。

### 三、描述性分析

描述性分析是一种利用少量的综合性描述指标来概括大量原始数据的统计分析方法。描述性分析的目的是用简洁有效的方式描述复杂烦琐的数据，以此来帮助用户快速了解数据的整体情况和特征，例如最大值、最小值、中位数、均值和方差等。常用的描述性分析方法有集中趋势分析、离中趋势分析和分布形状分析。

集中趋势是指变量观测值向中心集中的趋势，即在中心附近的观测值数目较多，而在远离中心位置的观测值数目较少，如图 1-5 (a) 所示。集中趋势指标主要用于反映数据水平的代表值或中心值。通过测定集中趋势可以反映现象总体的客观规律性，例如全国的人均 GDP、城市居民的平均工资等。按计算方法的不同，可将集中趋势指标分为两大类：数值平均数和位置平均数。数值平均数是根据样本数据中的所有观测值计算得到的，其特点是样本中任何一个观测值的变动都会在一定程度上影响数值平均数的计算结果。常用的数值平均数包括算术平均数、调和平均数和几何平均数，这些指标都容易受极端值的影响。位置平均数是根据样本数据中处于特殊位置上的个别或部分观测值确定的，不易受极端值的影响。常用的位置平均数包括众数、中位数和四分位数。

离中趋势是指一组数据中各观测值以不同程度偏离其中心 (平均数) 的趋势，如图

1-5(b)所示。离中趋势指标主要反映变量的观测值远离其中心值的程度,即离散程度。离散程度越大,表示变量观测值与集中趋势指标的偏差越大。离中趋势指标可以看作是对集中趋势指标的补充说明,它从另一个侧面说明了集中趋势指标值的代表程度。常用的离中趋势指标有极差、平均差、方差、标准差、异众比率、四分位差等。

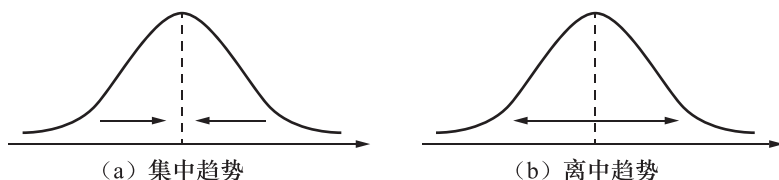


图 1-5 集中趋势与离中趋势示意图

常用的集中趋势指标和离中趋势指标大多隐含假设:数据服从对称分布。“集中”和“离中”的“中”隐含表示中心对称位置。当采用标准差作为离中趋势度量时,还会进一步隐含假设数据服从正态分布。但在实际应用中,这些假设可能并不满足,因此需要一些描述性指标来概括数据的分布形状。偏度和峰度是两种常用的描述数据分布形状的统计指标。

#### 四、探索性分析

传统的统计分析,包括描述性分析和推断性分析,通常需要先假设数据服从某种分布或是具有某种特性,然后再采用适合的指标或数学模型进行分析和预测。如果假设不合理,将导致指标和模型不适用,以致结果出现偏差或错误。当我们对数据不够了解或是缺乏足够的先验知识、问题比较难理解时,很难提出合理的假设。

探索性分析是一种数据驱动的分析方法,它不依赖于任何分布假设和复杂理论,完全从客观数据出发,通过可视化方法探索数据内在的规律和特性。这里的“探索性”是指对问题的理解会随着研究的深入不断变化。

探索性分析的基本思想是让数据自身“说话”,并通过可视化方式将其展示出来,同时借助一些交互式技术来帮助人们理解数据,其示意图如图 1-6 所示。这类方法假设人类是最好的模式识别者,能够识别出数据中的规律和特性,进而提出一些合理的假设或有价值的结论,为后续的推断性分析和基于模型与算法的分析奠定基础。

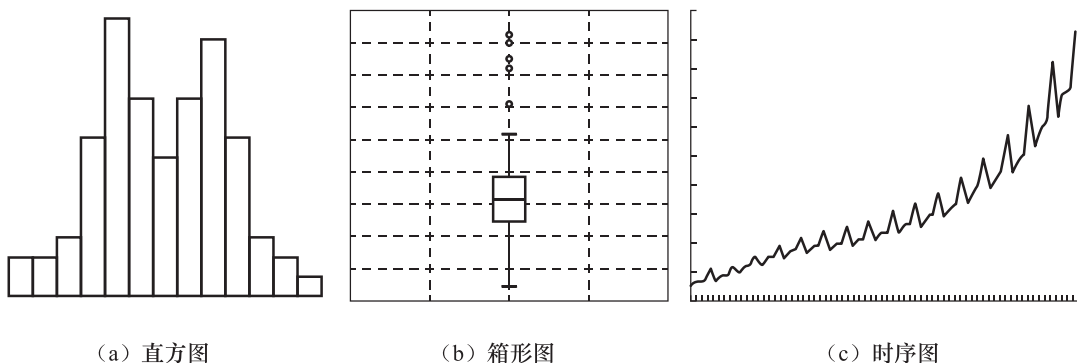


图 1-6 探索性分析示意图

探索性分析主要有以下三个目标。

(1) 帮助人们检查数据预处理的结果，发现错误、缺失和异常的数据，以辅助人们做好进一步的数据预处理。后续的数据分析，包括推断性分析、基于模型与算法的分析等，大多假设数据已经进行了较好的预处理，不存在错误、缺失和异常数据。但在实际应用中，通过调查、观测等方法所得到的数据大多杂乱无章，很难一次性处理好，因此需要通过探索性分析发现数据预处理中的不足，反复迭代，方可避免数据预处理中的错误和不足。

(2) 帮助人们理解数据，以获得对数据的直观感受，从杂乱的数据中，提取有用信息，获取数据背后的分析意义。通过各种可视化的图表展示数据的底层结构，帮助人们了解数据的内在规律，进而了解业务场景和业务特性。

(3) 帮助人们确定数据中的重要变量，提出合理的数据分布和模型假设，进而为后续的数学建模和推断性分析奠定基础。现有的各种数学模型和推断性分析方法，都显式或隐式地依赖于对数据分布和特性的一些假设。因此，在选择或构建数学模型和推断性分析方法之前，需要进行充分的探索性分析，以提出合理的假设。

## 五、推断性分析

推断性分析是指根据已知的（少量）样本观测数据对未知的（大量）总体特征做出科学判断的分析方法，其推断结果通常以概率形式表述。在实际应用中，组成研究对象总体的个体可能是大量的甚至是无限多的，或是单个个体的观测成本较高，导致客观上难以或无法对全部个体进行观测。推断性分析的目标是在只能获取研究对象部分个体信息的情况下，推断总体的特征或规律。例如，要检测一批商品总体的质量或使用寿命，不可能通过直接使用完所有商品再统计其质量或寿命，所以只能利用抽样数据进行推断性分析以得到相应的参数。

常见的推断性分析方法包括参数估计、假设检验和方差分析，如图 1-7 所示。

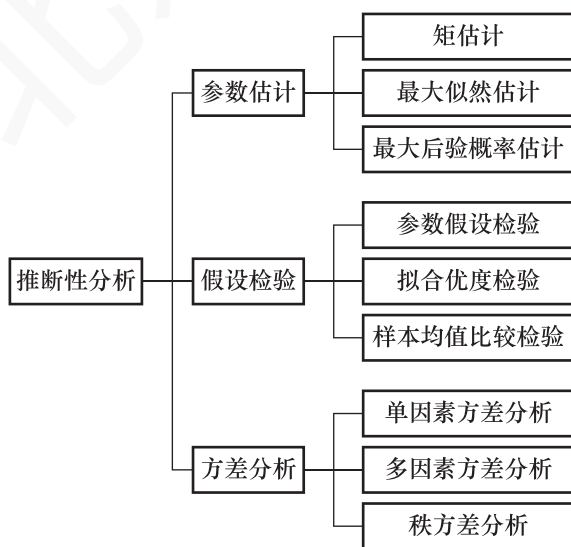


图 1-7 常见的推断性分析方法

参数估计是指通过已观测到的样本数据估计未观测到或是无法观测到的总体的参数。除了可以估计总体分布的参数,例如总体期望、总体方差等,还可以对各种(机器学习)模型的参数进行估计。常用的参数估计方法包括矩估计、最大似然估计、最大后验概率估计等。

假设检验是指根据样本数据检验关于总体的某个命题在统计上(即在一定的概率下)是否成立。假设检验的基本思想是根据样本数据的信息,利用反证法来检验关于总体的某个命题是否正确。在逻辑上运用反证法,在统计上依据的是小概率原理。假设检验常用于对产品参数、方法效果、总体间差异、总体分布的论断和假设进行检验。例如,针对一个产品的质量参数声明,可以通过参数检验对其真实性进行验证;针对一种方法(例如医疗诊断、营销策略等)的效果,可以通过样本均值比较对其进行显著性检验;针对待分析的一组数据是否服从某种分布,可以通过分布拟合对其进行检验。常用的假设检验方法包括参数检验、拟合优度检验和样本均值比较。

方差分析是指通过分析数据中差异的来源,找出对待研究问题或对象有显著影响的因素、各因素之间的交互作用以及显著影响因素的最佳水平。一个复杂的问题或研究对象往往存在许多相关的影响因素,且这些影响因素之间可能会互相制约又相互依存。为了对问题做出科学的判断,需要了解其各种影响因素。方差分析的基本思想是采用离差平方和来度量差异,并尝试把总差异按照各种指定的差异来源进行分解,即从总离差平方和中分解出可追溯到指定来源的部分离差平方和。常用的方差分析方法包括单因素方差分析、多因素方差分析和秩方差分析等。

## 六、基于模型与算法的分析

对于一些简单的问题,例如营销效果影响因素分析、商品受众分析等,人工通过探索性分析、描述性分析或简单的推断性分析就可以解决。但对于一些复杂的问题,例如垃圾邮件识别、信用风险预测、股票价格走势预测等,难以或是无法通过人工完成,则需要采用基于模型与算法的分析方法进行自动学习和推理。

数学模型是一种基于系统化的符号和数学表达式的问题描述方法,通过方程或其他数学工具定量地描述变量之间的关系。算法是指对解题方案的准确而完整的描述,由一系列解题指令构成。算法可以看作是用系统的方法描述解决问题的策略与机制,常用于数值计算、数据处理和自动推理,也可用于模型参数学习。

根据数据和问题不同,可以对模型和算法进行不同的划分。根据目标变量类型的不同,可以分为回归模型和分类模型。回归模型适用于目标变量是数值型变量的问题,而分类模型则适用于目标变量是类别型变量的问题。根据建模时是否考虑时间因素,可以分为横截面分析模型和时间序列分析模型。

针对同一问题,可以从不同的角度构建不同的模型,也可以采用不同的算法进行问题求解或是模型参数学习。例如,针对股票价格预测问题,可以采用基于横截面数据的传统多因子回归模型,也可以采用基于时间序列的自回归模型。基于同样的数据,可以采用线性回归模型,也可以采用非线性回归模型。



## 七、数据分析报告与应用

数据分析报告是建立在前期数据分析的基础上，给出明确的结论并提出合理的建议或解决方案的最终交付物。一份好的数据分析报告，应该有一个好的组织框架，并且图文并茂、层次清晰，能够让读者一目了然。

针对有些问题，通过数据的图形化表示就能够直观地展示出相应的结论。例如，图 1-8 展示的是一项四国 GDP 发展研究报告结果，从图中可以直观地看出中国的 GDP 在近几十年一直处于高速发展状态，并在 2007—2010 年间先后超过德国和日本，成为世界第二大经济体。此外，从图中还可以看出，自 1995 年起日本的整体经济一直处于增长停滞甚至是下行状态。

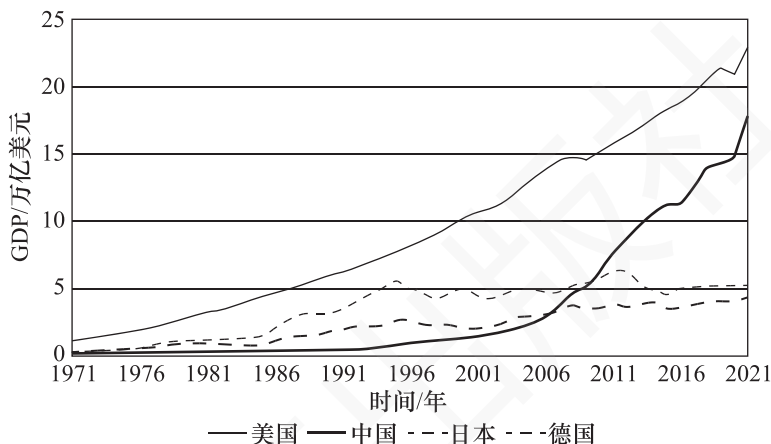


图 1-8 四国 GDP 发展研究报告结果

一份完整的数据分析报告至少应包含研究背景、问题定义、数据说明、分析方法、结果与结论、方案与建议等内容。研究背景部分应给出本次数据分析的动机和意义，即为什么要进行此次数据分析 (why)，其结果能够为谁 (who) 创造什么样的价值。问题定义部分应写明具体要分析什么 (what)，包括分析的目标和问题边界。数据说明部分应给出数据来源、收集方法、样本数量等基本信息，以说明使用了什么样的数据，可信度如何。分析方法部分应描述所使用的分析方法或模型的基本思想、分析步骤等内容，以说明是如何进行数据分析 (how)。结果与结论部分应给出基于收集的数据采用所选的分析方法或模型，得出的分析结果和相应的结论。方案与建议部分应基于定量的分析结果，并结合业务场景和相关的领域知识，给出针对目标问题的解决方案或合理建议，进而为具体的决策提供支持。

### 扩展阅读

#### 为什么“晚上 9 点钟洗澡的大学生成绩更好”？

有人在 ×× 电子科技大学做了一项研究，涉及 18 960 名本科生的匿名数据，覆盖了

5 个学期, 包括 3 380 567 次洗澡、20 060 881 次吃饭、3 466 020 次进出图书馆和 2 305 311 次在教学楼打水的记录。我们利用这些数据来刻画一个学生的行为特征, 例如努力程度和生活的规律性, 然后再看这些行为特征能否用来预测学生的学习成绩。

这里选择直接用进出图书馆的次数和在教学楼打水的次数来刻画学生的努力程度, 因为这两种行为与上课及上自习紧密相关。刻画生活的规律性要稍微复杂一点。如果要度量吃饭, 不仅要看时间分布是否集中, 还要看是否有序。例如, 今天吃早餐、吃午餐、吃晚餐, 第二天再吃早餐、吃午餐、吃晚餐, 这是有序的; 如果今天吃了早餐不吃午餐, 直接吃晚餐, 第二天不吃早餐, 吃午餐和晚餐, 这样就没有规律了。

从获取的两个学生的数据中可以看到: 一个生活很有规律的学生, 基本上都在晚上 9 点钟洗澡; 另外一个生活没什么规律的学生, 除了凌晨 2 点半到 5 点半不洗澡外, 其他时间好像随时都可以去洗澡。去食堂吃饭的情况也与之类似: 生活有规律的学生在 8 点左右、11 点到 12 点之间、17 点到 18 点之间去食堂刷卡吃饭, 可能要吃 10~20 分钟, 这和学校教学的作息时间是高度匹配的; 而生活没有规律的学生, 除了晚上 10 点到清晨 6 点(这段时间学校食堂也不开门), 其他时间随时都可以去食堂刷卡, 不一定是吃饭, 可能就是买根烤肠、买瓶水, 但明显没有什么规律。

充分利用研究得到的数据, 包括学生以前的考试成绩, 可以很精确地预测一个学生的期末考试成绩。学生的努力程度和生活规律性与成绩之间都有着很强的关联。除了吃饭、洗澡的时间, 本研究还分析了很多其他特征, 包括学生睡觉的时间、睡觉的规律性以及和其行为相似的同学们的成绩等, 这些都与 GPA(平均学分绩点)有关系。

举例来说, 一个学生上个月去图书馆 20 次, 在教学楼打水 30 次, 这个月只去了两次图书馆, 在教学楼只打了 5 次水, 那么我们就关注一下这个学生的情况了。在没有这种针对过程数据的分析手段时, 如果一个学生沉迷游戏, 第一学期考试可能勉强及格, 第二学期可能有一两科没及格, 这些现象并不会引起重视, 等到一两年后, 学生有多科不及格的时候, 再想让其改正就很困难了。行为数据分析的好处是能够及时通过异常变化发现问题, 而不会有很长时间的滞后。

教育数据分析是利用数据科学技术和方法来解决教育领域问题的一种新兴趋势。通过对学生学习行为的数据进行深入分析, 可以帮助教育工作者更好地了解学生的学习状态、行为特征和潜在问题, 从而提供个性化的教学指导和支持。

## 思考与练习

### 一、选择题

1. 以下选项中对数据定义最准确的是( )。
  - A. 数据是存储在计算机中的信息。
  - B. 数据是事实、观察或统计结果的集合。
  - C. 数据是用户输入到电子表格中的数字。
  - D. 数据是存储在数据库中的文本文件。



2. 按照研究对象的不同性质和计量的精确程度,可以采取不同的方式将变量分成四个层次,分别为( )。

- A. 称名数据 顺序数据 等距数据 比率数据
- B. 定类测量 定序测量 定距测量 定比测量
- C. 定类变量 定序变量 定距变量 定比变量
- D. 定类测量 定序测量 定距变量 定比变量

3. 常用的数据分析方法大多是以数理统计为基础,可以分为( )三大类。

- A. 描述性分析 探索性分析 推断性分析
- B. 集中趋势分析 离中趋势分析 分布形状分析
- C. 描述性分析 推断性分析 探索性分析
- D. 参数估计 假设检验 方差分析

4. 数据分析报告是建立在( )分析的基础上,给出明确的结论并提出合理的建议或解决方案的最终交付物。

- A. 前期数据 B. 中期数据 C. 后期数据 D. 末期数据

### 二、填空题

1. 数据泛指描述事物的\_\_\_\_\_,是构成\_\_\_\_\_和\_\_\_\_\_的原材料。
2. 由定距变量和定比变量形成的数据说明的是现象的数量特征,能够用数值来表示,也称为\_\_\_\_\_或\_\_\_\_\_。
3. 问题定义是确保数据分析过程有效性的一个基础环节,主要包含两部分内容:\_\_\_\_\_和\_\_\_\_\_。
4. 描述性分析是一种利用少量的综合性描述指标来概括大量\_\_\_\_\_的统计分析方法。
5. \_\_\_\_\_是指根据已知的(少量)样本观测数据对未知的(大量)总体特征做出科学判断的分析方法,其推断结果通常以概率形式表述。

### 三、简答题

1. 请解释什么是数据分析。
2. 请简述数据分析的一般流程,并解释每个部分的作用。
3. 请简述什么是问题定义以及如何问题进行问题定义。
4. 请简述数据预处理的目的是常用方法。

### 四、实训题

案例调研:请调研你感兴趣的行业或领域,整理一个应用数据分析的实际案例。