第 1 章

数据与信息概述

人类创造了数据。人类智慧的可贵之处在于，不仅可以发现自然界和人类社会的奥秘，还可以创造出客观世界不曾存在的新事物，如车轮、拉链、货币、艺术品、汽车和飞机等。其中数据无疑是人类最伟大的杰作之一。数据能够让人们探索一切未知的事物，并展示所有的可能。

有人将数据比喻为工业化经济的石油，也有人将数据比喻为信息化社会的矿藏，还有人将数据比喻为人类文明的土壤。石油、矿藏和土壤都是大自然赋予的，而数据作为人类创造的一种普遍性资源，具有更丰富的内涵与价值。本章将主要讲解数据和信息方面的知识。

1.1 数据的概念、特征和作用

1.1.1 数据的概念

在一般情况下，**我们将数据**（Data）**定义为记录客观事物的符号集合**。可以从以下几方面深入理解数据的概念。

1．数据具有普遍性

数据源于人类对客观事物的观察，由于客观事物是普遍存在的，因此数据也具有普遍性。随着人们数据处理能力的不断提高，数据无时不在，无处不在，并且成为人类生存发展的重要资源。同时，由于数据是人类从主观出发利用各种观测手段记录客观事物的符号体系，因此数据是主观认识与客观对象的综合体现。数据符号只有与客观事物相互联系才有实际意义。受到主观认识和观测手段的限制，数据会存在一定偏差。

2．数据是信息的表现形式和载体

一般而言，我们将数据视为一种未经加工的原始素材。数据在经过数据处理系统加工后，可以对决策产生有价值的信息。换句话说，**信息也可以视为经过加工处理的、有价值的数据**。数据处理系统示意图如图 1-1 所示。从目标结果角度看，数据是信息的表现形式和载体。

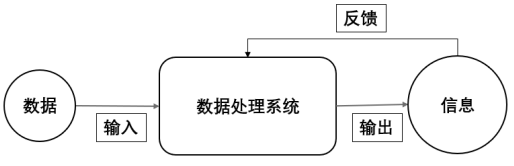


图 1-1　数据处理系统示意图

3．数据不等同于数字

数据不仅是数字，数字只是数据的组成部分。除了数字，数据还包括文字、图像、声音、视频等多种类型。同时，数据可通过不同介质来存储和展现。古人利用绳结的个数记录事物的数量，利用绳结的形状和大小区分事物的类别和重要性等；如今除广泛使用纸张作为数据介质外，人们还充分利用数字化技术将大量数据存储在计算机及各种电子存储系统中，并通过个人计算机、平板电脑、智能手机等终端设备展现，其主要介质有硬盘、U 盘和存储卡等数字设备。

1.1.2　数据的特征

通常数据具有以下特征。

1．数据的事实性

数据应具有事实性。只有可以展示事实的数据才有体现信息的根本价值，违背事实的数据是没有价值的。

2．数据的可复制性

数据可以被多次复制。随着数据存储技术的快速发展，一般数据复制工作不会造成数据的损耗，只需要消耗介质。数据复制工作的低成本性和便捷性使得数据可以迅速传播并被反复使用。

3．数据的共享性

当把一个苹果赠予别人时，赠予者便无法与被赠予者共享这个苹果，这是物质的不可共享性。当把一些电能输送到某座城市时，甲家庭消耗了此电能，乙家庭就无法再次消耗甲家庭消耗的电能，这是能量的不可共享性。但是，当把一份数据发送给其他人后，自身数据并没有发生任何改变，赠予者和被赠予者可以共享完全相同的数据，这是数据的共享性。

4．数据的时效性

数据的价值与时间紧密相关。有些数据具有明显的实时性，在产生之初价值最高，随着时间流逝价值逐渐降低，如电网监测数据、突发的重大新闻、战时敌方情报等。有些数据具有明显的历史性。逐步积累起来的历史数据可形成时间序列数据，从而完整地展现某事物在某段时期的发展轨迹，基于此可以发现问题、探索规律、预测未来。

5．数据的开放性与安全性

应该开放的数据没有开放，或者不该开放的数据被泄露，都会损害数据的价值。因此数据价值不仅取决于自身的信息内涵，还取决于数据的开放程度。数据开放性既是社会进步的条件，也是社会健康发展的表征，信息不对称性是诸多社会痼疾的根源。与此同时，对涉及国家利益、企业商业秘密和个人隐私的相关数据应依法保护，以保障数据的安全性。

1.1.3　数据的作用

20 世纪 50 年代中期，计算机技术开始在政府部门、工程管理、银行和商业等重大领域展开应用。与此同时，一些标志性的大型复杂工程不断上马并取得成功，如卫星发射计划、登月计划、大型客机设计制造等。这时一些欧美发达国家从事技术研究和业务管理的“白领工人”人数开始超过“蓝领工人”人数。美国未来学者阿尔文·托夫勒认为这是一个由信息爆炸引发的新一轮社会经济革命。发生这种革命的根本原因在于，数据处理及获取信息工作日趋重要。随后，互联网及移动互联网、大数据及人工智能等发展浪潮进一步推动了这一革命性进程。总结归纳数据的重要作用主要表现在以下方面。

1．数据是一种社会经济资源

在农业化社会，拥有了土地、牲畜、房屋和劳动力等物质资源，就掌握了社会经济发展的命脉；在工业化社会，拥有了石油、煤矿、电力与核能等能量资源，就掌握了社会经济发展的根本；在信息化社会，拥有了数据资源生产、传输、处理和利用的关键能力，就掌握了社会经济发展的核心。数据已经成为构成人类社会经济发展的第三种根本资源。

2．数据是认识世界的基本材料

数据是客观事物存在状态和发展规律的标志，人类只有通过数据才能认识客观事物的本质及其发展变化规律，才能真正感知丰富多彩的世界。同时由于客观事物是广泛联系的，所以数据是人们发现事物相互关系和因果逻辑的重要途径。数据是物质到意识、实践到认识、客体到主体、现象到本质的桥梁，是人类通过主观思维识知客观世界的基本材料。

3．数据是管理决策的依据，是决定企业竞争成败的关键

管理学家亨利·明茨伯格（Henry Mintzberg）明确指出管理活动的核心就是人际任务、信息任务和决策任务。管理活动是确定目标、制定计划、实施计划、监测评估、反馈调整的过程，是一个循环往复、螺旋上升的动态过程。科学管理决策中的每一个环节都需要大量的数据支持。

如果说在工业化社会人们信奉“时间是金钱，效率是生命”，那么在 21 世纪的信息化社会，就应树立“数据是金钱，决策是生命”的科学理念。及时捕获千变万化的数据，准确理解数据中蕴涵的深层含义，并迅速制定符合自身特点的决策发展战略，已成为决定企业竞争成败的关键。

总而言之，每个人、每个企业和每个国家只有具备敏锐的数据意识、清晰的数据思维、良好的数据处理习惯、突出的数据分析能力，才能在大数据时代和数字化世界生存发展。

1.2　数据的尺度与类型

数据的尺度是指数据度量的一般规则，它与数据的类型共同决定了研究数据的方法。

1.2.1 定性数据和定量数据

任何客观事物都有定性和定量两种基本属性。由于数据是记录客观事物的符号集合，所以数据有**定性数据**和**定量数据**之分。

定性数据是对客观事物各种质化属性的文字描述，如描述企业员工的性别、民族、籍贯、文化程度等的具体字符；定量数据是对客观事物各种量化属性的数字描述，如描述企业员工的年龄、销售额、工资等的具体数值。定性数据和定量数据可以刻画同一客观事物的不同方面。

1968 年，美国统计学家史蒂文斯（S.S.Stevens）按照数据的功能特点**将定性数据进一步分为定类尺度**（Nominal Scale）**数据与定序尺度**（Ordinal Scale）**数据，两者统称为分类型数据；将定量数据进一步分为定距尺度**（Interval Scale）**数据与定比尺度**（Ratio Scale）**数据，两者统称为数值型数据。**

1．定类尺度数据

定类尺度数据是按照无序的分类标准测度产生的数据。例如，性别中的男、女；职业中的工人、农民、军人等。定类尺度数据是一种最粗略的数据形式，可以使用具体数字指代，但并不改变数据的基本内涵，数字的次序和大小是没有意义的。例如，1 代表男，2 代表女。

2．定序尺度数据

定序尺度数据是按照有序的分类标准测度产生的数据，如受教育程度中的小学、初中、高中、大学等，产品质量中的一级品、二级品、次品等，可以使用具体数字指代。定序尺度数据不仅可以表示不同分类，还可以反映分类的大小顺序关系，数据的排序是有意义的。

需强调的是，定序尺度数据虽然具有一定的比较特征，但不能明确度量分类之间的数量关系，如无法测度小学和初中、初中和高中是否具有相同的数量差。

3．定距尺度数据

定距尺度数据是按照一定数量标准测度产生的数据，如气温、海拔等数值。这种数据不仅可以排序，还可以通过算术运算准确计算出各个数据间的差距等。例如，12 号的气温为 30℃，11 号的气温为 28℃，12 号气温比 11 号气温高 2℃。

4．定比尺度数据

定比尺度数据的主要数学特征是不仅可以进行算术运算，还可以进行比例运算。例如，甲同学成绩为 90 分，乙同学成绩为 30 分，甲同学成绩比乙同学成绩高 60 分，同时甲同学成绩是乙同学成绩的 3 倍。

定距尺度数据与定比尺度数据的根本区别在于，数据 0 的含义不同。定距尺度数据中的 0 和负数有具体的测量意义，如温度 0℃、海拔−5m 等。而定比尺度数据中的 0 表示“没有”或者“无”，负数没有意义。例如，不能说 90℃比 30℃暖和 3 倍，因为没有确定标准的“零”位，定距尺度数据无法进行比例运算。日常使用的摄氏度和华氏度都是定距尺度数据，而物理学中使用的热力学温度是定比尺度数据。

上述四种不同尺度的数据具有递进关系，计量层次上从低到高，测量精度上从粗略到精

确。高层次的数据具有低层次数据的全部特性，所以可以将高层次的数据转化为低层次的数据，如可以将考试成绩的百分制转化为优、良、中、及格和不及格等；反之则不可行。在数据处理中，高层次的数据包含的数学特性更多，适用的计算分析方法更多，因此数据层次越高其计算价值越高。

1.2.2　离散数据和连续数据

在定量数据中，数据依据取值特征可分为**离散数据**和**连续数据**。

离散数据通常是通过计数方式获得的一种定量数据。例如，某个学校有多少名学生，某学生学习了几门课程，等等。一般离散数据刻画的事物属性的取值允许跳跃式变化，如某学校的学生从 500 名增加到 800 名。

在一定数字区间内可以任意取值的数据称为连续数据。一般只有连续变化的事物属性才具有产生连续数据的条件，如某名学生的身高和体重、某个城市的气温和公路里程等。

有时为满足数据处理要求，需要对离散数据或连续数据重新进行分组，将其转换为定性数据中的定序尺度数据。例如，按子女人数分组研究居民家庭情况，可分为无子女家庭、独生子女家庭、二孩家庭、三孩家庭等。这种直接按原始数据值排序位置进行的分组称为**单项式分组**。又如，对于学生某门课程考试成绩，将 90～100 分的分组为 A，80～89 分的分组为 B，70～79 分的分组为 C，60～69 分的分组为 D，60 分以下的分组为 E。这种分组称为**组距分组**。组距分组把整个数据取值范围依次分为若干数据区间，区间的距离称为组距，各分组数据值由原始数据值落入的区间决定。

1.2.3　结构化数据和非结构化数据

1．结构化数据

通俗地讲，**结构化数据是指符合行列二维表格式的规范化数据**，符合 Excel 电子表格和关系型数据库数据表的组织格式要求。表 1-1 为某高校学生数据表，该表中的数据就是典型的结构化数据。

表 1-1　某高校学生数据表

学　号	姓　名	出 生 日 期	学　院	性　别	身　高/cm	体　重/kg
2003005	张三	2002-10-08	统计	男	183	76
2003015	李四	2002-07-15	统计	女	165	53
2002005	王五	2001-11-26	信息	男	176	68
2005016	刘六	2002-11-03	经济	男	172	65
……	……	……	……	……	……	……

结构化数据中的每行数据的属性及顺序都相同。每列数据的类型一致，长度基本整齐，标识基本规范。表 1-1 中的结构化数据可分别从统计学和关系型数据库两个学科角度做进一步讨论和解读。

1）从统计学角度讨论

在统计学中，称表 1-1 中的**一行为一个样本或一个个案**（Case），称表 1-1 中的**一列为一个变量**（Variable）**或一个维度**（Dimension）。如果将表 1-1 中的所有“身高”数据视为一个**总体**（Population），那么组成该总体的每个基本单元被称为**个体**，如“身高”数据中的“183”“165”等。当然，个体和总体是相对而言的。例如，一所高校相对于一名学生而言是总体，相对于一个省的所有高校而言就是个体。

2）从关系型数据库角度讨论

在关系型数据库理论中，一般称如表 1-1 所示二维表为**数据表**（Table），这是关系型数据库的基本逻辑表示方式；称数据表描述的对象（高校学生）为**实体**（Entity）。数据表不仅存储了具体的数据，还存储了描述这些数据的表头。表头被称为数据表的**表结构**（Table Structure）。数据表的一行被称为一条**记录**（Record）**或者一个元组**（Tuple），一列被称为一个**字段**（Field）或者一个**属性**（Attribute）。记录由字段名称和字段取值（Value）组成。一条记录中能唯一区别于其他记录的字段称为关键字段，简称**关键字**（Key），如学号。

数据表中的某个数据是由其对应的行列关系确定的，也就是说数据的语义是由所在行的关键字和所在列的字段名确定描述的，可借助函数直观地表示为 *F*(关键字,字段名)=数值。例如，*F*(学号="2003015",姓名)="李四"；*F*(学号="2005016",身高)=172；等等。

2．非结构化数据

与结构化数据相对的是非结构化数据。非结构化数据主要包括文字资料、图片、音频和视频信息等。例如，学生的奖惩情况可能包含长短不一的文字说明和数量不同的各种获奖证书图片等资料。非结构化数据在计算机数据库系统中需要用一些特殊技术进行存储处理。

如果简单地将结构化数据理解为有固定结构（统一的类型和长度等）的数据，将非结构化数据理解为没有结构的数据，那么半结构化数据就是指结构不固定的数据。常见的半结构数据有超级文本标记语言（Hyper Text Markup Language，HTML）、可扩展标记语言（Extensible Markup Language，XML）和 JS 语言对象标记法（Java Script Object Notation，JSON）等格式的数据，如表 1-2～表 1-4 所示。

表 1-2 所示为两名学生的姓名、性别、身高和体重的 HTML 格式数据。

表 1-2　两名学生的姓名、性别、身高和体重的 HTML 格式数据

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="UTF-8" />
    <title></title>
  </head>
  <body>
    <table border="1" cellspacing="10" cellpadding="20">
      <tr>
        <th>姓名</th>
        <th>性别</th>
```

续表

```
            <th>身高</th>
            <th>体重</th>
        </tr>
        <tr>
            <td>张三</td>
            <td>男</td>
            <td>183</td>
            <td>76</td>
        </tr>
        <tr>
            <td>李四</td>
            <td>女</td>
            <td>165</td>
            <td>53</td>
        </tr>
    </table>
  </body>
</html>
```

表 1-3 所示为两名学生的姓名、性别、身高和体重的 XML 格式数据。

表 1-3　两名学生的姓名、性别、身高和体重的 XML 格式数据

```
<student>
    <name>张三</name>
    <sex>男</sex>
    <height>183</height>
    <weight>76</weight>
</student>
```

```
<student>
    <name>李四</name>
    <sex>女</sex>
    <height>165</height>
    <weight>53</weight>
</student>
```

表 1-4 所示为描述两名学生有关属性的 JSON 格式数据。

表 1-4　描述两名学生有关属性的 JSON 格式数据

```
{   "student": {
      "name": "张三",
      "age": "18",
      "sex": "男",
      "address": {
         "province": "江苏省",
         "city": "南通市",
         "county": "崇川区"
      }
   }
}
```

```
{   "student": {
      "name": "李四",
      "weight": 53,
      "sex": "女",
      "address": {
         "province": "北京市",
         "city": "朝阳区",
         "county": "东大桥街道"
      }
   }
}
```

半结构化数据不符合关系型数据库或其他主流数据表的数据存储模式，但它通过相关辅助标记实现了数据内容的分隔和层次关系的表示。对于同一事物可以设置不同的属性，这些属性的数量、顺序和层次无须一致。半结构化数据是一种比较灵活的、具有自描述性的数据，已经作为数据结构的一种标准，在数据传输、存储、服务接口等方面有普遍应用。

在人类获取的数据中非结构化数据远多于结构化数据，且蕴含着更丰富的信息。随着人类数据技术（如互联网和物联网、云计算、大数据、人工智能等）的快速进步，人们对非结构化数据的转化处理和直接处理能力不断提高，并达到了前所未有的水平。

1.3 数据的表格化

采用简洁直观的方式组织、刻画和展示大批量数据是极必要的。在一般情况下，通过制作数据表格的方式组织数据，也称数据表格化；通过采用各种统计指标刻画数据；通过可视化图形展示数据；通过数据模型探索数据的内在规律。本节将重点讨论数据的表格化和统计指标。

相对于使用文字报告描述一批数据，使用表格描述数据更加直观、集中和易于比较，且对相关的数据逻辑模型设计和数据物理存储方式等有重要作用。

1.3.1 个体数据的表格化

在日常工作与生活中，我们经常会看到针对一个个体的数据的登记表，如某高校学生信息登记表（见表 1-5）、某人户口登记页、某企业员工信息登记表等。

表 1-5 某高校学生信息登记表

<table>
<tr><td rowspan="5">我的基本情况</td><td>姓名</td><td></td><td>性别</td><td></td><td>出生年月</td><td></td><td rowspan="5">相片</td></tr>
<tr><td>就读高中</td><td></td><td>家庭住址</td><td colspan="3"></td></tr>
<tr><td rowspan="2">住宿</td><td>□家里 □亲戚</td><td>地址</td><td colspan="3"></td></tr>
<tr><td>□校内 □校外</td><td>电话</td><td colspan="3"></td></tr>
<tr><td>个人手机</td><td colspan="2"></td><td>QQ</td><td colspan="3"></td></tr>
<tr><td rowspan="3">我的家庭</td><td>成员</td><td>姓名</td><td>年龄</td><td>学历</td><td>工作单位</td><td>职务</td><td>联系电话</td></tr>
<tr><td>父亲</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>母亲</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

类似于表 1-5 的表格常用于个体数据的采集和展现，但若将数据存入计算机数据库中，则需要将其转换成二维数据表的样式，如表 1-6 所示，包括姓名、性别等字段。

表 1-6 二维数据表示例

姓　名	性　别	出生年月	就读高中	家庭住址	住　宿	地　址	电　话	……

从计算机科学的数据仓库理论角度来看，表 1-6 是高校学生情况的实际反映，常被称为**数据粒度较细的事实表**（Fact Table），是记录所有学生当前状态的台账。表 1-1 也是一张事实表。表 1-5 是半结构化数据的具体体现，采用二维数据表存储并不是最理想的，可进一步考虑选用多张二维数据表存储或采用 JSON 等格式文件存储。

表 1-5 便于展现单个个体数据，并不适合展现批量个体数据。

1.3.2　批量汇总数据的表格化

在日常生活中，对于批量汇总数据大多采用多维统计表的形式来组织和呈现，如表 1-7 所示。

表 1-7　2021 年各学院学生情况统计表

学校名称：高校 1

学院	男			女		
	学生数/人	平均身高/cm	平均体重/kg	学生数/人	平均身高/cm	平均体重/kg
统计	655					
信息						
经济						
……						
合计						

资料来源：　　制表人：　　审核人：

多维统计表一般是由表标题、表栏、列栏、行栏、数值区和表注部分组成的。

- 表标题：概括说明了表格中统计数据的整体内容。例如，表 1-7 中的“2021 年”。
- 表栏：说明了表格中的统计数据共同具备的性质。例如，表 1-7 中的“高校 1”。
- 列栏：也称列维，说明了表格中某列数值共同具备的性质。例如，表 1-7 中的“男”和“女”，以及“学生数”、“平均身高”和“平均体重”。
- 行栏：也称行维，说明了表格中某行数值共同具备的性质。例如，表 1-7 中的各个学院的名称“统计”、“信息”和“经济”等。
- 数值区：是一批相关统计指标的数值部分的集中描述区域。
- 表注：说明了与统计表相关的其他辅助内容。例如，表 1-7 中的“资料来源”、“制表人”和“审核人”。

统计表数值区中某个数值的语义就是由对应的表标题、表栏、列栏和行栏部分共同确定的。例如，表 1-7 中的“655”的语义为“2021 年高校 1 统计学院男生人数”。

由此可见，**多维统计表是对批量数据提炼精简和直观展示的有效途径**。表中涉及的统计指标的概念将在下文进行讨论。

对多维统计表需进行如下说明。

（1）多维统计表的格式不是唯一的。

如表 1-7 所示的统计表还可以设计成如表 1-8 所示的多维统计表。

表 1-8　2021 年各学院学生情况统计表

学校名称：高校 1

		学生数/人	平均身高/cm	平均体重/kg
统计	男	655		
	女			
信息	男			
	女			
经济	男			
	女			
……	男			
	女			
合计				

资料来源：　　制表人：　　审核人：

多维统计表的格式实际上是将多个维度的统计指标，合理地布置在表标题、表栏、列栏和行栏中。其中，列栏和行栏可以使用多层结构表示多维数据。

（2）多维统计表是传统数据采集方式——统计报表制度的具体体现。

多维统计表可作为一种传统的数据采集形式，多用于多层级行政管理部门间的数据上报。一般上级单位设计并固定统计表的表标题、行栏和列栏，将表栏的统计时间、统计对象、数值区等留空，由下级部门填写上报。这种数据采集方式也称为统计报表制度。根据实际业务需要统计报表制度在时间上可分为年报、季报、月报、旬报、日报，甚至更细致的小时报等，这些统计报表需依赖计算机系统实现采集和汇总。

（3）多维统计表的计算机存储。

多维统计表的计算机存储往往需要采用关系型数据库中的二维数据表。该二维数据表本质是对如表 1-7 所示统计表多维列栏进行扁平化压缩，对表栏进行冗余化展开的结果。表 1-7 对应的二维数据表如表 1-9 所示。

表 1-9　表 1-7 对应的二维数据表

年　份	学校名称	学院名称	男学生数/人	男平均身高/cm	男平均体重/kg	女学生数/人	女平均身高/cm	女平均体重/kg
2021	高校 1	统计	655					
2021	高校 1	信息						
2021	高校 1	经济						
2021	高校 1	……						
2021	……							
……								

从计算机科学的数据仓库理论角度来看，相对表 1-1 而言，表 1-9 所示的学生统计表又称为**数据粒度较粗的快照表**（Snapshot Table），是如表 1-1 所示的事实表在某个时刻数据状态汇总的历史留存。针对快照表可以对不同时期的情况进行分析，并对事物的发展趋势进行预测。

1.3.3　统计指标

多维统计表中的数据是统计指标。**统计指标用于说明客观事物整体数量特征的概念和数值，其中概念部分称为统计指标名，数值部分称为统计指标值。统计指标可以概括表示批量数据的整体情况。**

例如，“2021 年高校 1 统计学院男学生人数 655”就是一个统计指标，其中统计指标名为“2021 年高校 1 统计学院男学生人数”，统计指标值为“655”。

理解统计指标的类型、结构和特征对后续学习数据科学是非常有益的。

1．统计指标的类型

统计指标的类型划分方式如下。

1）绝对指标和相对指标

绝对指标反映的是客观事物整体规模的绝对数量，如 2021 年某地区新生儿 2.35 万人。相对指标反映的是客观事物变化发展的相对程度，如 2021 年某高校男生人数比女生人数多 11%。

2）时点指标和时期指标

时点指标反映的是客观事物在某个时刻整体数量取值。例如，2021 年某地区新生儿 2.35 万人，这个统计指标的计算截止到 2021 年 12 月 31 日 24 时 0 分 0 秒。时期指标反映的是在一个连续时间段内整体数量取值。例如，2021 年北京市财政收入 x 万亿元，这个统计指标计算的是从 2021 年 1 月 1 日起到 2021 年 12 月 31 日止的财政收入累积。不同时间的同时期指标是可以合计的，但不同时间的同时点指标是不可以合计的。

3）截面数据和时间序列数据

截面数据是相同或近似相同时刻的一批统计指标的集合。例如，2021 年我国 31 个省、自治区、直辖市人口数的集合就是一个截面数据。这里的截面指的是省市自治区。

时间序列数据是一批在连续时间内的相同统计指标的集合。例如，2010 年 1 月至 2020 年 12 月我国月进出口贸易总额是一系列的月度指标序列，这就是一个时间序列。

2．统计指标的结构

对统计指标进一步进行结构分析发现，每个统计指标是由 6 个指标单元构成的，即统计时间（Time）、统计对象（Object）、统计分类（Kind）、统计指标名（Indicator）、统计指标值（Value）、计量单位（Unit）。用一个多元函数直观表示为 F(统计时间,统计对象,统计分类,统计指标名,计量单位)=统计指标值，即 $F(T,O,K,I,U)=V$。

例如，“2021 年高校 1 统计学院男生 655 人”的指标单元分别为统计时间：2021 年；统计对象：高校 1；统计分类：统计学院、男；统计指标名：学生人数；统计指标值：655；计量单位：人。

对于一个具体的统计指标而言，统计时间、统计对象、统计指标名、统计指标值和计量单位是必不可少的，否则将无法正确理解统计指标的语义；但统计分类既可以是退化的，也可以是多重的。

例如，“2021 年某地区新生儿 2.35 万人”的指标单元分别为统计时间：2021 年；统计对

象：某地区；统计分类：无；统计指标名：新生儿人数；统计指标值：2.35；计量单位：万人。

3．统计指标的特征

通过上述内容可知，统计指标具有多维性，以此出发进行深入探讨，可以发现统计指标的更多特性。

（1）质量统一性。

统计指标反映了客观事物的性质和数量，其中统计指标名是对性质的描述，统计指标值是对数量的刻画。所以统计指标具有质量统一性，在数据处理中简化或省略统计指标名可能导致数据含义不清、使用困难等问题。

（2）历史性。

统计指标可以按照时间维度积累数据。随着时间推移，历史数据成为进行发展分析、规律探索和趋势预测的基础，不会失去存在的价值。

（3）结构性。

统计指标的结构性是指统计指标的多维性和层次性。上文已对多维性有所介绍，而统计指标的层次性是由指标单元的层次性决定的。例如，统计对象“全国”向下可分为多个省、自治区、直辖市等；统计时间“某年”向下可分为 4 个季度、12 个月和 365 天等；统计分类“工业”向下可分为轻工业和重工业，其中“轻工业”向下又可分为纺织业、食品业和服装业等。

（4）动态变化性。

一般个体数据是对客观事物当前状态的记录，数据动态变化后数据的修改是覆盖式的，如某学生身高从 172cm 变为 174cm。历史性的统计指标是累积式的，各个时期的同一统计指标的含义（内涵和外延）经常发生变化。例如，重庆市原属于四川省，后被设立为直辖市，因此四川省的数据在重庆市被设立为直辖市前后的统计口径是不同的，在对不同时期数据进行处理时必须调整统计口径。

需要说明的是，在实际应用中很少使用一个或者几个统计指标研究分析具体的客观事物，一般会使用由一批相关的统计指标组成的整体，即**统计指标体系**，来多角度、多层次地进行描述处理。

1.4 数据的数字化

数据的表格化是数据在实际应用中的逻辑组织形式，而数据的数字化是数据在计算机系统或电子存储系统中的物理存储形式。

1.4.1 二进制与数字化

人类发明并习惯使用十进制数进行计算，并习惯用十进制数表示客观事物，如手机号、银行卡号和商品编码等。科学家在创造早期的机械式及继电式计算机系统时也习惯性地采用了十进制方式。这意味着这些机器必须用 10 个稳定状态的基本元件来表示十进制的 10 个基本数字

元素：0,1,2,…,9。但是 10 个稳态的元器件是很难设计开发的，这造成了系统极大的复杂性，并降低了系统的稳定性。

科学家进一步研究发现，大多元件一般只具有两个稳定状态。例如，电路的通和断、电压的高和低、电磁的正和负等。二进制数的采用使得机器系统的物理实现变得容易，运算变得简单高效，可靠性得到极大提高。这一革命性的设计思想在计算机等系统中得到广泛应用，缔造了 20 世纪最为光辉闪耀的科技成果之一。

二进制具有 2 个基本数字元素：0 和 1，同十进制一样，既可以进行数字运算，又可以通过编码来表示各种客观事物。

在数字运算上，二进制数采用逢二进一的方式，可以表示任何数字形式（如整数或小数等），并可转化为任意其他进制的数。在客观事物表示上，长度为 1 的二进制数可区别表示 2 个事物（0、1），长度为 2 的二进制数可区别表示 4 个事物（00、01、10、11），依次类推，长度为 *N* 的二进制数可区别表示 2 的 *N* 次方个事物。因此只要由 0 和 1 排列组合的符号串足够长，就可以区别表示世界上的万事万物。

二进制字符串中的每个 1 或 0 称作一个比特位（Binary Digit，bit），它是度量二进制数的最小单位。计算机原理中所说的一个字节（Byte）是计算机中组织和存储数据的基本单位，由 8 个比特位组成，1Byte=8bit。

我们所说的数据的数字化，就是将丰富多彩且千变万化的数据转变为一系列 0 和 1 的编码，并利用各种信息技术进行统一管理和统一处理的过程。基于此，人们能够将客观现实世界纳入计算机系统，并重构为一个全感知、全连接、全场景、全智能的数字世界，在更宽广的时空尺度上实现更高效的协同、更深层的融合，以及更有价值的创造。

1.4.2　文本的数字化

1．英文——ASCII

英文是典型的拼音文字，其基本字符集合由大小写字母、数字符号和一些特殊符号（@、+、*等）组成。英文的基本字符的个数不多，使用 7 个比特位编码就可以表达。但为与计算机系统的按字节处理方式一致，可在高位统一补 0。这样就可以设计一套二进制英文字符的计算机内部编码，我们称之为**机内码**。

为规范不同计算机厂商或不同国家地区的英文机内码设计方案，以使不同的计算机系统的数据可以互相识别，1967 年美国国家标准协会提出了美国信息交换标准代码（American Standard Code for Information Interchange，ASCII）方案，简称 **ASCII** 方案，并逐步成为国际标准。

例如，数字符号 7 的 ASCII 为 00110111。又如，大写字母 D 的 ASCII 为 01000100，小写字母 a 的 ASCII 为 01100001，小写字母 t 的 ASCII 为 01110100，英文“Data”的机内码就是由上述四个字节组成的二进制字符串。这种使用多个字节的延伸编码，统称 ANSI（American National Standards Institute，美国国家标准学会）编码。

2．中文——GBK

中文是典型的象形文字，其基本字符集合由众多汉字和一些特殊符号等组成。

自 1980 年，我国陆续颁布了国家汉字编码标准。例如，1995 年发布的 GBK 对常用的 6763 个汉字进行了编码。之后颁布的 GBK 扩充汉字编码共收录了 21 003 个汉字（包括部首和构件），883 个图形符号，其中吸收了港澳台地区的繁体汉字 BIG5 编码方案中的所有汉字，后续还增加了一些少数民族文字。这套标准的汉字编码方案使用两个字节表示一个汉字的机内码，最多可表示 256×256 = 65536 个基本汉字符号。

计算机键盘设置的特点使得汉字在输入时必须建立一套英文键盘输入规则与汉字的机内码间的对应方法，因此国内学者设计开发了全拼、双拼和五笔字型等汉字输入法。

二进制机内码虽然适合计算机内部的存储和处理，但并不便于直观展示，需要使用一种二进制编码方案来表示汉字或英文。于是人们创造了字形码，又称字模。

汉字的字形码有点阵和矢量两种表示方式。其中用行列的点阵方式表示字形时，一个汉字可以有一个16×16 的简易点阵（32 字节/汉字）、24×24 的普通点阵（72 字节/汉字）、32×32 的高密点阵（128 字节/汉字）或 48×48 的高密点阵（288 字节/汉字）。点阵的行列数越大，需要的存储字节数越多，字形越清晰。

形象地说，字形码就是将一个字符的笔画落在一个点阵网格上，其中落入笔画的网格记录为 1，未落入笔画的网格记录为 0，这样就得到了一批描述这个字符形状的有序的字节，该字节即这个字符的字形码。计算机字形码示意图如图 1-2 所示。

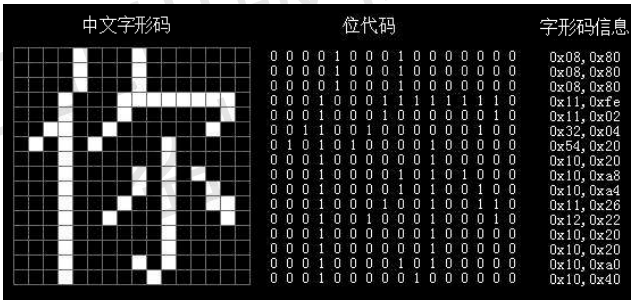


图 1-2　计算机字形码示意图

一个汉字从输入到输出的过程可以简要地描述为先根据汉字的输入法得到这个汉字的机内码，然后根据机内码找到这个汉字的字形码，最后根据字形码将汉字输出到屏幕上显示。

3．世界文字——Unicode

根据德国学者公布的研究成果，目前世界上有记录的文字种类大约有 5560 种。为避免大家各搞一套以至于互不兼容，国际标准化组织（International Organization for Standardization，ISO）提出了一套包含世界主要语言基本字符的二进制编码方案，称为**统一码**（Unicode）或**万国码**。这套方案将所有基本字符分配在 17 个区域，1 个区域称为 1 个平面（Plane）。ISO 最先定义公布了第一个基本平面的编码方案，这个平面中的字符编码用 2 个字节表示，其他辅助平面的编码方案正在陆续发布中，根据实际情况可采用 2 个字节、3 个字节或 4 个字节表示一个字符。

Unicode 在计算机系统实际存储和传输等处理中面临着以下难题。例如，针对 ASCII 文档，如果每个字符统一使用 2 个以上字节来表示，那么许多比特位将填补为 0，文档至少要扩大 1

倍的空间进行存储，同时网络传输会消耗更大；如果不这样处理，计算机系统就必须有效识别哪些字符是用 2 个字节表示的，哪些字符是用 1 个字节表示的。

全球化的互联网繁荣发展，让 Unicode 方案的推广应用更加迫切。为解决上述问题人们进一步设计了 UTF（UCS Transfer Format）改进方案，分为 UTF-8、UTF-16 和 UTF-32 等。其中，UTF-8 根据每个基本字符的具体情况可以使用 1～4 个字节变长度编码，既能保证文档的存储和传输效率，又能正确地表示各种文字的原貌，成为互联网应用最广泛的编码方案之一。例如，在使用 Windows 记事本软件存储文档时，有多种 Unicode 选项可供选择，如图 1-3 所示。

图 1-3　Windows 记事本软件的不同编码存储方式

1.4.3　数字的数字化

常用的数字一般分为整数和小数。在计算机系统中采用多少字节表示一个数字与计算机安装的操作系统和计算机语言等软件系统有关。

1．整数

一般计算机系统采用 2 个字节表示一般的短整数（Integer），采用 4 个字节表示一个取值范围更大的长整数（Long Integer）（某些 64 位计算机系统开始采用 8 个字节表示整数）。通过二进制最高比特位取 0 或者取 1 来区分正整数或负整数。如果这个比特位直接用来表示数值，就是无符号整数（Unsigned Integer）。

对于 2 字节 16 位的短型整数，如果二进制最高比特位为正负符号位，那么其表示的整数的取值范围是−32768～+32767；如果不设符号位，那么其表示的无符号整数的取值范围是 0～65535。同理，对于 4 字节 32 位的长型整数，有符号位和无符号位表示的整数范围分别是−2147483648～+2147483647 和 0～4294967295。

2．小数

一般在计算机系统中小数按浮点数的方式表示。浮点数（Float）是指小数点位置不固定的

小数。美国电气和电子工程师学会（Institute of Electrical and Electronics Engineers，IEEE）发布了一个解决方案并成为浮点数的行业标准。该方案规定了四种表示浮点数的方式，其中最普遍的是单精度浮点数（Single Float）使用 4 字节 32 位表示，双精度浮点数（Double Float）使用 8 字节 64 位表示。

IEEE 标准指出任意一个浮点数都可以表示为规范化的格式。例如，十进制数 0.234 通过计算可以转换表示为二进制数 0.001110111110，进一步可以表示为规范化形式：1.110111110 乘以 2 的−3 次方。

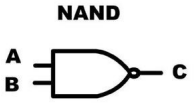
IEEE 标准规定，对于 32 位的单精度浮点数，最高比特位是表示正负的符号位，后续 8 位表示指数（如−3），剩下的 23 位表示具体有效小数（如 1110111110），所以单精度浮点数的取值范围为-3.4×10^{38}～$+3.4\times10^{38}$。对于 64 位的双精度浮点数，最高比特位仍是符号位，后续 11 位表示指数，剩下的 52 位表示具体有效小数，所以双精度浮点数的取值范围为-1.79×10^{308}～$+1.79\times10^{308}$。

3．数字的运算与数字电路

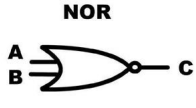
二进制数在计算机内部表示和存储为数字的快速计算铺平了道路。依据以下原则，计算机可进行基本的加减乘除运算，这里仅对其进行简述，更复杂的计算可参考相关专业资料。

- 加法：二进制数的加法按照比特位的运算规则进行，即 0+0=0、0+1=1、1+0=1、1+1=10。
- 减法：减去一个数字也就是加上这个数字的负数，可通过反码和补码来实现。
- 乘法：通过二进制数的移位和累加实现。
- 除法：通过二进制数的移位和累减实现。

可见，计算机的基本数字运算是以二进制数加法运算为基础的。科技人员采用现代半导体技术和工艺设计开发了各种数字存储器和数字电路。其中，构成数字电路的基本逻辑单元又称逻辑门，包括与门、或门、非门、异或门、与非门、或非门等。逻辑门通过控制高电平、低电平等状态来代表二进制数 1 或 0 的输入，并根据不同处理形成 1 或 0 的输出，从而实现比特位上的逻辑运算，如图 1-4 所示。通过组合设计逻辑门可以形成更复杂的集成数字电路，实现更复杂的逻辑运算和逻辑控制。



A	B	C
1	1	0
1	0	1
0	1	1
0	0	1



A	B	C
1	1	0
1	0	0
0	1	0
0	0	1

图 1-4　与非门（NAND）、或非门（NOR）符号与逻辑运算

以离散数学和布尔代数等为理论基础，基于二进制数运算体系设计并开发的逻辑数字电路等，使得计算机系统获得了数字运算能力，而且远超人类的运算速度和运算精度。许多学者认为这是计算机系统实现人工智能的一个重要标志。

1.4.4　多媒体的数字化

多媒体的数字化主要指图像、声音、视频的数字化。

1．图像的数字化

计算机系统中的图像从存储内容上可以分为位图和矢量图两大类。

1）位图

位图在计算机中是由一些按位置排列的、不同颜色的点组成的，这些点称为图像的像素点。例如，一张宽度为 1024 像素，高度为 768 像素的黑白图像包含1024×768 个像素点。可以使用若干字节表示每个像素点的灰度等属性。

存储位图的图像文件一般由描述这些像素点的基本信息和一些描述文件性质的信息组成。前者是像素文件内容的主体。描述一个像素点的字节越多，存储图像需要的空间就越大。为减少存储空间，提高网络传输效率，人们设计开发了许多图像格式的压缩算法。BMP 格式是 Windows 操作系统中的标准图像文件格式，由于很少使用压缩处理所以存储空间较大，但可以被多种常用的图形图像软件处理；JPEG 格式和 GIF 格式是一种高效的有损压缩位图文件格式；PNG 格式是一种采用无损压缩算法的位图文件格式。

在计算机应用中，人们常用位图数据进行图形的识别和处理。通常可使用将位图数据转换为一个二维数组或矩阵的方式对位图进行分析研究。

2）矢量图

矢量图一般是由一系列点和连线类的图形元素组成的。每个图形元素都是一个独立的实体对象，并带有位置坐标、层次、形状轮廓、颜色和大小等属性的数据描述。

矢量文件实际上存储着某个图像几何画法的参数，在逻辑上可使用 XML 语言、PostScript 语言或自定义描述规则进行定义，在物理上可将这些描述用 ASCII 等方式进行存储。因此，在一般情况下，矢量文件占用存储空间较小但生成速度稍慢。其最大优势在于，生成的图像与显示设备的分辨率无关，图像放大后不会失真。因此，矢量图适用于图形辅助设计、区划地图设计，以及文字、标志和版式设计等应用领域。常见的矢量图的格式有 Adobe Illustrator 软件使用的 AI 格式、CorelDraw 软件使用的 CDR 格式和 AutoCAD 软件使用的 DXF 格式等。

2．声音的数字化

声音是连续的，早期人们使用模拟信号描述声音数据，之后采用数字方式描述声音数据。在采用数字方式描述声音数据时，需要先按照一定的采样频率，即每秒从连续的声音信号中提取采集声音数据，并组成离散式的声音数据。采样频率的计量单位为次/秒，即赫兹（Hz）。采样频率越高，单位时间内采集的声音数据量越大，声音数据的质量越高。对一些简单的声音信号过度高频采样是一种浪费，一般对应的采样频率为电话语音：8000Hz；收音机广播：22 050Hz；音乐 CD：44 100Hz；数字电视和 DVD：48 000Hz；蓝光高清 DVD 等：96 000～192 000Hz。然后对每个采样点的声音波幅的数值进行量化编码处理，一般可使用 1 字节、2 字节或 3 字节按二进制形式存储。

例如，存储一个采样频率为 44.1kHz、2 字节 16 位量化波幅、双声道、1 分钟的声音数据，

大约需要 10.1MB 的空间。微软公司为 Windows 开发的 WAV 格式就采用了上述参数标准，其不足之处在于占用存储空间较大。因此，人们在有效保证声音质量的前提下采用了诸多压缩技术，极大降低了音频文件的存储空间，如 MP3 格式等。

3．视频的数字化

一般来说，视频数据是由一系列图像数据和与之同步的声音数据组成的。通过快速连续播放一幅幅静态独立图像，可以让观察者获得无跳跃感的视频信息。其中每一幅图像称为一帧。每秒播放的帧数越大，图像越流畅，视频质量越高。在一般情况下，电影的播放速度为每秒 24 帧，我国采用的 PAL 制式电视的播放速度为每秒 25 帧。

在计算机中每帧图像数据的存储可采用上述位图方式，声音数据的存储也可采用上述采样和量化编码的方式。由于视频数据量的规模很大，人们根据视频数据的特点设计开发了许多有效的压缩算法，形成了不同的视频存储格式。例如，在视频数据中，许多相邻帧的相似度较高，许多像素点的位置和色彩等数据变化很小或者没有变化，因此就能以某一帧为基础，将相似的帧作为一组进行帧间压缩，每一帧只存储组内对应上一帧有变化的数据，这种压缩算法在一定程度上减少了视频数据的存储空间。

常见的视频数据格式有 MPEG、AVI、MOV、ASF、WMV、RM、FLV 等。其中，人们常用的 MP4 格式是 MPEG 格式的一个版本，可对视频数据中的音频、视频等多种数据进行压缩编码，是一种有损的、压缩比较高的视频数据处理标准。

1.4.5　数字化转型与数字化经济

进入 21 世纪后，数据对人类经济社会的重要作用愈加明显。随着大数据时代的到来，数据的价值更加突出。有识之士认为，如果互联网是生产关系，计算是生产力，那么数据就是生产资料，被比喻为生产活动的根本条件；数据还可以上升到数据资源，被比喻为持续发展的基本要素；数据还可以上升到数据资产，被比喻为有价值、可交易的社会财富。

数字化是世界发展的趋势，数据科学是建立在这个数字化世界上的知识体系。随着以大数据、人工智能、移动互联网、云计算、区块链等为代表的数字化技术的快速发展，创新的数字化产品与服务将不断涌现，数字化平台将不断被推出，从而有效驱动企业数字化转型，全面拉动国家数字化经济与数字化治理。

我们讨论的数字化、数字化转型和数字化经济的主题都是研究和开发数据，是在研究如何从数据中创造价值，使之与土地、劳动、技术和资本等生产要素一起为人类社会经济可持续发展服务。对几个概念归纳总结如下。

数字化通俗而言就是将人类采集的数据转变为一系列的二进制 0、1 编码的形式，并利用信息技术进行统一管理和统一处理，进而将客观现实世界重构成一个数字世界，为企业的数字化转型服务，为政府发展数字化经济和实现数字化治理服务。**数字化转型**是以价值创新为目的，用数字技术驱动业务变革的企业发展战略。数字化转型对于企业而言是一个长期发展过程，不是通过一两个短期的信息技术项目可以实现的。**数字化经济**也称数字经济，将数字化的数据、

信息和知识作为关键生产要素，将现代通信网络作为重要载体，将信息技术应用作为提升效率的手段，将经济结构优化作为重要推动力的一系列经济活动。

1.5　信息与信息熵

从数据中获取信息是进行数据处理的基本目标。简明地将信息定义为经过加工处理的、有价值的数据是一个直观层面的定义，它无法解释深层次的信息概念的问题。例如，信息的价值究竟是什么，信息价值的大小如何定量地衡量，等等。本节将尝试回答这些问题。

1.5.1　信息熵：不确定性的度量

1948 年，美国贝尔实验室的科学家克劳德·香农（Claude Shannon，1916—2001）发表了《通信的数学理论》等相关论文，通过研究通信中数据传输等相关问题，首次对通信过程建立了数学模型；并借鉴物理热力学中关于熵的理论，提出了信息熵的概念，以及衡量信息熵的数学公式。

香农的信息论解决了信息量和不确定性的度量问题，奠定了现代信息论的基础，对数据传输技术及数据编码学、数据加密技术及密码学、数据压缩理论等做出了直接贡献，对计算机科学、信息科学、统计学、热物理学、语言学和机器学习等产生了深远影响。

信息论、系统论和控制论被称为 20 世纪最伟大的理论成果，对现代科学思维和科学技术的发展起到了积极作用。信息和信息熵不仅是通信领域中的基本概念，还被广泛应用到数据科学和其他学科领域中。

信息论将信息抽象地定义为事物不确定性的减少，并建立了信息定量测度的数学描述，即信息熵。通过香农的信息熵的数学公式理解信息熵的概念可能会感到吃力，这里先举一个简单的例子直观地说明信息熵的含义。

☞【案例】总统竞选

某国 3 位总统候选人各方面的能力相当，民意测验结果是各占 1/3，很难预料谁会成为最后的赢家，这时总统选举的不确定性最高。然而，某知名报刊披露某位候选人曾有负面行为，这时民意测验结果显示，此位候选人的支持率降至 20%，而另外 2 位候选人的支持率升至 40%，这时总统选举的不确定性相对降低了。如果降低原因全部源于该报刊发布的消息，那么不确定性的减少量就是这则消息的信息量。

这里的关键问题是信息量如何度量。回到通信系统场景下。如果信源可以发送 4 种信号，那么采用二进制编码信号，需使用 2 个比特位发送信号。如果信源可以发送 8 种信号，就需要使用 3 个比特位发送信号。如果信源可以发送 N 种信号，就需要使用 $\mathrm{lb}N$ [①]个比特位才能实现无损传输。于是定义信息量为

$$I = \mathrm{lb}N \tag{1.1}$$

① $\mathrm{lb}N$ 即 $\log_2 N$。

在此基础上，香农将不同信号的发送概率引入计算，提出了信息熵（Entropy）的计算公式：

$$E=-\sum_i P_i \mathrm{lb} P_i \tag{1.2}$$

以掷骰子为例介绍信息熵的计算。一个骰子有 6 个面，在规范情况下每个面朝上的概率相等，均为 P_i，$P_i=\frac{1}{6}$（$i=1,2,\cdots,6$）。根据式（1.2）计算出的掷骰子信息熵为

$$E=-\sum_{i=1}^{6} P_i \mathrm{lb} P_i=-6\times\frac{1}{6}\mathrm{lb}\frac{1}{6}=-\mathrm{lb}\frac{1}{6}=\mathrm{lb}6=\mathrm{lb}N=I$$

可见，信息量 I 是信息熵 E 在不同信号的发送概率相等时的特例。

香农的信息论的重要成就在于，可以通过信息熵度量信息的不确定性。

例如，对于掷骰子问题，假设骰子被动过手脚，每次掷骰子的结果都是数字 6 朝上，即数字 6 朝上的概率等于 1，其他数字朝上的概率等于 0，不具有任何不确定性。此时计算的信息熵等于 0，小于每面朝上概率相等，即不确定性最大时的信息熵。可见，信息熵越大，不确定性越大；信息熵越小，不确定性越小。不确定性是抽象概念，其因信息熵而具有了定量标准。

1.5.2 信息增益：不确定性减少的度量

信息熵及在信息熵的基础上计算出的条件熵、信息增益，展现了更大的价值。下文通过两个例子来说明。

☞【案例】总统竞选

利用信息熵，计算上文总统竞选案例中负面消息的信息量。

负面消息发布前的信息熵为 $-\frac{1}{3}\mathrm{lb}\frac{1}{3}-\frac{1}{3}\mathrm{lb}\frac{1}{3}-\frac{1}{3}\mathrm{lb}\frac{1}{3}=\mathrm{lb}3=1.585$，负面消息发布后的信息熵为 $-0.2\mathrm{lb}0.2-0.4\mathrm{lb}0.4-0.4\mathrm{lb}0.4=1.522$，则负面消息的信息量（不确定性的减少量）为 $1.585-1.522=0.063$。

☞【案例】超市购买行为研究

在信息熵的基础上可以派生计算**条件熵**和**信息增益**，通过此案例对其进行说明。

假设收集到某超市某商品购买情况数据如表 1-10 所示。其中，“是否购买”变量取 1 表示购买，取 0 表示未购买。

表 1-10　某超市某商品购买情况数据

顾客编号	性　别	婚姻状况	年　龄	是否购买
1081253	女	未婚	25	1
1060921	男	已婚	45	0
1017336	女	已婚	35	1
1080035	女	未婚	32	1
1089272	男	已婚	50	0
1017338	男	未婚	20	0

对数据进行初步统计分析发现购买和未购买的顾客各占 50%；100%的女性顾客购买了商品，100%的男顾客未购买商品；1/3 的已婚顾客购买了商品，2/3 的未婚顾客购买了商品。

- 计算“是否购买”变量的信息熵：$-\frac{1}{2}\text{lb}\frac{1}{2}-\frac{1}{2}\text{lb}\frac{1}{2}=1$。
- 计算“是否购买”变量的以“性别”为条件的条件熵：
  对于女性顾客有 $-1\text{lb}1-0=0$；对于男性顾客有 $-0-1\text{lb}1=0$。
  条件熵为 $P_{女}\times 0+P_{男}\times 0=\frac{1}{2}\times 0+\frac{1}{2}\times 0=0$，式中，$P_{女}$和$P_{男}$分别为女性和男性的人数占比。
- 计算“是否购买”变量的以“婚姻状况”为条件的条件熵：
  对于已婚顾客有 $-\frac{1}{3}\text{lb}\frac{1}{3}-\frac{2}{3}\text{lb}\frac{2}{3}$；对于未婚顾客有 $-\frac{2}{3}\text{lb}\frac{2}{3}-\frac{1}{3}\text{lb}\frac{1}{3}$。
  条件熵为 $P_{已婚}\times\left(-\frac{1}{3}\text{lb}\frac{1}{3}-\frac{2}{3}\text{lb}\frac{2}{3}\right)+P_{未婚}\times\left(-\frac{2}{3}\text{lb}\frac{2}{3}-\frac{1}{3}\text{lb}\frac{1}{3}\right)=-\frac{1}{3}\text{lb}\frac{1}{3}-\frac{2}{3}\text{lb}\frac{2}{3}$，式中，$P_{已婚}$和 $P_{未婚}$分别为已婚和未婚的人数占比。

可见，计算考察“性别”前后“是否购买”信息熵的变化量为 $1-0=1$；计算考察“婚姻状况”前后“是否购买”信息熵的变化量为 $1-\left(-\frac{1}{3}\text{lb}\frac{1}{3}-\frac{2}{3}\text{lb}\frac{2}{3}\right)$。通常称这里的变化量为**信息增益**。两个信息增益分别度量了考察“性别”和“婚姻状况”两个变量对“是否购买”不确定性的影响。显然，“性别”的信息增益大于“婚姻状况”的信息增益，这意味着“性别”对减少不确定性的作用大于“婚姻状况”，“性别”与“是否购买”的相关性更大。因此，在预测某个新顾客未来“是否购买”该商品时，应更多关注其“性别”。由此可知，在建立预测模型时，信息熵可作为评价变量重要性的依据。