
# 数据分析与数据挖掘

# 大数据应用到我们的生活中会怎样呢

什么是大数据！一段买披萨的对话让你秒懂！
https://www.bilibili.com/video/BV11V411s7eq/?spm_id_from=autoNext


“天烛”大数据疫情防控平台
统计截至:
本地疫情追踪
自动轮巡
开
确诊病例
现有确诊
较昨日 +0
累计确诊
较昨日 +0
本土无症状
较昨日 +0
治愈人数
较昨日 +5
今日密接/次密接流转情况
待流转
车辆派遣
转运中
已达隔离点
车牌号
接载人数
状态
24
前往接送点
13
前往隔离点
30
完成转运
13
30
各区确诊病例统计
湖里区
累计确诊人数
翔安区
累计确诊人数
集美区
累计确诊人数
前往接送点
确诊病例
封控区域
管控区域
确诊病例
封闭道路
检测点
封控区域
确诊病例
密接管理
风险区域
交通管控
核酸检测
舆情民意
回顾模式
各区中高风险区统计
（高风险区域数量/中风险区域数量）
全市：1/3
0/0
0/0
1/3
0/0
0/0
0/0
舆情动态
市民集中反映

# 智能运维大屏

Intelligent operation and maintenance of the large screen


党建云
电商云
发改云
法治云
扶贫云
工业云
平安凯里云
国土云
环保云
交通云
科技创新云
旅游云
民生云
农业云
人社云
消防云
市场监管云
教育云
医疗云
应急云
智慧凯里
BIG DATA

# 党建大数据服务平台

集团党委

党支部

支部 1

支部 1

支部 1

党员

284
集团党委

2845
党支部数

28455
党员数

50%
性别占比
0
100

学历占比
本科
专科
硕士
0
100
高中以下
博士

1000
500
200
100
50
0
18~30岁
40~50岁
60~70岁
70岁以上

学习任务发布
0 1 2 3 4 5 6 7 8 9 10
2018.1
2018.2
2018.3
2018.4

优秀党员

左永刚
所属支部
某某支部

单位/h
学习进度
600
400
200
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 月/2018年

单位/次
三会一课
600
400
200
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 月/2018年

81.1%
标准化达标率

党建活动：342 次
教育学习：2510 次
意见反馈：2700 条

党建活动


思想建设
组织建设
作风建设
制度建设
反腐倡廉建设

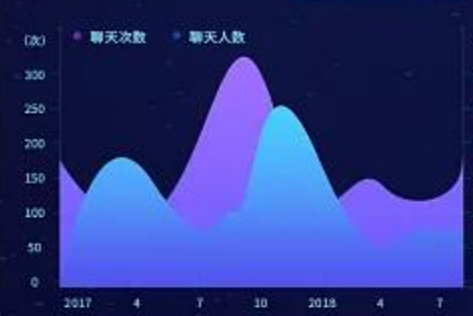
精准扶贫

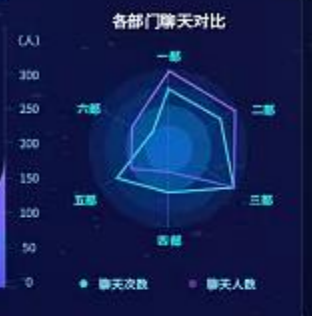
综合态势 | 营销态势 | 分析平台

# 移动营销大数据态势感知

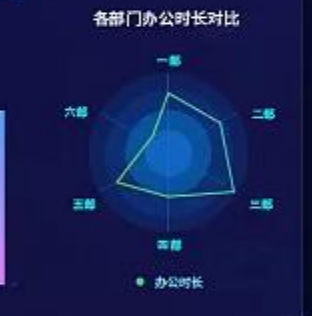
昨日 近一周 近一月

## 聊天分析



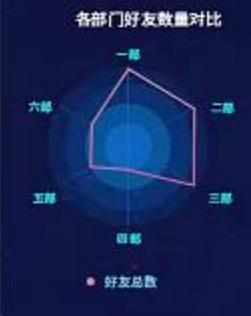



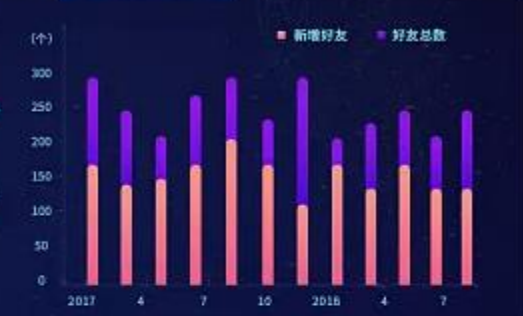

## 办公时长分析

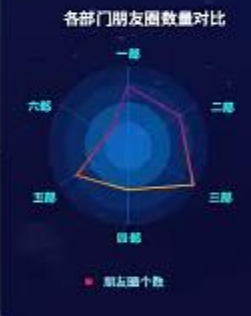


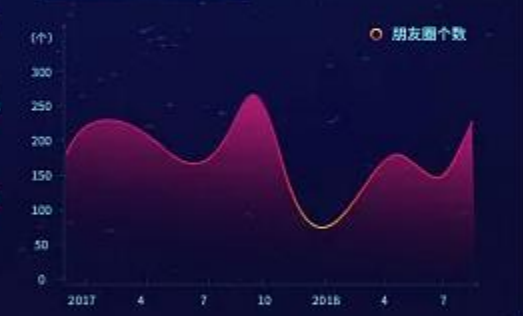
## 好友分析



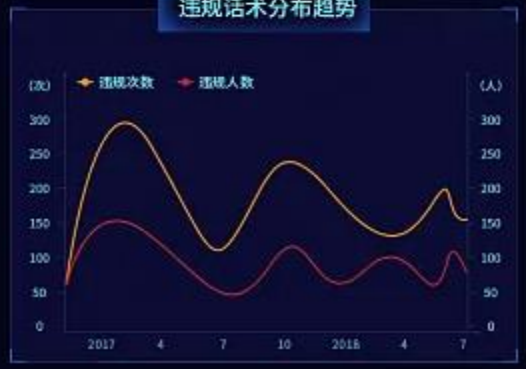

## 微信朋友圈分析



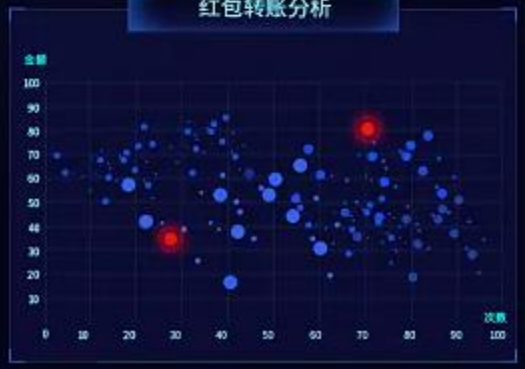

## 违规话术热词


## 违规话术分布趋势

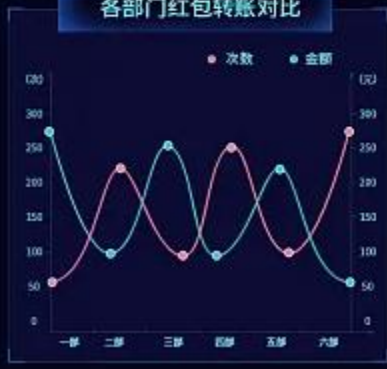



## 各部门话术违规情况对比


## 红包转账分析




## 各部门红包转账对比




GIS应急指挥调度系统
2019年07月02日 21:31:48
广州 28℃~19℃ 空气质量:优
限行：3 施工：3 红绿灯：15 停车位：316
全城 城东 城西 城南 城北
重点保护目标
重大危险源
应急物质库
消防设施
应急队伍
应急专家
气象指数
风速 风向 云量 湿度 气压
3m/s 东南 1.6% 1019.5hPa
水文参数
流速 流向
3m/s 东南
操作面板
启动预测
发送预测结果
影响分析
事件上报
方案修订
下发方案
拥堵 缓行 畅通

# 景区管理

综合分析 | 舆情分析 | 游客统计

景区管理 | 交通分析 | 2019-6-17 星期一 17:38:54

## 环境监测

地点 厦门 · 刚刚更新

阴 27°

优

北风 2级 降水量 0.0mm

湿度 80% 大气压强 993hpa

| | | |
|---|---|---|
| 今天 优 | | 25°/30° |
| 明天 优 | | 26°/30° |
| 后天 优 | | 26°/31° |

不宜 热 少发

## 监控警报

| | | | |
|---|---|---|---|
| 火情警告 | 1 | 人流警告 | 5 |
| 越界警告 | 3 | 车流警告 | 5 |

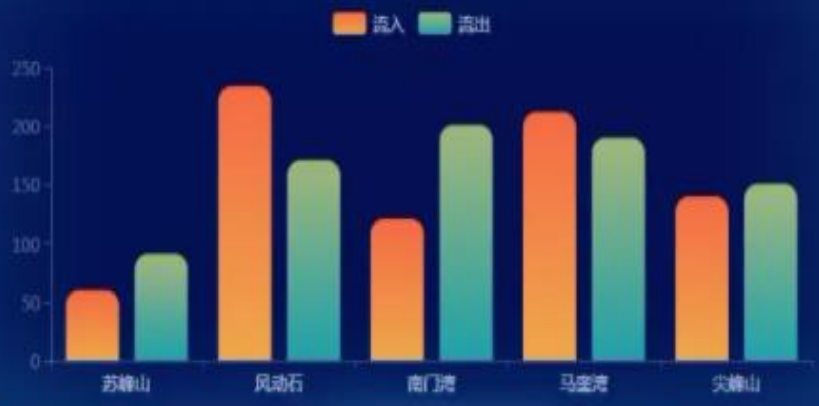
## 巡更记录表

全部 | 消火栓 | 灭火器 | 自动喷水灭火系统 | 导出报表

| 名称 | 负责人 | 巡更时间 | 状态 |
|---|---|---|---|
| 0001消火栓 | 布鲁斯李 | 2019.4.2 15:22:30 | 正常 |
| 0001消火栓 | 布鲁斯李 | 2019.4.2 15:22:30 | 正常 |
| 0001消火栓 | 布鲁斯李 | 2019.4.2 15:22:30 | 正常 |
| 0001消火栓 | 布鲁斯李 | 2019.4.2 15:22:30 | 正常 |
| 0001消火栓 | 布鲁斯李 | 2019.4.2 15:22:30 | 待巡更 |

## 视频监控

## 游客流入流出




## 景区设备点位地图

摄像头 公共厕所 停车场

## 旅游单位数量与产值分布

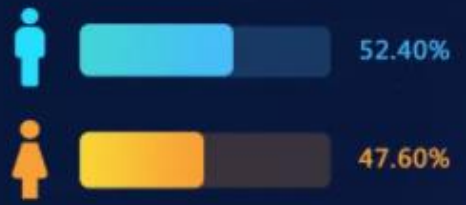
| 单位 | 景点 | 酒店 | 文创商店 | 餐饮 | 娱乐设备 |
|---|---|---|---|---|---|
| 数量 | 51 | 1532 | 63 | 253 | 485 |
| 产值(万元) | 105.5 | 198.2 | 208.7 | 130.6 | 47.7 |

数据概览 | 数据洞察 | 网络分析 | 用户画像

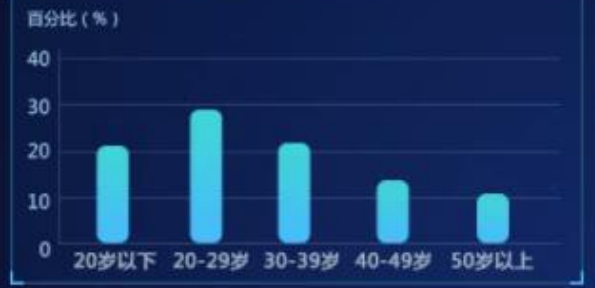
JSCN 江苏有线 股票代码：600959 | 广电宽带

# 广电宽带大数据智能分析

## 性别




## 年龄




## 区域


## 家庭特征


## 上网时段偏好


## 上网时段趋势


## 在线用户统计


## 视频网站偏好

| 排名 | 网站域名 | 点击量 |
|---|---|---|
| 1 | 爱奇艺 | 15800000 |
| 2 | 腾讯视频 | 14220000 |
| 3 | 哔哩哔哩 | 12798000 |
| 4 | 优酷网 | 11518200 |
| 5 | 乐视网 | 10366380 |

## 电商网站偏好

| 排名 | 网站域名 | 点击量 |
|---|---|---|
| 1 | 淘宝网 | 14220000 |
| 2 | 京东商城 | 11660400 |
| 3 | 亚马逊 | 9561528 |
| 4 | 苏宁易购 | 7840452 |
| 5 | 1号店 | 6429171 |

## 社交媒介偏好

| 排名 | 应用名称 | 点击量 |
|---|---|---|
| 1 | 微信 | 17800000 |
| 2 | QQ | 13350000 |
| 3 | 微博 | 10012500 |
| 4 | 百度贴吧 | 7509375 |
| 5 | 陌陌 | 5632031 |

## 网络游戏偏好

| 排名 | 游戏名称 | 点击量 |
|---|---|---|
| 1 | 王者荣耀 | 998000 |
| 2 | 绝地求生 | 898200 |
| 3 | 梦幻西游 | 808380 |
| 4 | 穿越火线 | 727542 |
| 5 | 坦克世界 | 654787 |

实时数据截止到2018年6月


腾讯位置大数据
首页
产品介绍
48,456,914,118
20:27:37
腾讯地图开放平台当日定位次数

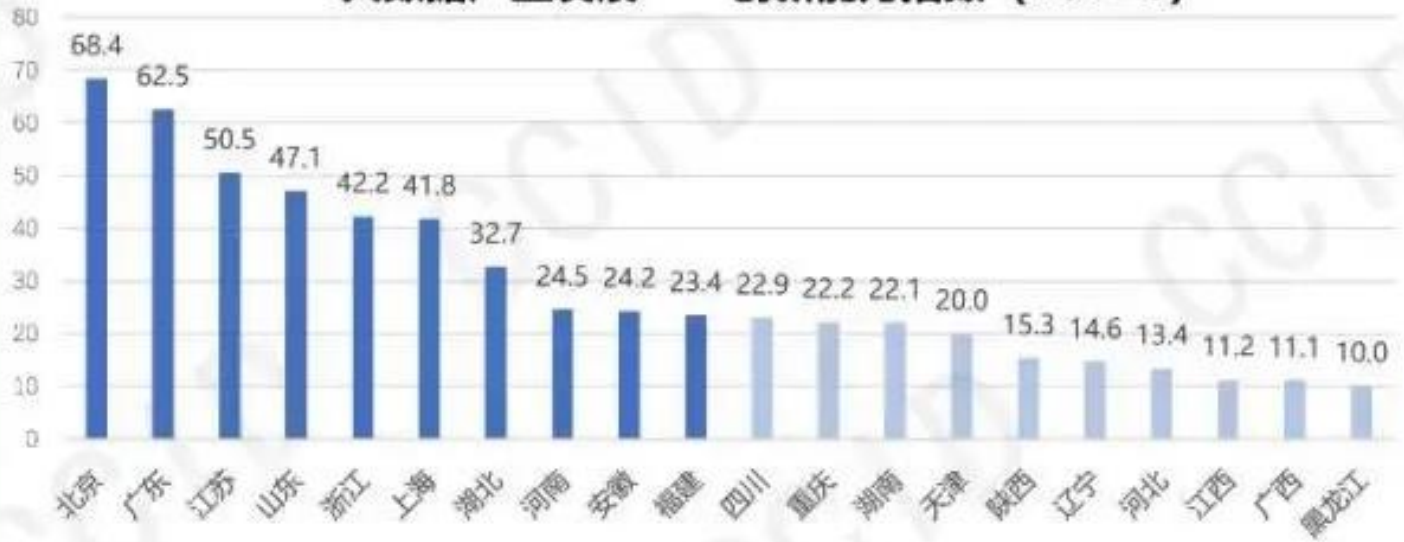
# 产业发展（续2）：创新能力南北双雄，数据资源东部集聚

- 创新能力方面，当前各省市均聚焦大数据等新兴产业领域，把突破关键核心技术作为数字经济政策的发展任务或实施工程，**推动产业链和创新链深度融合。**
  - **北京市**2021年大中型企业研究开发费用3030.6亿元，同比增长31.4%；PCT国际专利申请量10358件，增长25.1%；技术合同成交总额7005.7亿元，增长10.9%。
  - **广东省**2021年R&D经费投入居全国首位，每万人口发明专利拥有量68.76件；拥有6个双跨工业互联网平台；高新技术企业营业收入总额超10万亿元。

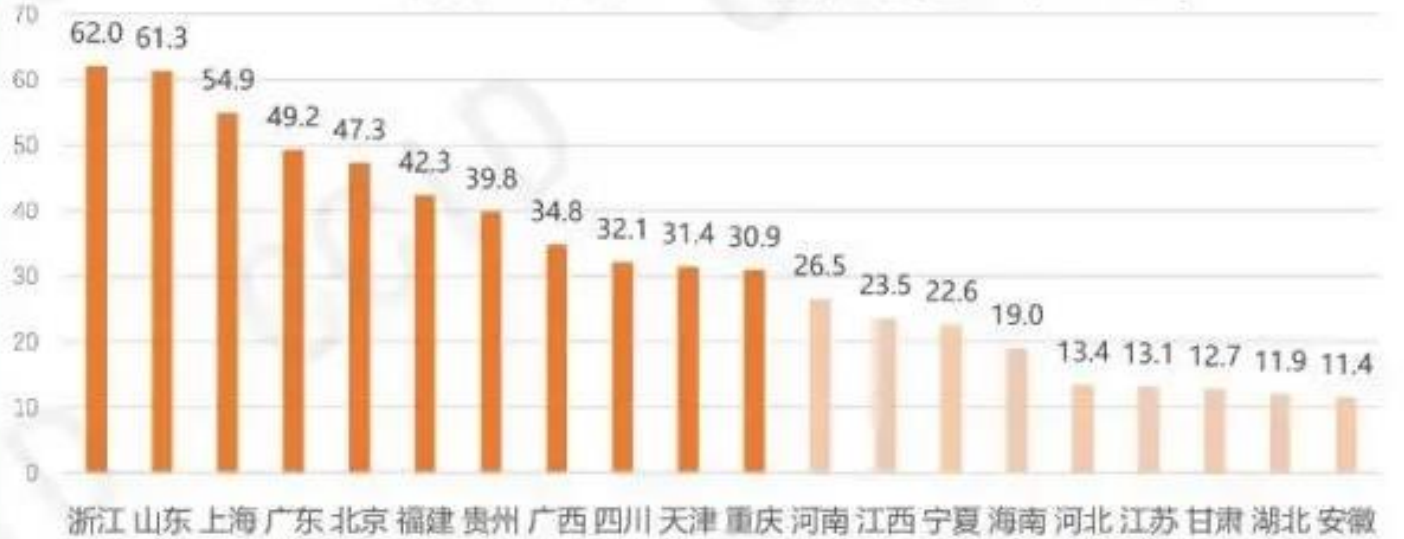
创新投入：全国R&D经费投入为2.79万亿，同比增长14.2%，与GDP之比达2.44%

- 数据资源方面，公共数据开放流通应用进程加快，地方政府数据开放平台数量和开放的有效数据集持续增长。北京、上海等地积极布局新型数据交易所，发展可信环境下的数据流通交易服务。
- 浙江、山东、上海、广东、北京等省市在数据资源体系建设方面全国领先，福建、贵州、广西、四川、天津、重庆等省市多维发力，位于前列。

## 大数据产业发展——创新能力指数（TOP20）

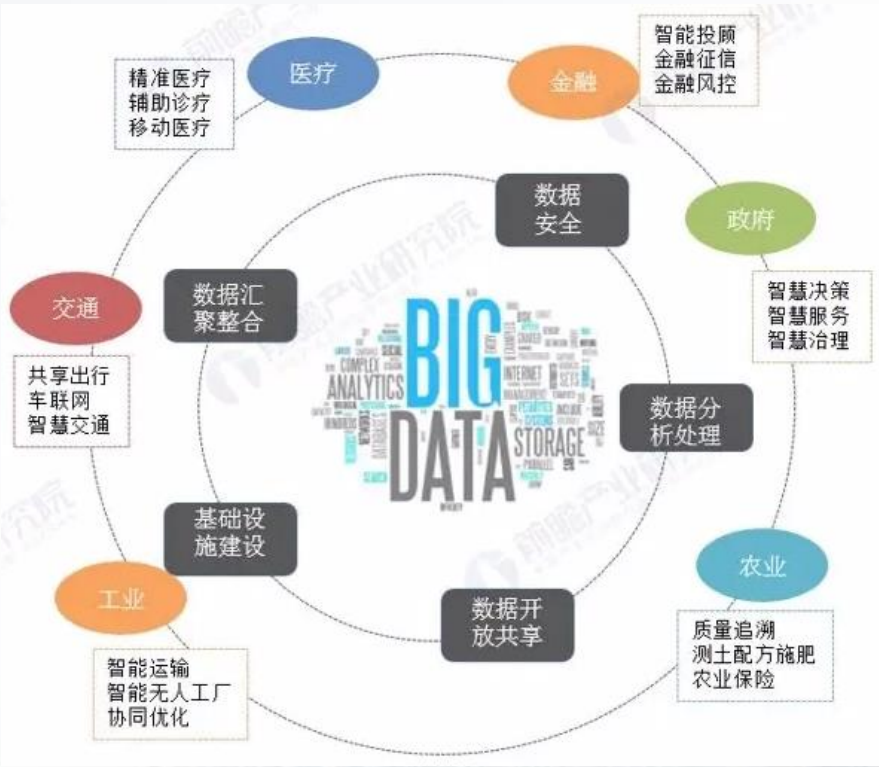



## 大数据产业发展——数据资源指数（TOP20）



# 进入大数据智能时代


## 中国大数据行业应用情况




未来已来,你来不来!

# 大数据与AI

- 两者相辅相成

人工智能的更全面更智慧发展需要依托大数据技术，需要大数据的支撑。

算法、算力与数据是人工智能（AI）发展的“三驾马车”，吴恩达等学者也常说：以数据为中心的AI，或数据驱动的AI。

第一：通过大数据来完成算法训练。
第二：通过大数据来辅助决策。
第三：通过大数据来扩展智能体的应用边界。

# 目录
CONTENTS

# Chapter 1.1

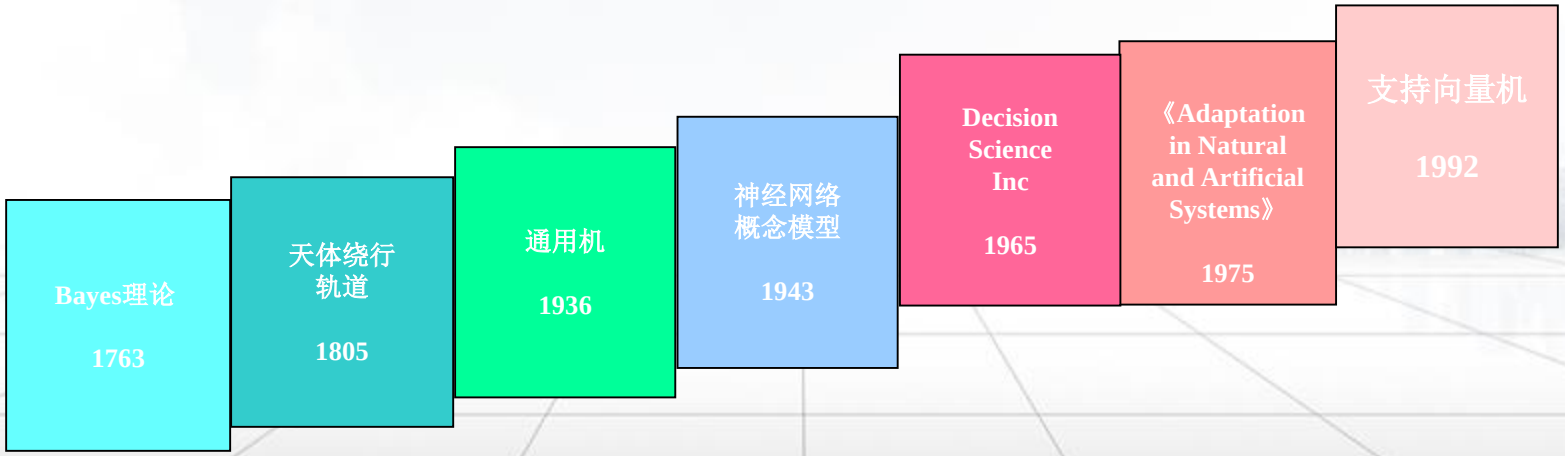
## 基本概念

# 问题？

- 大数据分析和挖掘的定义、目的、技术？
- 大数据分析和挖掘有何区别？

# 从“小”到“大”的数据分析处理

- Bayes理论（根据过去的数据预测未来发生的概率） 1763年
- 确定了天体绕行太阳的轨道（通过研究前人大量的观测数据） 1805年
- 通用图灵机（用机器计算） 1936年
- 神经网络的概念模型（用机器模拟神经元） 1943年
- Decision Science Inc （根据进化计算提供现实问题解决方案的公司） 1965年
- 《Adaptation in Natural and Artificial Systems》（《自然与人造系统的适应性》） 1975年
- 支持向量机（在统计学基础上发展起来的数据挖掘方法） 1992年



# 人工智能技术的典型应用事件


深蓝战胜国际象棋大师
KASPAROV, 1997


谷歌AlphaGo击败
欧洲围棋冠军樊麾, 2016


人工智能研究公司OPEN AI的第
三代自然语言模型GPT-3, 2020


沃森在美国Jeopardy (危险边缘)
节目击败人类选手, 2011

美交管局认定谷歌自动驾驶
系统为“驾驶员”, 2016

Neuralink公司发布的猴子用意
念玩乒乓球游戏的视频, 2021

# 数据信息，人类并没有充分利用

- **卫星遥感图像，目前用得上的不到5%，剩下的95%都被浪费了。**
- **虽然人类基因组测序已完成，但其中，现在能读懂的还不到10%，大部分仍是“天书”。全世界的生命科学界都把基因测序搬到中国。原因很简单——中国有足够大的样本。然而，在中国做完测序，数据却被对方拿走、分析。**

----2013年上海工博会“院士圆桌会议”

**物质、能量、环境资源不能浪费，**

**数据资源同样也不能浪费！---如何利用大数据?**

# 大数据带来的思考？

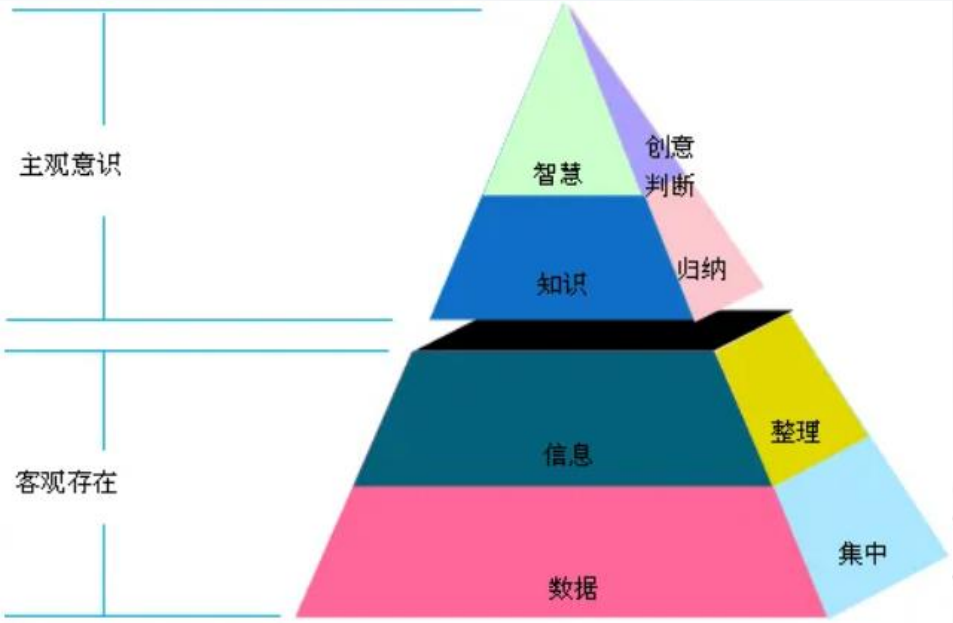
- **资源视角：** 大数据是新资源，体现了一种全新的资源观。
- **技术视角：** 大数据代表了新一代数据管理与分析技术。
- **理念视角：** 大数据打开了一种全新的思维角度。


# 数据中的知识发现

- **数据清理：** 消除噪声和删除不一致数据。
- **数据集成：** 多种数据源可以组合在一起，形成数据集市或数据仓库。
- **数据选择：** 从数据库中提取与分析任务相关的数据。
- **数据变换：** 通过汇总或聚集操作，把数据变换统一成适合挖掘的形式。
- **数据挖掘：** 使用智能方法提取数据模式。
- **模式评估：** 根据某种兴趣度量，识别代表知识的真正有趣的模式。
- **知识表示：** 使用可视化和知识表示技术，向用户提供挖掘的知识。

## 1. 数据分析

- 数据分析是指采用适当的统计分析方法对收集到的数据进行分析、概括和总结，对数据进行恰当地描述，提取出有用的信息的过程。
- **大数据分析的意义：**透过多维度、多层次的数据，以及历史关联数据，找到问题的症结，发现事实的真相。



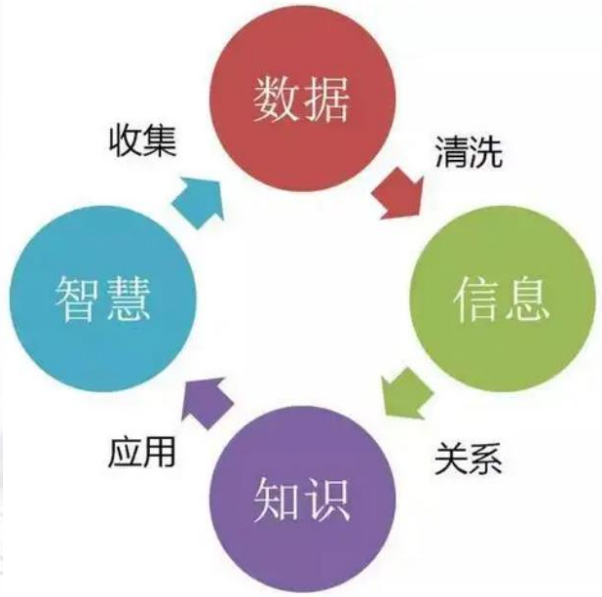

数据金字塔

为达到目标而运用知识的能力，创新/判断

用于解决问题的结构化信息，归纳/演绎

有价值的数据，处理/有逻辑

原始的数据，客观的抽象



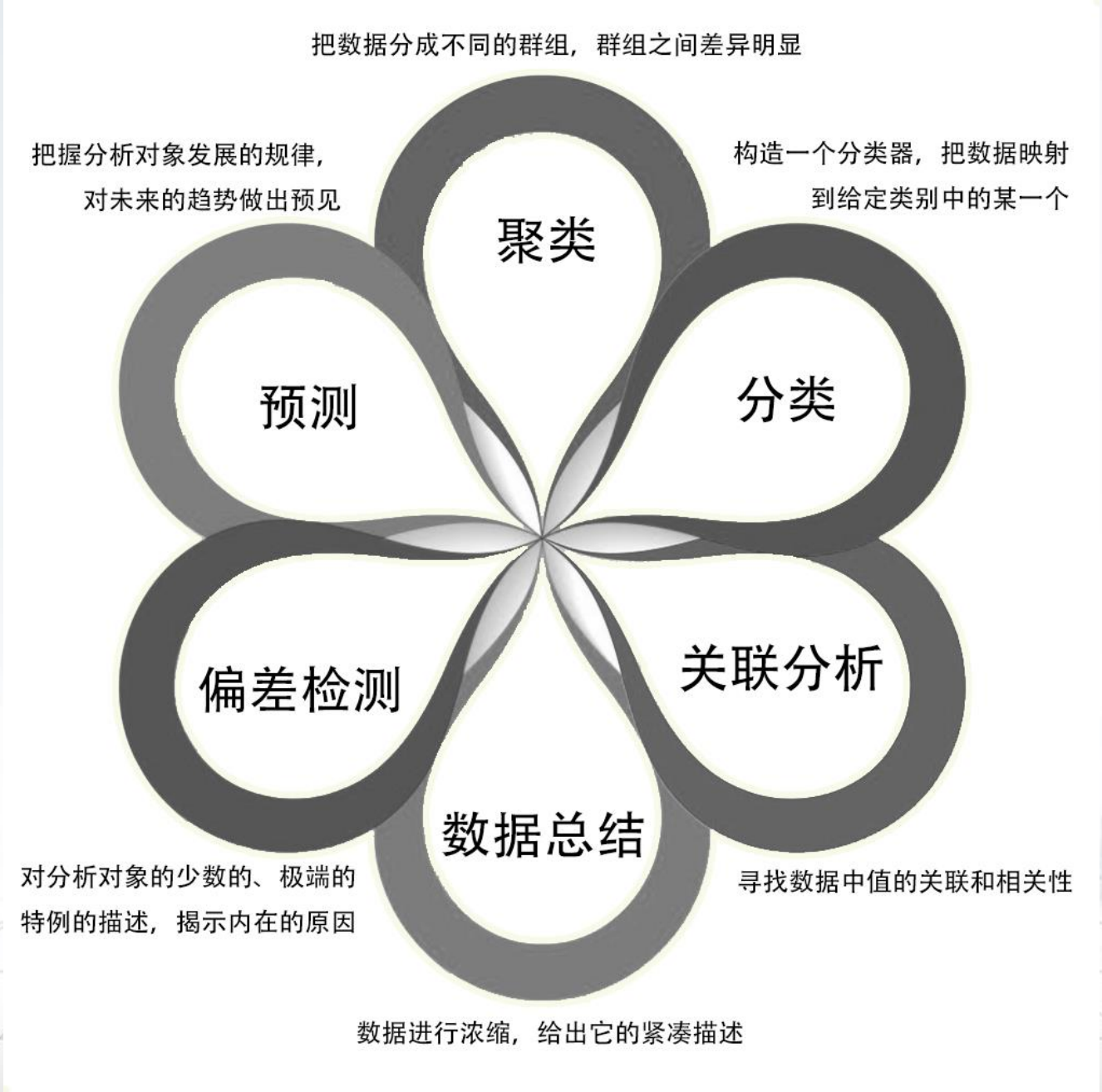
## 2. 数据挖掘


- 数据挖掘(Data Mining，DM)是指从海量的数据中通过相关的算法来发现隐藏在数据中的规律和知识的过程。
  - 汇聚了不同领域的研究人员，如数据库、人工智能、数理统计、并行计算等人才，将多领域专家学者引入到数据挖掘领域，形成好多新的挖掘热点。
  - 挖掘出来的知识必须是用户感兴趣的、能用得上的。
  - 发现的知识不具有通用性，也不是要去发现崭新的自然科学定理、数学公式，更不是定理的证明，实际上所发现的知识都是面向特定问题、特定领域的。
  - 挖掘的知识可以被用于信息管理、查询优化、决策支持、过程控制、预测未来等，还可以用于数据自身的维护。

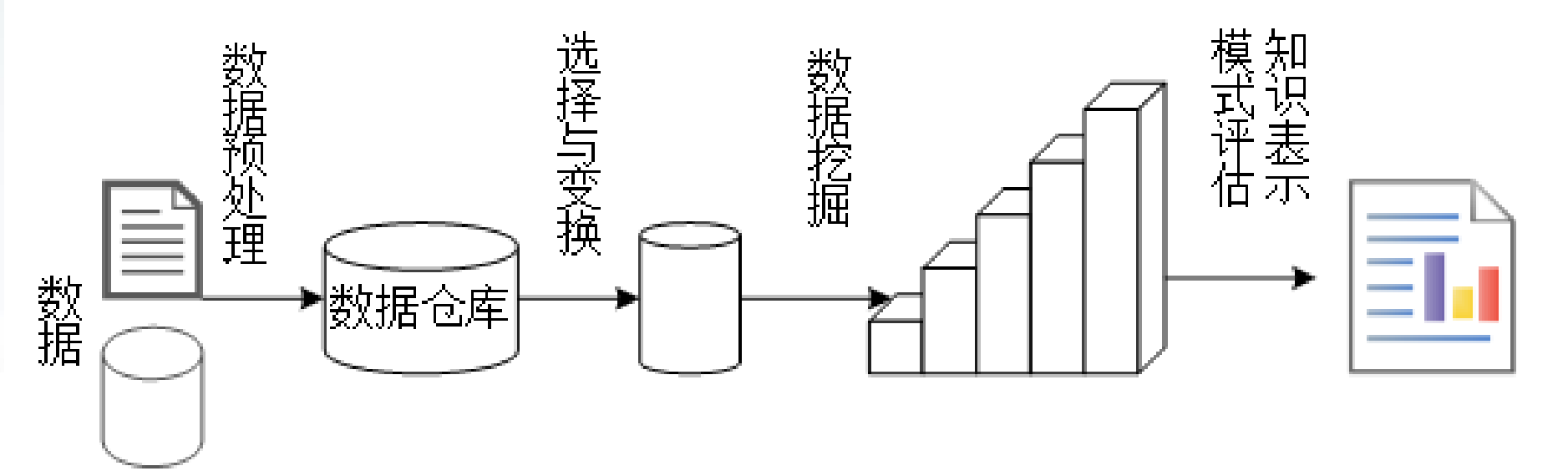
# 为什么进行数据挖掘?

- 数据的爆炸式增长: 从TB到PB
  - 丰富数据的主要来源
    - 商业: Web, 电子商务, 交易, 股票, ...
    - 科学: 遥感, 生物信息学, 科学仿真, ...
    - 社会与个人: 新闻, 数码相机, YouTube
  - 数据采集与数据可用性
    - 自动数据收集工具, 数据库系统, Web, 计算机化的社会
- 数据是丰富的，急需发现知识!

# 数据挖掘的主要功能分类



## 3. 知识发现（KDD）的过程




- 通常将数据挖掘视为数据中“知识发现”的同义词，也可以认为数据挖掘是知识发现中的一个步骤。

## 4. 数据分析与数据挖掘的区别

| 内容 | 数据分析 | 数据挖掘 |
| --- | --- | --- |
| 处理的数据量 | 不一定很大 | 海量 |
| 目标 | 比较明确 | 不明确的 |
| 侧重点 | 展现数据之间的关系 | 对未知的情况进行预测和估计 |

## 5. 数据分析与数据挖掘的联系

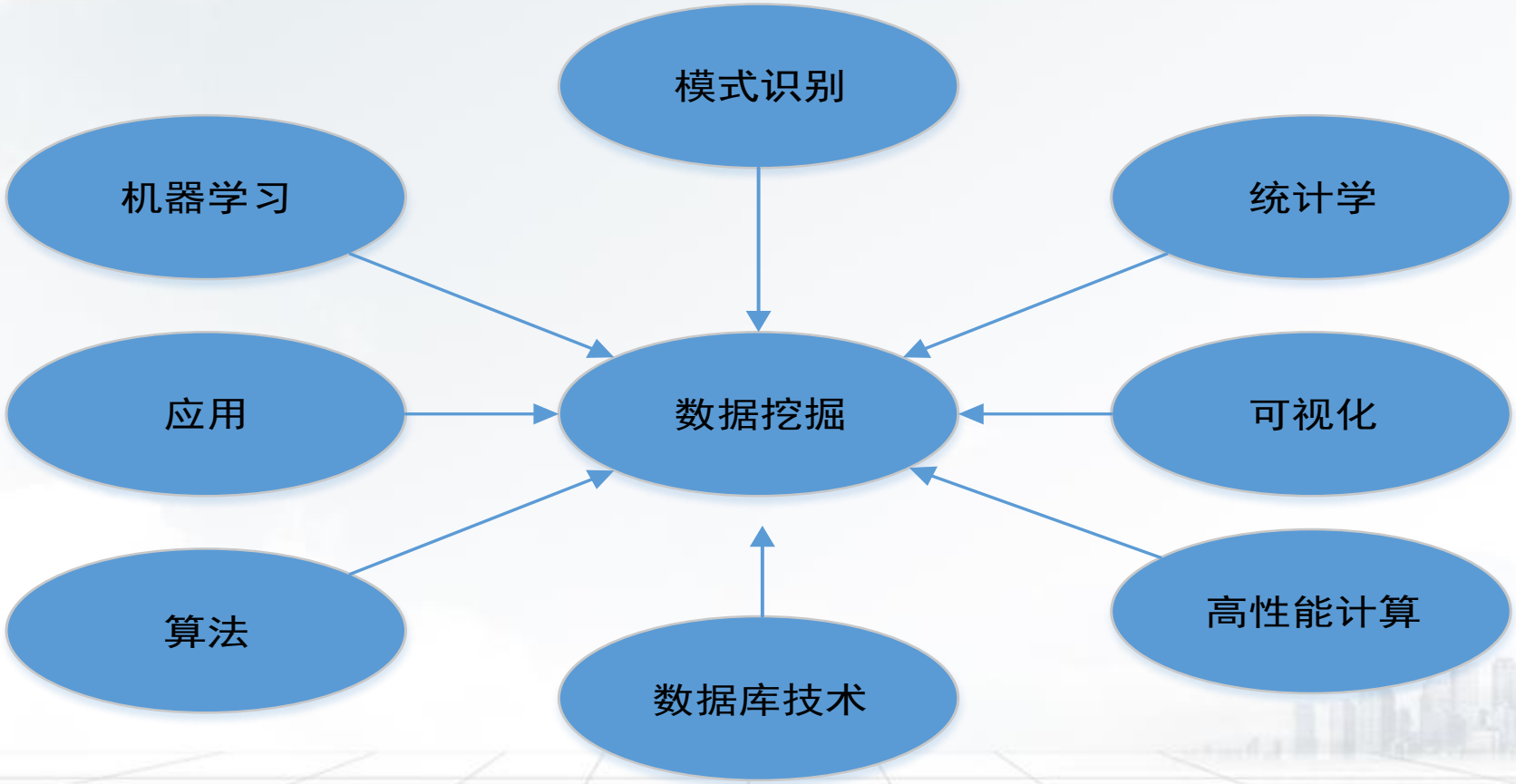
- 数据分析的结果往往需要进一步的挖掘才能得到更加清晰的结果，而数据挖掘发现知识的过程也需要对先验约束进行一定的调整而再次进行数据分析。
- 数据分析可以将数据变成信息，而数据挖掘将信息变成知识，如果需要从数据中发现知识，往往需要数据分析和数据挖掘相互配合，共同完成任务。

## 6. 数据分析与数据挖掘的主要方法

- **频繁模式：**从样本数据集中频繁出现的模式，是经常一起出现的模式。“模式”是一个比较抽象的概念，举个例子，比如在超市的交易系统中，记载了很多次交易，每一次交易的信息包括用户购买的商品清单。如果超市主管是个有心人的话，他会发现尿不湿，啤酒这两样商品在许多用户的购物清单上都出现了，而且频率非常高。尿不湿，啤酒同时出现在一张购物单上就可以称之为一种频繁模式，这样的发掘就可以称之为频繁模式挖掘。
- **离群点分析：**离群点是指全局或局部范围内偏离一般水平的观测对象。例如：当发现某个人的信用卡在不经常消费的地区短时间内消费了大量的金额，则可以认定这张卡的使用情况异常，可以作为离群点数据。
- **分类与回归：**包括决策树， 朴素贝叶斯分类，支持向量机，神经网络， 规则分类器， 基于模式的分类，逻辑回归 ...
- **聚类分析：**就是把一些对象划分为多个组或者“聚簇”，从而使得同组内对象间比较相似而不同组对象间差异较大。

# 7. 数据分析与数据挖掘使用的主要技术

数据挖掘是一门涉及面较广的交叉学科



## 8. 数据分析与数据挖掘的应用场景

- **商务智能：**通过数据挖掘等技术可以获得隐藏在各种数据中的有利信息，从而帮助商家进一步调整营销策略。
- **信息识别：**信息接受者从一定的目的出发，运用已有的知识和经验，对信息的真伪性、有用性进行辨识和甄别。
- **搜索引擎：**根据用户提供的关键词，在互联网上搜索用户最需要的内容。
- **辅助医疗：**对大量历史诊断数据进行分析和挖掘，有助于医生对病人的病情进行有效的判断。

# 9. 数据分析与数据挖掘存在的问题

- 数据类型的多样性
- 高维度数据
- 噪声数据
- 分析与挖掘结果的可视化
- 隐私数据的保护

# Chapter 1.2

## 数据的属性

## 1. 数据对象

- 数据集由数据对象组成。一个数据对象代表一个实体。

  例如：

  - 销售数据库：顾客、 商品、销售
  - 医疗数据库：患者、医生、诊断治疗
  - 选课数据库：学生、教师、课程
- 数据对象又称为样本、实例、数据点、对象或元组。
- 数据对象用属性描述。数据表的行对应数据对象; 列对应属性（关系数据）。

## 2. 属性（Attributes）

- **属性**(特征，变量)是一个数据字段，表示数据对象的一个特征。

例如：客户编号、姓名、地址等

商品编号、商品名、价格、种类等

## 3. 属性类型

- 标称属性(nominal)
- 二元属性(binary)
- 序数属性(ordinal)
- 数值属性(numeric)
  - 区间标度属性(interval-scaled)
  - 比率标度属性(ratio-scaled)

- 标称属性(nominal attribute)：类别，状态或事物的名字
  - 每个值代表某种类别、编码或状态，这些值不必具有有意义的序，可以看做是枚举的

    例如：头发颜色= {赤褐色，黑色，金色，棕色，褐色，灰色，白色，红色}
  - 也可以用数值表示这些符号或名称，但并不定量地使用这些数。

    例如： 婚姻状况，职业，ID号，邮政编码，

    可以用0表示未婚、1表示已婚

- 二元属性(binary attribute)：布尔属性，是一种标称属性，只有两个状态：0或1。
  - 对称的(symmetric): 两种状态具有同等价值，且具有相同的权重。

    例如：性别
  - 非对称的(asymmetric): 其状态的结果不是同样重要。

    例如：体检结果（阴性和阳性），惯例：重要的结果用1编码（如，HIV阳性）。

- 序数属性(ordinal attribute)，其可能的值之间具有有意义的序或者秩评定(ranking)，但是相继值之间的差是未知的。

  例如：尺寸={小，中，大}，军衔，职称

- 序数属性可用于主观质量评估

  例如：顾客对客服的满意度调查。0-很不满意；1-不太满意；1-基本满意；3-满意；4-非常满意

- **数值属性(numeric attribute)**:定量度量，用整数或实数值表示
  - 区间标度(interval-scaled)属性：使用相等的单位尺度度量。值有序，可以评估值之间的差，不能评估倍数。没有绝对的零点。

    例如：日期，摄氏温度，华氏温度
  - 比率标度(ratio-scaled)属性：具有固定零点的数值属性。值有序，可以评估值之间的差，也可以说一个值是另一个的倍数。

    例如：开式温标(K)，重量，高度，速度

## 离散属性VS连续属性

- **离散属性**(discrete Attribute)：具有有限或者无限可数个值。有时，表示为整型量。

  例如：邮编、职业或文库中的字集

  二进制属性是离散属性的一个特例

- **连续属性**(Continuous Attribute)：属性值为实数，一般用浮点变量表示。

  例如，温度，高度或重量，实际上，真实值只能使用一个有限的数字来测量和表示。

# Chapter 1.3

## 数据的基本统计描述

- 目的
  - 更好地识别数据的性质，把握数据全貌。
- 数据的基本统计描述
  - 中心趋势度量、数据分散度量、数据的图形表示
- 中心趋势度量
  - 均值、加权算数均值、中位数、众数、中列数
- 数据分散度量
  - 极差、分位数和四分位数、方差和标准差
- 数据的图形显示
  - 箱图、饼图、频率直方图、散点图

# 1. 中心趋势度量

## – 均值（Mean）

- 令$x_1$，$x_2$，…，$x_N$为某数值属性$X$的$N$个观测值，该值集合的均值如式(1-1)所示。

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \qquad (1\text{-}1)$$

## – 截尾均值

指在一个数列中，去掉两端的极端值后所计算的算术平均数。

# 1. 中心趋势度量

## – 加权算数平均数（Weighted Mean）

- 对于*i=1*，…，*N*，每个值$x_i$都有一个权重$w_i$。

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N} \qquad (1\text{-}2)$$

例：某同学的某一科的考试成绩：平时测验 80，期中 90，期末 95。科目成绩的计算方式是：平时测验占20%，期中成绩占30%，期末成绩占50%。这里，每个成绩所占的比重为权重。那么，

$$\bar{x} = \frac{80 \times 20\% + 90 \times 30\% + 95 \times 50\%}{20\% + 30\% + 50\%} = 90.5$$

## 1. 中心趋势度量

– 中位数(Median)：正中间的值

- 如果值有奇数个，取中间值，否则取中间两个数的平均值
- 有序数据值的中间值
- 如果观察值有偶数个，通常取最中间的两个数值的平均数作为中位数。

例：数据按递增排序为：33，45，60，65，70，77，80，90，100，100。有10个观测值，因此中位数不唯一。中间两个值为70和77，则中位数为

$$\frac{70+77}{2}=73.5$$

# 1. 中心趋势度量

## – 分组数据中位数(Grouped Median)

- 根据N/2确定中位数所在的组

$$M_e = L + \frac{\frac{N}{2} - S_{m-1}}{f_m} \times d \qquad (1\text{-}3)$$

$M_e$：中位数，$L$：中位数所在组的下限，$S_{m-1}$：中位数所在组以下各组的累计频数，$f_m$：中位数所在组的频数，$d$：中位数所在组的组距。

# 1. 中心趋势度量

## – 分组数据中位数

例：表1-1为某公司员工薪酬的分组数据，计算数据的近似分组数据中位数。

①判断中位数区间：

N=110+180+320+460+850+250+130+70+20+10=2400;

N/2=1200;

因为：110+180+320+460=1070<1200<1070+850=1920;

所以：1900~1999为对应区间。

②这里有：L=1900，N=2400，$S_{m-1}$=1070，$f_m$=850 中位数组

d=100，由式(1-3)得：

$$M_e = 1900 + \frac{\frac{2400}{2} - 1070}{850} \times 100 \approx 1915.29$$

因此，近似分组数据中位数为1915.29。

表1-1员工薪酬分组数据

| Salary | Frequency |
|---|---|
| 1500~1599 | 110 |
| 1600~1699 | 180 |
| 1700~1799 | 320 |
| 1800~1899 | 460 |
| 1900~1999 | 850 |
| 2000~2099 | 250 |
| 2100~2199 | 130 |
| 2200~2299 | 70 |
| 2300~2399 | 20 |
| 2400~2499 | 10 |

1070

1200

1920

# 1. 中心趋势度量

– 众数(Mode)：数据中出现最频繁的值

例：数据按递增序排序为：33，45，60，65，70，77，80，90，100，100。mode=100

- 可能最高频率对应多个不同值，导致多个众数
- 经验公式:

$$mean - mode = 3 \times (mean - median)$$

## 1. 中心趋势度量

– 中列数(Midrange)：数据集中最大值和最小值的算术平均值

例：数据按递增序排序为：33，45，60，65，70，77，80，90，100，100。

最小值和最大值分别为33和100，则中列数为

$$\frac{33+100}{2}=66.5$$

## 2. 数据分散度量

– 极差（又称全距，Range）：是集合中最大值与最小值之间的差距，即最大值减最小值后所得数据。

例：数据按递增序排序为：33，45，60，65，70，77，80，90，100，100。

100-33=67

## 2. 数据分散度量

- 分位数（Quantile）：取自数据分布的每隔一定间隔上的点，把数据划分成基本上大小相等的连贯集合。

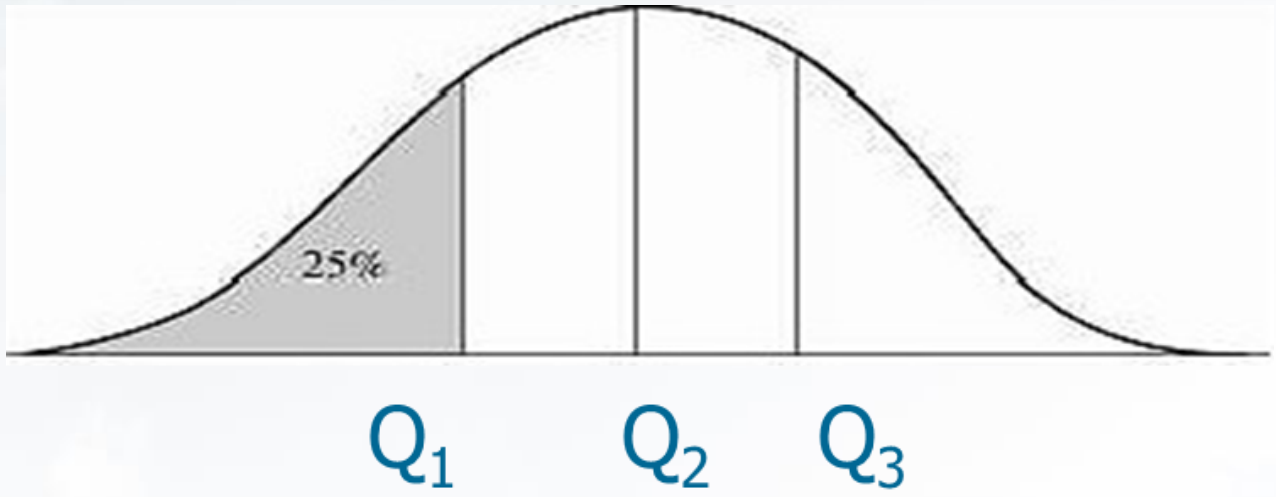



图1-1 某变量X的数据统计描述显示

- 给定数据分布的第k个q-分位数的值为x，使得小于x的数据值最多为k/q，而大于x的数据值最多为(q-k)/q，其中k是整数，使得0<k<q。这里有q-1个q-分位数。

## 2. 数据分散度量

– 四分位数（Quantile）：把数据分布划分成4个相等的部分，使得每部分表示数据分布的四分之一。这3个数据点称为四分位数。

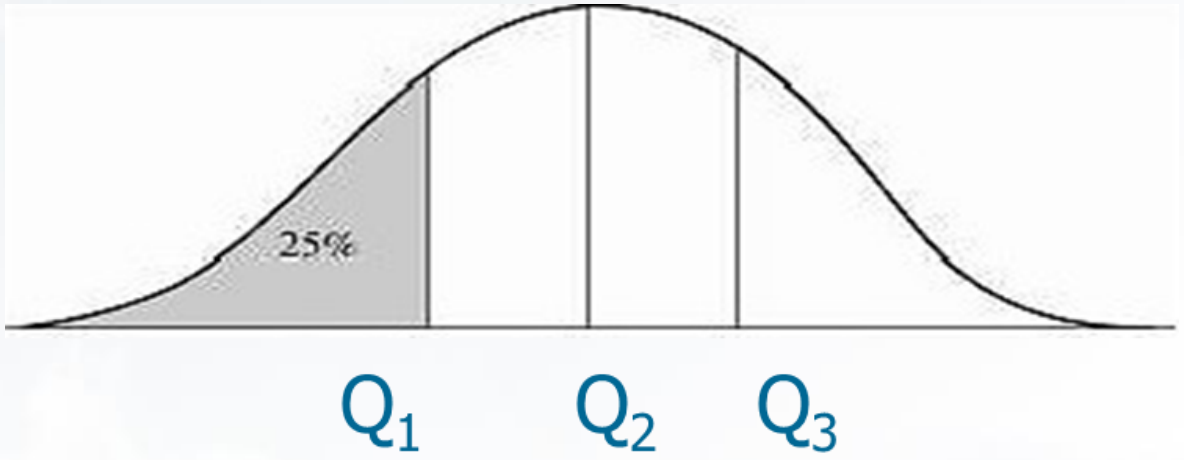



图1-1 某变量X的数据统计描述显示

– Q1：“下四分位数”；Q2：“中位数”；Q3：“上四分位数”。

## 2. 数据分散度量

– 四分位数极差（InterQuartile Range，IQR）：Q1和Q3之间的距离。

$$IQR = Q_3 - Q_1 \qquad (1\text{-}4)$$

确定四分位数的位置：

Q1的位置= (n+1)/4=(n+1) × 0.25

Q2的位置=2*(n+1)/4= (n+1) × 0.5

Q3的位置= 3*(n+1)/4=(n+1) × 0.75

n表示项数

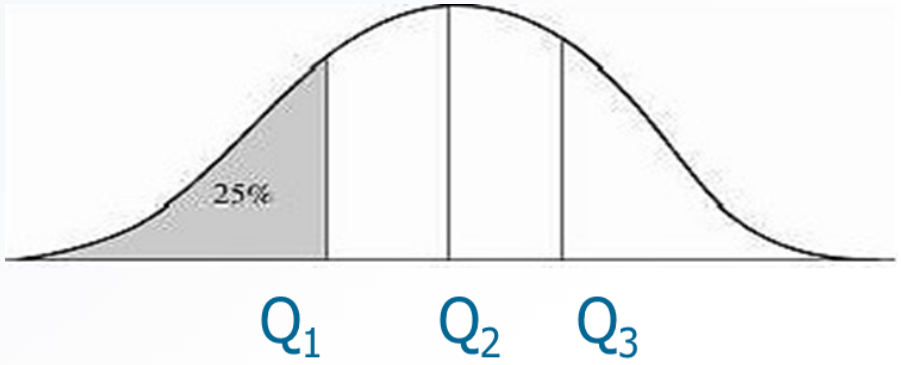



图1-1 某变量X的数据统计描述显示

## 2. 数据分散度量

– 四分位数极差（InterQuartile Range，IQR）：Q1和Q3之间的距离。

$$IQR = Q_3 - Q_1 \qquad (1\text{-}4)$$

另一种确定四分位数的位置：

Q1的位置=1+(n-1) × 0.25

Q2的位置=1+(n-1) × 0.5

Q3的位置=1+(n-1) × 0.75

n表示项数

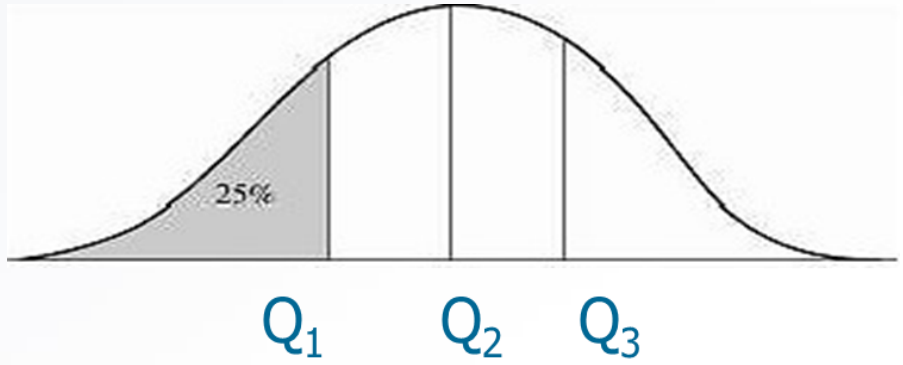



图1-1 某变量X的数据统计描述显示

## 2. 数据分散度量

– 四分位数极差（InterQuartile Range，IQR）：Q1和Q3之间的距离。

$$IQR = Q_3 - Q_1 \qquad (1\text{-}4)$$

例：由8人组成的旅游小团队年龄分别为：17，19，22，24，25，28，34，37，求其年龄的四分位差。

①计算Q1与Q3的位置：

Q1的位置=(n+1)/4=(8+1)/4=2.25；Q3的位置=3*(n+1)/4=3*(8+1)/4=6.75

**17，19，22，24，25，28，34，37**

②确定Q1与Q3的数值：

Q1=19+（22-19）*0.25=19.75；Q3=28+（34-28）*0.75=32.5

③计算四分位差：

IQR=Q3-Q1=32.5-19.75=12.75

## 2. 数据分散度量

- 方差（样本方差）：是每个数据分别与平均数之差的平方的平均数。
- 总体方差：

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = (\frac{1}{N}\sum_{i=1}^{N}x_i^2) - \bar{x}^2 \qquad (1\text{-}5)$$

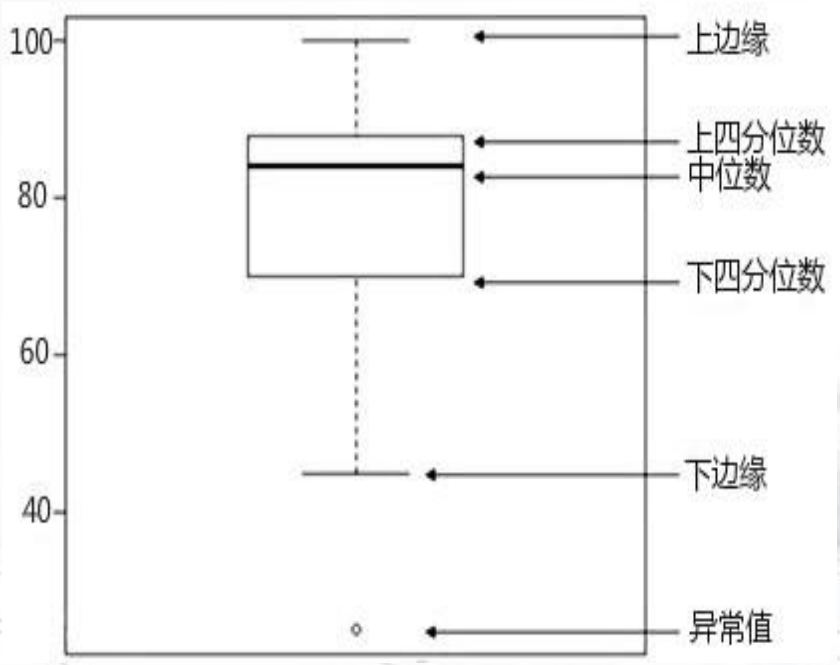
- 样本方差：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^{\,2} - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

- 标准差：方差的平方根

## 3. 数据的图形显示

- 盒图（又称箱线图，Box-plot），是一种用来描述数据分布的统计图形，可以表现观测数据的中位数、四分位数和极值等描述性统计量。
  - 用盒子表示数据
  - 盒子的端点在四分位数上，使得盒子长度为四分位数极差IQR
  - 中位数用盒内线标记
  - 盒子外线延伸到最小和最大的观测值
  - 离群点：绘制在离群阈值范围外的点



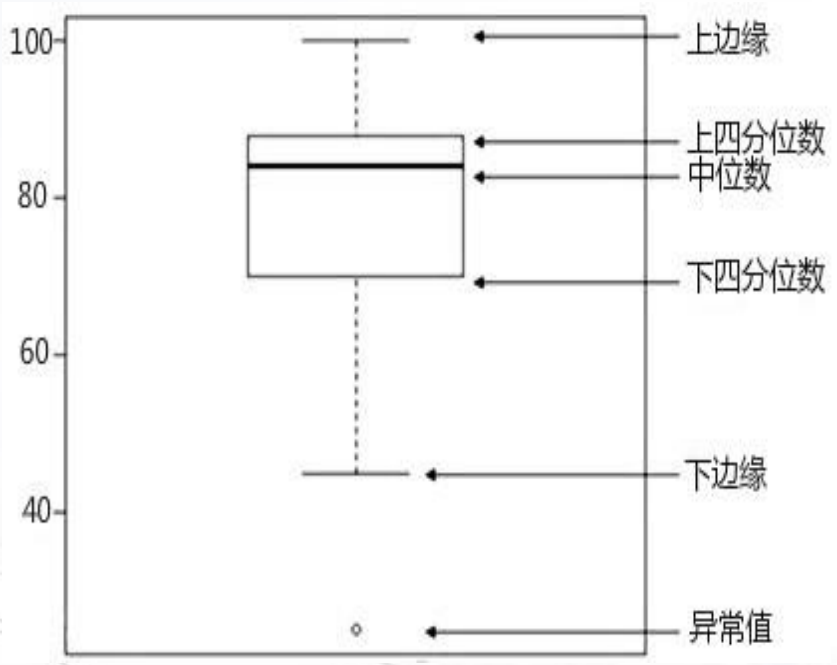
## 3. 数据的图形显示

- 盒图（又称箱线图，Box-plot），是一种用来描述数据分布的统计图形，可以表现观测数据的中位数、四分位数和极值等描述性统计量。

- 五数概括：min, Q1, median, Q3, max

- 盒图：分布直观表示，体现五数概括

- 离群点: 第三个四分位数之上或者第一个四分位数之下至少1.5 x IQR的值



## 3. 数据的图形显示

– 饼图（又称圆形图或饼形图，Pie Graph），通常用来表示整体的构成部分及各部分之间的比例关系。饼图显示一个数据系列中各项的大小与各项总和的比例关系。

例：使用饼图表示不同年龄区间的人参与某活动的情况

表1-4 某活动覆盖人群

| 年龄区间 | 参与人数 |
|---|---|
| 19岁及以下 | 270 |
| 20-29岁 | 1248 |
| 30-39岁 | 1080 |
| 40-49岁 | 280 |
| 50岁及以上 | 180 |

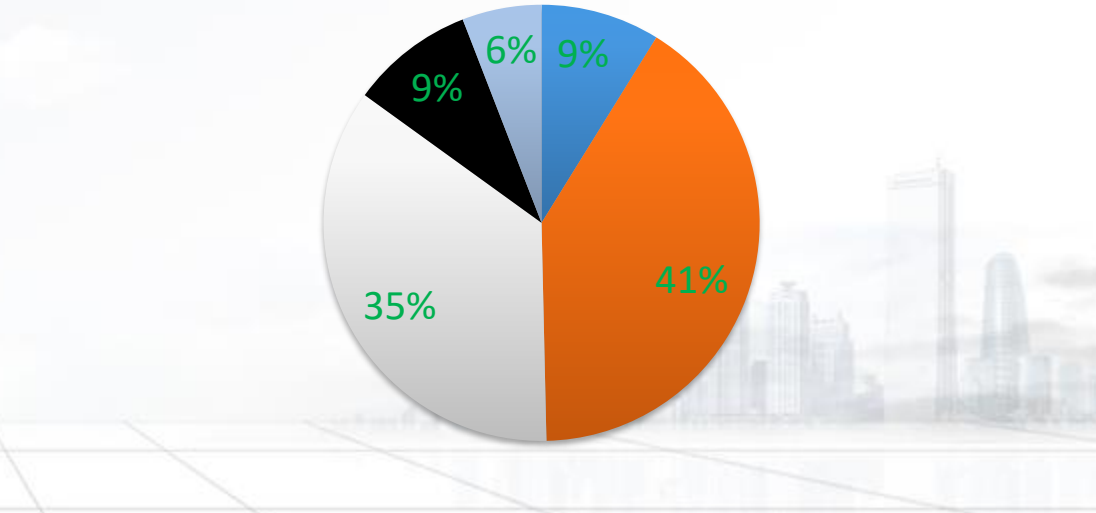



图1-4 某活动覆盖人群饼图

## 3. 数据的图形显示

– 频率直方图（又称频率分布直方图，Frequency Histogram），是在统计学中表示频率分布的图形。

例：使用直方图表示学生数学成绩的分布

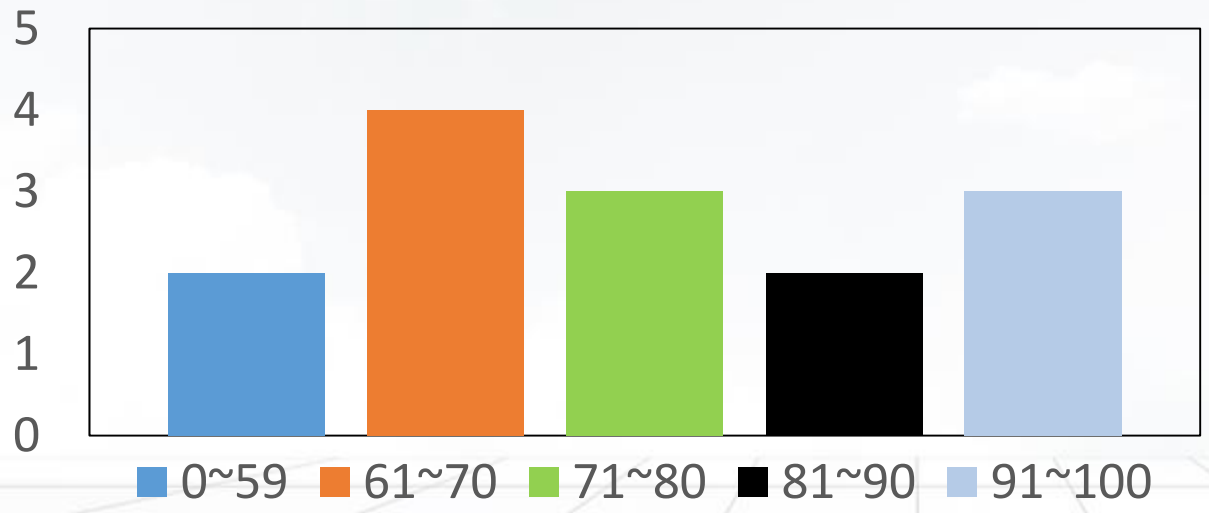



图1-5 学生成绩数据频率直方图

表1-5 学生数学成绩

| 学号 | 成绩 |
|---|---|
| 701 | 60 |
| 702 | 71 |
| 703 | 56 |
| 704 | 99 |
| 705 | 66 |
| 706 | 90 |
| 707 | 100 |
| 708 | 66 |
| 709 | 77 |
| 710 | 60 |
| 711 | 88 |
| 712 | 79 |
| 713 | 83 |
| 714 | 55 |

## 3. 数据的图形显示

– 散点图（Scatter Diagram）：将样本数据点绘制在二维平面或三维空间上，根据数据点的分布特征，直观地研究变量之间的统计关系以及强弱程度。

例：使用散点图表示物流收货天数和客户满意度之间的关系

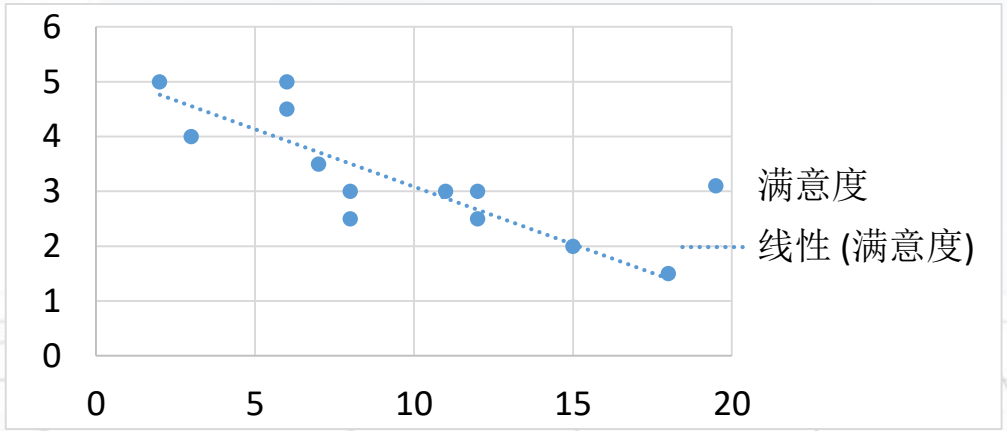


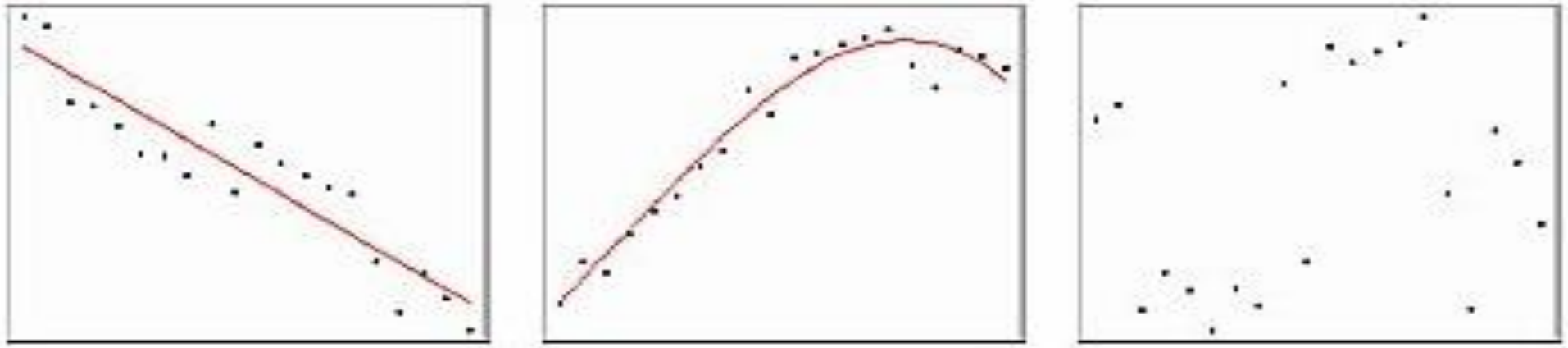

图1-7 物流收货天数和客户满意度散点图

表1-6 物流收货天数和客户满意度相关数据

| 物流收货天数 | 客户满意度 |
|---|---|
| 6 | 4.5 |
| 12 | 3 |
| 8 | 3 |
| 6 | 5 |
| 18 | 1.5 |
| 7 | 3.5 |
| 3 | 4 |
| 8 | 2.5 |
| 11 | 3 |
| 2 | 5 |
| 12 | 2.5 |
| 15 | 2 |

## 3. 数据的图形显示

– 散点图（Scatter Diagram）



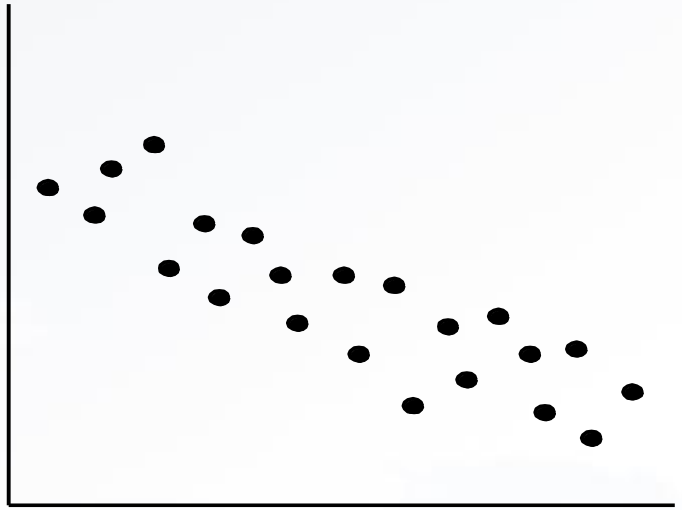
（a） 线性相关　　（b）非线性相关　　（c）不相关

图1-6 散点图中属性之间的相关性

## 3. 数据的图形显示

– 散点图（Scatter Diagram）

（d） 正相关



（e）负相关

图1-6 散点图中属性之间的相关性

## 基本统计图

### 盒图Boxplot

描述五数概括

### 直方图 Histogram

x-axis 表示数值大小,
y-axis 表示频率

### 饼图 Pie Graph

显示一个数据系列中各项的大小与各项总和的比例关系

### 散点图 Scatter plot

每个值视作一个坐标对，作为一个点画在平面上

# Chapter 1.4

# 数据的相似性与相异性

# 1.4 数据的相似性与相异性

- 相似性(Similarity)
  - 两个对象相似程度的数量表示
  - 数值越高表明相似性越大
  - 通常取值范围为[0,1]
- 相异性(Dissimilarity)(例如距离)
  - 两个对象不相似程度的数量表示
  - 数值越低表明相似性越大
  - 相异性的最小值通常为0
  - 相异性的最大值（上限）是不同的
- 邻近性(Proximity):相似性和相异性都称为邻近性

## 1. 数据矩阵与相异矩阵

- 数据矩阵：对象-属性结构
  - 行-对象：n个对象
  - 列-属性：p个属性
  - 二模矩阵(Two modes)
- 相异性矩阵：对象-对象结构
  - n个对象两两之间的邻近度
  - 对称矩阵
  - 单模(Single mode)

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

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

## 2. 标称属性的邻近性度量

– 相异性

$$d(i,j)=\frac{p-m}{p}=1-\frac{m}{p} \qquad (1\text{-}8)$$

- p是对象的属性总数，m是匹配的属性数目（即对象i和j状态相同的属性数）

– 相似性

$$sim(i,j)=1-d(i,j)=\frac{m}{p} \qquad (1\text{-}9)$$

## 2. 标称属性的邻近性度量

例：计算标称属性的相异性矩阵

表1-7 标称属性数据

| 对象标识符 | 属性 |
|---|---|
| 1 | A |
| 2 | B |
| 3 | C |
| 4 | A |

相异性计算：

$$d(2,1)=1-\frac{0}{1}=1 \quad d(3,1)=1-\frac{0}{1}=1$$
$$d(3,2)=1-\frac{0}{1}=1 \quad d(4,1)=1-\frac{1}{1}=0$$
$$d(4,2)=1-\frac{0}{1}=1 \quad d(4,3)=1-\frac{0}{1}=1$$

相异性矩阵：

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

### 3. 二进制属性的邻近性度量

– 相异性

- 对称的二进制属性

$$d(i,j)=\frac{p+n}{m+n+p+q}=\frac{p+n}{sum} \qquad (1\text{-}10)$$

- 非对称的二进制属性

$$d(i,j)=\frac{p+n}{m+n+p} \qquad (1\text{-}11)$$

– $m$是对象$i$ 和$j$ 都取1的属性数，$n$、$p$、$q$类似如表所示。

| $i$ \ $j$ | 1 | 0 | 合计 |
|---|---|---|---|
| 1 | $m$ | $n$ | $m+n$ |
| 0 | $p$ | $q$ | $p+q$ |
| 合计 | $m+p$ | $n+q$ | $sum$ |

– 相似性

$$sim(i,j)=1-d(i,j) \qquad (1\text{-}12)$$

### 3. 二进制属性的邻近性度量

例：计算二进制属性的相异性

表1-9 居民家庭情况调查表

| 姓名 | 婚否 | 买房否 | 买车否 |
|---|---|---|---|
| 张明 | Y | N | N |
| 李思 | N | Y | Y |
| 王刚 | Y | Y | N |

$$d(\text{张明},\text{李思}) = \frac{\text{NY}+\text{YN}}{\text{YY}+\text{YN}+\text{NY}+\text{NN}} = \frac{2+1}{0+1+2+0} = 1$$

$$d(\text{张明},\text{王刚}) = \frac{1+0}{1+0+1+1} = 0.33$$

$$d(\text{李思},\text{王刚}) = \frac{1+1}{1+1+1+0} = 0.67$$

### 4. 数值属性的相异性

- 欧几里得距离（Euclidean Distance）：又称直线距离。
  - 令$\boldsymbol{i}=(x_{i1},\ x_{i2},\ \cdots,\ x_{ip})$和$\boldsymbol{j}=(x_{j1},\ x_{j2},\ \cdots,\ x_{jp})$是两个被p个数值属性描述的对象。对象*i*和*j*之间的欧几里得距离为：

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2} \qquad (1\text{-}13)$$

- 曼哈顿距离（Manhattan Distance）：又称城市块距离。
  - 令$\boldsymbol{i}=(x_{i1},\ x_{i2},\ \cdots,\ x_{ip})$和$\boldsymbol{j}=(x_{j1},\ x_{j2},\ \cdots,\ x_{jp})$是两个被p个数值属性描述的对象。对象*i*和*j*之间的曼哈顿距离为：

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{ip}-x_{jp}| \qquad (1\text{-}14)$$

### 4. 数值属性的相异性

– 欧几里得距离和曼哈顿距离都满足如下数学性质：

①非负性：$d(i,j) \geq 0$：距离是一个非负的数值。

②同一性：$d(i,i) = 0$：对象到自身的距离为0。

③三角不等式：$d(i,j) \leq d(i,k) + d(k,j)$：从对象i到对象j的直接距离不会大于途经任何其他对象k的距离。

### 4. 数值属性的相异性

– 切比雪夫距离（Chebyshev Distance）：又称上确界距离，定义两个对象之间的上确界距离为其各坐标数值差的最大值。

$$d(i,j)=\lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if}-x_{jf}|^{h}\right)^{\frac{1}{h}}=\max_{f\to p}|x_{if}-x_{jf}| \qquad (1\text{-}16)$$

## 4. 数值属性的相异性

### – 闵可夫斯基距离（Minkowski Distance）

- 令$\boldsymbol{i} = (x_{i1}, x_{i2}, \cdots, x_{ip})$和$\boldsymbol{j} = (x_{j1}, x_{j2}, \cdots, x_{jp})$是两个被p个数值属性描述的对象。对象*i*和*j*之间的闵可夫斯基距离为：

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h} \qquad (1\text{-}15)$$

$h$=1即为曼哈顿距离.
$h$ =2即为欧几里德距离.
$h = \infty$即为切比雪夫距离.

### 4. 数值属性的相异性

例：数值属性的相异性计算

给定两个对象分别用元组(2，8，7，4)和(1，5，3，0)描述，计算这两个对象之间的欧几里得距离、曼哈顿距离、闵可夫斯基距离（h=4），以及切比雪夫距离。

- 欧几里得距离为：$d(i,j)=\sqrt{(2-1)^2+(8-5)^2+(7-3)^2+(4-0)^2}=\sqrt{42}=6.48$
- 曼哈顿距离为：$d(i,j)=|2-1|+|8-5|+|7-3|+|4-0|=1+3+4+4=12$
- 闵可夫斯基距离为：$d(i,j)=\sqrt[4]{|2-1|^4+|8-5|^4+|7-3|^4+|4-0|^4}=\sqrt[4]{594}\approx 4.94$
- 切比雪夫距离为：$d(i,j)=max\{|2-1|，|8-5|，|7-3|，|4-0|\}=max\{1，3，4，4\}=4$

### 5. 序数属性的邻近性度量

- 序数属性可以通过把数值属性的值域划分成有限个类别，对数值属性离散化得到。
- 相异性：

假设f是用于描述n个对象的序数属性，关于f的相异性计算步骤如下：

①第i个对象的f值为$x_{if}$，属性f有$M_f$个有序的状态，表示排位1，…，$M_f$。用对应的排位$r_{if}\in\{1，…，M_f\}$取代$x_{if}$。

②将对象的每个序数属性的值域映射到［0.0，1.0］上，以便每个属性都有相同的权重。通过用$z_{if}$代替第i个对象的$r_{if}$来实现数据规格化，其中

$$z_{if}=\frac{r_{if}-1}{M_f-1} \qquad (1\text{-}17)$$

③相异性可以用任意一种数值属性的距离度量计算，使用$z_{if}$作为第i个对象的f值。

## 5. 序数属性的邻近性度量

例：序数属性的相异性计算

表1-10 序数属性数据

| 对象标识符 | Test |
|---|---|
| 1 | excellent |
| 2 | good |
| 3 | ordinary |
| 4 | excellent |

Test有三个状态：ordinary、good和excellent，则$M_f=3$。

①把Test的每个值替换为它的排位，则4个对象的排位值分别为3、2、1、3。

②通过将排位1映射为0.0，排位2映射为0.5，排位3映射为1.0来实现对排位的规格化。

③使用欧几里得距离得到相异性矩阵：

$$\begin{bmatrix} 0 & & & \\ 0.5 & 0 & & \\ 1.0 & 0.5 & 0 & \\ 0 & 0.5 & 1.0 & 0 \end{bmatrix}$$

## 6. 余弦相似性

- 余弦相似性（又称余弦相似度，Cosine Similarity）：是基于向量的，它利用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。

  - 令向量$\boldsymbol{x}=(x_1,\ x_2,\ \cdots,\ x_p)$，向量$\boldsymbol{y}=(y_1,\ y_2,\ \cdots,\ y_p)$，两个向量的余弦相似性定义为：

$$sim(\boldsymbol{x},\boldsymbol{y})=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}=\frac{x_1y_1+x_2y_2+\cdots+x_py_p}{\sqrt{{x_1}^2+{x_2}^2+\cdots+{x_p}^2}\sqrt{{y_1}^2+{y_2}^2+\cdots+{y_p}^2}}$$

$$=\frac{\sum_{i=1}^{p}x_iy_i}{\sqrt{\sum_{i=1}^{p}{x_i}^2}\sqrt{\sum_{i=1}^{p}{y_i}^2}}\qquad(1\text{-}19)$$

  - 其中，$\|\boldsymbol{x}\|$是向量$\boldsymbol{x}=(x_1,\ x_2,\ \cdots,\ x_p)$的欧几里得范数，定义为$\sqrt{{x_1}^2+{x_2}^2+\cdots+{x_p}^2}$。同理，$\|\boldsymbol{y}\|$是向量$\boldsymbol{y}=(y_1,\ y_2,\ \cdots,\ y_p)$的欧几里得范数，定义为$\sqrt{{y_1}^2+{y_2}^2+\cdots+{y_p}^2}$。

## 6. 余弦相似性

例：余弦相似度的计算

给定两个向量$\boldsymbol{x} = (1，2，5，4)$和$\boldsymbol{y} = (2，3，5，1)$，计算两个向量的余弦相似度。

余弦相似度为：$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{1*2+2*3+5*5+4*1}{\sqrt{1+4+25+16}*\sqrt{4+9+25+1}} = \frac{37}{\sqrt{46}*\sqrt{39}} \approx 0.87$

# 感谢指导！

THANKS FOR YOUR ATTENTION