西安交通大學
XI'AN JIAOTONG UNIVERSITY

数据挖掘
第二章：数据预处理

刘均

陕西省天地网技术重点实验室

西安交通大学计算机学院

0 认识数据

1 为什么要与处理数据

2 数据清理

3 数据集成和变换

4 数据归约

基本要求：了解数据质量问题及其对挖掘的影响，掌握数据清理、集成和变换、归约等方法

本章内容

2.0 认识数据

2.1 为什么要预处理数据

2.2 数据清理

2.3 数据集成和变换

2.4 数据归约

2.0 认识数据

- **洞察数据有助于数据预处理与挖掘**
  - 数据由什么类型的属性或字段组成
  - 属性具有何种类型的属性值
  - 属性是离散的还是连续的
  - 数据分布特性
  - 数据可视化

2.0 认识数据 – 数据对象与属性类型

- **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
  - 顾客、商品、患者
  - 又称**样本、实例、数据点、元组**等
- **属性：**表示数据对象的一个特征
  - 维、特征、变量
  - 一个给定对象的一组属性称作**属性向量（特征向量）**
  - **属性的类型**由该属性可能具有的值的集合决定

2.0 认识数据 – 数据对象与属性类型

- **枚举类型（nominal attribute）：分类类型**
  - 属性值域是一个由符号、事物构成的有限集合
  - 头发颜色、婚姻状态、职业
  - 不具备有意义的序、不是定量的
  - 可用众数(mode)度量中心趋势
- **二元属性（binary attribute）：布尔属性**
  - 只有两个类别与状态：0与1， true与false
  - 对称的：两个状态分布或重要性相同。性别
  - 非对称的：两个状态分布或重要性不是相同的。HIV检验。

2.0 认识数据 – 数据对象与属性类型

- **序数类型（ordinal attribute）**
  - 属性值之间存在有意义的序，相继值之间差是定性的
  - 大中小、职位、军衔
  - 可通过把数值量的值域划分为有限个有序列性得到序数类型
  - 可用**众数**与**中位数**表示中心趋势

2.0 认识数据 – 数据对象与属性类型

- 数值属性（numeric attribute）
  - 可用整数或实数度量
  - 区间标度（interval-scaled）属性：用相同的单位尺度度量。
    - 可用众数、中位数、均值表示
  - 比例标度（ratio-scaled）属性：可用倍数表示。
    - 可用众数、中位数、均值表示

2.0 认识数据 – 数据对象与属性类型

- **离散属性：**具有有限个或无限可数个值
- **连续属性：**如果属性不是离散的，则它是连续的，用实数表示

2.0 认识数据 – 数据基本统计描述

- **动机：为了更好的理解数据**
  - 获得数据的总体印象
  - 识别数据的典型特征
  - 凸显噪声或离群点
- **度量数据的中心趋势**
  - 均值、中位数、众数（模）
- **度量数据的离散程度**
  - 四分位数、四分位数极差、方差等

2.0 认识数据 – 数据基本统计描述

- 算术平均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

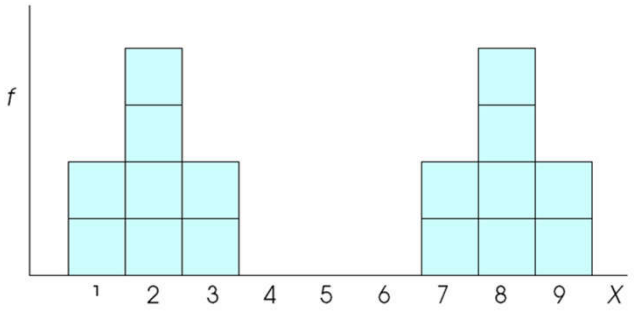
- 加权算术平均

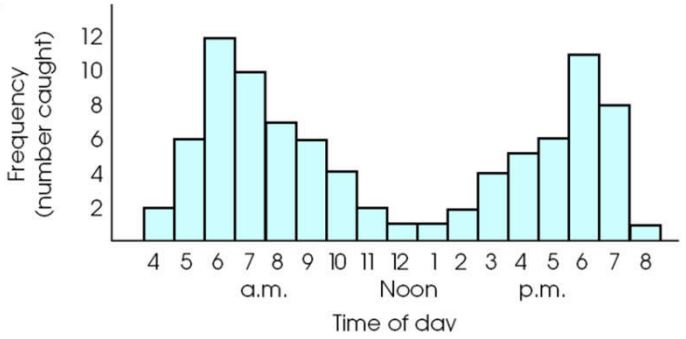
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- 截断均值（trimmed mean）：去掉高、低极端值得到的均值
  - e.g. 计算平均工资时，可以截掉上下各2%的值后计算均值，以抵消少数极端值的影响
- 中位数：有序集的中间值或者中间两个值平均

2.0 认识数据 – 数据基本统计描述

- **众数（Mode，模）：集合中出现频率最高的值**
  - 单峰的（unimodal，也叫单模态）、双峰的（bimodal）、三峰的（trimodal）；多峰的（multimodal）



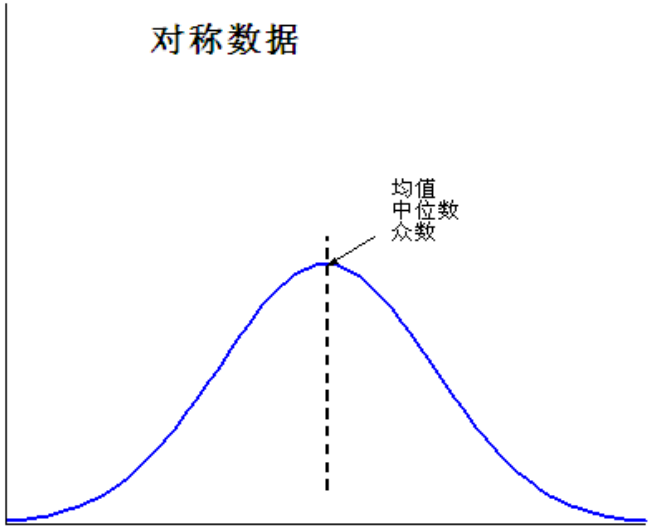


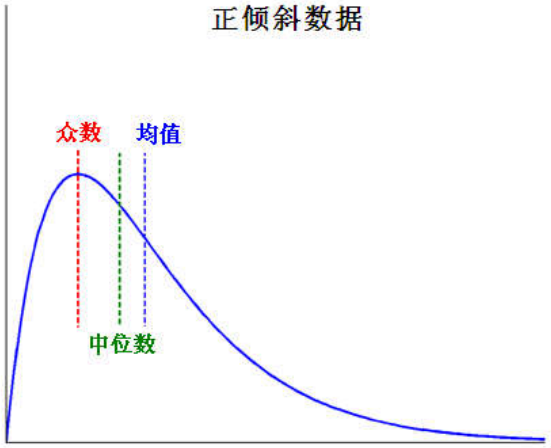
  - **对于适度倾斜（非对称的）的单峰频率曲线，可以使用以下经验公式计算众数**

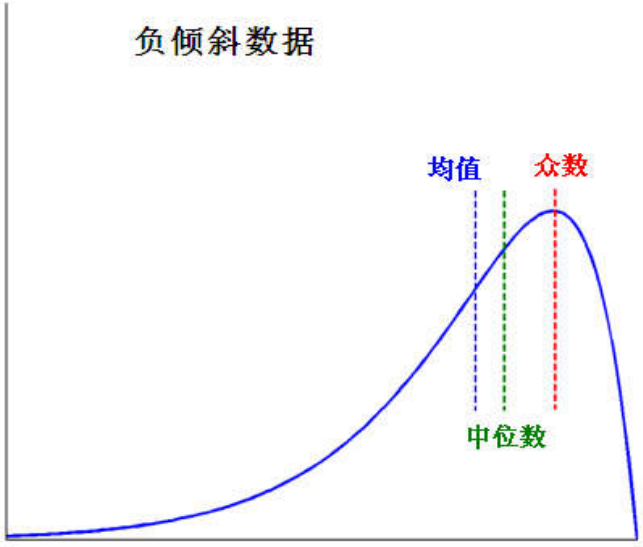
$$mean - mode = 3 \times (mean - median)$$

2.0 认识数据 – 数据基本统计描述

对称与正倾斜、负倾斜数据的中位数、均值和众数

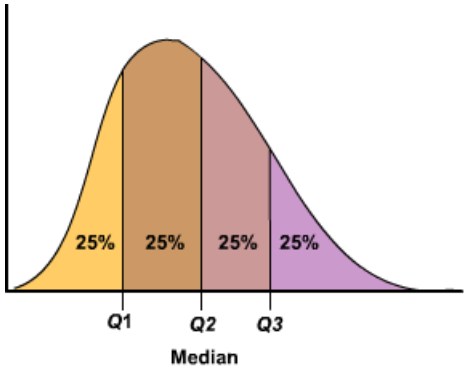






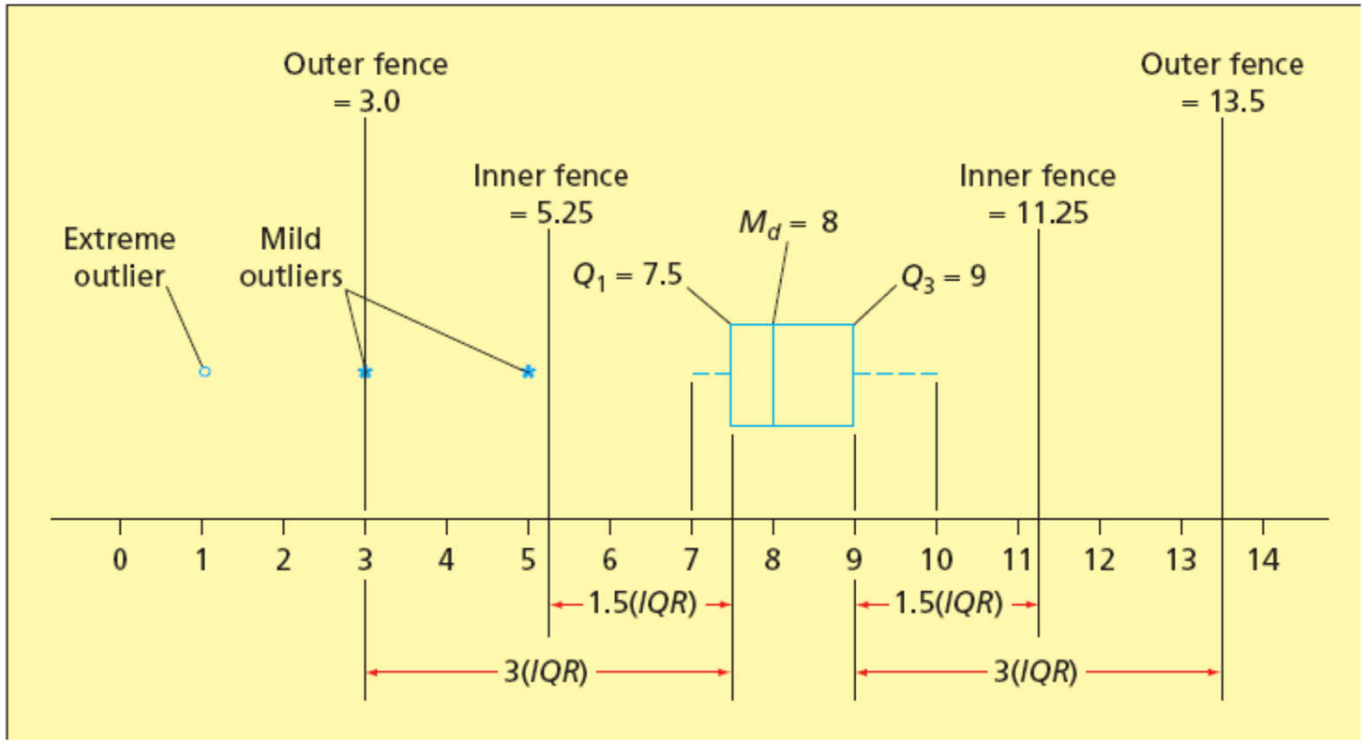
2.0 认识数据 – 数据基本统计描述

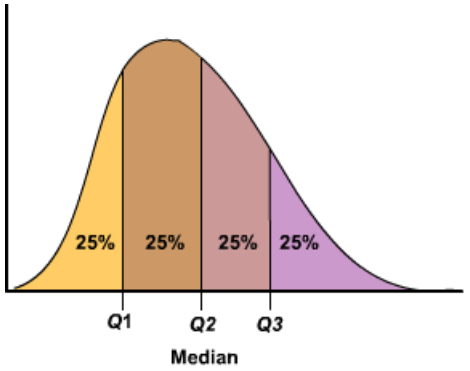
- **评估数值数据散布或发散的度量：极差、五数概括（基于四分位数）、中间四分位数极差和标准差**
  - 极差（range）：数据集的最大值和最小值之差
  - 百分位数(percentile)：第k个百分位数是具有如下性质的值x：k%的数据项位于或低于x
    - 中位数就是第50个百分位数
  - 四分位数：Q_1 (25th percentile), Q_3 (75th percentile)
  - 中间四分位数极差(IQR)： $IQR = Q_3 - Q_1$
  - 孤立点：通常挑出落在至少高于第三个四分位数或低于第一个四分位数 1.5 ×IQR处的值



2.0 认识数据 – 数据基本统计描述

- **评估数值数据散布或发散的度量：极差、五数概括（基于四分位数）、中间四分位数极差和标准差**





2.0 认识数据 – 数据基本统计描述

State	Employment Ratio (%)	State	Employment Ratio (%)	State	Employment Ratio (%)
Alabama	60.3	Kentucky	61.5	North Dakota	68.1
Alaska	68.8	Louisiana	59.4	Ohio	64.0
Arizona	63.3	Maine	65.1	Oklahoma	62.9
Arkansas	60.0	Maryland	67.2	Oregon	64.3
California	62.8	Mass.	66.5	Pennsylvania	61.6
Colorado	71.4	Michigan	66.0	Rhode Island	64.4
Connecticut	65.4	Minnesota	73.0	South Carolina	62.7
Delaware	64.7	Mississippi	58.0	South Dakota	71.1
Dist. Of Col.	63.4	Missouri	66.4	Tennessee	63.6
Florida	60.1	Montana	65.6	Texas	65.6
Georgia	66.8	Nebraska	71.0	Utah	69.6
Hawaii	63.2	Nevada	66.0	Vermont	69.9
Idaho	66.1	New Hamp.	70.3	Virginia	65.6
Illinois	66.7	New Jersey	64.1	Washington	66.9
Indiana	66.2	New Mexico	58.5	West Virginia	52.7
Iowa	70.1	New York	59.7	Wisconsin	70.1
Kansas	70.0	North Carolina	65.1	Wyoming	67.8

2.0 认识数据 – 数据基本统计描述

1. 52.7	11. 62.7	21. 64.4	31. 66.1	41. 68.8
2. 58	12. 62.8	22. 64.7	32. 66.2	42. 69.6
3. 58.5	13. 62.9	23. 65.1	33. 66.4	43. 69.9
4. 59.4	14. 63.2	24. 65.1	34. 66.5	44. 70
5. 59.7	15. 63.3	25. 65.4	35. 66.7	45. 70.1
6. 60	16. 63.4	26. 65.6	36. 66.8	46. 70.1
7. 60.1	17. 63.6	27. 65.6	37. 66.9	47. 70.3
8. 60.3	18. 64	28. 65.6	38. 67.2	48. 71
9. 61.5	19. 64.1	29. 66	39. 67.8	49. 71.1
10. 61.6	20. 64.3	30. 66	40. 68.1	50. 71.4
				51. 73

2.0 认识数据 – 数据基本统计描述

- 例子
  - 60th Percentile
    I = (60/100)*51 =30.6
    30.6 不是整数，选择整数31，故数值为 66.1
  - 33th Percentile
    I =(33/100)*51=16.83
    16.83 不是整数，选择整数17，故数值为 63.6
  - Q1:13th—62.9 ;
  - Q3: 38th—67.2
  - Q3−Q1=4.3
  - (62.9−1.5*4.3, 67.2+1.5*4.3)=(56.45,73.65)
  - The OUTLIER is 52.7

Lower Fence = Q_1 − 1.5(IQR)

Upper Fence = Q_3 + 1.5(IQR)

2.0 认识数据 – 数据基本统计描述

- 五数概括: min, Q_1, Median, Q_3, max
- 盒图：数据分布的一种直观表示
- 方差和标准差
  - 方差s^2：n个观测之x_1,x_2…x_n的方差是

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

  - 标准差s是方差s^2的平方根
    - 标准差s是关于平均值的离散的度量，因此仅当选平均值做中心度量时使用
    - 所有观测值相同则 s = 0，否则 s>0

2.0 认识数据 – 数据基本统计描述

- **例子：5个数据 30.8, 31.7, 30.1, 31.6, 32.1**
  - **平均值：31.26**

$$s^2 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})^2}{5-1}$$

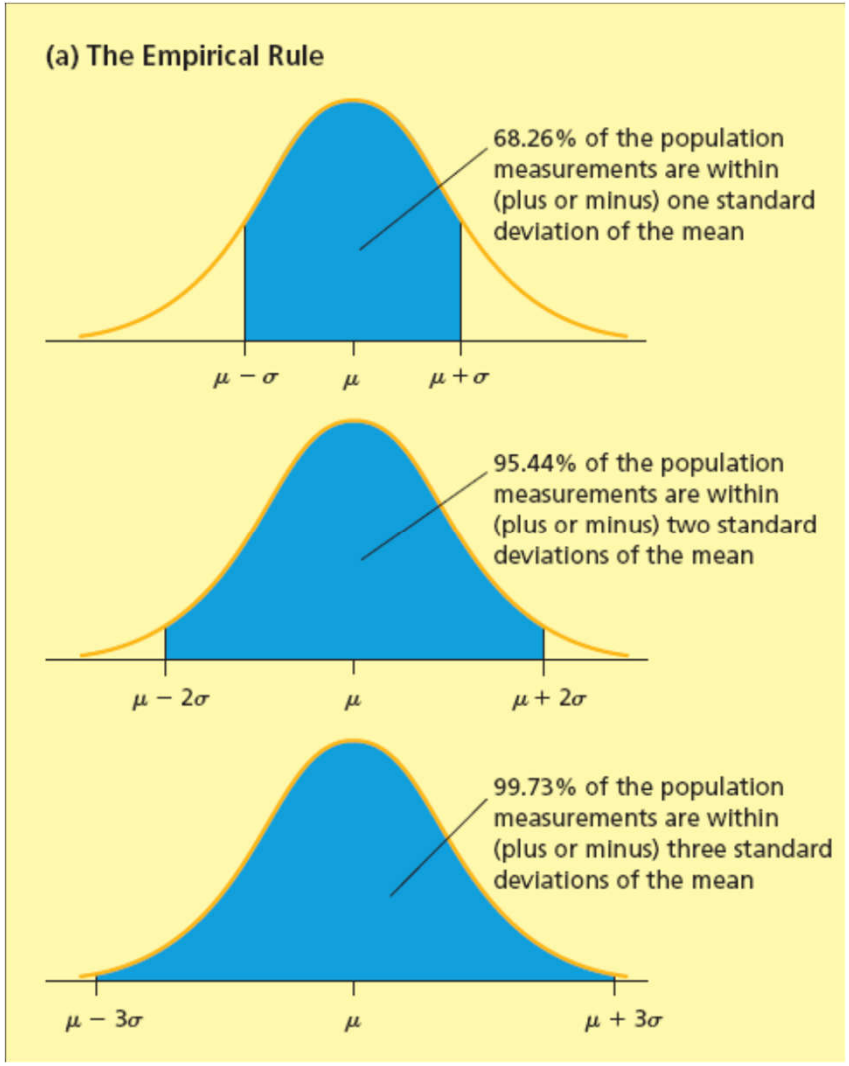
$$= \frac{(30.8-31.26)^2 + (31.7-31.26)^2 + (30.1-31.26)^2 + (31.6-31.26)^2 + (32.1-31.26)^2}{4}$$

$$= \frac{2.572}{4} = 0.643$$

$$s = \sqrt{s^2} = \sqrt{0.643} = 0.8019$$

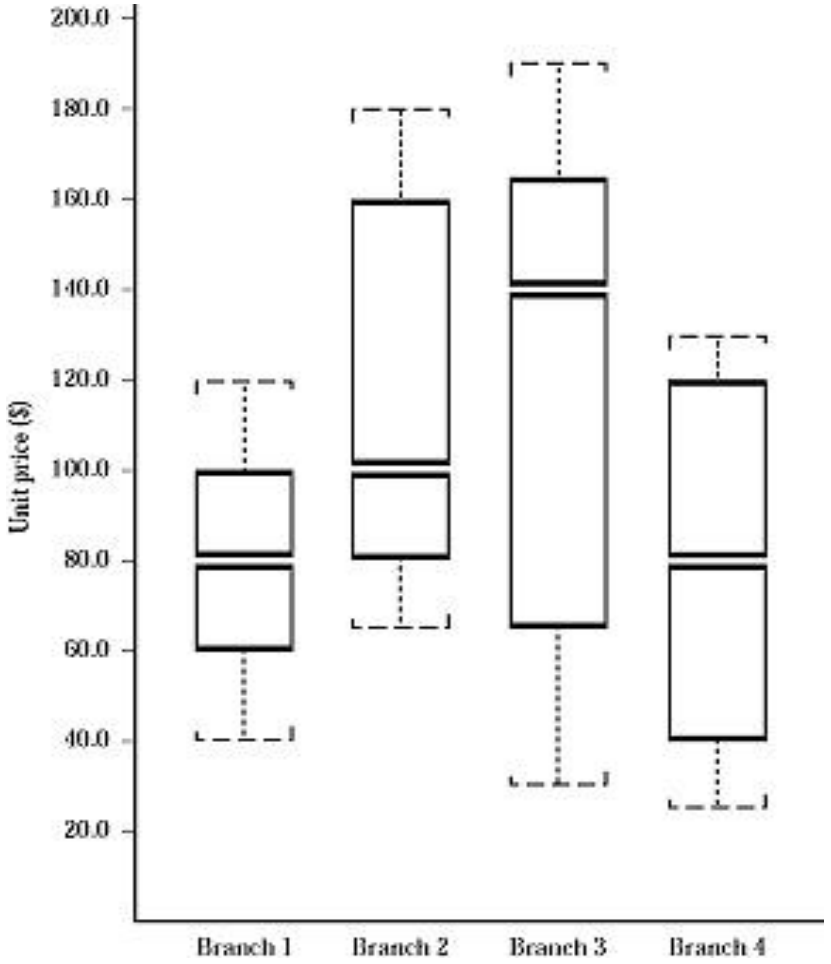
2.0 认识数据 – 数据基本统计描述

- ✓ 68.26% 的数据分布在 [μ ± s] = [31.6 ± 0.8] = [30.8, 32.4]
- ✓ 95.44%的数据分布在 [μ ± 2s] = [31.6 ± 1.6] = [30.0, 33.2]
- ✓ 99.73%的数据分布在 [μ ± 3s] = [31.6 ± 2.4] = [29.2, 34.0]
- ✓ 至少 $(1 - 1/k^2)$的数据分布在 $[\mu - ks, \mu + ks]$



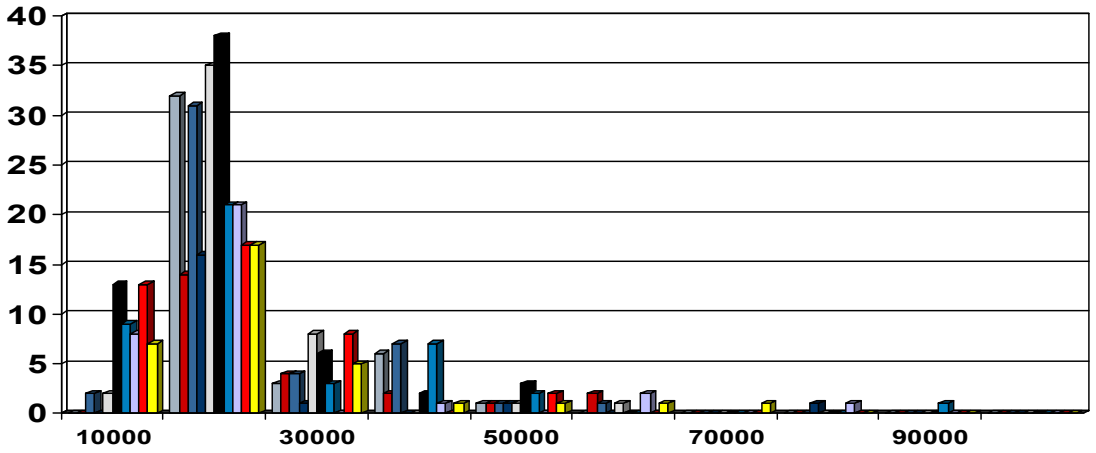
2.0 认识数据 – 数据基本统计描述

- **盒图：**数据分布的一种直观表示：
  - 端点在四分位数上，使得盒图的长度是IQR
  - 中位数M用盒内的线标记
  - 胡须延伸到最大最小观测值
- 该盒图为在给定时间段在AllElectronics的4个分店销售的商品单价的盒图
  - 分店1：中位数$80，Q1: $60，Q3: $100



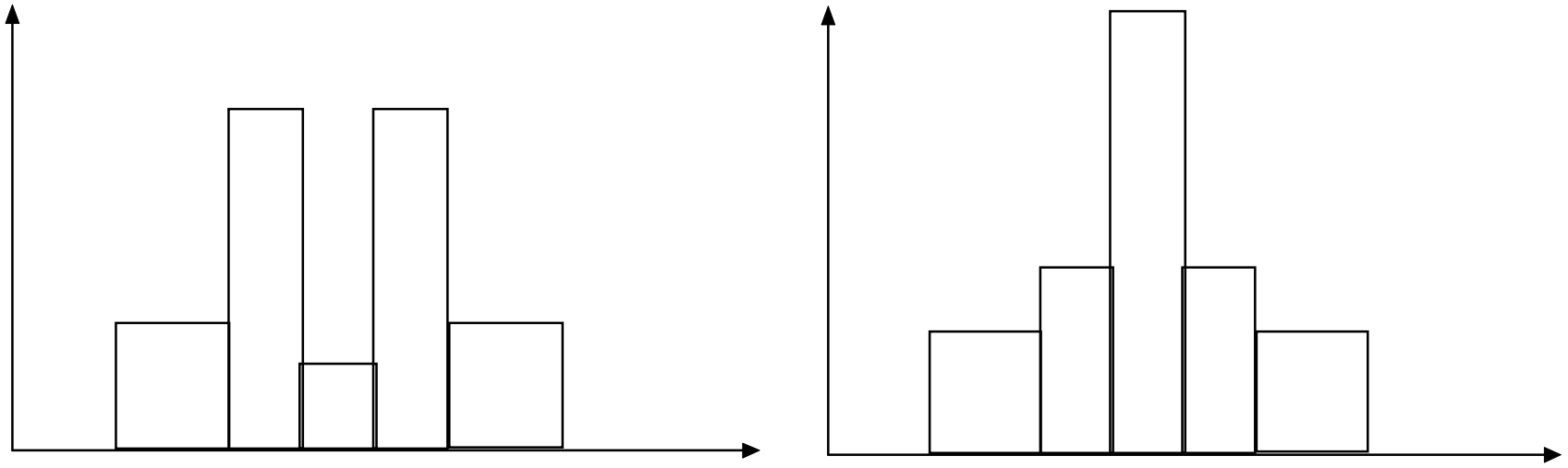
2.0 认识数据 – 数据基本统计描述

- **常用的显示数据汇总和分布的方法**
  - 直方图、分位数图、q-q图、散点图和局部回归曲线
- **直方图：一种单变量图形表示方法**
  - 将数据分布划分成不相交的子集或桶，通常每个桶宽度一致并用一个矩形表示，其高度表示桶中数据在给定数据中出现的计数或频率



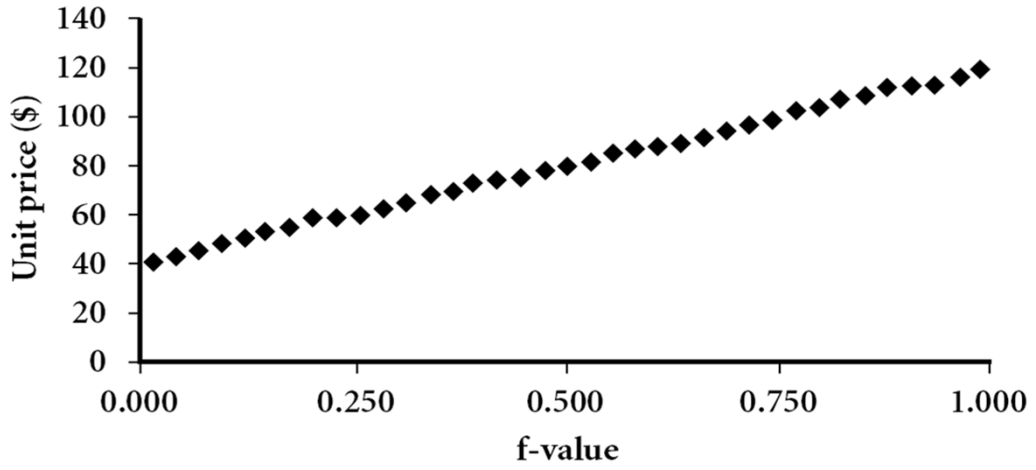
2.0 认识数据 – 数据基本统计描述

- **直方图能够比盒图展现更？的信息**
  - 两个直方图具有相同的min, Q1, median, Q3, max
  - 但是它们具有不同数据分布



2.0 认识数据 – 数据基本统计描述

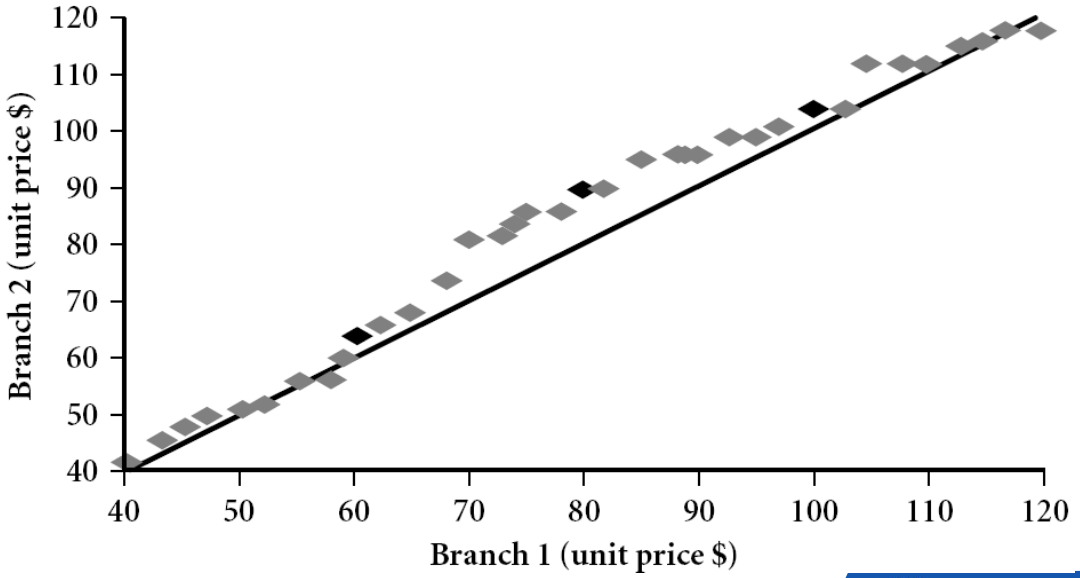
- **分位数图：**一种利用分位数信息观察单变量数据分布的简单有效方法
- 显示所有数据，允许用户评估总的情况和不寻常情况的出现
  - 设x_i是递增排序的数据，则每个x_i都有相对应的f_i，指出大约有$100\ f_i\ \%$的数据小于等于x_i



2.0 认识数据 – 数据基本统计描述

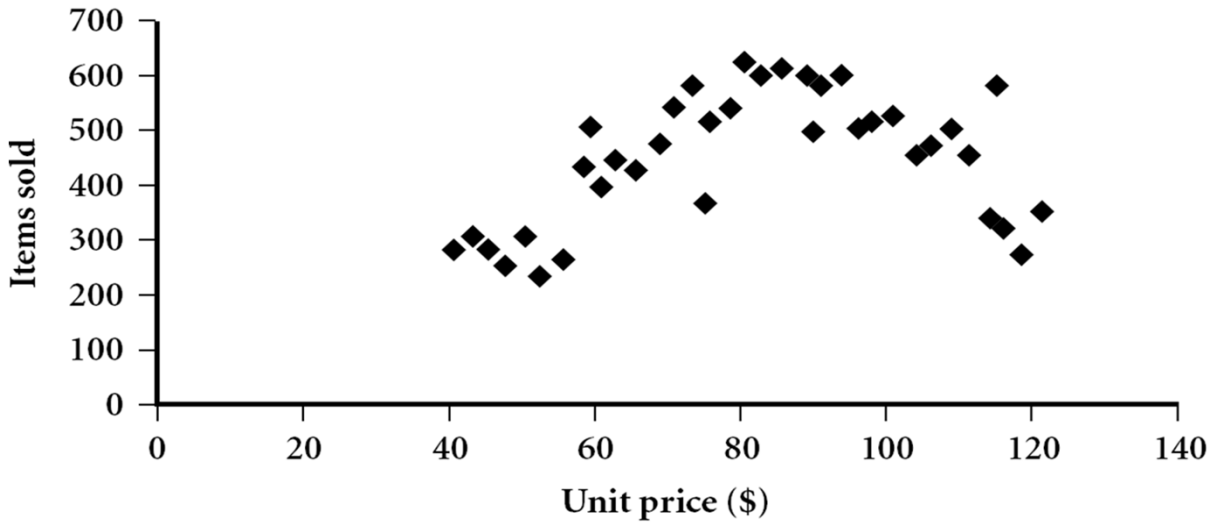
- **分位数－分位数图（Q-Q 图）：对着另一个单变量的分位数，绘制一个单变量分布的分位数**
- **允许用户观察是不是有从一个分布到另外一个分布的迁移**

Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



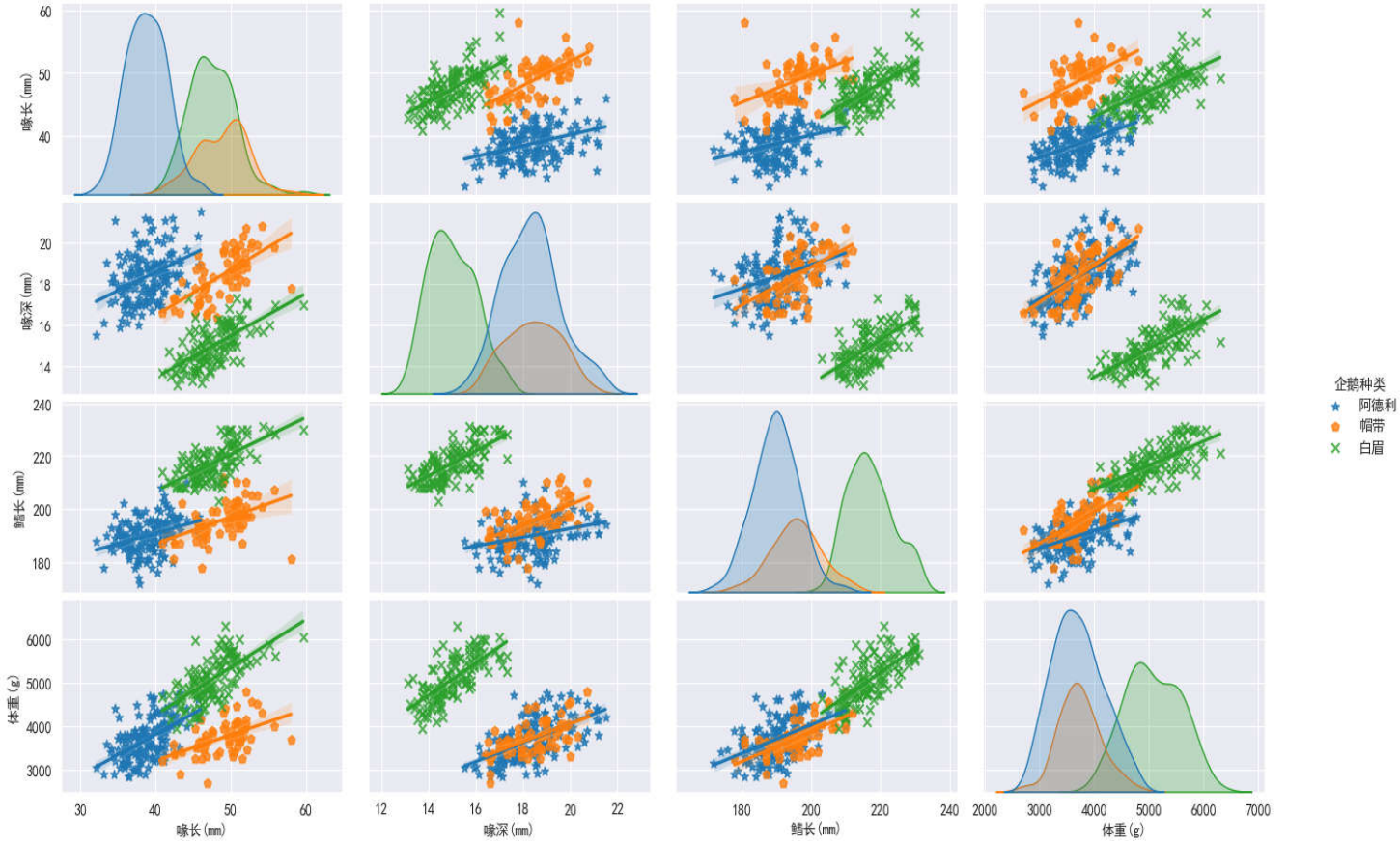
2.0 认识数据 – 数据基本统计描述

- **散点图：**确定两个量化的变量之间看上去是否有联系、模式或者趋势的最有效的图形方法之一
- 散点图中的每个值都被视作代数坐标对，作为一个点画在平面上
- 易于观察双变量数据在平面上的分布



2.0 认识数据 – 数据基本统计描述

散点图矩阵

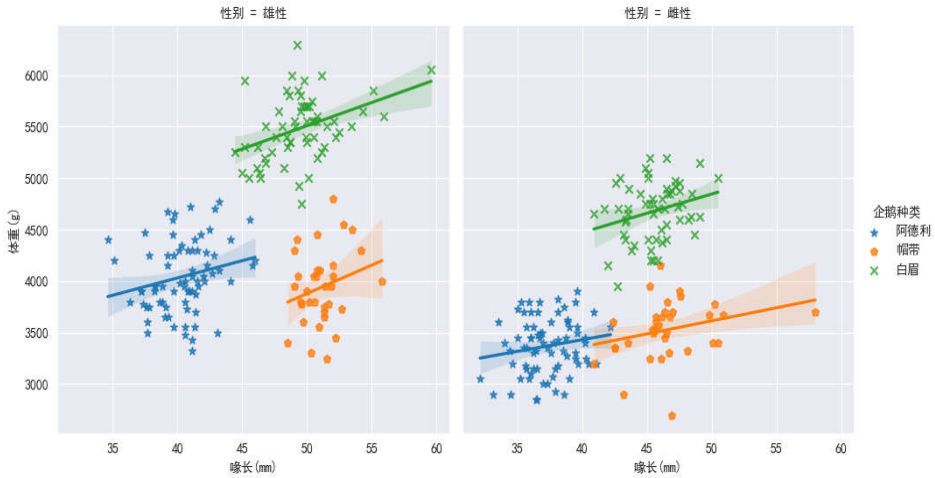


Palmer企鹅数据集包含3种（阿德利、白眉、帽带）共344只企鹅的数据

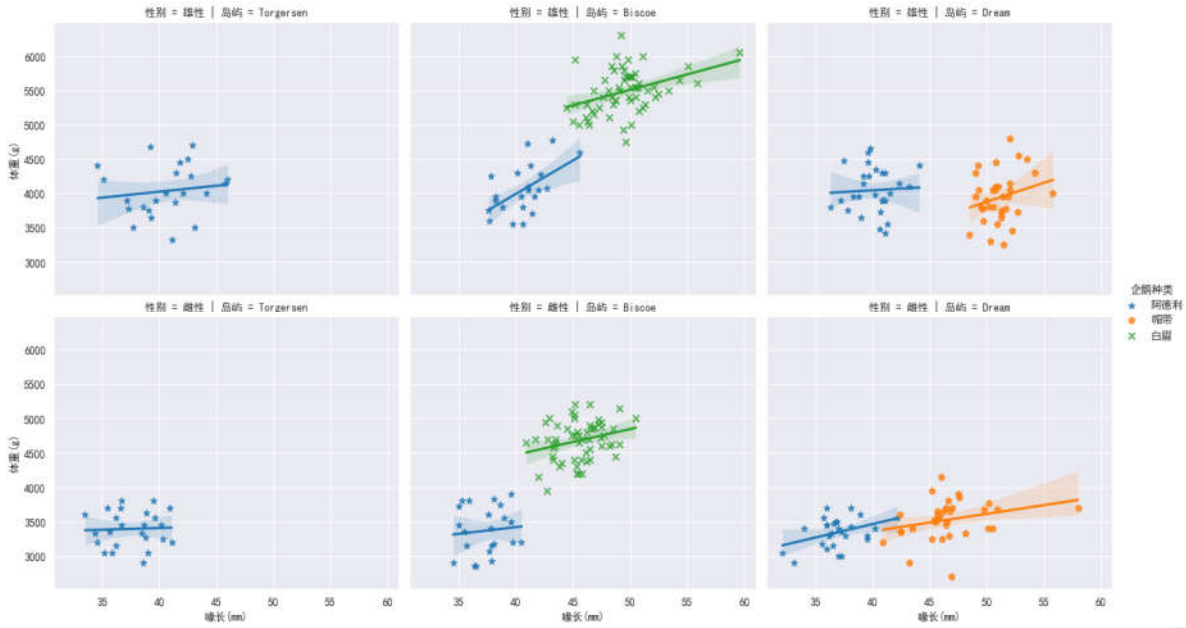
2.0 认识数据 – 数据基本统计描述

条件散点图

一个条件变量
（性别）

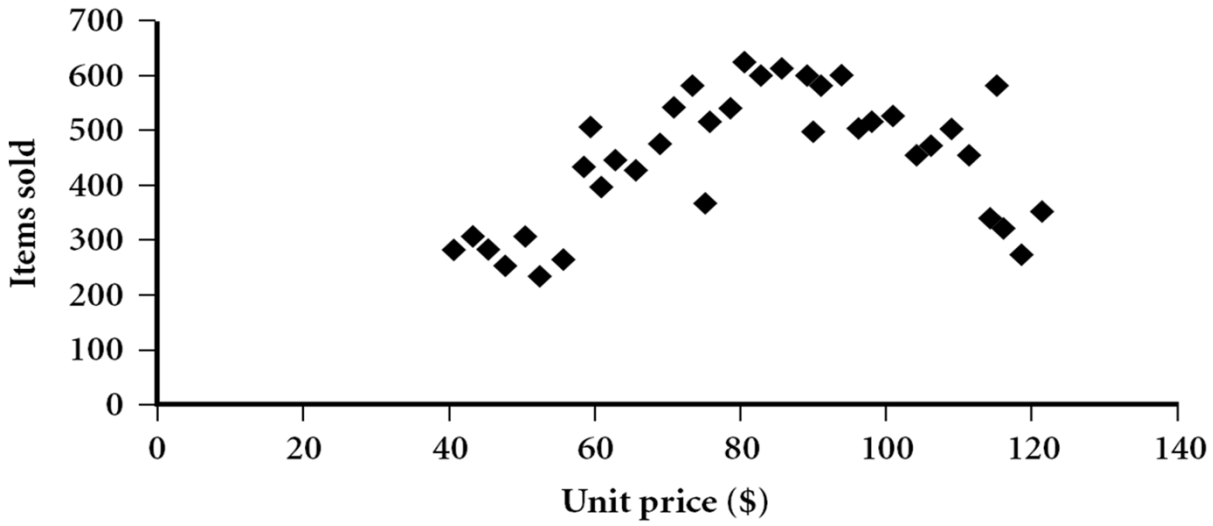


两个条件变量
（性别与岛屿）



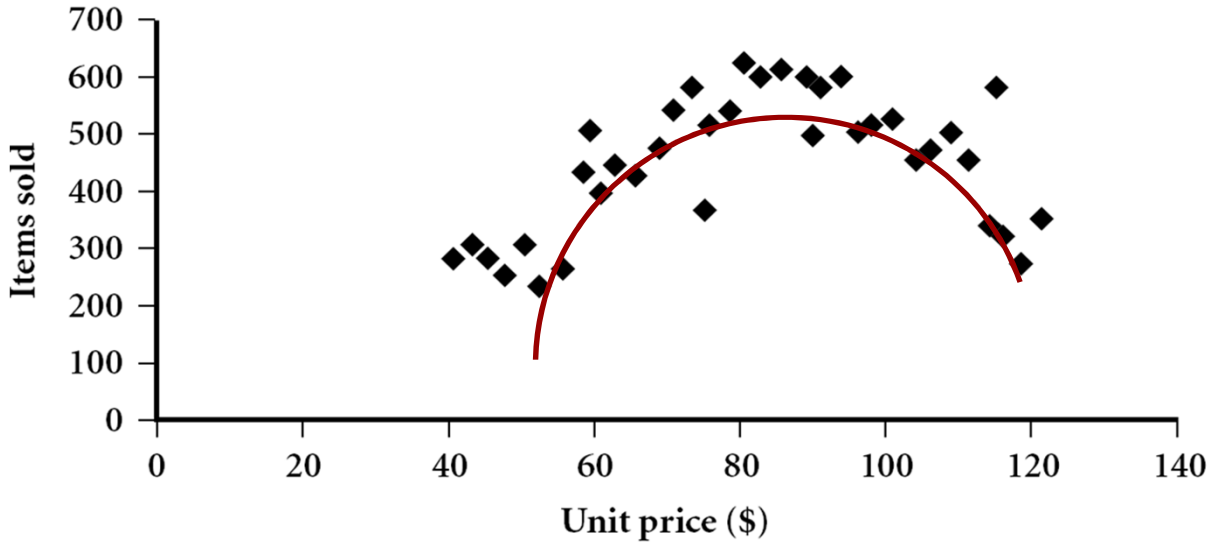
2.0 认识数据 – 数据基本统计描述

- **散点图：**确定两个量化的变量之间看上去是否有联系、模式或者趋势的最有效的图形方法之一
- 散点图中的每个值都被视作代数坐标对，作为一个点画在平面上
- 易于观察双变量数据在平面上的分布



2.0 认识数据 – 数据基本统计描述

- loess曲线为散点图添加一条平滑的曲线，以便更好的观察两个变量间的依赖模式
- Loess (local regression)意指“局部回归”，为了拟合loess曲线，需要两个参数：平滑参数α ，被回归拟合的多项式的阶λ

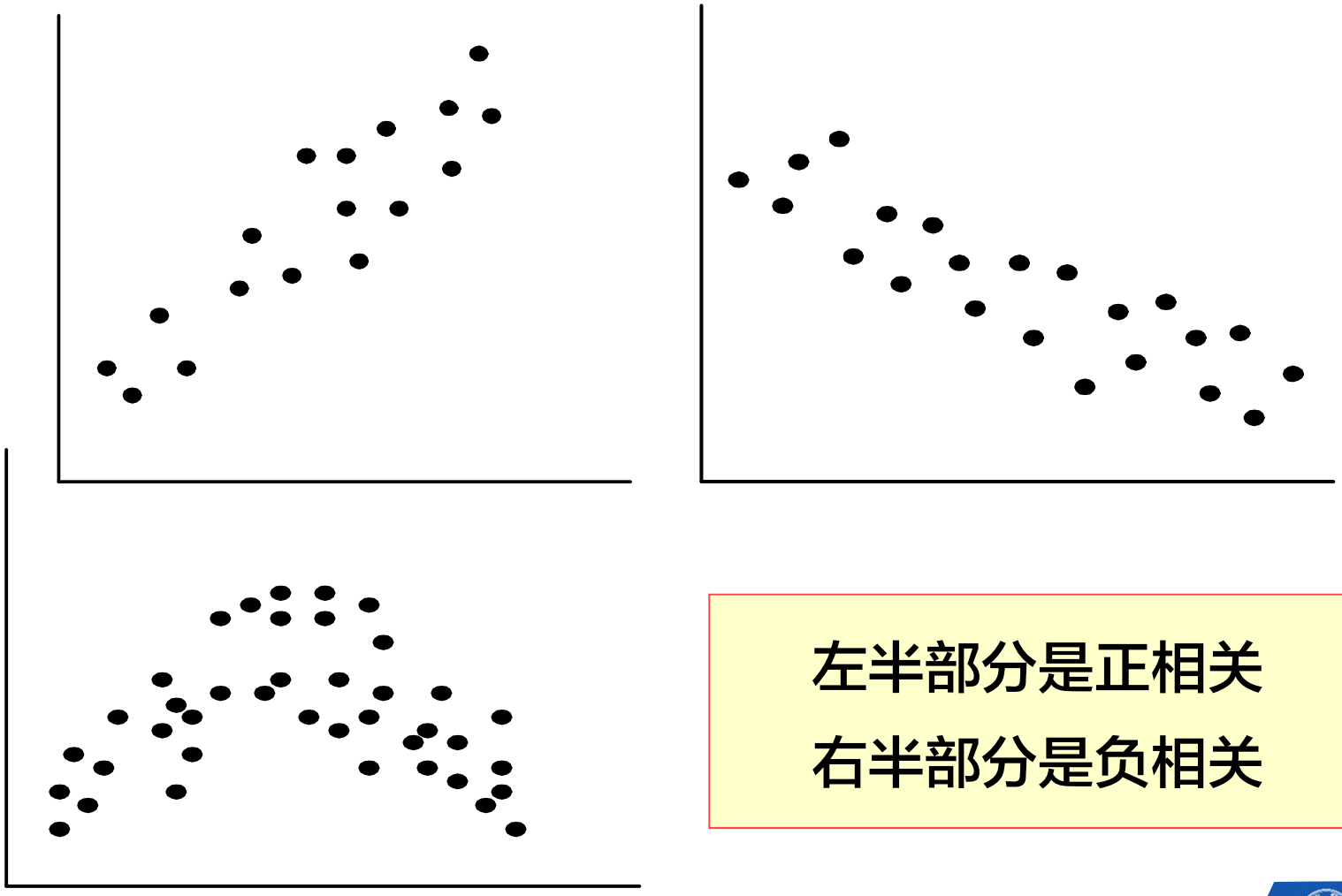


2.0 认识数据 – 数据基本统计描述

- 皮尔逊相关系数（Pearson correlation coefficient）度量两个变量*X*和 *Y*之间的线性相关程度，介于-1与1之间。

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

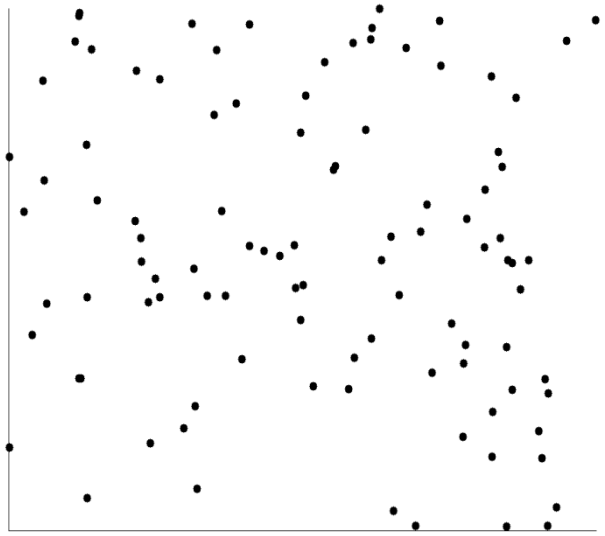
2.0 认识数据 – 数据基本统计描述

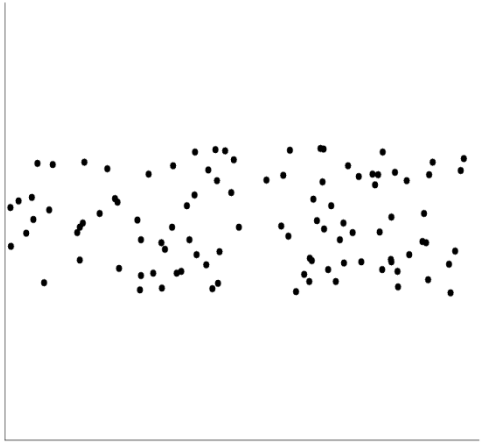


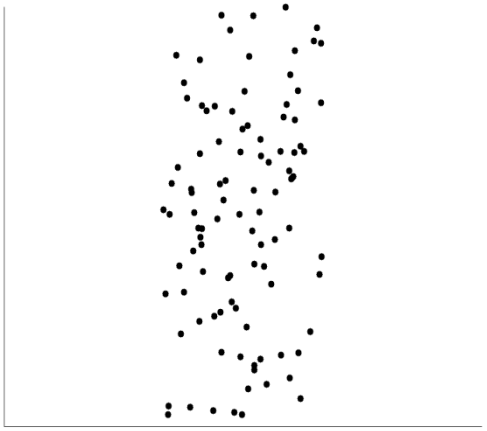
左半部分是正相关

右半部分是负相关

2.0 认识数据 – 数据基本统计描述



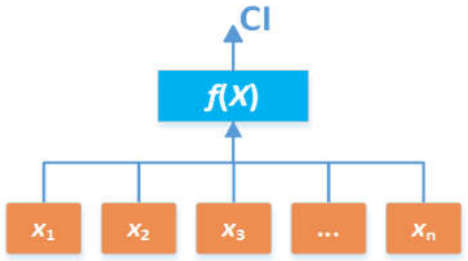




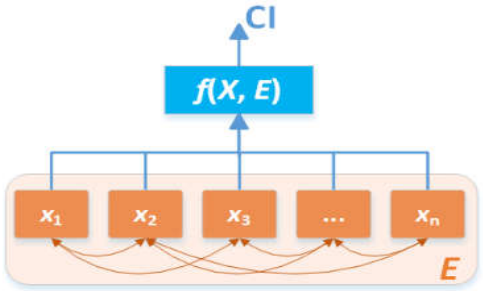
不相关数据

统计量与群智

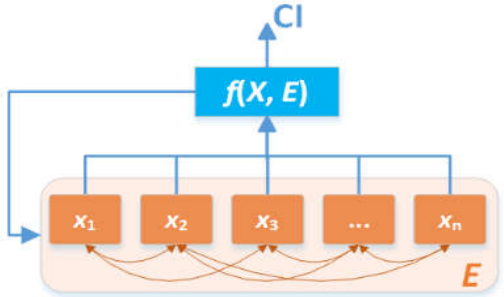
群体智能：群体在宏观层面涌现出的超过个体的智能水平。



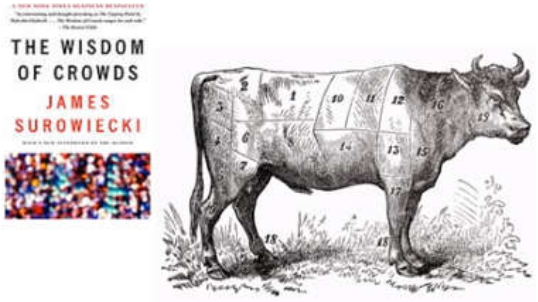
范式1：无交互、无反馈

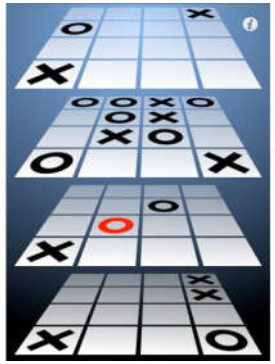


范式2：有交互、无反馈

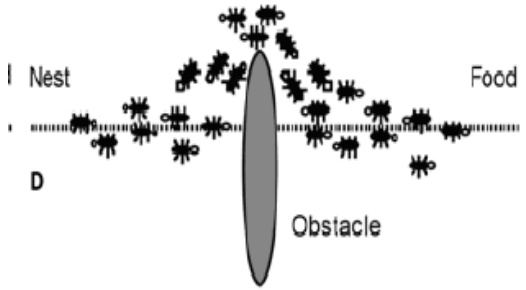


范式3：有交互、有反馈





DHT定理的合作证明



统计量与群智

- 设个个体的估值为 $x_1, x_2, .., x_i, .., x_n$ ，真实值为 x_T ，群体误差（Collective Error，CE）则是该估值的均值与真实值的平方离差（Squared Deviation），

  即 $(x_T - \bar{x})^2$ ， $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ；

- 群体多样性（Group Diversity，GD）是估值的方差，

  即 $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ；

- 平均个体误差（Average Individual Error，AIE）是指个体估值与真实值平方离差的均值，$\frac{1}{n}\sum_{i=1}^{n}(x_i - x_T)^2$ 。

  多样性预测定理：$CE = AIE - GD$

  - 当多样性较大时，有助于降低群体误差，促进群智的产生

统计量与群智

- 中位数在群智中起作用的一个重要原因：无论真实值是多少，群体估测值的中位数比至少50%的个体的估测值更接近真实值。

- 孔多塞陪审团定理（Condorcet's Jury Theorem）：如果群体中的个体独立进行决策，并且正确决策的概率大于0.5，那么群体按多数票策略正确决策的概率随着群体规模的增大而增大。

本章内容

2.0 认识数据

2.1 为什么要预处理数据

2.2 数据清理

2.3 数据集成和变换

2.4 数据归约

统计量与群智

- 现实世界的数据是“脏的”
  - 不完整(incomplete)
    - 缺少数据值；缺乏某些重要属性；仅包含汇总数据
  - 有噪声(noisy)
    - 包含错误或者孤立点(outliers)
  - 数据不一致(inconsistent)
    - e.g., 在编码或者命名上存在差异
    - e.g., Age=“42” Birthday=“03/07/1997”

GIGO (Garbage in, garbage out)原理：No quality data, no quality mining results!

2.1 为什么要预处理数据

- **广为认可的数据质量多维度量**
  - 精确度
  - 完整度
  - 一致性
  - 时效性
  - 可信度
  - 可增值性
  - 可解释性
  - 可访问性

2.1 为什么要预处理数据

■ 数据处理的主要任务

- **数据清理**
  - 填写空缺的值，平滑噪声数据，识别、删除孤立点，解决不一致性
- **数据集成**
  - 集成多个数据库、数据立方体或文件
- **数据变换**
  - 规范化和聚集
- **数据归约**
  - 得到数据集的压缩表示，但可得到相同或相近的结果
- **数据离散化**
  - 通过概念分层和数据离散化来规约数据，对数值型数据特别重要

本章内容

2.0 认识数据

2.1 为什么要预处理数据

2.2 数据清理

2.3 数据集成和变换

2.4 数据归约

2.2 数据清洗

- 业界对数据清理的认识
  - “数据清理是数据仓库构建中最重要的问题”— DCI survey
- 数据清理任务
  - 填写空缺值
  - 识别离群点和平滑噪声数据
  - 纠正不一致的数据
  - 解决数据集成造成的冗余

2.2 数据清洗

■ 空缺值

- ➢ **数据并不总是完整的**
  - 例如：数据库表中，很多条记录的对应字段没有相应值，比如销售表中的顾客收入
- ➢ **引起空缺值的原因**
  - 设备异常
  - 与其他已有数据不一致而被删除
  - 因为误解而没有被输入的数据
  - 在输入时，有些数据应为得不到重视而没有被输入
- ➢ **空缺值要经过推断而补上**

2.2 数据清洗

■ 如何处理空缺值

- 忽略元组：当类标号缺少时通常这么做（假定挖掘任务设计分类或描述），当每个属性缺少值的百分比很大时，它的效果非常差。
- 人工填写空缺值：工作量大，可行性低
- 使用一个全局变量填充空缺值：比如使用unknown或－∞
- 使用属性的平均值填充空缺值
- 使用与给定元组属同一类的所有样本的平均值
- 使用最可能的值填充空缺值：使用像Bayesian公式或判定树这样的基于推断的方法

2.2 数据清洗

■ 噪声数据

- ➢ 噪声：一个测量变量中的随机错误或偏差
- ➢ 引起噪声的原因
  - 数据收集工具的问题
  - 数据输入错误
  - 数据传输错误
  - 命名规则的不一致

2.2 数据清洗

■ 噪声数据处理

- ➢ **分箱(binning):**
  - 首先排序数据，并将他们分到等深的箱中
  - 可以按箱的平均值平滑、按箱中值平滑、按箱的边界平滑
- ➢ **回归：** 通过让数据适应回归函数来平滑数据
- ➢ **聚类：** 检测并且去除孤立点
- ➢ **计算机和人工检查结合：** 计算机检测可疑数据，然后对它们进行人工判断

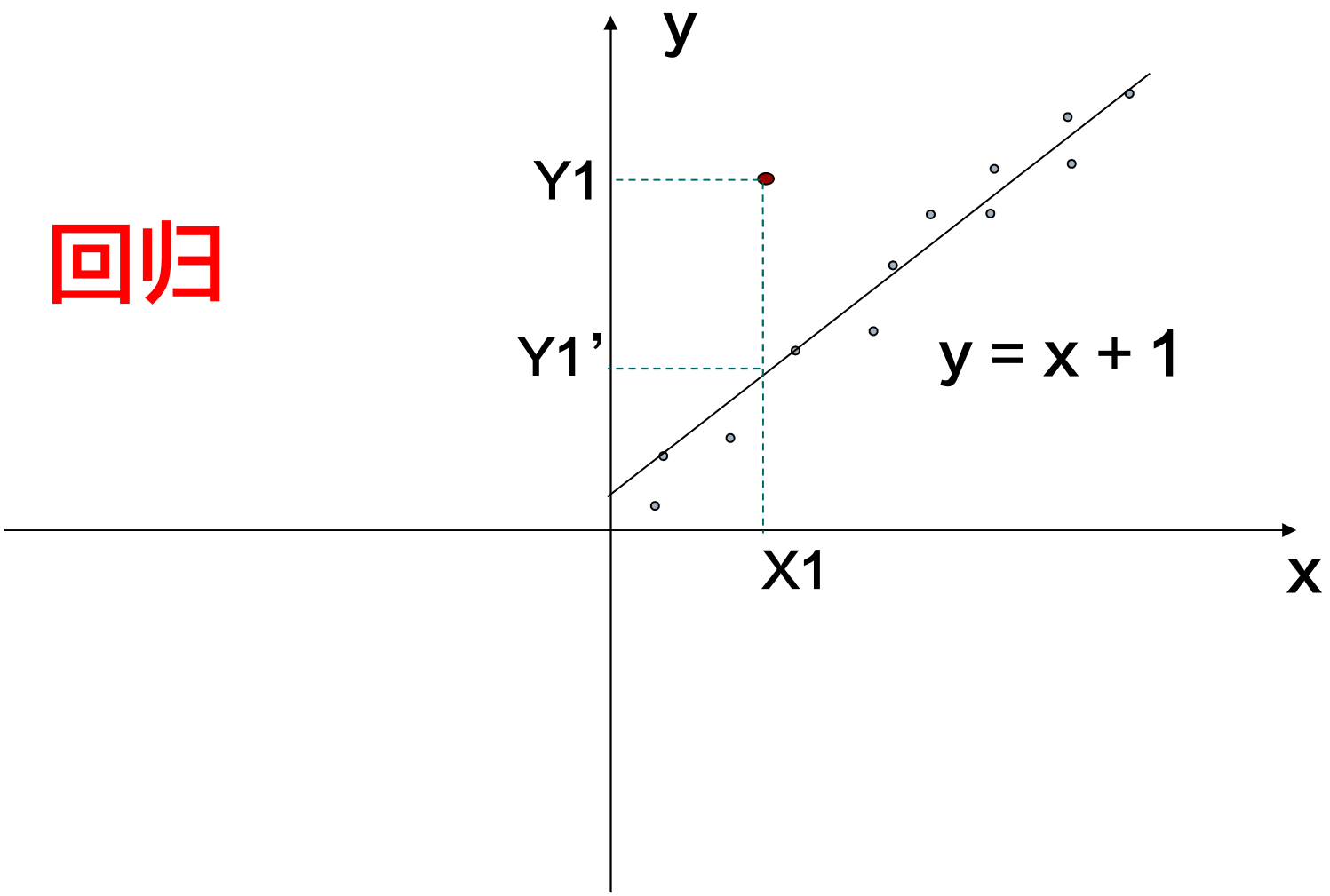
2.2 数据清洗

■ 分箱方法

- price的排序后数据：4，8，15，21，21，24，25，28，34
- 划分为（等深的）箱：
  - 箱1：4，8，15
  - 箱2：21，21，24
  - 箱3：25，28，34
- 用箱平均值平滑：
  - 箱1：9，9，9
  - 箱2：22，22，22
  - 箱3：29，29，29
- 用箱边界平滑：
  - 箱1：4，4，15
  - 箱2：21，21，24
  - 箱3：25，25，34

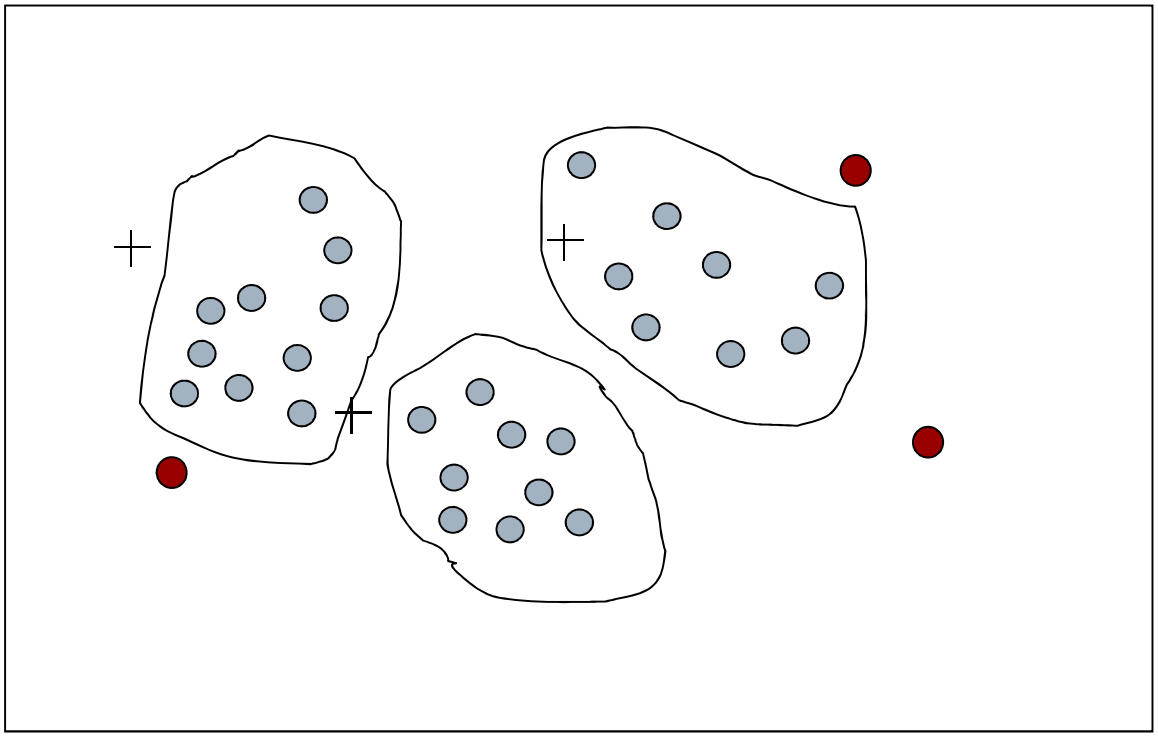
2.2 数据清洗

回归



2.2 数据清洗

聚类



- 通过聚类分析检测离群点，消除噪声：聚类将类似的值聚成簇。直观的，落在簇集合之外的值被视为离群点

本章内容

2.0 认识数据

2.1 为什么要预处理数据

2.2 数据清理

2.3 数据集成和变换

2.4 数据归约

2.3 数据集成和变换

- **数据集成：将多个数据源中的数据整合到一个一致的存储中**
- **模式集成：整合不同数据源中的元数据?**
  - e.g. A.cust_id = B.customer_no
- **实体识别问题：**
  - 匹配来自不同数据源的现实世界的实体
  - e.g. Bill Clinton = William Clinton
- **检测并解决数据值的冲突**
  - 对现实世界中的同一实体，来自不同数据源的属性值可能是不同的
  - 可能的原因：不同的数据表示，不同的度量等

2.3 数据集成和变换

■ 数据集成中的冗余数据处理

- 集成多个数据库时，经常会出现冗余数据

  对象识别：同一属性或对象在不同数据库中会有不同字段名

  可导出数据：一个属性可以由另外一个表导出，如“年薪”

- 有些冗余可以被相关分析检测到

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

皮尔逊系数

$-1 <= r_{xy} <= 1$，大于0时正相关，小于0时负相关。绝对值越接近于1，两要素关系越密切；越接近于0，越不密切。

- 将多个数据源中的数据集成起来，能减少或避免冗余与不一致性，从而可以提高挖掘的速度和质量。

2.3 数据集成和变换

■ 相关性分析用于属性选择

- ➢ 属性 A1与A2 是相互独立的吗?
  - 如果它们是相互依赖的，则可以任意去除一个
  - 如果A1 与类标签属性A2是独立的，可以从训练集中去除属性A1

ID	Outlook	Temperature	Humidity	Windy	Play
1	100	40	90	0	T
2	100	40	90	1	F
3	50	40	90	0	T
4	10	30	90	0	T
5	10	15	70	0	T
6	10	15	70	1	F
7	50	15	70	1	T
8	100	30	90	0	F
9	100	15	70	0	T
10	10	30	70	0	F
11	100	30	70	1	F
12	50	30	90	1	T
13	50	40	70	0	T
14	10	30	90	1	F

2.3 数据集成和变换

- 离散数据的相关性分析
  - X^2 (chi-square)测试(卡方测试)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

  - X^2的值越大，意味着两个变量相关的可能性越大
  - 期望值和观测值之间相差越大，值也将越大
  - 相关性不意味着因果关系
    - e.g. 我们发现一个地区的医院数和汽车盗窃数相关

2.3 数据集成和变换

■ 相关性分析用于属性选择

Outlook	Temperature
Sunny	High
Cloudy	Low
Sunny	High

Temperature → Outlook	High	Low	Outlook Subtotal
Sunny	2	0	2
Cloudy	0	1	1
Temperature Subtotal:	2	1	Total count in table =3

2.3 数据集成和变换

■ 零假设Ho： Outlook与Temperature无关。

Outlook	Temperature
Sunny	High
Cloudy	Low
Sunny	High

Temperature→ Outlook	High	Low	Outlook Subtotal
Sunny	3*2/3*2/3=1.33	3*2/3*1/3=0.67	2
Cloudy	3*1/3*2/3=0.67	3*1/3*1/3=0.33	1
Temperature Subtotal:	2	1	Total count in table =3

2.3 数据集成和变换

■ 零假设Ho： Outlook与Temperature无关。

The chi-squared formula is:

$$\text{Chi-squared } (X^2) = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \cdots\cdots + \frac{(o_n - e_n)^2}{e_n}$$

$$chi - squared(x^2) = \frac{(2-1.33)^2}{1.33} + \frac{(0.67)^2}{0.67} + \frac{(0-0.67)^2}{0.67} + \frac{(1-0.33)^2}{0.33}$$

$$=0.33+0.67+0.67+1.33=3$$

- 确定自由度(?)为(2−1)×(2−1)=1 (if table has n*m items, then freedom = (n−1)*(m−1))，选择显著水平α=0.05。

Degrees of Freedom	Probability, *p*				
	0.99	0.95	0.05	0.01	0.001
1	0.000	0.004	3.84	6.64	10.83
2	0.020	0.103	5.99	9.21	13.82

3<3.84!
不能拒绝零假设

2.3 数据集成和变换

- **数据变换：将数据转换或统一成适合挖掘的形式**
  - **规范化：**将数据按比例缩放，使之落入一个小的特定区间
    - 最小－最大规范化
    - z-score规范化
    - 小数定标规范化
  - **属性构造：**通过现有属性构造新的属性，并添加到属性集中；以增加对高维数据的结构的理解和精确度
  - **数据泛化：**沿概念分层向上汇总

2.3 数据集成和变换

- **最小－最大规范化**

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- **z-score规范化：** 最大最小值未知，或者离群点影响较大的时候适用

$$v' = \frac{v - mean_A}{standard\_dev_A}$$

- **小数定标规范化**

$$v' = \frac{v}{10^j}$$

其中，j是使 $Max(|v'|)<1$的最小整数

2.3 数据集成和变换

■ 虚拟变量陷阱（Dummy Variable Trap）

- ➢ 指虚拟变量之间是多重共线性（Multicollinearity），一个变量可以预测其它变量的值。

- **Dummy variable trap**

This model cannot be estimated (perfect collinearity)

$$wage = \beta_0 + \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

- ➢ 原特征有m个类别时，可转换成m−1个虚拟变量。

http://www.ce.memphis.edu/7012/Lecture17_MLR_Dummy.pdf

本章内容

2.0 认识数据

2.1 为什么要预处理数据

2.2 数据清理

2.3 数据集成和变换

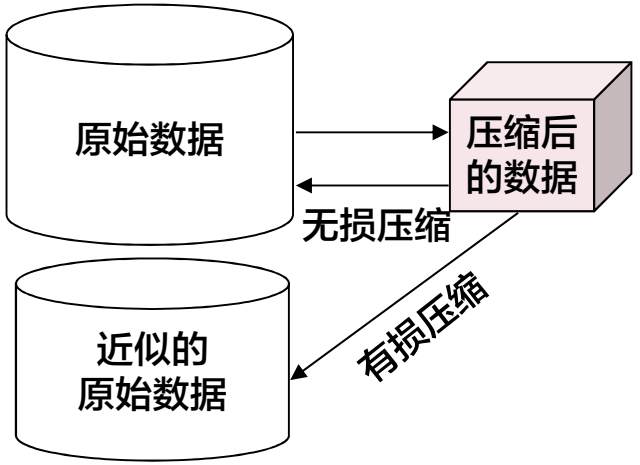
2.4 数据归约

2.4 数据归约

- 为什么需要进行数据规约?
  - 在海量数据上进行复杂数据分析与挖掘需要很大的时空开销
- 数据归约(data reduction)：用来得到数据集的归约表示，它小得多，但可产生相同或几乎相同的分析结果
- 常用的归约策略
  - 维归约，e.g. 移除不重要的属性
  - 数据压缩
  - 数值归约，e.g. 使用模型来表示数据
  - 离散化和概念分层产生
- 用于数据归约的时间不应当超过或“抵消”在归约后的数据上挖掘节省的时间

2.4 数据归约

■ 数据压缩



- ➢ 有损压缩 VS. 无损压缩
- ➢ 字符串压缩
  - 有广泛的理论基础和精妙的算法
  - 通常是无损压缩
  - 在解压缩前对字符串的操作非常有限
- ➢ 音频/视频压缩
  - 通常是有损压缩，压缩精度可以递进选择
  - 有时可以在不解压整体数据的情况下，重构某个片断
- ➢ 两种有损数据压缩的方法：小波变换和主要成分分析

2.4 数据归约 – 数值规约

- 维归约
  - 通过删除不相干的属性或维减少数据量
  - 属性子集选择（特征选择）
- 数值规约
  - 通过选择替代的、较小的数据表示形式来减少数据量
  - 有参方法
    - 使用一个参数模型估计数据，最后只要存储参数即可，不用存储数据（除了可能的离群点）
    - 常用方法：线性回归方法；多元回归；对数线性模型；
  - ✓ 无参方法
    - 不使用模型的方法存储数据
    - 常用方法：直方图，聚类，选样

2.4 数据归约 – 数值规约

■ 回归分析

- ➢ **线性回归**：数据被拟合为一条直线 $Y = wX + b$
  - 两个回归系数，w和b，由手头数据来进行估算
  - 通常适用**最小二乘法**来确定这条直线
- ➢ **多元回归**：线性回归的扩充，允许响应变量Y被建模为两个或多个预测变量的线性或非线性函数

$$Y = b_0 + b_1 X_1 + b_2 X_2.$$

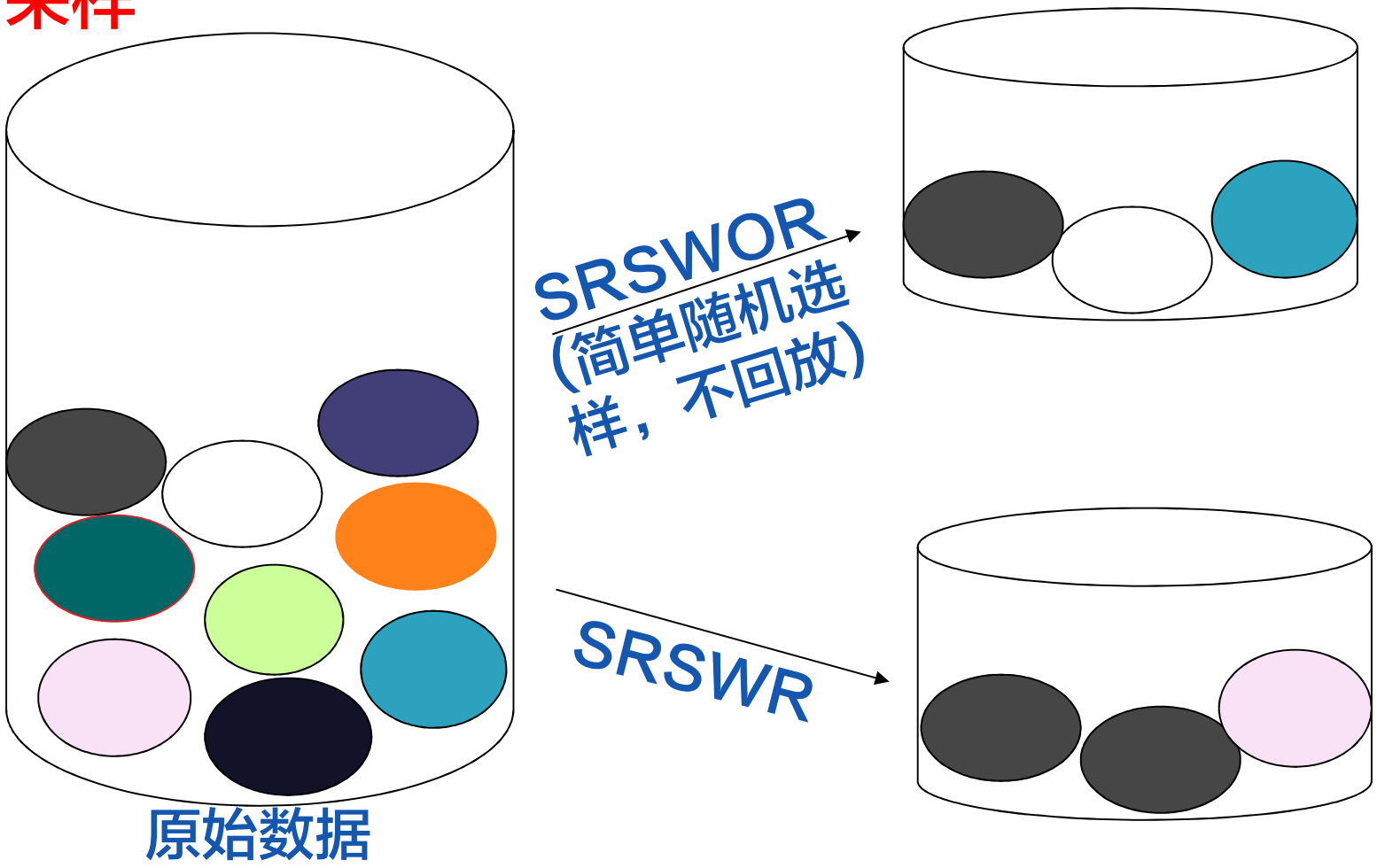
2.4 数据归约 – 数值规约

■ 采样

- ➢ 允许用数据的较小随机样本（子集）表示大的数据集
- ➢ 对数据集D的样本选择：
  - *s*个样本无放回简单随机抽样（Simple random sampling without replacement，SRSWOR）：由D 的N个元组中抽取s个样本（s<N）
  - s个样本有放回简单随机抽样（ Simple Random Sampling With Replacement，SRSWR）：过程同上，只是元组被抽取后，将被回放，可能再次被抽取
  - 聚类选样：D 中元组被分入 M 个互不相交的簇中，可在其中的m个簇上进行简单随机选择（SRS，m<M）

2.4 数据归约 – 数值规约

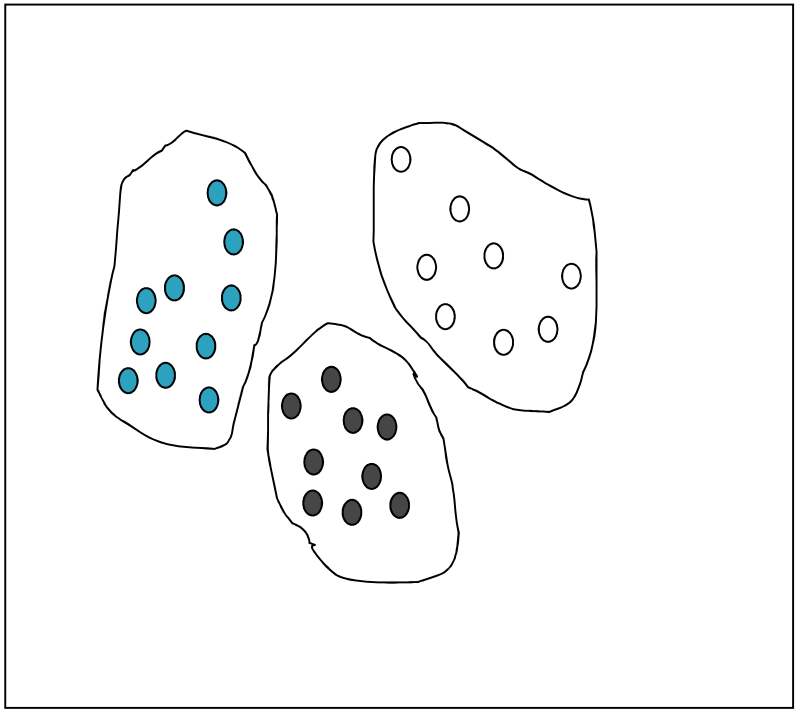
■ 采样



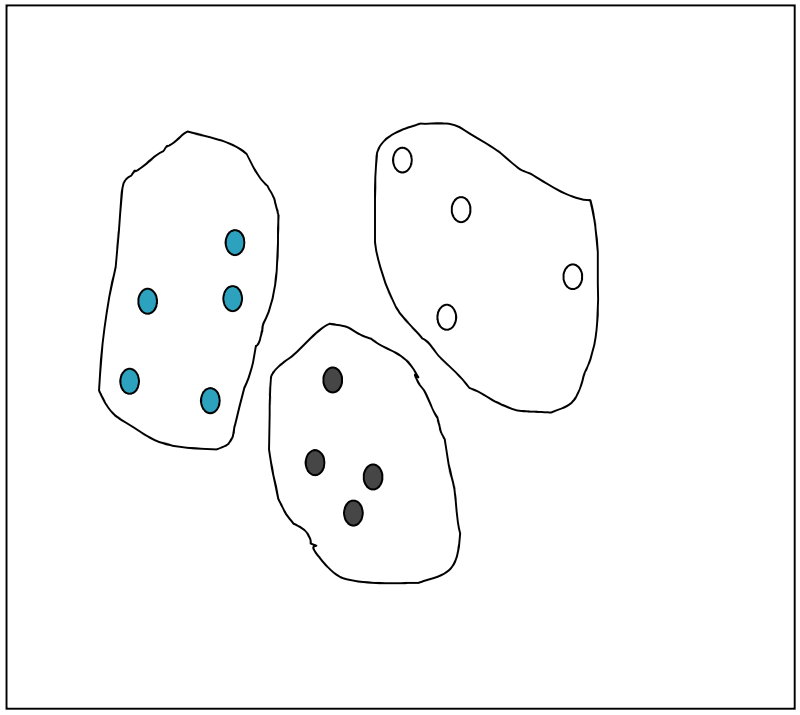
2.4 数据归约 – 离散化和概念分层

■ 采样——聚类采样

原始数据



聚类采样

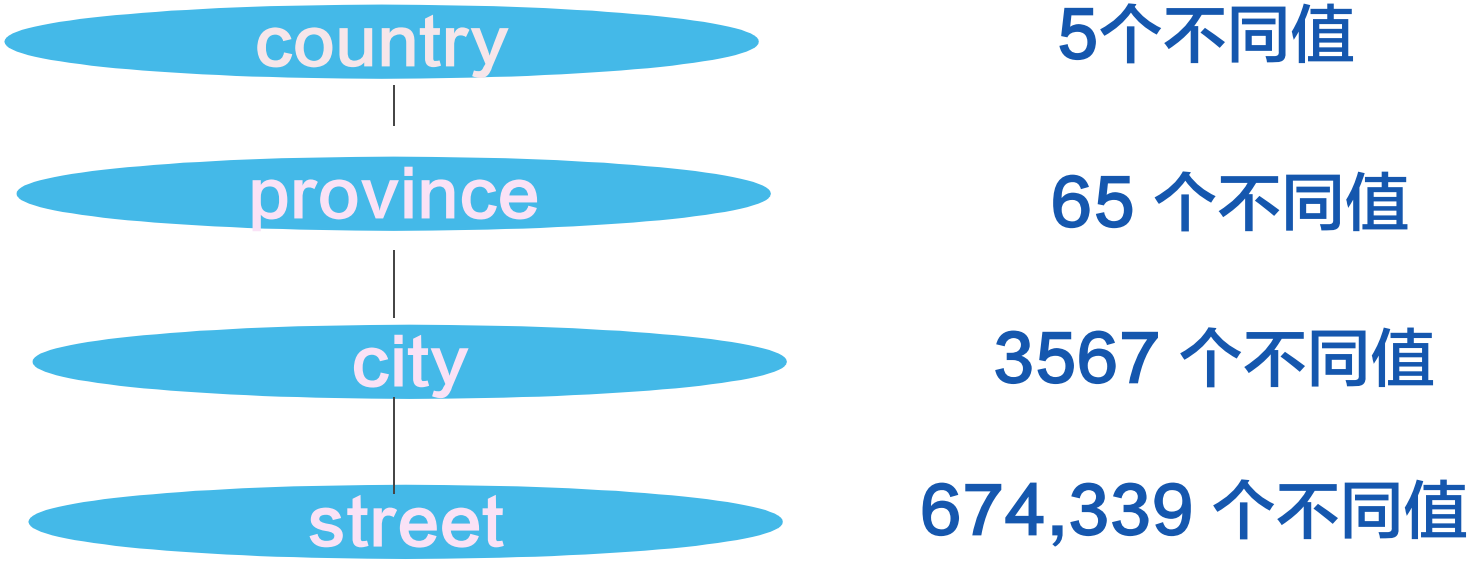


2.4 数据归约 – 离散化和概念分层

- **离散化**
  - 有些分类算法只接受离散属性值
  - 将连续属性的范围划分为区间，区间的标号可以代替实际的数据值，减少给定连续属性值的个数
- **概念分层：** 使用高层概念（比如：青年、中年、老年）来替代底层的属性值（比如：实际年龄）来规约数据
  - 聚类分析产生概念分层可能会将一个工资区间划分为：[51263.98, 60872.34]
  - 通常数据分析人员希望看到划分的形式为[50000，60000]

2.4 数据归约 – 离散化和概念分层

- 分类数据是指无序的离散数据，它有有限个值（可能很多个）。
- 根据在给定属性集中，每个属性所包含的不同值的个数，可以自动的生成概念分层；不同值个数最多的属性将被放在概念分层的最底层。



总结

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Basic statistical data description: central tendency, dispersion, graphical displays
- Data quality: accuracy, completeness, consistency, timeliness, believability, interpretability
- Data cleaning: e.g. missing/noisy values, outliers
- Data reduction: Dimensionality reduction, Numerosity reduction, Data compression
- Data transformation and data discretization: Normalization, Concept hierarchy generation