数据挖掘
第二讲· 认识数据

授课教师：别荣芳 教授、博导

助　　教：周也、王舒扬

Outline

数据集

- **Record（记录）**
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- **Graph and network（图和网络）**
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- **Ordered（序列数据）**
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- **Spatial, image and multimedia（时空、图像、多媒体数据）**
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

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
  - 销售数据库中的对象：顾客、商品或销售；医疗数据库的对象：患者；
  - 大学数据库的对象：学生、教授和课程。
- 数据对象用属性描述
- 数据对象又称样本、实例、数据点或对象
  - 如果数据对象存放在数据库中，则它们是数据元组。
  - 数据库的行对应于数据对象，而列对应于属性。

数据对象

RID	*age*	*income*	*student*	*credit_rating*	*buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	no
5	senior	Low	yes	fair	no
6	senior	Low	yes	excellent	yes
7	middle_aged	Low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	no
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	yes

数据对象

属性

数据对象

数据集

RID	age	income	student	credit_rating	buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	no
5	senior	Low	yes	fair	no
6	senior	Low	yes	excellent	yes
7	middle_aged	Low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	no
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	yes

属性（Attributes）

- **属性 （Attribute or dimensions, features, variables)**: 属性是一个数据字段，表示数据对象的一个特征.
  - *E.g., customer _ID, name, address*
- 属性类型:
  - **标称属性（Nominal）**： 符号或名称
  - **二元属性（Binary）**： 只有两个类别或状态
  - **序数属性（Ordinal）**： 数值之间具有有意义的序或秩
  - **数值属性（Numeric）**: 定量的、可度量的量
    - 区间标度（Interval-scaled）： 用相等的单位尺度度量
    - 比率标度（Ratio-scaled）： 具有固有零点的数值属性

标称属性

- 标称属性（nominal attribute）的值是一些符号或事物的名称。
- 每个值代表某种类别、编码或状态，因此标称属性又被看做是分类的（categorical）。
- 这些值不必具有有意义的序。在计算机科学中，这些值也被看做是枚举的（enumeration）。

标称属性例子

假设hair_color（头发颜色）和marital_status（婚姻状况）是两个描述人的属性。在我们的应用中，hair_color的可能值为黑色、棕色、淡黄色、红色、赤褐色、灰色和白色。属性marital_status的取值可以是单身、已婚、离异和丧偶。hair_color和marital_status都是标称属性。标称属性的另一个例子是occupation（职业），具有值教师、牙医、程序员、农民等。

二元属性

- 二元属性（binary attribute）是一种标称属性，只有两个类别或状态：0或1，其中0通常表示该属性不出现，而1表示出现。如果两种状态对应于true和false的二元属性又称布尔属性。
  - 对称的二元属性，如果它的两种状态具有同等价值并且携带相同的权重
  - 非对称的二元属性，如果其状态的结果不是同样重要的

二元属性例子

倘若属性smoker描述患者对象，1表示患者抽烟，0表示患者不抽烟。类似地，假设患者进行具有两种可能结果的医学化验。属性medical_test是二元的，其中值1表示患者的化验结果为阳性，0表示结果为阴性。

序数属性

- 序数属性（ordinal attribute）是一种属性，其可能的值之间具有有意义的序或秩评定（ranking），但是相继值之间的差是未知的。

序数属性例子

假设drink_size对应于快食店供应的饮料量。这个标称属性具有3个可能的值——小、中、大。这些值具有有意义的先后次序（对应于递增的饮料量）。然而，例如我们不能说“大”比“中”大多少。序数属性的其他例子包括grade（成绩，例如A+、A、A-、B+等）和professional_rank（职位）。职位可以按顺序枚举，如对于教师有助教、讲师、副教授和教授，对于军阶有列兵、一等兵、专业军士、下士、中士等。

数值属性

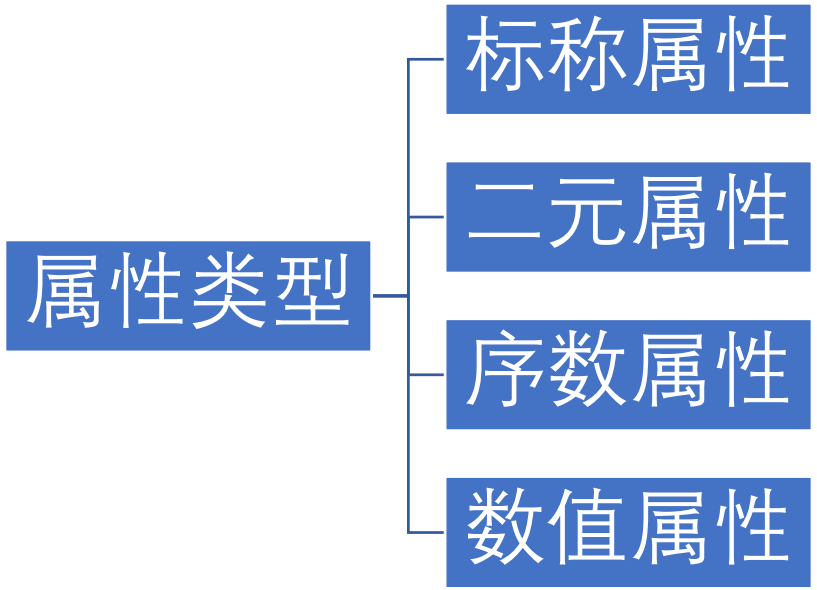
- 数值属性（numeric attribute）是定量的，即它是可度量的量，用整数或实数值表示。数值属性可以是区间标度的或比率标度的。
  - 区间标度（interval scaled）属性用相等的单位尺度度量。区间属性的值有序，可以为正、0或负。因此，除了值的秩评定之外，这种属性允许我们比较和定量评估值之间的差。
  - 比率标度（ratio scaled）属性是具有固有零点的数值属性。也就是说，如果度量是比率标度的，则我们可以说一个值是另一个的倍数（或比率）。

数值属性

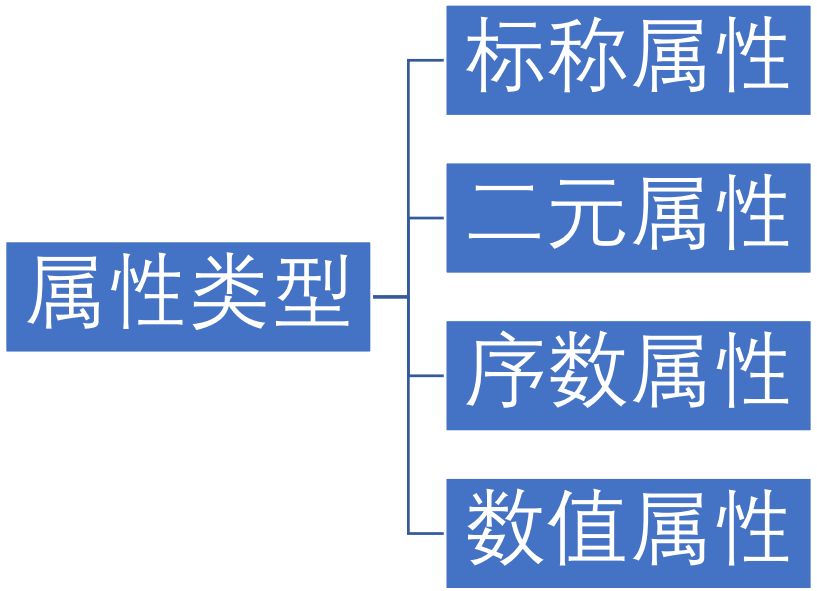
区间标度属性例子 temperature（温度）属性是区间标度的。假设我们有许多天的室外温度值，其中每天是一个对象。把这些值排序，则我们得到这些对象关于温度的秩评定。此外，我们还可以量化不同值之间的差。例如，温度20℃比5℃高出15℃。日历日期是另一个例子。例如，2002年与2010年相差8年。

比率标度属性例子 不像摄氏和华氏温度，开氏温标（K）具有绝对零点（0° K=-273.15℃）：在该点，构成物质的粒子具有零动能。比率标度属性的其他例子包括诸如工作年限（例如，对象是雇员）和字数（对象是文档）等计数属性。其他例子包括度量重量、高度、速度和货币量（例如，100美元比1美元富有100倍）的属性。

离散属性 VS 连续属性



离散属性 VS 连续属性



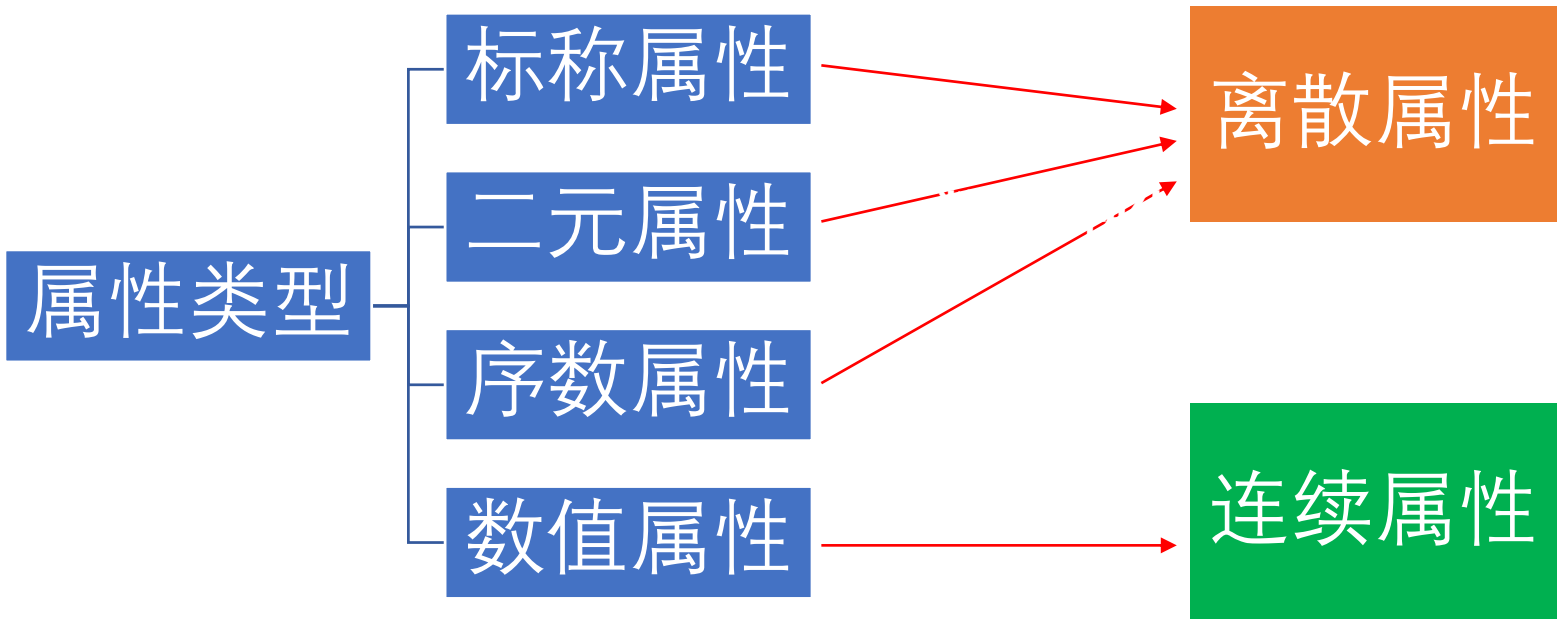
离散属性

具有有限或无限可数个值，可以用或不用整数表示

连续属性

如果属性不是离散的，则它是连续的

离散属性 VS 连续属性



具有有限或无限可数个值，可以用或不用整数表示

如果属性不是离散的，则它是连续的

Outline

Outline

数据的基本统计描述

中心趋势度量

度量数据分布的中部或中心位置

离散趋势度量

数据的散布的情况

数据统计描述的图形显示

中心趋势度量

- 代数度量（algebraic measure）
  - 均值（mean）

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

  - 加权均值（weighted mean）

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

  - 截尾均值（trimmed mean）
    - 去掉高、低极端值得到的均值

中心趋势度量

- 中位数（median）
  - 对于倾斜（非对称）数据，数据中心的更好度量是中位数（median）
  - 有序数据值的中间值
  - 有序集合（假设是N个不同值）中，如果N是奇数，则中位数是有序集合的中间值，如果N是偶数，则中位数是有序集合的中间两个数的平均值
  - 插值公式计算中位数的近似值

$$median = L_1 + (\frac{N/2 - (\sum freq)_l}{freq_{median}})width$$

其中L_1是中位数区间的下界，N是整个数据集的值的个数，$(\sum freq)_l$是低于中位数区间的所有区间的频率和，$freq_{median}$是中位数区间的频率，而width是中位数区间的宽度。

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

中心趋势度量

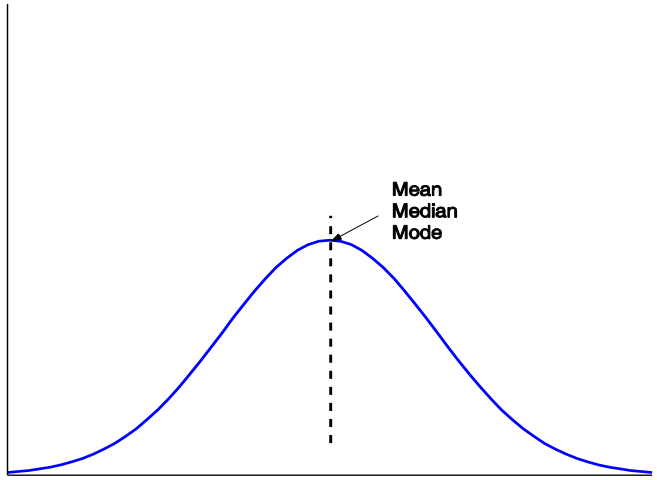
- 众数
  - 集合中出现频率最高的值
  - 可能有多值，单峰的（unimodal）双峰的（ bimodal）三峰的（ trimodal）
  - 对适度倾斜（非对称）的单峰频率曲线，已知均值和中位数，可以利用经验公式计算众数：

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

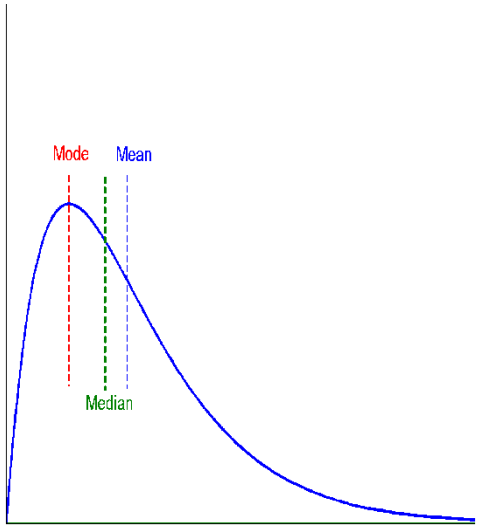
- 中列数
  - 数据的最大值和最小值的平均值

度量中心趋势

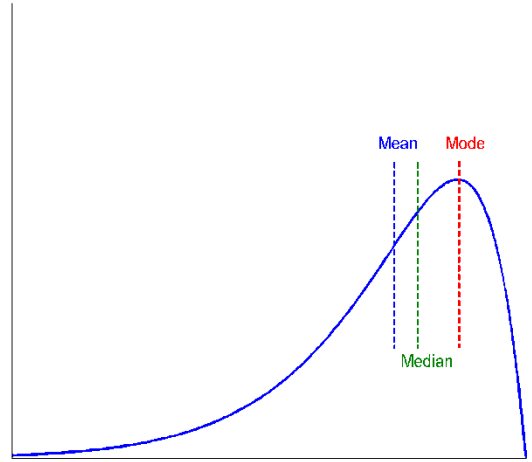
- 完全对称的数据：均值=中位数=众数
- 正倾斜数据：众数<中位数
- 负倾斜数据：众数>中位数



对称数据



正倾斜数据



负倾斜数据

数据的基本统计描述

中心趋势度量

度量数据分布的中部或中心位置

离散趋势度量

数据的散布的情况

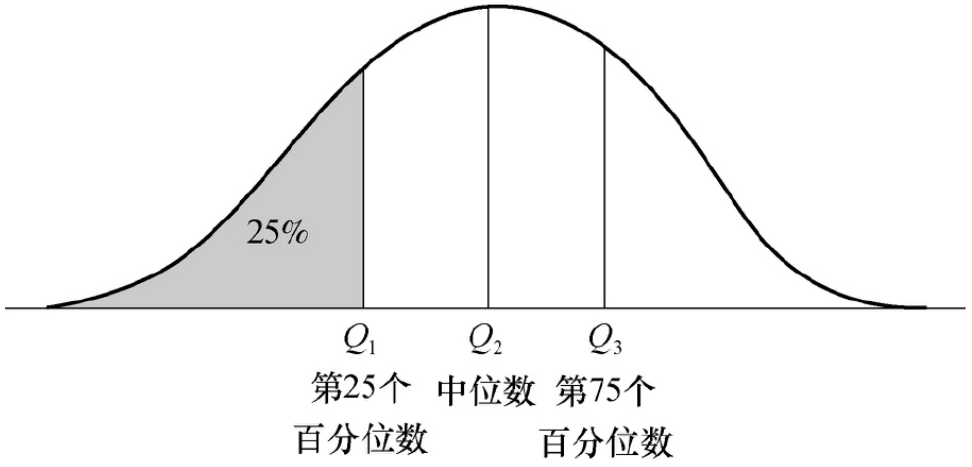
数据统计描述的图形显示

度量数据散布

- 考察和评估数值数据散布或发散的度量
  - 极差、四分位数、四分位数极差
  - 五数概括
  - 方差、标准差

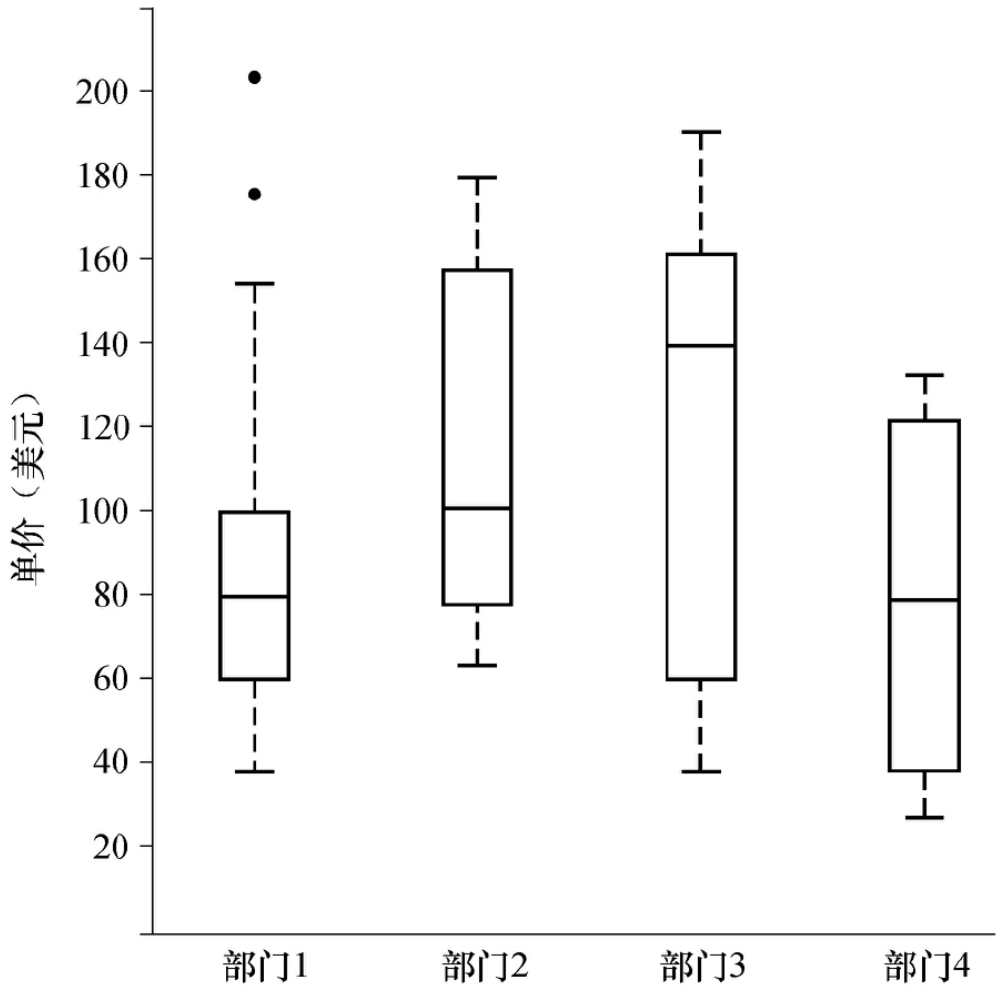
度量数据散布

- 极差
  - 最大值（max）和最小值（min）之差
- 分位数
  - 取自数据分布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
  - 2-分位数：中位数
  - 四分位数：3个数据点（Q_1，中位数，Q_3），把数据分布划分成4个相等的部分
  - 百分位数：100-分位数，把数据分布划分为100个大小相等的连贯集
- 四分位数极差
  - $IQR=Q_3-Q_1$



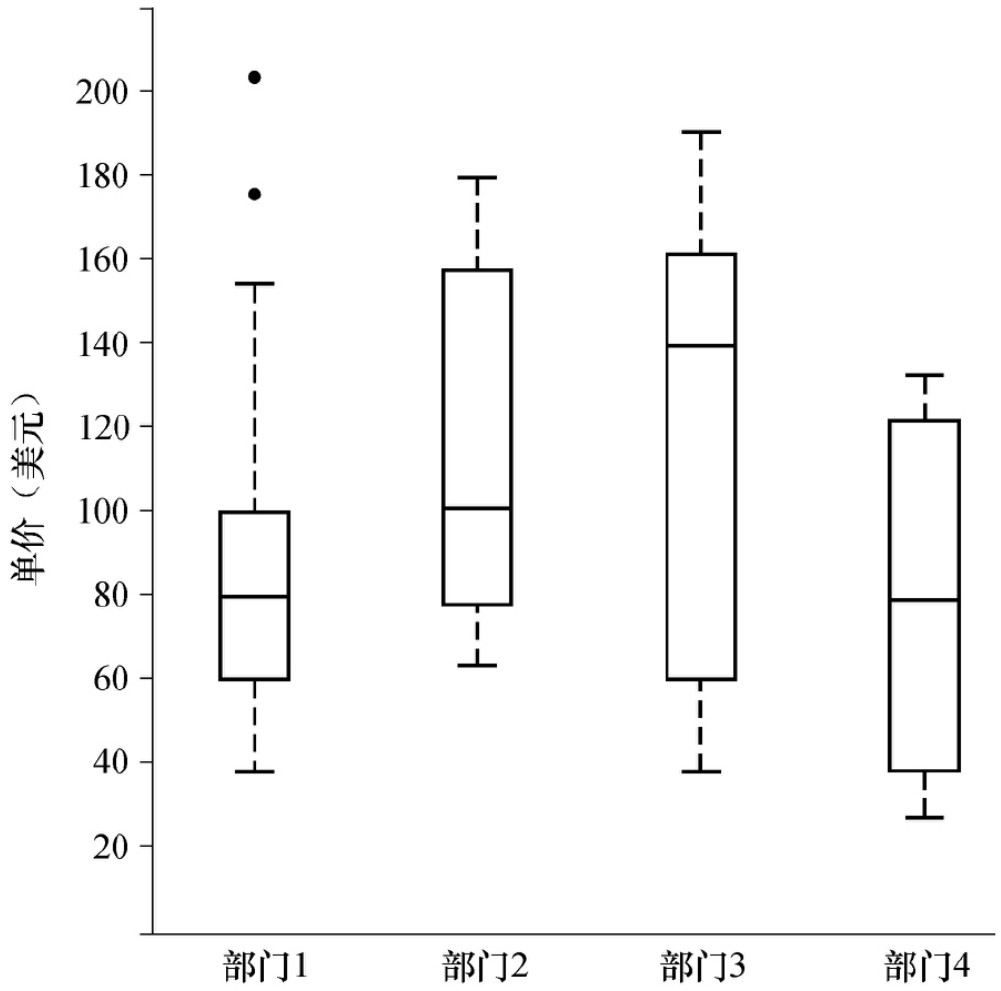
度量数据散布

- 五数概括（five-number summary）
  - 中位数、四分位数Q1和Q3，最小和最大观测值
  - 即Median, Q1 , Q3 , Minimum, Maximum
- 孤立点判定规则：
  - 落在至少高于Q3或低于Q1 的1.5*IQR 处的值
- 盒图
  - 一种流行的分布的直观表示，体现了五数概括
  - 盒的端点在四分位数上，盒的长度
    是中间四分位数极差(IQR)
  - 中位数用盒内的线标记
  - 盒外的两条线（称作胡须）延伸到最小和最大观测值



度量数据散布

- 当处理数量适中的观测值时，潜在的离群点需要在盒图中个别绘出
  - 仅当这些值超过四分位数不到 1.5×IQR 时，胡须扩展到最高和最低观测值。
  - 否则，胡须出现在四分位数的1.5×IQR之内的最极端的观测值处终止。
  - 其余情况个别地绘出。



度量数据散布

- 方差
  - N个观测值x_1，x_2，…，x_N的方差为：

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

$\bar{x}$ 是观测值的均值

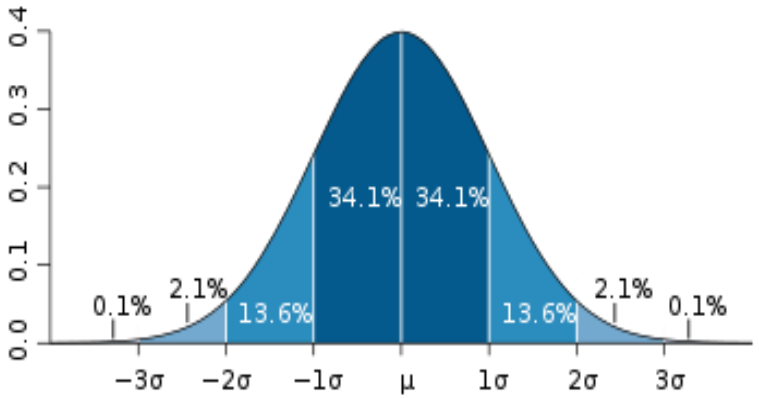
- 标准差
  - σ 是方差 σ^2 的平方根
  - σ 度量关于均值的发散，仅当选择均值作为中心度量时使用
  - 仅当不存在发散，即当所有的观测值都具有相同值时， σ =0，否则，σ >0

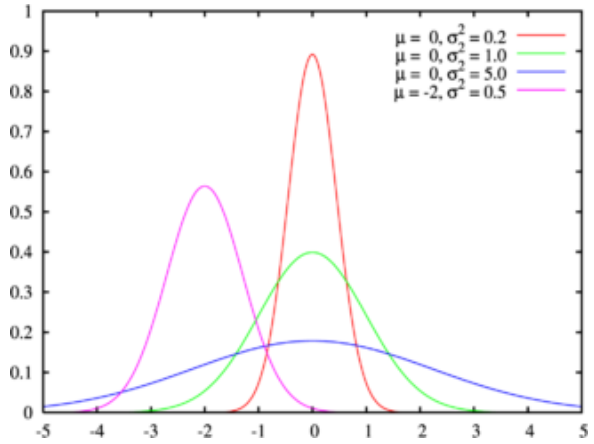
关于标准差

- 标准差的观念是由卡尔·皮尔逊（Karl Pearson)引入到统计中；
- 标准差是一组数值自平均值分散开来的程度的一种测量观念；
- 一个较大的标准差，代表大部分的数值和其平均值之间差异较大；
  - 一个较小的标准差，代表这些数值较接近平均值；
- 例如，两组数的集合｛0，5，9，14｝和｛5，6，8，9｝其平均值都是 7 ，但第二个集合具有较小的标准差。

正态分布的规则

- 深蓝区域是距平均值小于一个标准差之内的数值范围。在正态分布中，此范围所占比率为全部数值之 68% 。





- 根据正态分布，两个标准差之内（深蓝，蓝）的比率合起来为 95% 。
- 根据正态分布，三个标准差之内（深蓝，蓝，浅蓝）的比率合起来为 99% 。

数据的基本统计描述

中心趋势度量

度量数据分布的中心

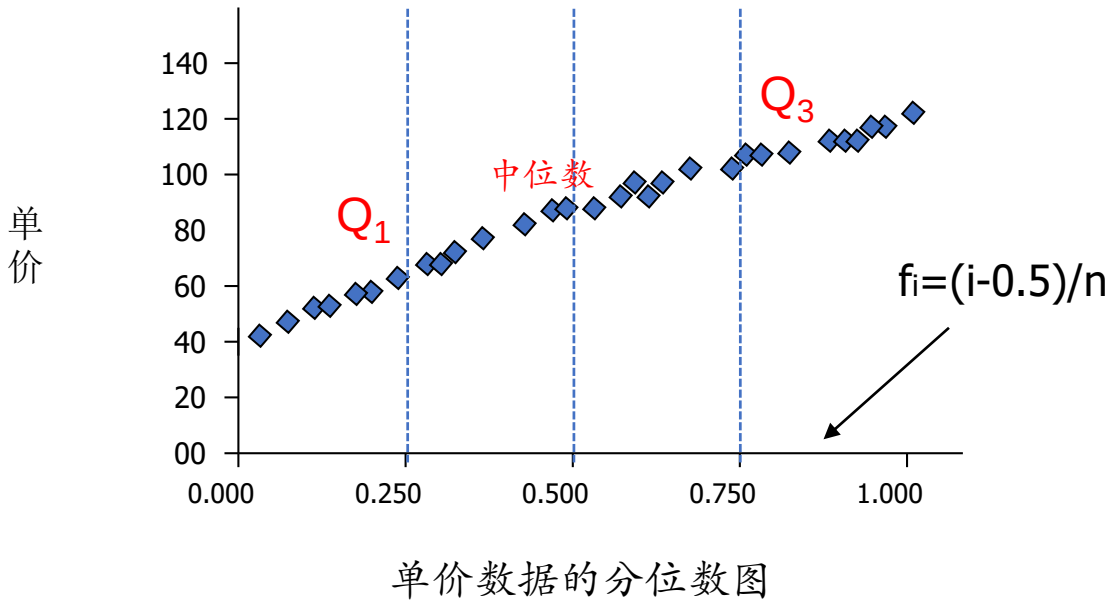
离散趋势度量

数据的散布的情况

数据统计描述的图形显示

分位数图

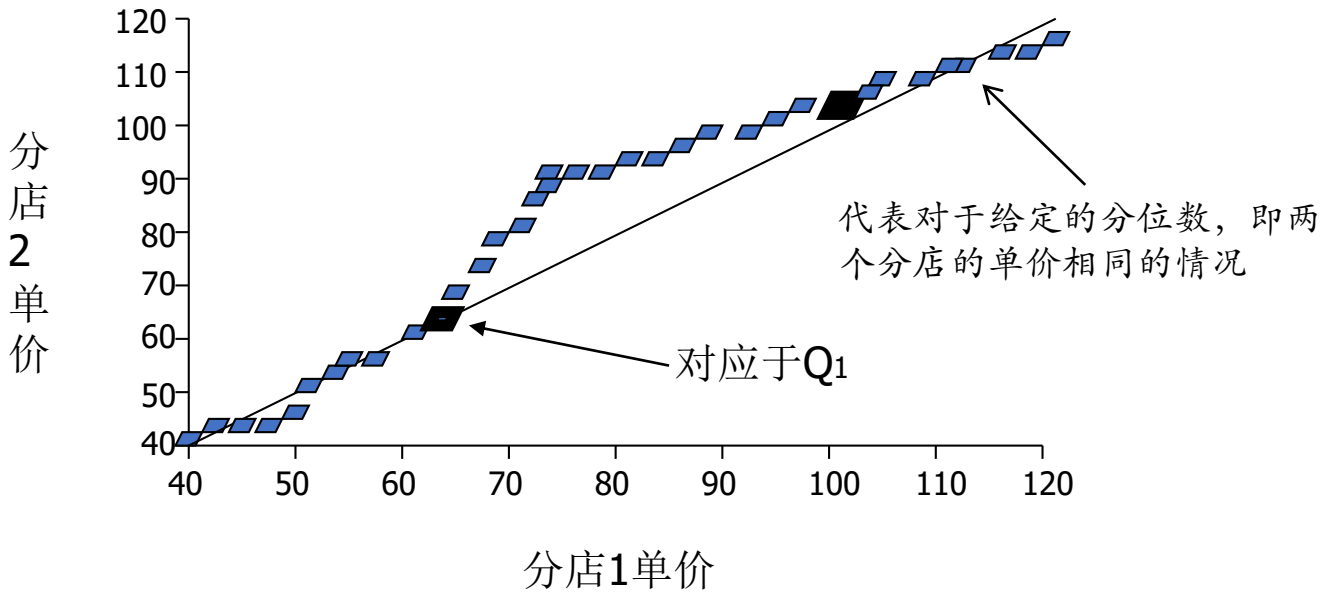
- 一种观察单变量数据分布的简单有效方法
- 显示给定属性的所有数据，允许用户评估总的情况和不寻常的出现；
- 绘出分位数的信息；
- 设x_i（i=1, 2…, n）是按递增序排序的数据，使得x_1是最小的观测值，而x_n是最大的。每个观测值x_i与一个百分数f_i配对，指出大约100×f_i%的数据小于x。



单价数据的分位数图

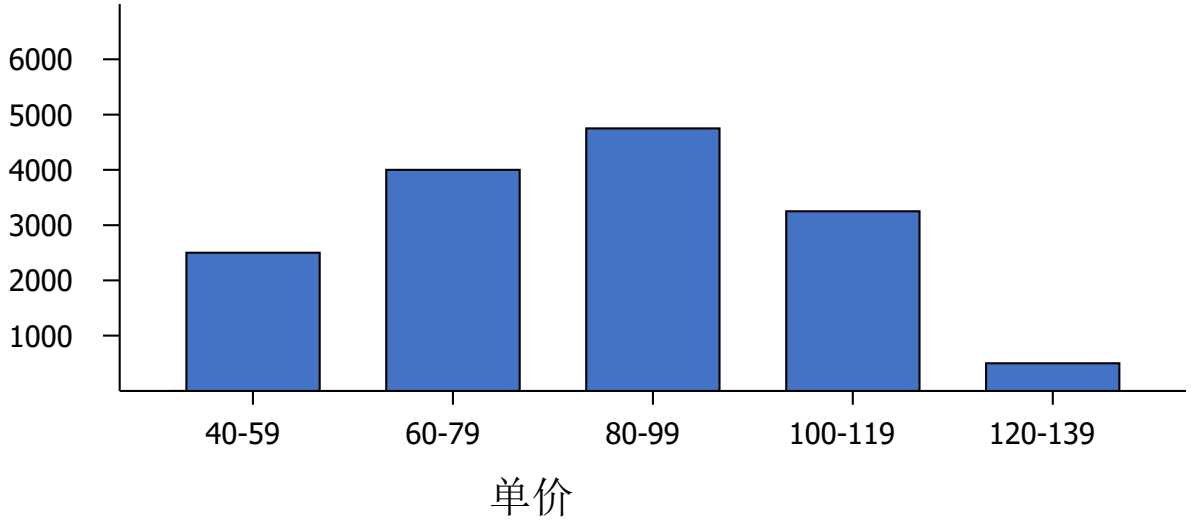
分位数-分位图（q-q图）

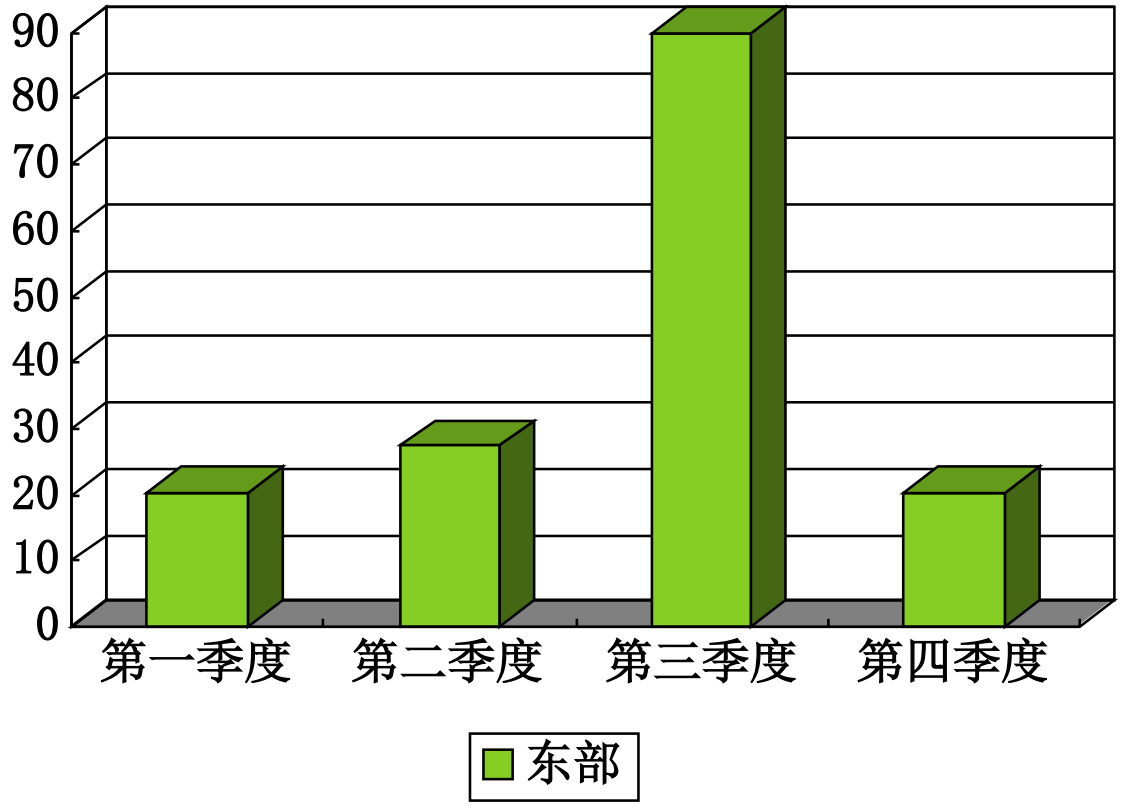
- 分位数-分位数图（quantile-quantile plot）或q-q图对着另一个对应的分位数，绘制一个单变量分布的分位数
- 可以观察从一个分布到另一个分布是否有漂移。



两个不同分店的单价数据q-q图

直方图

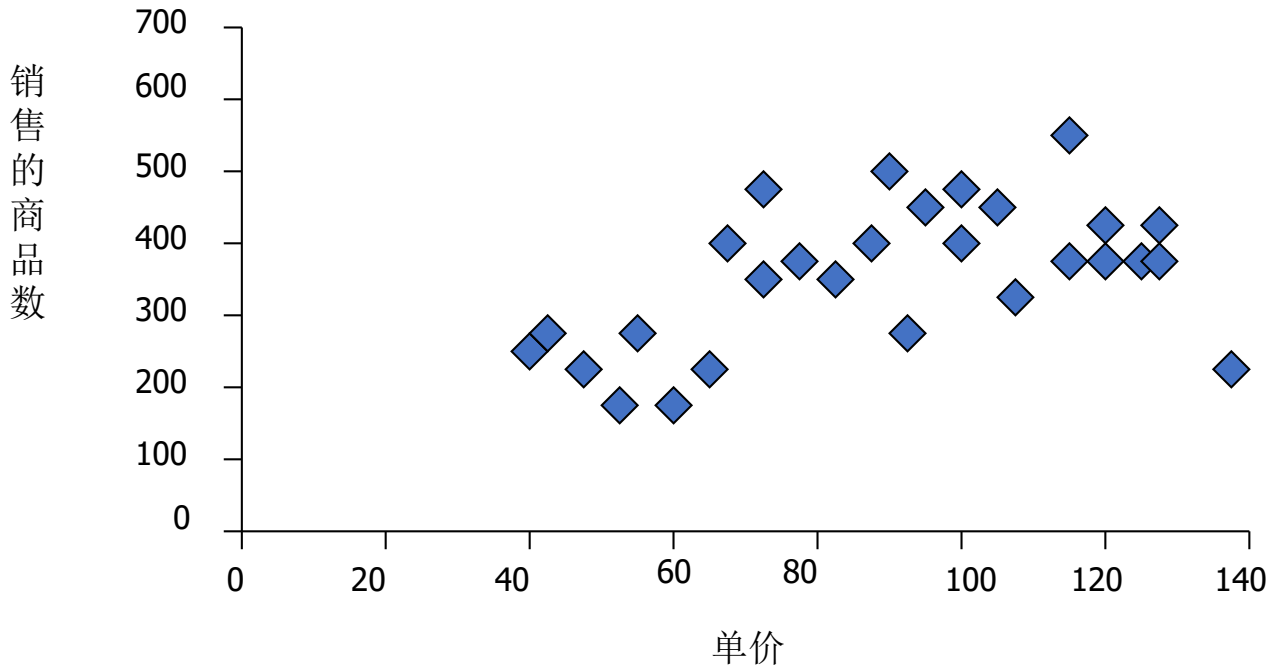

6000
5000
4000
3000
2000
1000
40-59
60-79
80-99
100-119
120-139
单价


90
80
70
60
50
40
30
20
10
0
第一季度
第二季度
第三季度
第四季度
东部

数据集的直方图

散布图

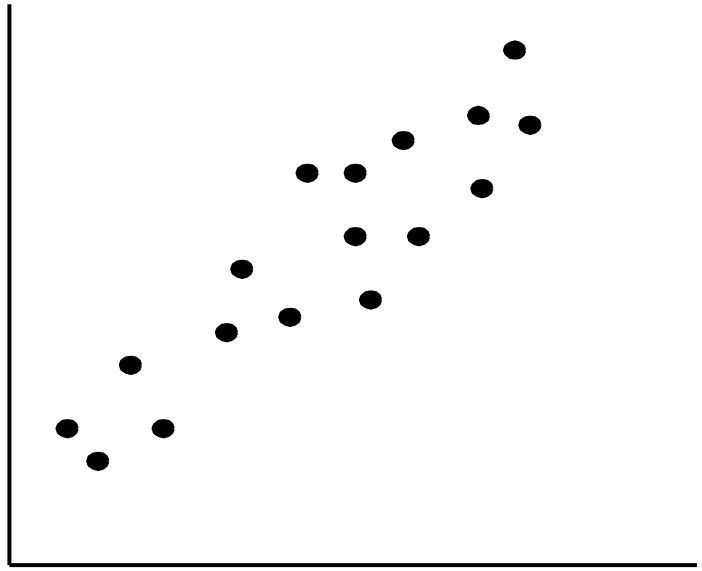
- 观察两个数值变量之间是否存在联系的图形方法之一



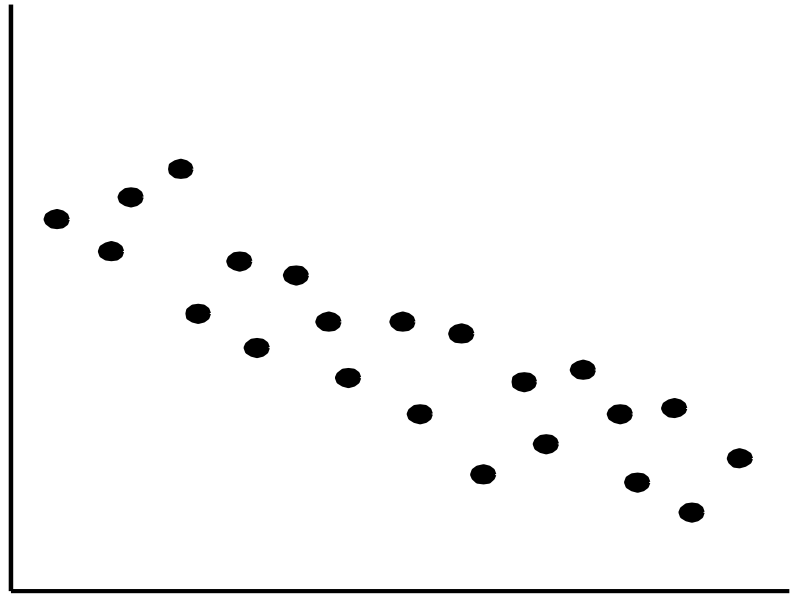
数据集的散布图

散布图

- 正相关 VS 负相关



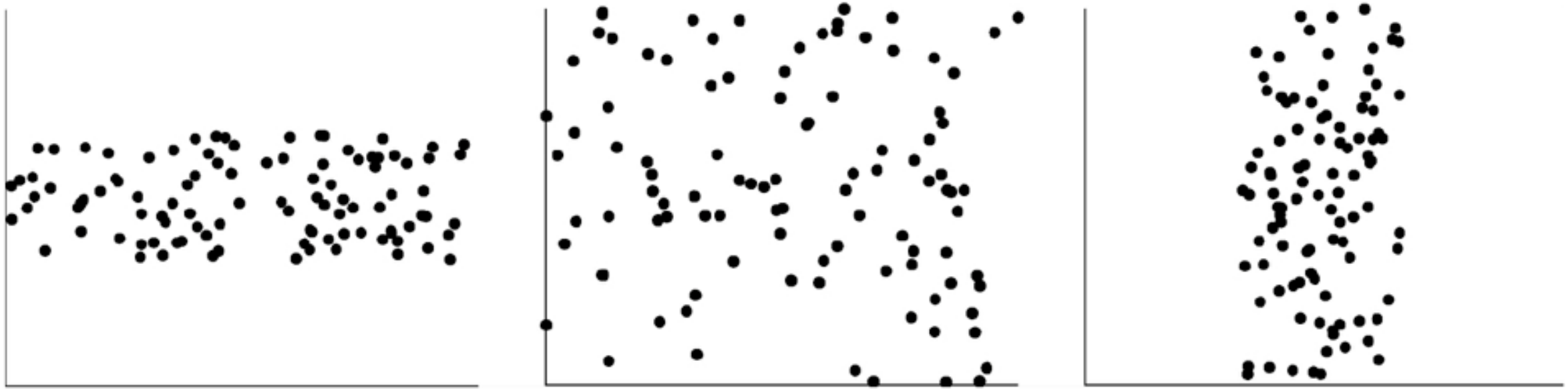
正相关



负相关

散布图

- 无相关性



三种情况，每个数据集中两个属性之间都不存在观察到的相关性

Outline

1 数据对象和属性类型

2 数据的基本统计描述

3 度量数据的相似度

4 数据可视化