数据分析及实践
Analysis and Practice of the Data

第三章 数据统计

刘 淇
Email: qiliuql@ustc.edu.cn

http://staff.ustc.edu.cn/~qiliuql/AD2021.html

3/29/2021

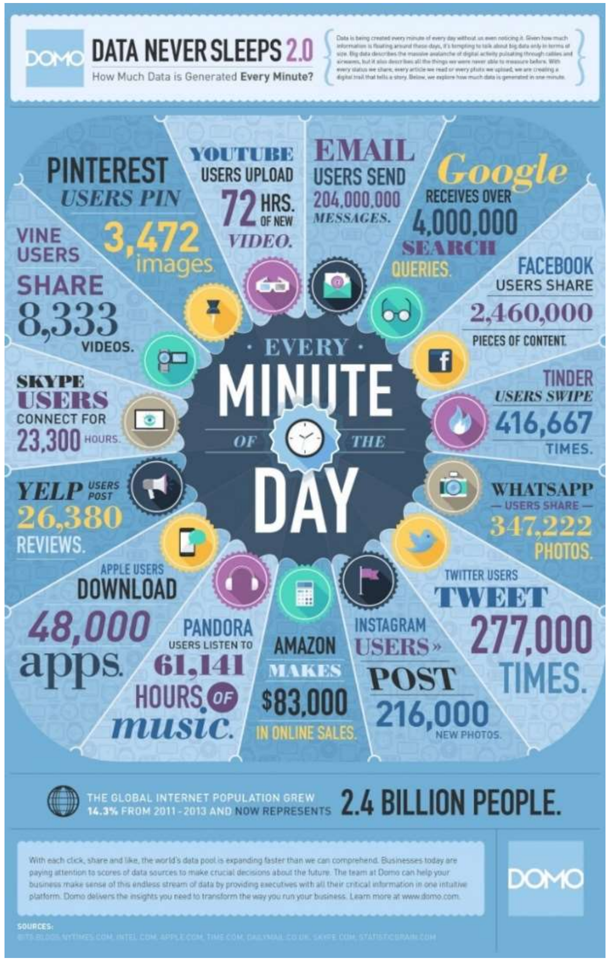
第二章数据分析入门小结

- 数据采集 Data Acquisition
  - 信息检索
  - 网络爬虫
- 数据预处理 Data Preprocessing
  - 数据清洗
  - 数据集成
  - 数据变换
  - 数据规约
- 特征工程 Feature engineering
  - 特征设计
  - 特征选择

Data

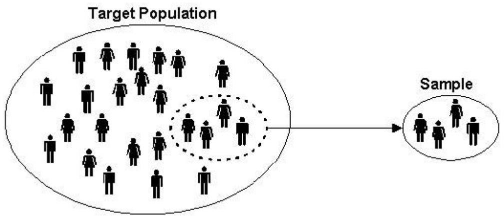
- 大数据
  - 数据量大
  - 类型繁多
  - 时效性高
  - 价值密度低

- 大数据由于本身特性，通常处理代价巨大，可先利用统计手段了解数据基本信息
- 在实际处理大数据前，还可先在抽样得到的小型数据集上对总体进行推断



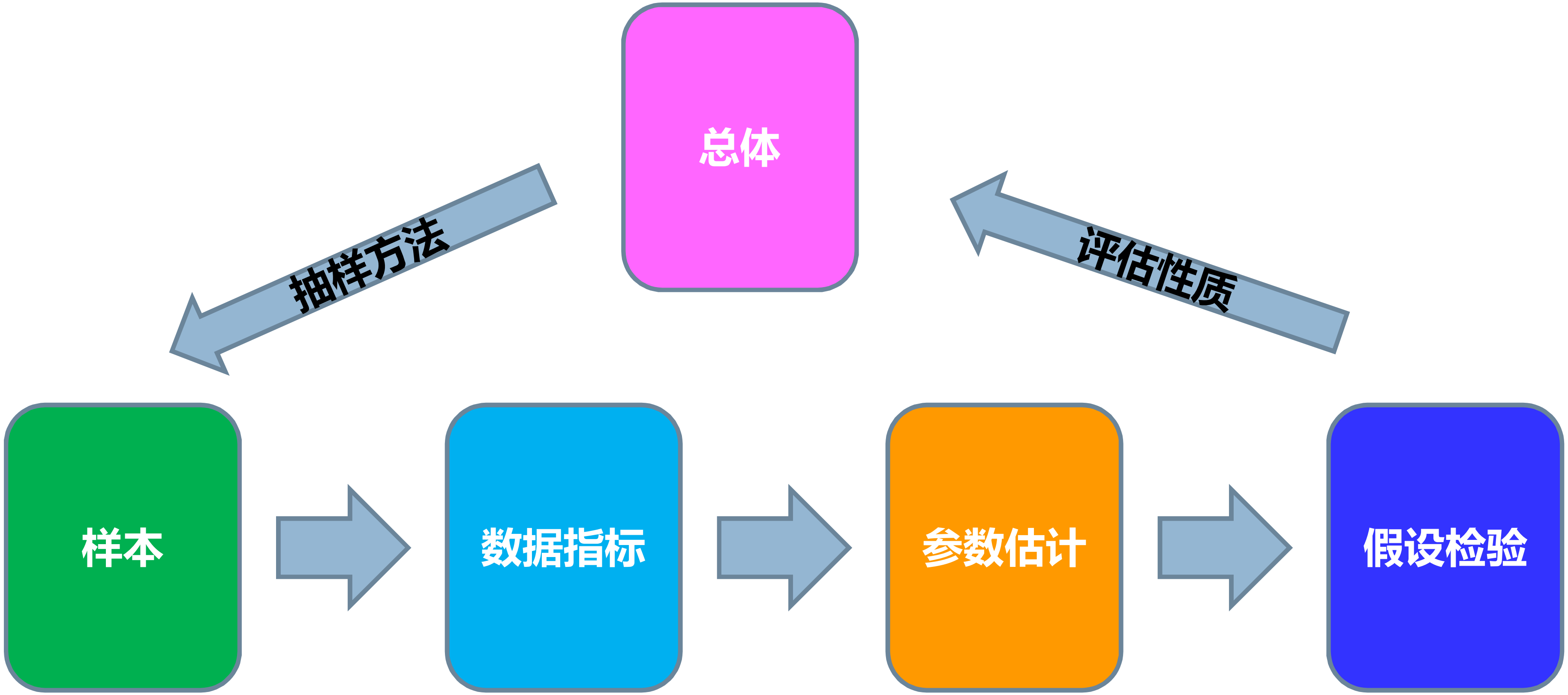
Data

- 总体：
  - 在每一个特定的大数据分析问题中，问题有关对象（个体）所构成的集合即为待研究问题的总体(Population)
  - 总体是由客观存在且有同一性质基础的多个个体结合而成的
  - 例如：
    - 对班级进行研究：全体同学是总体，每位同学是个体
    - 对社交网络进行研究：所有用户是总体，每位用户是个体
- 样本
  - 从总体中抽取若干个个体
  - 随机性与 独立性



- 本章介绍一些基本统计分析处理方法，获得对于样本总体特征的信息

Data



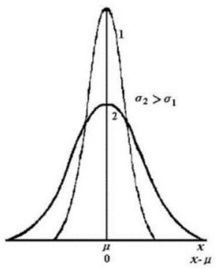
Data

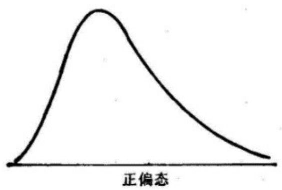
- 数据分布基本指标
- 参数估计
- 假设检验
- 抽样方法

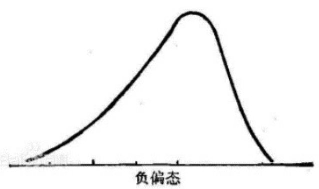
数据分布基本指标

- 在对大数据进行研究时，研究者往往希望知道所获得的数据的基本分布特征

- 数据分布的特征可以从三个方面进行测度和描述：
  - 描述数据分布的集中趋势：反映数据向其中心靠拢或聚集程度
  - 描述数据分布的离散程度：反映数据远离中心的趋势或程度
  - 描述数据分布的形状变化：反应数据分布的形状特征





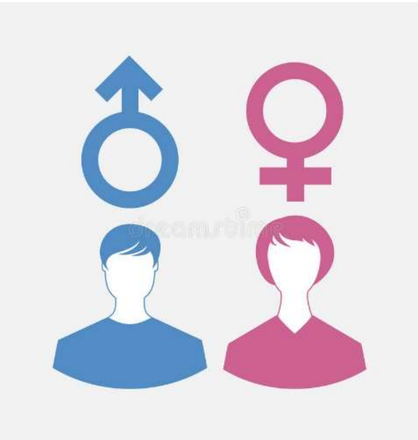


数据分布基本指标

- 集中趋势
  - 集中趋势反映了一组数据的中心点位置所在及该组数据向中心靠拢或聚集的程度。
- 四种最常用的反映数据集中趋势的指标：
  - 平均数
  - 中位数
  - 分位数
  - 众数

数据分布基本指标-集中趋势

- 平均数
  - 平均数也称均值(mean)，它是一组数据相加后除以数据的个数得到的结果，是集中趋势最主要的指标。
  - 主要适用于数值型数据，而不适用于分类数据和顺序数据。



数据分布基本指标-集中趋势

- 简单平均数(simple mean)
  - 根据未经分组数据计算得到的平均初即为简单平均数。
  - 若有一组数据，$x_1, x_2, x_3, \ldots, x_n$，简单平均数用μ 表示，则该数据的简单平均数为：

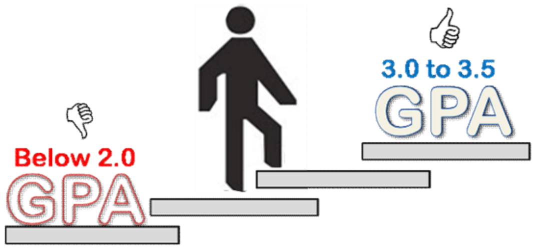
$$\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

数据分布基本指标-集中趋势

- 加权平均数(weighted mean)
  - 根据分组数据计算的平均数称为加权平均数。、
  - 若有一组数据被分为k组，各组的值分别用$M_1, M_2, M_3, \ldots, M_k$ 表示
  - 各组变量出现的频数分别用$f_1, f_2, f_3, \ldots, f_k$ 表示，则该组数据的加权平均数为：

$$\mu = \frac{M_1 f_1 + M_2 f_2 + M_3 f_3 + \ldots + M_n f_n}{f_1 + f_2 + f_3 + \ldots + f_n} = \frac{\sum_{i=1}^{k} M_i f_i}{n}$$

Grade	GPA
A	4.0
B	3.0
C	2.0
D	1.0
F	0.0



数据分布基本指标-集中趋势

- 几何平均数(geometric mean)
  - 几何平均数是n 个变量值乘积的n 次方根，用G 表示。
  - 若有一组数据被分为k组，各组的值分别用$M_1, M_2, M_3, \ldots, M_k$ 表示
  - 主要用于计算平均比率。当所掌握的变量值本身是比率的形式时，采用几何平均数更为合理。
  - 若有一组数据，$x_1, x_2, x_3, \ldots, x_n$，则该组数据的几何平均数为：

$$G = \sqrt[n]{x_1 \times x_2 \times x_3 \times \ldots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

数据分布基本指标-集中趋势

- 中位数
  - 中位数是一组数据排序后处于中间的变量值，用M_e 表示。
  - 中位数主要适用于测度顺序数据的集中趋势，也适用于数值型数据，但不适用于分类数据。
  - 当数据围绕其中心对称分布时，有简单平均数= 中位数.
  - 若有一组数据，$x_1, x_2, x_3, \ldots, x_n$，排序后的顺序为$x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$，则该数据的中位数为：

$$M_e = \begin{cases} x_{(\frac{n+1}{2})} & \text{n 为奇数;} \\ \frac{1}{2}\{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\} & \text{n 为偶数.} \end{cases}$$

均值易受噪声的影响
但中位数可能不唯一，
没有均值的数学性质好

数据分布基本指标-集中趋势

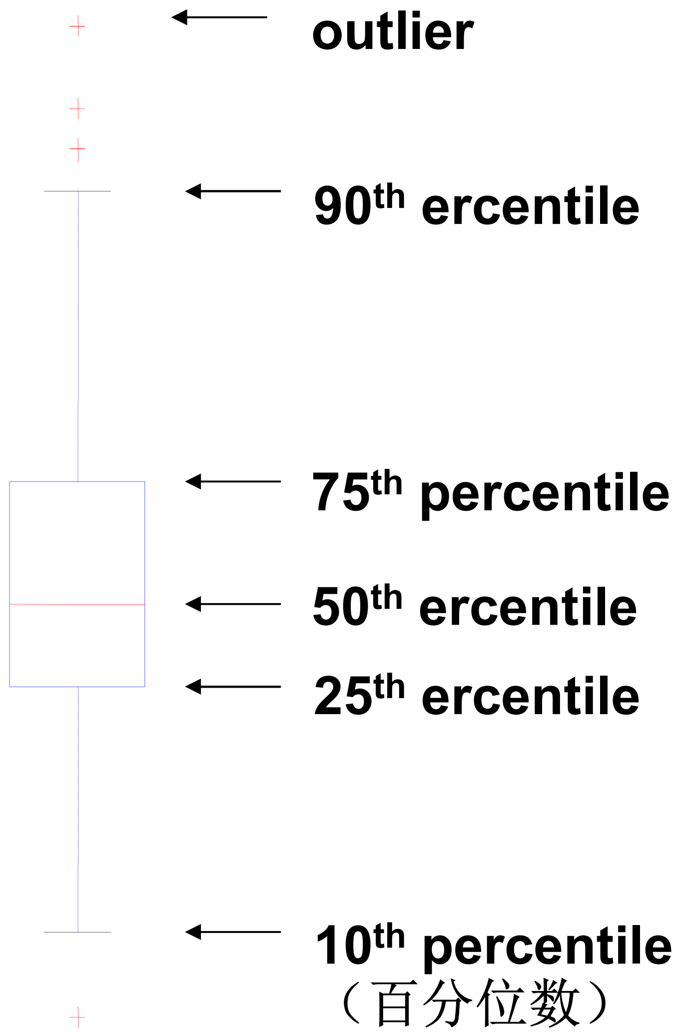
- 分位数
  - 中位数用1 个点将数据两等分，类似的，若用3 个点将数据四等分、9 个点将数据十等分、99 个点将数据一百等分，则对应等分点上的值为四分位数(quartile)、十分位数(decile) 和百分位数(percentile)。
  - 四分位数也称四分位点，它通过3 个点将数据等分成四个部分。不难看出，中间的四分位数就是中位数，所以通常所提到的四分位数是指处在25% 位置上的数值（下四分位数）和处在75% 位置上的数值（上四分位数）。

Q1 Q2 Q3

25%	25%	25%	25%

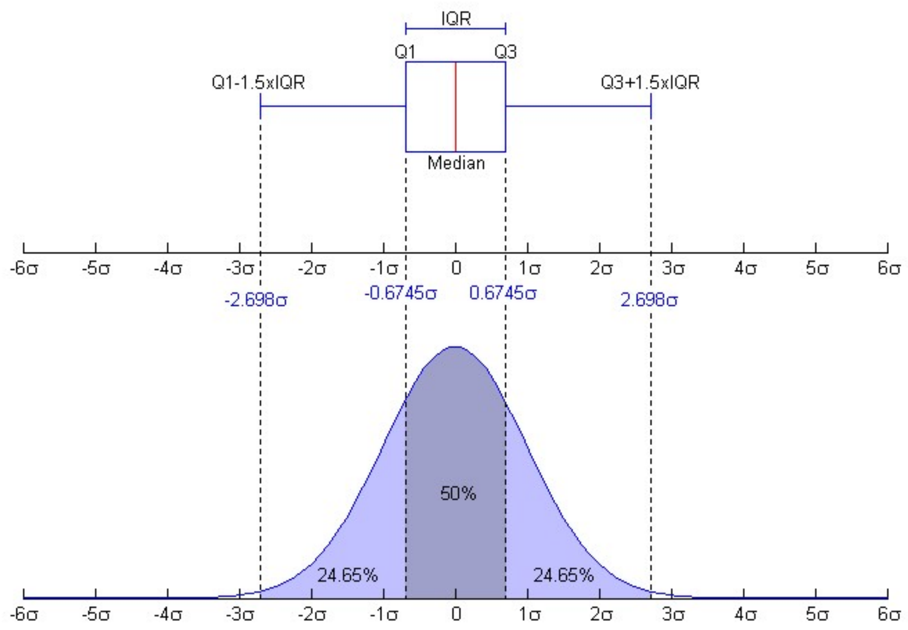
数据分布基本指标-集中趋势

- 分位数
- 箱图（Box Plots）
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



数据分布基本指标-集中趋势

- 分位数
- 箱图（Box Plots）
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



https://wiki.mbalib.com/wiki/%E7%AE%B1%E7%BA%BF%E5%9B%BE

课外实践：案例学习

- IRIS(鸢尾花) + sklearn特征工程案例
  - http://www.cnblogs.com/jasonfreak/p/5448385.html
  - 1. 数据集的描述与导入

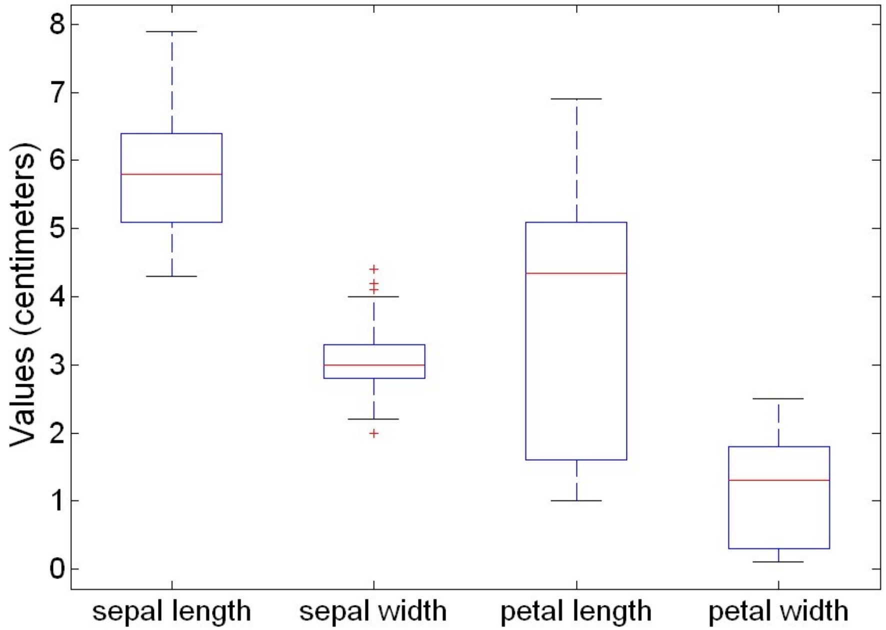
数据的特征: 花萼长度 | 花萼宽度 | 花瓣长度 | 花瓣宽度

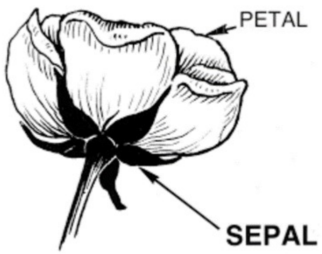
花的类别: 山鸢尾 | 杂色鸢尾 | 维吉尼亚鸢尾

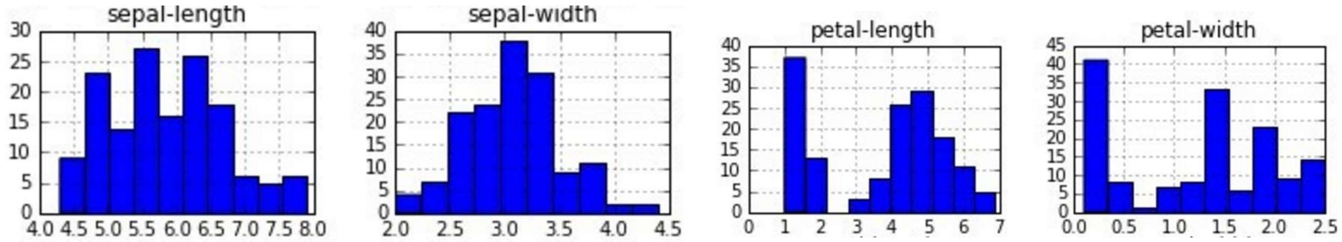
```python
from sklearn.datasets import load_iris

#导入数据集IRIS
iris = load_iris()

#特征矩阵
iris.data

#目标向量
iris.target
```

数据分布基本指标-集中趋势

- 分位数
  - IRIS(鸢尾花)



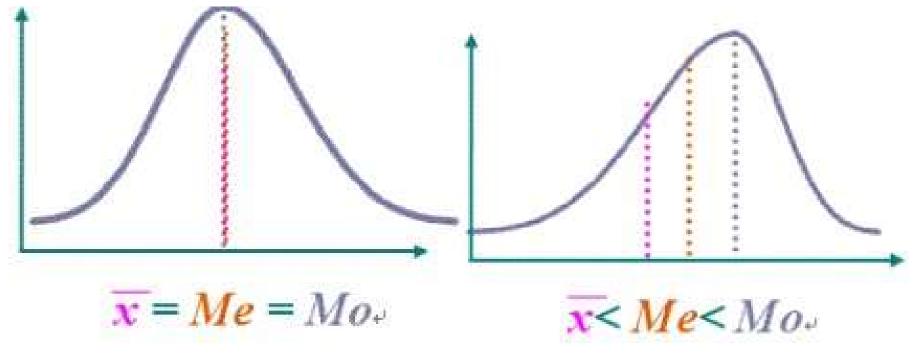




http://archive.ics.uci.edu/ml/datasets/Iris/

数据分布基本指标-集中趋势

- 众数
  - 众数(mode)用M_0 表示,是一组数据中出现次数最多的变量值。
  - 主要用于测度分类数据的集中趋势，也适用于作为数值型数据以及顺序数据集中趋势的测度值。
  - 不同于平均数的是，众数不会受到数据中极端值的影响，是具有明显集中趋势点的数值.
  - 通常，众数只有在数据量较大的情况下才有意义。



均值 中位数 众数

数据分布基本指标-离散程度

- 离散程度
  - 离散程度反映了各个数据属性值远离其中心值的程度，是数据分布的另一个重要特征。
  - 数据的离散程度越大，则集中趋势的测度值对该组数据的代表性就越差，反之亦然。

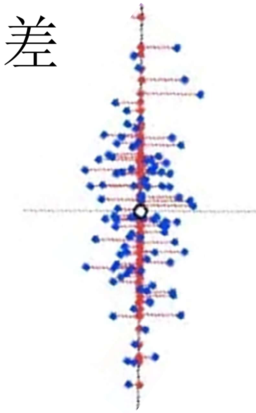
- 四种最常用的反映数据离散程度的指标：
  - 方差和标准差
  - 极差和四分位差
  - 异众比率
  - 变异系数

数据分布基本指标-离散程度

- 方差和标准差
  - 在数值型数据中, 刻画数据围绕其中心位置附近分布的数字特征时，最重要且最常用的是方差(variance) 和标准差(standard deviation)。

  - 方差是各个变量与均值之差平方的平均数
    - 通过平方的方法消去差值中的正负号，再对其进行平均。

  - 方差的平方根即为标准差，两个指标均能较好地反映出数值型数据的离散程度。



数据分布基本指标-离散程度

- 方差
  - 对于使用简单平均数作为数据中心的未分组数据数据，$x_1, x_2, x_3,$ …, x_n，总体方差为:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

  - 对于使用加权平均数作为数据中心的分组数据，该组数据的总体方差为：

$$\sigma^2 = \frac{\sum_{i=1}^{k}(M_i-\mu)^2 f_i}{N}$$

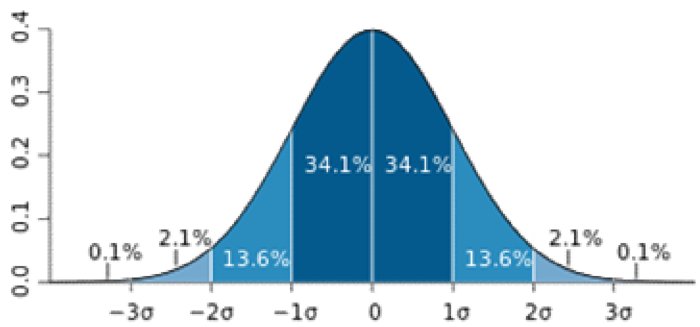
数据分布基本指标-离散程度

- 标准差
  - 标准差为方差的算数平方根，是具有量纲(与原数据有相同单位)的。
  - 它与变量值的计量单位相同，实际意义比方差更清楚。

  - 对于未分组数据和加权的分组数据来说，其标准差的计算公式分别为：

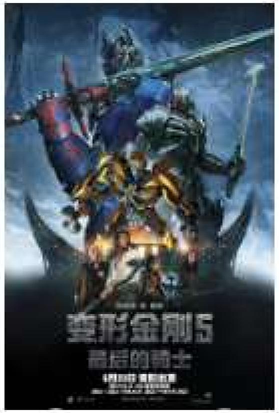
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(M_i-\mu)^2 f_i}{N}}$$

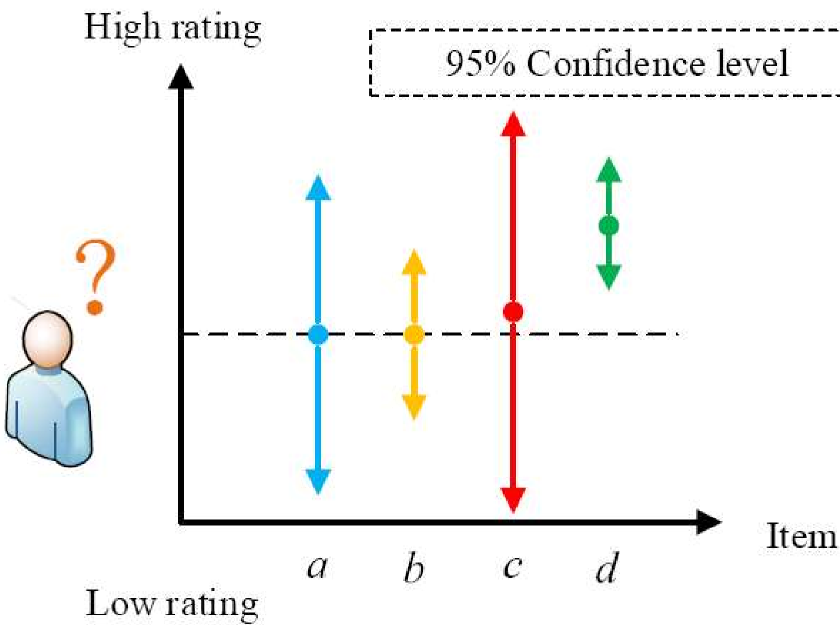


数据分布基本指标-离散程度

- 平均数和方差



虽然用户对电影**b**的评分平均数略低于**c**,但是**b**的评分方差较小



Chao Wang, Qi Liu, Runze Wu, Enhong Chen, Chuanren Liu, Xunpeng Huang, Zhenya Huang, Confidence-aware Matrix Factorization for Recommender Systems, AAAI'2018: 434-442, 2018.

数据分布基本指标-离散程度

- 极差和四分位差
  - 在顺序数据中，当中位数作为数据中心位置的指标时，一般可用极差或四分位差反映数据的离散程度。
  - 极差：
    - 一组数据的最大值和最小值之差被称为极差(range)，也被称为全矩，用R表示，是描述数据离散程度的最简单的测度值。
    - 若一组数据中的最大值为max(x_i)，最小值为min(x_i)，则该组数据的极差R为：

$$R = max(x_i) - min(x_i)$$

数据分布基本指标-离散程度

- 极差：
  - 一组数据的最大值和最小值之差被称为极差(range)，也被称为全矩，用R表示，是描述数据离散程度的最简单的测度值。
  - 若一组数据中的最大值为max(x_i)，最小值为min(x_i)，则该组数据的极差R为：

$$R = max(x_i) - min(x_i)$$

  - 极差即数据的振幅，振幅越大说明数据越分散，其直观意义非常明显。但由于极差只是利用了一组数据的两端信息，容易受极端值的影响，且不能反映出中间数据的分散状况、准确描述出数据的分散程度。