数据科学导论
Introduction to Data Science

第三章 数据统计基础

黄振亚，陈恩红，刘淇
Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

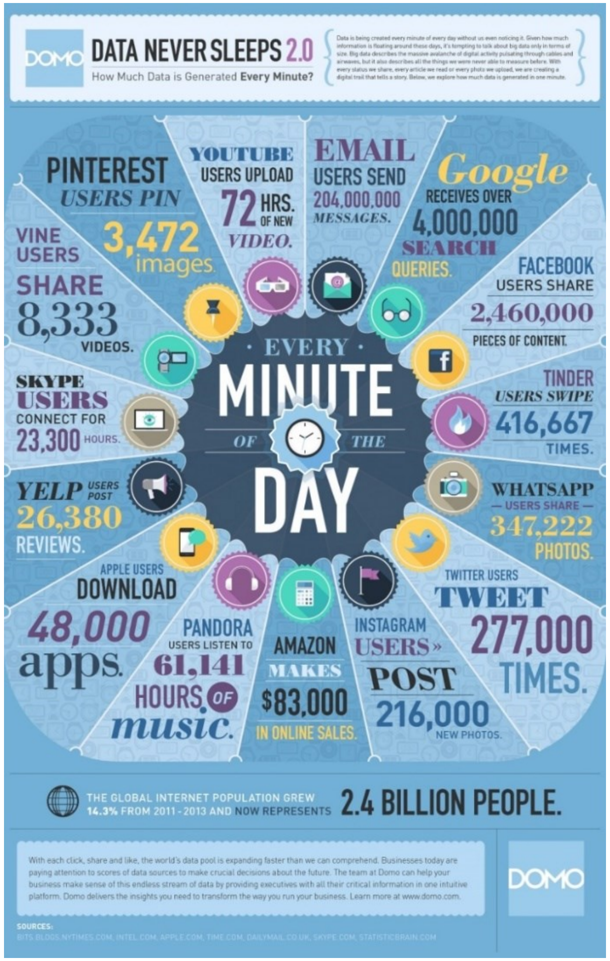
课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html

第二章数据分析基础小结

- 数据采集 Data Collection
  - 信息检索，网络爬虫
- 数据存储 Data Storage
- 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
  - 特征理解

Data

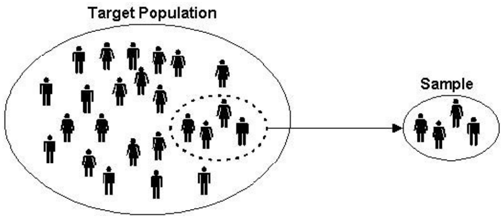
- 大数据
  - 数据量大
  - 类型繁多
  - 时效性高
  - 价值密度低



- 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

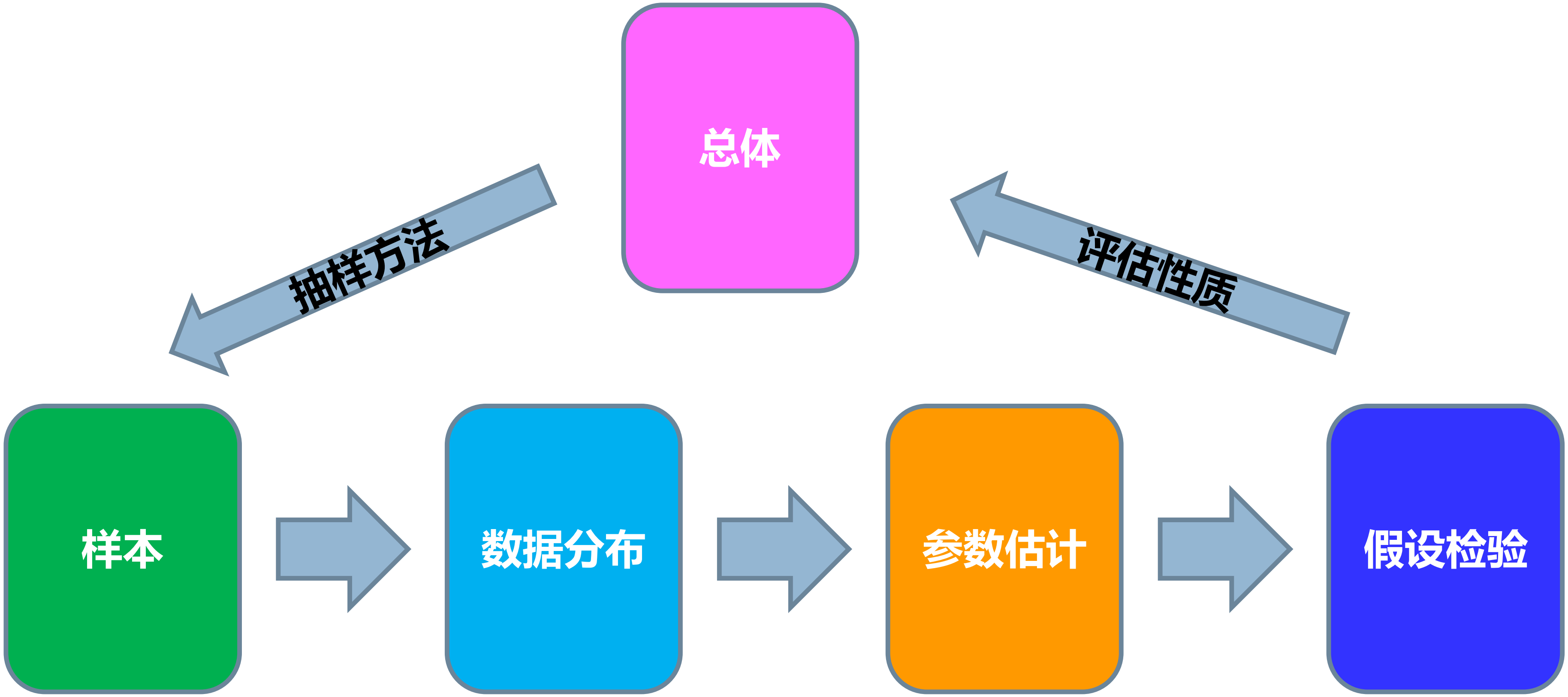
数据统计

- 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)
  - 总体由客观存在且具有同一性质基础的多个个体结合而成
  - 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

数据统计



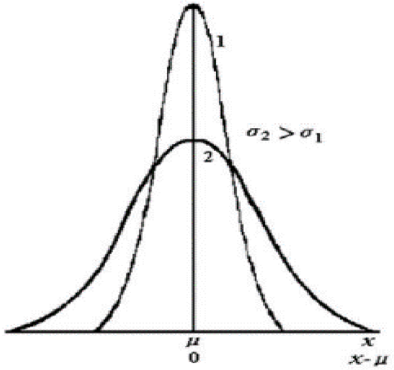
数据统计

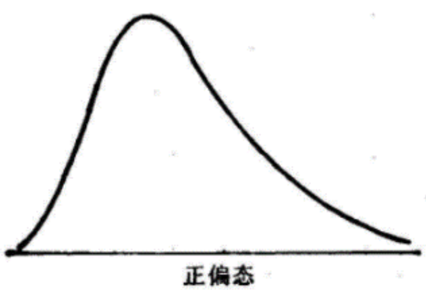
- 数据分布
- 参数估计
- 假设检验
- 抽样方法

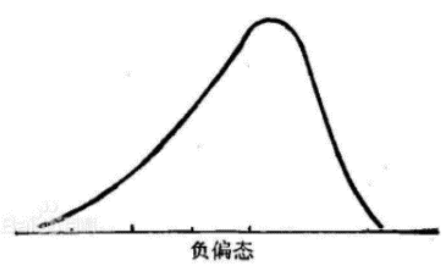
数据分布基本指标

- 在对大数据进行研究时，首先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







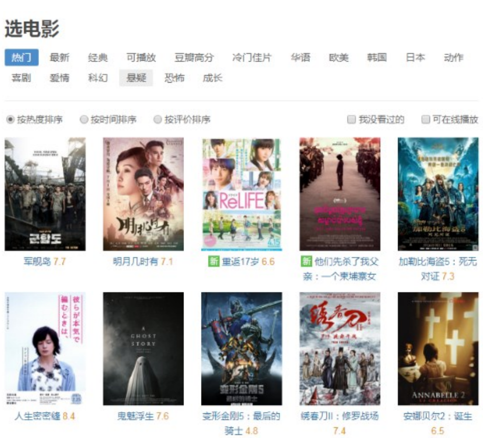
数据分布基本指标

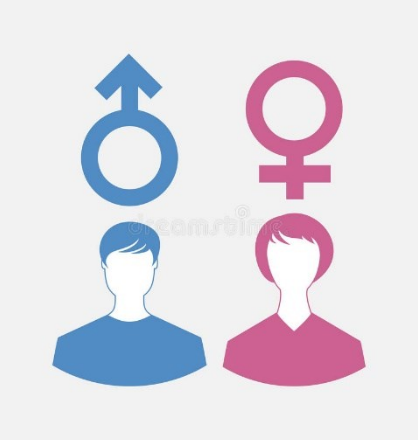
- 集中趋势
  - 集中趋势反映了一组数据的中心点位置所在及该组数据向中心靠拢或聚集的程度。

- 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): 算术平均数
  - 根据未经分组数据计算得到的平均数
  - 若有一组数据：$x_1, x_2, x_3, \cdots, x_n$, 则简单平均数为：

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

  - 特点：易受极端值的影响

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
  - 若有一组n个数据分为K组，各组的值表示为：$x_1, x_2, x_3, \cdots, x_K$,
  - 各组变量出现的频数表示为：$f_1, f_2, f_3, \cdots, f_k$,
  - 则该数据的加权平均数为：

$$\boldsymbol{\mu} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_K f_K}{f_1 + f_2 + f_3 + \cdots + f_K} = \frac{\sum_{i=1}^{K} x_i f_i}{n}$$

  - 特点：
    - 影响因素：组数值，频数
    - 频数越多，该组影响最大

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(geometirc mean)
  - 几何平均数是n个变量值乘积的n次方根
  - 适用范围
    - 平均比率：年利率、合格率等
  - 若一组数据$x_1, x_2, x_3, \cdots, x_n$，则该组数据的几何平均数为

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

  - 若数值为增长率

$$\boldsymbol{G = \sqrt[n]{(1+x_1)\times(1+x_1)\times(1+x_1)\times\cdots\times(1+x_n)} - 1}$$

  - 特点
    - 几何平均数受极端值的影响较算术平均数小
    - 如果变量值有负值，计算出的几何平均数就会成为负数或虚数
    - 几何平均数的对数是各变量值对数的算术平均数

数据分布基本指标-集中趋势

- 算术平均数 vs 几何平均数
- 例：一只股票价格第一年初价格为10元，第一年增长了100%，第二年下降了50%，计算两年平均增长率？

- 算术平均数
- $x = \frac{1-0.5}{2} = 0.25$

- 几何平均数
- $x = \sqrt[2]{(1+1)\times(1-0.5)} - 1 = 0$

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：

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

数据分布基本指标-集中趋势

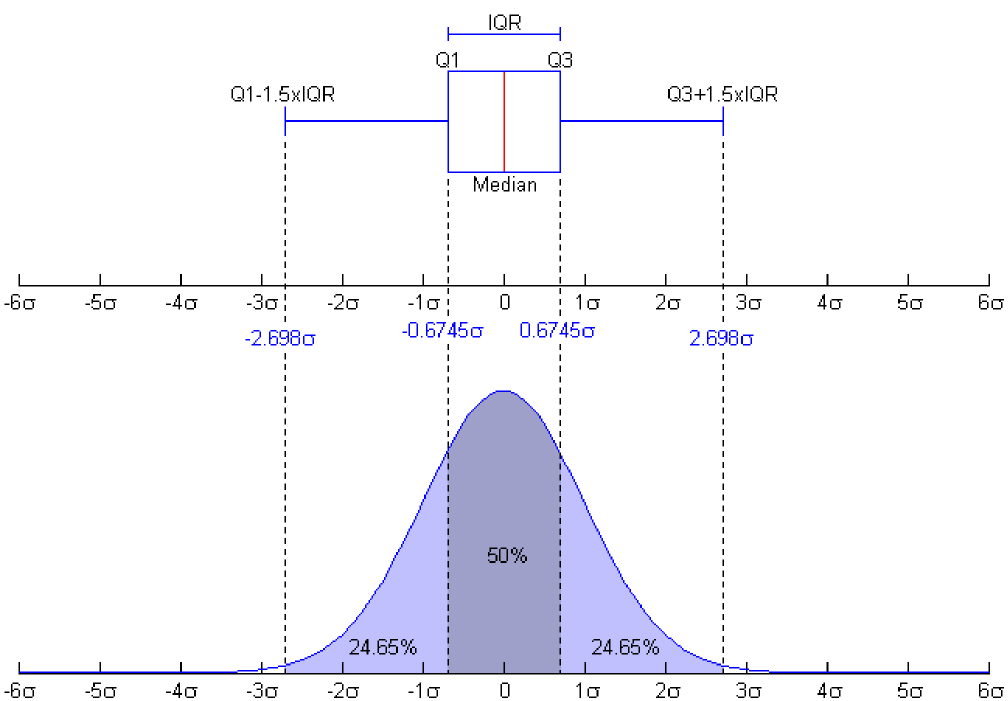
- 分位数
  - 中位数用1 个点将数据两等分
  - 类似的，若用3 个点将数据四等分、9 个点将数据十等分、99 个点将数据一百等分，则对应等分点上的值为四分位数(quartile)、十分位数(decile) 和百分位数(percentile)
  - 四分位数也称四分位点，它通过3 个点将数据等分成四个部分
    - 中间的四分位数就是中位数
    - 下四分位数：处在25% 位置上的数值，第一四分位数
    - 上四分位数：处在75% 位置上的数值，第三四分位数
    - 四分位距IQR：Q3-Q1

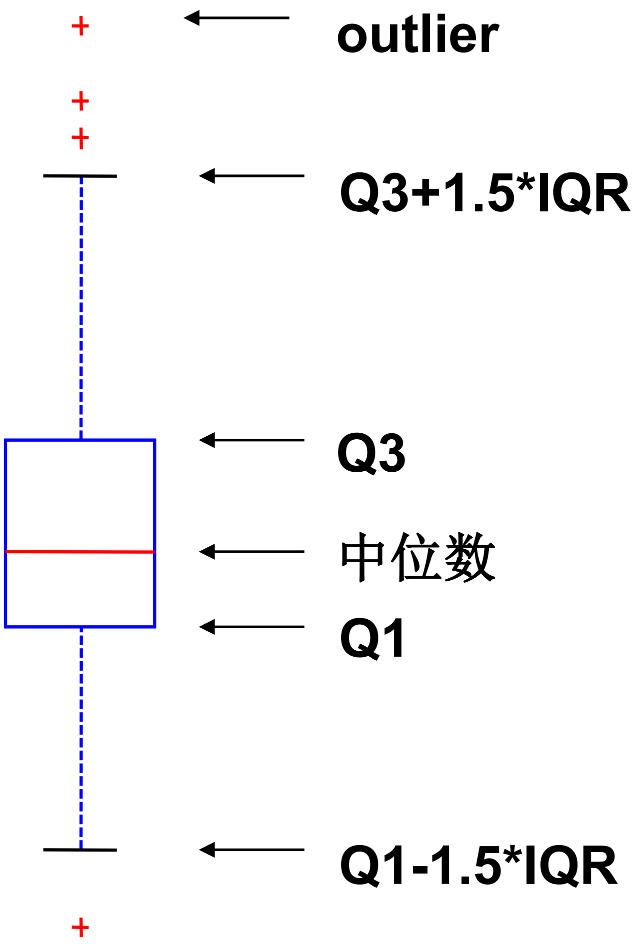
Q1 Q2 Q3

25%	25%	25%	25%

数据分布基本指标-集中趋势

- 箱图（Box Plots）
  - Invented by J. Tukey
  - 显示数据分布





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

数据分布基本指标-集中趋势

- 箱图（Box Plots）
  —研究应用
  - 相对稳定的方式描述数据分布
  - 不受异常值影响，识别了异常值

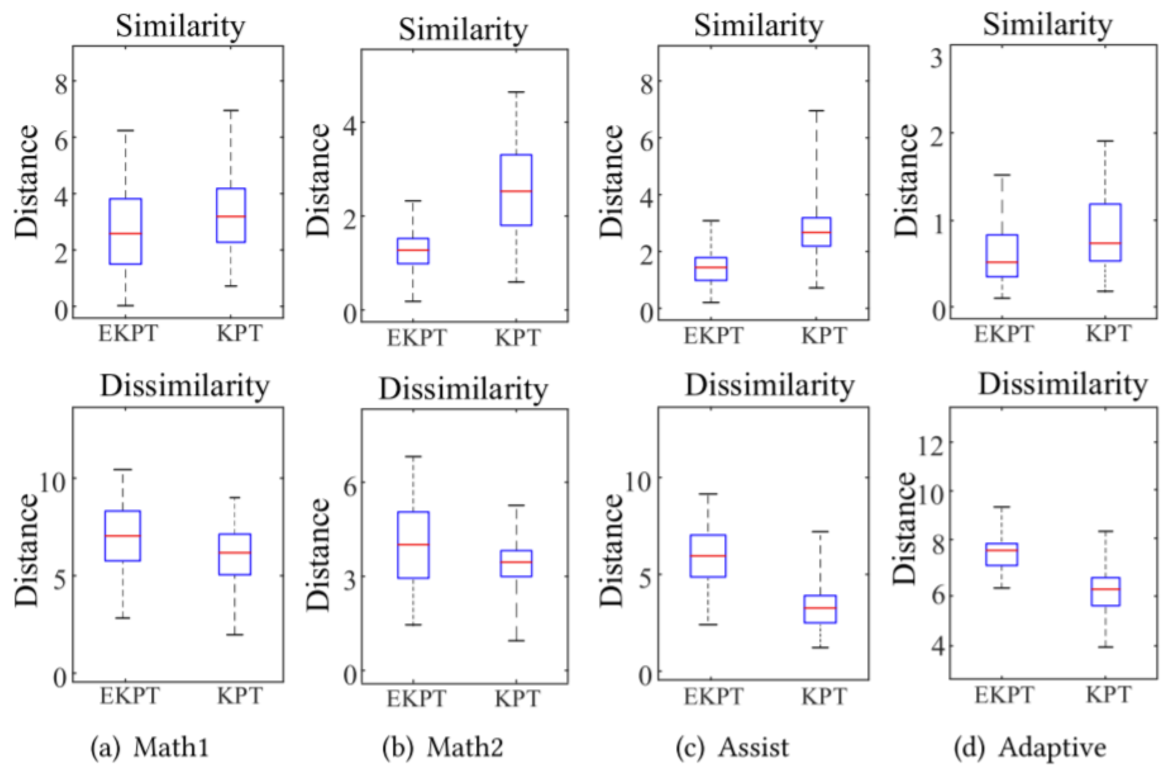


Fig. 11. Results comparison of exercise relationship with EKPT and KPT in all datasets.

✓ **Zhenya Huang, Qi Liu, Le Wu, Keli Xiao, Enhong Chen, Learning or Forgetting? A Dynamic Approach for Tracking the Knowledge Proficiency of Students, (ACM TOIS)**

数据分布基本指标-集中趋势

- 箱图（Box Plots）
  —研究应用
  - 相对稳定的方式描述数据分布
  - 不受异常值影响，识别了异常值

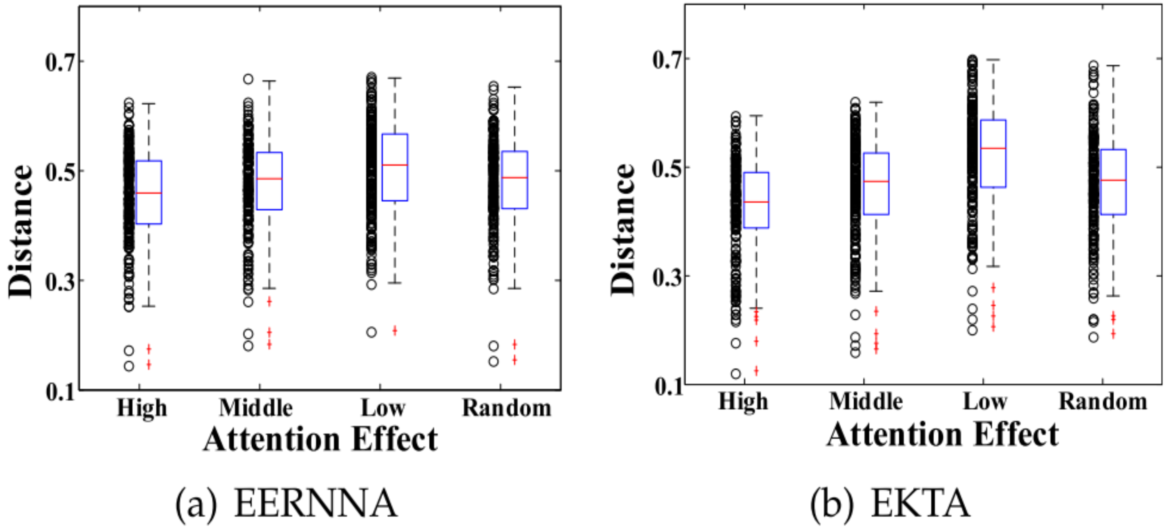


Fig. 12. Performance over different attention values in proposed models.

✓ **Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, Guoping Hu, EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction, IEEE TKDE**

课后实践—案例学习

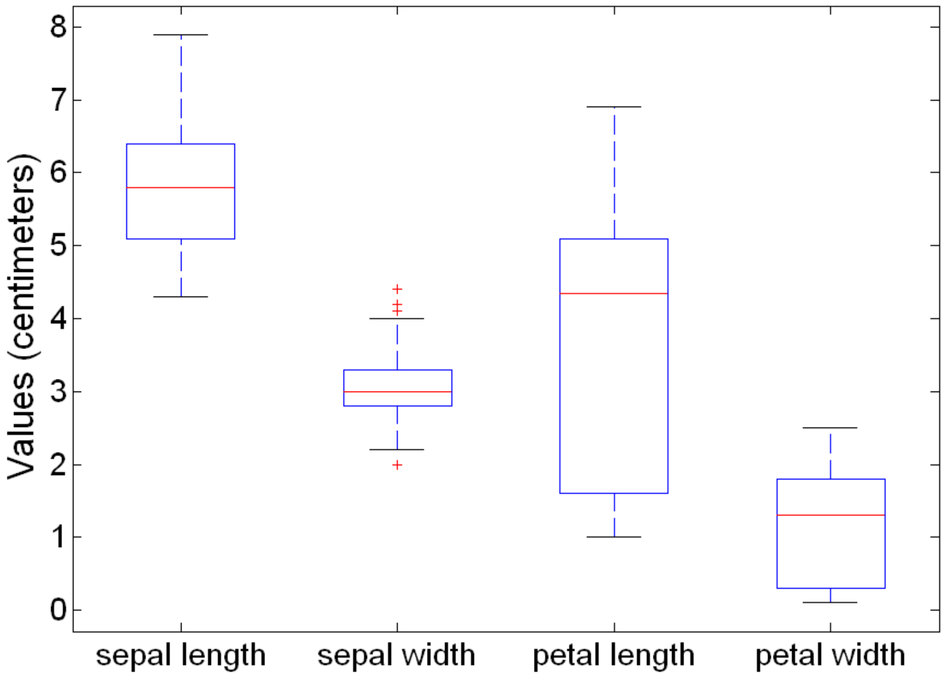
- IRIS(鸢尾花) + sklearn数据分析案例
  - http://www.cnblogs.com/jasonfreak/p/5448385.html
  - 1. 数据集的描述与导入

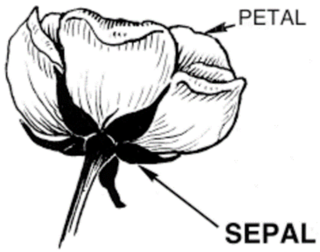
数据的特征:	花萼长度	花萼宽度	花瓣长度	花瓣宽度
花的类别:	山鸢尾	杂色鸢尾	维吉尼亚鸢尾	

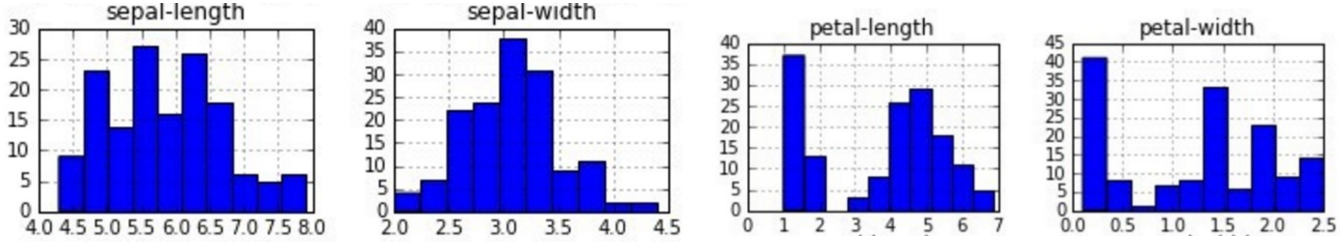
```python
from sklearn.datasets import load_iris

#导入数据集IRIS
iris = load_iris()

#特征矩阵
iris.data

#目标向量
iris.target
```

课后实践-集中趋势

- 分位数
  - IRIS(鸢尾花)



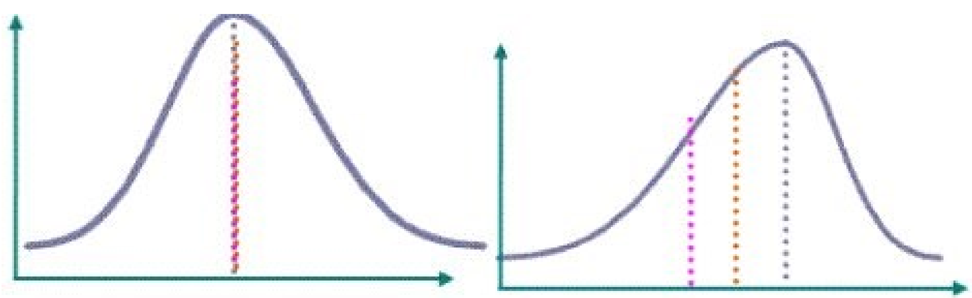




http://archive.ics.uci.edu/ml/datasets/Iris/

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。



$\bar{x} = Me = Mo$　　$\bar{x} < Me < Mo$

均值 中位数 众数

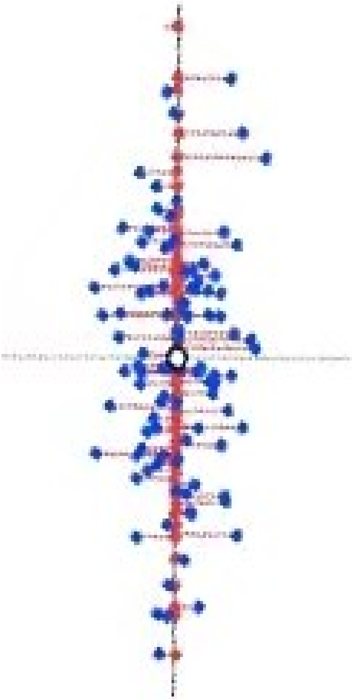
数据分布基本指标-离散程度

- 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)
  - **衡量平均数对数据的代表性**
  - 方差是各个变量与均值之差平方的平均数
  - 标准差为方差的平方根，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
  - 对于未分组数据$x_1, x_2, x_3, \cdots, x_N$，数据的算术平均数为$\mu$。数据的总体方差为

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

  - 对于已分为K组的N个数据，各组的值表示为：$x_1, x_2, x_3, \cdots, x_K$, 各组变量出现的频数表示为：$f_1, f_2, f_3, \cdots, f_k$, 数据的加权平均数为$\mu$，则数据的总体方差为

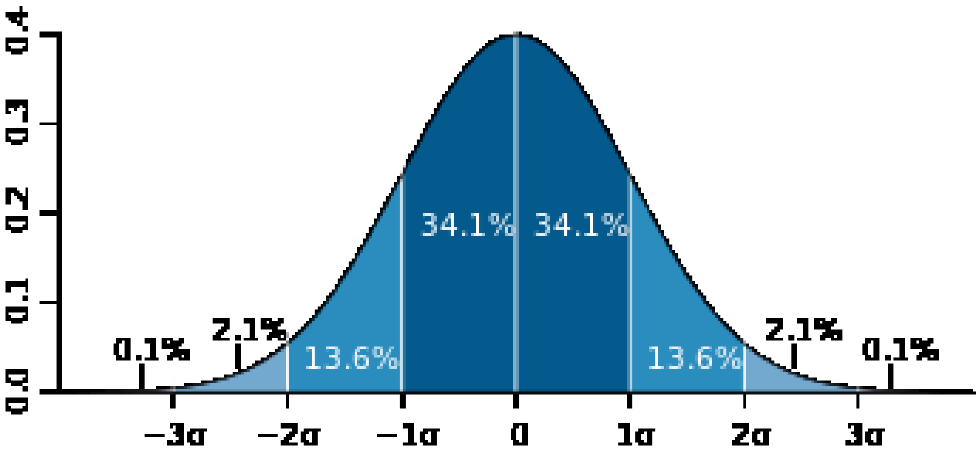
$$\sigma^2 = \frac{\sum_{i=1}^{K}(x_i - \mu)^2 f_i}{N}$$

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，具有量纲(与原数据有相同单位)
  - 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（K组）来说，其标准差的计算公式分别为：

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

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

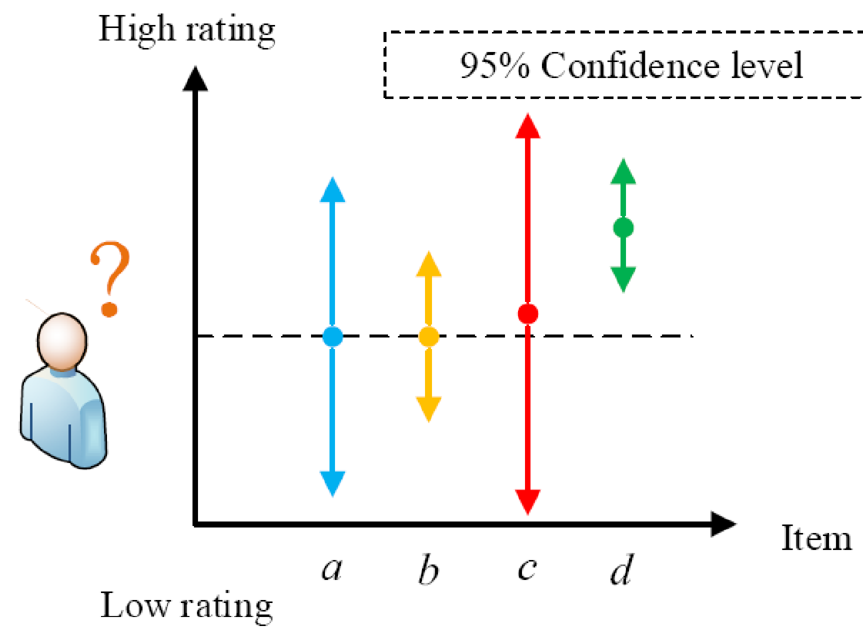


数据分布基本指标-离散程度

平均数和方差—研究应用

参数估计—区间估计

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, 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，当中位数为数据中心位置的指标时，可以用极差或者四分位差反映数据的离散程度
  - **衡量中位数对数据的代表性**

- 极差
  - 一组数据的最大值和最小值之差为极差(range)，也被称为全矩(R), 描述数据离散程度的最简单的测度值
  - 一组数据$x_1, x_2, x_3, \cdots, x_N$，则该组数据的极差为

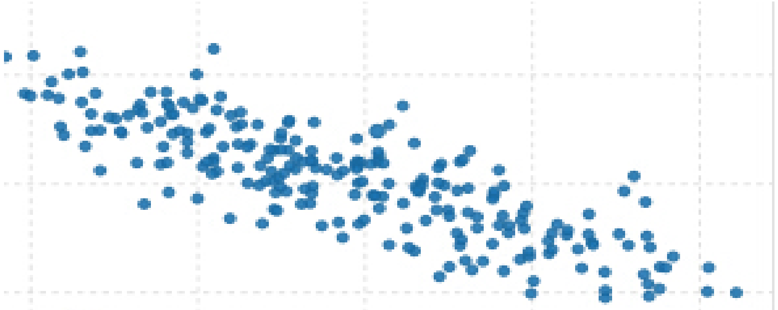
$$R = max(x_i) - min(x_i)$$

数据分布基本指标-离散程度

- 极差
  - 一组数据$x_1, x_2, x_3, \cdots, x_N$，则该组数据的极差为

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

  - 特点
    - 极差是数据的振幅，振幅越大表示数据越分散
    - 极差只利用了一组数据的两端信息，易受极端值影响。若大部分数据集中在一个较窄的范围，极端值的数据较少，则极差不能准确描述数据的分散程度，即不能反映中间数据的分散程度。

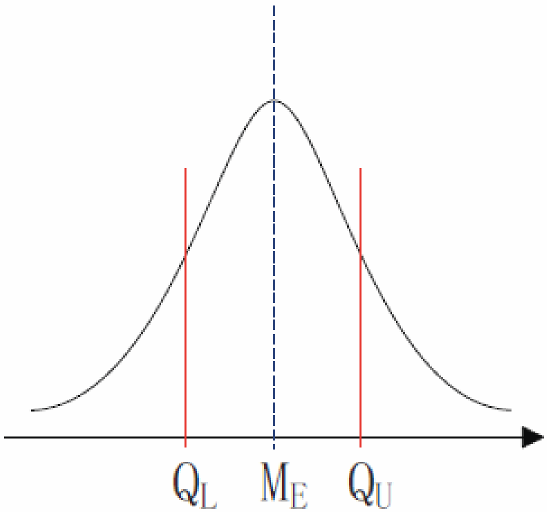


数据分布基本指标-离散程度

- 四分位差
  - 一组数据的上四分位数和下四分位数的差，也称为内矩
  - 若上四分位数为Q_U，下四分位数为Q_L, 则四分位差为

$$Q = Q_U - Q_L$$

  - 特点
    - Q是区间$[Q_L, Q_U]$的长度
    - 区间$[Q_L, Q_U]$含有50%的数据
    - 四分位数不会收到数据中极端值的影响



数据分布基本指标-离散程度

- 异众比率
  - 在以众数作为数据中的的分类数据中，异众比率(variation ratio) 是指非众数组的频数占总频数的比率，用V_r 表示。
  - 主要用于衡量众数对一组数据的代表性程度。
  - 除了对于分类数据外，对于数值型数据和顺序数据也也可以计算其异众比率。计算公式为：

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i}$$

  - 其中，Σf_i 为变量值的总频数，f_m 为众数组的频数。异众比率越大，众数组的频数占总频数比率越小，数据离散程度越高，众数作为其中心的代表性越差。

数据分布基本指标-离散程度

- 变异系数
  - 当需要比较两组数据离散程度大小的时候，如果两组数据的测量尺度相差太大，或者数据量纲的不同，直接使用标准差来进行比较不合适，此时就应当消除测量尺度和量纲的影响。
  - 变异系数（Coefficient of Variation）是原始数据标准差与原始数据平均数的比。计算公式为：

$$c_v = \frac{\sigma}{\mu}$$

  - 在进行数据统计分析时，如果变异系数大于15%，则要考虑该数据可能不正常。

反映单位均值上的离散程度，常用在两个总体均值不等的离散程度的比较上

缺陷：当平均值接近于0的时候，微小的扰动也会对变异系数产生巨大影响，因此造成精确度不足。变异系数无法发展出类似于均值的置信区间的工具。

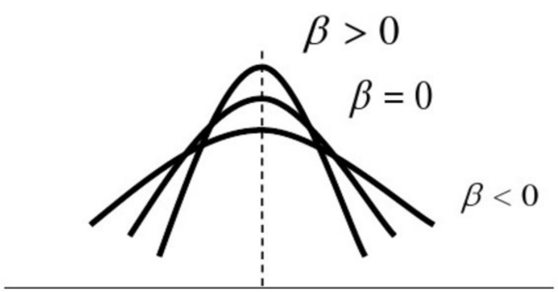
数据分布基本指标-形状变化

- 数据分布形态
  - 数据分布形态反映了一组数据分布的整体形状信息。
- 两种最常用的反映数据形状变化的指标：
  - 峰度
  - 偏度

数据分布基本指标-分布形态

- 峰度：度量数据在中心聚集程度
  - 峰度（Kurtosis）是描述总体中所有取值分布形态陡峭程度 or 平坦程度
  - 峰度的具体计算公式为：
  - 正态分布的峰度值为3
    - 个别软件将峰度值减3, 如：SPSS等
  - 与正态分布相比较
    - 峰度=0表示该总体数据分布与正态分布的陡缓程度相同
    - 峰度>0表示该总体数据分布与正态分布相比较为陡峭，为尖顶峰
    - 峰度<0表示该总体数据分布与正态分布相比较为平坦，为平顶峰

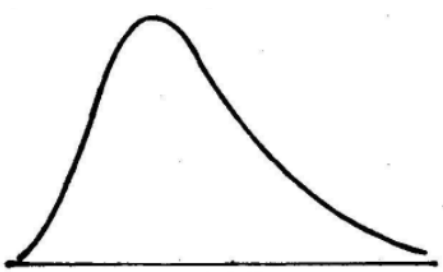
$$K=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right)^2}$$



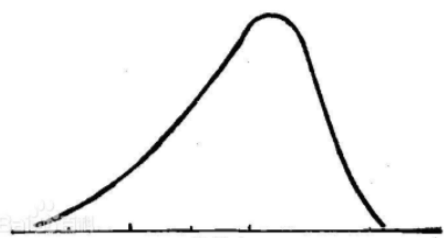
数据分布基本指标-分布形态

- 偏度
  - 偏度（Skewness）描述的是某总体取值分布的对称性
  - 偏度的具体计算公式为：
  - 正态分布的偏度值为0
  - 某个总体
    - 偏度=0表示数据分布形态与正态分布的偏斜程度相同
    - 偏度>0表示数据分布形态与正态分布相比为正偏或右偏，即有一条长尾巴拖在右边，数据右端有较多的极端值
    - 偏度<0表示数据分布形态与正态分布相比为负偏或左偏，即有一条长尾拖在左边，数据左端有较多的极端值。

$$SK=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{3}{2}}}$$



正偏态



负偏态

数据分布—建模应用

- 利用数据指标指导建模思路
  - 若均值与中位数接近，且偏度接近0，可知数据分布是近似对称的，建模时可考虑运用对称信息。
  - 若极差或四分位差较大，建模时需考虑数据是否有长尾现象

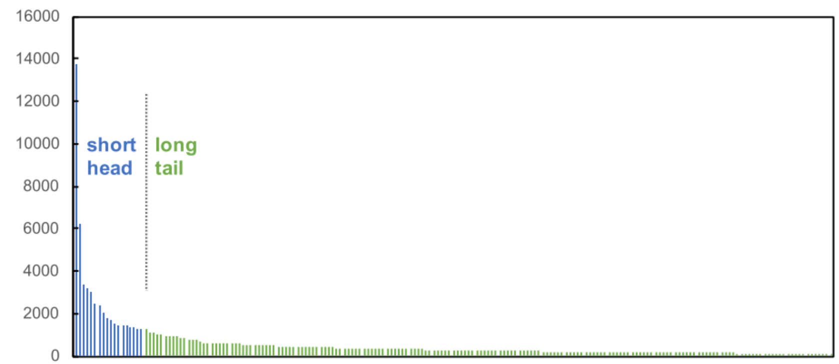
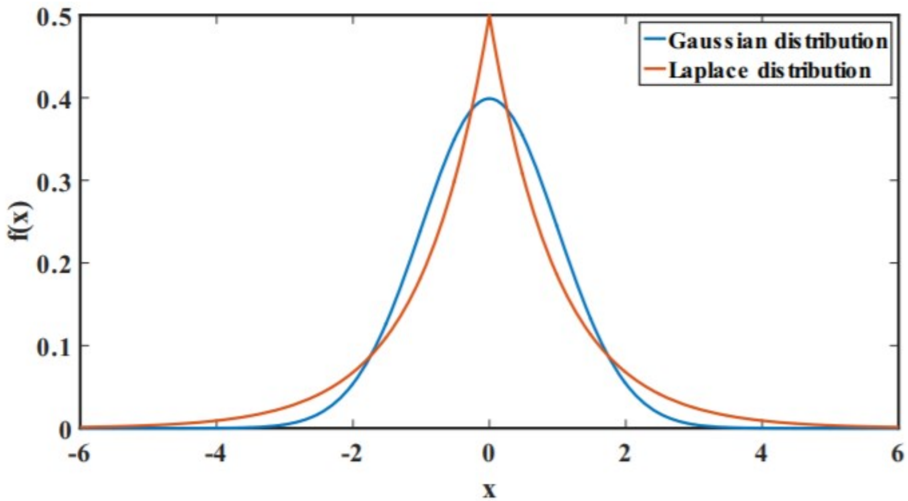


Fig. 1. The popularity of different items in which each item is presented in the Flickr dataset [25]. The horizontal axis denotes the index of items, and the vertical axis indicates the frequency of being interested.

✓ **Li, Jingjing, et al. On both cold-start and long-tail recommendation with social data. TKDE 2019**

数据分布—建模应用

- 利用数据指标指导建模思路
  - 峰度的应用
    - 正态分布
    - 拉普拉斯分布：更好的拟合0出现概率较大的稀疏数据



概率密度函数：

高斯分布 $$p(y)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Lapalace $$p(x)=\frac{1}{2\lambda}e^{-\frac{|x-\mu|}{\lambda}}$$

✓ Zhen Pan, Enhong Chen, Qi Liu, Tong Xu, Haiping Ma, Hongjie Lin. Sparse Factorization Machines for Click-Through Rate Prediction, ICDM'2016

数据分布—建模应用

- 利用数据指标指导建模思路
  - 泊松分布：
    - 基于位置社交网络(LBSN)的推荐系统(POI recommendation)
  - 幂律分布：对数空间下呈现出线性关系(80-20法则)
    - 例如：社交网络(Social Network), 图网络分析

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$$

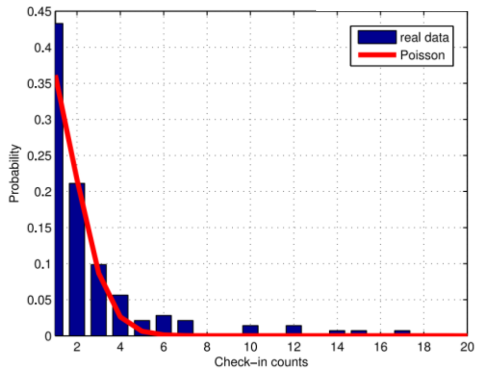
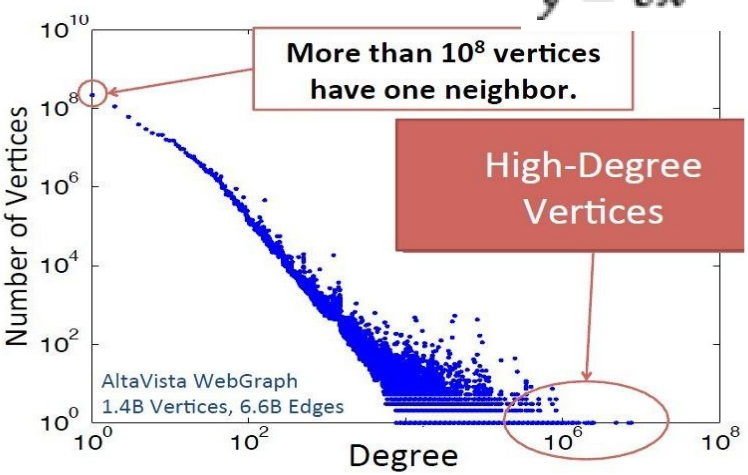


Fig. 3. The check-in counts distribution of a randomly selected user and a Poisson approximation of this distribution (Foursquare dataset).

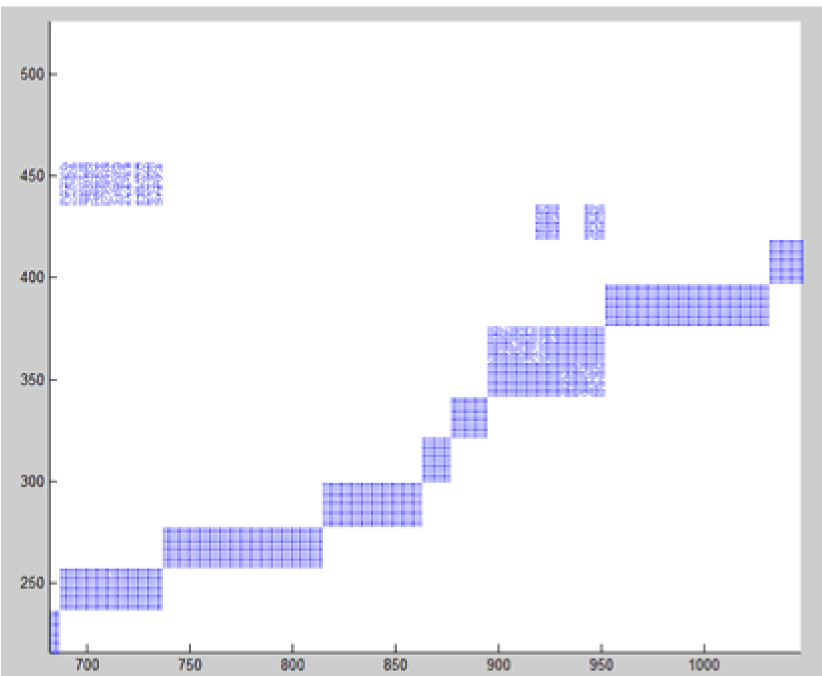
$$y=cx^{-r}$$

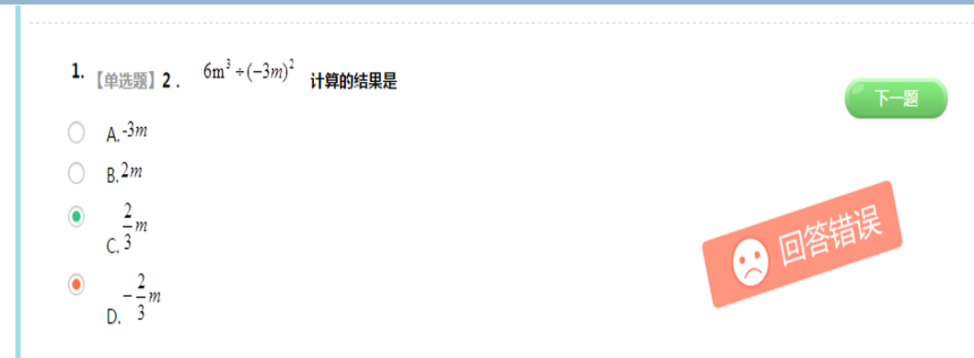


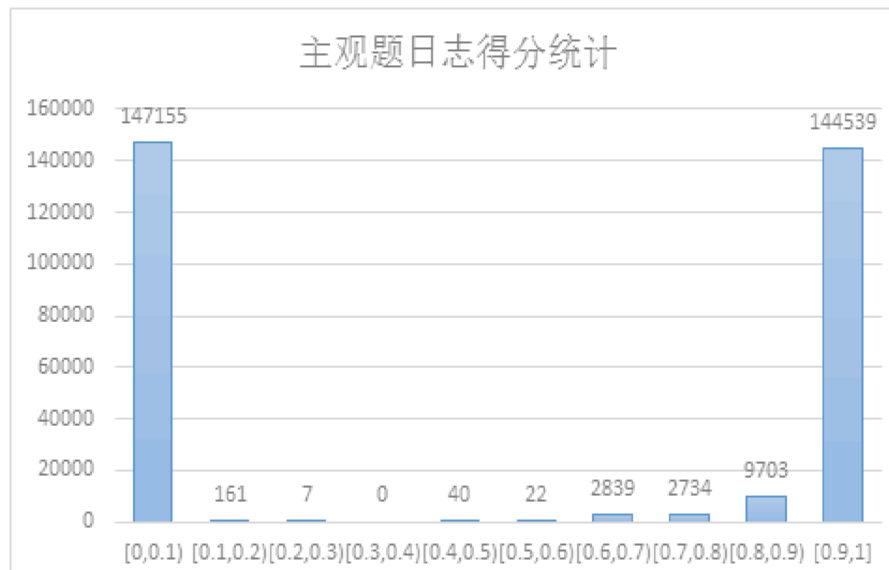
✓ Liu, Bin, et al. A general geographical probabilistic factor model for point of interest recommendation. TKDE 2014.
✓ Perozzi, Bryan, et al. Deepwalk: Online learning of social representations. KDD 2014.

数据分布—建模应用

- 其它指标和现象观察
  - 教育数据







- ✓ Qi Liu, Enhong ChenFuzzy Cognitive Diagnosis for Modelling Examinee Performance. ACM TIST
- ✓ Jinze Wu, Zhenya Huang, Qi Liu, Enhong Chen, Federated Deep Knowledge Tracing, WSDM'2021

数据分布—建模应用

- 以旅游套餐数据为例

(Tour style-Culture & History, Wildlife & Nature), 8 days, from $1260.00
This eastern travel experiences the biggest, boldest and brightest of American destinations. From New York City, Niagara to Cambridge and Washington DC. Experience American life in full and gain perspective among giant monuments, stunning skyscrapers, fascinating history and spectacular natural wonders. Day 1 New York: Enter a neon jungle at Times square, find a quiet corner in Central Park or watch the sunset from atop the Empire State Building. Days 2-3 Washington DC: See all the big names - the White House, the Lincoln Memorial, Washington Monument and Capitol Hill. Day 4 Finger Lakes: Finger Lakes, go swimming or hiking. Day 5 Niagara Falls: Niagara Falls is a favorite for lovers and lovers of nature alike. Days 6-7 Boston: Retrace the nation's revolutionary past by walking the Freedom Trail, or visit bustling North End for Italian feasts. Day 8 New York: Continue to buzzing New York and travel to Coney Island, the Met or see a Broadway show.
Accommodation: Multishare hostels/cabins. Size: 13 travelers per group. What to budget: Allow USD $160 for meals not included.........

Figure 1. An example of the travel package document, where the landscapes are represented by the words in red.

✓ **Qi Liu, Enhong Chen, Hui Xiong, Yong Ge, Zhongmou Li, Xiang Wu, A Cocktail Approach for Travel Package Recommendation, TKDE 2014**

数据分布—建模应用

- 以旅游套餐数据为例

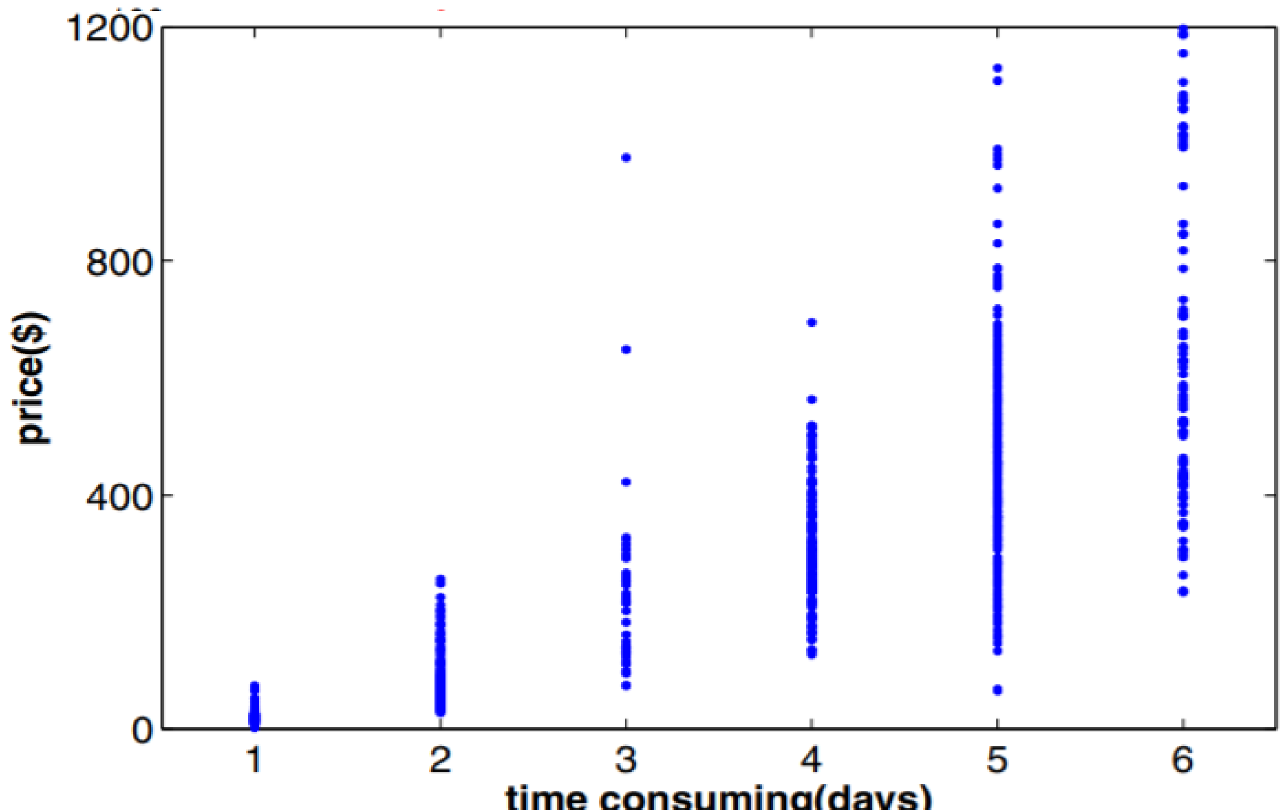


Figure 4. The relationship between the time cost and the financial cost in travel packages.

Data

- 数据分布基本指标
- 参数估计
- 假设检验
- 抽样方法

参数估计

- 参数(parameter)
  - 参数 是用来描述**总体数据特征**的度量
- 统计量(statistic)
  - 统计量 是用来描述**样本数据特征**的度量
    - 由试验计算得出，不依赖于任何其他未知的量（特别是不能依赖于总体分布中所包含的未知参数）
- 参数估计(parameter estimation)
  - 是统计推断的基本问题之一：用**样本统计量**估计总体的**参数**
    - 参数未知的真实
    - 统计量已知的估计
  - 例：掷骰子例子

参数估计

- 参数估计
  - **点估计：**用样本统计量$\hat{\theta}$的某个取值直接作为总体参数θ的估计值
    - 简单来说，直接以样本指标来估计总体指标
    - 总体的某个特征值，如数学期望、方差和相关系数等
  - 区间估计：从总体中抽取的样本，根据一定的正确度与精确度的要求，构造出适当的区间，以作为总体的分布参数(或参数的函数)的真值所在范围的估计
    - 用数轴上的一段经历或一个数据区间，表示总体参数的可能范围。这一段距离或数据区间称为区间估计的置信区间

参数估计

- 点估计(point estimate)
  - 点估计是用样本统计量$\hat{\theta}$的某个取值直接作为总体参数θ的估计值
    - 用样本均值x直接作为总体均值μ的估计值
    - 用样本方差s^2直接作为总体方差σ^2的估计值
- 点估计的常用方法
  - 矩估计
  - 最小二乘估计
  - 极大似然估计
  - 最大后验概率
  - 贝叶斯估计

参数估计—矩估计

- 矩估计
  - 原理：大数定律
  - 矩估计是基于“替换”思想，即用样本矩估计总体矩
    - 均值，方差
  - 随机变量的矩
    - K阶原点矩：$E\left(X^k\right)=\frac{1}{n}\sum_{i=1}^{n}X_i^k$
    - K阶中心距：$E\left([X-E(X)]^k\right)$
      - 一阶原点矩表示期望
      - 二阶中心矩表示方差
      - 三阶中心矩表示偏度
      - 四阶中心矩表示峰度

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

$$SK=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{3}{2}}} \qquad K=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right)^2}$$

课后练习：思考并推导矩估计与数据统计指标的关系

参数估计—矩估计

- 矩估计
  - 矩估计法的基本思想是替换原理，即用样本矩替换同阶总体矩。

设 X_1, X_2, ..., X_n 是来自总体 X 的样本，$X(f,\theta)$, $\theta \in \Theta$，其中 $\theta = \theta_1, \theta_2, \ldots, \theta_k$ 为未知分布参数，Θ 为 k 维欧式空间的一个子集。记 $\mu_i = E(X^i)$ 为总体第 i 阶原点矩，$m_i = \frac{1}{n}\sum_{j=1}^{n} x_j^i$ 为样本第 i 阶原点矩 $(i = 1, 2, \ldots, k)$。替换原理即为，若参数 θ_i 能表示为 $\theta_i = g_i(\mu_1, \mu_2, \ldots, \mu_k)(i = 1, 2, \ldots, k)$，其中 $g_1, g_2, \ldots, g_k$ 为 k 个多源的已知函数，则可用 m_i 替换 $\mu_i(i = 1, 2, \ldots, k)$，得到 $\hat{\theta}_i = g_i(m_1, m_2, \ldots, m_k)$，即为 θ_i 的估计 $(i = 1, 2, \ldots, k)$。

直接把样本当成总体然后求解参数

参数估计—矩估计

问 设总体X在$[a,b]$上服从均匀分布，$X_1,\ldots,X_n$是来自X的样本，试求a,b的矩估计量

解

有两个未知量，故我们需要列出 1 阶矩和 2 阶矩：

$$\begin{cases} \mu_1 = E(X) = (a+b)/2 \\ \mu_2 = E(X^2) = D(X) + [E(X)]^2 = \dfrac{(b-a)^2}{12} + \dfrac{(a+b)^2}{4} \end{cases}$$

解得

$$\begin{cases} a = \mu_1 - \sqrt{3(\mu_2 - \mu_1^2)} \\ b = \mu_1 + \sqrt{3(\mu_2 - \mu_1^2)} \end{cases}$$

由于样本k阶矩是k阶矩的无偏估计量，故用A_1,A_2代替μ_1,μ_2得到a,b的矩估计量为：

$$\begin{cases} A_1 = \dfrac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \\ A_2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases}$$

$$\begin{cases} a = \bar{X} - \sqrt{3\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)} = \bar{X} - \sqrt{\dfrac{3}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ b = \bar{X} + \sqrt{3\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)} = \bar{X} + \sqrt{\dfrac{3}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \end{cases}$$

参数估计—矩估计 vs MSE

- 举例：黑白球（矩估计）
  - 例：假如有一个罐子，里面有黑白两种颜色的球，数目多少不知，两种颜色的比例也不知。每次任意从已经摇匀的罐中拿1个球出来，记录球的颜色，然后把拿出来的球再放回罐中。假如在前面的100次重复记录中，有70次是白球。请问罐中白球所占的比例是多少？

解：用样本中白球比例的均值作为估计代替总体均值。

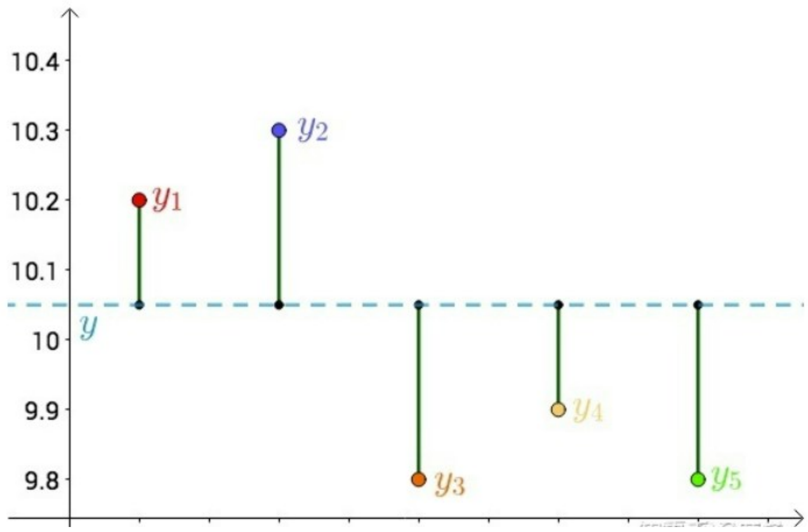
即估计结果为罐中白球所占的比例70% $= \frac{7}{10}$

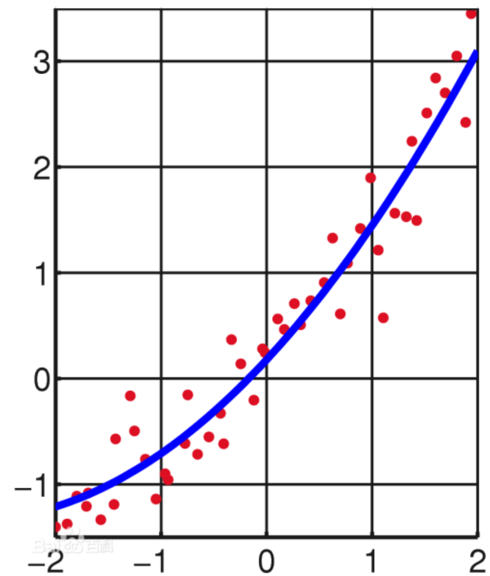
符合直观

(独立同分布，无偏估计)

参数估计—最小二乘估计

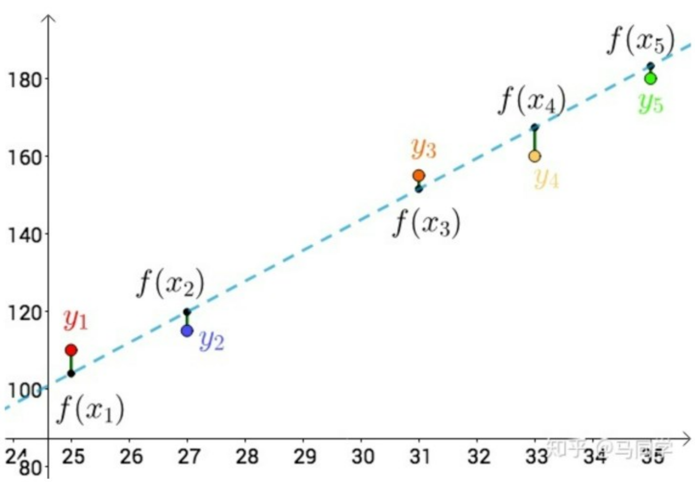
- 最小二乘估计(Least Square Estimate, LSE)
  - 参数估计量应该使得模型能最好地拟合样本数据，即估计值与观测值之差的平方和最小
  - 目标：最小化估计值θ与观测值$\hat{\theta}$之差的平方和
  - $\min L(\theta)=\sum_{i=1}^{N}(\theta-\hat{\theta}_i)^2$





参数估计—最小二乘估计

- 最小二乘估计(LSE)
  - 常用于线性回归分析做参数估计
  - 给定数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，假设模型$f(X|\theta)$
  - 例:在线性回归模型$f(X|\theta)=\theta_0+\theta_1 x+\theta_1 x^2+\cdots+\theta_n x^n$ $=\boldsymbol{\theta}^T\boldsymbol{X}$
  - 目标：$\min L(\theta)=\sum_{i=1}^{N}(f(X|\theta)-Y)^2$
  - 求解：一阶导数为0：$\frac{\partial L(\theta)}{\partial\theta_0}$=0，$\frac{\partial L(\theta)}{\partial\theta_1}$=0，…，$\frac{\partial L(\theta)}{\partial\theta_\mathrm{n}}$=0



课后学习：最小二乘矩阵求解方法

参数估计—最小二乘估计

- 举例：黑白球（最小二乘估计）—课堂练习
  - 问题：假如有一个罐子，里面有黑白两种颜色的球，数目多少不知，两种颜色的比例也不知。每次任意从已经摇匀的罐中拿1个球出来，记录球的颜色，然后把拿出来的球再放回罐中。假如在前面的100次重复记录中，有70次是白球。请问罐中白球所占的比例是多少？
  - **请使用最小二乘估计方法，求解上述问题**

假设：白球占比为$\boldsymbol{\theta}$

目标：最小化估计值θ与观测值$\hat{\theta}$之差的平方和

$$\min L(\theta) = \sum_{i=1}^{N}(\theta - \hat{\theta}_i)^2$$