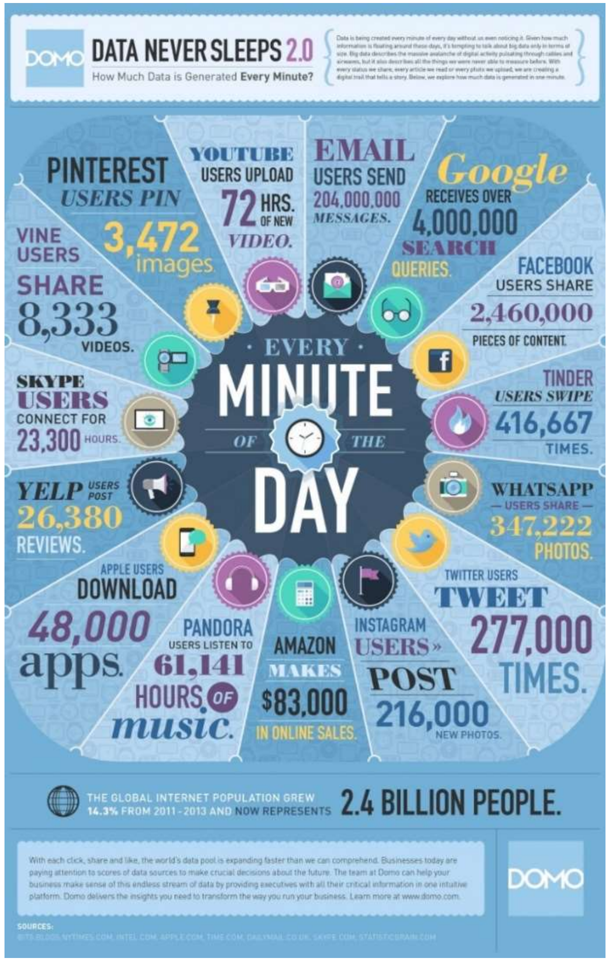
数据科学导论
Introduction to Data Science

第三章 数据统计

常标
Email: qiliuql@ustc.edu.cn
课程主页：
http://staff.ustc.edu.cn/~qiliuql/DS2017.html

Data

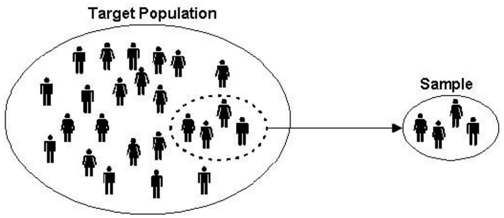
- 大数据
  - 数据量大
  - 类型繁多
  - 时效性高



- 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

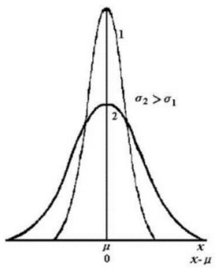
Data

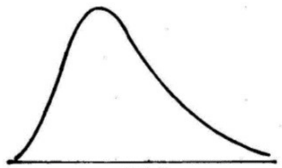
- 数据分布基本指标
- 参数估计
- 假设检验
- 抽样方法

数据分布基本指标

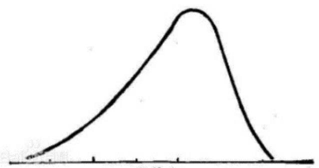
- 在对大数据进行研究时，研究者往往希望知道所获得的数据的基本分布特征

- 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





正偏态



负偏态

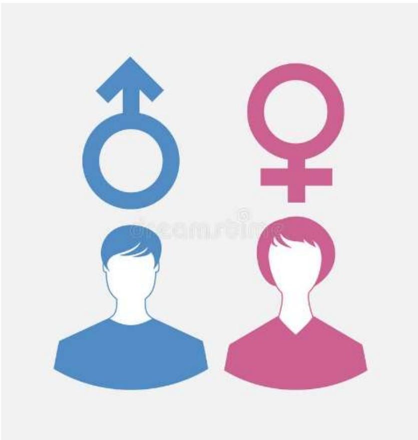
数据分布基本指标

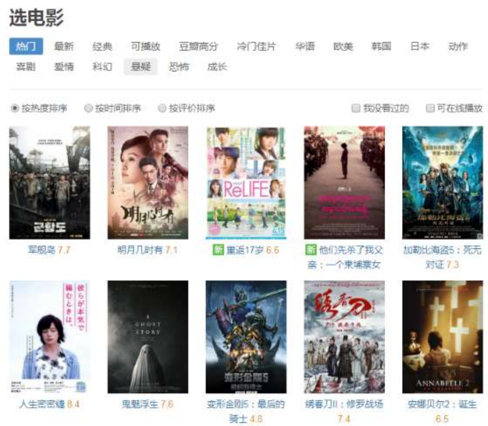
- 集中趋势
  - 集中趋势反映了一组数据的中心点位置所在及该组数据向中心靠拢或聚集的程度。

- 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

- 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

- 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：

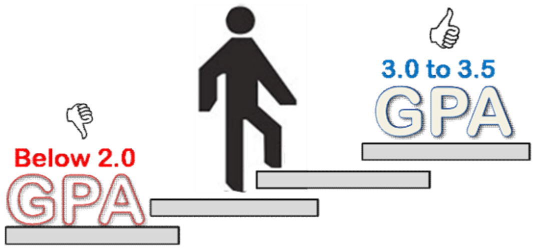
$$\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

数据分布基本指标

- 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

- 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：

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

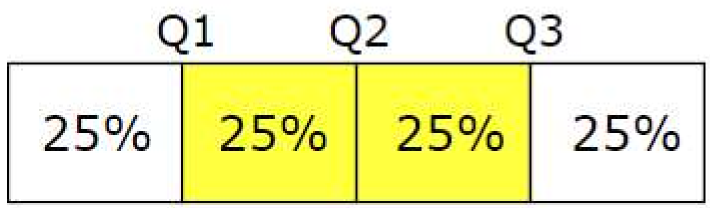
数据分布基本指标

- 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:

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

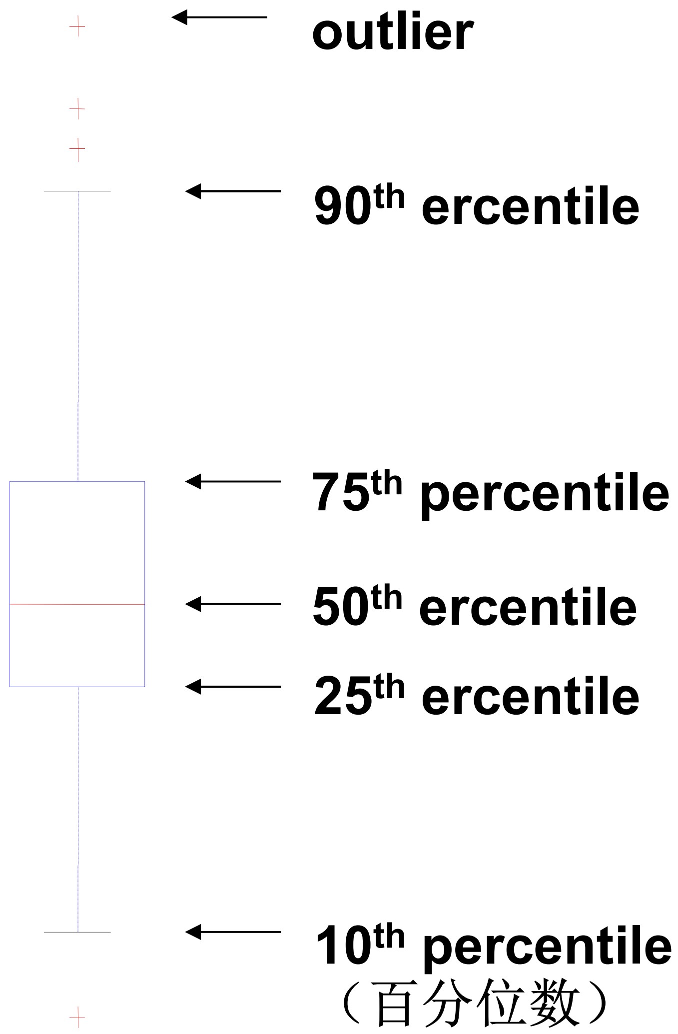
数据分布基本指标

- 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）。



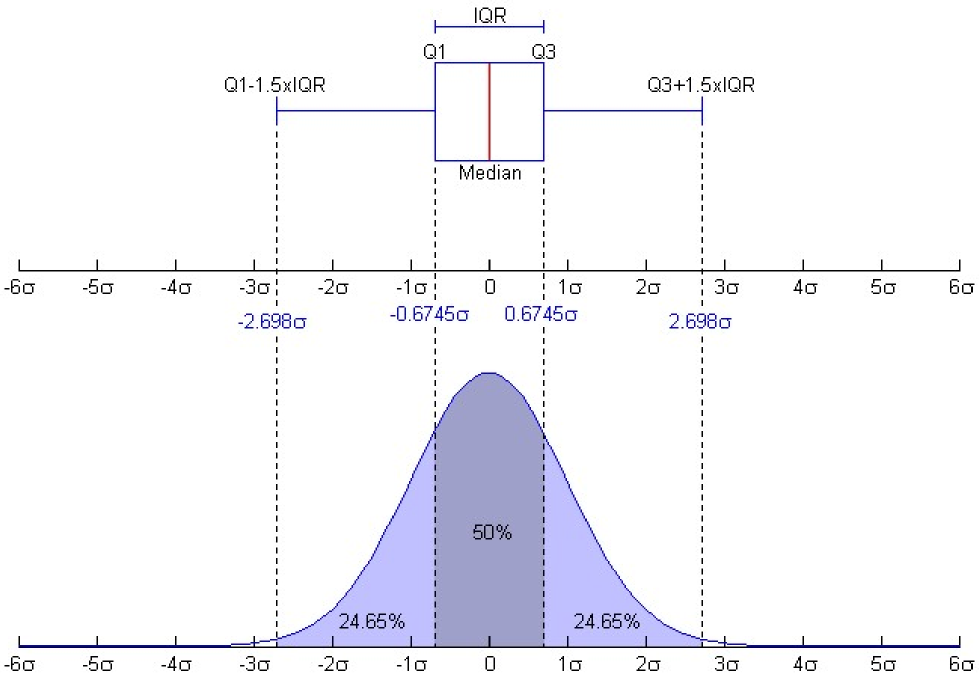
数据分布基本指标

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



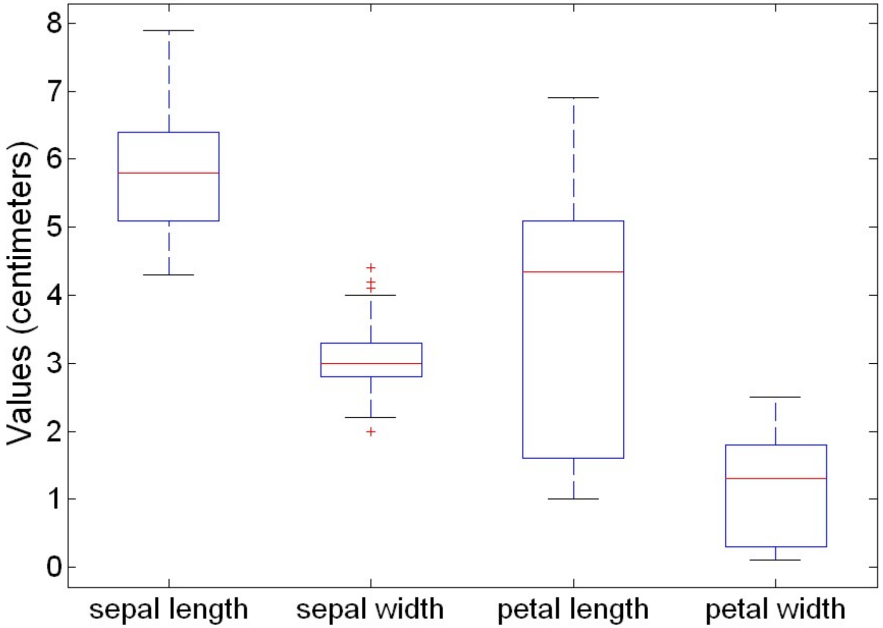
数据分布基本指标

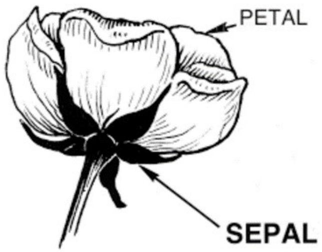
- 分位数
- Box Plots
  - Invented by J. Tukey
  - Another way of display the distribution of data
  - Following figure shows the basic part of a box pl

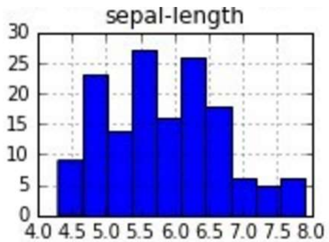


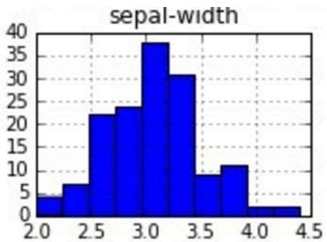
数据分布基本指标

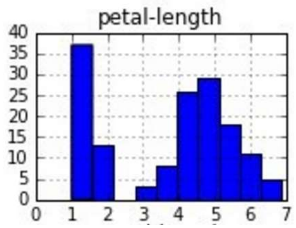
- 分位数

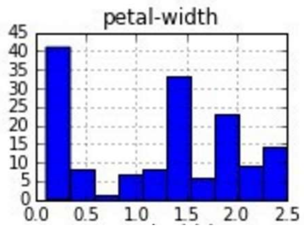






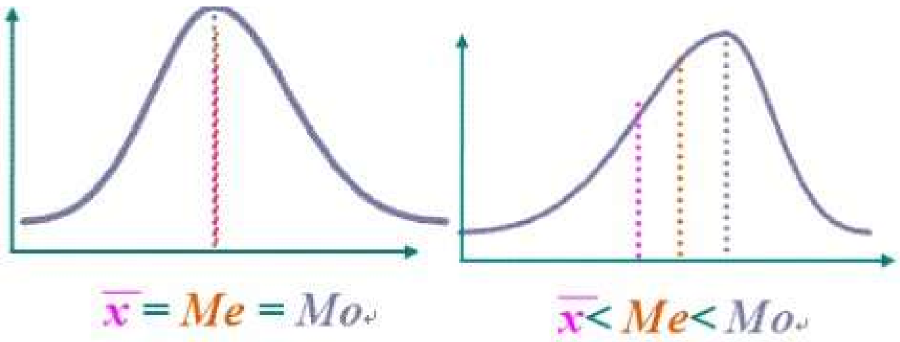






数据分布基本指标

- 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。

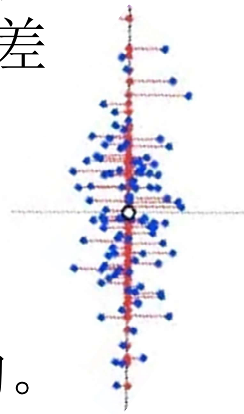


数据分布基本指标

- 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

- 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

- 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，$x_1, x_2, x_3, \ldots, x_n$，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：

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

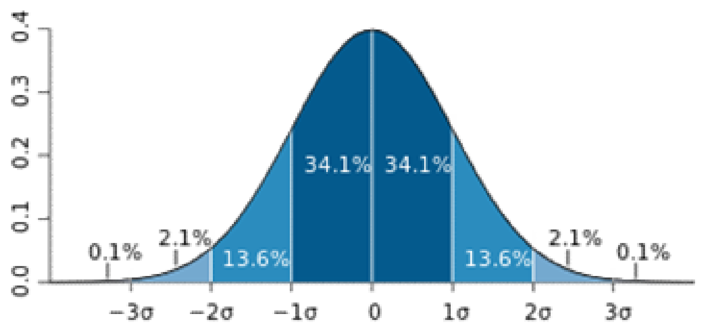
数据分布基本指标

- 标准差
  - 标准差为方差的算数平方根，是具有量纲的。
  - 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：

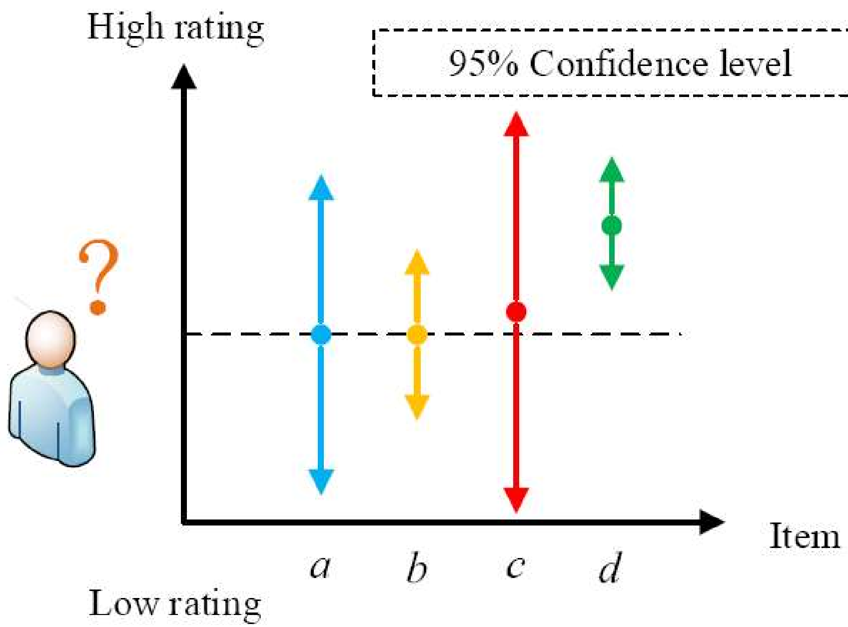
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

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



数据分布基本指标

- 平均数和方差



数据分布基本指标

- 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：

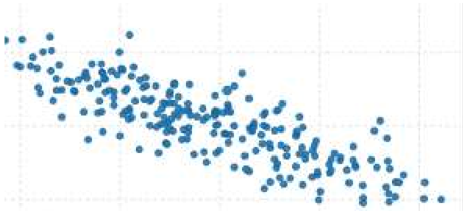
$$R = max(x_i) - min(x_i)$$

数据分布基本指标

- 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。

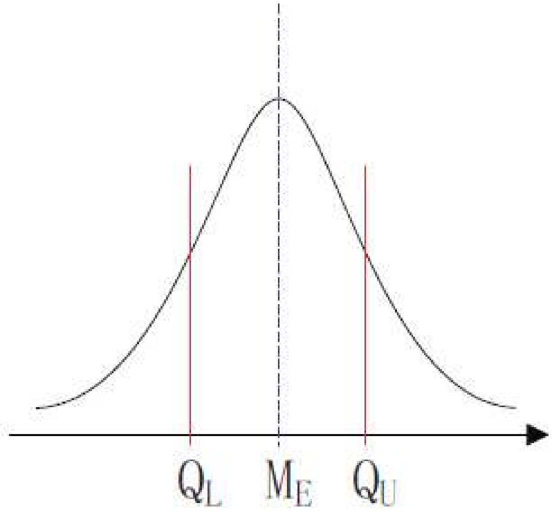


数据分布基本指标

- 四分位差：
  - 一组数据的上四分位数和下四分位数的差值被称为四分位差(quartile deviation)，也被称为内矩，用H表示。
  - 若一组数据的上四分位数为Q_U，下四分位数为Q_L，则该数据的四分位差H为：

$$Q = Q_U - Q_L$$

  - 从定义可以看出，H 是区间(Q_L, Q_U) 的长度。
  - 且区间(Q_L, Q_U)正好含有50%的数据。
  - 不同于极差，四分位差不会受到数据中极端情况的影响。



数据分布基本指标

- 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

- 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。

数据分布基本指标

- 形状变化
  - 形状变化反映了一组数据分布的整体形状信息。

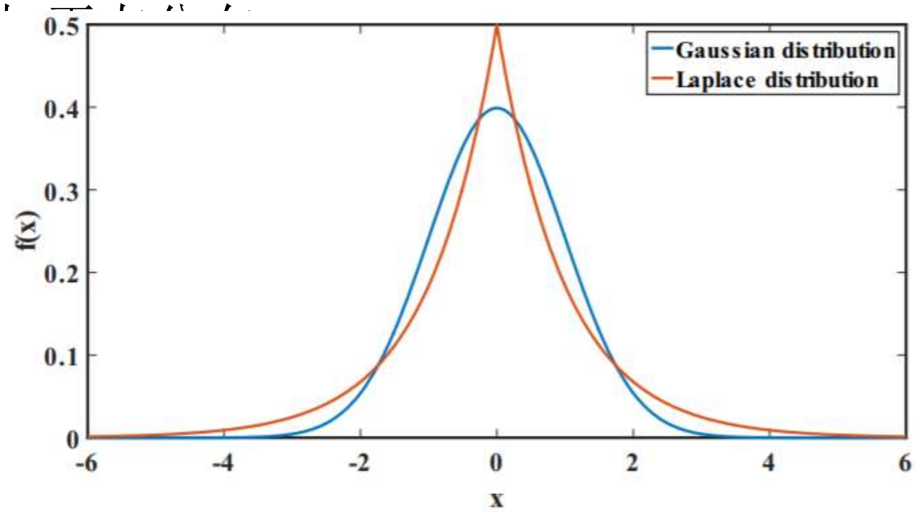
- 两种最常用的反映数据形状变化的指标：
  - 峰度
  - 偏度

数据分布基本指标

- 峰度
  - 峰度（Kurtosis）是描述总体中所有取值分布形态陡缓程度的统计量。
  - 峰度的具体计算公式为：

$$K = \frac{\sum_{i=1}^{k}(x_i - x)^4 f_i}{ns^4}$$

  - 正态分布的峰度值为3
  - 需要与正态分布相比较
    - 峰度为0表示该总体数据分布
      的陡缓程度相同
    - 峰度大于0表示该总体数据分
      布相比较为陡峭，为尖顶峰；
    - 峰度小于0表示该总体数据分
      布相比较为平坦，为平顶峰。

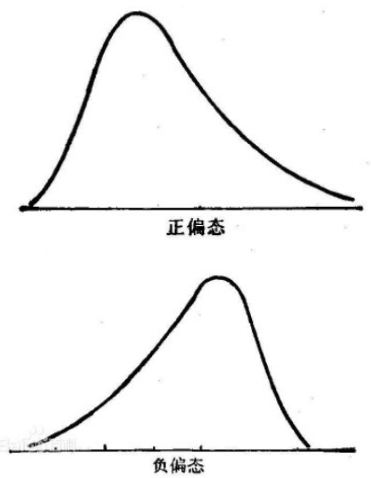


数据分布基本指标

- 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：

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

  - 正态分布的偏度值为0
  - 某个总体
    - 偏度为0表示其数据分布形态与正态分布的偏斜程度相同；
    - 偏度大于0表示其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
    - 偏度小于0表示其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。

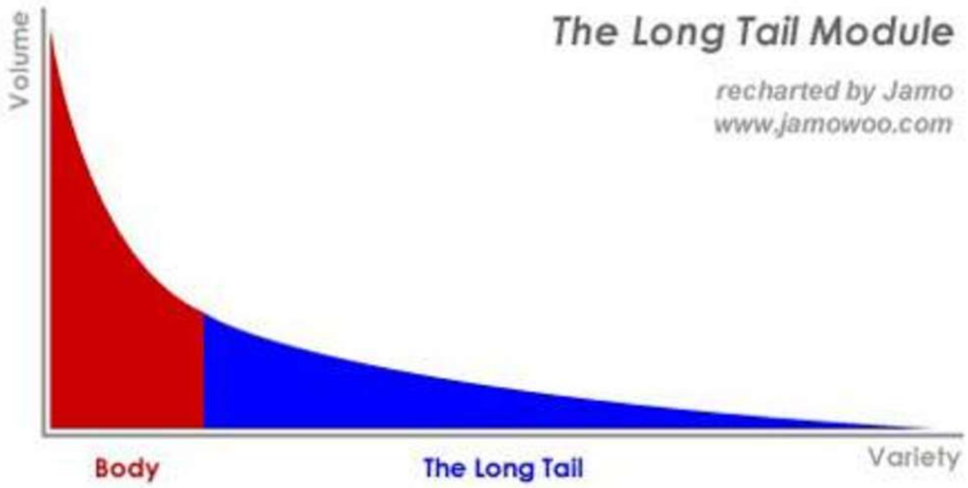


数据分布基本指标

- 数据指标指导建模思路

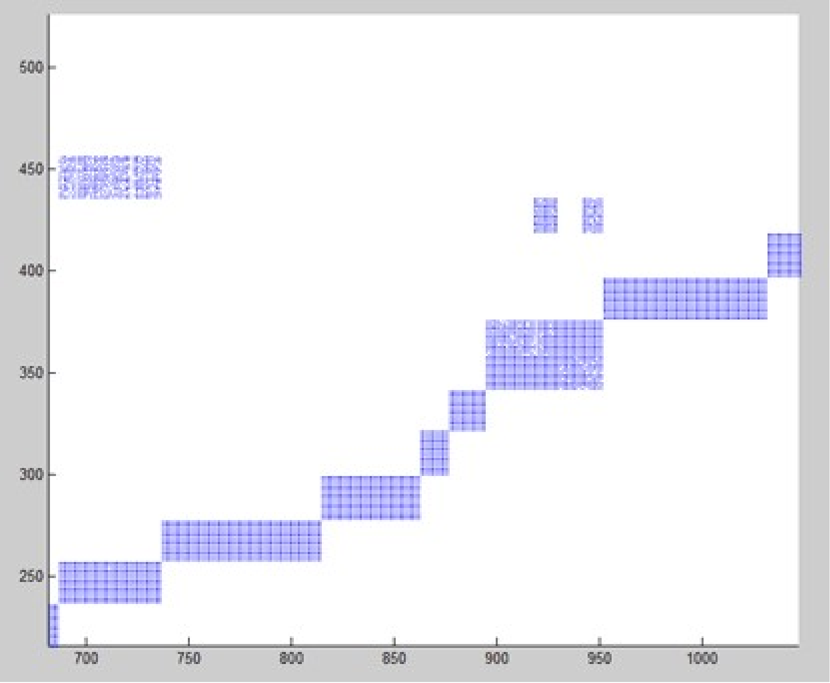
  - 若均值与中位数接近，且偏度接近0，可知数据分布是近似对称的，建模时可考虑运用对称信息。

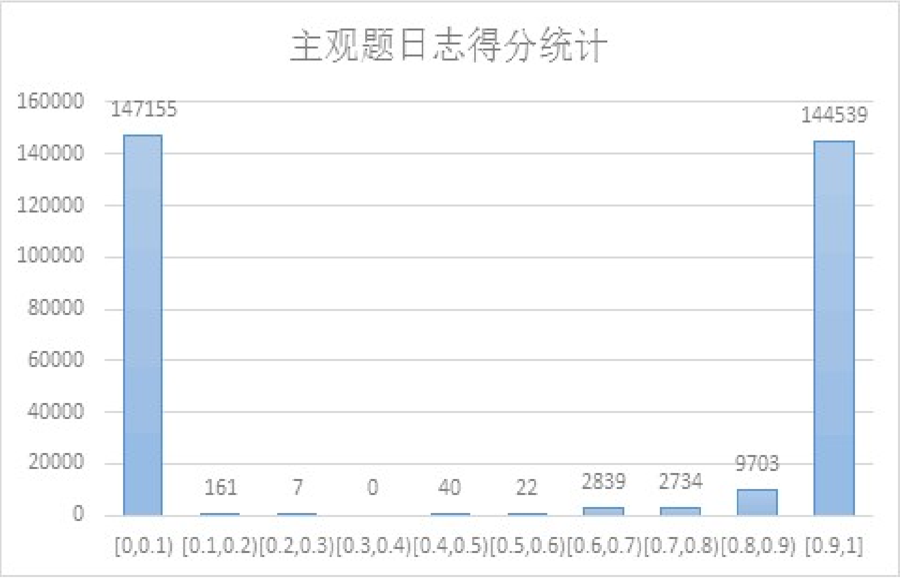
  - 若极差或四分位差较大，建模时需考虑数据是否有长尾现象。



数据分布基本指标

- 其它指标和现象观察
  - 教育数据





数据分布基本指标

- 以旅游套餐数据为例

Niagara Falls Discovery

(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.

数据分布基本指标

- 以旅游套餐数据为例

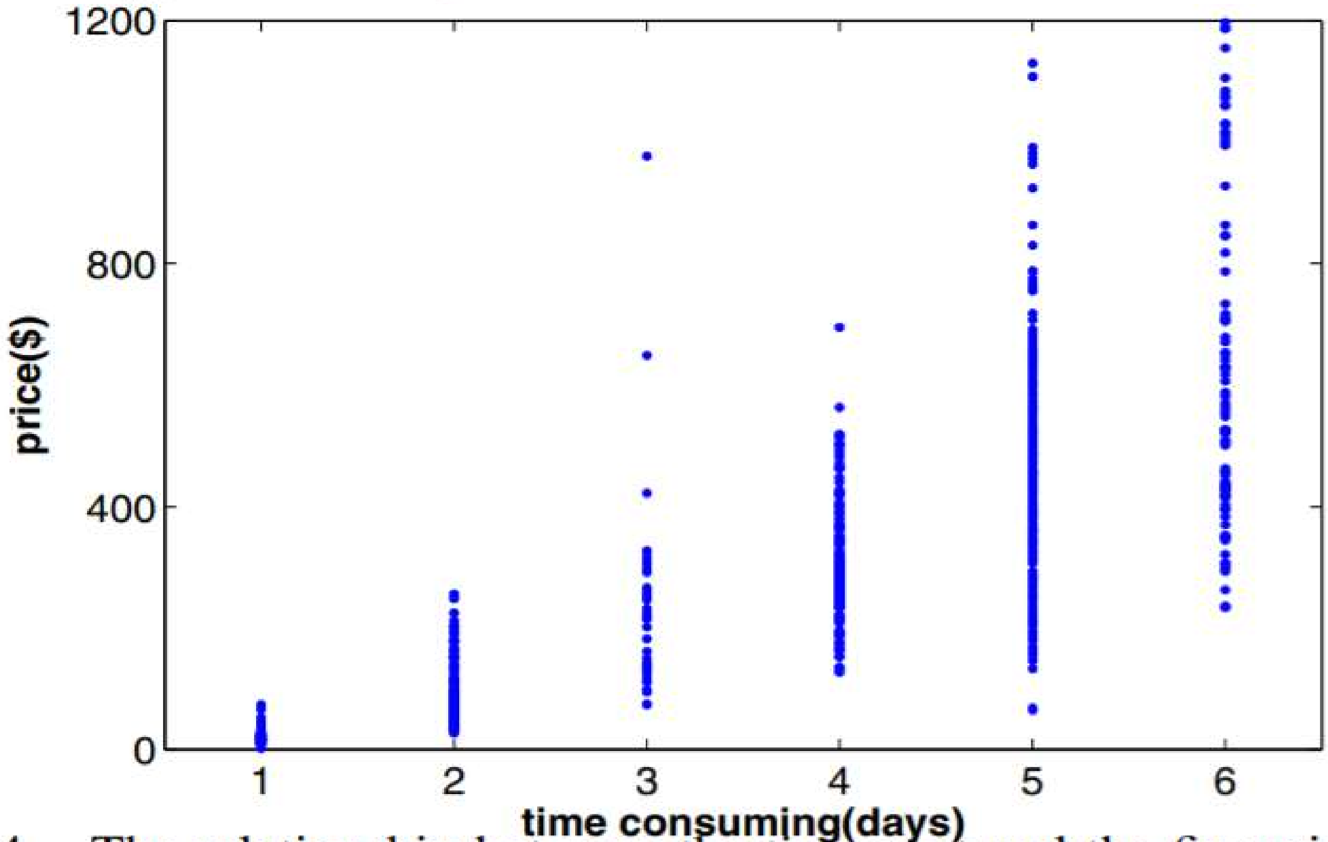


Figure 4. The relationship between the time cost and the financial cost in travel packages.

参数估计

- 统计指标能提供总体的信息吗？
  - 全面调查——可以
  - 抽样调查——不可以

- 通常在实际工作中，受限于人力物力的制约，几乎不可能调查全部个体，故常用抽样结果的统计量来估计总体的参数。

参数估计

- 参数
  - 参数 (parameter) 是用来描述总体特征的概括性数字度量。
- 统计量
  - 统计量 (statistic) 是用来表述样本特征的概括性数字度量，它完全由所抽取的样本计算得出，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
  - 参数估计 (parameter estimation) 是统计推断的基本问题之一，就是用样本统计量估计总体的参数。

参数估计

- 点估计
  - 点估计 (point estimate) 就是用样本统计量 $\hat{\theta}$ 的某个取值直接作为总体参数 θ 的估计值。

- 三个常用的点估计方法
  - 矩估计
  - 极大似然估计
  - 贝叶斯估计

参数估计

- 矩估计
  - 每一个随机变量X的矩都告诉你一些关于X分布的信息。
  - 随机变量X的矩：
    - K阶原点矩：$E(X^k)$
    - K阶中心矩：$E([X-E(X)]^k)$
    - K是正整数，且假设上述期望均存在。
  - 随机变量的一阶原点矩就是均值，二阶中心矩就是方差。

参数估计

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 $X_1, X_2, \ldots, X_n$ 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 $(i = 1, 2, \ldots, k)$。

参数估计

- 例子：黑白球（矩估计）
  - 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一个球出来，记录球的颜色，然后把拿出来的球再放回罐中。
  - 假如在前面的一百次重复记录中，有七十次是白球。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%。符合直观。

参数估计

- 极大似然估计
  - 极大似然估计 (Maximum likelihood estimation) 只适用于总体的分布类型已知的统计模型，是一种在大数据分析中较常见的估计方法，也称最大似然估计。

设 $X_1, X_2, \ldots, X_n$ 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。则样本 $(X_1, X_2, \ldots, X_n)$ 的分布为:

$$f(x_1,\theta)f(x_2,\theta)\ldots,f(x_n,\theta)$$

记为 $L(x_1, x_2, \ldots, x_n, \theta)$。

极大似然估计

固定 θ 时，$x_1, x_2, \ldots, x_n$ 是变元，L 是一个概率密度函数或概率函数。但如令变元固定在 X_1，X_2,…，X_n 处，让 θ 变化，则

$$L(\theta) = f(X_1, \theta)f(x_2, \theta), \ldots, f(x_n, \theta)$$

是一个固定在 Θ 上的函数，它被称为“似然函数”。直观上 $L(\theta)$ 表示由参数 θ 产生样本 $(X_1, X_2, \ldots, X_n)$ 的“可能性”大小。若把样本看成结果，把参数 θ 看成是导致这个结果的原因。现在已经有了结果，要反过来推算各种原因的概率，$L(\theta)$ 则是度量产生当前结果的各种原因的机会。参数 θ 并非事件或随机变量，是有一定的值的（虽然未知），因此并不称为概率，而改用“似然”这个词。

参数估计

极大似然估计

θ 的一个合理估计应该是的导致当前结果的机会（由 $L(\theta)$ 度量）达到最大值，由此我们可以给出定义：如 $\hat{\theta}$ 满足 $L(\hat{\theta}) = \max_{\theta\in\Theta} L(\theta))$，则称 $\hat{\theta}$ 为 θ 的极大似然估计。

当 $L(\theta)$ 关于 θ 可微时，引入对数似然函数 $l(\theta) = \ln L(\theta)$，由对数函数的凸性，$l(\theta)$ 与 $L(\theta)$ 在相同的位置取得极大值，可建立方程组（称为似然方程组），:

$$\frac{\partial l(\theta)}{\partial \theta_i} = \frac{\partial \ln L(\theta)}{\theta_i} = 0(i = 1, 2, \ldots, k)$$

如果该似然方程组有唯一的解，又能验证它是一个极大值点，则它必定使得 L 达到最大的点，即极大似然估计。

参数估计

- 例子：黑白球（极大似然估计）
  - 解：假设罐中白球的比例是θ，那么黑球的比例就是1- θ 。
  - 在一百次抽样中，七十次是白球的概率是P(Data | θ)。Data是所有的数据，包括$x_1, x_2, \dots, x_{100}$。
  - $P(Data|M) = \mathrm{P}(\mathrm{x}_1|\theta)\mathrm{P}(\mathrm{x}_2|\theta) \dots \mathrm{P}(\mathrm{x}_{100}|\theta) = \theta^{70}(1-\theta)^{30}.$
  - θ在取什么值的时候，P(Data | θ)的值最大呢？
  - 将上式对θ求导，可得θ =0.7。在边界点θ =0,1，P(Data | θ)=0。所以当θ =0.7时，P(Data | θ)的值最大。结果也是70%。

参数估计

- 贝叶斯估计
  - 矩估计和极大似然估计在根据统计量推断参数之前，对待估计的参数的可能取值范围没有任何的先验（prior）信息。
  - 贝叶斯估计则认为在抽样之前，试验人员或领域内专家已经对待估计参数有一定的认知，并认为这些“试验之前”就了解的信息是非常有用的。
  - 试验中通过把待估计参数 θ 看做一个随机变量并定义先验密度 $p(\theta)$ 来对其建模。

参数估计

- 贝叶斯估计
  - 先验密度
    - 先验密度 (prior density) 是根据领域内专家的经验，在抽样之前对待估计参数 θ 的可能取值的估计。试验中通过把待估计参数 θ 看做一个随机变量并定义先验密度 $p(\theta)$ 来对其建模。
  - 后验密度
    - 贝叶斯估计中集合专家的经验（先验密度，$p(\theta)$）和样本的信息（似然密度，$p(X|\theta)$），根据贝叶斯规则，得到 θ 的后验密度（posterior density）。
  - 贝叶斯学派也被称为经验学派，是和概率学派相对应的数理统计学派。

参数估计

- 贝叶斯估计
  - 在抽取样本之后，θ 可能的取值为：

$$p(\theta|X)=\frac{p(X|\theta)p(\theta)}{p(X)}=\frac{p(X|\theta)p(\theta)}{\int p(X|\theta')p(\theta')d\theta'}$$

  - 若试验中得到有效统计量 θ，就可以完全了解分布的概率密度函数（或概率函数）。
  - 若使用所有 θ 值的平均值，用它们的概率加权。则预测为：

$$y=\int g(x|\theta)p(\theta|X)d\theta$$

  - 如果后验概率没有很好的积分形式，上式的积分非常难求解。当总体积分难以求解时，可以把它看做一个一个的单点。

参数估计

- 贝叶斯估计
  - 贝叶斯估计(Bayes' estimator) 可被定义为后验密度的期望值：

$$\theta_{Bayes} = E[\theta|X] = \int \theta p(\theta|X) d\theta$$

  - 因为随机变量的最佳估计是该随机变量的均值，而贝叶斯估计中将待估计参数 θ 看做随机变量，所以在贝叶斯估计中取期望。

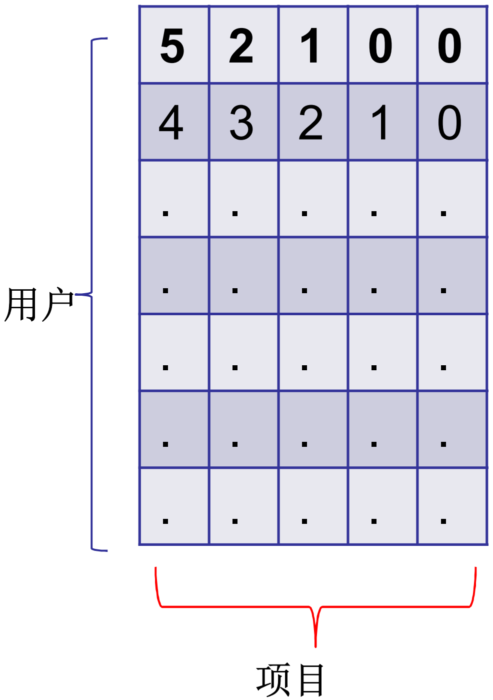
参数估计

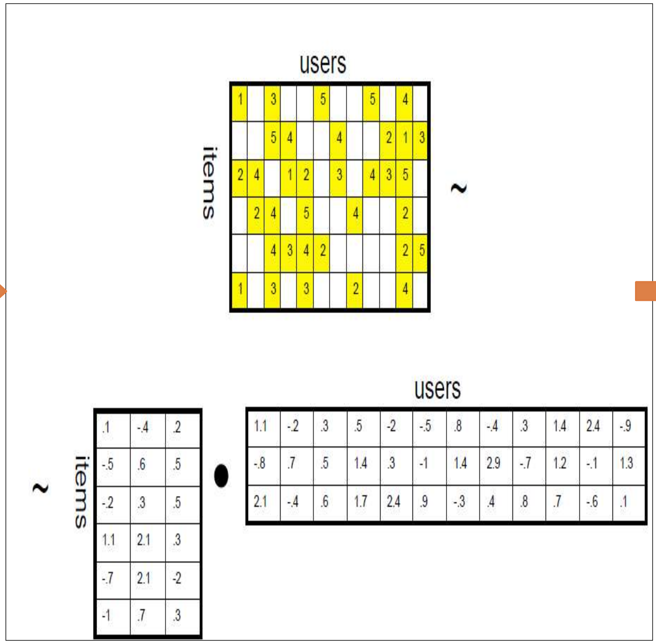
- 例子：黑白球（贝叶斯估计）
  - 若先验信息认为白球占比服从[0,1]上的均匀分布，如何估计罐中白球所占的比例？
  - 解： $P(\theta|Data)=\frac{P(Data|\theta)P(\theta)}{P(Data)}=\frac{\theta^{70}(1-\theta)^{30}P(\theta)}{P(Data)}$
  - 由于[0,1]上的均匀分布$P(\theta)$概率密度恒为1，且在比较$P(\theta|Data)$时可将$P(Data)$视为常数。故上式仍在θ=0.7时取得最大值。
  - 思考：若先验信息认为白球占比服从均值0.5，方差0.1的正态分布，如何估计罐中白球所占的比例？

  提示：通常可将先验分布的概率密度函数p($\boldsymbol{\theta}$)代入贝叶斯公式，求导寻找最大值。若导函数不易直接求解，可借助数值计算方法寻找最大值。

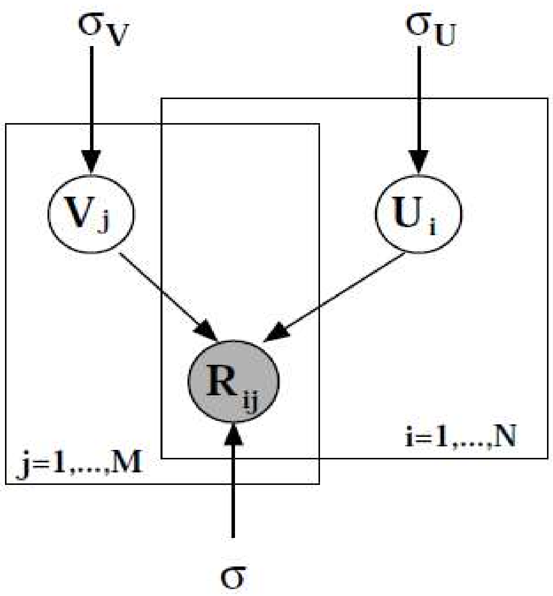
参数估计

- 基于矩阵分解的协同过滤算法
  - 面向评分预测的模型





基本分解思想

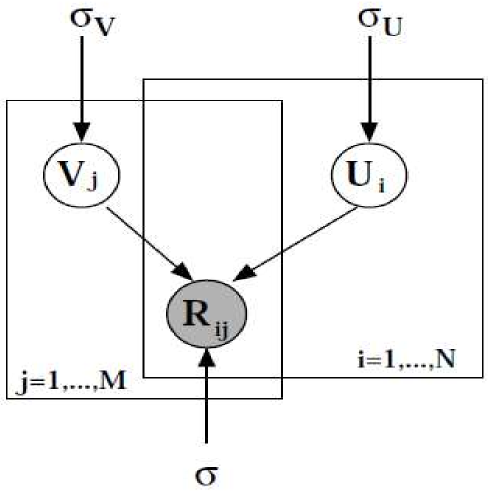


PMF（概率矩阵分解）

评分预测算法设计

- 基于矩阵分解的协同过滤算法
  - PMF Solution

$$p(R|U,V,\sigma^2) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}(R_{ij}|U_i^T V_j,\sigma^2)\right]^{I_{ij}}$$



  - How to get U and V?
    - The log-posterior of user and item features over fixed parameters

$$p(U,V\mid R,\sigma^2,\sigma_U^2,\sigma_V^2)$$

$$\propto \underbrace{\boxed{p(R\mid U,V,\sigma^2)}}_{\text{Likelihood!}} * \underbrace{\boxed{p(U\mid\sigma_U^2) * p(V\mid\sigma_V^2)}}_{\text{Prior}}$$

$$p(V|\sigma_V^2) = \prod_{j=1}^{M}\mathcal{N}(V_j|0,\sigma_V^2\mathbf{I})$$

$$p(U|\sigma_U^2) = \prod_{i=1}^{N}\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})$$

参数估计

- 如何评价点估计？
  - 由于估计值依赖于对样本的一次观测结果，因此所得到的估计值有一定的偶然性。
  - 在考虑估计值的优劣时，须从统计量的整体性能去评价。
    - 估计量的某种特性（如无偏性）
    - 具体的数量型指标（如均方误差）
  - 对点估计的比较是相对的，要从多个角度去考虑。

参数估计

- 点估计的优良性
  - 无偏性
  - 有效性
  - 相和性

参数估计

- 无偏性
  - 无偏性 (unbiasedness) 是指估计量抽样分布的数学期望等于被估计的总体参数。
  - 设总体参数为 θ，所选择的估计量为 $\hat{\theta}$，如果 $E(\hat{\theta})=\theta$，则称 $\hat{\theta}$ 为 θ 的无偏估计量
  - 从无偏性的定义可以看出，满足该性质的估计量没有系统偏差，即用$\hat{\theta}$ 估计 θ 时只存在随机误差。
  - 由于随机误差在多次反复的实验中会相互抵消，平均说来 $\hat{\theta}$ 可以准确的估计出 θ 。

参数估计

- 有效性
  - 有效性 (efficiency) 指对同一总体参数的两个无偏估计，有更小标准差的估计量更有效。
  - 设 $\hat{\theta}_1, \hat{\theta}_2$是 θ 的两个无偏估计，它们的抽样分布的方差分别用 $D(\hat{\theta}_1)$ 和 $D(\hat{\theta}_2)$ 表示。
  - 如果 $D(\hat{\theta}_1) <= D(\hat{\theta}_2)$，且至少对某个$\theta_0$使之成立严格不等式，就称 $\hat{\theta}_1$比 $\hat{\theta}_2$更有效。
  - 在 θ 的所有无偏估计中，方差最小的那个一无偏估计被称为一致最小方差无偏估计。

参数估计

- 相和性
  - 相合性 (consistency) 是指随着样本容量 n 的不断增加，点估计的值越来越接近被估计总体的参数，即 $\hat{\theta}$ 越来越接近 θ 。
  - 大样本量给出的估计量更接近于总体参数 θ 。

参数估计

- 点估计的数量型指标
  - 均方误差(MSE)
  - 均方根误差(RMSE)
  - 平均绝对误差(MAE)

参数估计

- 均方误差(MSE)
- $MSE = \frac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2$
  - 均方误差是指参数估计值与参数真值之差平方的期望值;
  - MSE可以评价数据的变化程度，MSE的值越小，说明预测模型描述实验数据具有更好的精确度。

参数估计

- 均方根误差(RMSE)
- $MSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2}$
  - 均方根误差是均方误差的算术平方根。
  - RMSE是对点估计最常用的评价指标。
- RMSE常见应用场景
  - 推荐系统
  - 得分预测
  - 回归

参数估计

- 平均绝对误差(MAE)
- $\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}|observed_t - predicted_t|$
    - 平均绝对误差是绝对误差的平均值。
    - 平均绝对误差能更好地反映预测值误差的实际情况。

参数估计

- 区间估计
  - 通过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。
  - 在物理、化学等实验领域中，区间估计比点估计更受青睐。
  - 在数据科学中，区间估计应用较少。