数据科学导论

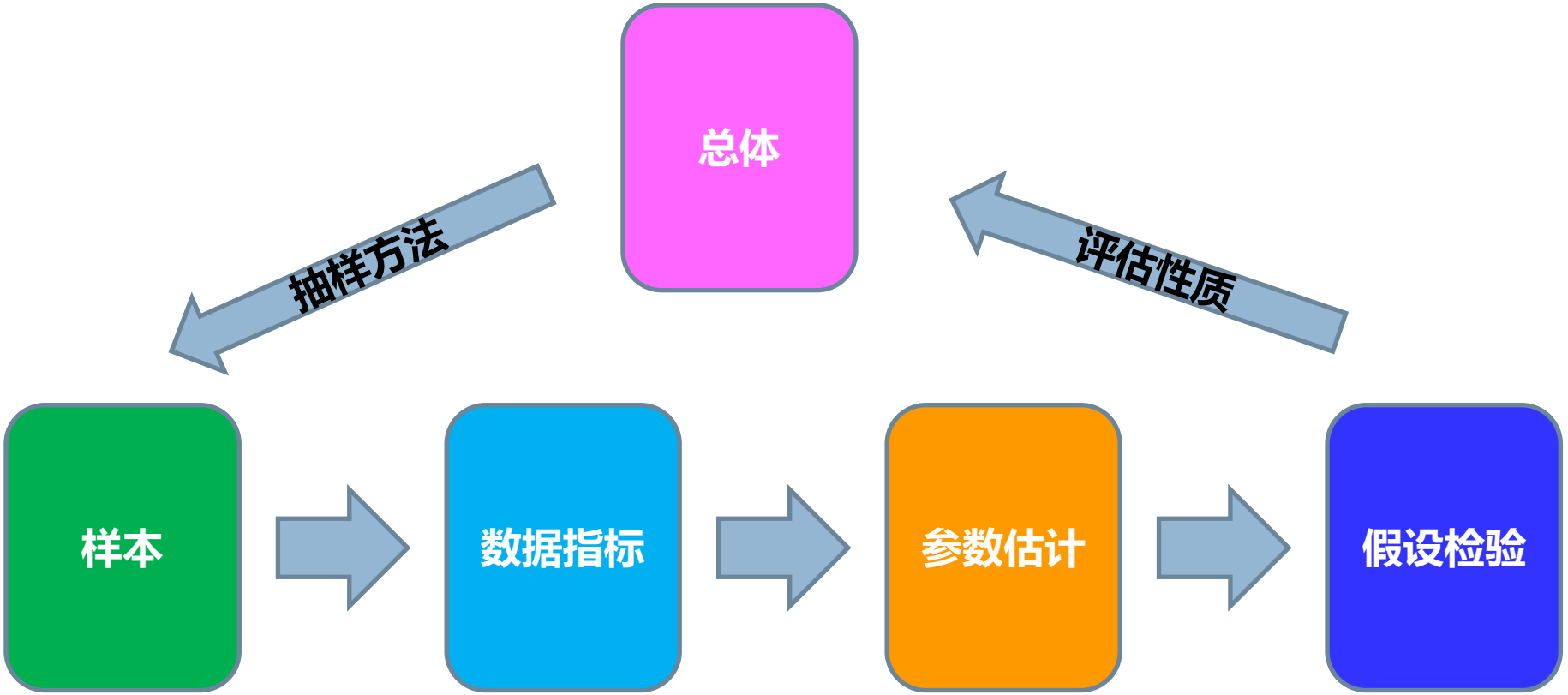
Introduction to Data Science

第三章 数据统计基础

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

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html

回顾：数据统计



数据统计

- 数据分布
- 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： n趋近于无穷，样本矩趋近于总体矩
    - 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：$E(X^k)=\frac{1}{n}\sum_{i=1}^{n}X_i^k$
    - K阶中心距：$E([X-E(X)]^k)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$
      - 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

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

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

参数估计—矩估计

- 矩估计
  - 原理：大数定律： n趋近于无穷，样本矩趋近于总体矩
    - 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：$E(X^k)=\frac{1}{n}\sum_{i=1}^{n}X_i^k$
    - K阶中心距：$E([X-E(X)]^k)=$
      - 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

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

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

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—

K阶原点矩：$E(X^k)=\frac{1}{n}\sum_{i=1}^{n}X_i^k$

K阶中心距：$E([X-E(X)]^k)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$

问 设总体X在$[a,b]$上服从均匀分布，$X_1,\dots,X_n$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：

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

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

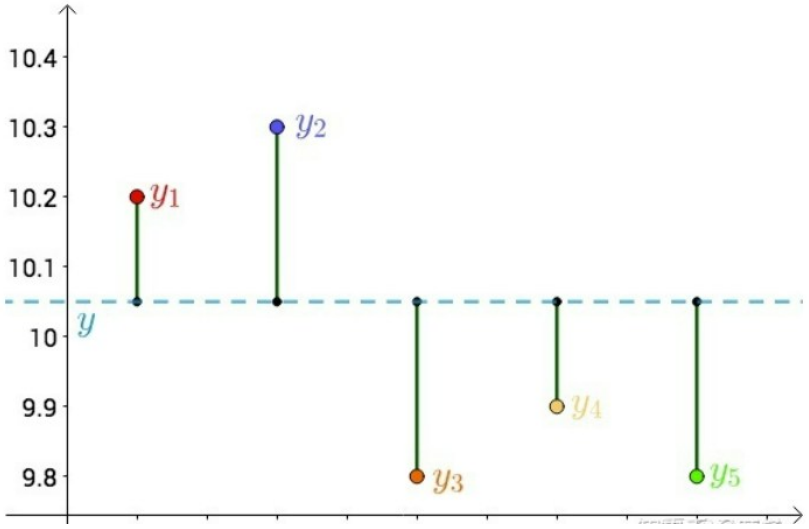
参数估计—矩估计

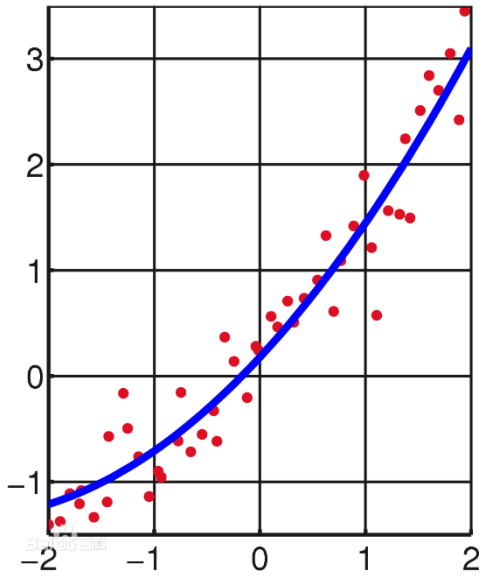
- 举例：黑白球（矩估计）
  - 例：假如有一个罐子，里面有黑白两种颜色的球，数目多少不知，两种颜色的比例也不知。每次任意从已经摇匀的罐中拿1个球出来，记录球的颜色，然后把拿出来的球再放回罐中。假如在前面的100次重复记录中，有70次是白球。请问罐中白球所占的比例是多少？

    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

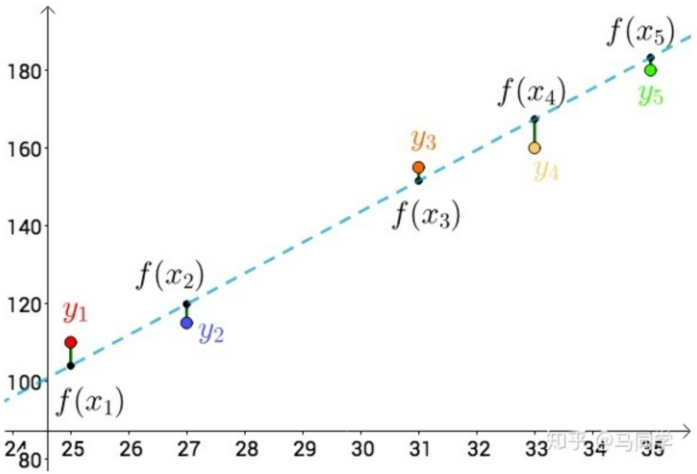
- 最小二乘估计(Least Square Estimate, LSE)
  - 总体的模型：用样本数据拟合总体的参数估计量，即估计值与观测值之差的平方和最小
  - 目标：最小化估计值θ与观测值$\hat{\theta}$之差的平方和
  - $\min L(\theta) = \sum_{i=1}^{N}(\theta - \hat{\theta}_i)^2$





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$f(X|\theta) = \theta_0 + \theta_1 x + \theta_1 x^2 + \cdots + \theta_n x^n = \boldsymbol{\theta^T X}$
  - 目标：$\min L(\theta) = \sum_{i=1}^{N} (f(X|\theta) - Y)^2$
  - 求解：一阶导数为0：$\frac{\partial L(\theta)}{\partial \theta_0} = 0$，$\frac{\partial L(\theta)}{\partial \theta_1} = 0$，$\cdots$，$\frac{\partial L(\theta)}{\partial \theta_n} = 0$



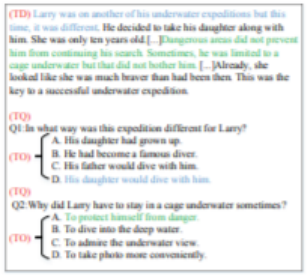
课后学习：最小二乘矩阵求解方法

参数估计—最小二乘估计

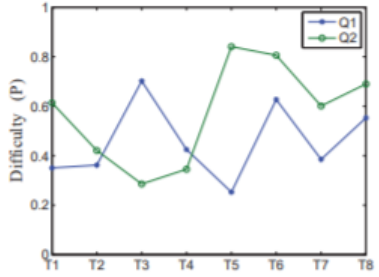
- 最小二乘估计—建模案例

Question Difficulty Prediction for READING Problems in Standard Tests

$$\mathcal{J}(\Theta) = \sum_{Q_i} (P_i - \mathcal{M}(Q_i))^2 + \lambda_\Theta ||\Theta_\mathcal{M}||^2, \qquad (5)$$



(a) A READING problem

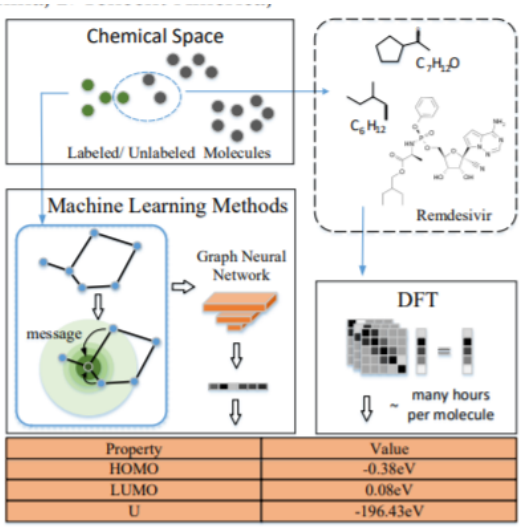


(b) Difficulties in tests

ASGN: An Active Semi-supervised Graph Neural Network for Molecular Property Prediction

Traditionally, MPGNN is trained in a supervised manner where all the labels are given and we usually use mean square loss (MSE) between predictions and labels $\boldsymbol{y}_i$ (i.e. the labeled properties in $\mathcal{D}_l$) to guide the optimization of the model parameters:

$$\mathcal{L}_p = \sum_{i=1}^{N_l} \|\boldsymbol{y}_i - f_\theta(\boldsymbol{z}_{\mathcal{G}_i})\|^2. \qquad (4)$$



- ✓ Zhenya Huang, Enhong Chen, Question Difficulty Prediction for READING Problems in Standard Tests, AAAI2017
- ✓ Zhongkai Hao, Zhenya Huang,, Qi Liu, Enhong Chen, ASGN: An Active Semi-supervised Graph Neural Network for Molecular Property Prediction, KDD 2020

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θ

目标：最小化估计值θ与观测值$\hat{\theta}$之差的平方和

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

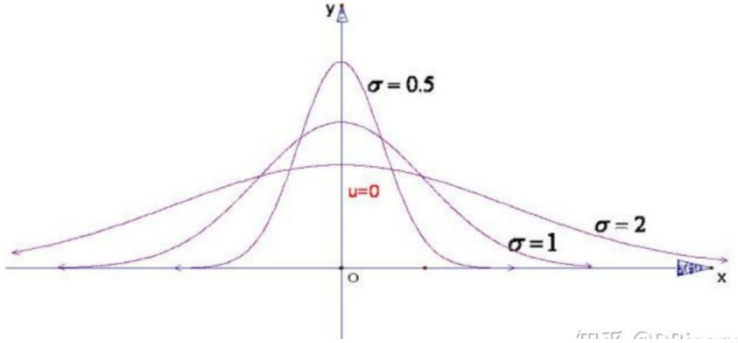
参数估计

- 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
- 点估计的常用方法
  - 矩估计
  - 最小二乘估计 LSE
  - **极大似然估计 MLE**
  - **最大后验估计 MAP**
  - **贝叶斯估计**

参数估计—极大似然估计

- 极大**似然**估计(Maximum Likelihood Estimate，MLE)
  - 思想：利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值
  - **模型已定，参数未知**
  - 目标：概率分布函数或者似然函数最大
    - 用似然函数取到最大值时的参数值作为估计值
  - 概率分布模型
    - 伯努利分布
    - 二项分布
    - 高斯分布
    - 泊松分布

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



参数估计—极大似然估计

- 极大似然估计(MLE)
  - MLE目标：用似然函数取到最大值时的参数值作为估计值
  - 设总体分布为$f(X|\theta)$，$x_1, x_2, x_3, \cdots, x_N$为样本。样本满足独立同分布，则他们的联合密度函数为：

$$\mathrm{L}(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

  - 其中，θ为未知参数。样本已经存在(观测)，即，$x_1, x_2, x_3, \cdots, x_n$是固定的。$\mathrm{L}(X|\theta)$是关于$\theta$的函数，称为似然函数
  - 目标：求参数θ，**使似然函数取极大值，称为极大似然估计**
  - 实践中，通常对似然函数取对数(log或ln)(连乘运算变为连加运算)，即**对数似然函数**。所以，极大似然估计问题可以写成

$$\ln \mathrm{L}(x_1, x_2, \cdots, x_n|\theta) = \sum_{i=1}^{\mathrm{n}} ln\big(f(x_i|\theta)\big)$$

参数估计—极大似然估计

- 例子: 扔硬币，X每次实验X_i服从伯努利分布
  - 参数为$\boldsymbol{\theta}$，假设为事件(正面向上)发生的概率

$$P(X_i \mid \theta)=\begin{cases}\theta, & X_i\text{为正面}\\ 1-\theta, & X_i\text{为反面}\end{cases}$$

- n次实验，共k次正面向上，采用MLE估计参数$\boldsymbol{\theta}$：

样本观测

正　反

k次　N-k次
5次　5次

产生观测样本的概率不同

目标：找到发生样本最大概率的参数

总体参数

	正	反
θ	0.5	0.5

	正	反
θ	0.2	0.8

	正	反
θ	0.9	0.1

参数估计—极大似然估计

- 例子: 扔硬币，X每次实验X_i服从伯努利分布
  - 参数为$\boldsymbol{\theta}$，假设为事件(正面向上)发生的概率

$$P(X_i \mid \theta)=\begin{cases}\theta, & X_i\text{为正面}\\ 1-\theta, & X_i\text{为反面}\end{cases}$$

- n次实验，共k次正面向上，采用极大似然估计估计参数$\boldsymbol{\theta}$：
  - 似然函数：$L(x_1,x_2,\cdots,x_n|\theta)=C_n^k\theta^k(1-\theta)^{n-k}$
  - 对数似然函数：$\ln L(X|\theta)=\ln C_n^k+k\ln\theta+(n-k)\ln(1-\theta)$
  - 求极值：$\frac{\partial L(\theta)}{\partial\theta}=0$，则：$\frac{k}{\theta}-\frac{n-k}{1-\theta}=0$
  - 参数$\boldsymbol{\theta}$的最大似然估计值: $\theta_{MLE}=\frac{k}{k+n-k}=\frac{k}{n}$

✓ 二项分布中每次事件发生的概率$\boldsymbol{\theta}$ = 做N次独立重复随机试验中事件发生的概率
✓ 例如：如果做20次实验，出现正面14次，反面6次:
  ✓ MLE得到参数值p为14/20 = 0.7

参数估计—极大似然估计

- 极大似然估计—高斯分布的参数
  - 例：给定$x_1, x_2, x_3, \cdots, x_N$为样本，已知样本来自于高斯分布$N(\mu, \sigma)$,估计参数$\mu, \sigma$

解：

- 高斯分布的概率密度函数：$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
- 带入样本，似然函数：$$L(X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$
- 对数似然：$$\ln L(X) = \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$
$$= -\frac{n}{2}\ln(2\pi\sigma^2) + -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$
- 求偏导估计参数：$$\mu = \frac{1}{n}\sum_i x_i \qquad \sigma^2 = \frac{1}{n}\sum_i (x_i-\mu)^2$$

与矩估计结果相同。

参数估计—矩估计 vs LSE vs MLE

- 举例：黑白球（极大似然估计）—课堂练习
  - 例：假如有一个罐子，里面有黑白球，数目不知，两种颜色比例也不知。每次任意从摇匀的罐中拿1个球出来，记录颜色，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：假设：白球占比为θ，黑球占比为1-θ

已知100次观测量：出现白球70次，黑球概率30次

则，极大似然估计的目标为：

$$\max L(x_1, x_2, \cdots, x_{100}|\theta) = C_{100}^{70}\theta^{70}(1-\theta)^{30}$$

取对数为：$\ln\left(L(x_1, x_2, \cdots, x_{100}|\theta)\right) = 70 \times ln\theta + 30 \times ln(1-\theta)$

令：$\frac{\partial L(\theta)}{\partial \theta} = 0$

则：$\theta = \frac{7}{10}$

矩估计、LSE、MLE结果均相同，但原理不同

参数估计—课后学习与思考

- 矩估计 vs LSE vs MLE—关联与区别
  - LSE可以通过高斯分布+MLE推算出来
  - LSE和MLE对应机器学习中的经验风险最小化
- MLE
  - 似然函数取对数后导数还是不好求：期望最大算法（EM）
  - 高斯混合模型
  - 机器学习中的交叉熵
  - 线性模型的极大似然估计方法
  - 逻辑斯蒂回归的极大似然估计方法

参数估计—最大后验估计

- 回顾扔硬币的例子
  - X每次实验X_i服从伯努利分布
  - 假设为事件(正面向上)发生的概率，参数为$\boldsymbol{\theta}$，
  - n次实验，共k次正面向上，目标为估计参数$\boldsymbol{\theta}$

- MLE的思想
  - $L(X|\theta)$ 似然函数取到最大值时的参数值作为估计值

样本观测

正 反

k次 N-k次
5次 5次

MLE

总体参数

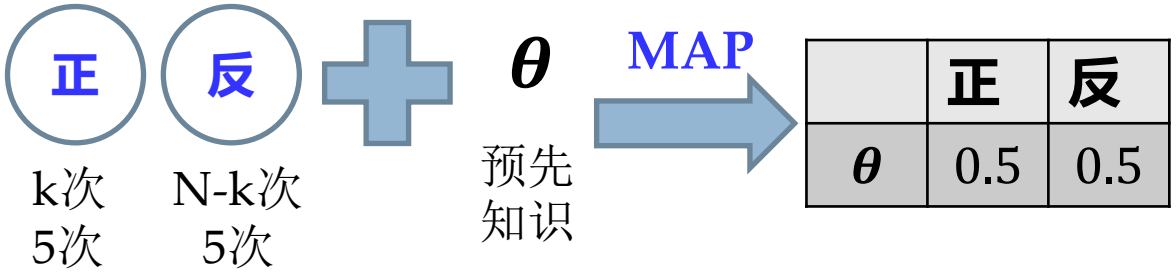
	正	反
$\boldsymbol{\theta}$	0.5	0.5

- 频率学派
  - 完全相信数据
  - 世界是确定的
  - 事件在多次重复实验中趋于稳定

参数估计—最大后验估计

- MLE是否有不足？—实验是否靠谱？
  - 实验对象(硬币)是否均匀？实验次数是否足够？
  - 实验环境是否有影响？。。。
- 最大后验估计(Maximum A Posteriori Estimation, MAP)
  - 目标：最大化在给定数据样本X的情况下模型参数的**后验概率**
    - 模型参数使得模型能够产生该数据样本的概率最大—**似然概率**
    - 但对于模型参数有了一个**假设**，加入了**先验知识**，即模型参数可能满足某种分布，即，估计不止依赖数据样本。

样本观测　　**参数假设**　　**总体参数**



- **贝叶斯学派**
  - 不能完全相信数据
  - 世界是不确定的
  - 数据量的增加，参数向数据靠拢—先验影响越小

参数估计—最大后验估计

- 贝叶斯公式：

$$P(A,B) = P(B) * P(A|B) = P(A) * P(B|A)$$

$$P(A|B) = \frac{P(B|A)}{P(B)} * P(A)$$

$$= \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\sim A)P(\sim A)}$$

$$\sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

- 因果概率

$$P(Cause \mid Effect) = \frac{P(Effect \mid Cause)P(Effect)}{P(Effect)}$$

参数估计

- 贝叶斯公式的理解—举例
  - 已知：
    - 临床案例发现：患者得 meningitis(脑膜炎) 导致 stiff neck(颈部僵硬) 的概率为 50%
    - 先验知识：患者得 meningitis 的概率为 1/50,000
    - 先验知识：患者得 stiff neck 的概率为 1/20
  - 问：如果患者得 stiff neck, 那么他患有meningitis的概率为?
    - 设 *M* 为患 meningitis（脑膜炎）的概率, *S* 为患 stiff neck 的概率:

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

注: 患有 meningitis 的后验概率 仍然非常小