数据科学导论

Introduction to Data Science

第 10 章
统计分析的原理

矩阵

线性代数

关系代数

概率论

统计

微积分

机器学习基础

数据科学的数学基础

- 矩阵（Matrix）是一个按照长方阵列排列的复数或实数集合。涉及到的机器学习应用有SVD、PCA、最小二乘法、共轭梯度法等。
- 线性代数是研究向量、向量空间、线性变换等内容的数学分支。向量是线性代数最基本的内容。中学时，数学书告诉我们向量是空间（通常是二维的坐标系）中的一个箭头，它有方向和数值。在数据科学家眼中，向量是有序的数字列表。线性代数是围绕向量加法和乘法展开的。
- 矩阵和线性代数是一体的，矩阵是描述线性代数的参数。它们构成了数据科学的庞大基石。

矩阵和线性代数

智力：10　敏捷：6　力量：2

标量

英雄A：(10，6，2)

英雄B：(3，4，10)

英雄C：(5，10，4)

英雄D：(6，9，5)

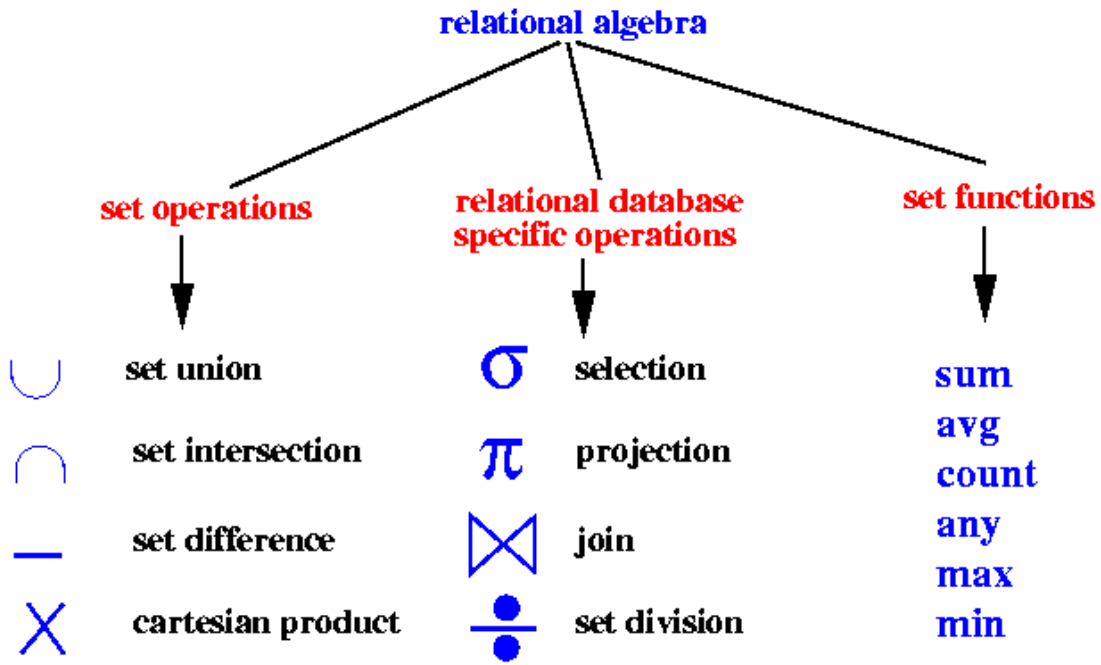
$$\begin{bmatrix} 10 & 6 & 2 \\ 3 & 4 & 10 \\ 5 & 10 & 4 \\ 6 & 9 & 5 \end{bmatrix}$$

矩阵

向量

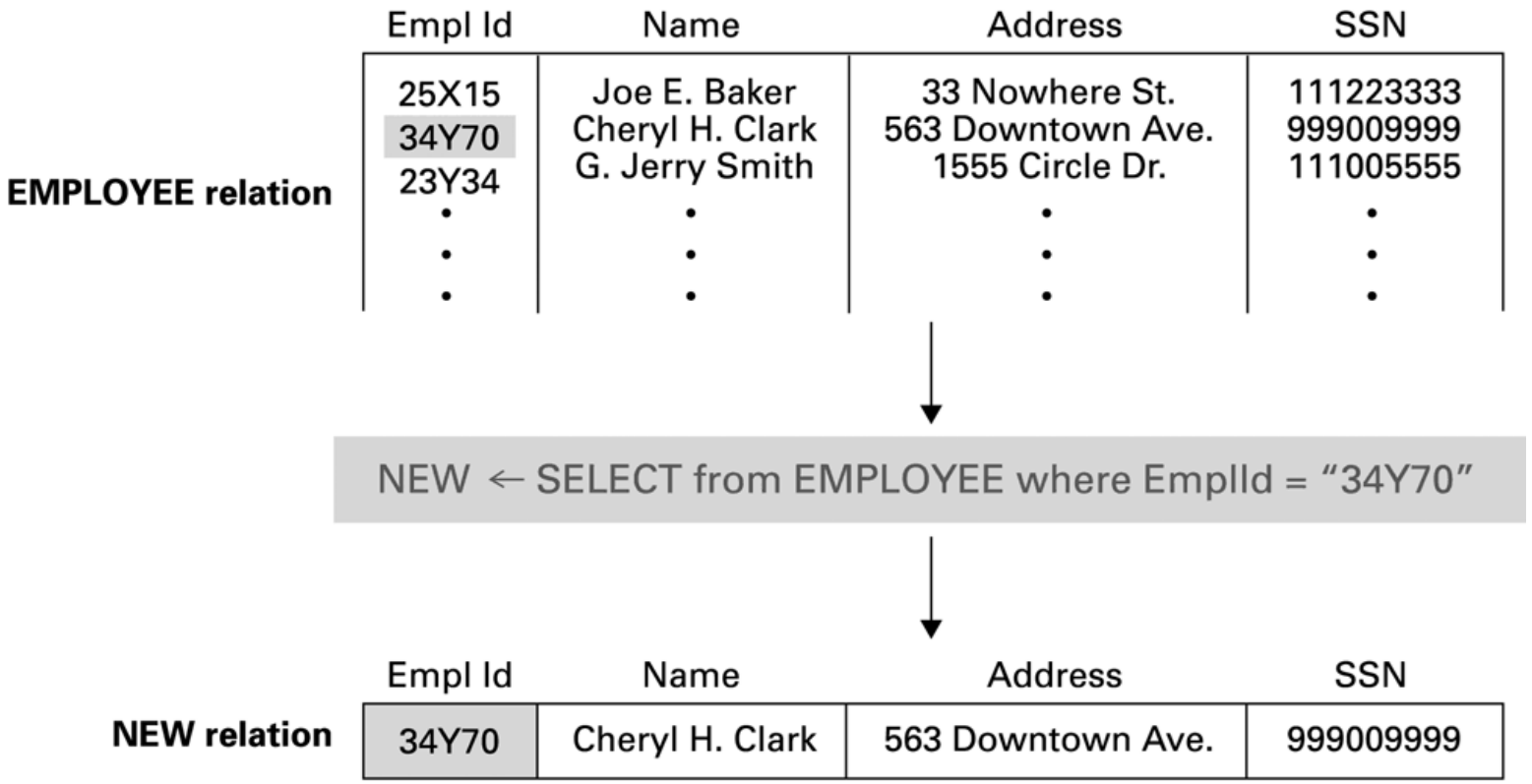
矩阵和线性代数

- 它是一种抽象的查询语言。基本的代数运算有选择、投影、集合并、集合差、笛卡尔积和更名。
- 关系型数据库就是以关系代数为基础，在SQL语言中都能找到关系代数相应的计算。



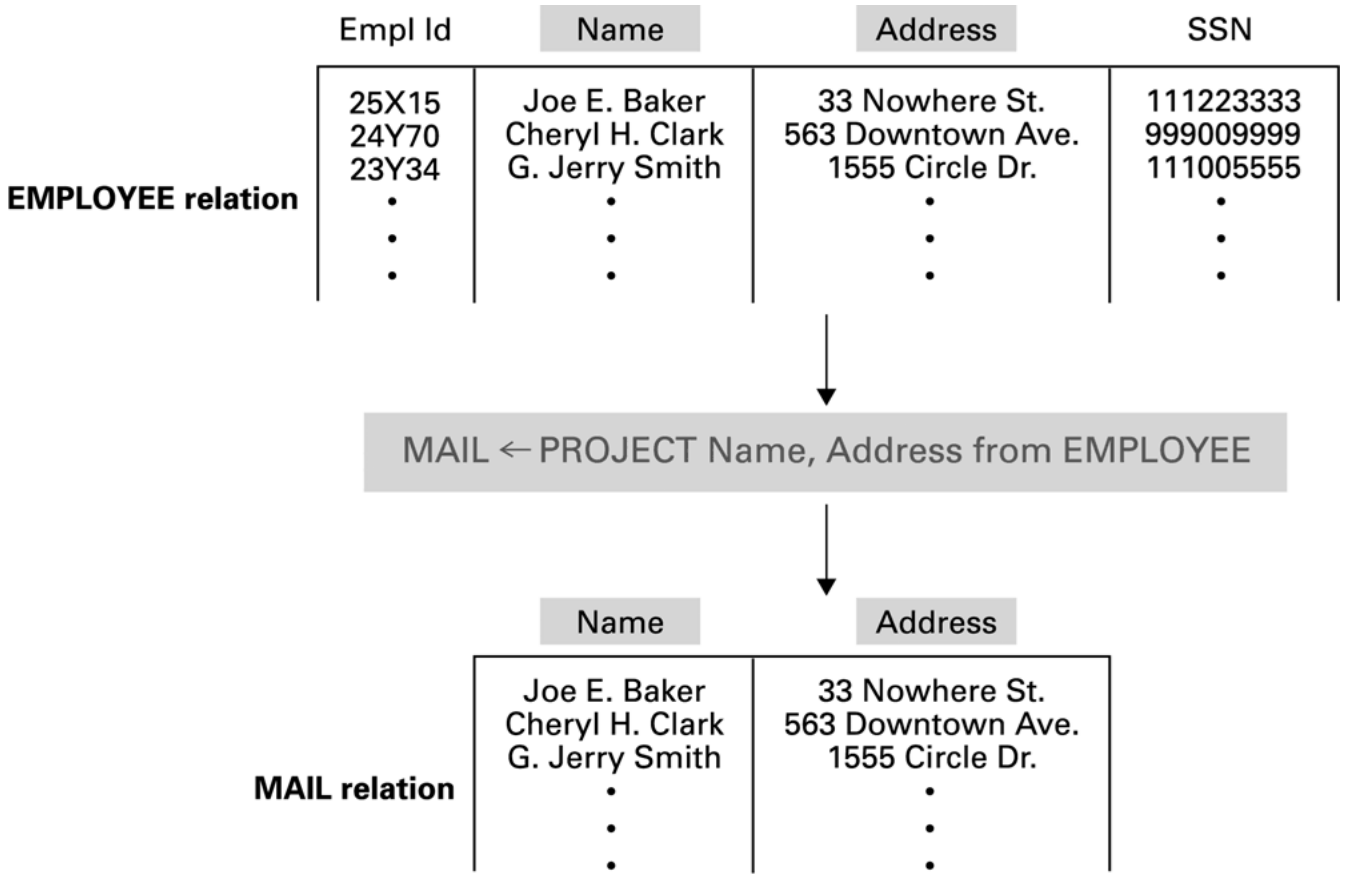
关系代数

The SELECT operation



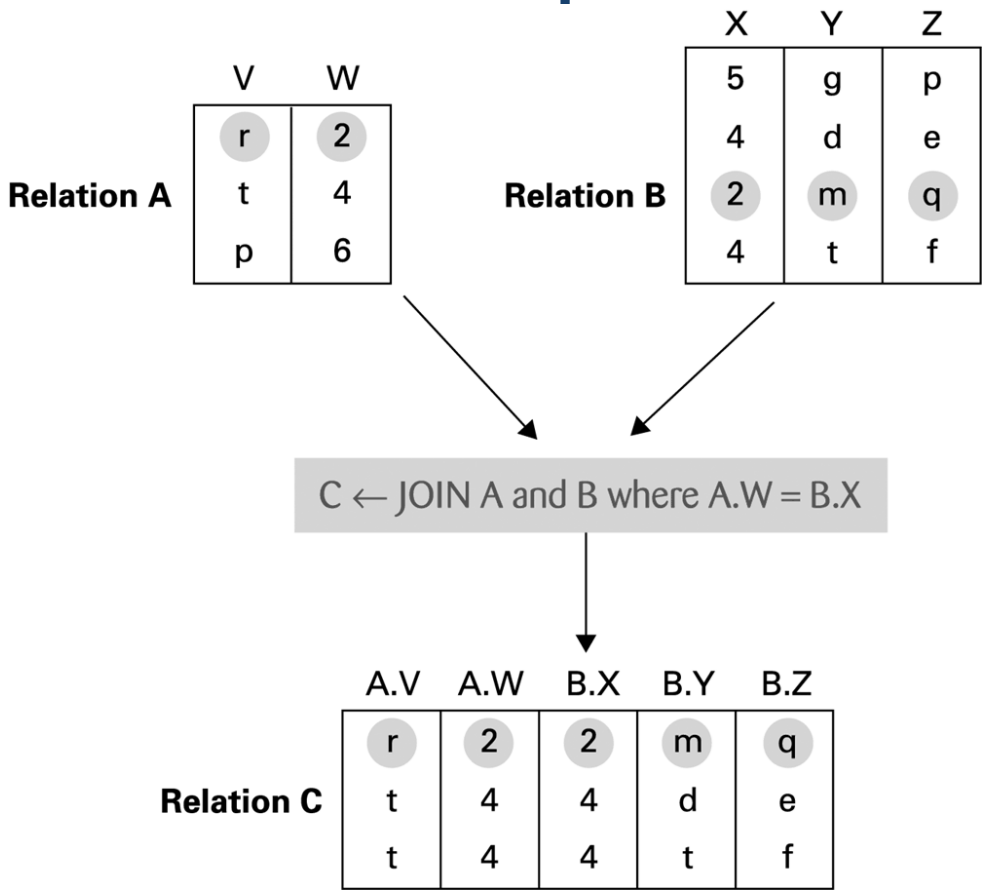
关系代数

The PROJECT operation



关系代数

The JOIN operation



关系代数

- Bayes Theorem（贝叶斯定理）
- Random Variables（随机变量）
- Cumulative Distribution Function（累计分布函数）
- Continues Distributions（连续分布）
- Probability Density Function（概率密度函数）
- ANOVA（方差分析）
- Central Limit Theorem（中心极限定理）
- Monte Carlo Method（蒙特卡罗方法）
- Hypothesis Testing（假设检验）
- p-Value（P值）
- Estimation（估计）
- Confidence interval（置信区间）
- Maximum Likelihood Estimate（极大似然估计）
- Kernel Density Estimate（核密度估计）
- Regression（回归）
- Covariance（协方差）
- Correlation（相关性）
- Pearson correlation coefficient（Pearson相关系数）
- Causation（因果性）
- Least Squares Fitting（最小二乘法）
- Euclidean Distance（欧氏距离）

概率论

- 统计学（Statistics）是通过搜索、整理、分析、描述数据等手段，以达到推断所测对象的本质，甚至预测对象未来的一门综合性科学。
- 事物的发展充满了不确定性，而统计学，既研究如何从数据中把信息和规律提取出来，找出最优化的方案；也研究如何把数据当中的不确定性量化出来。
- 大数据告知信息但不解释信息。打个比方，大数据是“原油”而不是“汽油”，不能被直接拿来使用。
- 大数据时代，统计学是数据分析的灵魂。

统计学

- 起源：用单个数或者数的小集合捕获可能很大值集的各种特征
  - 频率度量：众数
  - 位置度量：均值和中位数
  - 散度度量：极差和方差
  - 数据分布：频率表、直方图
  - 多元汇总统计：相关矩阵、协方差矩阵

汇总数据的初衷如公司的组织结构，高层期望看到工作概要，而不是细节

统计学：起源

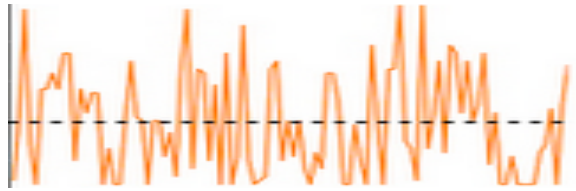
汇总数据指标的设计，源于非常朴素的思想

- 标准差：想设计一个指标，可以用来衡量数据集合的发散性，经过如下思考
  - 每个样本的偏差累加就可以衡量 $\sum(x_i - \hat{x_i})$
  - 偏差较大的值应该具有更大的权重 $\sum(x_i - \hat{x_i})^2$
  - 集合中数字越多，方差越大，应该与集合大小无关 $\frac{\sum(x_i-\hat{x_i})^2}{n}$
  - 量纲与原始数据不同，无法比 $\sqrt{\frac{\sum(x_i - \hat{x_i})^2}{n}}$
  - 最终结果，RMSE（均方根误差）

统计学：设计

$$RMSE = \sqrt{\frac{\sum(x_i - \hat{x}_i)^2}{n}}$$

貌似这个宽度就可以体现数据的波动性大小



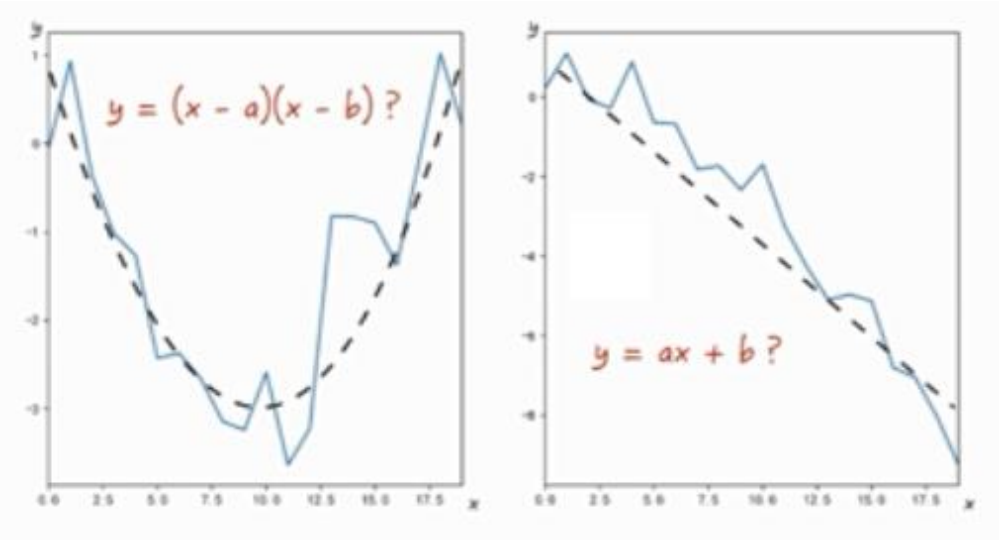
5次约会，每次迟到10分钟，与一次迟到50分钟，哪个更难接受?

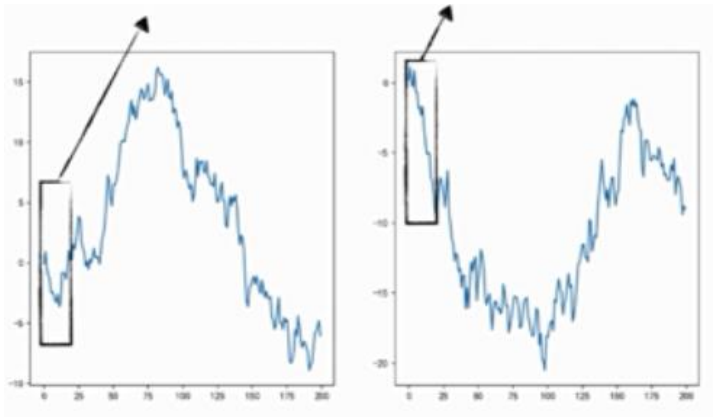
统计学：设计

- 统计分析是基于统计理论，是应用数学的一个分支。在统计理论中，随机性和不确定性由概率理论建模。统计分析技术可以分为：
  - Descriptive Statistics（描述性统计）：解释数据的一些特征；
  - Exploratory Statistics Analysis（探索性统计分析）：开始关注数据的内在规律；
  - Inferential Statistics（推断性统计）：怎样用已知数据来进行预测和判断。
- 例如多元统计分析：回归、因子分析、聚类和判别分析等。

统计分析方法

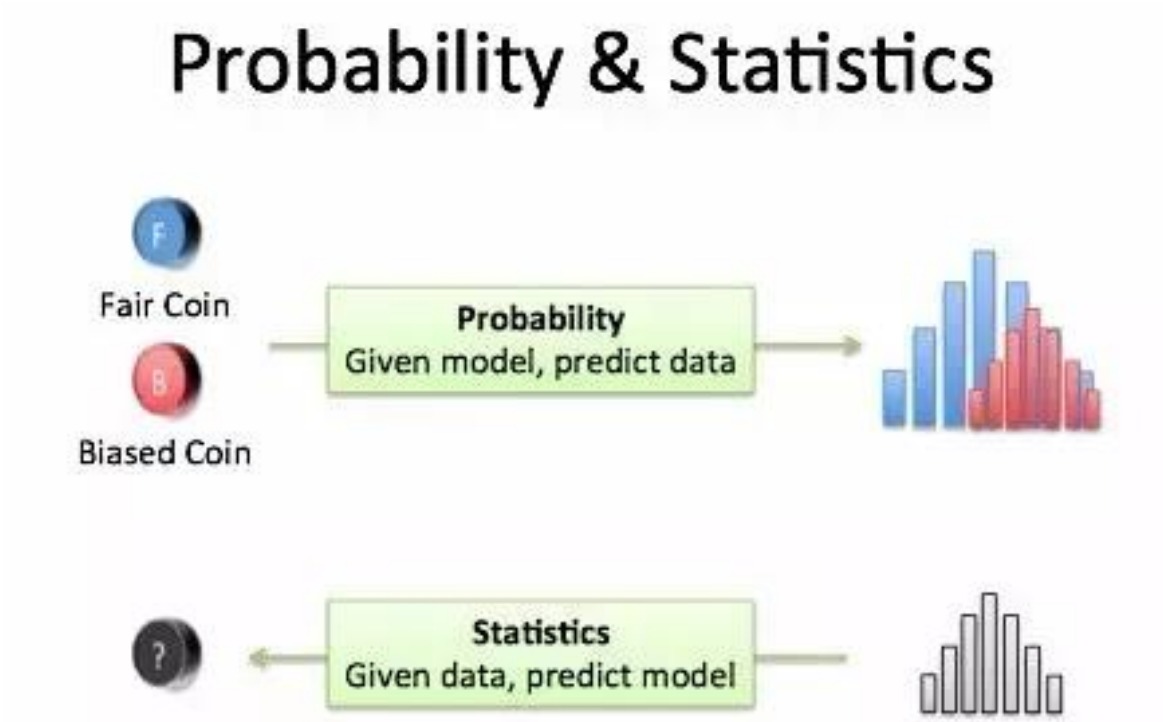
统计学家最关心的问题：如何避免掉入数据陷阱

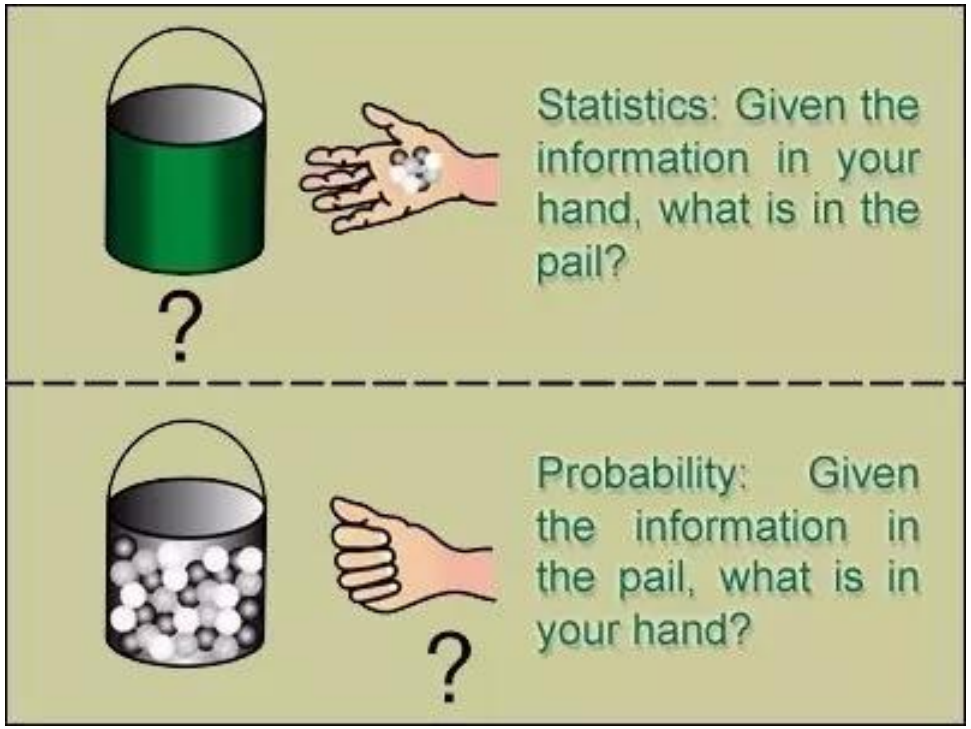




统计：不要过分关注局部

统计分析方法



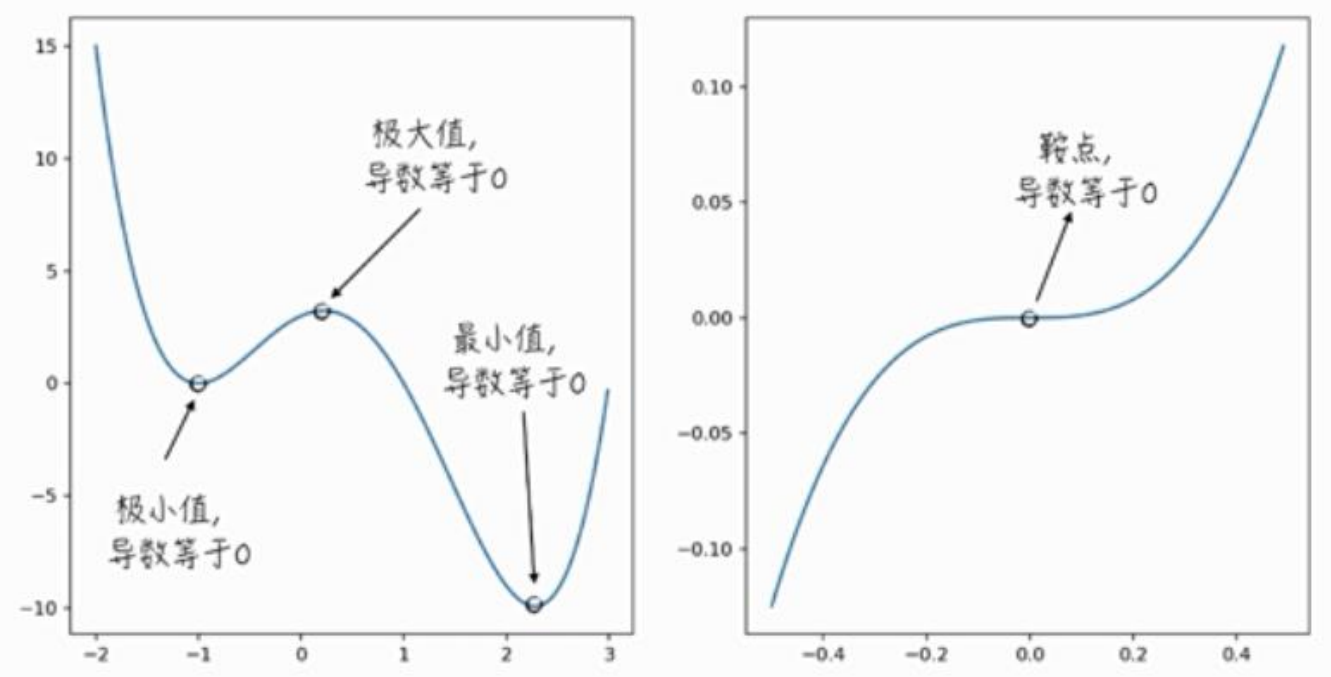


公众号文章：说说统计学、概率论和数理统计这些老梗，2017-02-17

统计学、概率论与数理统计

- 导数和积分
- 极限：变化的终点
- 复合函数
- 多元函数与偏导数
- 极值与最值

数据科学中，常常遇到寻求曲线最值点的问题



微积分

- Numerical Variable（数值变量）
  - 数值变量和分量变量
- Supervised Learning（监督学习）
  - 常见于KNN、线性回归、朴素贝叶斯、随机森林等
- Unsupervised Learning（非监督学习）
  - 常见于聚类、隐马尔可夫模型等
- Input space , Output space and Feature space（输入空间、输出空间、和特征空间）
- Training Data and Test Data（训练集和测试集）
- Cross validation（交叉验证）

- Classifier（分类）
- Prediction（预测）
- Regression（回归）
- Ranking（排序）
- Lift curve （Lift曲线）
- Receiver Operating Characteristic Curve（ROC曲线）
- Overfitting and underfitting（过拟合和欠拟合）
- Bias and Variance（偏差和方差）
- Classification Rate（分类正确率）
- Boosting（提升方法）
- Perceptron（感知机）
- Neural Networks（神经网络）

第 10 章
统计分析的原理

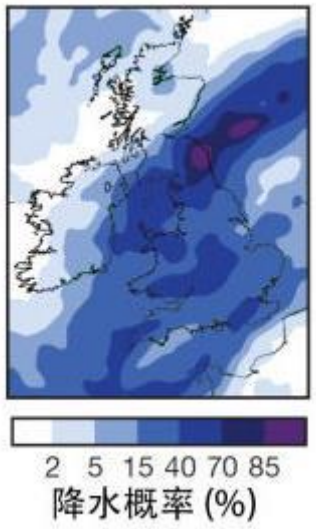
1 数据科学的数学基础

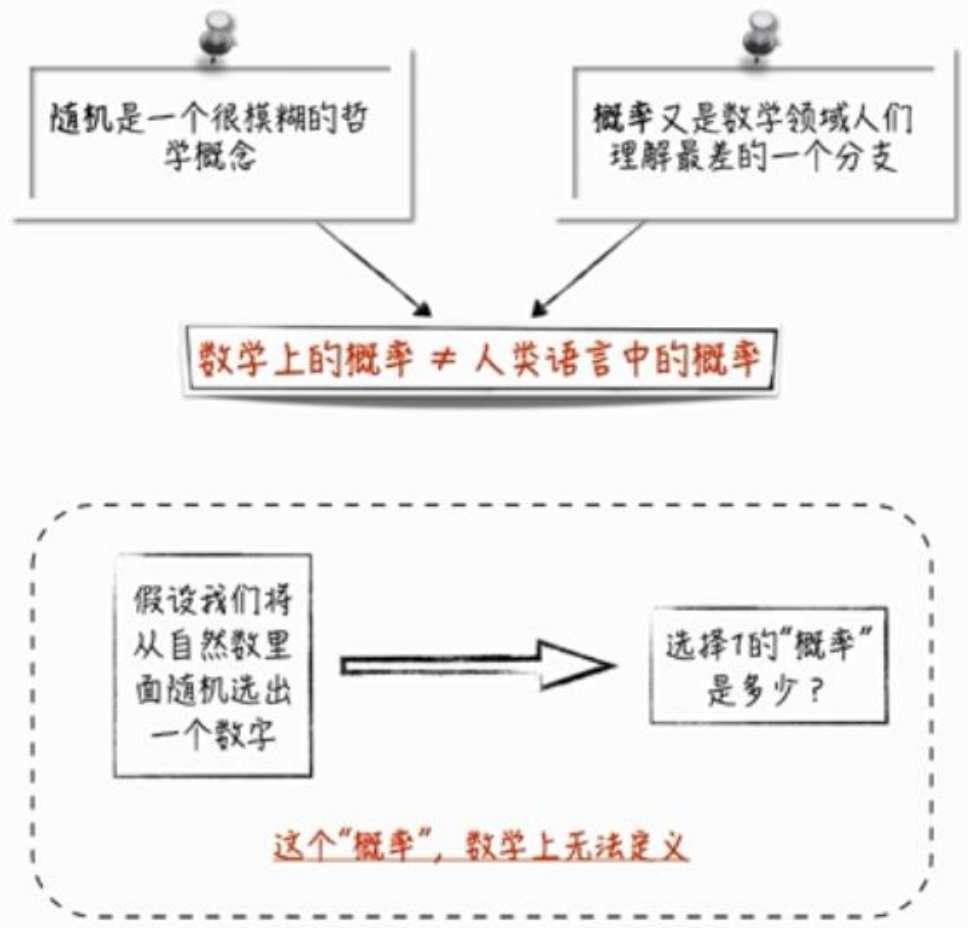
2 **概率与统计基础**

3 统计建模：线性回归模型

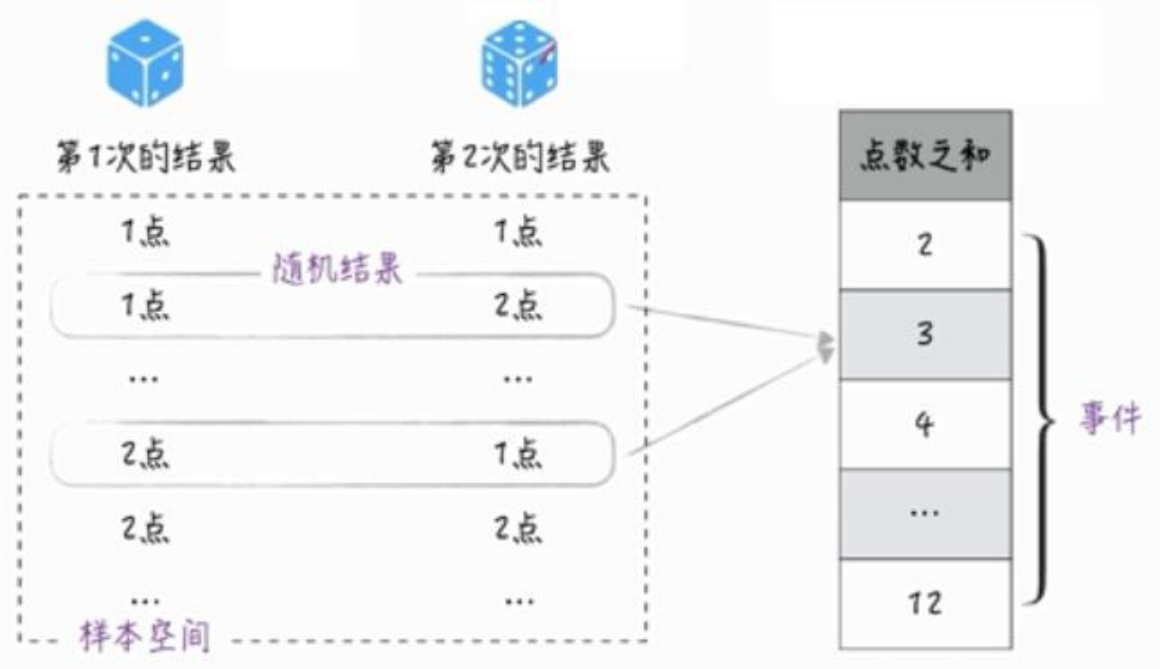
4 数据分析的工具

- 现实世界里，充满了各种随机事件
  - 彩票中奖、掷骰子的点数
- 概率是用来刻画随机的一种数学工具





什么是概率



- 在随机结果有限的情况下：
  - 定义样本空间S：所有随机结果ω组成的集合
  - 定义概率：满足如下三个条件的，从样本空间到实数的函数

$$P(A) \geq 0$$

$$P(S) = 1$$

$$P\left(\bigcup A_i\right) = \sum P(A_i)$$

什么是概率

- 假设一个班中：

	来自重庆	来自其他省份
喜欢吃辣	9	9
不喜欢吃辣	1	81

- 用事件A表示学生来自重庆
- 用事件B表示学生喜欢吃辣

已知某个学生喜欢吃辣，则他来自重庆的概率是多少？

条件概率

	来自重庆	来自其他省份
喜欢吃辣	9	9
不喜欢吃辣	1	81

- 用事件A表示学生来自重庆
- 用事件B表示学生喜欢吃辣

已知某个学生喜欢吃辣，则他来自重庆的概率是多少？

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{9/100}{18/100} = 0.5$$

$P(A) = 0.1$

他来自重庆？

$P(A|B) = 0.5$

他喜欢吃辣。
他来自重庆？

量化信息的价值！

条件概率

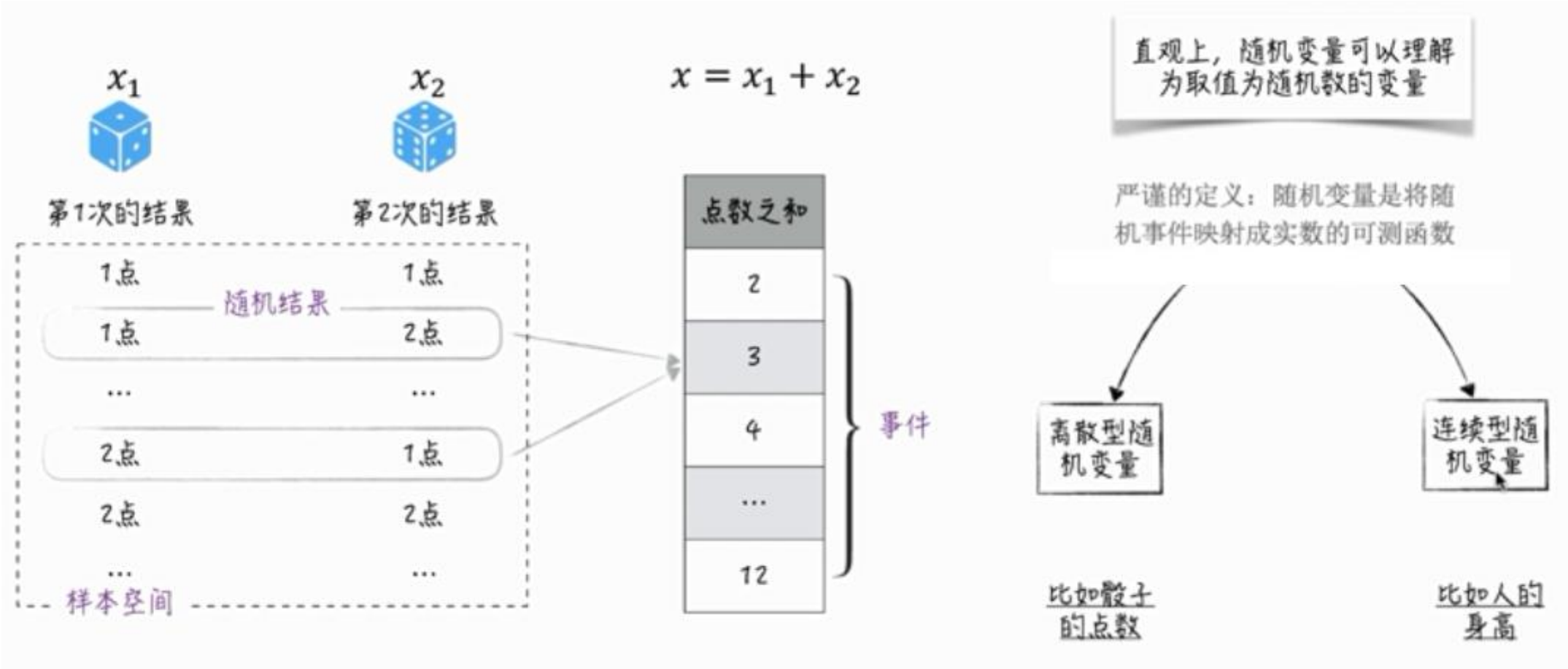
- 甲在15：00扔一个骰子
- 乙在15：01扔一个骰子
- 甲乙无任何交流联系

已知甲扔出了2点，则乙扔出5点的概率是多少？

乙扔出5点：$P(B)=\frac{1}{6}$ = 已知甲扔出2点，乙扔出5点：$P(B|A)=\frac{1}{6}$

若$P(A)=P(A|B)$，则称A, B为相互独立事件
（B事件的发生与否对A事件没有任何影响）

独立事件


x_1
x_2
$x = x_1 + x_2$
第1次的结果
第2次的结果
1点
1点
随机结果
1点
2点
…
…
2点
1点
2点
2点
…
…
样本空间
点数之和
2
3
4
…
12
事件
直观上，随机变量可以理解为取值为随机数的变量
严谨的定义：随机变量是将随机事件映射成实数的可测函数
离散型随机变量
连续型随机变量
比如骰子的点数
比如人的身高

随机变量

离散型随机变量

对于离散型的随机变量，使用概率分布函数来刻画它

x_1

$$P(x_1 = i) = 1/6; i = 1, \ldots, 6$$

连续型随机变量

对于连续型的随机变量，使用概率密度函数来刻画它

$$P(a \le x \le b) = \int_a^b f_x(t)dt$$

累积分布函数

$$C_x(a) = P(x <= a)$$

期望

$$E[x] = \sum_i P(x = x_i)x_i$$

$$E[x] = \int xf(x)dx$$

方差

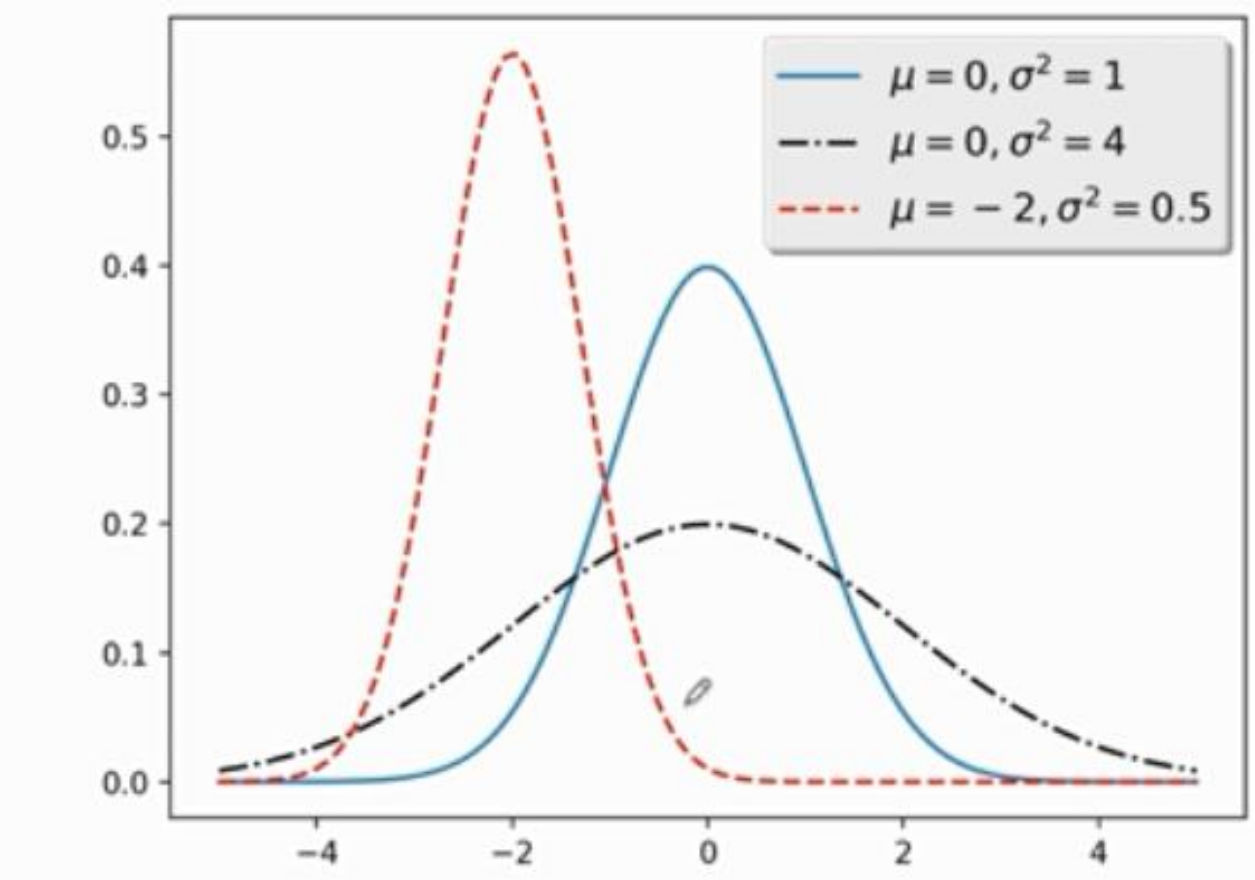
$$Var(x) = E[(x - E[x])^2]$$
$$= E[x^2] - (E[x])^2$$

协方差

$$Cov(x, y) = E[(x - E[x])(y - E[y])]$$
$$= E[xy] - E[x]E[y]$$

刻画随机变量的方法

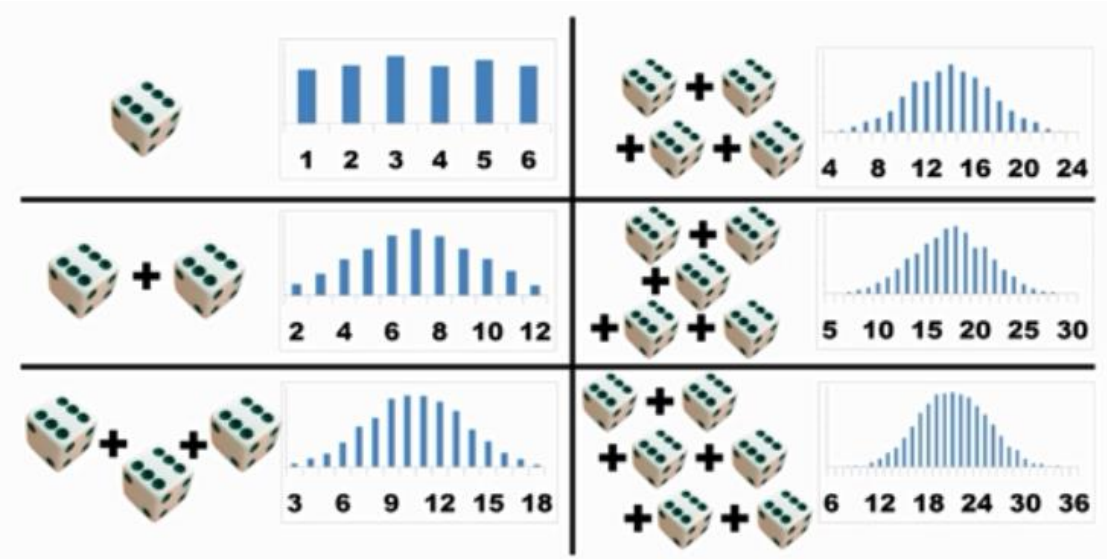
- 正态分布又称高斯分布，是最为重要的一种概率分布



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

正态分布

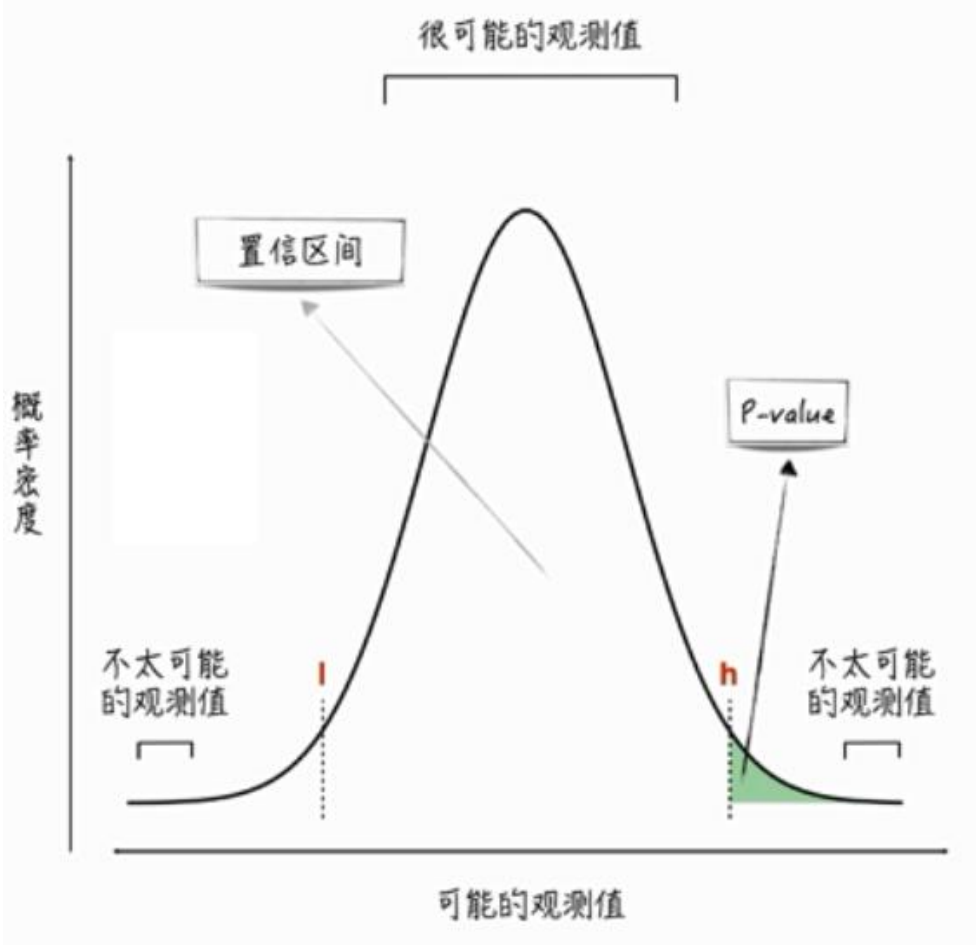
- N个独立同分布的随机变量相叠加，将越来越接近正态分布
- 在实际中，一个随机现象往往是多个随机因素的叠加
- 很多随机现象的分布都可以用正态分布来描述
  - 大学英语考试（College English Test）
  - 高斯板



中心极限定理

- 置信区间：概率值等于a，且以期望为中心的对称区域（在实际中a常常等于0.95）

- 对于置信区间的两个边界值，它们的P-value为$\frac{(1-a)}{2}$



置信区间

第 10 章
统计分析的原理

1 数据科学的数学基础

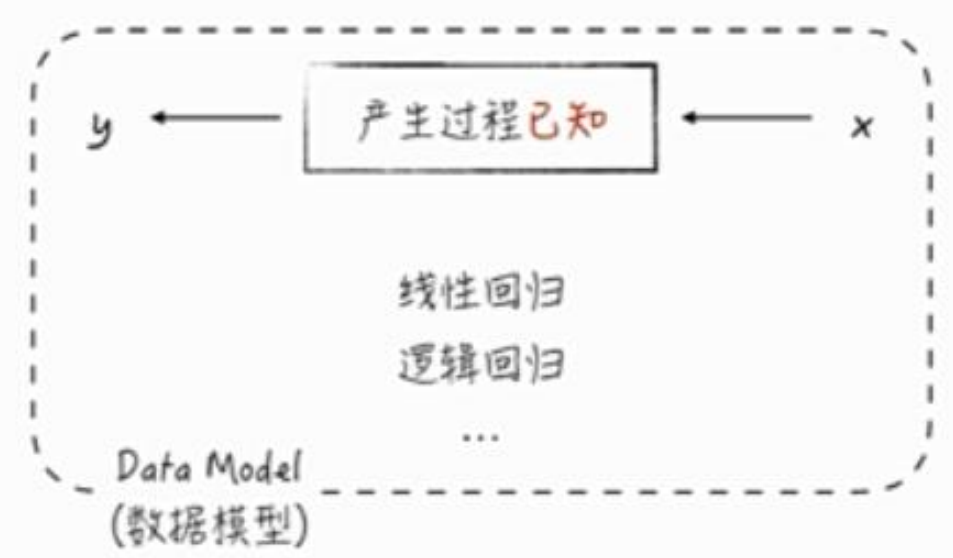
2 概率与统计基础

3 **统计建模：线性回归模型**

4 数据分析的工具

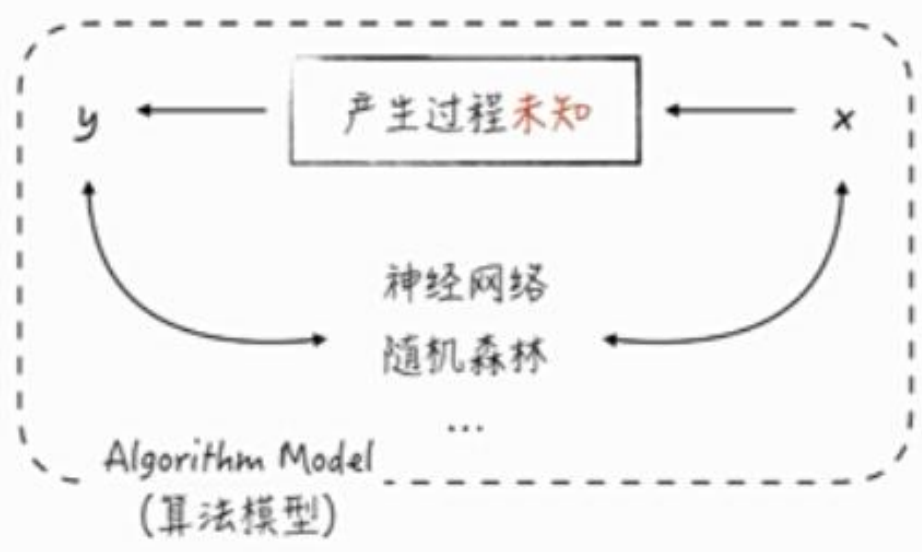
- 注重用数学的方法来搭建模型
  - 理论更加扎实
  - 模型容易理解和控制

统计模型



- 注重用工程的方法来搭建模型
  - 可以处理的场景更多
  - 模型的预测效果更好

机器学习模型



统计模型与机器学习模型

建模五部曲

以下是上帝视角

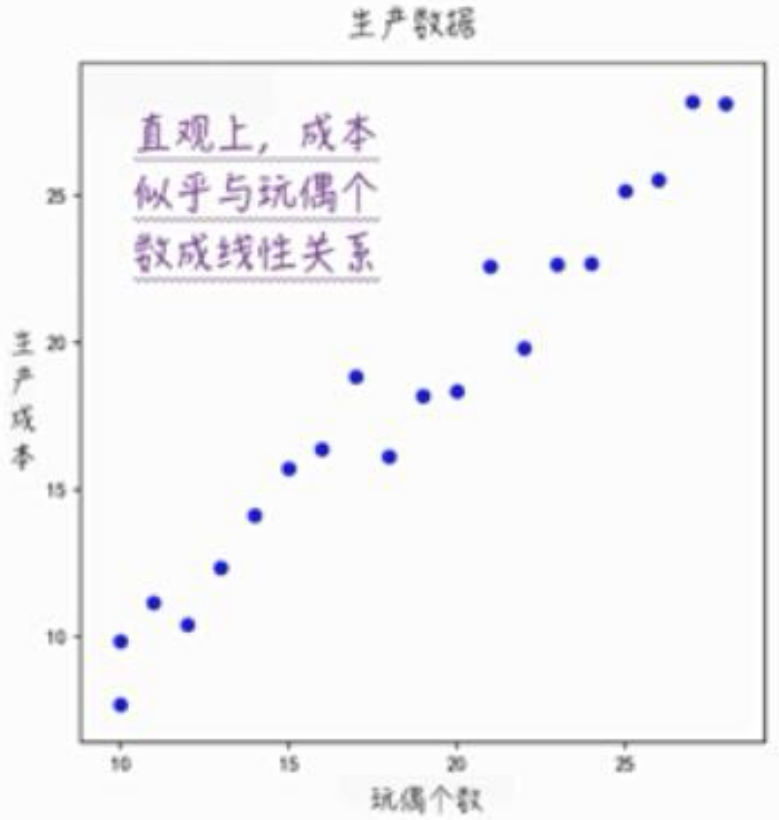
数据由"上帝之手"按如下的数学公式生成：

$$y_i = x_i + \varepsilon_i$$

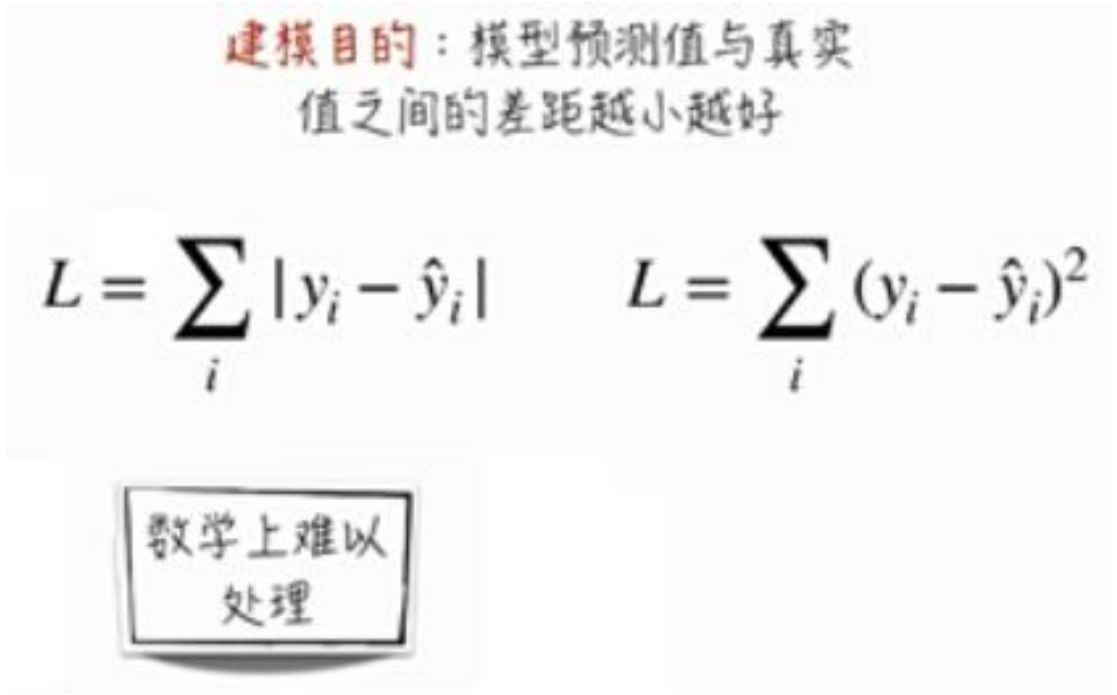
- 玩偶的单位成本等于1
- ε_i 表示生成的随机成本，它服从期望等于0、方差等于1的正态分布
- 随机成本与玩偶个数相互独立

生产记事本

日期	玩偶个数	成本	第几天
04/01	10	7.7	1
04/02	10	9.87	2
04/03	11	10.87	3
04/04	12	12.18	4
04/05	13	11.43	5
04/06	14	13.36	6
04/07	15	15.15	7
04/08	16	16.73	8
04/09	17	17.4	9
...	...	...	...



示例：线性回归



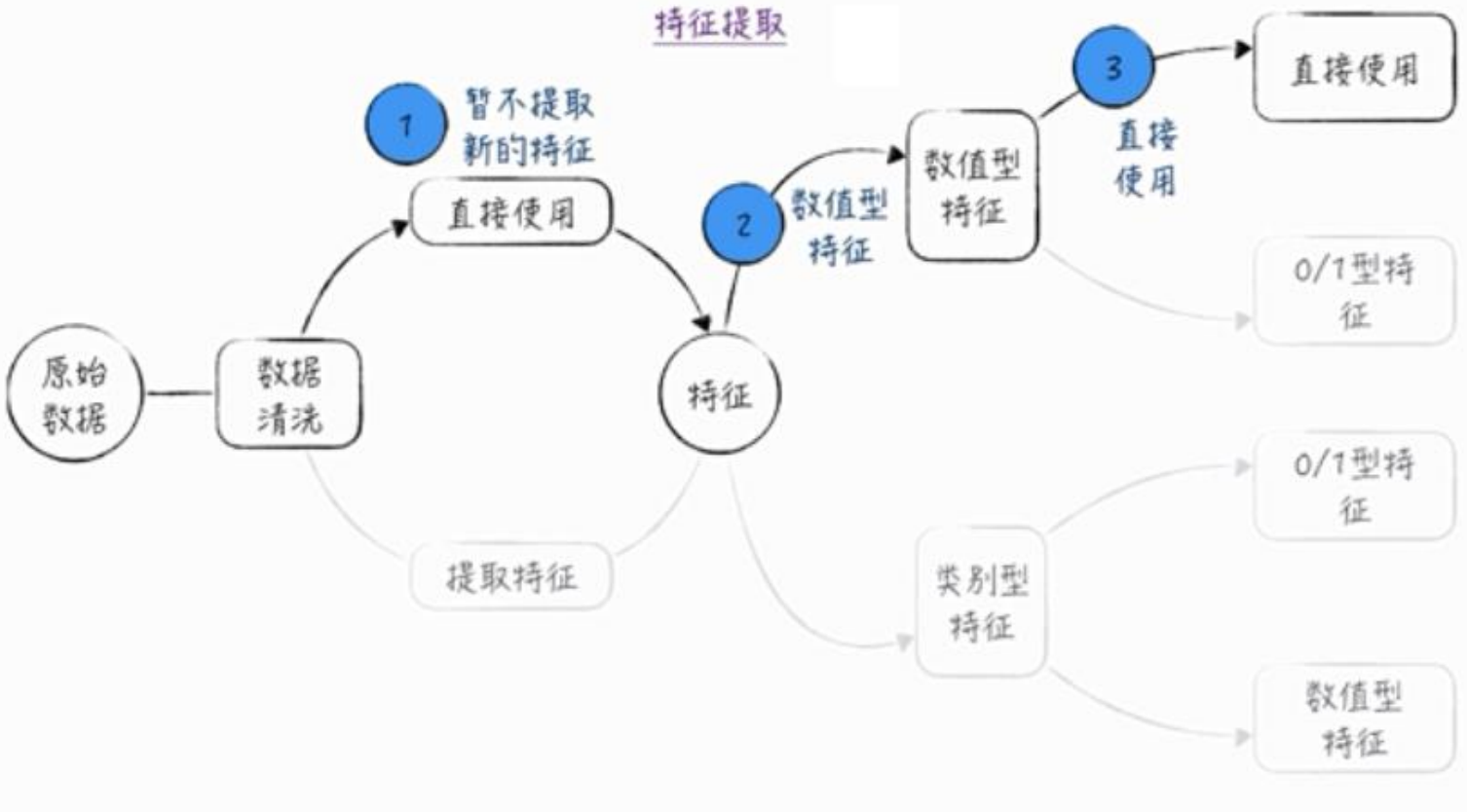
- 损失函数是统计模型的核心：
  - 即使相同的模型形式，不同的损失函数对应着不同的模型效果
- 如果把模型看成是人工智能，则损失函数就是它的价值观
- 模型的损失函数是由人定义的
  - 从理论上讲，技术已不再中立
  - 模型已经成为今天生活的主宰

线性回归：损失函数

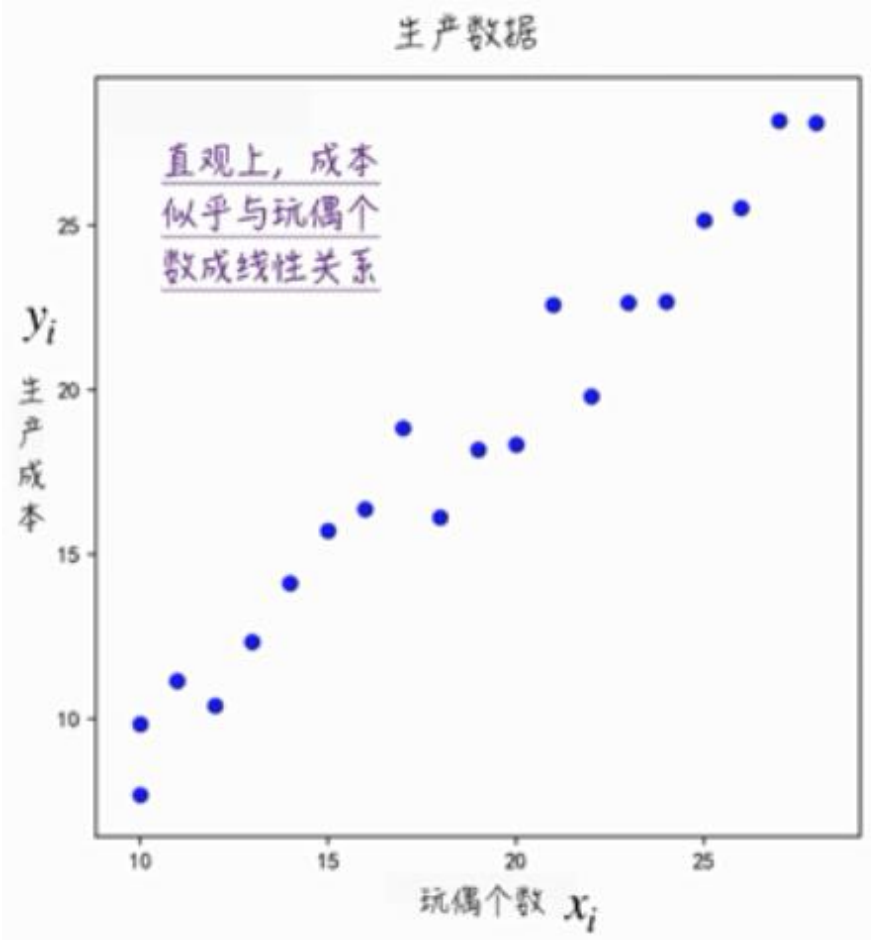
生产记事本

日期	玩偶个数	成本	第几天
04/01	10	7.7	1
04/02	10	9.87	2
04/03	11	10.87	3
04/04	12	12.18	4
04/05	13	11.43	5
04/06	14	13.36	6
04/07	15	15.15	7
04/08	16	16.73	8
04/09	17	17.4	9
...	...	...	...

直接使用



线性回归：特征提取



确定模型形式：线性模型

$$\hat{y}_i = ax_i + b$$

结合模型的损失函数

$$L = \sum_i (y_i - \hat{y}_i)^2$$

得出模型参数的估计公式

$$(\hat{a}, \hat{b}) = argmin_{a,b} \sum_i (y_i - ax_i - b)^2$$

线性回归：模型形式和参数估计

对于回归问题，常用的模型评估指标有两个：

- 均方差（MSE）：预测值与真实值的平均差距
- 决定系数（R^2）：数据变化被模型解释的比例

均方差

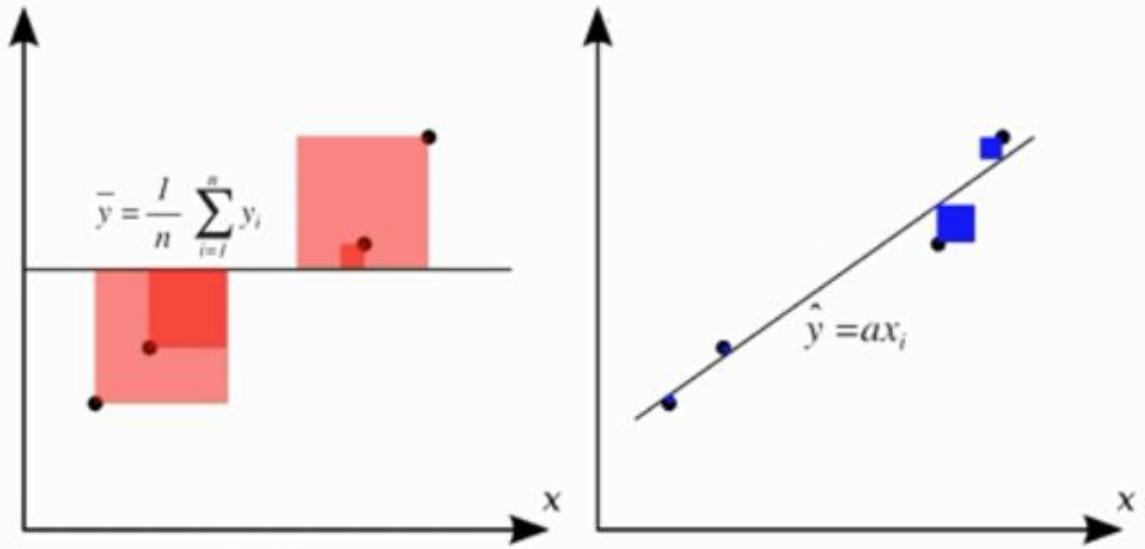
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}L$$

决定系数

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

决定系数示意图



$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

线性回归：模型效果评估

第 10 章
统计分析的原理

1 数据科学的数学基础

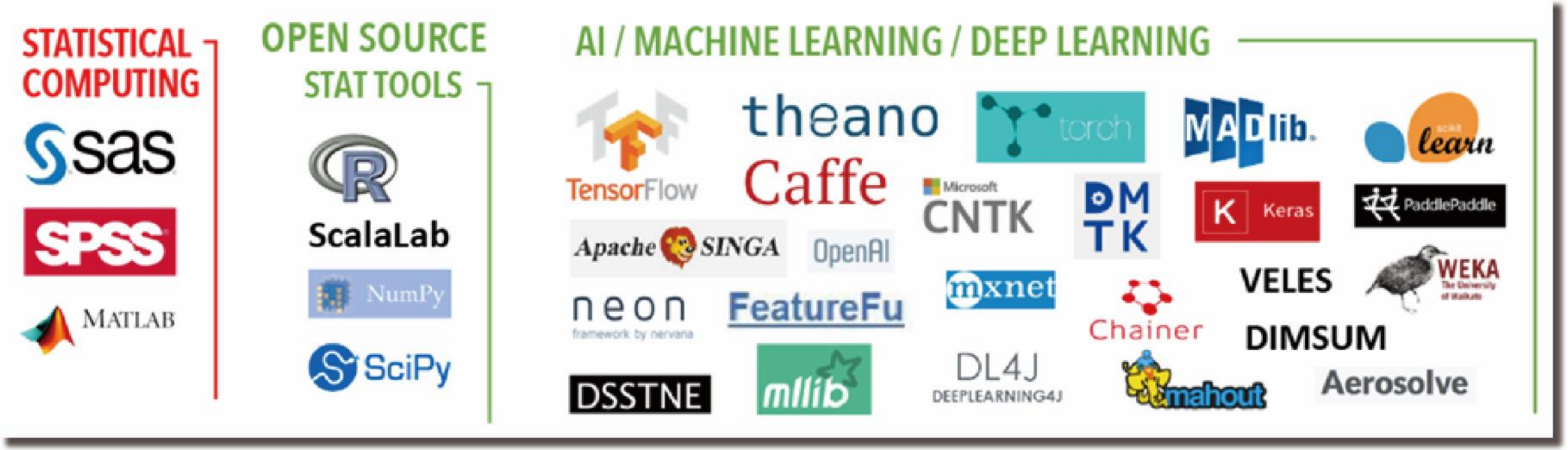
2 概率与统计基础

3 统计建模：线性回归模型

4 **数据分析的工具**

Python，R，SQL

编程语言


STATISTICAL COMPUTING
sas
SPSS
MATLAB
OPEN SOURCE STAT TOOLS
R
ScalaLab
NumPy
SciPy
AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow
theano
Caffe
torch
MADlib.
learn
Microsoft CNTK
DMTK
Keras
PaddlePaddle
Apache SINGA
OpenAI
mxnet
WEKA
neon
framework by nervana
FeatureFu
Chainer
VELES
DIMSUM
DSSTNE
mllib
DL4J
DEEPLEARNING4J
mahout
Aerosolve

数据分析工具

第 10 章
统计分析的原理

1 数据科学的数学基础

2 概率与统计基础

3 统计建模：线性回归模型

4 数据分析的工具