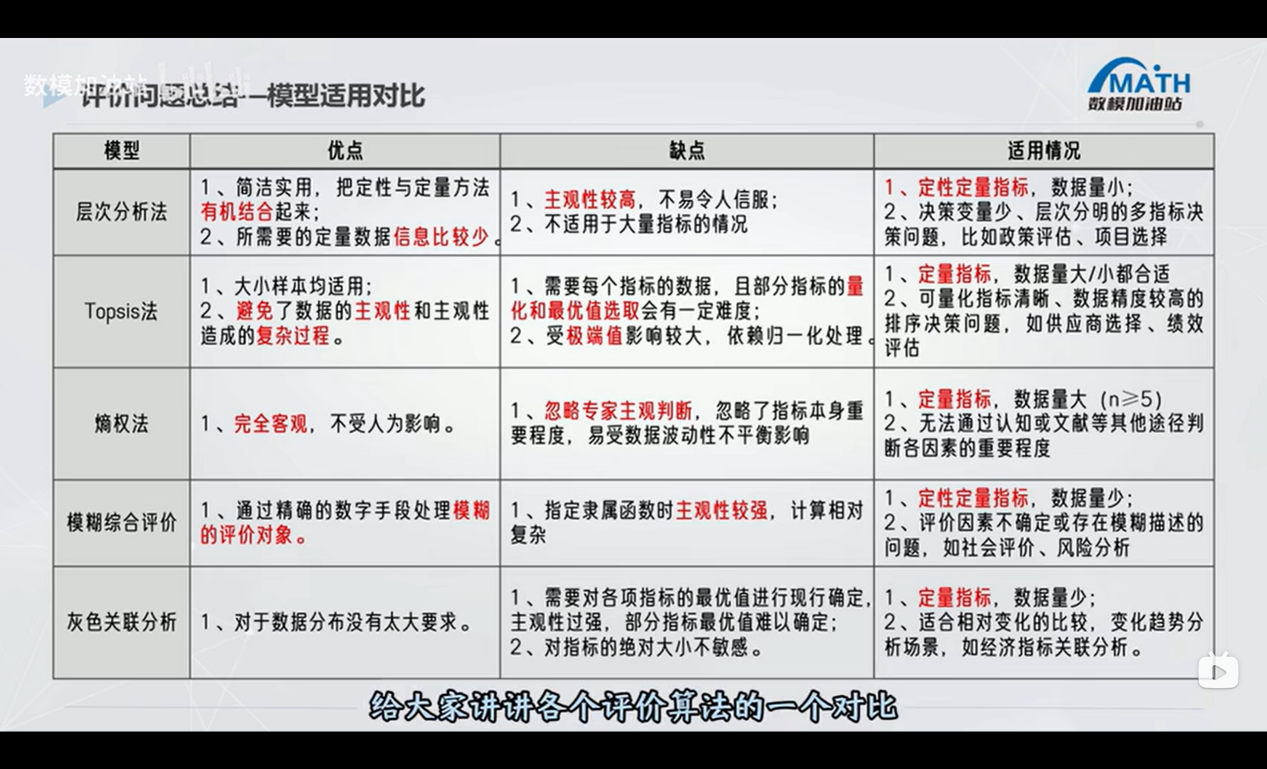
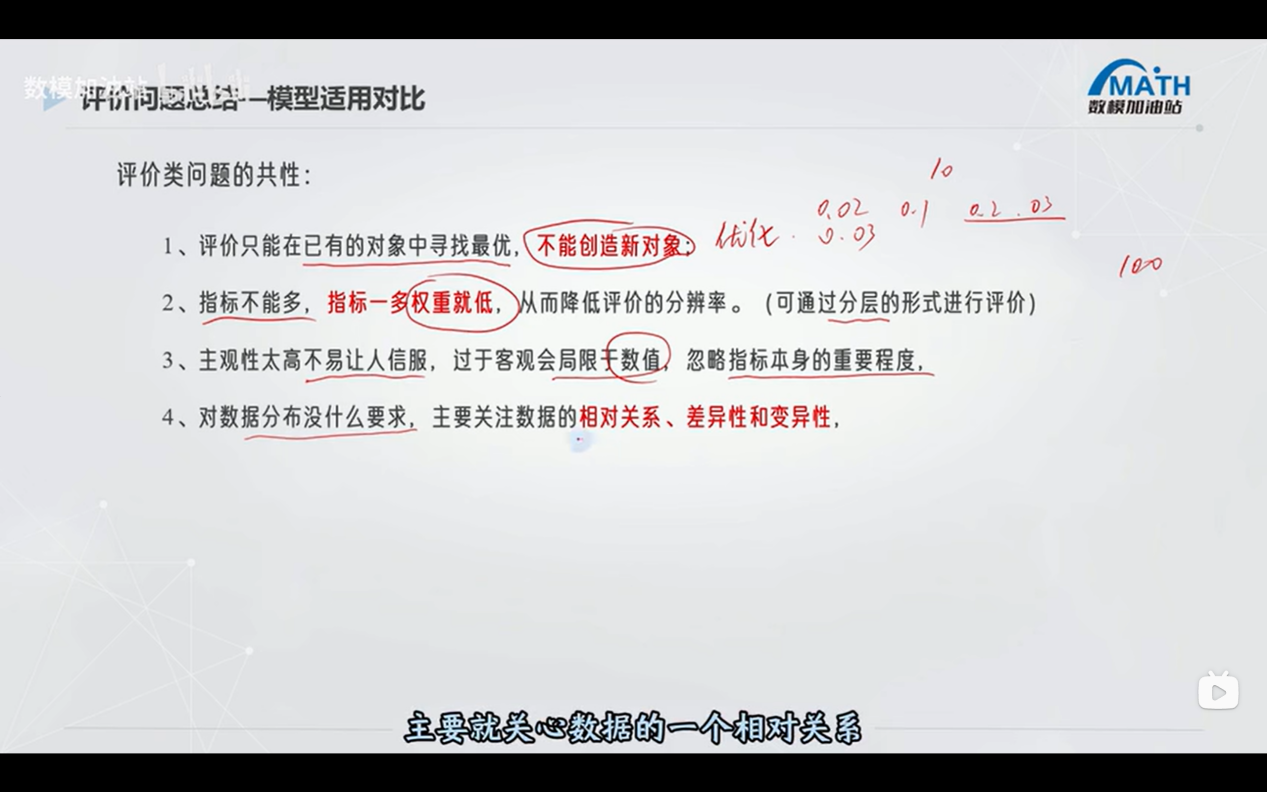
# Day1 2025年8月25日

# 评价类算法



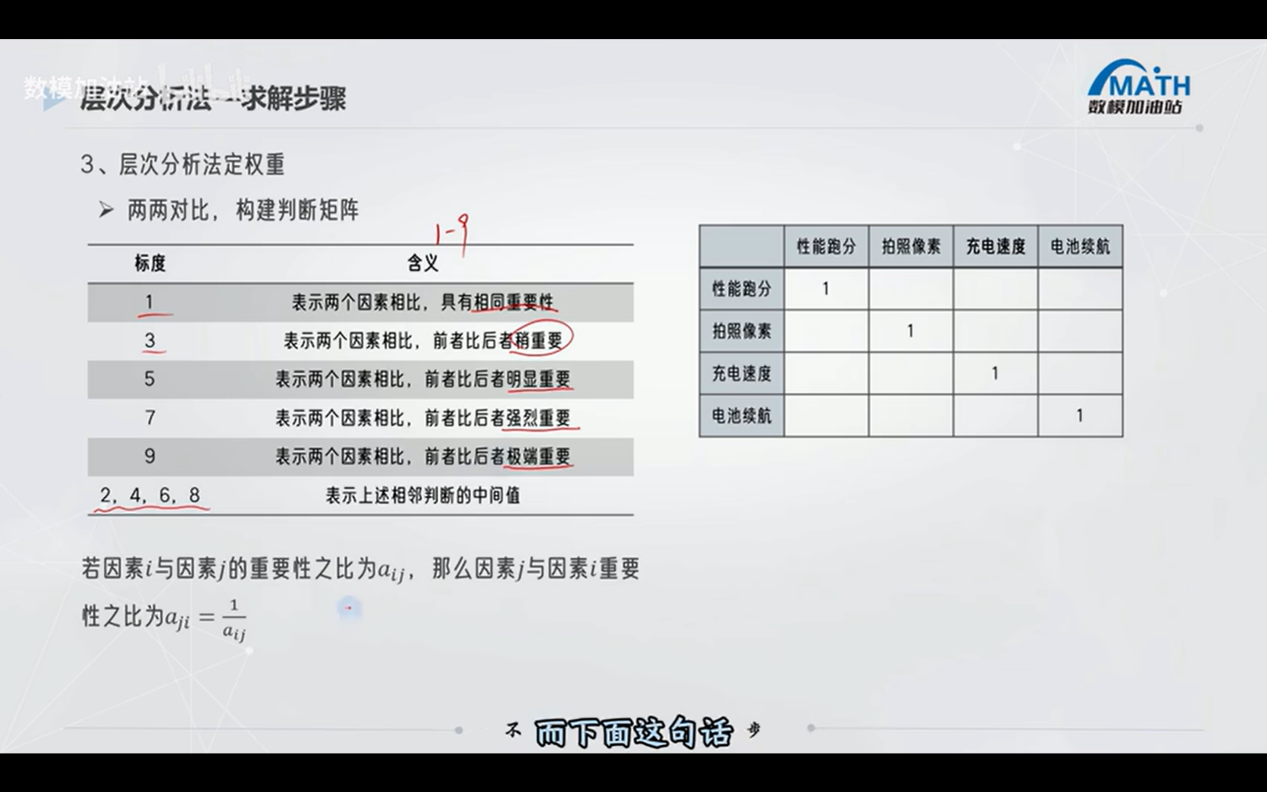


## 层次分析法

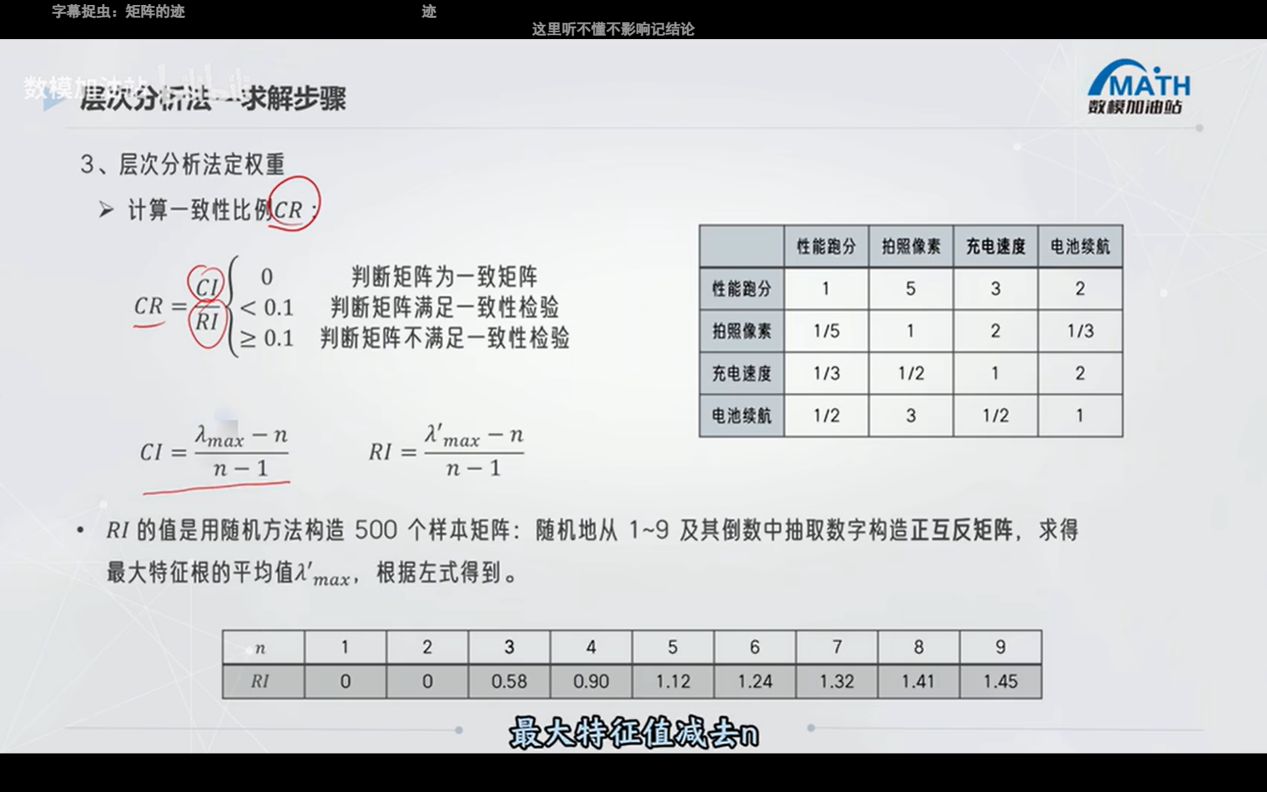
**人为打分，较为主观地得出标准权重**

归一化，统一量纲

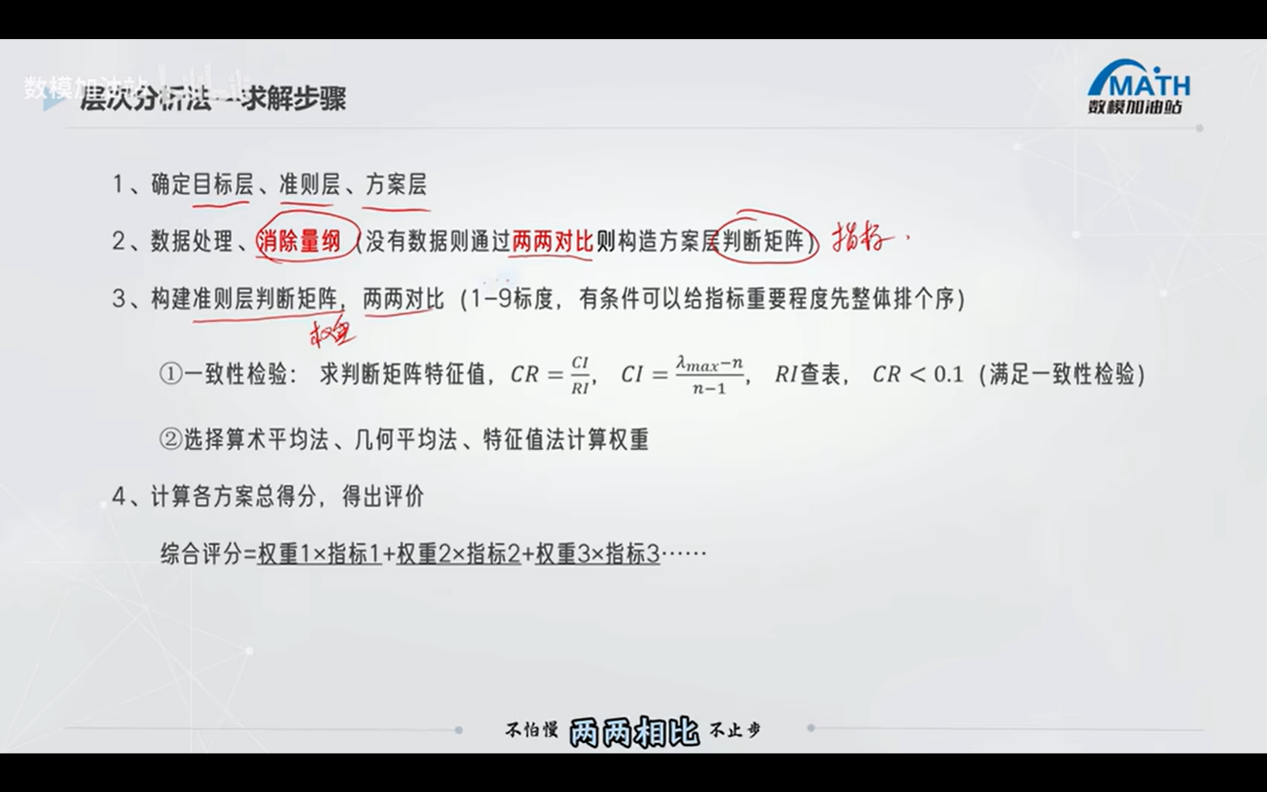
制作判断矩阵



检验矩阵一致性



然后将构造的判断矩阵进行列归一化，然后每列加和除以n得到权重



## TOPSIS

**不是一个计算标准权重的方法，而是通过已有原始数据和标准权重数据，得到最优/劣方案，从而计算方案优劣并排序的方法。**

### 输入原始矩阵

### 消除量纲，进行标准化处理

### 通过某些方法，如层次分析法等得到权重向量

### 标准化矩阵与权重向量相乘得到结果矩阵

### 从结果矩阵中构造出理想解和负理想解

### 然后计算每个方案与最优目标、最劣目标的欧氏距离，进行排序

### 就能对方案的好坏进行得到结论了

## 熵权法

**一种基于大数据本身的信息熵性质计算权重的方法，更客观，且要求数据满足一定要求**

### 数据收集与整理：构建评价矩阵（样本 × 指标）

### 数据标准化：消除不同指标间的量纲影响

### 计算指标的信息熵：基于标准化数据计算各指标的熵值

### 计算指标权重：根据熵值计算各指标的权重

### 计算综合得分：利用权重对各样本进行综合评价

# Day2 2025年8月26日

## 模糊综合评价

**通过将模糊不清、难以量化的评价指标（如 “服务质量”“产品满意度”）转化为可计算的数值，比如成绩低于60是0，高于90是1，60~90之间写一个矩形隶属函数。**

区别于经典数学中 “非此即彼” 的精确集合（如 “身高≥180cm 属于高个子”），模糊集合允许 “亦此亦彼” 的中间状态。例如 “年轻人” 是一个模糊集合：20 岁属于 “完全年轻”，35 岁属于 “部分年轻”，50 岁属于 “基本不年轻”。

需要三个关键集合 评价因素集 U 评语集 V 权重向量 A

### 确定评价对象与评价因素集 U：

### 确定评价什么，有哪些标准

### 建立评语集 V：

### 评价结果的等级标准，比如 优良中差 很满意满意一般不满意

### 确定各因素的权重向量 A：

### 根据实际需要选择一个确定权重的方法，专家法、层次分析法、熵权法等等

### 构建隶属度矩阵 R：

### 可以主观打分（专家打分法）、模糊统计（统计有大量打分样本）、隶属函数法（选择合适的隶属函数（如三角形、梯形、正态分布函数）计算隶属度)

### 模糊合成运算 综合评价结果向量 B

### 解析综合评价结果

写一个模糊综合评价求解问题的python程序，题目如图。要求读入本地的excel数据，对每个标准进行梯形隶属函数的处理，求出所有隶属值，再用熵权法确定每个标准的权重，最后求综合评价

## 灰色关联分析

**分析已有序列和参考序列（理想方案、目标值）的相似程度（关联度），在灰色场景下，对方案进行排序，将各指标对整体评估的影响进行量化分析**

白色：信息完全可知，系统行为完全可预测，譬如单摆系统

黑色：信息完全不可知，系统行为完全不可预测，我们对宇宙的起源一无所知

**灰色：我们知道一部分信息，对系统行为掌握了一部分规律，天气预报**

核心是计算变量之间的关联度：大气状态、天气系统、历史参考、边缘干扰——天气预报

对于一组数据，其中有一组母数据和几组子数据，先进行归一化处理，再依次用每组子数据比较母数据得到差值，用差值构造一个差值矩阵A，从中找到最大值和最小值，并带入公式得到关联度计算公式，依次处理每个子数据，得到该子数据对母数据的关联度矩阵

### 准备原始数据矩阵

### 归一化处理（消除量纲差异）

### 构造差值矩阵 A（量化子母数据差异）

### 提取差值矩阵的 “全局最大值” 和 “全局最小值”

### 计算关联度系数，构建关联度矩阵

# 概率论与数理统计

## 一元线性回归

**通过机器学习，利用原始数据点，拟合出一条自变量与因变量的直线**

**用决定系数来判断直线是否拟合，越接近1，就越说明每个点离直线的距离越小(实际上通过距离的平方之和计算)**

### 明确变量，收集数据集

### 假设出线性关系

### 最小二乘法确定最优拟合直线

### 得到回归方程

### 决定系数 评估拟合效果，越接近1则拟合效果越好

### 应用，代入新的自变量，计算对应的因变量、

### 或判断影响方向（正相关、负相关）和影响强度

## Day 3 2025年8月27日

## 多元线性回归

**这东西完全不是多个一元线性回归的简单堆砌，因为每个特征之间系数是不同的。**

**核心是 控制其他变量，估计单个变量的净效应，并 全局优化所有变量的联合解释力。**

**有的时候特征数据种类太多了，我们也没法保证每个特征都是有用的，这个时候可以对每个特征都进行一次一元线性回归的分析，查看是否线性相关，筛选一批相关的特征，然后正常做多元线性回归的分析。**

### 明确研究目标与变量定义

### 数据收集与预处理，比如对每个特征一元线性回归，清洗噪声特征、缺失值处理

### 构建多元线性回归模型，利用计算

### 模型结果分析与解读，整体拟合效果、自变量显著性、残差分析

## Day4 2025年8月28日

## 非线性回归

**有的时候拟合方程是非线性的，比如人的生长曲线等等，如果硬要使用线性方程拟合就无法体现一些关键信息，这就是线性回归的局限性，这里我们要用到非线性回归。**

**仍然是找y和x的关系，比如多项式回归**

**指数模型 对数模型等等**

**线性回归一般用最小二乘法即可估计参数，而非线性回归的参数估计较为复杂。对样本量要求也高，还容易不稳定、过拟合**

### 问题定义与数据准备

### 选择合适的非线性模型：

实际解题的时候注意用什么函数适合拟合当前的数据，可以先画个散点图、折线图，然后看变化，选好函数才能顺利拟合。

### 确定初始参数猜测：

非线性回归需要初始参数作为迭代起点，不合理的初始值可能导致拟合失败

### 执行非线性拟合

### 模型评估与验证

## 灰色预测分析

**主要用于小数据建模**

对于白色系统，我们往往清楚其机理和过程，比如我们知道饭后血糖变化规律呈单峰非对称曲线，知道年龄与身高的指数变化规律，但是对于某些灰色系统，我们的了解有限，无法预测大致变化规律，数据点也很少，没法暴力拟合出一条线，这个时候我们要使用灰色预测分析

**灰色总是意味着部分系统已知，部分系统未知的研究：信息不完整，样本规模小、处理不确定性**

**仍然可用天气预报举例子：我们知道今天的温度、湿度、云层的运动，但是不知道大气每个分子的运动。**

再比如：人口预测上，我们知道[2020,2025)四年的某城市人口数量，要是我们知道每天的人口就可以有1800个数据非线性回归了，现在只能依靠这四个点进行灰色预测分析未来人口

### 灰色预测（如最常用的 GM (1,1) 模型）的核心是通过数据累加生成、建立微分方程、求解模型参数、逆运算还原预测值

### 数据检验，观察是否可以建模

### 累加生成，可以减弱随机性

### 构建GM(1,1)模型，建立一阶微分方程

### 利用最小二乘法求解

### 逆运算，还原出原始序列的预测值

### 模型检验

## 时间序列预测

**利用数据在时间维度上的 “历史规律”（如趋势、周期性、波动性），预测未来某个时间点的数值。**

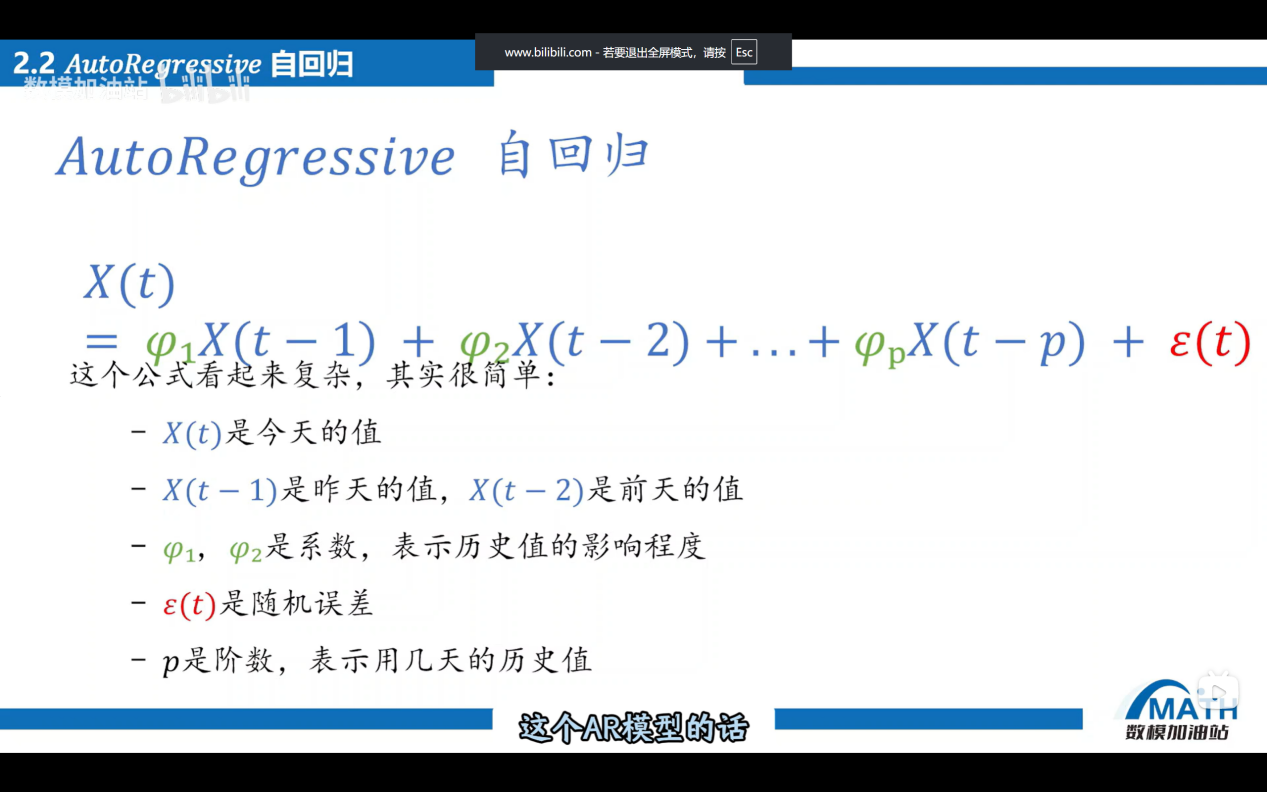
所有预测逻辑都围绕 “时间先后带来的关联性” 展开，而非单纯用 “时间值”（如 2024 年 1 月、第 10 天）直接计算结果。

**比起传统的回归，并非单纯使用时间的值（x互相独立），还考虑了时间的前后关系（过去影响未来），体现了随时间的变化。**

**分离不同性质对时间序列的影响，时间序列的变化由趋势、周期、季节等共同决定**

**平稳性的核心定义是：序列的均值、方差、自协方差不随时间推移而变化（即 “无趋势、无周期性波动、波动幅度稳定”）。**

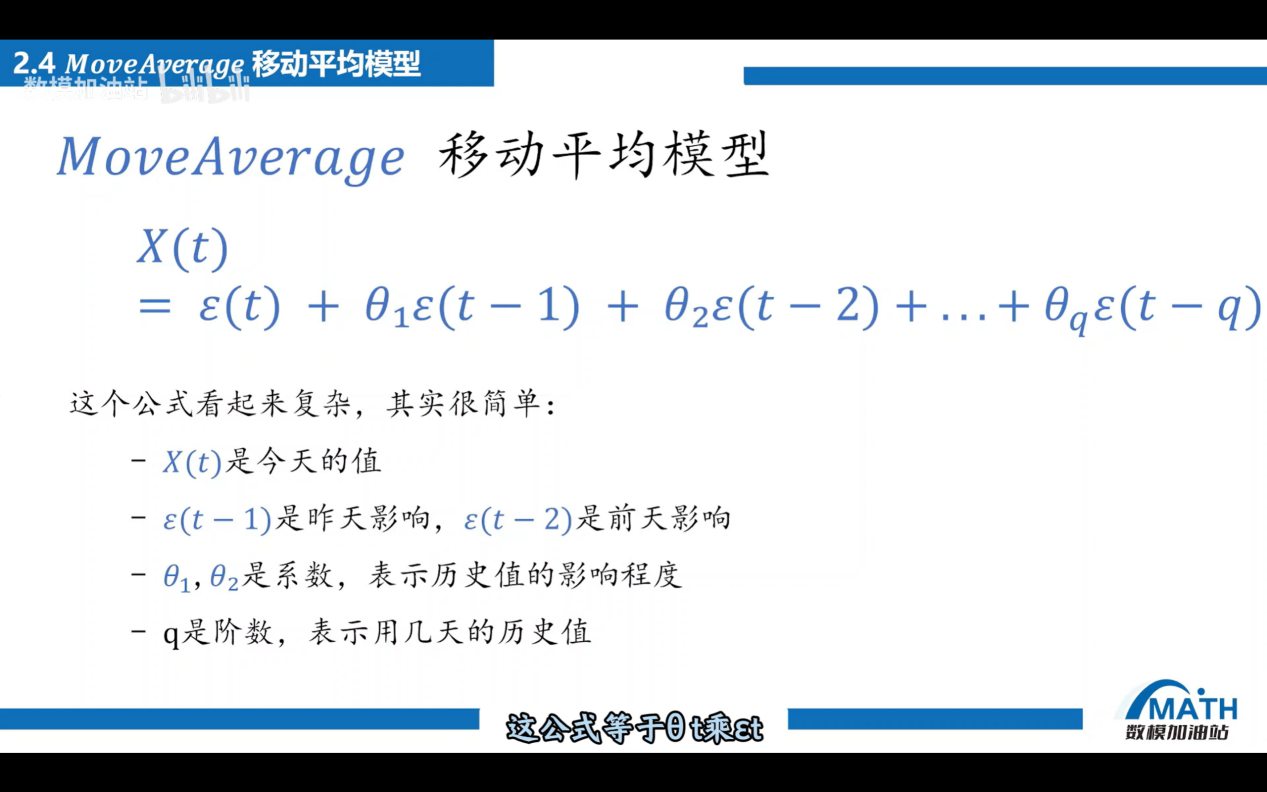
自回归：通过过去决定今天，今天的值取决于过去的值



有点像斐波那契

差分：将非平稳序列转化为平稳序列，比如对于股票，其价值是非平稳序列，但是收益率是平稳序列，我们只能用时间序列回归模型对平稳序列做研究

移动平均模型：



同样今天受到过去的影响，但是自回归采用昨天的**值**，移动平均采用昨天的**影响。**

### 数据准备和可视化

### 平稳性检验

### 确定差分次数

### 确定AR和MA阶数

### 模型拟合

### 预测计算

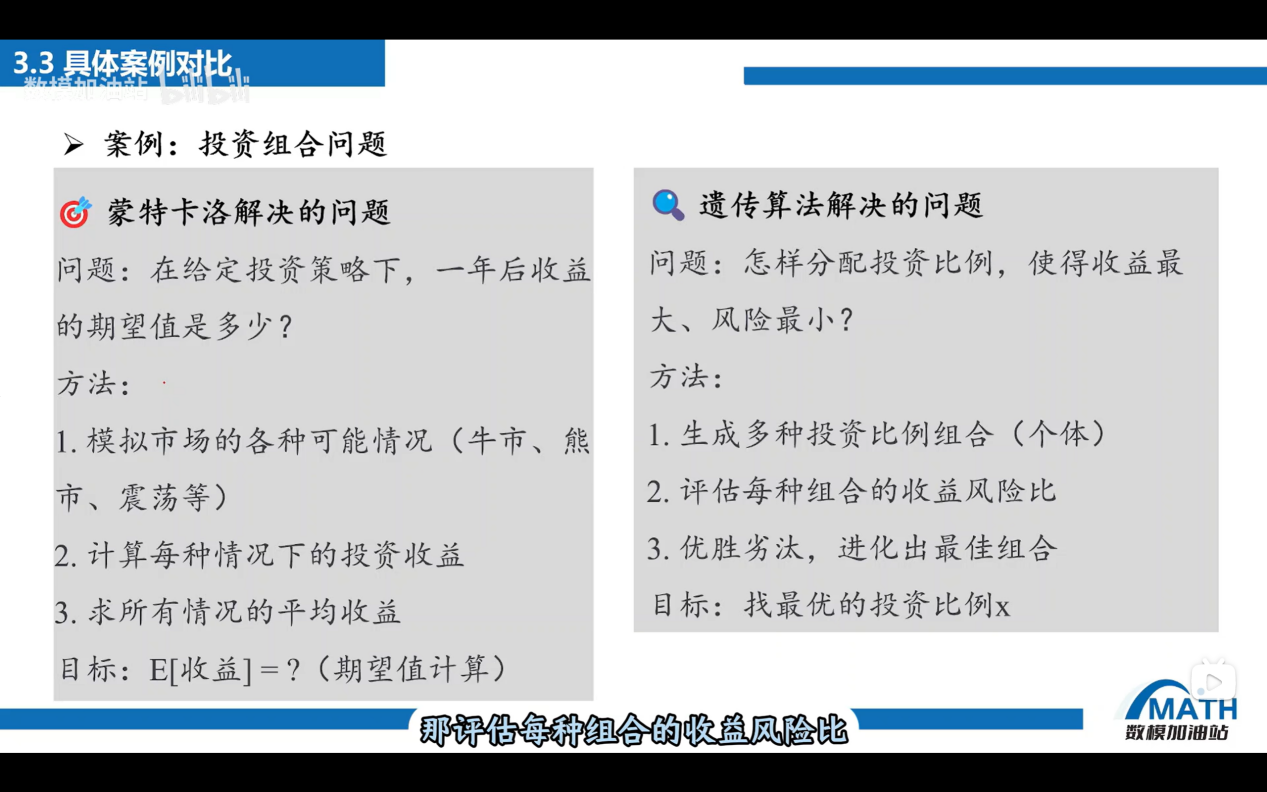
# Day5 2025年8月29日

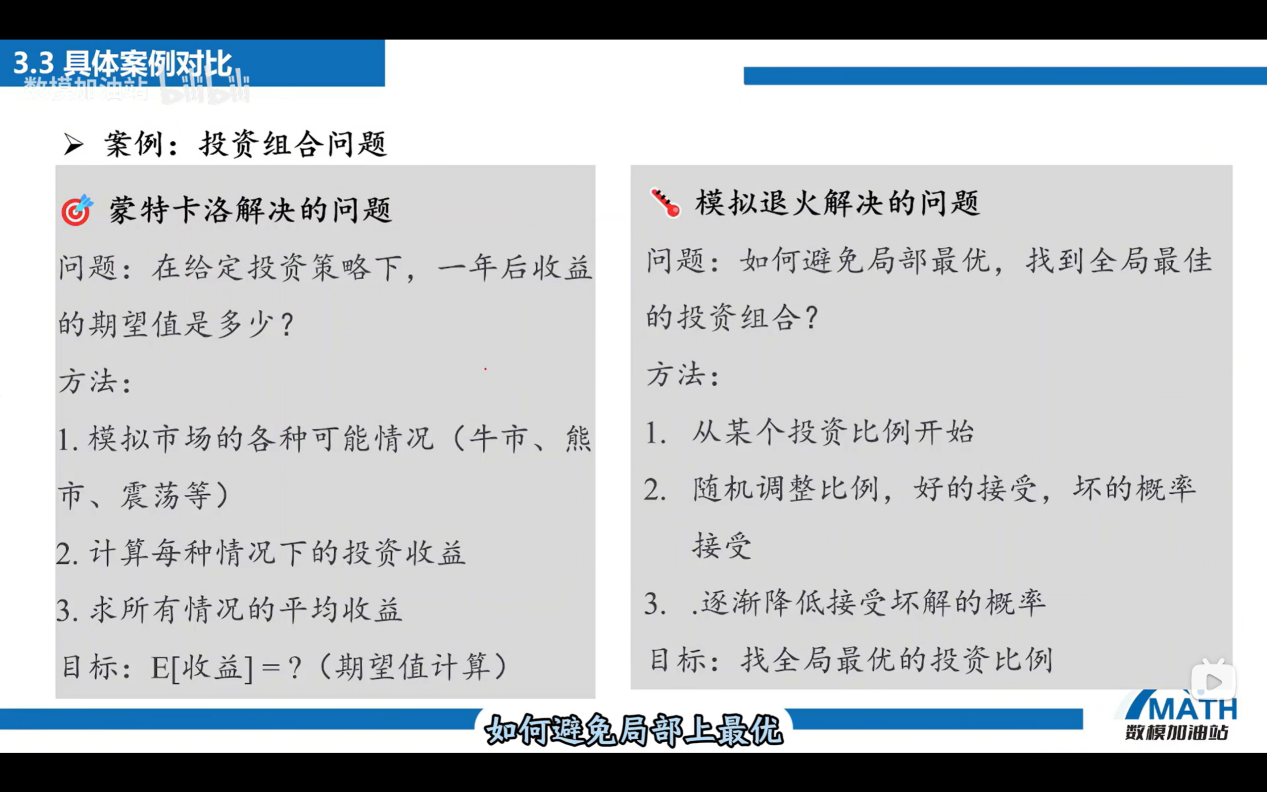
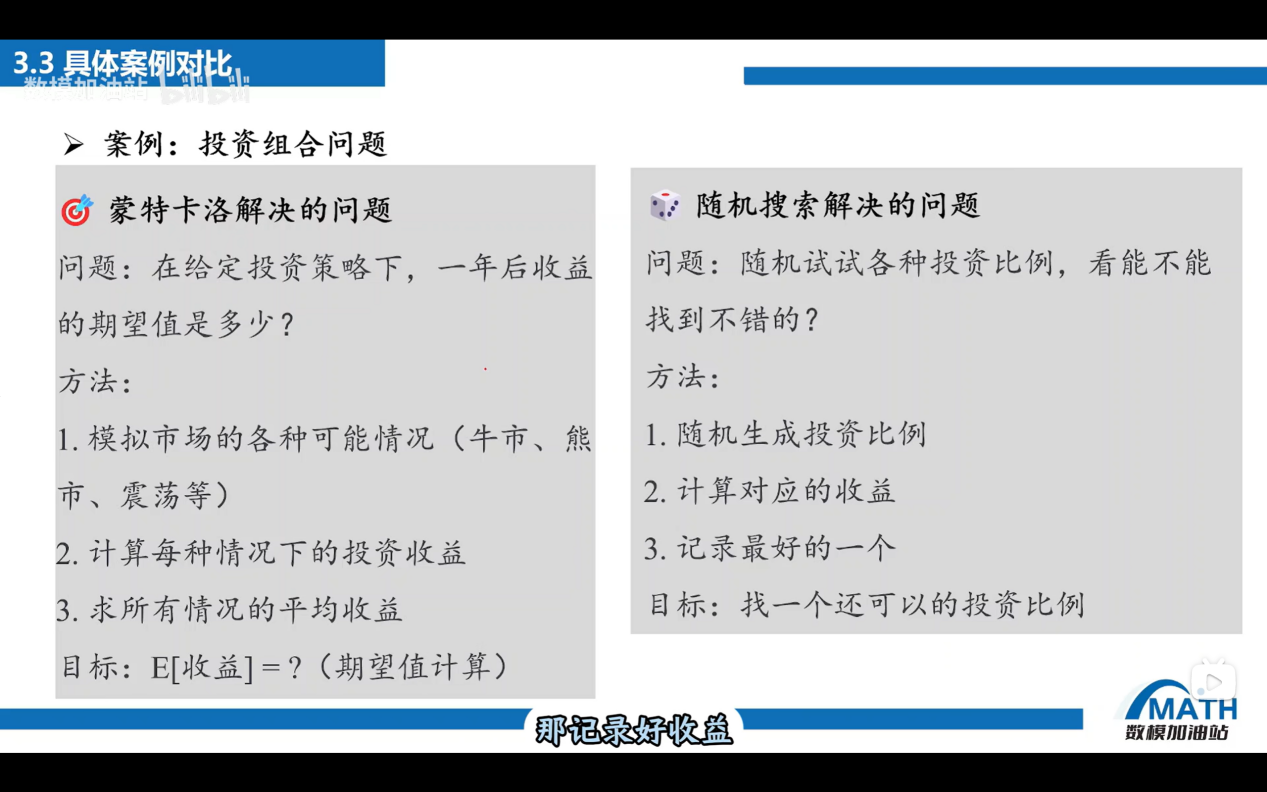
## 蒙特卡洛

没什么好说的，打一百万个点，然后算满足条件的点的数量，除以一百万就是结果；

这两天筑基了一下，专门学了np，pd，lin库的一些用法。

适合估计一些充满随机性的系统，太随机了不太好建别的模型，比如投资策略，用蒙特卡洛计算出期望收益





和其他的很暴力的优化算法，有类似这样的区别。**蒙特卡洛只是一种期望值计算方法，而不是优化算法。**

### 构建概率模型：将问题转化为概率/期望值计算问题

### 随机抽样：根据概率模型生成大量样本

### 统计计算：对样本统计、计算相关量

### 结果估计：用样本统计量估计目标量

## 马尔可夫算法

根据事物目前的状态，预测未来各个时刻的状态

比如这样

|  |  |  |  |
| --- | --- | --- | --- |
| 今天\明天 | 晴天 | 阴天 | 雨天 |
| 晴天 | 0.7 | 0.2 | 0.1 |
| 阴天 | 0.3 | 0.4 | 0.3 |
| 雨天 | 0.2 | 0.3 | 0.5 |

首先，分析对象要满足马尔科夫性，即**未来的状态只取决于现在，与过去无关**

**得到类似于以上表格的矩阵后，再确定初始状态，就可以用简单的矩阵运算计算特定时间的状态概率了**

### 状态定义

### 转移矩阵估计：量化 状态->状态 的可能性

### 初始状态分布

### 根据状态转移概率矩阵进行预测

### 模型验证

# 概率论

## 条件概率

其实就是这个东西：

在数学建模中这几种应用

| **模型类型** | **抽样规则** | **概率变化特点** | **核心场景** |
| --- | --- | --- | --- |
| 波利亚模型 | 放回 + 新增同色球 | 条件概率自增强，边际概率不变 | 传染病、推荐系统、进化 |
| 普通有放回抽样 | 仅放回，不新增 | 概率始终固定 | 质量检测（如产品合格率） |
| 无放回抽样 | 不放回，总球数减少 | 概率逐渐变化（无自增强） | 抽奖（如抽中奖后奖券作废） |

还可能混合点什么二项分布、超几何分布之类的，传染病模型，安全模型。

二项分布在n足够大时可以近似正态分布。

## 独立性

试验A的任一结果与试验B的任一结果相互独立，则两个试验相互独立

那么就有P(AB)=P(A)\*A(B) P(A|B)=P(A) P(B|A)=P(B)

在这样的一个系统里，我们经常遇到三种问题：

串联问题（有一个不行就不行），

并联问题（全都不行才不行），

表决问题（当系统故障零件的比例低于某个值时，系统仍然有效）

# Day6 2025年8月30日

## 常用分布

离散性随机变量：

### 二项分布：

### 抛硬币 可近似为泊松分布、正态分布

### 泊松分布：单位内发生事件次数，一个商场一个小时内有多少人进入

某小区快递柜运营方统计了过去 30 天的包裹丢失数据，发现平均每天丢失 2 个包裹。由于： 独立性：某天丢失包裹不会影响其他天的丢失概率（除非有系统性漏洞，但默认运营稳定）；

稀有性：一天内丢失 10 个及以上包裹的概率几乎为 0（实际 30 天内最多仅丢失 4 个）；

稳定性：30 天平均日丢失量稳定在 2 个，无明显波动。

因此，该小区快递柜 “每日包裹丢失次数” 这一随机变量 X，服从泊松分布 X ~ P (λ=2)（λ=2 即 “日平均丢失次数”）。

泊松近似与二项分布——n重伯努利的区别在于：

无限次数、无明确单次概率只有单位范围内概率、计算发生次数、均值即为方差，因为时间跨度一致

抛 50 次硬币，计算“正面朝上 15 次”的概率

路口平均每小时发生 2 起交通事故 计算“该路口 1 小时内发生 3 起事故”的概率）

### 超几何分布：

从 50 件（含 5 件次品）产品中不放回抽 10 件，抽到次品的数量 X，只能取 0,1,2,3,4,5

### 几何分布 无记忆性

连续抛硬币，直到第一次出现正面，需要抛的次数 X（X 可能取 1,2,3,...）

连续性随机变量

### 正态分布 时常需要将均值—方差正态分布标准化为标准正态分布

### 指数分布 无记忆性

某银行柜台办理一笔业务的时间服从指数分布，已知柜台平均每小时可办理15笔业务（即 ) 笔/小时，平均服务时间 分钟/笔）。问题：求一位顾客办理业务的时间超过5分钟的概率？

即顾客等待时间超过5分钟的概率约为28.65%。

### 伽马分布 系统抵御外界改变，在第k次失败的概率

### 贝塔函数

## 统计量计算

均值、中位数、众数：数据中心位置

极差、方差、标准差、四分位距：离散程度

偏度、峰度：形状特征

偏度（不对称程度，左长右短之类的）

峰度（与正态分布比较陡峭程度，数据集中于均值或否）

对于一些多重维度的统计量计算，我们可以使用类似于面积的**协方差**解决方法，

( Ai-E(A) ) \* ( Bi-E(B) )之和即为协方差的分子部分，可以看出当两个同号时等式为正，异号时等式为负，可以看出正相关/负相关。这里得出的协方差仍然带有量纲，可以通过一些方法标准化（**皮尔逊相关系数）**

## 分布近似和中心极限定理

各个分布间都有联系，很多分布可以互相转化，比如超几何分布的二项近似，二项分布的泊松近似，各种近似的正态近似

**无论总体服从何种分布（正态分布、均匀分布、偏态分布等），只要从总体中随机抽取 “足够大” 的样本（通常要求样本量 \(n \geq 30\)），那么这些样本的 “样本均值” 会近似服从 正态分布。**

可近似为正态分布：

超几何、泊松、二项···

有一些数据不符合正态分布，但是可以处理一下数据然后继续正态，比如右偏分布，偏大，我们就取个对数。

# 数理统计

## 参数估计

就是一些估计参数的方式，通过现象，估计出概率函数

比如 极大似然估计MLE，对于一些显而易见的现象，比如抛硬币，只要抛个一百次硬币，然后用正面/反面除以100就是概率了

还有的不是很显然，对于已有的原始数据，我们可以使用EM算法进行参数估计，看起来很难，其实就是先初始化一个猜出来的概率，然后估计Z，然后用Z估计概率，，，不断重复，一直套娃直至收敛

EM的问题是根据不同的初始化值会得到不同的收敛，即多个不同的极值点，这是大部分机器学习的问题。

贝叶斯估计...

## 假设检验