

# 机器学习与深度学习

## ——概率与统计



Personal Website: <https://www.miaopeng.info/>



Email: [miaopeng@stu.scu.edu.cn](mailto:miaopeng@stu.scu.edu.cn)



Github: <https://github.com/MMeowwhite>



Youtube: <https://www.youtube.com/@pengmiao-bmm>

# 目录章节

CONTENTS

**01** 引言

**02** 随机与分布

**03** 学习与推断

**04** 泛化与稳健

**05** 总结

# ► 引言

## ► 为什么需要有概率统计？

- 世界充满不确定性：例如自然波动（基因表达、神经活动、疾病发生），实验噪声（测量误差、仪器误差、样本误差）导致结果完全不可确定。量子力学中的海森堡不确定性原理：

$$\Delta x \Delta p \geq \frac{\hbar}{2}$$

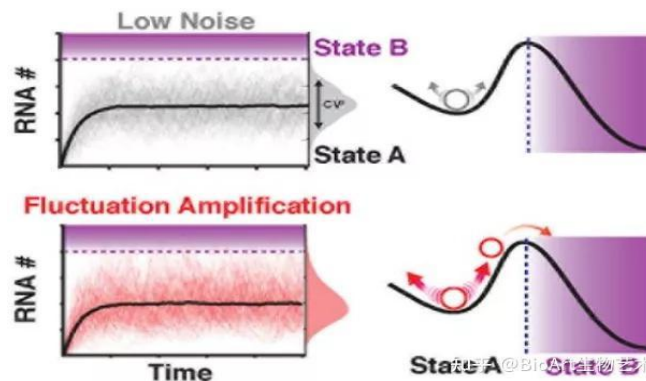
- 数据只是“样本”，我们通常接触到的不是全部信息，而是一部分数据。统计的任务就是从有限样本中推断出背后的真实规律。概率就是这种推断的语言。
- 决策需要风险评估，例如分类模型输出0.7的概率，应该如何做决策？概率帮我们量化不确定性，做更合理的选择。
- 复杂系统只能用分布描述，在基因网络、社会行为等模型中，单一数值不足以描述状态，我们需要概率分布（比如正态分布、泊松分布）来全面表征。

**概率是量化不确定性的语言，统计是基于数据推断规律的工具。**

# ► 现实中的不确定性

## ► 什么是不确定性呢？

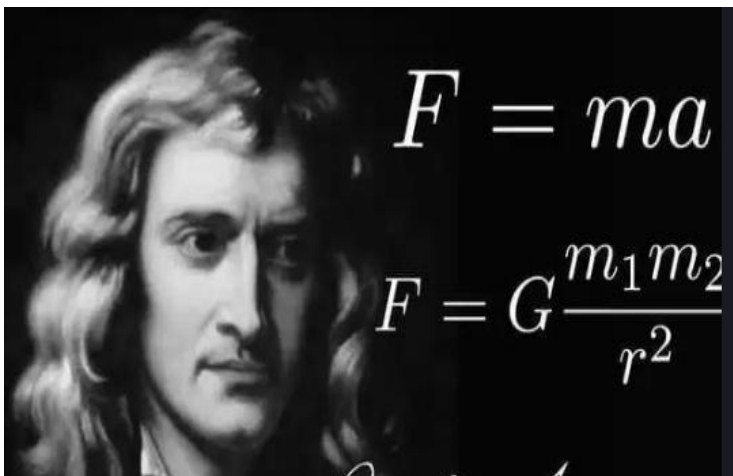
- 基因表达的波动性：同一基因在不同时间或不同细胞中的表达水平会出现随机变动，即基因表达并不是完全稳定或恒定的，而存在一定的随机性或“噪声”。



- 实验噪声：实验结果往往会收到各种误差的干扰（仪器误差、测量误差、人为误差等），导致每次测量出来的结果都不准确，存在一定的随机性。
- 股市预测：股市价格受宏观经济、政策、市场情绪等多重因素影响，具有高度波动性，即便完全掌握历史数据，未来价格仍存在内在随机波动。
- 天气预报：大气系统是典型的混沌系统，微小扰动会被放大，导致预测结果具有不可避免的内在随机性。

## ► 确定性 vs 随机性

### ► 确定性：



- 在已知初始位置 $x_0$ 、初始速度 $v_0$ 、以及作用力（或势能场）的条件下，通过微分方程可以解出任意时刻的位置 $x(t)$ 和速度  $v(t)$ 。

### ► 随机性：



- 从开始略微不同的初始条件摆杆将导致一个完全不同的轨迹。双杆摆是具有混沌方案最简单的动力系统之一。

# ► 点到分布

- 单数据点：
  - 我们测量一个细胞中特定基因的表达量，得到一个数值，比如5。但是这只是单次观测，代码当前时刻该基因表达水平。但这只是一个快照，无法全面反映基因表达的全貌。
- 多次观测与样本集合：
  - 实际上，基因表达会因细胞内环境、调控网络噪声等多种因素而波动。在同一条件下，我们测量大量细胞的该基因表达量，得到一组数据：

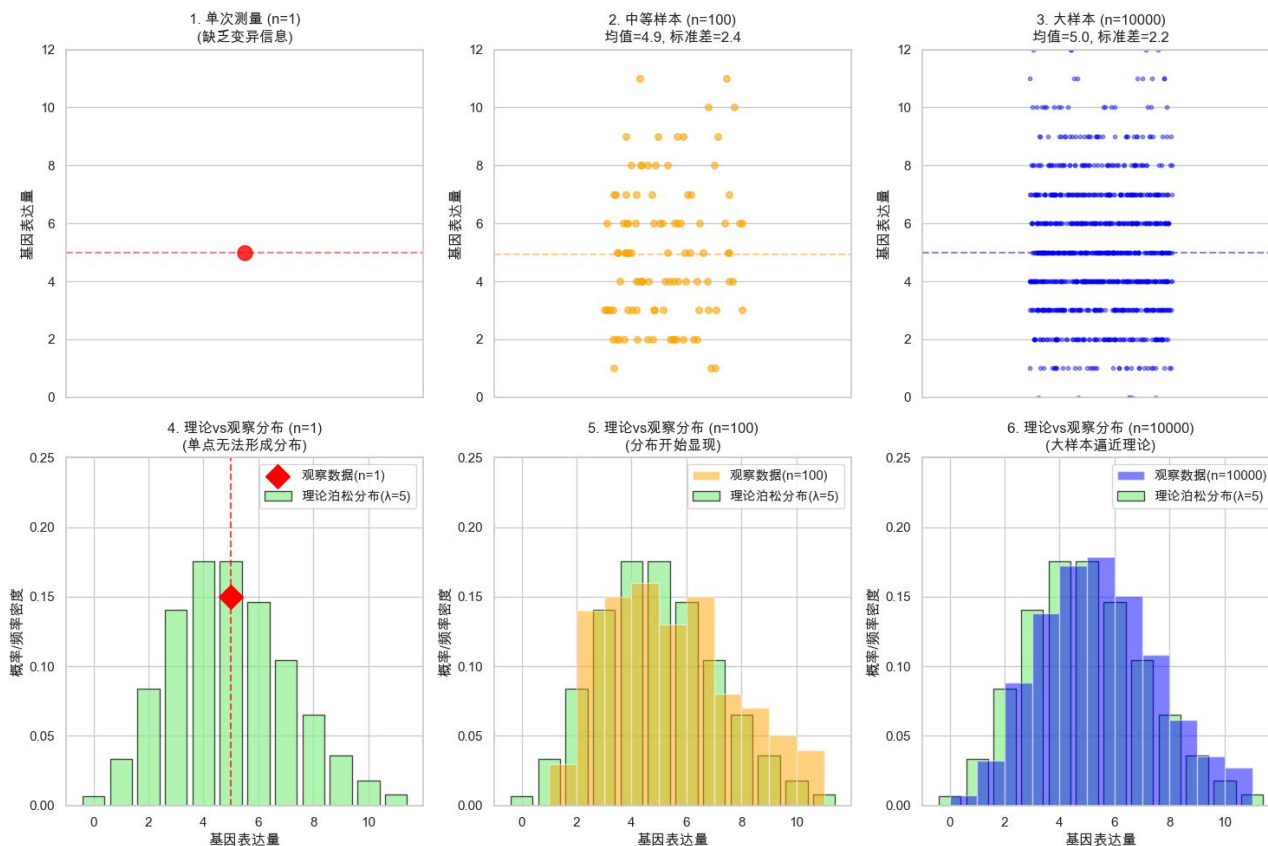
$$x_1, x_2, x_3, x_4, \dots$$

- 这是一组样本，代表了基因表达的多次测量。

| 点思维          | 分布思维           |
|--------------|----------------|
| 单次观测，确定性较强   | 多次观测，表达波动的整体信息 |
| 结果可能受随机噪声的影响 | 描述系统内在的随机性和变异性 |
| 直观但片面        | 全面、稳健、科学       |

# ► 点到分布

## ➤ 从点到分布，能够揭示波动性和不确定性：



- 通过统计分析，这组数据被描述为概率分布（如正态分布、泊松分布等）。
- 多样本数据不仅告诉我们平均表达水平，还展示表达的变异范围。
- 概率分布体现了基因表达的波动性和不确定性。

## ► 概率的直观含义

- 频率学派 (Frequentist)：视概率为“**重复实验的频率**”。直观类比：掷硬币无数次，看正面出现的比例。
- 贝叶斯学派 (Bayesian)：视概率为“**我对事件的信念**”。直观类比：一开始猜硬币可能偏向正面（先验），不断投掷（数据），逐步修正这个看法（后验）。



Pierre-Simon Laplace

概率的定义：事件在无限重复实验中出现的频率  $P(A) = \lim_{n \rightarrow \infty} \frac{\text{count}(A)}{n}$

参数：固定但未知（例如基因表达均值  $\mu$ ）。

数据：随机的。 在此处键入公式。

推断：通过假设检验、置信区间来对固定参数进行推断。

概率的定义：主观信念的强度，随新数据更新。

参数：也看作随机变量，有先验分布。

数据：用来更新我们对参数的认知。

推断：基于贝叶斯公式更新：  $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$

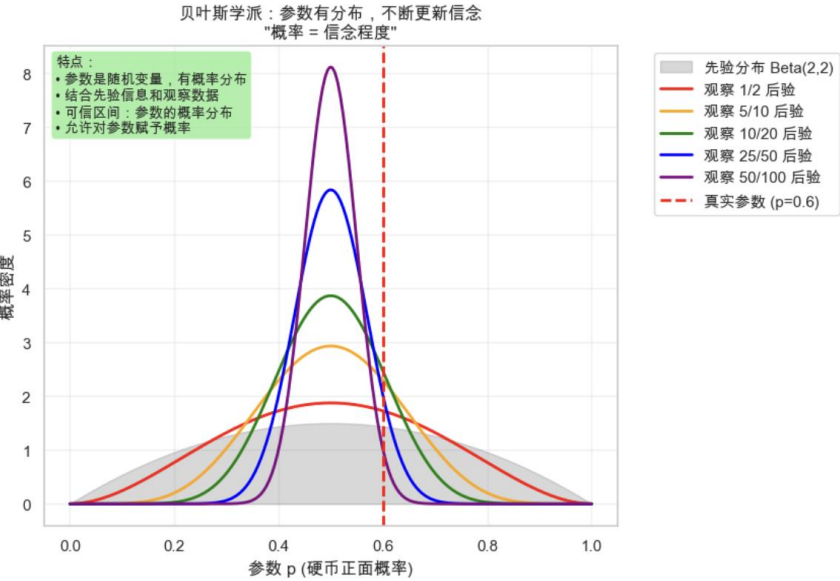
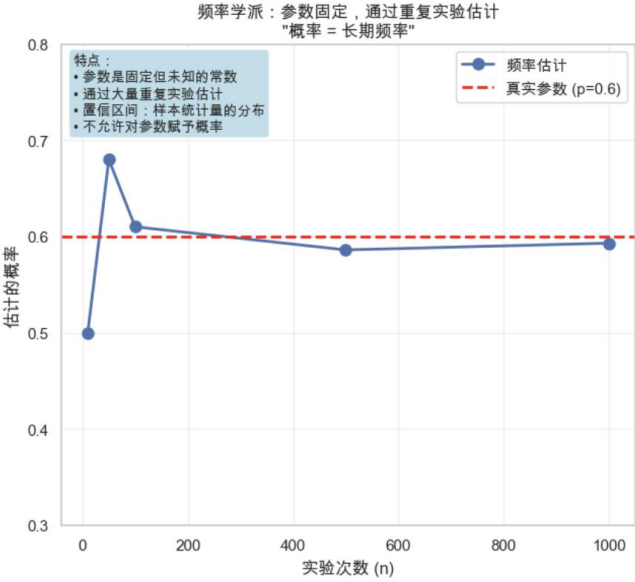


Thomas Bayes

频率学派：概率是长期频率；贝叶斯学派：概率是信念更新。



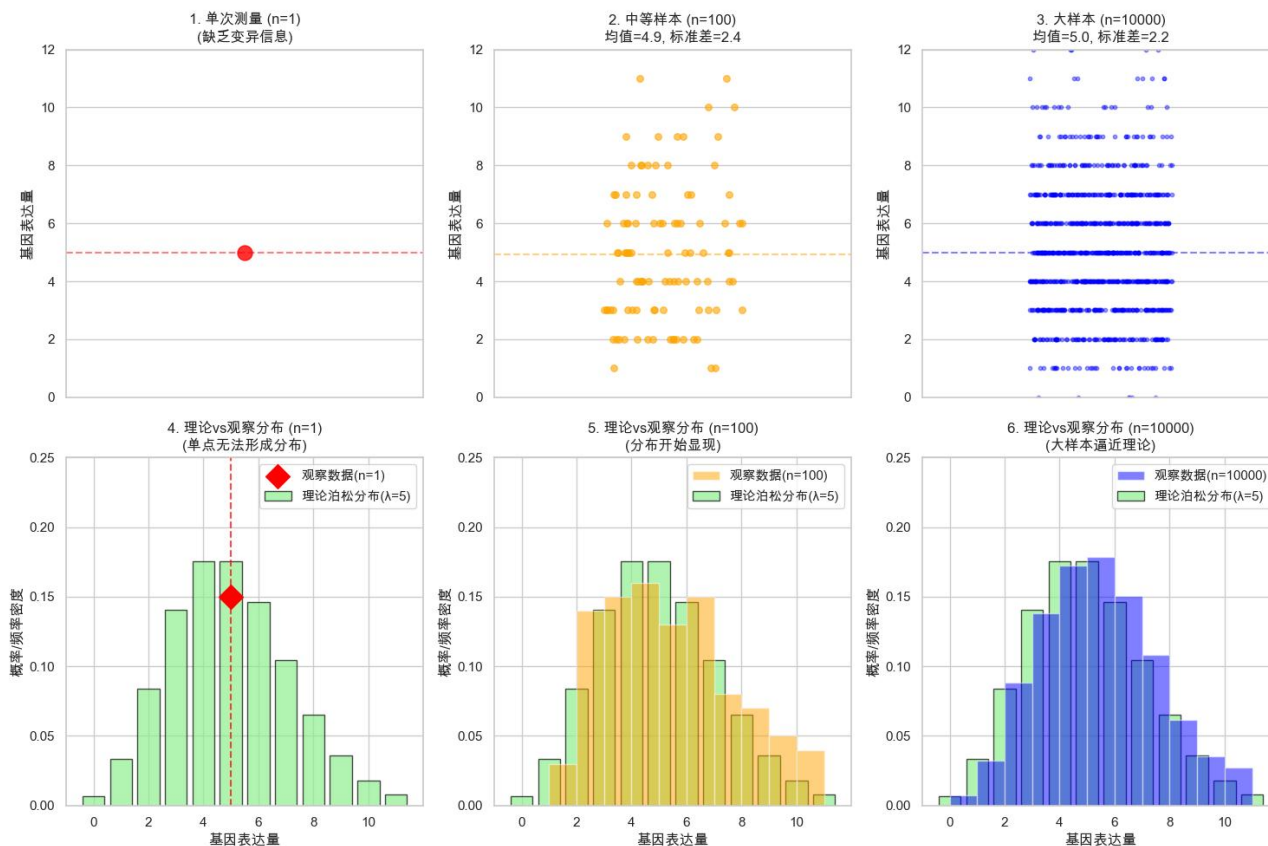
# ► 概率的直观含义



|      | 频率学派      | 贝叶斯学派      |
|------|-----------|------------|
| 概率含义 | 长期频率      | 信念程度       |
| 参数   | 固定未知      | 随机变量       |
| 结果   | 置信区间      | 后验分布       |
| 优势   | 传统成熟，计算搞笑 | 融合先验，灵活直观  |
| 劣势   | 无法利用先验知识  | 计算复杂，需设定先验 |

# ► 点到分布

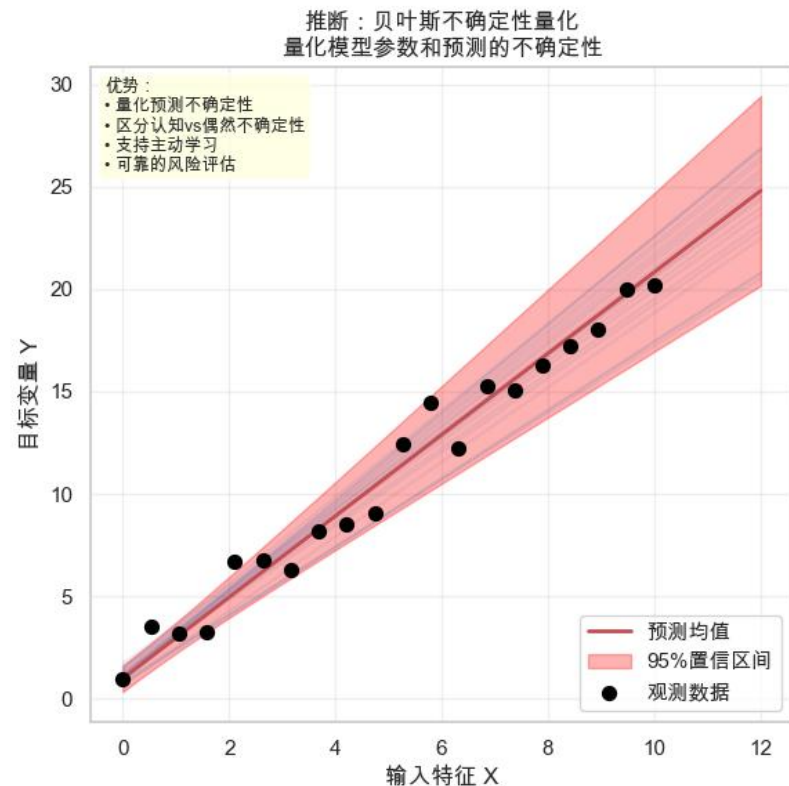
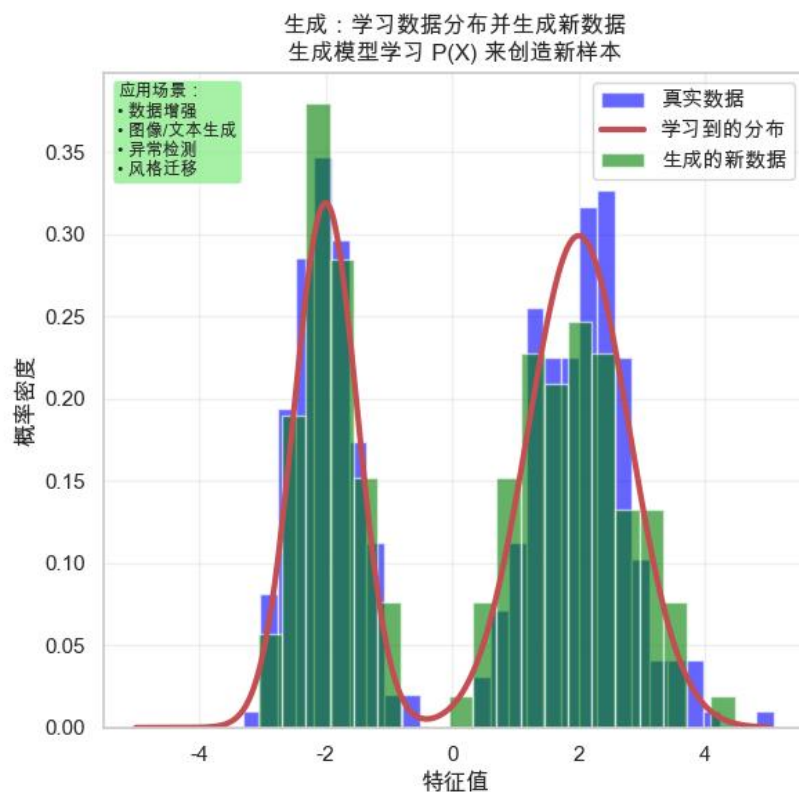
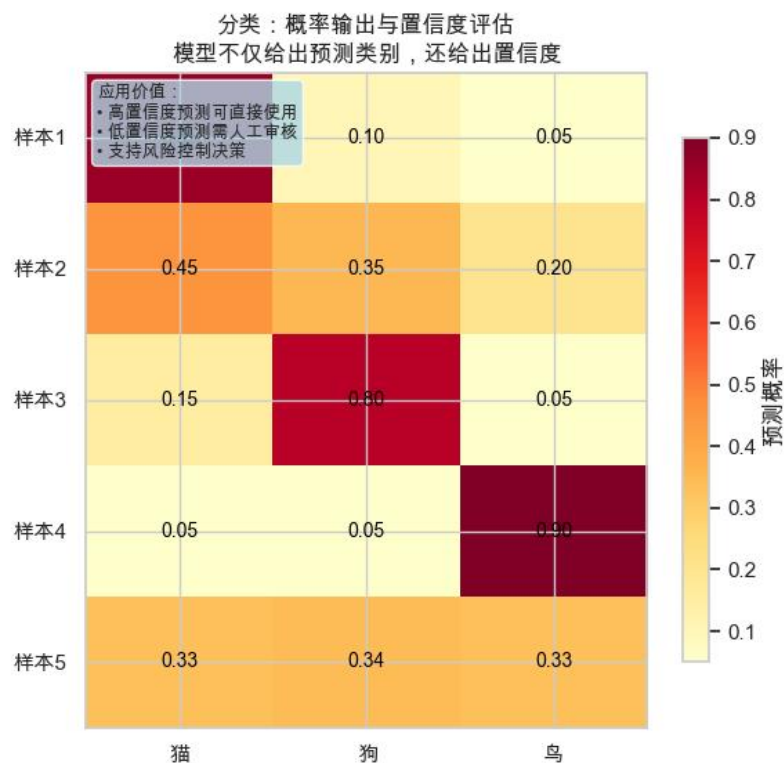
## ➤ 从点到分布，能够揭示波动性和不确定性：



- 通过统计分析，这组数据被描述为概率分布（如正态分布、泊松分布等）。
- 多样本数据不仅告诉我们平均表达水平，还展示表达的变异范围。
- 概率分布体现了基因表达的波动性和不确定性。

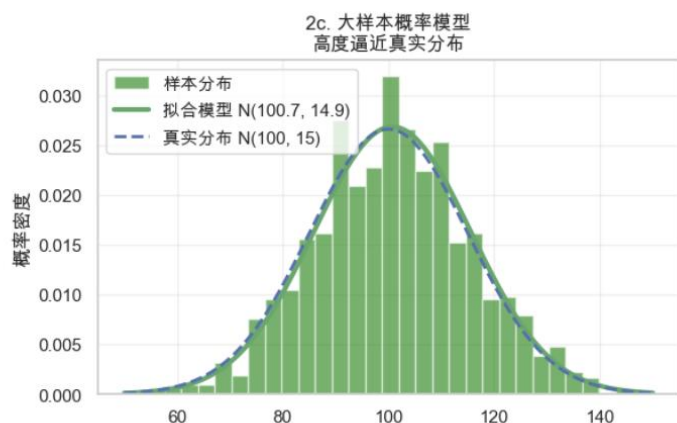
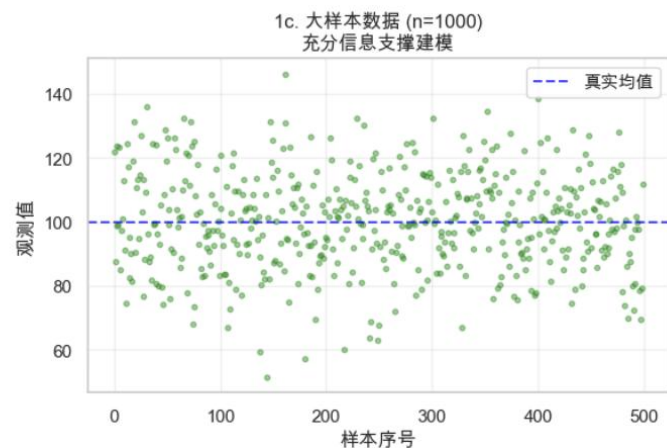
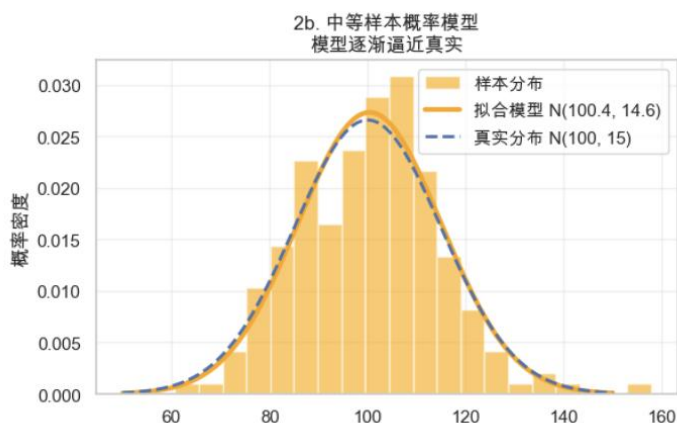
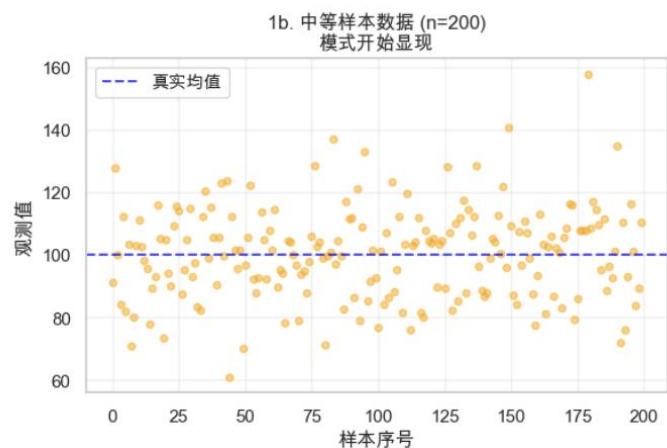
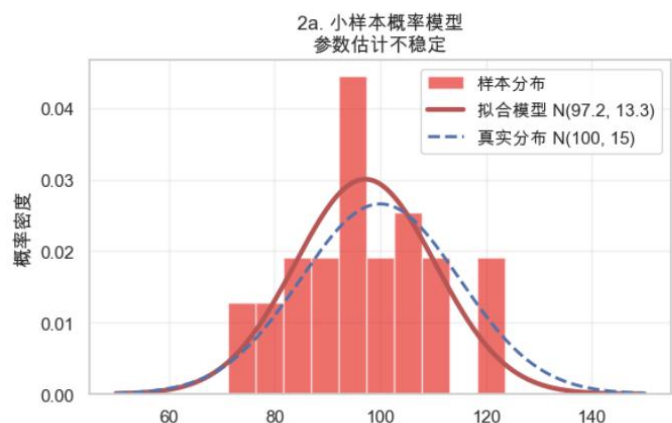
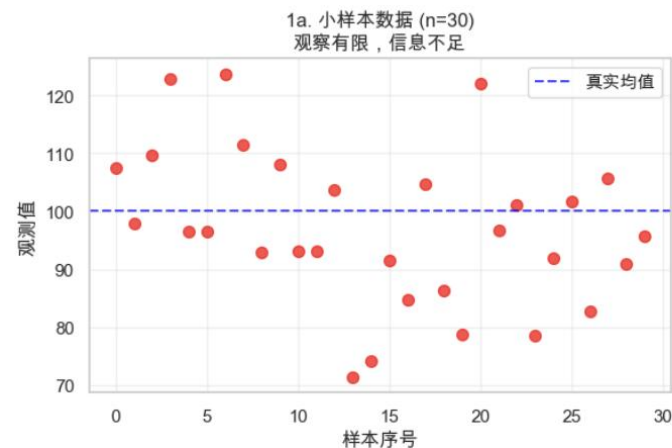
# ► 为什么机器学习离不开概率？

- 分类：机器学习不仅要给出类别，还要给出属于每个类别的概率，用于决策与置信度评估。
- 生成：生成模型本质上是在学习数据的概率分布，从分布中采样生成新数据。
- 推断：贝叶斯方法通过概率更新量化模型与预测中的不确定性。



# ► 统计：从样本到发现规律

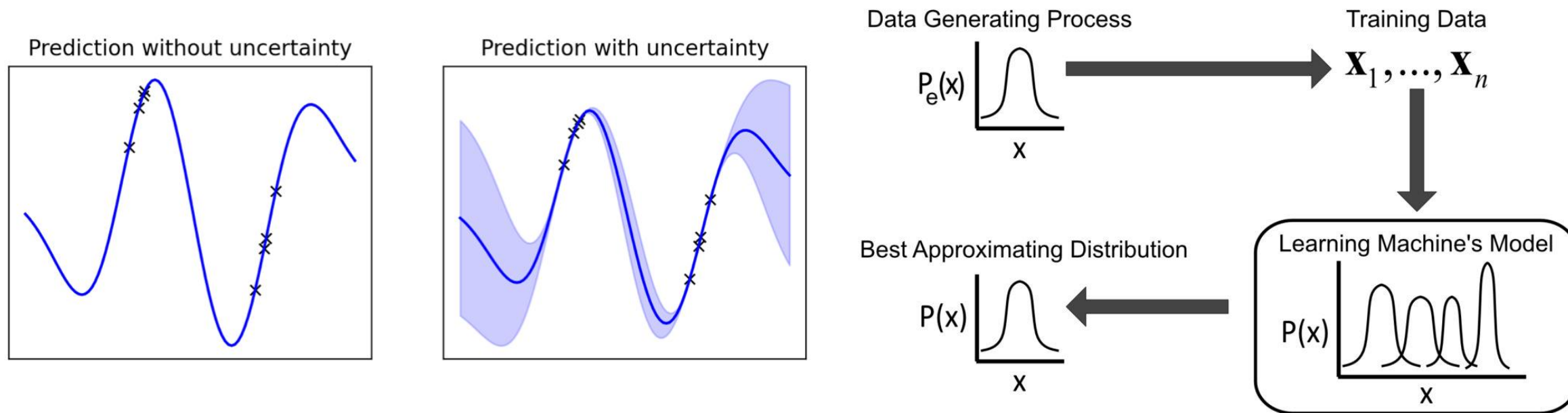
► 核心思想：样本数据 → 概率模型 → 推断真实规律。



统计就是通过有限样本，推断数据背后的整体规律与不确定性；随着样本数量的增多，这种推断会越来越接近真实。

## ► 思维转变：接受不确定性

- 不确定性是自然规律的一部分，我们无法完全消除它，但可以理解它。
- 科学的目标不是追求绝对确定，而是用概率与统计，量化和利用不确定性。

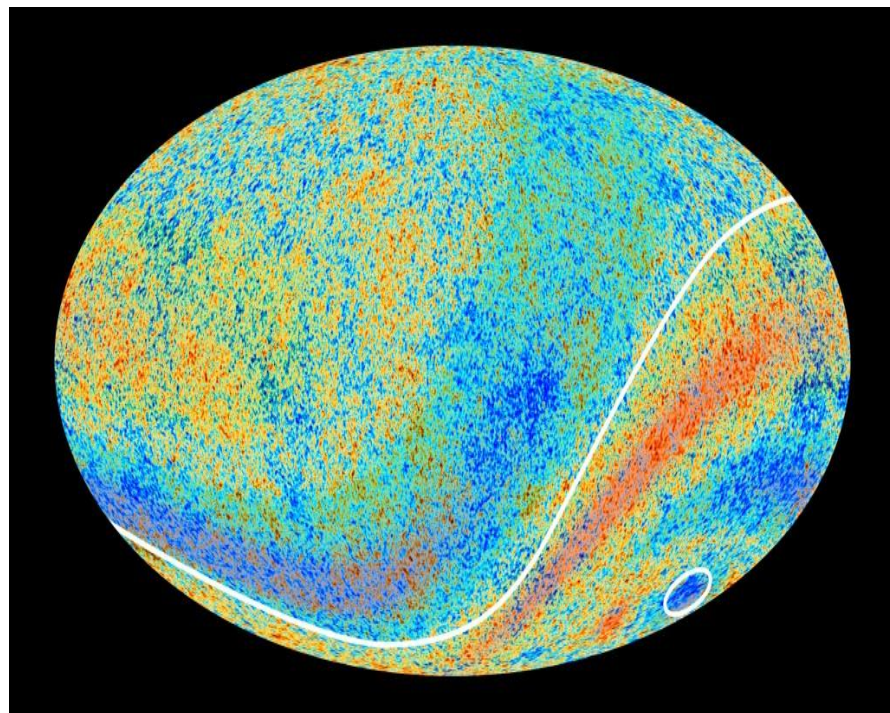
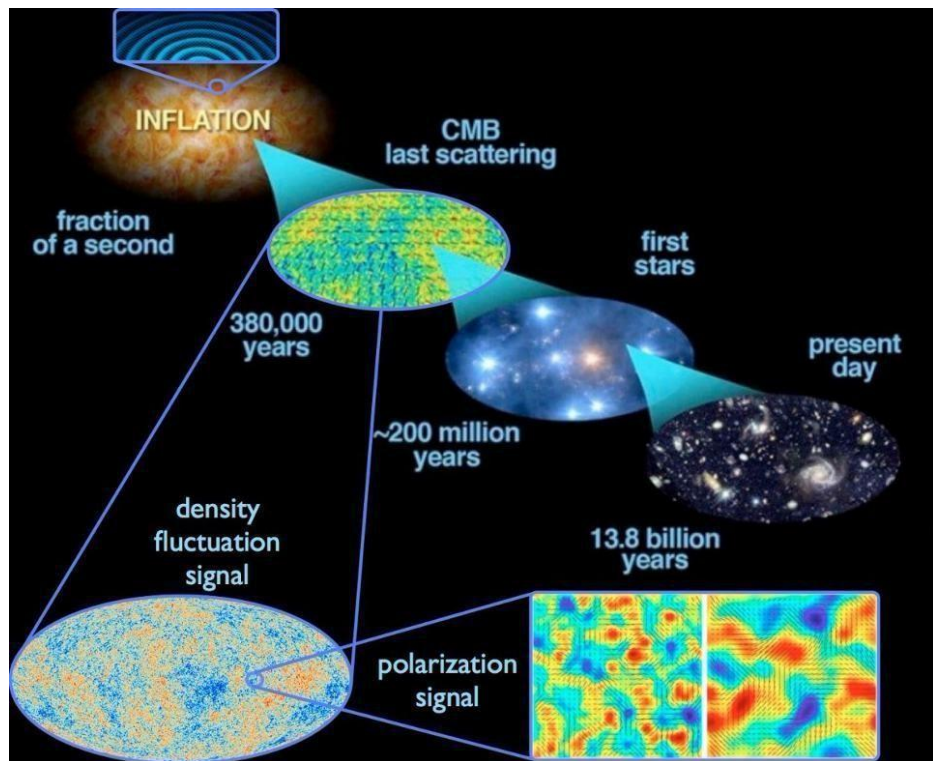


**科学不追求绝对确定，而是用概率刻画与利用随机性。**



## ► 小结

- 不确定性是研究的起点，概率是描述与推理的核心工具。



- 如何具体描述随机性？怎么使用随机性建立模型呢？

**科学不追求绝对确定，而是用概率刻画与利用随机性。**

# 目录章节

CONTENTS

01 引言

02 随机与分布

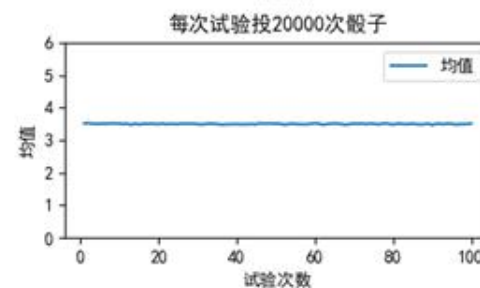
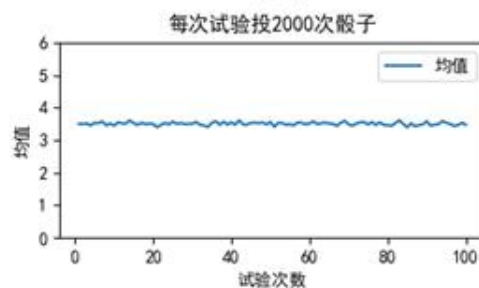
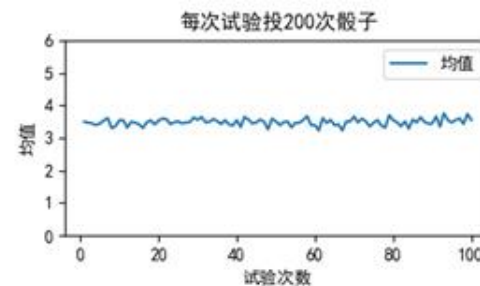
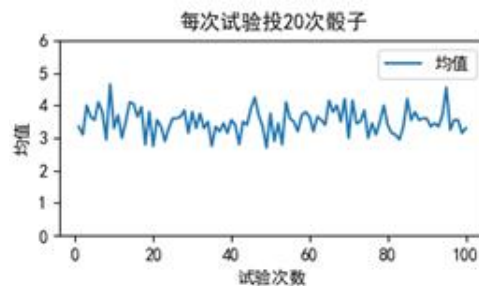
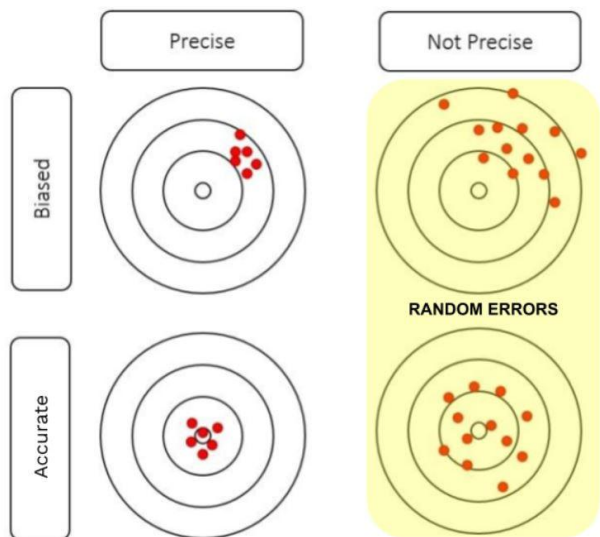
03 学习与推断

04 泛化与稳健

05 总结

## ► 如何描述随机性？

► 我们说某个现象是‘随机的’，到底是什么意思？它只是‘乱’，还是背后有规律？



► 数学抽象：随机性  $\neq$  无序，而是在不确定中有规律。

► 定义：

- 随机变量：把不确定结果映射为数值  $X : \Omega \rightarrow \mathbb{R}$
- 概率分布：描述每个可能结果的可能性  $P(X = x)$

概率学把随机性抽象为一个三元组： $(\Omega, F, P)$

样本空间  $\Omega$ ：所有可能结果的集合，例如掷骰子：  
 $\Omega = \{1, 2, 3, 4, 5, 6\}$

事件集  $F$ ：我们关心结果的集合，例如掷骰子结果是偶数： $\{2, 4, 6\}$

概率测度  $P$ ：为每个事件分配一个概率，满足  
 $P(\omega) = 1, 0 \leq P(A) \leq 1, P(\cup A_i) = \sum_i P(A_i)$



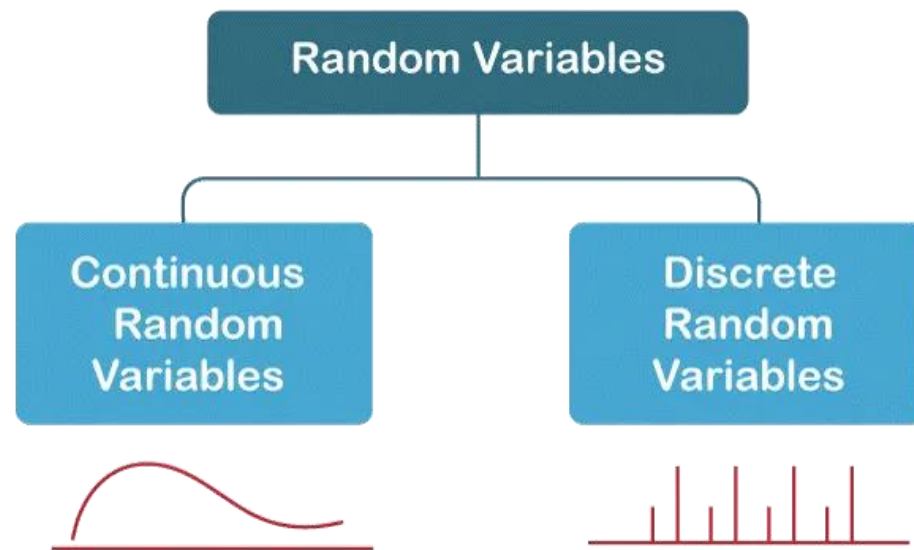
## ► 随机变量

► 什么是随机变量？把样本空间的每个结果映射为一个数（便于计算与建模）

$$X : \Omega \rightarrow \mathbb{R}$$

► 两类随机变量：

- 离散型：骰子点数、基因突变数。
- 连续型：身高分布、基因表达量。



**随机变量是连接“随机现象”和“数学世界”的桥梁。**

## ► 概率分布

- 什么是概率分布？描述随机变量所有可能取值及其发生概率的规律。
- 两类随机变量对应两类函数：
  - 离散型：用**概率质量函数（PMF）**  $P(X=x)$  进行描述。
  - 连续型：用**概率密度函数（PDF）**  $f(x)$  进行描述。
- 为什么重要？全面刻画不确定性（不均是均值/方差），是机器学习、深度学习建模的基础。
- 常见分布：
  - 离散型：伯努力分布、二项分布、泊松分布。
  - 连续型：正态分布、指数分布。

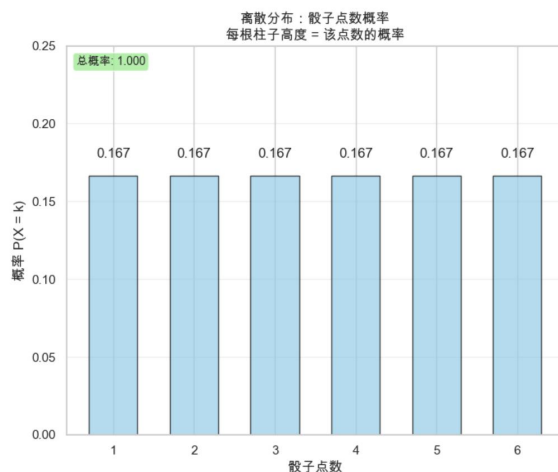
**概率分布就是描述一个随机现象中，所有可能结果出现的可能性大小的整体规律，它帮助我们**从“随机”中看见“秩序”。

# ► 概率质量函数 vs 概率密度函数

## ► 概率质量函数 (PMF)

- 适用于离散值（如掷骰子、性别、计数）
- 每个具体取值有明确概率：

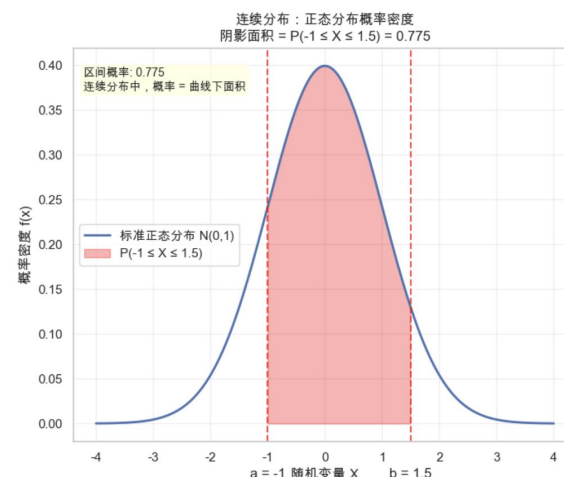
$$P(X = x_i) = p_i, \sum_i p_i = 1$$



## ► 概率密度函数 (PDF)

- 适用于连续值（如身高、温度）
- 概率是区间上的面积，单点概率为0：

$$P(a \leq X \leq b) = \int_a^b f(x)dx, \int_{-\infty}^{+\infty} f(x)dx = 1$$



**概率分布通过概率质量函数刻画离散结果的可能性，通过概率密度函数描绘连续取值的相对可能性，用来全面描述随机现象的不确定性。**

# ► 随机性的数值刻画

## ► 集中趋势：

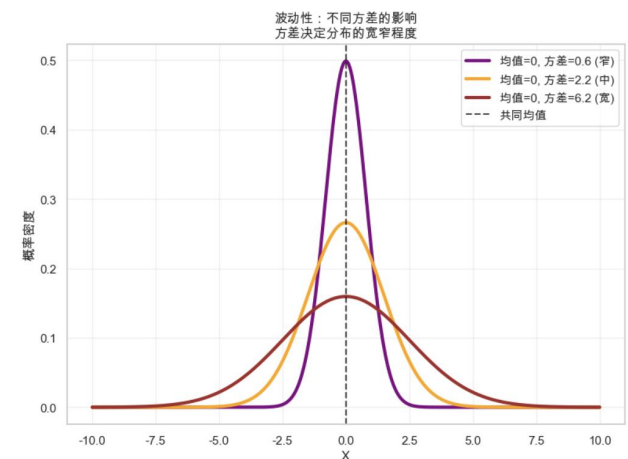
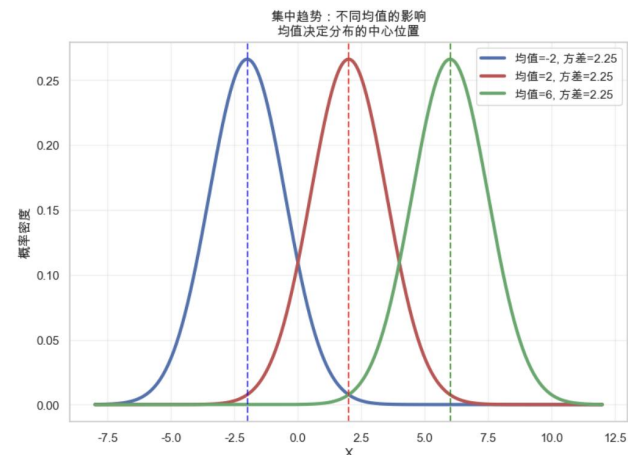
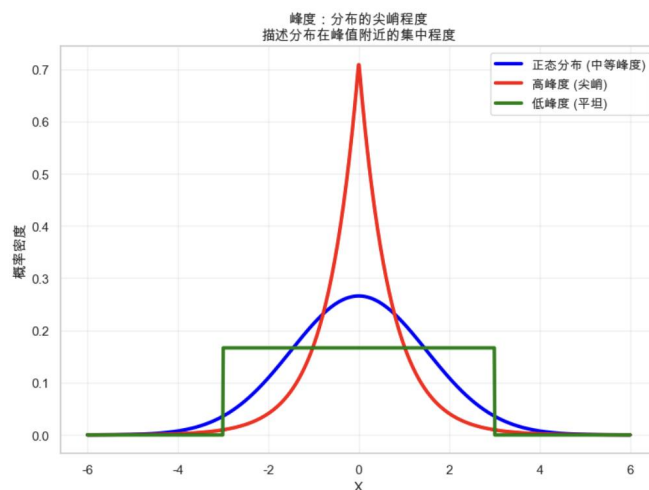
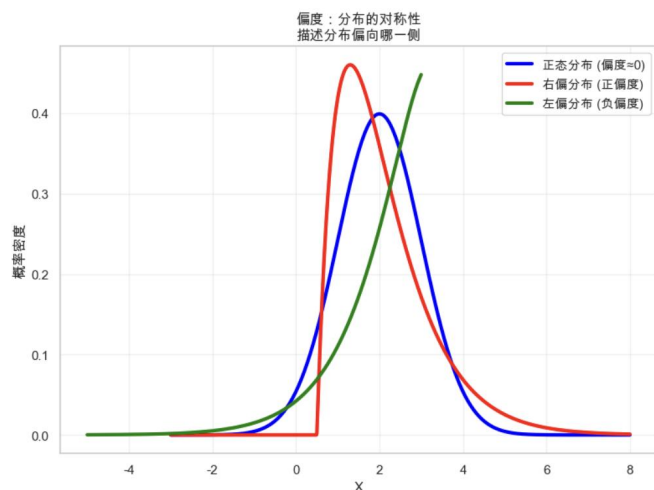
- 均值 (Expectation)：描述数据的中心位置。

## ► 波动性：

- 方差 (Variance)：衡量数据围绕均值的离散程度。

## ► 更复杂特征：

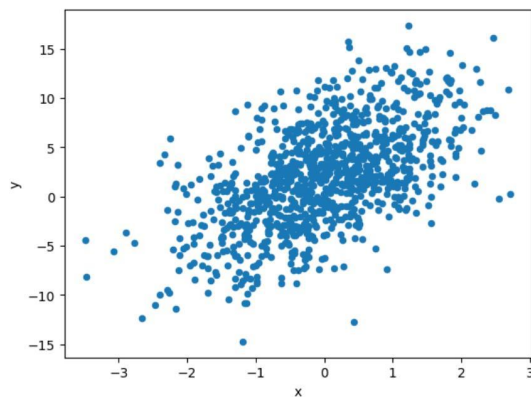
- 偏度 (Skewness)：描述分布的对称性。
- 峰度 (Kurtosis)：描述分布的尖峭程度。



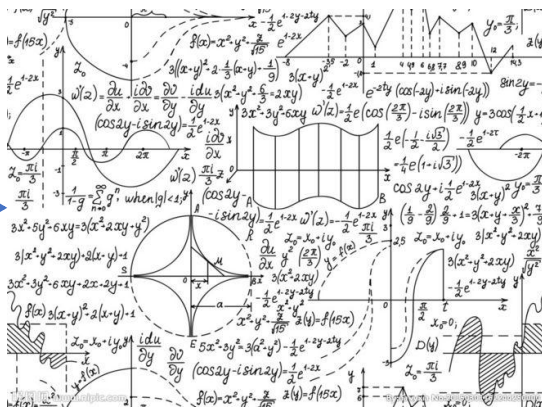
- 随机性的数值刻画：通过均值、方差等统计量概括数据的中心位置、波动程度及分布形态，从而定量描述随机现象。

# ► 从随机到模型：为什么需要建模？

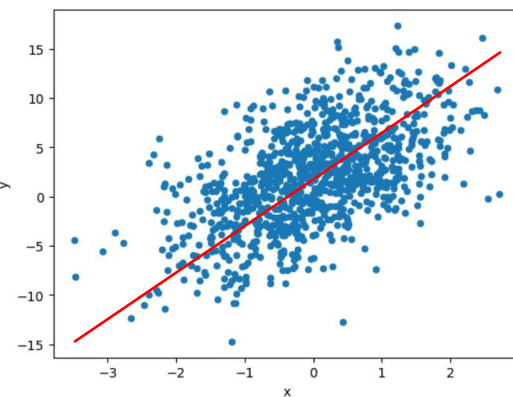
- 数据本身只是现象，模型才让我们理解并预测背后的规律。
- 目标：不仅需要描述事件，还能预测/生成事件【机器学习、深度学习】。
- 核心思路：从杂乱数据 → 建立数学模型 → 预测与理解。



杂乱数据



建立模型



预测理解

建模是将杂乱的随机数据抽象为可理解的规律，用于解释现象并预测未来。

# ► 参数化分布建模：用少量参数刻画随机性

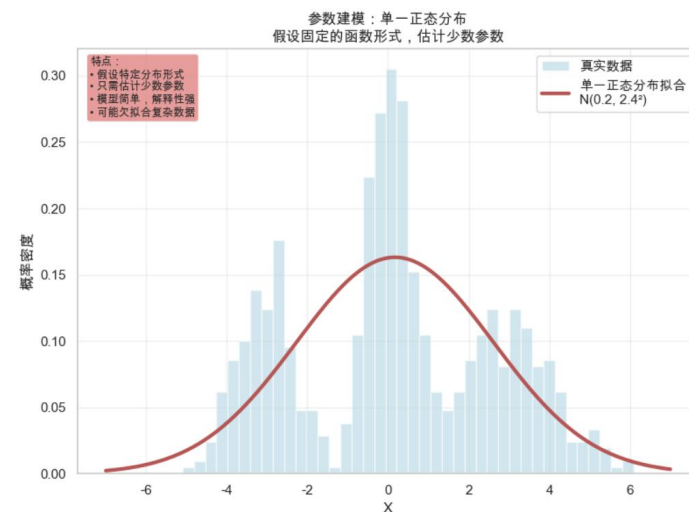
## ► 为什么要参数化？

- 现实数据复杂，但我们希望用少量参数（如均值、方差）抽象本质规律。
- 举例：正态分布只需 2 个参数（均值、方差）就能描述形状。

## ► 怎么做？

- **假设**：假设数据遵循某个分布（正态、泊松、指数等）。
- **估计参数**：用观测数据计算（最大似然 / 贝叶斯方法）。
- **建模**：得到可解释、可用于预测的分布模型。

- 优势：1）可解释：参数直接对应数据特性；2）可预测：分布模型能外推、模拟新数据；3）可比较：不同数据集可用同一参数框架比较。



**参数化建模，就是用少量有意义的参数，提炼随机数据的核心规律。**

# ► 非参数建模：让数据自己说话

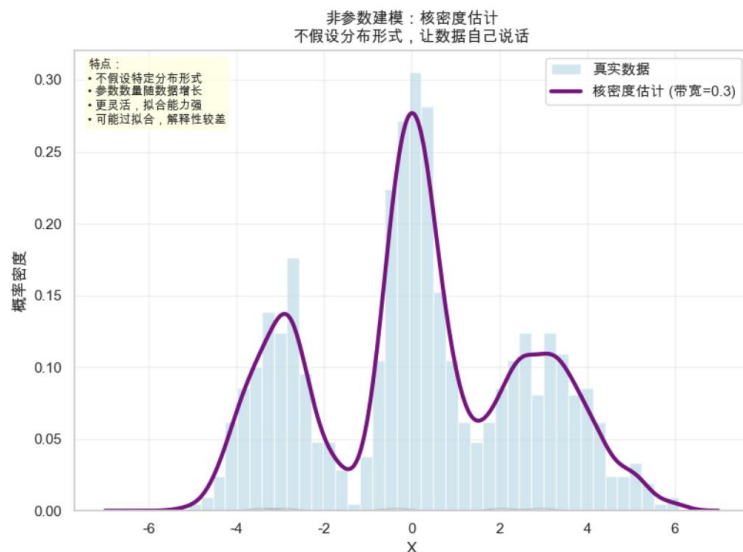
## ► 为什么需要非参数建模？

- 参数化模型有分布假设（如“数据是正态的”），但**真实数据常常不符合假设**。需要一种灵活的方法，直接从数据中“学”出规律。

## ► 核心思想：不预设具体分布形式，**让模型根据数据形状自适应**。即让数据自己说话，形状由数据驱动。

- 举例：核密度估计（KDE）、直方图、k近邻等。

## ► 优势：灵活、不受分布假设限制，能捕捉复杂结构；缺点：需要更多数据，解释性较弱，计算量大。



**非参数建模不假设分布形式，而是让数据自己决定模型形状。**



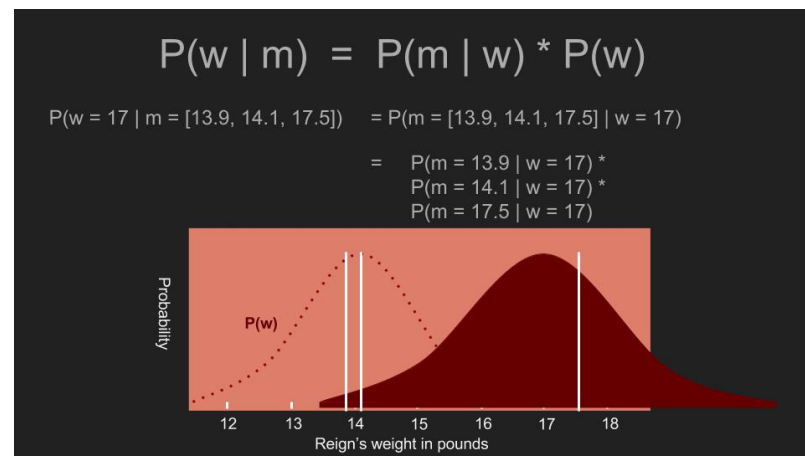
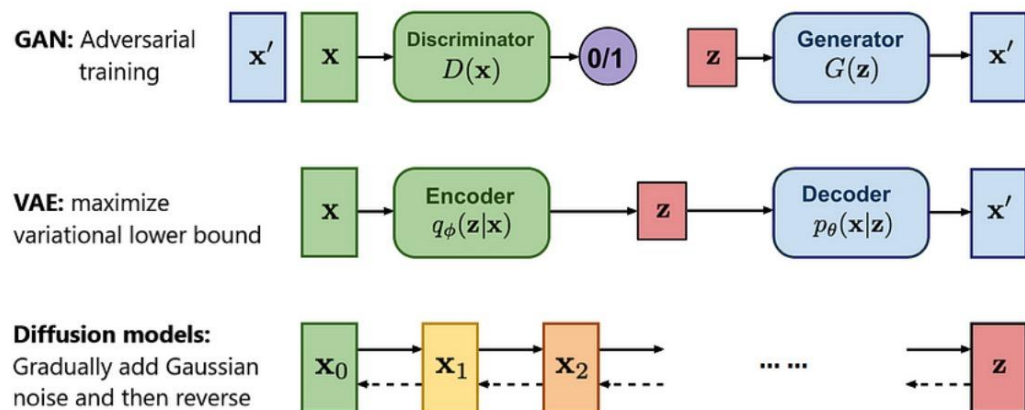
## ► 随机性的力量：生成与推断

► 随机性不是干扰，而是资源。

- 传统认知：随机选 = 杂乱无章的噪声。
- 新视角：随机性 = 帮助我们探索未知的工具。

► 生成模型：学习数据分布，采样新数据（GAN、VAE）。

► 推断模型：利用分布进行预测、置信区间、假设检验

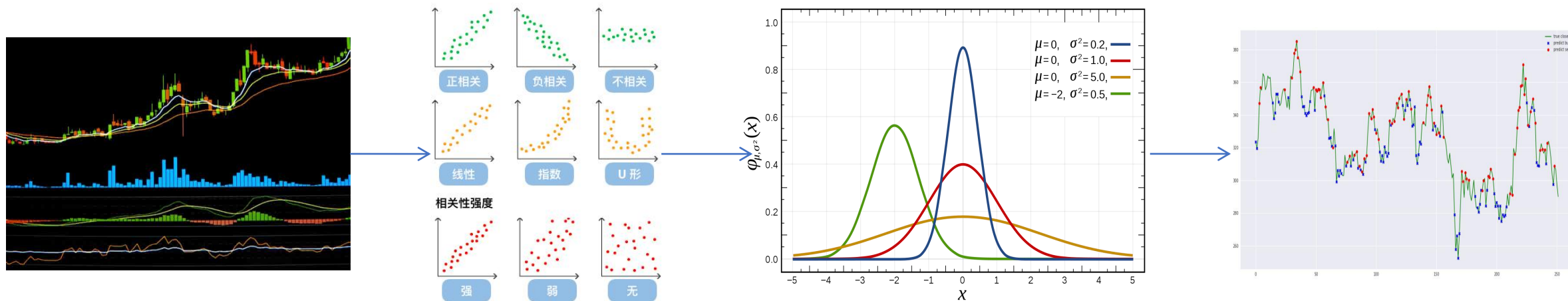


随机性不仅描述世界的不确定，更是我们生成新知识与模型的驱动力。



## ► 小结

- **从不确定性到随机性**：现实世界充满不可预测性，我们用随机变量来刻画这种不确定性。
- **从点到分布**：不再只看单个值，而是用概率分布全面描述随机现象的所有可能性及其概率。
- **随机性的数值刻画**：通过期望、方差等指标总结分布的集中趋势与波动性，量化不确定性。
- **分布的力量**：概率分布是建模的基础，支撑预测、推断与生成，是从数据到科学规律的桥梁。



从现实世界的不确定现象出发，用随机变量抽象描述，并通过概率分布刻画其规律，从而实现预测与决策。

# 目录章节

CONTENTS

01 引言

02 随机与分布

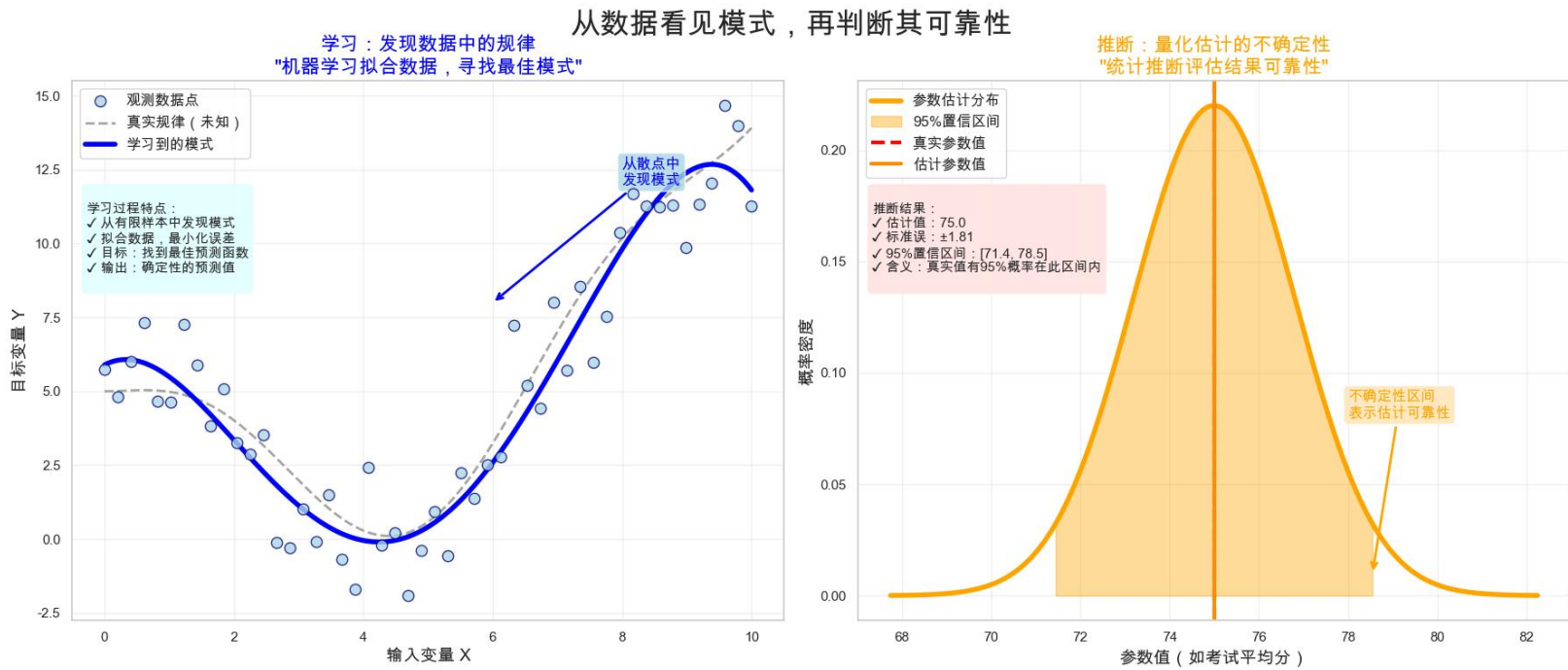
03 学习与推断

04 泛化与稳健

05 总结

# ► 如何从数据中学习与推断？

- 从描述随机性，走向利用随机性 —— 统计学习与推断让数据变成知识。
- 学习是从有限样本中发现潜在规律，把数据转化为模型；推断是在模型基础上，对未知整体特征与未来结果做出合理判断。



|    | 学习          | 推断          |
|----|-------------|-------------|
| 目标 | 从数据中发现模式和规律 | 评估估计结果的可靠性  |
| 方法 | 拟合算法，优化损失函数 | 概率论，假设检验    |
| 输出 | 预测模型，确定性结果  | 置信区间，不确定性量化 |
| 关注 | 模型性能，预测精度   | 统计显著性，可信度   |
| 例子 | 神经网络学习图像特征  | 药物疗效的显著性检验  |

学习与推断：让我们既看见模式，又能量化其可靠性与不确定性。

## ► 学习的核心问题与算法

- 核心问题：已知一批带不确定性的数据，如何找到最能解释数据的概率模型/分布？【**本质：数据更新对世界的认知**】
- 两大类学习任务：
  - **参数学习 (Parameter Learning)**：1) 最大似然估计 (MLE)：找到最可能产生数据的参数；2) 贝叶斯更新 (MAP/Posterior)：融合先验和数据，得到参数分布。
  - **结构学习 (Model Structure)**：学习概率图模型、潜变量模型的依赖关系。
- 代表性算法：
  - 频率学派：1) MLE (最大似然估计)；2) EM算法 (处理潜变量)
  - 贝叶斯学派：1) Bayes更新 (先验→后验)；2) MCMC、变分推断 (近似后验)

**概率学习的任务，就是在不确定性下，用数据学习最合理的分布及其参数。**

## ► 推断的核心问题与算法

- 核心问题：已知学到的概率模型（分布/参数），如何回答关于未知量的问题（预测、区间、假设）？【**本质：从不确定的分布中计算我们关心的结论**】
- 主要任务：
  - 点估计：给出最可能的值（均值、最大后验估计）
  - 区间估计：量化不确定性（置信区间 / 贝叶斯可信区间）
  - 假设检验：判断差异或效应是否显著
- 代表性算法：
  - 频率学派：1）t检验、卡方检验、方差分析（ANOVA）；2）置信区间构建。
  - 贝叶斯学派：1）后验推断（Posterior Inference）；2）MCMC采样、变分推断（计算复杂后验）

**推断的任务，就是在模型和数据的框架下，量化未知并回答科学问题。**

## ► 学习与推断的结合

➤ 学习是从数据中获取知识，推断是用知识回答问题，概率论是它们的桥梁。

➤ 为什么要结合？

- 仅学习：得到模型，但无法量化模型的可靠性。
- 仅推断：有不确定性量化，但没有数据驱动模型。
- 结合：通过概率分布建模，**既能学习数据规律，又能量化不确定性。**

➤ 核心思路：

- 贝叶斯框架：先学习后推断。即先利用数据更新模型的后验分布，再在后验分布上进行预测、决策。
- 频率学派 + 贝叶斯学派：在参数估计与假设检验间建立联系。

**学习让我们认识世界，推断让我们在不确定中行动；概率把两者连为一体。**

## ► 假设检验：提出问题 → 统计回答

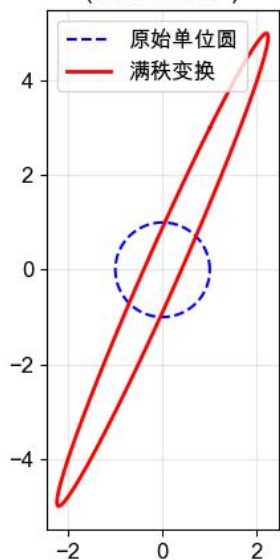
- 核心思路：提出一个假设，用数据来检验它是否成立。（不是证明真伪，而是用概率衡量“**是否有足够证据**”）
- 流程：
  - 提出假设：零假设 $H_0$ 是“没有差异/没有关系”，备择假设 $H_1$ 是存在差异或关系。
  - 选择检验方法：t检验、卡方检验、非参数检验等。
  - 计算统计量：量化样本与假设的差异。
  - P值：出现当前结果的概率有多大？
  - 做决策：P值 < 阈值（如0.05）→ 拒绝 $H_0$ ，
- 举例：药物是否有效？（实验组 vs 对照组），广告是否提高了点击率？（A/B测试）。

**假设检验是把“怀疑”变成“有数据支撑的判断”。**

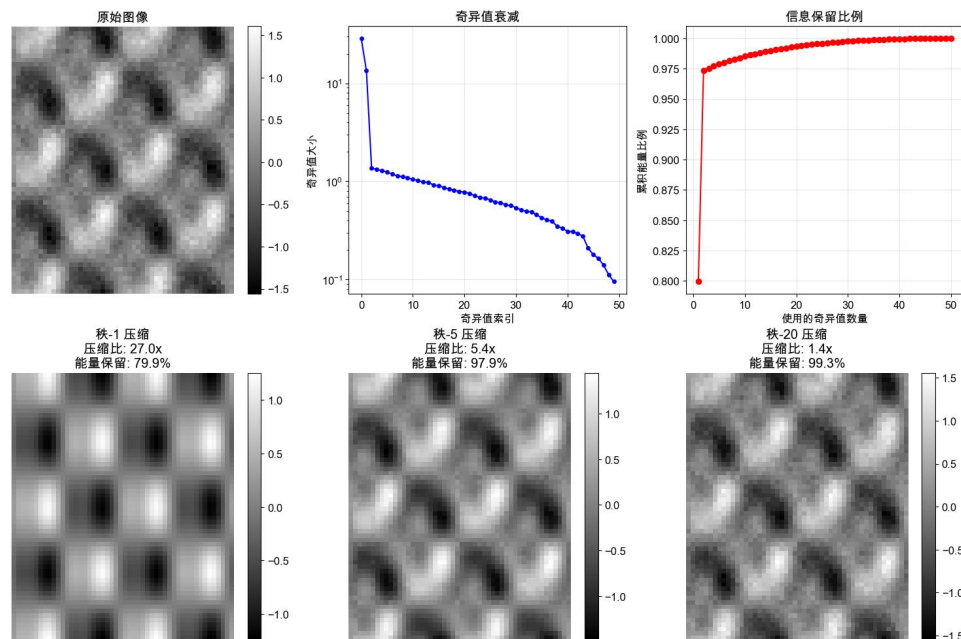
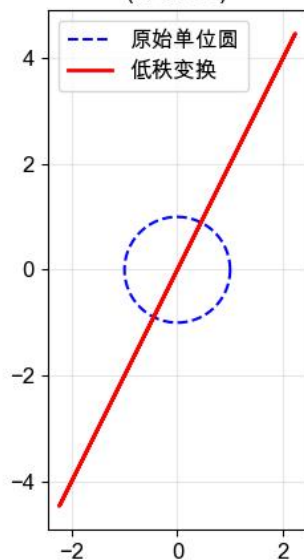
## ► 小结

- 直观理解：原矩阵就像存了所有像素，而 SVD 只存主要方向（奇异向量）和权重（奇异值）。而在还原时，用这些“主要方向”重新组合出一个近似矩阵，信息量大幅减少但结构尽量保留。
- 问题：一个矩阵变成三个矩阵，不是信息量变多了？如果只保留前 $k$ 个奇异值（最大的能量方向），并且对应的  $U_k(m \times k)$ 、 $V_k(n \times k)$ ，只保留前  $k$  列，原始矩阵存储量为 $mn$ ，压缩后的矩阵存储量为 $mk + k + nk$ ，比 $mn$ 小很多【当 $k \ll \min(m, n)$ 时】。

满秩矩阵：圆  $\rightarrow$  椭圆  
(保持2D特性)



低秩矩阵：圆  $\rightarrow$  直线  
(降为1D)





# 目录章节

CONTENTS

01 引言

02 随机与分布

03 学习与推断

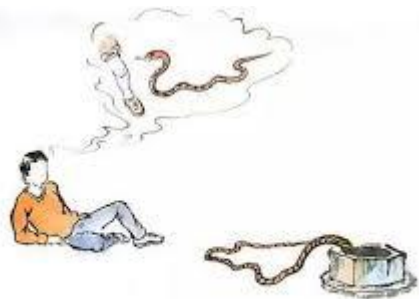
04 泛化与稳健

05 总结

# ► 泛化 (Generalization)

► 什么是泛化？模型不仅在训练数据上表现好，**还能在未见过的数据上保持预测能力。**

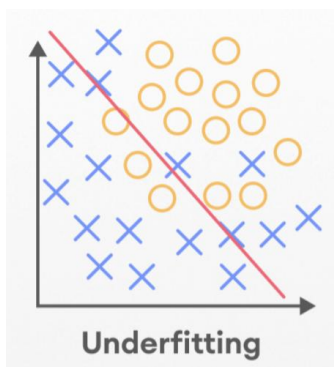
- 概率语言：我们关心的是真实分布 $P_{\text{true}}$ ，而不是仅仅拟合样本分布 $P_{\text{empirical}}$ 。



大脑在缺乏足够样本时，会过度更新对真实分布的认知，把“蛇的概率”泛化到了“所有类似形状的物体”。

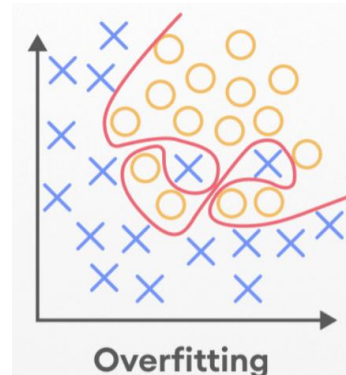
本质：这是**基于少量样本的主观贝叶斯更新**，但因为先验和样本量不足，导致后验分布极端偏移（“**过拟合**”到一次经历）。

► 欠拟合 (Underfitting) :



- 模型太简单或学习不充分，无法捕捉数据中的真实规律。
- 在训练集和测试集上都表现不好。
- 概率视角：模型的假设空间太小，先验约束过强，无法拟合真实分布。

► 过拟合 (Overfitting) :

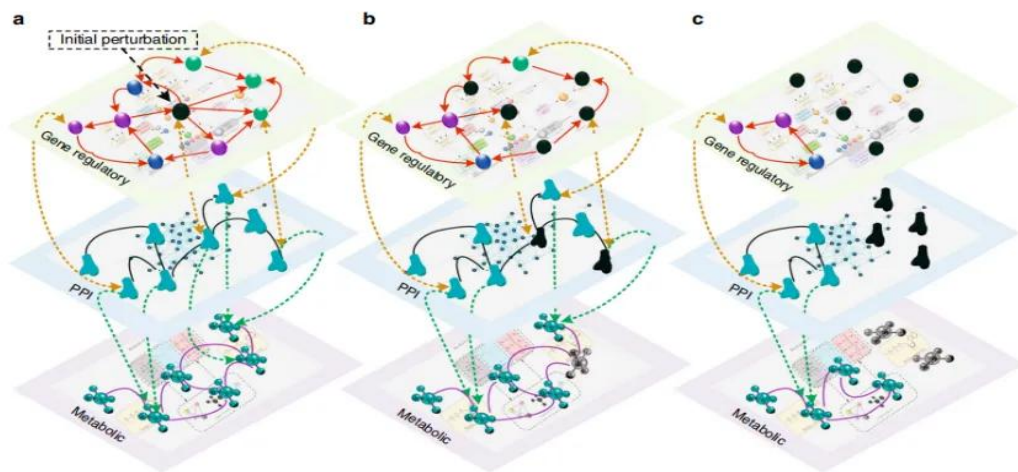


- 模型太复杂或过度训练，把噪声或偶然样本当成了规律。
- 在训练集表现好，但在测试集表现差。
- 概率视角：后验过度依赖有限样本，忽略了分布的真实不确定性。

**泛化：模型不仅在训练数据上表现好，更能在未知数据上保持稳定与准确的能力。**

## ► 稳健 (Robustness)

- 什么是稳健？**当数据存在噪声、异常值或分布漂移时，模型仍能保持稳定表现。**
  - 概率语言：学习的是分布的整体结构，而非依赖少量极端样本；可用分布鲁棒优化、贝叶斯置信区间等方法提升稳健性



多层生物分子网络中对模型鲁棒性重要的基因更倾向于是生物上的关键基因，敲除基因调控网络中代谢疾病相关基因会对代谢网络鲁棒性造成较大伤害

- 怎么增强模型的稳健性？

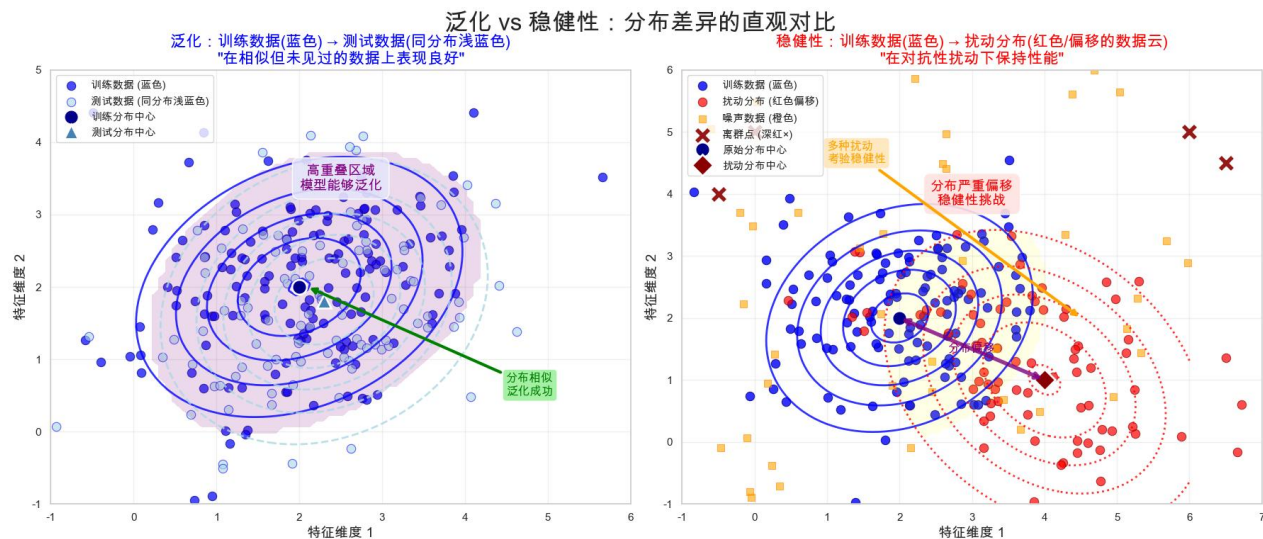
**噪声建模：**显式引入观测噪声（如高斯噪声、Laplace噪声），评估模型在扰动下的表现。

**分布不变性：**要求模型在不同分布（domain shift）下预测稳定，例如贝叶斯框架可通过先验+后验更新来适应变化。

**稳健：模型在面对噪声、异常或分布变化时仍能保持可靠性能的能力。**

# ► 为什么需要泛化与稳健？

- 泛化：解决“模型能否在**新数据**上保持表现”。
- 稳健：解决“模型能否在**噪声、扰动、极端条件**下保持表现”。
- 联系：两者都是在**训练数据之外**，保证模型表现的能力，只是关注的外部世界不同（普通新样本 vs 极端/扰动样本）。



- 现实类比：泛化：像学生做新题，题型变化不大，能举一反三；稳健：像学生遇到“刁钻题”或考试环境异常，依然能保持发挥。

泛化是“跨样本”的可靠性，稳健是“跨场景”的可靠性；两者共同决定模型在真实世界的可用性。

## ► 扩展：概率论中的泛化与稳健

### ➤ 泛化：

- 概率意义：训练样本只是总体分布 $P(X)$ 的有限采样，我们希望学到的模型不仅解释样本，还能对未来来自同一分布的数据表现良好。
- 数学表述：从样本分布推断总体分布 $P(X)$ （大数定律、泛化误差界）。
- 核心：解决“样本外”的表现问题。

### ➤ 稳健：

- 概率意义：当数据分布发生扰动（如噪声、极端值、轻微分布偏移），模型输出仍能保持稳定。
- 数学表述：在  $|P-Q| < \delta$  情况下，模型输出/推断结果的波动可控（稳健统计、Wasserstein ball、Huber M-estimator）。
- 核心：解决“数据偏差或干扰下”的可靠性。

## ► 扩展：数理统计中的泛化与稳健

### ► 泛化：

- 经验风险最小化（ERM）→ 期望风险最小化：

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^n L(f_{\theta}(x_i), y_i) \rightarrow \min_{\theta} \mathbb{E}_{(X,Y) \sim P} [L(f_{\theta}(X), Y)]$$

- 样本外表现就是泛化误差。
- 工具：一致性估计、大数定律、VC维、PAC-Bayes界。

### ► 稳健：

- 稳健估计：如 M-estimators（Huber 损失）、分位数回归，减少极端值影响。
- 分布鲁棒优化（DRO）：在邻域分布集合Q中最坏情况下优化：

$$\min_{\theta} \max_{Q \in \mathcal{Q}} \mathbb{E}_{(X,Y) \sim Q} [L(f_{\theta}(X), Y)]$$

|               |                                       |
|---------------|---------------------------------------|
| $\theta$      | 模型参数                                  |
| $Q$           | 某个具体分布（可能不等于训练分布）                     |
| $\mathcal{Q}$ | 分布不确定集合（如 Wasserstein-ball、KL-ball 等） |

- 工具：Influence Function、Breakdown Point、Wasserstein鲁棒性。

## ► 扩展：机器学习中的泛化与稳健

### ➤ 泛化：

- 训练集是总体分布 $P(X,Y)$ 的有限样本，模型需学得能在未见样本上保持良好性能的映射 $f(\theta)$ 。

- 数学表述：

目标是最小化真实风险： $R(\theta) = \mathbb{E}_{(X,Y) \sim P}[L(f_\theta(X), Y)]$

但实际可见的只是经验风险： $\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n [L(f_\theta(x_i), y_i)]$

- 核心任务：控制经验风险与真实风险的差距（VC 维、PAC-Bayes 界、正则化等）。

### ➤ 稳健：

- 数据可能存在噪声、对抗扰动或分布偏移，稳健模型需在这些“非理想”输入下仍保持稳定预测。
- 分布鲁棒优化（DRO）：在邻域分布集合 $\mathcal{Q}$ 中最坏情况下优化：

$$\min_{\theta} \max_{Q \in \mathcal{B}(P, \delta)} \mathbb{E}_{(X,Y) \sim Q}[L(f_\theta(X), Y)]$$

$\theta$  模型参数  
 $Q$  某个具体分布（可能不等于训练分布）  
 $\mathcal{B}(P, \delta)$   $P$  周围的分布邻域（如 Wasserstein ball）。

- 核心任务：当数据被扰动或分布轻微变化时，模型还能否可靠。



|      | 泛化（Generalization）        | 稳健（Robustness）                         |
|------|---------------------------|--|
| 目标   | 在未见样本上仍保持良好表现             | 在噪声/扰动/分布偏移下保持可靠                       |
| 概率视角 | 从训练样本分布推断总体分布，关注样本外数据的表现。 | 在真实分布附近的扰动分布中，保证模型输出稳定，关注分布偏移或噪声下的可靠性。 |
| 数学表述 | 控制经验风险与真实风险的差距            | 分布鲁棒优化（DRO）                            |
| 常用方法 | 正则化、交叉验证、PAC-Bayes、数据增强   | Huber 损失、对抗训练、DRO、稳健统计                 |
| 核心问题 | 样本外表现：新数据是否依然有效？          | 扰动下稳定性：数据异常是否影响结果                      |

泛化是模型在未知中保持准确，稳健是模型在扰动中保持坚韧。



# 目录章节

CONTENTS

01 引言

02 随机与分布

03 学习与推断

04 泛化与稳健

05 总结

## ► 总结

### ➤ 随机与分布：

- ✓ 数据来源于潜在随机过程。
- ✓ 样本服从某一未知真实分布。
- ✓ 分布决定模型训练与推断的基础。

### ➤ 学习与推断：

- ✓ 利用样本数据构建模型。
- ✓ 优化经验风险以逼近真实风险。
- ✓ 推断未见样本的输出结果。

### ➤ 泛化与稳健：

- ✓ 泛化：保持对未知样本的良好表现。
- ✓ 稳健：抵抗噪声和分布扰动保证稳定性。
- ✓ 两者共同提升模型在真实环境中的可靠性。

随机性与分布构建数据基础，学习与推断实现模型拟合，泛化保证模型在未知样本上的表现，稳健确保模型在扰动与分布偏移下的可靠性。

18.05 | Spring 2022 | Undergraduate

# Introduction To Probability And Statistics

Syllabus

Calendar

Instructor Insights ▼

Classes: Reading and In-class Materials

Problem Sets

Exams

R Studio Resources

Materials for Teachers

R: Information, Tutorials, and Sample Code ▼

Mathlets Applets

### Course Description

This course provides an elementary introduction to probability and statistics with applications. Topics include basic combinatorics, random variables, probability distributions, Bayesian inference, hypothesis testing, confidence intervals, and linear regression.

These same course materials, including interactive ... [Show more](#)

### Course Info

#### INSTRUCTORS

[Dr. Jeremy Orloff](#)

[Dr. Jennifer French Kamrin](#)


#### DEPARTMENTS

[Mathematics](#)

#### LEARNING RESOURCE TYPES

 Lecture Notes

 Problem Sets with Solutions


 Exams with Solutions

 Readings

 Activity Assignments with Examples

 Supplemental Exam Materials

 Tools

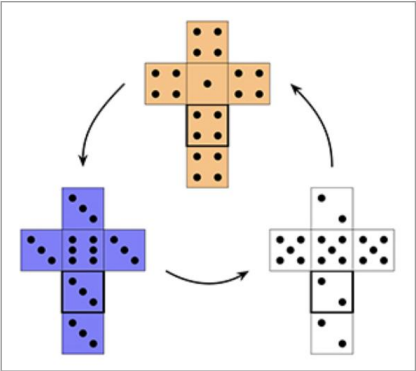
 Instructor Insights

#### TOPICS

▼ [Mathematics](#)

[Discrete Mathematics](#)

[Probability and Statistics](#)



Jon has three six-sided dice with unusual numbering. A game consists of two players each choosing a die. They roll once and the highest number wins. Which die would you choose? (Image from class slides.)

[Download Course](#)

# 感谢聆听



Personal Website: <https://www.miaopeng.info/>



Email: [miaopeng@stu.scu.edu.cn](mailto:miaopeng@stu.scu.edu.cn)



Github: <https://github.com/MMeowwhite>



Youtube: <https://www.youtube.com/@pengmiao-bmm>