

Ch00: 先导课程

Ch00: 先导课程

Introduction to Probability and Statistics

August 25, 2025

目录

1. 概率的起源: 点数分配问题
2. 概率和统计的发展历史
3. 概率统计的认识和使用
4. 概率和统计的联系和区别

概率的起源：点数分配问题

例 0.1 (点数分配问题-1) 两人进行一场赌博, 5 局 3 胜, 赌金为 1000; 假设当前比分为 2:1, 而比赛由某种原因不得不中止.

问题：最“公平合理”的奖金分配方式？

这个问题最早由 1495 年意大利数学家/修道士帕西奥尼 (Luca Pacioli), 持续了 150 年左右.

概率的起源：点数分配问题

例 0.2 (点数分配问题-2) 现在需要比较两个选手的竞技水平, 以一场胜负定输赢, 胜利者可以赢得所有奖金.

请问:

1. 一场定输赢是否公平?
2. 一场定输赢是否足以证明两个选手的水平高低?

概率的起源：点数分配问题

例 0.3 (点数分配问题-3) 为了进一步比较两个选手的竞技水平, 比赛方举行多轮比赛, 比如 5 局 3 胜 (定局赛 Best of 3 或者抢三 Race to 3), 领先对手 3 局即胜制 (净胜三局 Win by 3) 等. 现在比赛规定: 5 局 3 胜, 获胜者获得所有奖金. 比赛进行到一定的比分, 由某种原因不得不中止.

1. 假设当前比分为 2:1
2. 假设当前比分为 2:2

请问:

1. 面对上述两种情况, 坚持“选择一名选手获得全部奖金”的方案是否合理?
2. 面对上述两种情况, 最“公平合理”的奖金分配方式分别是?

点数分配问题——问问 GPT



这个问题涉及到在比赛被迫中止的情况下，如何合理分配奖金的问题。为了更好地回答这个问题，我们可以从概率和公平性的角度出发，分析不同比分下两名选手各自的获胜概率，并以此为基础提出合理的奖金分配方案。

1. 是否坚持“选择一名选手获得全部奖金”的方案合理？

在比赛进行到某一比分时中止，选择一名选手获得全部奖金并不一定是合理的。这种方式忽略了两名选手在比赛中的努力和未来可能的比赛结果，尤其是在双方比分接近的情况下，这样的决定可能会显得不公平。因此，在大多数情况下，坚持这种方案是不合理的。

2. 不同比分下的公平奖金分配方式

我们可以通过计算每名选手在比赛中止时的获胜概率，来确定合理的奖金分配方式。

1) 当前比分为 2:1

- 选手 A 已赢 2 局，选手 B 赢 1 局。
- 未来比赛可能有两种情况：
 - A 再赢 1 局，最终比分为 3:1，A 获胜。
 - B 连赢 2 局，最终比分为 3:2，B 获胜。

假设两名选手实力相当，赢得每一局的概率均为 50%。

- A 获胜的概率 $P(A) = P(A \text{再赢1局}) = 0.5$
- B 获胜的概率 $P(B) = P(B \text{连赢2局}) = 0.5 \times 0.5 = 0.25$

因此，在这种情况下，公平的奖金分配方式是：

- A 获得 $\frac{0.75}{1} \times \text{奖金}$ （因为 A 目前赢 2 局，有更高的获胜概率）
- B 获得 $\frac{0.25}{1} \times \text{奖金}$

2) 当前比分为 2:2

- 双方目前打成 2:2 平。
- 未来比赛只有一局，胜者为最终获胜者。

在这种情况下，由于只剩下一局，且双方实力相当，因此获胜概率均为 50%。

因此，最公平的奖金分配方式是：

- A 和 B 各获得 50% 的奖金。

结论

1. **“选择一名选手获得全部奖金”的方案**通常不合理，因为它忽略了比赛的未确定性和公平性。
2. **公平合理的奖金分配方式：**
 - 如果比分为 2:1，A 和 B 的奖金分配比例应为 75% 对 25%。
 - 如果比分为 2:2，A 和 B 各分得 50% 的奖金。

概率的起源：点数分配问题

例 0.4 (点数分配问题-4) 为了进一步比较两个选手的竞技水平, 比赛方举行多轮比赛, 比如 5 局 3 胜 (定局赛 Best of 3 或者抢三 Race to 3), 领先对手 3 局即胜制 (净胜三局 Win by 3) 等. 现在比赛规定: 5 局 3 胜, 获胜者获得所有奖金. 比赛进行到比分 2:1 时, 由某种原因不得不中止.

根据上述讨论, 坚持“选择一名选手获得全部奖金”的方案已不再合理, 且根据 GPT 的回答, 选手采用 3:1 的比例分配奖金. 请问该“公平合理”分配方式是否唯一, 是否存在其他方案?

解答问题：点数分配-4

问题：为了进一步比较两个选手的竞技水平，比赛方举行多轮比赛，比如 5 局 3 胜 (定局赛 Best of 3 或者抢三 Race to 3)，领先对手 3 局即胜制 (净胜三局 Win by 3) 等。现在比赛规定：7 局 4 胜，获胜者获得所有奖金。比赛进行到比分 2:1 时，由某种原因不得不中止。根据上述讨论，坚持“选择一名选手获得全部奖金”的方案已不再合理，且根据 GPT 的回答，选手采用 3:1 的比例分配奖金。请问该“公平合理”分配方式是否唯一，是否存在其他方案？

解答思路：根据两个选手接下来的胜率来确定奖金分配方式

- 假设每局对局两个选手的胜率是 1:1
- 假设每局对局两个选手的胜率是 2:1

解答问题：点数分配-4

问题：为了进一步比较两个选手的竞技水平，比赛方举行多轮比赛，比如 5 局 3 胜 (定局赛 Best of 3 或者抢三 Race to 3)，领先对手 3 局即胜制 (净胜三局 Win by 3) 等. 现在比赛规定: 7 局 4 胜, 获胜者获得所有奖金. 比赛进行到比分 2:1 时, 由某种原因不得不中止. 根据上述讨论, 坚持“选择一名选手获得全部奖金”的方案已不再合理, 且根据 GPT 的回答, 选手采用 3:1 的比例分配奖金. 请问该“公平合理”分配方式是否唯一, 是否存在其他方案?

解答思路：根据两个选手接下来的胜率来确定奖金分配方式, 假设每局对局两个选手的胜率是 1:1, 则按照既定方式完成比赛至少需要进行 2 场, 最多进行 4 场. 所以选手 A 赢的概率为: 以下三种情况求和; 然后选手 B 赢的概率呢?

- 进行 2 场, 选手 A 赢的概率: $(1/2)^2$
- 进行 3 场, 选手 A 赢的概率 $C_2^1(1/2)^2(1/2)$
- 进行 4 场, 选手 A 赢的概率 $C_3^2(1/2)^2(1/2)^2$

解答问题：点数分配-4

问题：为了进一步比较两个选手的竞技水平，比赛方举行多轮比赛，比如 5 局 3 胜 (定局赛 Best of 3 或者抢三 Race to 3)，领先对手 3 局即胜制 (净胜三局 Win by 3) 等. 现在比赛规定: 7 局 4 胜, 获胜者获得所有奖金. 比赛进行到比分 2:1 时, 由某种原因不得不中止. 根据上述讨论, 坚持“选择一名选手获得全部奖金”的方案已不再合理, 且根据 GPT 的回答, 选手采用 3:1 的比例分配奖金. 请问该“公平合理”分配方式是否唯一, 是否存在其他方案?

解答思路：根据两个选手接下来的胜率来确定奖金分配方式, 假设每局对局两个选手的胜率是 **2:1**, 则按照既定方式完成比赛至少需要进行 2 场, 最多进行 4 场. 所以选手 A 赢的概率为: 以下三种情况求和; 然后选手 B 赢的概率呢?

- 进行 2 场, 选手 A 赢的概率: $(2/3)^2$
- 进行 3 场, 选手 A 赢的概率: $C_2^1(2/3)^2(1/3)$
- 进行 4 场, 选手 A 赢的概率: $C_3^2(2/3)^2(1/3)^2$

概率的起源：点数分配问题

例 0.5 (点数分配问题-5) 为了进一步比较两个选手的竞技水平, 比赛方举行多轮比赛, 比如 5 局 3 胜 (定局赛 Best of 3 或者抢三 Race to 3), 领先对手 3 局即胜制 (净胜三局 Win by 3) 等. 现在比赛规定: 选手 A 赢 10 局获胜, 选手 B 赢 7 局获胜, 获胜者赢得所有奖金. 比赛进行到一定的比分, 由某种原因不得不中止. 假设当前比分为 6 : 3. 问题: 面对上述两种情况, 最“公平合理”的奖金分配方式分别是?

小结: 点数分配问题

比赛机制	奖金分配方式	
	根据剩余胜率	根据剩余胜率和先验概率
一局输赢	例 0.2	
定局赛	例 0.3	例 0.4
变局赛	例 0.5	

概率的起源

概率起源于公元 1650 年左右的法国, 萌芽于赌博

- 赌博流行且时尚, 不受法律限制
- 赌博变得更加复杂, 风险增大
- 有必要通过数学方法来计算胜率
- 法国贵族德梅根 (De Mere) 关心点数分配问题
- 克里斯蒂安·惠更斯 (Christiaan Huygens) 在《论赌博中的计算》中提出了点数分配问题的数学解法, 出现了期望的概念.

概率的形成和发展 (18 世纪)

贝努利 (James Bernoulli): 《推想的艺术》, 1713 年

- 大数定律
- 频率稳定性理论化
- 特殊问题到一般理论

棣谟佛 (Abraham de Moivre): 《机遇原理》, 1718 年

- 概率乘法法则
- 正态分布律
- 中心极限定理的一个特例

概率的进一步发展 (19 世纪)

拉普拉斯 (Pierre-Simon Laplace) :

- 《Theorie Analytique des Probabilities》, A mathematical theory of probability with an emphasis on scientific applications

Greats emerge.

- 高斯 (Carl F. Gauss)
- 麦克斯韦 (James C. Maxwell)
- 吉布斯 (Josiah W. Gibbs)

概率的日渐成熟 (20 世纪)

1900 年, 希尔伯特 (David Hilbert) 提出了著名的 23 个数学问题

- 概率公理化 (Axiomatic Probability)

柯尔莫哥洛夫 (Andrey Kolmogorov):

- published 《Foundations of the Theory of Probability》 or 《Grundbegriffe der Wahrscheinlichkeitsrechnung》, 1933
- 提出了概率公理化三要素:
 - 非负性、规范性、可列可加性
- 建立概率公理化理论体系, 利用基本性质来定义概率, 可媲美于欧几里得几何公理化

现代概率统计: 测度论 (Measure Theory)

Recommended Readings

- 《20 世纪统计怎样变革了科学: 女士品茶》by David Salsburg, 故事: 英国女士的下午茶, 内核: 近代数理统计中的试验设计法
- 《赤裸裸的统计学》by Charles Wheelan
- 《醉汉的脚步》by Leonard Mlodinow, 让生活漫游在随机性、偶然性和概率中
- 《简单统计学: 如何轻松识破一本正经的胡说八道》by Gary Smith, 一方面用简单的统计学原理揭穿生活中的各种数据骗局, 另一方面揭穿概率统计自身的骗局 (度量不确定性本身就带有不确定性)
- 《统计学的世界》by David S. Moore, 专业书籍的通俗读物

“有用的” 概率统计

- 1832 年, 霍乱袭击伦敦, 导致 6500 人死亡. 当时的医疗机构认为霍乱是由呼吸有毒气体引起的.
- 1849 年, 36 岁的医生约翰·斯诺 (John Snow) 发表了一篇论文《论霍乱的传播模式》, 认为霍乱是由引用污染水导致的.
- 斯诺考察了 1854 年霍乱流行前 7 个星期的所有病人死亡记录, 并且确定了由这两家水务公司提供水源的家庭.

	家庭数量	霍乱死亡数量	每一万户家庭的死亡数量
萨瑟克和沃克斯豪尔公司	40046	1263	315
兰贝斯公司	26107	98	37
伦敦其他地区	256423	1422	59

“相悖的” 概率统计

- 20 世纪 70 年代, 有人指控加州大学伯克利分校研究生院歧视女性申请人.

	申请人	录取率
男性	8842	44%
女性	4321	35%

“相悖的” 概率统计

- 法院启动了一项调查, 以确定哪些系的问题最为严重.

	总计		男性		女性	
系	申请人	录取率	申请人	录取率	申请人	录取率
1	933	64%	825	62%	108	82%
2	585	63%	560	63%	25	68%
3	918	35%	325	37%	593	34%
4	792	34%	417	33%	375	35%
5	584	25%	191	28%	393	24%
6	714	6%	373	6%	341	7%
总计	4526	39%	2691	45%	1835	30%

“相悖的” 概率统计 – 辛普森悖论

- 当聚合数据被分解时其中的模式发生逆转的现象.
- 分解聚合数据本质上是一种关于 (分子/分母) 数字的运算, 而这种运算是由数字的“定义”带来的.

$$\begin{array}{l} \text{e.g., } \frac{2}{3} \neq \frac{1}{2} + \frac{1}{1} \\ \text{but } \frac{2}{3} > \frac{1}{2} \quad \text{and} \quad \frac{2}{3} < \frac{1}{1} \end{array}$$

“相悖的” 概率统计：梵蒂冈谋杀率

梵蒂冈 (Vatican City) 是一个全球最小的国家, 常住人口大约在 800 人左右。但在某一年公布的“谋杀率排行榜”中, 梵蒂冈的谋杀率竟然名列前茅, 甚至超过了许多以暴力闻名的大国.

当年警长夫妇遭遇谋杀. 根据国际统计惯例, 谋杀率通常按“每 10 万人口中发生的谋杀数”来计算. 因此, 有

$$\text{谋杀率} = \frac{\text{谋杀人数}}{\text{总人口数}} \times 1 \text{ e}6 = 125$$

该结果可以与其他国家做对比

- 美国的年均谋杀率大约为 5-6
- 墨西哥某些高犯罪地区可能达到 30-40

概率统计的有用和相悖

有用的概率统计

- 点数分配案例中, 利用概率统计解决未完赛时奖金分配问题;
- 斯诺霍乱实验中, 利用对比实验和统计指标实现溯因;
- 出场顺序无碍竞赛公平
- ...

相悖的概率统计

- UCB 男女录取案例中, 归因: 辛普森悖论;
- 梵蒂冈谋杀率案例中, 归因: 条件改变概率;
- ...

概率统计时灵时不灵, 甚至是“任人打扮的”

“任人打扮的” 概率统计：选手可靠

问题: 选手 A 和选手 B 谁在关键时刻更可靠?

- 建模: 统计关键时刻选手 A 和 B 的得分情况

选手	得 1 分	得 5 分	得分率	逆转次数	逆转率
选手 A	9:9	0:2	81. 80%	0	0%
选手 B	0:0	2:9	22. 20%	1	100%

- 请问: 该表格是否足以支持论点?

“任人打扮的” 概率统计：选手可靠

问题: 选手 A 和选手 B 谁在关键时刻更可靠?

- 建模: 统计关键时刻选手 A 和 B 的得分情况

选手	得 1 分	得 5 分	得分率	逆转次数	逆转率
选手 A	9:9	0:2	81. 80%	0	0%
选手 B	0:0	2:9	22. 20%	1	100%

- 请问: 该表格是否足以支持论点?
 - 观点一: 选手 A 更可靠, 原因: 得分率 81. 80% : 22. 20%
 - 观点二: 选手 B 更可靠, 原因: 逆转率 0% : 100%
 - 观点三: ???

“任人打扮的” 概率统计：计算工资水平

问题: 请用恰当的“指标”反映某公司的工资水平. 如下的统计指标:

总经理	经理	部门骨干	职员	职员	职员	...
100w	50w	25w	18w	15w	21w	...

- 平均值: 37.5w
- 中位数: 25.5w
- 众数: 18w

你觉得哪个指标才能反映某公司的工资水平呢?

概率统计的“任人打扮”

有用的概率统计

- 点数分配案例中, 利用概率统计解决未完赛时奖金分配问题;
- 斯诺霍乱实验中, 利用对比实验和统计指标实现溯因;
- 出场顺序无碍竞赛公平

相悖的概率统计

- UCB 男女录取案例中, 归因: 辛普森悖论;
- 梵蒂冈谋杀率案例中, 归因: 条件改变概率;

概率统计时灵时不灵, 甚至是“任人打扮的”, why?!

- 选手 A 和选手 B 谁在关键时刻更可靠? (文无第一, 没有定论)
- 工资水平, 恰当的指标即可恰当地反映工资水准 (有失公平)

概率统计的“滥用”

案例一：多次抛投一枚质量均匀的硬币, 实验结果发现正面向上的次数接近总抛投次数的一半. 为什么？

案例二：请问可以据此来预测接下来抛投的结果吗？

案例三：如果用一台高速摄像机实时追踪硬币下落的过程, 是否可以对抛投结果进行精确预测？

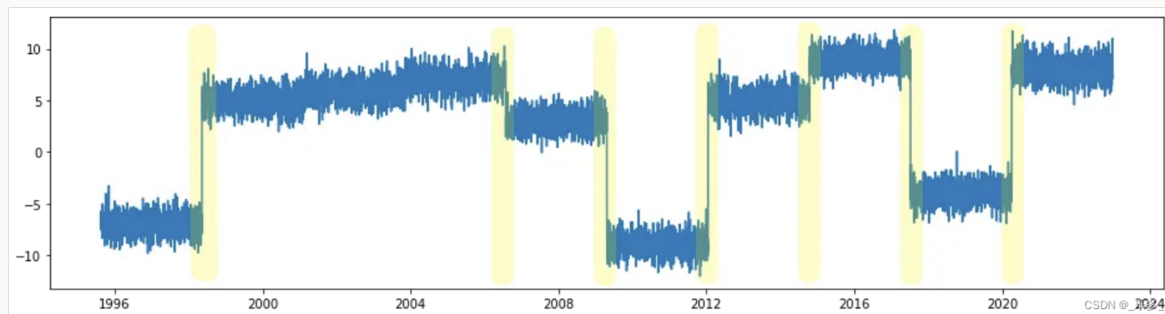
概率统计的特点：事后诸葛亮与大数定律

案例一中抛投硬币具有统计规律，即抛投硬币的随机试验符合大数定律。注意该规律只能反映“已经发生的”随机试验的统计规律，不确保一定能反映未来的规律。

□ 硬币抛投具有统计规律

实验次数	第一次		第二次		第三次		第四次		第五次	
	频数	频率	频数	频率	频数	频率	频数	频率	频数	频率
$n=100$	47	0.47	54	0.54	53	0.53	48	0.48	43	0.43
$n=1000$	503	0.503	498	0.498	510	0.510	469	0.469	512	0.512
$n=10000$	5008	0.5008	4974	0.4974	4918	0.4918	5074	0.5074	5003	0.5003

□ 时间序列的“突变” 统计规律只能反映已发生的规律



概率统计的特点：独立性

案例二中随机试验具有独立性 (即前后两次试验结果无关), 往往难以对下一次抛投结果做精准预测.

- 双色球
- 抽签
- ...

概率统计的特点：不确定性与观察粒度有关

在案例三中, 如果用一台高速摄像机实时追踪硬币下落的过程, 是否可以对抛投结果进行精确预测?

答案是确定的. 为什么呢?

因为高速摄像的实时追踪“消除”了抛投试验中的随机性 (或者随机因素).

思考: 随机因素有哪些?

概率统计的特点：不确定性与观测粒度和结果有关

举例：某女性出生于 1898 年 06 月 29 日中午 12 点

玄学观测粒度与预测：

- 粗粒度：只看太阳星座 – 巨蟹座（Cancer）
 - 性格：敏感、情感丰富、重视家庭、安全感；优点：富同理心、温柔体贴；缺点：情绪波动大、可能依赖环境
- 中粒度：太阳星座 + 月亮星座 – 大约落在天秤座（Libra）
 - 外显太阳：对家庭和亲密关系敏感
 - 内含月亮：情感渴望和谐、平衡、社交能力强
- 较细粒度：太阳星座 + 月亮星座 + 上升星座 – 大约落在狮子座（Leo）
 - 个人对外呈现的方式：自信、热情、具有表现力
 - 综合来看：情感温柔、追求和谐、表现自我和领导力
- 最细粒度：太阳星座 + 月亮星座 + 上升星座 + 行星分布

概率统计的特点：不确定性与观测粒度和结果有关

举例：某女性出生于 1898 年 06 月 29 日中午 12 点

玄学观测粒度与预测：

- 粗粒度：戊戌 戊午 癸亥
 - 日支亥水，内心柔韧，具智慧之象
 - 两个戊土克水，处境偏弱
 - 亥午暗合，通过正官产生偏财
- 细粒度：戊戌 戊午 癸亥 戊午
 - 日支亥水，内心柔韧，具智慧之象 (不变)
 - 三个戊土克水，处境偏弱 (加剧)
 - 亥午暗合，通过正官产生偏财 (产生矛盾)
 - 午午自刑，性格中容易有执拗或情绪反复 (新发现)

概率统计的特点：不确定性与观测粒度和结果有关

举例：某女性出生于 1898 年 06 月 29 日中午 12 点

玄学观测粒度与预测：

- 比较完整版：戊戌 戊午 癸亥 戊午
 - 家庭、早年环境较为稳重，土性稳重但略刚硬
 - 少年时期行动力、学习能力较强，情绪和外部压力较大 (压力主要来自于钱)
 - 青年时期需环境对自己要求多，内心容易紧张或压抑，要努力突破环境制约
 - 中年时期突破枷锁，比劫抗官杀，拥有很多姐妹，财运亨通
 - 个性方面，内心聪慧、柔韧、感性，个性柔弱中带有一定忍耐力
 - 形象方面，癸水正官象
 - 生平通过正官产生偏财

表观不确定性 VS 根本不确定性

- 表观不确定性 (Epistemic Randomness)
 - 经典物理, 即由于我们对系统的初始条件或演化规律了解不完全而产生的不确定性.
 - 抛硬币, 如果完全知道初速度、角动量、空气阻力等因素, 其结果就不再随机.
- 根本不确定性 (Ontic Randomness)
 - 即使观测再精确, 随机性/不确定性也无法消除.
 - 量子态的波函数坍缩 (Wavefunction Collapse) —— 量子力学本身的基本性质
 - 单光子通过偏振片的行为、放射性衰变 (铀-238) —— 即使无限提高观测能力, 也无法预判单个光子的行为、确切衰变时刻

概率统计的“滥用”

案例一：多次抛投一枚质量均匀的硬币, 实验结果发现正面向上的次数接近总抛投次数的一半. 为什么? (大数定律与统计规律)

案例二：请问可以据此来预测接下来抛投的结果吗? (独立性与不可预测性)

案例三：如果用一台高速摄像机实时追踪硬币下落的过程, 是否可以对抛投结果进行精确预测? (不确定性与观测粒度和结果有关)

概率与统计的联系与区别: 例 0.6

例 0.6 (Poker Hands)

- 由 52 张牌构成的扑克牌, Decks of 52 cards
 - 13 ranks: 2, 3, 4, ..., J, Q, K, A
 - 4 suits: S, H, C, D
- Gaming:
 - a one-pair hand consists of 5 cards
- **Questions:** the probability of a one-pair hand is
 - count less than point 12
 - all cards are “S”

概率与统计的联系与区别: 例 0.6

There are two ways to handle the problem of “count less than point 12”

- **combination-based:**

1. the target one-pair hand comprises $\{2, 2, 2, 2, 3\}$
2. possible counts 4
3. all combination $\binom{52}{5} = 2,598,960$
4. the probability $\frac{4}{2,598,960}$

- **sampling-based:**

1. build a trial by sampling a one-pair hand from 52 cards
2. repeat n trials
3. count the number m of “appropriate” trials
4. regard the frequency $\frac{m}{n}$ as the probability

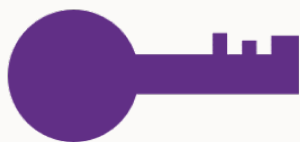
概率与统计的联系和区别

- **Probability** means possibility, a branch of mathematics concerned with analyzing random phenomena. The possibility indicates how likely an event is to occur, expressed as a number ranging $[0, 1]$ or $[0\%, 100\%]$.
 - logically self-contained
 - a few (mathematical or physical) rules for computing probabilities
 - one correct answer
- **Statistics** is a branch of applied mathematics that collects, describes, analyzes, and infers conclusions from quantitative data.
 - messier and more of an art
 - get experimental data and try to draw probabilistic conclusions
 - no single correct answer

Appendix: 概率与统计

- 当事物的 (数学、物理、...) 规律比较简单或者明确的时候, 我们倾向于使用 rules-based methods
 - 数学建模、物理建模、...
- 当事物的 (隐性) 规律比较复杂, 但具备一定的试验条件时, 我们倾向于使用 trial-based 或者 data-driven methods
 - 试验设计、...
- 在 AI 领域, 更侧重于后者, 或者两者结合的方法

学科定位



Probability and Statistics



Artificial Intelligence



Computer Science



Finance, Economy ...



Health-care



Physical Science



Daily Life



Military Science and Technology

培养方案

先修课程:

- Mathematical Analysis
- Advanced Algebra
- Computer Programming

后续课程:

- Machine Learning
- Statistical Learning
- Data Mining
- Design of Experiments, DOE

概率统计及其相关的顶级国际期刊与会议

- 期刊

- Annals of Statistics (AoS)
- Journals of the American Statistical Association (JASA)
- Annals of Probability
- [Journal of Machine Learning Research \(JMLR\)](#)

- 会议

- ICML: International Conference on Machine Learning
- [COLT: Annual Conference on Learning Theory](#)
- STOC: ACM Symposium on Theory of Computing
- FOCS: IEEE Symposium on Foundations of Computer Science