

概 率 论 与 数 理 统 计

Probability and Statistics

南 京 大 学 张 绍 群

更新：September 3, 2024

Course Information

Instructor: 张绍群

- Course web: LINK
- Time: 周二 3-4 节、周四 5-6 节, 1-17 周 (09.02 –12.29)
- Location: 南京大学苏州校区东校区, 南雍楼 - 西 426

Teaching Assistants:

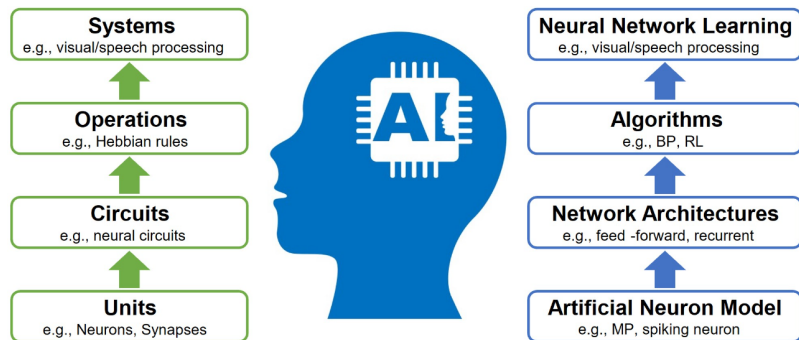
- Jia-Yi Chen, Jia-Wei Huang, Shuang Liang, Qian Sun

Please arrange an appointment, if you want to have a meeting with me.

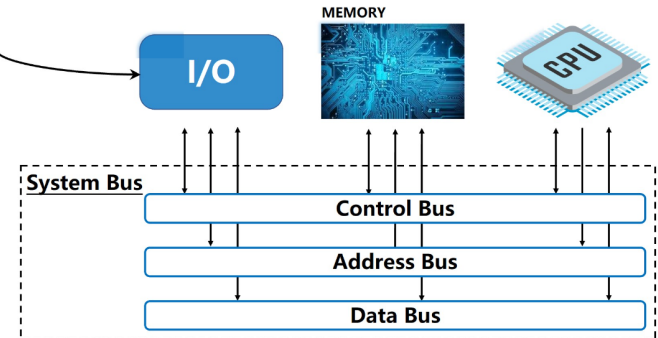
- Contact me with zhangsq@nju.edu.cn
- Office: 南雍楼东区 524

Information to Instructor

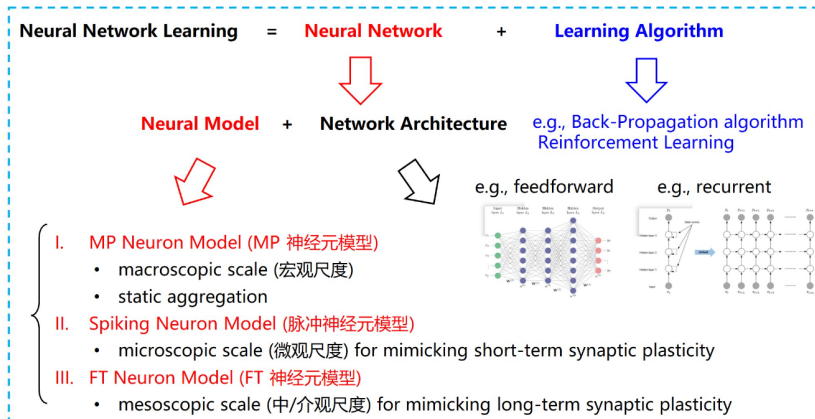
Our main studies on this topic are summarized in the figure below.



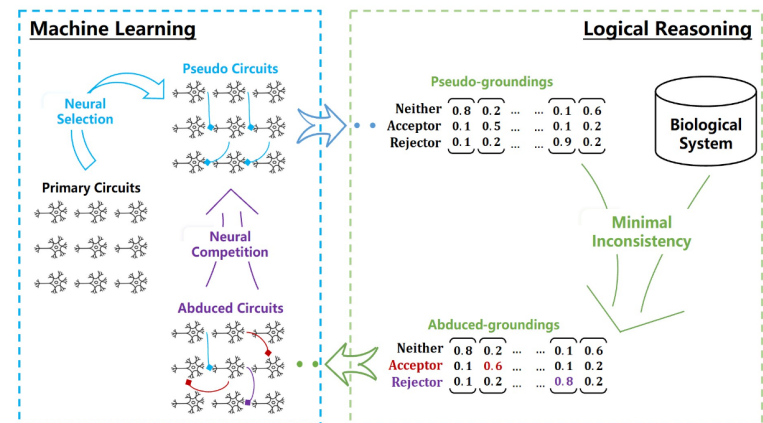
(A) Biological system colocates the operations with the physical substrate it is processed on.



(B) Modern computing, based on the von Neumann architecture, optimizes for generality such that learning algorithms are treated somewhat independently of the hardware they are processed on.

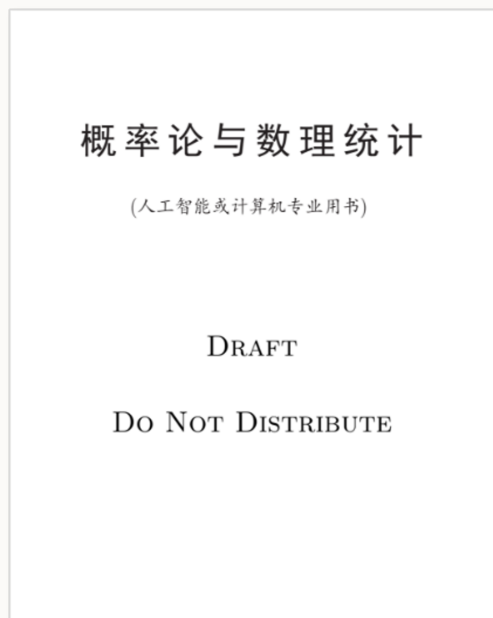


(C) Paradigm of neural network learning and Types of artificial neuron models.



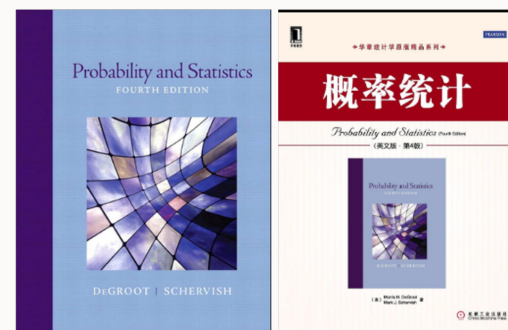
(D) Neural selection and competition during memory formation.

Course Information — Textbooks



概率论与数理统计

- 盛骤、谢式千等编
- 高等教育出版社



Probability and statistics

- M. H. DeGroot and M. J. Schervi
- 机械工业出版社

Period: 1-17 周 (09.02 – 12.29)

第一学期											
年	月	周次	日期	星期	一	二	三	四	五	六	日
二零二五年	九月	1	单	2	3	4	5	6	7	8	
		2	双	9	10	11	12	13	14	15	
		3	单	16	中秋	18	19	20	21	22	
		4	双	23	24	25	26	27	28	29	
		5	单	30							
	十月	6	双	7	8	9	10	11	12	13	
		7	单	14	15	16	17	18	19	20	
		8	双	21	22	23	24	25	26	27	
		9	单	28	29	30	31				
	十一月	10	双	4	5	6	7	8	9	10	
		11	单	11	12	13	14	15	16	17	
		12	双	18	19	20	21	22	23	24	
		13	单	25	26	27	28	29	30		
	十二月	14	双	2	3	4	5	6	7	8	1
		15	单	9	10	11	12	13	14	15	
		16	双	16	17	18	19	20	21	22	
		17	单	23	24	25	26	27	28	29	
	一月	考试	30	31							
			6	7	8	9	10	11	12		
寒		13	14	15	16	17	18	19			
		20	21	22	23	24	25	26			
二月	假	27	28	春节	30	31					

- 总计: 33 次课

 - 第 17 周安排 1-2 次答疑
 - 每 6-8 次课安排一次习题课
 - 大概 27 次讲授式课堂

2024-2025 学年第一学期校历及课时安排

Course Information — Lesson Plan

节次	内容	节次	内容
1	CH00. 导论	12	CH05. 二维随机向量和联合分布
2	CH01. 随机事件、频率与概率公理化	13	CH05. 随机变量的独立性和条件分布
3	CH01. 古典概型与几何概型	14	CH06. 多元随机向量的运算 - 1
4	CH01. 十二重计数	15	CH06. 多元随机向量的运算 - 2
E1	习题课	16	CH06. 多元随机向量的数字特征 - 1
5	CH02. 条件概率	17	CH06. 多元随机向量的数字特征 - 2
6	CH02. 全概率公式和贝叶斯公式	E4	习题课
7	CH02. 随机事件的独立性	18	CH08. 大数定律和中心极限定理
E2	习题课	19	CH09. 统计基础概念
8	CH03. 离散型随机变量	20	CH09. 抽样分布定理
9	CH03. 常见的离散型随机变量和分布	E5	习题课
Interactive Exercise - 1		21	CH10. 参数估计: 点估计
10	CH04. 连续型随机变量	22	CH10. 参数估计: 区间估计
11	CH04. 常见的连续型随机变量和分布	23	CH11. 假设检验
E3	习题课	E6	习题课
Interactive Exercise - 2		24	课程总结
期中考试		期末考试	

Course Information —— Lesson Plan

1. 课堂管理:

- 不点名
- 鼓励参与课堂互动

2. 课程管理:

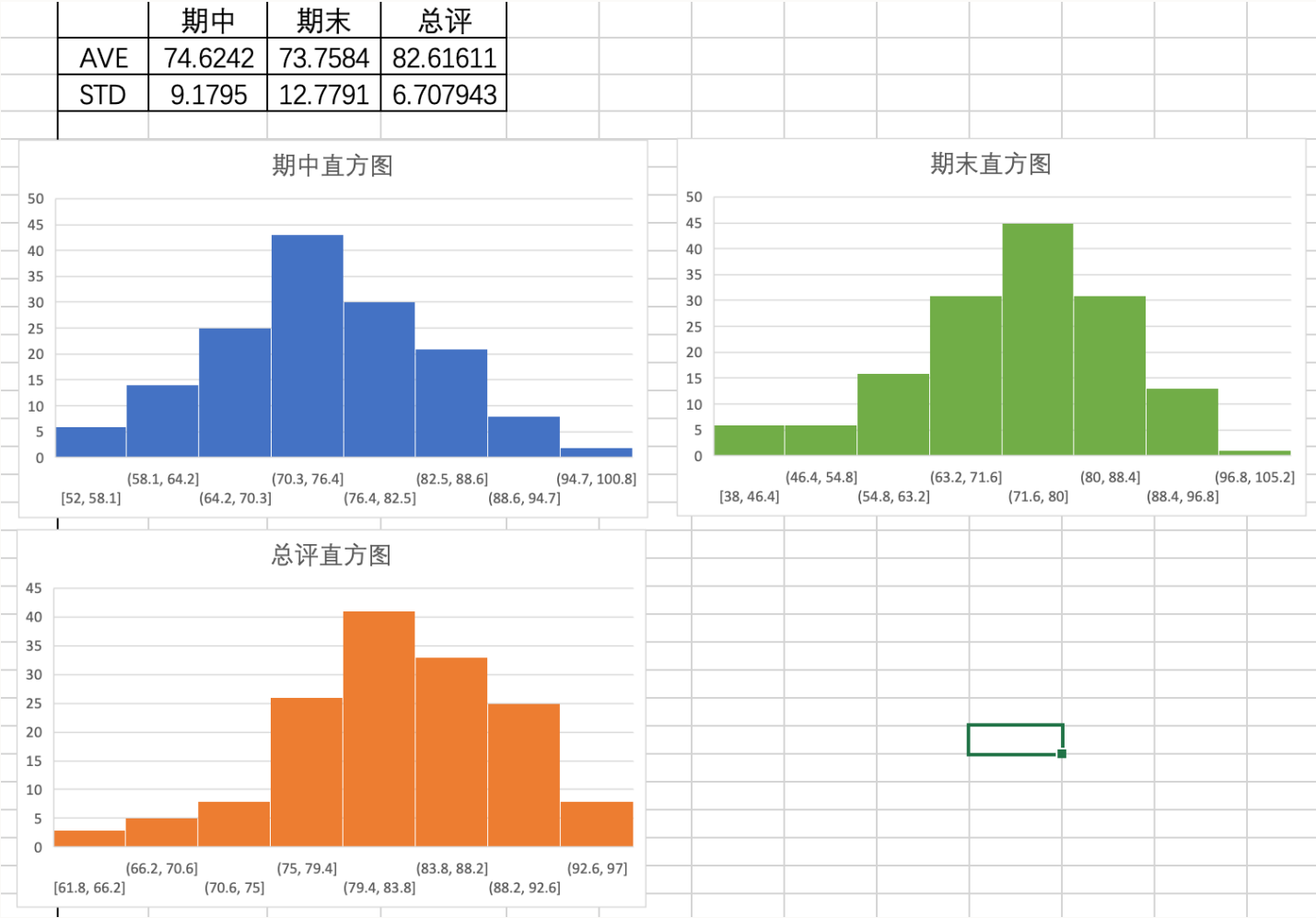
- 课件: 每次课前 1-2 天将新课件、课后 0-1 天内将更新课件上传至课程主页
- 作业: 每周 1-2 次; 发布在 QQ 群中或者课程主页上; 每周二上课前交 (上一周的) 作业;
- 思考和建议: 可以私发到我的邮箱 zhangsq@nju.edu.cn
- 鼓励大家可以跟我有课堂外的交流、甚至是合作 (前提是什么?)

Course Information —— 考评

总评成绩 = 平时成绩 (40%) + 期中考试 (20%) + 期末考试 (40%)

1. 平时成绩 $\approx$ 作业 (60% 基础 + 30% 强化 + 10% 思考)
2. 考试难度构成为 60% 基础: 30% 提升: 10% 挑战
3. 评分原则: 思路清晰与否 (过程分)、计算是否准确 (结果分)
4. 我不以师生课程评价为目标导向, 力求每月做一次课程反馈

Course Information ——— 考评



Course Information —— About our lesson

1. Motivations:

- handles fundamental terminologies (concepts, formulas, theorems, etc.)
- grasps the (intuitive) understanding of Probability and Statistics
- provides support for machine learning or artificial intelligence
- **NOT** focuses on exams, postgraduate entrance examinations, etc.

2. 一门知识密集型课程, 涉及大量的术语、公式、定理, 需要练习

3. 该课程将尽量 **脱离书本**, 关注于: 示例、理解、知识结构

- 很多的阅读、基础练习需要大家在课堂之外完成
- 推荐书籍, 安排习题、思考题, 习题课 supported by 助教
- 设置更多的交互式课堂 (Interactive Exercises), 与更广泛的知识建立联系 (AI、ML 等)

Ch00: 先导课程

Ch00: 先导课程

Introduction to Probability and Statistics

September 3, 2024

概率的起源

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

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

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), 优先赢得 7 局的获胜 (抢七) 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?

概率的起源——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。

概率的起源——问问 GPT

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——思考

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), 优先赢得 7 局的获胜 (抢七) 等。现在比赛规定: 优先赢得 10 局的人, 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。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?

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), 优先赢得 7 局的获胜 (抢七) 等。现在比赛规定: 甲赢 10 局获胜, 乙赢 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?

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。

$$\text{e.g., } \frac{2}{3} \neq \frac{1}{2} + \frac{1}{1}$$

“任人打扮的” 概率统计

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

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: ???

概率与统计：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：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

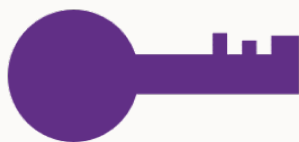
概率与统计的联系和区别

- **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