概 率 论 与 数 理 统 计

Probability and Statistics

南 京 大 学　张 绍 群

更新：September 4, 2023

Course Information

Instructor: 张绍群

- Course web: LINK
- Time: 周一 5-6 节、周四 1-2 节, 1-17 周（09.04–12.31）
- Location: 南京大学苏州校区东校区, 南雍楼-西 209

Teaching Assistants:

- Jia-Yi Chen
- Xin-Shuang Zhang

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

- Contact me with zhangsq@nju.edu.cn
- Office: where?

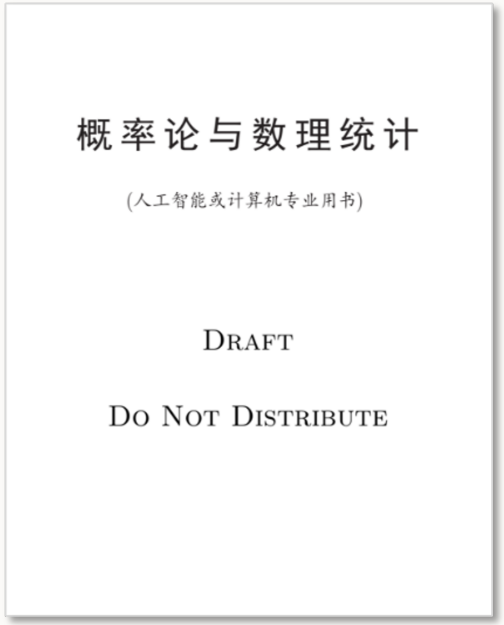
Period: 1-17 周 (09.04 – 12.31)

第一学期

年	月	周次	日期	星期一	二	三	四	五	六	日
二零二三年	九月	1	单	4	5	6	7	8	9	10
		2	双	11	12	13	14	15	16	17
		3	单	18	19	20	21	22	23	24
		4	双	25	26	27	28	中秋	30	
	十月	5	单	庆	节	4	5	6	7	国 8
		6	双	9	10	11	12	13	14	15
		7	单	16	17	18	19	20	21	22
		8	双	23	24	25	26	27	28	29
		9	单	30	31					
	十一月	10	双	6	7	1 8	2 9	3 10	4 11	5 12
		11	单	13	14	15	16	17	18	19
		12	双	20	21	22	23	24	25	26
		13	单	27	28	29	30			
	十二月	14	双	4	5	6	7	1 8	2 9	3 10
		15	单	11	12	13	14	15	16	17
		16	双	18	19	20	21	22	23	24
		17	单	25	26	27	28	29	30	31

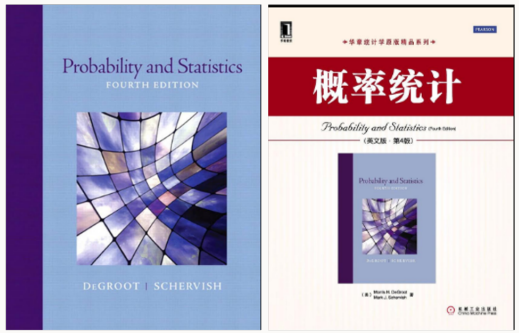
- 总计: 33 次课
- 第 17 周安排 1-2 次答疑
- 每 6-8 次课安排一次习题课
- 大概 27 次讲授式课堂

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

Course Information ——Lesson Plan

节次	内容	节次	内容
1	CH01. 随机事件及其运算	12	CH05. 二维联合分布函数
2	CH01. 频率与概率公理化	13	CH05. 二维离散型和连续型随机向量
3	CH01. 古典概型与几何概型	14	CH05. 随机变量的独立性
4	CH01. 组合计数	15	CH05. 条件分布
5	CH02. 条件概率	16	CH06. 多维随机向量的统计量 (期望、协方差)
6	CH02. 全概率公式和贝叶斯公式	17	CH06. 相关系数和条件期望
7	CH02. 随机事件的独立性	18	CH07. 集中不等式 (一)
8	CH03. 离散型随机变量	19	CH07. 集中不等式 (二)
9	CH04. 分布函数、概率密度函数	20	CH08. 大数定律
10	CH04. 统计量 (期望、方差)	21	CH08. 中心极限定理
11	CH04. 连续型随机变量的计算	22	CH09. 统计: 总体与样本
	Interactive Exercises	23	CH09. 抽样分布定理
	期中考试	24	CH10. 参数估计: 点估计
		25	CH10. 参数估计: 区间估计
		26	CH11. 假设检验
			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: 我会在每次课后 0-1 天内更新在课程主页上
   - 作业: 每周 1-2 次, 发布在 QQ 群中或者课程主页上, 下周 (周一或者周四) 上课前提交
   - 思考题: 可以私发到我的邮箱 zhangsq@nju.edu.cn
   - 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

1. 平时成绩 (40%)
   - 作业：40%
   - 思考题：+ 10%

2. 考试成绩 (60%)
   - 期中考试: 20%
   - 期末考试: 40%

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.

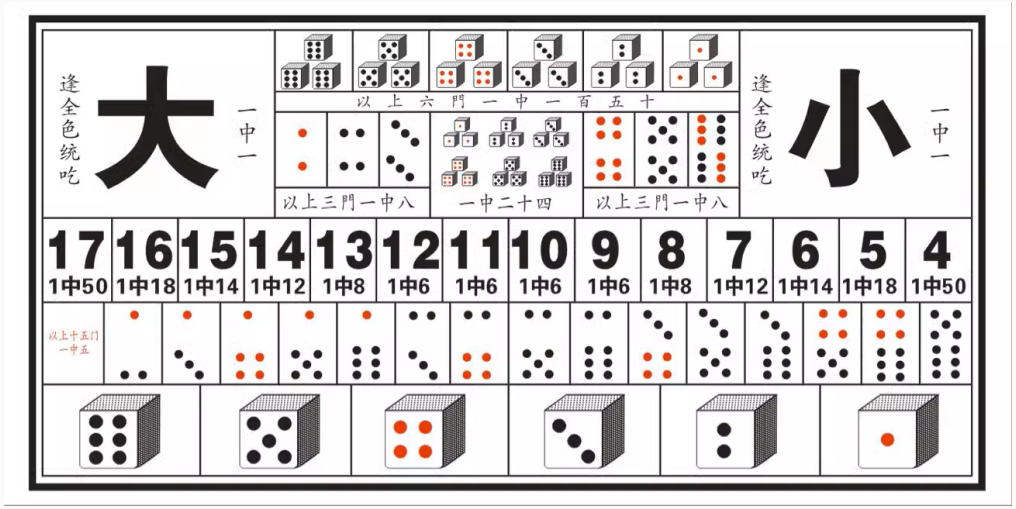
2. 一门知识密集型课程, 涉及大量的术语、公式、定理, 需要练习
   - 尝试双语教学, 可能在考试中使用双语考试 (e.g., 期中考试)

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

Course Information ——Exploration

1. **Thinking and Talking**
   - About Probability and Statistics
   - About the applications on AI and ML

2. **Interactive Exercises**



Ch00: 先导课程

Introduction to Probability and Statistics

September 4, 2023

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

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。

1. 假设当前比分为 6 : 3
2. 假设当前比分为 8 : 4

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 Moiver): 《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 Wahrscheinlichkeitsrechnun》, 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。

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

“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?
  - 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 0.5

例 0.5 (Poker Hands)

- 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 0.5

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

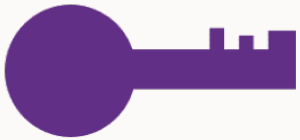
概率与统计的联系和区别

- **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:

- 数学分析
- 高等代数
- 计算机编程

后续课程:

- 机器学习、高级机器学习
- 统计学习
- 数据挖掘、试验设计 (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