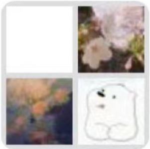
信息学中的概率统计

王若松

前沿计算研究中心
北京大学

课程介绍

- 课程名称：信息学中的概率统计
- 时间地点：1-16周 每周 周五3-4节，双周
- 考核方式：作业 30%，期中35%，期末 35%
- Office Hour：每周二2-3pm，静园五院206-
- 助教： 马允轩、王颖、张宇博、 …
- 习题课、助教office hour
- 邮箱：ruosongwang@pku.edu.cn



群聊: 2024Fall信息学中的概率统计

该二维码7天内(9月19日前)有效，重新进入将更新

课程大纲

概率部分

1. 概率论的基本概念
2. 随机变量及其分布
3. 多维随机变量及其分布
4. 尾不等式
5. 大数定律与中心极限定理

统计部分

1. 数理统计的基本概念
2. 参数估计
3. 假设检验
4. 方差分析和回归分析

定量分析（上下界）、在信息科学中的应用

参考书籍

1. 概率论与数理统计

作者：茆诗松

2. 概率论与数理统计

作者：何书元

3. 概率论与数理统计

作者：盛骤、谢式千、潘承毅

4. Introduction to Probability for Computing

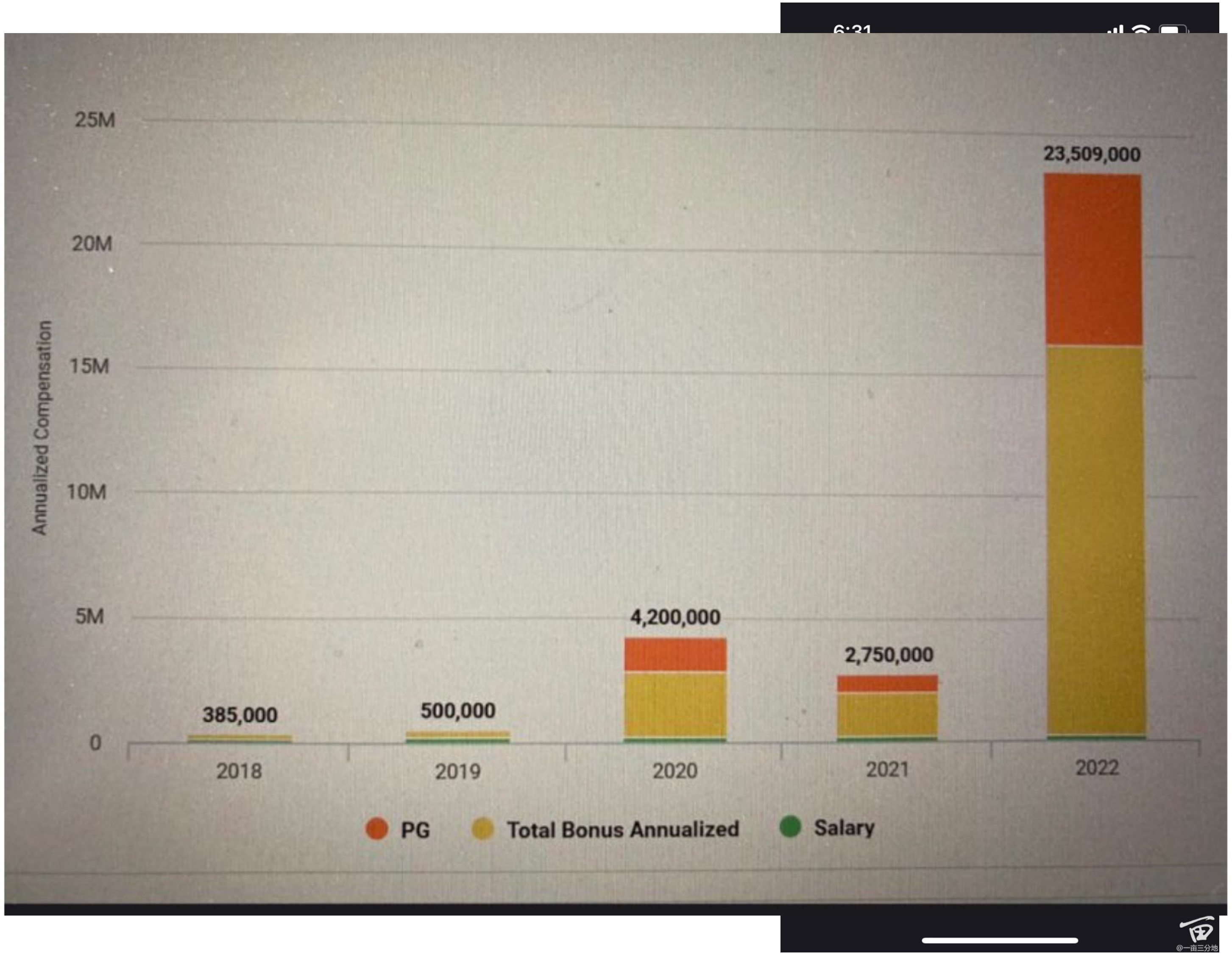
作者：Mor Harchol-Balter

为什么要学习概率统计？

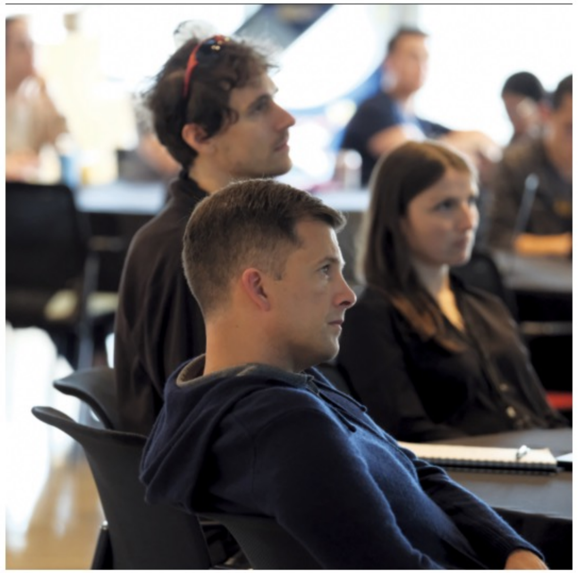
- 机器学习：**Stochastic** Gradient Descent, Diffusion **Probabilistic** Model
- 博弈论：**Mixed** Strategy
- 网络/系统：**Stochastic** Models in Queueing Theory
- 计算理论：**Probabilistic** Turing Machine
- 算法设计：**Randomized** Algorithm
- 图形学：**Stochastic** Progressive Photon Mapping
- …

Gary L. Miller@CMU
Miller-Rabin算法

为什么要学习概率统计？



Our Approach



Quantitative Research & Modeling covers a diverse set of disciplines, ranging from simple statistics, to complex theoretical mathematics, to cutting edge machine learning techniques.

概率论的基本概念

1. 随机事件和样本空间
2. 概率和频率、古典概率模型与几何概率模型
3. 概率的公理化
4. 条件概率
5. 事件的独立性

1. 随机事件和样本空间

- 随机现象：在一定的条件下, **并不总是出现相同结果**的现象
  - 掷硬币的结果、投骰子的结果
  - CPU的寿命、测量物理量的误差
  - 比赛的输赢、经济增长速度
- 确定性现象：只有一个结果
- 随机试验：在相同条件下可以重复的随机现象

1. 随机事件和样本空间

- **样本空间**：随机现象的一切可能基本结果组成的集合

- $S=\{e\}$, e为基本结果（**样本点**）
  - 掷硬币： $S=\{正面, 反面\}$
  - 勺园餐厅一天用餐的人数： $S=\{0,1,2,\ldots\}$
  - 对电子产品测试输入电压： $S=[0,220]V$
  - 测量误差： $S=\mathbb{R}$

- 思考：使用随机数生成器的计算机程序的样本空间

1. 随机事件和样本空间

- **随机事件（事件）**：样本空间的子集，样本点的集合

- 掷硬币：$S=\{正面,反面\}$，$A=\{正面\}$
- 勺园餐厅一天用餐的人数：$S=\{0,1,2,..,\}$，$A=\{用餐人数为奇数\}$
- 对电子产品测试输入电压：$S=[0,220]V$，$A=\{输入电压不超过110V\}$
- 测量误差：$S=\mathbb{R}$，$A=[-10,10]$

- 事件**发生**：事件中的某个样本点出现

1. 随机事件和样本空间

- **随机事件（事件）**：样本空间的子集，样本点的集合

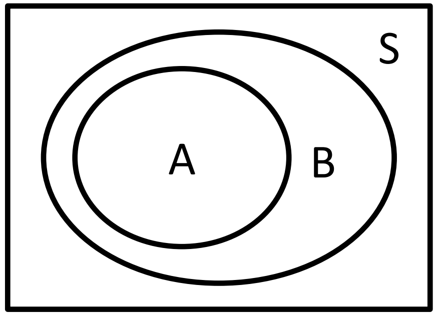
- **基本事件**：仅包含单个样本点的集合
- **必然事件**：样本空间本身
- **不可能事件**：空集

- 掷硬币
  - 必然事件：{正面, 反面}
  - 基本事件：{正面}，{反面}
  - 不可能事件：$\emptyset$

PEKING UNIVERSITY

1. 随机事件和样本空间

- 事件的关系：包含、相等、互不相容
- $A \subset B$ ：事件A发生时事件B也发生
- $A = B$ ：同时有$A \subset B$ 和 $B \subset A$
- 事件A与事件B互不相容：A与B没有相同的样本点

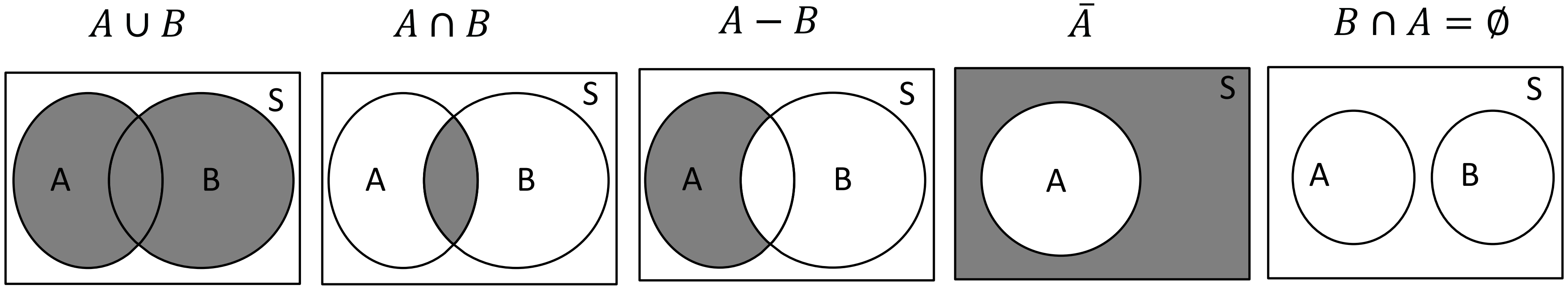


- 投骰子
  - $A = \{出现4点\},\ B = \{点数为偶数\},\ C = \{点数为奇数\}$
  - 则有$A \subset B, C$与B互不相容

- 对于任何事件A，有$\emptyset \subset A \subset S$

1. 随机事件和样本空间

- 事件的运算
- 事件的**并**：A和B至少一个发生，$A \cup B$
- 事件的**交**：A和B同时发生，$A \cap B$，也记作AB
- 事件的**差**：A发生B不发生，$A - B = A \cap \bar{B}$
- **对立（互补）**事件：$\bar{A} = S - A$



1. 随机事件和样本空间

- 交换律：$A \cup B = B \cup A, AB = BA$
- 结合律：$(A \cup B) \cup C = A \cup (B \cup C),\ (AB)C = A(BC)$
- 分配率：$(A \cup B) \cap C = AC \cup BC,\ (A \cap B) \cup C = (A \cup C)(B \cup C)$
- De Morgan公式：$\overline{A \cup B} = \bar{A} \cap \bar{B},\ \ \overline{A \cap B} = \bar{A} \cup \bar{B}$

- 掷两枚硬币，$A = \{$第一枚硬币正面朝上$\}, B = \{$第二枚硬币正面朝上$\}$
- $A \cup B, AB, \overline{A \cup B}, \overline{A \cap B}$ 分别的含义是什么？

2. 概率和频率

- **频率:** n次试验中事件发生的比例：$f_n(A) = \frac{n_A}{n}$
- n_A：事件A发生的次数, n：总试验次数
- 例：执行同一程序10次，成果求解8次，频率为80%

- 频率的性质
  - 非负性： $f_n(A) \in [0,1]$
  - 规范性： $f_n(S) = 1$，$f_n(\emptyset) = 0$
  - 有限可加性：对互不相容事件A_i, 有$f_n(\cup_i A_i) = \sum_i f_n(A_i)$
  - 单调性：对于事件$A \subset B$，有$f_n(A) \le f_n(B)$

- 频率$f_n(A)$随着n增大逐渐趋于一个稳定值，也即是事件A发生的概率

2. 概率和频率

- **概率**：事件频率的稳定值

- 概率的性质
  - 非负性：$P(A) \in [0,1]$
  - 规范性：$P(S) = 1$
  - 有限可加性：对互不相容事件A_i有$P(\bigcup_i A_i) = \sum_i P(A_i)$

2. 古典概率模型

- **有限性**：样本点个数有限
- **等可能性**：每个样本点概率相等
- 事件的概率$P(A) = \frac{|A|}{|S|}$

- 概率的性质
  - 非负性： $P(A) \in [0,1]$
  - 规范性： $P(S) = 1$
  - 有限可加性：对互不相容事件A_i有$P(\bigcup_i A_i) = \sum_i P(A_i)$

- 例：投骰子
  - $A = \{$点数为偶数$\}, P(A) = \frac{|A|}{|S|} = \frac{1}{2}$

2. 古典概率模型

- 随机图模型：有n个人，每一对人等概率认识或不认识。
- 求特定k个人两两均认识的概率？
- 求特定k个人两两均不认识的概率？

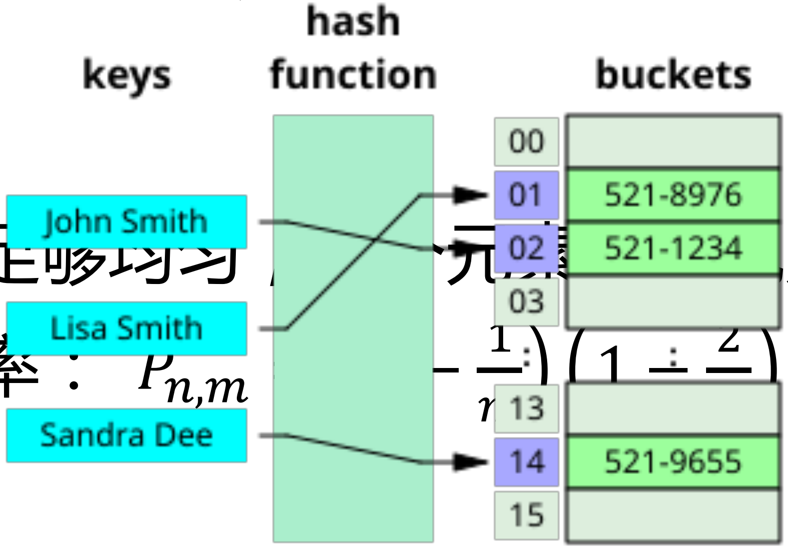
- 样本点数量：$2^{\frac{n(n-1)}{2}}$
- 特定k个人两两均认识的概率：$2^{-\frac{k(k-1)}{2}}$
- 特定k个人两两均不认识的概率：$2^{-\frac{k(k-1)}{2}}$

- 随机图应用：网络科学

2. 古典概率模型

- 球与桶模型：有n个球，每个球都等可能被放到m个桶中的任一个。求每个桶中至多有一个球的概率。

- 样本点数量： m^n
- 放法： $m(m-1)\cdots(m-n+1)$



- 哈希表：当哈希函数足够均匀 元素 为放进一个随机的桶中
- 哈希表没有冲突的概率： $P_{n,m} \cdots \left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right)\cdots\left(1-\frac{n-1}{m}\right)$

2. 古典概率模型

- $P_{n,m} = \left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right)\cdots\left(1-\frac{n-1}{m}\right)$
- $$\begin{aligned}\ln P_{n,m} &= \ln\left(1-\frac{1}{m}\right)+\ln\left(1-\frac{2}{m}\right)+\cdots+\ln\left(1-\frac{n-1}{m}\right)\\ &\le -\frac{1}{m}-\frac{2}{m}-\cdots-\frac{n-1}{m}\\ &= -\frac{(n-1)n}{2m}\end{aligned}$$

  $\ln(1+x)\le x$
- $P_{n,m} \le \exp\left(-\frac{(n-1)n}{2m}\right)$
- 作业： $P_{n,m} \ge \exp\left(-\frac{(n-1)n}{2m}\right)\left(1-O\left(\frac{n^3}{m^2}\right)\right)$

回顾：渐进记号

- $f(n) = O(g(n))$ 如果存在常数 M，使得$f(n) \leq Mg(n)$对充分大的n均成立
- $f(n) = \Omega(g(n))$ 如果存在常数 m，使得$f(n) \geq mg(n)$对充分大的n均成立
- $f(n) = \Theta(g(n))$ 如果 $f(n) = O(g(n))$ 且 $f(n) = \Omega(g(n))$

PEKING UNIVERSITY

2. 古典概率模型

- $P_{n,m} \le \exp\left(-\frac{(n-1)n}{2m}\right)$
- $P_{n,m} \ge \exp\left(-\frac{(n-1)n}{2m}\right)\left(1 - O\left(\frac{n^3}{m^2}\right)\right)$
  - 当$m \le M_1 n^2$时，$P_{n,m} \to 0$
  - 当$m \ge M_2 n^2$时，$P_{n,m} \to 1$

- 无冲突：$m = \Theta(n^2)$
- 生日悖论：23人中两人同生日的概率超过50%

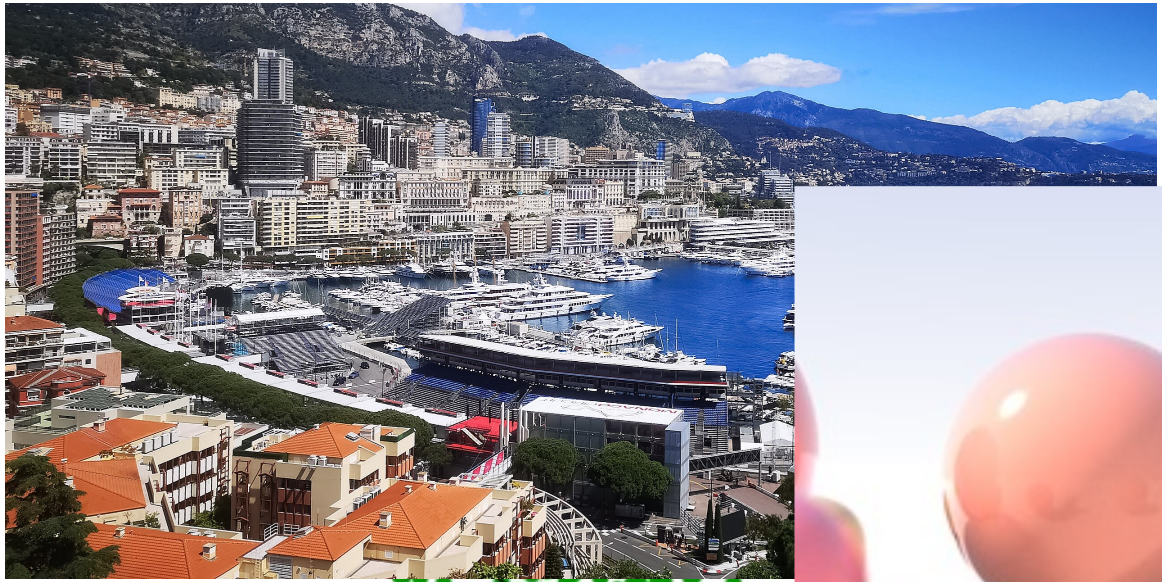
2. 几何概率模型

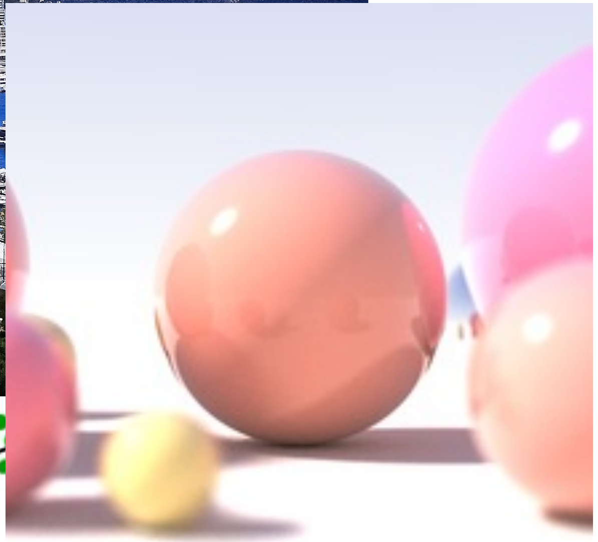
- 样本空间空间 S充满某个空间，度量用$m(S)$表示
- 事件 A 对应样本空间中的一个区域，概率为$P(A) = m(A)/m(S)$

- 概率的性质
  - 非负性： $P(A) \in [0, 1]$
  - 规范性： $P(S) = 1$
  - 有限可加性：对互不相容事件A_i有$P(\bigcup_i A_i) = \sum_i P(A_i)$

2. 几何概率模型

- 在 $[-1,1]\times[-1,1]$中等概率取一个点，落在单位圆中的概率？
- 样本空间$S=[-1,1]\times[-1,1]$
- $m(A)=\pi, m(S)=4$
- $P(A)=\pi/4$
- Monte Carlo方法：利用随机数进行数
- 应用：图形学渲染





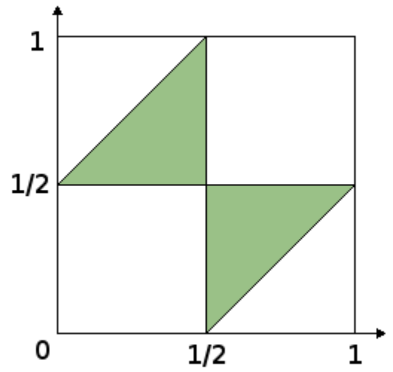
2. 几何概率模型

- 在一根长度为 1的木棍上随机取两个点将其分为三段, 求它们可以构成一个三角形的概率

- 设两个切点到端点距离分别为a和b，则三段的长度可写为

- $x=a, y=b, z=1-a-b$ 当 $a+b\leq 1$
- $x=1-a, y=1-b, z=a+b-1$ 当 $a+b\geq 1$

- 可以构成一个三角形等价于$x<\frac{1}{2}, y<\frac{1}{2}, z<\frac{1}{2}$



3. 概率的公理化

- 将样本空间的哪些子集看作事件？
  - 对集合操作封闭
  - 全部子集？

- **事件域**：对于样本空间S，事件的集合$F = \{A: A \subset S\}$满足
  - $S \in F$
  - 若$A \in F$，则$\bar{A} \in F$
  - 若$A_i \in F$，则$\cup_{i=1}^{\infty} A_i \in F$
- F为一个事件域，又称为$\boldsymbol{\sigma}$**-代数**

- 对集合操作封闭的集合的集合

3. 概率的公理化

- **概率**：设S为样本空间，F为S的某些子集组成的事件域。如果定义在F上的实值函数P满足
  - 对于任意$A \in F, P(A) \geq 0$
  - $P(S) = 1$
  - 对互不相容事件$A_1, A_2, A_3, \ldots$，有$P(\cup_{i=1}^{+\infty} A_i) = \sum_{i=1}^{+\infty} P(A_i)$
- 则称$P(A)$为事件A的概率，(S, F, P)为概率空间

3. 概率的公理化

- 有限可加性： $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$
- 补集：$P(\bar{A}) = 1 - P(A)$
- 空集是事件，且$P(\emptyset) = 0$
- 单调性：如果$A \subset B$，$P(A) \leq P(B)$
- 可减性：$P(A - B) = P(A) - P(AB)$
- 加法公式： $P(A \cup B) = P(A) + P(B) - P(AB)$
- 一般加法公式： $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i A_j A_k) + \cdots + (-1)^{n-1} P(A_1 A_2 \ldots A_n)$

3. 概率的公理化

- 有n个有编号的球，将他们随机打乱，使得每种排列均等概率出现。求随机打乱后，至少有一个球位置没有改变的概率。

- 设事件A_i为编号为i的球位置没有改变，则有
  - $P(A_i)=\frac{1}{n}$
  - 当$i\neq j, P(A_iA_j)=\frac{1}{n(n-1)}$
  - 当i,j,k均不相同时，$P(A_iA_jA_k)=\frac{1}{n(n-1)(n-2)}$
  - …

- $$P(A_1\cup A_2\cup\cdots\cup A_n)=1-\binom{n}{2}\cdot\frac{1}{n(n-1)}+\binom{n}{3}\cdot\frac{1}{n(n-1)(n-2)}-\cdots$$
$$=1-\frac{1}{2!}+\frac{1}{3!}-\frac{1}{4!}+\cdots$$

3. 概率的公理化

- 半可加性(Union Bound)： $P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$
  - 对 A_i没有任何要求

- 球与桶模型：有n个球，每个球都等可能被放到m个桶中的任一个。求每个桶中至多有一个球的概率。
- 使用Union Bound估计？
  - 事件$A_{i,j}$表示第i个球和第j个球被放入同一个桶中
  - $P\left(\cup_{i\neq j} A_{ij}\right) \leq \binom{n}{2} \cdot \frac{1}{m}$
  - $P_{n,m} = 1 - P\left(\cup_{i\neq j} A_{ij}\right) \geq 1 - \binom{n}{2} \cdot \frac{1}{m} = 1 - \frac{(n-1)n}{2m}$
  - 对比：$P_{n,m} \geq \exp\left(-\frac{(n-1)n}{2m}\right)\left(1 - O\left(\frac{n^3}{m^2}\right)\right)$

3. 概率的公理化

- 有限可加性： $P(\bigcup_{i=1}^{n} A_i) = \sum_i P(A_i)$
- 可列可加性：$P(\bigcup_{i=1}^{\infty} A_i) = \sum_i P(A_i)$
- 是否等价？

- $S = [0, \infty)$，$P(A) = \lim_{k \to \infty} \frac{1}{k} \lambda(A \cap (0, k))$，$A_i = [i-1, i)$
- $\lambda(A)$可以简单理解为A的长度

北京大学 PEKING UNIVERSITY

3. 概率的公理化

- $P(\emptyset)=0$的逆否命题：如果事件A满足 $P(A)>0$，则$A\neq\emptyset$
- 可用于证明特定数学对象的存在性（概率证法）

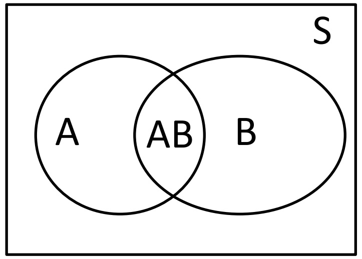
- 证明：对于任意 $k\geq 3$，存在一个$n=2^{\frac{k}{2}-1}$个人组成的关系网络，使得对于任意k个人，既非两两均认识，也非两两均不认识。

- 考虑n个人组成的随机图。事件A为存在 k个人两两均认识，事件B为存在 k个人两两均不认识。
  - $P(A)\leq\binom{n}{k}\cdot 2^{-k(k-1)/2}, P(B)\leq\binom{n}{k}\cdot 2^{-k(k-1)/2}$
  - $P(A\cup B)\leq 2\binom{n}{k}\cdot 2^{-\frac{k(k-1)}{2}}\leq 2\cdot 2^{\frac{k(k-2)}{2}}\cdot 2^{-\frac{k(k-1)}{2}}<1$
  - $P(\overline{A}\cap\overline{B})>0$

$\binom{n}{k}\leq n^k=2^{\frac{k(k-2)}{2}}$

4. 条件概率

- 对于概率空间(S,F,P)，事件B满足 $P(B)>0$，则事件B发生的条件下事件A发生的概率为 $P(A|B)=\frac{P(AB)}{P(B)}$



- 条件概率满足概率的三条性质
  - 对于任意$A\in F, P(A|B)\geq 0$
  - $P(S|B)=1$
  - 对互不相容事件$A_1,A_2,A_3,\ldots$，有$P(\cup_{i=1}^{+\infty}A_i|B)=\sum_{i=1}^{+\infty}P(A_i|B)$

- $P(\cup_{i=1}^{+\infty}A_i|B)=\frac{P((\cup_{i=1}^{+\infty}A_i)B)}{P(B)}=\frac{P(\cup_{i=1}^{+\infty}(A_iB))}{P(B)}=\frac{\sum_{i=1}^{+\infty}P(A_iB)}{P(B)}=\sum_{i=1}^{+\infty}P(A_i|B)$

4. 条件概率

- 例1 ：一个家庭中有两个小孩，已知至少一个是女孩，问两个都是女孩的概率是多少？

- 例2：一个小球被等概率放在三个盒子中的一个。已知2号盒子没有小球，1号盒子有小球的概率是多少？

4. 条件概率

- **乘法公式**
- $P(AB) = P(A|B)P(B) = P(B|A)P(A)$
- $P(A_1A_2 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \dots P(A_n|A_1A_2 \dots A_{n-1})$

- 例：一批零件共有100 个, 其中有10 个不合格品。从中一个一个取出, 求第三次才取得不合格品的概率是多少?

4. 条件概率

- **全概率公式**
- 事件$B_1, B_2, \dots B_n$ 为S的划分，即$B_1, B_2, \dots B_n$互不相容，且$\cup_{i=1}^{n} B_i = S$
- 若$P(B_i) > 0$，则$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$

- 当$n = 2$，$P(A) = P(B) \cdot P(A|B) + P(\overline{B}) \cdot P(A|\overline{B})$

- 例1 ：设在n张彩票中有一张奖券.　求第二人摸到奖券的概率是多少?
- 例2：假设患肺癌的概率为0. 1%，已知20%的人吸烟，且他们患肺癌的概率为0. 4%，不吸烟患肺癌的概率为多少？

4. 条件概率

- **贝叶斯公式**
- $P(A|B) = P(B|A) \cdot \frac{P(A)}{P(B)}$

- 例：某癌症发病率为百万分之一，某针对该癌症的检测试剂的准确率是99.9%。若检测结果为阳性，有多大概率患病？

- A表示患病， B表示检测结果为阳性
- $P(A) = 10^{-6}, P(B) = 10^{-6} \cdot 0.999 + (1 - 10^{-6}) \cdot 0.001$
- $P(A|B) = 0.999 \cdot \frac{P(A)}{P(B)} \approx \frac{1}{1001}$

4. 条件概率

- **条件概率的全概率公式**
- 事件$B_1, B_2, \dots B_n$ 为S的划分
- 若$P(B_i) > 0$且$P(C) > 0$ ，则$P(A|C) = \sum_{i=1}^{n} P(B_i|C)P(A|(B_iC))$
- $P(B_1|C)P(A|(B_1C)) + P(B_2|C)P(A|(B_2C)) = \frac{P(AB_1C)}{P(C)} + \frac{P(AB_2C)}{P(C)} = \frac{P(AC)}{P(C)} = P(A|C)$

- 例：保险公司认为世界上有两种人。第一种人每年有40%的概率出事故，第二种人每年有20%的概率出事故。人群中有30%的人是第一种人。若某个客户第一年出了事故，第二年出事故的概率？
- A表示为第一种人，A_1表示第一年出事故，A_2表示第二年出事故
- $P(A_2 \mid A_1) = P(A_2 \mid (AA_1)) \cdot P(A \mid A_1) + P\left(A_2 \mid (\bar{A}A_1)\right) \cdot P(\bar{A} \mid A_1)$
- $P(A \mid A_1) = P(A_1 \mid A) \cdot \frac{P(A)}{P(A_1)}$

5. 事件的独立性

- 对于事件A和B，若$P(AB) = P(A)P(B)$满足，则事件A和B**相互独立**
- 若$P(B) > 0$，等价于$P(A|B) = P(A)$

- 例1：若A与B互不相容且有$P(A) > 0$和$P(B) > 0$，是否有A与B相互独立？

- 例2：在随机图模型中，事件E_{ijk}表示 i, j, k三人两两认识。何时有E_{abc}与 E_{def}独立？
  - $P(E_{abc}) = P(E_{def}) = 1/8$
  - 当$|\{a,b,c,d,e,f\}| \geq 5, P(E_{abc} \cap E_{def}) = 1/64$
  - 当$|\{a,b,c,d,e,f\}| = 4, P(E_{abc} \cap E_{def}) = 1/32$
  - 当$|\{a,b,c,d,e,f\}| = 3, P(E_{abc} \cap E_{def}) = 1/8$

5. 事件的独立性

- 对于事件A和B，若$P(AB) = P(A)P(B)$满足，则事件A和B**相互独立**
- 若$P(B) > 0$，等价于$P(A|B) = P(A)$

- 例3： 在球与桶模型中，事件E_i表示第i个桶中没有球。何时有E_i与E_j独立？
  - $P(E_i) = P(E_j) = \left(1 - \frac{1}{n}\right)^m$
  - $P(E_i \cap E_j) = \left(1 - \frac{2}{n}\right)^m$

5. 事件的独立性

- 对于三个事件 A, B, C
  - 如果 $P(ABC) = P(A)P(B)P(C)$，是否有三个事件中任意两个事件均独立？
  - 如果三个事件中任意两个事件均独立，是否有$P(ABC) = P(A)P(B)P(C)$？

- 如何定义多个事件的独立性？

PEKING UNIVERSITY

5. 事件的独立性

- 对于三个事件A, B, C，若
  - $P(AB) = P(A)P(B)$
  - $P(AC) = P(A)P(C)$
  - $P(BC) = P(B)P(C)$
- 则称 A, B, C**两两独立**
- 若 A, B, C两两独立，且$P(ABC) = P(A)P(B)P(C)$，则称 A, B, C**相互独立**

- 对于n个事件 $A_1, A_2, \ldots, A_n$，如果对于$\{1, 2, \ldots, n\}$的**任意子集** I，都有
  - $P(\bigcap_{i\in S} A_i) = \prod_{i\in S} P(A_i)$
- 则称 $A_1, A_2, \ldots, A_n$**相互独立**

5. 事件的独立性

- 例1：若A, B, C相互独立，是否有$A \cup B$与C独立？
  - $P\big((A \cup B)C\big) = P(AC \cup BC) = P(AC) + P(BC) - P(ABC)$
    $= \big(P(A) + P(B) - P(AB)\big)P(C) = P(A \cup B)P(C)$

- 例2：两人轮流射击。甲的命中概率为p_1，乙的命中概率为p_2。先命中者获胜。求甲乙各自获胜的概率？
  - $p_1 + (1 - p_1)(1 - p_2)p_1 + (1 - p_1)^2(1 - p_2)^2 p_1 + \cdots = \frac{p_1}{1 - (1 - p_1)(1 - p_2)}$

5. 事件的独立性

- 做n次重复试验，每次试验结果均独立，称为**n重独立重复试验**
- 如果某个随机试验只有两个可能的结果A和$\bar{A}$，将试验独立地重复进行n次，称为**n重伯努利试验**

- 例：若$P(A) = 1/n$，求
  - 第i次试验A发生的概率
  - 在n次试验中A均不发生的概率
  - 在n次试验中A发生过至少一次的概率