案例分析: 概率计算

(习题课)

回答思考题、补充例题、复盘作业

September 28, 2023

三因徒问题: 例 0.43

例 0.43 (三因徒问题) 犯人 a, b, c 均被判为死刑, 法官随机赦免其中一人, 看守知道谁被赦免但不会说. 犯人 a 问看守: b 和 c 谁会被执行死刑? 看守的策略:

1. 若赦免 b, 则说 c
2. 若赦免 c, 则说 b
3. 若赦免 a, 则以 $1/2$ 的概率说 b 或 c

看守回答犯人 a: 犯人 b 会被执行死刑. 犯人 a 兴奋不已, 因为自己生存的概率为 $1/2$. 犯人 a 将此事告诉犯人 c. c 同样高兴, 因为他觉得自己的生存几率为 $2/3$.

那么谁才是正确的呢?

解答: 例 0.43

问题: 三犯人 a, b, c 均被判为死刑, 法官随机赦免其中一人, 看守知道谁被赦免但不会说. 犯人 a 问看守: b 和 c 谁会被执行死刑? 看守的回答策略为: i) 若赦免 b , 则说 c; ii) 若赦免 c , 则说 b; iii) 若赦免 a , 则以 1/2 的概率说 b 或 c ; 看守回答 a : 犯人 b 会被执行死刑. 犯人 a 兴奋不已, 认为自己生存的概率为 1/2. 犯人 a 将此事告诉犯人 c, c 同样高兴, 因为他觉得自己的生存几率为 2/3, 犯人 a 和犯人 c 中谁的想法是正确的?
解答:

- 事件“看守说犯人 b 会被执行死刑”的“原因”有两种情况, 即事件“赦免 c , 看守说 b 会被执行死刑”或者事件“赦免 a , 看守说以 1/2 的概率说 b 或 c 会被执行死刑”.
- 用事件 A, B, C 分别表示 a, b, c 被赦免, 则 $P(A) = P(B) = P(C) = 1/3$. 用事件 D 表示看守说犯人 b 被执行死刑, 则

$$P(D|A) = 1/2 \qquad P(D|B) = 0 \qquad P(D|C) = 1$$

由全概率公式有

$$P(D) = P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C) = 1/2$$

由贝叶斯公式有

$$P(A|D) = P(A)P(D|A)/P(D) = 1/3 \qquad \text{和} \qquad P(C|D) = P(C)P(D|C)/P(D) = 2/3$$

所以犯人 a 的想法是错误的, 犯人 c 的想法是正确的.

解答: 例 0.43

- 用事件 A, B, C 分别表示犯人 a, b, c 被赦免, 因为法官随机赦免, 所以

$$P(A) = P(B) = P(C) = 1/3\ .$$

- 用事件 D 表示看守人说犯人 b 被执行死刑, 根据看守的策略, 有

$$P(D \mid A) = 1/2\ , \quad P(D \mid B) = 0\ , \quad P(D \mid C) = 1\ .$$

- 我们要求解的是 "哪个犯人才是导致 D 事件发生的最大原因", 即求 max,

$$P(A \mid D) = \frac{P(A)P(D \mid A)}{P(D)}\ , \quad P(C \mid D) = \frac{P(C)P(D \mid C)}{P(C)}$$

- 根据上式我们还需要知道 $P(D)$, 可以根据全概率公式求得

$$P(D) = P(A)P(D \mid A) + P(B)P(D \mid B) + P(C)P(D \mid C) = 1/2\ .$$

- 最后得到, $P(A \mid D) = 1/3$ and $P(C \mid D) = 2/3$.

抛投不均匀硬币: 例 0.44

例 0.44 设一个箱子中有 $k+1$ 枚不均匀的硬币, 投掷第 i 枚硬币时正面向上的概率为 $i/k\,(i=0,1,2,\ldots,k)$. 现从箱子中任意取出一枚硬币, 并任意重复投掷多次, 若前 n 次正面向上, 求第 $n+1$ 次正面向上的概率.

解答: 例 0.44

问题: 设一个箱子中有 $k+1$ 枚不均匀的硬币, 投掷第 i 枚硬币时正面向上的概率为 i/k $(i=0,1,2,\ldots,k)$. 现从箱子中任意取出一枚硬币, 并任意重复投掷多次, 若前 n 次正面向上, 求第 $n+1$ 次正面向上的概率.

解答:

- 本题中, 事件 "前 n 次正面向上" 和 "第 $n+1$ 次正面向上" 在 "从箱子中任意取出一枚硬币反复抛掷多次" 发生的情况下是条件独立的, 即同一枚硬币抛掷的结果是独立事件.
- 用 A 表示第 $n+1$ 次投掷正面向上的事件, 用 B 表示前 n 次投掷正面向上的事件, 用 C_i 表示从箱子中取出第 i 枚硬币的事件 $(i=0,1,2,\ldots,k)$.
- 我们要求解的是 $P(A \mid B)$ [方法一].
  - 因为 $P(A \mid B) = P(AB)/P(B)$ 或者 $= \sum_i P(A \mid BC_i)P(C_i)$, 这里 $P(C_i) = \frac{1}{k+1}$
  - 其中, $P(A \mid BC_i) = P(AB \mid C_i)/P(B \mid C_i)$
  - 再其中, $P(B \mid C_i) = (i/k)^n$, $P(AB \mid C_i) = (i/k)^{n+1}$

- 所以

$$P(A \mid B) = \sum_i P(A \mid BC_i) P(C_i) = \sum_i \frac{(i/k)^{n+1}}{(i/k)^n} \frac{1}{k+1}$$

- 我们要求解的是 $P(A \mid B)$ [方法二].
  - 因为 $P(A \mid B) = P(AB)/P(B)$, 且

$$P(AB) = \sum_{i=0}^{k} P(C_i) P(AB|C_i) = \sum_{i=0}^{k} P(C_i) P(A|C_i) P(B|C_i) = \frac{1}{k+1} \sum_{i=0}^{k} \frac{i^{n+1}}{k^{n+1}}$$

以及

$$P(B) = \sum_{i=0}^{k} P(C_i) P(B|C_i) = \frac{1}{k+1} \sum_{i=0}^{k} \frac{i^n}{k^n}$$

由此可知

$$P(A|B) = \frac{\sum_{i=0}^{k} (i/k)^{n+1}}{\sum_{i=0}^{k} (i/k)^n}$$

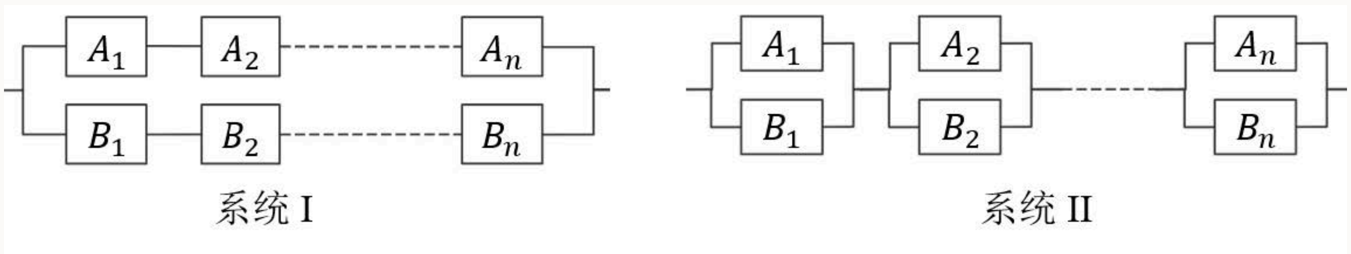
- 另外, 当 k 非常大或 $k \to \infty$ 时可利用积分近似

$$\frac{1}{k} \sum_{i=1}^{k} (i/k)^n \approx \int_0^1 x^n dx = \frac{1}{n+1} \qquad \text{和} \qquad \frac{1}{k} \sum_{i=1}^{k} (i/k)^{n+1} \approx \int_0^1 x^{n+1} dx = \frac{1}{n+2}$$

此时有 $P(A|B) \approx (n+1)/(n+2)$.

电路可靠性: 例 0.45

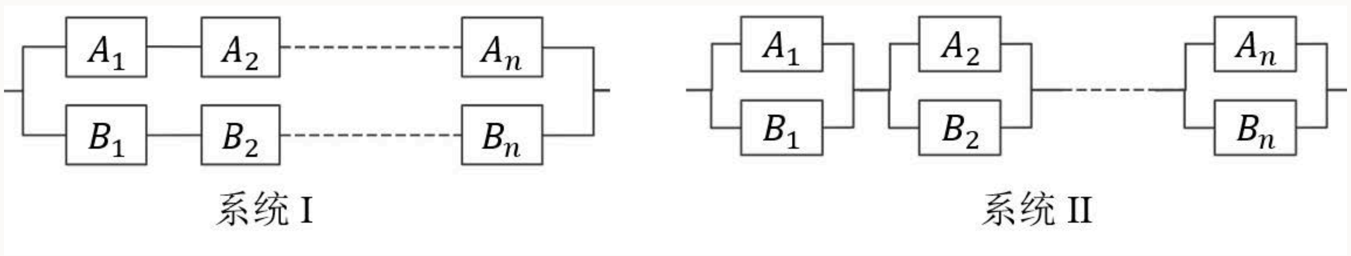
例 0.45 设构成系统的每个元件的可靠性均为 $p\,(0 < p < 1)$, 且各元件是否正常工作是相互独立的. 设有 $2n$ 个元件按下图所示, 两种不同连接方式构成两个不同的系统, 比较这两种系统的可靠性大小.



解答: 例 0.45

问题: 设构成系统的每个元件的可靠性均为 $p\ (0<p<1)$, 且各元件是否正常工作是相互独立的. 设有 $2n$ 个元件按下图所示, 两种不同连接方式构成两个不同的系统, 比较这两种系统的可靠性大小.

解答:



- 对于系统 I, 它能正常工作当且仅当系统中的两条通路至少有一条正常工作, 而每条通路正常工作当且仅当它的每个元件都能正常工作; 对于系统 II, 它能正常工作当且仅当每对并联元件能够正常工作.
- 用事件 A_i 和 B_i 表示图中对应元件正常工作 $(i=0,1,2,\ldots,n)$, 因此系统 I 的可靠

性为

$$
\begin{aligned}
&P((A_1A_2\dots A_n)\cup(B_1B_2\dots B_n))\\
&=P(A_1A_2\dots A_n)+P(B_1B_2\dots B_n)-P(A_1A_2\dots A_nB_1B_2\dots B_n)\\
&=2p^n-p^{2n}=p^n(2-p^n)
\end{aligned}
$$

系统 II 可靠性为

$$
P\left(\bigcap_{i=1}^{n}(A_i\cup B_i)\right)=\prod_{i=1}^{n}P(A_i\cup B_i)=(2p-p^2)^n=p^n(2-p)^n
$$

利用数学归纳法可证明当 $n\geq 2$ 时有 $(2-p)^n>2-p^n$ 成立, 由此可知系统 II 的可靠性更好.

多项式相等: 0.46

例 0.46 给定两个较复杂的多项式

$$F(x) = (x+2)^7(x+3)^5 + (x+1)^{100} + (x+2)(x+3) + x^{20}$$

$$G(x) = (x+3)^{100} - (x+1)^{25}(x+2)^{30} + x^{20} + (x-2)(x-3)\dots(x-100)$$

如何快速验证 $F(x) \equiv G(x)$?

解答: 例 0.46

问题: 如何快速验证 $F(x) \equiv G(x)$, 给定两个较复杂的多项式

$$F(x) = (x+2)^7(x+3)^5 + (x+1)^{100} + (x+2)(x+3) + x^{20}$$
$$G(x) = (x+3)^{100} - (x+1)^{25}(x+2)^{30} + x^{20} + (x-2)(x-3)\ldots(x-100)$$

解答:

- 若通过展开多项式合并同类项, 比较每项系数是否相同的方法来验证 $F(x) \equiv G(x)$, 则需要较高的计算时间开销.
- 设计一种利用独立随机性方法来验证 $F(x) \equiv G(x)$ 是否正确, 使得该方法验证结果为正确的概率较高, 同时降低计算时间.
  - 假设 $F(x)$ 或 $G(x)$ 的最高次项不超过 d, 考虑从集合 $[100d] = \{1, 2, \ldots, 100d\}$ 中等可能独立地随机选取 $k(< d)$ 个数 $r_1, r_2, \ldots, r_k$. 若存在 r_i 使得 $F(r_i) \neq G(r_i)$ 成立, 则返回 $F(x) \neq G(x)$, 否则返回 $F(x) \equiv G(x)$.
  - 为什么要用 "$[100d]$"?
  - 分析该方法的正确性:

1. 若多项式 $F(x) \equiv G(x)$, 则该方法得到 "正确" 的结果, 因为对于任意 $r_i \in [100d]$ 都有 $F(x) = G(x)$;
2. 若多项式 $F(x) \neq G(x)$ 且 $F(r_i) \neq G(r_i)$, 则该方法得到 "正确" 的结果, 因为存在一个 r_i 使得 $F(r_i) \neq G(r_i)$;
3. 若多项式 $F(x) \neq G(x)$ 但 $F(r_i) = G(r_i)$, 即存在 $r_i \in [100d]$ 使得 $F(r_i) = G(r_i)$ 成立, 此时 r_i 为多项式 $F(x) - G(x) = 0$ 的一个实数根. 根据代数知识可知最高次项不超过 d 的多项式 $F(x) - G(x) = 0$ 至多有 d 个实数根, 而 r_i 为 $[100d]$ 中等可能随机选取, 因此有

$$P(F(r_i) = G(r_i)) \leq \frac{d}{100d} = \frac{1}{100} .$$

进而有,

$$P\left(\bigcap_{i=1}^{k}\{F(r_i) = G(r_i)\}\right) = \prod_{i=1}^{k} P(F(r_i) = G(r_i)) \leq \frac{1}{100^k} .$$

矩阵乘法相等: 例 0.47

例 0.47 给定矩阵 $A, B, C \in \{0,1\}^{n\times n}(n \geq 10000000)$，如何快速验证 $AB = C$。

解答: 例 0.47

问题: 给定矩阵 $A, B, C \in \{0,1\}^{n\times n}(n \geq 10000000)$, 如何快速验证 $AB = C$.

解答:

- 若直接采用矩阵乘法计算 AB , 再与矩阵 C 进行比较, 计算复杂度开销为 $O(n^3)$
- 类似于验证多项式 $F(x) \equiv G(x)$ 的方法, 随机选取一个向量 $\bar{r} = (r_1, r_2, \ldots, r_n)^T$, 其中元素 $r_1, r_2, \ldots, r_n$ 都是从 $\{0,1\}$ 独立等可能随机选取所得, 通过验证事件

$$\text{存在一个向量 } \bar{r} \text{ 使得 } A(B\bar{r}) \neq C\bar{r}$$

  发生的概率极小来反向验证 $AB = C$.

- 这里, r 可以理解为 “函数” AB 和 C 的自变量.

解答: 例 0.47

- Freivalds (弗赖瓦尔兹) 算法

输入: 矩阵 $\mathbf{A}, \mathbf{B}, \mathbf{C}$
输出: 是/否　　　　　　　　%% 验证 $\mathbf{AB} \stackrel{?}{=} \mathbf{C}$

- -

For $i = 1 : k$
　　随机选择向量 $\bar{\mathbf{r}}_i = (r_{i1}, r_{i2}, \cdots, r_{in})$, 其每个元素是从 $\{0, 1\}$ 独立等可能随机采样所得
　　计算向量 $\bar{\mathbf{p}}_i = \mathbf{A}(\mathbf{B})\bar{\mathbf{r}}_i - \mathbf{C}\bar{\mathbf{r}}_i$
　　If $\{\bar{\mathbf{p}}_i$ 不是零向量$\}$ then
　　　　返回"否"
　　EndIf
EndFor
返回"是".

- 计算事件"存在一个向量 $\bar{r}$ 使得 $A(B\bar{r}) \neq C\bar{r}$"发生的概率. 设随机变量 $\bar{r_1}, \bar{r_2}, \ldots, \bar{r_k} \in \{0, 1\}^n$ 中每个元素都是从 $\{0, 1\}$ 独立等可能随机选取所得, 若 $AB \neq C$, 则有

$$P\left[\bigcap_{i=1}^{k}\{A(B\bar{r_i}) = C\bar{r_i}\}\right] \leq \frac{1}{2^k}$$

小结与发散

通过例 0.46 和例 0.47, 我们可以发现在证明一些数学的等式的时候可以通过 "设计随机试验 + 概率计算" 的方式来证明.

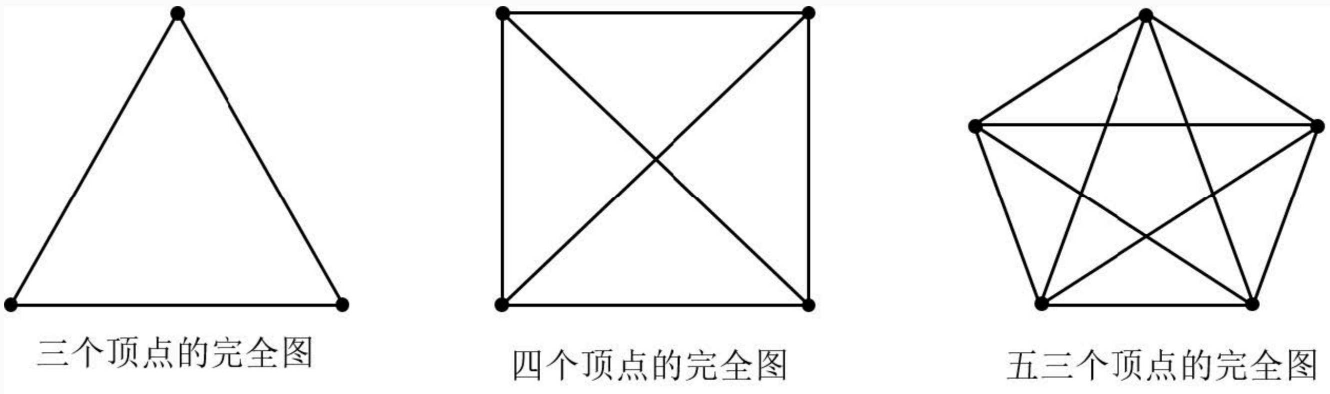
进一步, 我们可以思考如下问题要如何解决呢?

- 证明不等式 $2ab \le a^2 + b^2$
- 比较 n^k 和 $(k+1)^{n+1}$ 的大小?
- F 和 G 是两个不规则的图形, 比较这两者的面积大小?

这套方法有何利弊?

完全图着色 (边着色): 例 0.48

例 0.48 设平面上有 n 个顶点, 其中任意三个顶点不在同一条直线上, 用 $n(n-1)/2$ 条边将这些顶点连接起来的图称为 n 个顶点的完全图. 例如三个、四个、五个顶点的完全图如下所示.



现将图中的每条边都分别染成红色或蓝色, 给定两正整数 $n \geq 10$ 和 $k > n/2$, 是否存在一种染色方法, 使得图上任意 k 个顶点相对应的 $k(k-1)/2$ 条边不是同一颜色?

解答: 例 0.48

问题: 如上所述.

解答:

- 若通过穷举的方法, 则计算的开销较大
- 可以利用概率的方法证明至少存在一种染色方法使得任意 k 个顶点相对应的 $k(k-1)/2$ 条边不是同一颜色
  - 假设每条边都等可能独立地被染成红色或蓝色, 即每条边为红色或蓝色的概率均为 $1/2$. 从 n 个不同的顶点中选出 k 个顶点有 $\binom{n}{k}$ 种不同的选法, 分别对应于 $\binom{n}{k}$ 个包含 k 个顶点的子集, 这里将的子集分别标号为 $1, 2, \ldots, \binom{n}{k}$.
  - 用 E_i 表示第 i 个子集中 $k(k-1)/2$ 条边染成相同颜色的事件, 根据题意可得

$$P(E_i) = 2(1/2)^{k(k-1)/2} \qquad i = 1, 2, \ldots, \binom{n}{k}$$

  - 若存在 k 个顶点, 其对应的 $k(k-1)/2$ 条边是同一种颜色的事件可表示为 $\bigcup_{i=1}^{\binom{n}{k}} E_i$.

根据布尔不等式有

$$P\left(\bigcup_{i=1}^{\binom{n}{k}} E_i\right) \leq \sum_{i=1}^{\binom{n}{k}} P(E_i) = \binom{n}{k}(1/2)^{k(k-1)/2-1}$$

注意当 $n \geq 10$ 和 $k > n/2$ 时, 上式右侧是单调递减函数, 最大值在 $k = n/2 + 1$ 处取得, 有 $P\left(\bigcup_{i=1}^{\binom{n}{k}} E_i\right) < 1$. 因此事件 “完全图中任意 k 个顶点相对应的 $k(k-1)/2$ 条边不是同一颜色” 的概率大于零. 这意味着至少存在一种染色方法使得任意 k 个顶点相对应的 $k(k-1)/2$ 条边不是同一颜色.

隐私问题的调查 0.49

例 0.49 每个人都有一些隐私或秘密, 相关信息不希望被外人知晓. 对于具有社会普遍性的隐私问题, 需要对相关问题进行一些必要的调查. 需要设计一种调查方案, 使被调查者既愿意作出真实回答、又较好地保护个人隐私. 经过多年研究与实践, 心理学家和统计学家设计了一种巧妙的方案. 核心是如下两个问题:

[问题 A]: 你的生日是否在 7 月 1 日之前? [问题 B]: 你是否有抑郁倾向?

同时再准备一个箱子, 里面装有 m 个白球和 n 个红球, 被调查者随机抽取一球, 若抽中白色回答问题 A, 否则回答问题 B. 无论抽中哪个问题都只需回答“是”或“否”, 并将答案放入一个密封箱中 (假设在保护隐私的情况下, 学生诚实回答问题). 上述过程在一无人的房间内进行, 以保障被调查者的隐私.

若有 $N(N \geq 500)$ 位学生参加调查, 请估计出具有抑郁倾向的学生比例.

解答: 例 0.49

问题: 如上所述.

解答:

- 事件“答卷选择‘是’”的“原因”有两种情况, 即事件“学生抽到红球时, 选择‘是’”或者事件“学生抽到白球时, 选择‘是’”.
- 设有 N_y 张答卷选择“是”, 一个学生有抑郁倾向的概率为 p; 不妨假设一个学生的生日在 7 月 1 日之前的概率为 1/2, 根据全概率公式有

$$\begin{aligned}&P(\text{一个学生回答‘是’})\\=&P(\text{一个学生回答‘是’}|\text{红球})P(\text{红球})+P(\text{一个学生回答‘是’}|\text{白球})P(\text{白球})\end{aligned}$$

由此可得

$$\frac{N_y}{N}\approx\frac{m}{m+n}\times\frac{1}{2}+\frac{n}{m+n}\times p$$

进一步估计出具有抑郁倾向的学生比例 $p\approx(m+n)N_y/nN-m/2n$.