遍历问题

苏 淳　　江 涛

问题的提出

“遍历性” 是概率论所关注的一种随机性质, 其内容早已渗透到中学数学的教学之中. 例如, 在 2025 年厦门四检中就有这样一道题目:

例 1. 在一个不透明的口袋里装有大小形状完全相同的 n 个小球, 上面分别写有编号 $1, 2, \cdots, n$. 每次从口袋里随机抽取一个小球, 记录编号后放回, 直至取遍小球后立即停止. 以 X_n 记此时的摸球次数.

(1) 试求 X_n 的期望 EX_n;

(2) 证明:$EX_n > n\ln(n+1)$.

在这里, “取遍小球” 就是概率论中所说的 “遍历”. 一般来说

定义 1. *如果某种随机试验共有 n 种不同试验结果 $A_1, A_2, \cdots, A_n$, 它们的出现概率分别为 $p_1, p_2, \cdots, p_n$, 即有*

$$P(A_k) = p_k, \quad k = 1, 2, \cdots, n, \qquad \sum_{k=1}^{n} p_k = 1.$$

重复进行这种试验, 各次试验相互独立进行, 直到 n 种结果都有出现为止, 以 X_n 记所需的试验次数. 则 X_n 就是实现遍历所需的试验次数, 或者称为遍历的实现时刻.

我们有时也把 n 种不同的试验结果 $A_1, A_2, \cdots, A_n$ 称为 n 种不同的状态. 由于状态数目是有限的, 又由于每种状态的出现概率又但是正的, 所以实现遍历是迟早的事.

我们来看一个与遍历性有关的抛掷骰子的问题.

例 2. 抛掷一枚均匀的骰子, 直到抛出的点数中既有 3 的倍数又有非 3 的倍数时为止. 求该事件的发生概率.

以 A 表示一次抛掷中, 抛出的点数是 3 的倍数的事件, 则有 $P(A)=\frac{1}{3}$. 我们要求的是 A 与 $\overline{A}$ 都先后出现过的事件 G 的概率.

如果前 $i-1$ 次抛掷出现的都是事件 A, 而直到第 i 次抛掷时才出现事件 $\overline{A}\,(i\geq 2)$, 我们把这个事件记作 C_i, 那么就有

$$P(C_i)=P(\underbrace{A\cdots A}_{i-1\text{个}}\overline{A})=\left(\frac{1}{3}\right)^{i-1}\cdot\frac{2}{3},\quad i\geq 2.$$

而如果前 $j-1$ 次抛掷出现的是事件 $\overline{A}$, 而直到第 j 次抛掷时 A 才出现 $(j\geq 2)$, 我们把这个事件记作 D_j, 那么就有

$$P(D_j)=P(\underbrace{\overline{A}\cdots\overline{A}}_{j-1\text{个}}A)=\left(\frac{2}{3}\right)^{j-1}\cdot\frac{1}{3},\quad j\geq 2.$$

A 与 $\overline{A}$ 都先后出现过的事件 G 是所有这些 C_i 与 D_j 的并集, 而这些事件都是两两不交的, 所以

$$\begin{aligned}P(G) &= \sum_{i=2}^{\infty}P(C_i)+\sum_{j=2}^{\infty}P(D_j)\\ &= \frac{2}{3}\sum_{i=2}^{\infty}\left(\frac{1}{3}\right)^{i-1}+\frac{1}{3}\sum_{j=2}^{\infty}\left(\frac{2}{3}\right)^{j-1}=\frac{1}{3}+\frac{2}{3}=1.\end{aligned}$$

$P(G)=1$ 这个结论毫不奇怪, 只要我们不断抛掷下去, 就一定会既出现抛出的点数是 3 的倍数, 又出现抛出的点数不是 3 的倍数的现象, 那种永远只出现一种情况的事件的概率是 0, 意即几乎不可能发生.

有鉴于此, 人们在考察遍历性问题时, 一般不再关注遍历性能否实现, 转而关注何时能够实现, 更具体地说, 更加关注遍历实现时刻 X_n 的期望 EX_n.

从几何分布谈起

有限状态的遍历性问题, 都与几何分布有关. 几何分布是一种等待胜利的分布 (参阅 [1][2]), 而我们现在就是在等待一次次的胜利. 每出现一个原先没有出现过的新的状态就是一次新的胜利. 一直要等待到所有 n 个状态都先后出现为止. 因此, 遍历实现时刻 X_n 就是 n 次胜利的总的等待时间, 也就是总的试验次数.

在文章 [1] 和 [2] 中, 我们都给出了参数为 p $(0<p<1)$ 的几何分布随机变量 Y 的期望的计算方法, 并且求得

$$EY=\frac{1}{p},\qquad(*)$$

在这里, p 就是一次试验中所关注的结果的出现概率, 而 Y 就是直到该结果出现为止所进行的试验次数.

现在我们就来解答例 1. 以 Y_k 记在已经抽到 $k-1$ 个不同的小球之后, 等待第 k 个不同的小球出现的等待次数. 那么每个 Y_k 就都是一个服从几何分布的随机变量, 而总的等待次数 X_n 就是所有这些 Y_k 的和, 意即

$$X_n = Y_1 + Y_2 + \cdots + Y_n. \qquad ①$$

易知 $Y_1 = 1$, 这是因为第一次抽出的球, 不论是哪一个, 都是没有出现过的. Y_2 则不然了, 它必须等到不同于前面的球出现才行, 由于此时没有出现过的球有 $n-1$ 个, 所以 Y_2 服从参数为 $p_2 = \dfrac{n-1}{n}$ 的几何分布, 如此下去, 可知, 一般地, Y_k 服从参数为 $p_k = \dfrac{n-k+1}{n}$ 的几何分布. 于是, 根据 ① 和 $(*)$, 可知

$$EX_n = EY_1 + EY_2 + \cdots + EY_n = \left(\frac{1}{n} + \frac{1}{n-1} + \cdots + 1\right) n. \qquad (*)$$

这就解答了例 1 中的问题**(1)**. 为解答问题**(2)**, 只需注意

$$\frac{1}{k} > \frac{1}{x}, \quad k < x < k+1,$$

所以

$$\frac{1}{k} > \int_k^{k+1} \frac{1}{x} dx,$$

从而由 $(*)$ 知

$$EX_n = n\left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1} + \frac{1}{n}\right) > n\int_1^{n+1} \frac{1}{x} dx = n\ln(n+1).$$

$(*)$ 是一个广泛的结论, 从它可以派生出许多问题的答案:

(a)抛掷一枚均匀的硬币, 直到正面与反面都出现过才结束, 平均需要抛掷 3 次 (在 $(*)$ 式中以 $n=2$ 代入即可);

(b)抛掷一个均匀的四面体, 直到四个面都被抛出过才结束, 平均需要抛掷 $\dfrac{25}{3}$ 次 (在 $(*)$ 式中以 $n=4$ 代入即可);

(c)抛掷一枚均匀的骰子, 直到六个面都被抛出过才结束, 平均需要抛掷 14.7 次 (在 $(*)$ 式中以 $n=6$ 代入即可).

非均匀硬币

假设有一枚不均匀的硬币, 每次抛掷时, 抛出正面的概率是 p, 抛出反面的概率是 $q = 1-p$, 其中 $0 < p < 1$, $p \neq \frac{1}{2}$. 反复抛掷这枚硬币, 直到正面与反面都出现过才结束, 以 X 表示所需的抛掷次数. 我们要来讨论 X 的期望与分布.

以 A 表示抛出正面, 那么 $\overline{A}$ 就表示抛出反面. 如果第一次就抛出正面, 那么接下来就只需等候反面的出现, 这时的等待次数 X_1 服从参数为 q 的几何分布; 而如果第一次抛出反面, 那么接下来就要等候正面的出现, 这时的等待次数 X_2 依然服从参数为 p 的几何分布. 如果用示性函数 (参阅 [3]) 表示, 那么就有

$$X = I(A)(1 + X_1) + I(\overline{A})(1 + X_2).$$

从而就有

$$\begin{aligned} EX &= P(A)(1 + EX_1) + P(\overline{A})(1 + EX_2) \\ &= p\Big(1 + \frac{1}{q}\Big) + q\Big(1 + \frac{1}{p}\Big) \\ &= 1 + \frac{p}{q} + \frac{q}{p} = \frac{1}{pq} - 1. \end{aligned}$$

其中用到几何分布的期望公式 $(*)$, 以及第一次试验结果与接下来的等待次数之间的独立性.

如果要写出 $P(X = n)$ 的话, 那么它就等于

$$\begin{aligned} P(X = n) &= P(A)P(X_1 = n - 1) + P(\overline{A})P(X_2 = n - 1) \\ &= p \cdot p^{n-2}q + q \cdot q^{n-2}p = p^{n-1}q + q^{n-1}p, \quad n = 2, 3, \cdots. \end{aligned}$$

其实就是用到了全概率公式.

我们可以把所得的结论写成定理的形式.

定理 1. 如果某个随机试验只有两个不同的实验结果, 每次试验中它们的出现概率分别为 p 与 q $(0 < p < 1,\ p + q = 1)$, 而 X 是直到两种试验结果都出现时所进行的试验次数, 则 X 的分布律是

$$P(X = n) = p^{n-1}q + q^{n-1}p, \quad n = 2, 3, \cdots,$$

X 的期望是

$$EX = \frac{1}{pq} - 1. \qquad ②$$

一般情形

现在假设在独立重复进行的试验中不同的结果数目 $n \geq 3$. 以 $X_{i,j}$ 表示直到结果 A_i 与 A_j 都出现为止所需的试验次数. 我们以 $G(p)$ 表示参数为 p 的几何分布.

如果按第一次的试验结果分类, 则有

$$X_{i,j} = I(A_i)(1+X_j) + I(A_j)(1+X_i) + I(\overline{A}_i\overline{A}_j)(1+X'_{i,j}),$$

其中 $X_j \sim G(p_j),\ X_i \sim G(p_i),\ X'_{i,j}$ 与 $X_{i,j}$ 同分布, 它们都与第一次的实验结果独立. 从而

$$\begin{aligned} EX_{i,j} &= p_i\Big(1+\frac{1}{p_j}\Big) + p_j\Big(1+\frac{1}{p_i}\Big) + (1-p_i-p_j)(1+EX'_{i,j}) \\ &= \frac{p_i}{p_j} + \frac{p_j}{p_i} + 1 + (1-p_i-p_j)EX_{i,j}. \end{aligned}$$

所以

$$EX_{i,j} = \frac{1}{p_i+p_j}\left(\frac{p_i}{p_j} + \frac{p_j}{p_i} + 1\right) = \frac{1}{p_i} + \frac{1}{p_j} - \frac{1}{p_i+p_j}. \qquad ③$$

易见, 在只有两种结果的场合下, 由于 $p_i+p_j=1$, 此时 ③ 式划归 ② 式.

继续往下讨论. 现在假设在独立重复进行的试验中不同的结果数目 $n \geq 4$. 以 $X_{i,j,k}$ 表示直到结果 A_i, A_j 与 A_k 都出现为止所需的试验次数. 其它符号同前. 按第一次的试验结果分类, 得到

$$\begin{aligned} X_{i,j,k} = {} & I(A_i)(1+X_{j,k}) + I(A_j)(1+X_{i,k}) + I(A_k)(1+X_{i,j}) \\ & + I(\overline{A}_i\overline{A}_j\overline{A}_k)(1+X'_{i,j,k}), \end{aligned}$$

其中 $X'_{i,j,k}$ 与 $X_{i,j,k}$ 同分布, 根据后续的等待次数与第一次试验结果的独立性以及 ② 式, 得到

$$\begin{aligned} EX_{i,j,k} &= P(A_i)(1+EX_{j,k}) + P(A_j)(1+EX_{i,k}) + P(A_k)(1+EX_{i,j}) \\ &\quad + P(\overline{A}_i\overline{A}_j\overline{A}_k)(1+EX_{i,j,k}) \\ &= p_i(1+\frac{1}{p_j}+\frac{1}{p_k}-\frac{1}{p_j+p_k}) + p_j(1+\frac{1}{p_i}+\frac{1}{p_k}-\frac{1}{p_i+p_k}) \\ &\quad + p_k(1+\frac{1}{p_i}+\frac{1}{p_j}-\frac{1}{p_i+p_j}) + (1-p_i-p_j-p_k)(1+EX_{i,j,k}) \\ &= 1 + \left(\frac{p_i}{p_j}+\frac{p_i}{p_k}+\frac{p_j}{p_i}+\frac{p_j}{p_k}+\frac{p_k}{p_i}+\frac{p_k}{p_j}\right) - \left(\frac{p_i}{p_j+p_k}+\frac{p_j}{p_i+p_k}+\frac{p_k}{p_i+p_k}\right) \\ &\quad + (1-p_i-p_j-p_k)EX_{i,j,k} \\ &= 1 + \left(\frac{p_j+p_k}{p_i}+\frac{p_i+p_k}{p_j}+\frac{p_i+p_j}{p_k}\right) - \left(\frac{p_i}{p_j+p_k}+\frac{p_j}{p_i+p_k}+\frac{p_k}{p_i+p_k}\right) \\ &\quad + (1-p_i-p_j-p_k)EX_{i,j,k}. \end{aligned}$$

由此易得

$$(p_i+p_j+p_k)EX_{i,j,k} = 1 + \left(\frac{p_j+p_k}{p_i}+\frac{p_i+p_k}{p_j}+\frac{p_i+p_j}{p_k}\right) - \left(\frac{p_i}{p_j+p_k}+\frac{p_j}{p_i+p_k}+\frac{p_k}{p_i+p_k}\right),$$

即有

$$EX_{i,j,k}=\left(\frac{1}{p_i}+\frac{1}{p_j}+\frac{1}{p_k}\right)-\left(\frac{1}{p_j+p_k}+\frac{1}{p_i+p_k}+\frac{1}{p_i+p_j}\right)+\frac{1}{p_i+p_j+p_k}. \quad ③$$

由此并利用归纳法, 我们可以得到在独立重复随机试验中, 等到 n 个不同结果都先后出现的等待次数 X 的期望的计算公式:

$$EX=\sum_i\frac{1}{p_i}-\sum_{i<j}\frac{1}{p_i+p_j}+\sum_{i<j<k}\frac{1}{p_i+p_j+p_k}+\cdots+(-1)^{n+1}\frac{1}{p_1+p_2+\cdots+p_n}.$$

参考文献

[1] 苏淳: 头像问题, 许康华竞赛优学, 微信公众号, 2024-08-24

[2] 苏淳, 江涛: 通过条件期望求期望, 许康华竞赛优学, 微信公众号,

[3] 苏淳: 示性函数与抽屉原理, 许康华竞赛优学, 微信公众号,2024-03-17