详细解答

STEP I 1994，第 13 题

概率与统计 · 指示变量

STEP I 1994，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/94-s1-q13

关键思路

把一个计数写成若干指示变量（indicator variable，每个取值 1 或 0）之和，可以把难算的期望值（expectation）变成易算的，因为无论各部分是否独立（independent），期望值都是线性的。这里各次抽取肯定不独立，但对称性告诉我们每个 X_i 的期望值都相同。然后我们用第二种方法计算 $\mathrm{E}(X)$，即直接由 X 的分布出发；令两个答案相等，几乎不费工夫就得到一条二项式系数（binomial coefficient）恒等式。

解答

全文设 $1 \le n \le M$，并约定当 $b > a$ 或 $b < 0$ 时 $\binom{a}{b} = 0$。

第 (i) 部分

若第 i 个取出的代币是红色，X_i 为 1，否则为 0。因此在和 $X_1 + X_2 + \cdots + X_n$ 之中，每个取出的红色代币恰好贡献 1，每个非红色代币贡献 0。所以这个和数出的正是取出的 n 个代币中红色代币的数目，即 X。

第 (ii) 部分

由于 X_i 只取 0 和 1 两个值，

$$\mathrm{E}(X_i) = 0 \cdot \mathrm{P}(X_i = 0) + 1 \cdot \mathrm{P}(X_i = 1) = \mathrm{P}(\text{第 } i \text{ 个取出的代币是红色}).$$

? **想一想** 第 i 个代币是红色的机会，难道不取决于之前抽到甚么吗？

就条件概率而言，是的；就无条件概率而言，不是。试想像把代币编号，然后逐一全部抽出：M 个代币的每一种排列次序都同样可能出现。因此，位置 i 上的代币同样可能是 M 个代币中的任何一个，正如位置 1 上的代币一样。抽了 n 次便停下来，并不会改变位置 $i \le n$ 上发生的事。由于 M 个代币中有 m 个是红色，

$$\mathrm{E}(X_i) = \mathrm{P}(X_i = 1) = \frac{m}{M} \qquad \text{对每个} i = 1, \ldots, n.$$

（对于 $i = 2$，你可以用较长的方法验证：$\frac{m}{M} \cdot \frac{m-1}{M-1} + \frac{M-m}{M} \cdot \frac{m}{M-1} = \frac{m(M-1)}{M(M-1)} = \frac{m}{M}$。）

第 (iii) 部分

由第 (i) 部分及期望值的线性性质（无需独立性），

$$\mathrm{E}(X) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n) = n \cdot \frac{m}{M} = \frac{mn}{M},$$

正是所求。

第 (iv) 部分

取出的 n 个代币所成的集合，同样可能是 $\binom{M}{n}$ 个大小为 n 的子集中的任何一个。要恰好有 k 个红色代币，我们从 m 个红色代币中选 k 个，再从其余 $M-m$ 个中选 $n-k$ 个，所以

$$\mathrm{P}(X = k) = \frac{\binom{m}{k}\binom{M-m}{n-k}}{\binom{M}{n}}, \qquad k = 0, 1, \ldots, n.$$

按照关于二项式系数的约定，当 $k > m$（红色代币不足）或 $n - k > M - m$（其他代币不足）时，上式正确地等于 0。这就是*超几何分布*（hypergeometric distribution）。

第 (v) 部分

直接由第 (iv) 部分计算 $\mathrm{E}(X)$；$k = 0$ 的一项为零，所以

$$\mathrm{E}(X) = \sum_{k=1}^{n} k\,\mathrm{P}(X = k) = \frac{1}{\binom{M}{n}} \sum_{k=1}^{n} k \binom{m}{k}\binom{M-m}{n-k}.$$

令此式与第 (iii) 部分的结果相等，

$$\sum_{k=1}^{n} k \binom{m}{k}\binom{M-m}{n-k} = \frac{mn}{M}\binom{M}{n}.$$

最后，

$$\frac{n}{M}\binom{M}{n} = \frac{n}{M} \cdot \frac{M!}{n!\,(M-n)!} = \frac{(M-1)!}{(n-1)!\,(M-n)!} = \binom{M-1}{n-1},$$

所以右边等于 $m\binom{M-1}{n-1}$，这就是所求的恒等式。

验算

- 取 $M = 5$，$m = 2$，$n = 3$。第 (iv) 部分给出 $\mathrm{P}(X = 0, 1, 2) = \frac{1}{10}, \frac{6}{10}, \frac{3}{10}$，三者之和为 1，而 $\mathrm{E}(X) = \frac{6+6}{10} = \frac{6}{5} = \frac{mn}{M}$。恒等式则为 $1 \cdot 2 \cdot 3 + 2 \cdot 1 \cdot 3 = 12 = 2\binom{4}{2}$。
- 若 $n = M$（全部取出），则必有 $X = m$，而确实 $\mathrm{E}(X) = mM/M = m$。
- 我们对所有 $M \leq 6$ 的袋子，列出代币的每一种排列次序，验证了 $\mathrm{E}(X_i)$、$\mathrm{E}(X)$ 和 $\mathrm{P}(X = k)$；并对每个 $M < 30$ 验证了 (v) 中的恒等式。

常见错误

- 声称 $\mathrm{E}(X) = \sum \mathrm{E}(X_i)$ 需要独立性。其实不需要：期望值的线性性质永远成立。
- 把各次抽取当作独立，在 (iv) 中写出二项分布（binomial distribution）$\mathrm{B}(n, \frac{m}{M})$。这样得出的平均值是对的，但概率却是错的，因为抽取是不放回的。

STEP 之外

指示变量法是概率论中最有用的工具之一：它可以求出随机置换（random permutation）的不动点（fixed point）数目的期望值、掷硬币时连串（run）数目的期望值、随机图（random graph）中孤立顶点数目的期望值，而且还可藉 $\mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$ 推广至方差（variance）。以两种方法计算同一个量来证明组合恒等式，这种手法同样普遍适用。Blitzstein 与 Hwang 所著的 *Introduction to Probability*（概率论入门）全书贯彻使用指示变量，是观摩这两种想法实际运用的绝佳去处。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。