详细解答

STEP III 1988，第 16 题

概率与统计 · 超几何分布

> **STEP III 1988，第 16 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/88-s3-q16

关键思路

不放回抽取（drawing without replacement）令各次抽取互相依赖，逐次分析会十分繁琐。有两个想法可以绕过这个困难。第一，数算不计次序的选法，从而得到红球数目的分布。第二，利用对称性（symmetry）：假想把整个瓮以随机次序全部抽空，则每个球出现在任何指定位置的机会均等。把 X 写成指示变量（indicator）X_i 之和，便可求得期望值（mean）和方差（variance），完全不需要知道分布。我们只需要各对变量之间的协方差（covariance）。

解答

恰好有 k 个红球的概率

把各个球视为可区分的。抽出的 n 个球组成一个大小为 n 的集合，而每个这样的集合出现的机会均等，所以共有 $\binom{M}{n}$ 个等可能的结果。要组成一个恰好含 k 个红球的集合，须从 m 个红球中选 k 个，并从 $M-m$ 个绿球中选 $n-k$ 个。因此

$$\mathrm{P}(\text{恰好}k\text{ 个红球}) = \frac{\binom{m}{k}\binom{M-m}{n-k}}{\binom{M}{n}},$$

其中沿用惯常的约定：若 $b<0$ 或 $b>a$，则 $\binom{a}{b}=0$。（这已涵盖不可能发生的情况，例如 $n-k>M-m$。）这就是超几何（hypergeometric）分布。

第 (i) 部分

假想一直抽下去，直至瓮中再没有球为止。M 个（可区分的）球被抽出的次序是一个均匀随机的排列：全部 $M!$ 种次序出现的机会均等。抽了 n 次便停下来并不会改变首 n 个位置，所以我们可以用完整的排列来处理。

固定一个位置 i。某个指定的球位于位置 i 的排列共有 $(M-1)!$ 个，这个数目对每个球都相同。所以位置 i 上的球是 M 个球中任何一个的机会均等，而

$$\mathrm{P}(X_i = 1) = \frac{m\,(M-1)!}{M!} = \frac{m}{M},$$

正是所求。

> **想一想** 第二个球的颜色，不是明明取决于第一个球是甚么颜色吗?

在已知第一个球的情况下，确实如此。但我们要求的是*无条件*（unconditional）概率。当 $i = 2$ 时，由全概率定律（law of total probability）可得

$$\mathrm{P}(X_2 = 1) = \frac{m}{M}\cdot\frac{m-1}{M-1} + \frac{M-m}{M}\cdot\frac{m}{M-1} = \frac{m(M-1)}{M(M-1)} = \frac{m}{M}.$$

对称性论证则一次过为每个 i 完成了同样的计算。

第 (ii) 部分

固定两个位置 $i \neq j$（所以 $M \geq 2$）。数算这两个位置上都是红球的排列。位置 i 上的球有 m 个选择，然后位置 j 有 $m-1$ 个选择，其余的球则有 $(M-2)!$ 种排法。因此

$$\mathrm{P}(X_i = 1 \text{ 及 } X_j = 1) = \frac{m(m-1)(M-2)!}{M!} = \frac{m(m-1)}{M(M-1)},$$

正是所求。换句话说，位置 (i, j) 上的有序球对，是 $M(M-1)$ 个由相异的球组成的有序对中任何一个的机会均等。

X 的期望值和方差

X 数算首 n 个抽出的球当中红球的数目。根据期望值的线性性质（linearity of expectation）——这并不需要独立性（independence）——再加上 $\mathrm{E}(X_i) = \mathrm{P}(X_i = 1)$，

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{nm}{M}.$$

至于方差，可用 $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$。把和的平方展开，并利用 $X_i^2 = X_i$（因为 X_i 只取 0 或 1），

$$X^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j = X + \sum_{i\neq j} X_i X_j,$$

其中第二个和遍历满足 $i \neq j$ 的 $n(n-1)$ 个有序对 (i,j)。而 $X_iX_j = 1$ 当且仅当两个球都是红球，所以由第 (ii) 部分得 $\mathrm{E}(X_iX_j) = \dfrac{m(m-1)}{M(M-1)}$。因此，当 $M \geq 2$ 时，

$$\begin{aligned}\mathrm{Var}(X) &= \frac{nm}{M} + \frac{n(n-1)m(m-1)}{M(M-1)} - \frac{n^2m^2}{M^2} \\ &= \frac{nm}{M^2(M-1)}\Big[M(M-1) + (n-1)(m-1)M - nm(M-1)\Big].\end{aligned}$$

方括号内的式子展开后为

$$\begin{aligned}&M^2 - M + nmM - nM - mM + M - nmM + nm \\ &\quad = M^2 - nM - mM + nm = (M-m)(M-n).\end{aligned}$$

因此

$$\mathrm{Var}(X) = \frac{nm(M-m)(M-n)}{M^2(M-1)}.$$

（若 $M = 1$，则 $n \leq 1$，根本没有球对，而 X 是常数，所以 $\mathrm{Var}(X) = 0$。）

设 $p = m/M$ 会令结果更具启发性。这时

$$\mathrm{E}(X) = np, \qquad \mathrm{Var}(X) = np(1-p) \cdot \frac{M-n}{M-1}.$$

这正是二项分布（binomial distribution）的方差乘以一个修正因子，而当 $n \geq 2$ 时，这个因子小于 1。

验算

- $n = M$：我们把所有球都抽出，所以必定有 $X = m$。公式给出期望值 m 和方差 0，后者源于因子 $M - n$。
- $n = 1$：X 是单独一个伯努利（Bernoulli）变量，其 $p = m/M$，而方差化简为 $p(1-p)$。
- $M = 6$、$m = 2$、$n = 3$：公式给出 $\mathrm{Var}(X) = \frac{3\cdot2\cdot4\cdot3}{36\cdot5} = \frac{2}{5}$。把全部 6! 种次序逐一列举，结果完全吻合；$M \leq 6$ 的每一个情况亦然。

常见错误

- 把各个 X_i 当作互相独立，于是得出二项分布的方差 $np(1-p)$。其实它们是负相关（negatively correlated）的：早段抽到红球，会令其后抽到红球的机会降低。
- 忘记了当球对是有序的时候，二重和 $\sum_{i\neq j}$ 共有 $n(n-1)$ 项，而不是 $\binom{n}{2}$ 项。

STEP 之外

把一个计数写成指示变量之和，是概率论中最强大的技巧之一。即使分布本身根本无从写出，这个方法仍能给出期望值和方差。因子 $\frac{M-n}{M-1}$ 就是抽样调查中所用的有限总体修正（finite population correction）。它说明了为何一个访问 1000 人的民意调查，对一个国家所能提供的信息，与对一个无限总体所能提供的几乎一样多。Blitzstein 与 Hwang 所著的 *Introduction to Probability*（概率论入门教材）全书贯彻使用指示变量，并详细讨论超几何分布。

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。