详细解答

STEP III 2004，第 12 题

概率与统计 · 指示变量与错排

> **STEP III 2004，第 12 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/04-s3-q12

关键思路

把一个计数写成指示变量（indicator variable）之和，是概率论中最有用的技巧之一。指示变量的期望值（expectation）其实就是概率，所以 $\mathrm{E}(N)$ 唾手可得；方差（variance）则需要用到每一对变量之间的协方差（covariance），这些协方差很小但不是零，因为球衣并不是独立地分配的。一旦知道 N 的均值和方差都等于 1，参数为 1 的泊松分布（Poisson distribution）自然就是值得一试的模型。

解答

单一指示变量 C_1

球衣的 $m!$ 种分配方式每一种都等可能，所以球衣 1 分给 m 名球员中任何一人的可能性都相同。因此 $\mathrm{P}(C_1 = 1) = \frac{1}{m}$，而由于 C_1 只取 0 和 1 两个值，

$$\mathrm{E}(C_1) = 1 \cdot \mathrm{P}(C_1 = 1) = \frac{1}{m},$$

正如所求。因为 $C_1^2 = C_1$（因 $0^2 = 0$ 及 $1^2 = 1$），我们亦有 $\mathrm{E}(C_1^2) = \frac{1}{m}$，所以

$$\mathrm{Var}(C_1) = \frac{1}{m} - \frac{1}{m^2} = \frac{m-1}{m^2}.$$

C_1 与 C_2 的协方差

这里我们需要 $m \geq 2$。乘积 C_1C_2 恰好在球员 1 和 2 都拿到自己的球衣时等于 1，否则等于 0。已知球员 1 拿到球衣 1，余下的 $m-1$ 件球衣便随机分给其余 $m-1$ 名球员，所以

球员 2 拿到球衣 2 的概率是 $\frac{1}{m-1}$。因此

$$\mathrm{E}(C_1C_2) = \mathrm{P}(C_1 = 1,\, C_2 = 1) = \frac{1}{m} \cdot \frac{1}{m-1},$$

而

$$\begin{aligned}\mathrm{Cov}(C_1, C_2) &= \mathrm{E}(C_1C_2) - \mathrm{E}(C_1)\,\mathrm{E}(C_2) = \frac{1}{m(m-1)} - \frac{1}{m^2} \\ &= \frac{m-(m-1)}{m^2(m-1)} = \frac{1}{m^2(m-1)}.\end{aligned}$$

这个值是正数：知道球员 1 拿对了球衣，就等于从球员 2 可能拿到的球衣中剔除了一件错的，令其拿对的机会略为提高。

> **想一想** 为甚么协方差不是零？各球员的球衣不是分开选的吗？

并不是。球衣是每人分配一件的，所以球员 1 拿到甚么，会影响其他人剩下甚么可拿。各个 C_i 是相依的，这正是计算 $\mathrm{Var}(N)$ 时必须计入协方差项的原因。

N 的均值与方差

根据对称性，每个 C_i 的分布都与 C_1 相同，而每一对 C_i, C_j（其中 $i \neq j$）的联合分布（joint distribution）都与 C_1, C_2 相同（把球员和球衣一并重新编号，并不会改变这个随机分配）。利用期望值的线性性质（linearity of expectation），而这性质并不需要独立性，

$$\mathrm{E}(N) = \sum_{i=1}^{m} \mathrm{E}(C_i) = m \cdot \frac{1}{m} = 1.$$

至于和的方差，我们把所有方差加起来，再加上所有有序对 $i \neq j$ 的协方差，这样的有序对共有 $m(m-1)$ 个：

$$\begin{aligned}\mathrm{Var}(N) &= \sum_{i=1}^{m} \mathrm{Var}(C_i) + \sum_{i \neq j} \mathrm{Cov}(C_i, C_j) \\ &= m \cdot \frac{m-1}{m^2} + m(m-1) \cdot \frac{1}{m^2(m-1)} \\ &= \frac{m-1}{m} + \frac{1}{m} = 1.\end{aligned}$$

所以 $\mathrm{E}(N) = \mathrm{Var}(N) = 1$，正如所求。（若 $m = 1$，则必有 $N = 1$，于是 $\mathrm{Var}(N) = 0$；题目其实隐含假设了 $m \geq 2$，一如计算协方差时已经需要的那样。）

用哪一种近似？

无论 m 取甚么值，正态近似（Normal approximation）的效果都很差。N 只取整数值 $0, 1, 2, \ldots$，而由于均值为 1、标准差为 1，它的概率几乎全部集中在 $0, 1, 2, 3$ 这三四个值上。

这个分布在 0 处被截断，那里只比均值低一个标准差，所以它明显向右偏斜；而 $\mathrm{N}(1,1)$ 分布是对称的，并且在 $-\frac{1}{2}$ 以下有不可忽视的概率（即使作了连续性校正（continuity correction），也约有 7%），但 N 根本不可能取到那里的值。最关键的是，增大 m 也无济于事：对每一个 m，均值和方差都维持在 1，所以分布的形状永远不会展开成钟形。

相反，泊松近似则合情合理。N 所数的，是在为数甚多的 m 个事件之中“成功”的个数，每个事件的概率 $\frac{1}{m}$ 都很小，而这些事件之间只有很弱的相依性（协方差 $\frac{1}{m^2(m-1)}$ 极小）。计数近似服从泊松分布，正是在这种情况下发生的。此外，泊松变量的均值等于方差，而 N 恰好具有这个性质，所以 $\mathrm{Po}(1)$ 是自然的选择。

$m = 4$ 的情况

共有 $4! = 24$ 种等可能的分配方式。把一种分配写成球员 $1,2,3,4$ 所穿球衣的号码序列，则没有人穿上自己球衣的分配是

$$2143,\quad 2341,\quad 2413,\quad 3142,\quad 3412,$$
$$3421,\quad 4123,\quad 4312,\quad 4321.$$

（有系统地列举的方法：球员 1 穿的是球衣 2、3 或 4，而每种情况下都恰好有三种方法完成余下的分配，所以共有 $3 \times 3 = 9$ 种。）因此

$$\mathrm{P}(N = 0) = \frac{9}{24} = \frac{3}{8} = 0.375.$$

参数为 1 的泊松近似给出 $\mathrm{P}(N = 0) \approx \mathrm{e}^{-1}$。取 $\mathrm{e} \approx 2.72$，

$$\mathrm{e}^{-1} \approx \frac{100}{272} = \frac{25}{68}.$$

以真确值为基准的相对误差（relative error）是

$$\frac{\frac{3}{8} - \frac{25}{68}}{\frac{3}{8}} = 1 - \frac{25}{68} \cdot \frac{8}{3} = 1 - \frac{200}{204} = \frac{1}{51} \approx 0.0196,$$

即大约 2%，与题目所述相符。（如果改为以近似值为基准来量度误差，则得 $\frac{1}{50}$，恰好是 2%；无论用哪一种量法，对这么小的 m 来说，吻合程度都好得惊人。）

验算

- 当 $m = 2$ 时，只有两种可能：两人都拿对（$N = 2$）或两人都拿错（$N = 0$），概率各为 $\frac{1}{2}$：均值为 1，方差为 1。吻合。
- 用电脑枚举 $m = 2, \ldots, 7$ 的所有排列（permutation），确认了 $\mathrm{E}(C_1) = \frac{1}{m}$、$\mathrm{Var}(C_1) = \frac{m-1}{m^2}$、$\mathrm{Cov}(C_1, C_2) = \frac{1}{m^2(m-1)}$ 及 $\mathrm{E}(N) = \mathrm{Var}(N) = 1$ 全部准确成立，并找到 4 件物件的错排（derangement）共 9 个，所以 $\mathrm{P}(N = 0) = 0.375$，对比 $1/2.72 \approx 0.3676$。

- 留意 $N = m - 1$ 是不可能的：如果除一人外其他球员都拿对了球衣，最后那一人也必定拿对。这再次显示 N 与正态分布相去甚远。

常见错误

- 把各个 C_i 当作独立，结果得出 $\mathrm{Var}(N) = \frac{m-1}{m}$ 而不是 1。
- 只数了 $\binom{m}{2}$ 个协方差项而不是 $m(m-1)$ 个，即忘记了 $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$ 中的因子 2。
- 列举错排时粗心大意：有系统地数（3×3）可避免遗漏。

STEP 之外

这是经典的“配对”问题，又称错排问题。利用容斥原理（inclusion–exclusion）可以证明 $\mathrm{P}(N=0) = \sum_{k=0}^{m} \frac{(-1)^k}{k!}$，这个值极快地趋向 e^{-1}，而事实上当 $m \to \infty$ 时，N 依分布收敛（converges in distribution）于 $\mathrm{Po}(1)$。更广泛的原理是：由许多罕见而弱相依的事件所构成的计数，近似服从泊松分布；Chen–Stein 方法把这个原理严格地表述出来。Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论入门教材）一书，反复以指示变量和配对问题作为例子。

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。