详细解答

STEP III 样本试卷，第 16 题

概率与统计 • 负二项分布

> **STEP III 样本试卷，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/spec-s3-q16

关键思路

等到 r 个儿子出生，等同于先等一个儿子出生，然后重新开始等下一个，如此重复 r 次。因此 C 可以拆成 r 个分布相同的等候时间之和，余下的工作交给期望值的线性（linearity of expectation）便可。至于第 (ii) 部分，窍门是想像总统停止生育之后，婴儿仍然继续出生：这样一来，他在某个时刻之前是否已经停下，就变成一个关于固定数目的孩子的问题，再利用男孩与女孩之间的对称性（symmetry）便可解决。

解答

以下一律假设每次出生的婴儿是男孩的概率为 $\frac{1}{2}$，并且与其他各次出生互相独立（independent）。

第 (i) 部分

对于 $i = 1, 2, \ldots, r$，设 X_i 为第 $(i-1)$ 个儿子出生之后，直至第 i 个儿子出生为止（包括这个儿子在内）所生的孩子数目。（当 $i = 1$ 时，由最初开始数起。）于是

$$C = X_1 + X_2 + \cdots + X_r.$$

每个 X_i 所数的，都是在一轮全新的、独立而公平的出生之中，直至第一个男孩出现为止的出生次数，所以每个 X_i 的分布（distribution）都与 X 相同，其中

$$\mathrm{P}(X = k) = \left(\tfrac{1}{2}\right)^{k-1} \cdot \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^{k}, \qquad k = 1, 2, 3, \ldots,$$

因为 $X = k$ 表示先有 $k-1$ 个女孩，然后才有一个男孩。

要求出 $\mathrm{E}(X)$，可把几何级数（geometric series）$\sum_{k\geq 0} x^k = \dfrac{1}{1-x}$（适用于 $|x|<1$）求导，得到 $\sum_{k\geq 1} kx^{k-1} = \dfrac{1}{(1-x)^2}$。乘以 x 并代入 $x=\frac{1}{2}$，得

$$\mathrm{E}(X) = \sum_{k=1}^{\infty} k\left(\tfrac{1}{2}\right)^k = \frac{\frac{1}{2}}{\left(1-\frac{1}{2}\right)^2} = 2.$$

这个结果合乎常理：如果所有婴儿之中有一半是男孩，那么平均要等两次出生才会有一个男孩。

根据期望值的线性（无论各个 X_i 是否独立，这个性质都成立，不过这里它们确实是独立的），

$$\mathrm{E}(C) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_r) = 2r.$$

第 (ii) 部分

想像有一个无穷无尽的婴儿序列，各个婴儿的性别早已预先决定，彼此独立而且公平；总统按着这个序列生孩子，直至第 r 个儿子出生为止。这样做完全不会改变 C，却让我们即使在他提早停下的情况下，仍然可以谈论“首 $2r-1$ 个婴儿”。

? **想一想** 怎样只用序列中的首 $2r-1$ 个婴儿，来描述事件 $C<2r$?

由于 C 是整数，$C<2r$ 即是 $C\leq 2r-1$，也就是说，第 r 个儿子在第 $2r-1$ 个位置或之前出现。这种情况发生，当且仅当至少有 r 个男孩出现在首 $2r-1$ 个婴儿之中。所以

$$\mathrm{P}(C<2r) = \mathrm{P}(\text{至少有}r\text{ 个男孩出现在首}2r-1\text{ 个婴儿之中}).$$

现在设 B 和 G 分别为这 $2r-1$ 个婴儿之中男孩和女孩的数目，于是 $B+G=2r-1$。因为 $2r-1$ 是奇数，B 与 G 不可能相等，而且两者之中恰好有一个不小于 r：

- 若 $B\geq r$，则 $G=2r-1-B\leq r-1$；
- 若 $B\leq r-1$，则 $G\geq r$。

所以事件 $\{B\geq r\}$ 与 $\{G\geq r\}$ 互斥（mutually exclusive），而且两者合起来涵盖了所有可能情况，由此得到

$$\mathrm{P}(B\geq r) + \mathrm{P}(G\geq r) = 1.$$

但每次出生时，男孩和女孩出现的机会均等，所以把“男孩”和“女孩”这两个标签对调，并不会改变任何概率，因此 $\mathrm{P}(B\geq r) = \mathrm{P}(G\geq r)$。故此两者各为 $\frac{1}{2}$，而

$$\mathrm{P}(C<2r) = \tfrac{1}{2},$$

这正是所要证明的。

由此可见，家庭人数的期望值是 $2r$，但家庭人数少于 $2r$ 的情况却恰好占一半。家庭人数的分布是偏斜（skewed）的：略小于 $2r$ 的数值十分常见，而偶尔出现的一长串女孩则把平均值拉高。

验算

- 取 $r = 1$：$\mathrm{E}(C) = 2$，而 $C < 2$ 表示第一个婴儿是男孩，其概率为 $\frac{1}{2}$。
- 取 $r = 2$：$C < 4$ 表示 $C = 2$ 或 3，概率分别为 $\frac{1}{4}$ 和 $2 \cdot \frac{1}{8} = \frac{1}{4}$，总和为 $\frac{1}{2}$。
- 对 $r = 1, \ldots, 8$ 把 $c\binom{c-1}{r-1}2^{-c}$ 精确地求和，每次都得到 $\mathrm{E}(C) = 2r$ 和 $\mathrm{P}(C < 2r) = \frac{1}{2}$；以 $r = 3$ 和 200 000 个家庭进行模拟，得出平均值为 6.00，以及 $\mathrm{P}(C < 6) \approx 0.499$。

常见错误

- 以“孩子之中有一半是男孩，所以 $C = 2r$”作为论证。这句话就平均而言是对的，但 C 本身是随机的；第 (i) 部分需要把它分解成各段等候时间，而不能只靠猜测。
- 在第 (ii) 部分只考虑总统实际所生的孩子。当 $C < 2r - 1$ 时，根本就没有 $2r - 1$ 个孩子；正是把出生序列延伸至停止点之后，计数的论证才得以成立。
- 忘记了 $2r - 1$ 是奇数。正是这一点排除了男孩与女孩数目相同的情况，并使两个事件互补（complementary）。

STEP 之外

C 的分布就是负二项（negative binomial）分布：$\mathrm{P}(C = c) = \binom{c-1}{r-1}2^{-c}$，其中 $c \geq r$。把一个随机变量（random variable）写成若干个较简单部分之和，再运用期望值的线性，是大学一年级概率论中最有力的工具之一；而第 (ii) 部分把“等候时间”与“固定试验次数中的计数”联系起来的窍门，正是连结负二项分布与二项（binomial）分布的那种对偶性（duality）（在连续时间的情况下，它连结的是伽玛（gamma）分布与泊松（Poisson）分布）。Blitzstein 与 Hwang 合著的 *Introduction to Probability* 以大量例子讲解了这两个想法。

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。