详细解答

STEP III 2011，第 12 题

概率与统计 · 以概率生成函数处理复合分布

> **STEP III 2011，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/11-s3-q12

关键思路

关于一个分布的平均值（mean）和方差（variance），我们所需要的一切都藏在它的概率母函数（probability generating function, pgf）在 $t = 1$ 处的首两阶导数之中。因此，我们用连锁法则（chain rule）对复合函数 $G(H(t))$ 求导（两次），代入 $t = 1$，并利用 $H(1) = 1$。至于掷硬币的部分，N 服从几何分布，而每个 X_i 都是单次掷公平硬币的指示变量（indicator），所以两个概率母函数都很简单，而 $G(H(t))$ 是一个有理函数（rational function），可以展开成几何级数（geometric series）。

解答

一般结果

对于任何取非负整数值、概率母函数为 $P(t) = \mathrm{E}(t^Z)$ 的随机变量 Z，我们有 $P(1) = 1$，

$$P'(1) = \mathrm{E}(Z), \qquad P''(1) = \mathrm{E}(Z(Z-1)),$$

因此

$$\mathrm{Var}(Z) = P''(1) + P'(1) - (P'(1))^2.$$

记 $\mu = \mathrm{E}(X_i) = H'(1)$、$\sigma^2 = \mathrm{Var}(X_i)$，并以 $K(t) = G(H(t))$ 表示 Y 的概率母函数。

由连锁法则，$K'(t) = G'(H(t))\,H'(t)$。由于 $H(1) = 1$，

$$\mathrm{E}(Y) = K'(1) = G'(1)H'(1) = \mathrm{E}(N)\,\mathrm{E}(X_i),$$

这就是第一个结果。

再求导一次（先用乘积法则（product rule），再用连锁法则），

$$K''(t) = G''(H(t))(H'(t))^2 + G'(H(t))\,H''(t),$$

所以，再次利用 $H(1) = 1$，

$$K''(1) = G''(1)\mu^2 + \mathrm{E}(N)\,H''(1).$$

把上面的方差公式分别应用于 N 和 X_i，得

$$G''(1) = \mathrm{Var}(N) - \mathrm{E}(N) + (\mathrm{E}(N))^2, \qquad H''(1) = \sigma^2 - \mu + \mu^2.$$

? **想一想** 为甚么不直接由 $K''(1)$ 计算 $\mathrm{Var}(Y)$，而要先改写 $G''(1)$ 和 $H''(1)$?

因为我们想要的答案是以方差表示，而不是以阶乘矩（factorial moments）表示；通过 $\mathrm{Var}(N)$ 和 σ^2 来表达 $G''(1)$ 和 $H''(1)$，可以让多余的项互相抵消。记 $m = \mathrm{E}(N)$，

$$\begin{aligned}\mathrm{Var}(Y) &= K''(1) + K'(1) - (K'(1))^2 \\ &= (\mathrm{Var}(N) - m + m^2)\mu^2 + m(\sigma^2 - \mu + \mu^2) + m\mu - m^2\mu^2 \\ &= \mathrm{Var}(N)\mu^2 + m\sigma^2 + (-m\mu^2 + m\mu^2) + (m^2\mu^2 - m^2\mu^2) + (-m\mu + m\mu) \\ &= \mathrm{Var}(N)(\mathrm{E}(X_i))^2 + \mathrm{E}(N)\,\mathrm{Var}(X_i),\end{aligned}$$

即为所求。

掷硬币

*N 的概率母函数。*第一次掷出正面发生在第 n 次，当且仅当首 $n-1$ 次都是反面而第 n 次是正面，所以对 $n = 1, 2, 3, \ldots$，$\mathrm{P}(N = n) = (\frac{1}{2})^n$（这是几何分布（geometric distribution））。因此，当 $|t| < 2$ 时，

$$G(t) = \sum_{n=1}^{\infty} \left(\frac{t}{2}\right)^n = \frac{t/2}{1 - t/2} = \frac{t}{2 - t}.$$

*每个 X_i 的概率母函数。*若其后的第 i 次掷出正面，令 $X_i = 1$，否则令 $X_i = 0$。于是 $Y = X_1 + \cdots + X_N$，而 $H(t) = \frac{1}{2} + \frac{1}{2}t = \frac{1}{2}(1 + t)$。其后各次掷硬币彼此独立，亦与决定 N 的各次掷硬币独立，所以一般结果的条件成立。

Y 的概率母函数。

$$G(H(t)) = \frac{\frac{1}{2}(1 + t)}{2 - \frac{1}{2}(1 + t)} = \frac{1 + t}{3 - t}.$$

*平均值与方差。*对于 $p = \frac{1}{2}$ 的几何分布，$\mathrm{E}(N) = 1/p = 2$，$\mathrm{Var}(N) = (1 - p)/p^2 = 2$（或者对 G 求导：$G'(t) = 2/(2 - t)^2$、$G''(t) = 4/(2 - t)^3$，得 $G'(1) = 2$、$G''(1) = 4$，以及

$\mathrm{Var}(N) = 4 + 2 - 4 = 2$)。每个 X_i 都服从伯努利（Bernoulli）分布 $(\frac{1}{2})$，所以 $\mathrm{E}(X_i) = \frac{1}{2}$，$\mathrm{Var}(X_i) = \frac{1}{4}$。因此

$$\mathrm{E}(Y) = 2 \cdot \tfrac{1}{2} = 1, \qquad \mathrm{Var}(Y) = 2 \cdot \tfrac{1}{4} + 2 \cdot \tfrac{1}{4} = 1.$$

Y 的分布。当 $|t| < 3$ 时，我们展开

$$\frac{1+t}{3-t} = \frac{1+t}{3} \cdot \frac{1}{1-t/3} = \frac{1+t}{3} \sum_{k=0}^{\infty} \frac{t^k}{3^k} = \sum_{k=0}^{\infty} \frac{t^k}{3^{k+1}} + \sum_{k=0}^{\infty} \frac{t^{k+1}}{3^{k+1}}.$$

t^0 的系数只来自第一个和。当 $r \geq 1$ 时，t^r 的系数是 $\dfrac{1}{3^{r+1}} + \dfrac{1}{3^r} = \dfrac{4}{3^{r+1}}$。所以

$$\mathrm{P}(Y = 0) = \frac{1}{3}, \qquad \mathrm{P}(Y = r) = \frac{4}{3^{r+1}} \quad (r = 1, 2, 3, \ldots).$$

验算

- 各概率之和为 1：$\frac{1}{3} + 4\sum_{r\geq 1} 3^{-(r+1)} = \frac{1}{3} + 4 \cdot \frac{1}{6} = 1$。
- 直接计算 $\mathrm{P}(Y = 0)$：$\sum_{n\geq 1} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^n = \sum_{n\geq 1} 4^{-n} = \frac{1}{3}$。同理，$\mathrm{P}(Y = 1) = \sum_{n\geq 1} n\, 4^{-n} = \frac{4}{9}$，与 $4/3^2$ 吻合。
- 对 $n \leq 200$ 精确求和 $\mathrm{P}(N = n)\binom{n}{r}2^{-n}$，可重现 $r \leq 7$ 时的 $\mathrm{P}(Y = r)$，而所得平均值和方差均为 1.000；以 200 000 次模拟，得到 0.996 和 0.994。

常见错误

- 把 N 当作从 0 开始（概率母函数为 $1/(2-t)$）：这里 N 计算的是掷出正面的那一次，所以 $N \geq 1$。
- 忘记 $P''(1)$ 是 $\mathrm{E}(Z^2) - \mathrm{E}(Z)$ 而不是 $\mathrm{E}(Z^2)$，以致在方差公式中漏掉 $+P'(1)$。
- 对所有 $r \geq 0$ 都给出 $4/3^{r+1}$：$r = 0$ 的一项是不同的。

STEP 之外

Y 是一个随机和（random sum）（即复合分布（compound distribution）），而上述方差公式是全方差定律（law of total variance）$\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid N)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid N)\big)$ 的一个特例。概率母函数的复合也是分支过程（branching process）的核心：在分支过程中，$G(G(\cdots G(t)))$ 给出第 n 代的规模，而灭绝的概率是 G 的一个不动点（fixed point）。Grimmett 和 Welsh 的 *Probability: An Introduction*（《概率导论》）正是在这个程度上讲解母函数、随机和及分支过程。

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。