详细解答

STEP I 2018，第 12 题

概率与统计 · 放回抽样

> **STEP I 2018，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/18-s1-q12

关键思路

把每个正面次数写成两个指示变量（indicator variable）之和，每次抛掷对应一个。无论各次抛掷是否独立（independent），和的期望值（expectation）都等于期望值之和，所以平均值很容易求得。方差（variance）则取决于一个数，就是两次抛掷皆出现正面的概率，而这正是放回抽样（drawing with replacement）与不放回抽样（drawing without replacement）的分别所在。

解答

全文中，设 $p = \frac{1}{3}(p_1 + p_2 + p_3)$ 及

$$q = p_1p_2 + p_2p_3 + p_3p_1.$$

我们假设每枚硬币被选中的机会相等，而且硬币选定后，各次抛掷互相独立。

第 (i) 部分

设 C_k 为选中第 k 枚硬币的事件，故 $\mathrm{P}(C_k) = \frac{1}{3}$。由全概率定律（law of total probability），

$$\mathrm{P}(\text{正面}) = \sum_{k=1}^{3} \mathrm{P}(C_k)\,\mathrm{P}(\text{正面} \mid C_k) = \tfrac{1}{3}(p_1 + p_2 + p_3) = p.$$

第 (ii) 部分

设 X_1 和 X_2 分别为第一次和第二次抛掷出现正面的指示变量，故 $N_1 = X_1 + X_2$。由于第一枚硬币会被放回，第二次抛掷的情况与第一次完全相同，由第 (i) 部分得 $\mathrm{P}(X_1 = 1) =$

$\mathrm{P}(X_2=1)=p$。因此

$$\mathrm{E}(N_1)=\mathrm{E}(X_1)+\mathrm{E}(X_2)=2p.$$

求方差需要 $\mathrm{P}(X_1=1,\,X_2=1)$。两次选择互相独立，9 个有序硬币对 (j,k) 的概率各为 $\frac{1}{9}$，而给定该硬币对时，两次皆为正面的机会是 p_jp_k。所以

$$\mathrm{P}(X_1=1,\,X_2=1)=\frac{1}{9}\sum_{j=1}^{3}\sum_{k=1}^{3}p_jp_k=\left(\frac{p_1+p_2+p_3}{3}\right)^2=p^2=\mathrm{P}(X_1=1)\,\mathrm{P}(X_2=1).$$

同样的因式分解对其他结果亦成立，因此 X_1 与 X_2 互相独立，且 $N_1\sim\mathrm{B}(2,p)$。也可以直接计算：由于 $X_i^2=X_i$，

$$\begin{aligned}\mathrm{E}(N_1^2)&=\mathrm{E}(X_1)+\mathrm{E}(X_2)+2\,\mathrm{E}(X_1X_2)=2p+2p^2,\\ \mathrm{Var}(N_1)&=2p+2p^2-(2p)^2=2p(1-p),\end{aligned}$$

正如所求。

第 (iii) 部分

现在以不放回方式先后选出两枚硬币，设 Y_1、Y_2 分别为第一枚和第二枚选出的硬币出现正面的指示变量，故 $N_2=Y_1+Y_2$。每枚硬币成为第一枚被选中的机会相等，由对称性，每枚成为第二枚被选中的机会亦相等，所以与第 (i) 部分一样，$\mathrm{P}(Y_1=1)=\mathrm{P}(Y_2=1)=p$，且

$$\mathrm{E}(N_2)=2p.$$

? **想一想** 可以沿用第 (ii) 部分中 P(两次皆正面) 的计算吗?

不可以：同一枚硬币不能再被选中两次。现在有 6 个机会相等的有序对 (j,k)，其中 $j\neq k$，而每个无序对 $\{j,k\}$ 出现两次，所以

$$\mathrm{P}(Y_1=1,\,Y_2=1)=\frac{1}{6}\sum_{j\neq k}p_jp_k=\frac{2q}{6}=\frac{q}{3}.$$

一般而言这并不等于 p^2，故 Y_1 与 Y_2 并不独立。再次利用 $Y_i^2=Y_i$，

$$\begin{aligned}\mathrm{E}(N_2^2)&=\mathrm{E}(Y_1)+\mathrm{E}(Y_2)+2\,\mathrm{E}(Y_1Y_2)=2p+\tfrac{2}{3}q,\\ \mathrm{Var}(N_2)&=2p+\tfrac{2}{3}(p_1p_2+p_2p_3+p_3p_1)-4p^2.\end{aligned}$$

第 (iv) 部分

两式相减，

$$\mathrm{Var}(N_1)-\mathrm{Var}(N_2)=(2p-2p^2)-(2p+\tfrac{2}{3}q-4p^2)=2p^2-\tfrac{2}{3}q.$$

而 $9p^2 = (p_1+p_2+p_3)^2 = p_1^2+p_2^2+p_3^2+2q$，所以

$$\begin{aligned} 2p^2 - \tfrac{2}{3}q &= \tfrac{2}{9}(p_1^2+p_2^2+p_3^2+2q) - \tfrac{6}{9}q = \tfrac{2}{9}(p_1^2+p_2^2+p_3^2-p_1p_2-p_2p_3-p_3p_1) \\ &= \tfrac{1}{9}\Big[(p_1-p_2)^2+(p_2-p_3)^2+(p_3-p_1)^2\Big], \end{aligned}$$

最后一步成立，是因为展开这三个平方会得到 $2(p_1^2+p_2^2+p_3^2)-2q$。

实数的平方和是非负的，所以 $\mathrm{Var}(N_2) \le \mathrm{Var}(N_1)$。等号成立当且仅当三个平方皆为零，即当且仅当 $p_1 = p_2 = p_3$。证明完毕。

作为副产品，我们得到第 (iii) 部分答案的一个简洁形式：

$$\mathrm{Var}(N_2) = 2p(1-p) - \tfrac{1}{9}\Big[(p_1-p_2)^2+(p_2-p_3)^2+(p_3-p_1)^2\Big].$$

验算

- 若 $p_1 = p_2 = p_3 = p$，每枚硬币都一样，那么硬币怎样选出并无分别：两个次数都服从 $\mathrm{B}(2,p)$，方差相等，与等号成立的情况吻合。
- 取 $p_1 = \frac{1}{2}$、$p_2 = \frac{1}{3}$、$p_3 = \frac{1}{6}$，则 $p = \frac{1}{3}$，$q = \frac{11}{36}$。于是 $\mathrm{Var}(N_1) = \frac{4}{9}$，$\mathrm{Var}(N_2) = \frac{2}{3} + \frac{11}{54} - \frac{4}{9} = \frac{23}{54}$，少了 $\frac{1}{54}$，与 $\frac{1}{9}\left(\frac{1}{36}+\frac{1}{36}+\frac{1}{9}\right) = \frac{1}{54}$ 一致。亲手列出所有结果亦得出相同数值。
- 极端情况 $p_1 = p_2 = 1$、$p_3 = 0$：在第 (iii) 部分中，除非第 1 和第 2 枚硬币同时被选中，否则选出的两枚硬币中恰有一枚必定是正面，因此 N_2 以概率 $\frac{1}{3}$ 取值 2，其余情况取值 1，方差为 $\frac{2}{9}$。公式给出 $2p + \frac{2}{3}q - 4p^2 = \frac{4}{3} + \frac{2}{3} - \frac{16}{9} = \frac{2}{9}$。

常见错误

- 把第 (iii) 部分的两次抛掷当作独立，于是又得到 $2p(1-p)$；第 (iv) 部分的重点正正在于它们并不独立。
- 把 $\mathrm{Var}(N_1)$ 写成各枚硬币方差 $2p_k(1-p_k)$ 的平均；这忽略了随机选择硬币所带来的额外变异，而且在第 (ii) 部分中，每次抛掷可能用不同的硬币。

STEP 之外

第 (ii) 与第 (iii) 部分的分别，就是放回抽样与不放回抽样的分别。不放回抽样使各次抽取呈负相关（negatively correlated），从而降低方差；同样的效应在抽样调查中产生“有限总体校正”（finite population correction），亦解释了为何超几何分布（hypergeometric distribution）的方差比具相同平均值的二项分布（binomial distribution）小。第 (iv) 部

分的代数运算就是恒等式 $\sum p_k^2 - \sum_{j<k} p_j p_k = \frac{1}{2}\sum_{j<k}(p_j - p_k)^2$，它与“方差永不为负”这个事实关系密切。Blitzstein 和 Hwang 的 *Introduction to Probability*（《概率导论》）正是以这种精神讲解指示变量、协方差（covariance）和超几何分布。

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。