詳細解答

STEP III 1997，第 12 題

概率與統計 • 贈券收集問題

> **STEP III 1997，第 12 題**
>
> 繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：
>
> step.maths.org/questions/97-s3-q12

關鍵思路

第 (i) 部分是幾何分佈（geometric distribution），而它的概率生成函數（probability generating function）只需求導，便可給出平均值和方差：$\mathrm{E}(K) = \mathrm{G}'(1)$ 及 $\mathrm{E}(K(K-1)) = \mathrm{G}''(1)$。在第 (ii) 部分，竅門是把集齊全套的漫長等待拆成 52 段短等待，每段對應一張新牌。每段短等待都是一個「擲硬幣直至首次出現正面」的實驗，所以第 (i) 部分已經把工夫全做好了。

解答

第 (i) 部分

首次正面出現在第 k 次投擲，當且僅當頭 $k-1$ 次投擲都是反面，而第 k 次是正面。各次投擲互相獨立（independent），所以

$$\mathrm{P}(K = k) = q^{k-1}p, \qquad k = 1, 2, 3, \ldots$$

因此

$$\mathrm{G}(s) = \sum_{k=1}^{\infty} pq^{k-1}s^k = ps\sum_{j=0}^{\infty}(qs)^j = \frac{ps}{1-qs},$$

這裏用了等比級數（geometric series）求和，該級數在 $|qs| < 1$ 時收斂（特別是當 $p > 0$ 時，對接近 1 的 s 收斂）。這就是所要求的結果。

求導得

$$\mathrm{G}'(s) = \frac{p(1-qs)+psq}{(1-qs)^2} = \frac{p}{(1-qs)^2},$$
$$\mathrm{G}''(s) = \frac{2pq}{(1-qs)^3}.$$

由於 $\mathrm{G}'(s) = \sum k\,\mathrm{P}(K=k)s^{k-1}$ 及 $\mathrm{G}''(s) = \sum k(k-1)\mathrm{P}(K=k)s^{k-2}$，代入 $s=1$ 並利用 $1-q=p$，可得

$$\mathrm{E}(K) = \mathrm{G}'(1) = \frac{p}{p^2} = \frac{1}{p}, \qquad \mathrm{E}(K(K-1)) = \mathrm{G}''(1) = \frac{2pq}{p^3} = \frac{2q}{p^2}.$$

所以

$$\begin{aligned}\mathrm{Var}(K) &= \mathrm{E}(K(K-1)) + \mathrm{E}(K) - (\mathrm{E}(K))^2 \\ &= \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q+p-1}{p^2} = \frac{q}{p^2}.\end{aligned}$$

故此 $\mathrm{E}(K) = \dfrac{1}{p}$ 及 $\mathrm{Var}(K) = \dfrac{q}{p^2}$。

第 (ii) 部分

設 N_i 為已經見過恰好 $i-1$ 張不同的牌之後所抽的次數，計至並包括抽出第 i 張不同的牌的那一次。於是 $N_1 = 1$（第一張牌必定是新的），而

$$N = N_1 + N_2 + \cdots + N_{52}.$$

想一想 為甚麼每個 N_i 都是像 K 那樣的「首次正面」變量？

一旦見過 $i-1$ 張不同的牌，52 張牌之中仍有 $53-i$ 張未見過。由於抽牌後放回（with replacement），每次抽牌都與過去的結果獨立，而且抽出未見過的牌的概率都相同，即

$$p_i = \frac{53-i}{52}.$$

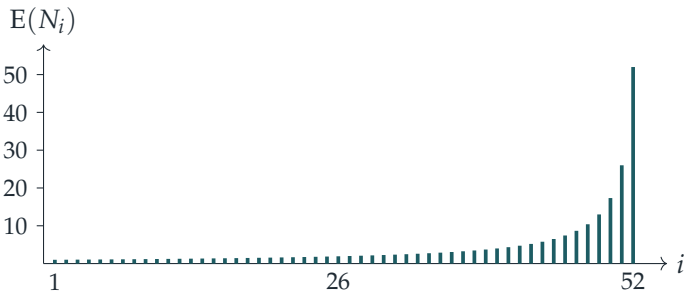
把這樣的一次抽牌稱為「正面」。那麼 N_i 就是計至並包括首次正面的抽牌次數，正正是第 (i) 部分中的變量 K，其中 $p = p_i$。因此

$$\mathrm{E}(N_i) = \frac{1}{p_i} = \frac{52}{53-i}.$$

期望值（expectation）是線性的（這裏毋須各個 N_i 互相獨立），所以

$$\begin{aligned}\mathrm{E}(N) &= \sum_{i=1}^{52} \frac{52}{53-i} = 52\left(\frac{1}{52} + \frac{1}{51} + \cdots + \frac{1}{2} + \frac{1}{1}\right) \\ &= 52\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{52}\right).\end{aligned}$$

早段的等待很短，最後幾段才是主導：單是最後一張牌，平均就要抽 52 次。



要作估算，可用 $1+\frac{1}{2}+\cdots+\frac{1}{n}\approx 0.5+\ln n$，取 $n=52$。至於 $\ln 52$：$2.7^2=7.29$ 而 $7.29^2\approx 53.1$，所以 $e^4\approx 53\approx 52$，得出 $\ln 52\approx 4$。因此

$$\mathrm{E}(N)\approx 52\,(0.5+4)=52\times 4.5=234.$$

所以平均而言，我們大約要抽 234 次，即約為整副牌張數的四倍半，才能見齊每一張牌。

驗算

- 在 (i) 中，$\mathrm{G}(1)=p/(1-q)=1$，任何概率生成函數都必須如此。若硬幣是公平的 $(p=\frac{1}{2})$，便得 $\mathrm{E}(K)=2$ 及 $\mathrm{Var}(K)=2$；而當 $p\to 1$ 時，方差 $q/p^2\to 0$，這正是第一次投擲幾乎必然是正面時應有的結果。
- 準確值是 $52H_{52}\approx 235.98$，而模擬 10^5 次完整收集所得的平均值是 235.7。我們的估計值 234 相當接近；那少許差額來自近似式 $0.5+\ln n$（較精確的是 $\ln n+0.577$），以及把 $\ln 52\approx 3.95$ 上捨入為 4。

常見錯誤

- 把 N_i 的成功概率弄顛倒：抽到*新*牌的機會是 $(53-i)/52$，它隨着收集到的牌增多而減小。
- 取 $\mathrm{Var}(K)=\mathrm{G}''(1)-\mathrm{G}'(1)^2$，忘記了 $+\mathrm{G}'(1)$ 這一項；$\mathrm{G}''(1)$ 是 $\mathrm{E}(K(K-1))$，而不是 $\mathrm{E}(K^2)$。
- 以為各個 N_i 必須互相獨立，才可以把期望值相加。無論獨立與否，期望值的線性性質都成立（不過在這裏它們其實是獨立的）。

STEP 之外

這就是優惠券收集問題（coupon collector's problem）：若有 n 種出現機會均等的類型，集齊全部所需的期望時間是 $nH_n \approx n\ln n$。大學本科課程會更進一步，證明這個時間高度集中在 $n\ln n$ 附近，波動的量級為 n；而概率生成函數亦會成為處理這類獨立變量之和的標準工具。Grimmett 與 Welsh 的 *Probability: An Introduction* 仔細講解生成函數和幾何分佈；Blitzstein 與 Hwang 的 *Introduction to Probability* 則討論優惠券收集問題及其多種變化。

Beyond STEP 詳細解答：https://beyondstep.keithmcnulty.org/zh-hant/。解答文字採用 CC0 1.0；題目本身版權屬 Cambridge University Press & Assessment 所有（©），本文件不予轉載。由機器從英文翻譯而成；尚未經審閱。