9.11 集齐一套

工作纸

题目

STEP II 1994，第 13 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/94-s2-q13

提示

请先自行尝试解题。如果遇到困难，请逐一阅读提示，每读完一个便回到题目再试。标有*停一停，试一试*的提示是关键步骤。每个提示之后均留有空位供你演算。

> **想一想**　如果每包各自独立地以概率 p 开出一张新卡，那么直至开出第一张新卡为止（连这一包计算在内），你开包数目的期望值（expectation）是多少？

阶段 1　等待新卡

涵盖：等待新卡

> **想一想**　已经有 r 张不同的卡时，下一包抽到新卡的概率是多少？利用几何分布（geometric distribution）的期望等待次数。

阶段 2 比较和与积分

涵盖：积分不等式

想一想 在 $r \le x \le r+1$ 上，$\frac{1}{x}$ 介乎哪两个值之间？在这个长度为 1 的区间上积分。

停一停，试一试 把你的不等式由 $r = 1$ 加到 $r = n - 1$。各个积分会连接成一个积分：是哪一个？

阶段 3　集齐全套

涵盖：集齐整套卡

想一想 所需的总包数，就是每张新卡的等待次数之和。把它们的期望值（expectation）相加，再利用第二阶段的结果，把这个和与 $\ln n$ 比较。

背后的数学

这个估计有多准确？

这个比较还可以再精确一些。图 9.11.1（见网上该节的完整版）中曲线与上方长方形之间的阴影空隙，总面积是有限的，所以 $H_n - \ln n$ 趋向一个极限，

$$\gamma = \lim_{n\to\infty}(H_n - \ln n) = 0.5772156649\ldots,$$

即欧拉常数。（没有人知道 γ 是否有理数。）因此包数的期望值是 $n\ln n + \gamma n + \frac{1}{2} + \cdots$。以 600 张贴纸的贴纸簿为例，$600 \times (\ln 600 + 0.577) \approx 4185$，这就是引子中所引述的数字。

时间都花到哪里去了？各个阶段极不平均。集齐全套的前一半只需大约 $n\ln 2 \approx 0.69n$ 包，而单是最后一张卡，平均就要 n 包。大部分工夫都花在最后几张卡上，这正是收藏者要互相交换重复卡的原因。

集中现象

期望值只是个开始：总数的变动有多大？总数的方差（variance）是 $\sum nr/(n-r)^2$，约为 $\frac{\pi^2}{6}n^2$（Section 2.4 中的和 $\sum 1/k^2 = \pi^2/6$ 在这里出现了）。所以标准差（standard deviation）约为 $1.28n$，当 n 很大时远小于平均值 $n\ln n$：总数集中在其平均值附近。艾狄胥（Paul Erdős）和雷尼（Alfréd Rényi）在 1961 年证明了一个更精确的结果：总数减去 $n\ln n$ 再除以 n，会有一个极限分布（limiting distribution），即耿贝尔分布（Gumbel distribution），这个分布也描述了许多随机量之中最大的一个（Section 12.25）。

无处不在的集券问题

这个问题可追溯至棣莫弗（Abraham de Moivre，1712 年）和拉普拉斯（Laplace）。如今它以许多不同的面貌出现。要检查一个系统的每个组件，需要做多少次随机测试？在一个网络上进行随机漫步（random walk），要多久才会走遍每个节点？电脑的杂凑表（hash table）要探查多少次？每一种情况的答案都涉及 $n \ln n$。在生物学中，同样的计算可用来估计要有多少个定序“读段”（reads）才能覆盖整个基因组。

延伸练习

1 **半套。**证明集齐 $\frac{n}{2}$ 张不同的卡（n 为偶数）所需包数的期望值是 $n(H_n - H_{n/2})$，并且这个值约为 $n \ln 2$。

2 **不同的卡。**证明开了 k 包之后，你手上不同的卡的数目的期望值是 $n(1-(1-\frac{1}{n})^k)$。（提示：对每一张卡，求出你拥有它的概率，然后相加。）大约要开多少包，才能得到全套的 95%？

3 **两位收藏者。**两个小孩收集同一套卡，并共用所开的各包，因此对其中一人来说是重复的卡，对另一人来说可能是新的。证明集齐两套所需包数的期望值约为 $n(\ln n + \ln \ln n)$。（这题较难；先说服自己这个数目远小于 $2n \ln n$。）

4 **更精确的界。**改用梯形法则（trapezium rule）的图像而不用长方形，证明 $H_n - \ln n$ 是递减的，而 $H_n - \ln n - \frac{1}{2n}$ 是递增的。由较小的 n 值推导出 $0.5 < \gamma < 0.7$。

延伸阅读

- **Geoffrey Grimmett & David Stirzaker, *One Thousand Exercises in Probability***（Oxford University Press）。集券问题的许多变化。
- **Julian Havil, *Gamma: Exploring Euler's Constant***（Princeton University Press）。关于 γ 和调和级数的一切，写得深入浅出。
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing***（Cambridge University Press）。集券问题及其在电脑科学中的应用。

完整的引导式解答见第 9.11 节：

https://beyondstep.keithmcnulty.org/zh-hans/ch09/coupon-collector.html

每星期一道 STEP 题目，完整讲解：The STEP Up，以电子邮件免费发送，订阅网址：keithmcnulty.substack.com