详细解答

STEP II 1994，第 13 题

概率与统计 · 赠券收集问题

> **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

本题在 *Beyond STEP* 第 9.11 节集齐一套中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch09/coupon-collector.html

关键思路

这是经典的“优惠券收集问题”（coupon collector problem），虽然资料库把它归入别的类别，它其实是纯粹的概率（probability）题。把集齐全套所需的漫长等待拆成 n 段较短的等待：先等第一张新卡，再等第二张，如此类推。每段短等待都服从几何分布（geometric distribution），所以平均值很容易求得，再用期望值的线性性质（linearity of expectation）把它们加起来。总和是 n 乘以一个调和和（harmonic sum），而题中的积分不等式正是用来告诉我们这个和有多大。

解答

等待一张新卡

我们假设每包内有 n 张卡的其中一张，每张出现的概率均为 $\frac{1}{n}$，而且各包之间互相独立（independent）。

假设我已经拥有 r 张不同的卡，其中 $0 \leq r \leq n-1$。打开一包得到新卡的概率是

$$p = \frac{n-r}{n} > 0,$$

得到旧卡的概率则是 $q = 1 - p = \frac{r}{n}$。设 X 为我直至第一次得到新卡为止（包括该包在内）所打开的包数。由于各包互相独立，$X = k$ 表示先有 $k-1$ 张旧卡，然后是一张新卡，

所以

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

因此

$$\mathrm{E}(X)=\sum_{k=1}^{\infty} kpq^{k-1}=p\,(1-q)^{-2}=\frac{p}{p^2}=\frac{1}{p},$$

其中用到当 $|q|<1$ 时 $\sum_{k\geq 1} kq^{k-1}=(1-q)^{-2}$（把几何级数（geometric series）求导，或者用二项式定理展开 $(1-q)^{-2}$）。这里 $0\leq q<1$，因为 $r\leq n-1$。所以

$$\mathrm{E}(X)=\frac{n}{n-r},$$

这正是所要证明的结果。

（一个更简洁的做法：第一包要么给出新卡，我们就此停下；要么以概率 q 令我们回到起点，所以 $\mathrm{E}(X)=1+q\,\mathrm{E}(X)$，由此得 $\mathrm{E}(X)=1/(1-q)=1/p$。这个做法假设了 $\mathrm{E}(X)$ 是有限的，而上面的级数已确认了这一点。）

积分不等式

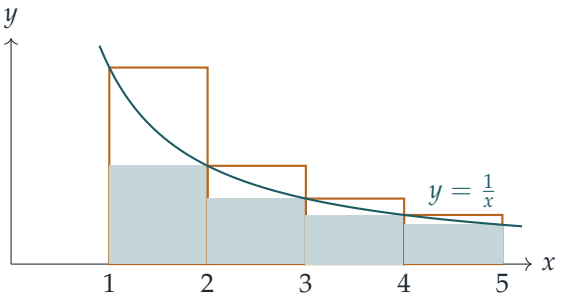
设 $r\geq 1$。当 $r\leq x\leq r+1$ 时，我们有

$$\frac{1}{r+1}\leq\frac{1}{x}\leq\frac{1}{r},$$

因为 $1/x$ 在 $x>0$ 时是递减的。在长度为 1 的区间上积分，这些不等式仍然成立，所以

$$\frac{1}{r+1}\leq\int_r^{r+1}\frac{1}{x}\,\mathrm{d}x\leq\frac{1}{r}.$$

在图中，曲线 $y=1/x$ 在 $x=r$ 与 $x=r+1$ 之间的面积，介乎高度为 $\frac{1}{r+1}$ 的矮长方形与高度为 $\frac{1}{r}$ 的高长方形之间。



（为清楚起见，纵轴的比例已经拉长。）涂上阴影的长方形位于曲线之下；只画出外框的长方形则高于曲线。

现在把 $r = 1, 2, \ldots, n-1$ 的不等式加起来，其中 $n \geq 2$。各个积分会连接成一个：

$$\sum_{r=1}^{n-1} \frac{1}{r+1} \leq \int_1^n \frac{1}{x}\,\mathrm{d}x \leq \sum_{r=1}^{n-1} \frac{1}{r}.$$

中间的一项是 $\ln n - \ln 1 = \ln n$，而在左边的和式中，我们把 $r+1$ 改记为 r，它由 2 取到 n。所以

$$\sum_{r=2}^{n} \frac{1}{r} \leq \ln n \leq \sum_{r=1}^{n-1} \frac{1}{r},$$

这正是所要证明的结果。

集齐全套

设 X_r 为我恰好拥有 r 张不同的卡期间所打开的包数，以给出第 $r+1$ 张卡的那一包作结。集齐全套所需的总包数是

$$T = X_0 + X_1 + \cdots + X_{n-1}.$$

想一想 各个 X_r 取决于之前发生过的事。我们仍然可以直接把它们的期望值相加吗?

可以。期望值的线性性质 $\mathrm{E}(A+B) = \mathrm{E}(A) + \mathrm{E}(B)$ 完全不需要独立性。（事实上，无论之前的经过如何，X_r 都服从 $p = \frac{n-r}{n}$ 的几何分布，所以第一部分的结果适用于每一个 X_r。）因此

$$\mathrm{E}(T) = \sum_{r=0}^{n-1} \frac{n}{n-r} = n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{1}\right) = n\sum_{k=1}^{n} \frac{1}{k},$$

其中记 $k = n - r$。

记 $H_n = \sum_{k=1}^{n} \frac{1}{k}$。不等式的两边分别给出

$$H_n = 1 + \sum_{k=2}^{n} \frac{1}{k} \leq 1 + \ln n, \qquad H_n = \sum_{k=1}^{n-1} \frac{1}{k} + \frac{1}{n} \geq \ln n + \frac{1}{n}.$$

所以

$$n\ln n + 1 \leq \mathrm{E}(T) \leq n\ln n + n.$$

上下界之间的差距小于 n，当 n 很大时，这与 $n\ln n$ 相比微不足道（两者之比 $1/\ln n \to 0$）。所以当 n 很大时，我预计大约要吃

$$n\ln n \text{ 包。}$$

例如，100 位数学家就意味着大约 460 包 Cruncho。

验算

- $r = 0$：第一包必定是新卡，而 $n/(n-0) = 1$。$r = n-1$：只欠某一张指定的卡，它出现的概率是 $\frac{1}{n}$，所以我们预计需要 n 包，与 $n/1$ 吻合。
- $n = 2$：$\mathrm{E}(T) = 2(1+\frac{1}{2}) = 3$，而上下界分别给出 $2\ln 2 + 1 \approx 2.39$ 和 $2\ln 2 + 2 \approx 3.39$。
- 以 $n = 20$ 模拟 10^5 次完整的收集过程，平均需要 71.9 包，而 $20H_{20} \approx 71.95$；上下界是 60.9 和 79.9。

常见错误

- 由 $r = 0$ 开始把积分不等式相加：界 $\frac{1}{r}$ 在该处不成立，而且 $\int_0^1 \frac{1}{x}\,\mathrm{d}x$ 发散（diverge）。求和必须由 $r = 1$ 开始。
- 以为各个 X_r 必须互相独立，才可以把期望值相加。线性性质并不需要独立性。
- 答 $\ln n$ 或 n：包数的期望值的增长速度与 $n\ln n$ 相若，比 n 快。

STEP 之外

差 $H_n - \ln n$ 递减至一个极限，即欧拉－马斯刻若尼常数（Euler–Mascheroni constant）$\gamma \approx 0.5772$，所以更精确的答案是 $\mathrm{E}(T) \approx n(\ln n + \gamma)$；当 $n = 100$ 时，这给出大约 519 包，而不是 460 包。像这样把和与积分作比较，就是级数的积分审敛法（integral test），在大学第一门分析课程的初段就会遇到。优惠券收集问题本身亦会在算法分析中再次出现，也会在研究要随机洗牌多少次才能把一副牌洗匀时出现。Grimmett 与 Welsh 合著的 *Probability: An Introduction*（概率入门教科书）是一个好去处，可以看到几何等待时间与期望值的线性性质如何一并运用。

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。