详细解答

STEP II 2014，第 13 题

概率与统计 • 离散随机变量

STEP II 2014，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/14-s2-q13

关键思路

这其实是乔装了的生日问题（birthday problem）。事件 $X \geq k$ 不过是说首 $k-1$ 个整数互不相同，而 $X = r$ 则是说首 $r-1$ 个互不相同，而第 r 个与其中一个相同。两个概率（probability）都是一些简单因子的乘积。至于那两条奇特的恒等式，则来自以两种方法计算同一个量：第 (i) 部分是总概率等于 1，第 (iv) 部分是 $\mathrm{E}(X)$ 的两条公式相等。

解答

全文中记

$$Q_m = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{m-1}{n}\right) \qquad (m \geq 1),$$

其中 $Q_1 = 1$（空积（empty product））。由于每个整数都与之前的整数独立（independent），并在 $\{1, \ldots, n\}$ 上均匀分布（uniform），Q_m 就是 $I_1, \ldots, I_m$ 互不相同的概率：已知首 j 个互不相同，下一个避开它们全部的概率是 $1 - \frac{j}{n}$。

$\mathrm{P}(X = 4)$ 的值

我们需要 I_1, I_2, I_3 互不相同，然后 I_4 等于这三个之中的一个：

$$\mathrm{P}(X = 4) = 1 \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdot \frac{3}{n} = \frac{3(n-1)(n-2)}{n^3}.$$

第 (i) 部分

同样地，$X = r$ 表示 $I_1, \ldots, I_{r-1}$ 互不相同（概率为 Q_{r-1}），而 I_r 是这 $r-1$ 个值之一（概率为 $\frac{r-1}{n}$，且独立）。所以

$$\mathrm{P}(X = r) = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{r-2}{n}\right)\frac{r-1}{n} = Q_{r-1}\,\frac{r-1}{n}, \qquad 2 \le r \le n+1.$$

想一想 为甚么这些概率加起来等于 1?

X 必定存在，并且介乎 2 与 $n+1$ 之间：第一个整数不可能是重复，所以 $X \ge 2$；而 $I_1, \ldots, I_{n+1}$ 共有 $n+1$ 个整数，却只取 n 个值，所以由抽屉原理（pigeonhole principle），到第 $n+1$ 个时必有某个值出现了两次，故 $X \le n+1$。（数列永不重复的机会是零，所以没有遗漏任何情况。）因此

$$\sum_{r=2}^{n+1} \mathrm{P}(X = r) = 1.$$

把 $r = 2, 3, 4, \ldots$ 的项逐一写出，得

$$\frac{1}{n} + \left(1 - \frac{1}{n}\right)\frac{2}{n} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\frac{3}{n} + \cdots = 1,$$

这正是所求的恒等式。级数（series）在 $r = n+1$ 终止，之后的任何项都是零，因为它们含有因子 $1 - \frac{n}{n}$。证毕。令 $m = r - 1$，上式可写成

$$\sum_{m=1}^{n} \frac{m}{n}\,Q_m = 1. \qquad (*)$$

第 (ii) 部分

由期望值（expectation）的定义，

$$\mathrm{E}(X) = \sum_{r=2}^{n+1} r\,\frac{r-1}{n}\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{r-2}{n}\right) = \sum_{m=1}^{n} \frac{m(m+1)}{n}\,Q_m.$$

第 (iii) 部分

$X \ge k$ 恰好在 $I_1, \ldots, I_{k-1}$ 之中没有重复时发生，即这 $k-1$ 个整数互不相同。所以

$$\mathrm{P}(X \ge k) = Q_{k-1} = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-2}{n}\right), \qquad 2 \le k \le n+2,$$

而 $\mathrm{P}(X \ge 1) = \mathrm{P}(X \ge 2) = 1$。（当 $k = n+2$ 时，乘积含有 $1 - \frac{n}{n} = 0$，理应如此。）

第 (iv) 部分

记 $p_r = \mathrm{P}(Y = r)$。由于 $r = \sum_{k=1}^{r} 1$，所以

$$\mathrm{E}(Y) = \sum_{r=1}^{N} r\, p_r = \sum_{r=1}^{N} \sum_{k=1}^{r} p_r = \sum_{k=1}^{N} \sum_{r=k}^{N} p_r = \sum_{k=1}^{N} \mathrm{P}(Y \geq k).$$

由于是有限和，交换求和次序并无问题：两个二重和都遍历所有满足 $1 \leq k \leq r \leq N$ 的数对 (k, r)。证毕。

> **想一想** 我们应该比较 $\mathrm{E}(X)$ 的哪两个表达式?

把这结果应用于 $Y = X$，它在 $\{1, \ldots, N\}$ 中取值，其中 $N = n+1$（取值 1 的概率只是 0 而已）。利用第 (iii) 部分，

$$\mathrm{E}(X) = \sum_{k=1}^{n+1} \mathrm{P}(X \geq k) = 1 + \sum_{k=2}^{n+1} Q_{k-1} = 1 + \sum_{m=1}^{n} Q_m.$$

现在令这式等于第 (ii) 部分的结果，把 $\frac{m(m+1)}{n} = \frac{m^2}{n} + \frac{m}{n}$ 拆开，并利用 $(*)$：

$$\sum_{m=1}^{n} \frac{m(m+1)}{n} Q_m = \sum_{m=1}^{n} \frac{m^2}{n} Q_m + \underbrace{\sum_{m=1}^{n} \frac{m}{n} Q_m}_{=1 \text{ 由}(*)}.$$

所以 $1 + \sum_{m=1}^{n} Q_m = \sum_{m=1}^{n} \frac{m^2}{n} Q_m + 1$，即

$$\sum_{m=1}^{n} Q_m \left(1 - \frac{m^2}{n}\right) = 0.$$

$m = 1, 2, 3$ 的项分别是 $\left(1 - \frac{1^2}{n}\right)$、$\left(1 - \frac{1}{n}\right)\left(1 - \frac{2^2}{n}\right)$ 和 $\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3^2}{n}\right)$，所以这正是所求的恒等式。与第 (i) 部分一样，$m = n$ 之后的每一项都含有因子 $1 - \frac{n}{n} = 0$。证毕。

验算

- 当 $n = 1$ 时，$X = 2$ 必然成立。第 (i) 部分给出 $\frac{1}{1} = 1$，而最后的和只有一项 $1 - 1 = 0$。
- 当 $n = 4$ 时，最后的和是 $\frac{3}{4} + \frac{3}{4} \cdot 0 + \frac{3}{4} \cdot \frac{1}{2} \cdot \left(-\frac{5}{4}\right) + \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{4} \cdot (-3) = \frac{24-15-9}{32} = 0$，而两条公式都给出 $\mathrm{E}(X) = \frac{103}{32}$。
- 当 $n = 365$（生日）时，$\mathrm{E}(X) = 1 + \sum Q_m \approx 24.62$。以电脑模拟 10^5 个数列，得平均值 24.61；而对 $n \leq 5$ 作完全穷举（exact enumeration），亦证实了上述每一条公式。

常见错误

- 写 $\mathrm{P}(X = r)$ 时用了 $r-1$ 个形如 $1 - \frac{j}{n}$ 的因子，而不是 $r-2$ 个。只有第二至第 $(r-1)$ 个整数须避开之前的值。
- 在 $\sum \mathrm{P}(X \geq k)$ 中漏掉 $k = 1$ 的项。这一项贡献 1，正好与 $(*)$ 互相抵消。

STEP 之外

尾和公式（tail-sum formula）$\mathrm{E}(Y) = \sum_k \mathrm{P}(Y \geq k)$ 对任何取非负整数值的随机变量（random variable）都成立，而它的连续版本 $\mathrm{E}(T) = \int_0^\infty \mathrm{P}(T > t)\,\mathrm{d}t$ 是大学一年级概率论的标准工具。当 n 很大时，$\mathrm{E}(X) \approx \sqrt{\pi n/2}$，这解释了为何杂凑函数（hash function）的碰撞（collision）只需约 $\sqrt{n}$ 次尝试便会出现（即“生日攻击”（birthday attack））。Grimmett 与 Welsh 的 *Probability: An Introduction*（《概率导论》）涵盖了这两个概念。

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。