详细解答

STEP I 2012，第 13 题

概率与统计 · 以指示变量求期望值

STEP I 2012，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/12-s1-q13

关键思路

首先要把条件样本空间（conditional sample space）弄清楚：一旦知道哪些数字被排除，这个数就只是一串五个数字，每个数字都从 $\{1,2,3,4,5\}$ 中选取，而每一串出现的可能性都相同。接着，与其求出“不同数字的个数”的整个分布（distribution），不如把这个个数写成指示变量（indicator variable）之和，每个可能出现的数字对应一个。期望值的线性性质（linearity of expectation）把问题化为求一个概率，而这个概率考虑其补事件（complement）便很容易求得。

解答

条件样本空间

由 10000 至 99999 的每个整数都是一串五个数字 $d_1d_2d_3d_4d_5$，其中 $d_1 \neq 0$，而每一串都恰好出现一次。条件说明每个数字都属于 $\{1,2,3,4,5\}$。由于 0 本来就已被排除，限制 $d_1 \neq 0$ 并没有增加任何约束，因此符合条件的整数恰好对应于

$$5^5 = 3125$$

串由 $\{1,2,3,4,5\}$ 中数字组成的五位数字串。

原本的选取是均匀的。以某事件为条件时，该事件内的各个结果仍然等可能，所以在给定条件下，这 3125 串中每一串的概率都是 $\frac{1}{3125}$。换句话说，这五个数字互相独立（independent），而且每个都在 $\{1,2,3,4,5\}$ 上均匀分布（uniform）：这 3125 个结果中，每一个的概率都是五个因子 $\frac{1}{5}$ 的乘积。

指示变量

设 N 为所选的数中不同数字的个数。对 $j = 1,2,3,4,5$，定义

$$I_j = \begin{cases} 1 & \text{若数字} j \text{ 至少出现一次,} \\ 0 & \text{否则。} \end{cases}$$

每个出现了的数字对总和恰好贡献 1，因此

$$N = I_1 + I_2 + I_3 + I_4 + I_5.$$

想一想 各个 I_j 并不互相独立。我们仍然可以把它们的期望值相加吗?

可以。期望值的线性性质 $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ 对任何随机变量（random variable）都成立，不论它们是否独立。因此

$$\mathrm{E}(N) = \sum_{j=1}^{5} \mathrm{E}(I_j) = \sum_{j=1}^{5} \mathrm{P}(\text{数字} j \text{ 出现}).$$

数字 j 不出现，当且仅当五个独立的数字每一个都是其余四个值之一，而每个数字发生这情况的概率都是 $\frac{4}{5}$。因此

$$\mathrm{P}(\text{数字} j \text{ 出现}) = 1 - \left(\tfrac{4}{5}\right)^5 = 1 - \frac{1024}{3125} = \frac{2101}{3125}.$$

由对称性，这个概率对 j 的五个值都相同，所以

$$\mathrm{E}(N) = 5 \cdot \frac{2101}{3125} = \frac{2101}{625} = 3.3616,$$

即为所求。

另一方法：完整分布

如果你较喜欢直接计数，恰好使用 k 个不同数字的字串数目，等于 $\binom{5}{k}$（选出那些数字）乘以填满五个位置而使所选的 k 个数字全部出现的方法数。后者可由容斥原理（inclusion–exclusion）求得。例如当 $k = 3$ 时，它等于 $3^5 - 3 \cdot 2^5 + 3 \cdot 1^5 = 150$。

k	1	2	3	4	5	总数
字串数	$5 \cdot 1$	$10 \cdot 30$	$10 \cdot 150$	$5 \cdot 240$	$1 \cdot 120$	
	5	300	1500	1200	120	3125

于是

$$\mathrm{E}(N) = \frac{1 \cdot 5 + 2 \cdot 300 + 3 \cdot 1500 + 4 \cdot 1200 + 5 \cdot 120}{3125} = \frac{10505}{3125} = 3.3616.$$

这方法行得通，但较长，出错的地方也较多。指示变量的方法只需要求一个概率。

验算

- 表中各数目之和为 $3125 = 5^5$，证实每一串都恰好被计算了一次。
- 答案合理：N 介乎 1 与 5 之间，而从五个值中抽取五次，预期会有相当多重复，因此一个略大于 3 的值是可信的。作为一般性的验算，若从 m 个值中抽取 n 次，同样的论证给出 $m\left(1-\left(1-\frac{1}{m}\right)^n\right)$，当 $n=1$ 时它等于 1，正如所料。
- 用电脑列举由 10000 至 99999 的所有整数，保留只使用数字 1–5 的那 3125 个，再求不同数字个数的平均值，得到的恰好是 $\frac{2101}{625} = 3.3616$。

常见错误

- 把样本空间当作全部 90000 个整数。条件把它缩减为 5^5 串等可能的字串。
- 以为期望值的线性性质需要独立性。其实不需要，而这正是指示变量如此有用的原因。
- 以对五个位置各加 $\frac{1}{5}$ 的方式计算 P(数字j 出现)，结果得出 1，因为这样会把 j 出现多于一次的字串重复计算了多次。应该使用补事件。

STEP 之外

这是经典的“占位”(occupancy)问题,又称集券问题(coupon-collector problem):把 n 个球随机投进 m 个盒子,N 为最终有球的盒子数目。指示变量的技巧 $\mathrm{E}(N) = m\left(1-\left(1-\frac{1}{m}\right)^n\right)$ 是大学一年级概率论中，以至分析杂凑（hashing）等算法时最有用的工具之一。精确的分布牵涉第二类斯特灵数（Stirling numbers of the second kind），它们已在上面的容斥计数中出现过。Blitzstein 与 Hwang 的 *Introduction to Probability*（《概率导论》）全书贯彻使用指示随机变量，是绝佳的下一步读物。

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。