解答例

STEP I 2012、第 13 問

確率・統計 ・ 指示変数を用いた期待値

> **STEP I 2012、問 13**
>
> 先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：
>
> step.maths.org/questions/12-s1-q13

鍵となる考え方

まず、条件付き標本空間（conditional sample space）を正しく把握する。どの数字が除外されるかがわかれば、この数は $\{1,2,3,4,5\}$ から選んだ数字を 5 個並べた列にすぎず、そのような列はどれも同様に確からしい。次に、異なる数字の個数の分布（distribution）全体を求めるのではなく、その個数を、現れうる数字ごとに 1 つずつ用意した指示変数（indicator variable）の和として表す。期待値の線形性（linearity of expectation）により、問題はただ 1 つの確率を求めることに帰着し、その確率は余事象（complement）を考えれば容易に求められる。

解答

条件付き標本空間

10000 から 99999 までの整数はどれも、$d_1 \neq 0$ を満たす 5 個の数字の列 $d_1d_2d_3d_4d_5$ であり、そのような列はそれぞれちょうど 1 回ずつ現れる。条件は、すべての数字が $\{1,2,3,4,5\}$ に属するということである。0 はもともと除外されているので、制約 $d_1 \neq 0$ は何も付け加えない。したがって、条件を満たす整数は、$\{1,2,3,4,5\}$ の数字からなる 5 桁の列と過不足なく対応し、その個数は

$$5^5 = 3125$$

である。

もとの選び方は一様（uniform）であった。ある事象で条件付けても、その事象に含まれる結果は同様に確からしいままである。よって、条件のもとでこれら 3125 個の列はそれぞれ確率 $\frac{1}{3125}$ をもつ。言い換えれば、5 個の数字は独立（independent）であり、

それぞれが $\{1,2,3,4,5\}$ 上の一様分布に従う。実際、3125 個の結果はどれも $\frac{1}{5}$ を 5 個掛けたものである。

指示変数

選ばれた数に含まれる異なる数字の個数を N とする。$j = 1,2,3,4,5$ に対して

$$I_j = \begin{cases} 1 & \text{数字} j \text{が少なくとも 1 回現れるとき,} \\ 0 & \text{それ以外のとき.} \end{cases}$$

と定める。現れる数字はそれぞれ和にちょうど 1 だけ寄与するので、

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

となる。

> **考えてみよう** I_j は独立ではない。それでもそれらの期待値を足し合わせてよいのだろうか？

よい。期待値の線形性 $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ は、独立であるかどうかにかかわらず、任意の確率変数（random variable）について成り立つ。したがって

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

である。

数字 j が現れないのは、独立な 5 個の数字のそれぞれが残りの 4 つの値のいずれかであるときであり、またそのときに限る。各数字についてその確率は $\frac{4}{5}$ である。よって

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

である。対称性（symmetry）により、これは j の 5 つの値すべてについて同じであるから、

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

となり、示すべきことが得られた。

別解：分布全体を求める

直接数え上げたい場合は、ちょうど k 個の異なる数字を用いる列の個数は、$\binom{5}{k}$（数字の選び方）と、選んだ k 個の数字がすべて現れるように 5 つの位置を埋める方法の数との積である。後者の個数は包除原理（inclusion–exclusion）から得られる。たとえば $k = 3$ のときは $3^5 - 3\cdot 2^5 + 3\cdot 1^5 = 150$ である。

k	1	2	3	4	5	合計
列の個数	$5\cdot 1$	$10\cdot 30$	$10\cdot 150$	$5\cdot 240$	$1\cdot 120$	
	5	300	1500	1200	120	3125

すると

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

となる。この方法でも答えは得られるが、手間がかかり、間違えやすい箇所も多い。指示変数を用いる方法では、確率を 1 つ求めるだけで済む。

検算

- 表の個数を合計すると $3125=5^5$ となり、どの列もちょうど 1 回ずつ数えられていることが確認できる。
- 答えは妥当である。N は 1 以上 5 以下であり、5 つの値から 5 回選ぶのだから、かなりの重複が予想される。したがって 3 を少し上回る値はもっともらしい。一般的な検算として、m 個の値から n 回選ぶ場合、同じ議論により $m\left(1-\left(1-\frac{1}{m}\right)^n\right)$ が得られ、これは $n=1$ のとき当然そうあるべきように 1 となる。
- 10000 から 99999 までのすべての整数をコンピュータで列挙し、数字 1–5 のみを用いる 3125 個を残して異なる数字の個数を平均すると、ちょうど $\frac{2101}{625}=3.3616$ となる。

よくある間違い

- 標本空間を 90000 個すべての整数とみなしてしまう。条件によって、標本空間は同様に確からしい 5^5 個の列に縮小される。
- 期待値の線形性には独立性が必要だと考えてしまう。必要ではない。まさにそのことが、指示変数がこれほど役に立つ理由である。
- P(数字jが現れる) を、5 つの位置それぞれについて $\frac{1}{5}$ を足して計算してしまう。こうすると 1 になるが、それは j が 2 回以上現れる列を何重にも数えているからである。余事象を用いること。

STEP のその先へ

これは古典的な「占有問題（occupancy problem）」、あるいはクーポン収集問題（coupon-collector problem）である。n 個の球を m 個の箱に無作為に投げ入れ、最終的に球が入っている箱の個数を N とする。指示変数を用いる技法による $\mathrm{E}(N) = m\left(1 - (1 - \frac{1}{m})^n\right)$ は、大学 1 年次の確率論や、ハッシュ法（hashing）などのアルゴリズムの解析において、最も有用な道具の 1 つである。正確な分布には第 2 種スターリング数（Stirling numbers of the second kind）が関わり、これは上の包除原理による数え上げにも現れていた。Blitzstein and Hwang, *Introduction to Probability*（確率論入門）は全編を通じて指示確率変数を用いており、次の一歩として最適である。

Beyond STEP 解答集：https://beyondstep.keithmcnulty.org/ja/。解答本文は CC0 1.0 である。問題そのものは © Cambridge University Press & Assessment に帰属し、ここには掲載していない。英語版から機械翻訳したものであり、まだ校閲を経ていない。