解答例

STEP II 2014、第 13 問

確率・統計 ・ 離散確率変数

> **STEP II 2014、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/14-s2-q13

鍵となる考え方

これは姿を変えた誕生日問題（birthday problem）である。事象 $X \geq k$ は、最初の $k-1$ 個の整数がすべて異なることにほかならず、$X = r$ は、最初の $r-1$ 個が互いに異なり、r 番目がそのうちのどれかと一致することを意味する。どちらの確率も簡単な因数の積で表される。そして二つの奇妙な恒等式は、同じ量を二通りに計算することから得られる。小問 (i) では確率の総和が 1 に等しいことを、小問 (iv) では $\mathrm{E}(X)$ を表す二つの式が一致することを用いる。

解答

以下を通じて

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

とおき、$Q_1 = 1$（空積（empty product））とする。各整数はそれ以前の整数と独立（independent）で、$\{1, \ldots, n\}$ 上の一様分布（uniform distribution）に従うから、Q_m は $I_1, \ldots, I_m$ がすべて異なる確率である。実際、最初の j 個が互いに異なるとき、次の整数がそのいずれとも一致しない確率は $1 - \frac{j}{n}$ である。

$\mathrm{P}(X = 4)$ の値

I_1, I_2, I_3 が互いに異なり、さらに I_4 がその三つのうちどれかに等しければよい。

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

小問 (i)

同様に、$X = r$ は、$I_1, \ldots, I_{r-1}$ が互いに異なり（確率 Q_{r-1}）、かつ I_r がこれら $r-1$ 個の値のどれかである（独立に確率 $\frac{r-1}{n}$）ことを意味する。よって

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

考えてみよう なぜこれらの確率の和は 1 になるのか。

X は必ず存在し、2 以上 $n+1$ 以下である。最初の整数が重複になることはありえないので $X \ge 2$ である。また、$I_1, \ldots, I_{n+1}$ は n 通りの値しかとらない $n+1$ 個の整数であるから、鳩の巣原理（pigeonhole principle）により、時刻 $n+1$ までにある値が 2 回現れている。よって $X \le n+1$ である。（数列がいつまでも重複しない確率は 0 なので、取りこぼしはない。）したがって

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

$r = 2, 3, 4, \ldots$ の項を書き並べると

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

となり、これが求める恒等式である。級数は $r = n+1$ で終わり、それより先の項は因数 $1 - \frac{n}{n}$ を含むので 0 である。以上で示された。$m = r - 1$ とおくと、これは

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

と書ける。

小問 (ii)

期待値（expectation）の定義より

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

小問 (iii)

$X \ge k$ が起こるのは、$I_1, \ldots, I_{k-1}$ の中に重複がないとき、すなわちこれら $k-1$ 個の整数がすべて異なるとき、かつそのときに限る。よって

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

であり、$\mathrm{P}(X \ge 1) = \mathrm{P}(X \ge 2) = 1$ である。（$k = n+2$ のとき、積は $1 - \frac{n}{n} = 0$ を含み、当然そうなるべき値になっている。）

小問 (iv)

$p_r = \mathrm{P}(Y = r)$ とおく。$r = \sum_{k=1}^{r} 1$ であるから

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

和は有限なので、和をとる順序を入れ替えてよい。どちらの二重和も、$1 \leq k \leq r \leq N$ を満たすすべての組 (k, r) にわたる和である。以上で示された。

> **考えてみよう** $\mathrm{E}(X)$ のどの二つの表し方を比べればよいか。

この結果を $Y = X$ に適用する。X は $N = n+1$ として $\{1, \ldots, N\}$ に値をとる（値 1 の確率が 0 であるだけである）。小問 (iii) を用いると

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

これを小問 (ii) の結果と等しいとおき、$\frac{m(m+1)}{n} = \frac{m^2}{n} + \frac{m}{n}$ と分けて $(*)$ を用いる。

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

よって $1 + \sum_{m=1}^{n} Q_m = \sum_{m=1}^{n} \frac{m^2}{n} Q_m + 1$、すなわち

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

$m = 1, 2, 3$ の項は $\left(1 - \frac{1^2}{n}\right)$、$\left(1 - \frac{1}{n}\right)\left(1 - \frac{2^2}{n}\right)$、$\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3^2}{n}\right)$ であるから、これはまさに求める恒等式である。小問 (i) と同様に、$m = n$ より先の項はすべて因数 $1 - \frac{n}{n} = 0$ をもつ。以上で示された。

検算

- $n = 1$ のとき、確実に $X = 2$ である。小問 (i) は $\frac{1}{1} = 1$ を与え、最後の和はただ一つの項 $1 - 1 = 0$ となる。
- $n = 4$ のとき、最後の和は $\frac{3}{4} + \frac{3}{4}\cdot 0 + \frac{3}{4}\cdot\frac{1}{2}\cdot\left(-\frac{5}{4}\right) + \frac{3}{4}\cdot\frac{1}{2}\cdot\frac{1}{4}\cdot(-3) = \frac{24-15-9}{32} = 0$ であり、どちらの式からも $\mathrm{E}(X) = \frac{103}{32}$ が得られる。
- $n = 365$（誕生日）のとき、$\mathrm{E}(X) = 1 + \sum Q_m \approx 24.62$ である。10^5 個の数列によるコンピュータシミュレーションでは平均 24.61 が得られ、$n \leq 5$ に対する厳密な数え上げにより、上のすべての式が確認された。

よくある間違い

- $\mathrm{P}(X = r)$ を、$1 - \frac{i}{n}$ の形の因数を $r-2$ 個ではなく $r-1$ 個として書いてしまう。それ以前の値を避けなければならないのは、2 番目から $(r-1)$ 番目までの整数だけである。
- $\sum \mathrm{P}(X \geq k)$ の $k = 1$ の項を忘れる。この項は 1 を寄与し、それがちょうど $(*)$ と打ち消し合う。

STEP のその先へ

裾和公式（tail-sum formula）$\mathrm{E}(Y) = \sum_k \mathrm{P}(Y \geq k)$ は、非負整数値をとる任意の確率変数（random variable）に対して成り立ち、その連続版 $\mathrm{E}(T) = \int_0^\infty \mathrm{P}(T > t)\,\mathrm{d}t$ は大学 1 年次の確率論における標準的な道具である。n が大きいとき $\mathrm{E}(X) \approx \sqrt{\pi n/2}$ であり、これがハッシュ関数（hash function）の衝突がわずか $\sqrt{n}$ 回程度の試行で現れる理由である（「誕生日攻撃（birthday attack）」）。Grimmett と Welsh の *Probability: An Introduction*（確率論入門）は、この両方の考え方を扱っている。

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。