解答例

STEP III 1997、第 12 問

確率・統計 ・ クーポン収集問題

STEP III 1997、問 12

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/97-s3-q12

鍵となる考え方

小問 (i) は幾何分布（geometric distribution）であり、その確率母関数（probability generating function）を微分すれば平均（mean）と分散（variance）が得られる：$\mathrm{E}(K) = \mathrm{G}'(1)$ および $\mathrm{E}(K(K-1)) = \mathrm{G}''(1)$ である。小問 (ii) での工夫は、全種類がそろうまでの長い待ち時間を、新しいカード 1 枚ごとの 52 個の短い待ち時間に分割することである。それぞれの短い待ち時間は「初めて表が出るまで硬貨を投げる」試行なので、あとは小問 (i) の結果がすべてを片づけてくれる。

解答

小問 (i)

初めて表が出るのが k 回目の投げであるのは、最初の $k-1$ 回がすべて裏で、k 回目が表であるとき、かつそのときに限る。各回の投げは独立（independent）なので、

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

したがって

$$\mathrm{G}(s) = \sum_{k=1}^{\infty} pq^{k-1}s^k = ps\sum_{j=0}^{\infty}(qs)^j = \frac{ps}{1-qs},$$

となる。ここで等比級数（geometric series）の和を求めたが、これは $|qs| < 1$ のとき収束（converge）する（特に $p > 0$ ならば s が 1 に近いとき収束する）。これが求める結果である。

微分すると、

$$G'(s) = \frac{p(1-qs)+psq}{(1-qs)^2} = \frac{p}{(1-qs)^2},$$
$$G''(s) = \frac{2pq}{(1-qs)^3}.$$

$G'(s) = \sum k\,P(K=k)s^{k-1}$ かつ $G''(s) = \sum k(k-1)P(K=k)s^{k-2}$ であるから、$s=1$ を代入し $1-q=p$ を用いると

$$E(K) = G'(1) = \frac{p}{p^2} = \frac{1}{p}, \qquad E(K(K-1)) = G''(1) = \frac{2pq}{p^3} = \frac{2q}{p^2}.$$

を得る。よって

$$\begin{aligned} \mathrm{Var}(K) &= E(K(K-1)) + E(K) - (E(K))^2 \\ &= \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q+p-1}{p^2} = \frac{q}{p^2}. \end{aligned}$$

以上より、$E(K) = \dfrac{1}{p}$、$\mathrm{Var}(K) = \dfrac{q}{p^2}$ である。

小問 (ii)

ちょうど $i-1$ 種類の異なるカードを見た時点以降、i 種類目の異なるカードが出る回までを含めた引く回数を N_i とする。すると $N_1 = 1$（最初のカードは必ず新しい）であり、

$$N = N_1 + N_2 + \cdots + N_{52}.$$

となる。

> **考えてみよう** なぜ各 N_i は K と同じような「初めて表が出る」型の確率変数なのだろうか。

$i-1$ 種類の異なるカードを見た時点で、52 枚のうち $53-i$ 枚はまだ見ていない。復元抽出（drawing with replacement）なので、各回の抽出は過去と独立であり、まだ見ていないカードが出る確率はつねに同じ

$$p_i = \frac{53-i}{52}.$$

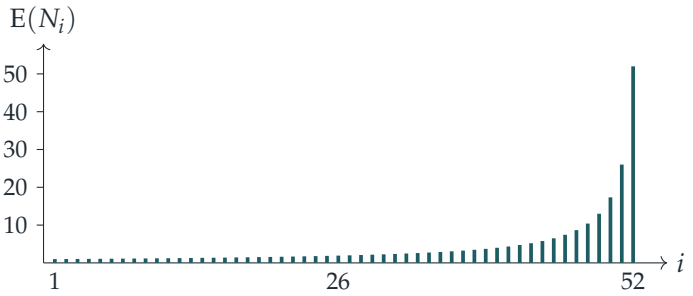
である。このような抽出を「表」と呼ぶことにする。すると N_i は初めて表が出る回までを含めた抽出回数であり、これはまさに $p = p_i$ としたときの小問 (i) の確率変数 K である。したがって

$$E(N_i) = \frac{1}{p_i} = \frac{52}{53-i}.$$

期待値（expectation）は線形である（このためには N_i たちの独立性は必要ない）から、

$$\begin{aligned}\mathrm{E}(N) &= \sum_{i=1}^{52} \frac{52}{53-i} = 52\left(\frac{1}{52}+\frac{1}{51}+\cdots+\frac{1}{2}+\frac{1}{1}\right)\\ &= 52\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{52}\right).\end{aligned}$$

初めのうちの待ち時間は短く、最後の数枚が大部分を占める。最後の 1 枚だけで平均 52 回の抽出を要するのである。



概算するために、$n = 52$ として $1+\frac{1}{2}+\cdots+\frac{1}{n} \approx 0.5+\ln n$ を用いる。$\ln 52$ については、$2.7^2 = 7.29$、$7.29^2 \approx 53.1$ より $\mathrm{e}^4 \approx 53 \approx 52$ なので、$\ln 52 \approx 4$ である。したがって

$$\mathrm{E}(N) \approx 52\,(0.5+4) = 52\times 4.5 = 234.$$

つまり、すべてのカードを見るには平均しておよそ 234 回、カードの枚数の約 4.5 倍の抽出が必要である。

検算

- (i) において $\mathrm{G}(1) = p/(1-q) = 1$ であり、確率母関数として当然そうなるべき値になっている。公平な硬貨（$p = \frac{1}{2}$）では $\mathrm{E}(K) = 2$、$\mathrm{Var}(K) = 2$ となる。また $p \to 1$ のとき分散 $q/p^2 \to 0$ となり、最初の投げでほぼ確実に表が出る場合として妥当である。
- 正確な値は $52H_{52} \approx 235.98$ であり、全種類収集を 10^5 回シミュレーションしたところ平均は 235.7 であった。我々の概算値 234 はこれに近い。わずかに小さいのは、近似 $0.5+\ln n$（より精密な近似は $\ln n + 0.577$）と、$\ln 52 \approx 3.95$ を 4 に切り上げたことによる。

よくある間違い

- N_i の成功確率を逆にしてしまうこと。**新しい**カードが出る確率は $(53-i)/52$ であり、集めたカードが増えるにつれて小さくなる。
- $+\mathrm{G}'(1)$ の項を忘れて $\mathrm{Var}(K) = \mathrm{G}''(1) - \mathrm{G}'(1)^2$ としてしまうこと。$\mathrm{G}''(1)$ は $\mathrm{E}(K(K-1))$ であり、$\mathrm{E}(K^2)$ ではない。
- 期待値を足し合わせる前に N_i たちが独立でなければならないと考えること。期待値の線形性（linearity of expectation）は独立性に関係なく成り立つ（ただし、ここでは実際に独立である）。

STEP のその先へ

これは**クーポン収集問題**（coupon collector's problem）である。n 種類が等確率で出るとき、すべてを集めるまでの時間の期待値は $nH_n \approx n\ln n$ である。大学の講義ではさらに進んで、この時間が $n\ln n$ の周りに鋭く集中し、ゆらぎは n のオーダーであることを示す。また確率母関数は、このような独立な確率変数の和を扱う標準的な道具となる。Grimmett and Welsh, *Probability: An Introduction*（確率論入門）は母関数と幾何分布を丁寧に扱っている。Blitzstein and Hwang, *Introduction to Probability*（確率論入門）はクーポン収集問題をさまざまな変形とともに論じている。

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。