9.11 全種類を集める

ワークシート

問題

STEP II 1994、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/94-s2-q13

ヒント

まずは自力で問題に取り組んでみよう。行き詰まったら、ヒントを**一つずつ**読み、そのたびに問題に戻ること。**手を止めて、やってみよう**と記したヒントが重要な段階である。各ヒントの後には計算用の余白がある。

> ? **考えてみよう**　各袋がそれぞれ独立に確率 p で新しいカードをくれるとき、最初の新しいカードが出るまで（その袋を含めて）開ける袋の数の期待値（expectation）はいくらか。

段階 1　新しいカードを待つ

扱う内容：新しいカードを待つこと

> ? **考えてみよう**　すでに r 種類の異なるカードを持っているとき、次の 1 袋から新しいカードが出る確率はいくらか。幾何分布（geometric distribution）における待ち時間の期待値（expectation）を用いよ。

段階 2 和と積分の比較

扱う内容：積分を用いた不等式

考えてみよう $r \le x \le r+1$ において、$\frac{1}{x}$ はどの 2 つの値の間にあるか。長さ 1 のこの区間で積分せよ。

手を止めて、やってみよう 得られた不等式を $r=1$ から $r=n-1$ まで足し合わせよ。各積分はつながって 1 つの積分になる。それはどの積分か。

段階 3 全種類をそろえる

扱う内容：全種類を集めること

> **考えてみよう** 必要な袋の総数は、新しいカードそれぞれを得るまでの待ち時間の和である。それらの期待値を足し合わせ、第 2 段階の結果を用いてその和を $\ln n$ と比較せよ。

背後にある数学

この評価はどれほど良いか

この比較はさらに精密にできる。図 9.11.1（オンライン版の節全体を参照）において曲線と上側の長方形の間にある影の付いた隙間は、面積の合計が有限なので、$H_n - \ln n$ はある極限に収束（converge）する。

$$\gamma = \lim_{n\to\infty} (H_n - \ln n) = 0.5772156649\ldots,$$

これが**オイラーの定数**である。（γ が有理数かどうかは誰も知らない。）したがって袋の数の期待値は $n\ln n + \gamma n + \frac{1}{2} + \cdots$ である。600 種類のステッカーアルバムでは $600 \times (\ln 600 + 0.577) \approx 4185$ となり、これが導入で挙げた数値である。

時間はどこに費やされるのだろうか。各段階はきわめて不均等である。セットの前半を集めるのにかかるのはわずか約 $n\ln 2 \approx 0.69n$ 袋であるのに対し、最後の 1 枚だけで平均 n 袋かかる。労力の大部分は最後の数枚に費やされる。コレクターたちがダブったカードを交換するのはこのためである。

集中

期待値は出発点にすぎない。総数はどれくらいばらつくのだろうか。総数の分散（variance）は $\sum nr/(n-r)^2$ であり、これはおよそ $\frac{\pi^2}{6}n^2$ である（Section 2.4 の和 $\sum 1/k^2 = \pi^2/6$ が現れる）。したがって標準偏差（standard deviation）はおよそ $1.28n$

であり、n が大きいときには平均 $n \ln n$ よりもずっと小さい。つまり総数はその平均の周りに**集中**（concentrated）している。ポール・エルデシュ（Paul Erdős）とアルフレッド・レーニ（Alfréd Rényi）は 1961 年に、より精密に、総数から $n \ln n$ を引いて n で割ったものが極限分布をもつことを示した。それが**ガンベル分布**（Gumbel distribution）であり、多数の確率的な量の最大値も記述する分布である（Section 12.25）。

至るところにクーポン

この問題はアブラーム・ド・モアブル（Abraham de Moivre、1712 年）とラプラス（Laplace）にさかのぼる。現在ではさまざまな姿で現れる。システムのすべての部品を確かめるには、無作為なテストが何回必要か。ネットワーク上のランダムウォークがすべての頂点を訪れるまでどれくらいかかるか。コンピュータのハッシュ表を何回探索しなければならないか。いずれの場合も答えには $n \ln n$ が現れる。生物学では、同じ計算によって、ゲノム全体を網羅するのに必要なシーケンシングの「リード」の数が見積もられる。

次の一歩

1. **セットの半分。** $\frac{n}{2}$ 種類の異なるカードを集めるのに必要な袋の数の期待値（n は偶数）が $n(H_n - H_{n/2})$ であること、そしてこれがおよそ $n \ln 2$ であることを示せ。

2. **異なるカードの数。** k 袋を開けた後に持っている異なるカードの種類数の期待値が $n(1 - (1 - \frac{1}{n})^k)$ であることを示せ。（**ヒント**：各カードについて、それを持っている確率を求め、足し合わせよ。）セットの 95% を集めるにはおよそ何袋必要か。

3. **二人のコレクター。** 二人の子どもが同じセットを集めており、一方にとってのダブりが他方にとっては新しいカードになりうるように袋を共有する。両方のセットを完成させるのに必要な袋の数の期待値がおよそ $n(\ln n + \ln \ln n)$ であることを示せ。（これは難しい。まず、それが $2n \ln n$ よりずっと小さいことを納得せよ。）

4. **より精密な評価。** 長方形の代わりに台形公式（trapezium rule）の図を用いて、$H_n - \ln n$ が減少し、$H_n - \ln n - \frac{1}{2n}$ が増加することを示せ。小さな n の値から $0.5 < \gamma < 0.7$ を導け。

さらに学ぶために

- **Geoffrey Grimmett & David Stirzaker, *One Thousand Exercises in Probability***（Oxford University Press）。クーポン・コレクター問題のさまざまな変種が収められている。

- **Julian Havil, *Gamma: Exploring Euler's Constant***（Princeton University Press）。γ と調和級数のすべてを、親しみやすい水準で解説している。
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing***（Cambridge University Press）。クーポン・コレクター問題と、そのコンピュータ科学における応用。

詳しい解説付きの解答は第 9.11 節にある：

https://beyondstep.keithmcnulty.org/ja/ch09/coupon-collector.html

毎週 1 問の STEP の問題を丁寧に解き明かす The STEP Up。メールで無料配信中：keithmcnulty.substack.com