解答例

STEP III 1988、第 16 問

確率・統計 ・ 超幾何分布

STEP III 1988、問 16

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/88-s3-q16

鍵となる考え方

非復元抽出（drawing without replacement）では各回の取り出しが互いに従属するので、1 回ずつ順に追っていく方法は煩雑になる。これを回避する考え方が二つある。第一に、順序を考えない選び方を数えることで、赤玉の個数の分布が得られる。第二に、対称性を利用する。壺の中の玉をすべてランダムな順序で取り出し尽くすと想像すれば、どの玉も、指定されたどの位置にも等しい確率で現れる。そこで X を指示変数（indicator variable）X_i の和として書けば、分布をまったく用いずに平均（mean）と分散（variance）が求まる。必要なのは各組の間の共分散（covariance）だけである。

解答

ちょうど k 個が赤である確率

玉はすべて区別できるものと考える。取り出した n 個の玉は大きさ n の集合をなし、そのような集合はどれも同様に確からしいので、同様に確からしい結果は $\binom{M}{n}$ 通りある。赤玉をちょうど k 個含む集合は、m 個の赤玉から k 個を、$M-m$ 個の緑玉から $n-k$ 個を選ぶことで作られる。したがって

$$\mathrm{P}(\text{ちょうど}k\text{ 個が赤}) = \frac{\binom{m}{k}\binom{M-m}{n-k}}{\binom{M}{n}},$$

となる。ただし、$b<0$ または $b>a$ のとき $\binom{a}{b}=0$ とする通常の約束に従う。（これにより、$n-k>M-m$ のような起こりえない場合も扱える。）これが超幾何分布（hypergeometric distribution）である。

小問 (i)

壺が空になるまで取り出しを続けると想像する。M 個の（区別できる）玉が出てくる順序は一様ランダムな並べ方であり、$M!$ 通りの順序はすべて同様に確からしい。n 回で取り出しをやめても最初の n 個の位置は変わらないので、全体の並べ方で考えてよい。

位置 i を固定する。特定の玉が位置 i にあるような並べ方の数は $(M-1)!$ であり、これはどの玉についても同じである。よって、位置 i にある玉は M 個の玉のいずれであることも同様に確からしく、

$$\mathrm{P}(X_i = 1) = \frac{m\,(M-1)!}{M!} = \frac{m}{M},$$

となり、示すべきことが示された。

> **考えてみよう** 2 番目の玉の色は、1 番目の玉が何であったかに依存するのではないか？

1 番目の玉が与えられたもとでは、確かに依存する。しかし求めたいのは**条件なしの**確率である。$i = 2$ の場合、全確率の公式（law of total probability）により

$$\mathrm{P}(X_2 = 1) = \frac{m}{M}\cdot\frac{m-1}{M-1} + \frac{M-m}{M}\cdot\frac{m}{M-1} = \frac{m(M-1)}{M(M-1)} = \frac{m}{M}.$$

を得る。対称性による議論は、これと同じ計算をすべての i について一度に行っているのである。

小問 (ii)

位置 $i \neq j$ を固定する（したがって $M \geq 2$ である）。両方の位置に赤玉がある並べ方を数える。位置 i の玉の選び方が m 通り、次に位置 j の玉の選び方が $m-1$ 通りあり、残りの玉の並べ方が $(M-2)!$ 通りある。したがって

$$\mathrm{P}(X_i = 1 \text{ かつ} X_j = 1) = \frac{m(m-1)(M-2)!}{M!} = \frac{m(m-1)}{M(M-1)},$$

となり、示すべきことが示された。言い換えれば、位置 (i, j) にある玉の順序対は、相異なる玉の $M(M-1)$ 通りの順序対のいずれであることも同様に確からしい。

X の平均と分散

X は最初に取り出した n 個のうちの赤玉の個数である。独立性を必要としない期待値の線形性（linearity of expectation）と、$\mathrm{E}(X_i) = \mathrm{P}(X_i = 1)$ であることから、

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{nm}{M}.$$

分散については $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$ を用いる。和の 2 乗を展開し、X_i は 0 または 1 であるから $X_i^2 = X_i$ となることを用いると、

$$X^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j = X + \sum_{i \neq j} X_i X_j,$$

となる。ここで第 2 の和は、$i \neq j$ を満たす $n(n-1)$ 個の順序対 (i, j) にわたってとる。$X_i X_j = 1$ となるのは両方の玉が赤であるときに限るので、小問 (ii) より $\mathrm{E}(X_i X_j) = \dfrac{m(m-1)}{M(M-1)}$ である。したがって、$M \geq 2$ のとき

$$\begin{aligned}\mathrm{Var}(X) &= \frac{nm}{M} + \frac{n(n-1)m(m-1)}{M(M-1)} - \frac{n^2m^2}{M^2} \\ &= \frac{nm}{M^2(M-1)}\Big[M(M-1) + (n-1)(m-1)M - nm(M-1)\Big].\end{aligned}$$

括弧の中を展開すると

$$\begin{aligned}&M^2 - M + nmM - nM - mM + M - nmM + nm \\ &\quad = M^2 - nM - mM + nm = (M-m)(M-n).\end{aligned}$$

となる。よって

$$\mathrm{Var}(X) = \frac{nm(M-m)(M-n)}{M^2(M-1)}.$$

である。（$M = 1$ ならば $n \leq 1$ であり、組は存在せず、X は定数なので $\mathrm{Var}(X) = 0$ である。）

$p = m/M$ とおくと見通しがよくなる。このとき

$$\mathrm{E}(X) = np, \qquad \mathrm{Var}(X) = np(1-p) \cdot \frac{M-n}{M-1}.$$

となる。これは二項分布（binomial distribution）の分散に補正因子を掛けたものであり、この因子は $n \geq 2$ のとき 1 より小さい。

検算

- $n = M$ の場合：すべての玉を取り出すので、確実に $X = m$ である。公式は平均 m、分散 0 を与える。これは因子 $M - n$ のためである。
- $n = 1$ の場合：X は $p = m/M$ の単一のベルヌーイ変数（Bernoulli variable）であり、分散は $p(1-p)$ に帰着する。
- $M = 6$、$m = 2$、$n = 3$ の場合：公式から $\mathrm{Var}(X) = \frac{3 \cdot 2 \cdot 4 \cdot 3}{36 \cdot 5} = \frac{2}{5}$ を得る。6! 通りの順序をすべて数え上げた結果はこれと正確に一致し、$M \leq 6$ のすべての場合についても一致する。

よくある間違い

- X_i を独立とみなし、二項分布の分散 $np(1-p)$ を答えてしまう。これらは負の相関をもつ。早い段階で赤玉が出ると、後で赤玉が出にくくなるのである。
- 組を順序対として数えるとき、二重和 $\sum_{i \neq j}$ の項数は $\binom{n}{2}$ ではなく $n(n-1)$ であることを忘れる。

STEP のその先へ

個数を指示変数の和として書くことは、確率論で最も強力な技法の一つである。分布そのものを書き下すのが絶望的な場合でも、平均と分散が得られる。因子 $\frac{M-n}{M-1}$ は、標本調査で用いられる**有限母集団修正**（finite population correction）である。これを見れば、1000 人を対象とする世論調査が、一国について、無限母集団についての場合とほぼ同程度の情報をもたらす理由がわかる。Blitzstein and Hwang, *Introduction to Probability*（確率論入門）は全編を通じて指示変数を用いており、超幾何分布を詳しく扱っている。

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。