解答例

STEP III 2004、第 12 問

確率・統計 ・ 指示変数と完全順列

STEP III 2004、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/04-s3-q12

鍵となる考え方

個数を指示変数（indicator variable）の和として表すことは、確率論で最も役に立つ手法の一つである。指示変数の期待値（expectation）は確率そのものなので、$\mathrm{E}(N)$ はただちに求まる。一方、分散（variance）を求めるには各組の間の共分散（covariance）が必要であり、シャツは独立（independent）に割り当てられるわけではないから、これらは小さいが 0 ではない。N の平均と分散がともに 1 に等しいとわかれば、パラメータ 1 のポアソン分布（Poisson distribution）が、まず試すべき自然なモデルとなる。

解答

一つの指示変数 C_1

シャツの割り当て方 $m!$ 通りはすべて同様に確からしいので、シャツ 1 は m 人の選手のどの人にも等しい確率で渡る。したがって $\mathrm{P}(C_1 = 1) = \frac{1}{m}$ であり、C_1 のとる値は 0 と 1 のみであるから、

$$\mathrm{E}(C_1) = 1 \cdot \mathrm{P}(C_1 = 1) = \frac{1}{m},$$

となり、題意が示された。また $C_1^2 = C_1$（$0^2 = 0$ かつ $1^2 = 1$ だから）なので $\mathrm{E}(C_1^2) = \frac{1}{m}$ も成り立ち、

$$\mathrm{Var}(C_1) = \frac{1}{m} - \frac{1}{m^2} = \frac{m-1}{m^2}.$$

C_1 と C_2 の共分散

ここでは $m \geq 2$ が必要である。積 C_1C_2 は、選手 1 と 2 がともに自分のシャツを着ているときにちょうど 1 となり、それ以外では 0 となる。選手 1 がシャツ 1 を着ている

という条件のもとでは、残りの $m-1$ 枚のシャツが他の $m-1$ 人の選手に無作為に配られるので、選手 2 がシャツ 2 を受け取る確率は $\frac{1}{m-1}$ である。よって

$$\mathrm{E}(C_1C_2) = \mathrm{P}(C_1 = 1,\, C_2 = 1) = \frac{1}{m} \cdot \frac{1}{m-1},$$

であり、

$$\begin{aligned}\mathrm{Cov}(C_1, C_2) &= \mathrm{E}(C_1C_2) - \mathrm{E}(C_1)\,\mathrm{E}(C_2) = \frac{1}{m(m-1)} - \frac{1}{m^2} \\ &= \frac{m-(m-1)}{m^2(m-1)} = \frac{1}{m^2(m-1)}.\end{aligned}$$

これは正である。選手 1 が正しいシャツを着ているとわかれば、選手 2 にとっての候補から間違ったシャツが 1 枚取り除かれるので、その確率がわずかに上がるのである。

考えてみよう なぜ共分散は 0 にならないのか。各選手のシャツは別々に選ばれるのではないのか。

そうではない。シャツは各選手に 1 枚ずつ割り当てられるので、選手 1 が何を受け取るかが、他の全員に残るシャツに影響する。C_i は**従属**（dependent）しており、まさにそのために $\mathrm{Var}(N)$ には共分散の項を含めなければならない。

N の平均と分散

対称性（symmetry）により、どの C_i も C_1 と同じ分布をもち、$i \neq j$ であるどの組 C_i, C_j も C_1, C_2 と同じ同時分布をもつ（選手とシャツの番号を同時に付け替えても、無作為な割り当ては変わらない）。独立性を必要としない期待値の線形性（linearity of expectation）により、

$$\mathrm{E}(N) = \sum_{i=1}^{m} \mathrm{E}(C_i) = m \cdot \frac{1}{m} = 1.$$

和の分散を求めるには、すべての分散と、$i \neq j$ である順序対すべてにわたる共分散を足し合わせる。順序対は $m(m-1)$ 個ある。

$$\begin{aligned}\mathrm{Var}(N) &= \sum_{i=1}^{m} \mathrm{Var}(C_i) + \sum_{i \neq j} \mathrm{Cov}(C_i, C_j) \\ &= m \cdot \frac{m-1}{m^2} + m(m-1) \cdot \frac{1}{m^2(m-1)} \\ &= \frac{m-1}{m} + \frac{1}{m} = 1.\end{aligned}$$

よって $\mathrm{E}(N) = \mathrm{Var}(N) = 1$ となり、題意が示された。（$m = 1$ のときは常に $N = 1$ となるので $\mathrm{Var}(N) = 0$ である。共分散の計算ですでに必要だったように、この問題は暗に $m \geq 2$ を仮定している。）

どの近似を用いるか

正規近似（Normal approximation）は、m の値にかかわらず不適切である。N がとるのは整数値 $0, 1, 2, \ldots$ のみであり、平均が 1、標準偏差（standard deviation）が 1 なので、確率のほぼすべてが $0, 1, 2, 3$ の 3 つか 4 つの値に集中している。分布は平均よりわずか標準偏差 1 つ分下の 0 で打ち切られているので、右に強く歪んで（skewed）いる。一方、$\mathrm{N}(1,1)$ 分布は対称であり、N が決してとりえない $-\frac{1}{2}$ より下の範囲に、連続性補正（continuity correction）を行ってもなお無視できない確率（約 7%）を割り当ててしまう。決定的なのは、m を大きくしても改善されないことである。平均と分散はどの m に対しても 1 のままなので、分布の形が広がって釣鐘型になることは決してない。

これに対し、ポアソン近似は妥当である。N は多数の m 個の事象のうち「成功」したものの個数を数えており、各事象の確率は小さな値 $\frac{1}{m}$ であり、しかもこれらの事象の従属性は弱い（共分散 $\frac{1}{m^2(m-1)}$ はごく小さい）。これはまさに、個数が近似的にポアソン分布に従う状況である。さらに、ポアソン分布に従う確率変数は平均と分散が等しく、N はちょうどこの性質をもつので、$\mathrm{Po}(1)$ が自然な選択となる。

$m = 4$ の場合

同様に確からしい割り当ては $4! = 24$ 通りある。割り当てを選手 $1, 2, 3, 4$ が着るシャツの番号の列として書くと、誰も自分のシャツを着ていないものは

$$2143,\quad 2341,\quad 2413,\quad 3142,\quad 3412,$$
$$3421,\quad 4123,\quad 4312,\quad 4321.$$

である。（系統的に列挙するには、選手 1 のシャツが 2、3、4 のいずれかであり、それぞれの場合について残りの決め方がちょうど 3 通りあるので、$3 \times 3 = 9$ となる。）したがって

$$\mathrm{P}(N = 0) = \frac{9}{24} = \frac{3}{8} = 0.375.$$

パラメータ 1 のポアソン近似では $\mathrm{P}(N = 0) \approx \mathrm{e}^{-1}$ となる。$\mathrm{e} \approx 2.72$ を用いると、

$$\mathrm{e}^{-1} \approx \frac{100}{272} = \frac{25}{68}.$$

真の値を基準とした相対誤差（relative error）は

$$\frac{\frac{3}{8} - \frac{25}{68}}{\frac{3}{8}} = 1 - \frac{25}{68} \cdot \frac{8}{3} = 1 - \frac{200}{204} = \frac{1}{51} \approx 0.0196,$$

であり、主張どおり約 2% である。（代わりに近似値を基準として誤差を測ると $\frac{1}{50}$、ちょうど 2% となる。いずれにせよ、これほど小さな m でこの一致は驚くほどよい。）

検算

- $m = 2$ のとき、起こりうるのは両方とも正しい（$N = 2$）か両方とも間違っている（$N = 0$）かのみで、それぞれ確率 $\frac{1}{2}$ である。平均は 1、分散は 1 となり、正しい。
- $m = 2, \ldots, 7$ についてすべての置換（permutation）を計算機で数え上げたところ、$\mathrm{E}(C_1) = \frac{1}{m}$、$\mathrm{Var}(C_1) = \frac{m-1}{m^2}$、$\mathrm{Cov}(C_1, C_2) = \frac{1}{m^2(m-1)}$、$\mathrm{E}(N) = \mathrm{Var}(N) = 1$ が厳密に成り立つことが確認された。また 4 個の対象の完全順列（derangement）は 9 個見つかり、$\mathrm{P}(N = 0) = 0.375$ と $1/2.72 \approx 0.3676$ の比較となった。
- $N = m - 1$ は起こりえないことに注意せよ。1 人を除く全員が正しいシャツを着ていれば、最後の 1 人も正しいシャツを着ていることになる。これも N が正規分布からほど遠いことを示すしるしである。

よくある間違い

- C_i を独立として扱うこと。そうすると 1 ではなく $\mathrm{Var}(N) = \frac{m-1}{m}$ となってしまう。
- 共分散の項を $m(m-1)$ 個ではなく $\binom{m}{2}$ 個しか数えないこと。すなわち $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$ における係数 2 を忘れること。
- 完全順列を不注意に列挙すること。系統的に数える（3×3）ことで数え落としを防げる。

STEP のその先へ

これは古典的な「マッチング問題」、すなわち完全順列の問題である。包除原理（inclusion–exclusion）を用いると $\mathrm{P}(N = 0) = \sum_{k=0}^{m} \frac{(-1)^k}{k!}$ が示され、これは極めて速く e^{-1} に近づく。実際、$m \to \infty$ のとき N は $\mathrm{Po}(1)$ に分布収束（converges in distribution）する。多数のまれで弱く従属した事象の個数はほぼポアソン分布に従うという、より一般的な原理は、Chen–Stein 法（Chen–Stein method）によって厳密に定式化される。Blitzstein と Hwang の *Introduction to Probability*（確率論入門）は、指示変数とマッチング問題を繰り返し登場する例として用いている。

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。