解答例

STEP III 2013、第 12 問

確率・統計 ・ 指示確率変数

STEP III 2013、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/13-s3-q12

鍵となる考え方

各 X_k は、位置 k で A の新しいブロックが始まるときにちょうど 1 となる。したがって S は、列に含まれる A のブロック（連、run）の個数を数えている。これらの変数は指示変数（indicator variable）なので、どの期待値（expectation）も、いくつかの指定された位置に指定された文字が入る確率にほかならない。そしてその確率は、関係する位置の個数だけで決まり、それらがどこにあるかにはよらない。分散（variance）を求めるには S^2 を展開し、積 X_iX_j を種類ごとに分類する。すなわち、添字が等しいもの、添字が隣り合うもの（これは起こりえないことがわかる）、そしてそれ以外である。

解答

以下では一貫して、$\binom{n}{a}$ 通りの並べ方がすべて同様に確からしいとする。ここから役に立つ帰結が得られる。任意の m 個の**相異なる**位置について、それらに p 個の A と q 個の B からなる（$p+q=m$）指定された並びが入る確率は、位置を一つずつ埋めていくことで求められる。たとえば

$$\Pr(\text{位置}r,s\text{ に入るのが}B,A) = \frac{b}{n}\cdot\frac{a}{n-1},$$

であり、これは相異なる位置 r と s がどこであっても成り立つ。各 X_k は 0 と 1 の値しかとらないので $\mathrm{E}(X_k) = \Pr(X_k = 1)$ であり、同様に $\mathrm{E}(X_iX_j) = \Pr(X_i = 1 \text{ かつ} X_j = 1)$ である。

小問 (i)

$i = 1$ のときは、最初の文字が A であればよいので

$$\mathrm{E}(X_1) = \frac{a}{n}.$$

$2 \leq i \leq n$ のときは、位置 $i-1$ に B、位置 i に A が入ればよいので

$$\mathrm{E}(X_i) = \frac{b}{n} \cdot \frac{a}{n-1} = \frac{ab}{n(n-1)}.$$

期待値の線形性（linearity of expectation）により、$n-1$ 個の添字 $i \geq 2$ について和をとると

$$\mathrm{E}(S) = \frac{a}{n} + (n-1) \cdot \frac{ab}{n(n-1)} = \frac{a}{n} + \frac{ab}{n} = \frac{a(b+1)}{n},$$

となり、示すべき式が得られた。

小問 (ii)

(a) $j \geq 3$ とする。このとき $X_1 X_j = 1$ となるのは、位置 1 に A、位置 $j-1$ に B、位置 j に A が入るときであり、かつそのときに限る。$j - 1 \geq 2$ なので、これらは相異なる三つの位置であり、そこに A が二つと B が一つ入る。順に埋めていくと

$$\mathrm{E}(X_1 X_j) = \frac{a}{n} \cdot \frac{b}{n-1} \cdot \frac{a-1}{n-2} = \frac{a(a-1)b}{n(n-1)(n-2)}.$$

(b) 次に $2 \leq i$ かつ $j \geq i+2$ とする。このとき $X_i X_j = 1$ となるのは、位置 $i-1$, i, $j-1$, j にそれぞれ B, A, B, A が入るときであり、かつそのときに限る。$j - 1 \geq i + 1 > i$ なので、これら四つの位置は相異なり、そこに A が二つと B が二つ入る。よって

$$\mathrm{E}(X_i X_j) = \frac{b}{n} \cdot \frac{a}{n-1} \cdot \frac{b-1}{n-2} \cdot \frac{a-1}{n-3} = \frac{a(a-1)b(b-1)}{n(n-1)(n-2)(n-3)}.$$

（ここで $a, b \geq 2$ より $n \geq 4$ なので、分母が 0 になることはない。）これは二重和に現れるどの組についても同じ値なので、組の個数を数えればよい。与えられた i に対して j は $i+2$ から n まで動くので、その値は $n-i-1$ 個である。i が 2 から $n-2$ まで動くと、この個数は $n-3, n-4, \ldots, 1$ という値をとる。したがって組の個数は

$$1 + 2 + \cdots + (n-3) = \tfrac{1}{2}(n-3)(n-2),$$

であり、

$$\sum_{i=2}^{n-2} \left(\sum_{j=i+2}^{n} \mathrm{E}(X_i X_j) \right) = \frac{(n-2)(n-3)}{2} \cdot \frac{a(a-1)b(b-1)}{n(n-1)(n-2)(n-3)} = \frac{a(a-1)b(b-1)}{2n(n-1)}.$$

(c) $\mathrm{Var}(S) = \mathrm{E}(S^2) - (\mathrm{E}(S))^2$ を用い、

$$S^2 = \sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j.$$

と展開する。X_i は 0 か 1 なので $X_i^2 = X_i$ であり、第一の和の期待値は $\mathrm{E}(S)$ である。

考えてみよう 小問 (b) で除外した、$j = i+1$ となる組はどうなるだろうか。

これらは何も寄与しない。常に $X_i X_{i+1} = 0$ だからである。$i = 1$ のとき、$X_1 = 1$ となるには位置 1 に A が必要だが、$X_2 = 1$ となるにはそこに B が必要である。$i \geq 2$ のとき、$X_i = 1$ となるには位置 i に A が必要だが、$X_{i+1} = 1$ となるにはそこに B が必要である。（A の二つのブロックが隣り合う位置から始まることはありえない。）

したがって、考慮すべき $i < j$ の組は $j \geq i+2$ を満たすものである。そのうち $i = 1$ のものは $j = 3, \ldots, n$ であり、$n-2$ 組あって、それぞれ小問 (a) で与えられる。$i \geq 2$ のものは、ちょうど小問 (b) の二重和である。よって

$$\begin{aligned}\sum_{i<j} \mathrm{E}(X_i X_j) &= (n-2)\cdot \frac{a(a-1)b}{n(n-1)(n-2)} + \frac{a(a-1)b(b-1)}{2n(n-1)} \\ &= \frac{a(a-1)b}{n(n-1)}\left(1 + \frac{b-1}{2}\right) = \frac{a(a-1)b(b+1)}{2n(n-1)},\end{aligned}$$

であり、したがって

$$\mathrm{E}(S^2) = \frac{a(b+1)}{n} + \frac{a(a-1)b(b+1)}{n(n-1)}.$$

$(\mathrm{E}(S))^2 = a^2(b+1)^2/n^2$ を引き、共通因数 $\dfrac{a(b+1)}{n^2(n-1)}$ をくくり出すと

$$\mathrm{Var}(S) = \frac{a(b+1)}{n^2(n-1)}\Big[n(n-1) + n(a-1)b - a(b+1)(n-1)\Big].$$

括弧の中を展開すると、項 nab と $-nab$ が打ち消し合い、

$$\begin{aligned}n^2 - n - nb - na + ab + a &= n(n-a-b) - n + ab + a \\ &= -n + ab + a \qquad \text{なぜなら} n = a+b \\ &= ab - b = b(a-1).\end{aligned}$$

が残る。よって

$$\mathrm{Var}(S) = \frac{a(b+1)}{n^2(n-1)}\cdot b(a-1) = \frac{a(a-1)b(b+1)}{n^2(n-1)},$$

となり、示すべき式が得られた。

検算

$a=3$、$b=2$ とすると $n=5$ であり、並べ方は 10 通りある。A のブロックを数えると、並べ方 $AAABB$、$BAAAB$、$BBAAA$ では $S=1$、$ABABA$ では $S=3$、残りの六つでは $S=2$ である。したがって $\mathrm{E}(S)=\frac{18}{10}=\frac{9}{5}$、$\mathrm{E}(S^2)=\frac{36}{10}$ であり、$\mathrm{Var}(S)=\frac{18}{5}-\frac{81}{25}=\frac{9}{25}$ となる。公式からは $\frac{3\cdot3}{5}=\frac{9}{5}$ と $\frac{3\cdot2\cdot2\cdot3}{25\cdot4}=\frac{9}{25}$ が得られる。コンピュータによる全数列挙により、$2\le a,b\le 6$ のすべてについて、問題中のすべての公式が厳密に成り立つことが確認された。

これらの公式は、問題で指定された範囲の外でも妥当な振る舞いをする。$a=2$、$b=1$ のとき、並べ方 AAB、ABA、BAA では $S=1,2,1$ なので $\mathrm{E}(S)=\frac{4}{3}$、$\mathrm{Var}(S)=2-\frac{16}{9}=\frac{2}{9}$ であり、$\frac{2\cdot2}{3}$ および $\frac{2\cdot1\cdot1\cdot2}{9\cdot2}$ と一致する。また $a=1$ のとき、分散の公式は 0 を与えるが、これは当然である。このときブロックは常にちょうど一つだからである。

よくある間違い

- X_1 をほかと同じように扱ってしまう。位置 0 は存在しないので、X_1 に必要なのは一文字だけであり、$\mathrm{E}(X_1)=a/n$ であって $ab/(n(n-1))$ ではない。
- X_i と X_{i+1} が同時に 1 になりえないことを忘れる、あるいは $\mathrm{E}(X_iX_{i+1})$ を 0 でない式で計算してしまう。
- (b) の組の個数を数え間違える。組の個数は $\frac{1}{2}(n-2)(n-3)$ であって、$\frac{1}{2}n(n-1)$ ではない。

STEP のその先へ

個数を指示変数の和として書き、平均を線形性から、分散を二つずつの積から求めるという方法は、大学の確率論で最も役に立つ技法の一つである。連、一致（matching）、ランダムな置換の不動点（fixed point）、ランダムグラフ（random graph）などは、この方法で解析される。ここでの S は、本質的には Wald–Wolfowitz の連検定（runs test）の統計量であり、これは列の非ランダム性を検出するために用いられる。Blitzstein and Hwang, *Introduction to Probability*（確率論入門）には、指示変数を用いる方法の例が数多く載っている。

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。