详细解答

STEP III 2013，第 12 题

概率与统计 · 指示随机变量

STEP III 2013，第 12 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/13-s3-q12

关键思路

每个 X_k 恰好在一段新的 A 从位置 k 开始时等于 1，所以 S 数的是这一行中 A 的段数（连串，runs）。这些变量都是指示变量（indicator variables），所以每个期望值（expectation）都只是若干指定位置放着指定字母的概率，而这个概率只取决于涉及多少个位置，与它们在哪里无关。至于方差（variance），我们把 S^2 展开，并把乘积 X_iX_j 分类：下标相等、下标相邻（结果是不可能发生的）以及其余情况。

解答

全文中，所有 $\binom{n}{a}$ 种排列的可能性均等。由此得到一个有用的结论：对任意 m 个不同的位置，它们放着某个指定图样（其中含 p 个 A 和 q 个 B，$p+q=m$）的概率，可以逐个位置填入来求得，例如

$$\Pr(\text{位置}r,s\text{ 分别为}B,A)=\frac{b}{n}\cdot\frac{a}{n-1},$$

不论不同的位置 r 和 s 是哪两个。由于每个 X_k 只取值 0 和 1，所以 $\mathrm{E}(X_k)=\Pr(X_k=1)$；同理 $\mathrm{E}(X_iX_j)=\Pr(X_i=1\text{ 且}X_j=1)$。

第 (i) 部分

当 $i=1$ 时，我们需要第一个字母是 A，所以

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

当 $2\le i\le n$ 时，我们需要位置 $i-1$ 是 B，而位置 i 是 A，所以

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

根据期望值的线性性质（linearity of expectation），对 $n-1$ 个下标 $i \geq 2$ 求和，

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

即为所求。

第 (ii) 部分

(a) 设 $j \geq 3$。那么 $X_1X_j = 1$ 恰好在位置 1 是 A、位置 $j-1$ 是 B 且位置 j 是 A 时成立。由于 $j-1 \geq 2$，这是三个不同的位置，放着两个 A 和一个 B。逐个填入，

$$\mathrm{E}(X_1X_j) = \frac{a}{n}\cdot\frac{b}{n-1}\cdot\frac{a-1}{n-2} = \frac{a(a-1)b}{n(n-1)(n-2)}.$$

(b) 现在取 $2 \leq i$ 及 $j \geq i+2$。那么 $X_iX_j = 1$ 恰好在位置 $i-1$, i, $j-1$, j 分别是 B, A, B, A 时成立。由于 $j-1 \geq i+1 > i$，这四个位置互不相同，放着两个 A 和两个 B，所以

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

（这里因为 $a, b \geq 2$，所以 $n \geq 4$，分母不会为零。）这个值对双重和中的每一对都相同，所以我们只需数出有多少对。对给定的 i，j 由 $i+2$ 取到 n，共 $n-i-1$ 个值；当 i 由 2 取到 $n-2$ 时，这个数依次取值 $n-3$, $n-4$, …, 1。所以对数为

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

而

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

(c) 我们利用 $\mathrm{Var}(S) = \mathrm{E}(S^2) - (\mathrm{E}(S))^2$，并展开

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

由于 X_i 取值 0 或 1，所以 $X_i^2 = X_i$，因此第一个和的期望值是 $\mathrm{E}(S)$。

想一想 那么 (b) 部分略去的 $j = i+1$ 的那些对又如何?

它们没有任何贡献：$X_iX_{i+1} = 0$ 恒成立。当 $i = 1$ 时，$X_1 = 1$ 要求位置 1 是 A，但 $X_2 = 1$ 要求该位置是 B。当 $i \geq 2$ 时，$X_i = 1$ 要求位置 i 是 A，但 $X_{i+1} = 1$ 要求该位置是 B。（两段 A 不可能在相邻的位置开始。）

所以真正有关的 $i<j$ 对，就是那些 $j\geq i+2$ 的。其中 $i=1$ 的有 $j=3,\dots,n$，共 $n-2$ 对，每一对都由 (a) 部分给出；$i\geq 2$ 的恰好就是 (b) 部分的双重和。因此

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

于是

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

减去 $(\mathrm{E}(S))^2=a^2(b+1)^2/n^2$，并抽出公因式 $\dfrac{a(b+1)}{n^2(n-1)}$，

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

展开括号，nab 与 $-nab$ 两项互相抵消，剩下

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

所以

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

即为所求。

验算

取 $a=3$、$b=2$，于是 $n=5$，共有 10 种排列。数出 A 的段数：排列 $AAABB$、$BAAAB$、$BBAAA$ 给出 $S=1$；$ABABA$ 给出 $S=3$；其余六种给出 $S=2$。所以 $\mathrm{E}(S)=\frac{18}{10}=\frac{9}{5}$，$\mathrm{E}(S^2)=\frac{36}{10}$，从而 $\mathrm{Var}(S)=\frac{18}{5}-\frac{81}{25}=\frac{9}{25}$。公式给出 $\frac{3\cdot3}{5}=\frac{9}{5}$ 及 $\frac{3\cdot2\cdot2\cdot3}{25\cdot4}=\frac{9}{25}$。以电脑穷举验证，对所有 $2\leq a,b\leq 6$，题中每条公式都完全正确。

这些公式在题目所述范围以外也表现合理。当 $a=2$、$b=1$ 时，排列 AAB、ABA、BAA 给出 $S=1,2,1$，所以 $\mathrm{E}(S)=\frac{4}{3}$，$\mathrm{Var}(S)=2-\frac{16}{9}=\frac{2}{9}$，与 $\frac{2\cdot2}{3}$ 及 $\frac{2\cdot1\cdot1\cdot2}{9\cdot2}$ 吻合。而若 $a=1$，方差公式给出 0，这理应如此，因为此时永远恰好只有一段。

常见错误

- 把 X_1 当作与其他变量一样处理：并没有位置 0，所以 X_1 只涉及一个字母，$\mathrm{E}(X_1)=a/n$，而不是 $ab/(n(n-1))$。

- 忘记 X_i 与 X_{i+1} 不可能同时为 1，或用一条非零的公式去计算 $\mathrm{E}(X_iX_{i+1})$。
- 数错 (b) 部分的对数：应有 $\frac{1}{2}(n-2)(n-3)$ 对，而不是 $\frac{1}{2}n(n-1)$ 对。

STEP 之外

把一个计数写成指示变量之和，再用线性性质求平均值，并由两两乘积求方差，是大学概率论中最有用的技巧之一；分析连串、配对、随机排列（random permutations）的不动点（fixed points）以及随机图（random graphs），都是用这个方法。这里的 S 本质上就是 Wald–Wolfowitz 连串检定（runs test）的统计量，用来检测一个序列是否并非随机。Blitzstein 与 Hwang 的 *Introduction to Probability*（《概率导论》）载有许多运用指示变量方法的例子。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。