詳細解答

STEP III 2013，第 12 題

概率與統計 • 指示隨機變量

> **STEP III 2013，第 12 題**
>
> 繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：
>
> step.maths.org/questions/13-s3-q12

關鍵思路

每個 X_k 恰好在一段新的 A 從位置 k 開始時等於 1，所以 S 數的是這一行中 A 的段數（連串，runs）。這些變量都是指示變量（indicator variables），所以每個期望值（expectation）都只是若干指定位置放着指定字母的概率，而這個概率只取決於涉及多少個位置，與它們在哪裏無關。至於方差（variance），我們把 S^2 展開，並把乘積 X_iX_j 分類：下標相等、下標相鄰（結果是不可能發生的）以及其餘情況。

解答

全文中，所有 $\binom{n}{a}$ 種排列的可能性均等。由此得到一個有用的結論：對任意 m 個不同的位置，它們放着某個指定圖樣（其中含 p 個 A 和 q 個 B，$p+q=m$）的概率，可以逐個位置填入來求得，例如

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

不論不同的位置 r 和 s 是哪兩個。由於每個 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

當 $i=1$ 時，我們需要第一個字母是 A，所以

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

當 $2\le i\le n$ 時，我們需要位置 $i-1$ 是 B，而位置 i 是 A，所以

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

根據期望值的線性性質（linearity of expectation），對 $n-1$ 個下標 $i \geq 2$ 求和，

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

即為所求。

第 (ii) 部分

(a) 設 $j \geq 3$。那麼 $X_1X_j = 1$ 恰好在位置 1 是 A、位置 $j-1$ 是 B 且位置 j 是 A 時成立。由於 $j-1 \geq 2$，這是三個不同的位置，放着兩個 A 和一個 B。逐個填入，

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

(b) 現在取 $2 \leq i$ 及 $j \geq i+2$。那麼 $X_iX_j = 1$ 恰好在位置 $i-1$, i, $j-1$, j 分別是 B, A, B, A 時成立。由於 $j-1 \geq i+1 > i$，這四個位置互不相同，放着兩個 A 和兩個 B，所以

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

（這裏因為 $a, b \geq 2$，所以 $n \geq 4$，分母不會為零。）這個值對雙重和中的每一對都相同，所以我們只需數出有多少對。對給定的 i，j 由 $i+2$ 取到 n，共 $n-i-1$ 個值；當 i 由 2 取到 $n-2$ 時，這個數依次取值 $n-3$, $n-4$, …, 1。所以對數為

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

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) 我們利用 $\mathrm{Var}(S) = \mathrm{E}(S^2) - (\mathrm{E}(S))^2$，並展開

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

由於 X_i 取值 0 或 1，所以 $X_i^2 = X_i$，因此第一個和的期望值是 $\mathrm{E}(S)$。

想一想 那麼 (b) 部分略去的 $j = i+1$ 的那些對又如何？

它們沒有任何貢獻：$X_iX_{i+1} = 0$ 恆成立。當 $i = 1$ 時，$X_1 = 1$ 要求位置 1 是 A，但 $X_2 = 1$ 要求該位置是 B。當 $i \geq 2$ 時，$X_i = 1$ 要求位置 i 是 A，但 $X_{i+1} = 1$ 要求該位置是 B。（兩段 A 不可能在相鄰的位置開始。）

所以真正有關的 $i<j$ 對，就是那些 $j\geq i+2$ 的。其中 $i=1$ 的有 $j=3,\ldots,n$，共 $n-2$ 對，每一對都由 (a) 部分給出；$i\geq 2$ 的恰好就是 (b) 部分的雙重和。因此

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

於是

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

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

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

展開括號，nab 與 $-nab$ 兩項互相抵消，剩下

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

所以

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

即為所求。

驗算

取 $a=3$、$b=2$，於是 $n=5$，共有 10 種排列。數出 A 的段數：排列 $AAABB$、$BAAAB$、$BBAAA$ 給出 $S=1$；$ABABA$ 給出 $S=3$；其餘六種給出 $S=2$。所以 $\mathrm{E}(S)=\frac{18}{10}=\frac{9}{5}$，$\mathrm{E}(S^2)=\frac{36}{10}$，從而 $\mathrm{Var}(S)=\frac{18}{5}-\frac{81}{25}=\frac{9}{25}$。公式給出 $\frac{3\cdot3}{5}=\frac{9}{5}$ 及 $\frac{3\cdot2\cdot2\cdot3}{25\cdot4}=\frac{9}{25}$。以電腦窮舉驗證，對所有 $2\leq a,b\leq 6$，題中每條公式都完全正確。

這些公式在題目所述範圍以外也表現合理。當 $a=2$、$b=1$ 時，排列 AAB、ABA、BAA 給出 $S=1,2,1$，所以 $\mathrm{E}(S)=\frac{4}{3}$，$\mathrm{Var}(S)=2-\frac{16}{9}=\frac{2}{9}$，與 $\frac{2\cdot2}{3}$ 及 $\frac{2\cdot1\cdot1\cdot2}{9\cdot2}$ 吻合。而若 $a=1$，方差公式給出 0，這理應如此，因為此時永遠恰好只有一段。

常見錯誤

- 把 X_1 當作與其他變量一樣處理：並沒有位置 0，所以 X_1 只涉及一個字母，$\mathrm{E}(X_1)=a/n$，而不是 $ab/(n(n-1))$。

- 忘記 X_i 與 X_{i+1} 不可能同時為 1，或用一條非零的公式去計算 $\mathrm{E}(X_iX_{i+1})$。
- 數錯 (b) 部分的對數：應有 $\frac{1}{2}(n-2)(n-3)$ 對，而不是 $\frac{1}{2}n(n-1)$ 對。

STEP 之外

把一個計數寫成指示變量之和，再用線性性質求平均值，並由兩兩乘積求方差，是大學概率論中最有用的技巧之一；分析連串、配對、隨機排列（random permutations）的不動點（fixed points）以及隨機圖（random graphs），都是用這個方法。這裏的 S 本質上就是 Wald–Wolfowitz 連串檢定（runs test）的統計量，用來檢測一個序列是否並非隨機。Blitzstein 與 Hwang 的 *Introduction to Probability*（《概率導論》）載有許多運用指示變量方法的例子。

Beyond STEP 詳細解答：https://beyondstep.keithmcnulty.org/zh-hant/。解答文字採用 CC0 1.0；題目本身版權屬 Cambridge University Press & Assessment 所有（©），本文件不予轉載。由機器從英文翻譯而成；尚未經審閱。