J. Appl. Prob. **47,** 367–377 (2010)
Printed in England

ON THE NUMBER OF RUNS FOR BERNOULLI ARRAYS

DJILALI AIT AOUDIA * AND
ÉRIC MARCHAND,** *Université de Sherbrooke*

Abstract

We introduce and motivate the study of $(n+1)\times r$ arrays X with Bernoulli entries $X_{k,j}$ and independently distributed rows. We study the distribution of $S_n = \sum_{j=1}^{r}\sum_{k=1}^{n} X_{k,j}X_{k+1,j}$, which denotes the number of consecutive pairs of successes (or runs of length 2) when reading the array down the columns and across the rows. With the case $r=1$ having been studied by several authors, and permitting some initial inferences for the general case $r>1$, we examine various distributional properties and representations of S_n for the case $r=2$, and, using a more explicit analysis, the case of multinomial and identically distributed rows. Applications are also given in cases where the array X arises from a Pólya sampling scheme.

Keywords: Bernoulli; multinomial; Pólya urn; probability generating function; runs

2010 Mathematics Subject Classification: Primary 60C05; 60E05; 60K99

1. Introduction

For an array X of Bernoulli random variables $X_{k,j}$, $k \geqslant 1$ and $j = 1,\dots,r$, such that the random vectors $\underline{X}_k = (X_{k,1},\dots,X_{k,r})^\top$ are independent, we study the distributional properties of $S_n = \sum_{k=1}^{n} \underline{X}_k^\top \underline{X}_{k+1}$, $n \geq 1$, which denotes the number of pairs of consecutive Bernoulli successes (or runs of length 2) in the array X when reading down the lines and across the columns. We may alternatively write

$$S_n = \sum_{j=1}^{r}\sum_{k=1}^{n} X_{k,j}X_{k+1,j} = \sum_{j=1}^{r} Z_j \tag{1}$$

with

$$Z_j = \sum_{k=1}^{n} X_{k,j}X_{k+1,j}.$$

As an illustration, the array

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix},$$

where $n=3$ and $r=5$, yields $Z_1=1$, $Z_2=0$, $Z_3=0$, $Z_4=2$, $Z_5=1$, and $S_3 = \sum_{j=1}^{5} Z_j = 4$.

Received 5 August 2009; revision received 14 January 2010.
* Current address: Département de Mathématiques et de Statistique, Université de Montréal, Montréal, QC H3C 3J7, Canada.
** Postal address: Département de Mathématiques, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada.
Email address: eric.marchand@usherbrooke.ca

Example 1. A key class of situations where S_n arises consists of $n+1$ draws with replacement from an urn with r colours of cardinalities $\alpha_1, \ldots, \alpha_r$ ($\alpha_i > 0$), where we define

$$X_{k,j} = 1_{\{k\text{th draw is of type } j\}}, \qquad 1 \le k \le n,\ 1 \le j \le r,$$

with $1_{\{\cdot\}}$ denoting the indicator function. We thus have $\underline{X}_k \sim$ Multinomial$(1, \theta_{k,1}, \ldots, \theta_{k,r})$, with $\theta_{k,j} = \alpha_j / \sum_{j=1}^r \alpha_j$, and S_n denoting the number of pairs of consecutive draws with matching colours. Observe that the independence of the vectors $\underline{X}_k$ is a consequence of the draws with replacement.

The univariate column case (i.e. $r = 1$ or as pertaining to the marginal distribution of Z_j) has generated much recent interest, analysis, and interpretation; see [4], [5], [6], [7], [9], and [10], among others. Although our main interest relates to cases where the Z_js are not independent, it is worthwhile here summarizing results that may be inferred in cases where the columns are independent.

Example 2. (*Independent columns and* $\mathrm{P}(X_{k,j} = 1) = a_j/(a_j + b_j + k - 1)$.) Whenever the columns are independent, representation (1) implies that an efficient analysis of S_n passes through the convolution of r independent one-dimensional problems. For instance, whenever the $X_{k,j}$s, $k \ge 1$, $j = 1, \ldots, r$, are independent, and $\mathrm{P}(X_{k,j} = 1) = a_j/(a_j + b_j + k - 1)$, we can show that the distribution of $S = \lim_{n\to\infty} S_n$ admits the following representation:

$$S \mid Y \sim \mathrm{Poi}(Y) \quad \text{with} \quad Y = \sum_{j=1}^r a_j L_j,$$

where $L_1, \ldots, L_r$ are independent random variables such that $L_j \sim \mathrm{Beta}(a_j, b_j)$. Indeed, this may be established by exploiting representation (1), the independence of the Z_js, and the characterization

$$Z_j \mid L_j \sim \mathrm{Poi}(a_j L_j)$$

(see, e.g. [5]), so that, for all $t \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{E}[t^S] &= \mathrm{E}[t^{\sum_{j=1}^n Z_j}] \\
&= \prod_{j=1}^r \mathrm{E}[t^{Z_j}] \\
&= \prod_{j=1}^r \mathrm{E}[\mathrm{E}[t^{Z_j}] \mid L_j] \\
&= \prod_{j=1}^r \mathrm{E}[\exp(a_j L_j (t-1))] \\
&= \mathrm{E}\left[\exp\left(\sum_{j=1}^r a_j L_j (t-1)\right)\right] \\
&= \mathrm{E}[\mathrm{e}^{Y(t-1)}].
\end{aligned}$$

Our introduction (for $r > 1$) and focus on the properties of S_n is, as far as we can tell in the face of a vast amount of literature on runs as defined above for $r = 1$, novel, and so are our main

https://doi.org/10.1239/jap/1276784897 Published online by Cambridge University Press

findings which pertain to the bidimensional case ($r = 2$) with additional implications for Pólya urns. In Section 2 we derive a general recurrence (Lemma 1) for the probability generating function (PGF) of S_n using a conditioning argument (as in [7, Proposition 1]). We then proceed to more explicit representations (Corollary 1) for the particular case of multinomial rows (as in Example 1), which will lead to an explicit expression (Theorem 1) for the probability mass function of S_n in the identically distributed case. The new family of distributions thus obtained, which is quite interesting on its own, may be viewed as one that contains (for fixed n) the $\mathrm{Bin}(n, \frac{1}{2})$ distribution and whose members admit a large n normal approximation (Theorem 2). Finally, although our setup does not encompass Pólya urn sampling schemes where the rows of X are dependent, we nevertheless derive by de Finetti's representation theorem novel implications for Pólya urns, as expanded upon in Section 3.

2. The bidimensional case ($r = 2$)

We begin by analyzing the general bidimensional case ($r = 2$) where

$$f_{i,j}^{(k)} = \mathrm{P}(\underline{X}_k = (i, j)), \qquad 1 \le k \le n+1,\ i, j = 0, 1.$$

Define S_n as in (1) above with $S_0 = 0$, i.e.

$$S_n = S_{n-1} + X_{n,1}X_{n+1,1} + X_{n,2}X_{n+1,2}, \qquad n \ge 1. \tag{2}$$

To derive a useful expression for the PGF $\varphi_{S_n}(t)(= \mathrm{E}[t^{S_n}])$ of S_n, we introduce the auxiliary random variables $W_{l,n}$, $l = 1, 2, 3$, defined as, for $n \ge 0$,

$$W_{1,n} := S_n + X_{n+1,1} + X_{n+1,2}, \qquad W_{2,n} := S_n + X_{n+1,1}, \qquad W_{3,n} := S_n + X_{n+1,2}. \tag{3}$$

We denote their PGFs by $\varphi_{W_{l,n}}$, and we also write

$$\underline{\varphi}_n(\cdot) = (\varphi_{S_n}(\cdot), \varphi_{W_{1,n}}(\cdot), \varphi_{W_{2,n}}(\cdot), \varphi_{W_{3,n}}(\cdot))^\top, \qquad n \ge 0.$$

The next result establishes an explicit recurrence and expression for $\underline{\varphi}_n(\cdot)$, $n \ge 1$.

Lemma 1. *We have, for $n \ge 1$ and $t \ge 0$,*

$$\underline{\varphi}_n(t) = M_{n+1}\underline{\varphi}_{n-1}(t) \tag{4}$$

and

$$\underline{\varphi}_n(t) = M_{n+1}\cdots M_2\underline{\varphi}_0(t) = M_{n+1}\cdots M_1\mathbf{1}, \tag{5}$$

where

$$M_n = \begin{bmatrix} f_{0,0}^{(n)} & f_{1,1}^{(n)} & f_{1,0}^{(n)} & f_{0,1}^{(n)} \\ f_{0,0}^{(n)} & t^2 f_{1,1}^{(n)} & t f_{1,0}^{(n)} & t f_{0,1}^{(n)} \\ f_{0,0}^{(n)} & t f_{1,1}^{(n)} & t f_{1,0}^{(n)} & f_{0,1}^{(n)} \\ f_{0,0}^{(n)} & t f_{1,1}^{(n)} & f_{1,0}^{(n)} & t f_{0,1}^{(n)} \end{bmatrix}$$

and $\mathbf{1} = (1, 1, 1, 1)^\top$.

https://doi.org/10.1239/jap/1276784897 Published online by Cambridge University Press

Proof. We condition on $\underline{X}_{n+1}$. For S_n, $n \geq 1$, we obtain, from (2), (3), and the independence of the $\underline{X}_k$s,

$$\mathcal{L}(S_n \mid \underline{X}_{n+1} = (1,1)) = \mathcal{L}(S_{n-1} + X_{n,1} + X_{n,2}) = \mathcal{L}(W_{1,n-1}),$$
$$\mathcal{L}(S_n \mid \underline{X}_{n+1} = (1,0)) = \mathcal{L}(S_{n-1} + X_{n,1}) = \mathcal{L}(W_{2,n-1}),$$
$$\mathcal{L}(S_n \mid \underline{X}_{n+1} = (0,1)) = \mathcal{L}(S_{n-1} + X_{n,2}) = \mathcal{L}(W_{3,n-1}),$$
$$\text{and} \quad \mathcal{L}(S_n \mid \underline{X}_{n+1} = (0,0)) = \mathcal{L}(S_{n-1}).$$

Since

$$\varphi_{S_n}(t) = \mathrm{E}[\mathrm{E}[t^{S_n} \mid \underline{X}_{n+1}]],$$

the above translates to

$$\varphi_{S_n}(t) = (f_{0,0}^{(n+1)}, f_{1,1}^{(n+1)}, f_{1,0}^{(n+1)}, f_{0,1}^{(n+1)})\underline{\varphi}_{n-1}(t)$$

(i.e. the scalar product of the first row of M_{n+1} and $\underline{\varphi}_{n-1}(t)$). The remaining system of equations is obtained along the same lines. Finally, (5) follows from (4), with $\underline{\varphi}_0(t)$ derived directly as $M_1\mathbf{1}$ from the definitions of $W_{l,0}$ in (3), and since $S_0 = 0$.

We now pursue further analysis of multinomially distributed rows $\underline{X}_k$, where, for all $k \in \{1, \ldots, n+1\}$,

$$f_{1,0}^{(k)} = 1 - f_{0,1}^{(k)} = p_k \quad \text{(say)}, \tag{6}$$

and where S_n reduces to $\sum_{k=1}^{n} 1_{\{X_{k+1,1} = X_{k,1}\}}$. This corresponds to nonhomogeneous (or nonidentically distributed) draws in Example 1. We do not assume for the time being that the $\underline{X}_k$s are identically distributed, but the more explicit results that follow later (e.g. part (b) of Corollary 1) do apply to cases where

$$f_{1,0}^{(k)} = 1 - f_{0,1}^{(k)} = p \tag{7}$$

for all $k \in \{1, \ldots, n+1\}$. We require the following result, a proof of which is given in [2].

Lemma 2. *If A is a 2×2 matrix with distinct eigenvalues λ_1 and λ_2, and I_2 is the 2×2 identity matrix, then, for all $n \geq 2$,*

$$A^n = \left(\frac{\lambda_2^n - \lambda_1^n}{\lambda_2 - \lambda_1}\right)A - \lambda_1\lambda_2\left(\frac{\lambda_2^{n-1} - \lambda_1^{n-1}}{\lambda_2 - \lambda_1}\right)I_2.$$

The next result consists of specializations of Lemma 1 to cases where (6) or (7) holds. In cases where the first two columns of M_n are zero vectors, we obtain a useful and more explicit representation for the PGF of S_n.

Corollary 1. (a) *Under assumption (6), we have, for all $n \geq 1$ and $t \geq 0$,*

$$\varphi_{S_n}(t) = p_{n+1}\varphi_{W_{2,n-1}}(t) + (1 - p_{n+1})\varphi_{W_{3,n-1}}(t) \tag{8}$$

with

$$\begin{bmatrix} \varphi_{W_{2,n}}(t) \\ \varphi_{W_{3,n}}(t) \end{bmatrix} = \begin{bmatrix} tp_{n+1} & 1 - p_{n+1} \\ p_{n+1} & t(1 - p_{n+1}) \end{bmatrix} \begin{bmatrix} \varphi_{W_{2,n-1}}(t) \\ \varphi_{W_{3,n-1}}(t) \end{bmatrix}$$

and $(\varphi_{W_{2,0}}(t), \varphi_{W_{3,0}}(t))^\top = (p_1 t + (1 - p_1), p_1 + (1 - p_1)t)^\top$.

https://doi.org/10.1239/jap/1276784897 Published online by Cambridge University Press

(b) *For all* $p \in [0, 1]$, $n \geq 1$, *and* $t \geq 0$, *set*

$$\lambda_1 = \frac{1}{2}\big(t + \sqrt{t^2 - 4p(1-p)(t^2-1)}\big), \qquad \lambda_2 = t - \lambda_1, \qquad \alpha_n = \frac{\lambda_1^n - \lambda_2^n}{\lambda_1 - \lambda_2}.$$

Under assumption (7), we have, for all $n \geq 1$ *and* $t \geq 0$,

$$\varphi_{S_n}(t) = \alpha_n[2p(1-p) + t(1 - 2p(1-p))] + \alpha_{n-1}p(1-p)(1-t^2). \tag{9}$$

Proof. (a) The result follows directly from Lemma 1, and since

$$W_{2,0} = X_{1,1} \quad \text{and} \quad W_{3,0} = X_{1,2}.$$

(b) From part (a) we have, for all $n \geq 1$ and under assumption (7),

$$\begin{bmatrix} \varphi_{W_{2,n-1}}(t) \\ \varphi_{W_{3,n-1}}(t) \end{bmatrix} = B^{n-1} \begin{bmatrix} \varphi_{W_{2,0}}(t) \\ \varphi_{W_{3,0}}(t) \end{bmatrix} = B^n \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where

$$B = \begin{bmatrix} tp & 1-p \\ p & t(1-p) \end{bmatrix}.$$

Observe that λ_1 and λ_2 are the eigenvalues of B, so that

$$B^n = \alpha_n B - (t^2 - 1)p(1-p)\alpha_{n-1}1_2,$$

by virtue of Lemma 2. From this we obtain

$$\begin{bmatrix} \varphi_{W_{2,n-1}}(t) \\ \varphi_{W_{3,n-1}}(t) \end{bmatrix} = \alpha_n \begin{bmatrix} tp + 1 - p \\ p + t(1-p) \end{bmatrix} + \alpha_{n-1} \begin{bmatrix} (1-t^2)p(1-p) \\ (1-t^2)p(1-p) \end{bmatrix},$$

and the result follows by applying (8).

We pursue an expansion of φ_{S_n} as given in (9) in order to derive an explicit form for the probability function of S_n.

Theorem 1. *Under assumption (7), we have the following assertions.*

(a) *For* $p = \frac{1}{2}$, *we have* $S_n \sim \mathrm{Bin}(n, \frac{1}{2})$.

(b) *Let* $\rho = 4p(1-p)$, *let* n *and* k *be nonnegative integers, and let*

$$a_{n,k}(\rho) = \left(\frac{\rho}{1-\rho}\right)^{\lceil n/2 \rceil - k - 1} \times \sum_{j=\lceil n/2 \rceil - k - 1}^{\lceil n/2 \rceil - 1} \binom{j}{\lceil n/2 \rceil - k - 1} \binom{n}{2j+1} (1-\rho)^j 1_{\{0 \leq k \leq \lceil n/2 \rceil - 1\}}.$$

For all $p \in (0, 1)$, $p \neq \frac{1}{2}$, *we have*

(i) *for odd* n,

$$\mathrm{P}(S_n = 2k) = \frac{\rho a_{n,k}(\rho)}{2^n},$$

$$\textit{and} \quad \mathrm{P}(S_n = 2k+1) = \frac{1}{2^n}[(2-\rho)a_{n,k}(\rho) + \rho a_{n-1,k}(\rho) - \rho a_{n-1,k-1}(\rho)];$$

https://doi.org/10.1239/jap/1276784897 Published online by Cambridge University Press

(ii) *for even n,*

$$\mathrm{P}(S_n = 2k) = \frac{1}{2^n}[(2-\rho)a_{n,k-1}(\rho) + \rho a_{n-1,k}(\rho) - \rho a_{n-1,k-1}(\rho)],$$
$$\textit{and} \quad \mathrm{P}(S_n = 2k+1) = \frac{\rho a_{n,k}(\rho)}{2^n}.$$

Proof. For part (a), evaluate (9) with $p = \frac{1}{2}$ to obtain

$$\alpha_n = \left(\frac{t+1}{2}\right)^n - \left(\frac{t-1}{2}\right)^n$$

and

$$\varphi_{S_n}(t) = \alpha_n\left(\frac{t+1}{2}\right) - \alpha_{n-1}\left(\frac{t-1}{2}\right)\left(\frac{t+1}{2}\right) = \left(\frac{t+1}{2}\right)^n,$$

which implies that $S_n \sim \mathrm{Bin}(n, \frac{1}{2})$. For part (b), begin with standard operations to express α_n as a polynomial in t. Write the λ_1 and λ_2 of Corollary 1 as $\lambda_1 = (t+\Delta)/2$ and $\lambda_2 = (t-\Delta)/2$, with $\Delta = \sqrt{\rho + t^2(1-\rho)}$, so that $\lambda_1 - \lambda_2 = \Delta$, and

$$\begin{aligned}
\alpha_n &= \frac{\lambda_1^n - \lambda_2^n}{\lambda_1 - \lambda_2} \\
&= \frac{1}{\Delta 2^n}\{(t+\Delta)^n - (t-\Delta)^n\} \\
&= \frac{1}{\Delta 2^n}\sum_{k=0}^{n}\binom{n}{k}t^{n-k}\Delta^k(1^k - (-1)^k) \\
&= \frac{1}{2^{n-1}}\sum_{k=0,\,k\text{ odd}}^{n}\binom{n}{k}t^{n-k}\Delta^{k-1} \\
&= \frac{1}{2^{n-1}}\sum_{j=0}^{\lceil n/2\rceil-1}\binom{n}{2j+1}t^{n-2j-1}(\rho + t^2(1-\rho))^j \\
&= \frac{1}{2^{n-1}}\sum_{j=0}^{\lceil n/2\rceil-1}\binom{n}{2j+1}t^{n-2j-1}\sum_{k=0}^{j}\binom{j}{k}\left(\frac{\rho}{1-\rho}\right)^k(1-\rho)^j t^{2j-2k} \\
&= \frac{1}{2^{n-1}}\sum_{k=0}^{\lceil n/2\rceil-1}t^{n-2k-1}\left(\frac{\rho}{1-\rho}\right)^k\sum_{j=k}^{\lceil n/2\rceil-1}\binom{j}{k}\binom{n}{2j+1}(1-\rho)^j \\
&= \frac{t^{1\{n\text{ even}\}}}{2^{n-1}}\sum_{k=0}^{\lceil n/2\rceil-1}t^{2k}\left(\frac{\rho}{1-\rho}\right)^{\lceil n/2\rceil-k-1} \\
&\qquad\qquad \times \sum_{j=\lceil n/2\rceil-k-1}^{\lceil n/2\rceil-1}\binom{j}{\lceil n/2\rceil-k-1}\binom{n}{2j+1}(1-\rho)^j \\
&= \frac{t^{1\{n\text{ even}\}}}{2^{n-1}}\sum_k a_{n,k}(\rho)t^{2k}, \qquad (10)
\end{aligned}$$

https://doi.org/10.1239/jap/1276784897 Published online by Cambridge University Press

with the change of variable $k \to \lceil n/2 \rceil - k - 1_{\{n \text{ even}\}}$ in the penultimate line. Making use of (9) and (10), we obtain, for odd n,

$$\begin{aligned}\varphi_{S_n}(t) &= \left\{\frac{1}{2^{n-1}}\left(\sum_k a_{n,k}(\rho)t^{2k}\right)\left(\frac{\rho}{2} + t\left(1 - \frac{\rho}{2}\right)\right)\right\} \\ &\quad + \left\{\frac{1}{2^{n-2}}\left(\sum_k a_{n-1,k}(\rho)t^{2k+1}\right)\left(\frac{\rho(1-t^2)}{4}\right)\right\} \\ &= \frac{1}{2^n}\left\{\sum_k \rho a_{n,k}(\rho)t^{2k} + \sum_k (2-\rho)a_{n,k}(\rho)t^{2k+1} + \sum_k \rho a_{n-1,k}(\rho)t^{2k+1}\right. \\ &\quad \left. - \sum_k \rho a_{n-1,k-1}(\rho)t^{2k+1}\right\},\end{aligned}$$

and the result follows by collecting terms in the representation $\varphi_{S_n}(t) = \sum_k t^{2k} \mathrm{P}(S_n = 2k) + \sum_k t^{2k+1} \mathrm{P}(S_n = 2k+1)$. Finally, a similar expansion leads to the stated result for even n.

The probabilities of the events $\{S_n = 0\}$ and $\{S_n = n\}$ are of particular interest.

Corollary 2. *Under assumption (7), we have, for all $n \geq 1$,*

$$\mathrm{P}(S_n = 0) = 2(p(1-p))^{(n+1)/2} 1_{\{n \text{ odd}\}} + (p(1-p))^{n/2} 1_{\{n \text{ even}\}},$$
$$\text{and} \quad \mathrm{P}(S_n = n) = p^{n+1} + (1-p)^{n+1}.$$

Proof. First, observe that $\mathrm{P}(S_n = 0) = \varphi_{S_n}(0)$. The result follows by evaluating (9) with $t = 0$, $\lambda_1 = -\lambda_2 = \sqrt{p(1-p)}$, $\alpha_n = (p(1-p))^{(n-1)/2}$ for odd n, and $\alpha_n = 0$ for even n. Now, consider $\mathrm{P}(S_n = n)$ for even n (part (b) of Theorem 1 with $k = n/2$). Then

$$\begin{aligned}\mathrm{P}(S_n = n) &= \frac{1}{2^n}[(2-\rho)a_{n,n/2-1}(\rho) + \rho a_{n-1,n/2}(\rho) - \rho a_{n-1,n/2-1}(\rho)] \\ &= \frac{1}{2^n}\left[(2-\rho)\sum_{j=0}^{n/2-1}\binom{n}{2j+1}(1-\rho)^j + 0 - \rho\sum_{j=0}^{n/2-1}\binom{n-1}{2j+1}(1-\rho)^j\right] \\ &= \frac{1}{2^{n+1}}\left[\left(\sqrt{1-\rho} + \frac{1}{\sqrt{1-\rho}}\right)\{(1+\sqrt{1-\rho})^n - (1-\sqrt{1-\rho})^n\}\right. \\ &\quad \left. + \left(\sqrt{1-\rho} - \frac{1}{\sqrt{1-\rho}}\right)\{(1+\sqrt{1-\rho})^{n-1} - (1-\sqrt{1-\rho})^{n-1}\}\right],\end{aligned}$$

by virtue of the identity

$$\sum_{j=0}^{s/2-1}\binom{s}{2j+1}w^j = \frac{1}{2\sqrt{w}}[(1+\sqrt{w})^s - (1-\sqrt{w})^s],$$

where s is an even positive integer and $w > 0$. Finally, with a little algebra, we obtain the equivalent form

$$\mathrm{P}(S_n = n) = \frac{1}{2^{n+1}}[(1+\sqrt{1-\rho})^{n+1} + (1-\sqrt{1-\rho})^{n+1}] = p^{n+1} + (1-p)^{n+1}.$$

A similar development yields the result for odd n.

https://doi.org/10.1239/jap/1276784897 Published online by Cambridge University Press

Remark 1. The probabilities of the events $\{S_n = 0\}$ and $\{S_n = n\}$, as well as the distribution of S_n for the constant case $p = \frac{1}{2}$ in (7), can also be evaluated by elementary combinatorial arguments. For instance, we can directly infer that

$$\mathrm{P}(S_n = n) = \mathrm{P}\left(\bigcap_{k=1}^{n+1}\{X_{k,1} = 1\}\right) + \mathrm{P}\left(\bigcap_{k=1}^{n+1}\{X_{k,1} = 0\}\right) = \prod_{k=1}^{n+1} p_k + \prod_{k=1}^{n+1}(1 - p_k),$$

under (6), and

$$\mathrm{P}(S_n = n) = p^{n+1} + (1 - p)^{n+1},$$

under (7). We refer the reader to [1] for additional details.

From both a combinatorial and probabilistic point of view, the probability distributions described in (9) and Theorem 1 form an interesting family on their own with parameter ρ, $\rho \in (0, 1]$, with the case $\rho = 1$ corresponding to a $\mathrm{Bin}(n, \frac{1}{2})$ distribution. Moreover, as shown below, all of these distributions may be described by a large n normal approximation which extends the well-known result applicable to the $\mathrm{Bin}(n, \frac{1}{2})$ distribution.

Theorem 2. *Under assumption (7), we have*

$$\frac{S_n - n(1 - \rho/2)}{\sqrt{n(\rho - 3\rho^2/4)}} \xrightarrow{\mathrm{D}} N(0, 1),$$

with $\rho = 4p(1 - p)$.

Proof. Observe that $S_n \stackrel{\mathrm{D}}{=} \sum_{k=1}^{n} Y_k$ with $Y_k = 1_{\{X_{k+1,1} = X_{k,1}\}} \sim \mathrm{Bernoulli}(1 - \rho/2)$. The Bernoulli sequence $Y_1, Y_2, \ldots$ is stationary and 1-dependent (i.e. Y_i and Y_j are independent for all $|i - j| > 1$), so that the results of Stein [11, Corollary 3.1] imply that

$$\frac{S_n - \mathrm{E}[S_n]}{\sqrt{\mathrm{var}(S_n)}} \xrightarrow{\mathrm{D}} N(0, 1). \tag{11}$$

Evaluations yield

$$\begin{aligned}
\mathrm{E}[S_n] &= n\left(1 - \frac{\rho}{2}\right),\\
\mathrm{var}(Y_1) &= \frac{\rho}{2}\left(1 - \frac{\rho}{2}\right),\\
\mathrm{cov}(Y_1, Y_2) &= \mathrm{P}(X_1 = X_2 = X_3) - (p^2 + (1 - p)^2)^2\\
&= (p^3 + (1 - p)^3) - (p^2 + (1 - p)^2)^2\\
&= \frac{\rho(1 - 2\rho)}{4},\\
\mathrm{var}(S_n) &= \mathrm{var}\left(\sum_{k=1}^{n} Y_k\right)\\
&= n\left\{\mathrm{var}(Y_1) + 2\left(1 - \frac{1}{n}\right)\mathrm{cov}(Y_1, Y_2)\right\}\\
&= n\left\{\frac{\rho}{2}\left(1 - \frac{\rho}{2}\right) + \left(1 - \frac{1}{n}\right)\frac{\rho(1 - 2\rho)}{2}\right\}\\
&= n\left(\rho - \frac{3\rho^2}{4}\right) + O(n^{-1}).
\end{aligned}$$

Finally, the above properties along with (11) yield the result.

https://doi.org/10.1239/jap/1276784897 Published online by Cambridge University Press

3. Applications for Pólya urns

By virtue of de Finetti's representation theorem for sequences of 0–1 exchangeable random variables, the results above under assumption (7) (i.e. Corollary 1 part (b) and Theorem 1) permit us to describe the distribution of S_n in (1) for $r = 2$ where the rows $(X_{k,1}, X_{k,2})$, $k \geq 1$, are no longer independent, but arise in the context of a Pólya urn sampling scheme. We briefly describe such a context, but refer the reader to [8] for a general reference. In such schemes, an urn initially contains b black balls and w white balls. At step k, a ball is drawn randomly and uniformly from the urn, and returned to the urn with s ($s > 0$) balls of the same colour. This generates the sequence of rows $(X_{k,1}, X_{k,2})$, $k \geq 1$, where $X_{k,1} = 1 - X_{k,2}$ is 1 or 0 according to whether the colour of the ball selected in the kth draw is black or white, respectively. Hence, S_n given in (1), and as described in Example 1, represents the number of consecutive pairs of draws with matching colours, among the first $n + 1$ draws. Now, in such a case, de Finetti's representation theorem (see, e.g. [3]) implies the representation

$$X_{1,1}, \ldots, X_{n+1,1} \quad \text{i.i.d. Bernouilli}(\theta), \quad \text{with} \quad \theta \sim \text{Beta}\left(\frac{b}{s}, \frac{w}{s}\right). \tag{12}$$

We thus obtain the following result.

Corollary 3. *Let*

$$c_{n,j,k} = \binom{j}{\lceil n/2 \rceil - k - 1}\binom{n}{2j+1} 1_{\{\lceil n/2 \rceil - k - 1 \leq j \leq \lceil n/2 \rceil - 1\}} 1_{\{0 \leq k \leq \lceil n/2 \rceil - 1\}}$$

and

$$B_{n,k,m} = \sum_j c_{n,j,k} \sum_{i=0}^{j-(\lceil n/2 \rceil - k - 1)} \binom{j - (\lceil n/2 \rceil - k - 1)}{i} (-1)^i 4^{j+m-i} \times \frac{(b/s)_{j+m-i}(w/s)_{j+m-i}}{((b+w)/s)_{2(j+m-i)}} \tag{13}$$

for positive integers n and k, and $m = 0$ or 1, where $(a)_x$ is the usual Pochhammer function defined as $(a)_0 = 1$, and $(a)_x = \prod_{i=0}^{x-1}(a+i)$ for $x = 1, 2, \ldots$. Then, for a Pólya urn as described above with parameters b, w, and s, we have

(a) *for odd n,*

$$\mathrm{P}(S_n = 2k) = \frac{B_{n,k,1}}{2^n},$$

$$\textit{and} \quad \mathrm{P}(S_n = 2k+1) = \frac{1}{2^n}[2B_{n,k,0} - B_{n,k,1} + B_{n-1,k,1} - B_{n-1,k-1,1}];$$

(b) *for even n,*

$$\mathrm{P}(S_n = 2k) = \frac{1}{2^n}[2B_{n,k-1,0} - B_{n,k-1,1} + B_{n-1,k,1} - B_{n-1,k-1,1}],$$

$$\textit{and} \quad \mathrm{P}(S_n = 2k+1) = \frac{B_{n,k,1}}{2^n}.$$

Proof. It follows directly from representation (12) and Theorem 1 that the probability function of S_n in the context here of a Pólya urn is given by the above equations with $B_{n,k,m} = \mathrm{E}[Z^m a_{n,k}(Z)]$, where $Z \stackrel{\mathrm{D}}{=} 4\theta(1-\theta)$ and $\theta \sim \text{Beta}(b/s, w/s)$. It remains to show that (13) is

https://doi.org/10.1239/jap/1276784897 Published online by Cambridge University Press

a valid expression for $B_{n,k,m}$. Indeed, we have

$$\begin{aligned}\mathrm{E}[Z^m a_{n,k}(Z)] &= \sum_j c_{n,j,k}\,\mathrm{E}[Z^{m+\lceil n/2\rceil-k-1}(1-Z)^{j-(\lceil n/2\rceil-k-1)}] \\ &= \sum_j c_{n,j,k} \sum_{i=0}^{j-(\lceil n/2\rceil-k-1)} \binom{j-(\lceil n/2\rceil-k-1)}{i}(-1)^i\,\mathrm{E}[(4\theta(1-\theta))^{j+m-i}] \\ &= B_{n,k,m}, \end{aligned} \tag{14}$$

by the evaluation

$$\mathrm{E}(Z^u(1-Z)^v)] = \frac{(a_1)_u(a_2)_v}{(a_1+a_2)_{u+v}} \quad \text{for } Z\sim \mathrm{Beta}(a_1,a_2),\ u,v\geq 0. \tag{15}$$

Remark 2. For the cases where $b = w$ (i.e. an equal number of black and white balls initially in the Pólya urn), the expression of $B_{n,k,m}$ in Corollary 3 is also equivalent to the simpler expression

$$B_{n,k,m} = \left(\frac{2b}{s}-1\right)_{m+\lceil n/2\rceil-k-1} \sum_j c_{n,j,k}\frac{(1/2)_{j-(\lceil n/2\rceil-k-1)}}{(2b/s-1/2)_{j+m}}.$$

This is verified by establishing that (i) $Z = 4\theta(1-\theta) \sim \mathrm{Beta}(2b/s-1, \frac{1}{2})$ whenever $\theta \sim \mathrm{Beta}(b', b')$, and (ii) by directly evaluating (14) via (15).

4. Concluding remarks

We have introduced a rich collection of problems relative to Bernoulli arrays. We have proceeded with explicit representations for the probability generating and mass functions in the case of multinomial rows of length $r = 2$. Various properties and implications have been discussed, including applications to Pólya urns. Further analysis or extensions for $r > 2$ with or without assumption (7) would be most useful and welcome.

Acknowledgements

We gratefully acknowledge NSERC of Canada, who supported the research work of Éric Marchand, as well l'Université de Sherbrooke, who provided partial financial support while Ait Aoudia was a postdoctoral fellow there. We thank an anonymous referee for constructive comments which improved the paper.

References

[1] Ait Aoudia, D. and Marchand, É. (2009). On the number of runs of Bernoulli arrays. Res. Rep. 76, Université de Sherbrooke. Available at: http://www.usherbrooke.ca/mathematiques/recherche/publications/rapports-recherche/.
[2] Ardakov, K. (1997). Powers and other functions of 2×2 matrices. *Math. Gazette* **4,** 434–431.
[3] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II. John Wiley, New York.
[4] Holst, L. (2007). Counts of failure strings in certain Bernouilli sequences. *J. Appl. Prob.* **44,** 824–830.
[5] Holst, L. (2008). The number of two-consecutive successes in a Hoppe–Polyá urn. *J. Appl. Prob.* **45,** 901–906.
[6] Huffer, W. F., Sethuraman, J. and Sethuraman, S. (2009). A study of counts of Bernoulli strings via conditional Poisson processes. *Proc. Amer. Math. Soc.* **137,** 2125–2134.
[7] Joffe, A., Marchand É., Perron, F. and Popadiuk, P. (2004). On sums of products of Bernoulli variables and random permutations. *J. Theoret. Prob.* **17,** 285–292.

[8] Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Applications*. John Wiley, New York.
[9] Móri, T. F. (2001). On the distribution of sums of overlapping products. *Acta Sci. Math.* **67,** 833–841.
[10] Sethuraman, J. and Sethuraman, S. (2004). On counts of Bernoulli strings and connections to rank orders and random permutations. *A Festschrift for Herman Rubin* (IMS Lecture Notes Monogr. Ser. **45**), Institute of Mathematical Statistics, Beachwood, OH, pp. 140–152.
[11] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Prob.*, Vol. II, University of California Press, Berkeley, pp. 583–602.