COUNTS OF FAILURE STRINGS IN CERTAIN BERNOULLI SEQUENCES

LARS HOLST

Department of Mathematics, Royal Institute of Technology
SE–100 44 Stockholm, Sweden

E-mail: lholst@math.kth.se

October 6, 2006

Abstract

In a sequence of independent Bernoulli trials the probability for success in the k:th trial is p_k, $k = 1, 2, \ldots$. The number of strings with given number of failures between two subsequent successes is studied. Explicit expressions for distributions and moments are obtained for $p_k = a/(a + b + k - 1)$, $a > 0$, $b \geq 0$. Also the limit behaviour of the longest failure string in the first n trials is considered. For $b = 0$ the strings correspond to cycles in random permutations.

Keywords: Binomial moments; Ewens Sampling Formula; Hoppe's urn; Poisson distribution; Poisson–Dirichlet distribution; Pólya's urn; random permutations; records; spacings; sums of indicators

AMS 2000 SUBJECT CLASSIFICATION: PRIMARY 60C05
SECONDARY 60K99

Running title: Counts of failure strings

1 Introduction

In an infinite sequence of independent Bernoulli trials the probability for success in the k:th trial is p_k for $k = 1, 2, \ldots$. A d–string is a string $SF \ldots FS$ of $d - 1$ failures between two subsequent successes. We will study the number of such strings. Explicit results are obtained for $p_k = a/(a+b+k-1)$, $a > 0$, $b \geq 0$. To our knowledge only special cases have been studied previously.

For $a = 1, b = 0$, that is $p_k = 1/k$, 1–strings correspond to double records in a record sequence. Hahlin (1995) proved that the total number of such records is Po(1) (Poisson distributed with mean 1). After that, an unpublished proof by Diaconis inspired a number of studies on 1–strings, see Chern *et al* (2000), Mori (2001), Joffe *et al* (2004) and the references therein.

Sethuraman and Sethuraman (2004) studied d–strings for $a = 1$, $b > 0$, and obtained the joint distribution of the number of d–strings for $d = 1, 2, \dots$. For $a > 0$ and $b = 0$, d–strings are closely connected with cycle lengths in random permutations, see e.g. Arratia, Barbour and Tavaré (2003) page 95.

In Section 2 we introduce notations and derive recursions for the binomial moments of the number of d–strings in a finite sequence for general p_k's. The special case $p_k = a/(a + k - 1)$, connected with random permutations, is studied in Section 3. In Section 4 we derive the joint distribution of the total number of d–strings, $d = 1, 2, \dots$, and study the limit behaviour of the longest failure string in the first n trials in an infinite Bernoulli sequence with $p_k = a/(a + b + k - 1)$.

2 General case: notations and moments

In the following $I_1, I_2, \dots$ is a sequence of independent Bernoulli random variables, I_k is Be(p_k), that is

$$P(I_k = 1) = 1 - P(I_k = 0) = p_k.$$

The number of d–strings in the first n trials is

$$M_{dn} = \sum_{k=1}^{n-d} I_k(1 - I_{k+1}) \cdots (1 - I_{k+d-1})I_{k+d}$$

with mean

$$E(M_{dn}) = \sum_{k=1}^{n-d} p_k(1 - p_{k+1}) \cdots (1 - p_{k+d-1})p_{k+d}.$$

Note that $M_{dn} = 0$ for $d \geq n$ and $\sum_{j=1}^{n-1} jM_{jn} \leq n - 1$. Implicitly, the following result gives the distribution of $(M_{1n}, \dots, M_{dn})$.

Proposition 2.1 *For the binomial moments*

$$f_n(r_1, \dots, r_d) = E\Big(\binom{M_{1n}}{r_1} \cdots \binom{M_{dn}}{r_d}\Big)$$

with $d \leq n - 1$ and $\sum_{j=1}^{d} jr_j \leq n - 1$, the recursion holds:

$$f_{n+1}(r_1,\dots,r_d) = f_n(r_1,\dots,r_d)$$
$$+p_{n+1}\big[f_n(r_1-1,r_2,\dots,r_d) - (1-p_n)f_{n-1}(r_1-1,r_2,\dots,r_d)\big]$$
$$+p_{n+1}(1-p_n)\big[f_{n-1}(r_1,r_2-1,r_3,\dots) - (1-p_{n-1})f_{n-2}(r_1,r_2-1,r_3,\dots)\big] + \dots$$
$$+p_{n+1}(1-p_n)\cdots(1-p_{n-d+2})$$
$$\times\big[f_{n-d+1}(r_1,\dots,r_{d-1},r_d-1) - (1-p_{n-d+1})f_{n-d}(r_1,\dots,r_{d-1},r_d-1)\big].$$

Proof. Using generating functions and the independence between the I_k's we get

$$E\Big[t_1^{M_{1,n+1}}\cdots t_d^{M_{d,n+1}}\Big]$$
$$= E\Big[t_1^{M_{1n}}\cdots t_d^{M_{dn}}\big(1+(t_1-1)I_nI_{n+1}\big)\big(1+(t_2-1)I_{n-1}(1-I_n)I_{n+1}\big)\cdots\Big]$$
$$= E\Big[t_1^{M_{1n}}\cdots t_d^{M_{dn}}\big(1+(t_1-1)I_nI_{n+1}+(t_2-1)I_{n-1}(1-I_n)I_{n+1}+\dots\big)\Big]$$
$$= E\Big[t_1^{M_{1n}}\cdots t_d^{M_{dn}}\Big] + (t_1-1)p_{n+1}E\Big[t_1^{M_{1n}}\cdots t_d^{M_{dn}}\big(1-(1-I_n)\big)\Big]$$
$$+(t_2-1)p_{n+1}E\Big[t_1^{M_{1n}}\cdots t_d^{M_{dn}}\big(1-(1-I_{n-1})\big)\big(1-I_n\big)\Big] + \dots$$
$$= E\Big[t_1^{M_{1n}}\cdots t_d^{M_{dn}}\Big] + (t_1-1)p_{n+1}\Big[E\Big(t_1^{M_{1n}}\cdots t_d^{M_{dn}}\Big) - (1-p_n)E\Big(t_1^{M_{1,n-1}}\cdots t_d^{M_{d,n-1}}\Big)\Big]$$
$$+(t_2-1)p_{n+1}(1-p_n)\Big[E\Big(t_1^{M_{1,n-1}}\cdots t_d^{M_{d,n-1}}\Big) - (1-p_{n-1})E\Big(t_1^{M_{1,n-2}}\cdots t_d^{M_{d,n-2}}\Big)\Big] + \dots .$$

Expansion in series around $t_1 = 1,\dots,t_d = 1$ proves the assertion. □

Including the string $SF\cdots F$ with $d-1$ failures after the last success in the count we get the random variable

$$N_{dn} = M_{dn} + I_{n-d+1}(1-I_{n-d+2})\cdots(1-I_n)$$

with mean

$$E(N_{dn}) = E(M_{dn}) + p_{n-d+1}(1-p_{n-d+2})\cdots(1-p_n).$$

Proposition 2.2 *For $p_{n+1} > 0$ it holds for the binomial moments:*

$$E\Big(\binom{N_{1n}}{r_1}\cdots\binom{N_{dn}}{r_d}\Big) = E\Big(\binom{M_{1n}}{r_1}\cdots\binom{M_{dn}}{r_d}\Big)$$
$$+\frac{1}{p_{n+1}}\Big[E\Big(\binom{M_{1,n+1}}{r_1}\cdots\binom{M_{d,n+1}}{r_d}\Big) - E\Big(\binom{M_{1n}}{r_1}\cdots\binom{M_{dn}}{r_d}\Big)\Big].$$

Proof. By the law of total probability we have

$$E\Big(t_1^{M_{1,n+1}}\cdots t_d^{M_{d,n+1}}\Big)=p_{n+1}E\Big(t_1^{N_{1n}}\cdots t_d^{N_{dn}}\Big)+(1-p_{n+1})E\Big(t_1^{M_{1n}}\cdots t_d^{M_{dn}}\Big),$$

from which the assertion follows by expansion in series. □

In an infinite sequence the total number of d–strings

$$M_{d\infty}=\sum_{k=1}^{\infty}I_k(1-I_{k+1})\cdots(1-I_{k+d-1})I_{k+d}<+\infty$$

with probability one if and only if

$$\sum_{k=1}^{\infty}p_k(1-p_{k+1})\cdots(1-p_{k+d-1})p_{k+d}<+\infty.$$

Indeed, by splitting the series for $M_{d\infty}$ into $d+1$ (independent) series this follows from the Borel-Cantelli lemmas, cf. Mori (2001) page 834.

3 The case $p_k=a/(a+k-1)$

Following Knuth (1992) we denote descending and ascending factorials by

$$x^{\underline{n}}=x(x-1)\cdots(x-n+1),\quad x^{\overline{n}}=x(x+1)\cdots(x+n-1)=\sum_{k=1}^{n}\begin{bmatrix}n\\k\end{bmatrix}x^k,$$

where $\begin{bmatrix}n\\k\end{bmatrix}$ is a cycle number or signless Stirling number of the first kind. We assume in the rest of this section that $p_k=a/(a+k-1)$ with $a>0$. Closed simple formulas can be obtained for the binomial moments. Note that $\sum_{j=1}^{n-1}jM_{jn}\le n-1$ and $\sum_{j=1}^{n}jN_{jn}=n$ (with probability one).

Proposition 3.1 *For $m=\sum_{j=1}^{d}jr_j\le n-1$:*

$$E\Big(\binom{M_{1n}}{r_1}\cdots\binom{M_{dn}}{r_d}\Big)=\frac{(n-1)^{\underline{m}}}{(a+n-1)^{\underline{m}}}\prod_{j=1}^{d}\frac{(a/j)^{r_j}}{r_j!}.$$

Proof. By telescoping sums we get

$$E(M_{dn})=\sum_{k=1}^{n-d}\frac{a}{a+k-1}\Big(1-\frac{a}{a+k}\Big)\cdots\Big(1-\frac{a}{a+k+d-2}\Big)\frac{a}{a+k+d-1}$$

$$= \frac{a}{d}\sum_{k=1}^{n-d}\Big[\big(1-\frac{a}{a+k+d-1}\big)-\big(1-\frac{a}{a+k-1}\big)\Big]\Big(1-\frac{a}{a+k}\Big)\cdots\Big(1-\frac{a}{a+k+d-2}\Big)$$

$$= \frac{a}{d}\Big(1-\frac{a}{a+n-d}\Big)\cdots\Big(1-\frac{a}{a+n-1}\Big) = \frac{a}{d}\frac{(n-1)^{\underline{d}}}{(a+n-1)^{\underline{d}}}.$$

Hence the assertion holds for $E(M_{dn})$. By elementary calculations the recursion in Proposition 2.1 is verified. The proof is finished by induction. □

Proposition 3.2 *For $d \le n$ and $m = \sum_{j=1}^{d} jr_j \le n$:*

$$E\Big(\binom{N_{1n}}{r_1}\cdots\binom{N_{dn}}{r_d}\Big) = \frac{n^{\underline{m}}}{(a+n-1)^{\underline{m}}}\prod_{j=1}^{d}\frac{(a/j)^{r_j}}{r_j!},$$

and for $\sum_{j=1}^{n} jx_j = n$:

$$P(N_{1n}=x_1,\ldots,N_{nn}=x_n) = \frac{n!}{a^{\overline{n}}}\prod_{j=1}^{n}\frac{(a/j)^{x_j}}{x_j!}.$$

Proof. Using Propositions 2.2 and 3.1 the first assertion follows from an elementary calculation.

We have using generating functions that

$$E\big(t_1^{N_{1n}}\cdots t_n^{N_{nn}}\big) = E\big((1+(t_1-1))^{N_{1n}}\cdots(1+(t_n-1))^{N_{nn}}\big)$$

$$= \sum E\Big(\binom{N_{1n}}{r_1}\cdots\binom{N_{nn}}{r_n}\Big)(t_1-1)^{r_1}\cdots(t_n-1)^{r_n}$$

$$= \sum\sum E\Big(\binom{N_{1n}}{r_1}\cdots\binom{N_{nn}}{r_n}\Big)(-1)^{r_1-x_1}\cdots(-1)^{r_n-x_n}\binom{r_1}{x_1}\cdots\binom{r_n}{x_n}t_1^{x_1}\cdots t_n^{x_n}.$$

As $\sum_1^n jN_{jn} = n$ the binomial moments disappears for $\sum_1^n jr_j > n$. Therefore for $\sum_1^n jx_j = n$ we have $r_j = x_j$ in the summation. Thus

$$P(N_{1n}=x_1,\ldots,N_{nn}=x_n) = E\Big(\binom{N_{1n}}{x_1}\cdots\binom{N_{nn}}{x_n}\Big),$$

proving the second assertion. □

The distribution of $(N_{1n},\ldots,N_{nn})$ is the famed Ewens Sampling Formula. Furthermore, N_{dn} is the number of d–strings in $1I_2I_3\ldots I_n1$. Using this, the last proposition can be derived by combinatorial arguments, cf. Arratia *et al* (2003) page 95. In that context N_{dn} is interpreted as the number of cycles of length d in a random

permutation of $1, 2, \ldots, n$ biased by a^{K_n}, where $K_n = \sum_{k=1}^{n} I_k$ is the number of cycles with the distribution

$$P(K_n = j) = \begin{bmatrix} n \\ j \end{bmatrix} \frac{a^j}{a^{\overline{n}}}, \quad j = 1, 2, \ldots, n.$$

The moment convergence

$$E\Big(\binom{M_{1n}}{r_1} \cdots \binom{M_{dn}}{r_d}\Big) \to \prod_{j=1}^{d} \frac{(a/j)^{r_j}}{r_j!}, \quad n \to \infty,$$

implies the following result, well known for a–biased random permutations, see Arratia *et al* (2003) page 96.

Proposition 3.3 *The number of strings $M_{1\infty}, M_{2\infty}, \ldots$ are independent Poisson random variables with $E(M_{d\infty}) = a/d$.*

4 The case $p_k = a/(a + b + k - 1)$

In this section we assume that $p_k = a/(a + b + k - 1)$ with $a > 0$ and $b > 0$. Clearly

$$M_{d\infty} = \sum_{k=1}^{\infty} I_k (1 - I_{k+1}) \cdots (1 - I_{k+d-1}) I_{k+d} < +\infty$$

with probability one. Mori (2001) derived the distribution of $M_{1\infty}$. For the special case $a = 1$ Sethuraman and Sethuraman (2004) obtained the joint distribution of $M_{1\infty}, M_{2\infty}, \ldots$. Using different methods we generalize their result to any $a > 0$. Let U be Beta(a, b), that is a random variable with density

$$f_U(u) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1 - u)^{b-1}, \quad 0 < u < 1.$$

Theorem 4.1 *Conditional on a Beta(a,b) random variable U, the number of strings $M_{1\infty}, M_{2\infty}, \ldots$ are independent Poisson random variables with*

$$E(M_{d\infty}|U) = \frac{a}{d}\,(1 - (1 - U)^d), \quad d = 1, 2, \ldots \quad .$$

Proof. We introduce the following mixture of Pólya's and Hoppe's urn models. An urn contains initially one white and one black ball of weights a and b respectively. Balls are drawn at random proportional to weights. The white and the black ball are replaced together with a new ball of a colour not present in the urn, other balls

are replaced together with one new ball of the same colour. All new balls have weight one. Obviously, the probability of drawing the white ball at drawing k is $p_k = a/(a+b+k-1)$.

Generate a sequence of W's and B's. We get a W if drawing the white ball or a ball of a colour emanating from a draw of the white, else we get a B. This sequence is as drawing from an ordinary Pólya urn. Note that the sequence is exchangeable. Therefore, by de Finetti's theorem the sequence can be thought of as having been generated by first observing the Beta(a,b) random variable U and then, conditional on the outcome $U=u$, generating a sequence of independent Be(u) random variables, with 1 corresponding to W and 0 to B.

In the *subsequence* of W's in the original sequence $I_1, I_2, \dots$ the probability of getting the white ball at the j:th trial is $p_j^* = a/(a+j-1)$. According to Proposition 3.3 the number of d–strings in the subsequence, $M_{d\infty}^*$, is Po(a/d) and $M_{1\infty}^*, M_{2\infty}^*, \dots$ are independent.

Recall the following well known fact. If the random variable ξ is Po(μ) and independent of the independent Be(p) random variables $\varepsilon_1, \varepsilon_2, \dots$, then $\sum_{j=1}^{\xi} \varepsilon_j$ and $\sum_{j=1}^{\xi}(1-\varepsilon_j)$ are independent Po(μp) and Po$(\mu(1-p))$ respectively.

Consider the 1–strings in the subsequence of W's. Each such 1–string is also a 1–string in the original sequence, provided it was not interrupted by a B. The probability for interruption is $1-u$. As $M_{1\infty}^*$ is Po(a) it follows from the fact above and the independence, that conditional on $U=u$, the total number of 1–strings in the original sequence, $M_{1\infty}$, is Po(au) and independent of $M_{1\infty}^* - M_{1\infty}$ which is Po$(a(1-u))$.

For the 2–strings we can argue in a similar way as above. Conditional on $U=u$, the random variable $M_{2\infty}$ is Poisson with mean

$$\frac{a}{2}\,u^2 + a(1-u)u = \frac{a}{2}(1-(1-u)^2)$$

and independent of $M_{1\infty}$. The argument extends, $M_{d\infty}$ is Poisson with (conditional) mean

$$\frac{a}{d}u^d + \frac{a}{d-1}\binom{d-1}{1}u^{d-1}(1-u) + \frac{a}{d-2}\binom{d-1}{2}u^{d-2}(1-u)^2 + \cdots + au(1-u)^{d-1}$$

$$= \frac{a}{d}\left(1-(1-u)^d\right)$$

and independent of $M_{1\infty}, M_{2\infty}, \dots, M_{d-1,\infty}$. □

Finally, consider long strings of failures. Let the last success in the first n trials occur at trial $n+1-A_{1n}$; if there is no success set $A_{1n}=0$. We have for $j=1,2,\ldots,n$

$$P(A_{1n}>j)=\Big(1-\frac{a}{a+b+n-j}\Big)\cdots\Big(1-\frac{a}{a+b+n-1}\Big)$$

$$=\frac{(b+n-j)^{\overline{j}}}{(a+b+n-j)^{\overline{j}}}=\frac{\Gamma(b+n)}{\Gamma(b+n-j)}\frac{\Gamma(a+b+n-j)}{\Gamma(a+b+n)}.$$

For $j,n\to\infty$ such that $j/n\to x$, $0<x<1$, Stirling's formula gives

$$P(A_{1n}/n>j/n)\to(1-x)^a,\quad n\to\infty,$$

that is A_{1n}/n converges in distribution to Beta$(1,a)$.

In a similar way we find for the number of trials between the last and the second last success, A_{2n}, that

$$(A_{1n},A_{2n})/n\to(U_1,(1-U_1)U_2),\quad n\to\infty,$$

in distribution, where U_1,U_2 are independent Beta$(1,a)$ random variables. The procedure can be repeated in like manner.

The limit behaviour of the long strings is as if $A_{1n},A_{2n},\ldots$ had been cycle lengths in an a–biased random permutation, see Arratia *et al* (2003) Section 5.4. The limit distribution of the size ordered A's is the Poisson–Dirichlet distribution with parameter a. In particular we have:

Theorem 4.2 *For the longest string of failures in the first n trials:*

$$\max(A_{1n},A_{2n},\ldots)/n\to L_1=\max(U_1,(1-U_1)U_2,(1-U_1)(1-U_2)U_3,\ldots),\ n\to\infty,$$

in distribution, where $U_1,U_2,\ldots$ are independent Beta$(1,a)$ random variables.

Various formulas connected with the random varible L_1 can be found in Arratia *et al* (2003) Section 5.5.

References

[1] Arratia, R., Barbour, A.D. and Tavaré, S. (2003). *Logarithmic Combinatorial Structures: a Probabilistic Approach.* European Mathematical Society Publishing House, ETH-Zentrum, Zürich.

[2] Chern, H.-H., Hwang, H.-K. and Yeh, Y.-N. (2000). Distribution of the number of consecutive records. *Random Structures & Algorithms,* **17** 169–196.

[3] Hahlin, L.O. (1995). Double Records. *Uppsala University Department of Mathematics Report,* **1995:12**. Licentiat thesis.

[4] Joffe, A., Marchand, E., Perron, F. and Popadiuk, P. (2004). On sums of products of Bernoulli variables and random permutations. *Journal of Theoretical Probability,* **17** 285–292.

[5] Knuth, D. (1992). Two notes on notations. *The American Mathematical Monthly,* **99** 403–422.

[6] Mori, T.F. (2001). On the distribution of sums of overlapping products. *Acta Scientiarum Mathematica (Szeged),* **67** 833–841.

[7] Sethuraman, J. and Sethuraman, S. (2004). On counts of Bernoulli strings and connections to rank orders and random permutations. In *A festschrift for Herman Rubin. IMS Lecture Notes Monograph Series,* **45** 140–152, Institute of Mathematical Statistics, Beachwood, Ohio.