Ann Inst Stat Math (2017) 69:489–495
DOI 10.1007/s10463-015-0549-2

Number of appearances of events in random sequences: a new generating function approach to *Type II* and *Type III* runs

Yong Kong[1]

Received: 5 November 2014 / Revised: 5 August 2015 / Published online: 30 December 2015
© The Institute of Statistical Mathematics, Tokyo 2015

Abstract Distributions of runs of length at least k (*Type II* runs) and overlapping runs of length k (*Type III* runs) are derived in a unified way using a new generating function approach. A new and more compact formula is obtained for the probability mass function of the *Type III* runs.

Keywords Runs statistics · Generating function · Asymptotic distributions · Factorial moments · Wilf-Zeilberger method

1 Introduction

Runs statistics have found many applications in various fields and have attracted attentions of many researchers (Balakrishnan and Koutras 2002). In the long history of study of run-related statistics, many of the results were obtained by ingenious combinatorial method. These traditional methods are ad hoc and not easy to generalize. Several unified methods have been devised to overcome the combinatorial difficulties. One of them is the finite Markov chain imbedding approach (Fu and Koutras 1994; Koutras and Alexandrou 1995). This approach projects the original problems into a Markov chain, and thus expresses the problem under study in terms of transition matrices of the Markov chain. The method is quite versatile and easy to be adapted to handle Markov-dependent multi-state trials. Another elegant approach is the method of

Electronic supplementary material The online version of this article (doi:10.1007/s10463-015-0549-2) contains supplementary material, which is available to authorized users.

✉ Yong Kong
yong.kong@yale.edu

[1] Department of Molecular Biophysics and Biochemistry, W.M. Keck Foundation Biotechnology Resource Laboratory, Yale University, 333 Cedar Street, New Haven, CT 06510, USA

Koutras (1997). The method derives the double generating function of the number of appearances of a pattern using the generating function of the waiting time for the rth appearance of the pattern, which is usually easier to obtain.

Recently, we used a different unified approach to study statistics of runs of multi-state systems by utilizing matrix formulation to derive the generating function (GF) of the whole system from the proper GFs of individual objects (Kong 2006, 2014, 2015). Originally, we applied the GF method mainly to "exact length runs" (Mood 1940), as the method naturally applies to this definition of runs by considering each run as a "block" that can be represented readily by the GF of the individual object (Kong 2006, 2015). The method is actually versatile enough to handle other definitions of runs: in Kong (2014), for example, the method is applied to *Type I* runs. In this paper, we use the method to study *Type II* and *Type III* runs. For definitions, see Balakrishnan and Koutras (2002, p.139). In addition to show that this new GF method can deal with different definitions of runs in a simple and unified way, we also derive a new and more compact formula for the probability mass function (pmf) of *Type III* runs (Eq. 9). Generalizations of the method to include interactions between neighboring runs and Markov-dependent processes have also been developed and will be published elsewhere.

1.1 Definitions and notation

Throughout the paper, we assume that the elements in the sequence are independent of each other, with probability p_i to appear for the ith object and $q_i = 1 - p_i$. The length of the sequence is denoted as n. The mth falling factorial powers are defined as $x^{(m)} = x(x-1)\dots(x-m+1)$ for $m > 0$, with $x^{(0)} = 1$. By this definition, we have $x^{(m)} = m!\binom{x}{m}$. The rth factorial moment of a random variable X is denoted as $E(X^{(r)})$, and pmf as $P(X)$. For a series $f(x)$ in powers of x, we use the usual notation $[x^n]f(x)$ to denote the coefficient of x^n in the series. The numbers of *Type II* and *Type III* runs are denoted as $G_{n,k}$ and $M_{n,k}$ respectively. When there is no ambiguity, we will choose the first object to discuss its distributions and the subscript in $G_{n,k}^{\{1\}}$, etc. will be omitted.

The method applies to multiple-state systems naturally, so we will not treat the binary case specially.

1.2 General strategy: bivariate GF, factorial moment, and pmf

Here, we outline the general strategy to use the GF method for different kind problems. We first write down the GFs of individual objects based on the particular problem we want to solve, then use these individual GFs to get the bivariate GF of the whole system using Thm (2) of Kong (2006), from which the rth factorial moment is obtained (Eqs. 1, 2), which in turn yields the pmf (Eq. 3).

The rth factorial moment of a random variable X, $E_n(X^{(r)}) = \sum_x x^{(r)} P(X = x)$ ($r \geq 0$), when considered as a sequence of n, has its generating function $\mathcal{E}_r(z)$ defined as $\mathcal{E}_r(z) = \sum_n E_n(X^{(r)}) z^n$. The following results can be used to obtain $\mathcal{E}_r(z)$ and $E_n(X^{(r)})$ from the bivariate GF $G(z, u)$ of the whole system:

$$\mathcal{E}_r(z) = r![u^r]G(z, u + 1), \tag{1}$$

$$E_n(X^{(r)}) = [z^n]\mathcal{E}_r(z) = r![z^n u^r]G(z, u + 1). \tag{2}$$

The proof of Eq. (1) can be obtained directly from the definition of $E_n(X^{(r)})$ by binomial expansion. The pmf $P(X = x)$ can be obtained from the expression of $E_n(X^{(r)})$ as:

$$P_n(X = x) = (-1)^x \sum_{r=x}^{\infty} (-1)^r \binom{r}{x} \frac{E_n(X^{(r)})}{r!}, \tag{3}$$

which can be proved from the relation $\mathcal{M}_n(t) = \mathcal{P}_n(t + 1)$, where $\mathcal{M}_n(t)$ is the traditional factorial moment generating function (pmgf) defined as $\mathcal{M}_n(t) = \sum_{r=0}^{\infty} \frac{E_n(X^{(r)})}{r!} t^r$, and $\mathcal{P}_n(t)$ is the probability generating function (pgf) defined as $\mathcal{P}_n(t) = \sum_x P_n(X = x) t^x = E_n(t^X)$.

2 *Type II* runs (runs of length *at least k*)

To get the pmf of *Type II* runs, we can use the following g_i with variable z tracking the number of total elements of the system and variable u tracking the number of runs of length at least k of the first object:

$$\begin{aligned} g_1 &= \sum_{i=1}^{k-1} (p_1 z)^i + u \sum_{i=k}^{\infty} (p_1 z)^i = \frac{p_1 z - (p_1 z)^k}{1 - p_1 z} + \frac{u(p_1 z)^k}{1 - p_1 z}, \\ g_j &= \sum_{i=1}^{\infty} (p_j z)^i = \frac{p_j z}{1 - p_j z}, \qquad j \neq 1. \end{aligned} \tag{4}$$

To get the joint distribution of $G_{n,k_1}^{\{1\}}$ and $G_{n,k_2}^{\{2\}}$, also called trinomial distributions of order (k_1, k_2), we can simply add an additional variable u_2 in the g_2 of Eq. (4) to track the number of runs of length at least k_2 for the second object,

$$g_2 = \sum_{i=1}^{k_2-1} (p_2 z)^i + u_2 \sum_{i=k_2}^{\infty} (p_2 z)^i = \frac{p_2 z - (p_2 z)^{k_2}}{1 - p_2 z} + \frac{u_2 (p_2 z)^{k_2}}{1 - p_2 z}.$$

Following the strategy mentioned above, all interesting properties of *Type II* runs can be obtained, including the bivariate GF $G(z, u)$ (or $G(z, u_1, u_2)$ for joint distributions), $\mathcal{E}_r(z)$, $E(G_{n,k}^{(r)})$, and pmfs as well as the mean and (co)variance. The mean and variance are linear in n when n is large (Hirano and Aki 1993) [see also (Balakrishnan and Koutras (2002), p.164)]. A large family of distributions with rational GFs have this

property, as stated by the singularity perturbation theorem (Flajolet and Sedgewick 2009, TheoremIX.9). The theorem also states that these distributions have Gaussians as the limiting distribution.

3 *Type III* runs (*overlapping* runs)

For *Type III* runs (*overlapping* runs) of order k, the following g_i can be used:

$$g_1 = \sum_{i=1}^{k-1}(p_1 z)^i + \sum_{i=1}^{\infty} u^i (p_1 z)^{i+k-1} = \frac{p_1 z - (p_1 z)^k}{1 - p_1 z} + \frac{u(p_1 z)^k}{1 - u p_1 z},$$
$$g_j = \sum_{i=1}^{\infty}(p_j z)^i = \frac{p_j z}{1 - p_j z}, \qquad j \neq 1. \tag{5}$$

The GF $\mathcal{E}_r(z)$ of factorial moments of order k *Type III* runs can be obtained as

$$\mathcal{E}_r(z) = \frac{r!\left[1 - p_1 z - q_1(1 - (p_1 z)^k)\right]^{r-1} p_1^k z^{r+k-1}}{(1 - p_1 z)^{r-1}(1 - z)^{r+1}}, \qquad r \geq 1. \tag{6}$$

From Eq. (6), the expression of the rth factorial moments of $M_{n,k}$ can be obtained.

Proposition 1 *The rth factorial moment of $M_{n,k}$ is given by*

$$E\left(M_{n,k}^{(r)}\right) = r! \sum_{i=0}^{r-1} \sum_{j=0}^{n-k-ki-1} \binom{n-k-ki-j-1}{r-2} \binom{r-1}{i} \binom{i+j+1}{j} p_1^{n-i-j} q_1^i \tag{7}$$

for $r \geq 1$. When $r = 0$, $E(M_{n,k}^{(0)}) = 1$.

Proof For $r \geq 1$, Eq. (6) can be rewritten as

$$\mathcal{E}_r(z) = \frac{r!(p_1 z)^{r+k-1}}{(1 - p_1 z)^{r-1}(1 - z)^2}\left[1 + \frac{q_1 (p_1 z)^k}{p_1(1 - z)}\right]^{r-1}.$$

The part on the right with power of $r-1$ can be expanded using binomial theorem, then the $(1 - z)^2$ in the denominator can be pulled in and combined with the $1/(1 - z)^i$ terms in the binomial expansion. After two more expansions of $1/(1 - z)^{i+2}$ and $1/(1 - p_1 z)^{r-1}$, Eq. (7) can be obtained. The case for $r = 0$ is trivial. □

A formula for $E(M_{n,k}^{(r)})$, credited to Charalambides, was mentioned in Balakrishnan and Koutras (2002, p.167), but its proof does not seem to be published in literature.

With Eq. (7), a closed form formula for $P(M_{n,k} = x)$ that involves three sums can be obtained directly. This expression, however, can be simplified. To simplify the expression further, we need the following identity of binomial coefficients, whose proof is in Appendix 1.

Lemma 1 *For $u \geq 0$ and $v \geq 0$, the following identity holds when $m \geq 0$,*

$$S_m = \sum_{r=u-1}^{m} \binom{m}{r}\binom{r+1}{u}\binom{r+2}{v}(-1)^r = (-1)^m \frac{(u+1)(uv-u+2m+2)m!}{(m-u+1)!(m-v+2)!(u+v-m)!}. \tag{8}$$

When $m = -1$, $S_m = -1$ only when $u = 0$ and $v \in \{0, 1\}$; otherwise $S_m = 0$.

Using this identity, we can now derive a simplified formula for the pmf of $M_{n,k}$.

Theorem 1 *The pmf of the* Type III *runs of length k is given by, when $x > 1$,*

$$P(x) = p_1^k \sum_{i=0}^{n-k} \sum_{j=a}^{b} \binom{n-k-j-ki+i}{i+1} \frac{(-1)^{j-x} j!(i+1)(ix-i+2j+2)}{(j-x+2)!(j-i+1)!(x+i-j)!} p_1^{-i+j+ki+1} q_1^i, \tag{9}$$

where the limits of inner sum are $a = \max\{0, x-2, i-1\}$ and $b = \min\{n-k-ki-1, x+i\}$. For $x = 0$, an extra term of $1-(n-k+1)p_1^k$ is added to Eq. (9). For $x = 1$, an extra term of $(n-k+1)p_1^k$ is added to Eq. (9).

Proof For a given pair of integers n and k, the maximum value $M_{n,k}$ can take is $n-k+1$, so $E(M_{n,k}^{(r)}) = 0$ when $r > n-k+1$. This sets the upper summation limit of r in Eq. (3) as $n-k+1$. Applying Eq. (3)– (7) by assuming that it applies to all $r \geq 0$ (we will make the correction of 1 for $x = 0$ later), we have

$$P(x) = (-1)^x \sum_{r=x}^{n-k+1} \binom{r}{x}(-1)^r \times \sum_{i=0}^{r-1} \sum_{j=0}^{n-k-ki-1} \binom{n-k-ki-j-1}{r-2}\binom{r-1}{i}\binom{i+j+1}{j} p_1^{n-i-j} q_1^i.$$

Interchanging the order of summation, and taking care of the summation limits, we obtain

$$P(x) = (-1)^x \sum_{i=0}^{n-k} \sum_{j=0}^{n-k-i-ik} \binom{i+j+1}{j} p_1^{n-i-j} q_1^i \times \sum_{r=i-1}^{n-1-k-j-ki} \binom{n-1-k-j-ki}{r}\binom{r+1}{i}\binom{r+2}{x}(-1)^r.$$

The upper summation limit of r takes value of -1 only when $i = 0$ and $j = n-k$. In this case, the innermost sum can only take nonvanishing values when $x = 1$ or $x = 0$. The triple summation for $x = 1$ or $x = 0$ when $i = 0$ and $j = n-k$ is $(-1)^{x+1}(n-k+1)p_1^k$. For other combinations of i and j, $n-1-k-j-ki \geq 0$ and

hence the identity of Eq. (8) in Lemma 1 holds. Using this identity leads to Eq. (9) after some variable changes. For $x = 0$, $E(M_{n,k}^{(0)}) = 1$ should be added. □

To our best knowledge, the formula in Eq. (9) is more compact than the published results, which usually have three or more summations (for example, Balakrishnan and Koutras 2002, pp. 155–156). The original Markov chain imbedding approach uses a square matrix of dimension $(n - k + 1)(k + 1)$, which becomes incredibly big when n increases. The improved Markov chain imbedding approach uses a recursion on a vector of dimension $(k + 1)$ (Koutras and Alexandrou 1995). For a given x, this improved method needs $O(n^{x+1})$ operations of multiplication of the vector with one of the two square matrices of dimension $(k+1)$, so overall the computational complexity is $O(n^{x+1}k^2)$. Equation (9) significantly reduces the computational complexity. The upper limit of the outer sum shows that fewer computations are needed as k increases. For a given k, the computational complexity is $O(n^3)$. In practice, Eq. (9) is much faster than the improved Markov chain imbedding method, especially for large n and k. In a Maple implementation, Eq. (9) takes 0.005 seconds to calculate $P(x = 12)$ when $n = 1200, k = 310$, and $p = 0.3$, while the same calculation takes the improved Markov chain imbedding method 1551 seconds. The Maple codes for both methods can be found in the supplementary document and can also be downloaded from http://graphics.med.yale.edu/runs/type_III.

The mean and variance for *Type III* runs can be obtained from Eq. (7) (Balakrishnan and Koutras 2002, p.166). Again we see that the mean and variance are linear functions of n when n becomes large, as predicted by the singularity perturbation theorem (Flajolet and Sedgewick 2009). As for *Type II* runs, the joint distributions for *Type III* runs can be readily obtained by introducing more tracking variables to the individual GF g_i's in Eq. (5).

Acknowledgements This work was supported in part by the Clinical and Translational Science Award UL1 RR024139 from the National Center for Research Resources, National Institutes of Health.

Appendix: Proof of Lemma 1

In this Appendix, we give the proof of Lemma 1.

Proof Using the Wilf–Zeilberger method (Petkovšek et al.1996), we can obtain the following linear recurrence equation for S_m:

$$\begin{aligned}&(m - v + 3)(m - u + 2)(uv - u + 2m + 2)S_{m+1}\\ &= (m + 1)(uv - u + 2n + 4)(m - u - v)S_m.\end{aligned}$$

To get explicit form of S_m, first assume that $v - 2 \geq u - 1$. From the recurrence we have

$$S_{m+1} = \frac{(-1)^{m-v+3}\,[(m+1)\cdots]\,[(u+2)\cdots(u+v-m)]\,(uv-u+2n+4)}{[(m-u+2)\cdots]\,[(m-v+3)\cdots 1]\,(uv+2v-u-2)} S_{v-2}.$$

When $m = v - 2$, there is only one term in the sum, which leads to

$$S_{v-2} = \binom{v-1}{u}(-1)^{v-2}.$$

After substitution of S_{v-2} and rearrangements, the identity in Eq. (8) of Lemma 1 is obtained for the case when $v - 2 \geq u - 1$ and $m \geq 0$. If we assume $v - 2 < u - 1$, the same result is obtained for $m \geq 0$. In this case the recurrence ends with $m = u - 1$, and we use the identity

$$S_{u-1} = \binom{u+1}{v}(-1)^{u-1}$$

to get the explicit form of S_m.

The case of $m = -1$ is trivial since in this case the only value u can take is $u = 0$; hence, the sum involves only one term when $r = -1$. This further restricts the values of v, which can only take $v = 0$ or $v = 1$ for the sum to take nonvanishing value. □

References

Balakrishnan, N., Koutras, M. V. (2002). *Runs and scans with applications*. New York, NY: Wiley.

Flajolet, P., Sedgewick, R. (2009). *Analytic combinatorics*. New York, NY: Cambridge University Press.

Fu, J. C., Koutras, M. V. (1994). Distribution theory of runs: A Markov chain approach. *Journal of the American Statistical Association*, *89*, 1050–1058.

Hirano, K., Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov chain. *Statistica Sinica*, *3*, 313–320.

Kong, Y. (2006). Distribution of runs and longest runs: A new generating function approach. *Journal of the American Statistical Association*, *101*, 1253–1263.

Kong, Y. (2014). Number of appearances of events in random sequences: A new approach to non-overlapping runs. *Communications in Statistics-Theory and Methods* (To appear).

Kong, Y. (2015). Distributions of runs revisited. *Communications in Statistics-Theory and Methods*, *44*, 4663–4678.

Koutras, M. (1997). Waiting times and number of appearances of events in a sequence of discrete random variables. In N. Balakrishnan (Ed.), *Advances in combinatorial methods and applications to probability and statistics* (pp. 363–384). Boston, MA: Birkhäuser.

Koutras, M. V., Alexandrou, V. A. (1995). Runs, scans and urn model distributions: A unified Markov chain approach. *Annals of the Institute of Statistical Mathematics*, *47*(4), 743–766.

Mood, A. M. (1940). The distribution theory of runs. *Annals of Mathematical Statistics*, *11*, 367–392.

Petkovšek, M., Wilf, H. S., Zeilberger, D. (1996). *A = B*. Wellesley, MA: A K Peters Ltd.