Moments of Absorption Time for a Conditioned Random Walk*

By W. A. Beyer and M. S. Waterman

A random walk on the set of integers $\{0,1,2,\ldots,a\}$ with absorbing barriers at 0 and a is considered. The transition times from the integers z $(0<z<a)$ are random variables with finite moments. The nth moment of the time to absorption at a, $D_{z,n}$, conditioned on the walk starting at z and being absorbed at a, is discussed, and a difference equation with boundary values and initial values for $D_{z,n}$ is given. It is solved in several special cases. The problem is motivated by questions from biology about tumor growth and multigene evolution which are discussed.

I. Introduction

We derive a formula for the mean absorption time for a random walk on a finite set of integers with absorbing barriers at both ends. The mean absorption time is conditioned on absorption at the right end. The time epoch of the steps of the walk is a random variable with values in $(0,\infty)$ and finite mean. This problem arose from two mathematical models in biology, which are discussed. A method is presented of obtaining higher moments. Stern [10] has derived similar formulae for the first moment in the special case of steps in unit times (gambler's ruin problem). A general formula is given for the higher moments in the symmetric gambler's ruin problem conditioned on absorption at an end point.

Dr. William A. Beyer, T-7 Mail Stop 233, Los Alamos Scientific Laboratory, P. O. Box 1663, Los Alamos, New Mexico 87545.

0022-2526/79/010083+08$01.75

II. Applications

The first mathematical model to which the formula (1) below has been applied concerns the growth of a cancer tumor. A tumor is a noninflamed abnormal mass of tissue. A cancer tumor is thought of as arising from one wayward cell that has lost the ability to control itself. This is the model used for certain simple cases. The wayward cell starts reproducing without regard to the presence of other cells. The cell progeny may all die before catastrophe overtakes the host organism, or may produce a family tree of descendants (called a clone) large enough to be noticeable to the host organism, in which case the descendants become a tumor. It is the tumor that is noticeable, not the early cells that died without progeny. It is of interest to estimate the time it takes for one wayward cell to reach tumor size.

Bell [1] has posed the problem of modeling tumor growth and answering the above and related questions. A set of models can be developed as follows. Suppose each cell has the probability p ($0 < p < 1$) of dividing to produce two identical new cells, and probability $q = 1 - p$ of dying. Of interest is the event of a cell becoming, by this chance mechanism, a macroscopic clone of N cells. To complete this model, division and death times must be specified. The question of interest is the time it takes for one wayward cell to reach a clone of N cells. The natural model for this problem is a birth and death process with linear growth.

A second biological model that also leads to a birth and death process with linear growth was suggested by Perelson and Bell [8]. This model is based on the evolution process of repeated copies of genes by unequal crossover. For information refer to the article by Smith [9]. The existence of multiple copies of genes is an important biological fact. For our purposes, we may regard a gene as a subsequence of a nucleotide sequence (N-sequence), where an N-sequence is a finite sequence over an alphabet of four letters. The gene is a code for some biological function. Figure 1(a) shows how an N-sequence might be diagrammed with the genes numbered as shown. The N-sequence doubles preparatory to mitosis [Fig. 1(b)]. Normally, the N-sequence pair then divides; one pair goes to one cell of the daughter cells, and the other N-sequence goes to the other daughter cell. Occasionally, the phenomenon of crossing over takes place. Homologous points, such as the dividing point of genes 2 and 3, will become crossed. Less often the crossing over will be unequal. For example, the dividing point of genes 1 and 2 in N-sequence 1 will cross over the dividing point of genes 2 and 3 in chromosome 2 [Fig. 1(c)]. The N-sequences then separate, as shown in Fig. 1(d). In N-sequence 2', gene 2 appears twice, whereas in N-sequence 1', gene 2 does not appear. Thus, in one of the N-sequences a gene has been duplicated. The process can be repeated. There exist genes that have 10^7 copies in one organism. If we are given the probabilities and time epoch distribution of such crossovers and know that 10^7 copies have been made, we use the methods shown below to estimate the time of appearance of the first gene from which the rest are copies.
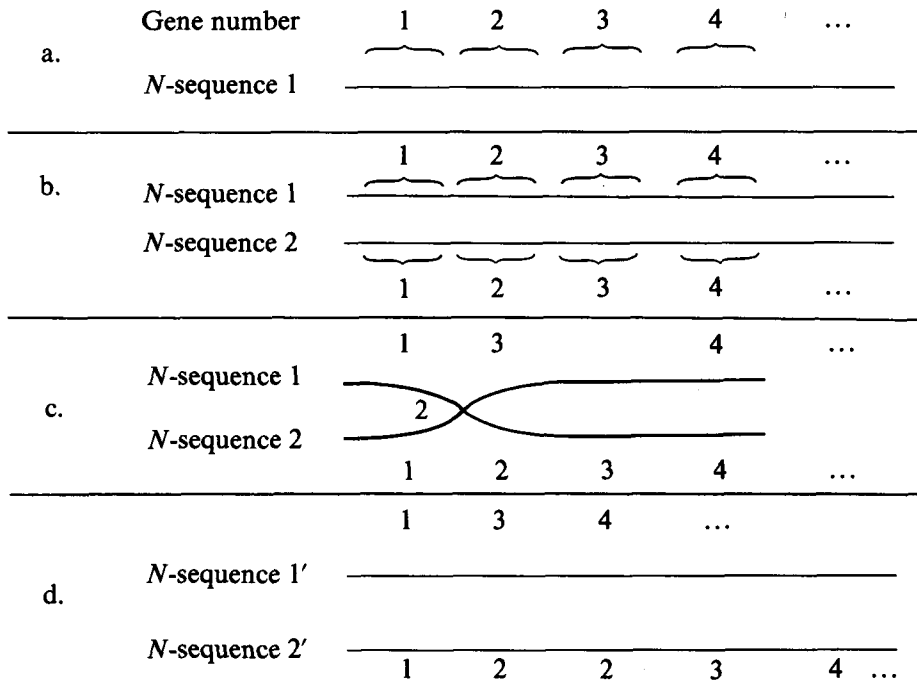
| Gene number | 1 | 2 | 3 | 4 | ... |

Figure 1. Gene duplication by unequal crossing over.

III. Random walk model and difference equation

Consider a random walk process on the set $\{0,1,2,\ldots,a\}$ with

$$\left.\begin{array}{l} P(z\to z+1)=p, \\ P(z\to z+1)=q=1-p, \end{array}\right\} \quad 0<z<a, \quad 0<p<1,$$

and 0 and a as absorbing barriers. If the arrival time at an integer z is t, then the transition time at z is $t+\tau_z$, where τ_z is a random variable depending on z with values in $(0,\infty)$ and with finite moments $E(\tau_z^n)$. If P_z is the probability of absorption at a of a random walk which starts at z, then from Feller [5, p. 345]:

$$P_z = \frac{(q/p)^z - 1}{(q/p)^a - 1}, \qquad p \neq q,$$

and

$$P_z = \frac{z}{a}, \qquad p = q.$$

Let $D_{z,n}$ be the nth moment of the duration of a random walk which starts at z, given that the walk terminates at a. By conditioning on the first step, it is seen

that $D_{z,n}$ satisfies the difference equation

$$p_z D_{z,n} = p p_{z+1} D_{z+1,n} + q p_{z-1} D_{z-1,n}$$

$$+ \sum_{i=0}^{n-1} \binom{n}{i} E(\tau_z^{n-i}) [p p_{z+1} D_{z+1,i} + q p_{z-1} D_{z-1,i}], \qquad (1)$$

with the boundary conditions

$$p_0 D_{0,n} = D_{a,n} = 0,$$

and the initial conditions

$$D_{z,0} = 1. \qquad (2)$$

(In every case in this note, $0 < z < a$, and n is a positive integer.) On putting

$$M_{z,n} = p_z D_{z,n},$$

one obtains from (1)

$$p M_{z+1,n} - M_{z,n} + q M_{z-1,n} = - \sum_{i=0}^{n-1} \binom{n}{i} E(\tau_z^{n-i})(p M_{z+1,i} + q M_{z-1,i}), \qquad (3)$$

with the boundary conditions

$$M_{0,n} = M_{a,n} = 0, \qquad (4)$$

and the initial conditions

$$M_{z,0} = p_z. \qquad (5)$$

IV. Case of birth and death processes

For linear birth and death processes,

$$E(\tau_z^n) = \frac{c^n n!}{z^n}, \qquad (7)$$

where c is a constant. For such processes, and for $p \neq q$,

$$D_{z,1} = \frac{c(p-q)^{-1}}{(p^a - q^a)} \left[- \sum_{i=1}^{a} \frac{(p^i - q^i)(p^{a-i} - q^{a-i})}{i} \right.$$

$$\left. + \frac{p^a - q^a}{p^z - q^z} \sum_{i=1}^{z} \frac{(p^i - q^i)(p^{z-i} - q^{z-i})}{i} \right]. \qquad (8)$$

Applications of (8) are given in [1] and [8]. For large a,

$$D_{1,1} = \frac{c}{|p-q|}(\log a + \gamma) + O\left(\frac{1}{a}\right),$$

where γ is Euler's constant.

For $p = q = \frac{1}{2}$,

$$D_{z,1} = c(a-z),$$

$$D_{z,2} = \frac{2c^2}{3}\left\{(2a-z)(a-z) + 3[F(z) - F(a)]\right\},$$

and

$$\sigma^2 = \frac{c^3}{3}\left(a^2 - z^2 - 6\sum_{i=z+1}^{a}\frac{1}{i}\right)$$

$$= \frac{c^2}{3}\left\{a^2 - z^2 - 6[F(a) - F(z)]\right\},$$

where $F(z)$ is the digamma function. Application of (3) to determine $M_{z,3}$ from the values of $M_{z,0}$, $M_{z,1}$, and $M_{z,2}$ requires the determination of

$$\Delta^{-2}\frac{1}{z}F(z),$$

which is not expressible in finite terms. For large n, the term involving the highest power of z in $D_{z,n}$ is

$$\frac{(-2cz)^n}{(n+1)!}.$$

V. Gambler's ruin problem

The gambler's ruin problem $[E(\tau_z) \equiv 1]$ is not relevant to the biological applications discussed above, but is discussed here because the formulae (1) or (3) subsume it. We consider first the case $p = q = \frac{1}{2}$. Then (3) reduces to

$$\Delta^2 M_{z,n} = -\sum_{i=0}^{n-1}\binom{n}{i}[M_{z+2,i} + M_{z,i}]$$

with

$$M_{0,n} = M_{a,n} = 0$$

and

$$M_{z,0} = \frac{z}{a}.$$

It is easy to show that

$$M_{z,n} = \sum_{j=1}^{2n+1} b_j^n(z)_j, \qquad n > 0,$$

where $(z)_j$ is the falling factorial $z(z-1)\cdots(z-j+1)$ and

$$b_1^0 = \frac{1}{a}.$$

Then

$$b_1^n = \frac{2}{a} \sum_{i=0}^{n-1} \binom{n}{i} \sum_{j=1}^{2i+1} b_j^i \sum_{k=0}^{\min(j,2)} \frac{(a)_{j-k+2}}{k!(j-k+2)_{2-k}}, \qquad n > 1,$$

$$b_2^n = -\sum_{i=0}^{n-1} \binom{n}{i}(b_1^i + b_2^i), \qquad n > 1,$$

$$b_l^n = -\frac{2}{l(l-1)} \sum_{i=0}^{n-1} \binom{n}{i} b_{l-2}^i - \frac{2}{l} \sum_{i=0}^{n-1} \binom{n}{i} b_{l-1}^i - \sum_{i=0}^{n-1} \binom{n}{i} b_l^i,$$

$$n > 2 \quad (3 \leqslant l \leqslant 2n-1),$$

$$b_{2n}^n = -\frac{b_{2n-2}^{n-1}}{2n-1} - b_{2n-1}^{n-1}, \qquad n \geqslant 2,$$

and

$$b_{2n+1}^n = -\frac{b_{2n-1}^{n-1}}{2n+1}, \qquad n > 1.$$

(Unassigned b's have the value zero.) For example,

$$b_1^1 = \frac{a}{3} - \frac{1}{3a},$$

$$b_2^1 = -\frac{1}{a},$$

and

$$b_3^1 = -\frac{1}{3a}.$$

(The values for b_1^1, b_2^1, and b_3^1 have been obtained by Stern [10].) In addition, one has

$$b_1^2 = \frac{7a^4 - 20a + 13}{45a},$$

$$b_2^2 = \frac{5}{3a} - \frac{2a}{3},$$

$$b_3^2 = \frac{17}{9a} - \frac{2a}{9},$$

$$b_4^2 = \frac{2}{3a},$$

and

$$b_5^2 = \frac{1}{15a}.$$

For large n, the term involving the highest power of z in $D_{z,n}$ is

$$\frac{(-z^2)^n}{(2n+1)!!}.$$

For $p \neq q$, (2)–(5) give, with $r = q/p$,

$$D_{z,1} = \frac{(p-q)^{-1}}{1-r^z} \left[(a-z)(r^z + 1) + \frac{2\alpha}{r^a - 1}(r^z - r^a) \right],$$

which Stern [10] also obtains.

VI. Remarks

1. The moments of the expected time to absorption at $z = 0$ conditioned on absorption at $z = 0$ are obtained by replacing z with $a - z$ and interchanging p and q in the above formulae. More details of the derivation of the above formulae are given in [2]. The solutions of these difference equations are best obtained by the methods in Jordan's book [6].

2. Everett and Stein [4] have investigated the combinatorial properties of the numbers $P_{a,l}$, which are the number of distinct walks that start at $z = 1$ and

arrive at a without $z = 0$ and have exactly l steps in the negative direction. In terms of $P_{a,l}$ one has for the classical ruin problem

$$D_{1,n} = \sum_{l=0}^{\infty} (a + 2l - 1)^n P_{a,l} p^{a+l-1} q^l.$$

Karni [7] has obtained a formula for the probability $\theta_{N,z}$ for a walk of N steps on a finite set of integers (with absorbing barriers at both ends) conditioned on absorption at a given end. The walk starts at the integer z. In terms of $\theta_{N,z}$, one has for the classical ruin problem

$$D_{z,n} = \sum_{N=0}^{\infty} N^n \theta_{N,z}.$$

References

1. G. I. BELL, Models of carcinogenesis as an escape from mitotic inhibitors, *Science* 192: 569–572 (1976).
2. W. A. BEYER and M. S. WATERMAN, A conditioned random walk with applications, Los Alamos Scientific Laboratory report LA-6990-MS, 1977.
3. W. A. BEYER and M. S. WATERMAN, Symmetries for conditioned ruin problems, *Math. Mag.* 50: 42–45 (1977).
4. C. J. EVERETT and P. R. STEIN, The combinatorics of random walk with absorbing barriers, *Discrete Math.* 17: 27–45 (1977).
5. W. FELLER, *An Introduction to Probability Theory and Its Applications*, Wiley, New York, 3rd ed., 1968.
6. C. JORDAN, *Calculus of Finite Differences*, Budapest, 1939; reprinted, Chelsea, New York, 1950.
7. E. KARNI, The probability distribution of the duration of the game in the classical ruin problem, *J. Appl. Probability* 14: 416–420 (1977).
8. A. S. PERELSON and G. I. BELL, Mathematical models for the evolution of multigene families by unequal crossing-over, *Nature* 265: 304–310 (1977).
9. G. P. SMITH, Evolution of repeated DNA sequences by unequal cross-over, *Science* 191: 528–535 (1976).
10. F. STERN, Conditional expectation of the duration in the classical ruin problem, *Math. Mag.* 48: 200–203 (1975).