Proc. Nat. Acad. Sci. USA
Vol. 71, No. 3, pp. 746–749, March 1974

The Moments of Stochastic Integrals and the Distribution of Sojourn Times*

(population genetics/random drift/finite populations/diffusion models)

THOMAS NAGYLAKI

Department of Medical Genetics, University of Wisconsin, Madison, Wisc. 53706

Communicated by James F. Crow, October 24, 1973

ABSTRACT **For a single diallelic locus in a finite population with any time-independent selection scheme, using the diffusion approximation, a formula is derived in terms of sojourn times for the moments of the integral of an arbitrary function of gene frequency along sample paths. Irreversible mutation and conditioned and unconditioned processes without mutation are treated. From this expression, the differential equation satisfied by the moments follows directly, and the exact probability distribution of sojourn times is deduced. An independent probabilistic proof of the last result based on the properties of time-homogeneous Markov processes is presented.**

In a recent paper (1), Maruyama and Kimura, relying on a theorem proved in Dynkin (2), developed a unified approach to the calculation of the moments of many distributions significant in the genetics of finite populations. These characteristics of stochastic changes in gene frequencies include sojourn times, times to fixation or loss, and total heterozygosities. The only probability distribution of this nature that has been derived analytically is that of the fixation time for a neutral gene (3).

Maruyama and Kimura consider a single diallelic locus, and denote the frequencies of the alleles A and a by x and $1 - x$. They discuss only the case without mutation, but irreversible mutation may be permitted without altering the basic formalism. Thus, $x = 0$ or $x = 1$ or both are absorbing barriers. Following ref. 1, we designate the frequency of A at time t for the particular sample path ω by $x(\omega, t)$. We assume all paths start at $t = 0$, and let $x(\omega, 0) \equiv p$. Since there is at least one absorbing barrier, for each sample path ω, the exit time from the interval (0,1), $\mathfrak{I}(\omega)$, is finite, and $x[\omega,\mathfrak{I}(\omega)] = 0$ or 1.

The nth moment of the integral of an arbitrary function of gene frequency, $f(x)$, along the sample path ω is

$$F^{(n)}(p) = E\left\{\left[\int_0^{\mathfrak{I}(\omega)} f[x(\omega,t)]dt\right]^n\right\}, \qquad [1]$$

where E indicates the expectation with respect to sample paths. If there is no mutation, frequently one desires to study paths conditioned on fixation or loss. Suppose $u_1(p)$ represents the probability that a path ultimately reaches $x = 1$. The nth moment comparable to $F^{(n)}(p)$ reads

$$K^{(n)}(p) = F_1^{(n)}(p)/u_1(p), \qquad [2]$$

with

$$F_1^{(n)}(p) = u_1(p)E\left\{\left[\int_0^{\mathfrak{I}(\omega)} f[x(\omega,t)]dt\right]^n \middle| x[\omega,\mathfrak{I}(\omega)] = 1\right\}. \qquad [3]$$

In terms of the time-independent drift and diffusion coefficients $M(p)$ and $V(p)$, we may write the backward diffusion operator as

$$L = M(p)\frac{\partial}{\partial p} + \frac{V(p)}{2}\frac{\partial^2}{\partial p^2}. \qquad [4]$$

In Eq. **4**, $M(p)$ is the expected change in gene frequency per unit time (generation), and $V(p) = p(1 - p)/(2N_e)$, where N_e denotes the variance effective population number.

Maruyama and Kimura (1) base their discussion on the theorem (2)

$$LF^{(n)}(p) + nf(p)F^{(n-1)}(p) = 0, n \geq 1. \qquad [5]$$

For two absorbing barriers, they give the boundary conditions

$$F^{(n)}(0) = F^{(n)}(1) = 0. \qquad [6]$$

If only $x = 1$ is an absorbing state, Eq. **6** must be replaced by the requirements that

$$\frac{d}{dp}F^{(n)}(0) \text{ be finite} \qquad [7a]$$

and

$$F^{(n)}(1) = 0. \qquad [7b]$$

The condition **7a** can be deduced from the work of Feller (**4**). For a single absorbing barrier at $x = 0$, 0 and 1 must be interchanged in Eq. **7**. The conditional moment $F_1^{(n)}(p)$ also satisfies Eqs. **5** and **6**(1). The moment-generating function

$$\Phi(\lambda;p) = \sum_{n=0}^{\infty}(\lambda^n/n!)F^{(n)}(p) \qquad [8]$$

satisfies

$$L\Phi(\lambda;p) + \lambda f(p)\Phi(\lambda;p) = 0, \qquad [9]$$

as does its conditioned analogue (1),

$$\Phi_1(\lambda;p) = \sum_{n=0}^{\infty}(\lambda^n/n!)F_1^{(n)}(p). \qquad [10]$$

In order to apply the theory expounded by Maruyama and Kimura, one must either solve Eq. **5** recursively or derive the moment-generating function from Eq. **9**. For higher

* Paper no. 1702 from the Genetics Laboratory, University of Wisconsin.

moments, the first method is clearly quite inconvenient, if not impracticable. The second technique, in general, is rather difficult; for the lower moments, manifestly unnecessarily so. We shall present an explicit integral formula for the moments $F^{(n)}(p)$ and $K^{(n)}(p)$ as functions of mean sojourn times. The latter have received considerable attention (5–8) and may be computed easily as follows. Set

$$G(x) = \exp\left\{-2\int^x [M(\xi)/V(\xi)]d\xi\right\} \quad [11]$$

and

$$g(a,b) = \int_a^b G(\xi)d\xi. \quad [12]$$

With no mutation, the respective probabilities of fixation and loss are

$$u_1(p) = g(0,p)/g(0,1) \quad [13a]$$

and

$$u_0(p) = g(p,1)/g(0,1). \quad [13b]$$

Let $\tau(p,y)\Delta y$ designate the total time the population spends in the interval I: $(y, y + \Delta y)$ for a particular sample path ω (starting at p, as stated above). We shall call $\tau(p,y)$ the sojourn time; the mean sojourn time is

$$\bar{\tau}(p,y) = E[\tau(p,y)]. \quad [14]$$

Finally, we introduce the Heaviside unit step function

$$\theta(\xi) = \begin{cases} 1, & \xi > 0 \\ 1/2, & \xi = 0. \\ 0, & \xi < 0 \end{cases} \quad [15]$$

If both fixation and loss are possible,

$$\bar{\tau}(p,y) = 2[V(y)G(y)]^{-1}[u_0(p)g(0,y)\theta(p - y) + u_1(p)g(y,1)\theta(y - p)]. \quad [16]$$

The mean sojourn times conditioned on fixation or loss read

$$\bar{\tau}_1(p,y) = \bar{\tau}(p,y)u_1(y)/u_1(p) \quad [17a]$$

and

$$\bar{\tau}_0(p,y) = \bar{\tau}(p,y)u_0(y)/u_0(p). \quad [17b]$$

With absorption possible only at $x = 1$,

$$\bar{\tau}(p,y) = 2[V(y)G(y)]^{-1} \times [g(p,1)\theta(p - y) + g(y,1)\theta(y - p)], \quad [18a]$$

while if it can occur only at $x = 0$,

$$\bar{\tau}(p,y) = 2[V(y)G(y)]^{-1}[g(0,y)\theta(p - y) + g(0,p)\theta(y - p)]. \quad [18b]$$

From the expressions in the *General Theory* section for $F^{(n)}(p)$ and $K^{(n)}(p)$, we shall deduce Eq. **5** directly. This seems desirable because Dynkin's proof (2) is neither brief nor elementary. In the section on *Sojourn Times*, we shall calculate the probability distribution of the sojourn time $\tau(p,y)$ in terms of its mean $\bar{\tau}(p,y)$ for all cases. This result will be verified by a probabilistic argument.

GENERAL THEORY

We rewrite Eq. **1** in the form

$$F^{(n)}(p) = E\left\{\prod_{i=1}^n \int_0^{3(\omega)} f[x(\omega,t_i)dt_i\right\}, \quad [19]$$

and note that the $n!$ permutations of the times t_i contribute equally to $F^{(n)}(p)$. Therefore, we may order them so that $t_i \geq t_{i-1}$. Introducing the probability $\phi(p, x; t)\Delta x$ that the gene frequency, with initial value p, is in the interval $(x, x + \Delta x)$ at time t, Eq. **19** becomes

$$F^{(n)}(p) = n! \prod_{i=1}^n \int_0^{t_{i+1}} dt_i \int_0^1 dx_i f(x_i)\phi(x_{i-1},x_i;t_i - t_{i-1}), \quad [20]$$

where we define $t_0 = 0$, $t_{n+1} = \infty$, $x_0 = p$, and average over sample paths by setting $x_i = x(\omega,t_i)$. Next, we change variables to the time intervals

$$t_i' = t_i - t_{i-1}, \quad [21a]$$

$$t_i = \sum_{j=1}^i t_j', \quad [21b]$$

and find

$$F^{(n)}(p) = n! \prod_{i=1}^n \int_0^1 dx_i \int_0^\infty dt_i' f(x_i)\phi(x_{i-1},x_i;t_i'). \quad [22]$$

Substituting

$$\bar{\tau}(\xi,\eta) = \int_0^\infty \phi(\xi,\eta;t)dt, \quad [23]$$

into Eq. **22**, we obtain our main result,

$$F^{(n)}(p) = n! \prod_{i=1}^n \int_0^1 dx_i\, f(x_i)\bar{\tau}(x_{i-1},x_i). \quad [24]$$

Equation **24** applies to all unconditional processes with at least one absorbing barrier. If we treat only simple paths that eventually reach $x = 1$, Eq. **20** must be modified to

$$F_1^{(n)}(p) = n! \prod_{i=1}^n \int_0^{t_{i+1}} dt_i \int_0^1 dx_i f(x_i)\phi(x_{i-1},x_i;t_i - t_{i-1})u_1(x_n), \quad [25]$$

whence the equation corresponding to Eq. **22** will read

$$F_1^{(n)}(p) = n! \prod_{i=1}^n \int_0^1 dx_i \int_0^\infty dt_i' f(x_i)\phi(x_{i-1},x_i;t_i')u_1(x_n). \quad [26]$$

The conditional mean sojourn time $\bar{\tau}_1(\xi,\eta)$ is given by Eqs. **17a** and **23**. Hence, the integrations over time may be performed successively to derive, recalling Eq. **2**,

$$K^{(n)}(p) = n! \prod_{i=1}^n \int_0^1 dx_i f(x_i)\bar{\tau}_1(x_{i-1},x_i). \quad [27]$$

As might have been expected, Eq. **27** has the same form as Eq. **24**. (For an alternative expression, see Eq. **32**.) Therefore, all results derived from the former will apply to conditioned processes if unconditioned mean sojourn times are replaced by conditioned ones. Observe that, since the mean sojourn times **16**, **17**, and **18** satisfy the appropriate boundary conditions **6** or **7**, so do $F^{(n)}(p)$ and $F_1^{(n)}(p)$ evaluated from our

expressions 24 and 27. Therefore, we have found the unique solution, in integral form, of the system of differential equations 5, 6, and 7.

Should we desire to calculate the total heterozygosity before absorption, we have merely to set $f(x) = 2x(1 - x)$. For absorption times, $f(x) = 1$. In the latter case, the integral over x_n may be performed at once:

$$\bar{T}(x_{n-1}) = \int_0^1 dx_n \bar{\tau}(x_{n-1},x_n), \quad [28]$$

where $\bar{T}(\xi)$ is the mean absorption time for paths starting at $x = \xi$ at $t = 0$. For example, the second moment of the distribution of absorption times is

$$\bar{T}^{(2)}(p) = 2\int_0^1 dx \bar{\tau}(p,x)\bar{T}(x). \quad [29]$$

To demonstrate Eq. 5, we apply the operator L to Eq. 24. Now, from Eq. 23 and the backward Kolmogorov equation,

$$\begin{aligned} L\bar{\tau}(p,x_1) &= \int_0^\infty \frac{\partial \phi}{\partial t}(p,x_1;t)dt \\ &= \phi(p,x_1;\infty) - \phi(p,x_1;0) \\ &= -\delta(x_1 - p), \ 0 < p, x_1 < 1, \end{aligned} \quad [30]$$

whence, integrating over x_1 in Eq. 24,

$$LF^{(n)}(p) = -nf(p)\left[(n-1)!\prod_{i=2}^{n}\int_0^1 dx_i f(x_i)\bar{\tau}(x_{i-1},x_i)\right], \quad [31]$$

with the understanding that, due to Eq. 30, in Eq. 31 $x_1 = p$. The bracket is just $F^{(n-1)}(p)$, so Eq. 5 follows. Using Eq. 17a in Eq. 27, we find,

$$F_1^{(n)}(p) = n!\prod_{i=1}^{n} dx_i f(x_i)\bar{\tau}(x_{i-1},x_i)u_1(x_n). \quad [32]$$

Of course, the product does not include $u_1(x_n)$ here. Hence, the proof for $F^{(n)}(p)$ holds equally for $F_1^{(n)}(p)$.

SOJOURN TIMES

With the identification (1)

$$f(x) = \delta(x - y), \quad [33]$$

where δ is the Dirac delta function, the theory developed above yields the moments of the distribution of $\tau(p,y)$. Substituting Eq. 33 into 24 yields

$$F^{(n)}(p) = n!\bar{\tau}(p,y)[\bar{\tau}(y,y)]^{n-1}, \ n \geq 1. \quad [34]$$

The sum in Eq. 8 is trivial;

$$\Phi(\lambda;p,y) = 1 + \frac{\lambda\bar{\tau}(p,y)}{1 - \lambda\bar{\tau}(y,y)}. \quad [35]$$

Now,

$$\Phi(\lambda;p,y) = \int_0^\infty e^{\lambda\tau}P(\tau;p,y)d\tau, \quad [36]$$

where $P(\tau;p,y)$ is the probability density of $\tau = \tau(p,y)$. Hence, inverting the simple Laplace transform 35, we find

$$P(\tau;p,y) = \left[1 - \frac{\bar{\tau}(p,y)}{\bar{\tau}(y,y)}\right]\delta(\tau) + \frac{\bar{\tau}(p,y)}{[\bar{\tau}(y,y)]^2}e^{-\tau/\bar{\tau}(y,y)}. \quad [37]$$

This distribution is valid for all cases discussed in this paper; for conditional processes, one simply inserts conditional mean sojourn times. Note also that, from Eq. 17,

$$\bar{\tau}_1(y,y) = \bar{\tau}_0(y,y) = \bar{\tau}(y,y). \quad [38]$$

Equation 35 was confirmed purely analytically by solving Eq. 9 for every possible process. To do this, one must use Eqs. 4, 6, 7, 11 to 18, and, crucially, 33. For conditional fixation at $x = 1$, the second moment calculated by Maruyama (9) may be simplified to our formula

$$K^{(n)}(p) = n!\bar{\tau}_1(p,y)[\bar{\tau}(y,y)]^{n-1}, \ n \geq 1, \quad [39]$$

for $n = 2$.

A few observations concerning Eq. 37 are instructive. The delta-function term allows for the possibility of not visiting the interval I: $(y, y + \Delta y)$ at all. If the process commences at y, that is, $p = y$, this term disappears, as it must:

$$P(\tau;y,y) = \frac{e^{-\tau/\bar{\tau}(y,y)}}{\bar{\tau}(y,y)}. \quad [40]$$

If the population must visit I, the term also vanishes. For $y > p$, this occurs if there is only one absorbing barrier at $x = 1$, or if fixation is conditioned there. For $y < p$, the single absorbing state or conditioning must be at $x = 0$. From Eqs. 16, 17, and 18 it is easy to check that all four instances do, indeed, satisfy $\bar{\tau}(p,y) = \bar{\tau}(y,y)$. If $p = 0$ and there are two unconditioned absorbing barriers, or conditioning at $x = 0$, or a single absorbing state at $x = 0$, the process is finished at $t = 0$ and Eqs. 16, 17, and 18 show $\bar{\tau}(p,y) = 0$. This statement is equally valid if $p = 0$ and $x = 0$ are replaced by $p = 1$ and $x = 1$, respectively. From Eq. 37, we obtain, then, as we must, $P(\tau;p,y) = \delta(\tau)$, indicating that the sojourn time necessarily vanishes. Of course, the delta-function term in Eq. 37 serves only to normalize the probability density; it does not contribute to any of its moments.

Finally, we shall derive Eq. 37 by a general probabilistic argument using the properties of time-homogeneous Markov processes. Clearly,

$$P(\tau;p,y) = Q(p,y)\delta(\tau) + [1 - Q(p,y)]P(\tau;y,y), \quad [41]$$

where $Q(p,y)$ is the probability of not visiting I before fixation or loss. Therefore,

$$\bar{\tau}(p,y) = \int_0^\infty \tau P(\tau;p,y)d\tau \quad [42]$$

$$= [1 - Q(p,y)]\bar{\tau}(y,y), \quad [43]$$

whence we see that the probability of a visit to I is

$$1 - Q(p,y) = \bar{\tau}(p,y)/\bar{\tau}(y,y). \quad [44]$$

It remains to determine $P(\tau;y,y)$. To this end, we modify the treatment of continuous sojourn times (those between an entrance to an interval and the next exit) for a discrete state space by Dynkin and Yushkevich (10). Let $R(\tau;y,y)$ be the probability that $\tau(y,y)$ is greater than some value τ; thus

$$R(\tau;y,y) = \int_\tau^\infty P(\tau';y,y)d\tau'. \quad [45]$$

For a time-homogeneous Markov process, evidently

$$R(\tau_1 + \tau_2;y,y) = R(\tau_1;y,y)R(\tau_2;y,y). \quad [46]$$

Consequently, for some constant k,

$$R(\tau;y,y) = e^{-k\tau}, \quad [47]$$

and

$$P(\tau;y,y) = -\frac{d}{d\tau} R(\tau;y,y)$$

$$= ke^{-k\tau}. \quad [48]$$

Substituting Eq. **48** into Eq. **42**, we find

$$\bar{\tau}(y,y) = 1/k, \quad [49]$$

and combining Eqs. **44**, **48**, and **49**, we obtain Eq. **37**.

I am grateful to Prof. James F. Crow for his constant interest and encouragement. I thank the pioneer and preëminent exponent of the role of random drift in population genetics and evolution, Prof. Sewall Wright, for generously spending countless hours during the past year communicating to me some of his uniquely broad and deep understanding. This research was supported by the National Institutes of Health (Grant GM-15422).

1. Maruyama, T. & Kimura, M. (1971) "Some methods for treating continuous stochastic processes in population genetics," *Jap. J. Genet.* **46**, 407–410.
2. Dynkin, E. B. (1965) *Markov Processes* (Springer-Verlag, Berlin), Vol. II, pp. 52–53.
3. Kimura, M. (1970) "The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population," *Genet. Res.* **15**, 131–133.
4. Feller, W. (1954) "Diffusion processes in one dimension," *Trans. Amer. Math. Soc.* **77**, 1–31.
5. Ewens, W. J. (1963) "The diffusion equation and a pseudo-distribution in genetics," *J. Roy. Stat. Soc.* **B25**, 405–412.
6. Ewens, W. J. (1964) "The pseudo-transient distribution and its uses in genetics," *J. Appl. Prob.* **1**, 141–156.
7. Ewens, W. J. (1969) *Population Genetics* (Methuen, London), pp. 52–55.
8. Kimura, M. (1969) "The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations," *Genetics* **61**, 893–903.
9. Maruyama, T. (1972) "The average number and the variance of generations at a particular gene frequency in the course of fixation of a mutant gene in a finite population," *Genet. Res.* **19**, 109–113.
10. Dynkin, E. B. & Yushkevich, A. A. (1969) *Markov Processes* (Plenum Press, New York), pp. 152–153.