Genet. Res., Camb. (1975), **25**, *pp.* 89–94
Printed in Great Britain

On sojourn times at particular gene frequencies

By EDWARD POLLAK and BARRY C. ARNOLD
Department of Statistics, Iowa State University, Ames, Iowa 50010, U.S.A.

(*Received* 25 *April* 1974)

SUMMARY

The distribution of visits to a particular gene frequency in a finite population of size N with non-overlapping generations is derived. It is shown, by using well-known results from the theory of finite Markov chains, that all such distributions are geometric, with parameters dependent only on the set of b_{ij}'s, where b_{ij} is the mean number of visits to frequency $j/2N$, given initial frequency $i/2N$. The variance of such a distribution does not agree with the value suggested by the diffusion method. An improved approximation is derived.

Consider a gene, say of type A, that is introduced into a diploid population that is of size N in every generation. We assume that there is no mutation. Then, if the initial frequency of A is $i/2N$, there is a proper random variable T_{ij}, which is the number of generations spent at frequency $j/2N$ in the progress of the population toward fixation or loss of the gene. The purpose of this note is to show how the entire distribution of T_{ij} may be derived, once the means of T_{ij} and T_{jj} are known. Thus, it will only be necessary to obtain approximations to $E(T_{ij})$ and $E(T_{jj})$ to be able to immediately write down an approximation to the probability that T_{ij} assumes a particular value. This contrasts to the approach of Maruyama & Kimura (1971) and Maruyama (1973), who set up a system of differential equations to compute the moments of the distribution of T_{ij}. Our approach also will be shown to be applicable to the problem of deriving the distribution of T_{ij}, given ultimate fixation, which was considered by Maruyama (1972).

We begin by noting that the underlying stochastic process is a finite Markov chain with states $E_i, i = 0, 1, \ldots, 2N$, associated with gene frequencies $i/2N$. We define f_{ij} to be the probability that there is, at some time, a visit to state E_j, given that the initial state is E_i. We suppose first that $i \neq j$. Then, to ensure that $T_{ij} = m \geqslant 1$, there must be an initial transition from E_i to E_j, followed by $m-1$ returns to E_j, and then a failure to return to E_j. Utilizing the Markov property of the process, we conclude that, if $i \neq j$, then

$$P(T_{ij} = m) = f_{ij} f_{jj}^{m-1} (1-f_{jj}). \tag{1}$$

It is also clear that

$$P(T_{ij} = 0) = 1-f_{ij}. \tag{2}$$

If $i = j$, then $T_{ii} = m \geqslant 1$ if and only if there are $m-1$ returns to the initial state, followed by a failure to return. Because E_i is now the initial state, it is, of course, impossible for the population to spend no generations at E_i. Hence, expressions (1) and (2) still apply, provided we replace f_{ij} by 1. Expressions (1) and (2) are well known in the theory of finite Markov chains.

https://doi.org/10.1017/S0016672300015494 Published online by Cambridge University Press

Either by consideration of the generating function of T_{ij}, or directly, one may verify that (cf. Kemeny & Snell, 1960)

$$E(T_{ij}) = f_{ij}/(1-f_{jj}), \tag{3}$$

$$\sigma^2(T_{ij}) = b_{ij}(2b_{jj}-b_{ij}-1), \tag{4}$$

if we introduce the notation b_{ij} for $E(T_{ij})$. It is convenient to rewrite the distribution of T_{ij} in terms of the b_{ij}'s. Thus, we have from (3) that $f_{ij} = b_{ij}/b_{jj}$, so that (1) and (2) take the form

$$P(T_{ij}=0) = 1-(b_{ij}/b_{jj}), \tag{5}$$

$$P(T_{ij}=m) = (b_{ij}/b_{jj}^2)\,[1-1/b_{jj}]^{m-1} \quad (m \geqslant 1), \tag{6}$$

which is how they appear in Kemeny & Snell (1960, p. 62). Note that in this formulation the count of b_{ii} includes the initial generation in E_i.

All the foregoing reasoning is still applicable if we consider the conditional distribution of T_{ij}, given that there is ultimate fixation. This is because, even with the conditioning, the process is still a finite Markov chain. The only change that needs to be made in (4), (5) and (6) is to replace b_{ij} by

$$b_{ij|F} = E[T_{ij}|\text{ultimate fixation}].$$

We note that if we replace b_{ij} by $b_{ij|F}$ in (4) it is not consistent with the formula given by Maruyama (1972) for the expected value of T_{ij}^2, given ultimate fixation. In his notation this expectation is written as $\Phi_1^{(2)}(x,y)/(2N)^2$, while what we have written as $b_{ij|F}$ is denoted by $\Phi_1(x,y)/2N$. It then follows from expression (7) in Maruyama's paper that, in our notation,

$$E[T_{ij}^2|\text{fixation}] = 2b_{ij|F}b_{jj|F}.$$

Also, Maruyama (1973) has obtained an expression which is

$$E(T_{1j}^2) = 2b_{1j}b_{jj}$$

in our notation, rather than $2b_{1j}b_{jj}-b_{1j}$, as implied by (4).

Nagylaki (1974) has derived the distribution of sojourn times by another method than the one used here and has obtained results consistent with those of Maruyama. However, his derivation is for a process that is continuous in its state space and time parameter. He has informed us that his methods also may be used to obtain the sojourn time distributions for a discrete time Markov chain and the results then agree with (5) and (6).

In the remainder of this paper we consider the Fisher–Wright model with no selection. In this case the probability of a transition from E_i to E_j in one generation is

$$p_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^j\left(1-\frac{i}{2N}\right)^{2N-j}. \tag{7}$$

It is well known (cf. Ewens, 1973) that in this case the diffusion approximation gives

$$\left.\begin{aligned} b_{ij} &= \frac{2(1-i/2N)}{1-j/2N} \quad (j < i)\\ &= \frac{2i}{j} \qquad\qquad (j > i).\end{aligned}\right\} \tag{8}$$

As Ewens (1973) has shown, $b_{ij|F}$ is equal to $(b_{ij}P_j)/P_i$, where P_i is the probability of ultimate fixation, given initial state E_i. With neutral genes $P_j/P_i = j/i$, so that

$$b_{ij|F} = 2, \quad j = i+1, \ldots, 2N \tag{9}$$

as found by Maruyama (1972) and Ewens (1973).

It seems natural to extend (8) to $i = j$, so that b_{ii} is set equal to 2. Professor Alan Robertson has suggested to us that this is wrong, since this implies that, when i and N are large, the mean number of visits to E_i after time 0 would be 1 rather than 2, as at E_{i-1} and E_{i+1}, which seems implausible. He suggests that the results of Maruyama and of Ewens refer to subsequent visits, not counting the initial state, and therefore in the present formulation $b_{ii} = b_{ii|F} \doteqdot 3$, so that if T^*_{ij} represents the number of visits to state E_j after time 0 and N is large,

$$P(T^*_{ij} = m|\text{ultimate fixation}) \doteqdot \tfrac{1}{3}(\tfrac{2}{3})^m \tag{10}$$

for all i, j and $m = 0, 1, 2, \ldots$. The variance of this conditional distribution is

$$\sigma^2(T^*_{ij})_{|F} \doteqdot 6. \tag{11}$$

The approximation to b_{ii} can be improved by considering how the b_{ij}'s are computed for the underlying Markov chain. Thus, there is a transition from E_i to E_j in n steps with probability $p^{(n)}_{ij}$. This is also the mean number of visits to state E_j at time n, so that

$$b_{ij} = \sum_{n=0}^{\infty} p^{(n)}_{ij}. \tag{12}$$

Expression (12) and the Chapman–Kolmogorov equations imply that

$$b_{ij} = p^{(0)}_{ij} + \sum_{n=1}^{\infty} \sum_{r=1}^{2N-1} p^{(n-1)}_{ir} p_{rj} = p^{(0)}_{ij} + \sum_{r=1}^{2N-1} b_{ir} p_{rj}, \tag{13}$$

$i = 1, \ldots, 2N-1$.

If $i = j$, it follows from (7), (8) and (13) that

$$\begin{aligned} b_{ii} = {} & \left[1 - \binom{2N}{i}\left(\frac{i}{2N}\right)^i \left(1 - \frac{i}{2N}\right)^{2N-i}\right]^{-1} \left\{1 - 2\binom{2N}{i}\left(\frac{i}{2N}\right)^i \left(1 - \frac{i}{2N}\right)^{2N-i}\right. \\ & + 2\sum_{r=1}^{i} \frac{(2N)!}{i!\,(2N-i-1)!}\left(\frac{r}{2N}\right)^i \left(1 - \frac{r}{2N}\right)^{2N-i-1} \frac{1}{2N} \\ & \left. + 2\sum_{r=i+1}^{2N-1} \frac{(2N)!}{(i-1)!\,(2N-i)!}\left(\frac{r}{2N}\right)^{i-1} \left(1 - \frac{r}{2N}\right)^{2N-i} \frac{1}{2N}\right\}. \end{aligned}$$

Approximating the sums by integrals, we have, if $p = i/2N$,

$$\begin{aligned} b_{ii} = {} & \left[1 - \binom{2N}{i}\left(\frac{i}{2N}\right)^i \left(1 - \frac{i}{2N}\right)^{2N-i}\right]^{-1} \left\{3 - 2\binom{2N}{i}\left(\frac{i}{2N}\right)^i \left(1 - \frac{i}{2N}\right)^{2N-i}\right. \\ & + \frac{2}{B(i+1, 2N-i)} \int_0^p x^i (1-x)^{2N-i-1}\,\mathrm{d}x - \frac{2}{B(i, 2N-i+1)} \int_0^p x^{i-1}(1-x)^{2N-i}\,\mathrm{d}x, \end{aligned} \tag{14}$$

https://doi.org/10.1017/S0016672300015494 Published online by Cambridge University Press

where $B(a, b)$ is the Beta-function with parameters a and b. It is known (see, for example, Abramowitz & Stegun (1968, p. 944)) that the difference between the two integrals in (14) is equal to $-2\binom{2N}{i}p^i(1-p)^{2N-i}$. Hence

$$b_{ii} \doteq \left(3-4\binom{2N}{i}p^i(1-p)^{2N-i}\right)\Big/\left(1-\binom{2N}{i}p^i(1-p)^{2N-i}\right). \tag{15}$$

If N is large and p is not near 0 or 1, we have from the central limit theorem that

$$b_{ii} \doteq (3-4(4N\pi p(1-p))^{-\frac{1}{2}})/(1-(4N\pi p(1-p))^{-\frac{1}{2}}). \tag{16}$$

As $N \to \infty$, b_{ii} approaches 3, as suggested by Robertson. Considered as a function of p, with given N, b_{ii} assumes its maximum value at $p = \frac{1}{2}$ and diminishes as $|p-\frac{1}{2}|$ increases. If, on the other hand, p is near 0, the Poisson approximation leads us to

$$b_{ii} \doteq (3-4\,\mathrm{e}^{-i}i^i/i!)/(1-\mathrm{e}^{-i}i^i/i!). \tag{17}$$

Expressions (8) and (16) are approximations, applicable when N is large. The b_{ij}'s can be computed directly by noting that (13) may also be put in the form

$$\mathbf{b}'_i = [b_{i1}, \ldots, b_{i,\,2N-1}] = \mathbf{e}'_i(\mathbf{I}-\mathbf{Q})^{-1}, \tag{18}$$

where $\mathbf{e}'_i$ is the $1 \times 2N-1$ vector with 1 in the ith position and zeros elsewhere and $\mathbf{Q}$ is the matrix of probabilities of transitions from transient states to transient states. Since the right side of (18) is the ith row of $(\mathbf{I}-\mathbf{Q})^{-1}$, it follows that the quantities b_{ij} are the elements of $(\mathbf{I}-\mathbf{Q})^{-1}$, as shown by Kemeny & Snell (1960, ch. III). If there are neutral genes, $b_{ij|F} = (jb_{ij})/i$. Some values of $b_{ij|F}$ have been computed for the case in which $2N = 50$ and there is no selection, and are displayed in Table 1.

Table 1. *Some values of $b_{ij|F}$ for $2N = 50$*

	j						
i	1	9	17	25	33	41	49
1	2·238	1·998	1·996	1·993	1·987	1·969	1·643
9	—	2·770	1·996	1·993	1·987	1·969	1·643
17	—	—	2·818	1·994	1·987	1·969	1·643
25	—	—	—	2·828	1·987	1·969	1·643
33	—	—	—	—	2·818	1·970	1·643
41	—	—	—	—	—	2·770	1·643
49	—	—	—	—	—	—	2·238

These results indicate that (9) gives a good approximation to $b_{ij|F}$ if j is not near 50. Approximate values of $b_{ii|F} = b_{ii}$, calculated from (16), are 2·828, 2·865, 2·873, 2·865 and 2·828 if i = 9, 17, 25, 33 and 41 respectively. They are too large, as might have been expected, because the diffusion approximation that was used in place of b_{ir} in (13) to obtain (16) is an overestimate.

If we substitute the expressions given by (9) and (16) for $b_{ij|F}$ and b_{ii} in (4) we have

$$\begin{aligned}\sigma^2(T_{ij})_{|F} &= b_{ij|F}(2b_{jj|F}-b_{ij|F}-1)\\ &\doteq 2\left\{\frac{6-8(4N\pi p(1-p))^{-\frac{1}{2}}}{1-(4N\pi p(1-p))^{-\frac{1}{2}}}-3\right\},\end{aligned} \tag{19}$$

for $j = i+1, \ldots, 2N$. As $N \to \infty$, the right side approaches 6. Considered as a function of p, with given N, it assumes its maximum value at $p = \frac{1}{2}$ and diminishes as $|p-\frac{1}{2}|$ increases. Approximate values of $\sigma^2(T_{ij})_{|F}$ computed from (19) when $i = 1$, $j = 9$, 17, 25, 33 and 41 are 5·312, 5·460, 5·492, 5·460 and 5·312 respectively.

Expression (19) is applicable if j is not too near 0 or $2N$ and indicates that, for such intermediate values of j, $\sigma^2(T_{ij})_{|F}$ will approach 6 as $N \to \infty$. If, on the other hand, p is near 0, (17) implies that $\sigma^2(T_{ij})_{|F}$ approaches a smaller value.

By using $b_{ij|F}$ in place of b_{ij} in (4), and the figures in Table 1, it is possible to compute $\sigma^2(T_{ij})_{|F}$. Some numerical values are given in Table 2.

Table 2. *Some values of* $\sigma^2(T_{ij})_{|F}$ *for* $2N = 50$

	j						
i	1	9	17	25	33	41	49
1	2·77	5·08	5·27	5·31	5·26	5·06	3·01
9	—	4·90	5·27	5·31	5·26	5·06	3·01
17	—	—	5·12	5·31	5·26	5·06	3·01
25	—	—	—	5·17	5·26	5·06	3·01
33	—	—	—	—	5·12	5·06	3·01
41	—	—	—	—	—	4·90	3·01
49	—	—	—	—	—	—	2·77

The approximate values given by (19) are thus overestimates, although, qualitatively, the prediction of the behaviour of $\sigma^2(T_{ij})_{|F}$ is not misleading. The variances $\sigma^2(T_{ij})_{|F}$ are not all approximately equal to 4 for $j \geqslant i$, as asserted by Maruyama (1972). Instead, if i is small, they increase toward a peak that is well above 5 as j increases toward 25 and then decline as j becomes still larger. This is what we would expect from (19).

We are grateful to Professor Alan Robertson for helpful suggestions and comments and to Dr Bud Meador, Charlotte Bentley and Pamela Doctor for help in computing Table 1.

This article is Journal Paper No. J-7887 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project 1669. Research was supported in part by National Institute of Health, Grant GM13827.

https://doi.org/10.1017/S0016672300015494 Published online by Cambridge University Press

REFERENCES

Abramowitz, M. & Stegun, I. A. (eds.) (1968). *Handbook of Mathematical Functions.* New York: Dover Publications, Inc.

Ewens, W. J. (1973). Conditional diffusion processes in population genetics. *Theoretical Population Biology* **4**, 21–30.

Kemeny, J. G. & Snell, J. L. (1960). *Finite Markov Chains.* New York: D. van Nostrand Company.

Maruyama, T. (1972). The average number and the variance of generations at particular gene frequency in the course of fixation of a mutant gene in a finite population. *Genetical Research* **19**, 109–113.

Maruyama, T. (1973). The variance of the number of loci having a given gene frequency. *Genetics* **73**, 361–366.

Maruyama, T. & M. Kimura (1971). Some methods for treating continuous stochastic processes in population genetics. *Japanese Journal of Genetics* **46**, 407–410.

Nagylaki, T. (1974). The moments of stochastic integrals and the distribution of sojourn times. *Proceedings of the National Academy of Sciences U.S.A.* **71**, 746–749.

https://doi.org/10.1017/S0016672300015494 Published online by Cambridge University Press