[1] University of North Carolina at Chapel Hill. Research partially supported by the Air Force Office of Scientific Research under Contract No. AFOSR-68-1415.

[2] University of Toronto. Research partially supported by a Public Health Service Genetics Training Program Grant, GM 00685, at the University of North Carolina.

A Diffusion Model for Random Drift with Variable Population Size

Woollcott Smith[1] *and* J. J. Hsieh[2]

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 816

April, 1972

A Diffusion Model for Random Drift with Variable Population Size

Woollcott Smith[1] *and* J. J. Hsieh[2]

SUMMARY

A diffusion model for random genetic drift with a variable population size is developed. The model is based on Karlin's Markov chain model for genetic drift in a haploid population. The variable size diffusion model is shown to be equivalent to Wright's constant size diffusion model with a random time change. Conditions are given for the population growth that will assure a positive probability that the two types of individuals remain in the population indefinitely. The mean time to fixation or loss in the variable size model is found under certain conditions to be the same as the mean time to fixation or loss in Wright's constant size diffusion model. A diffusion approximation to a branching process is used to construct an example of a variable size population model with the same mean time to fixation or loss as the constant size diffusion model.

---

[1] University of North Carolina at Chapel Hill. Research partially supported by the Air Force Office of Scientific Research under Contract No. AFOSR-68-1415.

[2] University of Toronto. Research partially supported by a Public Health Service Genetics Training Program Grant, GM 00685, at the University of North Carolina.

Section 1. INTRODUCTION

In this paper we develop a diffusion model to describe random fluctuations in gene frequencies over time in a finite random haploid population. The model is similar to the diffusion approximation of Wright's (1931) random drift model studied later by Wright (1945) and also by Kimura (1957), except that we allow for random changes in population sizes with time. In Section 2 of this paper we show that our random population size model is equivalent to the fixed population size model with a random time change. In Sections 3 and 4 we use this result to study the properties of the model and to investigate some special cases.

Wright (1931) and Fisher (1922, 1930) proposed a stochastic model to describe the random fluctuations of gene frequencies in a population of N haploid individuals. Suppose that there are two types (alleles) of a given gene in the haploid population, type A_1 and type A_2. Let X_n denote the number of individuals of type A_1 in the nth generation; there are then $N - X_n$ members of type A_2. We assume that selection, mutation and migration are all absent and that generation $n + 1$ is formed by N independent random samples of the nth generation with replacement. The process $\{X_n\}$ is a Markov chain with one-step transition probabilities

$$\Pr\{X_{n+1} = j | X_n = i\} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}. \tag{1}$$

States 0 and N are absorbing states corresponding respectively to loss and fixation of type A_1. One would like to find the properties of the distribution of the time to loss or fixation (i.e., the first passage time to states 0 or N); unfortunately, the most interesting properties such as the mean time to loss or fixation have not been obtained in closed form for this model.

Wright (1945) first investigated the diffusion approximation to the random drift model. Let $X(t)$ denote the proportion of members of type A_1 in the population. Assume that $X(t)$ is a diffusion with infinitesimal mean and variance equal to 0 and $x(1-x)/N$ respectively; the forward Kolmogorov diffusion equation for the probability density of the random variable $X(t)$, given that $X(0) = p$, is

$$\frac{\partial f(p,x;t)}{\partial t} = \frac{1}{2N} \frac{\partial^2 x(1-x) f(p,x;t)}{\partial x^2} . \tag{2}$$

The distribution of the first passage time to fixation or loss in the $X(t)$ diffusion was first obtained by Kimura (1957) by solving the backward diffusion equation. The mean first passage time can be obtained from Kimura's results or directly from the backward diffusion equation (Ewens, 1969). Let T_X denote the first passage time to exit boundaries 0 or 1 in the $X(t)$ diffusion, then

$$E[T_X | X(0) = p] = -2N[p\ell n p + (1-p)\ell n(1-p)]. \tag{3}$$

It has been noted by Feller (1951) that random drift models with a fixed population size are not realistic and may introduce bias. However, there are mathematical difficulties in analysing models with a population size that varies stochastically with time. Karlin (1968) has proposed an extension of Wright's random drift model to the case where the population size at time n, N_n, is a finite state Markov chain. The process $\{(N_n, X_n)\}$ is assumed to be a bivariate Markov chain. Under Karlin's model, if the unconditional progeny distribution is Poisson, then the one-step transition probabilities for the bivariate Markov chain are

$$\Pr\{N_{n+1} = j, X_{n+1} = \ell | N_n = i, X_n = m\} = p_{ij} \binom{j}{\ell} \left(\frac{m}{i}\right)^{\ell} \left(1 - \frac{m}{i}\right)^{j-\ell}, \tag{4}$$

where p_{ij} is the one-step transition probability for $\{N_n\}$. We should note

that this model assumes that the transitions in the Markov chain $\{N_n\}$ are independent of the genetic make-up of the population and that the population size does not affect the random sampling for the next generation. Again the mean first passage time to fixation or loss has not been obtained in closed form for this model: however, by investigating the eigenvalues of the one-step transition matrix, one can obtain as $t \to \infty$ the limiting behavior of the distribution of the first passage time (Karlin, 1968).

In Section 2, we generalize Wright's diffusion model to a variable population size situation. This generalization is analogous to Karlin's extension of the constant size Markov chain to the variable population size case. Since the analysis of this model is simpler then the Markov chain model, we are able to investigate a number of properties of our diffusion model that have not been investigated for other models.

Section 2. A RANDOM POPULATION SIZE MODEL

It will be convenient in this discussion to represent diffusion processes as stochastic differential equations. An introductory account of stochastic differential equations is given in Wong (1971).

The diffusion process, $\{X(t),\ t \geq 0\}$, for the Fisher-Wright fixed population size model can be represented by the stochastic differential equation

$$dX(t) = [\tfrac{1}{N} X(t)(1 - X(t))]^{\frac{1}{2}}\, dW(t), \qquad (5)$$

where N is the total initial population size and $\{W(t)\}$ is a standard Brownian motion with infinitesimal mean (drift) 0 and infinitesimal variance 1. $X(t)$ is a diffusion on the interval $(0, 1)$ with exit boundaries at 0 and 1, representing loss or fixation of type A_1.

The Fisher-Wright diffusion model can be extended to the situation where the population size varies with time. Let $N(t)$ denote the size of the total population at time t. $N(t)$ could be either a stochastic or a deterministic process. Let $Y(t)$ denote the proportion of individuals of type A_1 in the population at time t. We assume that $\{Y(t), t \geq 0\}$ is determined by the process $N(t)$ and the stochastic differential equation

$$dY(t) = [\frac{1}{N(t)} Y(t)(1 - Y(t))]^{\frac{1}{2}} dW(t), \tag{6}$$

where $\{W(t)\}$ is a standard Brownian motion as above. The process $W(t)$ is assumed to be independent of the population process $N(t)$. The independence of the $N(t)$ and $W(t)$ processes corresponds to the assumption in Karlin's model that the Markov chain for the total population size $\{N_n\}$ has transition probabilities that do not depend on the gene frequency and that the random sampling for the next generation is not affected by the population size.

Let $\{X_1(t), t \geq 0\}$ denote the diffusion process defined by (5) with population size $N = 1$. Roughly speaking, we wish to make a random time change on the t scale, say $M(t)$, so that the process $X_1(M(t))$, defined in the new time, obeys the stochastic differential equation given by (6). We use a result derived in a more general setting by Volkonskii (1961).

Theorem 1: *Let* $M(t)$ *denote the random time change*

$$M(t) = \int_0^t \frac{du}{N(u)}, \tag{7}$$

and let $X_1(t)$ *denote the diffusion defined in (5) with* $N = 1$. *Let* $Y(t)$ *be defined by*

$$Y(t) = X_1(M(t)); \tag{8}$$

then $\{Y(t)\}$ *is described by the stochastic differential equation (6) and the process* $\{N(t)\}$.

Proof: For a given realization of the population process, $\{N(t)\}$, from Volkonskii (1961, page 44) it follows that the process defined by (8) is a diffusion with zero drift and infinitesimal variance

$$\sigma(y,t) = \frac{1}{N(t)} y(1-y).$$

Thus the process $\{Y(t)\}$ is completely determined by the $\{N(t)\}$ process and the stochastic differential equation (6).

We now consider the time to loss or fixation in our random population size model. Let T_Y and T_1 denote the first passage times to the boundaries 0 or 1 in the $Y(t)$ and $X_1(t)$ diffusions respectively.

Theorem 2: *If* $\Pr\{\lim_{t\to\infty} M(t) = \infty\} = 1$; (9)

then, $T_Y = M^{-1}(T_1)$. (10)

where M^{-1} *is the inverse function of* M. *Let*

$$\bar{m}(t) = E[M^{-1}(t)], \tag{11}$$

then

$$E[T_Y] = E[\bar{m}(T_1)]. \tag{12}$$

Proof: Since $M(t)$ is a strictly increasing function of t, M is a one-to-one function so that its inverse M^{-1} exists. Now, (7) and (9) together imply that with probability 1 the range of $M(t)$ is $[0,\infty]$ and thus M^{-1} is a function that maps the interval $[0,\infty]$ into itself, and hence (10). Furthermore, T_X and M^{-1} are stochastically independent, since they depend only on the two independent processes $\{X_1(t)\}$ and $\{N(t)\}$ respectively, thus

$$E[T_Y] = E[E[M^{-1}(T_1)|T_1]] = E[\bar{m}(T_1)]. \tag{13}$$

The problem of finding the expected time to fixation or loss is reduced now to finding $\bar{m}(t)$ for the particular population process, $N(t)$, being investigated. From (7) we have

$$\frac{dM(t)}{dt} = \frac{1}{N(t)}, \tag{14}$$

and thus

$$\frac{dM^{-1}(t)}{dt} = N[M^{-1}(t)]. \tag{15}$$

Integrating (15) and taking the expectation we have

$$\bar{m}(t) = \int_0^t \bar{n}(u)du, \tag{16}$$

where

$$\bar{n}(t) = E[N(M^{-1}(t))]. \tag{17}$$

These results yield an immediate conclusion:

Corollary: If $\bar{n}(t)$ is an increasing function, then

$$E[T_Y] \geq \bar{m}[E(T_X)]. \tag{18}$$

If $\bar{n}(t)$ is a decreasing function then inequality (18) is reversed.

Proof: Equation (16) implies that if $\bar{n}(t)$ is an increasing function then $\bar{m}(t)$ is convex and the result follows from Jensen's inequality and (13). If $\bar{n}(t)$ is decreasing $\bar{m}(t)$ is concave and the inequality in (18) is reversed.

Let $N^*(t)$ denote the population process with the random time change $M^{-1}(t)$.,

$$N^*(t) = N[M^{-1}(t)]. \tag{19}$$

The evaluation of the $\{N^*(t)\}$ process is central to our problem. The case where $\{N(t)\}$ is a strong Markov process has been investigated by Volkonskii (1958). The main result is that if $N(t)$ is a strong Markov process then $\{N^*(t)\}$ is also a strong Markov process. In particular if $\{N(t)\}$ is a diffusion with infinitesimal mean $\mu(x)$ and infinitesimal variance $\sigma^2(x)$, then $N^*(t)$ is a diffusion with infinitesimal mean and variance

$$\mu^*(x) = x\mu(x), \quad \text{and} \quad \sigma^{*2}(x) = x\sigma^2(x). \tag{20}$$

Thus those population processes that are strong Markov processes can be easily handled by the methods outlined. Clearly if $N(t)$ is deterministic then $N^*(t)$ can be found using straightforward analysis.

Section 3. THE MEAN FIRST PASSAGE TIME IN THE CONSTANT AND VARIABLE POPULATION SIZE MODELS

Since the mean first passage for Wright's constant population size model is known, it is natural to ask under what conditions are the mean first passage times for the constant and variable size population models the same. If

$$\bar{n}(t) = N(0) = N_0 \tag{21}$$

in the variable size model then from (15) we have

$$\bar{m}(t) = N_0 t$$

which together with (13) yields

$$E[T_Y|N_0] = E[N_0 T_1] = E[T_X|N_0].$$

Thus the variable population size model and Wright's model, with $X(0) = Y(0) = p$ and $N(0) = N_0$, have the same mean first passage time if (21) holds.

Note that if (21) holds, the variance of the first passage time in the variable size model is greater than in the fixed size model, since

$$\begin{aligned} \mathrm{Var}[T_Y|N_0] &= \mathrm{Var}[M^{-1}(T_1)] \\ &= \mathrm{Var}_{T_1}[M^{-1}(T_1)|T_1] + E_{T_1}[\mathrm{Var}\, M^{-1}(T_1)|T_1] \\ &\geq \mathrm{Var}(\bar{m}(T)) = \mathrm{Var}\,[T_X|N_0]. \end{aligned}$$

The third line follows from dropping the last term on the left-hand side and again noting the independence of $\{M^{-1}(t)\}$ and T_1.

The problem now remains to find processes $\{N(t)\}$ which satisfy (21). With some ingenuity one can find processes which satisfy (21). Here we will restrict our attention to the class of diffusion processes with zero drift.

Theorem 3: *Let* $\{N(t)\}$ *denote a diffusion on the interval* (a,b), $0 \leq a < b \leq \infty$. *If the diffusion* $\{N^*(t)\}$ *defined by (19) has natural or exit boundaries at* a *and* b *and zero drift* $(\mu^*(x) = 0)$, *and if* $\sigma^*(x)$, *defined by (20), satisfies the following two regularity conditions:*

1. *There exists a* $K \geq 0$ *such that for all* x

$$\sigma^*(x) = K\sqrt{1 + |x|^2}. \tag{23}$$

2. *For every* $c > 0$, *there exists an* L_c *such that for* $|x| \leq c$, $|y| \leq c$,

$$|\sigma^*(x) - \sigma^*(y)| \leq L_c|x - y|; \tag{24}$$

then

$$E[T_Y|N_0] = E[T_X|N_0]. \tag{25}$$

Proof: From the result of Volkonskii (1958, p. 312, Theorem 3), $\{N^*(t)\}$ is a diffusion process with diffusion coefficient $\sigma^{*2}(x)$ given by (19). The forward diffusion equation for the $N^*(t)$ process is

$$\frac{\partial f(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2 \sigma^{*2}(x) f(x,t)}{\partial x^2}.$$

where $f(x,t)$ is the probability density function for $N^*(t)$.

Feller (1952) has shown that if $\sigma^*(x)$ is such that the boundaries a and b are either natural or exit boundaries, then the forward equation has a unique solution. If $\sigma^*(x)$ also satisfies the regularity conditions given in (23) and (24), then, from Skorokhod (1965, p. 120),

$$N^{\ddagger}(t) = N^{\ddagger}(0) + \int_0^t \sigma^*[N^{\ddagger}(u)]dW(u) \tag{26}$$

is a diffusion process with diffusion coefficient $\sigma^{*2}(x)$, and (26) has a unique solution with probability 1. Therefore, with probability 1, $\{N^{\ddagger}(t)\}$ defined by (26) represents the diffusion process $\{N^*(t)\}$. But $\{N^{\ddagger}(t)\}$ considered as a function of the upper limit t of the stochastic integral in (26) is a martingale (Doob, 1953, p. 444, Theorem 5.1). Thus (21) is satisfied and this implies that (25) holds. Our proof is complete.

We now consider a particular example of Theorem 3. This example is developed in detail in Hsieh (1972). Let $N(t)$ denote the diffusion approximation to the branching process with mean family size 1. This diffusion process has been investigated by Feller (1951, p. 235). Under the assumption that the mean family size is one, the infinitesimal mean and variance are

$$\mu(x) = 0 \quad \text{and} \quad \sigma^2(x) = x^{\frac{1}{2}} \tag{27}$$

respectively. $\{N_b(t)\}$ is a diffusion on the interval $(0,\infty)$ with an exit boundary at 0. Note that under these assumptions the population is certain to

die out, so that for this model the first passage time to fixation or loss, T_Y, is just the time to loss of type A_1 or type A_2, whichever comes first. This random drift process $\{Y_n(t)\}$ is completely defined by the stochastic differential equation (6) and by $\{N_b(t)\}$.

It is easy to show that the process $\{N_b(t)\}$ satisfies all the conditions of Theorem 3 and therefore the mean first passage to fixation or loss is the same as Wright's constant size model, given by (4). However, for this example one can also obtain the distribution of the first passage explicitly (Hsieh, 1972). We briefly outline the argument. Let

$$\left.\begin{aligned} Z_1(t) &= N_b(t)\, Y_b(t) \\ Z_2(t) &= N_b(t)(1 - Y_b(t)) \end{aligned}\right\}, \tag{28}$$

$Z_1(t)$ and $Z_2(t)$ are the number of individuals of type A_1 and A_2 respectively. It can be shown that $\{Z_1(t)\}$ and $\{Z_2(t)\}$ are independent diffusion processes with infinitesimal mean and variances given by (27). With probability one both processes will reach the exit boundary 0. Let T_1 and T_2 denote the first passage time to the boundary 0 in the $Z_1(t)$ and $Z_2(t)$ processes respectively. From a known result (Feller, 1951), the distributions of the random variables T_i, $i = 1,2$, are

$$F_i(t) = \Pr(T_i \le t \,|\, Z_i(0) = z_i) = \exp(-2z_i/t). \tag{29}$$

The first passage time T_Y can be represented by

$$T_Y = \min[T_1, T_2],$$

but in this example, T_1 and T_2 are independent and thus

$$F_{T_Y}(t) = \Pr(T_Y \le t \,|\, Z_1(0) = z_1,\ Z_2(0) = z_2) = \prod_{i=1}^{2} (1 - F_i(t)). \tag{30}$$

From (29) and (30) one can compute the moments of T_Y,

$$E[T_Y] = -2[z_1 \ln (z_1/(z_1+z_2)) + z_2 \ln (z_2/(z_1+z_2))]. \quad (31)$$

Since $p = z_1/N$ and $N = z_1 + z_2$, we see that (31) and (4) are identical. The higher moments of the random variables T_X and T_Y, however, are not identical. In fact for our independent branching process example all higher moments are infinite whereas in Wright's model all moments are finite. In Figure 1, the probability density function for the first passage time for the branching process example and for Wright's model are compared when $p = .5$ and $N = 100$.

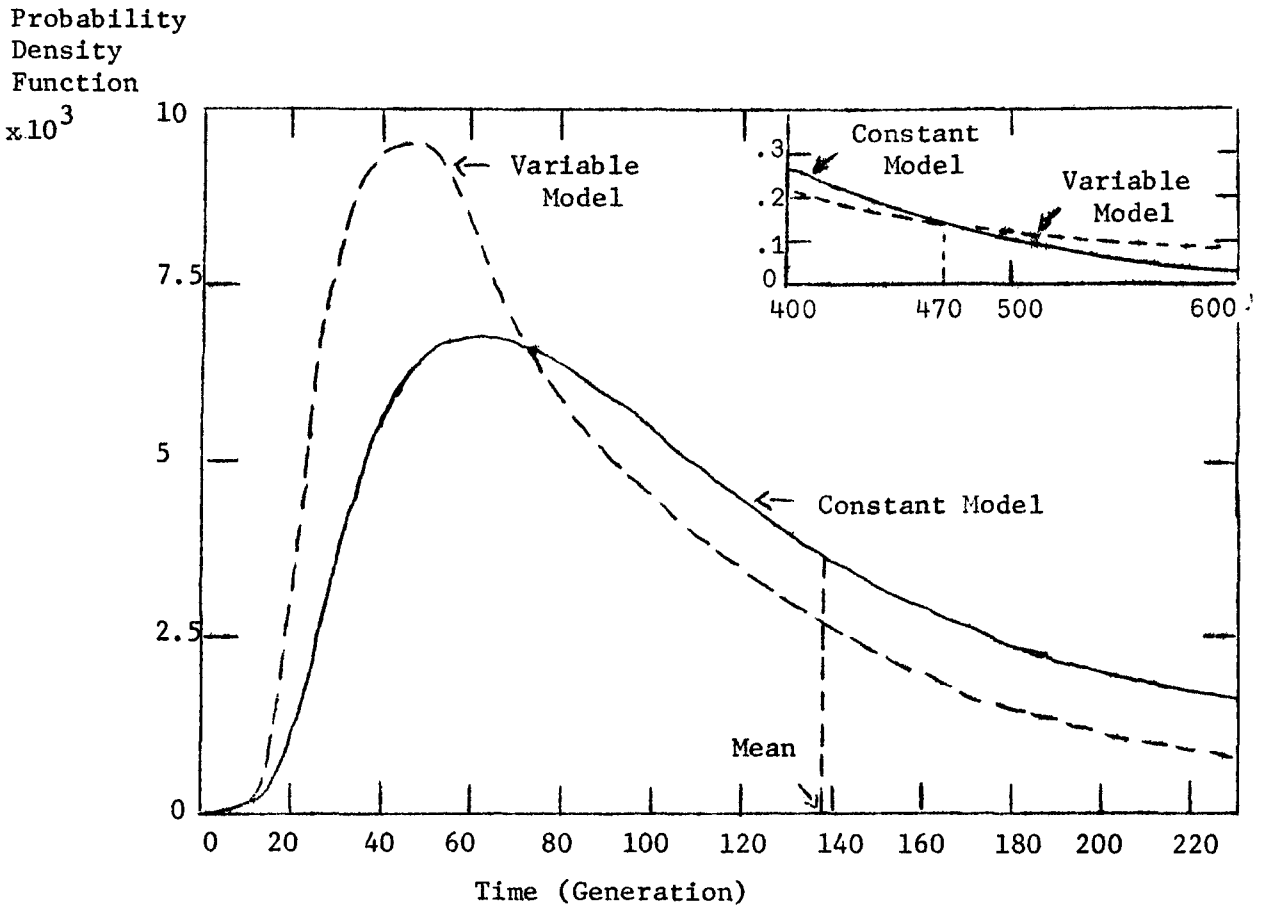

FIGURE 1: Graphs of the Two p.d.f.'s of the First-Passage Time to Fixation or Loss, T_X and T_Y, for $N = 100$ and $p = 0.5$.

Section 4. PROBABILITY THAT BOTH GENE TYPES REMAIN INDEFINITELY IN A GROWING POPULATION

In many natural populations we find that two or more gene types are present. Several mechanisms could account for this: mutation, polymorphism, etc. However, there is a rather simple explanation in terms of the model developed in Section 2. If the population grows at a fast enough rate, then the two gene types could remain in the population indefinitely. The question that then arises is how fast must the population grow for there to be a positive probability that neither of the two gene types will become extinct, that is

$$\Pr\{0 < \lim_{t \to \infty} Y(t) < 1\} > 0. \tag{32}$$

Since $M(t)$ is an increasing function of t, its limit as $t \to \infty$ exists,

$$\lim_{t \to \infty} M(t) = M(\infty), \tag{33}$$

where $M(\infty)$ is either a finite positive number or ∞. Consequently, from (8) the limit in (32) exists. If

$$\Pr\{M(\infty) < \infty\} > 0 \tag{34}$$

then there exists a $b < \infty$ and a $\delta > 0$ such that

$$\Pr\{M(\infty) < b\} = \delta.$$

From the independence of the $\{M(t)\}$ and $\{X_1(t)\}$ processes we have

$$\begin{aligned}\Pr\{0 < X_1(M(\infty)) < 1\} &\geq \Pr\{0 < X_1(M(\infty)) < 1 \text{ and } M(\infty) < b\} \\ &\geq \Pr\{0 < X_1(b) < 1 \text{ and } M(\infty) < b\} \\ &= \Pr\{0 < X_1(b) < 1\}\ \delta > 0.\end{aligned}$$

Thus when (34) holds there is a positive probability that neither of the gene types will become extinct in the population.

We consider now the special case where the population process $N(t)$ is deterministic and defined by

$$N(t) = N_0 + at^{\rho}, \quad \text{with} \quad N_0, a, \rho > 0.$$

From the definition of $M(t)$ given by (7) we see that if $\rho > 1$, then $M(\infty) < \infty$, and thus there is a positive probability that neither of the two types will become extinct in the population. Since for the deterministic case $M(\infty)$ is a constant, this probability depends only on the distribution of the time, T_1, until fixation or loss in the $\{X_1(t)\}$ process,

$$\Pr\{0 < X_1(M(\infty)) < 1\} = \Pr\{T_1 > M(\infty)\} = 1 - F_1(M(\infty)),$$

where $F_1(x)$ is cumulative distribution function for the random variable T_1. The function $F_1(x)$ is given in Kimura (1957).

ACKNOWLEDGMENT

We are indebted to Mr. Kelvin Lee for computing from Kimura's results the probability density function for the first passage time given in Figure 1.

REFERENCES

Doob, J.L. 1953. *Stochastic Processes.* New York: Wiley.

Ewens, W.J. 1969. *Population Genetics.* London: Methuen.

Feller, W. 1951. Diffusion processes in genetics. *Proc. Second Berkeley Symp. Math. Stat. Prob.:* 227-246.

Feller, W. 1952. The parabolic differential equations and the associated semi-group of transformations. *Ann. Math.* 55: 468-519.

Feller, W. 1954. Diffusion processes in one dimension. *Trans. Amer. Math. Soc.* 77: 1-31.

Fisher, R.A. 1922. On the dominance ratio. *Proc. Roy. Soc. Edinb.* 42: 321-341.

Fisher, R.A. 1930. *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press.

Hsieh, J.J. 1972. *A Bivariate Branching Process and Its Diffusion Approximation - A Stochastic Model for Variable Sized Haploid Populations,* Ph.D. Dissertation. University of North Carolina, Chapel Hill.

Karlin, S. 1968. Rates of approach to homozygosity for finite stochastic models with variable population size. *Amer. Natur.* 102: 443-445.

Kimura, M. 1957. Some problems of stochastic processes in genetics. *Ann. Math. Stat.* 28: 882-901.

Skorokhod, A.V. 1965. *Studies in the Theory of Random Processes.* (Translation from the Russian.) Reading, Pa.: Addison-Wesley.

Volkonskii, V.A. 1958. Random substitution of time in strong Markov processes. *Theory Prob. Application.* (Translation from the Russian.) 3: 310-326.

Volkonskii, V.A. 1961. Construction of non-homogeneous Markov processes by means of a random substitution of time. *Theory Prob. Applications* (Translation from the Russian.) 6: 42-51.

Wong, Eugene 1971. *Stochastic Processes in Information and Dynamical Systems.* New York: McGraw-Hill.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16: 97-159.

Wright, S. 1945. The differential equation of the distribution of gene frequencies. *Proc. Nat. Acad. Sci.* 31: 382-389.