Genet. Res., Camb. (1981), **37**, *pp.* 33–46
Printed in Great Britain

The detection of a recessive visible gene in finite populations*

By SAMUEL KARLIN† and SIMON TAVARE†‡
† *Department of Mathematics, Stanford University, Stanford, California 94305*
‡ *Department of Mathematics, University of Utah, Salt Lake City, Utah 84112* (permanent address).

(*Received* 1 *February* 1980, *and in revised form* 17 *June* 1980)

1. INTRODUCTION

Robertson, 1978, addressed the interesting problem of ascertaining the distribution of the time to detection of a recessive homozygote in a finite population. He was motivated in part by breeding and artificial selection practices. The same problem arises in the context of evolutionary processes and medical genetics since it refers to the time of first appearance as a homozygote of a new crossover event or a mutant gene.

The model investigated is as follows. Consider a finite population of N diploid individuals, comprising the two genotypes AA and Aa. Its composition changes over successive generations under the effects of random mating and finite sampling. Robertson studied the case of visible recessive homozygotes where the process of diploid formation continues as long as no recessive genotypes are formed, whereas if an aa-gene does appear, the process terminates. Due to the finite population size it is possible for loss of the a-allele to occur prior to formation of a recessive homozygote even in populations starting with many heterozygous carriers.

Under this framework, Robertson estimated the mean number of generations until the first visible recessive homozygote appears, using simulation techniques and matrix numerical methods. These times were calculated for a range of population sizes (N), and on the basis of these results, he astutely suggested the order of magnitude of $N^{\frac{1}{3}}$ generations for the expected time to detection.

The objective of this paper is to examine Robertson's model by analytic means, and to study a number of variations. We will now indicate briefly several issues that will be considered. Some further extensions are indicated in the discussion. (1) What is the role of the initial population composition? (2) What is the probability that the detection of a homozygous aa-genotype will occur before loss of the a-allele? (3) Evaluation of a number of functionals of the process, including the aggregate number of heterozygous carriers ever occurring in the population until detection of the first recessive homozygote. (4) The moments of the detection and loss times, and of the sojourn times of the process between two prescribed values y_1 and y_2.

It is also relevant to investigate two associated processes derived by appropriate conditionings. Specifically, the primary conditioned process that concerns us

* Supported in part by NIH Grant 5RO1 GM10452-17 and NSF Grant MCS79-24310.

restricts attention to those realizations of the model ending in detection of a recessive homozygote. Another pertinent conditioning focuses attention only on those realizations leading to loss of the a-allele. The construction of conditioned processes occurs intrinsically in the study of transitions among mutant lines in population genetic theory (e.g. Robertson, 1960; Kimura, 1971; Ewens, 1979, p. 125). A new aspect in the present context is that the first conditioning event can happen from any population state.

A specific functional of interest for the process conditioned on detection is to determine the distributional properties of the number of heterozygotes (carriers) at the time of detection.

In our analysis we employ the traditional method of diffusion approximation to the underlying discrete model. The novelty in the analysis is that the underlying discrete process has not only an absorbing state, but also a killing rate which depends on the population configuration. This killing rate corresponds to the event of detection of the homozygous recessive genotype. Such processes already occur in the population genetics literature, one example involving the determination of the probability that a recombinant type appears before fixation (Karlin, McGregor & Bodmer, 1966), another involving the formation of high order mutants (Karlin, 1973).

A classical result shows that when the population size N is moderately large with no selection, a new mutant occurring in the population will eventually be fixed with probability $1/2N$, having mean time to fixation about $4N$ generations, e.g. Kimura, 1970. Robertson & Narain, 1971, found that if the recessive homozygote is lethal, the expected time until its elimination requires about $\sqrt{N}$ generations, a much shorter time frame. The same order of magnitude holds for the case in which there is strong selection against recessive homozygotes (Guess & Levikson, private communication). The approximating diffusions for both these models are of the usual kind described completely by the associated infinitesimal drift and variance parameters. For the present detection model, the outcome indicating the first appearance of a recessive kills the breeding line in which it occurs and the corresponding diffusion is now governed by infinitesimal drift and variance effects as usual, but also an infinitesimal killing rate.

We would like to emphasize that the representation of the problem in terms of a diffusion process allows us to compute a wide variety of analytical results for the problem which are intractable (in closed form at least) for the discrete problem. Further, within the diffusion framework, the models are much simpler to generalize and analyse, and therefore we are led more readily to a fuller description of the process of gene formation and detection. Some of these extensions are indicated in the summary.

2. THE MODEL

We consider a monoecious population of N diploid individuals, and a single locus at which there are two possible alleles, denoted A and a. As explained in the introduction, we are interested in the time to formation of the first recessive homozygote, aa. It is convenient to let X_n be the number of heterozygotes at time n; then, assuming we have not yet seen a recessive homozygote and that $X_n = i$, the number of A-alleles at time n is $2N-i$, and the number of a-alleles is i.

If we assume random mating in the population, then, under the usual Wright-Fisher finite population reproduction scheme, the next generation will comprise $N-j$ AA, $j-k$ Aa, and k aa with probability

$$\frac{N!}{(N-j)!\,(j-k)!\,k!}\left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^{j-k}\left(1-\frac{i}{2N}\right)^{2(N-j)}\left(\frac{i^2}{4N^2}\right)^{k}. \qquad (1)$$

Since any recessive homozygotes are detected instantly, it follows from (1) that if $X_n = i$, then $X_{n+1} = j$ with probability

$$p_{ij} = \binom{N}{j}\left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^{j}\left(1-\frac{i}{2N}\right)^{2(N-j)}, \; i, j = 0, 1, \ldots, N. \qquad (2)$$

A transition probability matrix such as that given in (2) is called *sub-Markovian*, because the transition probabilities no longer sum to unity. This follows because the process can be killed by the appearance of a recessive homozygote.

In the discrete state space case, we can add on an extra state H, say, to transform the process into a normal transition probability matrix. The transition probabilities into the state H are given by

$$p_{iH} = 1-\sum_{j=0}^{N} p_{ij} = 1-\left(1-\frac{i^2}{4N^2}\right)^{N}, \quad 0 \leqslant i \leqslant N. \qquad (3)$$

while

$$p_{HH} = 1, \quad p_{Hi} = 0, \quad 0 \leqslant i \leqslant N. \qquad (4)$$

Of course, state 0 is absorbing, since once the population comprises only AA-genotypes it remains so thereafter.

What is the probability of never observing a recessive homozygote? This is just the probability of reaching state 0 before reaching state H, i.e. before the appearance of a recessive homozygote. In principle, this can be solved by finding the solutions $(u_0, u_1, \ldots, u_N)$ to the system of equations

$$u_i = \sum_{j=1}^{N} p_{ij}u_j + p_{i0}, \quad 1 \leqslant i \leqslant N, \qquad (5)$$

under the boundary condition $u_0 = 1$. This system can be solved numerically for small values of N, but it seems difficult to get explicit results for the u_i. Secondly, what is the mean time taken to observe a homozygous recessive, conditional on this event occurring? Or, what is similar, what is the mean time until the process stops, either by reaching the absorbing state 0 or the killing state H? For small values of N it is possible to use standard theory and numerical methods to get solutions to these problems. Some results along these lines are given by Robertson (1978).

In order to analyse this model further, we will resort to the method of diffusion approximation. This will enable us to determine the appropriate *time-scale* and *state-space scale* under which a recessive homozygote will be seen 'instantly', 'never', or when the process can be modelled in a way that allows both possibilities.

3. DIFFUSION APPROXIMATIONS

The analysis proceeds in the usual way, but with the added contingency of killing corresponding to the detection event. We have to scale the state space and time space to produce a limiting diffusion process. To this end, we define $\Delta X = X_{n+1} - X_n$. Given that no recessive homozygote appeared in generation n, we have

$$\begin{aligned} E(\Delta X | X_n = i) &= \sum_{j=0}^{N} j p_{ij} - i \sum_{j=0}^{N} p_{ij} \\ &= \frac{-i^2}{2N} + \text{lower order terms}, \end{aligned} \tag{6}$$

while

$$E((\Delta X)^2 | X_n = i) = i + \text{lower order terms}, \tag{7}$$

and

$$Pr(\text{killed in time } (n, n+1) | X_n = i) = \frac{i^2}{4N} + \text{lower order terms}. \tag{8}$$

We now need to find the correct time and state space scalings to compute the infinitesimal parameters of the associated diffusion process.

By examination of (6)–(8), we determine that the sequence of processes depending on the parameter N (the population size) defined by rescaling time and the state variables in the form

$$Y^{(N)}(t) = \frac{X([(2N)^{\frac{1}{2}}t])}{(2N)^{\frac{1}{2}}}, \quad t \geqslant 0 \tag{9}$$

will converge (as $N \to \infty$) to a limiting diffusion process $\{Y(t), t \geqslant 0\}$ having state space $[0, \infty)$ and continuous positive time parameter. The interpretation and identifications of the methodology are as follows. One unit of time, $t = 1$, for the process $Y^{(N)}$ corresponds to $(2N)^{\frac{1}{2}}$ generations of the original process $\{X_n\}$. Moreover, we are keeping track of fluctuations in heterozygote numbers to the order of $N^{\frac{1}{2}}$. (The factor $2^{\frac{1}{2}}$ that appears in (9) is for notational convenience later on.) The limiting diffusion $Y(t)$ has infinitesimal parameters analogous to (6)–(8) given by

$$\mu(x) = 0 \text{ mean coefficient}, \tag{10}$$

$$\sigma^2(x) = x \text{ variance coefficient}, \tag{11}$$

$$k(x) = \frac{x^2}{2} \text{ killing rate}. \tag{12}$$

Accordingly, for the approximating diffusion $Y(t)$, we have

$$\begin{aligned} &E(Y(t+h) - Y(t) | Y(t) = x) = o(h) \\ &E((Y(t+h) - Y(t))^2 | Y(t) = x) = xh + o(h) \\ &Pr(Y(t) \text{ is killed during } (t, t+h) | Y(t) = x) = \frac{x^2}{2} h + o(h) \end{aligned} \tag{13}$$

($o(h)$ means terms of smaller order than h).

It is important to remark that the boundary point 0 is an exit (or absorbing) state for the process $Y(t)$. This means that the approximating diffusion mirrors the behaviour of the underlying discrete process, where 0 is an absorbing state (reflecting the loss of the a-allele).

We make a few important preliminary statements. Firstly, as can be established by examining (6)–(8), the scalings used to derive the diffusion approximation are *unique*. A demonstration of the uniqueness of this scaling is presented in the appendix so as not to interrupt this discussion. From this, we conclude that if $X_0 = i \ll N^{\frac{1}{3}}$ (i much smaller than $N^{\frac{1}{3}}$), then the AA genotype is established 'instantly', i.e. fixation is effectively the only possibility, while if $i \gg N^{\frac{1}{3}}$, then detection of the aa-recessive homozygote occurs 'instantly', i.e. detection is effectively the only possibility. When $X_0 = i$ is of the order $i \approx N^{\frac{1}{3}}$ then there is a positive probability of fixation of AA *or* detection of the aa homozygote, and in this case, the time to detection will be of order $N^{\frac{1}{3}}$ generations.

4. ANALYSIS OF THE PROCESS

(i) *The detection probability.* Let us define $u(x, t) = \Pr\{Y \text{ process has not been killed by time } t | Y(0) = x\}$. From diffusion theory (e.g. Karlin & Taylor, 1980, ch. 15), $u(x, t)$ satisfies the differential equation

$$\frac{\partial u}{\partial t} = \frac{x}{2}\frac{\partial^2 u}{\partial x^2} - \frac{x^2}{2}u, \quad x > 0, \quad t > 0 \tag{14}$$

with initial condition $u(x, 0) = 1$ for all $x > 0$ and boundary condition $u(0, t) = 1$ for all $t > 0$. It is clear that $u(x, t)$ is a decreasing function of t, and, since

$$u(x, t) \geqslant 0,$$

the limit

$$u(x) = \lim_{t\to\infty} u(x, t) \tag{15}$$

exists, and is equal to the probability that the process is never killed (that is, that a homozygous recessive is never detected). Since 0 is an absorbing attainable boundary, we see that $u(x)$ must be the probability of fixation at 0, the probability that the population will comprise only A alleles. From (14), we find that $\partial u/\partial t \to 0$ as $t \to \infty$ and hence $u(x)$ satisfies the differential equation

$$u''(x) - xu(x) = 0 \tag{16}$$

subject to the boundary conditions $u(0) = 1$, $u(\infty) = 0$.

The equation (16) is known as Airy's equation, and occurs in studies of radio waves and light spectra (Airy, 1838). There are two standard solutions of this equation, the so-called Airy functions of the first and second kind. These are explicitly represented in the form

$$A(x) = \frac{x^{\frac{1}{2}}}{3}\left(I_{-\frac{1}{3}}\left(\frac{2x^{\frac{3}{2}}}{3}\right) - I_{\frac{1}{3}}\left(\frac{2x^{\frac{3}{2}}}{3}\right)\right), \tag{17a}$$

and

$$B(x) = \frac{x^{\frac{1}{2}}}{\sqrt{3}}\left(I_{-\frac{1}{3}}\left(\frac{2x^{\frac{3}{2}}}{3}\right)+I_{\frac{1}{3}}\left(\frac{2x^{\frac{3}{2}}}{3}\right)\right), \tag{17b}$$

where $I_v(.)$ is the modified Bessel function of order v. The behaviour of these functions is well studied, and has been tabulated (Miller, 1946). The function $A(x)$ strictly decreases from $A(0) < \infty$ to zero as x traverses 0 to ∞.

The probability $u(x)$ of AA-fixation is therefore given by

$$u(x) = \frac{A(x)}{A(0)}. \tag{18}$$

Hence the probability of detection is given by

$$v(x) = 1-u(x) = 1-\frac{A(x)}{A(0)}. \tag{19}$$

As an example of the use of (19), we give in Table 1 some values of the detection probability in populations of size N starting with one heterozygote. These are computed by taking $X_0 = 1 = x(2N)^{\frac{1}{3}}$ or $x = (2N)^{-\frac{1}{3}}$. The values of the Airy function are taken from Miller, 1946.

Table 1. *Probability of detection of a recessive homozygote starting from one a-allele in a population of size N. This uses v(x) given in (19)*

N	Diffusion	Numerical*
5	0·294	0·26
10	0·261	0·23
20	0·210	0·18
50	0·155	0·14
100	0·124	0·11
500	0·073	0·07

* Robertson, 1978, Table 1 based on the numerical solution of (5).

For small values of x, we compute the approximation

$$v(x) \approx 0{\cdot}7290x-0{\cdot}1667x^3,$$

obtained by retaining the first three terms from the expansion of (19). This approximation agrees with those in Table 1 for $N \geqslant 50$.

As another example of the use of (19), we compute an approximation to the number X_0 of heterozygotes necessary in the initial generation to ensure that the detection probability is at least $\frac{1}{2}$. We require x such that $1-A(x)/A(0) \geqslant \frac{1}{2}$. We find that $x \geqslant 0{\cdot}76$, and so, converting back in terms of the original process, we need X_0 at least $0{\cdot}96\ N^{\frac{1}{3}}$.

(ii) *The mean time to detection or loss.* It is easy to check that the mean time $M(x)$ to detection *or* loss solves the differential equation

$$\frac{x}{2}M''(x)-\frac{x^2}{2}M(x) = -1, \tag{20}$$

subject to the boundary conditions $M(0) = 0$, $M(\infty) = 0$. The required solution is

$$M(x) = 2\pi A(x)\int_0^x \frac{B(u)-\sqrt{3}A(u)}{u}\,du + 2\pi(B(x)-\sqrt{3}A(x))\int_x^\infty \frac{A(u)}{u}\,du, \quad (21)$$

where $A(x)$ and $B(x)$ are given in (17a) and (17b). For small values of x, (21) shows that $M(x) \approx Dx \ln x$, where D is a constant. From this we deduce that in the discrete model starting with 1 heterozygote, the order of magnitude of the time to loss *or* detection is $\ln (2N)$ generations. This is of the same order as the time to loss or fixation in the classical pure random drift model (cf. Kimura, 1970). Even though killing is possible, the predominant outcome from an initial single heterozygote is loss of the a-allele, and this happens in about $\ln (2N)$ generations.

(iii) *Aggregate number of heterozygote carriers.* One interesting functional of the process is the mean number of heterozygotes produced in the evolution of the population before loss or detection of allele a. This function, denoted by $H(x)$, gives one type of measure of the 'genetic costs' of having a deleterious mutant in the population. If T is time to loss or detection of the a-allele, then

$$H(x) = E\left(\int_0^T Y(u)du \,|\, Y(0) = x\right),$$

which is the natural analog of the discrete measure $E\left(\sum_{k=1}^{T} X_k | X_0 = i\right)$, the total number of heterozygotes that appear before detection or loss. It follows that

$$H(x) = 2\pi A(x)\int_0^x [B(u)-\sqrt{3}A(u)]\,du + 2\pi[B(x)-\sqrt{3}A(x)]\int_x^\infty A(u)\,du. \quad (22)$$

For small values of x we have $H(x) \approx xH'(0) = x(2\pi/3)\,[B'(0)-\sqrt{3}A'(0)] \approx 1{\cdot}88x$. To see what this implies for the discrete process, we note that $H(.)$ is in units of xt, and hence we approximate the discrete result by $(2N)^{\frac{2}{3}}H(.)$. We see that if we start with one heterozygote ($x = (2N)^{-\frac{1}{3}}$), then the expected total number of heterozygotes subsequently appearing in the population is given approximately by

$$H = 1{\cdot}88.\frac{1}{(2N)^{\frac{1}{3}}}.(2N)^{\frac{2}{3}} = 2{\cdot}37N^{\frac{1}{3}}.$$

(iv) *Other functionals.* The function

$$G(x, y) = \begin{cases} 2\pi[B(x)-\sqrt{3}A(x)]\dfrac{A(y)}{y}, & x \leqslant y \\ 2\pi\dfrac{A(x)}{y}[B(y)-\sqrt{3}A(y)], & y \leqslant x \end{cases} \quad (23)$$

is commonly called the Green function of the diffusion process and possesses the following interpretation. If we require the expected time the heterozygote numbers hover between y_1 and y_2 before detection or loss, the answer is given by

$$\int_{y_1}^{y_2} G(x, y)\,dy.$$

We have used (23) implicitly in deriving (21) and (22).

(v) *Mean time to detection, conditioned that detection occurs.* The problem that most interested Robertson was the mean time $M_K(x)$ for the process to end by detection, conditional on this event occurring. We find that

$$M_K(x) = \frac{2\pi A(0)A(x)}{A(0)-A(x)}\int_0^x \frac{B(u)-\sqrt{3}A(u)}{u}\left(1-\frac{A(u)}{A(0)}\right)du$$
$$+\frac{2\pi A(0)[B(x)-\sqrt{3}A(x)]}{A(0)-A(x)}\int_x^\infty \frac{A(u)}{u}\left(1-\frac{A(u)}{A(0)}\right)du. \qquad (24)$$

(The subscript K keeps in view the conditioning event of eventual detection.) For small values of x, $M_K(x)$ is approximately constant, from which we deduce the following. In populations of size N, the mean time to detection, conditional on this happening, is of order $CN^{\frac{1}{3}}$ generations, where C is given by

$$2^{\frac{1}{3}}\left\{\frac{2\pi[B'(0)-\sqrt{3}A'(0)]A(0)}{-A'(0)}\int_0^\infty \frac{A(u)}{u}\left(1-\frac{A(u)}{A(0)}\right)du\right\}.$$

The integral above can be evaluated explicitly using a limiting argument (based on the representation of Airy functions in terms of Bessel functions) on the integral given in Gradshteyn & Ryzhik, 1965, p. 693, (4). This shows that the integral has value $\Gamma(\frac{1}{3})/2(3^{\frac{5}{3}}) = 0{\cdot}214650$, and hence $C = 2{\cdot}090$. This is in close agreement with the result found by Robertson.

If we let $M_K^{(l)}(x)$ be the l-th moment of the detection time, then for very small initial numbers of heterozygotes, we may approximate $M_K^{(l)}(x)$ by $C_l = M_K^{(l)}(0)$, and hence in the discrete model, the l-th moment of the (conditional) detection time is of order $N^{l/3}\,2^{\frac{1}{3}}C_l$ generations.

(vi) *Aggregate numbers of heterozygotes conditioned on detection.* This is derived analogous to (iii). The result is

$$H_K(x) = \frac{2\pi A(0)A(x)}{A(0)-A(x)}\int_0^x [B(u)-\sqrt{3}A(u)]\left(1-\frac{A(u)}{A(0)}\right)du$$
$$+\frac{2\pi A(0)[B(x)-\sqrt{3}A(x)]}{A(0)-A(x)}\int_x^\infty A(u)\left(1-\frac{A(u)}{A(0)}\right)du. \qquad (25)$$

Its evaluation for *x small* (we can take $x = 0$) is the constant value

$$H_K(0) \approx \frac{2\pi A(0)[B'(0)-\sqrt{3}A'(0)]}{-A'(0)}\left(\frac{1}{3}-\frac{[A'(0)]^2}{A(0)}\right) = 1{\cdot}12.$$

Therefore, in terms of the discrete process the average total number of heterozygotes appearing before detection, given detection occurs and starting from 1 heterozygote, is approximately

$$H_K = 1{\cdot}12(2N)^{\frac{2}{3}} = 1{\cdot}78N^{\frac{2}{3}}.$$

The intuitive reason for the different order of magnitude between H and H_K may be explained as follows. Starting from one heterozygote, the over-whelming outcome is loss (as opposed to detection), which occurs rapidly. Those paths which lead to detection have to build up a substantial number of heterozygotes to avoid loss, and so the total number that occur is much larger.

(vii) *The maximum number of heterozygotes attained before detection conditioned on eventual detection.* For the conditioned process, with certain detection, the distribution of the maximum functional, i.e. the probability that from an initial state x the maximum exceeds $y > x$ prior to detection is

$$\frac{(B(x)-\sqrt{3}\,A(x))\,(A(0)-A(y))}{(B(y)-\sqrt{3}\,A(y))\,(A(0)-A(x))}.$$

In particular, starting from a single heterozygote the probability that the maximum exceeds y is approximately

$$\left(\frac{B'(0)-\sqrt{3}\,A'(0)}{-A'(0)}\right)\frac{A(0)-A(y)}{B(y)-\sqrt{3}\,A(y)} = 2\sqrt{3}\,\frac{A(0)-A(y)}{B(y)-\sqrt{3}\,A(y)}.$$

Table 2. *The probability of exceeding a level $\alpha(2N)^{\frac{1}{3}}$ heterozygotes with one initial heterozygote conditioned on detection*

α	Probability
0·1	0·998
0·4	0·964
0·7	0·890
1·0	0·782
1·3	0·652
1·7	0·468
2·0	0·343
3·0	0·086
3·5	0·037
4·0	0·015

(viii) *The rate of approach to loss or detection.* In order to assess the rate of approach to loss or detection, it is of interest to compute the leading eigenvalues of the process; that is, we wish to find the eigenvalues $\{\mu_n\}$ and corresponding eigenvectors $\{u_n(x)\}$ satisfying

$$xu_n''(x)-x^2u_n(x) = -2\mu_n u_n(x), \quad n = 1, 2, \ldots, \tag{26}$$

subject to $u_n(0) = 0$, $\int_0^\infty (1/x)\,u_n^2(x)\,dx < \infty$. It appears to be difficult to solve (26) explicitly, but it is possible to extimate $\{\mu_n\}$ and $\{u_n(x)\}$ using a numerical scheme (e.g. Rayleigh-Ritz method). The eigenvalue of particular interest in this context is μ_1, the smallest positive one. The rate of approach to loss or detection is then of the order $e^{-\mu_1 t}$. Using Rayleigh-Ritz, we obtained $\mu_1 = 1{\cdot}070$, the next two eigenvalues being $\mu_2 = 2{\cdot}740$, $\mu_3 = 4{\cdot}718$. Of course, the same eigenvalues apply to the model conditioned on detection.

To compare this to the discrete result, let μ be the largest non-unit eigenvalue of the discrete chain specified in (2) and (3). Recalling that one unit of diffusion time corresponds to $(2N)^{\frac{1}{3}}$ generations in discrete time, we see that

$$e^{-\mu_1} \approx \mu^{(2N)^{\frac{1}{3}}}, \quad \text{or} \quad \mu \approx \exp\{-\mu_1(2N)^{-\frac{1}{3}}\}. \tag{27}$$

In table (3) we compare the values of μ computed from (27) with the matrix results computed by Robertson.

Table 3. *Comparison of leading eigenvalue μ of discrete process estimated by diffusion method (27) and matrix results of Robertson (Table 1; $\lambda = 1-\mu$)*

$2N$	Diffusion method	Matrix method
10	0 61	0·62
20	0·67	0·68
40	0·73	0·74
100	0·79	0·80
200	0·83	0·83
1000	0·90	0·90

(ix) *The number of carriers at detection.* It can be shown that, given an initial value $Y(0) = x$, and conditional on detection occurring, the probability that detection occurs in the interval $(y, y+dy)$ is given by $w_x(y)\,dy$, where

$$w_x(y) = \begin{cases} \dfrac{\pi A(x) y(B(y) - \sqrt{3}A(y))}{1-u(x)}, & x \geqslant y \\ \dfrac{\pi(B(x) - \sqrt{3}A(x))\, yA(y)}{1-u(x)}, & x \leqslant y. \end{cases} \tag{28}$$

We can use (28) to ascertain the most likely number of carriers in the population when detection occurs. We will restrict our attention to the case in which we start with a very small number of heterozygotes (that is, we may take $x \to 0$ in (28)). In this case, we find that

$$w_0(y) = 2\sqrt{3}\pi A(0) yA(y), \quad y > 0. \tag{29}$$

The function $w_0(y)$ can be interpreted as a probability density function, with the interpretation that, for any interval $I = (a, b)$, the probability that detection occurs in the interval I is given by

$$\int_a^b w_0(y)\,dy = 2\sqrt{3}\pi A(0)\,[A'(b) - A'(a)]. \tag{30}$$

The density (29) has a maximum at the point y_0 satisfying the equation $yA'(y) + A(y) = 0$. Numerical solution yields $y_0 = 0{\cdot}885$. It follows that detection is most likely to occur at y_0, which corresponds to a frequency of $0{\cdot}885(2N)^{\frac{1}{3}} = 1{\cdot}12N^{\frac{1}{3}}$ heterozygotes in the discrete model.

5. DISCUSSION

Robertson, 1978 addressed in an analytic formulation the important problem of estimating the mean time to first appearance of a recessive visible gene in a finite population starting with a single heterozygote. This problem has interest with respect to artificial selection practices, in evolutionary studies concerned with the elapsed time to observation of new mutant types and for medical intervention and counselling programmes.

Robertson noted a number of examples in *Drosophila* of recessive visibles occurring in several selection lines, although only a few initial stocks are involved. Cases of rare genetic diseases exhibiting anomalous frequency estimates in certain population groups are sufficiently documented, often attributed to founder effects (e.g. Tay Sachs, Nieman Pick, Gaucher's syndrome), and are often only recognized several generations after the defective phenotype arose.

An essential problem motivated by the foregoing considerations is to estimate the distributional properties of the time to detection of a newly arising mutant (visible only as a recessive homozygote) in a finite population. By simulation and numerical methods Robertson suggested the conclusion that the mean time to uncover the existence of such recessive genes emanating from one heterozygote carrier is of relatively low dependence on population size, of the order $N^{\frac{1}{3}}$. (The occurrence of the gene more than once in the initial sample and the examination of more individuals than parents would undoubtedly reduce the times of detection.) The time scale $N^{\frac{1}{3}}$ stands in sharp contrast to the classical result on the expected time of establishment of a new mutant gene conditioned on its fixation which is about $4N$ generations. Stimulated by the work of Robertson, we investigated analytically his model and a number of extensions with the aid of diffusion process approximations. We confirmed his finding that the correct time scaling involves $N^{\frac{1}{3}}$ generations of the discrete process corresponding to unit time of the diffusion process, but then in this time scale, only fluctuations of heterozygote numbers of the order $N^{\frac{1}{3}}$ are discernible. That is to say, if the number of initial heterozygotes X_0 is much less than $N^{\frac{1}{3}}$, then only AA-fixation results (virtually instantly) and when X_0 is much larger than $N^{\frac{1}{3}}$, then quick detection of an aa-homozygote happens. For X_0 of the order $N^{\frac{1}{3}}$ one of both outcomes (loss versus detection) results and each has positive probability depending on the initial heterozygote numbers X_0.

In seeking an incisive analysis of these models there are two processes of prime relevance: (1) The realizations in the approximating diffusion process $\{Y(t), t \geqslant 0\}$ ($Y(t)$ = number of heterozygote carriers at time t) end in one of two mutually exclusive outcomes; either that of random elimination of the heterozygote carriers through repeated sampling, leading to AA-fixation, or detection of a visible recessive homozygote. The latter event is represented in the diffusion process by the operation of a killing rate depending on the numbers of heterozygotes, i.e. the state variable (cf. the studies of Karlin *et al.* [1967] on first recombination occurrences in a finite population). (2) The second diffusion process is derived as a conditioned diffusion from the original diffusion defined by restricting considerations only to the realizations of the process ending in killing (= detection). The time scale translated from this diffusion to the discrete case is again $N^{\frac{1}{3}}$ generations.

Both diffusions are remarkably tractable and relate to a classical differential equation analysed in studies on radio waves and light spectra. The basic solutions are known as the Airy functions of the first and second kind. They can be represented in terms of appropriate Bessel functions and in these terms are extensively tabulated.

We ascertained analytically the following functionals of these processes in terms of the initial heterozygote numbers x. (1) The probability $u(x)$ of loss of the recessive allele as against detection of a homozygote recessive, the latter occurring with probability $v(x)$ given explicitly by (19). (2) The expected time until either loss or detection. (Higher moments are also accessible.) (3) The cumulative number of heterozygotes over the population lifetime, i.e. until loss or detection.

We calculated for the conditioned process (conditioning on the detection outcome) the same functionals and also the distribution of the maximum heterozygote numbers attained prior to detection. It is worth highlighting several of these evaluations for an initial state corresponding to a single heterozygote. We determined then that the expected time until detection, conditioned on eventual detection, is $2{\cdot}09N^{\frac{1}{3}}$ generations. The aggregate average numbers of heterozygotes over the life of this process is $1{\cdot}78N^{\frac{1}{3}}$. During this period the probability exceeds $\frac{1}{2}$ that the maximum number of heterozygotes achieved a level exceeding $2{\cdot}02N^{\frac{1}{3}}$. The same calculations can be done for any initial population state and analytic formulas are available (see section 4, paragraphs (iv)–(vii)).

We have also tried to assess the effect on detection of examining more individuals than are used as parents in the subsequent generation. If we use N parents and examine M offspring (usually $M > N$) then the detection probability is increased, as expected. Suppose we are interested in ascertaining how many offspring we should examine to ensure that the detection probability is at least one half. Starting with one heterozygote in the initial parent population, we found that we need $M = 0{\cdot}88N^2$, meaning that we have to examine a very large number of offspring.

We have also analysed the effects of selection differentials in heterozygotes, and of recurrent mutation to the a-allele. The details of these analyses may be found in Karlin & Tavaré [1981], but the results may be summarized as follows. If we suppose that selection acts on the heterozygotes, the selection difference being s (which may be positive or negative), then the diffusion method shows that the scalings again have to be of order $N^{\frac{1}{3}}$, as long as the selection coefficient is of order at most $N^{-\frac{1}{3}}$. The 'usual' scaling of the selection coefficient in the Wright-Fisher model is N^{-1}; this result shows that we can have quite strong selection intensities before we significantly alter the probability of detection.

It is interesting that for all directions of selection (heterozygote advantage $s > 0$, or heterozygote disadvantage $s < 0$), the fixation probability is a monotone decreasing function of x. It is no longer exclusively convex for $s < 0$. The time to detection conditioned on this occurring, is again of order $N^{\frac{1}{3}}$ generations.

For models in which we allow mutation from the A-allele to the a-allele, and the mutation rate is taken to be of order N^{-1}, the correct order of magnitude for the time scale is again $N^{\frac{1}{3}}$. These results show that the order of magnitude $N^{\frac{1}{3}}$ is quite a robust result for a wide spectrum of genetic models concerned with detecting particular genotypes in finite populations.

We would like to thank Dr William G. Hill for discussions which stimulated our interest in this problem. We are also indebted to Dr Alan Robertson for many valuable comments on the manuscript.

REFERENCES

AIRY, SIR G. B. (1838). On the intensity of light in the neighbourhood of a caustic. *Transactions of Cambridge Philosophical Society* **6**, 379–402.

EWENS, W. J. (1979). *Mathematical Population Genetics*. Springer Verlag. New York.

GRADSHTEYN, I. S. & RYZHIK, I. M. (1965). *Tables of Integrals, Series and Products*, 3rd edition. New York: Academic Press.

KARLIN, S. (1973). Sex and infinity, a mathematical analysis of the advantages and disadvantages of genetic recombination. In *The Mathematical Theory of the Dynamics of Biological Populations* (ed. M. S. Bartlett and R. W. Hiorns), pp. 155–194. New York and London: Academic Press.

KARLIN, S., MCGREGOR, J. L. & BODMER, W. F. (1966). The rate of production of recombinants between linked genes in finite populations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. IV, University of California Press. pp. 403–414.

KARLIN, S. & TAVARÉ, S. (1981). The detection of particular genotypes in finite populations. *Theoretical Population Biology* (to appear).

KARLIN, S. & TAYLOR, H. M. (1980). *A second Course in Stochastic Processes*. New York: Academic Press.

KIMURA, M. (1970). Length of time required for a selectively neutral mutant to reach fixation through random drift in a finite population. *Genetical Research* **15**, 131–133.

KIMURA, M. (1971). Theoretical foundations of population genetics at the molecular level. *Theoretical Population Biology* **2**, 174–208.

MILLER, J. C. P. (1946). *The Airy Integral*. Mathematical tables. Part-vol. B. Cambridge University Press, England.

ROBERTSON, A. (1960). A theory of limits in artificial selection. *Proceedings of the Royal Society of London*, B **153**, 234–249.

ROBERTSON, A. (1978). The time to detection of recessive visible genes in small populations. *Genetical Research* **31**, 255–264.

ROBERTSON, A. & NARAIN, P. (1971). The survival of recessive lethals in finite populations. *Theoretical Population Biology* **2**, 24–50.

APPENDIX

Derivation of diffusion approximation (9)

In order to find the diffusion approximation described by (9) and (10)–(13), we ascertain values of the parameters α, β for which the processes

$$Y^{(N)}(t) = \frac{X([(2N)^{\alpha}t])}{(2N)^{\beta}}, \quad t \geq 0, \quad N \geq 2 \tag{31}$$

converges to a limiting diffusion $Y(t)$ as $N \to \infty$. (The coefficients 2 that appear in (31) are for notational ease only.) In (31), one unit of the discrete time scale corresponds to $\Delta t = (2N)^{-\alpha}$ time units on the new time scale. In order to evaluate which values of α, β (≥ 0) are admissible, we compute the following limits as $N \to \infty$ (i.e. $\Delta t \to 0$):

$$\mu(x) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} E(\Delta Y_N | Y_N(t) = x),$$

$$\sigma^2(x) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} E((\Delta Y_N)^2 | Y_N(t) = x), \tag{32}$$

and

$$k(x) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} Pr(Y_N(.) \text{ killed in } (t, t+\Delta t) | Y_N(t) = x)$$

where $\Delta Y_N \equiv \Delta Y_N(t) = Y^{(N)}(t+\Delta t) - Y^{(N)}(t)$ and $x = i/(2N)^{\beta}$. In the current problem, admissible values of α and β are those for which $k(x)$ and $\sigma^2(x)$ are finite and positive, and $\mu(x)$ is finite. We will show in this appendix that we must have $\alpha = \beta = \frac{1}{3}$, which reduces (31) to (9). Using (32), the resulting coefficients of the approximating process are given by (10)–(12).

Using the equations (6)–(8) and neglecting lower order terms, we arrive at

$$E(\Delta Y_N \mid Y^{(N)}(t) = x) = \frac{-x^2 \Delta t}{2(2N)^{1-\beta-\alpha}}, \tag{33a}$$

$$E((\Delta Y_N)^2 \mid Y^{(N)}(t) = x) = \frac{x \Delta t}{(2N)^{\beta-\alpha}}, \tag{33b}$$

and

$$\Pr\,(Y_N(.) \text{ killed in } (t, t+\Delta t) | Y_N(t) = x) = \frac{x^2 \Delta t}{2(2N)^{1-2\beta-\alpha}}. \tag{33c}$$

(32) combined with (33c) necessitates that $1-2\beta-\alpha = 0$, while (33b) shows that $\beta-\alpha = 0$. It follows that $\alpha = \beta = \frac{1}{3}$ is the (unique) required scaling. Finally, from (33a) and (31), $\mu(x) = \lim_{\Delta t \downarrow 0} (-x^2/2(2N)^{\frac{1}{3}}) = 0$, which is (12).