Reprinted from THEORETICAL POPULATION BIOLOGY
Vol. 19, No. 2, April 1981
Printed in Belgium

The Detection of Particular Genotypes in Finite Populations II. The Effects of Partial Penetrance and Family Structure*

SAMUEL KARLIN AND SIMON TAVARÉ†

Department of Mathematics, Stanford University, Stanford, California 94305

Received February 1, 1980

INTRODUCTION

The detection problem introduced by Robertson (1978) concerns the time taken to form the first recessive homozygote in finite populations. The diallelic model previously studied assumes that heterozygote carriers, Aa, are indistinguishable from the normal homozygote, AA. Various distributional properties of the time to detection or loss of the a-allele were determined using the traditional method of diffusion approximation to the relevant Wright–Fisher discrete process. The novelty in the diffusion structure here rests on the appearance of a killing rate corresponding to the aa-genotype detection events; see Karlin and Tavaré (1980). The analysis was extended in Karlin and Tavaré (1981a), referred to henceforth as Part I, to take account of various forms and levels of natural selection effects.

In this part we will study the effects of partial penetrance in heterozygotes or, equivalently, the effect of partial detection, resulting from a screening program that can in some cases detect the presence of the a-allele in heterozygous form. These considerations are appropriate in view of recent progress in biochemical techniques which permit greater ability to detect differences in phenotypically similar genotypes.

Utilizing the advancing technology, increasingly more genetic disease screening programs (e.g., sickle cell anemia, a number of thalassemia disorders, Tay Sachs syndrome, phenylketonuria) are available for purposes of identifying heterozygous carriers. In this perspective we envision the detection problem under the conditions where recessive homozygotes are instantly detected as before, but in addition, a heterozygous carrier can be

* Supported in part by NIH Grant GM10452-16 and NSF Grant MCS-80624-A01.
† Permanent address: Department of Mathematics, University of Utah, Salt Lake City, Utah 84112.

0040-5809/81/020215-15$02.00/0

ascertained or will express itself with probability α. If α is "significant" then the recessive allele will usually be first detected among heterozygotes. For α "very small," the heterozygotes are indistinguishable from normal homozygotes and allele a will usually be revealed with the appearance of a recessive homozygote. The distribution of the detection time differs sharply in these two cases. The detailed comparisons and contrasts of these models are set forth in the following section.

Another class of models takes account of family structure, motivated principally by artificial selection schemes (Section 2). 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. A quantitative assessment of this procedure is developed.

We also consider a model of N independent breeding individuals per generation which produce families of r offspring. All offspring are examined and the process is terminated if a recessive homozygote appears. This model can easily be extended to allow random family sizes.

The presence of family structure shows qualitatively the same distributional properties of the detection times as the case of families with one offspring. We also treat a mixed reproduction scheme in which any undetected recessive homozygotes are replaced by a random sample from a large population with fixed proportions of AA and Aa individuals.

One aim of this study is to assess the stability of the order of magnitude $N^{1/3}$ generations until detection that was first noted by Robertson (1978). As will be seen, in several of the problems studied here the time scale $N^{1/3}$ is no longer the only scaling that leads to an appropriate diffusion process with killing; further, the different orders of magnitude are often reflected by processes whose infinitesimal parameters are functionally different. In this analysis it is our intention to discuss qualitative aspects of these variants; the methods described in Karlin and Tavaré (1980) can readily be evaluated in the present context. Derivations of a variety of differential equations satisfied by relevant probabilistic functions are given in Karlin and Tavaré (1981b). Further discussion of the background of this problem appears in Part I.

1. Incomplete Detection Schemes

It is a common practice in medical genetics to screen an "at risk" population for the presence of carriers (for example, screening for carriers of Tay Sachs disease). In this section, we analyze the effects of such a screening system. Suppose that any heterozygote Aa that appears in the population can be detected with probability α, independently for different individuals. As

before, homozygous *aa*-genotypes are assumed to be visible, and so are detected as soon as they are formed. Another way to interpret the parameter α is to suppose that the a-allele has incomplete penetrance in heterozygotes, such a person being phenotypically distinguishable from AA with probability α. The reproduction process terminates with the detection of either a heterozygote or a recessive homozygote.

To model this problem, we proceed as follows. Let X_n denote the number of heterozygotes in the population of size N at time n. If $X_{n+1} = j$, the process continues as long as no heterozygote is detected (probability $(1-\alpha)^j$) and no recessive homozygotes are formed. The transition matrix of the chain $\{X_n\}$ is then given by

$$P_{ij} = \binom{N}{j} [q_i(1-\alpha)]^j p_i^{N-j}, \qquad i, j = 0, 1, \ldots, N, \tag{1}$$

where $q_i = (i/N)(1 - i/2N)$, $p_i = (1 - i/2N)^2$. Denoting the killing state by H, we have

$$P_{iH} = 1 - (p_i + q_i(1-\alpha))^N, \qquad i = 0, 1, \ldots, N. \tag{2}$$

As in Part I, we will try to approximate this chain by an associated diffusion process. It is clear that the resulting process should depend on the relative magnitude of the parameter α; we will try scalings of the form $\alpha = c(2N)^{-\gamma}$, $\gamma > 0$. Again, the time and state space scalings are of the form

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

the approximating diffusion then being given by $Y(t) = \lim_{N\to\infty} Y^{(N)}(t)$. Using the method illustrated in Part I, and some routine calculations based on (1) and (2), we can determine the relationship between γ and δ that results in a limiting diffusion process. As will be seen below, this diffusion is no longer functionally unique. In Table I, we give the mean, variance, and killing parameters, $\mu(x)$, $\sigma^2(x)$, $k(x)$, respectively, of the diffusions. The state space of $Y(t)$ is $[0, \infty)$.

It is possible to give a simple intuitive explanation of the parameters in Table I. In case (a), the magnitude of the detection probability α is too small to significantly affect the detection problem; the diffusion process is functionally identical to the case $\alpha = 0$ (cf. Karlin and Tavaré, 1980), and the process effectively terminates with detection of a recessive homozygote. In case (b), which we call *weak* detection, the detection probability is of large enough magnitude to ensure that the process ends by detection of *either* a heterozygote *or* a homozygote. In case (c), referred to as *strong detection*, the process effectively ends with detection of a heterozygote. In all three

TABLE I

Diffusion Limits for Incomplete Detection Model[a]

δ, γ	$\mu(x)$	$\sigma^2(x)$	$k(x)$
(a) $\delta = \frac{1}{3}$, $\gamma > \frac{2}{3}$	0	x	$x^2/2$
(b) $\delta = \frac{1}{3}$, $\gamma = \frac{2}{3}$ weak detection	0	x	$cx + x^2/2$
(c) $0 < \delta < \frac{1}{3}$, $\gamma = 2\delta$ strong detection	0	x	cx

[a] Time scale for diffusion is of order $(2N)^{\delta}$ generations. γ is the index of the detection probability, a.

cases, of course, the processes can terminate in loss of the a-allele from the population. The strong detection model no longer has a time scale of order $(2N)^{1/3}$, but is now of order $(2N)^{\delta}$, $0 < \delta < 1/3$. We analyze this case first.

(a) *Strong Detection*

The probability $u(x)$ of never detecting a carrier (Aa or aa) before loss of the a-allele is the solution of the equation

$$\frac{x}{2}\frac{d^2u}{dx^2} - cxu = 0; \qquad u(0) = 1, u(\infty) = 0. \tag{3}$$

Clearly,

$$u(x) = \exp(-\theta x); \qquad \theta = (2c)^{1/2}, \tag{4}$$

and so the probability of ending with detection is given by $v(x) = 1 - \exp(-\theta x)$. In order to approximate the probabilities for the underlying chain $\{X_n\}$, recall that $i \approx x(2N)^{\delta}$, $\alpha = c(2N)^{-2\delta}$, and so $(2c)^{1/2} x \approx (2\alpha)^{1/2} i$. Thus in large populations starting with i heterozygotes, the probability of detecting a carrier is approximately

$$v(i) = 1 - \exp[-(2\alpha)^{1/2} i]. \tag{5}$$

For example, if $i = 1$, $\alpha = 0.1$, then $v(1) = 0.361$.

Further analysis of this model proceeds most readily by evaluating the appropriate Green's function $G(x, y)$; formally, $G(x, y) = \int_0^{\infty} P(t, x\, y)\, dt$, where $P(t, x, y)$ is the transition density function of $Y(t)$. We obtain

$$\begin{aligned} G(x, y) &= e^{-\theta x}\frac{(e^{\theta y} - e^{-\theta y})}{\theta y}, \qquad 0 < y \leqslant x, \\ &= (e^{\theta x} - e^{-\theta x})\frac{e^{-\theta y}}{\theta y}, \qquad y \geqslant x. \end{aligned} \tag{6}$$

Using (6), the mean life time $M(x)$ starting from $Y(0)=x$ of the process is given by

$$M(x)=\frac{e^{-\theta x}}{\theta}\int_0^x \frac{e^{\theta t}-e^{-\theta t}}{t}\,dt+\frac{(e^{\theta x}-e^{-\theta x})}{\theta}\int_x^\infty \frac{e^{-\theta t}}{t}\,dt. \tag{7}$$

Recall that $\theta=(2c)^{1/2}=(2\alpha)^{1/2}(2N)^{\delta}=(2\alpha)^{1/2}/b$, where, for notational convenience, $b=(2N)^{-\delta}$. For $x=kb$ $(k=1, 2, 3,...)$, we see from (7) that

$$M(kb)=\frac{be^{-(2\alpha)^{1/2}k}}{(2\alpha)^{1/2}}\int_0^{kb}\frac{\{e^{(2\alpha)^{1/2}t/b}-e^{-(2\alpha)^{1/2}t/b}\}}{t}\,dt$$
$$+\frac{b\{e^{(2\alpha)^{1/2}k}-e^{-(2\alpha)^{1/2}k}\}}{(2\alpha)^{1/2}}\int_{kb}^\infty \frac{e^{-(2\alpha)^{1/2}t/b}}{t}\,dt.$$

The change of variable $z=(2\alpha)^{1/2}t/b$ reduces the integral above to

$$M(kb)=\frac{b}{(2\alpha)^{1/2}}\left[e^{-(2\alpha)^{1/2}k}\int_0^{(2\alpha)^{1/2}k}\frac{e^z-e^{-z}}{z}\,dz\right.$$
$$\left.+\{e^{(2\alpha)^{1/2}k}-e^{-(2\alpha)^{1/2}k}\}\int_{(2\alpha)^{1/2}k}^\infty \frac{e^{-z}}{z}\,dz\right].$$

Recalling that one unit of time in the diffusion corresponds to b^{-1} generations in the discrete model, we see that the mean time to loss or detection in the discrete process is given approximately by $b^{-1}M(kb)$, $X_0=k=1, 2,....$ For $k=1$, the mean is approximately 2.50 generations, a result that is again independent of population size. In Table II, we give the results of 5000 simulations of the discrete process for a variety of population sizes.

TABLE II

Simulation Results for Time to Detection or Loss for Discrete Model[a]

N	Mean	Variance
5	2.08 ± 0.04	2.22
10	2.15 ± 0.05	2.70
20	2.22 ± 0.05	2.87
50	2.35 ± 0.05	3.69
100	2.30 ± 0.05	3.59
500	2.33 ± 0.05	3.59

[a] $\alpha=0.1$; ± figures are approximate 95% confidence intervals. $X_0=1$.

The most interesting part of the detecting problem involves focusing attention only on those sample paths which result in detection. In what follows, *we will use the subscript D to denote properties of the process conditional on detection occurring as opposed to loss of the a-allele.* The conditional Green's functions $G_D(x, y)$ central to the analysis are given by

$$G_D(x, y) = G(x, y)\frac{v(y)}{v(x)}, \tag{8}$$

where $G(x, y)$, $v(x)$ are given explicitly by (4) and (6), respectively. The mean detection time is given by

$$M_D(x) = \frac{e^{\theta x} - e^{-\theta x}}{\theta(1 - e^{-\theta x})}\int_x^\infty \frac{e^{-\theta t}(1 - e^{-\theta t})}{t}\,dt + \frac{e^{-\theta x}}{\theta(1 - e^{-\theta x})}\int_0^x \frac{(e^{\theta t} - e^{-\theta t})(1 - e^{-\theta t})}{t}\,dt.$$

As $x \to 0$, we evaluate $M_D(0) = (2/\theta)\int_0^\infty [e^{-\theta t}(1 - e^{-\theta t})/t]\,dt = (2 \ln 2)/\theta$. It follows that in large populations starting with a very small number of heterozygotes, the mean time to detection is approximately given by $M_D = 2^{1/2}(\ln 2)/\alpha^{1/2}$ generations. For example, if $\alpha = 0.1$, $M_D = 3.10$. For comparison, we give some simulation results in Table III.

In the next paragraphs, we will mention briefly some other functionals of the process conditioned on detection which aid in our understanding of the detection process. The derivation of these quantities is based on results given in Karlin and Tavaré (1981b). The first of these involves the computation of

TABLE III

Simulation Results for Mean Time to Detection M_D Based on m Runs of Discrete Model That Ended in Detection[a]

N	m	Mean	Variance
5	1822	2.40 ± 0.07	2.58
10	1811	2.58 ± 0.08	3.12
20	1705	2.77 ± 0.09	3.65
50	1724	2.95 ± 0.10	4.56
100	1651	2.90 ± 0.11	4.80
500	1666	2.94 ± 0.10	4.70

[a] $\alpha = 0.1$, N = population size; ± figures are approximate 95% confidence intervals. $X_0 = 1$.

the average total number of heterozygotes that appear in the population before detection. Denoting this quantity by $H_D(x)$, we find that

$$H_D(x) = \int_0^\infty G_D(x, y) y \, dy,$$

while $H_D(0) = 1/\theta^2 = 1/2c$. The average number of heterozygote carriers before detection in the discrete population is then $H_D \approx 1/2a$, as long as the populaion size is large, and the initial generation comprises a very small number of heterozygotes.

As another measure of the "genetic cost" of the a-allele in the population, we can compute the distribution of the maximum number of heterozygotes that appear before detection. We have

$$\begin{aligned} &Pr(\max_{u>0} Y(u) > y \mid Y(0) = x, \text{ allele } a \text{ detected}) \\ &\quad = 1, \qquad y \leqslant x \\ &\quad = \frac{(e^{\theta x} - e^{-\theta x})(1 - e^{-\theta y})}{(e^{\theta y} - e^{-\theta y})(1 - e^{-\theta x})}, \qquad y \geqslant x. \end{aligned}$$

The mean maximum number of heterozygotes is

$$T_D(x) = x + \frac{e^{\theta x} - e^{-\theta x}}{1 - e^{-\theta x}} \int_x^\infty \frac{1 - e^{-\theta y}}{e^{\theta y} - e^{-\theta y}} \, dy,$$

with $T_D(0) = 2(\ln 2)/\theta$. This corresponds to about $2^{3/2}(\ln 2)\, a^{-1/2}$ individuals in populations starting with a very small number of heterozygotes.

One particularly descriptive function of the detection processes is the position at which detection takes place. This allows us to compute the most likely number of heterozygotes in the population when detection occurs. Defining P_D to be the place at which detection occurs, the density function $w_D(x; y)$ of P_D given $Y(0) = x$ is given by

$$\begin{aligned} w_D(x; y) &= \frac{\theta e^{-\theta x}}{1 - e^{-\theta x}} \frac{e^{\theta y} - e^{-\theta y}}{2}, \qquad y \leqslant x, \\ &= \theta \left(\frac{e^{\theta x} - e^{-\theta x}}{2} \right) \frac{e^{-\theta y}}{1 - e^{-\theta x}}, \qquad y \geqslant x \end{aligned} \tag{9}$$

(cf. Karlin and Tavaré, 1981b). Specializing to the case in which $x = 0$, we see that P_D has the exponential density

$$w_D(0; y) = \theta e^{-\theta y}, \qquad y > 0. \tag{10}$$

The shape of these densities backs up the conclusions of the previous sections, showing that detection will occur very quickly starting from any position, since the mode of the detection distribution is at x, the starting frequency.

Finally, we mention briefly some aspects of the time-dependent behavior of this model. It is possible to evaluate an explicit expression for the transition density $P(t, x, y)$ of the process $\{Y(t)\}$, and consequently for the process conditioned on detection. The rate of decline of this density is dominated by $e^{-\theta t}$, in that

$$P(t, x, y) = 4\theta^2 x e^{-\theta(x+y)} e^{-\theta t} + O(e^{-2\theta t}), \qquad t \to \infty \tag{11}$$

(Karlin and Tavaré, 1981b). The rate $e^{-\theta t}$ is that at which the time-dependent functionals decline. As a consequence of (11), we can derive the asymptotic distribution $a_D(y)$ given detection has not occurred yet of the process conditioned on ultimate detection. Denoting by ξ_D the time to detection, then

$$\begin{aligned} a_D(y)\, dy &= \lim_{t\to\infty} Pr\{Y(t) \in (y, y + dy) \mid \xi_D > t, Y(0) = x\} \\ &= 2\theta e^{-\theta y}(1 - e^{-\theta y})\, dy. \end{aligned} \tag{12}$$

We deduce from (12) that if the process has been running a long time, and detection has not yet taken place, then the average number of heterozygotes is approximately $3/2(2\alpha)^{1/2}$, with variance approximately $5/8\alpha$.

We comment before continuing with the weak detection model that in the case of strong detection, the functionals are effectively independent of population size; this result, which stands in marked contrast to the results of Part I, will be discussed further in the summary.

(b) *Weak Detection*

As shown in Table I, the infinitesimal parameters in this case are given by $\mu(x) = 0$, $\sigma^2(x) = x$, $k(x) = (x^2/2) + cx$. The probability of detection before loss is then given by

$$v(x) = 1 - \frac{A(x + 2c)}{A(2c)}, \tag{13}$$

where $A(x)$ is the Airy function of the first kind. Functionals of this process can be evaluated in the usual way using the Green's function, $G(x, y)$. These are given by

$$\begin{aligned} G(x, y) &= \frac{2\pi A(x + 2c)}{y} C(y; 2c), \qquad y \leqslant x \\ &= \frac{2\pi A(y + 2c)\, C(x; 2c)}{y}, \qquad y \geqslant x, \end{aligned} \tag{14}$$

where $C(y; 2a) = B(x + 2c) - [B(2c)/A(2c)]\, A(x + 2c)$ (compare Part I, formula (29)). The order of magnitude of the time to detection is proportional to $(2N)^{1/3}$ generations, the variance of this time being of order $(2N)^{2/3}$ generations, provided the population commences with a very small number of heterozygotes. This approximating process models the competition that arises between the two possible detection events: detection of a carrier, or detection of a "recessive" homozygote. Notice that the time scale of this process is "longer" than that of the strong detection case, where detection or loss occurs more rapidly.

2. Family Structure

In the following sections, we will try to assess the role of a variety of "family structure" models. The first of these is a testing procedure that might occur in artificial selection schemes.

(a) *Examining More Children Than Are Used as Parents*

We assume that the population comprises N breeding adults, and that at each reproduction point, M offspring are formed. We will assume that $M > N$. These offspring are examined, and if no homozygous recessives are found, N of them are chosen at random to be parents in the next generation. We would like to assess the importance of this sampling procedure on the detection problem.

A simple model is as follows:

AA	$N-i$	$M-k$	$N-j$
Aa	i	k	j
aa	0	0	0
	time	no "aa" detected	time
	n	in offspring	$n+1$

Let X_n be the number of heterozygotes at time n in the adult population. Given that no recessive homozygotes are detected in the offspring, we take a (hypergeometric) sample of size N to produce the next generation of breeding adults. Assuming that $X_n = i$, and that no homozygote recessives are produced in the offspring, there will be $M-k$ AA-individuals, k Aa-individuals with probability

$$P_{ik} = \binom{M}{k}\left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^k\left(1-\frac{i}{2N}\right)^{2(M-k)}.$$

After sampling N of these individuals, the number of heterozygotes in the next generation will be j with probability

$$P_{ij} = \sum_{k=j}^{M-N+j} \bar{P}_{ik} \binom{M-k}{N-j}\binom{k}{j}\Big/\binom{M}{N} \tag{15}$$

$$= \binom{N}{j}\left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^j\left(1-\frac{i}{2N}\right)^{2(N-j)}\left(1-\frac{i^2}{4N^2}\right)^{M-N}, \quad 0 \leqslant i, j \leqslant N,$$

while

$$P_{iH} = 1 - \left(1-\frac{i^2}{4N^2}\right)^M.$$

If we set $r = M/N \geqslant 1$, then a familiar diffusion analysis of the transition matrix (15) shows that the processes

$$Y_N(t) = \frac{X([(2N)^{1/3}\, t])}{(2N)^{1/3}}, \qquad t > 0,$$

will converge to a limiting diffusion $Y(t)$ as $N \to \infty$, with infinitesimal parameters given by

$$\mu(x) = 0, \qquad \sigma^2(x) = x, \qquad k(x) = rx^2/2. \tag{16}$$

The detection probability is then given by

$$v(x) = 1 - \frac{A(\gamma x)}{A(0)}, \qquad x \geqslant 0; \quad \gamma = r^{1/3} \geqslant 1, \tag{17}$$

which, for very small values of x, is approximated by $v(x) \sim 0.729\gamma x - 0.167\gamma^3 x^3$. We can use the result in (17) to determine how large r should be to ensure that, starting from 1 heterozygote in the initial generation, the process ends by detection (as opposed to loss) with probability at least 0.5. Setting $x = (2N)^{-1/3}$ shows that $v((2N)^{-1/3}) \geqslant 1/2$ if $A[(r/2N)^{1/3}] \leqslant A(0)/2 = 0.17751$, or $M \geqslant 0.874N^2$. Clearly, we need to examine a large number of children to alter significantly the detection probability.

Qualitatively, the conclusions that can be drawn in this sampling scheme correspond closely to the standard case $r = 1$ (Karlin and Tavaré, 1980). The time scale of events is still of order $(2N)^{1/3}$ generations. For example, in populations starting with a very small number of heterozygotes (taking $x = 0$), then conditioned on detection, the mean time to detection is approximately $2.090N^{1/3}/\gamma$ generations. Starting from the same initial position, and again conditioned on detection, we find that the mode of the distribution of the detection point (P_D before Eq. (9)) is at the point $0.885/\gamma$, corresponding

to detection being most likely to occur at a frequency of about $1.12N^{1/3}/\gamma$ heterozygotes. More formal evaluations of properties of the process are found using the Green's function

$$\begin{aligned} G(x, y) &= \frac{2\pi A(\gamma x)\, C(\gamma y; 0)}{\gamma y}, \qquad y \leqslant x, \\ &= 2\pi C(\gamma x; 0)\, \frac{A(\gamma y)}{\gamma y}, \qquad y \geqslant x. \end{aligned} \tag{18}$$

(b) *Screening Sibs*

Suppose now that the population comprises N breeding individuals, each of which produces a family of r offspring. For simplicity, we suppose that $r \geqslant 1$. We examine *all* offspring, again terminating the process if any recessive homozygotes are found. To continue to the next generation, we select a random individual from each family to become a parent in the next generation. It follows by examining the outcomes of random matings in the population comprising N adults, in which there are i heterozygotes, and $N - i$ AA-homozygotes, that

$$\begin{aligned} p_i &= Pr[\text{randomly selected individual is } AA \mid \text{process continues}] \\ &= \left(1 - \frac{i}{N}\right)^2 + \frac{1}{4}\left(\frac{i}{N}\right)^2 \left(\frac{3}{4}\right)^{r-1} + \frac{i}{N}\left(1 - \frac{i}{N}\right) \\ &= 1 - \frac{i}{N} + \frac{1}{4}\left(\frac{i}{N}\right)^2 \left(\frac{3}{4}\right)^{r-1}. \end{aligned}$$

The factor $(3/4)^{r-1}$ accounts for the fact that to continue the process, none of this individual's $r - 1$ sibs can be a recessive homozygote. In a similar way, we find that

$$\begin{aligned} q_i &= Pr[\text{randomly selected individual is } Aa \mid \text{process continues}] \\ &= \frac{i}{N}\left(1 - \frac{i}{N}\right) + \frac{1}{2}\left(\frac{i}{N}\right)^2 \left(\frac{3}{4}\right)^{r-1}. \end{aligned}$$

The transition matrix is then determined by

$$\begin{aligned} P_{ij} &= \binom{N}{j} (q_i)^j (p_i)^{N-j}, && 0 \leqslant i, \quad j \leqslant N; \\ P_{iH} &= 1 - \left(1 - \frac{i^2}{N^2}\left(1 - \left(\frac{3}{4}\right)^r\right)\right)^N, && 0 \leqslant i \leqslant N. \end{aligned} \tag{19}$$

Again, the approximating diffusion is that given by time scale and state space scale $(2N)^{1/3}$; the associated parameters are

$$\mu(x)=0, \quad \sigma^2(x)=x, \quad k(x)=Rx^2/2; \quad R=4(1-(\tfrac{3}{4})^r). \tag{20}$$

In this model, the results of (a) apply with $\gamma=R^{1/3}$. The generalization to random family sizes is immediate. If each family is of size k with probability c_k, $k=1, 2,...$, and $\Phi(s)=\sum_{k=1}^{\infty} s^k c_k$ is the probability generating function of $\{c_k\}$, then the parameter R of (20) is replaced by $R=4(1-\Phi(3/4))$. Of course, the same qualitative conclusions as earlier apply.

(c) *Mixed Sampling Scheme*

Suppose that at time n, the population comprises i heterozygotes and no recessive homozygotes. To model the structure of a hypothetical artificial selection scheme, we suppose that k recessive homozygotes are produced by the breeding population of size N. With probability β any one of these is detected *as a recessive homozygote*, while with probability $(1-\beta)$ the individual dies (perhaps from other causes), and has to be replaced. The process stops if we detect any recessive homozygotes. If not, such individuals have to be replaced (to make up the numbers in the breeding line). To achieve this, replacements are chosen from the population at large which is supposed to contain a proportion α of heterozygotes, and no recessive homozygotes. Once the population is up to the right size again, reproduction continues. The transition matrix for this process is determined by

$$\begin{aligned} P_{ij} &= \binom{N}{j} q_i^j p_i^{N-j}, \qquad 0 \leqslant i, \quad j \leqslant N; \\ P_{iH} &= 1-(1-r_i)^N, \qquad 0 \leqslant i \leqslant N, \end{aligned} \tag{21}$$

where

$$q_i=\frac{i}{N}\left(1-\frac{i}{2N}\right)+\frac{(1-\beta)\,\alpha i^2}{4N^2},$$

$$p_i=\left(1-\frac{i}{2N}\right)^2+\frac{(1-\beta)(1-\alpha)\,i^2}{4N^2}, \qquad r_i=\frac{\beta i^2}{4N^2}.$$

To approximate this process by a diffusion, we keep α fixed (since it reflects the heterozygote proportions in the population at large), but take β varying with N in such a way that $(2N)^\delta \beta \to b$ as $N \to \infty$, for some $\delta \geqslant 0$. Defining

$$Y_N(t)=\frac{X([(2N)^\epsilon t])}{(2N)^\epsilon}, \qquad N \geqslant 1, \qquad t \geqslant 0,$$

TABLE IV

Admissible Values for Diffusion Appproximation

	Mean $\mu(x)$	Variance $\sigma^2(x)$	Killing $k(x)$
(i) $\frac{1}{3} \leqslant \varepsilon < \frac{1}{2}$ $\delta = 3\varepsilon - 1$	0	x	$bx^2/2$
(ii) $\delta = \varepsilon = \frac{1}{2}$	$-x^2(1-\alpha/2)$	x	$bx^2/2$

we find the range of (ε, δ) values that result in a limiting diffusion $Y(t)$ that models the behavior of $\{X_n\}$. The results are summarized in Table IV.

Some simple qualitative conclusions can be derived from Table IV; we focus attention on case (i) only. Firstly, the functional form of the diffusion parameters is identical for *a range of time scalings* ε. The time scaling $\varepsilon = \frac{1}{3}$ (corresponding to the time scale in paragraphs (a) and (b)) corresponds to β, the detection probability, being *fixed* as a function of population size N. As ε increases from $\frac{1}{3}$ to $\frac{1}{2}$, β gets smaller (corresponding to replacements of more A-alleles in the population that is reproducing) and thus the time scale of events will be longer, as evidenced above. The analytic formulas derived in the previous sections (in particular, (17) and (18)) also apply with $\gamma = b^{1/3}$. Notice again that the time scalings are not unique, so care has to be taken in interpreting results for the discrete process.

3. Discussion

In this paper, we analyzed two essentially distinct problems involving the detection of visible genes. Motivated by recent advances in biochemical techniques that have led to greater ability to distinguish genes with apparently identical phenotypes (e.g., heterozygote Tay Sachs genotypes, Gaucher's syndrome), we have assessed the role of incomplete detection in heterozygotes. This could also be interpreted as incomplete penetrance, a heterozygote sometimes being phenotypically identical to the homozygote aa. Denoting the detection (or expressivity) probability by α, we found that the detection model is approximated by three parametrically distinct diffusion processes. These are determined by the order of magnitude of the detection probability with respect to the population size N. If α goes to zero like $c(2N)^{-\gamma}$, then there is a competition between detection of heterozygotes and homozygotes only when $\gamma = \frac{2}{3}$. In this case, the order of magnitude of the time scale is in units of $(2N)^{1/3}$ generations (compare to the selection model of Part I). For detection rates much larger than this order (that is, $0 < \gamma < \frac{2}{3}$),

the strong detection case effectively ends in detection of a heterozygote, the homozygotes playing little role in the problem. The order of magnitude of the detection time is then of the order $(2N)^{\delta}$, $0 < \delta < \frac{1}{3}$ $(\delta = \gamma/2)$. This particular case is especially tractable, and a variety of functionals describing the detection problem were analyzed in Section 1a. Indeed, for this problem a complete description of the time-dependent solutions is available, since one can compute explicitly the eigenstructure of the process. See Karlin and Tavaré (1981b) for the analysis of this problem, and some elaborations on the theoretical methods used to derive the results of this paper. In the final case of incomplete detection, the detection probability parameter γ is greater than $\frac{2}{3}$, the associated diffusion process ending essentially only by detection of a homozygote due to the small detection rate in heterozygotes. This parameter range is covered by the "random drift" results described in Karlin and Tavaré (1980), the time scale being again of order $(2N)^{1/3}$ generations.

The second class of models we investigated arose in the context of artificial selection schemes, and assessed the roles of family structure and the examination of more individuals than are used to breed. The latter situation again exhibited the stability of the $(2N)^{1/3}$ generation time scale, and the model can be fully analyzed in terms of the classical Airy functions. Starting with one heterozygote in the initial parent population, we found that we need $M \approx 0.87N^2$, meaning that we have to examine a very large number of offspring in order to insure the detection probability of allele a is at least one half. Generally the presence of family structure shows qualitatively the same behavior with adjusted detection rates in the basic model.

The final model examined involved a mixed reproduction structure in which any recessive homozygotes that were not detected are replaced by a random sample of individuals from a large population with fixed proportions of AA and Aa genotypes. In this case, a wide range of time scales was possible, depending on the probability of detection of a recessive homozygote *as a recessive homozygote*. Only if this rate is independent of population size does the time scale $(2N)^{1/3}$ govern detection. In other cases, rates up to $(2N)^{1/2}$ are possible.

References

CROW, J. F. AND KIMURA, M. 1970. "An Introduction to Population Genetics," Harper Row, New York.

EWENS, W. J. 1979. "Mathematical Population Genetics," Springer-Verlag, New York.

GRADSHTEYN, I. S. AND RYZHIK, I. M. 1965. "Tables of Integrals, Series and Products," 4th ed., Academic Press, New York.

JAMES, J. W. 1979. The time of detection of sex-linked recessives in small populations, *Genet. Res. (Camb.)* **34**, 11-17.

KARLIN S. AND TAVARÉ, S. 1980. The detection of recessive visible genes in finite populations, *Genet. Res.* (*Camb.*) **37**, 33–46.

KARLIN S., AND TAVARÉ, S. 1981a. The detection of particular genotypes in finite populations. I. Natural selection effects, *Theor. Pop. Biol.* **19**, 187–214.

KARLIN S., AND TAVARÉ, S. 1981b. Some diffusion stochastic processes with killing arising in population genetics, *J. Appl. Math.*, in press.

ROBERTSON, A. 1978. The time to detection of recessive visible genes in small populations. *Genet. Res.* (*Camb.*) **31**, 255-264.

Printed by the St. Catherine Press Ltd., Tempelhof 41, Bruges, Belgium