THE NUMBER OF MUTANT FORMS MAINTAINED IN A POPULATION

SAMUEL KARLIN and JAMES McGREGOR
STANFORD UNIVERSITY

1. Introduction

Experimental and empirical studies appear to indicate the existence of large numbers of polymorphisms in wild populations. It is desirable to formulate a model which qualitatively explains the relative importance of the pressures of selection, mutation and random drift in maintaining these polymorphisms. An equivalent problem investigated by Fisher [3] concerned the rate of mutation necessary in order to maintain a sufficient number of heterozygous loci which thus contribute toward genetic variance.

In recent years there has been much interest in the related problem of determining the number of alleles (alternative gene types) that can be maintained by mutation. It is this problem that the subsequent analysis is primarily aimed at; however, it directly adds insight into the question of clarifying the reasons for the large numbers of polymorphisms observed. Kimura and Crow [11] and Ewens [2] have investigated quantitatively the situation when each mutant form that arises is different from previous forms. Either due to selective disadvantages, mutation or migration pressures, or random sampling effects due to finite population size, the resulting subpopulation generated by each mutant form ultimately becomes extinct. The reasons for studying these models are set forth in Kimura and Crow at length and are not repeated here. For further bibliographic references on this subject we refer to Wright [14].

The model not formulated precisely but probably implicit in Kimura and Crow [11] and explicit in Ewens [2] is of the following structure. The population consists of $2N$ genes each capable of mutating in any generation with probability ν and thereby creating new allelic types. The fluctuations of the population size of a particular allelic line is assumed to be governed by the classical Fisher-Wright process. That is, if the number of representatives of the particular allele is i, then the probability that the number of the allele in question changes to j in the next generation is given by

$$P_{ij} = \binom{2N}{j}\left[\frac{i(1-\nu)}{2N}\right]^j\left[1-\frac{i(1-\nu)}{2N}\right]^{2N-j}, \qquad j = 0, 1, \cdots, 2N, \tag{1.1}$$

where $-i\nu$ is the expected decrease in i due to mutation to new alleles.

We expect on the average $2N\nu$ new allelic types to arise per generation. At

Research supported in part by Contract NIH USPHS 10452 at Stanford University.

equilibrium these new alleles must be balanced by a similar number of "old" alleles being lost because of random sampling. If at equilibrium there are $\bar{n}$ different alleles, on the average, and $\bar{t}$ is the mean number of generations that an allele exists in the population before being lost then the required balance is expressed by the relation

$$2N\nu = \frac{\bar{n}}{\bar{t}}. \tag{1.2}$$

The quantity $\bar{t}$ is evaluated in [2] by using the standard diffusion approximation to (1.1) valid only if N is large but $N\nu$ is *not* large. Ewens obtains the approximate formula

$$\bar{t} = 2\int_{(2N)^{-1}}^{1} x^{-1}(1-x)^{4N\nu-1}\,dx. \tag{1.3}$$

The analysis for the case of selectively neutral alleles as described above is partially extended with the aid of approximations derived from deterministic theory to the case of multiallelic heterotic genes. In a separate publication we shall discuss by the methods of this paper the possibility where some mutant alleles may be advantageous in the heterozygous state.

We shall approach the problem of determining the number of mutant lines maintained in a population by formulating two appropriate models. One model treats a nonstationary situation with varying total population size. In the second model we maintain a constant gene population size and study the equilibrium number of allelic types represented. The second model has several limitations which are discussed in the concluding section of this paper. The first model may be, perhaps, more pertinent in providing relevant insights into the nature of the problem of ascertaining the numbers of allelic lines maintained under the pressures of mutation, migration and selection.

In the first model (I) the creation of new allelic lines occur at random times and usually we assume each new mutant form to be deleterious so that its line of descendents will ultimately be lost. The combined population size of all alleles is random and either achieves a stochastic equilibrium or modulo random fluctuations grows to ∞. In the second model (II), the total population size is maintained at a constant level. Two variants in this model are treated. The first is a random walk on a simplicial lattice [7] and embodies the effects of birth, death and mutation. In the second model the population is transformed each generation allowing the possibility that the offspring number per parent is a random variable following a general distribution function. We will carefully formulate these models in sections 2 and 3 and describe some of the results. It will be clear that the Kimura and Crow version is a very special case of model I.

The connections of certain studies of Fisher concerning the maintenance of genetic variance by a balance of mutation and selection are briefly discussed in section 5. Section 4 presents a more detailed development of model II and some proofs and calculations of various quantities of interest. The elaboration of the structure and analysis of model I and its implications will be dealt with in a

separate publication. At that time we will duly demonstrate the flexibility and utility of model I with relevance to ecologic and genetic phenomena.

In section 6 we compare the results achieved in both cases and discuss some of their contrasting features and summarize some pertinent conclusions. Here we call attention to several important contributing factors for maintaining large numbers of alleles which seem to have not sufficiently been stressed in the previous literature of the subject.

2. The growth of mutant populations

There is a substantial literature dealing with stochastic processes describing the growth of a single population arising by mutation from a normal population. We refer to Bartlett [1], Kendall [10], and Harris ([4], chapter 5) and references therein. The studies presented below are of a different nature and concern the continued formation of new mutant populations. The principal variable of interest is the number of these populations which exist at any given time. Motivated by both ecologic and genetic phenomena, it is appropriate to set up the model in a general framework.

The model of this section proposed to describe the fluctuation of the number of mutant lines over time is composed from two processes:

(1) the stochastic process of formation of new allele (mutant) populations;

(2) the stochastic process which underlies the growth pattern of a particular mutant population.

We assume that new mutant lines arise over time according to a general stochastic process called the input process. (The origin of new lines may be ascribed to either migration or mutation forces.) More precisely, let $I(t)$ be the number of mutant lines coming into existence during the time interval $[0, t]$. The following three examples of $I(t)$ are of special interest.

(i) The number of mutants $I(t)$ is a Poisson process with parameter ν or, more generally, a variable time (inhomogeneous) Poisson process of intensity parameter $\nu(t)$. In this case, the probability of a new mutant line coming into existence during the time interval $(t, t + h)$ is $\nu(t)h + o(h)$ while the probability of no line being created is $1 - \nu(t)h + o(h)$. Moreover, the number of mutant lines formed during disjoint time intervals are independent random variables. The dependence of $\nu(t)$ on the time variable reflects the possibility of changing environmental conditions.

(ii) The number of mutants $I(t)$ is a renewal process, that is, the times between the successive starts of new mutant lines are independent positive random variables with common distribution function $G(\xi)$, $0 < \xi < \infty$. Thus, the times of the creation of new mutant lines are taken to occur at

$$\xi_1, \xi_1 + \xi_2, \xi_1 + \xi_2 + \xi_3, \cdots, \xi_1 + \cdots + \xi_n, \cdots, \tag{2.1}$$

where $\xi_1, \xi_2, \xi_3, \cdots$ are independent observations from the distribution law $G(\xi)$. In the case that $G(\xi)$ is a degenerate distribution corresponding to the value one then a new mutant line comes into existence each unit of time.

(iii) The number of mutants $I(t)$ is a general discrete time increasing point process of which a special example would be a Yule process, that is, in the latter case the times of the creation of new mutant lines coincide with the times of events of a Yule process.

The above formulation postulates that at each event of $I(t)$ a single new mutant line is formed. This could obviously be generalized such that the number of mutant lines coming into existence at a given moment may be more than one, in fact, possibly a random number. In this case the input process of example (ii) becomes a compound renewal process.

The example treated by Ewens [2] and Kimura and Crow [11] is the special case of (ii) where $G(\xi)$ is a degenerate distribution and, on an average, $2N\nu$ new mutant lines are formed each generation (one unit of time). The actual number is a binomial random variable with parameters $(2N, \nu)$.

The input process concept provides the mechanism underlying the generation of new mutant lines. The nature of the life of a specific mutant population is the other facet of the growth process. The simplest assumption is that the fluctuation of a specific mutant population follows the laws of a continuous time Markov chain $\mathcal{P}$ with transition probability function

$$P_{ij}(t), \qquad i, j = 0, 1, 2, \cdots . \tag{2.2}$$

Thus, if at time 0 a given mutant type is represented i times, then at time t its size is j with probability $P_{ij}(t)$. The 0 state is assumed to be absorbing which means that the particular mutant form in question becomes extinct once state 0 is entered. We usually assume that absorption into state 0 is a certain event.

Examples of $\mathcal{P}$ are as follows.

(a) Linear growth birth and death processes with individual infinitesimal birth and death rates λ and μ, respectively. In this case we usually require $\mu \geqq \lambda$ so that extinction occurs with certainty. Then the expected extinction time is finite if and only if $\mu > \lambda$. More generally, $\mathcal{P}$ could be any continuous time birth and death process where state 0 is a certain absorbing state.

(b) Continuous time branching processes whose infinitesimal expected mean change of population size corresponding to an initial single parent is nonpositive.

If time is discrete, then the specific Markov chain with transition probability matrix (1.1) could be chosen for (2.2).

We have described the two phases underlying the general mutation growth models. Some questions of interest to be investigated are of the following sort.

(1) Specifying an input process $\{I(t), t > 0\}$ and the individual growing process $\mathcal{P}$ for each mutant type, we should wish to determine the distribution function of the number N_t^* of different mutant lines existing at time t. More generally, what is the nature of the whole process $N_t^*, t > 0$? Letting $t \to \infty$ ordinarily leads to consideration of the random variable N_t^* under the condition that an equilibrium or stationary situation has been achieved.

(2) Motivated by certain genetic considerations we are also interested in determining the distribution function of the random variable $N_t^*(k)$, the number

of mutant populations at time t consisting of exactly k members, that is, $N_t^*(k)$ equals the number of mutant populations in state k at time t.

(3) We may also be interested in the random time required for all current mutant populations to disappear.

In the case where the input process is Poisson as in example (i), the solution of problem 2 is very simple and essentially classical. We have the following theorem.

THEOREM 2.1. *Let the input process be nonhomogeneous Poisson with intensity parameter $\nu(t)$. Let $\mathcal{P}$ be a continuous time Markov chain with transition probability matrix $\|P_{ij}(t)\|$. Let $N_t^*(k)$ denote the number of mutant populations existing at time t of size k. Then the random processes $N_t^*(k)$, $k = 1, 2, \cdots$ are independent nonhomogeneous Poisson processes with joint generating function*

$$\phi(t, z_1, z_2, \cdots) = \sum_{\bar{r}} z_1^{r_1} z_2^{r_2} \cdots \Pr\{N_t^*(i) = r_i, i = 1, 2, \cdots\} \tag{2.3}$$

$$= \exp\left[\sum_{k=1}^{\infty} (z_k - 1) \int_0^t P_{1k}(t - \tau)\nu(\tau)\, d\tau\right].$$

Subject to the hypothesis of the above theorem, it follows that

$$E[N_t^*(k)] = \int_0^t P_{1k}(t - \tau)\nu(\tau)\, d\tau. \tag{2.4}$$

If $\nu(t) = \nu$ is constant, the equilibrium distribution of $N_t^*(k)$, "$N^*(k) = \lim_{t\to\infty} N_t^*(k)$," (convergence in law) is Poisson with parameter

$$E[N^*(k)] = \nu \int_0^{\infty} P_{1k}(\tau)\, d\tau, \tag{2.5}$$

provided the integral exists which is certainly the case if 0 is an absorbing state. The expected number of alleles N^* in the stationary case is obviously

$$E(N^*) = \nu \int_0^{\infty} \sum_{k=1}^{\infty} P_{1k}(\tau)\, d\tau = \nu \int_0^{\infty} [1 - F(\tau)]\, d\tau, \tag{2.6}$$

where $F(\tau) = 1 - \sum_{k=1}^{\infty} P_{1k}(\tau)$ is the distribution function of the extinction time for a particular mutant line generated by a single initial parent. Notice that $E(N^*)$ is finite if and only if the distribution function $F(t)$ has finite mean. It is possible for $E[N^*(k)]$ to be finite while $E(N^*)$ is infinite. This situation is of interest and we will discuss some aspects of this phenomenon later.

The formula (2.6) is also valid when the input process is a renewal process, provided we interpreted ν as the reciprocal of the mean interarrival time. Higher moments of $N_t^*(k)$ can be obtained easily by iterating certain recursion relations. These calculations are also accessible when the input process is a renewal process or generalized Poisson process.

The mathematics needed to establish the result of theorem 2.1 has mostly been developed in the context of classical stochastic population growth models; see Bartlett [1] and Kendall [10]. We need merely adapt their analysis with minor modifications.

We next discuss briefly the situation where the input process is Poisson with

parameter ν and the growing process $\mathcal{P}$ is such that each mutant type ultimately becomes lost with probability 1 but the expected time until extinction is ∞. More specifically, we shall assume that the probability distribution $F(t)$ of the time until extinction (starting with a single initial parent) has the asymptotic growth behavior

$$1 - F(t) = \sum_{k=1}^{\infty} P_{1k}(t) \sim \frac{C}{t^{\alpha}}, \tag{2.7}$$

as $t \to \infty$, where $0 < \alpha \leqq 1$, and C is a constant.

For example, if $\mathcal{P}$ is a classical birth and death process with infinitesimal parameters $\lambda_n = n\lambda$, $\mu_n = n\lambda$, $n > 0$, then it is known [5] that

$$1 - F(t) \sim \frac{1}{\lambda t} \tag{2.8}$$

as $t \to \infty$. In this case we can easily determine the limiting growth behavior of $N_t^*(k)$ and $N_t^* = \sum_{k=1}^{\infty} N_t^*(k)$ as $t \to \infty$. Explicitly, we obtain that $N_t^*(k)$ is Poisson distributed with mean

$$\int_0^t \frac{(\lambda\tau)^{k-1}}{(1 + \lambda\tau)^{k+1}}\, d\tau = \lambda \int_0^{\lambda t/(1+\lambda t)} u^{k-1}\, du = \frac{\lambda}{k}\left(\frac{\lambda t}{1 + \lambda t}\right)^k. \tag{2.9}$$

Hence, as $t \to \infty$,

$$\lim_{t\to\infty} E[N_t^*(k)] = \frac{\lambda}{k}. \tag{2.10}$$

Thus, the number of populations with k members is of mean size λ/k.

What is more striking is the fact that the total number of existing mutant types N_t has an asymptotic normal distribution of mean $\log(1 + \lambda t)$ and variance $\log(1 + \lambda t)$. A more general result is as follows.

THEOREM 2.2. *If the input process is Poisson with parameter ν and $\mathcal{P}$ is a continuous time Markov chain such that the distribution function $F(t)$ of the time until extinction of a newly created mutant satisfies*

$$1 - F(t) \sim \frac{C}{t^{\alpha}} \tag{2.11}$$

for $t \to \infty$ (C is a constant) where $0 < \alpha \leqq 1$, then

$$N_t^* = \sum_{k=1}^{\infty} N_t^*(k) \tag{2.12}$$

has an asymptotic Normal distribution (as $t \to \infty$) with mean and variance

$$E(N_t^*) \sim \frac{C\nu t^{1-\alpha}}{1 - \alpha}, \qquad \mathrm{Var}\,(N_t^*) \sim \frac{C\nu t^{1-\alpha}}{1 - \alpha}, \tag{2.13}$$

when $0 < \alpha < 1$ and

$$E(N_t) \sim C\nu \log t, \qquad \mathrm{Var}\,(N_t) \sim C\nu \log t \tag{2.14}$$

when $\alpha = 1$.

PROOF. By theorem 2.1 we see that N_t^* is distributed as a Poisson variable with mean and variance given by

$$(2.15)\quad \nu \sum_{k=1}^{\infty} \int_0^t P_{1k}(\tau)\, d\tau = \nu \int_0^t [1 - F(\tau)]\, d\tau \sim \begin{cases} \dfrac{C\nu t^{1-\alpha}}{1-\alpha}, & 0 < \alpha < 1, \\ C\nu \log t, & \alpha = 1. \end{cases}$$

The asymptotic normality follows immediately since the mean parameter tends to ∞ as $t \to \infty$.

It is of interest to record some conditions which guarantee the validity of (2.11). Consider $\mathcal{P}$ to be a birth and death process with infinitesimal birth and death parameters $\lambda_n > 0$, $\mu_n > 0$, for $n = 1, 2, \cdots$, $\lambda_0 = 0$ and define

$$(2.16)\quad \pi_n = \frac{\lambda_1 \lambda_2 \cdots \lambda_n}{\mu_2 \mu_3 \cdots \mu_{n+1}}.$$

Suppose $\pi_n \sim Dn^{\gamma-1}$ and $1/\lambda_n \pi_n \sim En^{\beta-1}$ for γ and β positive where D and E are constants. It can be shown that $1 - F(t) \sim C/t^{\alpha}$ where $\alpha = \beta/(\gamma + \beta)$. In the special case of the linear growth birth and death process where $\lambda_n = n\lambda$ and $\mu_n = (n - 1 + \alpha)\lambda$, for $n = 1, 2, \cdots$, $0 < \alpha < 1$, then $\pi_n \sim D/n^{\alpha}$ and $1/\lambda_n \pi_n \sim En^{\alpha-1}$. Therefore $1 - F(t) \sim Ct^{-\alpha}$. It is frequently useful to replace (2.11) by the asymptotic relation $1 - F(t) \sim (C/t^{\alpha})L(t)$, for $0 < \alpha < 1$, where $L(t)$ is a slowly varying function. In this case we have

$$(2.17)\quad E(N_t^*) = \operatorname{Var}(N_t^*) \sim \frac{Ct^{1-\alpha}}{1-\alpha} L(t)$$

as $t \to \infty$ for $0 < \alpha < 1$.

It is worth interpreting and contrasting the nature of the results of theorems 2.1 and 2.2. Under the conditions of theorem 2.1, if $E(N^*) < \infty$, the number of different mutant lines maintained in the population achieves, as $t \to \infty$, a stable state (in the stochastic sense) which is Poisson distributed. The limiting total population size is random following a stationary distribution. On the other hand, subject to the hypothesis of theorem 2.2 the expected number of alleles grows to infinity owing to the fact that the average lifetime of each allelic type is infinite even though each individual allele is ultimately lost with probability one. This asserts, in particular, that even if the rate of formation of new alleles, that is, the mutation rate, is exceedingly small the number of existing alleles following an elapse of a sufficient duration of time is large and becomes infinite unless deterrents imply the nonapplicability of the model. In other words, *the property that each specific mutant population possesses a long lifetime may be a significant factor to account for the large number of alleles observed in nature.* A further discussion of the significance of this concept is given in section 6.

3. Model of mutant growth for a population of constant size

In the model described in section 2 the combined population size of all genes was not fixed although a stationary distribution exists if the mean absorption

time to state 0 for the process $\mathcal{P}$ is finite. We now formulate a model describing growth of different mutant types where the total population size of all genes is kept constant.

Assume tentatively that there are r different mutant types that can exist in a population consisting of N individuals where r is much larger than N. We can describe the make up of the population by an r-tuple

$$\overline{m} = (m_1, m_2, \cdots, m_r), \tag{3.1}$$

where m_i is the number of individuals of mutant type i. Thus, the m_i are non-negative integers such that $\sum_{i=1}^{r} m_i = N$.

We shall assume that the fluctuation of $\overline{m}$ is subject to the laws of some specified stochastic process $\mathcal{P}$. A natural case of this in the spirit of the models proposed by Moran [12] to study fluctuations of gene frequency is the following.

In each unit of time a single individual may change his type. (We could equally well postulate that the events of changes of state occur continuously in time at the events of a Poisson process with intensity parameter a function of population size; see [7] for a more detailed exposition of this formulation.) Thus, the following transitions are possible,

$$\overline{m} = (m_1, m_2, \cdots, m_r) \to (m_1, m_2, \cdots, m_i - 1, \cdots, m_j + 1, \cdots, m_r) \tag{3.2}$$

for all possibilities of i and j, $i \neq j$. We also permit the transition of $\overline{m}$ into itself, that is, no change of state. A transition is determined according to the following mechanism. First an individual is chosen at random to die where all possibilities are equally likely. Thus, an individual of type i is chosen to be replaced with probability m_i/N where m_i represents the current size of the ith type. If different mortality rates operate among the various types specified by the vector of relative survival rates $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_r)$, then in this case the probability that the ith type individual is to be replaced would be

$$\frac{m_i \lambda_i}{\sum_{k=1}^{r} \lambda_k m_k}. \tag{3.3}$$

The type of the new individual replacing the dead one is determined by selecting an allele at random from the original population and duplicating its kind. The probability of thereby creating a new jth type individual is m_j/N. We could treat fecundity differences by appropriately multiplying the m_j by suitable factors reflecting the relative fertilities similar to the manner of treating differing viabilities as described above. To ease the discussion, we shall assume, in this paper, no selection differences. However, we do permit the possibility of mutation. The individual just born may change into a different type, subject to the following probabilistic laws. Let β denote the probability that an individual of a given type changes into a specified different type. We postulate a symmetrical situation in that the probability of mutation is independent of the types involved. Then the probability of a chosen individual not changing his form is $1 - (r - 1)\beta$. Combining the above effects we calculate the probability of the transition

(3.2) as follows. The event described in (3.2) occurs if the chosen individual to be replaced is of the ith type (probability m_i/N), and the newly created individual is of type j. The last possibility happens either if a jth type individual is born who does not mutate or a kth type is born who mutates into a jth type. The probability of this event is

$$\frac{m_j}{N}[1 - (r-1)\beta] + \frac{(N - m_j)}{N}\beta. \tag{3.4}$$

Thus, the transition (3.2) has probability

$$\frac{m_i}{N}\left\{\frac{m_j}{N}[1 - (r-1)\beta] + \left(\frac{N - m_j}{N}\right)\beta\right\} = \frac{1 - r\beta}{N^2} m_i\left(m_j + \frac{N\beta}{1 - r\beta}\right). \tag{3.5}$$

The probability of a transition of $\overline{m}$ into itself is

$$1 - \left(\frac{1 - r\beta}{N^2}\right)\sum_{i \neq j} m_i\left(m_j + \frac{N\beta}{1 - r\beta}\right). \tag{3.6}$$

Assuming an equilibrium state is attained (that is, assuming the process in existence for a long time), we are interested in determining the distribution of the number N^* of alleles represented in the population. More generally, under the same conditions of equilibrium, we shall evaluate the expected number of alleles represented k times in the population. Thus, let $N^*(k)$ denote the number of components in $\overline{m}$ for which $m_i = k$. We find that as $r \to \infty$

$$E[N^*(k)] = \frac{1}{k}\frac{N\nu}{1 - \nu}\frac{\binom{\frac{N}{1-\nu} - (k+1)}{N-k}}{\binom{\frac{N}{1-\nu} - 1}{N}}, \tag{3.7}$$

where $\nu = r\beta$ is the probability of a mutation occurring per unit time. Similarly, we obtain under conditions of equilibrium

$$E(N^*) = \frac{N\nu}{1 - \nu}\left\{\frac{\Gamma'\left(\frac{N}{1-\nu}\right)}{\Gamma\left(\frac{N}{1-\nu}\right)} - \frac{\Gamma'\left(\frac{N\nu}{1-\nu}\right)}{\Gamma\left(\frac{N\nu}{1-\nu}\right)}\right\} \tag{3.8}$$

as $r \to \infty$, where Γ stands for the usual Gamma function and Γ' its derivative.

The assumption of $r \to \infty$ is reasonable and merely reflects the phenomenon that the alternative number of mutant types possible in the population is extremely large compared to actual population size; see Kimura and Crow [11] for justification of this assumption.

The validations of formulas (3.7) and (3.8) are given in section 4.

In order to underscore the growth behavior of $E(N^*)$ qualitatively it is appropriate to consider three principal cases. We summarize the results in tabular form. The values are asymptotically correct as indicated. The proofs of the

results asserted in table I are presented in section 4 and some discussion and interpretation of these formulas are covered in section 6.

TABLE I

GROWTH BEHAVIOR OF $E(N^*)$

Case	Orders of Magnitude of the Parameters	$E(N^*)$	Var N^*	Asymptotic Distribution of N^* when N is Large
1	$N\nu \to \infty$	$\sim \frac{N\nu}{1-\nu} \log \frac{1}{\nu}$	$\sim \frac{N\nu}{1-\nu} \log \frac{1}{\nu} - N\nu$	Probably Normal
2	$N\nu \to \lambda$, $N \to \infty$ $0 < \lambda < \infty$ λ not too large	$\sim \left[\lambda \log N - \lambda \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right]$	$\sim \lambda \log N$	Normal
3	$N\nu \log N \to C$ $0 \leqq C < \infty$	$1 + C$	C	

4. Mutation balance (continuation of model II)

We shall first compute the expected value of N^*. Since all types behave symmetrically, we find immediately that

$$E(N^*) = r \Pr\{m_1 \neq 0\}. \tag{4.1}$$

To evaluate $\Pr\{m_1 \neq 0\}$, we can lump all types (excluding type 1) as a single type and regard the process as that of a two type model with possible states $(m, N-m)$, $m = 0, 1, 2, \cdots, N$, where the first component denotes the number of type 1 individuals present. The possible transitions that may occur take the form:

$(m, N-m) \to (m+1, N-m-1)$ with probability

$$p_1 = \left(\frac{N-m}{N}\right)\left\{\left(\frac{N-m}{N}\right)\beta + \frac{m}{N}[1-(r-1)\beta]\right\};$$

$(m, N-m) \to (m-1, N-m+1)$ with probability

$$p_2 = \frac{m}{N}\left[\frac{m}{N}(r-1)\beta + \frac{N-m}{N}(1-\beta)\right],$$

$(m, N-m) \to (m, N-m)$ with probability $1 - p_1 - p_2$.

This process is a random walk on the state space $0, 1, \cdots, N$ and its stationary distribution $\{\pi_m\}_{m=0}^{\infty}$ is readily determined [6]. Explicitly the equilibrium probability of being in state m is

$$\pi_m = \frac{\binom{m+\alpha}{m}\binom{N-m+(r-1)(\alpha+1)-1}{N-m}}{\binom{N+r(\alpha+1)-1}{N}}, \qquad m = 0, 1, \cdots, N, \tag{4.2}$$

where

$$\alpha + 1 = \frac{N\beta}{1 - r\beta} = \frac{1}{r}\frac{N\nu}{1 - \nu} \tag{4.3}$$

since $r\beta = \nu$. Thus,

$$E(N^*) = r\left[1 - \frac{\binom{N + \frac{r-1}{r}\frac{N\nu}{1-\nu} - 1}{N}}{\binom{N + \frac{N\nu}{1-\nu} - 1}{N}}\right]. \tag{4.4}$$

We introduce the function

$$h(x) = \frac{\Gamma\left(\frac{N}{1-\nu} + x\right)}{\Gamma\left(\frac{N\nu}{1-\nu} + x\right)}. \tag{4.5}$$

Now expressing (4.4) in terms of Gamma functions we have

$$E(N^*) = r\left[1 - \frac{h\left(-\frac{1}{r}\frac{N\nu}{1-\nu}\right)}{h(0)}\right]. \tag{4.6}$$

So as $r \to \infty$ (that is, the universe of alternative allelic types is extremely large), we get

$$E(N^*) = \frac{N\nu}{1-\nu}\frac{h'(0)}{h(0)} = \frac{N\nu}{1-\nu}\left\{\frac{\Gamma'\left(\frac{N}{1-\nu}\right)}{\Gamma\left(\frac{N}{1-\nu}\right)} - \frac{\Gamma'\left(\frac{N\nu}{1-\nu}\right)}{\Gamma\left(\frac{N\nu}{1-\nu}\right)}\right\}. \tag{4.7}$$

The growth behavior of $E(N^*)$ splits into three main cases.

Case 1. $N\nu \to \infty$. We use the asymptotic formula $\Gamma'(z)/\Gamma(z) \sim \log z$, $z \to \infty$ in (4.7) which yields

$$E(N^*) \sim \frac{N\nu}{1-\nu}\log\frac{1}{\nu}. \tag{4.8}$$

Case 2. $N\nu \to \lambda$, $0 < \lambda < \infty$, with λ fixed, N large. For $\lambda > 0$, equation (4.7) becomes

$$E(N^*) \sim \lambda \log N - \lambda\frac{\Gamma'(\lambda)}{\Gamma(\lambda)}. \tag{4.9}$$

Case 3. $N\nu \to 0$, $N \to \infty$. Since $\Gamma'(z)/\Gamma(z)$ has a simple pole at $z = 0$ with residue -1, we find that

$$E(N^*) \sim 1 + N\nu \log N. \tag{4.10}$$

It is interesting to note that in the case that ν is of order $1/N \log N$, then the expected number of different mutant types is bounded. This means that few types exist and those present are represented in large numbers.

It is important to divide case 3 into two further subcases according as

Case 3a: $N\nu \log N \to c > 0$, $N \to \infty$;
Case 3b: $N\nu \log N \to 0$, $N \to \infty$.

We shall later calculate the variance and higher moments of N^*.

By using the same procedure we shall now compute $E[N^*(k)]$ (the expected number of types consisting of k members) under the conditions of equilibrium for $r \to \infty$. Clearly (see (4.1)), we have

$$E[N_r^*(k)] = r \Pr\{m_1 = k\}. \tag{4.11}$$

Consulting (4.2), we obtain

$$E[N_r^*(k)] = r \frac{\binom{\alpha + k}{k}\binom{N - k + \frac{r-1}{r}\frac{N\nu}{1-\nu} - 1}{N - k}}{\binom{N + \frac{N\nu}{1-\nu} - 1}{N}}, \tag{4.12}$$

where $\alpha + 1 = (1/r)(N\nu)/(1 - \nu)$. Letting $r \to \infty$ gives

$$E[N^*(k)] = \frac{1}{k}\frac{N\nu}{1-\nu}\frac{\binom{\frac{N}{1-\nu} - (k+1)}{N - k}}{\binom{\frac{N}{1-\nu} - 1}{N}} \qquad k = 1, 2, 3, \cdots. \tag{4.13}$$

The quantity (4.13) can be studied for the three cases $N\nu$ large, moderate, or tending to zero. We shall not enter into this investigation here.

Every mutant type is represented in some population, and therefore,

$$\sum_{k=1}^{N} k\, E[N^*(k)] = N, \tag{4.14}$$

or equivalently,

$$\frac{N\nu}{1-\nu}\sum_{k=1}^{N}\frac{\binom{\frac{N}{1-\nu} - (k+1)}{N - k}}{\binom{\frac{N}{1-\nu} - 1}{N}} = N. \tag{4.15}$$

The expected number of individuals belonging to a type represented k times is obviously

$$kE[N^*(k)] = \frac{N\nu}{1-\nu} \frac{\binom{\frac{N}{1-\nu} - (k+1)}{N-k}}{\binom{\frac{N}{1-\nu} - 1}{N}}. \tag{4.16}$$

The chance, on the average, of selecting at random an individual belonging to an allele with k representatives is then

$$\frac{1}{N}\left\{\frac{N\nu}{1-\nu} \frac{\binom{\frac{N}{1-\nu} - (k+1)}{N-k}}{\binom{\frac{N}{1-\nu} - 1}{N}}\right\}. \tag{4.17}$$

We next compute the probability of homozygosity F, that is, the probability that two genes (individuals) chosen at random are of the same type. The chance that the first is from a type represented k times is

$$\frac{\nu}{1-\nu} \frac{\binom{\frac{N}{1-\nu} - k - 1}{N-k}}{\binom{\frac{N}{1-\nu} - 1}{N}}, \tag{4.18}$$

and for this contingency the chance of homozygosity is k/N. By the law of total probabilities, we get

$$\Pr\{\text{homozygosity}\} = F = \sum_{k=1}^{N} \frac{k}{N} \frac{\nu}{1-\nu} \frac{\binom{\frac{N}{1-\nu} - k - 1}{N-k}}{\binom{\frac{N}{1-\nu} - 1}{N}}. \tag{4.19}$$

A simple calculation yields

$$F = \sum_{k=1}^{N} \frac{k}{N} \frac{\nu}{1-\nu} \frac{\binom{\frac{\nu N}{1-\nu} + N - k - 1}{N-k}}{\binom{\frac{N}{1-\nu} - 1}{N}} \tag{4.20}$$

$$= \left(\frac{\nu}{1-\nu}\right) \frac{1}{N} \frac{\binom{\frac{N}{1-\nu}}{N-1}}{\binom{\frac{N}{1-\nu} - 1}{N}} = \frac{1}{N\nu + 1 - \nu}.$$

Observe that for $\nu \to 0$ this is of the order

$$F = \frac{1}{N\nu + 1}, \tag{4.21}$$

which agrees with the classical formula (one must replace N by $2N$ which is the usual gene population size associated with N diploid individuals and then another factor of 2 enters arising from the fact that only one individual is altered at a time).

We now return to the study of

$$E[N^*(k)] = \frac{1}{k} \frac{N\nu}{1-\nu} \frac{\binom{\frac{N}{1-\nu} - k - 1}{N-k}}{\binom{\frac{N}{1-\nu} - 1}{N}} \tag{4.22}$$

(see (4.13)), for k and N large. Specifically, we shall consider $k/N \sim x$, $k = [Nx]$, x fixed, and let $\nu N = \lambda$, $0 < \lambda < \infty$, be fixed. Exploiting the asymptotic properties of the Gamma function, we obtain

$$E[N^*(k)] \sim \frac{1}{N} \frac{\lambda}{x} \frac{\Gamma[N(1-x) + \lambda]}{\Gamma[N(1-x) + 1]} \frac{\Gamma(N+1)}{\Gamma(N+\lambda)} \tag{4.23}$$

$$\sim \frac{1}{N} \frac{\lambda}{x} (1-x)^{\lambda-1} = \frac{\lambda}{x} (1-x)^{\lambda-1}\, dx$$

since $dx \sim 1/N$. Thus, the expected number of types whose frequency in the population occurs between x_1 and x_2 is approximately

$$\lambda \int_{x_1}^{x_2} \frac{1}{x} (1-x)^{\lambda-1}\, dx, \tag{4.24}$$

provided $\lambda = N\nu$ is fixed and finite and N is large.

Notice that for $x_1 = 1/N$ and $x_2 = 1$, the expression is approximately the expected number of distinct types in the population and this is of the order

$$\lambda \int_{1/N}^{1} \frac{1}{x}(1-x)^{\lambda-1}\,dx \sim N\nu \log N, \tag{4.25}$$

which agrees with (4.9) for $N\nu = \lambda$ finite. We emphasize again that this relation is valid provided $N\nu$ is not large.

Variance of N^*. Our next task is to compute the variance of N^*. Consider the equilibrium population involving N genes (individuals). We define

$$Z_i = \begin{cases} 1 & \text{if the } i\text{th type is present,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.26}$$

Clearly,

$$Z_1 + Z_2 + \cdots + Z_r = N_r^*. \tag{4.27}$$

We have calculated the mean of N_r^* in equation (4.7). We will now evaluate $E[(N_r^*)^2]$. Thus,

$$E[(N_r^*)^2] = rE(Z_1^2) + r(r-1)E(Z_1Z_2). \tag{4.28}$$

Now (see (4.5) and (4.6)),

$$E(Z_1^2) = E(Z_1) = 1 - \frac{h\left(-\frac{1}{r}\frac{N\nu}{1-\nu}\right)}{h(0)}. \tag{4.29}$$

By virtue of the symmetry of the problem, we have

$$\begin{aligned} E(Z_1Z_2) &= 1 - \Pr\{\text{either } m_1 = 0 \text{ or } m_2 = 0\} \\ &= 1 - 2\Pr\{m_1 = 0\} + \Pr\{m_1 = m_2 = 0\}. \end{aligned} \tag{4.30}$$

In order to calculate $\Pr\{m_1 = m_2 = 0\}$, we can consider the first two types identified and the other types all lumped together. This leads to a Moran type mutation model (see Karlin and McGregor [6]) with mutation parameters $\alpha_1 = (r-2)\beta$, $\alpha_2 = 2\beta$. Using the form of the stationary distribution analogous to (4.2), we find that

$$\Pr\{m_1 = m_2 = 0\} = \frac{\binom{N + \frac{r-2}{r}\frac{N\nu}{1-\nu} - 1}{N}}{\binom{N + \frac{N\nu}{1-\nu} - 1}{N}} = \frac{h\left(-\frac{2}{r}\frac{N\nu}{1-\nu}\right)}{h(0)}. \tag{4.31}$$

Thus, (4.30) becomes

$$E(Z_1Z_2) = \frac{1}{h(0)}\left[h(0) - 2h\left(-\frac{1}{r}\frac{N\nu}{1-\nu}\right) + h\left(-\frac{2}{r}\frac{N\nu}{1-\nu}\right)\right]. \tag{4.32}$$

As $r \to \infty$, we get

$$\lim_{r\to\infty} r(r-1)\,E(Z_1Z_2) = \left(\frac{N\nu}{1-\nu}\right)^2 \frac{h''(0)}{h(0)}. \tag{4.33}$$

Combining, we see that the second moment of N^* is

$$E[(N^*)^2] = \left(\frac{N\nu}{1-\nu}\right)^2 \frac{h''(0)}{h(0)} + \frac{N\nu}{1-\nu}\frac{h'(0)}{h(0)}, \tag{4.34}$$

and generally,

$$E[(N^*)^k] = \left(\frac{N\nu}{1-\nu}\right)^k \frac{h^{(k)}(0)}{h(0)} + \sum_{\ell=1}^{k-1} \left(\frac{N\nu}{1-\nu}\right)^\ell c_\ell \frac{h^{(\ell)}(0)}{h(0)}. \tag{4.35}$$

Now as $N \to \infty$ and $N\nu \to \infty$ it is not difficult to see that

$$\frac{h^{(\ell)}(0)}{h(0)} \sim \left(\log\frac{1}{\nu}\right)^\ell + o(1) \tag{4.36}$$

where the $o(1)$ term goes to zero provided $N\nu \to \infty$. It follows that

$$E\left[\frac{N^*}{\left(\frac{N\nu}{1-\nu}\right)\log\frac{1}{\nu}}\right]^k \to 1, \tag{4.37}$$

as $N \to \infty$, for every $k = 1, 2, \cdots$. This suggests the result

$$\frac{N^*}{\frac{N\nu}{1-\nu}\log\frac{1}{\nu}} \to 1 \tag{4.38}$$

with probability 1, which is certainly valid *in probability* owing to a standard theorem.

The variance

$$\mathrm{Var}\,(N^*) = \left(\frac{N\nu}{1-\nu}\right)^2 \left\{\frac{h''(0)}{h(0)} - \left[\frac{h'(0)}{h(0)}\right]^2\right\} + \frac{N\nu}{1-\nu}\frac{h'(0)}{h(0)} \tag{4.39}$$

is asymptotically δN as $N \to \infty$, where

$$\delta = \frac{\nu}{1-\nu}\left[\log\frac{1}{\nu} - (1-\nu)\right], \tag{4.40}$$

which ensues as a consequence of the identity

$$\begin{aligned}\frac{h''(0)}{h(0)} - \left[\frac{h'(0)}{h(0)}\right]^2 &= \left\{\frac{\Gamma''\left(\frac{N}{1-\nu}\right)}{\Gamma\left(\frac{N}{1-\nu}\right)} - \left[\frac{\Gamma'\left(\frac{N}{1-\nu}\right)}{\Gamma\left(\frac{N}{1-\nu}\right)}\right]^2\right\} \\ &\quad - \left\{\frac{\Gamma''\left(\frac{N\nu}{1-\nu}\right)}{\Gamma\left(\frac{N\nu}{1-\nu}\right)} - \left[\frac{\Gamma'\left(\frac{N\nu}{1-\nu}\right)}{\Gamma\left(\frac{N\nu}{1-\nu}\right)}\right]^2\right\} \\ &\sim -\frac{(1-\nu)^2}{\nu}\frac{1}{N},\end{aligned} \tag{4.41}$$

since

$$\frac{\Gamma''(z)}{\Gamma(z)} - \left[\frac{\Gamma'(z)}{\Gamma(z)}\right]^2 \sim \frac{1}{z}, \tag{4.42}$$

as $z \to \infty$. Therefore,

$$\text{Var } N^* \sim \frac{N\nu}{1-\nu}\left[\log\frac{1}{\nu} - (1-\nu)\right]. \tag{4.43}$$

(Of course, $\log(1/\nu) - (1-\nu) > 0$ for $0 < \nu < 1$ as it should be.) This suggests the validity of a central limit theorem. We shall explore this question in a separate publication.

In the case that $N\nu \to \lambda$, with $0 < \lambda < \infty$, we have

$$\frac{h''(0)}{h(0)} - \left[\frac{h'(0)}{h(0)}\right]^2 \sim \frac{1-\nu}{N} - \left\{\frac{\Gamma''(\lambda)}{\Gamma(\lambda)} - \left[\frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right]^2\right\}. \tag{4.44}$$

An estimate of the variance of N^* then reduces to

$$\begin{aligned}\text{Var } N^* &\sim \lambda^2\left\{\frac{1}{N} - \frac{\Gamma''(\lambda)}{\Gamma(\lambda)} + \left[\frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right]^2\right\} \\ &\quad + \lambda\left\{\log N - \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right\} \sim \lambda \log N.\end{aligned} \tag{4.45}$$

Case 4. If $N\nu \log N \to 0$ as $N \to \infty$, $\nu \to 0$, then it is easy to verify that $E(N^*) \to 1$ and Var $N^* \to 0$. This means that essentially all genes are of one allelic type. In the case that $N\nu \log N \to c > 0$ a simple calculation establishes that $E(N^*) \to 1 + c$ and Var $(N^*) \to c$.

5. Model II—case of general fertility

In this section we discuss the case of model II where the transition probability matrix governing the changes of

$$\bar{m} = (m_1, m_2, \cdots, m_r) \tag{5.1}$$

is that induced by conditioning a direct product branching process with offspring distribution generated by the probability generating function $f(s) = \sum_{k=0}^{\infty} a_k s^k$, $a_k \geqq 0$, $1 = f(1)$. In this setup the full population of N individuals may change in one transition. The interpretation of β and $\nu = r\beta$ is the same as before. We determine $P_{\bar{m},\bar{n}}$ the transition probability matrix as follows. For the states $\bar{m} = (m_1, m_2, \cdots, m_r)$ and $\bar{n} = (n_1, n_2, \cdots, n_r)$ where m_i, n_j are non-negative integers and $\sum m_i = \sum n_j = N$, we have

$$P_{\bar{m},\bar{n}} = \frac{\text{coeff } s_1^{n_1} s_2^{n_2} \cdots s_r^{n_r} \text{ in } \prod_{\alpha=1}^{r} f^{m_\alpha}\{s_1\beta + \cdots + s_{\alpha-1}\beta + s_\alpha[1-(r-1)\beta] + s_{\alpha+1}\beta + \cdots + s_r\beta\}}{\text{coeff } w^N \text{ in } f^N(w)}. \tag{5.2}$$

The rationale and scope of (5.2) are elaborated in Karlin and McGregor [9] to which we refer the reader. The special case where $f(s) = e^{\lambda(s-1)}$ leads to the transition probability matrix

$$(5.3)\quad P_{\bar{m},\bar{n}} = \frac{N!}{n_1!\, n_2! \cdots n_r!} \prod_{\alpha=1}^{r} \left\{ \frac{(N - m_\alpha)}{N} \beta + \frac{m_\alpha}{N} [1 - (r-1)\beta] \right\}^{n_\alpha},$$

which is the binomial sampling model commonly employed in analyzing fluctuation of gene frequency engendered by genetic drift (see Wright [12] and Fisher [3]).

Under conditions of equilibrium and where $r \to \infty$, we can determine for this model the expected number of mutant types in the population and other related quantities of interest.

Let $\{\pi_{\bar{m}}\}$ denote the stationary distribution of the process. We are interested in computing $E(N_7^*)$, $E[N_7^*(k)]$ and other similar quantities. We will have to distinguish two cases according as $N\nu$ is large, or $N\nu$ moderate.

Case 1. $N\nu \to \infty$. Clearly,

$$(5.4)\quad E(N_7^*) = r \Pr\{m_1 \neq 0\},$$

and therefore it suffices to evaluate $\Pr\{m_1 \neq 0\}$. Because of symmetry, we can lump all the other types together as a single type and regard the Markov chain on the state space consisting of all pairs $(m, N - m)$, $m = 0, 1, \cdots, N$, with probability transition matrix

$$(5.5)\quad P_{nm} = \frac{\text{coeff } s^m t^{N-m} \text{ in } f^m\{[1 - (r-1)\beta]s + (r-1)\beta t\}\, f^{N-m}[\beta s + (1-\beta)t]}{\text{coeff of } w^N \text{ in } f^N(w)}.$$

We denote the stationary distribution by $\{\pi_m\}_{m=0}^N$, that is, π_m is the probability in an equilibrium state that the population consists of m individuals of type 1, and $N - m$ individuals of the other types. We would like to evaluate

$$(5.6)\quad \Pr\{m_1 \neq 0\} = \pi_1 + \pi_2 + \cdots + \pi_N = 1 - \pi_0.$$

Bounds for (5.5) can be achieved in terms of the successive moments of the probability distribution $\{\pi_m\}$ as follows. Let

$$(5.7)\quad U_k = \sum_{m=k}^{N} (m)_k \pi_m, \qquad (m)_k = m(m-1)\cdots(m-k+1),$$

denote the factorial moments of $\{\pi_m\}_{m=0}^N$. Plainly

$$(5.8)\quad U_1 > 1 - \pi_0.$$

Notice that

$$(5.9)\quad U_1 - \frac{U_2}{2!} = \pi_1 + \pi_2 + \sum_{m=3}^{N} \left[\binom{m}{1} - \binom{m}{2} \right] \pi_m < \pi_1 + \pi_2 < 1 - \pi_0,$$

and generally

$$(5.10)\quad U_1 - \frac{U_2}{2!} + \frac{U_3}{3!} - \cdots \pm \frac{U_\ell}{\ell!} = \sum_{k=1}^{\ell} \pi_k + \sum_{m=\ell+1}^{N} \left[\sum_{k=1}^{\ell} (-1)^{k+1} \binom{m}{k} \right] \pi_m.$$

Observe that

$$(5.11)\quad \left[\sum_{k=1}^{\ell}(-1)^{k+1}\binom{m}{k}\right] = 1 - (-1)^{\ell}\binom{m-1}{\ell} \begin{matrix} \leqq 0 \text{ for } m \geqq \ell+1, \ell \text{ even}, \\ \geqq 0 \text{ for } m \geqq \ell+1, \ell \text{ odd}. \end{matrix}$$

Hence, we have the inequalities

$$(5.12)\quad U_1 - \frac{U_2}{2} + \frac{U_3}{3!} - \cdots (-1)\frac{U_{2\ell}}{(2\ell)!} < 1 - \pi_0 < U_1 - \frac{U_2}{2} + \frac{U_3}{3!} - \cdots + \frac{U_{2\ell-1}}{(2\ell-1)!}$$

for all $\ell = 1, 2, \cdots$.

Now the factorial moments can be determined recursively. We illustrate the method in the case of the first two moments and record the general formula.

We need the following transformation properties of $P = \|P_{nm}\|$ (see Karlin and McGregor [9])

$$(5.13)\quad \sum_{m=0}^{N} P_{nm}m = n[1 - (r-1)\beta] + (N-n)\beta,$$

$$(5.14)\quad \begin{aligned}\sum P_{nm}m(m-1) = {} & \{n(n-1)[1-(r-1)\beta]^2 \\ & + 2n(N-n)[1-(r-1)\beta]\beta \\ & + (N-n)(N-n-1)\beta^2\}\lambda_2 \\ & + \mu_2\{n[1-(r-1)\beta)]^2 + (N-n)\beta^2\},\end{aligned}$$

$$(5.15)\quad \lambda_2 = \frac{\text{coeff } w^{N-2} \text{ in } f^{N-2}(w)\,[f'(w)]^2}{\text{coeff } w^N \text{ in } f^N(w)}, \qquad \mu_2 = \frac{\text{coeff } w^{N-2} \text{ in } f^{N-1}(w)\,f''(w)}{\text{coeff } w^N \text{ in } f^N(w)},$$

and generally,

$$(5.16)\quad \begin{aligned}\sum P_{nm}(m)_k = {} & [\lambda_k(1-\nu)^k + \beta b_{k,1} + \beta^2 b_{k,2} + \cdots + \beta^k b_{k,k}](n)_k \\ & + (1-\nu)^k \sum_{i=1}^{k-1} a_{ki}(n)_{k-i} + C_{k-1}(\beta, n),\end{aligned}$$

where a_{ki} are suitable coefficients that can be computed routinely, if necessary, and

$$(5.17)\quad \lambda_k = \frac{\text{coeff } s^{N-k} \text{ in } [f(s)]^{N-k}\,[f'(s)]^k}{\text{coeff } s^N \text{ in } [f(s)]^N}, \qquad k = 2, 3, \cdots, N.$$

The term $C_k(\beta, n)$ is a polynomial of degree $k-1$ in the variable n such that the coefficient of n^ℓ has at least a factor β^ℓ, $\ell = 0, 1, \cdots, k-1$.

Note, using (5.16), we see that

$$(5.18)\quad \begin{aligned}U_k = \sum_{m=0}^{N}(m)_k\pi_m &= \sum_{m=0}^{N}(m)_k \sum_{m=0}^{N}\pi_n P_{nm} = \sum_{m=0}^{N}\pi_n \sum_{m=0}^{N} P_{nm}(m)_k \\ &= \lambda_k(1-\nu)^k U_k + (1-\nu)^k \sum_{i=1}^{k} a_{ki} U_{k-i} + \tilde{C}(\beta, U_1, \cdots, U_k),\end{aligned}$$

$k = 1, 2$, and $\tilde{C}(\beta, U_1, \cdots, U_k)/\beta$ tends to zero as $r \to \infty$ where $r\beta = \nu$ is fixed.

We use the relations (5.18) to evaluate $\lim_{r\to\infty} rU_r$. Thus,

$$U_1 = U_1(1-\nu) + N\beta \qquad \text{or} \qquad U_1 = \frac{N}{r}$$

so that

$$\lim_{r\to\infty} rU_1 = N. \tag{5.19}$$

In a similar manner we deduce

$$\lim_{r\to\infty} rU_2 = \frac{(1-\nu)^2 N\mu_2}{1-\lambda_2(1-\nu)^2}, \tag{5.20}$$

and generally,

$$\lim_{r\to\infty} rU_k = \frac{(1-\nu)^k\left[\sum_{i=1}^{k-1} a_{ki}\left(\lim_{r\to\infty} rU_{k-i}\right)\right]}{1-\lambda_k(1-\nu)^k}. \tag{5.21}$$

Combining (5.12), (5.19), and (5.20), we get the following estimates

$$\lim_{r\to\infty} rU_1 - \frac{\lim_{r\to\infty} rU_2}{2} \leqq E(N^*) \leqq \lim_{r\to\infty} rU_1, \tag{5.22}$$

or, what is the same,

$$N - \frac{(1-\nu)^2\mu_2 N}{2[1-\lambda_2(1-\nu)^2]} \leqq E(N^*) \leqq N. \tag{5.23}$$

Improved estimates can be obtained by exploiting (5.12) and the evaluations (5.21).

Case 2. $N\nu \to \lambda$ *for* $0 < \lambda < \infty$, $N \to \infty$. We next discuss the case where $N\nu \to \lambda$ is moderate. We shall use the standard diffusion approximation to the process which fits well for N large and λ not too big. In this case the stationary distribution $\{\pi_m\}$ of the frequency state m/N can be approximated by

$$\pi_m \sim Cx^{-1+2N\beta/\gamma}(1-x)^{-1+2N(r-1)\beta/\gamma}\,dx, \qquad [Nx] = m, \tag{5.24}$$

where $[h]$ denotes the greatest integer not exceeding h, while

$$C = \frac{\Gamma(2Nr\beta/\gamma)}{\Gamma(2N\beta/\gamma)\Gamma(2N(r-1)\beta/\gamma)} \tag{5.25}$$

and $\gamma = f''(1)$ provided the scaling is such that $f'(1) = 1$. In the case that $f'(1) \neq 1$ then γ has to be defined differently (see [8]). The quantity $1 - \pi_0$ is therefore approximable by

$$\Gamma(2N\nu/r\gamma)^{-1}\int_{1/N}^{1} x^{-1+2N\nu/r\gamma}(1-x)^{-1+2N\nu/\gamma}\,dx \sim \frac{2N\nu}{r\gamma}\int_{1/N}^{1} x^{-1}(1-x)^{-1+2N\nu/\gamma}\,dx \tag{5.26}$$

as $r \to \infty$. It follows that

$$\lim_{r\to\infty} r\Pr\{m_1 \neq 0\} = \frac{2N\nu}{\gamma}\int_{1/N}^{1} x^{-1}(1-x)^{-1+2N\nu/\gamma}\,dx = E(N^*) \tag{5.27}$$

which agrees with the formula (4.24) apart from the factor $1/\gamma$. The factor $1/\gamma$ reflects the random nature of the fertility distribution and this is 1 in the case when $f(s)$ is Poisson.

In the case of isoheterotic alleles the approximate formula for $E(N^*)$ has to be modified and takes the form

$$E(N^*) = \frac{2N\nu}{\gamma} \int_{1/N}^{1} x^{-1}(1-x)^{-1+2N(\nu+s)/\gamma} e^{2Nsx/\gamma}\, dx \tag{5.28}$$

where $1 + s$ is the selective advantage of any heterozygote relative to selective coefficient 1 for homozygotes.

6. Connections to Fisher's theory of balance of mutation and genetic drift

Fisher [3] was interested in the problem of determining the rate of mutation necessary in order to maintain a sufficient degree of heterozygosity which thus contributes toward the genetic variance. More precisely, consider a fixed number L of loci, at each of which there is initially one mutant and $2N - 1$ nonmutant genes. Fisher implicitly assumes that the fluctuations of these specific mutant populations, relative to a given locus, follow the laws of the Markov chain on the states $0, 1, \cdots, N$, with transition probability matrix

$$P_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}. \tag{6.1}$$

Each locus will eventually become homozygous (that is, either lost or fixed) due to random elimination if no further mutation occurs. Let L be chosen so that on the average, one locus becomes homozygous per generation. To compensate for this, at each generation a new locus is introduced having one mutant and $2N - 1$ nonmutants. The problem is to evaluate L and at equilibrium to determine the mean number b_i of loci having i, $i = 1, 2, \cdots, 2N - 1$, mutants.

It is convenient to alter the model as follows. Instead of feeding in one new mutant each generation, we start each new locus with a random number of mutants with possible values $1, 2, \cdots, 2N - 1$ whose probabilities are $c_1, c_2, \cdots, c_{2N-1}$ where $\{c_i\}_{i=1}^{2N-1}$ is the conditional limiting distribution of the state of the Markov chain given fixation has not occurred. The vector $c = \{c_i\}_{i=1}^{2N-1}$ is the left eigenvector associated with the eigenvalue $\lambda_2 = 1 - 1/2N$ of the matrix defined in (6.1) normalized so that $\sum_{i=1}^{2N-1} c_i = 1$.

Approach to absorption for the Markov chain induced by the transition probability matrix (5.5) occurs at a rate $1/2N$, that is, in each generation on the average $1/2N$ of the existing populations become fixed. We choose $L = 2N$, and then because each new locus is started with a size following the distribution law $\{c_i\}_{i=1}^{2N-1}$ the expected number of loci becoming fixed in each generation is one.

An essentially equivalent formulation is as follows. Consider a single locus at which there is initially one mutant and thus $2N - 1$ nonmutants. After a

variable number of generations, fixation of either mutant or nonmutant occurs. Let the mean number of generations for this to happen be K and let the mean number of generations for which the number of mutants is i be d_i. By definition $K = \sum_1^{2N-1} d_i$ and it is not difficult to see by a standard ergodic argument that $K = L$, $d_i = b_i$, also.

Now consider the setup of model I of creating one new mutant line each generation where initially there are $L = 2N$ mutant lines.

The computation of $E[N^*(k)]$, the expected number of mutant lines with k representatives existing in statistical equilibrium starting initially with N lines, is, of course, the same as evaluating the quantity b_k. Fisher succeeds in estimating b_k by passing to a related branching process. He determines b_k approximately to be a/k (a is a suitable constant). This agrees with the result of theorem 2.2 particularly in reference to the example of linear growth (see (2.10)).

It is important to emphasize that the analysis of Fisher's problem after appropriate identifications reduces to that of the analysis of model I. That is, the problem of ascertaining the number of loci with k heterozygotes out of a population of N individuals is mathematically equivalent to the problem of determining the number of different alleles maintained by a balance of mutation and random elimination as set forth in model I.

7. Conclusions and discussion

The existence of excessive numbers of polymorphism in natural populations is well established by empirical and laboratory tests. A large multiplicity of alleles at a locus has been observed frequently for many different characteristics in plant, insect, and human populations. The usual models involving the effects of mutation rates alone fail to account for these polymorphisms. For example, for populations of size 10^3 it would be necessary to ascribe a mutation rate as high as 10^{-3} which is completely inconsistent with known mutation rates. Much discussion and controversy has evolved in recent years in attempting to explain the reasons for the preponderance of multiallelic phenomena. We have formulated two different models to study this problem and now we sum up some of the qualitative ideas and conclusions derived from them.

In the second model where population size is kept constant, we obtained the estimate $\sim(2N\nu/\gamma) \log N$ for the expected number $E(N^*)$ of different alleles provided $N\nu = \lambda$ is not excessively large, that is, $\lambda \leqq 20$, while N is large, $N > 10^3$ (N is the total number of distinct genes in the population) and γ is the square of the coefficient of variation of the number of offspring per parent per generation. Under the above circumstances the numbers of different alleles represented is approximately $E(N^*) \sim 2\lambda c \log N$ where $\lambda = N\nu$ and $c = 1/\gamma$.

Obviously $E(N^*)$ is large of the order of magnitude of $\log N$ if N is large provided $2\lambda c$ is not too small. The variance of N^* is also of the order $\lambda c \log N$ and therefore the actual number of alleles at a given locus will fluctuate between

$$2\lambda c \log N \pm 2(\lambda c \log N)^{1/2} \tag{7.1}$$

with probability exceeding 0.95. Furthermore, since the number of different characteristics of populations which are genetically controlled are very large, we would expect some loci (not many) where the number of alleles is the order of magnitude

$$2\lambda c \log N + 2(\lambda c \log N)^{1/2}. \tag{7.2}$$

The above discussion points out that substantial numbers of alleles (of the order log N) at a locus can occur if N is sufficiently large provided $N\nu = \lambda$ is not too small. This number is increased at some occasional loci due to random fluctuations. The magnitude of the coefficient of variation of the number of progeny per parent per generation contributes also an influencing factor.

The analysis of model I introduces a new and perhaps more basic set of ideas emphasizing the important fact that the lifetime distribution of the population of a given allele may be a significant factor in accounting for large numbers of different alleles represented. Specifically, if the life of a specific allele in the population is very long, then even with very small mutation (and/or migration) rates the number of different alleles in existence becomes large if a long time has elapsed.

Selection effects are reflected in the nature of the lifetime distribution associated with a specific allele population. A heterotic allele possesses a lifetime distribution of large mean, while harmful alleles have relatively short expected lifetimes.

It is plausible and readily established by analytic investigations of standard models describing fluctuation of the numbers of rare mutants in plant and insect population that the lifetime of each allele is usually very long (for all practical purposes infinite mean lifetime) even for the case where ultimate extinction of the particular allele in question is certain. This is definitely true for selectively neutral alleles; a slight selective disadvantage of an allele would imply a finite expected lifetime which usually is still of long duration. On the other hand, a heterotic allele would entail a small positive probability of the allele being permanently established and so its mean lifetime is infinite. For the case of selectively neutral alleles where the expected lifetime of an allele is infinite but ultimate extinction is a certain event, it is a consequence of the results of model I that if the process is in effect a long time then even for exceptionally small mutation rates the expected number of alleles in existence is large and is an increasing function of the length of time of the operation of the process. A change of environmental conditions can occur invalidating the preceding model resulting in a substantial decrease of the population size of all genes, whereas during the same time period the number of different alleles is diminished by a much smaller factor owing to the long lifetime of each allele type.

Since environmental conditions do constantly change, *the study of stationary or equilibrium models is unable to account for the large number of alleles.* It is

precisely the transient character of the process coupled with the long lifetime distribution of individual allele types that invalidate the usual postulate of an equilibrium state. In many natural populations the population size is undergoing radical fluctuations and cannot be regarded as in an equilibrium situation. For such a case model I seems to be a reasonable representation and if the process has been going on for a long time both $E(N^*)$ and Var (N^*) are large, which is in accordance with the observations.

Another contributing source of confusion in discussions of genetic phenomena is the insistence of studying exclusively the changes in the frequencies of an allele rather than actual population sizes. Generally, changes in frequencies cannot be observed unless a sufficient number of generations have passed. Because most available data does not reflect a sufficient lapse of time (or generations), it is difficult to assess when a polymorphism is indeed a polymorphism. We further refer to Karlin and McGregor [7] where we have described the significance and fundamental character of the time factor relevant to describing variability of population size and frequencies of a given mutant type.

REFERENCES

[1] M. S. BARTLETT, *An Introduction to Stochastic Processes*, Cambridge, Cambridge University Press, 1955.
[2] W. J. EWENS, "The maintenance of alleles by mutation," *Genetics*, Vol. 50 (1964), pp. 891–898.
[3] R. A. FISHER, *The Genetical Theory of Natural Selection*, New York, Dover, 1958.
[4] T. E. HARRIS, *The Theory of Branching Processes*, Berlin, Springer-Verlag, 1963.
[5] S. KARLIN and J. MCGREGOR, "Linear growth, birth and death processes," *J. Math. Mech.*, Vol. 7 (1958), pp. 643–662.
[6] ———, "On a genetics model of Moran," *Proc. Cambridge Philos. Soc.*, Vol. 58 (1962), pp. 299–311.
[7] ———, "On some stochastic models in genetics," *Stochastic Models in Medicine and Biology*, Madison, University of Wisconsin Press, 1964, pp. 245–271.
[8] ———, "Direct product branching processes and related Markoff chains," *Proc. Nat. Acad. Sci. USA*, Vol. 51 (1964), pp. 598–602.
[9] ———, "Direct product branching processes and related induced Markoff chains, I. Calculations of rates of approach to homozygosity," *Bernoulli, Bayes, Laplace Anniversary Volume*, New York, Springer-Verlag, 1965, pp. 111–145.
[10] D. G. KENDALL, "Les processus stochastiques de croissance en biologie," *Ann. Inst. H. Poincaré*, Vol. 13 (1952), pp. 43–108.
[11] M. KIMURA and F. CROW, "The number of alleles that can be maintained in a finite population," *Genetics*, Vol. 49 (1964), pp. 725–738.
[12] P. A. P. MORAN, *The Statistical Processes of Evolutionary Theory*, Oxford, Clarendon Press, 1962.
[13] S. WRIGHT, "Evolution in Mendelian populations," *Genetics*, Vol. 16 (1931), pp. 97–159.
[14] ———, "On the number of self-incompatibility alleles maintained in equilibrium by a given mutation rate in a population of given size: a reexamination," *Biometrics*, Vol. 16 (1960), pp. 61–85.