Some Problems in Mathematical Genetics

A thesis for the degree of

Master of Science in Statistics.

Chia Ah Bah
Department of Statistics,
School of General Studies,
Australian National University.

June, 1966.

Contents

Chapter 1

Correcting Absorption Probabilities and Mean Absorption Times in Finite Genetic Models.

Chapter 2

Correcting Stationary Distribution.

Chapter 3

Correcting Absorption Probabilities by using the Pseudo-Transient Function.

Contents (cont.)

Chapter 4

The Maintenance of Alleles by Mutation.

Chapter 5

Initial Progress of new Gametes.

Acknowledgements.

I would like to express my sincere thanks to my supervisor, Dr W.J. Ewens for suggesting practically all the problems, his guidance and encouragement.

I am indebted to the A.N.U. for the research scholarship that made this work possible and to Mrs A. Debnam for her excellent typing of the thesis.

Summary

The primary aims of this thesis are (i) to ascertain the accuracy of diffusion methods in approximating certain finite genetic models and (ii) to examine some problems in mathematical genetics. A measure of the accuracy of the diffusion methods is demonstrated by finding the leading terms between the diffusion approximations and the true values. This has already been done by Ewens [1964a, 1964b] by making use of a function which he called the pseudo-transient distribution (P.T.D.). However, an alternative approach is presented here. This approach is simpler than that of Ewens and is treated in an Applied-Mathematics manner, the emphasis being mainly on the results and the method.

We justify our method by considering genetic models in which exact expressions are known for the various functions of interest and by comparing the exact correction terms obtained from the exact values by algebraic method with those obtained by the present method [see Chapter 1. section 1.3 and 1.5]. We notice that in section 1.3 the exact correction term [c.f. equations (1.35) and (1.36)] is obtained for the absorption probability in the Moran's model but this is not the case for the mean absorption time [c.f. equations (1.93) (1.94) and (1.95)]. This, we believe, is due to the fact of the inexactness of the boundary conditions [equation (1.76)] encountered there. However, (we have not considered this example in the thesis) in the random walk model with state space $\{0,1,2,\ldots,N\}$ and transition probabilities given by

(11)

$$P_{i,i-1} = \frac{1}{2} - \frac{C}{N}$$

$$P_{i,i+1} = \frac{1}{2} + \frac{C}{N}$$

$$P_{ii} = 0$$

$$P_{oo} = P_{NN} = 1$$

$$P_{ij} = 0 \qquad \text{for } . |i - j| \geq 2$$

where c is some positive constant, and in which the boundary conditions [equation (1.75)] present no difficulty, we can show that the corrected diffusion approximation $U^*(p)$ to the mean absorption time is given by

$$U^*(p) = -\frac{p}{c} + \frac{1}{c}\left[\frac{e^{2c}-e^{2c(1-p)}}{e^{2c}-1}\right] + \frac{2c^2}{3N^2(e^{2c}-1)}\left[\frac{e^{2c(1-p)}-e^{2c}}{e^{2c}-1} + pe^{2c(1-p)}\right].$$

Here the term involving N^{-2} is identical with that obtained from the exact value by algebraic method.

Chapter 2 is concerned with correcting stationary distributions in non-absorbing Markov genetic processes. In Chapter 3 we give an alternative application of the P.T.D. and shows that Ewens' method [1964a] is in fact equivalent to our present method. We also examine a conjecture by Ewens [1964a]. The number of different mutant forms that can be maintained in a population has recently attracted a great deal of research (Kimura and Crow [1964], Ewens [1964c]) and in Chapter 4 we extend Ewens' result [1964c] to the case when the population consists of two types of alleles. A generalisation to the case when the population consists of m types of alleles is also given. Finally in the last chapter we consider a problem in linkage and show that in a deterministic model there are no instances in which "looser" linkage is beneficial in the establishment of new gametes.

CHAPTER 1.

Correcting Absorption Probabilities and Mean Absorption Times in Finite Genetic Models.

1.1 Introduction

Diffusion approximations have been used in many finite genetic models mainly for two reasons. The first reason is that in certain genetic models, especially those classified under the heading of "Wright's models", no explicit expression has yet been found for many quantities of interest. The second reason is that even though exact expressions may have been found in the genetic models, diffusion approximations are preferred because they are less complicated and are easier to handle.

In this chapter we shall examine how close the diffusion approximations are to their true values. In these models in which no explicit expression is known, a measure of the accuracy of the diffusion approximations will be found by deriving the leading term of the difference between the diffusion approximation and the true value. This has already been done by Ewens [1964a, 1964b] by making use of a function which he called the pseudo-transient distribution. (P.T.D.); more will be said about an alternative application of the P.T.D. in Chapter 3. However, in this chapter we shall derive the leading term of the difference between the diffusion approximation and the true value by a method which does not involve the application of the P.T.D. In cases where exact expressions have been found, a measure of the accuracy of the diffusion approximation can also be found, of course,

by a purely algebraic method. We shall, in such cases, compare the exact algebraic correction and that derived by our present method and illustrate the accuracy of the latter.

1.2 Absorption Probabilities: Wright's model.

We consider the following model due to Wright. The model consists of a haploid population of fixed size N in which there are two genotypes A and a. We let s be the selective advantage of A over a where s is non-negative. ($s = 0$ corresponds to no selection.) The model then assumes that if at any generation, the number of A individuals is i, then the probability p_{ij} of getting exactly j A individuals in the next generation is given by

$$p_{ij} = \binom{N}{j}\left(\frac{i+si}{N+si}\right)^{j}\left(\frac{N-i}{N+si}\right)^{N-j} \qquad i,j = 0,1,2,\ldots N \qquad (1.1)$$

The expression p_{ij}, apart from the constants N and s, involves only i and j. Thus the number of A individuals between successive generations is a Markovian variate whose transition matrix P has typical element p_{ij}

$$P = [p_{ij}] \qquad i,j = 0,1,2,\ldots,N \qquad (1.2)$$

Now if E_i denotes the state "The number of A individuals = i" $(i = 0,1,2,\ldots,N)$ then E_0 and E_N are absorbing, all other states are transient, and eventually the process becomes absorbed in E_0 or E_N. No explicit formula has yet been found for absorption

probabilities for arbitrary initial state E_k. Thus some approximation is necessary and the approximation used is to replace the discrete stochastic process (1.2) by a continuous process. In order for the approximation to be valid, s must be of the order N^{-1} and we write

$$s = \frac{\alpha}{N} \text{ where } \alpha = 0(1) \tag{1.3}$$

Now if p is the frequency of the allele A so that the set of values taken by p is given by $\{p = \frac{i}{N} ; i = 0,1,2,\ldots N\}$, then it is easy to show that, if N is sufficiently large, the change in p per generation is sufficiently small to allow us to treat p as a continuous variate.

We want to find the probability that the process becomes absorbed in E_0 given initially that the process is in E_k, where $k = Np$ is the number of A individuals. Denote this probability by $P_0(p)$. Then, if $g(\delta p;p)$ is the probability density of the change from p to $p + \delta p$ in consecutive generations, we have

$$P_0(p) = \int P_0(p + \delta p)g(\delta p;p)d(\delta p) \tag{1.4}$$

where the integration is over all possible values of δp. (c.f. Kimura [1962]) Equation (1.4) is a special form of the Chapman-Kolmogorov equation in the theory of stochastic processes and is a consequence of the Markovian property of the random variable.

Expanding $P_0(p + \delta p)$ in the right hand side of the above

equation in terms of δp and writing

$$E[(\delta p)^k] \equiv \int (\delta p)^k g(\delta p:p)d(\delta p) = \mu_k(p) \quad k = 1,2,\ldots \quad (1.5)$$

and assuming that we can interchange the order of integration and summation, we obtain

$$\sum_{n=1}^{\infty} \frac{\mu_n(p)}{n!} \frac{d^n}{dp^n} P_0(p) = 0 \quad (1.6)$$

Thus the absorption probability $P_0(p)$ may be formed by solving (1.6) with the boundary conditions:

$$P_0(0) = 1 \text{ and } P_0(1) = 0 \quad (1.7)$$

In the above Wright's model, it can be shown by using moments of the binomial distribution and after some algebra, that

$$\begin{aligned}
\mu_1(p) &= E[x_{t+1}-x_t \mid x_t = p] = \frac{\alpha p(1-p)}{N} - \frac{\alpha^2 p^2(1-p)}{N^2} + 0(N^{-3}) \\
\mu_2(p) &= E[(x_{t+1}-x_t)^2 \mid x_t = p] \\
&= \frac{p(1-p)}{N} + \frac{\alpha p(1-p)[1-2p+\alpha p(1-p)]}{N^2} + 0(N^{-3}) \\
\mu_3(p) &= \frac{p(1-p)(1-2p)+3\alpha p^2(1-p)^2}{N^2} + 0(N^{-3}) \\
\mu_4(p) &= \frac{3p^2(1-p)^2}{N^2} + 0(N^{-3}) \\
\mu_k(p) &= 0(N^{-3}) \quad \text{for} \quad k = 5,6,\ldots
\end{aligned} \quad (1.8)$$

Now since $\mu_k(p)$ are of the order $N^{-\gamma}$ where γ is some positive integer and N is large, we may as a first approximation $\bar{P}_0(p)$ to $P_0(p)$ ignore terms of order smaller than N^{-1}. Hence we obtain from equation (1.6) the following differential equation (d.e.)

$$\bar{\mu}_1(p)\frac{d\bar{P}_0}{dp} + \frac{1}{2}\bar{\mu}_2(p)\frac{d^2\bar{P}_0}{dp^2} = 0 \qquad (1.9)$$

where

$$\begin{aligned} \bar{\mu}_1(p) &= \frac{\alpha p(1-p)}{N} \\ \bar{\mu}_2(p) &= \frac{p(1-p)}{N} \end{aligned} \qquad (1.10)$$

and subject to the boundary conditions: $\bar{P}_0(0) = 1$ and $\bar{P}_0(1) = 0$ Equation (1.9) reduces to

$$2\alpha\frac{d\bar{P}_0}{dp} + \frac{d^2\bar{P}_0}{dp^2} = 0 \ , \qquad (1.11)$$

whose solution is

$$\bar{P}_0(p) = \frac{e^{-2\alpha p} - e^{-2\alpha}}{1 - e^{-2\alpha}} \ . \qquad (1.12)$$

This is the diffusion approximation to the required absorption probability (see p. 6 of Ewens [1964a]).

We now wish to improve on the diffusion approximation $\bar{P}_0(p)$. To do this, we take into account terms of order N^{-2} in our d.e. and ignore terms of smaller order than this. We let

$$P_0^*(p) = \bar{P}_0(p) + \frac{q(p)}{N} \qquad (1.13)$$

be a more accurate approxmiation than $\bar{P}_0(p)$. Our purpose is to

determine the unknown function $q(p)$ to the order of accuracy required in the problem. Then from equation (1.6) we see that $P_0^*(p)$ satisfies the d.e.

$$\mu_1^*(p)\frac{dP_0^*}{dp} + \frac{1}{2}\mu_2^*(p)\frac{d^2P_0^*}{dp^2} + \frac{1}{3!}\mu_3^*(p)\frac{d^3P_0^*}{dp^3} + \frac{1}{4!}\mu_4^*(p)\frac{d^4P_0^*}{dp^4} = 0 \tag{1.14}$$

with the boundary conditions:

$$P_0^*(0) = 1 \quad \text{and} \quad P_0^*(1) = 0 \tag{1.15}$$

Here

$$\begin{aligned}
\mu_1^*(p) &= \frac{\alpha p(1-p)}{N} - \frac{\alpha^2 p^2(1-p)}{N^2} \\
\mu_2^*(p) &= \frac{p(1-p)}{N} + \frac{\alpha p(1-p)[1-2p+\alpha p(1-p)]}{N^2} \\
\mu_3^*(p) &= \frac{p(1-p)(1-2p)+3\alpha p^2(1-p)^2}{N^2} \\
\mu_4^*(p) &= \frac{3p^2(1-p)^2}{N^2}
\end{aligned} \tag{1.16}$$

Using equations (1.12) and (1.13) in (1.14) and ignoring terms of order N^{-3} we obtain, after simplification the d.e.

$$2\alpha\frac{d}{dp}q(p) + \frac{d^2}{dp^2}q(p) = \frac{4}{3}(p+1)\cdot\frac{\alpha^3 e^{-2\alpha p}}{1-e^{-2\alpha}} \tag{1.17}$$

with the boundary conditions:

$$q(0) = q(1) = 0 \tag{1.18}$$

Solving the above d.e. we obtain

$$\begin{aligned}
q(p) = [&\frac{2\alpha+1-(6\alpha^2+4\alpha+1)e^{-2\alpha}}{6(1-e^{-2\alpha})^2} - \frac{2\alpha^2 p(p+2)+2\alpha p+2\alpha+1}{6(1-e^{-2\alpha})}]e^{-2\alpha p} \\
&+ \frac{\alpha(3\alpha-1)e^{-2\alpha}}{3(1-e^{-2\alpha})^2}
\end{aligned} \tag{1.19}$$

Now $q(p)$ can be expressed in the form

$$q(p) = \frac{A + B}{1-e^{-2\alpha}} \qquad (1.20)$$

where

$$A = \frac{1}{6}\left[\frac{e^{-2\alpha}-e^{-2\alpha p}}{e^{-2\alpha}-1}\right]\{4\alpha^2(p+\tfrac{1}{2}p^2) + 2\alpha[(1+p)e^{-2\alpha p} - 1] + e^{-2\alpha p} - 1\}$$

$$B = \frac{1}{6}\left[\frac{e^{-2\alpha p}-1}{e^{-2\alpha}-1}\right]\left[(1+2\alpha+2\alpha p)e^{-2\alpha p} - (6\alpha^2-4\alpha^2 p-2\alpha^2 p^2+4\alpha-1)e^{-2\alpha}\right]$$

Thus a more accurate formula than equation (1.12) is

$$P_0^*(p) = \frac{e^{-2\alpha p}-e^{2\alpha}}{1-e^{-2\alpha}} + \frac{1}{N} \circ \frac{A + B}{1-e^{-2\alpha}} \qquad (1.21)$$

which agrees with Ewens' result obtained by the application of the P.T.D. (see equation (6.8). Ewens [1964b]). The above approach is much simpler than that of Ewens.

1.3 Absorption Probabilities: Moran's model

Moran [1958] introduced a genetic model to study the effect of solution on the fluctuations of gene frequencies. There is a haploid population of of fixed size N with only two genotypes A and a in which the selective advantage of A over a is s. An individual is randomly chosen to die and is replaced by a new individual which is either A or a with probabilities $\frac{(1+s)j}{N+sj}$ and $\frac{N-j}{N+sj}$ respectively, where j is the number of A individuals before the birth-death event. The number of A individuals is then $j-1$, j, $j+1$ with probabilities $p_{j,j-1}$, p_{jj} and $p_{j,j+1}$ respectively, where

$$
\begin{aligned}
p_{j,j-1} &= \frac{j}{N} \cdot \frac{N-j}{(N+sj)} \equiv \pi_j \\
p_{j,j+1} &= \frac{N-j}{N} \cdot \frac{(1+s)j}{(N+sj)} \equiv \eta_j \\
p_{jj} &= 1 - \pi_j - \eta_j \\
p_{jk} &= 0 , \qquad |k-j| \geq 2
\end{aligned}
\tag{1.22}
$$

Therefore, the number of A individuals is a Markovian variate with transition matrix defined by equation (1.22). The states E_0 and E_N are absorbing whereas the other states E_i $(i = 1,2,\ldots,N-1)$ are transient. Thus, eventually the system will enter either E_0 or E_N. We have not introduced mutation in the model. However, if two-way mutation were introduced in the model, the states E_0 and E_N are no longer absorbing and eventually the process will settle down to equilibrium. In this case the quantity of interest to us is then the stationary distribution. We shall discuss correcting diffusion approximations to stationary distributions in the next chapter.

In the above model exact expressions for the absorption probabilities are known. Moran [1958] showed that the probability P_0 that the A individuals were eventually eliminated was given by

$$
P_0 = \frac{(1+s)^{N-K} - 1}{(1+s)^N - 1} \tag{1.23}
$$

where K was the initial number of A individuals in the population.

We shall now obtain the diffusion approximation and the corrected diffusion approximation to P_0. In this way we are able to illustrate the accuracy of the diffusion approximation as

well as the accuracy of the method employed in correcting diffusion approximation. For the approximation to be valid we must have s to be of the order N^{-1} and using the same notation as equation (1.3) we write

$$s = \frac{\alpha}{N}$$

where $\alpha = 0(1)$. We also assume that N is large. Thus we can apply equation (1.6) to find the absorption probability $P_0(p)$.

Now for Moran's model, it is easy to show that

$$\begin{aligned}
\mu_1(p) &= \frac{\alpha p(1-p)}{N^2}\left(1 - \frac{\alpha p}{N} + \frac{\alpha^2 p^2}{N^2}\right) + 0(N^{-5}) \\
\mu_2(p) &= \frac{p(1-p)}{N^2}\left[2 + \frac{\alpha(1-2p)}{N} + \frac{\alpha^2 p(2p-1)}{N^2}\right] + 0(N^{-5}) \qquad (1.24) \\
\mu_n(p) &= \frac{1}{N^2}\mu_{n-2}(p) \qquad n = 3,4,\ldots
\end{aligned}$$

As a first approximation $\bar{P}_0(p)$ to $P_0(p)$ we ignore terms of order smaller than N^{-2} . Then from equation (1.6) we find that $\bar{P}_0(p)$ satisfies the d.e.

$$\bar{\mu}_1(p)\frac{d\bar{P}_0}{dp} + \frac{1}{2}\bar{\mu}_2(p)\frac{d^2\bar{P}_0}{dp^2} = 0 \qquad (1.25)$$

where

$$\begin{aligned}
\bar{\mu}_1(p) &= \frac{\alpha p(1-p)}{N^2} \\
\bar{\mu}_2(p) &= \frac{2p(1-p)}{N^2}
\end{aligned} \qquad (1.26)$$

subject to the obvious boundary conditions:

$$\bar{P}_0(0) = 1 \quad \text{and} \quad \bar{P}_0(1) = 0$$

Solving the above d.e. we get

$$\overline{P}_o(p) = \frac{e^{\alpha(1-p)}-1}{e^{\alpha}-1} \qquad (1.27)$$

To find a more accurate formula than $\overline{P}_o(p)$ we now include terms of order N^{-3} but exclude terms of order smaller than N^{-3}. We let

$$P_o^*(p) = \overline{P}_o(p) + \frac{q(p)}{N} \qquad (1.28)$$

be a more accurate formula than equation (1.27) where $q(p)$ is the unknown function which we wish to determine. Then from equation (1.6), $P_o^*(p)$ satisfies the d.e.

$$\mu_1^*(p)\frac{dP_o^*}{dp} + \frac{1}{2}\mu_2^*(p)\frac{d^2P_o^*}{dp^2} = 0 \qquad (1.29)$$

where

$$\mu_1^*(p) = \frac{\alpha p(1-p)}{N^2}\left(1 - \frac{\alpha p}{N}\right) \qquad (1.30)$$

$$\mu_2^*(p) = \frac{p(1-p)}{N^2}\left[2 + \frac{\alpha(1-2p)}{N}\right]$$

subject to the boundary conditions

$$P_o^*(0) = 1 \quad \text{and} \quad P_o^*(1) = 0 \qquad (1.31)$$

Substituting equation (1.28) into equation (1.29) and ignoring terms of order smaller than N^{-3}, we obtain the following d.e.

$$\alpha\frac{dq}{dp} + \frac{d^2q}{dp^2} = \frac{-\alpha^3 e^{\alpha(1-p)}}{2(e^{\alpha}-1)} \qquad (1.32)$$

Equation (1.31) implies that the boundary condition for the above d.e. are

$$q(0) = q(1) = 0 \qquad (1.33)$$

The solution to the d.e. (1.32) is

$$q(p) = \frac{\alpha^2 e^{\alpha}}{2(e^{\alpha}-1)} \left[\frac{e^{\alpha(1-p)}-1}{e^{\alpha}-1} - (1-p)e^{-\alpha p} \right]$$

$$= \frac{\alpha^2 e^{\alpha}}{2(e^{\alpha}-1)} \left[\overline{P}_0(p) - (1-p)e^{-\alpha p} \right] \qquad (1.34)$$

Hence, a more accurate formula than $\overline{P}_0(p)$ is

$$P_0^*(p) = \overline{P}_0(p) + \frac{\alpha^2 e^{\alpha}}{2N(e^{\alpha}-1)} \left[\overline{P}_0(p) - (1-p)e^{-\alpha p} \right] \qquad (1.35)$$

By putting $p = \frac{K}{N}$ and $s = \frac{\alpha}{N}$ in equation (1.23), we can express P_0 in the form

$$\frac{(1+\alpha/N)^{N(1-p)}-1}{(1+\alpha/N)^N - 1} = \overline{P}_0(p) + \frac{\alpha^2 e^{\alpha}}{2N(e^{\alpha}-1)} \left[\overline{P}_0(p) - (1-p)e^{-\alpha p} \right] + O(N^{-3}) \qquad (1.36)$$

by using the fact that

$$\left(1 + \frac{x}{N}\right)^N = e^x - \frac{x^2 e^x}{2N} + \left(\frac{x^3}{3} + \frac{x^4}{8}\right) \frac{e^x}{N^2} + \ldots \qquad (1.37)$$

Comparing equations (1.35) and (1.36) we notice that $P_0^*(p)$ is accurate up to the order of N^{-1} illustrating that for large N, the diffusion approximation should give us a very good approximation to its true value. Ewens [1964a] obtained the same correction term, equation (1.34), by an apparently different method involving the use of the pseudo-transient distribution. We shall discuss this method in detail in Chapter 3 and show that his method is, in fact, equivalent to our present method, which explains why we obtain the same result.

1.4 Mean Absorption Times: Wright's Model.

In the above genetic models we have discussed diffusion approximation and corrected diffusion approximation to the probability

that eventually the population will consist entirely of a individuals. Now another quantity which is of interest in such models is the mean absorption time. This is the mean time taken for the population to become entirely A or a individuals. Again no general explicit formula for the mean absorption time has yet been found in the Wright's model and we shall use a method similar to the one employed in sections 1.2 and 1.3 to find the diffusion approximation and its corrected value.

For simplicity, we consider the Markov chain, equation (1.2), for the case $s = 0$ we measure time in units of N generations. Denote the mean absorption time by $U(p)$. Then, if $g(\delta p;p)$ is the probability density of the change from p to $p + \delta p$ in consecutive generations, we have

$$U(p) = \int U(p + \delta p)g(\delta p;p)d(\delta p) + \frac{1}{N} \qquad (1.38)$$

where the integration is over all possible values of δp. Equation (1.38) is the continuous analogue of the discrete case. The term $\frac{1}{N}$ is due to the fact that we have now measure time in units of N generations.

Expanding $U(p + \delta p)$ in the right hand side of the above equation in terms of δp, and using the same notation as in equation (1.5) and with usual assumption, we have

$$\frac{1}{N} + \sum_{k=1}^{\infty} \frac{\mu_k(p)}{k!} \frac{d^k}{dp^k} U(p) = 0 \qquad (1.39)$$

Hence, $U(p)$ may be obtained by solving the above d.e. with the boundary conditions:

$$U(0) = U(1) = 0 \tag{1.40}$$

Now for Wright's model, with $s = 0$ (i.e. $\alpha = 0$) we obtain from equation (1.8),

$$\begin{aligned}
\mu_1(p) &= 0 \\
\mu_2(p) &= \frac{p(1-p)}{N} + 0(N^{-3}) \\
\mu_3(p) &= \frac{p(1-p)(1-2p)}{N^2} + 0(N^{-3}) \\
\mu_4(p) &= \frac{3p^2(1-p)^2}{N^2} + 0(N^{-3}) \\
\mu_k(p) &= 0(N^{-3}) , \qquad k = 5,6,\ldots
\end{aligned} \tag{1.41}$$

We let $\overline{U}(p)$ be a first approximation to $U(p)$ when we ignore terms of order smaller than N^{-1}. Then from equation (1.39) we see that $\overline{U}(p)$ satisfies the d.e.

$$\frac{1}{2}\bar{\mu}_2(p)\frac{d^2\overline{U}}{dp^2} + \frac{1}{N} = 0 \tag{1.42}$$

where $$\bar{\mu}_2(p) = \frac{p(1-p)}{N}$$

and with the boundary conditions:

$$\overline{U}(0) = \overline{U}(1) = 0 \tag{1.43}$$

On solving the above d.e., we obtain

$$\overline{U}(p) = -2[p\ell np + (1-p)\ell n(1-p)] \tag{1.44}$$

This is the diffusion approximation for the mean absorption times (see Ewens [1964b])

In order to find a more accurate formula than equation (1.44) we now take into account terms of order N^{-2} but exclude terms of order smaller than N^{-2}. We let

$$U^*(p) = \bar{U}(p) + \frac{q(p)}{N} \tag{1.45}$$

be a more accurate formula than $\bar{U}(p)$. From equation (1.39) we observe that the corrected diffusion approximation $U^*(p)$ satisfies the d.e.

$$\frac{1}{N} + \frac{1}{2}\mu_2^*(p)\frac{d^2U^*}{dp^2} + \frac{1}{3!}\mu_3^*(p)\frac{d^3U^*}{dp^3} + \frac{1}{4!}\mu_4^*(p)\frac{d^4U^*}{dp^4} = 0 \tag{1.46}$$

where

$$\begin{aligned} \mu_2^*(p) &= \bar{\mu}_2(p) = \frac{p(1-p)}{N} \\ \mu_3^*(p) &= \frac{p(1-p)(1-2p)}{N^2} \\ \mu_4^*(p) &= \frac{3p^2(1-p)^2}{N^2} \end{aligned} \tag{1.47}$$

subject to the boundary conditions:

$$U^*(0) = U^*(1) = 0 \tag{1.48}$$

For convenience, we write

$$\phi(p) = p\ell np + (1-p)\ell n(1-p) \tag{1.49}$$

Then equation (1.45) becomes

$$U^*(p) = -2\phi(p) + \frac{q(p)}{N} \tag{1.50}$$

Now substituting equation (1.50) into equation (1.46) and ignoring

terms of order smaller than N^{-2} , we obtain, after simplification, the following d.e.

$$\frac{6d^2q}{dp^2} = 3p(1-p)\frac{d^4\phi}{dp^4} + 4(1-2p)\frac{d^3\phi}{dp^3} \qquad (1.51)$$

Now equation (1.48) implies that the boundary conditions for the above d.e. are

$$q(0) = q(1) = 0 \qquad (1.52)$$

However, if we solve the d.e. (1.51) with the boundary conditions (1.52), we find that it is impossible to evaluate the two constants introduced by the d.e. Various methods have been employed to overcome this difficulty but they are in general too complicated to be of any use. Perhaps the simplest method is to assume that at extremes values of p i.e. at $p = \frac{1}{N}$ or $p = 1 - \frac{1}{N}$, we have

$$\begin{aligned} U^*(\tfrac{1}{N}) &= \overline{U}(\tfrac{1}{N}) \\ U^*(1-\tfrac{1}{N}) &= \overline{U}(1-\tfrac{1}{N}) \end{aligned} \qquad (1.53)$$

This assumption seems fairly reasonable because at such extreme values of p , the mean absorption time tends to infinity as $N \longrightarrow \infty$, and we would expect that any difference between U^* and $\overline{U}$ at these values would be relatively small compared with any difference between U^* and $\overline{U}$ at moderate values of p . Equation (1.53) then implies that we may modify the boundary conditions for the d.e. (1.51) to

$$q(\tfrac{1}{N}) = q(1-\tfrac{1}{N}) = 0 \qquad (1.54)$$

We now define $q(0) = q(1) = 0$ so that we have $U^*(0) = U^*(1) = 0$. Solving the d.e. (1.51) we obtain

$$6q(p) = -2[(1-p)\ell np + p\ell n(1-p)] + Ap + B \qquad (1.55)$$

where A and B are constants to be determined by the boundary conditions (1.54). Application of equation (1.54) gives

$$A = 0$$

$$B = -2\ell nN + \frac{2\ell nN}{N} + \frac{3}{N}\ell n(1 - \frac{1}{N}) \qquad (1.56)$$

To the order of accuracy required, we may ignore the last two terms in equation (1.56). Hence equation (1.55) becomes

$$q(p) = -\frac{1}{3}[(1-p)\ell np + p\ell n(1-p) + \ell nN] \qquad (1.57)$$

Thus the corrected diffusion approximation for the mean absorption time is

$$U^*(p) = \begin{cases} -2[p\ell np + (1-p)\ell n(1-p)] - \frac{1}{3N}[(1-p)\ell np + p\ell n(1-p) + \ell nN] \\ \qquad\qquad \text{for } \frac{1}{N} \leq p \leq 1 - \frac{1}{N} \\ 0 \qquad \text{for } p = 0,1 . \end{cases} \qquad (1.58)$$

where time is measured in units of N generations. We now compare this result with that obtained by Ewens [1964b] . In our notations, Ewens' result is

$$U^*(p) = -2[p\ell np + (1-p)\ell n(1-p)] - \frac{1}{3N^2}\left[\sum_{j=1}^{k} \frac{(N-k)(N^2-Nj+j^2)}{(N-j)^2 j} + \sum_{j=k+1}^{N-1} \frac{k(N^2-Nj+j^2)}{j^2(N-j)}\right] \qquad (1.59)$$

Writing $x = \frac{j}{N}$ and $p = \frac{k}{N}$, we can approximate the two sums by the following integrals.

$$\sum_{j=1}^{k} \frac{(N-k)(N^2-Nj+j^2)}{(N-j)^2 j} \doteq N(1-p) \int_{\frac{1}{N}}^{p} \frac{1-x+x^2}{x(1-x)^2}\,dx$$
$$\sum_{j=k+1}^{N-1} \frac{k(N^2-Nj+j^2)}{j^2(N-j)} \doteq Np \int_{p}^{1-\frac{1}{N}} \frac{1-x+x^2}{x^2(1-x)}\,dx \qquad (1.60)$$

Evaluating the two integrals and substituting their values into equation (1.59) we get

$$U^*(p) = -2[p\ell np + (1-p)\ell n(1-p)] - \frac{1}{3N}[(1-p)\ell np + p\ell n(1-p) + \ell n\, N + 1] \qquad (1.61)$$

Thus the difference between the two results is essentially the constant term $-\frac{1}{3N}$, which is negligible. Since the two results do not differ significantly, the present method of correcting diffusion approximation is preferred to that of Ewens [1964b] because of its simplicity.

1.5 Mean Absorption time: Moran's model

In this section we consider the Moran's model defined by equation (1.22). Explicit formula for the mean absorption time has been found. Watterson [1961] showed that the mean absorption time m_K in the case $s = 0$ was given by

$$m_K = (N-K)\sum_{j=1}^{K} \frac{N}{N-j} + K\sum_{j=1}^{N-K-1} \frac{N}{N-j} \qquad (1.62)$$

where K was the initial number of A individuals in the population. Ewens [1963b] obtained a more general formula than equation (1.62). He showed that the mean absorption time in the case $s \neq 0$ was given by

(using the same notations as equation (1.22))

$$m_K = \frac{(1+s)^{N-K}-1}{(1+s)^N-1} \sum_{i=1}^{K} \frac{(1+s)^i-1}{s\pi_i} + \frac{(1+s)^N-(1+s)^{N-K}}{(1+s)^N-1} \sum_{i=K+1}^{N-1} \frac{(1+s)^N-(1+s)^i}{\eta_i[(1+s)^N-(1+s)^{N-1}]} \qquad (1.63)$$

Equation (1.62) can be obtained from equation (1.63) by letting $s \longrightarrow 0$.

For simplicity, we shall only consider the diffusion approximation and its corrected value for the mean absorption time in the case $s = 0$, and compare them with equation (1.62). From equation (1.24) with $\mathcal{S} = 0$ (i.e. $\alpha = 0$) ,

$$\mu_{2n+1}(p) = 0 \qquad n = 0,1,2,\ldots$$

$$\mu_{2n}(p) = \frac{1}{N^2}\mu_{2n-2}(p) \qquad n = 2,3,4,\ldots \qquad (1.64)$$

$$\mu_2(p) = \frac{2p(1-p)}{N^2} + 0(N^{-5})$$

We measure time in units of N^2 birth-death events. The d.e. corresponding to equation (1.39) is then

$$\frac{1}{N^2} + \sum_{k=1}^{\infty} \frac{\mu_k(p)}{k!} \frac{d^k U(p)}{dp^k} = 0 \qquad (1.65)$$

where the $\mu_k(p)$ are given by equation (1.64). As a first approximation $\overline{U}(p)$, we ignore terms of order smaller than N^{-2} . Then $\overline{U}(p)$ satisfies the d.e.

$$\frac{1}{N^2} + \frac{1}{2}\mu_2(p)\frac{d^2\overline{U}}{dp^2} = 0 \qquad (1.66)$$

where $\bar{\mu}_2(p) = \frac{2p(1-p)}{N^2}$ subject to the boundary conditions:

$$\bar{U}(0) = \bar{U}(1) = 0 \tag{1.67}$$

The solution to equation (1.66) is

$$\bar{U}(p) = -[p\ell np + (1-p)\ell n(1-p)] \tag{1.68}$$

A more accurate formula than equation (1.68) can be obtained if we now include terms of order N^{-4} but exclude terms of order smaller than this. Let

$$U^*(p) = \bar{U}(p) + \frac{q(p)}{N^2} \tag{1.69}$$

be such a formula where $q(p)$ is the unknown function. Then from equation (1.65), $U^*(p)$ satisfies the d.e.

$$\frac{1}{N^2} + \frac{1}{2}\mu_2^*(p)\frac{d^2U^*}{dp^2} + \frac{1}{4!}\mu_4^*(p)\frac{d^4U^*}{dp^4} = 0 \tag{1.70}$$

subject to the boundary conditions:

$$U^*(0) = U^*(1) = 0 \tag{1.71}$$

Here
$$\mu_2^*(p) = \bar{\mu}_2(p) = \frac{2p(1-p)}{N^2}$$
$$\mu_4^*(p) = \frac{2p(1-p)}{N^4} \tag{1.72}$$

Substituting equations (1.69) and (1.72) into equation (1.70) and using equation (1.66), we obtain, after ignoring terms of order smaller than N^{-4} and simplification, the d.e.

$$\frac{d^2q}{dp^2} = -\frac{1}{12}\frac{d^4\bar{U}}{dp^4} \tag{1.73}$$

Integrating the above d.e. twice, we obtain

$$q(p) = -\frac{1}{12}\frac{d^2\overline{U}}{dp^2} + Ap + B$$

$$= \frac{1}{12p(1-p)} + Ap + B \qquad (1.74)$$

where A and B are constants to be determined by the boundary conditions:

$$q(0) = q(1) = 0 \qquad (1.75)$$

It is obvious that we cannot determine the two constants from the given boundary conditions. Hence some modification is necessary and we suggest the following modified boundary conditions: [see equation (1.54)]

$$q(\tfrac{1}{N}) = q(1 - \tfrac{1}{N}) = 0 \qquad (1.76)$$

Applying the boundary conditions (1.76) to equation (1.74), we get

$$A = 0$$

$$B = -\frac{N}{12(1-1/N)} \qquad (1.77)$$

$$= -\frac{N}{12}\left(1 + \frac{1}{N} + \frac{1}{N^2} + \ldots\right)$$

Because of the order of accuracy, we can only take B to be $-\frac{N}{12} - \frac{1}{12}$. Hence, a more accurate formula than $\overline{U}(p)$ is given by

$$U^*(p) = -[p\ell np + (1-p)\ell n(1-p)] + \frac{1}{N^2}\left[\frac{1}{12p(1-p)} - \frac{1}{12} - \frac{N}{12}\right] \qquad (1.78)$$

We have also considered various other methods of finding the constants. The following is one of them. We know that if m_i

is the mean absorption time where i is the initial number of A individuals in the population, then m_i satisfies the following difference equation

$$m_i = \pi_i m_{i-1} + (1 - \pi_i - \eta_i)m_i + \eta_i m_{i+1} + \frac{1}{N^2} \qquad (1.79)$$

where time is measured in units of N^2 birth-death events. From equation (1.22) with $s = 0$, we have

$$\eta_i = \pi_i = \frac{i(N-i)}{N^2} \qquad (1.80)$$

Hence equation (1.79) becomes

$$m_i = \frac{i(N-i)}{N^2} m_{i-1} + [1 - \frac{2i(N-i)}{N^2}]m_i + \frac{i(N-i)}{N^2} m_{i+1} + \frac{1}{N^2} \qquad (1.80)$$

The above equation is valid for $i = 1,2,\ldots,N$ if we define $m_0 = m_{N+1} = 0$ Putting $i = 1$ in equation (1.80) and simplifying, we get

$$2m_1 - m_2 = \frac{1}{N-1} \qquad (1.81)$$

We shall use the above equation as our boundary condition. Now the solution to the d.e. (1.42) is

$$\overline{U}(p) = -[p\ell np + (1-p)\ell n(1-p)] + Ap + B \qquad (1.82)$$

Instead of using equation (1.43) to determine the constants A and B, [clear that $A = 0$ because of the fact that $\overline{U}(p) = \overline{U}(1-p)$] we use equation (1.81) to determine B i.e. we want to choose B to satisfy the equation

$$2\bar{U}(\tfrac{1}{N}) - \bar{U}(\tfrac{2}{N}) = \frac{1}{N-1} \qquad (1.83)$$

This gives

$$B = \frac{1-2\ell n 2}{N} \qquad (1.84)$$

if we ignore terms of order smaller than N^{-1}. Hence, in this case, we have

$$\bar{U}(p) = -[p\ell n + (1-p)\ell n(1-p)] + \frac{1-2\ell n 2}{N}, \quad \frac{1}{N} \leq p \leq 1 - \frac{1}{N} \qquad (1.85)$$

Now since $U^*(p) = U(p) + \frac{q(p)}{N^2}$ is a more accurate formula than $\bar{U}(p)$ we obviously want $U^*(p)$ to satisfy the equation

$$2U^*(\tfrac{1}{N}) - U^*(\tfrac{2}{N}) = \frac{1}{N-1} \qquad (1.86)$$

Thus this is equivalent to requiring $q(p)$ satisfying the equation

$$2q(\tfrac{1}{N}) - q(\tfrac{2}{N}) = 0 \qquad (1.87)$$

Now in equation (1.74), we know that $A = 0$ because of fact that $U^*(p) = U^*(1-p)$. We now determine B in equation (1.74) by using (1.87). This gives

$$B = -\frac{N}{6} - \frac{1}{12} \qquad (1.88)$$

if we ignore $O(N^{-1})$. Thus the corrected diffusion approximation is then given by

$$U^*(p) = -[p\ell np + (1-p)\ell n(1-p)] + \frac{1}{N^2}\left[\frac{1}{12p(1-p)} - \frac{1}{12} + (1 - 2\ell n 2 - \frac{1}{8})N\right] \qquad (1.89)$$

which differs slightly from equation (1.78). Let us now compare equations (1.78) and (1.89) with equation (1.62). For convenience, write

$$S_j = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{j}$$

Then

$$m_K = N^2[S_N - (1 - \frac{K}{N})S_{N-K} - \frac{K}{N} S_K] \qquad (1.90)$$

$$= N^2[S_N - (1-p)S_{N(1-p)} - pS_{Np}] \quad \text{since} \quad p = \frac{K}{N}$$

where time here is measured in birth-death events. Now using the fact that

$$S_n = 1 + \frac{1}{2} + \ldots + \frac{1}{n} \sim \ell n\, n + \gamma - \frac{1}{2n} + \frac{B_2}{2n^2} - \frac{B_4}{4n^4} - \ldots \qquad (1.91)$$

where γ is Euler's constant and $B_{2\gamma}$ are Bernoullis' numbers ($B_2 = \frac{1}{6}$) [see Knopp, page 538] we can show that

$$S_N - (1-p)S_{N(1-p)} - pS_{Np} \sim \overline{U}(p) - \frac{1}{2N} + \frac{1}{12N^2p(1-p)} - \frac{1}{12N^2} + \ldots \qquad (1.92)$$

Thus we expect $q(p)$ to be given by

$$q(p) = \frac{1}{12p(1-p)} - \frac{1}{12} - \frac{N}{2} \qquad (1.93)$$

However, the "$q(p)$" in equations (1.78) and (1.89) are given respectively by

$$q_1(p) = \frac{1}{12p(1-p)} - \frac{1}{12} - \frac{N}{12} \qquad (1.94)$$

and

$$q_2(p) = \frac{1}{12p(1-p)} - \frac{1}{12} + (1 - 2\ell n2 - \frac{1}{8})N \qquad (1.95)$$

Since $1 - 2\ell n2 - \frac{1}{8} \doteqdot - 0{\cdot}51$, obviously equation (1.89) is a better result than equation (1.78). We do not know any simpler method of overcoming the difficulty about the boundary conditions (1.75). A detailed examination of these boundary conditions would be very desirable.

CHAPTER 2.

Correcting Stationary Distributions.

2.1 Introduction

In this chapter we shall be concerned with finite genetic models that admit stationary distributions. In such models the quantities of interest are obviously the stationary distributions. Once again, exact expressions have not yet been found for these quantities in certain models and it is the purpose of this chapter to examine the accuracy of the diffusion approximations to these exact expressions. Throughout our discussion, we shall assume that under certain conditions it is possible to replace the discrete stochastic process by a continuous stochastic process and that the process is time homogeneous. Also, we shall only be concerned with models in which the change in gene frequency x is Markovian.

2.2 General Theory

Let $f(x)$ be the stationary distribution. Now if $g(\delta x;x)$ is the probability density that the gene frequency changes from x to $x + \delta x$ between two consecutive generations (or birth-death events) given that x is the gene frequency in the first of the two generations (or birth-death events), then

$$f(x) = \int f(x - \delta x)g(\delta x;x - \delta x)d(\delta x) \qquad (2.1)$$

where the integration is taken over all possible values of δx .

Equation (2.1) is a special form of the Kolmogorov-Chapman equation in the Theory of stochastic processes and is a consequence of the Markovian nature of the gene frequency.

Now expanding the integrand in the R.H.S. of equation (2.1) in terms of δx, we have

$$f(x - \delta x)g(\delta x; x - \delta x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} (\delta x)^n \frac{d^n}{dx^n} [f(x)g(\delta x; x)] \tag{2.2}$$

Assuming that we may interchange orders of integration and summation etc, we can express equation (2.1) in the following form:

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \int (\delta x)^n \frac{d^n}{dx^n} [f(x)g(\delta x; x)]d(\delta x) \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [f(x) \int (\delta x)^n g(\delta x; x)d(\delta x)] \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\mu_n(x)f(x)] \end{aligned} \tag{2.3}$$

where we have now written

$$\mu_n(x) = \int (\delta x)^n g(\delta x; x)d(\delta x) \qquad n = 0,1,2,\ldots \tag{2.4}$$

Thus $f(x)$ may be found by solving the d.e. (2.3) subject to the condition

$$\int_0^1 f(x)dx = 1 \tag{2.5}$$

However, in practice, it seems very difficult, if not impossible, to solve the d.e. (2.3) directly. Hence, some approximation is necessary. Generally, if N is the fixed population size considered

in the model, $\mu_n(x)$ are of the orders $N^{-\gamma}$ where γ are some positive integers. For our purpose we may assume that the orders of $\mu_n(x)$ go in pairs in the sense that $\mu_1(x)$ and $\mu_2(x)$ are of the same order, say N^{-m}, where m is some positive integer; $\mu_3(x)$ and $\mu_4(x)$ are of the same but smaller order than N^{-m} and so on. Therefore, for a first approximation $\bar{f}(x)$ to $f(x)$ we may ignore terms of orders smaller than N^{-m}. Thus we ignore $\mu_n(x)$ $(n \geq 3)$. Now $\mu_1(x)$ and $\mu_2(x)$ can be expressed in the form

$$\mu_i(x) = \bar{\mu}_i(x) + \frac{a_i(x)}{N} \quad \frac{b_i(x)}{N^2} + \ldots \qquad (i = 1,2) \qquad (2.6)$$

where $\bar{\mu}_i(x)$ is the term involving only N^{-m} and $\frac{a_i(x)}{N}$ involving only $N^{-(m+1)}$ etc. Then from equation (2.3) we see that $\bar{f}(x)$ satisfies the following d.e. after ignoring terms of orders smaller than N^{-m}

$$\bar{f}(x) = \sum_{n=0}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n(x)\bar{f}(x)] \qquad (2.7)$$

which reduces to

$$\frac{1}{2} \frac{d^2}{dx^2} [\bar{\mu}_2(x)\bar{f}(x)] - \frac{d}{dx} [\bar{\mu}_1(x)\bar{f}(x)] = 0 \qquad (2.8)$$

Integrating equation (2.8) once, we obtain

$$\frac{1}{2} \frac{d}{dx} [\bar{\mu}_2(x)\bar{f}(x)] - \bar{\mu}_1(x)\bar{f}(x) = 0 \qquad (2.9)$$

The constant of integration is zero because the L.H.S. has the interpretation of probability flux. (see Ewens [1963 a]). Hence, equation (2.9) may be called the flux equation of $\bar{f}(x)$. To solve

the above d.e. we multiply through by the integrating factor

$$[\Psi(x)]^{-1} \equiv \exp[-2\int^{x} \frac{\bar{\mu}_1(y)}{\bar{\mu}_2(y)}\,dy] \quad . \tag{2.10}$$

Hence,

$$\bar{f}(x) = \frac{C\Psi(x)}{\bar{\mu}_2(x)} \tag{2.11}$$

where C is a constant such that

$$\int_0^1 \bar{f}(x)dx = 1$$

Thus,

$$\bar{f}(x) = [\int_0^1 \frac{\Psi(x)}{\bar{\mu}_2(x)}dx]^{-1} \frac{\Psi(x)}{\bar{\mu}_2(x)} \tag{2.12}$$

which is the diffusion approximation to the stationary distribution.

We now wish to improve on the diffusion approximation $\bar{f}(x)$. This can be done by taking into account terms of order $N^{-(m+1)}$ [which we have ignored in equation (2.3)] and ignoring terms of orders smaller than this. We let

$$f^*(x) = \bar{f}(x) + \frac{q(x)}{N} \tag{2.13}$$

be a more accurate formula than $\bar{f}(x)$ where $q(x)$ is the unknown function we wish to determine. Now the appropriate d.e. satisfied by the corrected diffusion approximation $f^*(x)$ depends on the orders of the terms $\mu_3(x)$ and $\mu_4(x)$. Since we assume that the orders of $\mu_3(x)$ and $\mu_4(x)$ are equal but smaller than that of $\mu_1(x)$ or $\mu_2(x)$, the orders of $\mu_3(x)$ and $\mu_4(x)$ are either $N^{-(m+1)}$ or smaller than this. Suppose the orders of $\mu_3(x)$ and

$\mu_4(x)$ are smaller than $N^{-(m+1)}$. Then from equation (2.3) we can deduce that $f^*(x)$ satisfies the d.e.

$$\frac{1}{2}\frac{d}{dx}[\mu_2^*(x)f^*(x)] - \mu_1^*(x)f^*(x) = 0 \qquad (2.14)$$

where

$$\mu_i^*(x) = \bar{\mu}_i(x) + \frac{a_i(x)}{N} \quad (i = 1,2) \qquad (2.15)$$

Equation (2.14) is of the same form as equation (2.9). Hence we may interpret equation (2.14) as the corrected flux equation of $f^*(x)$. Now suppose the orders of $\mu_3(x)$ and $\mu_4(x)$ are $N^{-(m+1)}$. We can express $\mu_3(x)$ and $\mu_4(x)$ in the form

$$\mu_i(x) = \mu_i^*(x) + \frac{a_i(x)}{N} + \frac{b_i(x)}{N^2} + \ldots \quad (i = 3,4) \qquad (2.16)$$

where $\mu_i^*(x)$ is the term involving only $N^{-(m+1)}$, $\frac{a_i(x)}{N}$ is the term involving $N^{-(m+2)}$ and so forth. Then from equation (2.3), after ignoring terms of orders smaller than $N^{-(m+1)}$, we observe that $f^*(x)$ satisfies the d.e.

$$\sum_{n=1}^{4} \frac{(-1)^n}{n!}\frac{d^n}{dx^n}[\mu_n^*(x)f^*(x)] = 0 \qquad (2.17)$$

which on integration gives

$$-\mu_1^*(x)f^*(x) + \frac{1}{2}\frac{d}{dx}[\mu_2^*(x)f^*(x)]$$

$$= \frac{1}{3!}\frac{d^2}{dx^2}[\mu_3^*(x)f^*(x)] - \frac{1}{4!}\frac{d^3}{dx^3}[\mu_4^*(x)f^*(x)] \qquad (2.18)$$

The constant of integration is zero because the expression

$$\sum_{n=1}^{4} \frac{(-1)^n}{n!} \frac{d^{n-1}}{dx^{n-1}}[\mu_n^*(x)f^*(x)]$$

may be interpreted as the corrected probability flux. Now without any loss of order of accuracy, we may replace $f^*(x)$ in the R.H.S. of equation (2.18) by $\bar{f}(x)$ to obtain the equation

$$\mu_1^*(x)f^*(x) + \frac{1}{2}\frac{d}{dx}[\mu_2^*(x)f^*(x)] = \phi(x) \qquad (2.19)$$

where $$\phi(x) = \frac{1}{3!}\frac{d^3}{dx^3}[\mu_3^*(x)\bar{f}(x)] - \frac{1}{4!}\frac{d^3}{dx^4}[\mu_4^*(x)\bar{f}(x)] \qquad (2.20)$$

Substituting $f^*(x) = \bar{f}(x) + \frac{q(x)}{N}$ and $\mu_i^*(x) = \bar{\mu}_i(x) + \frac{a_i(x)}{N}$ $(i = 1,2)$ into equation (2.19) and using the fact that $\bar{f}(x)$ satisfies equation (2.9), we obtain the following equation for $q(x)$.

$$\frac{1}{N}\{\frac{1}{2}\frac{d}{dx}[\bar{\mu}_2(x)q(x)] - \bar{\mu}_1(x)q(x)\} = \phi(x) + r(x) + \gamma_1(x) \qquad (2.21)$$

where $$r(x) = \frac{1}{N}\{a_1(x)\bar{f}(x) - \frac{1}{2}\frac{d}{dx}[a_2(x)\bar{f}(x)]\}$$

$$\gamma_1(x) = \frac{1}{N^2}\{a_1(x)q(x) - \frac{1}{2}\frac{d}{dx}[a_2(x)q(x)]\}$$

Now $\gamma_1(x)$ is of smaller order than the other terms and therefore to the order of accuracy required, we may neglect it. Hence equation (2.21) becomes

$$\frac{1}{N}\{\frac{1}{2}\frac{d}{dx}[\bar{\mu}_2(x)q(x)] - \bar{\mu}(x)q(x)\} = \phi(x) + r(x) \qquad (2.22)$$

However, if we were to use equation (2.18) instead of equation (2.19), we would have the extra term

$$\gamma_2(x) \equiv \frac{1}{N}\{\frac{1}{3!}\frac{d^2}{dx^2}[\mu_3^*(x)q(x)] - \frac{1}{4!}\frac{d^3}{dx^3}[\mu_4^*(x)q(x)]\}$$

on the R.H.S. of equation (2.21). $\gamma_1(x)$ and $\gamma_2(x)$ are of the same order and are therefore ignored. Hence in both cases, we obtain the same equation for $q(x)$. Thus, taking into account equations (2.14) and (2.19), we may say that, in general, the corrected diffusion approximation $f^*(x)$ satisfies an equation of the form

$$-\mu_1^*(x)f^*(x) + \frac{1}{2}\frac{d}{dx}[\mu_2^*(x)f^*(x)] = \phi(x) \quad (2.23)$$

where $\phi(x)$ may be interpreted as identically zero or the expression given by equation (2.20). It is therefore sufficient to consider only equation (2.22). Now the solution to equation (2.22) is

$$\frac{q(x)}{N} = \frac{2\Psi(x)}{\bar{\mu}_2(x)}\int^x [\phi(y) + r(y)][\Psi(y)]^{-1}dy + \frac{d\,\Psi(x)}{\bar{\mu}_2(x)} \quad (2.24)$$

where $$\Psi(x) = \exp[2\int^x \frac{\bar{\mu}_1(y)}{\bar{\mu}_2(y)}\,dy]$$

and d is a constant such that

$$\int_0^1 q(x)dx = 0$$

Hence, in models in which the corrected diffusion approximation $f^*(x)$ satisfies equation (2.20), a more accurate formula than $\bar{f}(x)$ is given by

$$f^*(x) = \frac{\Psi(x)}{\bar{\mu}_2(x)}\left[\int_0^1 \frac{\Psi(x)}{\bar{\mu}_2(x)}\,dx\right]^{-1} + \frac{2\Psi(x)}{\bar{\mu}_2(x)}\int^x [\phi(y) + r(y)][\Psi(y)]^{-1}dy + \frac{d\Psi(x)}{\bar{\mu}_2(x)} \quad (2.25)$$

An example of such a model is the Wright's model considered in the next section. Now the case in which $f^*(x)$ satisfies equation (2.14) can be deduced from equation (2.25) by putting

$$\phi(x) \equiv 0 .$$

Thus the corrected diffusion approximation $f^*(x)$ is given by the formula

$$f^*(x) = \frac{\Psi(x)}{\mu_3(x)} \left[\int_0^1 \frac{\Psi(x)}{\mu_2(x)}\, dx \right]^{-1} + \frac{2\Psi(x)}{\mu_2(x)} \int^x r(y)[\Psi(y)]^{-1} dy + \frac{d\Psi(x)}{\mu_2(x)} \qquad (2.26)$$

An example of such a case is Moran's model considered in section 2.4.

2.3 Stationary Distribution: Wright's Model.

In section 1.2 we considered the model introduced by Wright to study the effect of selection on the fluctuations of gene frequencies. The model can be extended to the study of the effect of mutation as well. This is done by incorporating two-way mutation in the model. Suppose α_1 is the probability that A mutates to a and α_2 is the probability that a mutates to A . $(\alpha_i > 0 , \ i = 1,2)$. Then the transition probabilities of the model are given by the matrix

$$P = [p_{ij}] \qquad (i,j = 0,1,2,\ldots,N) \qquad (2.27)$$

where $$p_{ij} = \binom{N}{j} p_i^j (1 - p_i)^{N-j}$$

and $$p_i = \frac{(1+s)i}{N+si} (1 - \alpha_1) + \frac{N-i}{N+si} \alpha_2 \qquad (2.28)$$

It is clear that the number of A individuals between successive generations is still a Markovian variate but the states E_0 and E_N are no longer absorbing states. Thus eventually the process will settle down to equilibrium and from the theory of Markov Chains, the stationary distribution $\underset{\sim}{v}$ is given by the equation

$$\underset{\sim}{v} = \begin{bmatrix} \underset{\sim}{v}'P \\ \\ v_0 \\ v_1 \\ \cdot \\ \cdot \\ \cdot \\ v_N \end{bmatrix} \qquad (2.29)$$

where

However, explicit expressions have not yet been found for v_i and diffusion method is used to obtain approximate values for the stationary distribution. In order to use diffusion methods, we must assume that α_1, α_2 and α are of the order N^{-1}. We write

$$\alpha_i = \frac{\beta_i}{N} \qquad (i = 1,2. \quad \beta_i > 0)$$

$$\alpha = \frac{\alpha}{N} \qquad (\alpha \geq 0)$$

Now it can be shown, after some algebra, that

$$\mu_1(x) = E[x_{t+1} - x_t \mid x_t = x] = \frac{\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2}{N}$$

$$- \frac{\alpha x}{N^2}[\beta_1 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2] + 0(N^{-3})$$

$$\mu_2(x) = \frac{1}{N^2}[\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2][1 + \beta_2 + (\alpha - \beta_1 - \beta_2 - 2)x - \alpha x^2] + \frac{x(1-x)}{N} + O(N^{-3})$$

$$\mu_3(x) = \frac{x(1-x)}{N^2}\{1 + 3\beta_2 + [3(\alpha - \beta_1 - \beta_2) - 2]x - \alpha x^2\} + O(N^{-3}) \qquad (2.30)$$

$$\mu_4(x) = \frac{3x^2(1-x)^2}{N^2} + O(N^{-3})$$

$$\mu_n(x) = O(N^{-3}) \qquad (n \geq 5)$$

Now for a first approximation to the stationary distribution, we ignore terms of order smaller than N^{-1}. Hence the diffusion approximation $\bar{f}(x)$ satisfies the d.e. [see equation (2.9)].

$$-\bar{\mu}_1(x)\bar{f}(x) + \frac{1}{2}\frac{d}{dx}[\mu_2(x)\bar{f}(x)] = 0 \qquad (2.31)$$

where

$$\bar{\mu}_1(x) = \frac{\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2}{N}$$

$$\bar{\mu}_2(x) = \frac{x(1-x)}{N} \qquad (2.32)$$

From equation (2.12) the solution to the above d.e. is given by

$$\bar{f}(x) = \frac{e^{2\alpha x} x^{2\beta_2 - 1}(1-x)^{2\beta_1 - 1}}{\int_0^1 e^{2\alpha x} x^{2\beta_2 - 1}(1-x)^{2\beta_1 - 1} dx} \qquad (2.33)$$

In order to obtain a more accurate formula than equation (2.32) we now include terms of order N^{-2} but exclude terms of order smaller than this. We let

$$f^*(x) = \bar{f}(x) + \frac{q(x)}{N}$$

be the corrected diffusion approximation to equation (2.32). Then

$f^*(x)$ satisfies the d.e. [see equation (2.23)]

$$-\mu_1^*(x)f^*(x) + \frac{1}{2}\frac{d}{dx}[\mu_2^*(x)f^*(x)] = \phi(x) \qquad (2.34)$$

where $$\phi(x) = \frac{1}{3!}\frac{d^2}{dx^2}[\mu_3^*(x)\overline{f}(x)] - \frac{1}{4!}\frac{d^3}{dx^3}[\mu_4^*(x)\overline{f}(x)]$$

$$\mu_1^*(x) = \frac{\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2}{N} - \frac{\alpha x}{N^2}[\beta_1 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2]$$

$$\mu_2^*(x) = \frac{1}{N}[\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2][1 + \beta_2 + (\alpha - \beta_1 - \beta_2 - 2)x - \alpha x^2] + \frac{x(1-x)}{N}$$

$$\mu_3^*(x) = \frac{x(1-x)}{N^2} \cdot \{1 + 3\beta_2 + [3(\alpha - \beta_1 - \beta_2) - 2]x - \alpha x^2\}$$

$$\mu_4^*(x) = \frac{3x^2(1-x)^2}{N^2}$$

The equation for $q(x)$ is then [see equation (2.22)]

$$\frac{1}{N}\left\{\frac{1}{2}\frac{d}{dx}[\bar{\mu}_2(x)q(x)] - \bar{\mu}_1(x)q(x)\right\} = \phi(x) + r(x) \qquad (2.35)$$

where $$r(x) = \frac{1}{N}\left\{a_1(x)\overline{f}(x) - \frac{1}{2}\frac{d}{dx}[a_2(x)\overline{f}(x)]\right\}$$

$$a_1(x) = -\frac{\alpha x}{N}[\beta_1 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2]$$

$$a_2(x) = \frac{1}{N}[\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2][1 + \beta_2 + (\alpha - \beta_1 - \beta_2 - 2)x - \alpha x^2]$$

and $\bar{\mu}_1(x)$, $\bar{\mu}_2(x)$ are the expressions given by equation (2.32). The solution to the above d.e. is given by equation (2.24) and after simplification, we obtain

$$q(x) = \frac{1}{6}\frac{e^{2\alpha x}x^{2\beta_2-1}(1-x)^{2\beta_1-1}\cdot\int^x v(y)y^{-2}(1-y)^{-2}dy}{\int_0^1 e^{2\alpha x}x^{2\beta_2-1}(1-x)^{2\beta_1-1}dx}$$

$$+ Ke^{2\alpha x}x^{2\beta_2-1}(1-x)^{2\beta_1-1} \qquad (2.36)$$

Hence K is a constant such that

$$\int_0^1 q(x)dx = 0$$

and $v(y)$ is a polynomial of the form

$$v(y) = C_0 + C_1y + C_2y^2 + C_3y^3 + C_4y^4 + C_5y^5 + C_6y^6$$

where $C_0 = 5\beta_2 - 10\beta_2^2 + \alpha p_0(\alpha,\beta_1,\beta_2)$

$$C_1 = 34\beta_2^2 + 14\beta_1\beta_2 - 12\beta_2 + \alpha p_1(\alpha,\beta_1,\beta_2)$$

$$C_2 = 9\beta_2 - 3\beta_1 - 42\beta_1\beta_2 - 38\beta_2^2 - 4\beta_1^2 + \alpha p_2(\alpha,\beta_1,\beta_2)$$

$$C_3 = 14\beta_1^2 + 14\beta_2^2 + 28\beta_1\beta_2 - 2\beta_1 - 2\beta_2 + \alpha p_3(\alpha,\beta_1,\beta_2) \qquad (2.37)$$

$$C_4 = \alpha p_4(\alpha,\beta_1,\beta_2)$$

$$C_5 = \alpha^2 p_5(\alpha,\beta_1,\beta_2)$$

$$C_6 = \alpha^3 p_6(\alpha,\beta_1,\beta_2)$$

It is easy to show that

$$p_0(\alpha,\beta_1,\beta_2) = 1$$

$$p_6(\alpha,\beta_1,\beta_2) = 8$$

whereas the other $p_i(\alpha, \beta_1, \beta_2)$ are function of α, β_1 and β_2. Now the formula for $q(x)$ when there is no selection in the model can be obtained from equation (2.36) by putting $\alpha = 0$. Hence, in this case we have

$$q(x) = \frac{1}{6} \cdot \frac{\Gamma(2\beta_1+2\beta_2)}{\Gamma(2\beta_1)\Gamma(2\beta_2)} \cdot x^{2\beta_2-1}(1-x)^{2\beta_1-1} \int^x v(y)y^{-2}(1-y)^{-2}dy + Kx^{2\beta_2-1}(1-x)^{2\beta_1-1} \qquad (2.38)$$

Here $v(y) = C_0 + C_1y + C_2y^2 + C_3y^3$ since $C_4 = C_5 = C_6 = 0$ and

$$C_0 = 5\beta_2 - 10\beta_2^2$$

$$C_1 = 34\beta_2^2 + 14\beta_1\beta_2 - 12\beta_2$$

$$C_2 = 9\beta_2 - 3\beta_1 - 42\beta_1\beta_2 - 38\beta_2^2 - 4\beta_1^2$$

$$C_3 = 14\beta_1^2 + 14\beta_2^2 + 28\beta_1\beta_2 - 2\beta_1 - 2\beta_2$$

It is easy to show that

$$\int^x v(y)y^{-2}(1-y)^{-2}dy = A_1 \ell nx - A_2 \ell n(1-x) - \frac{A_3}{x} \quad \frac{A_4}{1-x}$$

where

$$A_1 = 14\beta_2^2 + 14\beta_1\beta_2 - 2\beta_2$$

$$A_2 = 2\beta_1 - 14\beta_1\beta_2 - 14\beta_1^2$$

$$A_3 = 5\beta_2 - 10\beta_2^2$$

$$A_4 = 10\beta_1^2 - 5\beta_1 - 18\beta_2$$

Hence equation (2.38) becomes

$$q(x) = \frac{1}{6} \cdot \frac{\Gamma(2\beta_1+2\beta_2)}{\Gamma(2\beta_1)\Gamma(2\beta_2)} \cdot x^{2\beta_2-1}(1-x)^{2\beta_1-1}$$
$$\cdot \left[(14\beta_2^2+14\beta_1\beta_2-2\beta_2)\ell nx - (2\beta_1-14\beta_1\beta_2-14\beta_1^2)\ell n(1-x) - \frac{5\beta_2-10\beta_2^2}{x} - \frac{10\beta_1^2-5\beta_1-18\beta_2}{1-x}\right] + Kx^{2\beta_2-1}(1-x)^{2\beta_1-1} \tag{2.39}$$

where K is the constant to be found from the condition

$$\int_0^1 q(x)dx = 0$$

However, we may have

$$\int_0^1 q(x)dx = \infty$$

irrespective of the values of K. In this case it seems "reasonable" to modify the condition to

$$\int_{1/2N}^{1-1/2N} q(x)dx = 0 \tag{2.40}$$

2.4 Moran's Model.

We now introduce two-way mutation into the Moran's model considered in section 1.3. We let α_1 be the probability of A mutating to a and α_2 be the probability of a mutating to A. Then the transition probabilities of this model are given by the matrix

$$P = [p_{ij}] \qquad (i,j = 0,1,2,\ldots,N) \tag{2.41}$$

where

$$p_{i,j-1} = \frac{i}{N}(1 - p_i)$$

$$p_{i,i+1} = \frac{N-i}{N} \cdot p_i$$

$$p_{ii} = 1 - p_{i,i-1} - p_{i,i+1}$$

$$p_{ij} = 0 , \qquad |i - j| \geq 2$$

$$p_i = \frac{i(1+s)}{N+si} \cdot (1 - \alpha_1) + \frac{N-i}{N+si} \cdot \alpha_2$$

Explicit expressions for the stationary distribution are in fact known for Moran's model for any value of s (see Moran [1962] p. 133) The interest in corrected stationary distribution is therefore to see how much closer such corrected distributions are to the true distribution than are the diffusion approximations. (see, for example section 1.3, the corresponding analysis for absorption probabilities in Moran's model when there is no mutation) For the method to be valid, we must have

$$\alpha_i = \frac{\beta_i}{N} \qquad (i = 1,2 \quad \beta_i > 0 \quad \text{and} \quad \beta_i = 0(1))$$

$$s = \frac{\alpha}{N} \qquad (\alpha \geq 0)$$

Now it is easy to show that

$$\mu_1(x) = E[x_{t+1} - x_t \mid x_t = x]$$

$$= \frac{\beta_2+(\alpha -\beta_1-\beta_2)x- \alpha x^2}{N^2} - \frac{\alpha x}{N^3} [\beta_1 + (\alpha -\beta_1-\beta_2)x - \alpha x^2] + 0(N^{-4})$$

$$\mu_2(x) = E[(x_{t+1} - x_t)^2 \mid x_t = x]$$

$$= \frac{2x(1-x)}{N^2} + \frac{(1-2x)}{N^3}[\beta_2 + (\alpha -\beta_1-\beta_2)x - \alpha x^2] + 0(N^{-4})$$

$$\mu_3(x) = \frac{1}{N^3} \cdot \mu_1(x)$$

$$\mu_4(x) = \frac{1}{N^2} \cdot \mu_2(x)$$

Generally, $$\mu_{2n-1}(x) = \frac{1}{N^3} \cdot \mu_{2n-3}(x) \qquad (n \geq 2)$$

$$\mu_{2n}(x) = \frac{1}{N^2} \cdot \mu_{2n-2}(x) \qquad (n \geq 2)$$

As a first approximation to the stationary distribution we ignore terms of order smaller than N^{-2}. Then the diffusion approximation $\bar{f}(x)$ satisfies the d.e.

$$-\bar{\mu}_1(x)\bar{f}(x) + \frac{1}{2}\frac{d}{dx}[\bar{\mu}_2(x)\bar{f}(x)] = 0 \qquad (2.42)$$

where $$\bar{\mu}_1(x) = \frac{\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2}{N^3}$$

$$\bar{\mu}_2(x) = \frac{2x(1-x)}{N^2}$$

whose solution is

$$\bar{f}(x) = \left[\int_0^1 e^{\alpha x} x^{\beta_2 - 1}(1-x)^{\beta_1 - 1} dx\right]^{-1} e^{\alpha x} x^{\beta_2 - 1}(1-x)^{\beta_1 - 1} \qquad (2.43)$$

To improve on equation (2.43) we now include terms of order N^{-3} but exclude terms of order smaller than this. We let

$$f^*(x) = \bar{f}(x) + \frac{q(x)}{N}$$

be a more accurate formula than $\bar{f}(x)$. Then since $\mu_n(x)$ $(n \geq 3)$ are of the order N^{-4}, we find that $f^*(x)$ satisfies the d.e. [see equation (2.14)]

$$- \mu_1^*(x)f^*(x) + \frac{1}{2}\frac{d}{dx}[\mu_2^*(x)f^*(x)] = 0 \qquad (2.44)$$

where $\mu_1^*(x) = \frac{\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2}{N^2} - \frac{\alpha x}{N^3}[\beta_1 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2]$

$$\mu_2^*(x) = \frac{2x(1-x)}{N^2} + \frac{(1-2x)}{N^3}[\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2]$$

Hence from equation (2.24) with $\phi(y) \equiv 0$, we have

$$\frac{q(x)}{N} = \frac{2\Psi(x)}{\bar{\mu}_2(x)}\int^x r(y)[\Psi(y)]^{-1}dy + \frac{d\Psi(x)}{\bar{\mu}_2(x)} \qquad (2.45)$$

where $$r(y) = \frac{1}{N}\{a_1(y)\bar{f}(y) - \frac{1}{2}\frac{d}{dy}[a_2(y)\bar{f}(y)]\} \qquad (2.46)$$

and according to our notations in section 2.2, $a_1(y)$ and $a_2(y)$ are given by

$$\begin{aligned} a_1(y) &= -\frac{ay}{N}[\beta_1 + (\alpha - \beta_1 - \beta_2)y - \alpha y^2] \\ a_2(y) &= \frac{(1-2y)}{N}[\beta_2 + (\alpha - \beta_1 - \beta_2)y - \alpha y^2] \end{aligned} \qquad (2.47)$$

and $$\Psi(x) = e^{\alpha x} x^{\beta_2 - 1} (1-x)^{\beta_1 - 1}.$$

Using equation (2.46) and (2.47) in equation (2.45) and simplifying, we obtain

$$\begin{aligned} q(x) = {} & \left[\int_0^1 e^{\alpha x} x^{\beta_2 - 1}(1-x)^{\beta_1 - 1} dx\right]^{-1} e^{\alpha x} x^{\beta_2 - 1}(1-x)^{\beta_1 - 1} \\ & \frac{1}{2}\{\frac{\beta_2^2}{x} + \frac{\beta_1^2}{1-x} - \frac{(1-2x)[\beta_2 + (\alpha - \beta_1 - \beta_2)x - \alpha x^2]}{x(1-x)} + \alpha(2\beta_1 + 2\beta_2 - \alpha)x \\ & - 2\alpha_2(2\alpha - \beta_1 - \beta_2)\ell n x + [2\beta_1(\beta_1 + \beta_2) + 2\alpha(\beta_1 - \beta_2)]\ell n(1-x)\} \\ & + K e^{\alpha x} x^{\beta_2 - 1}(1-x)^{\beta_1 - 1} \end{aligned} \qquad (2.48)$$

where K is a constant such that

$$\int_0^1 q(x)dx = 0$$

or, if it should be impossible to choose K to satisfy this condition, we use the "reasonable" approximation

$$\int_{1/2N}^{1-1/2N} q(x)dx = 0$$

The formula for $q(x)$ when there is no selection in the model can be found by putting $\alpha = 0$ in the above equation. This gives

$$q(x) = \frac{1}{2}\cdot\frac{\Gamma(\beta_1+\beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)}\cdot x^{\beta_2-1}(1-x)^{\beta_1-1}\cdot\{\frac{\beta_1^2x+\beta_2^2(1-x)-(1-2x)(\beta_2-\beta_1x-\beta_2x)}{x(1-x)}$$

$$+ 2\beta_2^2\,\ell nx + 2\beta_1^2\,\ell n(1-x) + 2\beta_1\beta_2\,\ell nx(1-x)\}$$

$$+ K\,x^{\beta_2-1}(1-x)^{\beta_1-1}. \qquad (2.49)$$

Thus, in the case of no selection, a more accurate formula than

$$\bar{f}(x) = \frac{\Gamma(\beta_1\ \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)}\cdot x^{\beta_2-1}(1-x)^{\beta_1-1} \qquad (2.50)$$

is

$$f^*(x) = \frac{\Gamma(\beta_1+\beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)}\cdot x^{\beta_2-1}(1-x)^{\beta_1-1} + \frac{q(x)}{N} \qquad (2.51)$$

where $q(x)$ is given by equation (2.49). As an example, let $\beta_1 = \beta_2 = 1$. Then equation (2.49) gives

$$q(x) = 2 + 2\ell n\, x(1-x) \qquad (0 < x < 1) \qquad (2.52)$$

Thus a more accurate formula than $\bar{f}(x) = 1$ is

$$f^{*}(x) = 1 + \frac{2+2\ell n\ x(1-x)}{N} \qquad (0 < x < 1) \qquad (2.53)$$

We notice that the correction term $q(x)$ breaks down at $x = 0$ and $x = 1$ but nevertheless, the corrected values are much closer to the true values than are the diffusion approximations as illustrated by numerical examples in Ewens [1964 a].

Chapter 3.

Correcting Absorption Probabilities by using the Pseudo-Transient Function.

3.1 Introduction.

The idea of pseudo-transient distribution was introduced by Ewens [1963a] in his discussion of the behaviour of a variate x $(0 \leq x \leq 1)$ whose density function at time t, $\overline{f}(x;t)$: satisfies the forward Kolmogorov diffusion equation

$$\frac{\partial}{\partial t}\overline{f}(x;t) = \tfrac{1}{2}\frac{\partial^2}{\partial x^2}[\,\bar{\mu}_2(x)\overline{f}(x;t)] - \frac{\partial}{\partial x}[\,\bar{\mu}_1(x)\overline{f}(x;t)] \quad , \qquad (3.1)$$

where $\bar{\mu}_1(x)$ and $\bar{\mu}_2(x)$ are of the form

$$\bar{\mu}_1(x) = x(1-x)p(x)$$

$$\bar{\mu}_2(x) = cx(1-x) \; .$$

Here $p(x)$ is a polynomial (possibly zero or constant) and c is a constant. The boundaries $x = 0$ and $x = 1$ are known as "exit" boundaries and with probability unity, one or the other boundary will be reached in finite time. Thus if the process starts with initial value $x = p$ $(p \neq 0,1)$, then eventually the process will reach either one of the boundaries and once it reaches there, it remains fixed at that boundary. The above process does not admit a non-trivial stationary distribution. However, if we were to consider an associated

"return" process, which is precisely the same as the above process except that once one of the boundaries is reached, the process is immediately re-started with the same initial value $x = p$, then this "return" process does admit a stationary distribution. Ewens called the stationary distribution of the "return" process the pseudo-transient distribution (P.T.D.) because in a sense it describes the transient behaviour of the original non-return process. He showed that the P.T.D. was (for a detailed derivation and discussion of the P.T.D. see Ewens' paper [1963a])

$$\bar{n}(x)[\bar{U}(p)]^{-1} \tag{3.2}$$

where

$$\bar{n}(x) = \begin{cases} \dfrac{2\bar{P}_0(p)\psi(x)}{\bar{\mu}_2(x)} \displaystyle\int_0^x [\psi(y)]^{-1}dy \quad , & 0 \leq x < p \\ \dfrac{2\bar{P}_1(p)\psi(x)}{\bar{\mu}_2(x)} \displaystyle\int_x^1 [\psi(y)]^{-1}dy \quad , & p < x \leq 1 \end{cases} \tag{3.3}$$

Here $\bar{U}(p)$ is the mean absorption time and $\bar{P}_0(p) = 1 - \bar{P}_1(p)$ is the probability that the boundary $x = 0$ is reached eventually. $\psi(x)$ is defined to be

$$\psi(x) = \exp[2\int^x \frac{\bar{\mu}_1(y)}{\bar{\mu}_2(y)}\, dy] \quad . \tag{3.4}$$

The function $\bar{n}(x)$ is called the pseudo-transient function (P.T.F.) and it is clear that

$$\int_{x_1}^{x_2} \bar{n}(x)dx \tag{3.5}$$

represents the mean time the process spends in an arbitrary interval (x_1,x_2) $(0 \leq x_1 < x_2 \leq 1)$ before it is finally absorbed at $x = 0$ or $x = 1$.

P.T.F. $\bar{n}(x)$ will be used to correct absorption probabilities in the two models considered in Chapter 1. A detailed description of the method will be given in the next two sections.

3.2 Moran's Model

We consider the model introduced in section 1.3. The exact probability $P_0(p)$ that eventually the number of A individuals becomes zero is given by the expression

$$P_0(p) = \frac{(1+\frac{\alpha}{N})^{N(1-p)} - 1}{(1+\frac{\alpha}{N})^N - 1}$$

$$= \bar{P}_0(p) + \frac{\alpha^2 e^{\alpha}}{2(e^{\alpha}-1)N}[\bar{P}_0(p) - (1-p)e^{-\alpha p}] + O(N^{-2}) \tag{3.6}$$

where $\bar{P}_0(p)$ is the diffusion approximation to the absorption probability [see equation (1.36)]. We have shown in section 1.3 that the correction term for the diffusion approximation $\bar{P}_0(p)$ is the second term in equation (3.6). We shall now reobtain this result by using the P.T.F.

Let $\bar{n}(x)$ be the diffusion approximation to the P.T.F.

Then $\bar{n}(x)$ satisfies the d.e. [see equation 2(2.8)]

$$\tfrac{1}{2}\frac{d^2}{dx^2}[\bar{\mu}_2(x)\bar{n}(x)] - \frac{d}{dx}[\bar{\mu}_1(x)\bar{n}(x)] = 0 \tag{3.7}$$

where the $\bar{\mu}_i(x)$ are given by equation (1.26). Integrating the above d.e. once, we obtain

$$\tfrac{1}{2}\frac{d}{dx}[\bar{\mu}_2(x)\bar{n}(x)] - [\bar{\mu}_1(x)\bar{n}(x) = \begin{cases} \bar{P}_0(p) & , \quad 0 < x < p \\ -\bar{P}_1(p) & , \quad p < x < 1 \end{cases} \tag{3.8}$$

The R.H.S. takes the values $\bar{P}_0(p)$ and $-\bar{P}_1(p)$ because the L.H.S. has the interpretation of a probability flux. (see Ewens [2963a]) .

Solving the above d.e., we obtain

$$\bar{n}(x) = \begin{cases} \bar{n}_1(x) = 2\dfrac{\bar{P}_0(p)\psi(x)}{\bar{\mu}_2(x)}\displaystyle\int_0^x [\psi(y)]^{-1}dy \quad , & 0 < x < p \\ \bar{n}_2(x) = 2\dfrac{\bar{P}_1(p)\psi(x)}{\bar{\mu}_2(x)}\displaystyle\int_x^1 [\psi(y)]^{-1}dy \quad , & p < x < 1 \end{cases} \tag{3.9}$$

where $\psi(x)$ is defined by equation (3.4). It is easy to verify that $\bar{n}(x)$ is continuous at $x = p$.

We let

$$n^*(x) = \begin{cases} \bar{n}_1(x) + \dfrac{q_1(x)}{N} \quad , & 0 < x < p \\ \bar{n}_2(x) + \dfrac{q_2(x)}{N} \quad , & p < x < 1 \end{cases} \tag{3.10}$$

be a more accurate approximation to the P.T.F. and suppose that the

corrected diffusion approximation $n^*(x)$ satisfies the following corrected flux equation

$$\frac{1}{2}\frac{d}{dx}[[\mu_2^*(x)n^*(x)] - \mu_1^*(x)n^*(x) = \begin{cases} P_0^*(p) , & 0 < x < p \\ -P_1^*(p) , & p < x < 1 \end{cases} \qquad (3.13)$$

where the $\mu_i^*(x) = \bar{\mu}_i(x) + \frac{a_i(x)}{N}$ $(i = 1,2)$ are given by equation (1.30) and $P_0^*(p) = 1 - P_1^*(p) = \bar{P}_0(p) + \frac{h_0(p)}{N}$. Our aim is to find $h_0(p)$. Using equation (3.10) and solving the above d.e. we obtain

$$\frac{q_1(x)}{N} = \frac{2\psi(x)}{\bar{\mu}_2(x)} \int_0^x [\frac{h_0(p)}{N} + r_1(y)][\psi(y)]^{-1}dy , \quad 0 < x < p \qquad (3.12)$$

$$\frac{q_2(x)}{N} = \frac{2\psi(x)}{\bar{\mu}_2(x)} \int_x^1 [-\frac{h_0(p)}{N} + r_2(y)][\psi(y)]^{-1}dy , \quad p < x < 1$$

where $r_i = \frac{1}{N}\{a_1(y)\bar{n}_i(y) - \frac{1}{2}\frac{d}{dy}[a_2(y)\bar{n}_i(y)]\}$ $(i = 1,2)$. Now assuming that $n^*(x)$ is continuous at $x = p$, i.e. assuming that $q_1(p) = q_2(p)$, we get

$$\frac{h_0(p)}{N} = \frac{\int_p^1 r_2(y)[\psi(y)]^{-1}dy - \int_0^p r_1(y)[\psi(y)]^{-1}dy}{\int_0^1 [\psi(y)]^{-1}dy}$$

and after simplification we obtain

$$h_0(p) = \frac{\alpha^2 e^{\alpha}}{2(e^{\alpha}-1)}[\bar{P}_0(p) - (1 - p)e^{-\alpha p}] \qquad (3.14)$$

which is identical to equation (1.34). The above method and example

are due to Ewens [1964a] .

We now show that the above method is in fact equivalent to the method used in section 1.3. This explains why we obtain the same result in both cases. Consider equation (3.11). The solution to that equation can be shown to be

$$n^*(x) = \begin{cases} \dfrac{2P_0^*(p)\psi^*(x)}{\mu_2^*(x)} \displaystyle\int_0^x [\psi^*(y)]^{-1}dy , & 0 < x < p \\ \dfrac{2P_1^*(p)\psi^*(x)}{\mu_2^*(x)} \displaystyle\int_x^1 [\psi^*(y)]^{-1}dy , & p < x < 1 \end{cases} \tag{3.15}$$

where

$$\psi^*(x) = \exp[2\int^x \frac{\mu_1^*(y)}{\mu_2^*(y)}\, dy\ ] \tag{3.16}$$

The assumption that $n^*(x)$ is continuous at $x = p$ implies that

$$P_0^*(p) = \frac{\int_p^1 [\psi^*(x)]^{-1}dx}{\int_0^1 [\psi^*(x)]^{-1}dx} \tag{3.17}$$

which is the solution of the following d.e.

$$\mu_1^*(p)\frac{dP_0^*(p)}{dp} + \tfrac{1}{2}\,\mu_2^*(p)\frac{d^2P_0^*(p)}{dp^2} = 0 \tag{3.18}$$

subject to the conditions:

$$P_0^*(0) = 1 \quad \text{and} \quad P_0^*(1) = 0$$

We know that the diffusion approximation $\overline{P}_0(p)$ satisfies the d.e. [see equation (1.25)]

$$\mu_1(p)\frac{dP_0}{dp} + \tfrac{1}{2}\mu_2(p)\frac{d^2\bar{P}_0}{dp^2} = 0 \qquad (3.19)$$

and that the corrected diffusion approximation $P_0^*(p)$ actually satisfies equation (3.18). [see equation (1.29)] . Thus in finding the corrected diffusion approximation $P_0^*(p)$ where we employ the method of correcting $\bar{n}(x)$ and assuming the continuity of $n^*(x)$ at $x = p$, we are actually saying that $P_0^*(p)$ satisfies the d.e. (3.18). Hence it is not surprising that we obtained the same correction term. However, if we apply the same analysis to the Wright's model in section 1.2, we will get a correction term which differs slightly from the one given in equation (1.21). Ewens [1964a] conjectured that the reason is that the corrected diffusion approximation $n^*(x)$ is not continuous at $x = p$. We now examine the analysis of the Wright's model.

3.3 Wright's Model.

The diffusion approximation $\bar{n}(x)$ to the true P.T.F. is given by equation (3.9) where here $\mu_1(x)$, $\mu_2(x)$ and $\bar{P}_0(p)$ are the expressions given by equations (1.10) and (1.12) respectively. We now assume that the corrected diffusion approximation

$$n^*(x) = \begin{cases} \bar{n}_1(x) + \dfrac{q_1(x)}{N} & , \quad 0 < x < p \\ \bar{n}_2(x) + \dfrac{q_2(x)}{N} & , \quad p < x < 1 \end{cases} \qquad (3.20)$$

satisfies the following corrected flux equation [see equation (2.19)]

$$- \mu_1^*(x)n^*(x) + \tfrac{1}{2}\frac{d}{dx}\,[\mu_2^*(x)n^*(x)] - \frac{1}{3!}\frac{d^2}{dx^2}\,[\mu_3^*(x)\bar{n}(x)]$$

$$+ \frac{1}{4!}\frac{d^3}{dx^3}\,[\mu_4^*(x)\bar{n}(x)]$$

$$= \begin{cases} P_0^*(p) , & 0 < x < p \\ -P_1^*(p) , & p < x < 1 \end{cases} \qquad (3.21)$$

Here $\mu_i^*(x) = \bar{\mu}_i(x) + \frac{a_i(x)}{N}$ $(i = 1,2)$, $\mu_3^*(x)$ and $\mu_4^*(x)$ are given by equation (1.16) and

$$P_0^*(p) = \bar{P}_0(p) + \frac{h_0(p)}{N} \qquad (3.22)$$

where $h_0(p)$ is the unknown function we wish to determine. Using equations (3.20) and (3.22) in equation (3.21) and the fact that $\bar{n}(x)$ satisfies equation (3.8), we obtain, after simplification

$$\frac{1}{N}\{\tfrac{1}{2}\frac{d}{dx}[\bar{\mu}_2(x)q_1(x)] - \bar{\mu}_1(x)q_1(x)\}$$

$$= \frac{h_0(p)}{N} + y_1(x) + r_1(x) + \gamma_1(x) + \beta_1(x) , \quad 0 < x < p$$

$$\frac{1}{N}\{\tfrac{1}{2}\frac{d}{dx}[\bar{\mu}_2(x)q_2(x)] - \bar{\mu}_1(x)q_2(x)\}$$

$$= -\frac{h_0(p)}{N} + y_2(x) + r_2(x) + \gamma_2(x) + \beta_2(x) , \quad p < x < 1 \qquad (3.23)$$

where

$$y_i(x) = \frac{1}{3!}\frac{d^2}{dx^2}\,[\mu_3^*(x)\bar{n}_i(x)] - \frac{1}{4!}\frac{d^3}{dx^3}\,[\mu_4^*\bar{n}_i(x)] \qquad (i = 1,2)$$

$$r_i(x) = \frac{1}{N}\{a_1(x)\bar{n}_i(x) - \tfrac{1}{2}\frac{d}{dx}\,[a_2(x)\bar{n}_i(x)]\} \qquad (i = 1,2)$$

$$\gamma_i(x) = \frac{1}{N^3}\{a_1(x)q_i(x) - \tfrac{1}{2}\frac{d}{dx}[a_2(x)q_i(x)]\} \qquad (i = 1,2)$$

$$\beta_i(x) = \frac{1}{N}\{\frac{1}{3!}\frac{d^2}{dx^2}[\mu_3^*(x)q_i(x)] - \frac{1}{4!}\frac{d^3}{dx^3}[\mu_4^*(x)q_i(x)] \qquad (i = 1,2)$$

Now $\gamma_i(x)$ and $\beta_i(x)$ are of smaller order terms and can therefore be ignored. Hence equation (3.23) becomes

$$\frac{1}{N}\{\tfrac{1}{2}\frac{d}{dx}[\bar{\mu}_2(x)q_i(x)] - \bar{\mu}_1(x)q_i(x)\} = y_i(x) + r_i(x) + \begin{cases} \dfrac{h_o(p)}{N} & \text{for } i = 1 \\ -\dfrac{h_o(p)}{N} & \text{for } i = 2 \end{cases} \qquad (3.24)$$

Solving the above d.e. we obtain

$$\frac{q_1(x)}{N} = \frac{2\psi(x)}{\bar{\mu}_2(x)}\int_0^x [y_1(t) + r_1(t) + \frac{h_o(p)}{N}][\psi(t)]^{-1}\,dt \,, \qquad 0 < x < p$$

$$\frac{q_2(x)}{N} = \frac{2\psi(x)}{\bar{\mu}_2(x)}\int_0^1 [y_2(t) + r_2(t) - \frac{h_o(p)}{N}][\psi(t)]^{-1}\,dt \,, \qquad p < x < 1 \qquad (3.25)$$

To determine $h_o(p)$ we assume that $n^*(x)$ is continuous at $x = p$ i.e. we assume that $q_1(p) = q_2(p)$. Hence we obtain

$$\frac{h_o(p)}{N} = \frac{\int_p^1 [y_2(t) + r_2(t)][\psi(t)]^{-1}dt - \int_o^p [y_1(t) + r_1(t)][\psi(t)]^{-1}dt}{\int_o^1 [\psi(t)]^{-1}dt} \qquad (3.26)$$

Evaluation of equation (3.26) gives

$$h_0(p) = \frac{\alpha}{(e^{-\alpha}-1)}\left[\tfrac{1}{3}(-p + 2\alpha p + \alpha p^2)e^{-2\alpha p} - \tfrac{1}{3}(1 - \overline{P}_0(p))(-1 + 3\alpha)e^{-2\alpha}\right] \qquad (3.27)$$

which differs slightly from the correction term given in equation (1.21). Thus it is clear that if we were to solve equation (3.21) directly and obtain $P_0^*(p)$ by assuming the continuity of $n^*(x)$ at $x = p$, the corrected diffusion approximation $P_0^*(p)$ so obtained will not satisfy the d.e. [see equation (1.14)]

$$\mu_1^*(p)\frac{dP_0^*}{dp} + \tfrac{1}{2}\mu_2^*(p)\frac{d^2P_0^*}{dp^2} + \frac{1}{3!}\mu_3^*(p)\frac{d^3P_0^*}{dp^3} + \frac{1}{4!}\mu_4^*(p)\frac{d^4P_0^*}{dp^4} = 0$$

with the boundary conditions:

$$P_0^*(0) = 1 \quad \text{and} \quad P_0^*(1) = 0 \ .$$

Hence either our assumption that the corrected diffusion approximation $n^*(x)$ satisfies equation (3.21) or our assumption that $n^*(x)$ is continuous at $x = p$, (or both) must be false. However, we now believe that it is the first assumption that is incorrect. The correct equations should be derived in the following manner.

We know that if $f(x)$ is the stationary distribution and $g(\delta x:x)$ is the probability density that the gene frequency changes from x to $x + \delta x$ between two consecutive generations given that x is the gene frequencey in the first of the two generations, then

$$f(x) = \int f(x - \delta x)g(x:x - \delta x)d(\delta x) \qquad (3.28)$$

where integration is over all possible values of δx . Since the P.T.D. is a stationary distribution, the above equation is still valid. Now let

$$f(x) = \begin{cases} f_1(x) , & 0 < x < p \\ f_2(x) , & p < x < 1 \end{cases} \tag{3.29}$$

be the P.T.D. Then for $0 < x < p_2$ we have

$$f_1(x) = \int_{\delta x \geq x-p} f_1(x - \delta x)g(\delta x; x - \delta x)d(\delta x)$$

$$+ \int_{\delta x \geq x-p} f_2(x - \delta x)g(\delta x; x - \delta x)d(\delta x) \tag{3.30}$$

Now expanding $f_i(x - \delta x)g(\delta x; x - \delta x)$ in terms of δx , we have

$$f_i(x - \delta x)g(\delta x; x - \delta x) = \sum_{n=0}^{\infty} (-1)^n \frac{(\delta x)^n}{dx^n} \frac{d^n}{dx^n} [f_i(x)g(\delta x; x)]$$

$$(i = 1,2) \tag{3.31}$$

Substituting equation (3.31) into equation (3.30) and assuming that we may interchange the orders of integration and summation, we obtain

$$f_1(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\mu_n^+(x)f_1(x)]$$

$$+ \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\mu_n^-(x)f_2(x)] \tag{3.32}$$

where we have now written

$$\mu_n^+(x) = \int_{\delta x \geq x-p} (\delta x)^n g(\delta x : x)\, d(\delta x) , \quad n = 0,1,2\ldots$$

$$\mu_n^-(x) = \int_{\delta x < x-p} (\delta x)^n g(\delta x : x)\, d(\delta x) , \quad n = 0,1,2\ldots \tag{3.33}$$

We observe that

$$\mu_n(x) = \mu_n^+(x) + \mu_n^-(x) \tag{3.34}$$

[for the definition of $\mu_n(x)$ see equation (2,4)] and

$$\mu_0^+(x) = 1 - \mu_0^-(x) = \Pr(\delta x \geq x - p) = 1 - \Pr(\delta x < x - p) \tag{3.35}$$

We assume that for $n \geq 3$, $\mu_n^+(x)$ and $\mu_n^-(x)$ are small compared with μ_i^+ and $\mu_i^-(x)$ $(i = 1,2)$ so that we may ignore them and that to the order of accuracy required, we may replace μ_i^+ and $\mu_i^-(x)$ $(i = 1,2)$ by $\bar{\mu}_i^+(x)$ and $\bar{\mu}_i^-(x)$. Thus from equation (3.32), the approximation $\bar{f}_i(x)$ to $f_i(x)$ $(i = 1,2)$ satisfies the equation

$$\bar{f}_1(x) = \mu_0^+(x)\bar{f}_1(x) + \mu_0^-(x)\bar{f}_2(x) + \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^+(x)\bar{f}_1(x)]$$

$$+ \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^-(x)\bar{f}_2(x)] \tag{3.36}$$

which, after using equation (3.35), reduces to

$$0 = [\bar{f}_2(x) - \bar{f}_1(x)]\Pr(\delta x < x - p) + \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^+(x)\bar{f}_1(x)]$$

$$+ \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^-(x)\bar{f}_2(x)] \tag{3.37}$$

Similarly, [by replacing $f_1(x)$ in the L.H.S. of equation (3.30) by $f_2(x)$] we have for $p < x < 1$,

$$0 = [\bar{f}_1(x) - \bar{f}_2(x)]\,\Pr(\delta x \geq x - p) + \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^{+}(x)\bar{f}_1(x)] + \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^{-}(x)\bar{f}_2(x)] \tag{3.38}$$

Thus the diffusion approximation

$$\bar{n}(x) = \begin{cases} \bar{n}_1(x) , & 0 < x < p \\ \bar{n}_2(x) , & p < x < 1 \end{cases}$$

should satisfy the following equations:

$$0 = [\bar{n}_2(x) - \bar{n}_1(x)]\Pr(\delta x < x - p) + \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^{+}(x)\bar{n}_1(x)] + \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^{-}(x)\bar{n}_2(x)] ,$$

$$\text{for } 0 < x < p . \tag{3.39}$$

and

$$0 = [\bar{n}_1(x) - \bar{n}_2(x)]\Pr(\delta x \geq x - p) + \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^{+}(x)\bar{n}_1(x)] + \sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n^{-}(x)\ \bar{n}_2(x)] ,$$

$$\text{for } p < x < 1 . \tag{3.40}$$

and not equations of the form

$$\sum_{n=1}^{2} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\bar{\mu}_n(x)\bar{n}_i(x)] = 0 \quad (i = 1,2) , \tag{3.41}$$

which we had used to obtain $\bar{n}(x)$. We observe that equation (3.41) can be obtained from equations (3.39) and (3.40) by putting

$$\Pr(\delta x < x - p) = 0 \tag{3.42}$$

and

$$\Pr(\delta x \geq x - p) = 0 \quad . \tag{3.43}$$

For putting $\Pr(\delta x < x - p) = 0$ in equation (3.39) implies that values of $x < p$ can only be reached from values of $x \leq p$ and not from values of $x > p$. This implies

$$\mu_n^-(x) = \int_{\delta x < x-p} (\delta x)^n g(\delta x : x) d(\delta x) = 0 \quad .$$

Hence

$$\mu_n(x) = \mu_n^+(x)$$

Similarly, equation (3.43) implies that

$$\mu_n^+(x) = 0$$

and therefore

$$\mu_n(x) = \mu_n^-(x) \quad .$$

Now the correct equations satisfied by the corrected diffusion approximation

$$n^*(x) = \bar{n}(x) + \frac{q(x)}{N} \tag{3.44}$$

$$= \begin{cases} n_1^*(x) = \bar{n}_1(x) + \dfrac{q_1(x)}{N} , & 0 < x < p \\ n_2^*(x) = \bar{n}_2(x) + \dfrac{q_2(x)}{N} , & p < x < \end{cases}$$

can be derived similarly from equation (3.32) and they are:

$$[n_2^*(x) - n_1^*(x)]\Pr(\delta x < x - p) + \sum_{n=1}^{4} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\mu_n^{*+}(x) n_1^*(x)] + \sum_{n=1}^{4} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\mu_n^{*-}(x) n_2^*(x)] = 0 \tag{3.45}$$

$$[n_1^*(x) - n_2^*(x)]\Pr(\delta x \geq x - p) + \sum_{n=1}^{4} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\mu_n^{*+}(x) n_1^*(x)] + \sum_{n=1}^{4} \frac{(-1)^n}{n!} \frac{d^n}{dx^n} [\mu_n^{*-}(x) n_2^*(x)] = 0 . \tag{3.46}$$

Thus our error lies in the assumption that values of $x < p$ can only be reached from values of $x \leq p$ and not from values of $x > p$ and that values of $x > p$ can only be reached from values of $x > p$ and not from values of $x \leq p$. This is true in the Moran's model, which is why we obtain the correct answer for the correction term. However, in practice, it seems very difficult, if not impossible, to solve equations (3.32), (3.40), (3.45) and (3.46).

Chapter 4

The maintenance of alleles by mutation

4.1 Introduction

In this section we shall discuss briefly some of the results of Kimura and Crow [1964] and Ewens [1964c] . We shall show later how we could extend Ewens' result to the case when the population consists of two types of alleles. A generalisation to the case of m types (series) of alleles will also be given.

In their papers, the above-named authors have investigated quantitatively an extreme case of a system in which "the wild type allele is not a single entity, but rather a population of different isoalleles that are indistinguishable from each other by any ordinary procedure". They assume that all new alleles which arise by mutation are different from any new allele which exists or has existed in the population. Both approaches are entirely different in that Ewens, using diffusion approximation and the so-called Wright's model to describe the behaviour of the newly-formed alleles, [see equation (4.15)] showed that the mean number $\bar{n}$ of alleles maintained in a diploid population of size N with mutation rate u of the order N^{-1} was given, in the case of selectively neutral genes, approximately by

$$\bar{n} = 4Nu\int_{1/2N}^{1} x^{-1}(1-x)^{4Nu-1}dx \qquad (4.1)$$

whereas Kimura and Crow were interested in the quantity called the

and $A_i \neq A_j$ $a_i \neq a_j$ for $j \neq i$. Hence, if at any generation, the population consists of r different alleles of type A, denoted by $A_1, A_2, \ldots, A_r$ and k different alleles of type a represented by say, $a_1, a_2, \ldots, a_k$, then if any allele of type A, A_m $(1 \leq m \leq r)$ mutates, it mutates to either A_{r+1} or a_{k+1}

Define $D_A(t)$ to be the number of different alleles of type A in generation t. $D_a(t)$ is defined similarly. $D_A(t)$ and $D_a(t)$ are random variables. Let $\bar{n}_A(t)$ and $\bar{n}_a(t)$ be the mean values of the two random variables. Then our problem is to find the values of $\bar{n}_A$ and $\bar{n}_a$, which are defined to be the limits of $\bar{n}_A(t)$ and $n_a(t)$ as t tends to infinity. In other words we are interested in finding the mean values of the numbers of different alleles of the two types A and a when the process is in equilibrium.

Suppose at equilibrium, N_A and N_a are the numbers of alleles of types A and a respectively in the population. Then we expect on the average, $u(1-\pi)N_A + v\lambda N_a$ new alleles of type A to arise per generation by the mutation, where π and λ are defined by equations (4.5) and (4.8). Since at equilibrium these new alleles must be balanced by a similar number of old alleles of type A being lost by drift or mutation, we obtain, if $\bar{t}_A$ is defined to be the mean number of generations that any new allele of type A exists in the population before being lost, the relation

$$\frac{\bar{n}_A}{\bar{t}_A} = uN_A - u\pi N_A + v\lambda N_a \qquad (4.3)$$

Similarly, we have

$$\frac{\bar{n}_a}{\bar{t}_a} = vN_a - v\lambda N_a + u\pi N_A \qquad (4.4)$$

Hence if N_A , N_a , $\bar{t}_A$ and $\bar{t}_a$ are known, the values of $\bar{n}_A$ and $\bar{n}_a$ follow immediately.

To find N_A and N_a , we define $N_A(t)$ to be the number of alleles of type A in the t-th generation. $N_a(t)$ is defined similarly. We also require the following definitions. Suppose an allele of type A , say A_m , mutates and that $1 - \pi$ $(0 \leq \pi \leq 1)$ is the probability that it mutates to new allele of type A and π is the probability that it mutates to new allele of type a . Then since the mutation rate of A is u , the probability that any allele A mutating to new allele A , denoted by $P(A \longrightarrow \text{new } A)$, is given by

$$P(A \longrightarrow \text{new } A) = u(1 - \pi) \qquad (4.5)$$

and the probability that any allele A mutating to new allele of type a is given by

$$P(A \longrightarrow \text{new } a) = u\pi \qquad (4.6)$$

Similarly, we suppose that

$$P(a \longrightarrow \text{new } a) = v(1 - \lambda) \qquad (4.7)$$

$$P(a \longrightarrow \text{new } A) = v\lambda \qquad (4.8)$$

where $0 \leq \lambda \leq 1$. Then the expected numbers of alleles of types

A and a in the next generation are given respectively by the following equations:

$$E[N_A(t+1)] = N_A(t) + v\lambda N_a(t) - u\pi N_A(t) \tag{4.9}$$

$$E[N_a(t+1)] = N_a(t) + u\pi N_A(t) - v\lambda N_a(t) \tag{4.10}$$

Now

$$N_A = \lim_{t\to\infty} N_A(t) = \lim_{t\to\infty} N_A(t+1)$$

$$N_a = \lim_{t\to\infty} N_a(t) = \lim_{t\to\infty} N_a(t+1)$$

Assuming that $\lim_{t\to\infty} E[N_A(t+1)] = E[\lim_{t\to\infty} N_A(t+1)] = E[N_A] = N_A$ etc, and letting $t \longrightarrow \infty$ in the above equations, we obtain

$$N_A = \frac{v\lambda}{u\pi} N_a \tag{4.11}$$

Since $N_A + N_a = 2N$, we have

$$N_A = \frac{2v\lambda N}{v\lambda+u\pi} \tag{4.12}$$

$$N_a = \frac{2u\pi N}{v\lambda+u\pi} \tag{4.13}$$

It now remains to find expressions for $\bar{t}_A$ and $\bar{t}_a$. In order to find expressions for these quantities we have to set up a model to describe the behaviour of any newly formed allele. We denote by $X_{A_i}(t)$ the proportion of the alleles of type A_i in the population at generation t. The notation $X_{a_j}(t)$ is given a similar interpretation. Suppose that

$$\sum_{i=1}^{r} X_{A_i}(t) + \sum_{j=1}^{k} X_{a_j}(t) = 1$$

Let $X(t) = (X_{A_1}(t), X_{A_2}(t), \ldots, X_{A_r}(t), X_{a_1}(t), \ldots, X_{a_k}(t))$

The model then assumes that given $X(t)$, the distribution of $2NX(t+1)$ is multinomial. For example, if in the t-th generation there are:

i_m alleles of type A_m $(m = 1,2,\ldots,r)$

and

j_m alleles of type a_m $(m = 1,2,\ldots,k)$

in the population where

$$\sum_{m=1}^{r} i_m + \sum_{m=1}^{k} j_m = 2N \ .$$

and in the $(t+1)$ generation there are:

i'_m alleles of type A_m $(m = 1,2,\ldots,r)$

and

j'_m alleles of type a_m $(m = 1,2,\ldots,k)$

and α, β, γ and δ new mutants in the population where

α is the number of new A's derived by mutation from A

β is the number of new a's derived by mutation from A

γ is the number of new a's derived by mutation from a

δ is the number of new A's derived by mutation from a

such that

$$\sum_{m=1}^{r} i_m' + \sum_{m=1}^{k} j_m' + \alpha + \beta + \gamma + \delta = 2N .$$

Then the transition probability of going from the state of i_1A_1's,...,i_rA_r's , j_1a_1's,...,j_ka_k's to the state of $i_1'A_1$'s,...,$i_r'A_r$'s,.,$j_1'a_1$'s,...,$j_k'a_k$'s ; α , β , γ , δ which we denote by $P(i,j \longrightarrow i',j';\ \alpha,\beta,\gamma,\delta)$ is given by

$$P(i,j \longrightarrow i',j';\ \alpha,\beta,\gamma,\delta) = \frac{(2N)!}{\alpha!\ \beta!\ \gamma!\ \delta!\ \prod_{m=1}^{r} i_m''!\ \prod_{m=1}^{k} j_m''!}$$

$$\prod_{m=1}^{r} \left[\frac{i_m(1-u)}{2N}\right]^{i_m'} \prod_{m=1}^{k} \left[\frac{j_m(1-v)}{2N}\right]^{j_m'}$$

$$\left[\frac{i_1+\ldots+i_r}{2N} \cdot u(1-\pi)\right]^{\alpha} \left[\frac{i_1+\ldots+i_r}{2N} \cdot u\pi\right]^{\beta}$$

$$\left[\frac{j_1+\ldots+j_k}{2N} \cdot v(1-\lambda)\right]^{\gamma} \left[\frac{j_1+\ldots+j_k}{2N} \cdot v\lambda\right]^{\delta} . \qquad (4.14)$$

We observe that $\{X(t)\}$ is a discrete Markov process and also that each $\{X_{A_i}(t)\}$ or $\{X_{a_j}(t)\}$ is a Markov process. (Note that the latter result is not true if selection operates, so that the present approach cannot be used in such circumstances.) Since the distribution of $2NX(t+1)$ given $X(t)$ is multinomial, the distribution of $2NX_{A_i}(t+1)$ given $X_{A_i}(t)$ is therefore binomial and the probability $p_{i_m i_m'}$ that the number of genes of the allele of type A_i changes from i_m to i_m' in consecutive generations is given by

$$p_{i_m i'_m} = \binom{2N}{i'_m} \left[\frac{i_m(1-u)}{2N}\right]^{i'_m} \left[\frac{2N - i_m(1-u)}{2N}\right]^{2N-i'_m} \qquad (4.15)$$

In the above equation, the decrease $-i_m u$ is due to mutation to new alleles in the next generation. Since the initial value of any newly-formed allele is necessarily unity, the Markov chain characterised by equation (4.15) and the above initial condition are sufficient to determine the value of $\bar{t}_{A_i}$, which is equal to $\bar{t}_A$.

Now the Markov chain characterised by equation (4.15) is often classified under the heading of Wright's model. As mentioned in Chapter 1, exact expressions for many quantities of interest in such model are usually unknown and in practice, it seems very difficult to find exact expression for $\bar{t}_A$. Hence some approximation is necessary and the approximation used is to replace the discrete Markov process by a continuous diffusion process. The approximation is valid provided u is of the order N^{-1}. Ewens [1964c] showed that the diffusion approximation $\bar{t}_A^*$ to $\bar{t}_A$ was given by the expression

$$\bar{t}_A^* = 2\int_{1/2N}^{1} x^{-1}(1-x)^{4Nu-1}dx \quad . \qquad (4.16)$$

By exactly the same argument we can show that the diffusion approximation to $\bar{t}_a$ is given by

$$\bar{t}_a^* = 2\int_{1/2N}^{1} x^{-1}(1-x)^{4Nv-1}dx \quad . \qquad (4.17)$$

Hence, using equations (4.3), (4.4), (4,12), (4.13), (4.16)

(4.17), we show that the mean numbers of different alleles of types A and a maintained in a diploid population of fixed size N with mutation rates of A and a equal to u and v respectively in the case where u and v are of the order N^{-1}, are given approximately by

$$\bar{n}_A = \frac{4uv\lambda N}{v\lambda+u\pi}\int_{1/2N}^{1} x^{-1}(1-x)^{4Nu-1}dx \qquad (4.18)$$

$$\bar{n}_a = \frac{4uv\pi N}{v\lambda+u\pi}\int_{1/2N}^{1} x^{-1}(1-x)^{4Nv-1}dx \qquad (4.19)$$

Now, as a special case, suppose $u = v$, $\lambda = 1$, and $\pi = 0$. Then the above equations reduce to

$$\bar{n}_A = 4Nu\int_{1/2N}^{1} x^{-1}(1-x)^{4Nu-1}dx \qquad (4.20)$$

$$\bar{n}_a = 0$$

which agree with Ewen's result (equation (4.1)).

Intuitively, we know that as N tends to infinity, $\bar{n}_A$ and $\bar{n}_a$ both tend to infinity. However, it might be of interest to know the rate of divergence of $\bar{n}_A$ and $\bar{n}_a$ i.e. we like to know how fast the two quantities $\bar{n}_A$ and $\bar{n}_a$ tend to infinity. To investigate the rate of divergence, we examine the integrals in equations (4.18) and (4.19). It can be shown that if terms of order N^{-1} are ignored, we have

$$\int_{1/2N}^{1} x^{-1}(1-x)^{4Nu-1}dx = \ell n\ 2N + \sum_{r=1}^{\infty}\frac{(-1)^r}{r}\binom{4Nu-1}{r} \qquad (4.21)$$

$$\int_{1/2N}^{1} x^{-1}(1-x)^{4Nv-1}dx = \ell n\ 2N + \sum_{r=1}^{\infty} \frac{(-1)^r}{r} \binom{4Nv-1}{r} \tag{4.22}$$

Hence equations (4.18) and (4.19) become

$$\bar{n}_A = \frac{4uv\lambda N}{v\lambda+u\pi}\left[\ell n\ 2N + \sum_{r=1}^{\infty} \frac{(-1)^r}{r} \binom{4Nu-1}{r}\right] \tag{4.23}$$

$$\bar{n}_a = \frac{4uv\pi N}{v\lambda+u\pi}\left[\ell n\ 2N + \sum_{r=1}^{\infty} \frac{(-1)^r}{r} \binom{4Nv-1}{r}\right] \tag{4.24}$$

Now suppose for given u and v, $4Nu-1$ and $4Nv-1$ are some positive integers. Then the two series

$$\sum_{r=1}^{\infty} \frac{(-1)^r}{r} \binom{4Nu-1}{r} \quad \text{and} \quad \sum_{r=1}^{\infty} \frac{(-1)^r}{r} \binom{4Nv-1}{r}$$

terminate. Hence as $N \longrightarrow \infty$, we obtain the following asymptotic results from equations (4.23) and (4.24)

$$\bar{n}_A \sim \frac{4uv\lambda N}{v\lambda+u\pi} \ell n\ 2N = 2uN_A \ell n\ 2N \tag{4.25}$$

$$\bar{n}_a \sim \frac{4uv\pi N}{v\lambda+u\pi} \ell n\ 2N = 2vN_a \ell n\ 2N \tag{4.26}$$

Since u and v are of the order N^{-1}, we observe that $\bar{n}_A$ and $\bar{n}_a$ are therefore of the order $\ell n\ N$. Thus the rate of divergence of these two quantities is rather "slow". If we interpret $\frac{\bar{n}_A}{\bar{n}_A+\bar{n}_a}$ as the proportion of the mean number of different alleles of type A in the population and with a similar interpretation for

$\frac{\bar{n}_a}{\bar{n}_A+\bar{n}_a}$, then we have

$$\frac{\bar{n}_A}{\bar{n}_A+\bar{n}_a} \sim \frac{uN_A}{uN_A+vN_a} \tag{4.27}$$

$$\frac{\bar{n}_a}{\bar{n}_A+\bar{n}_a} \sim \frac{vN_a}{uN_A+vN_a} \tag{4.28}$$

In fact when the mutation rates of A and a are equal i.e. when $u = v$, we observe that the proportion of the mean number of different alleles of type A is asymptotically equal to the equilibrium frequency $\frac{N_A}{2N}$ of the allele A and that the proportion of the mean number of different alleles of type a in the population is asymptotically equal to the equilibrium frequency $\frac{N_A}{2N}$ of the allele a .

Roughly speaking, we can say that [see equation (4.25)] for large and fixed value of N , the mean number $\bar{n}_A$ of different alleles of type A in the equlibrium population is a bilinear function of the mutation rate u of the allele A and the number of alleles of type A in the population in the sense that if N_A is kept fixed, then $\bar{n}_A$ increases linearly with respect to u and if u is kept constant, then $\bar{n}_A$ increases linearly with respect to N_A . Similarly, we can say that $\bar{n}_a$ is a bilinear function of v and N_a . [See equation (4.26)] From equations (4.12) and (4.13) we see that for given u and v , N_A and N_a are functions of λ and π . In fact N_A is monotonic increasing with respect to λ . Since $N_A + N_a = 2N$, N_a decreases correspondingly with increasing values

of λ . However, N_a is monotonic increasing with respect to π . Hence, increase in λ implies increase in $\bar{n}_A$ and decrease in $\bar{n}_a$ whereas increase in π implies increase in $\bar{n}_a$ and decrease in $\bar{n}_A$. All the above agree with what would be expected on common-sense ground. We illustrate some of the above observations with two numerical examples.

Example 1

Suppose $2N = 10^6$, $u = 10^{-6}$, $v = 4 \times 10^{-6}$, $\lambda = 0{\cdot}1$ and $\pi = 0{\cdot}1$. Then

$$N_A = \frac{2v\lambda N}{v\lambda + u\pi} = 0{\cdot}8 \times 10^6$$

Now for increasing values of u , N_A can be kept constant by decreasing the values of π i.e. by keeping $u\pi$ fixed.

The following table shows the corresponding values of $\bar{n}_A$ and u where N_A is kept constant and equal to $0{\cdot}8 \times 10^6$.

Table 1.

u	10^{-6}	2×10^{-6}	3×10^{-6}	4×10^{-6}
$\bar{n}_A$	19·71	38·72	55·34	71·88

We observe that for fixed N_A , $\bar{n}_A$ increases somewhat less than linearly with u .

Example 2.

The following table shows the values of $\bar{n}_A$ and $\bar{n}_a$ for various values of λ and π with $2N = 10^6$, $u = 10^{-6}$ and $v = 4 \times 10^{-6}$.

Table 2.

u	10^{-6}	10^{-6}	10^{-6}	10^{-6}
v	4×10^{-6}	4×10^{-6}	4×10^{-6}	4×10^{-6}
λ	0·1	0·1	0·5	0·5
π	0·1	0·5	0·1	0·5
N_A	$0{\cdot}8 \times 10^6$	$\frac{4}{9} \times 10^6$	$\frac{20}{21} \times 10^6$	$0{\cdot}8 \times 10^6$
N_a	$0{\cdot}2 \times 10^6$	$\frac{5}{9} \times 10^6$	$\frac{1}{21} \times 10^6$	$0{\cdot}2 \times 10^6$
$\bar{n}_A$	19·71	10·95	23·47	19·71
$\bar{n}_a$	17·97	49·91	4·28	17·97

From columns 1 and 2, we notice that increase in π implies increase in $\bar{n}_a$ and decrease in $\bar{n}_A$. Columns 2 and 3 show that increase in λ implies increase in $\bar{n}_A$ and decrease in $\bar{n}_a$. Columns 1 and 4 show that the values of $\bar{n}_A$ and $\bar{n}_a$ are identical. This is obvious since simultaneous increase or decrease in values of λ and π in such a way that N_A and N_a remain constant does not alter the values of $\bar{n}_A$ and $\bar{n}_a$.

We define the two functions $f_A(x)$ and $f_a(x)$ by

$$f_A(x) = \frac{4uv\lambda N x^{-1}(1-x)^{4Nu-1}}{v\lambda + u\pi} \quad , \quad \frac{1}{2N} \leqq x \leqq 1 \tag{4.29}$$

$$f_a(x) = \frac{4uv\pi N x^{-1}(1-x)^{4Nv-1}}{v\lambda + u\pi} \quad , \quad \frac{1}{2N} \leqq x \leqq 1 \quad . \tag{4.30}$$

Then we have

$$\overline{n}_A = \int_{1/2N}^{1} f_A(x)dx \tag{4.31}$$

$$\overline{n}_a = \int_{1/2N}^{1} f_a(x)dx \tag{4.32}$$

Now if $1/2N \leqq x_1 < x_2 \leqq 1$, then it is interesting to know what interpretation one can give to the expression

$$\overline{n}_A(x_1,x_2) = \int_{x_2}^{x_1} f_A(x)dx \tag{4.33}$$

and

$$\overline{n}_a(x_1,x_2) = \int_{x_1}^{x_2} f_a(x)dx \quad . \tag{4.34}$$

However, if one remembers that the mean number of generations for which a given allele of type A having frequency in the range (x_1,x_2) before being lost is

$$2\int_{x_1}^{x_2} x^{-1}(1-x)^{4Nu-1}dx$$

generations, then the expression $\overline{n}_A(x_1,x_2)$ represents the mean number of different alleles of type A having frequency between

x_1 and x_2. Similarly, $\bar{n}_a(x_1,x_2)$ may be interpreted to be the mean number of different alleles of type a which have frequency between x_1 and x_2.

We observe from equations (4.29) and (4.30) that the functions $f_A(x)$ and $f_a(x)$ are essentially of the form

$$f(x) = kx^{-1}(1-x)^{\alpha}, \qquad \frac{1}{2N} \leq x \leq 1 \tag{4.35}$$

where k is a positive constant and α, a constant which may be positive or negative depending upon the given mutation rates u and v, and is negative if either u or v is small enough. A detailed discussion of the above function is already given in Ewens' paper [1964c] with obvious modification and we shall therefore not discuss it here.

Now the diffusion approximations to the equilibrium frequencies of the alleles A and a, which are equal to $\frac{v\lambda}{v\lambda+u\pi}$ and $\frac{u\pi}{v\lambda+u\pi}$ respectively, are given by

$$\int_{1/2N}^{1} x f_A(x)dx = \frac{v\lambda}{v\lambda+u\pi} + O(N^{-1}) \tag{4.36}$$

$$\int_{1/2N}^{1} x f_a(x)dx = \frac{u\pi}{v\lambda+u\pi} + O(N^{-1}) \tag{4.37}$$

The small error is due to the approximation of the discrete Markov process by the continuous diffusion process. The expression

$$\int_{1/2N}^{1} x^2[f_A(x) + f_a(x)]dx = \frac{v\lambda}{v\lambda+u\pi} \cdot \frac{1}{4Nu+1} + \frac{u\pi}{v\lambda+u\pi} \cdot \frac{1}{4Nv+1} + O(N^{-1}) \tag{4.38}$$

is the diffusion approximation to the coefficient of inbreeding or the probability that an individual chosen at random be homozygous. (c.f. equation (10) of Ewens' paper [1964c])

4.3 A generalisation.

It is now a simple matter to generalise our results to the case when the diploid population of fixed size N consists of m types of alleles $A_1, A_2, \ldots, A_m$. Suppose that the mutation rate of A_i is u_i $(i = 1,2,\ldots,m)$. We assume that all new mutants are different from any allele which exists or has existed in the population and that these newly-formed alleles belong to one of the m series (or groups):

$$A_i = \{A_{i1}, A_{i2}, \ldots, A_{ij}, \ldots\} \qquad (i = 1,2,\ldots m) .$$

so that if an allele of type A_i, say A_{ir}, mutates, it mutates to some new allele A_{ij} of the same type with probability π_{ii} and that it mutates to new allele of type A_j $(i \neq j)$ with probability π_{ij} with $\sum_{j=1}^{m} \pi_{ij} = 1$. Hence the probability that any allele of type A_i mutating to new allele of type A_j is given by the expression

$$P(A_i \longrightarrow \text{new } A_j) = u_i \pi_{ij} , \quad (i,j = 1,2,\ldots,m) .$$

At equilibrium, let N_{A_i} be the number of alleles of type A_i so that $\sum_{i=1}^{m} N_{A_i} = 2N$.

Define $\bar{n}_{A_i}$ to be the mean number of different alleles of type A_i in each generation and $\bar{t}_{A_i}$ to be the mean number of generations that any new allele of type A_i exists in the population before being lost. Then the balancing equations which are analogous to equations (4.3) and (4.4) are

$$\frac{\bar{n}_{A_i}}{\bar{t}_{A_i}} = \sum_{j=1}^{m} N_{A_j} u_j \pi_{ji} \quad (i = 1,2,\ldots,m) \tag{4.39}$$

Hence, if expressions for $\bar{t}_{A_i}$ and N_{A_i} can be found, the values of $\bar{n}_{A_i}$ follows immediately. Now by similar argument given in the previous section, we can approximate t_{A_i} by the diffusion process value $\bar{t}^*_{A_i}$ in the case when u_i is of the order N^{-1}. Hence

$$\bar{t}^*_{A_i} = 2\int_{1/2N}^{1} x^{-1}(1-x)^{4Nu_i - 1}\,dx . \tag{4.40}$$

In order to find N_{A_i} we observe that by balancing the number of new mutants

$$u_i N_{A_i} = \sum_{j=1}^{m} u_j \pi_{ji} N_{A_j} \quad (i = 1,2,\ldots,m) \tag{4.41}$$

Since

$$\sum_{i=1}^{m} u_i N_{A_i} = \sum_{i=1}^{m} \sum_{j=1}^{m} u_j \pi_{ji} N_{A_j}$$

$$= \sum_{j=1}^{m} u_j N_{A_j} \sum_{i=1}^{m} \pi_{ji}$$

$$= \sum_{j=1}^{m} u_j N_{A_j} ,$$

there are in fact only $m - 1$ distinct equations and without loss of generality we can take the $m - 1$ distinct equations to be

$$\sum_{\substack{j=1 \\ j\neq 1}}^{m} u_j \pi_{ji} N_{A_j} - u_i(1 - \pi_{ii}) N_{A_i} = 0 \quad (i = 1,2,\ldots,m-1) . \qquad (4.42)$$

The other equation is provided by

$$N_{A_1} + N_{A_2} + \ldots + N_{A_m} = 2N \qquad (4.43)$$

Thus we have m independent equations in m unknowns $(N_{A_1}, N_{A_2}, \ldots N_{A_m})$ and therefore in principle we are able to solve them. Rewriting the m equation in matrix notation we have

$$\begin{bmatrix} -u_1(1-\pi_{11}) & u_2\pi_{21} & \cdots & u_{m-1}\pi_{m-1,1} & u_m\pi_{m1} \\ u_1\pi_{12} & -u_2(1-\pi_{22}) & \cdots & u_{m-1}\pi_{m-1,2} & u_m\pi_{m2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ u_1\pi_{1i} & u_2\pi_{2i} & -u_i(1-\pi_{ii}) & u_{m-1}\pi_{m-1,i} & u_m\pi_{mi} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ u_1\pi_{1,m-1} & u_2\pi_{2,m-1} & \cdots & -u_{m-1}(1-\pi_{m-1,m-1}) & u_m\pi_{m-1,m-1} \\ 1 & 1 & \cdots & 1 & 1 \end{bmatrix} \begin{bmatrix} N_{A_1} \\ N_{A_2} \\ \cdots \\ \cdots \\ N_{A_3} \\ \cdots \\ \cdots \\ N_{A_{m-1}} \\ N_{A_m} \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 2N \end{bmatrix}$$

or more briefly in the form

$$\Lambda \underset{\sim}{x} = \underset{\sim}{b} \qquad (4.44)$$

where Λ is the required matrix,

$$\underset{\sim}{x} = \begin{bmatrix} N_{A_1} \\ \cdot \\ \cdot \\ \cdot \\ N_{A_m} \end{bmatrix}$$

$$\underset{\sim}{b} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 2N \end{bmatrix}$$

Since the determinant $|\Lambda| \neq 0$, we obtain on solving the above matrix equation

$$\underset{\sim}{x} = \Lambda^{-1}\,\underset{\sim}{b} \tag{4.45}$$

where $$\Lambda^{-1} = \frac{\text{adj } \Lambda}{|\Lambda|} .$$

Now

$$\text{adj } \Lambda = \begin{bmatrix} c_{11} & c_{21} & \cdots & c_{m1} \\ c_{12} & c_{22} & \cdots & c_{m2} \\ c_{1m} & c_{2m} & \cdots & c_{mm} \end{bmatrix}$$

where c_{ij} are the co-factors of $|\Lambda|$. Clearly

$$|\Lambda| = \sum_{i=1}^{m} c_{mi} .$$

Hence we obtain

$$N_{A_i} = \frac{2c_{mi}N}{\sum_{i=1}^{m} c_{mi}} \quad (i = 1,2,\ldots,m) \tag{4.46}$$

Therefore the mean number of different alleles of type A_i maintained in the diploid population of size N with mutation rate u_i to entirely new alleles in the case u_i of order N^{-1} , is given by

$$\overline{n}_{A_i} = \frac{4u_i c_{mi} N}{\sum_{i=1}^{m} c_{mi}} \int_{1/2N}^{1} x^{-1}(1-x)^{4Nu_i - 1}\,dx \quad (i = 1,2,\ldots,m) \tag{4.47}$$

Example 1.

Suppose $m = 3$. Then our matrix equation is

$$\begin{bmatrix} -u_1(1-\pi_{11}) & u_2\pi_{21} & u_3\pi_{31} \\ u_1\pi_{12} & -u_2(1-\pi_{22}) & u_3\pi_{33} \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} N_{A_1} \\ N_{A_2} \\ N_{A_3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2N \end{bmatrix}$$

and

$$c_{31} = u_2u_3[\pi_{21}\pi_{32} + \pi_{31}(1-\pi_{22})]$$

$$c_{32} = u_2u_3[(1-\pi_{11})\pi_{32} + \pi_{12}\pi_{31}]$$

$$c_{33} = u_1u_2[(1-\pi_{11})(1-\pi_{22})-\pi_{12}\pi_{21}]$$

and the values of $\bar{n}_{A_i}$ follow immediately from equation (4.47).

Chapter 5

Initial Progress of New Gametes

5.1 Introduction

In population genetic models the determination of non-trivial equilibrium points and the conditions for their stability are usually of great interest to the population geneticists because these play a vital role in the theory of evolution. However, an equally important aspect is the conditions for the establishment of new gametes (genes) and a great deal of research has also been done in this (Bodmer and Parsons [1960], [1962]).

Now it is well known in the two-locus theory that closer linkage is generally beneficial in the establishment of new gametes. In this chapter we propose to examine whether there are any circumstances under which "loose" linkage is beneficial in increasing the frequencies of new gametes.

5.2 The two-locus theory.

We consider the two-locus case, each with two alleles A, a and B , b respectively. Suppose that the two loci are linked and R is the coefficient of recombination between the two loci where R lies between 0 and $\frac{1}{2}$ inclusive. The case $R = 0$ corresponds to complete linkage and the case $R = \frac{1}{2}$ corresponds to non-linkage between the two loci.

There are 4 gametes which can be formed with A , a and B , b .

These are AB , ab , Ab and aB . Let c_1 , c_2 , c_3 , c_4 be the frequencies of the above gametes respectively. $(\Sigma_{i=1}^{4} c_i = 1)$

Denote the ten genotypes and their fitnesses by the following array:

	AB	ab	Ab	aB
AB	ω_{11}	ω_{12}	ω_{13}	ω_{14}
ab		ω_{22}	ω_{23}	ω_{24}
Ab			ω_{33}	ω_{34}
aB				ω_{44}

so that, for example,

ω_{12} is the fitness of the coupling heterozygote $\frac{AB}{ab}$

ω_{24} is the fitness of the repulsion heterozygote $\frac{Ab}{aB}$.

We assume that $\omega_{12} \neq \omega_{24}$ in general. The mean fitness of the population $\overline{W}$ is then given by

$$\overline{W} = \sum_{i,j} \omega_{ij} c_i c_j \quad .$$

By considering the frequency with which each genotype produces any particular gamete and combining the results from all the genotypes, we can obtain the expected gametic frequencies in one generation in terms of those in the previous generation. Let c_1' , c_2' , c_3' and c_4' be the gametic frequencies of the next generation $(\Sigma_{i=1}^{4} c_i' = 1)$. Then with the aid of the following table:

Table 1

zygotic genotype	fitness	c_1	c_2	c_3	c_4
AB/AB	ω_{11}	1	0	0	0
ab/ab	ω_{22}	0	1	0	0
Ab/Ab	ω_{33}	0	0	1	0
aB/aB	ω_{44}	0	0	0	1
AB/aB	ω_{14}	$\frac{1}{2}$	0	0	$\frac{1}{2}$
ab/Ab	ω_{23}	0	$\frac{1}{2}$	$\frac{1}{2}$	0
ab/aB	ω_{24}	0	$\frac{1}{2}$	0	$\frac{1}{2}$
AB/ab	ω_{12}	$\frac{1}{2}(1-R)$	$\frac{1}{2}(1-R)$	$\frac{1}{2}R$	$\frac{1}{2}R$
Ab/aB	ω_{34}	$\frac{1}{2}R$	$\frac{1}{2}R$	$\frac{1}{2}(1-R)$	$\frac{1}{2}(1-R)$
Aβ/Ab	ω_{13}	$\frac{1}{2}$	0	$\frac{1}{2}$	0

we can set up the following four equations

$$\begin{aligned}
\overline{W}c_1' &= \omega_{11}c_1^2 + \omega_{13}c_1c_3 + \omega_{14}c_1c_4 + (1-R)\omega_{12}c_1c_2 + R\omega_{34}c_3c_4 \\
\overline{W}c_2' &= \omega_{22}c_2^2 + \omega_{23}c_2c_3 + \omega_{24}c_2c_4 + (1-R)\omega_{12}c_1c_2 + R\omega_{34}c_3c_4 \\
\overline{W}c_3' &= \omega_{33}c_3^2 + \omega_{13}c_1c_3 + \omega_{23}c_2c_3 + R\omega_{12}c_1c_2 + (1-R)\omega_{34}c_3c_4 \\
\overline{W}c_4' &= \omega_{44}c_4^2 + \omega_{14}c_1c_4 + \omega_{24}c_2c_4 + R\omega_{12}c_1c_2 + (1-R)\omega_{34}c_3c_4
\end{aligned} \qquad (5.1)$$

If we write

$$W_i = \omega_{i1}c_1 + \omega_{i2}c_2 + \omega_{i3}c_3 + \omega_{i4}c_4 \qquad (i = 1,2,3,4)$$

then equation (5.1) can be expressed as

$$\bar{W}_i' = W_i c_i \pm R(\omega_{34} c_3 c_4 - \omega_{12} c_1 c_2) \quad (i = 1,2,3,4) \tag{5.2}$$

where "+" sign is used for $i = 1,2$ and the "-" sign for $i = 3,4$. We may interpret W_i $(i = 1,2,3,4)$ as the mean fitnesses of the gametes AB , ab , Ab and aB respectively. Equations (5.1) are the fundamental equations used in deriving the non-trivial equilibrium points of the population. We shall not discuss these equations at all. A detailed discussion of these equations is given by Bodmer and Parson [1962] and Moran [1964].

5.3 Conditions for the increase of gametes AB and aB .

Suppose locus A - a is kept polymorphic by the superiority in fitness of the heterozygote Aa while the allele b is predominant at the second locus. We wish to investigate the conditions under which the gametes AB and aB will eventually increase. To do this we assume that initially the frequencies of these gametes are small and that before the introduction of B , genotypes $\frac{ab}{ab}$, $\frac{Ab}{Ab}$ and $\frac{Ab}{ab}$ were at equilibrium so that the frequencies of ab and Ab are given by (c.f. Bodmer and Parsons [1960])

$$c_2 = p = \frac{\omega_{23}-\omega_{33}}{2\omega_{23}-\omega_{22}-\omega_{33}} \quad (\omega_{23} > \omega_{33})$$

$$c_3 = q = 1 - p = \frac{\omega_{23}-\omega_{22}}{2\omega_{23}-\omega_{22}-\omega_{33}} \quad (\omega_{23} > \omega_{22}) \; . \tag{5.3}$$

Suppose now there exists a small quantity of the allele B so that

$$\begin{aligned} c_1 &= \delta x_1 \\ c_2 &= p - \delta p \\ c_3 &= q - \delta q \\ c_4 &= \delta x_4 \end{aligned} \tag{5.4}$$

where δx_1 , δx_4 , δp and δq are some small positive numbers such that $\delta x_1 + \delta x_4 = \delta p + \delta q$. The mean fitness W of the equilibrium population before the introduction of B is given by

$$W = \omega_{22}p^2 + \omega_{33}q^2 + 2pq\omega_{23} \tag{5.5}$$

and it is easy to show that

$$\overline{W} = W + \text{small order terms.}$$

Now using equations (5.4) and ignoring second order terms like δx_1^2 , δx_4^2 , $\delta p \delta q$ etc, we obtain

$$\begin{aligned} c_1' &= W^{-1}(\omega_{12}p + \omega_{13}q - R\omega_{12}p)\delta x_1 + W^{-1}R\omega_{34}q\delta x_4 \\ c_4' &= W^{-1}R\omega_{12}p\delta x_1 + W^{-1}(\omega_{24}p + \omega_{34}q - R\omega_{34}q)\delta x_4 \end{aligned} \tag{5.6}$$

Expressing the above equations in matrix notation, we get

$$\begin{bmatrix} c_1' \\ c_4' \end{bmatrix} = M \begin{bmatrix} \delta x_1 \\ \delta x_4 \end{bmatrix}$$

where

$$M = \begin{bmatrix} W^{-1}(W_1 - R\omega_{12}p) & W^{-1}R\omega_{34}q \\ W^{-1}R\omega_{12}p & W^{-1}(W_4 - R\omega_{34}q) \end{bmatrix}$$

$$W_1 = \omega_{12}p + \omega_{13}q$$

$$W_4 = \omega_{24}p + \omega_{34}q \quad .$$

Now the gametes AB and aB will eventually increase in frequency if and only if the larger eigenvalue of M is greater than unity. The equation for the eigenvalues of M is

$$\lambda^2 - W^{-1}[W_1 + W_4 - R(\omega_{12}p + \omega_{34}q)] + W^{-2}[W_1W_4 - R(W_1\omega_{34}q + W_4\omega_{12}p)] = 0 \qquad (5.8)$$

whose roots λ_1 , λ_2 are given by

$$\lambda_1, \lambda_2 = (2W)^{-1}[W_1 + W_4 - R(\omega_{12}p + \omega_{34}q) \pm \sqrt{[W_1 + W_4 - R(\omega_{12}p + \omega_{34}q)]^2 - 4W_1W_4 + 4R(W_1\omega_{34}q + W_4\omega_{12}p)}] \quad . \qquad (5.9)$$

It is easy to verify that

$$\sqrt{[W_1 + W_4 - R(\omega_{12}p + \omega_{34}q)]^2 - 4W_1W_4 + 4R(W_1\omega_{34}q + W_4\omega_{12}p)}$$

$$= \sqrt{[W_1 - W_4 + R(\omega_{34}q - \omega_{12}p)]^2 + 4R^2\omega_{12}\omega_{34}pq}$$

and the expression under the square root on the R.H.S. is non-negative and hence the roots λ_1, λ_2 are real. Thus the larger eigenvalue is

$$\lambda_1 = (2W)^{-1}[W_1 + W_4 - R(\omega_{12}p + \omega_{34}q) + \sqrt{[W_1 + W_4 - R(\omega_{12}p + \omega_{34}q)]^2 - 4W_1W_4 + 4R(W_1\omega_{34}q + W_4\omega_{12}p)}] \qquad (5.10)$$

Now $\lambda_1 > 1$ if and only if

$$\sqrt{[W_1 + W_4 - R(\omega_{12}p + \omega_{34}q)]^2 - 4W_1W_4 + 4R(W_1\omega_{34}q + W_4\omega_{12}p)}$$

$$> 2W - (W_1 + W_4) + R(\omega_{12}p + \omega_{34}q) \quad . \tag{5.11}$$

Hence the solution to our problem reduces to solving the inequality (5.11). For simplicity we first assume that the double heterozygotes $\frac{AB}{ab}$ and $\frac{Ab}{aB}$ have the same fitness i.e. that $\omega_{12} = \omega_{34}$. Then the inequality (5.11) becomes

$$\sqrt{[W_1 + W_4 - R\omega_{12}]^2 - 4W_1W_4 + 4R\omega_{12}(W_1q + W_4p)}$$

$$> 2W - (W_1 + W_4) + R\omega_{12} \quad . \tag{5.12}$$

To solve the above the inequality, we observe that for given W , W_1 , W_4 we have either

$$\text{(i)} \quad 2W - (W_1 + W_4) < 0$$

$$\text{or} \quad \text{(ii)} \quad 2W - (W_1 + W_4) \geq 0 \quad .$$

Suppose that (i) is true. Then it is obvious that inequality (5.12) holds if R is sufficiently small since the L.H.S. of (5.12) is always non-negative. In fact inequality holds if

$$R < \frac{W_1+W_4-2W}{\omega_{12}} \quad . \tag{5.13}$$

We observe from equation (5.13) that if $W_1 + W_4 - 2W > \frac{1}{2}\omega_{12}$, then irrespective of the values of R , the frequencies of the gametes AB and aB will eventually increase. Intuitively, we expect that if

$W_1 > W$, $W_4 > W$, then the frequencies of the gametes AB and aB will eventually increase irrespective of the values of R . However, this may not be so because if we could have say,

$$W_1 = W + \varepsilon_1 \omega_{12}$$

$$W_4 = W + \varepsilon_4 \omega_{12}$$

where $\varepsilon_1, \varepsilon_4 > 0$

and such that $\varepsilon_1 + \varepsilon_4 < \frac{1}{4}$,

then $W_1 + W_4 - 2W = (\varepsilon_1 + \varepsilon_4)\omega_{12}$.

Thus for the gametes AB and aB to increase eventually, we must have

$$R < \frac{W_1 + W_4 - 2W}{\omega_{12}} = \varepsilon_1 + \varepsilon_4 < \frac{1}{4}$$

Now suppose that (ii) is true. Then since both sides of the inequality (5.11) are non-negative, we may square them and on simplification, we obtain

$$R\omega_{12}(W - qW_1 - pW_4) < -(W - W_1)(W - W_4) \qquad (5.14)$$

Then clearly in some cases, λ_1 is greater than unity if and only if R is sufficiently small. In fact if the following conditions are satisfied:

(1) $W > W_1$ and $W < W_4$ or $W < W_1$ and $W > W_4$

(2) $W > qW_1 + pW_4$

then $\lambda_1 > 1$ if and only if

$$R < R_0 = \frac{-(W-W_1)(W-W_4)}{\omega_{12}(W-qW_1-pW_4)} \qquad (5.15)$$

Remeber that W_1 and W_4 are the mean fitnesses of the gametes AB and aB respectively and hence we deduce that if one of the mean fitnesses of the gametes AB, aB is greater than the mean fitness W of the equilibrium population before the introduction of B while the other mean fitness (W_1 or W_4) is less than W and if

$$W > \max[\tfrac{1}{2}(W_1 + W_4) , qW_1 + pW_4]$$

and the coefficient of linkage between the two loci is tighter than the value given by R_0, then the frequencies of the gametes AB and aB will eventually increase.

We observe that if W is greater than W_1 and W is greater than W_4, then inequality (5.14) cannot be satisfied by any value of R. In other words this says that if the mean fitness of the equilibrium population before the introduction of B is greater than the mean fitnesses of the gametes AB, aB, then irrespective of the values of R, the gametes AB and aB will not increase in frequency. This agrees with what would be expected on common-sense ground.

We have shown that under certain circumstances, closer linkage is beneficial in the establishment of new gametes. However, it is of interest to know whether there are any circumstances under which "loose" linkage is beneficial in establishing new gametes.

This is equivalent to asking the question whether equation (5.14) can ever assume the form

$$R > k$$

where k is some positive number. From equation (5.14) we observe that this would require

$$\begin{aligned} &(1) \quad W - qW_1 - pW_4 < 0 \\ &(2) \quad (W - W_1)(W - W_4) > 0 \ . \end{aligned} \tag{5.16}$$

Jointly, this requires the condition:

$$W_1 > W \text{ and } W_4 > W \ . \tag{5.17}$$

But this contradicts our assumption that $2W \geq W_1 + W_4$ since equation (5.17) implies $2W < W_1 + W_4$. Hence there are no situations in which increased values of R are beneficial in increasing the new gametes when the fitnesses of the coupling heterozygote and the repulsion heterozygote are equal.

We now investigate whether the above conclusion is still valid if we allow the double heterozygotes to have different fitnesses i.e. we assume that $\omega_{12} \neq \omega_{34}$. To investigate this we go back to equation (5.11) which is the condition for the new gametes to increase. The condition is that λ_1 is greater than unity if and only if

$$\sqrt{[W_1 + W_4 - R(\omega_{12}p + \omega_{34}q)]^2 - 4W_1W_4 + 4R(W_1\omega_{34}q + W_4\omega_{12}p)}$$

$$> 2W - (W_1 + W_4) + R(\omega_{12}p + \omega_{34}q) \qquad (5.18)$$

As in the case when the two fitnesses of the double heterozygotes are the same, there are two cases to be considered here:

$$\text{either} \quad (i) \quad 2W - (W_1 + W_4) < 0$$

$$\text{or} \quad (ii) \quad 2W - (W_1 + W_4) \geq 0 \ .$$

Now if $2W - (W_1 + W_4) < 0$, then equation (5.18) is satisfied if

$$R < \frac{W_1 + W_4 - 2W}{\omega_{12}p + \omega_{34}q} \qquad (5.19)$$

since the L.H.S. of equation (5.18) is always non-negative. And we observe that it is impossible for equation (5.18) to assume the form

$$R > k \qquad (5.20)$$

where k is some positive constant. Thus there are no circumstances under which increased values of R are beneficial in increasing the frequencies of the new gametes. Now if we interpret $\frac{1}{2}(W_1 + W_4)$ to be the average mean fitness of the gametes AB and aB , then we can say that if the mean fitness W of the equilibrium population before the introduction of B is less than the average mean fitness of the e new gametes AB and aB , then the frequencies of the new gametes will eventually increase if the coefficient of linkage between the two loci is tighter than the value given by equation (5.19).

Now if $2W - (W_1 + W_4) \geq 0$, then squaring both sides of equation (5.18) and simplifying, we obtain

$$R[\omega_{34}q(W - W_1) + \omega_{12}p(W - W_4)] < -(W - W_1)(W - W_4) . \qquad (5.21)$$

For equation (5.21) to assume the form of equation (5.20), we require

(1) $(W - W_1)(W - W_4) > 0$

(2) $\omega_{34}q(W - W_1) + \omega_{12}p(W - W_4) < 0$.

Jointly, this is equivalent to the condition:

$$W_1 > W \text{ and } W_4 > W \qquad (5.22)$$

But this contradicts our assumption that $2W \geq W_1 + W_4$ since equation (5.22) implies that $2W < W_1 + W_4$. Hence, even if we allow different fitnesses for the two different double heterozygotes, it still does not alter our conclusion that there are no circumstances under which "looser" linkage is beneficial for establishing new gametes.

The above analysis shows that whether the mean fitness W of the equilibrium population is greater (or equal to) or less than the average mean fitness $\frac{1}{2}(W_1 + W_4)$ of the new gametes, there are situations in which closer or tighter linkage is beneficial in the establishment of new gametes but there are no circumstances under which "looser" linkage is beneficial in increasing the frequencies of the new gametes.

References

Bodmer, W.F. & Parsons, P.A. [1960], The initial progress of new genes with various genetic systems. Heredity 15, pp 283-299.

Bodmer, W.F. and Parsons, P.A. [1962], Linkage and Recombination in Evolution Advances in genetics 11, pp 1 - 100.

Ewens, W.J. [1963a], The diffusion equation and a pseudo-distribution in Genetics. J.R. Statist. Soc. (B) Vol. 25, No. 2, pp 405-412.

Ewens, W.J. [1963b], The mean time for absorption in a genetic process. Aust. J. Maths. Vol. III part 3 pp 357-383.

Ewens, W.J. [1963c], Numerical results and diffusion approximations in a genetic process. Biometrika 50 pp 241-249.

Ewens, W.J. [1964a], Correcting diffusion approximation in finite genetic models. Stanford University Technical Report No 1. April 1st.

Ewens, W.J. [1964b], The pseudo-transient distribution and its uses in genetics. J. Applied Prob. Vol 1. pp 141-156.

Ewens, W.J. [1964c], The maintenance of alleles by mutation. Genetics, Vol 50, No 5 pp 891-898.

Feller, W. [1957], An introduction to Prob. Theory and its application. Vol 1, 2nd Edition. John Wiley.

Kimura, M. [1962], On the probability of fixation of mutant genes in a population. Genetics Vol 47 No 6 pp 713-719 .

Kimura, M. [1964], Diffusion methods in population genetics. J. Applied Prob. Vol 1 No. 2 pp 177-232.

Kimura, M and Crow, J.F. [1964], The number of alleles that can be maintained in a finite population. Genetics Vol 49 pp 725-738.

Moran, P.A.P. [1958], Random process in genetics. Proc. Camb. Phil. Soc. 54 pp 60-71.

Moran, P.A.P. [1962], The statistical processes of Evolutionary Theory. Oxford, Claren Press.

Moran, P.A.P. [1964], On the non-existence of adaptive topographies. Ann. Hum. Genet. Lond. Vol. 27 pp 383-393.

Watterson, G.A. [1961], Markov chains with absorbing states: a genetic example. Ann. Math. Statist. 32, pp 716-729 .

Knopp, K. [1959], Theory and Application of infinite series. 2nd Edition London, Blackie and Son.