Genet. Res., Camb. (1969), **13**, *pp.* 143–163
With 2 text-figures
Printed in Great Britain

On the theory of artificial selection in finite populations*

By W. G. HILL†
Statistical Laboratory, Iowa State University, Ames, Iowa, 50010, U.S.A.

(*Received* 4 *June* 1968)

1. INTRODUCTION

In a simple type of artificial selection programme individuals are ranked on their own phenotype for some quantitative trait and the highest ranking individuals are selected to be parents of the next generation. Prediction equations for changes in gene frequency with this form of selection have been derived for models in which the population is assumed to be infinitely large (Haldane, 1931; Kimura, 1958; Griffing, 1960; Latter, 1965). The truncation point, which is the value on the phenotypic scale exceeded only by selected individuals, can be assumed to be constant for specified gene frequencies and genotypic effects if the population size is infinite. However, in a finite population the truncation point must be a random variable with its value dependent on the genotypes and environmental deviations of the individuals actually present in the population. Kojima (1961) has derived formulae for expected changes in gene frequency at a single locus in finite populations, but an assumption of his model is that the effects of individual genes on the quantitative trait are small relative to the phenotypic standard deviation. Curnow & Baker (1968) have extended Kojima's results to repeated cycles of selection by using a beta distribution to approximate the distribution of gene frequencies.

In this paper a rather restricted model is analysed exactly. Predictions of changes in gene frequency are obtained for the case where there is selection on the basis of the individual phenotype (mass selection), but the quantitative trait is affected by the genotype at only one locus and by random environmental deviations. The theory is developed initially for a single cycle of selection, but is then extended to cover repeated generations of selection in a finite monecious diploid population in which there is random mating. Some of the formulae obtained are evaluated numerically for the case of normally distributed environmental deviations.

These numerical results are used to check some approximate methods which may be used to study changes in gene frequency in finite populations. These approximations involve infinite population models or assumptions of genes with small effect on the quantitative trait. In particular, some of the theory of limits to artificial selection in finite populations (Robertson, 1960; Allan & Robertson, 1964; Hill & Robertson, 1966) has been based on results of Kimura (1957) for the chance of fixation of single genes. Kimura used a haploid model and adopted a diffusion equa-

* Journal Paper No. J-6036, Iowa Agricultural and Home Economics Experiment Station, Ames, Iowa, Project No. 1669. Supported by National Institute of Health, Grant No. G.M. 13827.

† Present address: Institute of Animal Genetics, Edinburgh, 9.

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

tion, which is continuous in time and gene frequency. In extending these results to artificial selection programmes Robertson (1960) had to use results from infinite population theory to compute selective values of the genes affecting the metric trait.

This paper thus falls into two separate parts. In the first a mathematical theory of response to artificial selection for single loci is developed and in the second numerical checks are made to test the accuracy of more simple, approximate, formulae.

2. THEORETICAL ANALYSIS

A generation, which comprises one cycle of selection, may be considered in two successive stages. In the initial stage a sample of say, M, individuals is obtained at random from reproduction among the parents. These M individuals are a sample from a conceptual population of infinite size, comprised of all possible progeny genotypes and phenotypes from the given set of parents, with the probability distribution of genotypes among these M individuals depending on the mating system and parental genotypic frequencies. In the second stage the M individuals are ranked on phenotype and the top ranking N, say, are selected to be parents of the next generation. The second stage, namely of selection, will be discussed first as this is more difficult. Thus we consider a subpopulation of M individuals each of which has specified genotype, but not phenotypic value.

(i) *Single stage of selection from a finite sample with specified genotypes*

For simplicity let us assume that there are only two kinds of genotype, denoted A_1 and A_2. These may be regarded as either haploid individuals or the only two genotypes segregating in a backcross to a homozygous line. Extension to three or more genotypes is straightforward and will be given later.

The phenotypic values of individuals of genotype A_1, for example, are random variables because there are chance environmental effects and, in general, because of segregation at other loci, but these loci are assumed neutral in the model. Let us assume that the phenotypic values have continuous probability density functions and cumulative distribution functions given by

$$\begin{aligned} A_1{:}f_1(x), \quad F_1(x), \quad -\infty < x < \infty; \\ A_2{:}f_2(x), \quad F_2(x), \quad -\infty < x < \infty. \end{aligned}$$

The mean of each distribution can be interpreted as the appropriate genotypic value, and deviations from the mean come from environmental effects.

Let us assume that in some sample of M individuals there are M_1 of type A_1 and M_2 of type A_2, with $M_1 + M_2 = M$. The N individuals with the best phenotype are selected and among these the numbers of A_1 and A_2 individuals will depend on the actual phenotypes of the M individuals. Thus we wish to compute the conditional probability of selecting $N_1 A_1$ and $N_2 = N - N_1$, A_2 individuals, conditional on M_1 and M_2 and also, of course M and N. Let this probability be denoted $p(N_1|M_1)$, where, for $N_1, N_2 \geqslant 0$, N_1 must lie in the range

$$\max(0, N - M_2) \leqslant N_1 \leqslant \min(M_1, N),$$

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

where max and min denote the greater and smaller terms, respectively, in their arguments. The probability $p(N_1|M_1)$ will now be derived using order statistics for mixed distributions.

Imagine that the poorest individual selected has phenotype in the range x to $x+dx$, with N_1A_1 and N_2A_2 selected. Then either the N_1th largest of the A_1's is in $[x, x+dx]$, so that N_2 of the A_2's have phenotype above $x+dx$ and $(M_2-N_2)A_2$'s have phenotype below x *or* the N_2th ranking of the A_2's is in $[x, x+dx]$ with N_1A_1's above $x+dx$ and $(M_1-N_1)A_1$'s below x. For $dx \to 0$ these events are mutually exclusive. From the theory of order statistics we know that the probability that the N_1th largest A_1 from the sample of M_1 lies in $[x, x+dx]$ is

$$\frac{M_1!}{(N_1-1)!\,(M_1-N_1)!}[F_1(x)]^{M_1-N_1}[1-F_1(x)]^{N_1-1}f_1(x)\,dx$$

and the probability that only N_2A_2's have phenotype superior to $x+dx$ is, as $dx \to 0$,

$$\frac{M_2!}{N_2!\,(M_2-N_2)!}[F_2(x)]^{M_2-N_2}[1-F_2(x)]^{N_2}.$$

These probabilities are independent. Also, summing over the two alternatives that the Nth ranking is A_1 or A_2, we obtain the probability that the Nth ranking lies in $[x, x+dx]$ with N_1A_1 and N_2A_2 selected, which is

$$\begin{aligned}
&\frac{M_1!}{(N_1-1)!\,(M_1-N_1)!}[F_1(x)]^{M_1-N_1}[1-F_1(x)]^{N_1-1}f_1(x)\,dx \\
&\qquad\qquad \times \frac{M_2!}{N_2!(M_2-N_2)!}[F_2(x)]^{M_2-N_2}[1-F_2(x)]^{N_2} \\
&+\frac{M_2!}{(N_2-1)!\,(M_2-N_2)!}[F_2(x)]^{M_2-N_2}[1-F_2(x)]^{N_2-1}f_2(x)\,dx \\
&\qquad\qquad \times \frac{M_1!}{N_1!(M_1-N_1)!}[F_1(x)]^{M_1-N_1}[1-F_1(x)]^{N_1}.
\end{aligned}$$

Integrating over x and simplifying, we obtain

$$p(N_1|M_1) = \binom{M_1}{N_1}\binom{M_2}{N_2}\int_{-\infty}^{\infty}\{[F_1(x)]^{M_1-N_1}[F_2(x)]^{M_2-N_2}[1-F_1(x)]^{N_1-1}[1-F_2(x)]^{N_2-1}$$

where
$$\times\{N_1[1-F_2(x)]f_1(x)+N_2[1-F_1(x)]f_2(x)\}\}\,dx, \qquad (1)$$

$$\max(0, N-M_2) \leqslant N_1 \leqslant \min(M_1, N) \quad \text{and} \quad M_2 = M-M_1, N_2 = N-N_1.$$

Generalization to $g > 2$ genotypes with distributions $F_1(x), \ldots, F_g(x)$ is immediate. The Nth largest individual may be from each alternative type. If there are $M_1, \ldots, M_g$ with $\sum_{i=1}^{g} M_i = M$ in the original sample of each type, the probability that $N_1, \ldots, N_g$ are selected becomes

$$\begin{aligned}
p(N_1, \ldots, N_g|M_1, \ldots, M_g) = \int_{-\infty}^{\infty}\prod_{i=1}^{g}\binom{M_i}{N_i}[F_i(x)]^{M_i-N_i}[1-F_i(x)]^{N_i} \\
\times \sum_{j=1}^{g} N_j[1-F_j(x)]^{-1}f_j(x)\,dx. \qquad (2)
\end{aligned}$$

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

Thus we have obtained a general equation for the distribution of the genotypes of individuals selected on the basis of phenotype from a finite population.

If the genotypes all have identical distributions, $F_1(x) = F_2(x) = \ldots = F_g(x)$, equation (2) reduces to

$$p(N_1, \ldots, N_g|M_1, \ldots, M_g) = \prod_{i=1}^{g} \binom{M_i}{N_i} \Big/ \binom{M}{N}$$

which is the hypergeometric distribution. So with neutral genes, the problem is reduced to one of random sampling of N out of M without replacement.

If there is complete dominance at a single locus having only two alleles, A and a, some simplification of the formulae for three genotypes is possible if we assume that the distributions of environmental deviations as well as genotypic values are the same for both genotypes carrying the dominant allele, A. Letting the subscripts 1, 2 and 3 refer to AA, Aa and aa individuals, respectively, we have $F_1(x) = F_2(x)$ and thus

$$p(N_1, N_2, N_3|M_1, M_2, M_3) = p(N_1+N_2|M_1+M_2) \binom{M_1}{N_1} \binom{M_2}{N_2} \Big/ \binom{M_1+M_2}{N_1+N_2}, \qquad (3)$$

where $p(N_1+N_2|M_1+M_2)$ is obtained by appropriate substitution in equation (1).

In the haploid case of equation (1) the expected frequency of A_1 among the selected individuals is, of course,

$$E(N_1/N|M_1) = \frac{1}{N} \sum_{\max(0,N-M_2)}^{\min(M_1,N)} N_1 p(N_1|M_1)$$

and in the diploid case the expected frequency of the allele A is

$$\frac{1}{N} \sum_{R} (N_1 + N_2/2)\, p(N_1, N_2, N_3|M_1, M_2, M_3), \qquad (4)$$

where R denotes all possible combinations of N_1, N_2 and N_3 such that

$$N_1 + N_2 + N_3 = N \quad \text{and} \quad N_1, N_2, N_3 \geqslant 0.$$

(ii) *The complete cycle of selection*

Our analysis has so far only been in terms of selection from a finite sample with specified genotypes. The distribution of these M genotypes available for selection will depend on the genotypes of their parents, the mating system, fertility differences among the parents and viability differences of the individuals prior to artificial selection. Let us consider just the case of three genotypes and assume that the probability that there are M_1, M_2 and M_3 individuals of genotype AA, Aa and aa available for selection is $\pi(M_1, M_2, M_3|S)$, where S specifies the parental genotypes, mating system, etc., and $M_1 + M_2 + M_3 = M$. If individual selection is practised among these M individuals, the probability $Q(N_1, N_2, N_3|S)$ of selecting N_1, N_2 and N_3 of type AA, Aa and aa, respectively, is

$$Q(N_1, N_2, N_3|S) = \sum_{C} p(N_1, N_2, N_3|M_1, M_2, M_3)\, \pi(M_1, M_2, M_3|S), \qquad (5)$$

where $p(N_1, N_2, N_3|M_1, M_2, M_3)$ is given by (2) and summation (C) is taken over all values of M_1, M_2 and M_3 such that $M_1 + M_2 + M_3 = M$.

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

In a regular breeding system in a monecious population in which each generation N individuals breed M progeny, from which the best N are selected, we can replace S in equation (5) by the numbers N_{1t}, N_{2t}, N_{3t} of genotypes at generation t, thus obtaining the transition probability $Q(N_{1,t+1}, N_{2,t+1}, N_{3,t+1}|N_{1t}, N_{2t}, N_{3t})$. Under our model this is independent of t.

(iii) *Repeated cycles of selection with random mating*

The model which will be considered in detail is where there is random mating, including random selfing, among N monecious diploid individuals in each generation, and there are no fertility or viability differences. Then the progeny are multinomially distributed, and

$$\pi(M_1, M_2, M_3|S) = \binom{M}{M_1 M_2 M_3}\left(\frac{i}{2N}\right)^{2M_1}\left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^{M_2}\left(1-\frac{i}{2N}\right)^{2M_3},$$

where there are $i = 2N_{1t} + N_{2t}A$ alleles among the parents at generation t. The gene frequency is $i/2N$. Thus

$$Q(N_{1,t+1}, N_{2,t+1}, N_{3,t+1}|N_{1t}, N_{2t}, N_{3t}) = Q(N_{1,t+1}, N_{2,t+1}, N_{3,t+1}|i)$$

so that with no selection and random mating the distribution of genotypes among individuals of the next generation is a function only of i and not the genotypic frequencies. Now we can construct a transition probability matrix, $\mathbf{P}$, for changes in gene frequency from generation to generation, and can ignore the genotypic distribution of both the parental and progeny populations. We also assume that these transition probabilities are independent of t; i.e. that the distribution of genotypic values does not change with time, nor does the mating system. Let $\mathbf{P}$, with elements $(p_{ij}), i, j = 0, \ldots, 2N$ be the conditional probability that there are jA alleles among the N parents at generation $t+1$ given that there were i among the parents at generation t. Thus

$$p_{ij} = \sum_{N_1, N_2, N_3 \text{ for } 2N_1 + N_2 = j} Q(N_1, N_2, N_3|i) \quad (i, j = 0, \ldots, 2N),$$

where summation is taken over all combinations of N_1, N_2, N_3 such that there are $2N_1 + N_2 = jA$ alleles among the parents of the next generation. Combining all the relevant formulae we obtain

$$p_{ij} = \sum_{N_1, N_2, N_3 \text{ for } 2N_1 + N_2 = j} \sum_{C} \binom{M}{M_1 M_2 M_3}\left[\frac{i}{2N}\right]^{2M_1}\left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^{M_2}\left[1-\frac{i}{2N}\right]^{2M_3}$$
$$\times \int_{-\infty}^{\infty} \prod_{h=1}^{3} \binom{M_h}{N_h}[F_h(x)]^{M_h - N_h}[1 - F_h(x)]^{N_h} \sum_{k=1}^{3} N_k[1 - F_k(x)]^{-1} f_k(x)\, dx. \quad (6)$$

Changes in the distribution of gene frequency for several cycles of selection can be obtained by repeated multiplication of the matrix $\mathbf{P}$.

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

3. NUMERICAL EVALUATION OF THE FORMULAE FOR NORMALLY DISTRIBUTED ENVIRONMENTAL DEVIATIONS

Let us assume that environmental deviations are normally distributed about the genotypic value, an assumption which is made in most theoretical predictions of selection advance. The integrals of equations (1) and (2) cannot then be evaluated without recourse to numerical methods unless, of course, $F_1(x) = F_2(x) = F_3(x)$ for the two allele diploid model we shall investigate. For the case of additive gene action on the quantitative trait let the phenotypic values of AA individuals have a normal distribution with mean $\mu + \alpha\sigma$ and variance σ^2, i.e. have the $N(\mu + \alpha\sigma, \sigma^2)$ distribution and similarly let $N(\mu + \alpha\sigma/2, \sigma^2)$ and $N(\mu, \sigma^2)$ be the distributions of Aa and aa individuals, respectively, where $-\infty < \mu < \infty$, $-\infty < \alpha < \infty$ and $0 < \sigma^2 < \infty$. But $p(N_1, N_2, N_3|M_1, M_2, M_3)$ is dependent only on α in this model so that equation (2) can be evaluated using the normal distributions $N(\alpha/2, 1)$, $N(0, 1)$ and $N(-\alpha/2, 1)$ for AA, Aa and aa individuals. Thus α is the difference between the phenotypic values of the two homozygotes as a proportion of the environmental standard deviation. Letting $\phi(x)$ and $\Phi(x)$ denote the density and distribution functions of the standardized normal distribution, $N(0, 1)$, equation (2) for the additive model becomes

$$p(N_1, N_2, N_3|M_1, M_2, M_3) = \int_{-\infty}^{\infty} \prod_{i=1}^{3} \binom{M_i}{N_i} [\Phi(x - \alpha + \tfrac{1}{2}i\alpha)]^{M_i - N_i} [\Phi(-x + \alpha - \tfrac{1}{2}i\alpha)]^{N_i} \times \sum_{j=1}^{3} \{N_j [\Phi(-x + \alpha - \tfrac{1}{2}j\alpha)]^{-1} \phi(x - \alpha + \tfrac{1}{2}j\alpha)\} dx \quad (7)$$

since, by symmetry, $1 - \Phi(x) = \Phi(-x)$.

With a model of complete dominance the alternative homozygotes have also been assumed to differ by $\alpha\sigma$ units in genotypic value and to have normally distributed phenotypic values. Thus from equations (1) and (3) we have

$$p(N_1, N_2, N_3|M_1, M_2, M_3) = \binom{M_1}{N_1}\binom{M_2}{N_2}\binom{M_3}{N_3} \int_{-\infty}^{\infty} [\Phi(x - \tfrac{1}{2}\alpha)]^{M_1 + M_2 - N_1 - N_2} [\Phi(x + \tfrac{1}{2}\alpha)]^{M_3 - N_3} \times [\Phi(-x + \tfrac{1}{2}\alpha)]^{N_1 + N_2 - 1} [\Phi(-x - \tfrac{1}{2}\alpha)]^{N_3 - 1} \times [(N_1 + N_2)\, \Phi(-x - \tfrac{1}{2}\alpha)\, \phi(x - \tfrac{1}{2}\alpha)\, dx + N_3\, \Phi(-x + \tfrac{1}{2}\alpha)\, \phi(x + \tfrac{1}{2}\alpha)\, dx]. \quad (8)$$

The probabilities $p(N_1, N_2, N_3|M_1, M_2, M_3)$ are much less quickly computed for all N_1, N_2, N_3 with equation (7) than equation (8). In the latter numerical integration need only be performed for the range of possible values of $N_1 + N_2$.

Equations (7) and (8) were integrated by Simpson's rule over the region

$$-5{\cdot}12 \leqslant x \leqslant 5{\cdot}12$$

using an I.B.M. 360/50 computer with double-precision arithmetic. The values of $\Phi(x)$ were previously tabulated in the computer in the same way. Since

$$p(N_1, N_2, N_3|M_1, M_2, M_3)$$

is a probability mass function it must sum to unity over the range of N_1, N_2 and N_3

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

possible for specified M_1, M_2 and M_3. The step length for integration was taken sufficiently small that

$$|\Sigma p(N_1, N_2, N_3|M_1, M_2, M_3) - 1| < 10^{-7}.$$

The range $-5{\cdot}12 \leqslant x \leqslant 5{\cdot}12$ was found to be adequately wide, since quantities like $[\Phi(x)]^{M-N}[\Phi(x)]^N \phi(x)$ are very small unless x is close to zero.

In Table 1 some examples of the form of $p(N_1, N_2, N_3|M_1, M_2, M_3)$ are given for the case of additive gene action. The expectations of the genotypic and gene frequencies among the selected individuals are also shown. In the next section we shall use the exact results obtained by numerical integration to check various approximate formulae for selection advance in both single and repeated cycles of selection.

Table 1. *Probabilities of selecting each possible combination of genotypes in a single stage of selection for an additive gene*

($M_1 = 4$, $M_2 = 8$, $M_3 = 4$ and $N = 4$.)

			$p(N_1, N_2, N_3\|M_1, M_2, M_3)$		
N_1	N_2	N_3	$\alpha = 0$	$\alpha = 0{\cdot}2$	$\alpha = 0{\cdot}8$
0	0	4	0·000549	0·000283	0·000028
0	1	3	0·017582	0·010634	0·001775
0	2	2	0·092308	0·065634	0·018227
0	3	1	0·123077	0·102878	0·047574
0	4	0	0·038462	0·037795	0·029133
1	0	3	0·008791	0·006232	0·001657
1	1	2	0·105494	0·087919	0·038877
1	2	1	0·246154	0·241161	0·177522
1	3	0	0·123077	0·141750	0·173873
2	0	2	0·019780	0·019321	0·013590
2	1	1	0·105494	0·121134	0·141784
2	2	0	0·092308	0·124597	0·242910
3	0	1	0·008791	0·011830	0·021990
3	1	0	0·017582	0·027813	0·086063
4	0	0	0·000549	0·001019	0·004994
$E(N_1/N)$			0·250000	0·282543	0·383159
$E(N_2/N)$			0·500000	0·498837	0·481673
$E(N_3/N)$			0·250000	0·218620	0·135168
$E[(N_1+N_2/2)/N]$			0·500000	0·531961	0·623996

4. COMPARISON OF RESULTS FROM EXACT AND APPROXIMATE METHODS

(i) *Single stage of selection from a sample with specified genotypes*

If there are M_1, M_2 and M_3 individuals of genotype AA, Aa and aa respectively from which selection is made, the expected gene frequency in selected individuals is given by (4). However, as $M \to \infty$ the average gene frequency among selected individuals is readily computed, for the trunaction point T is no longer a random variable. With additive gene action T must satisfy the following equation on the standardized scale

$$M_1 \Phi(-T + \tfrac{1}{2}\alpha) + M_2 \Phi(-T) + M_3 \Phi(-T - \tfrac{1}{2}\alpha) = N \qquad (9)$$

as $M \to \infty$ where, for example $\Phi(-T + \frac{1}{2}\alpha) = 1 - \Phi(T - \frac{1}{2}\alpha)$ is the proportion of AA

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

individuals which have phenotype superior to T and are selected. The mean frequency q', of A alleles among the selected individuals is then

$$q' = \frac{M_1}{N}\Phi(-T+\tfrac{1}{2}\alpha)+\frac{M_2}{2N}\Phi(-T). \tag{10}$$

Equation (10) is easily evaluated and requires little computation.

Table 2. *Expected change in gene frequency when selecting N individuals from a population of size M in exact Hardy–Weinberg frequencies*

(The change in gene frequency is tabulated for infinite M, and changes at other values of M as a percentage difference, $P = [(q'-q)_m/(q'-q)_\infty - 1] \times 100$.)

		Additive				Complete dominant			
	M ...	8	16	32	$\to\infty$	8	16	32	$\to\infty$
q	α		P		$q'-q$		P		$q'-q$
					(1) $N/M = \frac{1}{4}$				
0·25	0·2	—	0·91	0·36	0·024192	—	0·83	0·33	0·035937
	0·8	—	1·17	0·48	0·098819	—	0·74	0·28	0·140654
0·5	0·2	2·30	0·78	0·30	0·031713	1·65	0·51	0·18	0·030604
	0·8	2·23	0·73	0·27	0·123099	−0·27	−0·32	−0·21	0·104206
0·75	0·2	—	0·65	0·24	0·023391	—	0·33	0·09	0·011176
	0·8	—	0·29	0·07	0·086674	—	0·61	−0·35	0·034964
					(2) $\alpha = 0{\cdot}4$				
q	N/M								
0·25	$\frac{1}{2}$	—	1·58	0·72	0·029723	—	1·62	0·74	0·044569
	$\frac{1}{8}$	—	−0·84	−0·60	0·064769	—	−1·20	−0·75	0·093064
	$\frac{1}{16}$	—	—	−2·40	0·078694	—	—	−2·51	0·110752
0·5	$\frac{1}{2}$	3·81	1·59	0·72	0·039630	3·71	1·55	0·71	0·039431
	$\frac{1}{8}$	—	−1·18	−0·75	0·081449	—	−1·74	−0·89	0·071782
	$\frac{1}{16}$	—	—	−2·45	0·096986	—	—	−2·43	0·081990
0·75	$\frac{1}{2}$	—	1·58	0·72	0·029723	—	1·41	0·64	0·014637
	$\frac{1}{8}$	—	−1·45	−0·87	0·057791	—	−1·88	−1·05	0·024820
	$\frac{1}{16}$	—	—	−2·48	0·067622	—	—	−2·31	0·027830

In Table 2 predictions of expected change in gene frequency from a single stage of selection using the finite population and infinite population methods are compared for both additive and completely dominant gene action. The configurations of genotypic frequency among the M individuals are chosen such that $M_2^2 = 4M_1M_3$, with the original frequency q being $q = (M_1+M_2/2)/M$. Thus for $q = 0{\cdot}25$ and $M = 32$ we have $M_1 = 2, M_2 = 12$ and $M_3 = 18$. These genotypic frequencies are those corresponding to the Hardy–Weinberg equilibrium frequencies, but since we are only considering one sample they may be assumed to have occurred by chance. Other possible configurances have not been considered separately. In Table 2 the predicted changes in gene frequency $(q'-q)$ computed with the infinite population model (equation (10) for additive gene action) are given, and the expected changes using the finite model expressed as a proportion of these. The results of the table

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

indicate that the infinite population model gives a very close prediction of the response expected from finite populations. Even when as few as 2 individuals out of 16 are chosen the error scarcely exceeds 2 % of the mean change in gene frequency.

(ii) *Complete cycle of selection with random mating*

In a complete generation or cycle of selection there is sampling of progeny followed by selection of parents for the next generation. We shall only consider the case where the genotypes among the M progeny are multinomially distributed with expected frequencies q^2, $2q(1-q)$ and $(1-q)^2$ for AA, Aa and aa individuals, where q is the frequency of A among the parents. General formulae for this model, assuming a random mating monecious population, have been given in an earlier section. The expected gene frequency $E(q')$ among the parents of the next generation is, for complete dominance and integral $2Nq$,

$$E(q') = E(j/2N|q = i/2N) = \frac{1}{2N}\sum_{j=0}^{2N} jp_{ij} \tag{11}$$

where p_{ij} is given in equation (6). The model has been restricted by assuming that there are N parents in each of the two generations, but relaxation of this assumption is straightforward. Also integration has been carried out only for the case of complete dominance so that equation (8) could be used to reduce computation time.

An approximate method for obtaining $E(q')$ has been given by Kojima (1961). He showed that for small values of α (the gene effect in standard deviations) such that α^2, α^3, etc. could be ignored relative to α, the mean change in gene frequency

$$\left.\begin{aligned} \delta q &= E(q') - q \\ \text{is} \qquad \delta q &\sim k\alpha q(1-q)^2 \end{aligned}\right\} \tag{12}$$

for complete dominance. Kojima calls k a 'generalized selection differential', and Pike (1969) has shown that if the phenotypic values are normally distributed about the genotypic values k becomes the mean of the highest N order statistics in a sample of size M from a single standardized normal distribution.

As M becomes infinitely large the value of k can be obtained directly from tables of the normal distribution, and may be denoted i, the standardized selection differential. Thus $\lim_{M\to\infty} k = i$ for N/M constant. Equation (12) is then the well known approximate formula for the change in gene frequency with truncation selection (Haldane, 1931; Kimura, 1958; Griffing, 1960; Latter, 1965) in which $i\alpha$ is the selective value of the allele A. Latter (1965) has studied the errors associated with this approximation for predicting changes in gene frequency in infinite population. The exact values for q' in infinite population can, of course, be obtained from (10).

In Table 3 the approximate and exact methods are compared for a choice of values of the parameters α, N/M, q and M. Predictions of change of gene frequency, δq, have been computed for the exact method (equation (11)) and are tabulated as a proportion of the change predicted by the simple form $\delta q = k\alpha q(1-q)^2$. The values of k were obtained from tables of the expectations of order statistics from the normal distribution, which are given to 10 decimal places by Teichroew (1956). In the

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

limiting case of $N \to \infty$, k equals i in the approximate formulae, and in the exact formulation the finite population predictions are replaced by the exact infinite predictions of equation (10), where M_1, M_2 can be replaced by Mq^2 and $2Mq(1-q)$. The values for $N \to \infty$ are thus tests of the infinite model approximations, but at the same time serve as limiting values for the finite model approximations of Kojima in which it is assumed that gene effects (α) are small.

Table 3. *Response from one full cycle of selection for complete dominance*

(M progeny are taken at random from parents in Hardy–Weinberg equilibrium with gene frequency q, and N are selected. The response (δq) is tabulated as a percentage deviation from $k\alpha q(1-q)^2$, i.e. as $[\delta q/k\alpha q(1-q)^2 - 1] \times 100$, where k is the mean of first N from M order statistics from the standardized normal distribution.)

		M			
q	α	4	8	16	$\to \infty$
		$N/M = \frac{1}{2}$			
0·25	0·2	−0·27	−0·23	−0·20	−0·18
	0·8	−4·23	−3·60	−3·20	−2·73
0·5	0·2	−0·32	−0·31	−0·30	−0·29
	0·8	−4·87	−4·73	−4·66	−4·58
0·75	0·2	−0·41	−0·47	−0·51	−0·55
	0·8	−6·27	−7·12	−7·64	−8·22
		$\alpha = 0{\cdot}4$			
q	N/M		8	16	$\to \infty$
0·25	$\frac{1}{4}$	—	+0·09	+0·24	+0·42
	$\frac{1}{8}$	—	—	+0·32	+0·84
0·5	$\frac{1}{4}$	—	−7·24	−7·60	−7·99
	$\frac{1}{8}$	—	—	−11·91	−12·50
0·75	$\frac{1}{4}$	—	−11·48	−12·16	−21·90
	$\frac{1}{8}$	—	—	−18·22	−19·32

We see in Table 3 that the approximation for the finite model is rarely much poorer than with an infinite population. Since essentially the same assumptions about the size of α are made in each case, we should not be surprised to observe that a poor fit between the predictions is only found with finite populations for values of α and selection intensity (i.e. N/M) which lead to inadequate approximation in infinite population.

Kojima (1961) also derived formulae for the variance of change in gene frequency based on the same assumptions as the mean change. For complete dominance this is

$$V(\delta q) \sim \frac{q(1-q)}{2N}[1 + k\alpha(1-q)(1-3q)]. \tag{13}$$

Some checks on the accuracy of (13) have been made against the exact finite population prediction, obtained by finding $E(q')^2$ by extension of equation (11). Again, as we would expect the approximate and exact methods agree well except at the highest values of $\alpha(0{\cdot}8)$ and selection intensity.

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

Attention should perhaps be drawn to the fact that when we calculated the expected change in gene frequency from a population with specified numbers M_1, M_2, M_3 of each genotype a good approximation was obtained using an infinite population prediction (10). For small α equation (10) reduces to $q' = q + i\alpha q(1-q)^2$ for complete dominance, where $q = (2M_1 + M_2)/2M$ and i is the infinite population standardized selection differential. However, when the M individuals in the population are themselves a sample of genotypes then the selection differential should be calculated as k from order statistics for the appropriate *finite* population size. In the latter case we obtain a reasonable fit using order statistics from the normal distribution because the combined (binomial) distribution of genotypic values and (normal) distribution of environmental values is close to normal in form.

(iii) *Chance of fixation of single genes*

The theory of limits in artificial selection in finite populations developed by Robertson (1960) is based on the concept of the chance of fixation, $u(q_0)$, which is the probability that an allele with initial frequency q_0 will eventually be fixed in the population. Using a diffusion equation (the Kolmogorov backward equation), Kimura (1957, 1962) showed that, for example, the chance of fixation of a dominant allele with selective value was

$$u(q_0) = \int_0^{q_0} e^{Ns(1-x)^2}\,dx \Big/ \int_0^1 e^{Ns(1-x)^2}\,dx. \tag{14}$$

To describe the response to artificial selection the selective value has been taken as $s = i\alpha$ (Robertson, 1960; Hill & Robertson, 1966). The model used in (14) is continuous and haploid in form, so it seemed necessary to check the accuracy of (14) for diploids with artificial selection and discrete generations. Previously Ewens (1963) has made numerical tests on the errors resulting from use of the diffusion equation, but only for haploid individuals and additive gene action.

The approximate results were obtained by numerical integration of (14), using Simpson's rule, where s was replaced by $i\alpha$ and also by $k\alpha$, with k computed for a few pairs of N and M values.

Exact results for our model of a diploid monecious random mating population with stationary transition probabilities were obtained from the matrix $\mathbf{P}$ (equation (6)). A vector $\mathbf{v}_{(0)}$ with elements $v_{i(0)}$ was first constructed, where $v_{i(0)} = i/2N$, $i = 0, \ldots, 2N$. Then successive products $\mathbf{v}_{(1)} = \mathbf{P}\mathbf{v}_{(0)}, \mathbf{v}_{(2)} = \mathbf{P}\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(t)} = \mathbf{P}\mathbf{v}_{(t-1)}$ were computed. An element $v_{i(t)}$ is therefore the expected gene frequency at time t for an initial frequency of $i/2N$. Iteration was continued for at least $6N$ generations so that the ratio of changes in gene frequency in successive generations

$$(v_{i(t)} - v_{i(t-1)})/(v_{i(t-1)} - v_{i(t-2)})$$

became sufficiently constant that the chance of fixation, $\lim_{t\to\infty} v_{i(t)}$, could be predicted to 5 decimal places by fitting an exponential curve to the last 3 values of $v_{i(t)}$ by the δ^2 method (Aitken, 1926). This iterative method of obtaining the chance of fixation was preferred to more direct methods, since expected gene frequencies at inter-

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

mediate generations were required for further tests on approximate methods which will be described in the next section.

In Tables 4 and 5 comparisons are shown of the chance of fixation computed by the exact method and by the diffusion approximation. Results are given for different values of $N\alpha$. Positive values imply that the dominant allele is favoured by selection,

Table 4. *Chance of fixation* $u(q_0) \times 10^4$ *for a dominant gene with truncation selection computed exactly by matrix iteration and by diffusion approximation*

(The selective value for the diffusion equation is $k\alpha$, with k computed from order statistics for specified values of N. The chance of fixation is tabulated for the diffusion results with $N \to \infty$, D_∞, others by difference from D_∞.)

		Matrix N				Diffusion N		
$N\alpha$	q_0	2	4	10	∞	10	4	2
		$[D_\infty - u(q_0)] \times 10^4$			$D_\infty \times 10^4$	$[D_\infty - u(q_0)] \times 10^4$		
				$N/M = 0{\cdot}5$				
0·2	0·25	54	—	13	2864	15	—	64
	0·5	66	—	15	5403	15	—	68
	0·75	48	—	11	7748	10	—	42
0·4	0·25	114	62	26	3254	20	73	134
	0·5	130	70	29	5812	31	75	138
	0·75	90	48	20	7990	18	43	81
0·8	0·25	252	136	57	4098	66	158	292
	0·5	245	131	55	6621	60	145	270
	0·75	159	83	34	8447	33	80	149
1·6	0·25	560	287	118	5853	130	312	583
	0·5	396	202	81	8043	94	229	437
	0·75	224	110	43	9180	45	112	216
3·2	0·25	—	436	159	8430	134	337	—
	0·5	—	156	55	9556	60	155	—
	0·75	—	66	22	9850	23	61	—
6·4	0·25	—	—	80	9837	31	—	—
	0·5	—	—	6	9989	4	—	—
	0·75	—	—	1	9998	1	—	—
−0·2	0·5	−68	—	−15	4607	−15	—	−65
−0·4	0·5	−167	−73	−30	4229	−28	−68	−126
−0·8	0·5	−271	−142	−58	3531	−50	−121	−226
−1·6	0·5	−517	−255	−101	2404	−73	−178	−340
−3·2	0·5	—	−381	−135	1088	−68	−168	—
−6·4	0·5	—	—	−119	0238	−29	—	—
				$N/M = 0{\cdot}25$				
0·8	0·25	216	117	—	5148	—	156	296
	0·5	225	118	—	7510	—	124	238
	0·75	155	79	—	8916	—	64	123
1·6	0·25	307	160	—	7595	—	220	430
	0·5	220	114	—	9146	—	122	243
	0·75	134	65	—	9684	—	52	105
−0·8	0·5	−204	−103	—	2819	—	−97	−187
−1·6	0·5	−265	−125	—	1501	—	−109	−214

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

negative values that the recessive allele is favoured. The continuous model with $s = i\alpha$ can be regarded as the limiting case as $N \to \infty$, with $N\alpha$ remaining constant. We find in the tables that there is mostly quite good agreement between the exact and approximate predictions of chance of fixation. As we must expect, the fit is poorest at low values of N and high values of α, for given $N\alpha$, especially when the initial frequency of the favoured allele is low (Table 5). The diffusion approximation method with $s = i\alpha$ generally overestimates the total change in expected gene frequency, $|u(q_0) - q_0|$. Thus when s is replaced by $k\alpha$ for the appropriate M and N values a better fit is obtained since $k < i$; but, except for the smallest N values, this correction may not be thought worthwhile. The values of N ($\leqslant 10$) used in this study are less than in many animal selection experiments or programmes so, in practice, k and i may differ by very little.

Table 5. *Chance of fixation* $\times 10^4$

(Computed from exact transition matrix (TM) with $N = 10$ and $M = 20$, and by diffusion approximation (DA) with selective value computed from order statistics. Diffusion result is shown as difference $D = DA - TM$.)

q_0 ...	0·05		0·1		0·5		0·9		0·95	
$N\alpha$	TM	D	TM	D	TM	D	TM	D	TM	D
0·4	720	−2	1398	−3	5783	−2	9196	+1	9598	+1
1·6	1697	+4	3046	+2	7962	−13	9668	+1	9835	0
6·4	5472	+455	7911	+338	9983	+2	9999	0	9999	0
−0·4	336	+1	692	+2	4259	−2	8799	−4	9399	−2
−1·6	86	+2	197	+2	2505	−28	8226	−28	9109	−16
−6·4	0	0	1	0	349	−82	6187	−241	8384	−138

(iv) *Rate of selection advance with repeated cycles of selection; simple transition probability matrices*

A further consequence of the diffusion approximation to the selection process for single genes in finite populations is that the distribution of gene frequencies among replicate lines, and therefore the mean gene frequency also, is a function of only Ns and the initial frequency, provided that time is measured on a scale proportional to N. This result was pointed out by Robertson (1960) and it leads to a considerable simplification of the description of the rate of advance. The chance of fixation (as $t \to \infty$) is then a function of only Ns and q_0, and we see in Table 4 that this still holds reasonably well when we compare the exact values for the chance of fixation computed for the same $N\alpha$ and N/M, but different N.

As a measure of the rate of advance we shall use the 'half-life' of the change in gene frequency, which is the time taken for the mean gene frequency to get half way from its initial to its limiting value (Robertson, 1960). Half-lives were calculated by linear interpolation between the two successive generations which had mean gene frequency spanning the half-way frequency and have been expressed proportional to the parental population size, N, in the relevant tables.

If M and N become large an excessive amount of numerical integration is re-

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

quired in order to evaluate the matrix **P** (equation (6)), and it becomes very difficult to carry out the computation of **P** with sufficient accuracy. Therefore it seems desirable to have a more efficient, if approximate, method for computing intermediate gene frequencies and half-lives. A simple type of transition matrix was constructed and compared with the exact matrix **P** for artificial selection in a monecious random mating population. In this simple matrix it is assumed that the gene frequency among the parents of the next generation is binomially distributed with mean $q + sq(1-q)^2$ from (12). Let us denote this matrix **B**, with elements (b_{ij}), $i,j = 0, \ldots, 2N$, which define the same transition probabilities as do the elements of **P**. Thus b_{ij} is the (approximate) probability that there are jA alleles among the N parents at generation $t+1$, given that there were i at generation t. **B** is assumed independent of t. The elements of **B** are obtained by adopting a haploid type of model. We assume that the expected gene frequency in generation $t+1$ is

$$\frac{i}{2N} + \frac{si}{2N}\left(1 - \frac{i}{2N}\right)^2$$

for complete dominance. The selective value s can be replaced by $k\alpha$ in Kojima's (1961) formulation, as we have seen in equation (12). The $2N$ alleles among the parents of the next generation are then obtained by sampling from the binomial distribution. Thus

$$b_{ij} = \binom{2N}{j}\left[\frac{i}{2N} + \frac{si}{2N}\left(1 - \frac{i}{2N}\right)^2\right]^j \left[1 - \frac{i}{2N} - \frac{si}{2N}\left(1 - \frac{i}{2N}\right)^2\right]^{2N-j}. \qquad (15)$$

Expected gene frequencies in the intermediate stages of selection and chances of fixation were obtained by repeated iteration of the matrix **B** in the same manner as described for the matrix **P**.

In Tables 6 and 7 comparison is made of the chances of fixation and half-lives, respectively, computed using matrices **P** and **B**. In **B** the selective value s is set equal to $k\alpha$ for the appropriate value of N. We find a rather better fit in Table 6 between the pairs of matrix results than we observed between the results from the diffusion and the exact method in Table 4. For the half-lives the agreement between results for different values of N and constant $N\alpha$ improves as N increases with either method. Also, for large N and small α the approximate and exact methods agree more closely with each other. This pattern of results could be predicted to some extent because the continuous model assumptions are less severely violated at large N, ignoring terms in α^2, $\alpha^3, \ldots$ becomes less serious for small α, and because the change of k with N is smaller as N becomes larger. In Tables 6 and 7 effects of population size on the genetic sampling process and on selection intensity are confounded since the parameter k is used. For constant values of Ns, but differing N, half-lives have been computed using the simplified matrix **B** and a few results are given in Table 8. Again, although some wide discrepancies occur at the higher Ns value, there is probably sufficient agreement for practical purposes because approximate values of half-lives (or other measures of rate of advance) are all we are likely to need when planning or interpreting selection experiments. Also, it should be pointed out that the apparently large discrepancies between predicted half-lives (Table 7) using

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

Table 6. *Chance of fixation* $\times 10^4$

(Computed by exact transition probability matrix method (P) and approximate matrix method (B) with $N/M = 0{\cdot}5$. Results are shown as deviations $P_N - P_{10}$ or $B_N - P_{10}$ from exact method with $N = 10$.)

	N ...	2		4		10	
$N\alpha$	q_0	B	P	B	P	B	P
0·4	0·25	−103	−88	−46	−36	−4	3228
	0·5	−119	−101	−53	−42	−5	5783
	0·75	−81	−70	−34	−28	−3	7970
1·6	0·25	−588	−442	−295	−169	−65	5735
	0·5	−480	−315	−234	−121	−55	7962
	0·75	−268	−181	−124	−67	−27	9137
6·4	0·25	—	—	—	—	−33	9757
	0·5	—	—	—	—	−8	9983
	0·75	—	—	—	—	−3	9997
−0·4	0·5	+86	+107	+31	+42	−4	4259
−1·6	0·5	+124	+416	+15	+154	−54	2505
−6·4	0·5	—	—	—	—	−114	349

Table 7. *Half-lives* ($\times 1000/N$ *generations*)

(Computed by exact transition probability matrix method (P) and approximate matrix method (B) with $N/M = 0{\cdot}5$. Results are shown as deviations $P_N - P_{10}$ or $B_N - B_{10}$ from exact method with $N = 10$.)

	N ...	2		4		10	
$N\alpha$	q_0	B	P	B	P	B	P
0·4	0·25	−197	−156	−76	−57	−9	1231
	0·5	−173	−131	−74	−51	−10	1412
	0·75	−147	−94	−62	−37	−10	1701
1·6	0·25	−241	−92	−115	−45	−30	1331
	0·5	−235	−69	−113	−34	−30	1414
	0·75	−199	−5	−102	−12	−34	1644
6·4	0·25	—	—	—	—	−16	645
	0·5	—	—	—	—	−22	607
	0·75	—	—	—	—	−40	762
−0·4	0·5	−84	−122	−30	−48	+8	1272
−1·6	0·5	+7	−61	+16	−24	+18	994
−6·4	0·5	—	—	—	—	−8	400

Table 8. *Half-lives* $\times 1000/N$ *generations computed with different* N *and constant* Ns *using the simplified transition matrix* B

	q_0 ...	0·25		0·5		0·75	
	N ...	8	32	8	32	8	32
Ns	1	1239	1349	1398	1456	1639	1691
	4	756	791	711	757	857	900
	−1	794	820	1076	1090	1443	1446
	−4	250	278	476	482	853	835

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

matrices **P** and **B** may have no practical significance for small N values. For example, with $N = 2$ half-lives of $1{\cdot}345N$ and $1{\cdot}179N$ ($N\alpha = 1{\cdot}6, q_0 = 0{\cdot}5$) both imply simply that the mean gene frequency at generation 2 is less than half-way to its expected limit and that at generation 3 more than half-way. It is possible to construct matrices other than **B** which give better approximations to the exact results. For example, diploid selection can be included in terms of selective values, and still not require numerical integration. However, the extra computation involved in setting up such matrices does not seem justified by the small increase in precision obtained. Curnow & Baker (1968) have developed an alternative method of predicting the selection advance by approximating the gene frequency distribution by a beta distribution. The accuracy of this method has recently been checked by Pike (1969) and found to be satisfactory for all but the smallest population size (4) checked.

(v) *Optimum intensity of artificial selection*

The optimum intensity of selection in an artificial selection programme has been discussed by Dempster (1955) and Robertson (1960), who pointed out that, for fixed M, the selection limit would be maximized if $N/M = 0{\cdot}5$. This conclusion is based on the diffusion equation model and assumes that N is very large so that the limit is a function of Ni. For the normal distribution $i = z/(N/M)$ where z is the ordinate of the standardized normal distribution at the truncation point. Thus $Ni = Mz$ so Ni is maximized when z is maximized at $N/M = 0{\cdot}5$. However, even in finite populations, it will now be shown that sufficient conditions for Nk to be mazimized when $N/M = 0{\cdot}5$ are for the distribution of phenotypic values to be unimodal and symmetric. Let $x_1, \ldots, x_M$ be the expected values of the order statistics as deviations from the mean of a symmetric distribution, then

$$\sum_{i=1}^{N} x_i + \sum_{i=N+1}^{M} x_i = 0$$

and by symmetry, $x_i = -x_{M-i+1}$. Substituting, we obtain

$$\sum_{i=1}^{N} x_i = \sum_{i=N+1}^{M} x_{M-i+1}$$

$$= \sum_{i=1}^{M-N} x_i$$

or

$$Nk_N = (M-N)\,k_{M-N},$$

where k_N and k_{M-N} are the means of the best N and $M-N$, respectively, ordered individuals. Therefore, as long as the approximation from the diffusion equation that Nk is a sufficient parameter holds fairly well we expect the limit to be maximized when half the population is selected and to be symmetric about this proportion.

Some checks on this prediction were made using the exact model with $M = 16$, $q_0 = 0{\cdot}5$ and $\alpha = 0{\cdot}4$ or $-0{\cdot}4$, with the limit computed for $N = 2$, 4, 6, 8, 10, 12 and 14 individuals selected each generation. Results are shown in Fig. 1, where we

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

observe that the curve of chance of fixation against N departs very little from symmetry. Of course the rates of approach to the limit differ widely. This is illustrated in Fig. 2, in which the mean gene frequency is plotted against number of generations (t) for $M = 16$, $q_0 = 0{\cdot}5$ and $\alpha = 0{\cdot}2$. Also in Fig. 2 we find that the time scale fits

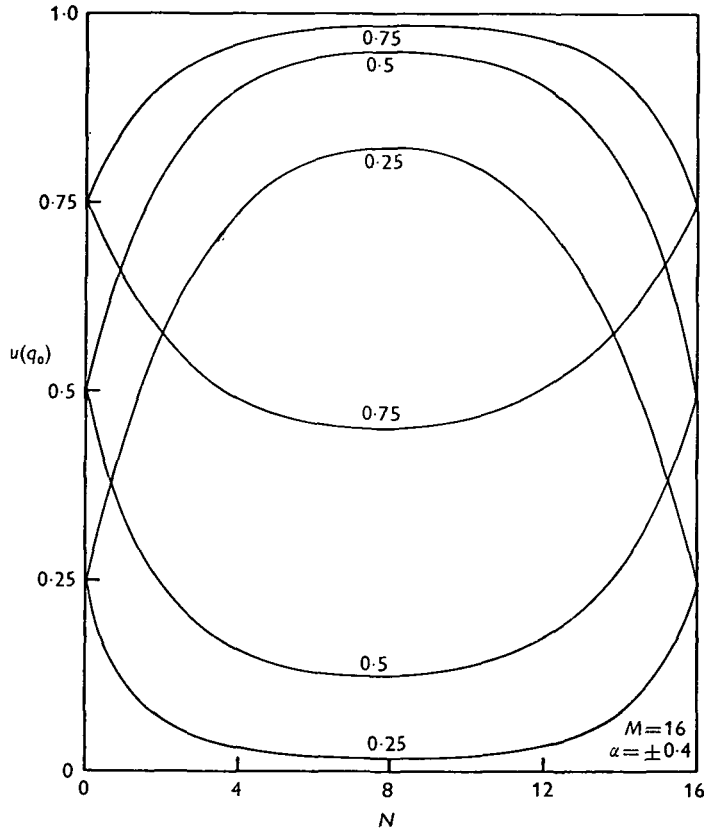

Fig. 1. The effect of selection intensity on the selection limit. The chance of fixation $u(q_0)$ is computed by the exact method and plotted for $q_0 = 0{\cdot}25$, $0{\cdot}5$ and $0{\cdot}75$, for different numbers of individuals (N) selected from 16 recorded every generation.

well with the diffusion theory in that it is inversely proportional to N. Thus we expect the same mean frequency after cN generations with population size N as with $c(M-N)$ generations with population size $M-N$, where c is a positive constant. In the example of Fig. 2 let us compare $N = 4$ with $N = 12$, where gene frequencies for some values of c are as follows:

Population size	c ...	0·5	1·0	2·0	4·0	$\to \infty$
			Mean gene frequency			
4		0·55348	0·59337	0·64871	0·70424	0·72718
12		0·55366	0·59365	0·64919	0·70530	0·74129

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

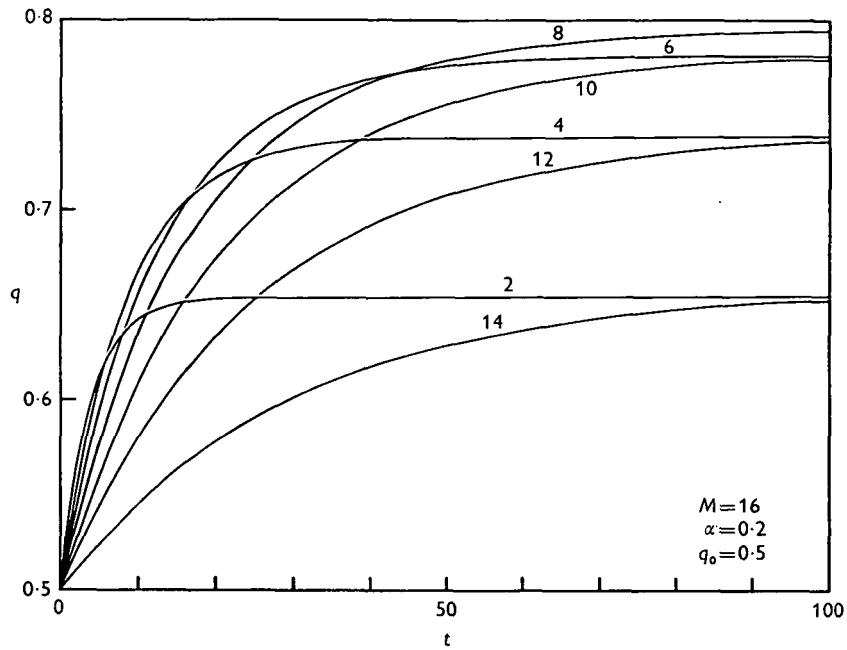

Fig. 2. The effect of selection intensity on the rate of selection advance. The mean gene frequency, q, is computed by the exact method and plotted against the number of generations (t) of selection for different numbers of individuals selected from 16 recorded every generation.

5. DISCUSSION

The formulae developed for the simple model of artificial selection in finite populations have enabled us to make checks on some of the simplifying assumptions in the theory of selection limits. The population sizes which have been tested (usually $N \leqslant 10$) are smaller than would normally be encountered in breeding schemes in which much emphasis is placed on selection within lines. Thus, in practice, we would expect population sizes to be larger and the diffusion approximation to fit better than in the examples given here. Therefore, in view of our results, we can probably conclude that the diffusion equation gives an adequate approximation for the model of a single gene in a random mating monecious population which is analysed in this paper.

At the same time, this single locus model is unlikely ever to be realized for a quantitative trait in nature, nor are monecious populations of direct interest in livestock improvement. The limitations of this approach therefore appear to rest mostly on the model adopted. However, this study should be viewed as an initial attempt to test the adequacy of some simple theory for describing artificial selection in finite populations.

A comparison of monecious and diecious models has been made by Hill & Robertson (1968) for the case of natural selection acting on viability differences at a single locus with complete dominance or heterozygote advantage. The populations

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

comprised 10 parents in the monecious model, and 5 of each sex in the diecious case. The alternative models led to similar results for the change in fitness (Hill & Robertson, 1968) and mean gene frequency (unpublished), and it was considered that the monecious model was adequate for descriptive purposes.

Even if there are many loci segregating for the quantitative trait the algebraic theory developed in the first part of this paper can, in principle, still be used for single generations of selection. One alternative approach is merely to evaluate the probability of selection of each possible genotype. As in the single locus situation we have discussed, the only variation of phenotypic value about genotypic value would be attributed to environmental deviation. The expected change of gene frequency at a single locus can then be obtained by summation of selection probabilities over all possible genotypes. Alternatively, segregation at other loci can be included as variation in the phenotypic values about the genotypic values of the locus with which we are concerned. Thus the distribution of phenotypic values will be the distribution of the sum of, say, normally distributed environmental deviations and perhaps binomially distributed genetic differences. If there is no linkage disequilibrium, epistasis or genotype-environment interaction the distributions of phenotype may only differ in mean. If there are many independent genes of small effect which influence the trait, the phenotypes may be almost exactly normally distributed. The variance will be equal to the total phenotypic variance for the trait, less that actually contributed by the locus under consideration. This approximation has been used in infinite population theory by Griffing (1960) and Latter (1965), and by Kojima (1961) for finite populations.

However, when there are repeated cycles of selection and several loci affect the selected trait it may be difficult to justify the assumption that the transition probabilities of matrix **P**, for example, are stationary. Selection and inbreeding will change the frequencies and variance at each locus, so that the distribution of phenotypic values for a specified genotype, and therefore the selective values, will not remain constant over generations. The extent to which selective values will change in the presence of other loci will, of course, depend on their initial frequencies, effects and linkage relationships. The general tendency would seem to be for selective values to increase as other loci approach fixation as a result of selection or drift. At the same time, as Robertson (1960) has mentioned, the environmental variance may rise due to inbreeding, and may partially compensate for the reduction in genetic variance. Also, it is clear that genes of large effect and low initial frequency of the favourable allele are most likely to be lost from the population in the first few generations. If they survive to this stage their frequency is unlikely still to be low, and they will become fixed eventually. Thus 'decisions' about the fate of such genes, and essentially all genes with relatively large Ns value, are taken in early generations before the phenotypic variance can have changed appreciably, so that a theory developed for single loci may given satisfactory predictions in such cases. It will be less satisfactory for genes of smaller effect when fixation takes longer, but further work on this topic is clearly required.

When selection is practised in finite populations initially in equilibrium tight

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

linkage leads to an excess of the repulsion phase (Hill & Robertson, 1966; Latter, 1966). Thus a theory based on single loci overestimates the expected selection gain in this case.

SUMMARY

The effect of selection on individual performance for a quantitative trait is studied theoretically for populations of finite size. The trait is assumed to be affected by environmental error and by segregation at a single locus. Exact formulae are derived to predict the change in gene frequency at this locus, initially by finding the probability distribution of the numbers of each genotype selected from a finite population of specified genotypic composition. Assuming that there is random mating and no natural selection the results are extended to describe repeated cycles of artificial selection for a monecious population. The formulae are evaluated numerically for the case of normally distributed environmental errors.

Using numerical examples comparisons are made between the exact values for the predicted change in gene frequency with values obtained using approximate, but simpler, methods. Unless the gene has a large effect (α) on the quantitative trait, relative to the standard deviation of the environmental errors, the agreement between exact and approximate methods is satisfactory for most predictive purposes. The chance of fixation after repeated generations of selection is also evaluated using the exact method, and by means of a diffusion approximation and simple transition probability matrix methods. Except for very small values of population size (N) and large α the results from the diffusion equation agree closely with those from the exact method. Similar results are found from tests made of the prediction from the diffusion equation that the limit is only a function of $N\alpha$ for a given intensity of selection and initial frequency, and that the rate of advance in gene frequency is proportional to $1/N$ for the same set of parameters.

I am grateful to Professors O. Kempthorne and Alan Robertson for their helpful suggestions and comments on the manuscript.

REFERENCES

Aitken, A. C. (1926). On Bernoulli's numerical solution of algebraic equations. *Proc. Roy. Soc. Edinb.* **46**, 289–305.

Allan, J. S. & Robertson, A. (1964). The effect of initial reverse selection upon total selection response. *Genet. Res.* **5**, 68–79.

Curnow, R. N. & Baker, L. H. (1968). The effect of repeated cycles of selection and regeneration in populations of finite size. *Genet. Res.* **11**, 105–112.

Dempster, E. R. (1955). Genetic models in relation to animal breeding problems. *Biometrika* **11**, 525–536.

Ewens, W. J. (1963). Numerical results and diffusion approximation in a genetic process. *Biometrika* **50**, 241–249.

Griffing, B. (1960). Theoretical consequences of truncation selection based on the individual phenotype. *Aust. J. Biol. Sci.* **13**, 307–343.

Haldane, J. B. S. (1931). A mathematical theory of natural and artificial selection. VII. Selection intensity as a function of mortality rate. *Proc. Camb. Phil. Soc.* **27**, 131–136.

Hill, W. G. & Robertson, A. (1966). The effect of linkage on the limits to artificial selection. *Genet. Res.* **8**, 269–294.

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press

HILL, W. G. & ROBERTSON, A. (1968). The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60** (in Press).

KIMURA, M. (1957). Some problems of stochastic processes in genetics. *Ann. Math. Statist.* **28**, 882–901.

KIMURA, M. (1958). On the change of population fitness by natural selection. *Heredity* **12**, 145–167.

KIMURA, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719.

KOJIMA, K. (1961). Effects of dominance and size of population on response to mass selection. *Genet. Res.* **2**, 177–188.

LATTER, B. D. H. (1965). The response to artificial selection due to autosomal genes of large effect. I. Changes in gene frequency at an additive locus. *Aust. J. Biol. Sci.* **18**, 585–598.

LATTER, B. D. H. (1966). The interaction between effective population size and linkage intensity under artificial selection. *Genet. Res.* **7**, 313–323.

PIKE, D. J. (1969). A comparison of two methods for predicting changes in the distribution of gene frequency when selection is applied repeatedly to a finite population. *Genet. Res.* **13**, 117–126.

ROBERTSON, A. (1960). A theory of limits in artificial selection. *Proc. Roy. Soc.* B **153**, 234–249.

TEICHROEW, D. (1956). Tables of expected values of order statistics and products of order statistics from samples of size 20 and less from the normal distribution. *Ann. Math. Statist.* **27**, 410–426.

https://doi.org/10.1017/S0016672300002858 Published online by Cambridge University Press