Genet. Res., Camb. (1968), **11**, *pp.* 105–112
Printed in Great Britain

The effect of repeated cycles of selection and regeneration in populations of finite size

By R. N. CURNOW
Department of Applied Statistics, University of Reading, England
and L. H. BAKER
Pioneer Hi-bred Corn Company, Des Moines, Iowa, U.S.A.

(*Received* 21 *December* 1966)

1. INTRODUCTION

Kojima (1961) derived approximate formulae for the mean and the variance of the change in gene frequency from a single cycle of selection applied to a population of finite size. He used these formulae to derive the expected gain from the selection. Kojima's paper was the first to take full account of the variable effects of selection in finite populations. In this paper, we use and extend Kojima's formulae to derive a simple approximate method for studying the effects of repeated cycles of selection and regeneration in a finite population. The method yields not only the first two moments of the gene frequency distribution, the expected gain from selection and the probabilities of fixation but also the variability of gain, i.e. the variation to be expected between the gains made in identical replicate populations. This latter variation is of interest in the study of evolution; in the design and interpretation of selection experiments and in devising plant and animal breeding improvement programmes. Reference will be made to the alternative of using transition matrices.

2. THE GENETIC MODEL AND THE SELECTION PROCEDURE

Using Kojima's notation, the population consists initially of N diploid individuals in which the three single locus genotypes, AA, Aa and aa, occur with frequencies U_1, U_2, and U_3 respectively. We shall assume that the population is a random-mating population and that, following each selection, the selected individuals are mated at random. Therefore $U_1 = q^2$, $U_2 = 2q(1-q)$ and $U_3 = (1-q)^2$, where q is the frequency of the A allele.

Kojima assumed that the distribution of the phenotypic values of the character being used for selection would be similar for each genotype, differing only in mean. We shall denote the three means as follows:

Genotype	Frequency	Mean
AA	q^2	a
Aa	$2q(1-q)$	ha
aa	$(1-q)^2$	o

The homozygote difference is a and the degree of dominance is measured by h. We shall assume that the phenotypic distribution for each of the three genotypes is normal. Because the genetic effects at a single locus are assumed small relative to the total variation, the variance for each genotype will be taken equal to the total phenotypic variance, σ^2. We shall assume that there are no epistatic effects and no linkage. We shall assume that the heritability of the character is sufficiently low that the total phenotypic variance can be assumed constant. This should be a reasonable approximation in the early cycles of a selection programme unless the selection is very intense or the population very small. The genetic values at a locus will be measured as a proportion of the phenotypic standard deviation, σ, i.e. a/σ will henceforth be written as a.

At each selection, the n individuals with the highest value of the character will be selected. The selected individuals are then mated at random to produce the next generation and so on.

Kojima showed that in a single selection, the change in gene frequency, Δq, had mean and variance given by

$$E(\Delta q) = akq(1-q)\,[h+q(1-2h)] \tag{1}$$

and

$$V(\Delta q) = \frac{q(1-q)}{2n}\{1+ak[(1-2h)q(1-q)+(1-2q)\,(h+q-2hq)\}, \tag{2}$$

where k is the mean of the top n values in a random sample of size N from a standard normal distribution. Values of k can be calculated from tables given in *Biometrika Tables for Statisticians*, volume 1 (1958). Kojima discusses formulae (1) and (2) in some detail. All terms of order a^2 and higher are, of course, being ignored.

We shall need one further result that can be derived from Kojima's formulation of the problem. This is the probability that an allele, A say, becomes fixed in a single selection. Kojima shows that the number of AA individuals selected has a binomial distribution with

$$\text{parameters } n \text{ and } p = q^2\left[1+\frac{d_1\phi(Y_0)}{P}\right],$$

where d_1 is the deviation of the AA mean from the population mean in standard deviation units. $\phi(Y_0)$ and $(1-P)$ are the values of the standard normal frequency function and distribution function at Y_0, the value of the best individual that is not selected. For given Y_0, the probability that all the individuals selected are AA and therefore that A becomes fixed is

$$p^n = q^{2n}\left[1+\frac{d_1\phi(Y_0)}{P}\right]^n.$$

Averaging this probability over the distribution of Y_0—a random variable since n is fixed—leads to analytical difficulties. Instead, as an approximation, we shall replace $\phi(Y_0)/P$ within the bracket by its expected value. This expected value can be shown to be k as defined above. The approximate probability of fixing A is then given by

$$P_1(q) = q^{2n}[1+ka(1-q)\,(1+q-2hq)]^n. \tag{3}$$

The corresponding formula for the probability of fixing a is

$$P_0(q) = (1-q)^{2n}[1-kaq(q+2h-2hq)]^n. \tag{4}$$

$P_1(q)$ and $P_0(q)$ cannot be satisfactorily approximated by using only the first two terms of the binomial expansion of p^n because this assumes that na as well as a is small.

3. CYCLES OF SELECTION

Equations (1) and (2) can be used at any stage of a programme of continued selection provided that q is the gene frequency in the preceding generation. This permits an iterative approach to the study of the effects of cycles of selection. We shall use symbols without dashes to indicate parameters in one generation and the same symbols with dashes for the corresponding parameters in the next generation.

From (1),

$$\begin{aligned} E(q') &= q + E(\Delta q) \\ &= q + akq(1-q)[h+q-2hq]. \end{aligned} \tag{5}$$

Denoting the rth moment about zero of the distribution of gene frequencies by μ_r and averaging over the distribution of q, (5) becomes

$$\mu_1' = \mu_1 + ak[h(\mu_1-\mu_2) + (1-2h)(\mu_2-\mu_3)]. \tag{6}$$

Similarly,
$$E[(q')^2] = q^2 + 2qE(\Delta q) + E(\Delta q)^2.$$

Using (1) and (2) and averaging over the distribution of q this gives, to the order of a single-locus genetic effect,

$$\begin{aligned} \mu_2' = \mu_2 + \frac{\mu_1-\mu_2}{2n} &+ 2ak[h(\mu_2-\mu_3) + (1-2h)(\mu_3-\mu_4)] \\ &+ \frac{ak}{2n}[(1-2h)(2\mu_2-5\mu_3+3\mu_4) + h(\mu_1-3\mu_2+2\mu_3)]. \end{aligned} \tag{7}$$

Equations (6) and (7) cannot yet be used for iteration since they contain the higher moments μ_3 and μ_4. Further equations could be derived for μ_3' and μ_4' but they would involve still higher moments. A particular form for the gene frequency distribution has to be assumed to relate μ_3 and μ_4 to the lower moments, μ_1 and μ_2. We shall assume that the gene frequency distribution is made up of finite probabilities, P_0 and P_1, that $q = 0$ and $q = 1$ and a beta-distribution with parameters l and m for $0 < q < 1$. Symbolically,

$$\begin{aligned} &\Pr(q=1) = P_1 \\ &f(q) = \frac{(1-P_0-P_1)\,q^{l-1}(1-q)^{m-1}}{B(l,m)} \quad (0<q<1) \\ &\Pr(q=0) = P_0, \end{aligned} \tag{8}$$

where $B(l,m)$ is the usual beta-function. The beta-distribution is extremely flexible. The true gene-frequency distribution is concentrated at the $(2n+1)$ points $r = i/2n$ $(i = 0, 1, \ldots, 2n)$. A continuous distribution is probably a sufficiently good

https://doi.org/10.1017/S001667230001123X Published online by Cambridge University Press

approximation providing that n is not too small. Selfing ($n = 1$) will be discussed later. The distribution has four parameters: P_1, P_0, l and m. We already have two iterative equations, (6) and (7). They involve the first four moments of the distribution. These moments can be written in terms of the four parameters above by evaluating the moments of distribution (8). This gives

$$\mu_r = \frac{P_1 + (1-P_0-P_1)\, l(l+1)\ldots(l+r-1)}{(l+m)\,(l+m+1)\ldots(l+m+r-1)} \quad (r = 1, 2, 3, 4). \tag{9}$$

We need two more iterative equations. These are obtained by relating the probabilities of fixation in one generation (P_0', P_1') to those in the previous generation (P_0, P_1). Using (3) and (4),

$$\begin{aligned} P_1' &= P_1 + \frac{(1-P_0-P_1)}{B(l,m)} \int_0^1 P_1(q)\, q^{l-1}(1-q)^{m-1}\, dq \\ &= P_1 + \frac{(1-P_0-P_1)\, B(l+2n, m)}{B(l,m)} \left\{ 1 + \frac{nkam(2l+m+4n+1-2hl-4hn)}{(l+m+2n)\,(l+m+2n+1)} \right\}. \end{aligned} \tag{10}$$

Terms involving $a^2, a^3, \ldots$, do not need to be considered because they are not multiplied by terms of order $n^2, n^3, \ldots$, as they were in the binomial expansion of p^n mentioned previously. Similarly,

$$P_0' = P_0 + \frac{(1-P_0-P_1)\, B(l, m+2n)}{B(l,m)} \left\{ 1 - \frac{nkal(l+2hm+4hn+1)}{(l+m+2n)\,(l+m+2n+1)} \right\}. \tag{11}$$

The process of iteration can be described as follows, where the numbers are the relevant equations.

Starting values a, h, k, n, G (number of generations), q.

	μ_1	μ_2	μ_3	μ_4	P_0	P_1	l	m
Initially	q	q^2	q^3	q^4	0	0	$\left[l = \frac{mq}{1-q} = \infty\right]$	
Generation g	μ_1	μ_2	μ_3	μ_4	P_0	P_1	l	m
		$\mu_1'(6)$				$P_0'(11)$		
		$\mu_2'(7)$				$P_1'(10)$		
					l',	m' (9)		
					μ_3',	μ_4' (9)		
Generation $(g+1)$	μ_1'	μ_2'	μ_3'	μ_4'	P_0'	P_1'	l'	m'

The values of P_0 and P_1 following the first selection can most easily be calculated from equations (3) and (4). The values of the beta-function in (10) and (11) will only be difficult to evaluate when n is large. A reasonable approximation then might be

$$\frac{B(l+2n, m)}{B(l,m)} \simeq \frac{\Gamma(l+m)}{\Gamma(l)} \frac{1}{(2n)^m},$$

where Γ is the usual gamma function. There is a similar approximation for the other ratio.

https://doi.org/10.1017/S001667230001123X Published online by Cambridge University Press

4. GENETIC MEAN AND BETWEEN-POPULATION VARIANCE IN GENETIC MEAN

The population mean after random mating in units of phenotypic standard deviations, when the gene frequency is q, is

$$aq[q+2h(1-q)].$$

This has a mean value of

$$a[\mu_2+2h(\mu_1-\mu_2)]. \tag{12}$$

The variance among replicate populations in mean performance in units of the phenotypic variance is

$$a^2E\{q[q+2h(1-q)]\}^2-(\text{mean})^2$$
$$=a^2\{(1-2h)^2\mu_4+4h^2\mu_2+4h(1-2h)\mu_3\}-(\text{mean})^2. \tag{13}$$

Both these quantities can be calculated after each cycle of selection. With no linkage or epistasis, the means and variances can be added over loci to study changes in the genetic value of the population. Values of the variance between replicate populations should be useful in devising and interpreting selection programmes; in studying evolutionary processes and in designing and analysing selection experiments.

5. APPLICATION OF METHOD

Formulae (1)–(7) and (10)–(11) assume that the selected individuals each act as both male and female parents of the next generation. It is not difficult to show that the following changes are needed in the iterative equations (6), (7), (10) and (11) when the 'selection intensities' and numbers selected are k_m and n_m, and k_f and n_f for males and females respectively.

(6) For k read $\bar{k}=\frac{1}{2}(k_m+k_f)$,

(7) $$\mu_2'=\mu_2+\frac{\mu_1-\mu_2}{2n_e}+2a\bar{k}[h(\mu_2-\mu_3)+(1-2h)(\mu_3-\mu_4)]$$
$$+\frac{a\bar{\bar{k}}}{2n_e}[(1-2h)(2\mu_2-5\mu_3+3\mu_4)+h(\mu_1-3\mu_2+2\mu_3)],$$

where n_e is the effective population size,

$$n_e=\frac{4n_mn_f}{n_m+n_f},$$

and $\bar{\bar{k}}$ is the weighted mean of the k values;

$$\bar{\bar{k}}=\frac{n_fk_m+n_mk_f}{n_m+n_f}.$$

(10) and (11) For n read n_m+n_f and for nk read $n_mk_m+n_fk_f$.

The methods of this paper have been used to consider how best to improve a composite population of maize as a source of inbred lines (Baker & Curnow, 1968). In this application the number of males was considered infinite and the selection applied to females only ($n_m=\infty$, $k_m=0$). With an infinite number of males, no gene

https://doi.org/10.1017/S001667230001123X Published online by Cambridge University Press

will become fixed so that $P_0 = P_1 = 0$ throughout. Values for the genetic variation between replicate populations of different sizes were calculated. The possible use of this variation in a breeding programme was discussed quantitatively for the first time. A computer is needed to carry out the calculations but the method has clear advantages over simulation studies. Simulation might provide reasonably good estimates of mean progress but the large sampling fluctuations involved make the variance in mean progress much more difficult to estimate. Simulation will still probably be needed to study the effects of multiple alleles, linkage and epistasis. However, Kojima's method could be extended to cover these situations. Kojima's formulation already covers the possibility of non-normality in the phenotypic distributions and deviations from random mating.

All the results obtained by the methods described in this paper could be derived using transition matrices (see Allan & Robertson (1964), Ewens (1963), Hill & Robertson (1966) and Robertson (1960) for examples of the use of transition matrices). The elements of the matrices are the probabilities of moving from one gene frequency to another in a single cycle of selection. The transition matrix approach is possible because all the first- and second-order statistics derived by Kojima's more correct procedure could be derived by calculating selective advantages for the genotypes based on selection above a fixed point (truncation selection) but with the usual 'selection intensity' i replaced by k. Kojima (1961) does show that the selective advantages for non-additive genes in finite populations have not always been calculated correctly (for a discussion of this, see Hill, 1968). The method described in this paper has some advantage in computer time over transition matrices, particularly for large population sizes. Work is in progress on checking the adequacy of the approximations, i.e. the beta-distribution assumption and the approximation used in calculating the probabilities of fixation, $P_1(q)$ and $P_0(q)$. The problem considered in Baker & Curnow (1968) would be difficult to solve for any population size using transition matrices. This is because the selection was applied only to the females and the males were assumed to be infinite in number. In this situation the size of the transition matrices will increase geometrically with the number of cycles of selection. Any differentiation between the sexes, whether or not it involves differences in the selection intensities or population sizes, will lead to a considerable increase in the size of the transition matrices involved.

Transition matrices have been used to check the results in Baker & Curnow (1968) for the first two cycles of selection when $n_f = 1$. Even for this low value of n_f, the approximation of the three- and nine-point gene frequency distributions in the first two cycles by a beta-distribution led to very little error in terms of genetic mean and variance. Later cycles should be even better until approximation errors accumulate or the gene frequencies change to values more sensitive to the approximation.

The next section, on selfing, shows how analytical results can sometimes be obtained using an approach that is essentially that of transition matrices.

6. SELFING

With selfing ($n = 1$), only three gene frequencies are possible and the assumption of a beta distribution of gene frequencies between $q = 0$ and $q = 1$ clearly invalid. Let the probabilities that the selected individual at the sth generation is AA, Aa and aa be $P_s(AA)$, $P_s(Aa)$ and $P_s(aa)$, respectively. If the $(s-1)$th generation individual is Aa then the probabilities that the sth generation individual is AA, Aa and aa are P_1, $1-P_0-P_1$ and P_0 respectively, where, from (3) and (4),

$$P_1 = \tfrac{1}{4}[1+\tfrac{1}{2}ak(\tfrac{3}{2}-h)]$$

and

$$P_0 = \tfrac{1}{4}[1-\tfrac{1}{2}ak(\tfrac{1}{2}+h)].$$

The probabilities in succeeding generations are related by the following formulae:

$$\begin{aligned} P_{s+1}(AA) &= P_s(AA)+P_s(Aa)\,P_1, \\ P_{s+1}(Aa) &= P_s(Aa)\,[1-P_0-P_1], \\ P_{s+1}(aa) &= P_s(aa)+P_s(Aa)\,P_0. \end{aligned}$$

If initially

$$P_0(AA) = U_1, \quad P_0(Aa) = U_2, \quad P_0(aa) = U_3 \quad (U_1+U_2+U_3 = 1),$$

then

$$P_s(AA) = U_1+\frac{U_2P_1}{P_0+P_1}[1-(1-P_0-P_1)^s],$$

$$P_s(Aa) = U_2(1-P_0-P_1)^s,$$

and

$$P_s(aa) = U_3+\frac{U_2P_0}{P_0+P_1}[1-(1-P_0-P_1)^s].$$

The values of U_1, U_2 and U_3 can be chosen appropriately if the first plant to be selfed is chosen at random or by some selection procedure. The mean gene frequency at the sth generation is

$$U_1+\tfrac{1}{2}U_2+\frac{(P_1-P_0)}{2(P_0+P_1)}U_2[1-(1-P_0-P_1)^s].$$

The genetic mean and between-population variance in genetic mean after random mating are easily derived. The probability that A becomes fixed at the sth generation is

$$P_{s-1}(Aa)\,P_1 = U_2P_1(1-P_0-P_1)^{s-1},$$

with a corresponding expression for the fixation of the other allele. After an infinite number of generations the probabilities of fixation are

$$U_1+\frac{U_2P_1}{P_0+P_1} \quad \text{and} \quad U_3+\frac{U_2P_0}{P_0+P_1},$$

for the A and a alleles respectively. Clearly one or the other allele will eventually be fixed.

The mean time to fixation for A is $(U_2P_1)/(P_0+P_1)^2$ and for a is $(U_2P_0)/(P_0+P_1)^2$.

https://doi.org/10.1017/S001667230001123X Published online by Cambridge University Press

The half-life (i.e. the time at which the mean gene-frequency is half-way to its final value) is given by the solution to the equation

$$(1-P_0-P_1)^s = \tfrac{1}{2},$$

i.e.

$$s = \frac{-\log 2}{\log (1-P_0-P_1)}.$$

The half-life is independent of the initial genotypic frequencies U_1, U_2 and U_3.

SUMMARY

Kojima's (1961) approximate formulae for the mean and variance of the change in gene frequency from a single cycle of selection applied to a finite population are used to develop an iterative method for studying the effects of repeated cycles of selection and random mating. This is done by assuming a particular, but flexible and probably realistic, approximate form for the distribution of gene frequencies at each generation.

The method gives for each generation the first two moments of the gene frequency distribution, the expected gain from selection, the probabilities of fixation and also the variability of gain. The variability of gain is of considerable importance in evolution, selection experiments and in plant and animal breeding programmes.

Kojima's (1961) formulae have been extended to allow for differentiation between males and females. Hence different selection intensities and population sizes for the two sexes can be studied. Selfing with selection is considered separately. Extensions to cover simple examples of multiple alleles, linkage and epistasis are possible. Reference is made to previous work using transition matrices.

The iterative procedures described have been used to compare different selection schemes for composite populations of maize (Baker & Curnow, 1968).

Work is in progress on testing, applying and extending these methods.

We are grateful to Dr Alan Robertson and Mr D. J. Pike for helpful comments on an earlier draft of this paper.

REFERENCES

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

Baker, L. H. & Curnow, R. N. (1968). On selection aimed at improving a composite population of maize as a source of inbred lines. (In preparation.)

Biometrika Tables for Statisticians (1958). 2nd edn. Ed. E. S. Pearson and H. O. Hartley, vol. I. Cambridge University Press.

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

Hill, W. G. (1968). The rate of selection advance for non-additive loci. *Genet. Res.* **11** (in the Press).

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

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

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