*Genet. Res., Camb.* (1986), **48**, *pp.* 41–50 *With* 11 *text-figures* *Printed in Great Britain*



# Models of long term artificial selection in finite population

WILLIAM G. HILL AND JONATHAN RASBASH

*Institute of Animal Genetics, University of Edinburgh, West Mains Road, Edinburgh EH9 3JN*

*(Received 4 November 1985 and in revised form 4 February 1986)*

**Summary**

The effects of population size and selection intensity, which are in the breeder's control, are investigated for ranges of values of quantities outside his control, namely the number, initial distribution of frequencies and effects of genes influencing the trait. Two alleles are assumed to be initially segregating at each locus, with no linkage, dominance or epistasis. The effects are assumed to follow a gamma distribution, using a wide range of its two parameters which specify both mean gene effect or selective value and the shape of the distribution, or the ratio of Wright's effective number to actual number of genes. The initial gene frequencies ($q$) are assumed to be either 0·5 at all loci, uniformly distributed over the range 0–1, or to have a U-shaped distribution, proportional to $[q(1-q)]^{-1}$ such as derives from neutral mutation, with gene effect and frequency distributions independent. The mean and variance of selection response and limits, in the absence of new mutation, are derived.

The shape of the distribution of effects is not usually important even up to the selection limit. With appropriate parametrization, the influence of the initial frequency distribution is small over a wide range of parameters. For reasonable choices of parameters, the effects of changing population size from those typically used in animal breeding programmes are likely to be small, but not negligible. For the initial U-shaped frequency distribution, further increases in population size are always expected to give a greater limit, regardless of present value, but not for the other distributions.

## 1. Introduction

Predictions of response to artificial selection for a few generations can be made using estimates of heritabilities, correlations and variances which are readily obtained (albeit often with large sampling errors) from straightforward analyses of covariance among relatives. Long term responses depend on many variables, however, that can not usually be estimated in the base population, namely: the number, frequency and effect of each gene influencing each trait of interest, together with the interactions introduced by epistasis, the correlations induced by linkage disequilibrium, the strength of natural selection opposing directional selection and the rate of occurrence and distribution of new mutations. The task of predicting long term responses has therefore either to be given up as impossible, or simple models constructed and the predictions made from them compared with data from selection experiments and breeding programmes. The concern of the breeder is to predict how the way he manages his population now influences its responses in the future, for example, what is the effect of increasing the size of the population on the total response? Even some simple models may make the process of decision making more reliable. Factors which influence this decision are, on the one hand, the costs of maintaining the population and controlling its management, which argue for small populations and, on the other hand, the possible inbreeding depression and loss of variation leading to reduced long term responses and limits, which argue for large populations. There is a further conflict in that high selection intensity, i.e. picking only few extreme individuals, increases short term but reduces long term response. These arguments and much of the basic theory were explained and developed by Robertson (1960).

In this paper, models of the quantitative trait are analysed in more detail, with the specific intention of reducing the number of relevant but generally unknown parameters to a minimum, either by reparameterization or by showing that others are not likely to be important. The main formulae are in terms of probability distributions of gene effects and frequencies

Table 1. *Definition of symbols*

| | |
|---|---|
| $t$ | generation number |
| $N$ | effective population size |
| $i$ | selection intensity |
| $h^2$ | heritability |
| $\sigma$ | phenotypic standard deviation |
| $a$ | gene effect, measured as difference between hommozygotes |
| $q$ | frequency of favourable allele |
| $s$ | $= ia/\sigma$, selective value |
| $\phi(a, q)$ | joint distribution of gene frequency and effects |
| $f(a)$ | distribution of gene effects |
| $g(q)$ | distribution of gene frequencies |
| $\alpha, \beta$ | parameters of the gamma distribution of gene effects |
| $\alpha^*$ | $= \alpha\sigma/(Ni)$, parameter of gamma distribution of $Ns$ values |
| $n$ | number of segregating genes influencing the trait |
| $n_e$ | $= n\beta/(\beta+1)$, effective number of genes |
| $n_\tau$ | total number of genes (segregating or fixed) influencing the trait |
| $M$ | effective population size prior to selection |
| $\mu$ | mutation rate per locus |
| $R_t$ | response at generation $t$ |
| $R_\infty$ | response limit |
| $u(Ns, q)$ | fixation probability of gene |
| $A^*$ | $= Nih/\sqrt{n_e}$ |

rather than by specifying any particular values. Even so, for many tastes, the models will be too simplistic. In this paper we deal with variability existing in the population at the outset; subsequently we shall incorporate the effects of mutation.

## 2. Model

### (i) *Basic assumptions*

Artificial selection is assumed to be practised by truncation selection with intensity $i$ on a trait with phenotypic standard deviation $\sigma$, so the selection differential is $i\sigma$, and the effective population size is $N$. These and other symbols are summarized in Table 1. In this study we shall make the basic assumptions that at all loci gene action is additive with no epistasis, there are two alleles, the allele conferring higher value on the trait having frequency $q$, there is a difference of $a$ in value between the homozygotes, and there are $n$ unlinked loci segregating that influence the trait. The case of multiple alleles will not be considered here: two situations in which the two-allele model is appropriate are where the population derives from a cross of two inbred lines or where population size ($M$) and mutation rate ($\mu$) have been sufficiently low previously ($4M\mu < 1$). The selective value of the gene is given by $s = ia/\sigma$, approximately (Falconer, 1981). Genes are assumed to start and remain in linkage equilibrium.

Let $\phi(a, q)$ define the joint distribution of gene frequency and effects in the base population. The heritability of the trait is given by

$$h^2 = n/(2\sigma^2) \int_{a=0}^{\infty} \int_{q=0}^{1} a^2 q(1-q)\, \phi(a, q)\, da\, dq; \qquad (1)$$

the expected total response to selection at generation $t$ is

$$R_t = n \int_{a=0}^{\infty} \int_{q=0}^{1} aE[q_t - q|a, q]\, \phi(a, q)\, da\, dq; \qquad (2)$$

and the variance of response among replicate lines selected from the same base population is

$$V(R_t) = n \int_{a=0}^{\infty} \int_{q=0}^{1} a^2 V(q_t - q|a, q)\, \phi(a, q)\, da\, dq, \qquad (3)$$

assuming linkage disequilibrium can be ignored. This variance is thus conditional on the distribution of effects and frequencies in the base population. The variance would be subtantially higher if it were measured among lines selected from different populations having, for example, the same heritability but a different origin. The latter case of unconditional variance is not completely defined because it depends on distributions of effects and frequencies among populations, so we do not consider it here.

### (ii) *Distribution of effects and frequencies*

The joint distribution of gene effects and frequency is generally unknown. Under some assumptions, notably that of genes neutral with respect to fitness before selection starts, the two distributions can be assumed to be independent. Thus we shall assume $\phi(a, q) = f(a)\, g(q)$.

The choice of distribution, $f(a)$, of gene effects is arbitrary, but one with a suitable range of properties is the gamma distribution (previously used by Kumura, 1979, except to describe the distribution of effects of deleterious genes). The density function is

$$f(a) = \alpha^\beta e^{-\alpha a} a^{\beta-1}/\Gamma(\beta) \quad (0 < a < \infty), \qquad (4)$$

where $\Gamma(\beta)$ is the gamma function. Its moments are: $E(a) = \beta/\alpha$, $E(a^2) = \beta(\beta+1)/\alpha^2$, and $V(a) = \beta/\alpha^2$. The parameter $\beta$ can be regarded as a measure of the equality of effects at different loci. Consider Wright's measure of effective number of loci ($n_e$) which compares the range ($K$) to the variance ($V$), i.e. $n_e = K^2/(8V)$ from a line cross with all genes at frequency 0·5. As $K = nE(a)$ and $V = (n/8)E(a^2)$, so $n_e/n = \beta/(\beta+1)$. Regardless of the actual gene frequencies the ratio $\beta/(\beta+1)$ will be defined as equal to the ratio of the effective to actual numbers of loci. With reference to distributions used previously in similar studies (Hill, 1982), the exponential distribution is given by $\beta = 1$, the gamma (half) distribution by $\beta = \frac{1}{2}$, the case of equal gene effects by $\beta \to \infty$, and the geometric distribution by $\beta \to 0$. Examples are given in Fig. 1. Thus the gamma spans a wide range of possibilities and, in particular, when $\beta$ is small implies that

https://doi.org/10.1017/S0016672300024642 Published online by Cambridge University Press

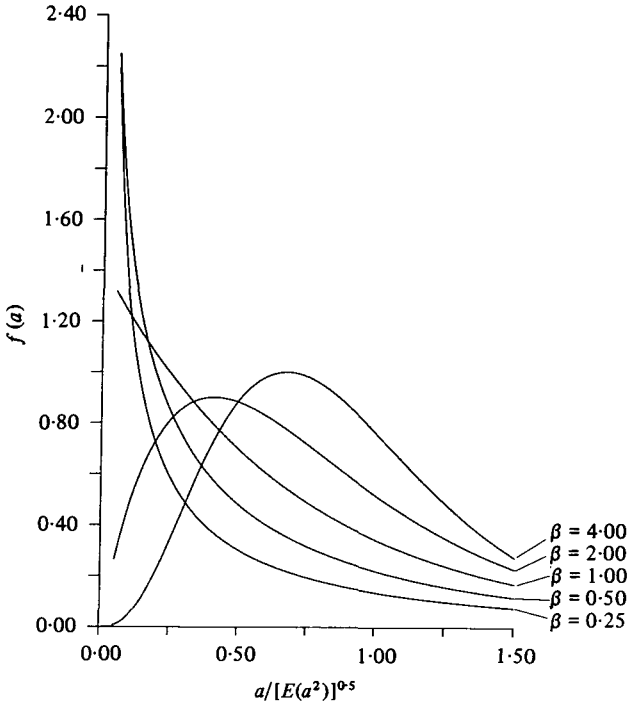


Fig. 1. Examples of the gamma distribution, $f(a) = \alpha^{\beta}e^{-\alpha a}a^{\beta-1}/\Gamma(\beta)$, for a range of $\beta$, with values expressed as $a/[E(a^2)]^{\frac{1}{2}}$, where $E(a^2) = \beta(\beta+1)/\alpha^2$. The parameter $\alpha$ describes scale rather than shape.

genes of small effect are much more common than those of large effect. The normal distribution of effects with mean zero and variance $V(a)$ has $K = 0{\cdot}798\ n\sqrt{V(a)}$ and $n_e/n = 0{\cdot}637$, equivalent to $\beta = 1{\cdot}75$. As will be demonstrated later, the gamma with this parameter or, more simply $\beta = 2$, provides a sufficient approximation.

The distribution of gene frequencies depends on the population history and we shall consider three special cases.

(*a*) *All genes at initial frequency* 0·5. Such a distribution would occur in a cross between two completely homozygous lines.

(*b*) U-*shaped distribution.* The distribution of gene frequencies describes that in a population maintained previously without selection for the metric trait, and with all genes neutral with respect to fitness. The distribution will then reflect the generation of new alleles by mutation and their loss by drift. We assume the product $M\mu$ of previous effective population size ($M$) and per-locus mutation rate ($\mu$) is sufficiently small that not more than two alleles are segregating at each locus. The frequency spectrum is then

$$g(q) = 4M\mu(1-q)^{4M\mu-1}q^{-1} \tag{5}$$

(Kimura & Crow, 1964), and for small $M\mu$,

$$g(q) \propto 1/[q(1-q)]$$

and $E[q(1-q)] \to 2M\mu$ (one half of the heterozygosity). Note that, in our model, the frequency distribution of genes influencing the trait is symmetric about 0·5.

If $n_\tau$ denotes the total number of loci affecting the trait in the genome (as opposed to $n$, used previously, to denote the number segregating), the expected heritability of the trait is then, from (1),

$$h^2 = M\mu n_\tau E(a^2)/\sigma^2. \tag{6}$$

although the actual value of heritability in any single line will depend on the frequency and effects of the genes segregating in it.

(*c*) *Uniform distribution.* A limiting case of analytical convenience and some interest is the uniform distribution of frequencies, $g(q) = 1$, $0 < q < 1$. This distribution would apply if all genes derived from mutation, crosses among populations or other sources very many generations ago, and subsequently the population had been maintained at a small effective size without appreciable selection and further mutation, the uniform being the asymptotic distribution for neutral genes (Kimura, 1955). For this case, $E[q(1-q)] = \frac{1}{6}$.

## 3. Analysis

### (i) *Consequences of diffusion approximation*

Under the assumptions of the diffusion approximation (see, for example, Crow & Kimura, 1970) the distribution of gene frequency $q_t$ at generation $t$ is a function of the initial gene frequency $q$, the products of effective population size and selective value, $Ns$, and generation number in units of population size, $t/N$, a property used by Robertson (1960). Therefore it follows that, if the gene frequencies are distributed independently of the gene effects and thus selective values, quantities such as the distribution of gene frequencies, mean gene frequencies and responses to selection are functions of the joint distributions of $q$ and $Ns$ and of $t/N$. In the case we consider where effects of initial frequencies are independent, expected responses, for example, are a function of $g(q)$, $f^*(Ns)$ and $t/N$, where $f^*$ denotes the density of $Ns$. We have assumed the gene effects, $a$, have a gamma distribution (4) with parameters $\alpha$ and $\beta$, and therefore $Ns = Nia/\sigma$ also has a gamma distribution with parameters $\alpha^* = \alpha\sigma(Ni)$ and $\beta$. Therefore $E(Ns) = \beta/\alpha^*$ and, because there is a linear relation between $Ns$ and $a$, the 'shape' of the distribution (Fig. 1) is unchanged. Thus, in summary, quantities such as expected responses are functions of $g(q)$, $\alpha^*$, $\beta$ and $t/N$.

After this reparametrization, (2) becomes

$$R_{t/N} = (n\sigma/Ni)\int_{Ns=0}^{\infty}\int_{q=0}^{1} Ns\ E[q_{t/N}-q|Ns,q] \times f^*(Ns)\,g(q)\,d(Ns)\,dq \tag{7}$$

and (3) becomes

$$V(R_{t/N}) = (n\sigma^2/N^2i^2)\int_{Ns=0}^{\infty}\int_{q=0}^{1} (Ns)^2\,V[q_{t/N}-q|Ns,q] \times f^*(Ns)\,g(q)\,d(Ns)\,dq. \tag{8}$$

https://doi.org/10.1017/S0016672300024642 Published online by Cambridge University Press

We shall consider ways of making these expressions more tangible in terms of observed parameters later. It follows that the coefficient of variation of response, $SE(R_{t/N})/R_{t/N}$ is a function of $\alpha^*$, $\beta$ and $g(q)$, and inversely proportional to $\sqrt{n}$.

(ii) *Evaluation at selection limits*

When all the variation initially present is lost a selection limit is reached in the absence of new mutation. The fixation probability, the mean gene frequency over replicate populations at the limit, is given from the diffusion approximation by

$$u(Ns, q) = [1-\exp(-2Nsq)]/[1-\exp(-2Ns)] \quad (9)$$

Kimura (1957). It has been shown previously that this approximation holds well for truncation selection even in very small populations (Hill, 1969). The expected change and variance of change in gene frequency are

$$\lim_{t\to\infty} E[q_t - q|Ns, q] = u(Ns, q) - q$$

and

$$\lim_{t\to\infty} V[q_t - q|Ns, q] = u(Ns, q)[1-u(Ns, q)]$$

which can be inserted into (7) and (8), respectively.

We first integrate over the distribution of $Ns$ and, to simplify the formulae, let $x = Ns$, and $\psi(q)$ be the function of $q$ in (7) after integrating out $Ns$. Thus

$$\psi(q) = \int_{x=0}^{\infty} [(1-e^{-2xq})/(1-e^{-2x}) - q]\, xf^*(x)\, dx.$$

Following Kimura (1979) in expanding $(1-e^{-2x})^{-1}$ in a series and inserting the gamma distribution for $f^*(x)$,

$$\psi(q) = \int_{x=0}^{\infty} [(1-e^{-2xq})(1+e^{-2x}+e^{-4x}+\ldots) - q] \times [\alpha^{*\beta} e^{-\alpha^* x} x^{\beta}/\Gamma(\beta)]\, dx.$$

Noting that

$$\int_{x=0}^{\infty} x^{\beta} e^{-cx}\, dx = \Gamma(\beta+1)/c^{\beta+1},$$

$$\psi(q) = (\beta/\alpha^*)\left\{1-q+\sum_{j=0}^{\infty}[(1+2(j+1)/\alpha^*)^{-\beta-1} - (1+2(j+q)/\alpha^*)^{-\beta-1}]\right\}. \quad (10)$$

A similar formula can be obtained for the equivalent quantity for the variance of response for insertion into (8), namely

$$\eta(q) = [\beta(\beta+1)/\alpha^{*2}] \sum_{j=0}^{\infty} [-j(1+2j/\alpha^*)^{-\beta-2} + (2j+1)(1+2(j+q)/\alpha^*)^{-\beta-2} - (j+1)(1+2(j+2q)/\alpha^*)^{-\beta-2}]. \quad (11)$$

Because, as $Ns \to 0$, $u(Ns, q) - q$ approaches $Nsq(1-q)$ (Robertson, 1960), it follows that

$$\lim_{\alpha^*\to\infty} \psi(q) = q(1-q)\int x^2 f^*(x)\, dx = q(1-q) \times \beta(\beta+1)/\alpha^{*2} = q(1-q)E[(Ns)^2].$$

For small $Ns$, $u(Ns, q)[1-u(Ns, q)]$ approaches $q(1-q)$, so $\eta(q)$ is proportional to $\psi(q)$. As $Ns$ values increase favourable genes become certain to be fixed, so as $\alpha^* \to 0$, $\psi(q)$ approaches $E[Ns](1-q)$ and $\eta(q)$ approaches 0.

(*a*) $q = 0{\cdot}5$. For the case where $q = 0{\cdot}5$ at al loci, it follows immediately from (7), (8), (10) and (11) that

$$R_{\infty} = (n\sigma/Ni)(\beta/\alpha^*)\left[\tfrac{1}{2} + \sum_{j=1}^{\infty}(-1)^j(1+j/\alpha^*)^{-\beta-1}\right] \quad (12)$$

and

$$V(R_{\infty}) = (n\sigma^2/N^2i^2)[\beta(\beta+1)/\alpha^{*2}] \times \left[\sum_{j=1}^{\infty}(-1)^{j+1} j(1+j/\alpha^*)^{-\beta-2}\right]. \quad (13)$$

(*b*) *q uniform.* For the case of a uniform distribution of initial frequency, integration of (10) and (11) over $g(q) = 1$ in (7) and (8) leads to

$$R_{\infty} = (n\sigma/Ni)(\beta/\alpha^*) \times \left\{\tfrac{1}{2} - \alpha^*/(2\beta) + \sum_{j=0}^{\infty}[1+2(j+1)/\alpha^*]^{-\beta-1}\right\} \quad (14)$$

and

$$V(R_{\infty}) = (n\sigma^2/N^2i^2)[\beta(\beta+1)/\alpha^{*2}] \times \left\{\sum_{j=1}^{\infty} j[1+2j/\alpha^*]^{-\beta-2} + \alpha^*/(\beta+1) \times \left[\tfrac{1}{2} + \sum_{j=1}^{\infty}[1+2j/\alpha^*]^{-\beta-1}\right]\right\}. \quad (15)$$

(*c*) *q with* U *distribution.* Numerical integration was required to evaluate (7) and (8), using Patterson's (1968) method. In order to speed convergence, eq. (10), for example, was evaluated with terms in the summation paired as shown for $q > \frac{1}{2}$, which cancel when evaluated at $q = 1$, while for $q < \frac{1}{2}$ they were paired so as to cancel when evaluated at $q = 0$. Because the prior gene frequency spectrum (5) rather than distribution was used, all results were scaled relative to other quantities, such as initial heterozygosity, using the formula $g(q) \propto 1/[q(1-q)]$ in each case.

(iii) *Evaluation at intermediate generations*

Numerical methods using transition probability matrices were used to evaluate the expressions at intermediate generations. Let $\mathbf{P}(N, Ns)$ denote the square matrix with elements $p_{jk}$ defining the Wright-Fisher stochastic process (Ewens, 1979, Chapter 3).

$$p_{jk} = \binom{2N}{k}(q+\Delta q)^k(1-q-\Delta q)^{2N-k} \quad (0 \leqslant j, k \leqslant 2N)$$

where $q = j/2N$ and $\Delta q = sq(1-q)$ is the expected change in gene frequency, and $s = ia/\sigma$ as before. Let $\mathbf{v}(t, N, Ns)$ denote the vector whose elements $v_{j(t)}$ are

https://doi.org/10.1017/S0016672300024642 Published online by Cambridge University Press

the expected frequencies at generation $t$ of genes with initial frequency $j/(2N)$. Thus

$$v_{j(0)} = j/2N$$

and

$$\mathbf{v}(t, N, Ns) = \mathbf{P}(N, Ns)\,\mathbf{v}(t-1, N, Ns). \quad (16)$$

Iteration was carried out each generation, and $E(q_{t/N} - q)$ is given by the elements of $\mathbf{v}(t/N, N, Ns) - \mathbf{v}(0, N, Ns)$. For some initial gene frequency vector $\mathbf{g}'$, describing $g(q)$, where $g_j$ = Prob (initial frequency = $j/2N$), then

$$\int_{q=0}^{1} E(q_{t/N} - q)\, g(q)\, dq$$

is approximated by

$$\mathbf{g}'[\mathbf{v}(t/N, N, Ns) - \mathbf{v}(0, N, Ns)].$$

Since $v$ was obtained for integral values of $t$, only the corresponding specific values of $t/N$ were used, e.g. $t/N = 0{\cdot}1, 0{\cdot}2, \ldots$ for $N = 10$. Variances of response were computed by iteration on a vector $\mathbf{w}$, where $w_{j(0)} = (j/2N)^2$.

Again, assuming independence, integration over the distribution of selective values, $s$, was done using Simpson's rule on iterated results for a range of $s$ values, typically $s = 0, 0{\cdot}046875, \ldots, 1{\cdot}5$; but in some cases, to improve precision, the range was split into two parts. Convergence was checked by comparison of two successive halvings of the $s$ interval.

In more simple terms the method was to evaluate the expected value of gene fequency for successive generations by transition matrix iteration for a range of selective values, and then to integrate these values over the distribution of selective values. The two stages were done separately so that the iteration results were used for each selective value distribution. This method has the great advantages over Monte Carlo simulation in that the results involve no sampling error and, if several distributions of effects are being used, involve less computation. The inference from the diffusion approximation that the composite parameters $t/N$ and $Ns$ (or the derived parameters $\alpha^*$ and $\beta$) were sufficient, rather than $t$, $N$ and $s$, meant that results needed to be computed only for a single value of population size.Typically this was $N = 20$. Various checks on the approximation were made. The only problems arose when $E(s)$ was large, such that a significant proportion of the $s$ values exceeded 1·5. In such cases larger values of $N$ were used.

(iv) *Alternative parametrizations*

As far as possible it is important to obtain expressions in terms of quantities which can be estimated or guessed at. We have used, so far, the parameters $\alpha^*$ and $\beta$, which are functions of the distribution of $Ns$, and do not satisfy this criterion. Noting, from (1), that with independent gene effects and frequency,

$$h^2 = (n/2\sigma^2)\, E(a^2)\, E[q(1-q)],$$

it follows that

$$E(a^2) = 2h^2\sigma^2/\{nE[q(1-q)]\} = \beta(\beta+1)/\alpha^2$$

and

$$\alpha^* = \alpha\sigma/Ni = [1/(Nih)]\,\{n\beta(\beta+1)\,E[q(1-q)]/2\}^{\frac{1}{2}}. \quad (17)$$

Alternatively, in terms of the effective number of genes, $n_e = n\beta/(\beta+1)$,

$$\alpha^* = [(\beta+1)/(Nih)]\,\{n_e\,E[q(1-q)]/2\}^{\frac{1}{2}}$$
$$= (\beta+1)\{E[q(1-q)]/2\}^{\frac{1}{2}}/A^* \quad (18)$$

where $A^* = Nih/\sqrt{n_e}$. Thus results can be expressed in terms of $A^*$ and $n_e/n$ instead of $\alpha^*$ and $\beta$, the only intangible being $n_e$. The use of $n_e$ rather than $n$ is justified, as we shall see, by the greater robustness of some results. If all gene frequencies are 0·5, $(\beta+1)/\alpha^* = A^*\sqrt{8}$ and if they are uniformly distributed, $(\beta+1)/\alpha^* = A^*\sqrt{12}$. This reparametrization is less meaningful for the mutation-derived U-shaped distribution.

As a reference point, it is convenient to express results in terms of the expected response, $R_1$, in the first generation of selection, further scaled by population size. From (7), noting that $R_1 = ih^2\sigma$

$$R_{t/N}/(NR_1) = n/(N^2i^2h^2) \times \text{function of } (\alpha^*, \beta, g(q)).$$

Noting further that $N^2i^2h^2/n$ is a function of $A^*$ and $\beta$, it follows that

$$R_{t/N}/(NR_1) \text{ is a function of } (A^*, \beta \text{ and } g(q)).$$

Similarly $SE(R_{t/N})/(NR_1)$ is a function of these parameters and is inversely proportional to $n_e$. These terms express response and its standard error in relation to the initial response. A useful reparametrization which expresses the response (or its standard error) in absolute terms, is $R_{t/N}/(h\sigma/\sqrt{n_e})$ which is seen from (7) to be a function of the same parameters as $R_{t/N}/(NR_1)$.

## 4. Results

(*a*) $q = 0{\cdot}5$. For simplicity and as a reference point, we consider this case first. The asymptotic response (when all genes are fixed) is shown in Fig. 2. At low values of $A^* = Nih/\sqrt{n_e}$, when $N_s$ (population size × selective values) are small for all loci, the response is given by $2Nih^2\sigma$ (Robertson, 1960), so when expressed as in the graph $R_\infty/(h\sigma\sqrt{n_e}) \to 2A^*$. At high values of $A^*$, $R_\infty/(h\sigma\sqrt{n_e}) \to \sqrt{2}$ for all distributions of gene effects, as all favourable genes become fixed. A log scale is used for the plot of $A^*$, so the effects of a doubling of population size, for example, give the same increase in the abscissa over the whole range. Whilst increase in population size gives a linear increase in response at low values of $A^*$, for $A^* > 10$ the effect of an increase in population size is very small for any distribution of gene effects. If all effects are equal ($\beta \to \infty$) little increase is achieved after $A^* = 1$, further gains are

https://doi.org/10.1017/S0016672300024642 Published online by Cambridge University Press

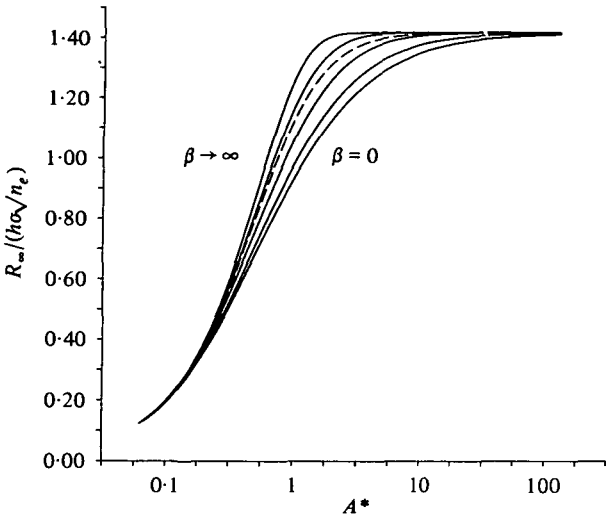


Fig. 2. Response at fixation, expressed as $R_\infty/(h\sigma\sqrt{n_e})$, for initial frequency 0·5 at all loci plotted against $A^* = Nih/\sqrt{n_e}$, for a range of $\beta$ values with gamma distribution (solid lines) and for normal distribution (dashed lines). Values of $R_\infty/h\sigma\sqrt{n_e})$ increase monotonically for the values of $\beta$ shown, namely $\beta = 0{\cdot}0$, 0·25, 1·0, 4·0 and $\to \infty$.

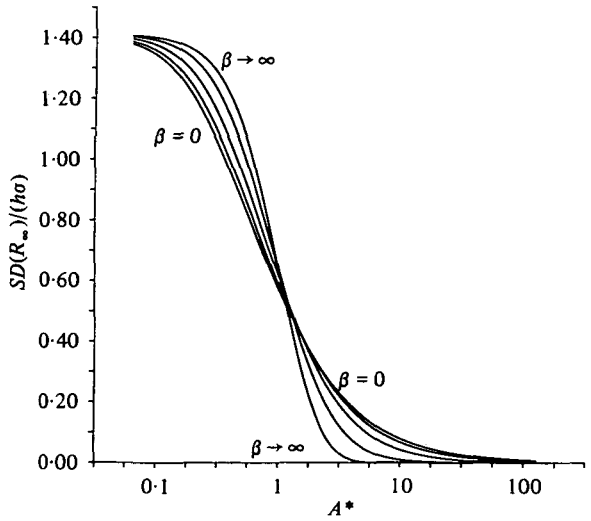


Fig. 3. Standard deviation of response at fixation expressed as $SD(R_\infty)/(h\sigma)$, for initial frequency 0·5 at all loci plotted against $A^* = Nih/\sqrt{n_e}$ for a range of $\beta$ values. Coefficients of variation of response can be deduced from Figs. 2 and 3 and are inversely proportional to $\sqrt{n_e}$. Values of $SD(R_\infty)/(h\sigma)$ increase monotonically with $\beta$ for low $A^*$ for the values shown, namely $\beta = 0{\cdot}0$, 0·25, 1·0, 4·0 and $\to \infty$ and decrease monotonically with $\beta$ for high $A^*$.

made for the most extreme case ($\beta \to 0$) up to $A^* = 10$ and a little beyond. Results for the normal distribution are shown to fall between $\beta = 1$ and $\beta = 4$ for the gamma distribution, and it is obvious that they do not differ much from those for $\beta = 2$ (not shown for clarity); examples of the normal are not given subsequently for this reason. In general, reparametrization leads to rather small differences according to $\beta$; for this case of $q = \frac{1}{2}$ the differences are generally larger than for other gene frequency distributions.

Let us consider an example to put the parameters into focus. For a typical trait with $h^2 = 0{\cdot}36$ and typical mass selection intensity of $i = 1{\cdot}67$, $A^* = N/\sqrt{n_e}$. Assume an exponential distribution of gene effects, $\beta = 1$ and therefore $N_e = n/2$, and that, say 200 genes affect the trait to some extent. Thus $A^* = N/10$ and for all but experimental populations $N$ is typically 50 or more, giving $A^* > 5$ with all useful variation eventually fixed and a final limit of almost $1{\cdot}4\,h\sigma\sqrt{n_e} \sim 8\sigma$.

The standard deviation of the asymptotic response among conceptual replicate populations selected from the same base is shown in Fig. 3. It is expressed in terms of $h\sigma$, and $SD(R_\infty)$ ranges from $1{\cdot}4h\sigma$ if $A^*$ is very small down to approaching zero as $A^*$ increases. For $A^* = 5$, $\beta = 1$ and $n_e = 100$, it is seen that $SD(R_\infty) = 0{\cdot}1h\sigma = 0{\cdot}06\sigma$ for $h^2 = 0{\cdot}36$. Thus the coefficient of variation of response, $CV(R_\infty)$, among replicates would be very small, of the order of $0{\cdot}06/8 \sim 0{\cdot}7\%$. It follows that $CV(R_\infty)$ can only be substantial when both $A^*$ and $n_e$ are small, i.e. $Nih$ is very small, implying in most situations a very small population size. The parameter of the distribution of gene effects, $\beta$, has a minor but not negligible influence on the variance. When $A^*$ is small variability is higher when

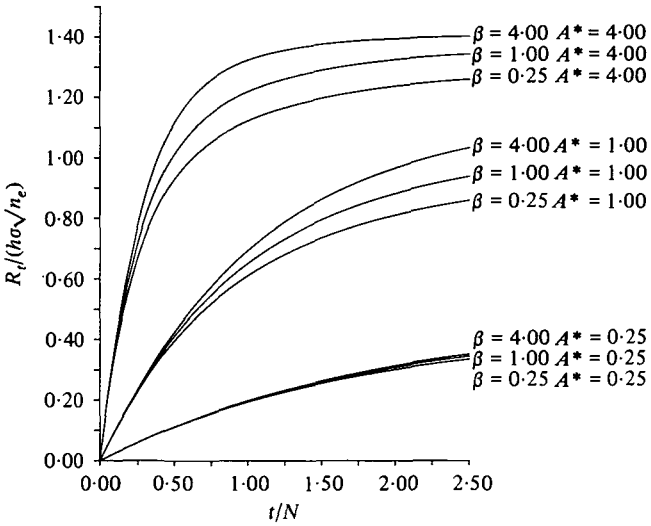


Fig. 4. Response, expressed as $R_t/(h\sigma\sqrt{n_e})$ for initial frequency 0·5 at all loci, plotted against generations, expressed as $t/N$, for a range of values of $A^*$ and $\beta$.

$\beta$ is small, suggesting consequences of non-certain fixation of genes of large effect. When $A^*$ is large the reverse is true, suggesting the residual source of variation is due to non-certain fixation of genes of small effect.

The expectation and standard deviation of response prior to fixation are shown in Figs. 4 and 5 respectively for three widely different $A^*$ values and, although a narrower range of $\beta$ values than given in the previous figures, probably wide enough to include the situation in nature. In both cases the value of $\beta$ (i.e. the relation between $n_e$ and $n$) is seen not to be very critical, whereas that of $A^*$ affects the limit, the

https://doi.org/10.1017/S0016672300024642 Published online by Cambridge University Press

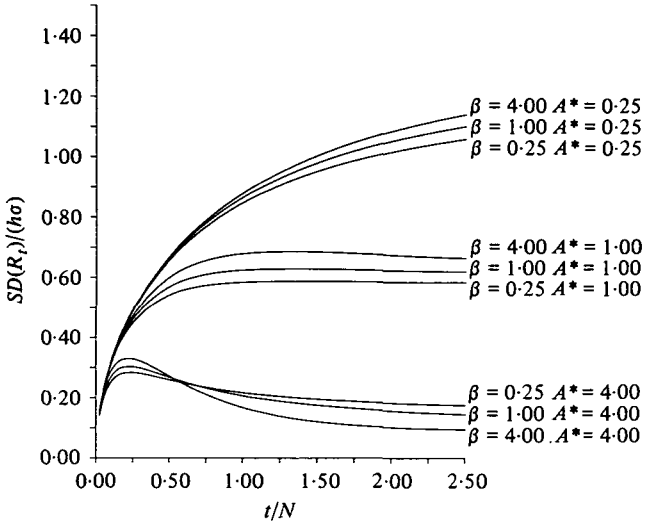


Fig. 5. Standard deviation of response, expressed as $SD(R_t)/(h\sigma)$, for initial fequency 0·5 at all loci, plotted against $t/N$ for a range of $A^*$ and $\beta$ as in Fig. 4. Further explanation in caption to Fig. 3.

rate of approach to the limit and the variation of response. The exception is in the early generations, when $V(R_t) \sim th^2\sigma^2/N$ and $SD(R_t/h\sigma) \sim \sqrt{(t/N)}$ for all values of parameters. The half-lives of the response, for $\beta = 1$, are approximately $0{\cdot}2N$, $0{\cdot}75N$ and approaching $1{\cdot}1N$ for $A^* = 4$, 1 and 0·25, respectively, the latter value being that for the case of very small $N_s$ values when pure drift predominates (Robertson, 1960). In this case of initial frequencies of 0·5, the variation among lines increases and then decreases over generations if $A^*$ is large, as useful genes approach fixation in all replicates; this is not a phenomenon found for the other distributions of gene effects considered. The simple, pure drift, formula of $V(R_t) = th^2\sigma^2/N$ applies for only about $0{\cdot}1N$ generations at $A^* = 4$, i.e. only 5 generations or so for our previous example with $N = 50$.

(*b*) *Other distributions.* No problems arise in parametrising the initial uniform distribution in terms of $A^*$ in the same way as for the case of a gene frequency of 0·5 at all loci, but when we consider the U-shaped case this is not possible because of the problems of defining the actual or effective number of genes. It is therefore convenient to use an alternative scaling, in terms of $(\beta+1)/\alpha^*$, which is a measure of the distribution of $Ns$ (population size × selective values), and the response to selection in the first generation, which equals $R_1 = ih^2\sigma$ and is not dependent on the gene frequency and effect distributions. The parametrization $(\beta+1)/\alpha^*$ rather than $E(Ns) = \beta/\alpha^*$ or $[E(Ns)^2]^{\frac{1}{2}} = [\beta(\beta+1)]^{\frac{1}{2}}/\alpha^*$ was used as it appears to remove most differences among the models.

Plots of the expected response at fixation are given in Fig. 6. For small values of $(\beta+1)/\alpha^*$ the limit approaches $2NR_1$. The graphs differ very little among the distributions for values of $(\beta+1)/\alpha^*$ less than about 0·5 and the slopes of the curves are very similar

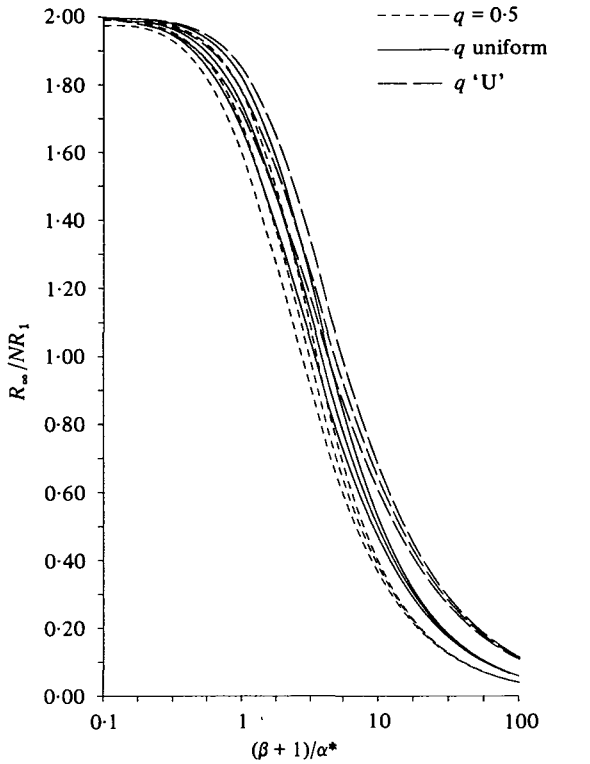


Fig. 6. Response at fixation, expressed in terms of the response in the first generation as $R_\infty/(NR_1)$, plotted against $(\beta+1)/\alpha^*$ for alternative values of $\beta$ (0.25, 1·0, 4·0) and initial frequency distributions, 0·5 at all loci (---), uniform (—) and U-shaped (––). In each case values of $R_\infty/(NR_1)$ are highest for $\beta = 4{\cdot}0$, intermediate for $\beta = 1{\cdot}0$ and lowest for $\beta = 0{\cdot}25$.

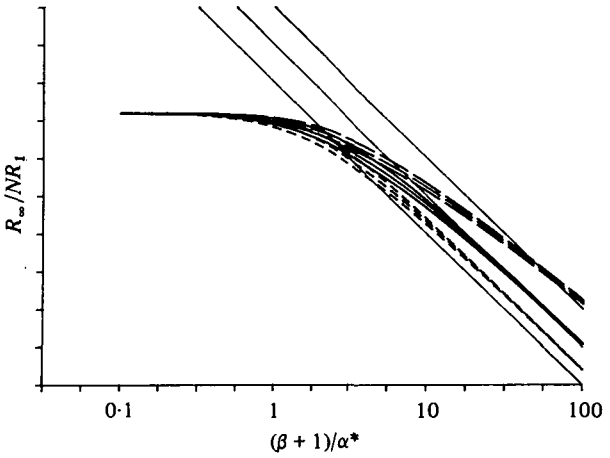


Fig. 6*a*. As Fig. 6, but using a logarithmic plot for both axes. Lines of slope −1 are also drawn.

over the whole range. The magnitude of $\beta$ is unimportant in all cases. Consider now the consequences of a doubling of population size from $N = 50$ to 100 with $\beta = 1$: for $(\beta+1)/\alpha^* = 5$, the values of $R_t/(R_1 N)$ are 0·8, 0·95 and 1·1 for $q = 0{\cdot}5$, uniform and U-shaped respectively, and for $(\beta+1)/\alpha^* = 10$ they are 0·4, 0·5 and 0·65 respectively, which when doubled to allow for the change in $N$ are 0·8, 1·0 and 1·3 implying, respectively, no further increase, a small increase and a distinct increase in response, respectively. This is seen more clearly in Fig. 6*a* using a log–log plot. A

https://doi.org/10.1017/S0016672300024642 Published online by Cambridge University Press

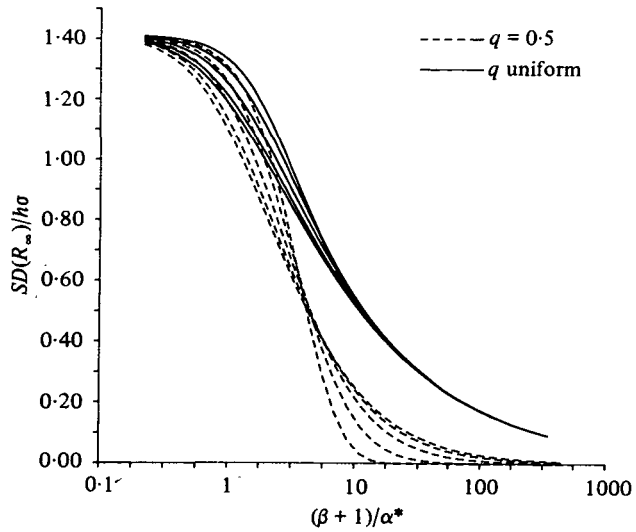


Fig. 7. Standard deviation of response at fixation, expressed as $SD(R_\infty)/(h\sigma)$, for gene frequency either 0·5 at all loci or uniformly distributed, plotted against $(\beta+1)/\alpha^*$ for a range of $\beta$ values (0, 0·25, 1·0, 4·0 and $\to \infty$). For both gene frequency distributions $SD(R_\infty)/(h\sigma)$ increases monotonically with $\beta$ when $(\beta+1)/\alpha^*$ is small.

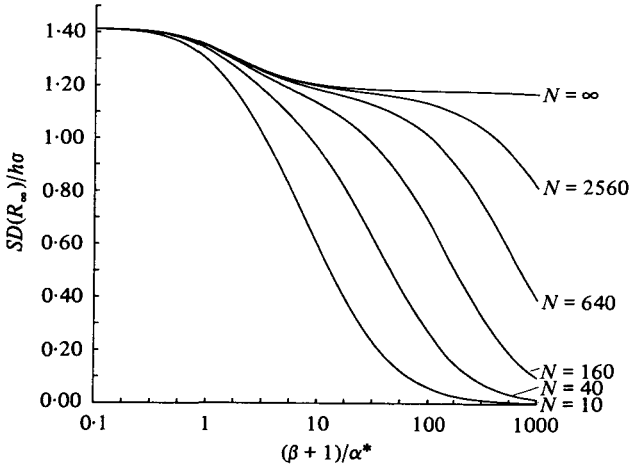


Fig. 8. Standard deviation of response at fixation, expressed as $SD(R_\infty)/(h\sigma)$, for U-shaped gene frequency distribution with $\beta = 1$ plotted against $(\beta+1)/\alpha^*$ for a range of $N$ values.

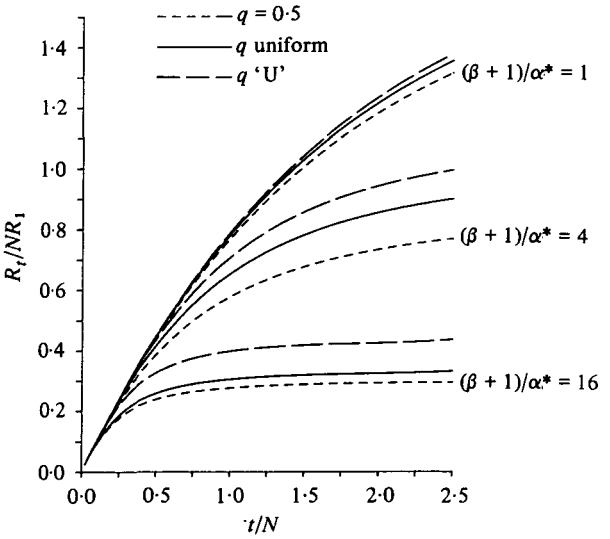


Fig. 9. Response in terms of that in the first generation, expressed as $R_t/(NR_1)$, plotted against generations, expressed as $t/N$, for $\beta = 1$, alternative initial frequency distributions (as Fig. 6) and values of $(\beta+1)/\alpha^*$.

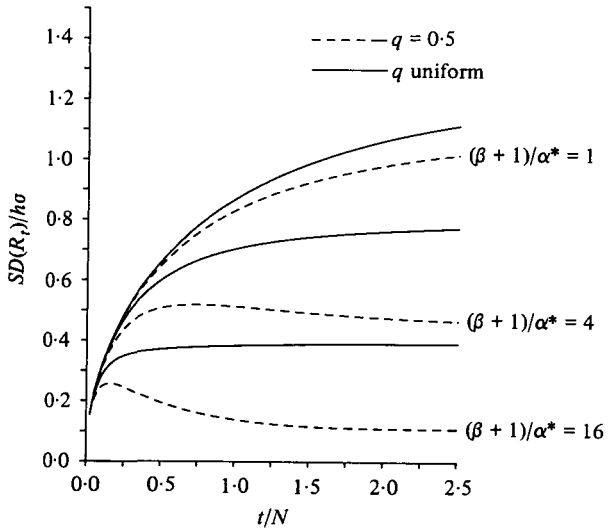


Fig. 10. Standard deviation of response, expressed as $SD(R_t)/(h\sigma)$, plotted against generations, expressed as $t/N$, for gene frequency 0·5 at all loci and uniformly distributed, $\beta = 1$ and a range of values of $(\beta+1)/\alpha^*$.

doubling of population size (or strictly $Ni$) leads to an increase in response providing the slope of the line is less negative than $-1$. This graph shows that, in theory, an increase in population size continues to increase response in the U-shaped case, by roughly 10% for each doubling at high values of $Ni$.

The variance of response at the limit has different properties for the U-shaped case, so we first contrast just the $q = \frac{1}{2}$ and uniform models, results being given in Fig. 7. It is, however, in all cases most convenient to express values in terms of $h\sigma$. Whilst at low values of $(\beta+1)/\alpha^*$ the results are similar for $q = \frac{1}{2}$ and uniformly distributed, they diverge widely for $(\beta+1)/\alpha^*$ values in excess of unity, especially when $\beta$ is large (i.e. genes of equal effect). For gene frequencies of 0·5, fixation of all the favourable genes is occurring in all lines, whereas it is not in the uniform case. Contrast also the expected responses and their standard deviation for the two distributions: as seen in Fig. 6, for $(\beta+1)/\alpha^* = 10$ there is only a 20% or so greater expected response in the uniform case, but a three or fourfold greater standard deviation of this response.

These differences are exaggerated for the U-shaped distribution (Fig. 8). The first problem is that the values are population size dependent in contrast to the other distributions and in apparent conflict with the diffusion approximation. The reason is that the maximum height of the frequency density is at extreme gene frequencies, notably at $1/(2N)$, when the fixation probability approaches $2Nsq$ for large $Ns$, and $s$ for

https://doi.org/10.1017/S0016672300024642 Published online by Cambridge University Press

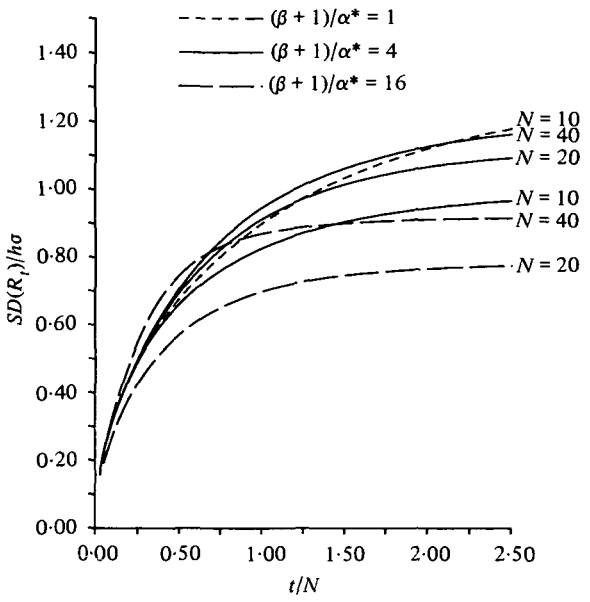


Fig. 11. Standard deviation of response, expressed as $SD(R_t)/(h\sigma)$, plotted against generations, expressed as $t/N$, for U-shaped gene frequency distribution, $\beta = 1$ and a range of values of $(\beta+1)/\alpha^*$ and $N$.

$q = 1/(2N)$. Such genes contribute almost all the variance of response as $(\beta+1)/\alpha^*$ becomes larger. Some of these peculiarities of the U-shaped distributions are not of practical importance, however, for at finite values of $N$, such as used in most breeding programmes or experiments, the variance rapidly declines as $(\beta+1)/\alpha^*$ is increased.

Examples of the expected response during the selection process are given in Fig. 9 for the three distributions, in each case with $\beta = 1$. The pattern is seen to be rather similar, although always ranking U > uniform > 0·5 in terms of the response at any generation relative to that in the first (or equivalently in terms of $h\sigma$ for the same $i$ value). Corresponding curves for the standard deviation are shown in Figs. 10 and 11. The differences in $SD$ between the distributions are seen to occur quite early when $(\beta+1)/\alpha^*$ is larger, and the '$N$-dependency' in the U-shaped case is clearly seen.

## 5. Discussion

The utility of these analyses and results depends on two fundamental assumptions: that the models are relevant and that reasonable assumptions about the necessary parameters can be made.

The models are undoubtedly over simplistic, in that dominance, epistasis, linkage, natural selection, mutation and multiple alleles are ignored, but they are intended to be a starting point. Infinitely many models of epistasis can be set up but we have little knowledge on which to construct them. In populations initially in linkage equilibrium and pertaining to mammals or birds, previous analyses have shown linkage effects on limits are small (Robertson, 1970); and even in the case of a line cross very tight linkage will be required to make much difference, although this needs to be investigated further. Natural selection has been ignored in the analysis: if its effects are as a different trait at the level of individual genes with selective values remaining constant, this merely changes the relationships between heritability, gene number and selective value; if its effects are dependent on the mean level of the populations, e.g. by stabilizing selection effects, further analysis is required (Zeng & Hill, 1986). Mutation has important implications on the role of population size in long term selection, but the theory for incorporation of mutation into the general models of this paper will be given in a later paper. The analysis of multiple alleles will also be deferred, as resort has to be made to simulation because the necessary analytical tools are not available.

The shape of the distribution of gene effects has been assumed to be of gamma form, but since it can take a wide range of shapes (Fig. 1) and since for most results this shape did not have an important effect, we need not trouble about this. The gene frequency distributions can arise in well defined situations; their deficiency is that each is symmetric about 0·5 and imply no selection on the trait prior to the analysis being undertaken. It is clear that if the mean frequency of the favourable alleles is greater than 0·5, then changes in population size (or $Ni$) will have less effect than when the mean is less than 0·5. Thus we can assume that if there is any relevant selection history or if there is any history of population bottlenecks, such that the U-shaped distribution discussed here will have lost some of the weight from the extreme frequency values, our results will exaggerate the effects of population size on long term response. The independence of gene frequencies and effects also implies no prior selection: if there has been such selection for the trait there is likely to be a positive correlation between gene effect and frequency, again reducing the effects of population size in a subsequent selection programme.

Some of the parameter values, namely $N$, $i$, $h^2$ and $\sigma$ cause few problems in estimation. Much attention is given to estimating heritability, and an error of a factor of two would be considered large, but it is nothing as compared to the guesswork in estimating numbers or effective numbers of genes. In view of the many pathways that lead to final product and of the many different degrees of primary effect of a gene substitution on, say, its properties as an enzyme, we must expect a very wide range of effects of genes on any metric trait. Further it seems reasonable to assume that most genes have at least a little effect on most traits. Thus we subscribe to models in which $n$ is relatively large, approaching the number of genes in the genome, and $n_e$ very much smaller, i.e. nearer $n_e/n = \frac{1}{5}$, corresponding to $\beta = \frac{1}{4}$ than to $n_e/n = \frac{1}{2}$, corresponding to $\beta = 1$. Thus we are unlikely to have effective numbers



https://doi.org/10.1017/S0016672300024642 Published online by Cambridge University Press

of genes of more than 400 and perhaps no more than 50. Estimates of $n_e$ from line crosses and selection experiments abound, but they can be objected to for various reasons. Wright's (1952) method leads to underestimates through linkage, as does the assay method of Jinks & Towey (1976); estimates from chromosome analysis (Thoday, 1961) are highly dependent on the work put into them; and analyses of selection experiments depend on assumptions that all useful genes have been fixed (Falconer, 1981). In contrast, estimates of numbers in segregating populations from selection experiments (Dudley, 1977) may be biased by subsequent mutation, likely in the Illinois corn experiment analysed by Dudley (Hill, 1982).

Fortunately, number of loci enter the formula as $\sqrt{n_e}$, so we are unlikely to have to consider values outside the range $4\sqrt{n_e} < 20$, only a fivefold range! On this basis, typical values of $A^* = N_i h/\sqrt{n_e}$ are around $0{\cdot}1N$. Our results suggest that for $A^*$ values in excess of 5, further increases in population size do not have large effect. This of course applies to the $q = \frac{1}{2}$ and uniform cases; that of the U-shaped distribution has to be considered rather differently.

Finally, it should be re-emphasised that we are dealing here with use of the existing variation. As has been shown earlier (Hill, 1982), with increasing generation number, mutation becomes of increasing importance.

This work was supported by a grant from the Agricultural and Food Research Council.

## References

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper & Row.

Dudley, J. W. (1977). Seventy-six generations of selection for oil and protein percentage in maize. In *Proceedings of the International Conference on Quantitative Genetics* (ed. E. Pollak, O. Kempthorne and T. B. Bailey, Jr.), pp. 459–473. Ames: Iowa State University Press.

Ewens, W. J. (1979). *Mathematical Population Genetics*. Berlin: Springer-Verlag.

Falconer, D. S. (1981). *Introduction to Quantitative Genetics*, 2nd edn. London: Longman.

Hill, W. G. (1969). On the theory of artificial selection in finite populations. *Genetical Research* **13**, 143–163.

Hill, W. G. (1982). Predictions of response to artificial selection from new mutations. *Genetical Research* **40**, 255–278.

Jinks, J. L. & Towey, P. (1976). Estimating the number of genes in a polygenic system by genotype assay. *Heredity* **37**, 69–81.

Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences USA* **41**, 144–150.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *Annals of Mathematical Statistics* **28**, 882–901.

Kimura, M. (1979). Model of effectively neutral mutations in which selective contrast is incorporated. *Proceedings of the National Academy of Sciences USA* **76**, 3440–3444.

Kimura, M. & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Patterson, T. N. L. (1968). The optimum addition of points to quadrature formulae. *Mathematical Computing* **22**, 847–856.

Robertson, A. (1960). A theory of limits in artificial selection. *Proceedings of the Royal Society of London* B **153**, 234–249.

Robertson, A. (1970). A theory of limits in artificial selection with many linked loci. In *Mathematical Topics in Population Genetics* (ed. K. Kojima), pp. 246–268. Berlin: Springer-Verlag.

Thoday, J. M. (1961). Location of polygenes. *Nature, London* **191**, 368–370.

Wright, S. (1952). The genetics of quantitative variability. In *Quantitative Inheritance* (ed. E. C. R. Reeve and C. H. Waddington), pp. 5–41. London: Her Majesty's Stationery Office.

Zeng, Z.-B. & Hill, W. G. (1986). The selection limit due to conflict between truncation and stabilizing selection with mutation. *Genetics* (submitted).