Genet. Res., Camb. (1969), **13**, *pp.* 117–126
Printed in Great Britain

A comparison of two methods for predicting changes in the distribution of gene frequency when selection is applied repeatedly to a finite population

BY DEREK J. PIKE

Department of Applied Statistics, University of Reading

(*Received* 24 *April* 1968)

1. INTRODUCTION

Over the last ten years a number of methods have been proposed for predicting changes in the distribution of gene frequency when sampling and selection are applied to finite populations. This paper is an attempt to compare two of these methods—those proposed by Robertson (1960) and Curnow & Baker (1968).

Kimura (1957) treated changes in gene frequency as stochastic processes, and described these processes by equations similar to Fokker–Planck diffusion equations. In this approach Kimura assumed completely overlapping generations and a continuous distribution of gene frequency. Robertson (1960) used Kimura's approach to derive a theory of limits in artificial selection. Robertson's paper also used probability transition matrices in this field for the first time. These matrices had elements which were the probabilities of changing from one gene frequency to another in a single generation of sampling and selection. Multiplication of a transition matrix by the probability distribution vector of the gene frequency in one generation gives the vector for the next generation. Allan & Robertson (1964), Hill & Robertson (1966) and Ewens (1963) have used the concept of transition matrices to investigate a number of selection problems.

Kojima (1961) derived formulae for the expected gain—to the first order in single-locus genetic effects—from a cycle of sampling and selection applied to a population of finite size. This involved deriving approximate formulae for the mean and variance of the change of gene frequency. Curnow & Baker (1968) avoided the use of transition matrices by assuming at each generation a beta distribution for the gene frequency. This is a distribution lying strictly between 0 and 1, with discrete probabilities added at the end-points to represent the fixation probabilities of the two alleles. Using this distribution together with Kojima's equations, Curnow & Baker's method provides a means of calculating the genetic mean and variance, the expected gene frequency and the fixation probabilities after each generation of selection. The parameters of the beta distribution at each generation are chosen to make its mean and variance equal to the mean and variance of the gene frequency distribution. The particular beta distribution used will therefore change from generation to generation.

Although Robertson's approach is analytically more exact than that of Curnow & Baker, it tends to be more demanding of the computer, both in store required and

https://doi.org/10.1017/S0016672300002834 Published online by Cambridge University Press

time taken, especially when a large number of individuals are selected. The time consuming computations are the calculation of the transition matrix, and its multiplication by vectors of gene frequency. The Curnow–Baker approach avoids these computations, resulting in a reduction of computer time. Furthermore, transition matrices cannot be utilized in situations where there are unequal numbers of males and females. Special difficulties are encountered when there is an infinite number of one sex and selection is practised on the other sex—as in Baker & Curnow (1968). In this paper Baker & Curnow consider an infinite population of males which at each generation is produced from the selected parents of the previous generation. In this situation the transition matrix process is not directly applicable. Also it would appear that transition matrix methods are difficult to apply in the more general situation of selection of unequal numbers of males and females.

This paper is an attempt to compare the genetic values obtained by the two methods for a range of selection intensities, population sizes and initial gene frequencies; and thereby to assess the range of validity of the Curnow–Baker method. Essentially Curnow & Baker are attempting to fit a particular form of continuous distribution to a set of discrete probability values. The number of these values is dependent solely upon the number of individuals selected at each stage. Consequently we should expect that if more individuals are selected the beta distribution will provide a better approximation to the discrete probabilities. Now we must decide how many must be selected before we are prepared to accept the degree of approximation inherent in the Curnow–Baker method.

2. THEORY

The theory underlying the two methods is more or less exactly as given in Kojima (1961), Robertson (1960) and Curnow & Baker (1968). An outline of the theory is given here together with a proof of the validity of a substitution made by Kojima.

For the transition matrix approach we construct a transition probability matrix, T, with elements corresponding to changes in gene frequency from generation to generation. We have a finite population consisting of N dipoid individuals sampled from an infinite population. From these we select n individuals. The elements (t_{ij}); $i,j = 0, 1, 2, \ldots, 2N$; of the transition matrix are assumed to be time independent and represent the probability that there are j A alleles among the n selected individuals in one generation, conditional upon there having been i in the previous generation. The model used to determine the transition probabilities is a diploid one with the probability of selecting any particular combination of genotypes being multinomial. Of the n selected individuals n_1 are AA, n_2 are Aa and n_3 are aa. Thus in the case of complete dominance

$$t_{ij} = \Sigma \binom{n}{n_1 n_2 n_3} [q^2\{1+2s(1-q)^2\}]^{n_1} [2q(1-q)\{1+2s(1-q)^2\}]^{n_2} [(1-q)^2\{1-2sq(2-q)\}]^{n_3},$$

where s is the selective value of the A allele, $q = i/2n$ and the summation is carried out over n_1, n_2, and n_3 conditional upon $2n_1+n_2 = j$ and $n_1+n_2+n_3 = n$.

Hill (1968) provides a more exact transition matrix approach to the study of selection in finite populations. He states that if in a finite population of size N there are N_1 of type AA, N_2 of type Aa, and N_3 of type aa, then the probability of selecting n individuals with n_1 of type AA, n_2 of type Aa and n_3 of type aa is not multinomial but hypergeometric. He then computes a transition probability matrix, in a similar way to that shown above, and carries out the calculations using this matrix. This approach is not used here since there are certain numerical integration problems connected with it and it does not produce appreciably different results from more approximate methods, as Hill shows.

The theory of the Curnow–Baker method will now be summarized. Using Kojima's notation we assume an initial infinite population of diploid individuals and that the character, Y, on which selection is to be based is normally distributed over this population with frequency function $\phi(Y)$. In this population the three genotypes AA, Aa, aa have mean values a, d, $-a$ and occur with frequencies U_1, U_2, U_3. We shall assume a random mating population. The values of the character for each of the three genotypes is assumed to be normally distributed—AA with frequency function $\phi_1(Y)$, Aa with frequency function $\phi_2(Y)$, and aa with frequency function $\phi_3(Y)$. The overall distribution, $\phi(Y)$, can also be assumed approximately normal because the three constitutent genotypic distributions are assumed normal and are nearly coincident; and the mixture of three nearly coincident normal distributions is also approximately normal. The origin and scale have been chosen so that $\phi(Y)$ has zero mean and unit variance. The quantities d_1, d_2, d_3 are the deviations of the means of the distribution of AA, Aa, aa from the overall mean, given in standard deviation units.

If a random sample of N observations from $\phi(Y)$ is ranked, the frequency function of the $(n+1)$th highest value, Y_0, is

$$f(Y_0) = \frac{N!}{n!(N-n-1)!}P^n(1-P)^{N-n-1}\phi(Y_0),$$

where

$$P = \int_{Y_0}^{\infty}\phi(Y)\,dY = \Phi(-Y_0).$$

Writing Pr for the 'the probability that',

$$\Pr[\text{best } n \text{ are } n_1AA_1,\ n_2Aa,\ n_3aa \mid (n+1)\text{th} = Y_0]$$

is multinomial, with parameters n and p_i where

$$p_i = \frac{U_iP_i}{P} \quad\text{and}\quad P_i = \int_{Y_0}^{\infty}\phi_i(Y)\,dY.$$

Following Kojima we can show that

$$E(n_i) = nU_i[1+kd_i],$$

where

$$k = E\left[\frac{\phi(Y_0)}{P}\right] = \int_{-\infty}^{\infty}\frac{\phi(Y_0)}{P}f(Y_0)\,dY_0 = \int_{-\infty}^{\infty}\frac{N!}{n!(N-n-1)!} \times p^{n-1}(1-P)^{N-n-1}\phi^2(Y_0)\,dY_0.$$

Kojima substitutes for k the mean of the top n order statistics in a sample of size N from a standardized normal distribution. The validity of this substitution can be proved as follows. The mean of the top n order statistics in a sample of size N is

$$\mu = \frac{1}{n}\sum_{r=N-n+1}^{N}\int_{-\infty}^{\infty}\frac{N!}{(r-1)!\,(N-r)!}P^{N-r}(1-P)^{r-1}Y\phi(Y)\,dY.$$

To prove Kojima's statement we must show that if $\phi(Y)$ is the frequency function of a standard normal variate, then $k = \mu$. The proof is as follows. Consider the following summation in the expression for μ:

$$\sum_{r=N-n+1}^{N}\frac{N!}{(r-1)!\,(N-r)!}P^{N-r}(1-P)^{r-1} = N\sum_{r-1=N-n}^{N-1}\binom{N-1}{r-1}P^{N-r}(1-P)^{r-1}.$$

From Kendall & Stuart (1963, §5.7), the above expression is the remainder after the first $N-n-1$ terms of the binomial expansion, and this is

$$N\int_0^1\frac{(1-P)^{N-n}\,(1-t)^{N-n-1}}{(N-n-1)!}\;\frac{(N-1)!}{(n-1)!}\;[P+(1-P)t]^{n-1}\,dt.$$

Putting $t = 1-x/(1-P)$ this reduces to

$$\frac{N!}{(N-n-1)!\,(n-1)!}\int_0^{1-P}x^{N-n-1}(1-x)^{n-1}\,dx.$$

Therefore, on substitution

$$\mu = \frac{1}{n}\int_{-\infty}^{\infty}Y\phi(Y)\,\frac{N!}{(N-n-1)!\,(n-1)!}\left\{\int_0^{1-P}x^{N-n-1}(1-x)^{n-1}\,dx\right\}dY.$$

Integrating by parts, using $\int Y\phi(Y)\,dY = -\phi(Y)$, we obtain

$$\mu = -\frac{1}{n}\left[\left\{\int_0^{1-P}x^{N-n-1}(1-x)^{n-1}\,dx\right\}\frac{N!}{(N-n-1)!\,(n-1)!}\phi(Y)\right]_{-\infty}^{\infty}$$

$$+\frac{1}{n}\int_{-\infty}^{\infty}\phi(Y)\,\frac{N!}{(N-n-1)!\,(n-1)!}(1-P)^{N-n-1}P^{n-1}\left(-\frac{dP}{dY}\right)dY.$$

The first term is zero since $\phi(\infty) = \phi(-\infty) = 0$; and $-dP/dY = \phi(Y)$. Therefore

$$\mu = \int_{-\infty}^{\infty}\frac{N!}{n!(N-n-1)!}P^{n-1}(1-P)^{N-n-1}\phi^2(Y)\,dY = k;$$

and the proof is complete.

Having proved that $k = \mu$ when the underlying distribution is normal, it is interesting to consider what information we can obtain concerning the nature of $\phi(Y)$ by assuming $k = \mu$. For example, does it imply normality? It has been shown that the additional assumption that $\phi(Y)$ is symmetric implies that $\phi(Y)$ is in fact normally distributed.

3. COMPUTATIONAL PROCEDURE

Computer programs were written to carry out both the transition matrix and the Curnow–Baker calculations. The print-out from each program contained the genetic mean and variance and the expected gene frequency; and also the probabilities of fixation of the a and A alleles at each generation. It was on the basis of these results that the comparisons were made. In addition, the first two moments of the gene frequency distribution for the transition matrix method were taken and the first two standardized cumulants (skewness and kurtosis) were calculated from them assuming that the distribution was a beta distribution. These cumulants were then compared with the actual values of skewness and kurtosis obtained from the transition matrix method. From this we should be able to see how well the assumption of a beta distribution for the distribution of gene frequency fitted the transition matrix results.

The sets of calculations carried out were as follows:

Number selected, n	4	8
Sample size, N	8, 16, 32, 48	16, 32, 48
Initial gene frequency, q	0·25, 0·50, 0·75	0·25, 0·50, 0·75

All selections were continued for fifteen generations. The genetic effects at the locus were considered to be either additive or completely dominant. The difference between the mean values of the two homozygotes in phenotypic standard deviation units was taken as 0·13 as in Baker & Curnow. Selection of two individuals was not carried out as it required the Curnow–Baker method to fit a beta distribution to three discrete values and so was almost bound to give inaccurate results. Selection of more than eight individuals was contemplated, but this work was carried out on an Elliott 803 computer and considerable time would have been required to consider selection of more than ten or fifteen individuals.

4. RESULTS AND DISCUSSION

Tables 1–8 show the genetic means and variances calculated by the two methods. Before attempting a comparison between the transition matrix and the Curnow–Baker results, a brief general picture of them is given. Graphs drawn from the results for the two methods showed little difference in general trend.

In the additive case all selections show an increasing genetic mean levelling off by fifteen generations. In the case of complete dominance, an initial drop in genetic mean is observed due to the effect of inbreeding depression. On continuing the selection process, or on performing more intense selection, it is seen that the inbreeding depression is overcome by the selection causing the genetic mean to begin to rise again. The genetic variance shows a trend similar to that for the additive genetic mean. There is a difference between the additive and complete dominance cases in that for dominance a much wider range of first generation variances is observed, for the three gene frequencies studied, than in the additive case. By the time fifteen generations have been reached, however, the range of values is comparable for the two cases.

In comparing the results for the transition matrix and Curnow–Baker methods it is best to consider the results for selection of four individuals and for selection of eight separately. The only difference between the additive and complete dominance cases, in terms of comparison, is that there is slightly more discrepancy between the two methods in the case of complete dominance. The significant point to be made, however, is the reduction in the discrepancy for selection of eight individuals compared with the selection of four.

For all the results presented here for selection of four individuals there is very close agreement between the two methods for selection with an initial gene frequency of 0·50. This is true for both genetic mean and variance. For initial gene frequency of 0·25 the Curnow–Baker method gave low values of the genetic mean and high values of genetic variance; with sometimes a discrepancy of 10 % or more. For initial gene frequency of 0·75 it gave high values of the genetic mean and low values of genetic variance; with discrepancies less than half those for 0·25. The reason for this becomes apparent as soon as we look at the probabilities of fixation, for here we see what is perhaps the only failing of the Curnow–Baker method. Too high a proportion of the genes is fixed, particularly those near to fixation anyway, because of the fitting of a continuous distribution to a small number of discrete points. This means that, when selecting for a good allele at a low frequency, too high a proportion becomes fixed at the lower end of the gene frequency scale giving rise to low values of the genetic mean and high values of the genetic variance. By contrast, with a high initial gene frequency too high a proportion becomes fixed at the upper end of the gene frequency scale giving rise to high values of the genetic mean and low values of genetic variance. This problem does not arise with initial gene frequencies midway between the two fixation points.

The results for the selection of eight individuals show considerable reduction in discrepancies between the two methods. The fixation problem is still apparent for low initial gene frequencies but is much less marked. The results for initial gene frequencies of 0·50 and 0·75 are similar to within about 1 %. Thus it would appear that, in the selection of eight individuals, the fitting of a beta distribution to fifteen discrete points provides a very accurate approximation to the results obtained by using the transition matrix method. The approximation will improve still further with selection of more than eight individuals.

The other calculations to be discussed concern the skewness and kurtosis calculated, assuming a beta distribution, from the first two moments of the gene frequency distribution derived by the transition matrix method. Comparison of these with the skewness and kurtosis actually calculated by the transition matrix method reinforced the conclusions already drawn. There was quite good agreement between the values of the kurtosis, but this tells us little more than that the distributions flatten out in a similar way with continued selection. The values of the skewness showed the tendency of the beta distribution to overestimate the positive skewness for a low initial gene frequency and to overestimate the negative skewness for a high

https://doi.org/10.1017/S0016672300002834 Published online by Cambridge University Press

initial gene frequency. It is this overestimation which in each case leads to fixation of too high a proportion of the genes in the Curnow–Baker method.

In conclusion, it would appear that the one real fault in the Curnow–Baker method is its tendency to fix too high a proportion of the genes, particularly when the initial gene frequency is near to a fixation point. (It should be noted here that this problem does not invalidate Baker & Curnow's (1968) results for gene frequencies of 0·1, 0·2, and 0·3. This is because they had an infinite population of males and selection was only practised on females. Therefore there was no fixation. The beta distribution was used without any additional points corresponding to fixation.) The problem of overestimating fixation is overcome to a large extent when more individuals are selected. The results here would lead to a rejection of the Curnow–Baker method for selection of as few as four individuals or when very small initial gene frequencies are considered. However, the method is much more accurate for selection of eight or more individuals. Furthermore, selection of eight individuals using the Curnow–Baker method requires only a third as much time as the transition matrix method. Thus with the understanding of the restrictions in its application we are led to conclude that the Curnow–Baker method provides a very useful tool for estimating changes in the distribution of gene frequency in selection from finite populations.

SUMMARY

Robertson (1960) used probability transition matrices to estimate changes in gene frequency when sampling and selection are applied to a finite population. Curnow & Baker (1968) used 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 to develop an iterative procedure for studying the effects of repeated cycles of selection and regeneration. To do this they assumed a beta distribution for the unfixed gene frequencies at each generation.

These two methods are discussed and a result used in Kojima's paper is proved. A number of sets of calculations are carried out using both methods and the results are compared to assess the accuracy of Curnow & Baker's method in relation to Robertson's approach.

It is found that the one real fault in the Curnow–Baker method is its tendency to fix too high a proportion of the genes, particularly when the initial gene frequency is near to a fixation point. This fault is largely overcome when more individuals are selected. For selection of eight or more individuals the Curnow–Baker method is very accurate and appreciably faster than the transition matrix method.

I am grateful to Professor R. N. Curnow and Dr W. G. Hill for much helpful advice and to the Science Research Council under whose financial assistance this work was carried out.

REFERENCES

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

Baker, L. H. & Curnow, R. N. (1968). Choice of population size and use of variation between replicate populations in plant breeding selection programmes. (In preparation.)

https://doi.org/10.1017/S0016672300002834 Published online by Cambridge University Press

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.

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

Hill, W. G. (1968). On the theory of artificial selection in finite populations. *Genet. Res.* **13**, 143–163.

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

Kendall, M. G. & Stuart, A. (1963). *The Advanced Theory of Statistics*, Volume 1. London: Griffin.

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

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. Roy. Soc.* B **153**, 234–249.

Table 1. *Selection of four individuals—additive case. Genetic mean* × 10^3

		Generation number					
	Initial	1		5		15	
Sample size	gene frequency	CB	TM	CB	TM	CB	TM
8	0·25	−60	−60	−47	−46	−35	−30
	0·50	6·1	6·1	23	23	41	41
	0·75	70	70	82	82	95	93
16	0·25	−57	−57	−35	−33	−15	−6·2
	0·50	10	10	38	39	65	65
	0·75	73	73	92	91	109	106
48	0·25	−54	−54	−20	−15	10	22
	0·50	15	15	55	56	90	89
	0·75	76	76	102	101	121	118

CB = Curnow–Baker Method. TM = Transition Matrix Method.

Table 2. *Selection of four individuals—additive case. Genetic variance* × 10^4

		Generation number					
	Initial	1		5		15	
Sample size	gene frequency	CB	TM	CB	TM	CB	TM
8	0·25	20	17	87	75	142	141
	0·50	23	21	80	79	131	131
	0·75	14	15	40	47	62	70
16	0·25	21	17	95	82	152	151
	0·50	24	21	75	74	108	109
	0·75	14	14	31	38	36	46
48	0·25	23	17	103	88	156	150
	0·50	24	21	65	65	74	78
	0·75	13	14	22	28	14	25

Table 3. *Selection of eight individuals—additive case. Genetic mean* × 10^3

		Generation number					
Sample size	Initial gene frequency	1		5		15	
		CB	TM	CB	TM	CB	TM
16	0·25	−60	−60	−42	−42	−14	−11
	0·50	6·4	6·4	28	28	60	60
	0·75	70	70	84	84	104	103
48	0·25	−56	−56	−20	−19	37	41
	0·50	12	12	52	52	101	99
	0·75	74	74	99	99	122	121

Table 4. *Selection of eight individuals—additive case. Genetic variance* × 10^4

		Generation number					
Sample size	Initial gene frequency	1		5		15	
		CB	TM	CB	TM	CB	TM
16	0·25	9·0	8·3	49	44	117	112
	0·50	11	11	45	44	78	80
	0·75	8·0	7·5	24	24	30	34
48	0·25	9·7	8·6	56	50	109	105
	0·50	11	10	38	37	32	38
	0·75	7·9	7·1	15	16	6·9	10

Table 5. *Selection of four individuals—complete dominance case. Genetic mean* × 10^3

		Generation number					
Sample size	Initial gene frequency	1		5		15	
		CB	TM	CB	TM	CB	TM
8	0·25	−14	−12	−17	−8.7	−20	−12
	0·50	63	63	58	58	53	53
	0·75	110	109	104	100	99	96
16	0·25	−8·2	−6·0	1·0	12	7·3	20
	0·50	66	67	72	73	79	78
	0·75	111	110	109	106	112	108
48	0·25	−1·2	1·7	22	36	41	56
	0·50	70	72	87	87	105	101
	0·75	112	111	114	112	122	118

Table 6. *Selection of four individuals—complete dominance case. Genetic variance* × 10^4

		Generation number					
	Initial	1		5		15	
Sample size	gene frequency	CB	TM	CB	TM	CB	TM
8	0·25	40	32	114	100	150	149
	0·50	21	19	69	69	121	121
	0·75	3·7	5·0	19	29	54	65
16	0·25	41	32	114	97	153	146
	0·50	21	18	54	54	86	90
	0·75	3·4	4·5	13	21	28	42
48	0·25	41	30	107	87	136	121
	0·50	20	15	36	37	40	53
	0·75	3·2	4·0	7·5	14	7·9	21

Table 7. *Selection of eight individuals—complete dominance case. Genetic mean* × 10^3

		Generation number					
	Initial	1		5		15	
Sample size	gene frequency	CB	TM	CB	TM	CB	TM
16	0·25	−9·1	−8·8	8·2	11	25	30
	0·50	67	67	74	74	85	84
	0·75	112	112	110	109	112	110
48	0·25	1·4	1·0	43	46	86	86
	0·50	73	73	92	93	115	113
	0·75	113	113	115	115	123	122

Table 8. *Selection of eight individuals—complete dominance case. Genetic variance* × 10^4

		Generation number					
	Initial	1		5		15	
Sample size	gene frequency	CB	TM	CB	TM	CB	TM
16	0·25	18	16	69	62	116	110
	0·50	10	9·4	32	31	50	55
	0·75	2·3	2·1	8·3	10	16	22
48	0·25	18	16	52	48	55	58
	0·50	9·1	8·1	18	18	9·1	15
	0·75	2·2	1·8	4·7	5·8	4·0	6·5

https://doi.org/10.1017/S0016672300002834 Published online by Cambridge University Press