INBREEDING AND VARIANCE EFFECTIVE NUMBERS IN POPULATIONS WITH OVERLAPPING GENERATIONS

JOSEPH FELSENSTEIN

Department of Genetics, University of Washington, Seattle, Washington 98105

Received January 18, 1971

THE concept of effective population number originated with SEWALL WRIGHT (1931, 1938). He and others have calculated effective population numbers for a variety of models of population reproduction. In particular, KIMURA and CROW (1963) have calculated the variance effective number for a population of constant size in which there are overlapping generations and age-dependent birth and death rates. NEI has presented (NEI and IMAIZUMI 1966) a different formula as a correction to KIMURA and CROW. In this paper, I will argue that the KIMURA & CROW formula is incorrect and the NEI formula is not precisely defined. I will derive equations for both inbreeding and variance effective numbers in models of population reproduction in which birth and death rates are age specific.

THE MORAN MODEL

One of the ways in which we can check the KIMURA and CROW and NEI formulas is to compare them with the effective number in a case in which the variance effective number can be calculated exactly. MORAN (1962) has stated stochastic models of genetic drift in which generations overlap. In its simplest form, this model is as follows: In each unit of time, one haploid individual, chosen at random, gives birth to a single offspring. Immediately afterwards, an individual chosen at random dies. The newborn individual may not be the one which dies, but all other individuals including the parent of the newborn are at risk. It is known that in this model the probability that two randomly chosen individuals are not identical by descent declines at a rate of $2/N^2$ per unit of time (MORAN 1962). Since a generation is N time units in this model, the probability of non-identity declines at a rate of $2/N$ per generation. Thus the "inbreeding" effective number in the MORAN model should be $N/2$, where N is the number of haploid individuals.

However, we cannot compare this directly with the KIMURA & CROW and NEI results, since both of those give variance effective numbers. For this comparison, we must calculate the variance of gene frequency per unit time in a MORAN model. If p and $1-p$ are the frequencies of the two alleles, the probability that p will increase by $1/N$ in one unit of time is $p(1-p)$, and there is also the same probability that p will decrease by $1/N$. Thus the variance of gene frequency is

$$\mathrm{Var}(\delta p) = \frac{2}{N^2}\, p(1-p).$$

Since this is the variance in one unit of time, which is $1/N$ of a generation, to calculate the variance effective number we must equate the variance to

$$\frac{1}{NN_e}p(1-p),$$

where N_e is the variance effective number. This gives $N_e = N/2$, which is equal to the inbreeding effective number.

KIMURA and CROW give as the variance effective number $N_e = N^2N_1T$, where T is the generation time (in this case $T = N$) and N_1 is the number of newborns per unit time (in this case $N_1 = 1$), so that for the MORAN model

$$N_{e(\mathrm{KC})} = N^2/N = N$$

which is off by a factor of two. More recently, KIMURA and CROW discovered that their formula was incorrect, and have informed me that they had intended to publish a retraction and correction. NEI uses the formula $N_e = N_mT$, where N_m is described as "the number of individuals who are born during time interval dy and able to reach the mean reproductive age or, more accurately, participate in the reproduction." As before, T is the generation time.

In this case it is not obvious what N_m is. The mean reproductive age is N, so that $(1-1/N)^{N-1} \simeq e^{-1} \simeq 0.37$ of the individuals born reach that age. However, all of the individuals born are exposed to a risk of reproduction, but (for large N) only half of them actually reproduce before dying. Thus we may or may not get the correct answer from NEI's formula depending on how we interpret N_m, and the more complicated the situation to which we apply it, the less clear it will be which is the interpretation to use. If we take N_m to be the number of individuals reaching the reproductive ages, N_m will be 1, and we get $N_{e(\mathrm{N})} = N$, which is off by a factor of two.

The KIMURA and CROW and NEI formulas are derived for overlapping generations with age-specific birth and death rates. The MORAN model is a special case of this general situation. It is easily verified that although the KIMURA and CROW and NEI formulas are derived for diploid models, they would be expected to apply equally well to haploid models.

THE MODEL

In order to calcuate an effective population number we must first state the model for which it is to be derived. This model has the advantage of maintaining both population size and age distribution exactly constant, but this property is obtained only at the cost of assuming that the births and deaths of different individuals are not independent of one another.

Suppose that we have a haploid population in which exactly N_1 individuals are born in each unit of time. Exactly N_2 of these survive to age 2, N_3 to age 3, and so on. The probability of survival to age k can be calculated to be $l_k = N_k/N_1$. The deaths cannot be independent events, since we assume that it is always true that *exactly* N_k survive to age k. Assuming equal probabilities of death for all individuals of a given age, the N_k individuals of age k at time t are a random sample without replacement from the N_{k-1} individuals of age $k-1$ at time $t-1$. When we consider cases in which the N's are very large, sampling without replacement is nearly the same as sampling with replacement, so that the lack of realism in assuming nonindependence of deaths will not have serious consequences.

We always obtain *exactly* N_1 newborns, so birth events also cannot be independent. For each newborn, we assume that it has probability p_k of coming from a parent of age k, and its parent is randomly chosen from among those in the particular age group. The different newborns are selected independently, which is not the same as each potential parent deciding independently how many offspring it will have during the next time interval. Again, the lack of independence of offspring numbers will not be serious if the N's are large. The births per individual of age k will be expected to be

$$b_k = N_1 p_k / N_k = p_k / l_k,$$

so that

$$\sum_k l_k b_k = 1.$$

It will be of interest to define reproductive values for individuals of different ages in this model. The reproductive value of an individual of age k is proportional to its expected average long-term contribution to the gene pool through offspring born at or following its present age. These values are standardized by setting the value of a newborn at unity. Since the model population is not growing, our counterpart to the equation of FISHER (1958) is

$$v_k = \frac{1}{l_k} \sum_{j \geq k} l_j b_j .$$

But $l_j b_j = p_j$. If we let $q_j = p_j + p_{j+1} + p_{j+2} + \ldots$ be the fraction of reproduction which takes place on or after age j, then

$$v_k = q_k / l_k .$$

We can also calculate the generation time. The formula used here will be

$$T = \sum_i l_i b_i i = \sum_i i p_i .$$

Note that this is precisely

$$T = \sum_i q_i$$

so that the total reproductive value of a population is

$$V = \sum_i N_i v_i = \sum_i N_1 l_i v_i = N_1 \sum_i l_i v_i ,$$

and

$$V = N_1 \sum_i q_i = N_1 T . \qquad (1)$$

With these preliminary calculations aside, we can turn to deriving effective numbers.

INBREEDING EFFECTIVE NUMBER

Although it does not make sense to talk of an inbred individual in a haploid population, it is possible to calculate an "inbreeding" effective number based on the rates of change of probabilities of identity by descent. In particular, let us follow the probabilities that an individual of age i, selected at random, is *not* identical by descent to an individual of age j selected at random *with replacement*. This means that if $i = j$, it is possible for the two randomly selected individuals to be the same individual. The probabilities will be called the H_{ij}. From the definition it is obvious that $H_{ij} = H_{ji}$.

Suppose that we know the values of the H_{ij} at a given time. We wish to calculate the H_{ij}', the values of the H_{ij} after one unit of time. Consider first the case in which $i = j = 1$, so that we are drawing two newborns. These have a chance $1-1/N_1$ of being different individuals. If they are, there is a probability p_i that the first is the offspring of a parent aged i in the previous time interval, and there is a probability p_j that the second is the offspring of a parent aged j. If the two offspring are not identical by descent, this can only result from their being descendants of parents who were not identical by descent. When we happen to sample the same offspring twice, which occurs with probability $1/N_1$, this obviously makes a zero contribution to the probability that the two individuals are *not* identical by descent. Putting all of this together,

$$H_{11}' = (1 - \frac{1}{N_1}) \sum_{ij} p_i p_j H_{ij} \,. \qquad (2)$$

Next consider the case in which $i > 1$ and $j = 1$. The individual aged i was aged $i - 1$ in the previous time interval, and the newborn individual had probability p_k of being the offspring of a parent of age k. So when $i > 1$,

$$H_{i1}' = \sum_k p_k H_{i-1,k} \,. \qquad (3)$$

There is a similar equation for the case in which $i = 1$. When i and j are both greater than one, and $i \neq j$, we have

$$H_{ij}' = H_{i-1,j-i} \,. \qquad (4)$$

Finally, suppose that $i = j > 1$. If the two individuals sampled happen to be distinct, their probability of nonidentity by descent is

$$H_{ii}'/(1-1/N_i) \,.$$

But this must be the same as the probability that two distinct individuals of age $i-1$ in the previous time interval were not identical by descent, which was

$$H_{i-1,i-1}/(1-1/N_{i-1}) \,.$$

Equating these, we have when $i > 1$

$$H_{ii}' = (\frac{1-1/N_i}{1-1/N_{i-1}}) H_{i-1,i-1} \,. \qquad (5)$$

Equations (2) through (5) give us a set of linear equations calculating the H_{ij}' in terms of the H_{ij}. All of the H_{ij} will decline towards zero. It is not ruled out that some will decline faster than others, but asymptotically we will be in a situation in which those H_{ij} are effectively zero, and the rest of the H_{ij} will all decline at the same rate, which is determined by the largest eigenvalue of the matrix of coefficients of the linear equations. Asymptotically, then, all of the nonzero H_{ij} will decline at the same rate. Noting that a generation is T units of time, we will define the inbreeding effective population number N_{eI} by letting the asymptotic decline be

$$H_{ij}' = (1 - \frac{1}{N_{eI}T}) H_{ij} \qquad (6)$$

for all i and j, so that roughly speaking, the H_{ij} decline by $1/N_{eI}$ per generation.

Any weighted average of the H_{ij} should asymptotically decline at the same rate as the individual H_{ij}'s. We will use this property to calculate N_{eI}. We will follow the weighted average:

$$H = \sum_{ij} \left(\frac{q_i}{T}\right) \left(\frac{q_j}{T}\right) H_{ij} .$$

This weights each age class by the proportion of the total reproductive value which is contained in that age class. The total reproductive value in age i is $N_1 q_i$, and the total reproductive value in the population is $N_1 T$. In the next time interval, H becomes

$$H' = \frac{1}{T^2} \sum_{ij} q_i q_j \; H_{ij}' .$$

We simply substitute the expressions given in (2) through (5) for the H_{ij}'. Then

$$H' = \frac{1}{T^2} \Big[q_1^2 \left(1 - \frac{1}{N_1}\right) \sum_{ij} p_i p_j H_{ij} + \sum_{i>1} q_i \sum_j p_j H_{i-1,j} + \sum_{j>1} q_j \sum_i p_i H_{j-1,i} + \sum_{i \neq j} q_i q_j H_{i-1,j-1} + \sum_{i>1} q_i^2 \left(\frac{1 - 1/N_i}{1 - 1/N_{i-1}}\right) H_{i-1,i-1} \Big] . \tag{7}$$

We note that $q_1 = 1$, and also approximate:

$$\frac{1 - 1/N_i}{1 - 1/N_{i-1}} \cong 1 - \frac{1}{N_i} + \frac{1}{N_{i-1}} \tag{8}$$

ignoring terms of order $1/N^2$. We can use (8) and rearrange (7) to get

$$H' \cong \frac{1}{T^2} \Big[\sum_{ij} (q_{i+1} q_{j+1} + q_{i+1} p_j + q_{j+1} p_i + p_i p_j) H_{ij} - \frac{1}{N_1} \sum_{ij} p_i p_j H_{ij} - \sum_{i>1} q_i^2 H_{i-1,i-1} \left(\frac{1}{N_i} - \frac{1}{N_{i-1}}\right) \Big] .$$

Note that

$$q_{i+1} q_{j+1} + q_{i+1} p_j + q_{j+1} p_i + p_i p_j = (q_{i+1} + p_i)(q_{j+1} + p_j) = q_i q_j .$$

We must also make another approximation by assuming that the H_{ij} are all nearly equal to H, the differences at most involving terms of order $1/N$. This enables us to replace some of the H_{ij} by H and obtain

$$H' = \frac{1}{T^2} \Big[\sum_{ij} q_i q_j H_{ij} - \frac{H}{N_1} \Big(1 + \sum_{i>1} q_i^2 \left(\frac{1}{l_i} - \frac{1}{l_{i-1}}\right)\Big)\Big], \tag{9}$$

or

$$H' = H \Big[1 - \frac{1}{N_1 T^2} \Big(1 + \sum_{i>1} q_i^2 \left(\frac{1}{l_i} - \frac{1}{l_{i-1}}\right)\Big)\Big].$$

The coefficient of H must be equal to $1 - 1/N_{eI}T$, so that, approximately,

$$N_{eI} = \frac{N_1 T}{1 + \sum_{i=1}^{\infty} q_{i+1}^2 \left(\frac{1}{l_{i+1}} - \frac{1}{l_i}\right)} .$$

Note that

$$\frac{1}{l_{i+1}} - \frac{1}{l_i} = \frac{1}{l_{i+1}} \left(1 - \frac{l_{i+1}}{l_i}\right) = \frac{d_i}{l_{i+\mathrm{i}}}$$

where d_i is the probability of death at the end of age i. The reproductive value of an individual of age $i + 1$ is

$$v_{i+1} = \frac{q_{i+1}}{l_{i+1}} ,$$

so that

$$v_{i+1}{}^2 = \frac{q_{i+1}{}^2}{l_{i+1}{}^2} .$$

If $s_i = 1 - d_i = l_{i+1}/l_i$, we end up with

$$N_{eI} = \frac{N_1 T}{1 + \sum\limits_{i=1}^{\infty} l_i s_i d_i v_{i+1}{}^2} . \qquad (10)$$

The numerator of (10) is simply the total reproductive value in the population (see equation (1)). The second term in the denominator is roughly the probability of death of an individual while it still has reproductive value. It will reflect infant mortality before the reproductive ages as well as part of the deaths of adults during the reproductive period. After the end of the reproductive ages, v_i will be zero, so that the deaths of older individuals will make no contribution to this term.

Ironically, it is difficult to check (10) by considering the MORAN model because the present model departs widely from the assumptions of the MORAN model. In the MORAN model, the age distribution of the population will fluctuate wildly, since each age cohort is represented by at most one individual, who has a constant risk of dying. In the present model, the age distribution is held rigidly stable by the lack of independence of deaths. Despite this large difference in the models, formula (10) works surprisingly well when applied to the MORAN model. In that case, all individuals have the same reproductive value, so that $v_i = 1$. We have

$$l_i = (1 - \frac{1}{N})^{i-1} ,$$

$$s_i = 1 - \frac{1}{N} ,$$

$$d_i = \frac{1}{N} ,$$

$$N_1 = 1 ,$$

and

$$T = N .$$

Then

$$N_{eI} = \frac{N}{1 + \sum\limits_{i=1}^{\infty} \left(1 - \frac{1}{N}\right)^i \frac{1}{N}} = \frac{N}{2 - \frac{1}{N}} ,$$

which is very nearly equal to the correct value of $N/2$.

To see whether the approximations affect the validity of (10), I have carried out a series of computer iterations. For a given set of values of the N_i and the p_i, the H_{ij} were initially assumed to all be 1. Iterations were done using (2) through (5) to calculate new values for the H_{ij}. These values were then normalized so that $H_{11} = 1$, a step which was taken to prevent the numbers from becoming too small for the computer to handle. After a number of iterations (in no case more than 500), the ratios of the H_{ij} to each other had stabilized, as had their rate of decline. The "true" effective population number was calculated on the assumption that asymptotically with time,

$$H_{11}' = H_{11}\left(1 - \frac{1}{N_{eI}T}\right) ,$$

so that

$$N_{eI} = \frac{1}{T\left(1 - \frac{H_{11}'}{H_{11}}\right)} . \qquad (11)$$

Table 1 shows comparisons of the true values of N_{eI} and those calculated from (10). The examples given are ones with or without infant mortality and with or without mortality of adults of reproductive age. In general, the two effective numbers are within ± 1 of each other, which is well within the bounds of the statistical errors which would result from the estimation of birth and death rates in any real population. While it might be possible to improve on some of the approximations used in deriving (10), there seems little point in doing so as long as we are applying (10) to populations larger than 100 individuals.

It is of interest to know what the effective number would be in a population whose age structure was that of a real human population. For this purpose we can make use of vital statistics for the U. S. A. white female population for 1967. The ages can be made discrete by pretending that all births take place when the mother reaches ages 5, 10, 15, 20, 25, 30, 35, 40, or 45. In this case the life-table values are those for survival to those ages, and the birth rates are obtained as follows. For age 25, the single-year birth rates for ages 24 and 25 are averaged, and the result is multiplied by five, since the next births will not be until age 30. The resulting number is then divided by 2.056, since only female births are being counted, and there are 1.056 male births for each female birth. The resulting life table and birth rates are presented in Table 2. Note that l_1 is 1 since we consider an individual to have entered the first age-class as soon as it is born.

To use equation (10) we must further modify the birth rates by dividing them by 1.21004 so that the new rates will satisfy

$$\sum_i l_i b_i = 1 .$$

The generation time in the resulting model is 5.26092 time units, which is 26.3 years. Applying equation (10), we obtain

$$N_{eI} = 5.1181\, N_1$$

where N_1 is the number of newborns per time interval. Since a time interval is five years, we can also write this as

$$N_{eI} = 25.59\, B ,$$

where B is the number of newborns per year. If the expected length of life of a

TABLE 1

Results of computer check of equation (10) by iterating equations (2)–(5)

Life table (l_i)																							Effective number	
Prereproductive ages								Reproductive ages								Postreproductive ages				N_1	T	N_1T	Expected	Observed
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	20	12.5	250	250.0	250.237
																				200	12.5	2500	2500.0	2499.981
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.2	0.2	0.2	0.2	20	11.5454	230.909	166.806	167.437
																				200	11.5454	2309.091	1668.060	1668.684
1.0	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	20	12.5	250	150.0	150.239
																				200	12.5	2500	1500.0	1500.234
1.0	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.60	0.55	0.50	0.45	0.40	0.35	0.30	0.25	0.25	0.25	0.25	0.25	20	11.88235	237.647	119.443	119.883
																				200	11.88235	2376.47	1194.435	1194.867
Age-specific birth rates were proportional to:																								
0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0					

TABLE 2

Life table (l_i) *and birth rates* (b_i) *derived from vital statistics for U. S. white females for 1967*
Each time unit is five years

i	l_i	b_i
1	1.00000	0.0000
2	0.97891	0.0000
3	0.97754	0.0208
4	0.97486	0.3515
5	0.97179	0.4157
6	0.96841	0.2602
7	0.96382	0.1377
8	0.95662	0.0542
9	0.94550	0.0069
10	0.92822	0.0000

Data from *Vital Statistics of the U. S., 1967* (Table 5–2 of volume II, Part A, and Table 1–16 of volume I).

newborn female is 75.1 years, then the total census number in a population of constant size will be 75.1 B. If N is the census number, we then have

$$N_{eI} = \left(\frac{25.59}{75.1}\right) N = 0.34\, N\,. \qquad (12)$$

Most of this reduction of effective number below the census number is due to the inclusion in N of a large number of individuals who are past the reproductive ages and thus are reproductively dead. Another factor reducing effective size is the death of individuals who have not yet passed out of the reproductive ages. However, the death rates in those ages are so low in industrialized countries that this turns out to have very little effect on the effective number.

If we use KIMURA and CROW's formula, we would have

$$N_{e(\mathrm{KC})} = N^2/(N_1 T) = \left(\frac{N}{N_1 T}\right) N$$
$$= \left(\frac{75.1}{26.3}\right) N = 2.86\, N$$

which is much larger than (12), being over nine times too high.

NEI's formula is somewhat difficult to apply. About 0.97 of the newborns survive to the mean reproductive age, so that

$$N_{e(\mathrm{N})} = (26.3)(0.97)\, B$$
$$= 25.51\, B \cong 0.34\, N$$

which is very close to (12), differing only in the third or fourth significant figure. It can in fact be shown that NEI's formula will give the correct result if all mortality is either prereproductive or postreproductive, in which case the NEI formula is not ambiguous. With mortality during reproductive ages, the difficulties of NEI's formula are more the result of vagueness of definition than inaccuracy of the results.

These effective numbers must be taken with a grain of salt. Humans are diploid and have two sexes, and the generation time and mean length of life are different for the two sexes. Furthermore, the variance of family sizes cannot be adequately described by a model in which the probability of having an offspring is completely independent of marital status or the number of previous children.

VARIANCE EFFECTIVE NUMBER

Before trying to relax some of the restrictive assumptions of our model, we should check the variance effective number, N_{eV}, to see whether it differs greatly from the inbreeding effective number N_{eI}. The object of this section will be to show that the variance effective number is the same as the inbreeding effective number. The model is the same as before. Consider an allele A. Let $x_i^{(t)}$ be the gene frequency of A among individuals of age i at time t. Consider the covariance between $x_i^{(t)}$ and $x_j^{(t)}$. It is readily shown that this covariance is the same as the covariance of the gene frequency in samples of one gene each drawn from ages i and j. The observed gene frequencies in those samples are 1 or 0 according to whether or not the gene sampled turns out to be A.

There is probability $1-H_{ij}^{(t)}$ that the two genes chosen are identical by descent, that is, that they are descendants of the same gene in the initial generation. If they are copies of the same gene, their covariance is simply their variance, which is $x(1-x)$. If t is sufficiently large, we do not need to put a subscript on this x, since the age group from which a gene is chosen will have no effect on its probability of being A. This is not to imply that there will be no differences in gene frequencies between age groups, but merely that after a large amount of time, knowledge of the gene frequencies in the initial generation does not permit us to predict which age groups at time t will have higher or lower gene frequencies.

With probability $H_{ij}^{(t)}$, the two genes sampled are not identical by descent, in which case their covariance is zero. Then

$$\mathrm{Cov}\,(x_i^{(t)},\ x_j^{(t)}) = (1-H_{ij}^{(t)})\ x(1-x)\ . \qquad (13)$$

This equation also holds in the case where $i=j$. In that case, the possibility that the two samples turn out to be exactly the same gene is included in the definition of $H_{ii}^{(t)}$, which is the probability that two genes sampled *with replacement* from age i are not identical by descent.

In a haploid population with discrete generations and a variance effective number of N_{eV}, the variance of gene frequency follows

$$\mathrm{Var}\,(x^{(t)}) = \left(1-\frac{1}{N_{eV}}\right)\mathrm{Var}\ (x^{(t-1)}) + \left(\frac{1}{N_{eV}}\right)x(1-x)\ ,$$

where the variances are, as in (13), the variances of the actual gene frequencies around their predicted value x. We can solve for N_{eV}:

$$N_{eV} = \frac{x(1-x)-\mathrm{Var}(x^{(t-1)})}{\mathrm{Var}(x^{(t)})-\mathrm{Var}(x^{(t-1)})}\ . \qquad (14)$$

If we could decide on some average gene frequency which could stand in place of $x^{(t)}$, we could also use (14) to calculate a variance effective number in the case of overlapping generations.

Let $x^{(t)}$ be some weighted average of the $x_i^{(t)}$, so that

$$x^{(t)} = \sum_i a_i x_i^{(t)}, \text{ where } \sum_i a_i = 1 \ .$$

Then

$$\mathrm{Var}\ (x^{(t)}) = \sum_{ij} a_i a_j \ \mathrm{Cov}\ (x_i^{(t)},\ x_j^{(t)}) \tag{15}$$

Equation (14) was based on the change in variance in one generation. In the case of overlapping generations, a time unit is $1/T$ of a generation. We can substitute (15) into (14) if we also replace N_{eV} by $N_{eV}T$:

$$N_{eV}T = \frac{\sum\limits_{ij} a_i a_j (x(1-x) - \mathrm{Cov}(x_i^{(t-1)}, x_j^{(t-1)}))}{\sum\limits_{ij} a_i a_j (\mathrm{Cov}(x_i^{(t)}, x_j^{(t)}) - \mathrm{Cov}(x_i^{(t-1)}, x_j^{(t-1)}))}$$

Assuming that t is large, we can use (13) and factor out an $x(1-x)$ to get

$$N_{eV}T = \frac{\sum\limits_{ij} a_i a_j (H_{ij}^{(t-1)})}{\sum\limits_{ij} a_i a_j (H_{ij}^{(t-1)} - H_{ij}^{(t)})} \ .$$

Since t is large $H_{ij}^{(t)} = \lambda H_{ij}^{(t-1)}$ for all i and j, so that

$$N_{eV}T = \frac{\sum\limits_{ij} a_i a_j H_{ij}^{(t-1)}}{\sum\limits_{ij} a_i a_j H_{ij}^{(t-1)} (1-\lambda)} = \frac{1}{1-\lambda} \ .$$

But $\lambda = 1 - 1/(N_{eI}T)$, where N_{eI} is the inbreeding effective number, so that

$$N_{eV} = \mathrm{N}_{eI} \ .$$

In their discussion of the inbreeding and variance effective numbers in discrete-generation models, KIMURA and CROW (1963) found that they were the same when population sizes were constant and when the parental generation was created by random mating. The first condition holds here. Although this model is haploid, the random mating may in some sense be equivalent to the independent sampling of individuals to be parents of different offspring.

GROWING POPULATIONS

The model used here has assumed that the population size remains constant. Human populations are manifestly not constant in size, which makes the applicability of the above effective number formulas questionable. When we alter the model to allow for growth (or decline) in population numbers, we must decide whether we will have the growth be deterministic or random. If it is to be deterministic, so that the exact number of offspring to be born in the next time interval is known in advance, we cannot allow the numbers of offspring of different parents to be independent. If we wish to have independence of offspring numbers, we must allow the numbers of births, and hence the classes of the age distribution, to be random variables.

Suppose that there are two types of haploid individuals, A_1 and A_2. Let the probability that an individual of age i gives birth to a single offspring be b_i. Let the probability of dying at the end of age interval i be d_i. Let l_i be the probability that the individual survives to the beginning of age i, so that

$$l_1 = 1$$

and

$$l_i = d_{i-1} l_{i-1} .$$

When the age distribution stabilizes, the population size will grow geometrically, being multiplied by a quantity λ once every time interval. We can calculate λ as the solution to

$$\sum_{i=1}^{\infty} l_i b_i \lambda^{-i} = 1 .$$

The generation time (the mean age of mothers of newborns in a population with a stable age distribution) is given by

$$T = \sum_{i=1}^{\infty} i l_i b_i \lambda^{-i} .$$

For each age we can calculate a reproductive value

$$v_i = \frac{\lambda^{i-1}}{l_i} \sum_{j \geq i} l_j b_j \lambda^{-j}. \tag{16}$$

These reproductive values always have $v_1 = 1$. They also have the property that no matter what the age distribution, the total reproductive value of the population after a unit of time has passed has an expectation of λ times the present total reproductive value. If $V = \Sigma n_i v_i$

$$\mathrm{E}(V') = \lambda V ,$$

irrespective of the n_i.

In this model we have an analogue to equation (1). The total reproductive value of the population is $V = \Sigma n_i v_i$. If the age distribution is stable and B is the number of newborns in the population,

$$N_i = B\, l_i \lambda^{-(i-1)}.$$

Using (16),

$$V = \sum_i B\, l_i \lambda^{-(i-1)} \frac{\lambda^{i-1}}{l_i} \sum_{j \geq i} l_j b_j \lambda^{-j}$$

$$= B \sum_i \sum_{j \geq i} l_j b_j \lambda^{-j} = B \sum_j \sum_{i=j}^{j} l_j b_j \lambda^{-j}$$

$$= B \sum_j j l_j b_j \lambda^{-j} ,$$

so that

$$V = B\, T. \tag{17}$$

We will derive a variance effective population number for this model. To do this we must define gene frequency. We weigh each individual by its reproductive value. If n_{i1} is the number of individuals of allele 1 who are of age i, the total reproductive value of A_1 individuals will be

$$V_1 = \sum_i n_{i1} v_i .$$

The gene frequency of A_1, x, is given by

$$x = \frac{V_1}{V_1+V_2} = \frac{V_1}{V} .$$

We use a large-sample approximation to get the variance of x:

$$\begin{aligned}\mathrm{Var}(x') &= \frac{(V_2')^2}{(V')^4}\,\mathrm{Var}(V_1') + \frac{(V_1')^2}{(V')^4}\,\mathrm{Var}(V_2') \\ &\quad - \frac{2V_1'V_2'}{(V')^4}\,\mathrm{Cov}(V_1', V_2') \;.\end{aligned} \qquad (18)$$

To calculate V_1', we note that the number of A_1 individuals in age class $i+1$ will be binomially distributed with parameters n_{i1} and s_i, where s_i is the probability of survival at the end of age i, that is, $1-d_i$. The number of newborns will be a sum of binomial variates, the ith of which has parameters n_{i1} and b_i. Then

$$\mathrm{Var}(V_1') = \sum_i n_{i1}b_i(1-b_i)v_1^2 + \sum_i n_{i1}s_id_iv_{i+1}^2 \;. \qquad (19)$$

There will be a similar formula for $\mathrm{Var}(V_2')$, with the n_{i1} replaced by n_{i2}. Since births and deaths in the A_1 subpopulation are independent of those in the A_2 subpopulation, $\mathrm{Cov}(V_1', V_2') = 0$.

The value of (19) will depend on the age distribution, n_{i1}. However, if the population is large, the n_{i1} will not be far from a stable age distribution. In that case, if B_1' is the number of A_1 newborns expected in the next time interval.

$$n_{i1} = B_1'l_i\lambda^{-i} \;.$$

Substituting that, and factoring out B_1'

$$\begin{aligned}Var(V_1') &= B_1'(v_1^2 \sum_i l_ib_i\lambda^{-i} - v_1^2 \sum_i l_ib_i^2\lambda^{-i} \\ &\quad + \sum_i l_is_id_i\lambda^{-i}\, v_{i+1}^2).\end{aligned}$$

So that

$$\begin{aligned}\mathrm{Var}(V_1') &= B_1'(1+K) \\ \mathrm{Var}(V_2') &= B_2'(1+K),\end{aligned} \qquad (20)$$

where

$$K = -\sum_i l_ib_i^2\,\lambda^{-i} + \sum_i l_is_id_i\lambda^{-i}v_{i+1}^2. \qquad (21)$$

The first term in (21) comes from the fact that the number of offspring born during one time interval to an individual of age i is a binomial variate with parameters 1 and b_i. If the number of offspring were not binomial, but Poisson with parameter b_i, this term would be zero. The second term depends on the death rates in those age classes which have reproductive value. It will be zero if no individual ever dies until it reaches the end of the reproductive period.

Substituting (20) into (18),

$$\mathrm{Var}(x') = \frac{(V_2')^2\mathrm{B}_1'\,(1+K)}{(V_1'+V_2')^4} + \frac{(V_1')^2B_2'\,(1+K)}{(V_1'+V_2')^4} \;.$$

If B' is the total number of births expected to occur in the population during the next time interval, then if the A_1 and A_2 subpopulations are in a stable age distribution,

$$B_1' = B'x,$$

and

$$B_2' = B'(1-x)$$

so that

$$\mathrm{Var}(x') = \frac{B'(1+K)}{(V')^2}\,(x(1-x)^2 + (1-x)x^2) \ ,$$

or

$$\mathrm{Var}(x') = \frac{B'(1+K)}{(V')^2}\,x(1-x) \ . \tag{22}$$

Before using (22) to obtain a variance effective number, we must show that changes in x in successive time intervals are independent. Otherwise (22) would give a misleading impression of the magnitude of longer term changes in gene frequency. It is a property of the total reproductive value of a population or subpopulation that its expected value after one unit of time is λ times its present value, irrespective of how the population arrived in its present situation. Knowing that x has just increased tells us nothing about whether the ratio of V_1 to V_1+V_2 is expected to increase or decrease during the next interval. Thus, successive changes in x have a zero covariance, so that (22) will adequately reflect long-term genetic drift. Other ways of defining x, weighting the age classes by anything other than their total reproductive values, will not have this property.

There are many ways in which effective population number could be calculated, corresponding to different choices of the idealized population with which the "real" population is to be compared. If we take as the idealized population one which has discrete generations and a constant population size of N_{eV}, it will have a gene-frequency variance per generation given by:

$$\mathrm{Var}(x') = \left(\frac{1}{N_{eV}}\right)x(1-x) \ .$$

Equation (22) gives the present gene-frequency variance per unit of time in the case of overlapping generations. We can prorate it to obtain a variance per generation:

$$\mathrm{Var_T}(x') = \left(\frac{B'T(1+K)}{(V')^2}\right)\,x(1-x) \ .$$

Equating the variances and solving for N_{eV}, keeping in mind that to good approximation $V' = B'T$,

$$N_{eV} = \frac{V'}{1+K} = \frac{B'T}{1+K} \ . \tag{23}$$

It is interesting to note that KIMURA and CROW (1963) found that the variance effective number in discrete-generations models reflected the number of offspring rather than the number of parents. Equation (23) shows the same behavior in

TABLE 3

Reproductive values corresponding to the birth rates and life table given in Table 2

i	v_i
1	1.0000
2	1.0594
3	1.1003
4	1.1233
5	0.8161
6	0.4321
7	0.1889
8	0.0586
9	0.0066
10	0.0000

that the number is related to V' rather than V.

Writing K out in full gives us

$$N_{eV} = \frac{B'T}{1 + \sum_i l_i s_i d_i v_{i+1}^2 \lambda^{-i} - \sum_i l_i b_i^2 \lambda^{-i}} \; . \tag{24}$$

The last term in the denominator will be omitted if the number of offspring per parent during one time interval is Poisson rather than binomial. If it is omitted, (24) is almost exactly the same as (10), except for the factors λ^{-i}.

Table 3 shows the reproductive values, v_i, calculated from the birth rates given in Table 2. The intrinsic rate of increase per five-year period, λ, is 1.037086. The generation time T is 5.2106 time units, or 26.053 years. If we ignore the last term in the denominator in (24), we get $K = .0293$ so that

$$N_{ev} = \left(\frac{5.2106}{1.0293}\right) B' = 5.062\, B'$$

Since B' is the number of births expected during the next five years, and since the population is not growing very rapidly, we have approximately

$$N_{eV} = 25.31\, N_0 \; ,$$

where N_0 is the number of births per year. This is close to the result obtained when the birth rates were reduced so that the population did not grow.

DIPLOIDY

All of the preceding models have been haploid. It is obviously of interest to know whether the effective number formulas will also apply to diploid models. Consider a model with only one sex, which is diploid. Aside from the diploidy, the model will be the same as the first model given above, in which population size is constant. There is a probability p_k that a gene in a newborn came from an individual of age k, in which case it was randomly sampled (with replacement) from among the $2N_k$ genes at that age.

In calculating inbreeding effective population number, the H_{ij} are defined exactly the same as before, except that the entities sampled with replacement are now genes instead of individuals. For the H_{ii} we note that there is a probability $1/2N_i$ that the same gene is sampled twice, $1/2N_i$ that two different genes from the same individual are sampled, and $1-(1/N_i)$ that the genes are sampled from different individuals. Because these individuals were created by sampling genes at random from a population of parents, two genes in the same individual have the same probability of nonidentity by descent as two genes from different individuals. Then

$$\begin{aligned} H_{ii} &= \frac{1}{2N_i} \times 0 + \frac{1}{2N_i} h_{ii} + \left(1 - \frac{1}{N_i}\right) h_{ii} \\ &= \left(1 - \frac{1}{2N_i}\right) h_{ii} \ , \end{aligned}$$

where h_{ii} is the probability that two distinct genes from age i are not identical by descent. The occurrence of random deaths between ages i—1 and i will not affect h_{ii}, so that

$$h_{ii}' = h_{i-1,\, i-1}$$

and

$$H_{ii}' = \frac{\left(1 - \frac{1}{2N_i}\right)}{\left(1 - \frac{1}{2N_{i-1}}\right)} H_{i-1,\, i-1} \tag{25}$$

which is to be compared with (5). For the other H_{ij}, the equations are easily derived:

$$H_{11}' = \left(1 - \frac{1}{2N_1}\right) \sum_{ij} p_i p_j H_{ij} \ , \tag{26}$$

$$H_{i1}' = \sum_k p_k H_{i-1,\, k} \quad (i \neq 1) \ , \tag{27}$$

and

$$H_{ij}' = H_{i-1,\, j-1} \quad (i \neq 1, j \neq 1, i \neq j). \tag{28}$$

These equations are identical with (2) through (5), except that the N_i are everywhere multiplied by two. Asymptotically, the H_{ij} will decline geometrically. The decline must be equated to

$$H_{ij}' = \left(1 - \frac{1}{2N_{eI}T}\right) H_{ij} \ ,$$

which is the same as (6) except for a factor of two.

Since both N_{eI} and the N_i are multiplied by two, this factor will ultimately cancel out. We can make the same approximations as before, ending up with

$$N_{eI} = \frac{N_1 T}{1 + \sum\limits_{i=1}^{\infty} l_i s_i d_i v_{i+1}^2} \ .$$

No attempt will be made here to derive effective population number formulas for more complex diploid models. I suspect that in general, we will not go far wrong using effective number formulas derived from haploid models to calculate effective numbers for diploid populations.

I wish to thank J. F. CROW for helpful criticisms, and M. NEI for catching a serious error in an earlier version of this paper. Research and publication of this work has been supported by U.S. Atomic Energy Commission Contract AT (45-1) 2065, and Task Agreement Number 5 of U.S. Atomic Energy Commission Contract AT (45-1) 2225, both with the University of Washington.

SUMMARY

Existing formulas for the effective number of a population with overlapping generations were tested and found wanting. New formulas were derived. The formulas of KIMURA and CROW and of NEI were tested by calculating them for the overlapping-generation model of MORAN, for which the rates of inbreeding and increase of gene-frequency variance are known. The KIMURA & CROW formula gave too large a number, and application of the NEI formula was difficult because of ambiguities in its statement. A haploid model in which population number remains constant was defined. Both inbreeding and variance effective numbers were calculated for that model. Both of these numbers were equal. Each was equal to the number of births per year times the generation time, divided by $1+K$, where K is approximately the probability that an individual dies while it still has reproductive value. For birth and death rates similar to those of a human population in an advanced industrial society, this means that effective number is about one-third of census number.—The variance effective number was also calculated for a growing haploid population. It was shown that a diploid population of constant number would have the same inbreeding effective number as the corresponding haploid population with the same birth and death structure. In all of these calculations it was necessary to weigh individuals by their reproductive values in calculating gene frequencies. The numerator of the effective number formula, the product of the number of births per year and the generation length, is equal to the total reproductive value of the population if it is in a stable age distribution.

LITERATURE CITED

FISHER, R. A., 1958 *The Genetical Theory of Natural Selection*. Second edition. Dover, New York.

KIMURA, M. and J. F. CROW, 1963 The measurement of effective population number. Evolution **17**: 279–288.

MORAN, P. A. P., 1962 *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

NEI, M. and Y. IMAIZUMI, 1966 Genetic structure of human populations. II: Differentiation of blood group gene frequencies among isolated populations. Heredity **21**: 183–190, 344.

U.S. Department of Health, Education, and Welfare. Public Health Service, Health Services and Mental Health Administration, National Center for Health Statistics, 1969 *Vital Statistics of the United States, 1967*. (3 vols.) U.S. Government Printing Office, Washington, D.C.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16**: 97–159. ——, 1938 Size of population and breeding structure in relation to evolution. Science **87**: 430–431.