Amer J Hum Genet 24:1–10, 1972

The Effective Number of a Population with Overlapping Generations: A Correction and Further Discussion

JAMES F. CROW[1] AND MOTOO KIMURA[2]

The concept of effective population number, introduced by Wright [1], compresses a great deal of the relevant information about the genetic structure of a population into a single number and has been widely used to measure random drift in natural populations. Various ways in which the effective number can be defined for populations with discrete generations have been discussed [1–9]. The problem is much more troublesome when generations overlap, as in the human population, and until recently only a few tentative steps have been taken toward a general solution [5, 10, 11]. In particular, we are now convinced that the formula that we put forth in 1963 [5, p. 286] is wrong or, at best, irrelevant to the more important problems in human evolution. Unfortunately, it has been referred to and used by others, most recently by MacCluer and Schull [12]. We should like to retract it before it does further mischief.

Our formula, for a population with a stable age structure and constant size, gave the effective number as $N_e = NE/\tau$, where N is the census number, E is the mean length of life, and τ is the mean age of reproduction. This is an appropriate number in answering a question such as proposed in the following. Suppose a human population is censused at two periods separated by 30 years. What size population with discrete generations and binomial distribution of progeny would have the same amount of gene frequency drift as the change in gene frequency in the censused population? Many individuals will be included in both censuses, and their gene frequency does not change; so the average change will be less and therefore equivalent to that which would occur in a larger population with complete turnover. However, one is usually more interested in the long-range effects of random drift, and the duration of life in postreproductive ages is irrelevant.

A much more useful formula is that given by Nei and Imaizumi [11]. They define the effective number for a stable population as

$$N_e = N_m \tau, \tag{1}$$

where N_m is the number born per year who are "able to reach the mean reproductive age or, more accurately, participate in the reproduction."

Received July 17, 1970; revised May 3, 1971.

Contribution no. 1403 from the Genetics Laboratory, University of Wisconsin, and no. 780 from the National Institute of Genetics, Mishima.

[1] Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706.

[2] National Institute of Genetics, Mishima, Shizuoka-ken, 411, Japan.

A detailed study has been made by Felsenstein [13, 14]. For a haploid population at equilibrium for age distribution and total size, he gives the effective number as

$$N_e = \frac{B\tau}{1+K}, \tag{2}$$

where B is the number of births per census period (e.g., per year), τ is the mean age of mothers of newborns, and

$$K = \sum_{i=1}^{\infty} l_i s_i d_i v_{i+1}^2. \tag{3}$$

In this formula l_i is the probability of surviving to age i, s_i is the probability of surviving from i to $i+1$, d_i is the probability of dying during this interval, and v_i is Fisher's [15] reproductive value at age i. K is roughly the probability of death of an individual while it still has reproductive value.

If there is no death in the reproductive period, the two formulas are in agreement. Felsenstein [14] has extended his formula to include populations that are growing but have reached a stable age distribution and has also discussed diploid populations.

ANOTHER DEFINITION OF EFFECTIVE POPULATION NUMBER

We should like to suggest a formula which is more precisely defined than that of Nei and Imaizumi (although very similar to it), which is usually somewhat simpler to apply than that of Felsenstein, and which should serve as a useful approximation for populations where there is not a large departure from a stable age distribution and where death rates during the reproductive period are moderate. The definition we suggest for N_e is

$$N_e = \bar{l} \int_0^{\infty} N_y v_y dy = \bar{l} V, \tag{4}$$

where N_y is the number in the population of age y, v_y is Fisher's reproductive value at age y, V is the total reproductive value of the population, and $\bar{l}$ is the probability that a newborn will survive into the reproductive period. A more precise definition of $\bar{l}$ will be discussed later. This may also be written as

$$N_e = N_0 \tau \bar{l}, \tag{5}$$

where N_0 is the number of births in one census unit (say, 1 year), and τ is the average age of reproducing mothers. Note the similarity between formula (5) and Nei's formula (1).

The formulas apply strictly to populations in a stable age distribution, although formula (4) may be useful as an approximation for not-too-wide departures from this. We justify the formulas on somewhat intuitive grounds. A precise formulation and derivation would require such procedures as were used by Felsenstein.

DERIVATION OF THE FORMULAS

For convenience of reference, we give here the definitions of the various symbols that will be used.

$N_y dy$ = the number of individuals in the age interval y to $y + dy$ at time t;

l_y = the probability of surviving from birth to age y;

$b_y dy$ = the expected number of births to an individual in the age interval y to $y + dy$;

v_y $= \dfrac{1}{e^{-my} l_y} \displaystyle\int_y^\infty e^{-mx} l_x b_x dx$, Fisher's measure of reproductive value at age y;

m = the Malthusian parameter of population increase, defined by

$$1 = \int_0^\infty e^{-mx} l_x b_x dx;$$

τ $= \int_0^\infty N_y b_y y dy \Big/ \int_0^\infty N_y b_y dy$, the mean age of reproduction in the population at time t; and

$\bar{l}$ $= \int_0^\infty l_y^2 b_y e^{-my} dy$, the probability of surviving into the reproductive period (this definition will be discussed later).

We assume that the population has attained a stable age distribution and therefore implicitly assume that the age-specific birth and mortality rates remain constant. Later, we shall discuss the situation when these assumptions are removed.

The population is enumerated at time t, at which time there are N_y individuals of age y. The probability of surviving from birth to age x is l_x and from birth to age y is l_y, so the probability of surviving from age y to age x is l_x/l_y. The probability of surviving from age y to age x and reproducing at that time is $l_x b_x / l_y$, of which a fraction $\bar{l}$ of those born will survive to the reproductive period. Therefore, during the remainder of their lifetimes, the N_y individuals who were age y at time t are expected to have

$$\frac{N_y \bar{l}}{l_y} \int_y^\infty l_x b_x \, dx = N_y w_y \bar{l}$$

births who survive into the reproductive period.

The sampling variance of the number of A genes contributed by parents of age y at time t is approximately $2\bar{l} N_y w_y p_y (1 - p_y)$, where p_y is the frequency of allele A in parents of age y. Summed over all ages, the variance of the number of A genes contributed is

$$V_{n_A} = 2\bar{l} \int_0^\infty N_y w_y p_y (1 - p_y) dy. \qquad (6)$$

The value of p_y is not exactly constant because of random differences in gene frequency in the individuals dying between time t and the time of reproduction. More important, the value of p_y is not the same for different y. However, if N is moderately large, these differences are small. Furthermore, the differences in gene frequency in the different age cohorts become randomized in future generations because of differences in the ages of reproduction. So, to a good approximation, we can replace p_y by $\bar{p}$, leading to

$$V_{n_A} = 2\bar{p}(1-\bar{p})\bar{l}\int_0^\infty N_y w_y \, dy. \tag{7}$$

This is the variance associated with sampling genes from the population alive at time t. It is equivalent to the variance of the gene frequency change if the parent genes were completely replaced by a random sample of $2\int N_y w_y \bar{l}\, dy$ progeny genes. The variance-effective number is the size of a progeny population with discrete generations and binomial sampling of gametes that has an equivalent sampling variance. So we equate (7) with $2\bar{p}(1-\bar{p})N_{eV}$, where N_{eV} is the variance-effective number, leading to

$$N_{eV} = \bar{l}\int_0^\infty N_y w_y dy. \tag{8}$$

This, however, is the variance-effective number appropriate not to the time t, but to the times of the various future births to individuals alive at time t. If the population is growing exponentially at rate m (or decreasing at a rate $-m$), it will have changed by a factor $e^{m(x-y)}$ during the time that a cohort ages from y to x. We can adjust for this by computing for each future birth its "present value," that is, the number of births occurring now that would make the same average contribution to later generations as a birth x units later. In the changing population, the appropriate weight is $e^{-m(x-y)}$.

Using this, equation (8) is modified to become

$$\begin{aligned} N_{eV} &= \bar{l}\int_0^\infty \frac{N_y}{l_y}\int_y^\infty e^{-m(x-y)} l_x b_x dx \, dy \\ &= \bar{l}\int_0^\infty \frac{N_y}{l_y e^{-my}}\int_y^\infty e^{-mx} l_x b_x dx \, dy \\ &= \bar{l}\int_0^\infty N_y v_y dy, \end{aligned} \tag{9}$$

where v_y is the reproductive value at age y. Alternatively,

$$N_{eV} = \bar{l}V, \tag{9a}$$

where V is the total reproductive value of the population.

Since the population is assumed to have a stable age distribution, $N_y = N_0 l_y e^{-my}$, and

$$
\begin{aligned}
N_{eV} &= N_0\bar{l}\int_0^\infty\int_y^\infty e^{-mx}l_xb_xdx\,dy \\
&= N_0\bar{l}\int_0^\infty e^{-my}l_yb_yy\,dy \qquad (10) \\
&= N_0\bar{l}\,\tau,
\end{aligned}
$$

where τ is the mean age of reproduction in the population at time t, providing another way of expressing the formula. This is very similar to the formula of Nei and Imaizumi.

It remains to define $\bar{l}$. We suggest that this be defined as the average probability in a cohort of surviving to age x, weighted by the proportion of total reproduction that occurs at that age and with each birth expressed as its present value. Thus,

$$
\bar{l} = \frac{\int_0^\infty l_x(l_xb_xe^{-mx})\,dx}{\int_0^\infty l_xb_xe^{-mx}\,dx} = \int_0^\infty l_x^2b_xe^{-mx}dx. \qquad (11)
$$

This definition is rather arbitrarily chosen, and we make no claim for its exactness. A more rigorous and exact definition would require a detailed treatment such as that of Felsenstein [14]. Later, we shall show that in numerical examples the formulas are in approximate agreement.

Equations (9), (9a), and (10) give the variance-effective number at time t. The variance-effective number applies to the progeny generation rather than the parent; in other words, it measures the gene-frequency drift in the period immediately prior to time t. This is apparent in discrete-generation models where, under idealized assumptions, the effective number is the same as the population number in the progeny generation (cf., [5]).

Therefore, the sampling variance given by $\bar{p}(1-\bar{p})/2N_{eV}$ is the variance of gene-frequency drift during the generation just before time t, that is, from $t-\tau$ to t, since τ is approximately the average number of years since the birth of individuals who reproduce at time t. Usually (as in the human population), time is measured in units that are shorter than a reproductive lifetime. The variance of gene-frequency drift from time $t-\delta t$ to t is given by

$$
V_{\delta\bar{p}} = \frac{\bar{p}(1-\bar{p})\,\delta t}{2\,N_{eV}\,\tau}. \qquad (12)
$$

In discrete-generation models, the inbreeding-effective number is the same as the variance-effective number if the expectation of progeny is the same for each individual at birth [4, 5]. We might therefore expect that the formulas developed here would also give the inbreeding-effective number. There are complications, however.

Mating is not at random among the age groups because of a tendency for mates to be near the same age; but, because of differences in the age of reproduction, the age stratification of the descendants of a particular individual disappears in a few generations. If we take a fairly long view over several generations, the correlation in age between descendants of a single ancestor is not great enough to affect significantly the probability that a mating pair come from the same remote ancestor. The inbreeding-effective number is then the number of parents weighted by their expectation of producing progeny who survive to the reproducing ages. So equations (9), (9a), and (10) are also appropriate as approximations to the inbreeding-effective number.

On the other hand, the inbreeding- and variance-effective numbers apply to different generations. The variance-effective number, as we have said, is appropriate to random gene-frequency drift in the past, the present generation being regarded as offspring. The increase in homozygosity measured by the inbreeding-effective number occurs in future generations. The present generation is regarded as the parent generation (or the grandparent if self-fertilization is prohibited), and the actual increase in homozygosity begins one (or two) generation later.

If the census is taken periodically (yearly, say) the discrete-time analogues of equations (9) and (11) are

$$N_{eV} = \bar{l} \sum_{y=0}^{\infty} N_y v_y, \tag{13}$$

$$\bar{l} = \sum_{y=0}^{\infty} l_y^2 b_y \lambda^{-(y+1)}, \tag{14}$$

$$1 = \sum_{y=0}^{\infty} l_y b_y \lambda^{-(y+1)}, \tag{15}$$

$$v_y = \frac{\lambda^y}{l_y} \sum_{x=y}^{\infty} l_x b_x \lambda^{-(x+1)}, \tag{16}$$

where N_y is the number of individuals of age y, l_x is the probability of surviving to age x, b_x is the probability of giving birth in the interval from age x to age $x + 1$, and λ is defined by equation (15). The value of l_0 is taken as one and b_0 is the probability of giving birth between the time of birth and the first birthday.

When the population is of constant size, $m = 0$ or $\lambda = 1$, and the formulas are simplified accordingly. In most human populations, the death rate during the reproductive period is low, the age-specific birth rate is not greatly skewed, and the growth rate is a few percentage points per year. In these circumstances, $\bar{l}$ is well approximated by the probability of surviving to the mean age of reproduction, and we have the same formula as Nei. In some of the numerical examples that were tried, a somewhat better approximation was obtained by using for $\bar{l}$ the probability of surviving to the median age of reproduction, that is, to the age when half the reproduction of a cohort has occurred.

COMPARISON WITH FELSENSTEIN'S FORMULA

As would be expected, there is very little difference among the formulas when the proportion of deaths during the reproductive period is small. In the United States, where current survival during the reproductive period appoaches 100%, our formula, Felsenstein's, and Nei's give results that are practically identical.

The formulas would be expected to be most discrepant when the reproductive age extends over a large period and when most of the deaths occur during this period. We have compared our formula with Felsenstein's in a few models where births and deaths occur at all ages. In the first comparison, survival was assumed to be exponential with death at a rate $1/N$ for each time unit. If the birth rate is constant, $1/N$, this is equivalent to Moran's [10] model. He showed that in this case the effective number is $N/2$; both our formula and Felsenstein's agree. In this case, $l_y = e^{-y/N}$, $b_y = dy/N$, $\tau = 1$, $\bar{l} = 1/2$, and $N_{eV} = N_0\tau/2 = N/2$. With the same death rates but with birth rate $2/N$, the population grows continuously, and both formulas give $N_e = (2/3)N_0\tau$.

When birth is an exponential function of age such that N remains constant, $b_y = (2/N)e^{-y/N}$, we get $(1/3)N_0\tau$ and Felsenstein gets $(3/8)N_0\tau$. If $b_y = (4/N)e^{-y/N}$, so that the population is growing, the Felsenstein formula gives $(5/6)N_0\tau$ and ours gives $(4/5)N_0\tau$. Finally, if $l_y = e^{-y^2/N}$ and $b_y = (2y/N)dy$, both formulas give $N_0\tau/2$.

NONEQUILIBRIUM AGE DISTRIBUTION

In human populations, the conditions determining the equilibrium are usually changing faster than the rate of approach to equilibrium, so that most populations are not in the age ratios that have been assumed in the foregoing discussion. Is there anything that we can do to get a reasonable approximation under these circumstances?

One possibility is to replace $e^{-m(x-y)}$ in equation (9) by B/B_{x-y}, where B is the number of births per time unit at time t, and B_{x-y} is the number $x - y$ units later. Equation (9) then becomes

$$N_{eV} \approx B\bar{l}\int_0^\infty \frac{N_y}{l_y}\int_y^\infty \frac{l_x b_x\,dx}{B_{x-y}}\,dy. \tag{17}$$

The equation for $\bar{l}$ would require a corresponding modification, but, unless there are a number of deaths during the reproductive period, this is not likely to make much difference.

If the records are available, then l_x and b_x can be taken directly from the vital statistics for the appropriate years. In most experimental or observational studies, the uncertainties of the data make it unnecessary to have further refinements in the formulas.

OTHER COMPLICATIONS

One assumption which we have made and which is very likely to be incorrect is that the expectation of progeny is the same for each individual in the population of a given age. Departure from this assumption appears to be the major reason for the discrepancies pointed out by MacCluer and Schull [12], as noted by Nei [16]. The proper correction for the continuous case has not been worked out, but approximate answers can be gotten by considering the situation in a population with discrete generations. For this case, the correction for nonequal expectation of progeny has been given for a population of constant size by Wright [27], but can be extended to more general cases. The variance-effective number of generation t is given by

$$N_{eV} = \frac{2N_t - (\bar{k}/2)}{1 + (V_k/\bar{k})}, \tag{18}$$

and the inbreeding-effective number by

$$N_{eI} = \frac{N_{t-1}\bar{k} - 1}{\bar{k} - 1 + (V_k/\bar{k})}, \tag{19}$$

where $\bar{k}$ and V_k are the mean and variance of the number of progeny per parent [5]. Slight modifications are required if there are separate sexes. If the mortality is small during the reproductive period, this correction could be applied to the effective numbers obtained by equations (9), (10), and (17) to give an approximation of the effective number.

If the parameters $\bar{k}$ and V_k are different in the two sexes, as might be true in a polygamous or promiscuous population, then a separation of equations (18) and (19) can be made for each sex and the two combined by Wright's well-known formula

$$N_e = \frac{4N_{ef}N_{em}}{N_{ef} + N_{em}}, \tag{20}$$

where N_{ef} and N_{em} refer to the numbers computed for the male and female sexes separately.

Finally, if the mean and variance of family size are known only for children counted at the time of birth (or some other early age) and with their subsequent survival not recorded, as is frequently the case with census data, then the key ratio $V_k/\bar{k}$ may be corrected to its expected adult value, provided we assume equal expectation of survival for each individual. Let the subscript b refer to the time of birth (or age of enumeration) and a to the adult. Then the adult ratio is obtained by

$$\frac{V_{ka}}{\bar{k}_a} = 1 + \frac{\bar{k}_a}{\bar{k}_b}\left[\frac{V_{kb}}{\bar{k}_b} - 1\right]. \tag{21}$$

This can also be used to correct for emigration from the population [16]. For a derivation and discussion of modifications when the survival within a sibship is correlated, see Crow and Morton [4].

Nei [16] has applied these principles to the data of MacCluer and Schull [12] and has shown that they provide estimates of the effective population number that agree satisfactorily with the computer simulations.

SUMMARY

An approximate formula is proposed for the effective number, N_e, of a population such as the human population where generations overlap and reproduction occurs at various ages. When y is the age (e.g., in years), the variance-effective number is given by

$$N_{eV} = \bar{l}\int_0^\infty N_y v_y dy = \bar{l}V,$$

where N_y is the number of age y and v_y is Fisher's [15] reproductive value at age y; V is the total reproductive value of the population, and $\bar{l}$ is the weighted mean probability of surviving into the reproductive period,

$$l = \int_0^\infty l_y^2 b_y e^{-my} dy.$$

The formula can also be written

$$N_{eV} = N_0\bar{l}\tau,$$

where τ is the mean age of reproduction. These are also a good approximation to the inbreeding-effective number.

These formulas all assume that the population has attained a stable age ratio and that the age-specific birth and mortality rates are constant. Modifications when these conditions are not met are discussed. We also retract an earlier, incorrect formula.

ACKNOWLEDGMENT

We should like to thank Joseph Felsenstein for a number of useful comments and for the opportunity to see his manuscript. However, he should be absolved of any responsibility for our results or for the intuitive manner of derivation; he would have preferred a more rigorous approach.

REFERENCES

1. Wright S: Evolution in Mendelian populations. *Genetics* 16:97–159, 1931
2. Wright S: Statistical genetics in relation to evolution, in *Actualités scientifiques et industrielles*, no. 802, Exposés de biométrie et de la statistique biologique XIII, Paris, Hermann, 1939, pp 5–64

3. Crow JF: Breeding structure of populations. II. Effective population number, in *Statistics and Mathematics in Biology,* edited by Kempthorne O, Ames, Iowa State College Press, 1954, pp 543–556
4. Crow JF, Morton NE: Measurement of gene frequency drift in small populations. *Evolution* 9:202–214, 1955
5. Kimura M, Crow JF: The measurement of effective population number. *Evolution* 17:279–288, 1963
6. Watterson GA: The application of diffusion theory to two population genetic models of Moran. *J Appl Probability* 1:233–246, 1964
7. Nei M, Murata M: Effective population size when fertility is inherited. *Genet Res* 8:257–260, 1966
8. Cockerham CC: Variance of gene frequencies. *Evolution* 23:72–84, 1969
9. Ewens WJ: *Population Genetics.* London, Methuen, 1969
10. Moran PAP: *The Statistical Processes of Evolutionary Theory.* Oxford, Clarendon, 1962
11. Nei M, Imaizumi Y: Genetic structure of human populations. II. Differentiation of blood group gene frequencies among isolated populations. *Heredity* 21:183–190, 344, 1966
12. MacCluer JW, Schull WJ: Estimating the effective size of human populations. *Amer J Hum Genet* 22:176–183, 1970
13. Felsenstein J: The effective size of a population with overlapping generations. *Genetics* 61, suppl.:18, 1969
14. Felsenstein J: Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* 68:581–597, 1971
15. Fisher RA: *The Genetical Theory of Natural Selection.* Oxford, Clarendon, 1930
16. Nei M: Effective size of human populations. *Amer J Hum Genet* 22:694–695, 1970