GENETIC DRIFT AND ESTIMATION OF EFFECTIVE POPULATION SIZE

MASATOSHI NEI AND FUMIO TAJIMA

Center for Demographic and Population Genetics, University of Texas at Houston, Houston, Texas 77025

Manuscript received December 15, 1980
Revised copy received April 10, 1981

ABSTRACT

The statistical properties of the standardized variance of gene frequency changes (a quantity equivalent to WRIGHT's inbreeding coefficient) in a random mating population are studied, and new formulae for estimating the effective population size are developed. The accuracy of the formulae depends on the ratio of sample size to effective size, the number of generations involved (t), and the number of loci or alleles used. It is shown that the standardized variance approximately follows the χ^2 distribution unless t is very large, and the confidence interval of the estimate of effective size can be obtained by using this property. Application of the formulae to data from an isolated population of *Dacus oleae* has shown that the effective size of this population is about one tenth of the minimum census size, though there was a possibility that the procedure of sampling genes was improper.

EFFECTIVE population size is one of the important parameters that determine the population dynamics of genes. At the present time, however, we know very little about the effective size of natural populations. Therefore, any attempt to estimate this size deserves special attention. In some species of insects, effective population size can be estimated from the rate of allelism of lethal genes WRIGHT, DOBZHANSKY and HOVANITZ 1942; NEI 1968), but this method cannot be used in other organisms since the system of balanced lethal chromosomes required for surveying lethal genes is not available. KRIMBAS and TSAKAS (1971) used the relationship between the amount of gene-frequency change in a population and effective size for estimating the effective size of an olive fly population in Greece. Recently, PAMILO and VARVIO-AHO (1980) examined the validity of their method under a certain scheme of gene sampling; they concluded that the estimate obtained by KRIMBAS and TSAKAS' method is subject to a serious error. However, their conclusion is heavily dependent on the scheme of gene sampling they assumed. If we consider a different scheme of gene sampling, a quite different conclusion is obtained. Furthermore, it is possible to improve KRIMBAS and TSAKAS' formula and estimate the effective population size from data on gene-frequency changes. The main purpose of this paper is to study the sampling property of KRIMBAS and TSAKAS' formula and present improved methods. The statistical errors associated with the estimates obtained by these methods will also be investigated. The formulae obtained will be applied to data from olive flies.

MATHEMATICAL THEORY

Consider a random-mating population of effective size N, and suppose that the allele frequencies in the 0th and tth generations are determined by sampling S_0 and S_t individuals, respectively. Let x_i and y_j be the frequencies of an allele at the ith locus in the samples from the 0th and tth generations, respectively. KRIMBAS and TSAKAS (1971) proposed the following formula for estimating N from this type of allele-frequency data for n loci (one allele from each locus).

$$\hat{N}_{KT} = t/[2\{F_a - (\frac{1}{2S_0} + \frac{1}{2S_t})\}] \quad , \tag{1}$$

where F_a is a measure of standardized variance of gene frequency changes (a quantity equivalent to WRIGHT's inbreeding coefficient) and given by

$$F_a = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i(1 - x_i)} \quad , \tag{2}$$

and the expectation of F_a was *assumed* to be approximately

$$\frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t}{2N} \quad . \tag{3}$$

Conducting an extensive numerical computation, PAMILO and VARVIO-AHO (1980) have shown that, under their scheme of gene sampling, formula (1) gives a totally erroneous estimate when t is small. However, they did not investigate the reason for this. In the following, we shall show that the poor performance of KRIMBAS and TSAKAS' formula is caused by the fact that the expectation of F_a is not given by (3).

In studying the expectation of F_a, it is important to note that the effective population size of a population is often substantially smaller than the actual size, N_A, and that there are two different possible ways of sampling genes from the population. The first scheme is that of PAMILO and VARVIO-AHO (1980), which assumed that N_A is equal to N and that the gene frequency is determined by sampling S individuals out of N. It is also assumed that sampling S individuals in a particular generation does not affect the effective population size. The latter assumption holds if the individuals are sampled after reproduction or if they are returned to the population after examination of genotypes. In human populations, this certainly applies, but in other organisms this may not necessarily be true.

In the second scheme, we assume that N_A is much larger than N, and that individuals for determining gene frequencies and those for the next generation are sampled separately from the population of N_A individuals. This scheme is similar to that considered by SCHAFFER, YARDLEY and ANDERSON (1977) and WILSON (1980) in their studies of the effect of selection on gene-frequency changes. In the following, we consider these two sampling schemes separately.

Sampling scheme I: We now consider the first sampling scheme mentioned above. Let p be the frequency of an allele in generation 0. It is not simple to com-

pute the exact expectation of F_a (SOURDIS and KRIMBAS 1980), but the approximate value can be obtained in the following way.

$$\bar{F}_a \approx \frac{E(x-y)^2}{E[x(1-x)]} = \frac{E(x-p)^2 + E(y-p)^2}{p(1-p) - E(x-p)^2} . \tag{4}$$

We note that $E(x-p)^2$, *i.e.*, the variance of x, is

$$E(x-p)^2 = \frac{p(1-p)}{2S_0}\left(1 - \frac{2S_0 - 1}{2N-1}\right) , \tag{5}$$

since $2S_0$ genes are sampled from a total of $2N$; thus, sampling is hypergeometric (see FELLER 1957). To evaluate the variance of y, *i.e.*, $E(y-p)^2$, we first consider the case of $t = 1$. In this case, the individuals in generation 0 produce a large gamete pool from which $2N$ genes are sampled to produce the individuals in generation 1. The sampling is obviously binomial. The gene frequency in generation 1 is then determined by sampling $2S_1$ genes from a pool of $2N$ genes. This two-step sampling is, however, equivalent to one-step binomial sampling of $2S_1$ genes from the parental gamete pool. Therefore,

$$E(y-p)^2 = p(1-p)/(2S_1) . \tag{6}$$

$E(y-p)^2$ for the case of $t > 1$ can now be easily derived, if we note that $E(y-p)^2$ for the $(t-1)$th generation is $p(1-p)\left[1 - \left(1 - \frac{1}{2N}\right)^{t-1}\right]$ (*e.g.*, CROW and KIMURA 1970). Since the sampling in generation t is binomial with a size of $2S_t$, we have

$$E(y-p)^2 = p(1-p)\left[1 - \left(1 - \frac{1}{2N}\right)^{t-1}\left(1 - \frac{1}{2S_t}\right)\right] . \tag{7}$$

Therefore, $\bar{F}_a$ is approximately given by

$$\bar{F}_a = \frac{1}{1-R}\left[\frac{2(N-S_0)}{2S_0(2N-1)} + 1 - \left(1 - \frac{1}{2N}\right)^{t-1}\left(1 - \frac{1}{2S_t}\right)\right] , \tag{8}$$

where $R = 2(N-S_0)/[2S_0(2N-1)]$. When N is large and $N >> S_0$, S_t, the above formula can be approximately written as

$$\bar{F}_a = \left[\frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t-2}{2N}\right] \Big/ \left[1 - \frac{1}{2S_0} + \frac{1}{2N}\right] . \tag{9}$$

Furthermore, if S_0 is sufficiently large, F_a can further be approximated by

$$\bar{F}_a = \frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t-2}{2N} . \tag{10}$$

The validity of our formulae can be checked by examining the special case of $S_0 = S_t = N$. In this case, our formula (9) or (10) gives $t/(2N)$ as expected; whereas, (3) gives $(t+2)/(2N)$, which is incorrect under the present scheme of gene sampling. It is clear from this comparison that the error involved in (1) is particularly large when t is small. It is also noted that when S_0, $S_t \ll N$ and t is

small, a large part of F_a is due to the sampling error at the gene-frequency survey rather than to random genetic drift.

Theoretically, (8) gives a more accurate value of $\bar{F}_a$ than (9), but for estimating N from the value of F_a, (9) is more convenient. Replacing $\bar{F}_a$ by F_a in this equation, we get the following formula for estimating N.

$$\hat{N}_a = \frac{t - 2 - F_a}{2\left[F_a\left(1 - \frac{1}{2S_0}\right) - \left(\frac{1}{2S_0} + \frac{1}{2S_t}\right)\right]} . \tag{11}$$

When S_0 and t are large, the formula obtainable from (10) can also be used. Namely,

$$\hat{N}_a = \frac{t - 2}{2\left[F_a - \left(\frac{1}{2S_0} + \frac{1}{2S_t}\right)\right]} . \tag{12}$$

We note that (12) cannot be used when $t = 2$. Actually, when t is small, both (11) and (12) do not give a reliable estimate of N, since in this case the contribution of S^0 and S_t to F_a is very large.

PAMILO and VARVIO-AHO computed the expected value of F_a by using the Markov chain method and estimated N from this expected value by using (1) for several values of N and S $(= S_0 = S_t)$. Their results are reproduced in Table 1.

TABLE 1

Estimates of effective population size obtained from PAMILO *and* VARVIO-AHO*'s (1980)* $\bar{F}$ *values*

		$S_0 = S_t = 20$			$S_0 = S_t = 40$		
N	t	$\bar{F}_a$	$\hat{N}_{KT}$	$\hat{N}_a$	$\bar{F}_a$	$\hat{N}_{KT}$	$\hat{N}_a$
100	1	0.0470	−165	125	0.0203	−107	103
	2	0.0520	512	−37	0.0253	3492	778
	4	0.0618	169	95	0.0353	197	101
	8	0.0813	128	101	0.0546	135	103
	16	0.1191	116	105	0.0924	119	105
250	1	0.0504	1117	611	0.0235	−335	285
	2	0.0524	409	−24	0.0255	1992	−70
	4	0.0564	311	195	0.0295	446	238
	8	0.0644	279	232	0.0374	323	250
	16	0.0800	267	249	0.0530	285	255
500	1	0.0516	309	−1696	0.0246	−1190	724
	2	0.0526	382	−20	0.0256	1727	−46
	4	0.0546	433	301	0.0276	777	437
	8	0.0586	465	416	0.0316	610	481
	16	0.0665	484	470	0.0395	553	498
1000	1	0.0522	224	−588	0.0251	4202	2398
	2	0.0527	370	−19	0.0256	1617	−46
	4	0.0537	540	413	0.0266	1236	778
	8	0.0557	702	690	0.0286	1107	921
	16	0.0597	826	849	0.0326	1054	971

The estimate ($\hat{N}_a$) of N by (11) is also presented. It is clear from this table that $\hat{N}_a$ is generally much closer to N than is $\hat{N}_{KT}$ except for $t = 2$. When $t = 1$, $\hat{N}_a$ is not a good estimate of N at all, but it is still closer to N than is $\hat{N}_{KT}$ except for $S_0 = S_t = 20$ with N equal to 500 and 1000. It is noted that $\hat{N}_a$ as an estimate for N is good when S/N and t are large. The poor performance of $\hat{N}_a$ when S/N and t are small is due to the approximation we made in the derivation of (9). Our formula is certainly better than (3), but a small error generated in this approximation process apparently affects the accuracy of $\hat{N}_a$ considerably.

From the theoretical point of view, (11) has one deficiency. Namely, when S_0 is small, the estimate is considerably affected by sampling error. Particularly when x_i is 0, F_a becomes ∞. This happened in KRIMBAS and TSAKAS' (1971) data analysis. To avoid $F_a = \infty$, they simply replaced x_i by y_i whenever $x_i = 0$. $F_a = \infty$ also occurred in PAMILO and VARVIO-AHO's (1980) and SOURDIS and KRIMBAS' (1980) computation of $\overline{F}_a$, but they ignored this event simply because the probability of the event was extremely small. (Actually they defined $\overline{F}_a$ by excluding the case of $F_a = \infty$.) However, a better way to avoid this difficulty is to use the following quantity in place of F_a.

$$F_b = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\frac{x_i + y_i}{2}\left(1 - \frac{x_i + y_i}{2}\right)} . \tag{13}$$

F_b is generally expected to be a better measure of standardized variance of gene-frequency changes than F_a, since the sampling error of $(x_i + y_i)/2$ would be smaller than that of x_i, unless t is large. The expectation of $(x+y)[2-(x+y)]/4$ is $p(1-p) - [E(x-p)^2 + E(y-p)^2]/4$. Therefore, if we make the simplifying assumption of $(1/2S_0 + 1/2S_t) >> (t-2)/2N$, N can be estimated by

$$\hat{N}_b = \frac{(t-2)(1+F_b/4)}{2\left[F_b\left\{1 - \frac{1}{4}\left(\frac{1}{2S_0} + \frac{1}{2S_t}\right)\right\} - \left(\frac{1}{2S_0} + \frac{1}{2S_t}\right)\right]} . \tag{14}$$

Another measure of standardized variance is

$$F_c = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{(x_i + y_i)/2 - x_i y_i} . \tag{15}$$

One nice property of (15) is that the expectation of $(x+y)/2 - xy$ is $p(1-p)$, so that N can be estimated by

$$\hat{N}_c = \frac{t-2}{2\left[F_c - \left(\frac{1}{2S_0} + \frac{1}{2S_t}\right)\right]} . \tag{16}$$

In practice, this gives an estimate close to that from (14). Another advantage of (15) is that it has a smaller variance than F_a and F_b because F_c takes on a narrower range of values than do F_a and F_b. Obviously, the minimum values of

these quantities are all 0. On the other hand, the maximum value of F_a for a single allele (locus) is ∞, as mentioned earlier; whereas, the maximum values of F_b and F_c are 4 and 2, respectively. Indeed, our computer simulation, which will be presented later, has shown that F_c has the smallest variance. Therefore, for the purpose of estimating N, F_c seems to be the best.

In the above formulation, we assumed that the effective population size remains the same for all generations. In many cases, this assumption will not hold, but our formulae can easily be modified to accommodate these cases. For example, when N varies with generation, $\bar{F}_a$ in (8) is given by

$$\bar{F}_a = \frac{1}{1-R}\left[R + 1 - \prod_{j=1}^{t-1}\left(1-\frac{1}{2N_j}\right)\left(1-\frac{1}{2S_t}\right)\right],$$

where N_j is the effective size of the jth generation, and $R = 2(N_0 - S_0)/[2S_0(2N_0 - 1)]$. When $N_j >> S_j$ and S_j is sufficiently large, the above formula is approximated by

$$\bar{F}_a = \left[\frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t-1}{2\tilde{N}} - \frac{1}{2N_0}\right] \Big/ \left[1 - \frac{1}{2S_0} + \frac{1}{2N_0}\right],$$

where $\tilde{N}$ is the harmonic mean of $N_1, N_2, \ldots,$ and N_{t-1}. Therefore, if N_0 is close to $\tilde{N}$, (11) can be used for estimating $\tilde{N}$. Similarly, (14) and (16) will estimate the harmonic mean $\tilde{N}$ when N varies with generation.

Sampling scheme II: In this scheme, it is assumed that the actual population size is larger than the effective size, and S individuals for the gene-frequency survey and N individuals for the next generations are independently sampled from N_A individuals, as mentioned earlier. Therefore,

$$E(x-p)^2 = \frac{p(1-p)}{2S_0}\left(1 - \frac{2S_0 - 1}{2N_A - 1}\right),$$

whereas $E(y-p)^2$ is

$$p(1-p)\left[1 - \left\{1 - \frac{1}{2N}\left(1 - \frac{2N-1}{2N_A-1}\right)\right\}\left(1-\frac{1}{2N}\right)^{t-1}\left(1-\frac{1}{2S_t}\right)\right].$$

Thus, the expectation of F_c is given by

$$\bar{F}_c = \frac{1}{2S_0} + \frac{1}{2S_t} + \frac{1}{2N}\left(t - \frac{2N}{N_A}\right), \tag{17}$$

approximately. It is clear that if $N_A = N$, (17) reduces to (10), but if $N \ll N_A$, it becomes approximately equal to (3). The value of N/N_A would vary considerably with organism. In man it is expected to be close to 1, but in many small organisms, such as Drosophila, N could be considerably smaller than N_A. Recently, MALPICA and BRISCOE (in preparation) estimated the effective sizes of six cage populations in *D. melanogaster* from information on the allelism rate of lethal genes. Each of the cage populations contained about 5000 adult flies, but the estimate of average effective size was about 600. This suggests that the N/N_A ratio in laboratory

populations is about 0.1. A similar N/N_A value was obtained by PROUT (1954) and MURATA (1970) in their experimental populations of Drosophila. Unfortunately, we do not know the N/N_A ratio in natural populations, but in small organisms it is likely to be smaller than 0.1. It is also interesting to note that, if we consider one generation before the 0th generation and define p as the gene frequency in the gamete pool of this generation, N_A is effectively ∞ and we have (3). Therefore, the validity of (3) depends partly on the scheme of gene sampling and partly on the definition of p. Of course, if any information about N/N_A is available, (17) would give a better estimate of N. At any rate, if we assume $N \ll N_A$, N can be estimated by (1) with a proper estimate of F. We suggest that F_c be used in place of F_a in (1), since F_c has a better statistical property. Namely,

$$\hat{N}_{c'} = t/\left[2\left\{F_c - \left(\frac{1}{2S_0} + \frac{1}{2S_t}\right)\right\}\right] . \tag{18}$$

In the above formulation, we assumed the same value of N for all generations. When N varies with generation, (18) again estimates the harmonic mean of N.

COMPUTER SIMULATION

Since our formulae involve some approximations and actual estimates are affected by sampling errors, we have conducted a computer simulation to see the accuracy of the estimates. In this simulation, we used sampling scheme I, since this scheme is more complicated than sampling scheme II and likely to lead to more erroneous results than the latter. We assumed $N = 100$, and the gene frequency change in a population was followed up to the 8th generation, starting from a given gene frequency (p). The gene-frequency change was simulated by the usual Monte Carlo method. At the 0th and tth generations, S_0 and S_t individuals were sampled, and the sample allele frequencies (x and y) were determined. We used three different values of S_0 ($= S_t$), *i.e.*, 20, 40 and 100. From these allele-frequency data, we computed F_a, F_b and F_c by using single-locus data ($n = 1$). This was repeated 5000 times, and using the means ($\bar{F}$) of F_a, F_b and F_c over 5000 replications, the estimates of N were computed. The results obtained for generations 1, 4 and 8 for the case of $p = 0.5$ are presented in Table 2, excluding those for F_b. The results for F_b are not presented, because they are similar to those for F_c, except that F_b has a somewhat larger variance than F_c. In the case of $S = 20$, the $\hat{N}$ value from F_a is an overestimate for $t = 1$, but an underestimate for $t = 4$. Table 2 also gives the standard deviation (s) of F and $s^2/\bar{F}^2$. These values are smaller for F_c than for F_a. The values for F_c were also smaller than those for F_b. This indicates that formula (16), which makes use of F_c, is better than the others.

As mentioned above, the $\hat{N}$ values in Table 2 are based on the means of F over 5000 replications that correspond to 5000 loci or independent alleles. In practice, of course, the number of independent alleles or loci that can be used for estimating N is much smaller, and thus the estimate is expected to be subject to sampling error. To see this point, we examined the distributions of F_c's based on

TABLE 2

Means ($\bar{F}$) *of* F_a *and* F_c *and estimates* ($\hat{N}$) *of effective population size*

	S	t	$\bar{F}$	s	$s^2/\bar{F}^2$	$\hat{N}$
F_a	20	1	0.0476	0.0698	2.15	148
		4	0.0638	0.0898	1.98	80
		8	0.0822	0.1136	1.91	98
	40	1	0.0203	0.0286	1.99	102
		4	0.0361	0.0504	1.95	92
		8	0.0555	0.0752	1.83	100
	100	1	0.0052	0.0075	2.11	104
		4	0.0193	0.0266	1.90	108
		8	0.0387	0.0540	1.95	105
F_c	20	1	0.0447	0.0615	1.89	95
		4	0.0601	0.0797	1.76	99
		8	0.0775	0.1010	1.70	109
	40	1	0.0198	0.0274	1.90	97
		4	0.0353	0.0482	1.86	97
		8	0.0543	0.0719	1.75	102
	100	1	0.0052	0.0075	2.11	104
		4	0.0193	0.0266	1.90	108
		8	0.0387	0.0540	1.95	105

These results were obtained by a Monte Carlo simulation with 5000 replications. The actual effective size was 100, and the initial gene frequency was 0.5. s is the standard deviation of F.

single-locus and 20-loci data. The F_c for 20 loci was computed by sampling 20 loci at random from the 5000 "loci" (replications) with replacement. The results for the case of $S = 40$ and $t = 8$ are given in Figure 1. This figure indicates that single-locus F_c has a large variation and that the distribution of $F_c/\bar{F}_c$ approximately follows the χ^2 distribution with one degree of freedom. Since the sum of χ^2's is again a χ^2, the $nF_c/\bar{F}_c$ for n loci is expected to follow the χ^2 distribution with n degrees of freedom. Indeed, our statistical test confirmed this suggestion. However, the χ^2 distribution applies only approximately, and, as t increases, the deviation from the χ^2 distribution increases. This can be seen from the $s^2/\bar{F}^2$ value in Table 2. This value should be exactly 2, if $F_c/\bar{F}_c$ follows the χ^2 distribution. It is indeed roughy 2 for $t = 1$, but tends to be smaller than 2 for $t = 8$. This is particularly so when S is small. In Figure 1, the distribution of F_c for 20 loci is somewhat narrower than the $(\bar{F}_c/n)\chi^2_{(n)}$ distribution. This occurred because the $s^2/\bar{F}^2$ for the single-locus F_c is not 2, but 1.75 in this case. Nevertheless, for practical purposes, we can assume that $F_a/\bar{F}_a$, $F_b/\bar{F}_b$ and $F_c/\bar{F}_c$ all approximately follow the χ^2 distribution, unless t is large. The property that $s^2/\bar{F}^2$ is approximately 2 is also seen in PAMILO and VARVIO-AHO's study of F_a. The same property has been noted by LEWONTIN and KRAKAUER (1973) when the inbreeding coefficient is computed from a number of subpopulations.

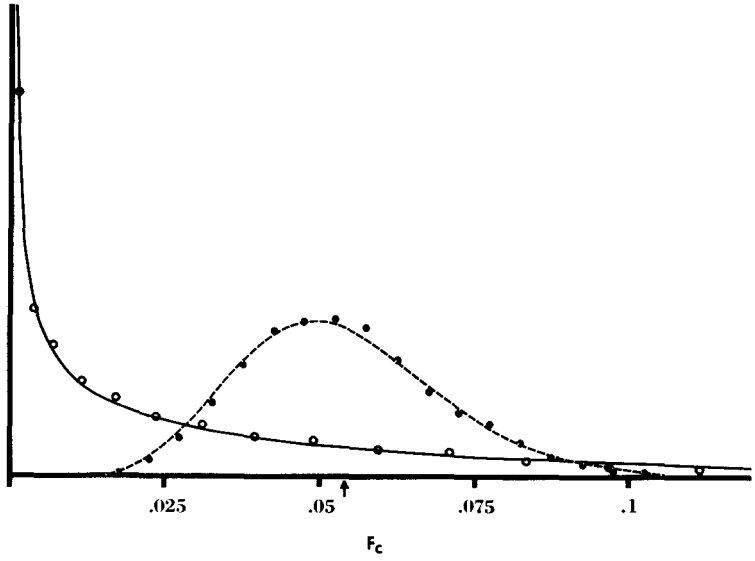

FIGURE 1.—Distribution of F_c based on single-locus (one allele) and 20-loci (20 independent alleles) data for the case of $N = 100$, $S = 40$ and $t = 8$. Open circles: single-locus F_c. Closed circles: 20-loci F_c. These results were obtained by computer simulation. The curves represent the distributions of $(\bar{F}_c/n)\chi^2_{(n)}$, where $\chi^2_{(n)}$ is the χ^2 value with n degrees of freedom. The arrow sign indicates $\bar{F}_c$.

Our simulation for the case of $p \neq 0.5$ has shown that $\bar{F}_a$ is much more sensitive to the variation of p than are $\bar{F}_b$ and $\bar{F}_c$. When $t = 1$ and $S = 20$, $\bar{F}_a$ increases as p decreases from 0.5 to 0.1. Thus, the $\bar{F}_a$ for $p = 0.1$ was 0.058 as compared with $\bar{F}_a = 0.048$ for the case of $p = 0.5$. However, the $\bar{F}_a$ for $p = 0.05$ was 0.041. The value of $s^2/\bar{F}_a^2$ also increased as $\bar{F}_a$ increased. For example, the value for $p = 0.1$ was 4.43. On the other hand, the values of $\bar{F}_b$ and $\bar{F}_c$ remained more-or-less the same as p decreased from 0.5 to 0.05. For example, the $\bar{F}_c$ for $p = 0.1$ was 0.046. However, the $s^2/\bar{F}^2$ value decreased gradually as p decreased in both F_b and F_c.

Table 3 shows the results for F_a and F_c for the cases of $t = 1$, 4 and 8, and $p = 0.1$. Comparison of this table with Table 2 indicates that the mean and standard deviation of F_a for $p = 0.1$ are substantially larger than those for $p = 0.5$, unless $S = N$. Furthermore, the value of $s^2/\bar{F}^2$ is considerably greater than 2, indicating that the variance of $F_a/\bar{F}_a$ is greater than that for the χ^2 distribution. The values of $\bar{F}_c$ for $p = 0.1$ are, however, more-or-less the same as those for $p = 0.5$ in all generations examined. The only difference between the two cases is that the values of $s^2/\bar{F}^2$ for $p = 0.1$ tend to be smaller than those for $p = 0.5$. From this point of view, therefore, F_c seems to have a better statistical property. The sampling properties of F_b were nearly the same as those of F_c, but the variance of F_b was almost always slightly larger than that of F_c. It is clear from this study that the approximation of the distribution of $F/\bar{F}$ by the χ^2 distribution when p is smaller (or larger) than 0.5 is not as good as that for the case of $p = 0.5$.

TABLE 3

Means ($\overline{F}$) *of* F_a *and* F_c *and estimates of effective population size*

	S	t	$\overline{F}$	s	$s^2/\overline{F}^2$	$\hat{N}$
F_a	20	1	0.0581	0.1224	4.43	−79
		4	0.0758	0.1703	5.05	40
		8	0.0975	0.2301	5.57	66
	40	1	0.0245	0.0500	4.17	633
		4	0.0374	0.0661	3.13	82
		8	0.0607	0.1301	4.59	85
	100	1	0.0049	0.0071	2.10	98
		4	0.0200	0.0292	2.12	100
		8	0.0403	0.0646	2.57	99
F_c	20	1	0.0455	0.0599	1.73	111
		4	0.0589	0.0740	1.58	113
		8	0.0734	0.0868	1.40	128
	40	1	0.0202	0.0279	1.92	104
		4	0.0337	0.0437	1.68	115
		8	0.0530	0.0637	1.44	107
	100	1	0.0049	0.0071	2.05	99
		4	0.0201	0.0264	1.72	99
		8	0.0396	0.0496	1.57	101

These results were obtained by a Monte Carlo simulation with 5000 replications. The actual effective size was 100, and the initial gene frequency was 0.1. s is the standard deviation of F.

This is particularly so with F_c. Nevertheless, the χ^2 approximation is useful for getting a rough idea about the reliability of the estimate of N, so that we shall use this approximation in the following discussion.

Because of the nature of the approximate χ^2 distribution, F_c has a large variance unless a large number of loci are used. From Figure 1, it is clear that, even if 20 loci are used, the F_c may deviate considerably from the expected value, and consequently the estimate of $\hat{N}$ may deviate substantially from the true value. Figure 2 shows the relationship between F_c and $\hat{N}$ for the case of $S = 40$ and $t = 8$. If F_c is 0.055, $\hat{N}$ will be exactly 100. However, as F_c decreases from this value, $\hat{N}$ gradually increases and reaches ∞ when $F_c = 0.025$. When F_c further decreases, $\hat{N}$ suddenly becomes $-\infty$ and gradually increases up to -120 when $F_c = 0$. On the other hand, if F_c increases from 0.055, $\hat{N}$ gradually declines. The probability distribution of F_c for $t = 8$ is given in Figure 2 under the assumption that $F_c/\overline{F}_c$ is χ^2-distributed. This distribution shows that in the case of $n = 20$ $\hat{N}$ becomes negative with a probability of about 0.02 and becomes larger than 200 or negative with a probability of about 0.22. It may become smaller than 50 with a probability of about 0.05. Therefore,when the number of genes used is relatively small, the

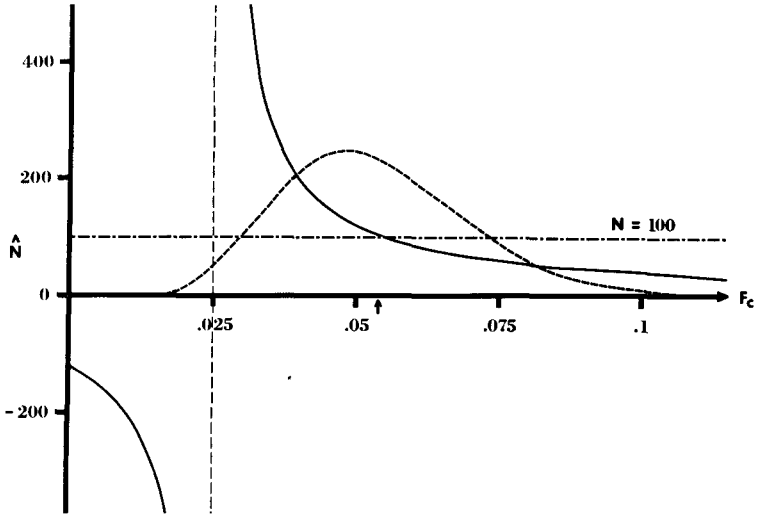

FIGURE 2.—Relationship (solid line) between F_c and $\hat{N}$ for the case of $N = 100$, $S = 40$ and $t = 8$. The broken-line curve represents the distribution of F_c based on data from 20 independent alleles under the assumption of χ^2 distribution. The arrow indicates $\bar{F}_c$.

estimate of $\hat{N}$ may be drastically different from the true value.

The sampling error associated with $\hat{N}$ is large when N is large and t is small. For given values of N and t, it is large when S is small. For example in the case of $N = 100$, $S = 20$, $t = 8$ and $n = 20$, $\hat{N}$ becomes negative with a probability of about 0.12. In general, if we use (18), F_c must be larger than $1/(2S_0) + 1/(2S_t)$ for $\hat{N}$ to be non-negative. Therefore, if sample size is small, $\hat{N}$ becomes negative with a high probability. On the other hand, if t is large or N is small, F_c is likely to be large, so that the effect of sampling error is small.

When N, S and t are given, the only way to increase the reliability of $\hat{N}$ is to increase the number of loci or alleles. If we assume that $nF_c/\bar{F}_c$ for n loci follows the χ^2 distribution with n degrees of freedom, the probability $[P(F_c < F_s)]$ that F_c is smaller than a specific value, F_s, can be computed from the table of χ^2 distribution. For example, if we want to know the probability that F_c leads to a negative value of $\hat{N}$, we set $F_s = 1/(2S_0) + 1/(2S_t)$, which is 0.025 for $S_0 = S_t = 40$. Therefore, $\chi_s^2 \equiv nF_s/\bar{F}_c = 0.460n$ for the case of $N = 100$ and $t = 8$, since $\bar{F}_c = 0.0543$ (Table 2). Thus, $P(F_c < F_s)$ or $P(\chi^2 < \chi_s^2)$ is 0.02 for $n = 20$ and 0.002 for $n = 40$. This indicates a large number of loci must be used to avoid a negative estimate. To make $\hat{N}$ lie within a narrower range, of course, an even larger number of loci is required. For example, the probability of $\hat{N}$ being larger than 300 or negative is computed by setting $F_s = 0.035$ (see Figure 2). This probability $[P(\chi^2 < 0.652n)]$ is 0.15 for $n = 20$ and 0.045 for $n = 40$. Namely, even if we use 40 genes, $\hat{N}$ can be larger than 300 (or negative) with a probability of

about 0.045. Of course, most investigators would be interested in getting only a rough idea of the effective size, and three-fold overestimates may not be very serious.

As mentioned earlier, we did not use sampling scheme II in our simulation. However, the sampling property of F for this case is expected to be similar to that for sampling scheme I if a proper adjustment is made for t. It is clear from (10) and (17) that, when $N \ll N_A$ the expected variance of gene-frequency changes $(x - y)$ for the case of sampling scheme II is larger than that for sampling scheme I by $2 \times (1/2N)$ for a given value of t. Therefore, the distribution of F for t for the former is expected to be essentially the same as that for $t + 2$ for the latter. This means that the approximation of the distribution of $nF/\bar{F}$ by the χ^2 for sampling scheme II is slightly less satisfactory, but for practical purposes it will not matter very much, unless t is extremely large.

SAMPLING ERROR OF $\hat{N}$

It is now clear that the estimate of $\hat{N}$ has a large sampling error. LEWONTIN and KRAKAUER (1973) presented a formula for computing the sampling variance of $\hat{N}$ when this was estimated by (1), assuming that $nF/\bar{F}$ has a χ^2 distribution. However, because of the complex relationship between F and $\hat{N}$ as shown in Figure 2, it is not meaningful to compute the sampling variance of $\hat{N}$ unless n is very large. In this case, a better way to determine the magnitude of the sampling error of $\hat{N}$ is to use the distribution of F_c, rather than that of $\hat{N}$. Since the distribution of $nF_c/\bar{F}_c$ is approximately χ^2, we can determine the F_c values that give the 2.5% (or 0.5%) and 97.5% (or 99.5%) cumulative probabilities. We can then compute the $\hat{N}$ values that correspond to these F_c values. The true value of N is expected to be somewhere between these two $\hat{N}$ values with probability 95% (or 99%). In practice, since nF_c/F_c tends to have a variance smaller than that of χ^2, the confidence interval thus determined tends to be an overestimate. A somewhat similar procedure for estimating the upper bound of $\hat{N}$ has been discussed by WILSON (1980).

NUMERICAL EXAMPLE

In September or October of 1966, 1967 and 1968, KRIMBAS and TSAKAS (1971) studied the gene frequencies at two esterase loci in an isolated population (Haghioi Apostoli) of the olive fly, *Dacus oleae*, near Athens, Greece. This population of olive flies infest an orchard of about 2000 olive trees and was subjected to an extensive spray of insecticide in 1968. Consequently, the census size of this population was reduced substantially in this year. The two esterase loci A and B were both highly polymorphic and included 18 and 13 electrophoretic alleles, respectively. However, some alleles were not observed in all years apparently because of limited sample size. From the allele frequencies presented in KRIMBAS and TSAKAS' paper, we computed F_c and estimates of N. When there were $n + 1$

alleles at a locus, we treated n alleles as independent alleles. This treatment seems to be satisfactory as an approximation, though they are not completely independent (E. POLLAK, personal communication). KRIMBAS and TSAKAS state that there are about four generations in one year in this fly. We have therefore assumed $t = 4$ for each year period. In the case of a two-year period (1966–1968), $t = 8$ was used. In autumn, the actual size of this population is apparently very large compared with the effective size, so that we used (18) to estimate N, assuming $N \ll N_A$.

The results obtained are presented in Table 4. The $\hat{N}$ values for A + B were obtained by using the average of F over the two loci and the harmonic means of S_0 and S_t. Table 3 also includes the 95% confidence interval of $\hat{N}$, computed by the method in the foregoing section. For all of $\hat{N}$'s, this confidence interval is very large. In the extreme case of locus A in 1966–1967, it has an interval of 240 to ∞. If we combine data from the two loci, the reliability of the estimate of N increases, but the confidence interval is still substantially large. Although the sampling errors of the estimates are very large, the $\hat{N}$'s for 1967–1968 are smaller than those for 1966–1967. This is probably due to the application of insecticide in 1968. However, the combined estimate of N for the period of 1966–1968 is quite large, so that the difference between the two periods could also be due to random error.

Comparisons of our estimates of N with those of KRIMBAS and TSAKAS indicate that both estimates are more or less the same, though the former tend to be larger

TABLE 4

F *and estimates* ($\hat{N}$) *of* N *obtained from esterase loci A and B in an isolated population of the olive fly* Dacus oleae (KRIMBAS *and* TSAKAS' *1971 data*)

F_c locus used	n^*	F_c	$\hat{N}$
1966–1967			
A	16	0.0061	583(240;∞)
B	12	0.0047	1056(314;∞)
A + B	28	0.0055	722(332;7408)
1967–1968			
A	17	0.0148	168(86;538)
B	11	0.0116	234(100;1984)
A + B	28	0.0135	189(108;446)
1966–1968			
A	14	0.0161	291(145;965)
B	11	0.0075	762(313;12069)
A + B	25	0.0123	400(225;961)

* Number of alleles per locus minus one.

$\hat{N}$ was obtained by (18). The figures in parentheses give the 95% confidence intervals. The standard error of F is approximately given by $F\sqrt{2/n}$ (see text).

than the latter in the 1966–1967 period. This rough agreement is of course expected, since the formulae used are both based on sampling scheme II. The effect of the assumption of sampling scheme can be seen by computing $\hat{N}_e$ by using (16). It becomes $\frac{1}{2}$ of $\hat{N}$ in Table 3 in the case of $t = 4$ and $\frac{3}{4}$ of $\hat{N}$ in the case of $t = 8$. Therefore, the effect is substantial in the present case. However, as mentioned earlier, sampling scheme II is likely to be more realistic than sampling scheme I in the present case.

Our estimates of $\hat{N}$ seem to be very small compared with the actual size of the Haghioi Apostoli population. KRIMBAS and TSAKAS state that the population size of olive flies becomes minimum in winter and, at that time, there are on average two flies per tree. If this estimate is correct, the minimum census size of the Haghioi Apostoli population is about 4000. If we assume that all of these individuals participate in mating and that the distribution of progeny size is approximately Poisson, we would expect the minimum effective size of this population to be about 4000. Our estimate (the value for 1966–1968) is one order of magnitude lower than this value.

There are two possible explanations for this discrepancy. One is that some of the individuals do not participate in mating and that the progeny size has a large variance compared with the mean. As mentioned earlier, the allelism rate of lethal genes in cage populations of Drosophila suggests that the effective size is considerably lower than the census size, even in an artificially controlled population. In natural populations, some adults may not mate at all, or, even if they mate, their progeny may not survive. If this happens, it is possible that our estimate of the effective size is roughly correct. Another explanation is local sampling of allele frequencies. The Haghioi Apostoli orchard is about 2.5 km long and 0.5 km wide (ZOUROS and KRIMBAS 1969), but the olive flies were apparently sampled from a small number of trees rather than from the entire orchard. KRIMBAS and TSAKAS (1971) state: "In the third (1968) sample, olive fruits bearing larvae were collected from three different locations, and the adults which hatched were electrophoresed." It is therefore possible that most of the flies sampled came from a relatively small number of broods produced by a limited number of parents. If this is the case, the estimated allele frequencies would not represent the frequencies of the entire population, even if the sample size is large. If this type of local sampling is conducted in the 0th and tth generations, the estimated allele-frequency changes $[F_c - 1/(2S_0) - 1/(2S_t)]$ will be larger than the actual changes in the entire population, and the effective population size will be underestimated. The effect of local sampling will be larger if there is local differentiation of allele frequencies in the population. In this connection, it is interesting to note that the estimate of WRIGHT's F_{ST} is often larger when a small area is used as a unit of subpopulation than when a large area is used (NEI and IMAIZUMI 1966b). At any rate, it is possible that our estimate of N is underestimated, not because of the deficiency of the statistical method, but because of inadequate sampling of allele frequencies.

DISCUSSION

We have seen that the estimate of effective population size is subject to a large sampling error unless a large number of genes and a larger number of generations are used. This is unfortunate, but we must accept it, since it accrues from the nature of stochastic change of gene frequencies. There are three ways to reduce the sampling error: (1) increase the number of independent alleles or loci used, (2) increase the number of generations involved, and (3) increase the ratio of sample size to effective size. In practice, however, most estimates obtained from the current scale of study would be subject to a large sampling error, since the number of genes that can be used is generally small. Nevertheless, even a crude estimate of effective size seems to be valuable, since we know very little at the present time about the effective size of natural populations. Furthermore, in the near future, the number of marker alleles that can be used for estimating N is expected to increase, since polymorphic alleles are now detectable by the restriction enzyme technique.

In the present study, we have made a number of simplifying assumptions. One of them is discrete generations. In many organisms, generations overlap, and strictly speaking our formulation does not apply to these organisms. However, if we know the generation time and the number of individuals becoming adult per unit time, the gene-frequency change in the population can be treated as though generations were discrete (*e.g.*, NEI and IMAIZUMI 1966a; HILL 1979). Therefore, our assumption does not seem to lead to serious error.

The second assumption we have made is no selection. Some polymorphic genes are certainly subject to selection; even if the genes themselves are selectively neutral, their behavior in populations may be affected by other adaptive genes that are closely linked to them. However, E. POLLAK (personal communication) has shown that the effect of selection on F is generally minor unless selection intensity is very large. Indeed, compared with the large sampling error of $\hat{N}$, the effect of selection seems to be generally much smaller, as long as a relatively short period of time is considered.

The third assumption is that the population under consideration is isolated and that no migration occurs from outside populations. This assumption would not hold in many natural populations. If a population is subdivided into many subpopulations and the gene-frequency changes are surveyed in one specific subpopulation, the estimate of effective size will be strongly affected by the migration rate among subpopulations. If the migration rate is very small, the estimate will be close to the effective size of the subpopulation studied. On the other hand, if it is very high (close to 1) and a large number of generations are included, the estimate would be close to the effective size of the entire population, provided that genes are sampled from the entire population. Therefore, it is important to know the population structure in the estimation of effective population size.

In some cases, investigators are interested in estimating the effective size of a subdivided population without knowing the migration pattern in the population.

In that case, we recommend that the estimate of effective size be computed by using gene-frequency data sampled from the entire area of the habitat of the population. The estimate thus obtained would be close to the effective size defined by WRIGHT (1943) for a subdivided population, irrespective of the migration pattern. Furthermore, this procedure would also eliminate the effect of local sampling in a single random mating population.

We thank R. C. LEWONTIN, E. POLLAK, P. PAMILO, C. B. KRIMBAS and two anonymous reviewers for their valuable comments. This study was supported by grants from the Public Health Service and the National Science Foundation.

LITERATURE CITED

CROW, J. F. and M. KIMURA, 1970 *An Introduction to Population Genetics Theory*. Harper and Row, New York.

FELLER, W., 1957 *An Introduction to Probability Theory and Its Applications*, Vol. I. John Wiley, New York.

HILL, W. G., 1979 A note on effective population size with overlapping generations. Genetics **92**: 317–322.

KRIMBAS, C. B. and S. TSAKAS, 1971 The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control—selection or drift? Evolution **25**: 454–460.

LEWONTIN, R. C. and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74**: 175–195.

MURATA, M., 1970 Frequency distribution of lethal chromosomes in small populations of *Drosophila melanogaster*. Genetics **64**: 559–571.

NEI, M., 1968 The frequency distribution of lethal chromosomes in finite populations. Proc. Natl. Acad. Sci. U.S. **60**: 517–524.

NEI, M. and Y. IMAIZUMI, 1966a Genetic structure of human populations. II. Differentiation of blood group gene frequencies among isolated populations. Heredity **21**: 183–190. ——, 1966b Genetic structure of human populations. III. Differentiation of ABO blood group gene frequencies in small areas of Japan. Heredity **21**: 461–472.

PAMILO, P. and S. VARVIO-AHO, 1980 On the estimation of population size from allele frequency changes. Genetics **95**: 1055.

PROUT, T., 1954 Genetic drift in irradiated experimental populations of *Drosophila melanogaster*. Genetics **39**: 529–546.

SCHAFFER, H. E., D. YARDLEY and W. W. ANDERSON, 1977 Drift or selection: A statistical test of gene frequency variation over generations. Genetics **87**: 371–379.

SOURDIS, J. and C. B. KRIMBAS, 1980 On Pamilo and Varvio-Aho's note about the estimation of effective population size. Genetics **96**: 561–563.

WILSON, S. R., 1980 Analyzing gene-frequency data when the effective population size is finite. Genetics **95**: 489–502.

WRIGHT, S., 1943 Isolation by distance. Genetics **28**: 114–138.

WRIGHT, S., TH. DOBZHANSKY and W. HOVANITZ, 1942 Genetics of natural populations. VIII. The allelism of lethals in the third chromosome of *Drosophila pseudoobscura*. Genetics **27**: 363–394.

ZOUROS, E. and C. B. KRIMBAS, 1969 The genetics of *Dacus oleae*. III. Amount of variation of two esterase loci in a Greek population. Genet. Research **14**: 249–258.

Corresponding editor: B. S. WEIR