Letter to the Editor

NOTE ON GENETIC DRIFT AND ESTIMATION OF EFFECTIVE POPULATION SIZE

Recently POLLAK (1983) proposed a new method for estimating the effective population size from allele frequency changes and compared his method with NEI and TAJIMA'S (1981) method. Although both methods are very similar, the former uses

$$F_{K1} = \frac{1}{K-1} \sum_{i=1}^{K} \frac{(x_i - y_i)^2}{(x_i + y_i)/2} \tag{1}$$

as a measure of standardized variance of gene frequency changes for a locus, whereas the latter uses

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

where K is the number of alleles, and x_i and y_i are the observed frequencies of the ith allele in the 0th and tth generations, respectively. When there are data from different loci, weighted means of F_{K1} and F_c, *i.e.*, $F_{K1} = \Sigma_j(K_j - 1)F_{K1j}/\Sigma_j(K_j - 1)$ and $F_c = \Sigma_j K_j F_{cj}/\Sigma K_j$, are used, where subscript j refers to the jth locus. Once F_{K1} or F_c is obtained, the effective size is estimated by formula (16) or (18) in NEI and TAJIMA (1981).

In this connection it should be noted that in NEI and TAJIMA'S (1981) definition of F_c, $K - 1$ is used in place of K. In their computation, however, K F_c's are first computed by eliminating one allele at a time from allele A_1 to allele A_K, and the average is used as the final value of F_c. This average is equal to (2). For example, in the case of three alleles, we can compute three F_c's, *i.e.*, $F_{c12} = (F_{c1} + F_{c2})/2$, $F_{c13} = (F_{c1} + F_{c3})/2$ and $F_{c23} = (F_{c2} + F_{c3})/2$, where $F_{ci} = (x_i - y_i)^2/[(x_i + y_i)/2 - x_i y_i]$. The average is

$$\begin{aligned} F_c &= (F_{c12} + F_{c13} + F_{c23})/3 \\ &= (F_{c1} + F_{c2} + F_{c3})/3. \end{aligned}$$

This shows that NEI and TAJIMA'S F_c is identical with (2). However, note that the number of degrees of freedom for computing the χ^2 is $K - 1$ rather than K.

At any rate, when he compared (1) and (2), POLLAK (1983) concluded that F_c is superior to F_{K1} for $K = 2$ but inferior for $K \geq 3$. This conclusion is based on the following observations. (1) The expectation of F_{K1} is approximately equal to that of F_c. (2) The maximum value of F_{K1} is $4/(K - 1)$, whereas that of F_c is 2. (3) When $K = 2$, the variance of F_c is approximately equal to that of F_{K1}. (4) The

variance of F_{K1} is smaller than that of F_c when the initial frequencies vary substantially with $K \geq 3$. However, his formulations involve some approximations. In this note we show that observations (2) and (3) are incorrect and present more accurate formulas for the variances of F_{K1} and F_c.

Equation (2) can be written as

$$F_c = \frac{4}{K}\left\{1 - \sum_{i=1}^{K} \frac{x_i y_i (2 - x_i - y_i)}{x_i + y_i - 2x_i y_i}\right\}. \tag{3}$$

Thus, the maximum value of F_c is $4/K$, which is always smaller than that $[4/(K-1)]$ of F_{K1}.

In order to obtain the variances of F_{K1} and F_c, POLLAK (1983) considered the variance of the numerator of (1) or (2), but ignored the variance of the denominator and the covariance between the numerator and denominator. He did this because he was interested in the case of relatively large sample sizes. When sample sizes are small, however, we must consider all of these components. (When gene frequency data for many different alleles or loci are used, even a small sample size gives a fairly reliable estimate of N.) In this case the variance of F_{K1} becomes

$$V(F_{K1}) \approx \frac{2F^2}{K-1} - \frac{F(4G - F^2)}{4(K-1)^2}\left(\sum_{i=1}^{K} \frac{1}{p_i} - K^2\right), \tag{4}$$

approximately, where p_i is the frequency of the ith allele in the population at generation 0, and F and G are quantities dependent on the sampling scheme used. If we use sampling scheme I of NEI and TAJIMA (1981), they become

$$F = \frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t-2}{2N},$$

$$G = \left(\frac{1}{2S_0} - \frac{1}{2N}\right)\left(\frac{1}{2S_0} - \frac{2}{2N}\right) + \left(\frac{1}{2S_t}\right)^2 + \frac{3}{2S_t}\frac{t-1}{2N} + \frac{(t-1)(3t-4)}{2}\left(\frac{1}{2N}\right)^2,$$

where S_0 and S_t are the sample sizes in the 0th and tth generations, respectively, and N is the effective population size. In their sampling scheme II we have

$$F = \frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t}{2N},$$

$$G = \left(\frac{1}{2S_0}\right)^2 + \left(\frac{1}{2S_t}\right)^2 + \frac{3}{2S_t}\frac{t}{2N} + \frac{t(3t-1)}{2}\left(\frac{1}{2N}\right)^2.$$

On the other hand, the variance of F_c is approximately given by

$$\begin{aligned} V(F_c) \approx{}& \frac{2F(F-2H)}{K}\left\{1 + \frac{1}{K}\sum\sum_{i \neq j} \frac{p_i p_j}{(1-p_i)(1-p_j)}\right\} \\ &- \frac{F(4G-F^2)}{K^2}\left\{\sum_{i=1}^{K} \frac{(1/2 - p_i)^2}{p_i(1-p_i)} - \sum\sum_{i \neq j} \frac{(1/2 - p_i)(1/2 - p_j)}{(1-p_i)(1-p_j)}\right\}, \end{aligned} \tag{5}$$

where $H = \left(\frac{1}{2S_0} - \frac{1}{2N}\right)\left(\frac{1}{2S_t} + \frac{t-1}{2N}\right)$ in sampling scheme I and $H = \frac{1}{2S_0} \cdot \left(\frac{1}{2S_t} + \frac{t}{2N}\right)$ in sampling scheme II. The moments of gene frequency changes necessary for obtaining (4) and (5) are given by NEI and TAJIMA (1981) and POLLAK (1983), except the following ones [*cf.*, pp. 332–335 in CROW and KIMURA (1970)].

$$E[(x_i - p_i)^3 + (y_i - p_i)^3] \approx p_i(1 - p_i)(1 - 2p_i)G,$$

$$E[(x_i - p_i)^2(x_j - p_j) + (y_i - p_i)^2(y_j - p_j)] \approx p_ip_j(2p_i - 1)G.$$

Before we compare $V(F_c)$ with $V(F_{K1})$, let us examine the accuracies of (4) and (5), since these equations still involve some approximations. For this purpose, we use the results obtained from a computer simulation by NEI and TAJIMA (1981) for sampling scheme I. We compare the quantities defined as

$$k_{K1} = (K - 1)V(F_{K1})/[E(F_{K1})]^2, \tag{6a}$$

$$k_c = (K - 1)V(F_c)/[E(F_c)]^2. \tag{6b}$$

The observed values of (6a) and (6b) obtained from the computer simulation and the expected values from (4) and (5) are presented in Table 1. There are some differences between the expected and observed values. However, if we note that in POLLAK'S (1983) formulas $k_{K1} = k_c = 2$ for all cases of $K = 2$, the expected values obtained from (4) and (5) are much closer to the observed values than POLLAK'S. Particularly when p_i deviates from 0.5 or when t is large, (4) and (5) give much better values of k_{K1} and k_c than POLLAK'S formulas.

Table 1 also shows that, unlike POLLAK'S conclusion, k_c is always smaller than or equal to k_{K1} when $K = 2$. Since the expectations of F_c and F_{K1} are more or less the same, this indicates that F_c is a better quantity than F_{K1} for estimating effective population size. When $K \geq 3$, k_c is again smaller than k_{K1} if $p_i = 1/K$. This is because, in this case, $V(F_{K1})$ and $V(F_c)$ become

$$V(F_{K1}) \approx 2F^2/(K - 1),$$

$$V(F_c) \approx 2F(F - 2H)/(K - 1).$$

Therefore, $V(F_c) < V(F_{K1})$. When the initial frequencies vary considerably, however, k_c is usually slightly larger than k_{K1}. Several examples for $K = 3$ are shown in Table 2. Therefore, POLLAK'S observation (4) seems to be correct.

From this study, we can conclude that when a majority of loci studied have only two alleles, F_c is preferable to F_{K1}. If a majority of loci have more than two alleles and their frequencies deviate from $1/K$ considerably, then F_{K1} is slightly better than F_c. In any case, however, the difference between the variances of F_{K1} and F_c is very small, so that both methods can be used.

In this connection BRUCE WEIR has suggested that the following quantity (F_d), which is equivalent to LATTER'S (1973) ϕ^*, might give a better estimate of N than F_c.

TABLE 1

Observed and expected values of k_{K1} *and* k_c *for sampling scheme I*

				k_{K1}		k_c	
p_1	p_2	$S_0 = S_t$	t	Observed	Expected	Observed	Expected
0.5	0.5	20	1	2.01	2.00	1.89	1.96
			4	1.91	2.00	1.76	1.95
			8	1.87	2.00	1.70	1.94
		40	1	1.96	2.00	1.90	1.98
			4	1.96	2.00	1.86	1.98
			8	1.88	2.00	1.75	1.97
		100	1	2.13	2.00	2.11	2.00
			4	1.96	2.00	1.90	2.00
			8	2.06	2.00	1.95	2.00
0.1	0.9	20	1	1.83	1.98	1.73	1.93
			4	1.68	1.83	1.58	1.77
			8	1.49	1.67	1.40	1.59
		40	1	1.97	1.97	1.92	1.95
			4	1.75	1.85	1.68	1.83
			8	1.52	1.68	1.44	1.65
		100	1	2.06	1.97	2.05	1.97
			4	1.76	1.84	1.72	1.84
			8	1.63	1.66	1.57	1.66

Observed values were obtained from NEI and TAJIMA'S (1981) computer simulation. $N = 100$ and $K = 2$ are assumed.

TABLE 2

Theoretical values of k_{K1}, k_c *and* k_d

				$S_0 = S_t = 100$ $N = 1000$			$S_0 = S_t = 20$ $N = 100$		
p_1	p_2	p_3	Sampling scheme	k_{K1}	k_c	k_d	k_{K1}	k_c	k_d
1/3	1/3	1/3	I	2.00	1.99	1.99	2.00	1.94	1.94
			II	2.00	1.99	1.99	2.00	1.93	1.93
0.2	0.3	0.5	I	2.00	2.02	2.11	1.97	1.95	2.05
			II	2.00	2.02	2.11	1.97	1.93	2.03
0.1	0.4	0.5	I	1.99	2.07	2.56	1.87	1.96	2.48
			II	1.98	2.07	2.55	1.86	1.94	2.46
0.1	0.1	0.8	I	1.97	2.09	2.30	1.72	1.81	2.02
			II	1.96	2.09	2.29	1.69	1.77	1.98

$K = 3$ and $t = 8$ are assumed.

$$F_d = \sum_{i=1}^{K} (x_i - y_i)^2 \Big/ \sum_{i=1}^{K} [(x_i + y_i)/2 - x_i y_i]. \quad (7)$$

This is because REYNOLDS, WEIR and COCKERHAM'S (1983) computer simulation has shown that this gives a less biased estimate of inbreeding coefficient than F_c when p_i deviates from $1/K$ and t is large. However, the theoretical variance of F_d has not been determined. We have, therefore, derived a formula for this variance, which is given by

$$V(F_d) \approx \frac{2F(F - 2H)[\Sigma p_i^2 - 2\Sigma p_i^3 + (\Sigma p_i^2)^2]}{(1 - \Sigma p_i^2)^2} - \frac{F(4G - F^2)[\Sigma p_i^3 - (\Sigma p_i^2)^2]}{(1 - \Sigma p_i^2)^2}. \quad (8)$$

The numerical values of $k_d = (K - 1)V(F_d)/[E(F_d)]^2$ in comparison with k_{K1} and k_c are given in Table 2. When $p_i = 1/K$, k_d is virtually the same as k_c. However, as p_i deviates from $1/K$, k_d becomes larger than k_c, and the difference can be substantial. This is in agreement with the results of REYNOLDS, WEIR and COCKERHAM (1983) from computer simulation, in which $V(F_d)$ was shown to be considerably larger than $V(F_c)$, although the smaller bias of F_d resulted in a smaller mean squared error for F_d than for F_c. (Simulation of REYNOLDS, WEIR and COCKERHAM also shows that, for $t = 20$ and $K = 2$, $V(F_d)$ is smaller than $V(F_{K1})$ when $p_i = 0.5$ but larger than $V(F_{K1})$ when p_i deviates from 0.5 considerably.) This result is in agreement with our theoretical prediction. Note that the t of REYNOLDS, WEIR and COCKERHAM corresponds to our $2t$ because they considered two populations rather than one.) We can, therefore, conclude that F_c is better than F_d from this point of view. It should also be noted that in the case of estimation of effective population size the t value used is generally small (about 10 or less), and, in this case, the bias of the estimate of F obtained from F_c is very small even if p_i deviates considerably from $1/K$ (see Table 3 of NEI and TAJIMA 1981). Furthermore, F_c has the advantage that it is approximately distributed as a χ^2 variate, so that the confidence interval of the estimate of N can easily be estimated. If we consider all these factors, F_c seems to be better than F_d.

LITERATURE CITED

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

LATTER, B. D. H., 1973 Measures of genetic distance between individuals and populations. pp. 27–39. In: *Genetic Structure of Populations*, Edited by N. E. MORTON. University Press of Hawaii, Honolulu, Hawaii.

NEI, M. and F. TAJIMA, 1981 Genetic drift and estimation of effective population size. Genetics **98:** 625–640.

POLLAK, E., 1983 A new method for estimating the population size from allele frequency changes. Genetics **104:** 531–548.

REYNOLDS, J., B. S. WEIR and C. C. COCKERHAM, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics **105:** 767–779.

FUMIO TAJIMA and MASATOSHI NEI
Center for Demographic and Population Genetics
The University of Texas at Houston
P.O. Box 20334
Houston, Texas 77225