Effect of non-random sampling on the estimation of parameters in population genetics

FUMIO TAJIMA
Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan

(*Received 24 April 1995 and in revised form 26 June 1995*)

Summary

The amount and pattern of genetic variation in a population can be estimated from genes or DNA sequences sampled from the population. Although random sampling is assumed in almost all cases, we often do not know whether sampling is random or not. Using a simple non-random sampling model, the effects of non-random sampling on the estimation of parameters in population genetics were investigated. This non-random sampling model assumes that *n* genes are randomly sampled with replacement from *m* genes which were randomly sampled from a large random mating population, and various degrees of non-randomness can be generated by changing the value of *m*. The results obtained show that the effect of non-random sampling on the number of alleles and the number of segregating sites is substantially large whereas the effect of non-random sampling on heterozygosity and the average number of nucleotide differences is negligibly small unless non-randomness is extremely large. The effects of non-random sampling on the tests of neutrality were also investigated, and the results obtained indicate that the effect of non-random sampling is stronger on Fu and Li's tests than on Tajima's test.

1. Introduction

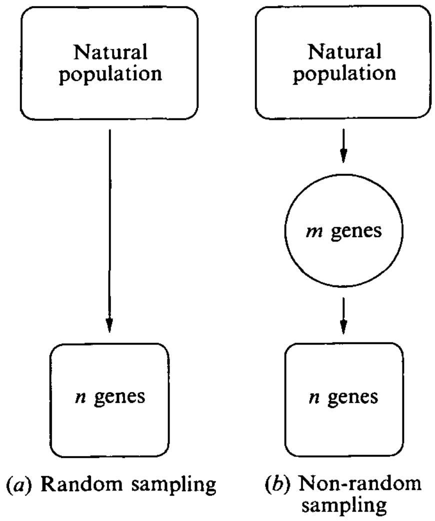
In order to identify the mechanism of the maintenance of genetic variability, we must first know the amount and pattern of genetic variation. The amount and pattern of genetic variation in a population can be estimated from genes or DNA sequences sampled from the population. In most, if not all, of the cases, random sampling is assumed. However, we often do not know whether genes are randomly sampled or not. For example, when fruit flies are collected by using a trap with banana in a day, some of the flies may be near relatives of some others. In the extreme case, all of them may be the offspring of a single pair of flies. In this case, all the flies might have the same mitochondrial DNA, ignoring newly arisen mutations and heteroplasmy, even when a large number of flies are collected. In another case, genes are sampled from animals in zoos without knowing the origins of these animals. They may have come from the same troop or family. Strictly speaking, random sampling means that each gene has an equal and independent chance of being sampled. Thus, most, if not all, of the samplings employed for population genetics study are not random. In some cases, for some unknown reason, sampling may deviate from randomness substantially. For example, Roy *et al.* (1994) suggest the possibility of non-random sampling in the studies of gray-wolf and coyote populations. Therefore, it is quite important to know the effect of non-random sampling on the estimation of parameters which can be used for identifying the mechanism of the maintenance of genetic variation. If the effect is negligibly small, then we do not have to seriously consider how genes are sampled from a population. On the other hand, if the effect is substantial, then we have to be careful with sampling.

In this paper, using a simple non-random sampling model, I will examine the effect of non-random sampling. Although there might be a large number of models for non-random sampling, I will consider only one simple model. Because of its simplicity, this model will be applicable to various cases. Throughout this paper, we consider only a large random mating population, since the main purpose of this paper is to investigate the effect of non-random sampling.

2. Model

In this model, *m* genes are first sampled from the population at random, and *n* genes are then randomly sampled from the *m* genes with replacement. Fig. 1

https://doi.org/10.1017/S0016672300034704 Published online by Cambridge University Press



(*a*) Random sampling (*b*) Non-random sampling

Fig. 1. Random sampling model (*a*) and non-random sampling model (*b*).

shows this sampling scheme together with random sampling scheme. It might be clear that, as m increases, non-random sampling approaches random sampling. In the case where all individuals sampled from the population are the offspring of a single pair of diploid individuals, we assume $m = 4$ for locus on autosome, $m = 3$ for that of X chromosome, and $m = 1$ for that of Y chromosome as well as for that of mitochondrial DNA under this model. This means that the value of m depends on the location of locus even for the same sampling. In this model, various degrees of non-randomness can be generated by changing the value of m. Thus, we can call $1/m$ the index of non-randomness. It should be noted here that this non-random sampling model assumes that each gene is sampled with an equal probability, but samples are not independent.

3. Theory

Using this model, we will examine the effect of non-random sampling on the estimates of parameters, which include allele frequency, heterozygosity, the number of alleles, the average number of nucleotide differences, and the number of segregating (or polymorphic) sites. We will also examine the effect on the tests of neutrality developed by Tajima (1989*b*) and Fu & Li (1993).

(i) *Allele frequency*

When the frequency of allele is p in the population, the probability distribution of allele frequency, x, in a sample of n genes is given by

$$\mathrm{Prob}\{x = i/n\} = \frac{n!}{i!(n-1)!}p^{i}(1-p)^{n-i} \tag{1}$$

for random sampling (see Fig. 1 *a*). On the other hand, considering binomial sampling twice (compound distribution), the probability distribution of x for non-random sampling (see Fig. 1 *b*) can be given by

$$\mathrm{Prob}\{x = 0\} = \sum_{j=0}^{m-1}(1-j/m)^{n}w_{j}, \tag{2a}$$

$$\mathrm{Prob}\{x = i/n\} = \frac{n!}{i!(n-i)!}\sum_{j=1}^{m-1}(j/m)^{i}(1-j/m)^{n-i}w_{j} \quad \text{for} \quad 1 \leqslant i \leqslant n-1, \tag{2b}$$

$$\mathrm{Prob}\{x = 1\} = \sum_{j=1}^{m}(j/m)^{n}w_{j}, \tag{2c}$$

where w_j is given by $w_j = m!p^{j}(1-p)^{m-j}/\{j!(m-j)!\}$. From (2*a*), (2*b*) and (2*c*), we can obtain the expectation and variance of x for non-random sampling, which are given by

$$\mathrm{E}(x) = p \quad \text{and} \quad \mathrm{V}(x) = \frac{m+n-1}{mn}p(1-p). \tag{3}$$

As m increases, $\mathrm{V}(x)$ approaches $p(1-p)/n$, which is the variance of x for random sampling. It is clear from (3) that unbiased estimates of allele frequency can be obtained even by non-random sampling.

Table 1. *Standard deviations of allele frequency, where* p = 0·5 *is assumed*

	m				
n	4	20	100	500	Random sampling
10	0·285	0·190	0·165	0·160	0·158
20	0·268	0·156	0·122	0·114	0·112
50	0·257	0·131	0·086	0·074	0·071
100	0·254	0·122	0·071	0·055	0·050
500	0·251	0·114	0·055	0·032	0·022

Standard deviations of x obtained by (3) are shown in Table 1, where $p = 0·5$ is assumed. We can see from this table that the effect of non-random sampling on the estimate of allele frequency depends on the ratio of m to n. Namely, if m is substantially larger than n, the variance of x for non-random sampling is close to that for random sampling. In fact, the ratio of variance of x for non-random sampling to that for random sampling is given by $1+n/m-1/m$. This means that if n is small, we do not have to seriously consider how genes are sampled from a population unless m is extremely small. On the other hand, if n is large, we have to be careful with sampling.

(ii) *Heterozygosity*

Heterozygosity in a population can be defined as

$$H = 1-\sum_{i} p_{i}^{2}, \tag{4}$$

where p_i is the frequency of the ith allele in the population. We usually estimate heterozygosity by

Table 2. *Expectations and standard deviations of heterozygosity, where* $p_1 = 0{\cdot}7$, $p_2 = 0{\cdot}2$ *and* $p_3 = 0{\cdot}1$ *are assumed*

	$m = 20$		$m = 100$		Random sampling	
n	$\mathrm{E}(\hat{H})$	$\surd\mathrm{V}(\hat{H})$	$\mathrm{E}(\hat{H})$	$\surd\mathrm{V}(\hat{H})$	$\mathrm{E}(\hat{H})$	$\surd\mathrm{V}(\hat{H})$
10	0·437	0·188	0·455	0·169	0·460	0·164
20	0·437	0·150	0·455	0·122	0·460	0·113
50	0·437	0·126	0·455	0·085	0·460	0·070
100	0·437	0·117	0·455	0·069	0·460	0·049
500	0·437	0·109	0·455	0·054	0·460	0·022

$$\hat{H} = \frac{n}{n-1}(1-\sum_i x_i^2), \tag{5}$$

where x_i is the frequency of the ith allele in a sample of n genes. Nei & Roychoudhury (1974) and Nei (1978) showed that (5) gives the unbiased estimate of H when n genes are sampled at random from the population. When sampling is not random, however, the expectation of $\hat{H}$ is given by

$$\mathrm{E}(\hat{H}) = \frac{m-1}{m}H, \tag{6}$$

which can be obtained by using

$$\mathrm{E}(x_i^2) = (m+n-1)\{p_i(1-p_i)+p_i^2\}/(mn)$$

[see (3)]. Therefore, $\hat{H}$ gives an underestimate of H under non-random sampling, and the amount of bias depends only on m. Namely, the amount of underestimation is substantial only when m is very small.

In the case of random sampling, Nei (1978) showed that the sampling variance of $\hat{H}$ is given by

$$\mathrm{V}(\hat{H}) = \frac{2}{n(n-1)}[\sum_i p_i^2-(\sum_i p_i^2)^2 + 2(n-2)\{\sum_i p_i^3-(\sum_i p_i^2)^2\}]. \tag{7}$$

On the other hand, the sampling variance of $\hat{H}$ for non-random sampling is shown in the Appendix.

The standard deviations of $\hat{H}$ as well as the expectations of $\hat{H}$ are shown in Table 2, where $p_1 = 0{\cdot}7$, $p_2 = 0{\cdot}2$ and $p_3 = 0{\cdot}1$ are assumed. In this table, the expectations and standard deviations of $\hat{H}$ for random sampling are also shown. We can see from this table that the effect of non-random sampling on the expectation of $\hat{H}$ is negligible unless m is very small, whereas the effect on the standard deviation is strong when m/n is small. From these results, we can conclude that, even when genes are not randomly sampled, heterozygosity can be used except when m is very small, although the standard error of $\hat{H}$ might be larger than that for random sampling especially when m/n is small.

(iii) *Number of alleles*

Ewens (1972) showed that when mutants are selectively neutral (Kimura, 1968, 1983), the expectation and variance of the number of alleles, n_a, in a sample of n genes chosen at random from the population are given by

$$\mathrm{E}(n_a) = \sum_{i=1}^{n}\frac{M}{M+i-1} \quad \text{and} \quad \mathrm{V}(n_a) = \sum_{i=1}^{n}\frac{(i-1)M}{(M+i-1)^2}, \tag{8}$$

where $M = 4Nv$, N is the effective population size, and v is the mutation rate per gene per generation. In the case of non-random sampling, the expectation and variance of n_a in a sample of n genes can be obtained as follows.

Let us consider the case where i genes among m genes are chosen at least once, and denote by $Q(i,n)$ the probability that i genes among m genes are chosen at least once, given that n genes are chosen from m genes with replacement. When $n = 1$, obviously we have $Q(1,1) = 1$. When $n = 2$, the probability that the gene sampled at the second sample is the same as that of the first sample is $1/m$, whereas the probability that the gene sampled at the second sample is not the same as that of the first sample is $1-1/m$. Therefore, we have

$$Q(1,2) = Q(1,1)/m = 1/m$$

and

$$Q(2,2) = (1-1/m)\,Q(1,1) = 1-1/m.$$

In general, the probability that the gene sampled at the nth sample is the same as one of i genes already sampled before the nth sample is i/m, and the probability that the gene sampled at the nth sample is different from $i-1$ genes already sampled before the nth sample is $1-(i-1)/m$. Therefore, we have

$$Q(i,n) = \frac{i}{m}Q(i,n-1)+\left(1-\frac{i-1}{m}\right)Q(i-1,n-1). \tag{9}$$

Solving (9), we obtain

$$Q(i,n) = \mathrm{S}_n^{(i)}\frac{(m-1)!}{m^{n-1}(m-i)!} \tag{10}$$

for $1 \leqslant i \leqslant m$ and $i \leqslant n$, otherwise $Q(i,n) = 0$, where $\mathrm{S}_n^{(i)}$ is the Stirling number of the second kind. Then, the expectation of n_a for non-random sampling is given by

$$\mathrm{E}(n_a) = \sum_{i=1}^{a} Q(i,n)\sum_{j=1}^{i}\frac{M}{M+j-1}, \tag{11}$$

where a is the smaller one of m and n. The variance of n_a for non-random sampling is given by $\mathrm{V}(n_a) = \mathrm{E}(n_a^2)-\{\mathrm{E}(n_a)\}^2$, where $\mathrm{E}(n_a)$ is given by (11) and $\mathrm{E}(n_a^2)$ is given by

$$\mathrm{E}(n_a^2) = \sum_{i=1}^{a} Q(i,n)\left\{\sum_{j=1}^{i}\frac{(j-1)M}{(M+j-1)^2}+\left(\sum_{j=1}^{i}\frac{M}{M+j-1}\right)^2\right\}, \tag{12}$$

in which a is the smaller one of m and n.

Computations of (11) and (12) are cumbersome, so that the following approximations might be useful.

Table 3. *Expectations and standard deviations of the number of alleles observed in a sample of* n *genes from a population, where* M = 1 *is assumed*

	$m = 20$		$m = 100$		Random sampling	
n	$\mathrm{E}(n_a)$	$\surd\mathrm{V}(n_a)$	$\mathrm{E}(n_a)$	$\surd\mathrm{V}(n_a)$	$\mathrm{E}(n_a)$	$\surd\mathrm{V}(n_a)$
10	2·713	1·097	2·884	1·159	2·929	1·174
	(2·720)	(1·092)	(2·885)	(1·158)		
20	3·162	1·267	3·504	1·385	3·598	1·415
	(3·167)	(1·264)	(3·506)	(1·384)		
50	3·518	1·389	4·264	1·627	4·499	1·695
	(3·519)	(1·388)	(4·266)	(1·627)		
100	3·592	1·413	4·733	1·763	5·187	1·885
	(3·592)	(1·413)	(4·734)	(1·762)		
500	3·598	1·415	5·180	1·884	6·793	2·269
	(3·598)	(1·415)	(5·181)	(1·883)		

Note: Values in parentheses are approximate ones obtained by (14*a*) and (14*b*).

Table 4. *Expectations and standard deviations of the average number of nucleotide differences among a sample of* n *genes, where* M = 10 *is assumed*

	$m = 20$		$m = 100$		Random sampling	
n	$\mathrm{E}(k\mid m)$	$\surd\mathrm{V}(k\mid m)$	$\mathrm{E}(k\mid m)$	$\surd\mathrm{V}(k\mid m)$	$\mathrm{E}(k\mid m)$	$\surd\mathrm{V}(k\mid m)$
10	9·500	5·648	9·900	5·653	10·000	5·655
		(2·500)		(2·529)		(2·534)
20	9·500	5·332	9·900	5·330	10·000	5·331
		(1·667)		(1·689)		(1·693)
50	9·500	5·166	9·900	5·160	10·000	5·160
		(1·018)		(1·032)		(1·035)
100	9·500	5·114	9·900	5·107	10·000	5·107
		(0·711)		(0·722)		(0·723)
500	9·500	5·074	9·900	5·066	10·000	5·065
		(0·315)		(0·320)		(0·320)
∞	9·500	5·065	9·900	5·056	10·000	5·055
		(0·000)		(0·000)		(0·000)

Note: Values in parentheses are $\surd\mathrm{V}_s(k\mid m)$.

The expected number of genes chosen at least once among m genes, given that n genes are chosen from m genes with replacement, can be given by

$$A = \sum_{i=1}^{a} iQ(i,n) = m\{1-(1-1/m)^n\}, \tag{13}$$

where a is the smaller one of m and n. Then, the expectation and variance of n_a are approximately given by

$$\mathrm{E}(n_a) = \sum_{i=1}^{b} \frac{M}{M+i-1} + B \tag{14a}$$

$$\mathrm{V}(n_a) = \sum_{i=1}^{b} \frac{(i-1)M}{(M+i-1)^2} + B(1-B), \tag{14b}$$

where B is given by $(A-b)M/(M+b)$ and b is the largest integer which is smaller than or equal to A.

Numerical examples are shown in Table 3, where $M = 1$ is assumed. In the computation of $Q(i,n)$, (9) was used rather than (10) since $\mathrm{S}_n^{(i)}$ becomes very large. We can see from this table that the effect of non-random sampling on the number of alleles is substantial, especially when n is large. In fact, as n increases, the expectation and variance of n_a approach

$$\mathrm{E}(n_a) = \sum_{i=1}^{m} \frac{M}{M+i-1} \quad \text{and} \quad \mathrm{V}(n_a) = \sum_{i=1}^{m} \frac{(i-1)M}{(M+i-1)^2}, \tag{15}$$

which are the same as those for random sampling with a sample size of m. From these results, we can conclude that we have to be careful with sampling in the case where the number of alleles are used for measuring the amount of genetic variation.

(iv) *Average number of nucleotide differences*

The average number of (pairwise) nucleotide differences among a sample of n DNA sequences can be used for measuring the amount of genetic variation at the DNA level, and can be defined as

$$k = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij}, \tag{16}$$

where k_{ij} is the number of nucleotide differences between the ith and jth DNA sequences. Tajima (1983) showed that, when mutants are selectively neutral and when there is no recombination, the expectation and variance of k are given by

$$\mathrm{E}(k) = M \quad \text{and}$$

$$\mathrm{V}(k) = \frac{n+1}{3(n-1)}M + \frac{2(n^2+n+3)}{9n(n-1)}M^2. \tag{17}$$

As shown in the Appendix, the expectation of k for non-random sampling is given by

$$\mathrm{E}(k\mid m) = \frac{m-1}{m}M. \tag{18}$$

We can see from this equation that the effect of non-random sampling on the expectation of the average number of nucleotide differences is not substantial unless m is extremely small, as on the expectation of heterozygosity. The variance of k for non-random sampling is shown in the Appendix.

Numerical examples are shown in Fig. 2 and Table 4, where $M = 10$ is assumed. In Fig. 2, the distributions of k are given for $m = 20$ and 100 as well as for random sampling, where $n = 100$ is assumed. These distributions were obtained by computer simulation, the method of which is shown in the Appendix. We can see from this figure that the effect of non-random sampling on the distribution of k is negligible. In Table 4, the expectation and variance of k are given for $m = 20$ and 100 as well as those for random sampling. We can see from this table that the effect of

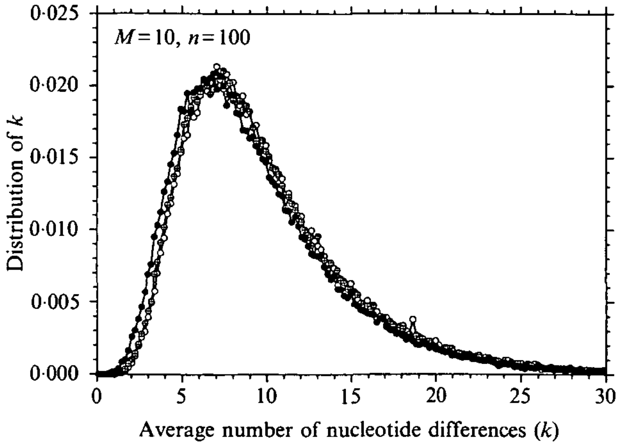


Fig. 2. Distribution of the average number of nucleotide differences (k), obtained by computer simulation. The averages (the standard deviations) of k for $m = 20$ (●), $m = 100$ (⊕) and random sampling (○) are 9·490 (5·128), 9·884 (5·110) and 10·025 (5·125), respectively.

non-random sampling on the expectation and variance of k is negligibly small. Therefore, we can conclude that, when we use the average number of nucleotide differences for measuring the amount of DNA polymorphism, we do not have to seriously consider how genes are sampled from a population.

(v) *Number of segregating sites*

The number of segregating (or polymorphic) nucleotide sites among a sample of DNA sequences is another measure of DNA polymorphism. Watterson (1975) showed that, when mutants are selectively neutral and when there is no recombination, the expectation and variance of the number of segregating sites, S, among a sample of n DNA sequences are given by

$$\mathrm{E}(S) = a_1(n)\,M \quad \text{and} \quad \mathrm{V}(S) = a_1(n)\,M + a_2(n)\,M^2, \tag{19}$$

where $a_1(n)$ and $a_2(n)$ are given by

$$a_1(n) = \sum_{i=1}^{n-1} (1/i) \quad \text{and} \quad a_2(n) = \sum_{i=1}^{n-1} (1/i^2) \quad \text{for } n \geqslant 2, \tag{20}$$

and $a_1(1) = a_2(1) = 0$. When n is large ($n > 5$), the approximations of $a_1(n)$ and $a_2(n)$ are available (Tajima, 1993*a*, *b*).

When sampling is not random, the expectation and variance of S can be obtained in the same way as in the case of the number of alleles. Namely, the expectation of S can be obtained by

$$\mathrm{E}(S) = \sum_{i=1}^{a} Q(i,n)\,a_1(i)\,M, \tag{21}$$

where a is the smaller one of m and n, and $Q(i,n)$ is defined by (10). Since $\mathrm{E}(S^2) = \mathrm{V}(S) + \{\mathrm{E}(S)\}^2 = a_1(n)\,M + [\{a_1(n)\}^2 + a_2(n)]\,M^2$ for random sampling, the expectation of S^2 for non-random sampling can be given by

$$\mathrm{E}(S^2) = \sum_{i=1}^{a} Q(i,n)\,(a_1(i)\,M + [\{a_1(i)\}^2 + a_2(i)]\,M^2). \tag{22}$$

Then, the variance of S can be obtained by $\mathrm{V}(S) = \mathrm{E}(S^2) - \{\mathrm{E}(S)\}^2$. As in the case of the number of alleles, using A defined as (13), the expectation and variance of S can be approximately given by

$$\mathrm{E}(S) = \{a_1(b) + (A-b)/b\}\,M, \tag{23a}$$

$$\mathrm{V}(S) = \{a_1(b) + (A-b)/b\}\,M + \{a_2(b) + (A-b)/b^2\}\,M^2, \tag{23b}$$

where b is the largest integer which is smaller than or equal to A. In order to estimate M, we often use $\hat{M} = S/a_1(n)$, since it gives the unbiased estimate of M under the assumption that mutants are selectively neutral, there is no recombination, and the population is panmictic (Watterson, 1975). The expectation and variance of $\hat{M}$ are given by $\mathrm{E}(\hat{M}) = \mathrm{E}(S)/a_1(n)$ and $\mathrm{V}(\hat{M}) = \mathrm{V}(S)/\{a_1(n)\}^2$, and it is now clear that $\hat{M}$ gives an underestimate of M when sampling is not random.

Table 5. *Expectations and standard deviations of* $\hat{\mathrm{M}}$ *estimated by using the number of segregating sites among a sample of* n *genes from a population, where* M = 10 *is assumed*

	$m = 20$		$m = 100$		Random sampling	
n	$\mathrm{E}(\hat{M})$	$\surd\mathrm{V}(\hat{M})$	$\mathrm{E}(\hat{M})$	$\surd\mathrm{V}(\hat{M})$	$\mathrm{E}(\hat{M})$	$\surd\mathrm{V}(\hat{M})$
10	9·142	4·728	9·824	4·765	10·000	4·772
	(9·177)	(4·705)	(9·828)	(4·759)		
20	8·690	3·869	9·722	3·924	10·000	3·935
	(8·707)	(3·857)	(9·728)	(3·919)		
50	7·733	3·109	9·464	3·194	10·000	3·214
	(7·736)	(3·106)	(9·468)	(3·191)		
100	6·840	2·696	9·112	2·801	10·000	2·834
	(6·840)	(2·695)	(9·114)	(2·800)		
500	5·224	2·056	7·614	2·162	10·000	2·244
	(5·224)	(2·056)	(7·615)	(2·160)		

Note: Values in parentheses are obtained by using approximations (23*a*) and (23*b*).

Numerical examples are shown in Table 5, where $M = 10$ is assumed. In this table, the expectation and standard deviation of $\hat{M}$ are given, and we can see that the effect of non-random sampling on $\hat{M}$ estimated by using the number of segregating sites is substantial when n is large. In fact, as n increases, the expectation and variance of S approach $\mathrm{E}(S) = a_1(m)\,M$ and $\mathrm{V}(S) = a_1(m)\,M + a_2(m)\,M^2$. The distributions of $\hat{M}$ obtained by computer simulation are shown in Fig. 3 (see the Appendix for the method of simulation). We can see from this figure that the effect of non-random sampling on the distribution of $\hat{M}$ are substantial. From these results, we can conclude that we have to be careful with sampling, when we use the number of

https://doi.org/10.1017/S0016672300034704 Published online by Cambridge University Press

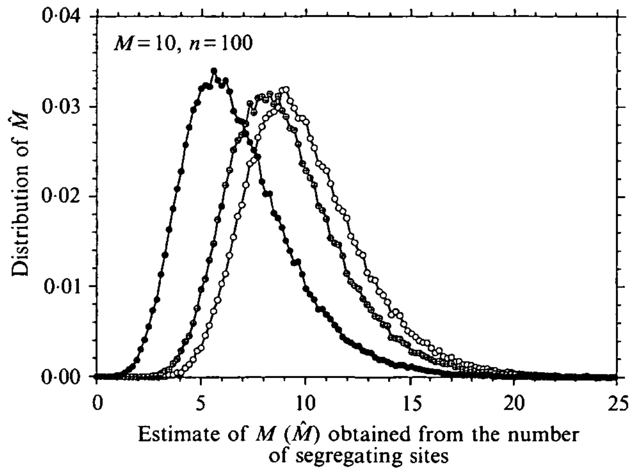


Fig. 3. Distribution of $\hat{M}$ estimated from the number of segregating sites, obtained by computer simulation. The averages (the standard deviations) for $m = 20$ (●), $m = 100$ (⊕) and random sampling (○) are 6·836 (2·697), 9·102 (2·801) and 10·017 (2·839), respectively.

segregating sites to estimate the amount of DNA polymorphism.

(vi) *Distribution of nucleotide frequencies*

When mutants are selectively neutral and when there is no recombination, Tajima (1989*b*) showed that the expected number of nucleotides whose frequency is i/n in a sample of n DNA sequences is given by

$$G_n(i) = \frac{n}{i(n-i)} M, \qquad (24)$$

where $1 \leqslant i \leqslant n-1$. In the case of non-random sampling, following (2*b*), we have

$$G_n(i \mid m) = \frac{n!}{i!(n-i)!} \sum_{j=1}^{m-1} (j/m)^i (1-j/m)^{n-i} G_m(j)$$

$$= \frac{n!}{i!(n-i)!\,m} \sum_{j=1}^{m-1} (j/m)^{i-1} (1-j/m)^{n-i-1} M, \quad (25)$$

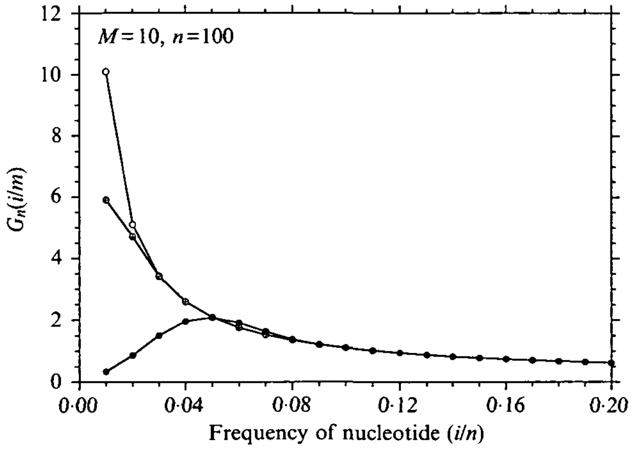


Fig. 4. Distribution of the expected number of nucleotides whose frequency is i/n, given that n DNA sequences are sampled from a population; $m = 20$ (●), m = 100 (⊕) and random sampling (○).

where $1 \leqslant i \leqslant n-1$. It should be noted that $G_n(i) = G_n(n-i)$ and $G_n(i \mid m) = G_n(n-i \mid m)$.

Numerical examples are shown in Fig. 4, where $M = 10$ and $n = 100$ are assumed. In this figure, $G_n(i \mid m)$'s for $m = 20$ and 100 as well as $G_n(i)$ are shown only when $0 < i/n \leqslant 0{\cdot}2$, since they are close to each other when $0{\cdot}1 < i/n < 0{\cdot}9$. We can see from this figure that non-random sampling affects the expected number of nucleotides whose frequency is close to 0 or 1, whereas it does not affect the expected number of nucleotides whose frequency is not close to 0 or 1. In other words, the effect of non-random sampling is substantial only on rare variants.

(vii) *Test of neutrality*

Under the assumptions of random mating population and no recombination, we can test the neutral mutation hypothesis (Kimura, 1968, 1983) from DNA sequences sampled from a natural population. Tajima's (1989*b*) method is based on the difference between the average number of nucleotide differences, k, and the estimate of M, $\hat{M} = S/a_1(n)$, obtained from the number of segregating sites, since the expectations of k and $\hat{M}$ are both M under these assumptions. As shown before, however, the expectations are not equal to M when sampling is not random. Since the effect of non-random sampling is stronger on S than on k, it might be expected that Tajima's method cannot be used when sampling is not random.

Recently, Fu & Li (1993) developed the other tests of neutrality. In addition to k and S, their method uses the number of singletons (S_1), which is defined as the number of nucleotides whose frequency is $1/n$ in a sample of n DNA sequences. Under the above assumptions, the expectation of S_1 is given by

$$E(S_1) = G_n(1) = \frac{n}{n-1} M, \qquad (26)$$

when sampling is random. Thus, we can estimate M by $\hat{M} = (n-1)S_1/n$. In fact, Fu & Li's method is based either on the difference between $S/a_1(n)$ and $(n-1)S_1/n$ or on the difference between k and $(n-1)S_1/n$. [Note that Fu & Li (1993) also developed the other methods which can be used when an outgroup sequence is available.] When sampling is not random, the expectations of k, S and S_1 can be obtained by using (25). Namely, the expectation of k can be given by

$$E(k) = \sum_{i=1}^{n-1} \frac{i(n-i)}{n(n-1)} G_n(i \mid m) = \frac{m-1}{m} M, \qquad (27)$$

which is identical with (18), the expectation of S can be given by

$$E(S) = \sum_{i=1}^{n-1} G_n(i \mid m)/2 = \sum_{i=1}^{m-1} \frac{m}{2i(m-i)} \times \{1-(1-i/m)^n - (i/m)^n\} M, \quad (28)$$

https://doi.org/10.1017/S0016672300034704 Published online by Cambridge University Press

Table 6. *Expectations of* k, S/a$_1$(n) *and* (n−1)S$_1$/n, *where* M = 10 *is assumed*

	m = 20			*m* = 100		
n	E(*k*)	$\frac{E(S)}{a_1(n)}$	$\frac{n-1}{n}E(S_1)$	E(*k*)	$\frac{E(S)}{a_1(n)}$	$\frac{n-1}{n}E(S_1)$
10	9·500	9·142	7·900	9·900	9·824	9·556
20	9·500	8·690	5·955	9·900	9·722	9·078
50	9·500	7·733	2·255	9·900	9·464	7·745
100	9·500	6·840	0·326	9·900	9·112	5·846
500	9·500	5·224	0·000	9·900	7·614	0·337

Note: $E(k) = E(S)/a_1(n) = (n-1)E(S_1)/n = 10$ for random sampling.

which leads to (21), and the expectation of S_1 can be given by

$$E(S_1) = G_n(1\mid m) = \frac{n}{m^{n-1}}\sum_{i=1}^{m-1} i^{n-2}M. \qquad (29)$$

Numerical examples are shown in Table 6, where E(*k*), $E(S)/a_1(n)$ and $(n-1)E(S_1)/n$ are given for $M = 10$. Note that all these values are equal to $M = 10$ for random sampling. We can see from this table that the effect of non-random sampling is substantial on $(n-1)S_1/n$, whereas non-random sampling does not affect *k* very much.

In order to know the effect of non-random sampling on the tests of neutrality, I have conducted computer simulation, the method of which is shown in the Appendix. In this simulation, three test statistics were compared, which are $D = d/\sqrt{V(d)}$, where $d = k - S/a_1(n)$ (Tajima, 1989*b*), $D^* = d^*/\sqrt{V(d^*)}$, where $d^* = S/a_1(n) - (n-1)S_1/n$ (Fu & Li, 1993), and $F^* = f^*/\sqrt{V(f^*)}$, where $f^* = k - (n-1)S_1/n$ (Fu & Li, 1993). The results of computer simulation are shown in Table 7 and Fig. 5, where $M = 10$ is assumed. In Table 7 the means and standard deviations of *D*, *D** and *F** are given. Figs 5*a*, *b* show the distributions of *D* and *D**, where $n = 50$ is assumed. It

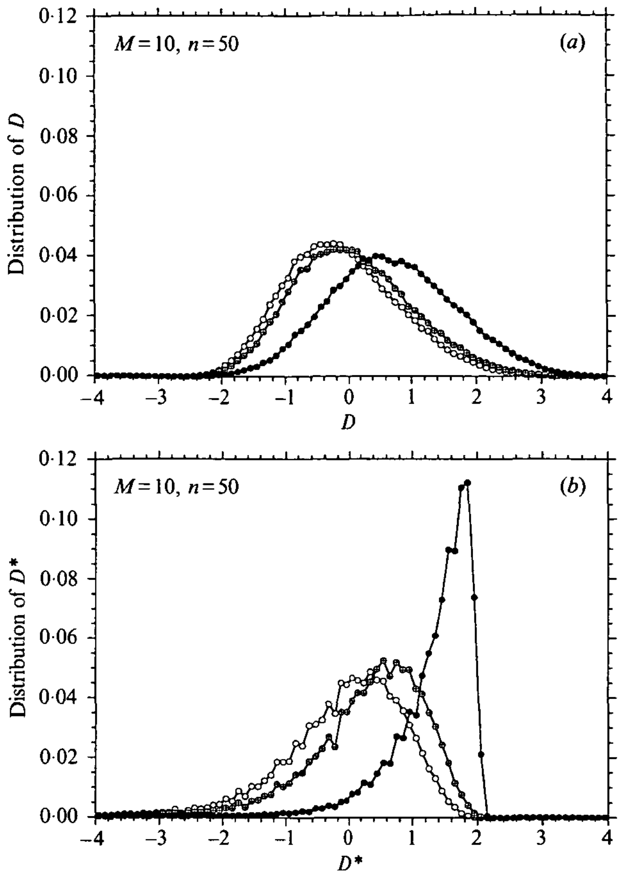


Fig. 5. Distributions of (*a*) *D* and (*b*) *D**, obtained by computer simulation; *m* = 20 (●), *m* = 100 (⊕), and random sampling (○).

Table 7. *Means and standard deviations* (S.DS) *of* D, D* *and* F* *obtained by using computer simulation, where* M = 10 *is assumed*

		D		*D**		*F**	
n	*m*	Mean	S.D.	Mean	S.D.	Mean	S.D.
10	20	0·124	0·942	0·161	0·957	0·172	1·052
	100	−0·031	0·902	−0·021	0·920	−0·027	1·012
	Random	−0·074	0·892	−0·070	0·912	−0·080	1·003
20	20	0·285	0·948	0·459	0·946	0·475	1·024
	100	−0·016	0·905	0·035	0·938	0·023	1·015
	Random	−0·090	0·889	−0·082	0·932	−0·098	1·008
50	20	0·671	0·983	1·279	0·769	1·266	0·848
	100	0·061	0·922	0·276	0·946	0·237	0·992
	Random	−0·102	0·898	−0·089	0·966	−0·111	1·007
100	20	1·091	1·016	1·880	0·407	1·884	0·623
	100	0·171	0·938	0·706	0·892	0·589	0·929
	Random	−0·096	0·903	−0·078	0·981	−0·102	0·995
500	20	2·044	1·137	2·093	0·302	2·512	0·709
	100	0·719	1·034	2·306	0·363	1·896	0·678
	Random	−0·091	0·907	−0·060	1·000	−0·091	0·970

https://doi.org/10.1017/S0016672300034704 Published online by Cambridge University Press

can be seen from these table and figures that the effect of non-random sampling is not very large in the case of D, whereas the effect is quite large in the cases of D^* and F^*. This is because the effect of non-random sampling is much stronger on $(n-1)\,S_1/n$ than on k or $S/a_1(n)$ (see Table 6). Therefore, we can conclude that we have to be very careful with sampling to apply the tests of neutrality, especially when D^* or F^* is used.

4. Discussion

In this paper, using a simple non-random sampling model, we have examined the effects of non-random sampling on the estimates of the amount of genetic variation in a population and on the tests of neutrality. The results indicate that the effect of non-random sampling on heterozygosity and the average number of nucleotide differences is small (see Tables 2 and 4) and that the effect on the number of alleles and the number of segregating sites is substantial (see Tables 3 and 5). Therefore, we can conclude that we have to be very careful with sampling in the case where the number of alleles or the number of segregating sites is used to estimate the amount of polymorphism, whereas we do not have to seriously consider how genes are sampled when heterozygosity or the average number of nucleotide differences is used to estimate the amount of polymorphism. In the case where we test the neutral mutation hypothesis by using the amount of DNA polymorphism, we have to be very careful with sampling, especially when Fu & Li's (1993) methods are used (see Table 7 and Fig. 5). We also note that all the three test statistics, D, D^* and F^*, tend to be positive under non-random sampling. Therefore if they are significantly smaller than zero, we do not have to consider the effect of non-random sampling as a cause of deviation. On the other hand, if they are significantly larger than zero, we have to consider whether sampling is random or not.

Although I used this simple non-random model, there is no reason to believe that non-random sampling always occurs in this way. To obtain the general effect of non-random sampling, we have to study the other types of non-random sampling.

When a population is subdivided and when migration rates among subpopulations are small, sampling biases might be expected if genes are sampled only from one subpopulation. This sampling is not random, since each gene does not have an equal change of being sampled from the population. In some cases, the average number of nucleotide differences, k, among DNA sequences sampled from one subpopulation gives the unbiased estimate of $4Nv$, where N is the effective size of entire population and v is the mutation rate per generation (Li, 1976; Slatkin, 1987; Strobeck, 1987), although this is not always the case (Tajima, 1989*a*, 1990). Furthermore, the number of segregating sites, $S/a_1(n)$, gives a biased estimate of $4Nv$, even when k gives the unbiased estimate (Tajima, 1989*a*). This means that the effect of non-random sampling is stronger on S than on k, as in the simple non-random sampling model.

In some cases, genes are sampled based on information about genetic variation. For example, genes are sampled and sequenced, based on allelic information revealed by electrophoresis (Kreitman, 1983). For theoretical studies under this sampling strategy, see Hudson & Kaplan (1986).

In population genetics, random sampling is assumed in almost all cases. As has been shown in this paper, we often obtain biased estimates of parameters if sampling is not random. Although the most efficient way to eliminate biases is random sampling, we may not be able to avoid non-random sampling in some cases. In such cases, if we can identify non-randomness from data, we will be able to avoid misinterpreting data. One possible way is to test whether or not Hardy–Weinberg law holds. Another possible way is to examine linkage disequilibrium. In both cases, however, additional information is necessary.

References

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Fu, Y.-X. & Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Hudson, R. R. & Kaplan, N. L. (1986). On the divergence of alleles in nested subsamples from finite populations. *Genetics* **113**, 1057–1076.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.

Li, W.-H. (1976). Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. *Theoretical Population Biology* **10**, 303–308.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590.

Nei, M. & Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics* **76**, 379–390.

Roy, M. S., Geffen, E., Smith, D., Ostrander, E. A. & Wayne, R. K. (1994). Patterns of differentiation and hybridization in North American wolflike canids, revealed by analysis of microsatellite loci. *Molecular Biology and Evolution* **11**, 553–570.

Slatkin, M. (1987). The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32**, 42–49.

Strobeck, C. (1987). Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**, 149–153.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Tajima, F. (1989*a*). DNA polymorphism in a subdivided population: the expected number of segregating sites in the two subpopulation model. *Genetics* **123**, 229–240.

https://doi.org/10.1017/S0016672300034704 Published online by Cambridge University Press

Tajima, F. (1989*b*). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Tajima, F. (1990). Relationship between migration and DNA polymorphism in a local population. *Genetics* **126**, 231–234.

Tajima, F. (1993*a*). Measurement of DNA polymorphism. In *Mechanisms of Molecular Evolution* (ed. N. Takahata and A. G. Clark), pp. 37–59. Sunderland, MA: Sinauer Associates.

Tajima, F. (1993*b*). Statistical analysis of DNA polymorphism. *Japanese Journal of Genetics* **68**, 567–595.

Watterson, G. A. (1975). On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology* **7**, 256–276.

Appendix

(i) *Sampling variance of the estimate of heterozygosity*

Following Nei & Roychoudhury (1974), the sampling variance of the estimate of heterozygosity ($\hat{H}$) for non-random sampling can be given by

$$V(\hat{H}) = \frac{n^2}{(n-1)^2}[\sum_i \mathrm{E}(x_i^4) + \sum_{i \neq j} \mathrm{E}(x_i^2 x_j^2) - \{\sum_i \mathrm{E}(x_i^2)\}^2]. \tag{A 1}$$

Using (2), $\mathrm{E}(x_i^4)$, $\mathrm{E}(x_i^2 x_j^2)$ and $\mathrm{E}(x_i^2)$ can be obtained as

$$\begin{aligned}\mathrm{E}(x_i^4) = \{&(n-1)(n-2)(n-3)\,\mathrm{E}(y_i^4)\\ &+6(n-1)(n-2)\,\mathrm{E}(y_i^3)\\ &+7(n-1)\,\mathrm{E}(y_i^2)+\mathrm{E}(y_i)\}/n^3,\end{aligned} \tag{A 2a}$$

$$\begin{aligned}\mathrm{E}(x_i^2 x_j^2) = [&(n-1)(n-2)(n-3)\,\mathrm{E}(y_i^2 y_j^2)\\ &+(n-1)(n-2)\{\mathrm{E}(y_i^2 y_j)+\mathrm{E}(y_i y_j^2)\}\\ &+(n-1)\,\mathrm{E}(y_i y_j)]/n^3,\end{aligned} \tag{A 2b}$$

$$\mathrm{E}(x_i^2) = \{(n-1)\,\mathrm{E}(y_i^2)+\mathrm{E}(y_i)\}/n \tag{A 2c}$$

for $i \neq j$, where $\mathrm{E}(y_i)$, $\mathrm{E}(y_i^2)$, $\mathrm{E}(y_i^3)$, $\mathrm{E}(y_i^4)$, $\mathrm{E}(y_i y_j)$, $\mathrm{E}(y_i^2 y_j)$, $\mathrm{E}(y_i y_j^2)$ and $\mathrm{E}(y_i^2 y_j^2)$ are given by

$$\mathrm{E}(y_i) = p_i, \tag{A 3a}$$

$$\mathrm{E}(y_i^2) = \{(m-1)p_i^2+p_i\}/m, \tag{A 3b}$$

$$\mathrm{E}(y_i^3) = \{(m-1)(m-2)p_i^3+3(m-1)p_i^2+p_i\}/m^2 \tag{A 3c}$$

$$\begin{aligned}\mathrm{E}(y_i^4) = \{&(m-1)(m-2)(m-3)p_i^4+6(m-1)(m-2)p_i^3\\ &+7(m-1)p_i^2+p_i\}/m^3,\end{aligned} \tag{A 3d}$$

$$\mathrm{E}(y_i y_j) = (m-1)p_i p_j/m, \tag{A 3e}$$

$$\mathrm{E}(y_i^2 y_j) = \{(m-1)(m-2)p_i^2 p_j+(m-1)p_i p_j\}/m^2, \tag{A 3f}$$

$$\mathrm{E}(y_i y_j^2) = \{(m-1)(m-2)p_i p_j^2+(m-1)p_i p_j\}/m^2, \tag{A 3g}$$

$$\begin{aligned}\mathrm{E}(y_i^2 y_j^2) = \{&(m-1)(m-2)(m-3)p_i^2 p_j^2\\ &+(m-1)(m-2)p_i p_j(p_i+p_j)+(m-1)p_i p_j\}/m^3.\end{aligned} \tag{A 3h}$$

(ii) *Expectation and variance of the average number of nucleotide differences among a sample of n genes*

When mutants are selectively neutral and when there is no recombination, the expectation and variance of the average number of nucleotide differences (k) among a sample of n genes are given by

$$\mathrm{E}(k) = M \quad \text{and} \quad \mathrm{V}(k) = \frac{n+1}{3(n-1)}M + \frac{2(n^2+n+3)}{9n(n-1)}M^2, \tag{A 4}$$

which can be obtained from the following formulae:

$$\mathrm{E}(k_{ij}) = M, \tag{A 5a}$$

$$\mathrm{E}(k_{ij}^2) = M+2M^2, \tag{A 5b}$$

$$\mathrm{E}(k_{ij}k_{ir}) = M/2+4M^2/3, \tag{A 5c}$$

$$\mathrm{E}(k_{ij}k_{rs}) = M/3+11M^2/9, \tag{A 5d}$$

where i, j, r and s are mutually different (Tajima, 1983).

In order to obtain the expectation and variance of k under non-random sampling, we have to consider how genes are sampled from m genes. To distinguish the expectations under non-random sampling from those under random sampling, we define $\mathrm{E}(x)$ under non-random sampling as $\mathrm{E}(x\,|\,m)$. If two genes sampled from m genes come from the same gene in m genes, then we have $\mathrm{E}(k_{ij}\,|\,m) = \mathrm{E}(k_{ii}) = 0$, whereas if they come from different genes, then we have $\mathrm{E}(k_{ij}\,|\,m) = \mathrm{E}(k_{ij}) = M$ from (A 5*a*). Since the probabilities of having the former and latter events are $1/m$ and $1-1/m$, respectively, we have

$$\mathrm{E}(k_{ij}\,|\,m) = \frac{\mathrm{E}(k_{ii})}{m}+\frac{m-1}{m}\mathrm{E}(k_{ij}) = \frac{m-1}{m}M. \tag{A 6a}$$

In the same way, we have

$$\mathrm{E}(k_{ij}^2\,|\,m) = \frac{\mathrm{E}(k_{ii}^2)}{m}+\frac{m-1}{m}\mathrm{E}(k_{ij}^2) = \frac{m-1}{m}(M+2M^2), \tag{A 6b}$$

$$\begin{aligned}\mathrm{E}(k_{ij}k_{ir}\,|\,m) &= \frac{\mathrm{E}(k_{ii}^2)}{m^2}+\frac{m-1}{m^2}\mathrm{E}(k_{ij}^2)+\frac{2(m-1)}{m^2}\mathrm{E}(k_{ii}k_{ij})\\ &\quad+\frac{(m-1)(m-2)}{m^2}\mathrm{E}(k_{ij}k_{ir})\\ &= \frac{m-1}{2m}M+\frac{2(m-1)(2m-1)}{3m^2}M^2,\end{aligned} \tag{A 6c}$$

$$\begin{aligned}\mathrm{E}(k_{ij}k_{rs}\,|\,m) &= \frac{\mathrm{E}(k_{ii}^2)}{m^3}+\frac{m-1}{m^3}\mathrm{E}(k_{ii}k_{jj})+\frac{2(m-1)}{m^3}\mathrm{E}(k_{ij}^2)\\ &\quad+\frac{4(m-1)}{m^3}\mathrm{E}(k_{ii}k_{ij})+\frac{2(m-1)(m-2)}{m^3}\mathrm{E}(k_{ii}k_{jr})\\ &\quad+\frac{4(m-1)(m-2)}{m^3}\mathrm{E}(k_{ij}k_{ir})\\ &\quad+\frac{(m-1)(m-2)(m-3)}{m^3}\mathrm{E}(k_{ij}k_{rs})\\ &= \frac{(m-1)(m+1)}{3m^2}M+\frac{(m-1)(11m^2-7m+6)}{9m^3}M^2.\end{aligned} \tag{A 6d}$$

https://doi.org/10.1017/S0016672300034704 Published online by Cambridge University Press

Then, the expectation of k is given by

$$\mathrm{E}(k\,|\,m) = \mathrm{E}(k_{ij}\,|\,m) = \frac{m-1}{m}M. \qquad \text{(A 7)}$$

The variance of k can be obtained from

$$\mathrm{V}(k\,|\,m) = \frac{2}{n(n-1)}\{\mathrm{E}(k_{ij}^2\,|\,m) + 2(n-2)\,\mathrm{E}(k_{ij}\,k_{ir}\,|\,m) + \frac{(n-2)(n-3)}{2}\mathrm{E}(k_{ij}\,k_{rs}\,|\,m)\} - \{\mathrm{E}(k\,|\,m)\}^2 \qquad \text{(A 8)}$$

(Tajima, 1983). Substituting (A 6*b*), (A 6*c*) and (A 6*d*) into (A 8), we finally obtain

$$\mathrm{V}(k\,|\,m) = \frac{(m-1)\{mn(n+1)+(n-2)(n-3)\}}{3m^2n(n-1)}M + \frac{2(m-1)\{m(m+1)(n^2+n+3)+3(n-2)(n-3)\}}{9m^3n(n-1)}M^2. \qquad \text{(A 9)}$$

As n increases, $\mathrm{V}(k\,|\,m)$ approaches

$$\mathrm{V}_{\mathrm{st}}(k\,|\,m) = \frac{(m-1)(m+1)}{3m^2}M + \frac{2(m-1)(m^2+m+3)}{9m^3}M^2, \qquad \text{(A 10)}$$

which can be called the stochastic variance (Tajima, 1983). The sampling variance can be obtained by $\mathrm{V}_{\mathrm{s}}(k\,|\,m) = \mathrm{V}(k\,|\,m) - \mathrm{V}_{\mathrm{st}}(k\,|\,m)$.

(iii) *Method of computer simulation*

The following computer simulation, which is similar to that of Tajima (1989*b*), was conducted in this study. First, the process in which n genes are randomly sampled from m genes is generated. Let $w(m,i)$ be the number of times the ith gene among m genes are chosen in this process. Then, $w(m,i)$'s for $i = 1,2,3,\ldots,$ and m can be generated by choosing one of 1, 2, 3, ..., and m with equal probability with replacement n times. Secondly, the process in which m genes are randomly sampled from a population is generated. In this process, the theory of gene genealogy is used. Two of 1, 2, 3, ..., and m are chosen with equal probability without replacement. If i and j are chosen ($i < j$), then $w(m-1,i)$ is computed by $w(m,i)+w(m,j)$, and $w(m-1,k)$ is computed by $w(m,k)$ for $k<j$ and $k \neq i$ or by $w(m,k+1)$ for $k \geqslant j$. In the same way, two of 1, 2, 3, ..., and $m-1$ are chosen at random and $w(m-2,i)$'s for $i = 1,2,3,\ldots,$ and $m-2$ are generated. We repeat this process until $w(2,1)$ and $w(2,2)$ are obtained. Now we generate mutations. The number of mutations, $z(j)$, occurred in $w(j,1)$, $w(j,2)$, $w(j,3)$, ..., and $w(j,j)$, is generated by assuming that $z(j)$ follows the geometric distribution,

$$P[z(j)] = \frac{(j-1)\,M^{z(j)}}{(M+j-1)^{z(j)+1}} \qquad \text{(A 11)}$$

(Watterson, 1975). Then, the number of mutations in each $w(j,i)$, $z(j,i)$, is generated by choosing one of $w(j,i)$'s at random with replacement $z(j)$ times. Under the theory of gene genealogy, $z(j)$ and $z(j,i)$ correspond to the number of mutations in j branches and the number of mutations in the ith branch among j branches between j DNA sequences and $j-1$ DNA sequences, respectively [see Tajima (1989*b*)], and $w(j,i)$ is the number of genes in the ith branch among j branches. Therefore, the average number of nucleotide differences can be computed by

$$k = 2\sum_{j=2}^{m}\sum_{i=1}^{j}\frac{w(j,i)\,[n-w(j,i)]}{n(n-1)}z(j,i), \qquad \text{(A 12)}$$

the number of segregating sites can be computed by

$$S = \sum_{j=2}^{m}\sum_{i=1}^{j}u(j,i)\,z(j,i), \qquad \text{(A 13)}$$

where $u(j,i) = 1$ if $1 \leqslant w(j,i) \leqslant n-1$ and $u(j,i) = 0$ otherwise, and the number of singletons can be computed by

$$S_1 = \sum_{j=2}^{m}\sum_{i=1}^{j}v(j,i)\,z(j,i), \qquad \text{(A 14)}$$

where $v(j,i) = 1$ if $w(j,i) = 1$ or $w(j,i) = n-1$ and $v(j,i) = 0$ otherwise. Finally, D, D^* and F^* are computed from k, S and S_1. For each set of parameter values, simulations are conducted 100000 times to obtain the means, variances and distributions of k, S, S_1, D, D^* and F^*. The method of computer simulation for random sampling is the same as that of Tajima (1989*b*), and simulations are also conducted 100000 times for each set of parameter values.