ESTIMATION OF AVERAGE HETEROZYGOSITY AND GENETIC DISTANCE FROM A SMALL NUMBER OF INDIVIDUALS

MASATOSHI NEI

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

Manuscript received November 1, 1977
Revised copy received February 13, 1978

ABSTRACT

The magnitudes of the systematic biases involved in sample heterozygosity and sample genetic distances are evaluated, and formulae for obtaining unbiased estimates of average heterozygosity and genetic distance are developed. It is also shown that the number of individuals to be used for estimating average heterozygosity can be very small if a large number of loci are studied and the average heterozygosity is low. The number of individuals to be used for estimating genetic distance can also be very small if the genetic distance is large and the average heterozygosity of the two species compared is low.

STUDYING the sampling variance of heterozygosity and genetic distance, NEI and ROYCHOUDHURY (1974) concluded that for estimating average heterozygosity and genetic distance a large number of loci rather than a large number of individuals per locus should be used when the total number of genes to be examined is fixed. Recently, GORMAN and RENZI (unpublished) have shown that in lizards even a single individual from each species provides genetic distance estimates that are quite useful for constructing dendrograms, provided that the genetic distances between species are sufficiently large. They also confirmed NEI and ROYCHOUDHURY'S (1974) theoretical conclusion that a relatively reliable estimate of average heterozygosity can be obtained from a small number of individuals if a large number of loci are examined. In this note I shall extend NEI and ROYCHOUDHURY'S (1974) study and present a further theoretical basis for GORMAN and RENZI'S unpublished observations. I will also present statistical methods for obtaining unbiased estimates of average heterozygosity and genetic distance.

The first problem I would like to discuss is the magnitude of systematic bias introduced by a small sample size when the ordinary method of estimating average heterozygosity and genetic distance is used. Let p_i be the frequency of the ith allele at a locus in a population and x_i be the corresponding allele frequency in a sample from the population. The population heterozygosity at this locus is $\zeta = 1 - \Sigma p_i^2$, where Σ stands for summation over all alleles. The average heterozygosity per locus (H) is defined as the mean of ζ over all structural loci in the genome. Theoretically, we assume that there are an infinite number of structural loci. We are then interested in estimating H by surveying r loci and n (diploid) individuals per locus. Thus, there are two sampling processes involved, *i.e.*, sampling of loci from the genome and sampling of genes ($2n$ genes) from the population at each locus. We assume that each of these samplings is conducted at random. Usually, H is estimated by a sample average heterozygosity, $\hat{H}_1$, which is the average of $1 - \Sigma x_i^2$ over the r loci studied. Under the assumption of multinomial sampling of genes, the expectation of Σx_i^2 for a particular locus is given by $\Sigma p_i^2 + (1 - \Sigma p_i^2)/2n$ (*e.g.*, CROW and KIMURA 1970). Therefore, the expectation of $\hat{H}_1$ is

$$\begin{aligned} E_g E_s(\hat{H}_1) &= E_g[\zeta - \zeta/2n] \\ &= H - H/2n \ , \end{aligned} \tag{1}$$

where E_g and E_s are the expectation operators with respect to the distribution of ζ among loci and the multinomial distribution of x_i, respectively.

For a single locus, an unbiased estimate of ζ is given by

$$h = 2n(1 - \Sigma x_i^2)/(2n - 1) \ , \tag{2}$$

whereas the corresponding unbiased estimate of H is

$$\hat{H} = \sum_{k=1}^{r} h_k/r \ , \tag{3}$$

where h_k is the value of h for the kth locus. Here n may vary from locus to locus. The estimate (3) generally has a larger expected squared deviation from H than $\hat{H}_1$ (NEI and ROYCHOUDHURY 1974; MITRA 1976), but if a few individuals are studied for a large number of loci, the systematic bias in (1) seems to be much more serious. NEI and ROYCHOUDHURY (1974) were aware of this bias, but did not particularly recommend formula (3), since the sample size employed at that time was generally large.

The ordinary estimate of genetic distance also has a systematic bias. Let p_i and q_i be the frequencies of the ith allele in populations X and Y, respectively, and x_i and y_i be the corresponding sample allele frequencies. NEI's (1972) genetic (standard) distance is defined as

$$D = -\ln[G_{XY}/\sqrt{G_X G_Y}] \ , \tag{4}$$

where G_X, G_Y, and G_{XY} are the means of Σp_i^2, Σq_i^2, and $\Sigma p_i q_i$ over all loci in the genome, respectively. The usual method of estimating D is to replace population gene identities, G_X, G_Y, and G_{XY}, by sample gene identities, J_X, J_Y, and J_{XY}, which are the averages of Σx_i^2, Σy_i^2, and $\Sigma x_i y_i$ over the r loci studied, respectively. Namely, it is estimated by $\hat{D}_1 = -\ln[J_{XY}/\sqrt{J_X J_Y}]$. When r is sufficiently large, the expectation of $\hat{D}_1$ is given by

$$E_g E_s(\hat{D}_1) \approx -\ln[E_s(J_{XY})/\sqrt{E_s(J_X)E_s(J_Y)}] \qquad \text{(Li and Nei 1975)}$$

$$= -\ln \frac{G_{XY}}{\sqrt{[G_X + (1-G_X)/2n_X][G_Y + (1-G_Y)/2n_Y]}}$$

$$\approx D + \frac{1-G_X}{4n_X G_X} + \frac{1-G_Y}{4n_Y G_Y}\ , \qquad (5)$$

where n_X and n_Y are the numbers of individuals sampled from population X and Y, respectively, and $(1-G_X)/(2n_XG_X)$ and $(1-G_Y)/(2n_YG_Y)$ are assumed to be small compared with unity, which is true in almost all cases. Here $E_s(\hat{D}_1)$ is the operator of taking the expectation of D_1 for r (given) loci with respect to the multinomial samplings of genes, whereas E_g refers to taking the expectation of $E_s(\hat{D}_1)$ with respect to sampling of r loci from the genome. Since average heterozygosity $(H \equiv 1-G)$ is generally 0.2 or less, the bias introduced by a small sample size in $\hat{D}_1$ is of the same order of magnitude as that for $\hat{H}_1$. However, $\hat{D}_1$ tends to give an overestimate of D, rather than an underestimate. It is noted that when $D = 0$, $G_X = G_Y = G$, and $n_X = n_Y = n$, $E(\hat{D}_1)$ is approximately $(1-G)/(2nG)$. Namely, even if the two populations are genetically identical with each other, the sample genetic distance can be larger than 0 when the sample size is small. Nei (1973) has called this *spurious distance.*

In many lizard species, the average heterozygosity is of the order of 0.06 (Gorman and Renzi, unpublished). Therefore, the expected magnitude of the bias when a single individual is sampled from each of the two species to be compared is about 0.03. This magnitude of bias is not important if D is large, say more than 0.15, but becomes serious when D is very small. On the other hand, if n_X and n_Y are 100, the expected bias is about 0.0003, which is generally negligible.

An unbiased estimate of D may be obtained by substituting the unbiased estimates of G_X and G_Y for J_X and J_Y. Namely,

$$\hat{D} = -\ln[\hat{G}_{XY}/\sqrt{\hat{G}_X\hat{G}_Y}]\ , \qquad (6)$$

where $\hat{G}_X$ and $\hat{G}_Y$ are the averages of $(2n_XJ_X - 1)/(2n_X - 1)$ and $(2n_YJ_Y - 1)/(2n_Y - 1)$ over the r loci studied, respectively, and $\hat{G}_{XY} = J_{XY}$. It is noted that, unlike $\hat{D}_1$, $\hat{D}$ can be negative, though its absolute value should not be large. This negative value is caused by sampling error and will occur only very rarely if n_X and n_Y are large. A negative value of $\hat{D}$ creates a problem in constructing a dendrogram. I suggest that all negative values of $\hat{D}$ should be replaced by 0 in this case.

Let us now consider the sampling variance of the unbiased estimate of average heterozygosity. It should be noted that this variance consists of two components, *i.e.*, interlocus variance and intralocus variance (Nei and Roy-

CHOUDHURY 1974). The former arises because of the fact that population heterozygosity varies greatly from locus to locus. This is caused by the evolutionary forces such as mutation, selection, and random genetic drift. The intralocus variance is generated primarily by the process of sampling a finite number of genes from the population. The underlying statistical model for the decomposition of the total sampling variance is as follows: For the kth locus, the observed heterozygosity (the unbiased estimate: $h_k = 2n(1 - \Sigma x_i^2)/(2n - 1)$) may be written as

$$h_k = \zeta_k + s_k \ , \tag{7}$$

where ζ_k is the population heterozygosity $(1 - \Sigma p_i^2)$, and s_k is the sampling error with mean $= 0$ and variance $V_s(h_k)$. Therefore, the variance, $V(h)$, of h_k over all loci (the entire genome) is

$$V(h) = V_\zeta(h) + V_s(h) \ , \tag{8}$$

where $V_\zeta(h)$ is the variance of ζ_k and $V_s(h)$ is the expectation of $V_s(h_k)$ over all loci. Here we have assumed that there are linkage equilibria among different loci and genes are sampled independently at each locus. Note that the variance components in (8) are slightly different from those of NEI and ROYCHOUDHURY (1974), since they considered the sample heterozygosity, $1 - \Sigma x_i^2$. If we note that the unbiased estimate $(\hat{H})$ of H is a simple average of heterozygosities for all individual loci, its variance is given by

$$V(\hat{H}) = V(h)/r \ . \tag{9}$$

To evaluate the effect of the number of individuals on the accuracy of the estimate of average heterozygosity, we have to know the relative magnitudes of $V_\zeta(h)$ and $V_s(h)$ in (8). To get a rough idea, I consider an equilibrium population in which the effects of mutation and random genetic drift are balanced with the same mutation rate for all loci, assuming no selection. If we use the infinite-allele model, the interlocus variance is given by $V_\zeta(h) = 2M/(M + 1)^2(M + 2)(M + 3)$, where $M = 4Nv$, in which N and v are the effective population size and the mutation rate per locus per generation, respectively (WATTERSON 1974; STEWART 1976; LI and NEI 1975). In practice, the value of $V_\zeta(h)$ seems to be slightly larger than that given by the above formula presumably because of interlocus variation in mutation rate and some other effects (NEI *et al.* 1976), but for our purpose it does not matter. (The stepwise mutation model gives a smaller interlocus variance than the infinite-allele model.) The intralocus variance of the unbiased estimate of heterozygosity for a locus may be obtained by modifying NEI and ROYCHOUDHURY's (1974) formula for the variance of $1 - \Sigma x^2$. It becomes

$$V_s(h_k) = \frac{1}{n(2n-1)}\left[\Sigma p_i^2 - (\Sigma p_i^2)^2 + 4(n-1)\{\Sigma p_i^3 - (\Sigma p_i^2)^2\}\right] \ . \quad (10)$$

The expectation of this variance over all loci can be obtained by evaluating the expectations of Σp_i^2, $(\Sigma p_i^2)^2$, and Σp_i^3 with respect to the allele frequency distributions. These expectations have been evaluated by LI and NEI (1975). Using their results, we have

$$V_s(h) = \frac{2M(M+4) + 8(n-1)M}{2n(2n-1)(M+1)(M+2)(M+3)} \ . \quad (11)$$

The values of $V_\xi(h)$ and $V_s(h)$ for various values of M and n are given in Table 1. It is clear that for $n = 1$, $V_s(h)$ is larger than $V_\xi(h)$ but $V_s(h)$ rapidly decreases as n increases. With $n = 10$, $V_s(h)$ is nearly one-tenth of $V_\xi(h)$. This clearly indicates that in order to reduce the sampling error of average heterozygosity we must examine a large number of loci rather than a large number of individuals per locus. Of course, if one wants to study not only the average heterozygosity but also the allele frequency distribution for each locus, he must examine a large number of individuals.

However, some warning against using an extremely small number of individuals should be mentioned. The above argument assumes that a large number of loci are available for study. In practice, technical difficulties often limit the number of loci studied. In fact, less than 30 loci were studied in most recent protein surveys. This number is small; ideally, more than 50 loci should be used to obtain a reliable estimate of average heterozygosity for the total genome. If this cannot be done technically, a large number of individuals studied per locus still helps to reduce the standard error of average heterozygosity. In Table 1, for example, if H is 0.167, the expected intralocus variance ($V_s(h)$) is 0.09943 for $n = 1$. Thus, if one individual is examined for 25 loci, the expected standard

TABLE 1

Effects of sample size (n = *number of individuals*) *on the intralocus variance* [V_s(h)] *of heterozygosity*

			$V_s(h)$				
M	H	$V_\xi(h)$	$n = 1$	$n = 2$	$n = 10$	$n = 20$	$n = 50$
0.02	0.020	0.00630	0.01292	0.00430	0.00068	0.00033	0.00013
0.06	0.057	0.01694	0.03646	0.01206	0.00189	0.00092	0.00036
0.1	0.091	0.02539	0.05725	0.01885	0.00295	0.00143	0.00056
0.2	0.167	0.03946	0.09943	0.03235	0.00501	0.00243	0.00095
0.4	0.286	0.05002	0.15406	0.04902	0.00744	0.00361	0.00142

$M = 4Nv$. $H \equiv M/(1+M)$ = the expected heterozygosity. $V_\xi(h)$ = interlocus variance of heterozygosity.

error of average heterozygosity estimate becomes $(0.009943/25)^{1/2} = 0.06$, neglecting the effect of interlocus variation. This is more than a third of H. On the other hand, if 50 individuals are studied for 25 loci, it becomes 0.0062, which is 1/27 of H.

A similar study can be made about the effect of the number of individuals on the estimate of genetic distance. For this purpose, however, it is simpler to work with the minimum distance rather than the standard distance (see NEI and ROYCHOUDHURY 1974). The minimum distance for the kth locus is defined as $\zeta_k = (\Sigma p_i^2 + \Sigma q_i^2)/2 - \Sigma p_i q_i$, and the distance for all loci (D_m) is the arithmetic mean of this quantity. An unbiased estimate of single locus genetic distance is given by

$$d_k = \frac{2n_X \Sigma x_i^2 - 1}{2(2n_X - 1)} + \frac{2n_Y \Sigma y_i^2 - 1}{2(2n_Y - 1)} - \Sigma x_i y_i \ , \tag{12}$$

whereas the unbiased estimate of D_m is given by

$$\hat{D}_m = \sum_{k=1}^{r} d_k / r \ . \tag{13}$$

As with h_k, d_k may be written as $d_k = \zeta_k + s_k$, where s_k is the sampling error with mean $= 0$ and variance $V_s(d_k)$. Again modifying NEI and ROYCHOUDHURY's (1974) formula, the intralocus variance, $V_s(d_k)$, becomes

$$\begin{aligned} V_s(d_k) &= \{V_s(j_X) + V_s(j_Y)\}/4 + V_s(j_{XY}) \\ &\quad - \mathrm{Cov}_s(j_X, j_{XY}) - \mathrm{Cov}_s(j_Y, j_{XY}) \ , \end{aligned} \tag{14}$$

where $V_s(j_X) = V_s(h_k)$ for population X, $V_s(j_Y) = V_s(h_k)$ for population Y, and

$$\begin{aligned} V_s(j_{XY}) &= \{(1 - 2n_X - 2n_Y)(\Sigma p_i q_i)^2 + (2n_X - 1)\Sigma p_i^2 q_i + (2n_Y - 1)\Sigma p_i q_i^2 \\ &\quad + \Sigma p_i q_i\}/(4n_X n_Y) \ , \\ \mathrm{Cov}_s(j_X, j_{XY}) &= 2\{\Sigma p_i^2 q_i - (\Sigma p_i^2)(\Sigma p_i q_i)\}/(2n_X) \ , \\ \mathrm{Cov}_s(j_Y, j_{XY}) &= 2\{\Sigma p_i q_i^2 - (\Sigma q_i^2)(\Sigma p_i q_i)\}/(2n_Y) \ . \end{aligned}$$

The variance of d_k over all loci is

$$V(d) = V_\zeta(d) + V_s(d) \ , \tag{15}$$

where $V_\zeta(d)$ and $V_s(d)$ are the variance of ζ_k and the mean of $V_s(d_k)$ over loci, respectively. Evaluation of the exact value of $V_\zeta(d)$ is complicated, but it can be shown that it increases with increase of the mean distance, $\bar{\zeta}_k \equiv D_m$ (LI and NEI 1975). If the mutation-drift balance is maintained in each of the two populations throughout the evolutionary process with $4Nv = 0.1$, then $V_\zeta(d)$ is 0.00410 for $D_m = 0.018$ and 0.11156 for $D_m = 0.168$. On the other hand, $V_s(d)$ is of the same order of magnitude as $V_s(h)$ when D_m is small but decreases slowly as D_m

increases. (The property of $V_s(d)$ is virtually the same as that of the intralocus variance of *sample minimum distance*, which was studied by NEI and ROYCHOUDHURY 1974). Therefore, it is clear that if D_m is as large as 0.168 and a large number of loci are examined, the number of individuals per locus can be very small. On the other hand, if D_m is as small as 0.018, a considerable number of individuals must be examined. Needless to say, the variance of $\hat{D}_m$ is given by $V(d)/r$.

The sampling variance of the unbiased estimate of standard genetic distance ($\hat{D}$) and its components can be obtained again by modifying NEI and ROYCHOUDHURY's (1974) formulae. That is, if we replace J_X, J_Y, and J_{XY} in their formulae (22) and (23) by $\hat{G}_X$, $\hat{G}_Y$, and $\hat{G}_{XY}$, respectively, they are immediately obtained. However, I shall not present the results here, since they are too complicated. (They are incorporated into our new computer program.) On the other hand, the relative values of the components corresponding to $V_\zeta(d)$ and $V_s(d)$ in (15) can be evaluated by LI and NEI's (1975) method. The results obtained are virtually the same as those for $\hat{D}_1$, and thus support GORMAN and RENZI's (unpublished) empirical finding. It should be noted, however, that the number of individuals to be examined depends also on the level of heterozygosity (Table 1). More individuals should be examined when heterozygosity is high than when it is low.

When a dendrogram for a group of species is constructed from genetic distance estimates, the reliability of the topology of the dendrogram depends on the differences in genetic distance among different pairs of species. If these differences are small, the genetic distances must be estimated accurately. Namely, a considerable number of individuals should be examined for each locus. On the other hand, if the differences are large, even a single individual may be sufficient for obtaining the correct topology of a dendrogram. In fact, this is exactly what GORMAN and RENZI (unpublished) observed with the *Anolis roquet* and *A. bimaculatus* group species. Another factor that affects the dendrogram is the level of heterozygosity. As discussed above, the standard error of genetic distance is large when average heterozygosity is high. Thus, in organisms with average heterozygosity higher than 0.1 a relatively large number of individuals should be examined to construct a reliable dendrogram.

Our formulae for obtaining unbiased estimates of average heterozygosity and genetic distance apply to any sample size and are superior to sample average heterozygosity and genetic distance, as long as many loci are used. However, the difference between the biased and unbiased estimators is very small when the number of individuals used is large, say more than 50. A computer program for obtaining the unbiased estimates of average heterozygosity and (standard) genetic distance and their standard errors is available by writing to the author.

I would like to thank GEORGE C. GORMAN for showing me his unpublished manuscript. This work was supported by grants from the National Science Foundation and the Public Health Service.

LITERATURE CITED

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

Li, W. H. and M. Nei, 1975 Drift variances of heterozygosity and genetic distance in transient states. Genet. Res. **25**: 229–248.

Mitra, S., 1976 More on Nei and Roychoudhury's sampling variances of heterozygosity and genetic distance. Genetics **82**: 543–545.

Nei, M., 1972 Genetic distance between populations. American Naturalist **106**: 283–292.

Nei, M., 1973 The theory and estimation of genetic distance. pp. 45–54. In: *Genetic Structure of Populations.* Edited by N. E. Morton, University Hawaii Press, Honolulu.

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

Nei, M., P. A. Fuerst and R. Chakraborty, 1976 Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. Nature **262**: 491–493.

Stewart, F. M., 1976 Variability in the amount of heterozygosity maintained by neutral mutations. Theoret. Popul. Biol. **9**: 188–201.

Watterson, G. A., 1974 Models for the logarithmic species abundance distributions. Theor. Pop. Biol. **6**: 217–250.

Corresponding editor: B. S. Weir