SAMPLING VARIANCES OF HETEROZYGOSITY AND GENETIC DISTANCE[1]

MASATOSHI NEI AND A. K. ROYCHOUDHURY

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

Manuscript received April 10, 1973

Mathematical formulae for the sampling variances of average heterozygosity and NEI's genetic distance are developed. These sampling variances are decomposed into their two components, i.e. the inter-locus and intra-locus variances. The relationship between the number of loci and the number of individuals per locus to be examined for estimating average heterozygosity and genetic distance is also discussed. The utility of the inter-locus variance of heterozygosity for studying the mechanism of maintenance of genetic variability in populations is indicated.

THE genetic variability of a population is usually measured by the average heterozygosity per locus, while the gene differences between two populations may be measured by the genetic distance recently proposed by NEI (1971, 1972). The main purpose of this paper is to investigate the sampling variances of these quantities.

We first present a general formula for each of these sampling variances and then study the components of the variance. As will be seen later, the sampling variance is made up of two components, i.e., the inter-locus and intra-locus variances. The inter-locus variance depends on the genetic structure of populations, which is determined by all sorts of evolutionary forces, such as mutation, selection and random genetic drift. It is, therefore, difficult to quantify its magnitude except for some special cases. On the other hand, the intra-locus variance solely depends on the sample size and gene frequencies of the locus studied, so that the mathematical formulation is easier. Knowledge of the intra-locus variance is required to compute the standard errors of heterozygosity and genetic distance for a particular locus or to estimate the magnitude of inter-locus variance.

On the basis of the formulae for the sampling variances of heterozygosity and genetic distance, we shall also discuss the number of loci and the number of individuals to be studied for estimating these quantities.

VARIANCE OF HETEROZYGOSITY

Consider a random mating population in which l alleles $(A_1, A_2, \ldots, A_l)$ are segregating at a locus. Let p_i be the frequency of the i-th allele in the population ($\Sigma p_i = 1$, where the summation is over i). We assume that genotype frequencies are in Hardy-Weinberg proportions. Thus, the population homozygosity and heterozygosity are given by $g = \Sigma p_i^2$ and $1 - g$, respectively. We further assume

[1] This study was supported in part by Public Health Service Grant GM 20293.

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

that there is no dominance and all heterozygotes are identifiable, as is usually the case with isozyme variations. (See DISCUSSION for the effect of dominance.) Then, if n genes ($n/2$ individuals in diploid organisms) are sampled at random from the population, the probability that there are n_1A_1 genes, n_2A_2 genes, . . , and n_lA_l genes in the sample is given by

$$P(n_1, n_2, \ldots, n_l) = \frac{n!}{n_1!n_2!\ldots n_l!} p_1^{n_1} p_2^{n_2} \ldots p_l^{n_l} \tag{1}$$

where $\Sigma n_i = n$ and the population size is assumed to be much larger than the sample size.

Let $x_i (= n_i/n)$ be the sample gene frequency of the i-th allele. Clearly, this is the maximum likelihood estimate of p_i. The sample homozygosity and heterozygosity are given by $j_X = \Sigma x_i^2$ and $h = 1 - j_X$, respectively. The expectation of sample homozygosity $\{E(\Sigma x_i^2)\}$ is $\Sigma p_i^2 + (1 - \Sigma p_i^2)/n$ or $g + (1 - g)/n$, since $E(x_i^2) = (1 - 1/n)p_i^2 + p_i/n$. Therefore, j_X is not an unbiased estimate of g, though it is *asymptotically* unbiased when n is large. The unbiased estimate is given by

$$\hat{g} = (n\Sigma x_i^2 - 1)/(n - 1). \tag{2}$$

This is identical to the formula given by MORTON *et al.* (1971). In most population surveys, however, n is fairly large, so that the bias of j_x is generally negligible. Furthermore, as is shown in the APPENDIX, the expectation of squared deviation of j_X from g, i.e., $E(j_X - g)^2$, is smaller than that of $\hat{g}$, $E(\hat{g} - g)^2$, when n is small, although j_x is statistically biased. For these reasons, we shall use j_x and h as the estimates of homozygosity and heterozygosity at a locus, respectively. Clearly, the variance of j_X is equal to that of h.

The average heterozygosity (H) of a population is defined as the average of h over all loci. In practice, only a limited number of loci are examined for polymorphism. If r loci are examined, the average heterozygosity is estimated by

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

where h_k is the estimate of heterozygosity of the k-th locus. The sampling variance of $\hat{H}$ may be obtained by

$$V(\hat{H}) = V(h)/r \tag{4}$$

in which $V(h)$ is the expected variance of h and estimated by

$$V(h) = \sum_{k=1}^{r} (h_k - \hat{H})^2/(r - 1). \tag{5}$$

In deriving (4) we assumed that h's at different loci are not correlated. This assumption seems to be satisfactory, unless there are linkage disequilibria.

Our empirical data for protein and blood group loci in man (NEI and ROYCHOUDHURY, unpublished) indicate that the distribution of h is inverse J-shaped. (This is theoretically expected.) Therefore, the exact test of significance for $\hat{H}$ is difficult. However, if r is large, the distribution of $\hat{H}$ will be approximately normal

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

because of the central limit theorem. Thus, an ordinary statistical test may be made by using the variance obtained by (4).

Let us now consider the following statistical model

$$h_k = \gamma_k + s_k \tag{6}$$

where h_k is the observed heterozygosity at the k-th locus, $\gamma_k = (1 - g_k)(1 - 1/n)$, in which $1 - g_k$ is the population heterozygosity at this locus, and s_k is a random variable with mean $= 0$ and variance $V_s(h_k)$. Clearly, the expectation of γ_k over all loci is $\bar{\gamma} = H(1 - 1/n)$, while that of s_k is 0, in which n is assumed to be the same for all loci. Therefore, the variance $(V(h))$ of h_k over all loci is

$$V(h) = V_\gamma(h) + V_s(h) \tag{7}$$

where $V_\gamma(h)$ is the variance of γ_k, and $V_s(h)$ is the expectation of s_k^2 and estimated by

$$V_s(h) = \sum_{k=1}^{r} V_s(h_k)/r \tag{8}$$

if there are r loci studied. We call $V_\gamma(h)$ and $V_s(h)$ as the inter-locus and intra-locus variances of h, respectively. Since $\gamma_k = (1 - g_k)(1 - 1/n)$, $V_\gamma(h)$ may be written as

$$V_\gamma = (n-1)^2 V_g(h)/n^2 \tag{9}$$

where $V_g(h)$ is the variance of $1 - g$ (and also of g) among loci.

We now derive the formula for the intra-locus variance of h. Mathematically it is convenient to study the intra-locus variance $(V_s(j_X))$ of j_X rather than h itself. Clearly, at a particular locus,

$$\begin{aligned} V_s(j_X) &= E\{(\Sigma x_i^2)^2\} - \{E(\Sigma x_i^2)\}^2 \\ &= \sum_i E(x_i^4) + \sum_{i \neq j} E(x_i^2 x_j^2) - \{\sum_i E(x_i^2)\}^2 \end{aligned} \tag{10}$$

Using (1), we can determine the following moments:

$$E(x_i) = p_i$$

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

$$E(x_i^3) = \{(n-1)(n-2)p_i^3 + 3(n-1)p_i^2 + p_i\}/n^2$$

$$E(x_i^4) = \{(n-1)(n-2)(n-3)p_i^4 + 6(n-1)(n-2)p_i^3 + 7(n-1)p_i^2 + p_i\}/n^3$$

$$E(x_i x_j) = (n-1)p_i p_j/n \qquad i \neq j$$

$$E(x_i^2 x_j) = \{(n-1)(n-2)p_i^2 p_j + (n-1)p_i p_j\}/n^2 \qquad i \neq j$$

$$E(x_i^2 x_j^2) = \{(n-1)(n-2)(n-3)p_i^2 p_j^2 + (n-1)(n-2)p_i^2 p_j + (n-1)(n-2)p_i p_j^2 + (n-1)p_i p_j\}/n^3 \qquad i \neq j$$

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

Substituting these moments into (10), we have

$$V_s(j_X) = \frac{2(n-1)}{n^3}\{(3-2n)(\Sigma p_i^2)^2 + 2(n-2)\Sigma p_i^3 + \Sigma p_i^2\} \tag{11}$$

In practice, the population parameters Σp_i^2 and Σp_i^3 are not known, so that they must be estimated by sample moments. The intra-locus variance of j_X is then estimated by

$$\frac{2(n-1)}{n^3}\{(3-2n)j_X^2 + 2(n-2)\Sigma x_i^3 + j_X\} \ . \tag{12}$$

Note that $V_s(j_X)$ is 0 if the locus is monomorphic, as it should be.

Since $V_s(h) = V_s(j_X)$, the above formula may be used for obtaining the intra-locus variance of heterozygosity at a single locus. If this is computed for all the loci studied, the average intra-locus variance may be obtained by (8). Then, it is possible to estimate $V_\gamma(h)$ or $V_g(h)$.

VARIANCE OF GENETIC DISTANCE

In recent years, several authors have proposed different measures of genetic distance between populations (SANGHVI 1953; CAVALLI-SFORZA and EDWARDS 1967; BALAKRISHNAN and SANGHVI 1968; ROGERS 1972; and others). Most of these measures are, however, constructed from the statistical point of view, and it is not clear what biological unit they are going to measure. In contrast to these measures, the genetic distance proposed by NEI (1971, 1972) is intended to estimate the number of net codon differences per locus between populations. He has devised three different estimates of this number, i.e., the minimum (D_m), standard (D) and maximum (D') distances. For the biological meanings of these estimates or distances, the reader may refer to NEI (1972, 1973a,b) and NEI and ROYCHOUDHURY (1972). They are all based on the identities of genes within and between populations.

Let p_i and q_i be the frequencies of the *i-th* allele at a locus in populations X and Y, respectively. The identities of two randomly chosen genes in X and Y are then $g_X = \Sigma p_i^2$ and $g_Y = \Sigma q_i^2$, respectively. The identity of two genes, chosen at random one from X and one from Y, is $g_{XY} = \Sigma p_i q_i$. The three distance measures are then defined as:

$$\text{Minimum:} \quad D_m = (G_X + G_Y)/2 - G_{XY}, \tag{13a}$$

$$\text{Standard:} \quad D = -\log_e(G_{XY}/\sqrt{G_X G_Y}), \tag{13b}$$

$$\text{Maximum:} \quad D' = -\log_e(G'_{XY}/\sqrt{G'_X G'_Y}), \tag{13c}$$

where G_X, G_Y and G_{XY} are the arithmetic means of g_X, g_Y and g_{XY}, respectively, over all loci, including monomorphic ones, while G_X', G_Y' and G_{XY}' are the geometric means.

In practice, of course, all the three distance measures are estimated by using the sample gene frequencies instead of population gene frequencies. In the following we denote by x_i and y_i the sample gene frequencies of the i-th allele in popu-

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

lations X and Y, respectively. Thus, g_X, g_Y and g_{XY} are estimated by $j_X = \Sigma x_i^2$, $j_Y = \Sigma y_i^2$ and $j_{XY} = \Sigma x_i y_i$, respectively. We know that j_X and j_Y are not unbiased estimates, but for the reasons mentioned earlier we use these estimates. On the other hand, j_{XY} is an unbiased estimate of g_{XY}.

The three distance measures mentioned above are then estimated by

$$\text{Minimum:} \quad \hat{D}_m = (J_X + J_Y)/2 - J_{XY}, \tag{14a}$$

$$\text{Standard:} \quad \hat{D} = -\log_e(J_{XY}/\sqrt{J_X J_Y}), \tag{14b}$$

$$\text{Maximum:} \quad \hat{D}' = -\log_e(J'_{XY}/\sqrt{J'_X J'_Y}), \tag{14c}$$

where J_X, J_Y and J_{XY} are the arithmetic means of j_X, j_Y and j_{XY}, respectively, over all loci including monomorphic ones, while J'_X, J'_Y and J'_{XY} are the geometric means. Obviously, J'_{XY} is 0 if one of j'_{XY}'s is 0; then the maximum estimate is meaningless. Actually, $\hat{D}'$ always tends to be an overestimate, and it is safe not to use this estimate if anyone of j_{XY}'s is small compared with unity (NEI 1972). However, when local races of a species are compared, there is generally not much difference between $\hat{D}_m$, $\hat{D}$ and $\hat{D}'$. Note also that $\hat{D}$ and $\hat{D}'$ are not unbiased estimates even if the unbiased estimates of g_X and g_Y are used instead of j_X and j_Y in the formulae. Nevertheless, they are asymptotically unbiased and mathematically simple. For these reasons we prefer these estimates.

The estimate of minimum genetic distance may be written as

$$\hat{D}_m = \sum_{k=1}^{r} d_k/r, \tag{15}$$

where d_k is the value of $d = (j_X + j_Y)/2 - j_{XY}$ for the k-th locus and r is the number of loci examined. Note also that d may be written as $\Sigma(x_i - y_i)^2/2$, where i denotes the i-th allele. The sampling variance of $\hat{D}_m$ is then computed by

$$V(\hat{D}_m) = V(d)/r \tag{16}$$

in which

$$V(d) = \sum_{k=1}^{r} (d - \hat{D}_m)^2/(r-1).$$

This variance may be used for a statistical test of the significance of $\hat{D}_m$.

In analogy to h_k, let d_k be

$$d_k = \gamma_k + s_k \tag{17}$$

where $\gamma_k = (g_X + g_Y)/2 - g_{XY} + (1 - g_X)/(2n_X) + (1 - g_Y)/(2n_Y)$ for the k-th locus, in which n_X and n_Y are the sample sizes for populations X and Y, respectively, and s_k is a random variable with mean $= 0$ and variance $V_s(d)$. Since γ_k and s_k are not correlated with each other, we have

$$V(d) = V_\gamma(d) + V_s(d), \tag{18}$$

where $V_\gamma(d)$ and $V_s(d)$ are the expected variances of γ_k and s_k over all loci, respectively.

We now determine the intra-locus variance of d, i.e. $V_s(d)$, considering one locus. Clearly, $V_s(d)$ is

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

$$V_s(d) = \{V_s(j_X) + V_s(j_Y)\}/4 + V_s(j_{XY}) - \mathrm{Cov}_s(j_X,j_{XY}) - \mathrm{Cov}_s(j_Y,j_{XY}) \quad (19)$$

where V_s and Cov_s denote the intra-locus variance and covariance, respectively. Note that the sampling covariance between j_X and j_Y is 0. $V_s(j_X)$ and $V_s(j_Y)$ in the above expression are obtained by (12). $V_s(j_{XY})$ and $\mathrm{Cov}_s(j_X,j_{XY})$ may be written as

$$V_s(j_{XY}) = E\{(\sum_i x_i y_i)^2\} - \{E(\sum_i x_i y_i)\}^2$$
$$= E\{\sum_i x_i^2 y_i^2 + \sum_{i \neq j} x_i y_i x_j y_j\} - \{E(\sum_i x_i y_i)\}^2 .$$
$$\mathrm{Cov}_s(j_X,j_{XY}) = E\{\sum_i x_i^3 y_i + \sum_{i \neq j} x_i^2 x_j y_j\} - E(\sum_i x_i^2) E(\Sigma x_i y_i).$$

Since the genes in populations X and Y are sampled independently, we have

$$E(x_i y_i) = p_i q_i$$
$$E(x_i^2 y_i^2) = p_i q_i\{(n_X - 1)(n_Y - 1)p_i q_i + (n_X - 1)p_i + (n_Y - 1)q_i + 1\}/(n_X n_Y)$$
$$E(x_i^3 y_i) = p_i q_i\{(n_X - 1)(n_X - 2)p_i^2 + 3(n_X - 1)p_i + 1\}/n_X^2$$
$$E(x_i^2 x_j y_j) = (n_X - 1)p_i p_j q_j\{(n_X - 2)p_i + 1\}/n_X^2$$
$$E(x_i x_j y_i y_j) = (n_X - 1)(n_Y - 1)p_i p_j q_i q_j/(n_X n_Y)$$

where $i \neq j$. Therefore,

$$V_s(j_{XY}) = \{(1 - n_X - n_Y)(\Sigma p_i q_i)^2 + (n_X - 1)\Sigma p_i^2 q_i + (n_Y - 1)\Sigma p_i q_i^2 + \Sigma p_i q_i\}/(n_X n_Y)$$
$$\mathrm{Cov}_s(j_X,j_{XY}) = 2(n_X - 1)\{\Sigma p_i^2 q_i - (\Sigma q_i^2)(\Sigma p_i q_i)\}/n_X^2 .$$

Similarly,

$$\mathrm{Cov}_s(j_Y,j_{XY}) = 2(n_Y - 1)\{\Sigma p_i q_i^2 - (\Sigma q_i^2)(\Sigma p_i q_i)\}/n_Y^2 .$$

Putting these into (19), the intra-locus variance of d is obtained. In practice, of course, the population moments Σp_i^2, Σp_i^3, $\Sigma p_i q_i$, etc. must be replaced by the sample moments Σx_i^2, Σx_i^3, $\Sigma x_i y_i$, etc., respectively. The intra-locus variance of d estimated from r loci is then given by

$$V_s(d) = \sum_{k=1}^{r} V_s(d_k)/r \quad (20)$$

where $V_s(d_k)$ is the variance of d at the k-th locus.

It is not easy to get the exact sampling variances of $\hat{D}$ and $\hat{D}'$, but the asymptotic variances when sample size is large are easily obtained. Namely, the asymptotic variance of $\hat{D}$ is

$$V(\hat{D}) = \left(\frac{\partial \hat{D}}{\partial J_X}\right)^2 V(J_X) + \left(\frac{\partial \hat{D}}{\partial J_Y}\right)^2 V(J_Y) + \left(\frac{\partial \hat{D}}{\partial J_{XY}}\right)^2 V(J_{XY})$$
$$+ 2\left(\frac{\partial \hat{D}}{\partial J_X}\frac{\partial \hat{D}}{\partial J_Y}\right)\mathrm{Cov}(J_X,J_Y) + 2\left(\frac{\partial \hat{D}}{\partial J_X}\frac{\partial \hat{D}}{\partial J_{XY}}\right)\mathrm{Cov}(J_X,J_{XY})$$
$$+ 2\left(\frac{\partial \hat{D}}{\partial J_Y}\frac{\partial \hat{D}}{\partial J_{XY}}\right)\mathrm{Cov}(J_Y,J_{XY}), \quad (21)$$

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

approximately. Note that $\mathrm{Cov}(J_X,J_Y)$ is not 0 in this case, since g's in populations X and Y are generally correlated. Since J_X, J_Y and J_{XY} are the arithmetic means of j_X, j_Y and j_{XY}, respectively, $V(J_X)$, $V(J_Y)$, $\mathrm{Cov}(J_X,J_Y)$, etc, are easily obtained by the observed values of j_X's, j_Y's and j_{XY}'s. For example, $V(J_X)$ is the same as $V(\hat{H})$ in (4), and

$$\mathrm{Cov}(J_X,J_Y) = \sum_{k=1}^{r} (j_{X(k)} - J_X)(j_{Y(k)} - J_Y)/\{r(r-1)\}.$$

Therefore,

$$V(\hat{D}) = \frac{\mathrm{V}(J_X)}{4J_X^2} + \frac{\mathrm{V}(J_Y)}{4J_X^2} + \frac{\mathrm{V}(J_{XY})}{J_{XY}^2} + \frac{\mathrm{Cov}(J_X,J_Y)}{2J_XJ_Y} - \frac{\mathrm{Cov}(J_X,J_{XY})}{J_XJ_{XY}} - \frac{\mathrm{Cov}(J_Y,J_{XY})}{J_YJ_{XY}}. \tag{22}$$

Again, this may be used for the significance test of $\hat{D}$.

The intra-locus variance $(V_s(\hat{D}))$ of $\hat{D}$ is also obtained in the same way. Since the intra-locus variances $(V_s(J_X), V_s(J_Y)$, etc.) of J_X, J_Y, etc. are $\Sigma V_s(j_X)/r^2$, $\Sigma V_s(j_Y)/r^2$, etc., respectively, where the summation is over different loci, we have

$$V_s(\hat{D}) = \left\{\frac{\Sigma V_s(j_X)}{4J_X^2} + \frac{\Sigma V_s(j_Y)}{4J_Y^2} + \frac{\Sigma V_s(j_{XY})}{J_{XY}^2} - \frac{\Sigma \mathrm{Cov}_s(j_X,j_{XY})}{J_XJ_{XY}} - \frac{\Sigma \mathrm{Cov}_s(j_Y,j_{XY})}{J_YJ_{XY}}\right\}/r^2. \tag{23}$$

On the other hand, $\hat{D}'$ may be written as

$$\hat{D}' = \sum_{k=1}^{r} d_k'/r \tag{24}$$

where d_k' is the value of $-(\log_e j_X + \log_e j_Y)/2 + \log_e j_{XY}$ for the k-th locus. Thus, $V(\hat{D}')$ is obtained by formula (16), simply replacing d_k by d_k' and $\hat{D}_m$ by $\hat{D}'$. Similarly, the intra-locus variance $(V_s(d'))$ of d' over all loci may be estimated by (20), replacing $V_s(d_k)$ by

$$V_s(d_k') = \frac{V_s(j_X)}{4j_X^2} + \frac{V_s(j_Y)}{4j_Y^2} + \frac{V_s(j_{XY})}{j_{XY}^2} - \frac{\mathrm{Cov}_s(j_X,j_{XY})}{j_Xj_{XY}} - \frac{\mathrm{Cov}_s(j_Y,j_{XY})}{j_Yj_{XY}}$$

for the k-th locus.

As noted earlier, $\hat{D}'$ is a poor estimate when any one of j_{XY}'s is small. This is reflected in the above formula; when $j_{XY} \to 0$, $V(D')$ diverges. The values of j_X and j_Y never become 0 in practice.

Computer programs for estimating the sampling variances of $\hat{H}$, $\hat{D}_m$, $\hat{D}$ and $\hat{D}'$ and their components have been developed. They are available by writing to the authors.

DISCUSSION

In the foregoing sections we assumed that genotype frequencies are in Hardy-Weinberg proportions. If this assumption is not fulfilled, our estimate of hetero-

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

zygosity is no longer true. For example, in a selfing population there are virtually no heterozygotes at equilibrium, so that our H does not measure the proportion of heterozygous loci in an individual. However, it is a good measure of genic variation of the population. In this sense H may be called the *heterogeneity index* or *gene diversity* and used in any population. It is equal to the probability of nonidentity of two randomly chosen genes. NEI's measure of genetic distance has been defined in terms of identities of genes, so that it is affected neither by nonrandom mating nor by natural selection.

Failure of the assumption of the Hardy-Weinberg equilibrium, however, affects the sampling variances of both heterogeneity index and genetic distance. This is because the multinomial distribution (1) no longer holds. If sample size remains the same, inbreeding is expected to increase the variances of these quantities. For example, in a completely inbred population, homozygous genotypes rather than genes are sampled according to a multinomial distribution, so that sampling of N individuals in this population is equivalent to sampling of N genes ($N/2$ individuals) in a Hardy-Weinberg population. Therefore, to compute the intra-locus variances of heterogeneity index and genetic distance, n (the number of genes examined) should be replaced by the number of individuals examined. Of course, if the degree of inbreeding is small, our formulae should hold approximately.

The intra-locus variances of heterogeneity index and genetic distance are also affected by dominance. Dominance is generally expected to increase these variances, but the effect should not be large, unless the frequencies of recessive genes are very small.

In a Hardy-Weinberg population it is possible to estimate the average heterozygosity by examining the proportion of heterozygotes directly, provided that all heterozygotes are recognizable. The sampling variance of this estimate at a locus may be obtained by $\hat{h}(1-\hat{h})/N$, where $\hat{h}$ is the proportion of heterozygotes in the sample and N is the number of individuals examined. Biologically, however, this estimate is subject to several difficulties. First, if the population size is small, the genotype frequencies in the population may deviate considerably from the Hardy-Weinberg proportions due to the sampling error at the time of fertilization. Generally speaking, gene frequencies are more stable than genotype frequencies in a finite population. Second, if strong natural selection operates in a developmental stage before observation, the proportion of heterozygotes would be distorted and the observed proportion will no longer correspond to the theoretical formulation of heterozygosity as determined by mutation rate, selection and population size (KIMURA and OHTA 1971). Third, if there is inbreeding, the proportion of heterozygotes is a poor measure of genetic heterogeneity of a population. For these reasons, this method cannot be recommended for a general use.

In planning a survey on the genetic heterogeneity of a population, it is important to know how many loci and how many individuals per locus should be examined when the total number is fixed. Theoretically, this problem may be solved by minimizing the sampling variance of $\hat{H}$, i.e,

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

$$V(\hat{H}) = \{V_\gamma(h) + V_s(h)\}/r$$

with the constraint of rn equal to a constant (the total number of genes to be studied), though the sampling variance is not the sole criterion in this case. But the differentiation of $V(\hat{H})$ with respect to n yields a cubic function of n, so that the general solution is not simple. In practice, however, $V_s(h)$ is generally much smaller than $V_\gamma(h)$ unless n is extremely small. Since the $V_\gamma(h)$ does not decrease with increasing n, this indicates that n can be relatively small when average heterozygosity is to be estimated.

Let us examine this problem by using the data obtained by AVISE and SELANDER (1972). These authors studied the protein polymorphisms in three cave and nine surface populations of the characid fish *Astyanax mexicanus*. For illustration, let us use the data for populations 3 (cave) and 4 (surface). The number of protein loci examined was 17, and the number of individuals examined ($n/2$) was 45 for all protein loci in population 3 and 79 in population 4. From the gene frequency data given in their paper, we can get the estimate ($\hat{H}$) of average heterozygosity, which becomes 0.0962 in population 3 and 0.1384 in population 4. On the other hand, the estimates of $V_\gamma(h)$ and $V_s(h)$ are 0.02407767 and 0.00069046, respectively, for population 3 and 0.03061028 and 0.00072046 for population 4. Thus, in both populations the estimate of $V_\gamma(h)$ is much larger than that of $V_s(h)$. In population 4, we recomputed the estimates of $V_\gamma(h)$ and $V_s(h)$, assuming that $n/2$ was 20 rather than 79. They are 0.02856154 and 0.0027692, respectively. Therefore, even with 20 individuals per locus, the estimate of $V_\gamma(h)$ is still larger than $V_s(h)$. This indicates that, for the purpose of estimating average heterozygosity, it would have been better to examine 67 loci (or even 30 loci) and 20 individuals per locus rather than 17 loci and 79 individuals per locus. Namely, if $r = 67$ and $V_\gamma(h)$ and $V_s(h)$ remain the same, the the standard error ($\sqrt{V(\hat{H})}$) of $\hat{H}$ is expected to be 0.0216, while with the original values of $r = 17$ and $n = 79$ it is 0.0429. On the other hand, the expected amount (H/n) of bias of $\hat{H}$ is 0.09% for $n/2 = 79$ and 0.35% for $n/2 = 20$, if H is assumed to be equal to 0.1384. Thus, the bias is very small even for $n/2 = 20$.

It is noted that in many studies on average heterozygosity so far conducted, the number of loci examined is rather small, while the number of individuals per locus is large. For estimating the average heterozygosity per locus, however, it is better to examine a large number of loci rather than a large number of individuals per locus, unless $V_\gamma(h)$ is extremely small. Of course, the actual sample size depends on the purpose of the survey. For example, if one is interested in testing the Hardy-Weinberg equilibrium as well as in estimating average heterozygosity, a relatively large number of individuals should be examined for each polymorphic locus. Note also that if n is too small, the bias of the estimate $\hat{H}$ becomes large.

In the above discussion we were concerned only with average heterozygosity. But a similar argument can be made about the optimum size for measuring genetic distance between populations. Our empirical studies with human popula-

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

tions indicate that the single-locus genetic distance (d) varies considerably with locus. This suggests that a large number of loci should be used to estimate the average genetic distance per locus. In a computer simulation of the Brownian motion/Yule process evolution, KIDD and CAVALLI-SFORZA (1971) also noted the importance of measuring a large number of "characters" in reconstructing evolutionary trees.

Some special comments are, however, necessary on the genetic distance between closely related populations. Our three distance measures are all non-negative, and thus the sampling variation of gene frequencies may produce non-zero estimates of distance even if the two populations under comparison are identical. The expected magnitude of this spurious distance when the two populations are identical may be evaluated by the method given by NEI (1973b). If the observed value of genetic distance is of the same order of magnitude as the spurious distance, the distance is not significant. In such a case the hypothesis $D_m \neq 0$ may be tested more appropriately by the ordinary χ^2 method, if the sample size is sufficiently large. In the present case the χ^2 for a locus is computed by

$$\chi^2 = n_X n_Y \sum_i \frac{(x_i - y_i)^2}{x_i n_X + y_i n_Y}$$

with the number of degrees of freedom equal to the number of alleles minus one. The test of the hypothesis $D_m \neq 0$ may be made by using the sum of these χ^2's for all loci studied. If $D_m \neq 0$, clearly $D \neq 0$, and $D' \neq 0$.

In the present paper we have shown how the inter-locus variance of heterozygosity or genetic distance can be estimated. Estimates of the inter-locus variance of heterozygosity permit some inference about the mechanism of maintenance of genetic variability in populations. STEWART (unpublished) worked out the theoretical variance of population heterozygosity when neutral mutations and genetic random drift are balanced. It is given by

$$\mathrm{Var}(h) = \frac{2\theta}{(1+\theta)^2(2+\theta)(3+\theta)}$$

where $\theta = 4Nu$, in which N is the effective population size and u is the mutation rate per locus per generation. Therefore, if we know θ, $\mathrm{Var}(h)$ can be obtained. One test of the neutral mutation theory is to compare this theoretical variance with the observed value. An estimate of θ may be obtained by $\hat{H}/(1-\hat{H})$, since the expectation of $\hat{H}$ is $\theta/(1+\theta)$. EWENS (1972) proposed a method for estimating θ from the actual number of alleles in a given sample of genes. However, if there are any deleterious genes segregating in the population, his method is expected to give an overestimate, even if such genes exist in low frequency and contribute very little to genetic variability.

In the previous data of AVISE and SELANDER (1972), the estimate of average heterozygosity is 0.0962 in population 3 and 0.1384 in population 4. Thus, the estimate of θ becomes 0.1064 in population 3 and 0.1607 in population 4. Hence, $\mathrm{Var}(h)$ is estimated to be 0.0266 in population 3 and 0.0349 in population 4. On the other hand, the estimate of $V_g(h)$ is 0.0246 in population 3 and 0.0310 in

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

population 4. Therefore, the expected and observed values of the variance of heterozygosity agree with each other surprisingly well. Of course, this may be coincidental and is not the proof for the neutral mutation hypothesis, since certain kinds of selection and varying mutation rates may produce the same effect. Apparently, more data should be analyzed before we make any conclusion from this sort of study.

We thank Drs. Ranajit Chakraborty and Stephen George for their helpful discussions.

APPENDIX

EXPECTED SQUARED DEVIATIONS OF j_X AND $\hat{g}$ FROM g

In the text we have defined g, j_X and $\hat{g}$ as follows:

$$g = \Sigma p_i^2, \quad j_X = \Sigma x_i^2, \quad \hat{g} = \frac{n\Sigma x_i^2 - 1}{n-1}$$

Thus,

$$(\hat{g} - g)^2 = \{\frac{n}{n-1}(j_X - g) + \frac{g-1}{n-1}\}^2$$
$$= (\frac{n}{n-1})^2(j_X - g)^2 + (\frac{g-1}{n-1})^2 + \frac{2n(j_x - g)(g-1)}{(n-1)^2}$$

The expectation of $(\hat{g} - g)^2$ is

$$E(\hat{g} - g)^2 = (\frac{n}{n-1})^2 E(j_X - g)^2 + (\frac{g-1}{n-1})^2 - \frac{2n(1-g)}{(n-1)^2} E(j_X - g)$$
$$= (\frac{n}{n-1})^2 E(j_X - g)^2 - (\frac{1-g}{n-1})^2 \tag{A1}$$

since $E(j_X) = g + (1-g)/n$, so that $E(j_X - g) = (1-g)/n$. Noting $n^2/(n-1)^2 = 1 + (2n-1)/(n-1)^2$, we have

$$E(\hat{g} - g)^2 - E(j_X - g)^2$$
$$= \frac{1}{(n-1)^2}\{(2n-1)E(j_X - g)^2 - (1-g)^2\}. \tag{A2}$$

$E(j_X - g)^2$ may be written as

$$E(j_X - g)^2 = E(j_X - \overline{j_X})^2 + (\overline{j_X} - g)^2$$
$$= E(j_X - \overline{j_X})^2 + (\frac{1-g}{n})^2$$

where $\overline{j_X} = E(j_X)$ and $E(j_X - \overline{j_X})^2$ is equal to $V_s(j_X)$ in (11) in the text. Therefore, we obtain

$$E(\hat{g} - g)^2 - E(j_X - g)^2 = [4(2n-1)(n-2)(\Sigma p_i^3 - g^2)$$
$$+ (1-g)\{(n^2 + 3n - 2)g - n(n-1)\}]/[n^3(n-1)]. \tag{A3}$$

If $E(\hat{g} - g)^2 - E(j_X - g)^2$ is positive, then j_X is a better estimate of g than $\hat{g}$. If it is negative, $\hat{g}$ is a better estimate. The first term inside the bracket in (A3) is always positive or 0, since

$$\Sigma p_i^3 - (\Sigma p_i^2)^2 = (\Sigma p_i^3)(\Sigma p_i) - (\Sigma p_i^2)^2$$
$$= \sum_{i \neq j} p_i p_j (p_i^2 - p_i p_j)$$
$$= \sum_{i>j} p_i p_j (p_i - p_j)^2 \geq 0.$$

The second term is positive if

$$g > \frac{n^2 - n}{n^2 + 3n - 2} \tag{A4}$$

Empirical data indicate that g is generally close to 1, the average being about 0.9. Therefore, the second term is also positive in a majority of cases if n is small. For example, if $g = 0.9$, then the second term is positive if n is smaller than 36. Then, j_X is a better estimate than $\hat{g}$. If n is large,

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024

it is not clear which is a better estimate, because of the second term in (A3). In this case, however, the difference between $E(\hat{g} - g)^2$ and $E(j_x - g)^2$ is very small, since the denominator $\{n^3(n - 1)\}$ in (A3) rapidly increases with increasing n. In general, therefore, j_X seems to be a better estimate than $\hat{g}$.

LITERATURE CITED

AVISE, J. S. and R. K. SELANDER, 1972 Evolutionary genetics of cave dwelling fishes of the genus *Astyanax*. Evolution **26**: 1–19.

BALAKRISHNAN, V. and L. D. SANGHVI, 1968 Distance between populations on the basis of attribute data. Biometrics **24**: 859–865.

CAVALLI-SFORZA, L. L. and A. W. F. EDWARDS, 1967 Phylogenetic analysis: models and estimation procedures. Am. J. Human Genet. **19**: 233–257.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theoret. Popul. Biol. **3**: 87–112.

KIDD, K. K. and L. L. CAVALLI-SFORZA, 1971 Number of characters examined and error in reconstruction of evolutionary trees. pp. 335–346. In: *Mathematics in the Archaeological and Historical Sciences. Proc. of the Anglo-Romanian Conf.* Edited by F. R. HODSON, D. G. KENDALL, and P. TAUTU. University of Edinburgh Press, Edinburgh.

KIMURA, M. and T. OHTA, 1971 *Theoretical Aspects of Population Genetics*. Princeton University Press, Princeton.

MORTON, N. E., S. YEE, D. E. HARRIS and R. LEW, 1971 Bioassay of kinship. Theoret. Popul. Biol. **2**: 507–524.

NEI, M., 1971 Identity of genes and genetic distance between populations. Genetics **68**: s47. ——, 1972 Genetic distance between populations. Am. Naturalist **106**: 283–292. ——, 1973a A new measure of genetic distance. *Genetic Distance*. Edited by J. F. CROW. Plenum Press, New York. (In press.) ——, 1973b The theory and estimation of genetic distance. *Genetic Structure of Populations*. Edited by N. E. MORTON. Univ. of Hawaii Press, Honolulu. (In press.)

NEI, M. and A. K. ROYCHOUDHURY, 1972 Gene differences between Caucasian, Negro and Japanese populations. Science **177**: 434–436.

ROGERS, J. S., 1972 Measures of genetic similarity and genetic distance. *Studies in Genetics VII*. Univ. of Texas Publ. **7213**: 145–153.

SANGHVI, L. D., 1953 Comparison of genetic and morphological methods for a study of biological differences. Amer. J. Phys. Anthrop. **11**: 385–404.

Corresponding Editor: R. LEWONTIN

Downloaded from https://academic.oup.com/genetics/article/76/2/379/5990983 by guest on 20 April 2024