THEORETICAL POPULATION BIOLOGY 11, 307–325 (1977)

# Drift Variances of $F_{ST}$ and $G_{ST}$ Statistics Obtained from a Finite Number of Isolated Populations

MASATOSHI NEI AND ARAVINDA CHAKRAVARTI

*Center for Demographic and Population Genetics, University of Texas at Houston, Texas 77030*

Received March 31, 1976

Approximate formulas for the mean and variance of the $F_{ST}$ or $G_{ST}$ statistic in a finite number of isolated populations are developed under the effect of random genetic drift. Computer simulation has shown that the approximate formulas give a fairly accurate result unless the initial frequency of one of the alleles involved is close to 1 and $t/2N$ is large, where $N$ is the effective size of a subpopulation and $t$ is the number of generations. It is shown that when the number of subpopulations ($s$) is small, the mean of $F_{ST}$ or $G_{ST}$ depends on the initial gene frequencies as well as on $s$. When the initial frequencies of all alleles are nearly equal to each other and the number of subpopulations is large, the distribution of $F_{ST}$ in the early generations is bell-shaped. In this case Lewontin and Krakauer's $k$ parameter is approximately 2 or less. However, if one of the initial allele frequencies is close to 1, the distribution is skewed and leptokurtic, and the $k$ parameter often becomes larger than 2 in later generations. Thus, even under pure random genetic drift, Lewontin and Krakauer's test of selective neutrality of polymorphic genes in terms of $F_{ST}$ is not always valid. It is also shown that Jacquard's approximate formula for $k$ generally gives an overestimate.

As a measure of the degree of genetic differentiation of populations, Wright (1943, 1951) introduced the following quantity:

$$F_{ST} = \frac{\sigma_x^2}{\bar{x}(1 - \bar{x})}, \tag{1}$$

where $\bar{x}$ and $\sigma_x^2$ are the mean and variance of gene frequency among populations. Recently, Nei (1973) extended this quantity to the case of multiple alleles and called it the coefficient of gene differentiation. This coefficient is defined as

$$G_{ST} = (H_T - H_S)/H_T, \tag{2}$$

where $H_S$ is the average gene diversity within populations and $H_T$ is the average gene diversity in the total population. The phrase *gene diversity* refers to the heterozygosity expected under Hardy–Weinberg equilibrium, irrespective of the actual genotype frequencies in the population. The difference $D_{ST} = H_T - H_S$





ISSN 0040-5809

is called the interpopulational gene diversity. If there are only two alleles at a locus, (2) reduces to (1). When there are multiple alleles, $G_{ST}$ is a weighted mean of $F_{ST}$, i.e., $G_{ST} = \sum_m \bar{x}_m(1 - \bar{x}_m) F_{STm} / \sum_m \bar{x}_m(1 - \bar{x}_m)$, where $F_{STm}$ is the value of $F_{ST}$ for the $m$th allele. Originally, $G_{ST}$ was intended to be applied to the average gene diversity for a large number of loci, but it can also be computed for each locus separately.

Wright's theory of $F$ statistics is based on the assumption of an infinitely large number of populations. If there is no migration among populations and the initial gene frequency is the same for all populations, the $F_{ST}$ in the $t$th generation is given by

$$\begin{aligned} F_{ST} &= 1 - \left(1 - \frac{1}{2N}\right)^t \\ &\approx 1 - e^{-t/2N}, \end{aligned} \tag{3}$$

where $N$ is the effective size of a population. In this case $G_{ST}$ is also given by the same formula (Nei, 1975). Clearly, $F_{ST}$ and $G_{ST}$ have no variance by definition.

In nature or in laboratory experiments, however, the number of populations is often very small. This is particularly so if we consider independent isolated populations (Robertson, 1975a,b). When the number of populations is small, $F_{ST}$ or $G_{ST}$ is subject to random genetic drift and is expected to show considerable variation. Thus, it is important to know the variances as well as the means of these quantities.

In the present paper we derive approximate formulas for the means and variances of $F_{ST}$ and $G_{ST}$ and test the accuracy of these formulas by means of computer simulation. We also study the distributions of these statistics under restricted conditions by using computer simulation. It is shown that Lewontin and Krakauer's (1973) test of selective neutrality of polymorphic genes in terms of $F_{ST}$ is not always valid even under pure random genetic drift. In this paper, we consider only the case of isolated populations. The effect of migration will be studied in an accompanying paper (Nei *et al.*, 1977).

## Mean and Variance of $G_{ST}(F_{ST})$

Since $F_{ST}$ is a special case of $G_{ST}$, we consider the latter quantity except when special remarks are necessary about $F_{ST}$. Suppose that $s$ subpopulations of effective size $N$ are derived from a foundation stock and in the subsequent generations no migration occurs among the subpopulations. (In the case of no migration this is, of course, equivalent to sampling $s$ subpopulations from a large number of independent subpopulations. If there is any migration, however, this is no longer true.) In the present paper we assume that the initial gene frequencies are the same for all subpopulations and random mating occurs in each sub-

population. To derive the mean and variance of $G_{ST}$, we consider an infinite number of independent populations (or loci) each consisting of $s$ subpopulations and regard the $G_{ST}$ for each population (or locus) as a random variable.

Let $x_{im}$ be the frequency of the $m$th allele at a locus in the $i$th subpopulation. The homozygosity or gene identity in this subpopulation is then given by $J_i = \sum_{m=1}^{r} x_{im}^2$, where $r$ is the number of alleles at the locus. On the other hand, the gene identity between the $i$th and $j$th subpopulations is defined as $J_{ij} = \sum_m x_{im}x_{jm}$. The gene identity within subpopulations ($J_S$) is the average of $J_i$ over subpopulations, and the heterozygosity or gene diversity within subpopulations is

$$H_S = 1 - J_S, \tag{4a}$$

where

$$J_S = \sum_{i=1}^{s} J_i/s. \tag{4b}$$

The gene identity in the total population ($J_T$) is computed by using the mean gene frequencies in the total population.

$$J_T = \sum_{m=1}^{r} \bar{x}_m{}^2,$$

where $\bar{x}_m$ is $\sum_i x_{im}/s$. Therefore, the gene diversity in the total population is

$$H_T = 1 - J_T, \tag{5a}$$

where

$$J_T = \left\{\sum_{i=1}^{s} J_i + \sum_{i \neq j} J_{ij}\right\}\Big/ s^2. \tag{5b}$$

$G_{ST}$ is then computed by (2).

It is noted that when $r = 2$, $H_T - H_S$ is $2\sum_i x_{im}^2/s - 2\bar{x}_m{}^2 = 2\sigma_{x1}^2$, where $\sigma_{x1}^2$ is the variance of gene frequency $x_{i1}$ among populations. On the other hand, $H_T = 2\bar{x}_1(1 - \bar{x}_1)$. Therefore, $G_{ST}$ becomes identical with $F_{ST}$, as mentioned earlier.

In his theoretical work, Nei (1975) studied a quantity defined as $E(G_{ST})^* = E(H_T - H_S)/E(H_T)$, where $E(\cdot)$ is the expectation operator. The reason for this is that he was interested in the ratio of the average of $H_T - H_S$ over loci to that of $H_T$. In the present paper we are interested in the mean and variance of $G_{ST}$ among replications of the same set of subpopulations or among independent loci. Unfortunately, it is difficult to get the exact mean and variance of $G_{ST}$, but approximate formulas may be obtained by using the Taylor expansion. The mean is given by

$$E(G_{ST}) = 1 - \frac{E(H_S)}{E(H_T)} + \frac{\mathrm{Cov}(H_S, H_T)}{E^2(H_T)} - \frac{E(H_S)\,V(H_T)}{E^3(H_T)} \tag{6}$$

approximately, where $V(\cdot)$ and $\mathrm{Cov}(\cdot, \cdot)$ refer to the variance and covariance, respectively. On the other hand, the approximate variance of $G_{ST}$ becomes

$$V(G_{ST}) = \frac{E^2(H_S)}{E^2(H_T)}\left[\frac{V(H_S)}{E^2(H_S)} + \frac{V(H_T)}{E^2(H_T)} - \frac{2\mathrm{Cov}(H_S, H_T)}{E(H_S)E(H_T)}\right]. \tag{7}$$

In (6) the third- and higher-order terms in the Taylor expansion are neglected, whereas in (7) the second- and higher-order terms are omitted. (In both formulas the terms involving the third and higher moments of $H_S$ and $H_T$ are neglected.) Thus, the accuracy of the above formulas depends on the magnitudes of the remainder terms. Analytical evaluation of the remainder terms are, however, very difficult, since they involve many terms of the higher moments of $H_S$ and $H_T$, which are in turn a complicated function of the initial allele frequencies, population size, generation time, and the number of subpopulations. For example, the fourth moments of $H_S$ and $H_T$ are a function of the first to eighth moments of allele frequencies. Nevertheless, from the studies of normal distributions, it is clear that the above formulas are quite accurate if $H_S$ and $H_T$ show a bell-shaped distribution. In the present case $H_S$ and $H_T$ are expected to show a bell-shaped distribution in the early generations if the initial frequencies of all alleles are more or less the same or if $s$ is sufficiently large. In the later generations, however, a bell-shaped distribution is not guaranteed. Later we examine the applicability of (6) and (7) by using computer simulations.

At any rate, if we know the means, variances, and covariance of $H_S$ and $H_T$, we can compute $E(G_{ST})$ and $V(G_{ST})$. In practice, it is simpler to work with $J_S$ and $J_T$ rather than with $H_S$ and $H_T$, using the relations $E(H_S) = 1 - E(J_S)$, $E(H_T) = 1 - E(J_T)$, $V(H_S) = V(J_S)$, $V(H_T) = V(J_T)$, and $\mathrm{Cov}(H_S, H_T) = \mathrm{Cov}(J_S, J_T)$.

From (4b) and (5b), it can be shown that

$$E(J_S) = E(J_i), \tag{8}$$

$$E(J_T) = \{E(J_i) + (s-1)E(J_{ij})\}/s, \tag{9}$$

$$V(J_S) = V(J_i)/s, \tag{10}$$

$$V(J_T) = \{V(J_i) + 4(s-1)\mathrm{Cov}(J_i, J_{ij}) + 2(s-1)V(J_{ij}) + 4(s-1)(s-2)\mathrm{Cov}(J_{ij}, J_{ik})\}/s^3, \tag{11}$$

$$\mathrm{Cov}(J_S, J_T) = \{V(J_i) + 2(s-1)\mathrm{Cov}(J_i, J_{ij})\}/s^2. \tag{12}$$

In the above formulas the subscripts of $J$ are used only to distinguish between the gene identities for the same population and for different populations; obviously $E(J_i) = E(J_j)$, $V(J_{ij}) = V(J_{kl})$, etc., for any combination of $i$, $j$, $k$, and $l$.

Our task is now reduced to finding out the moments of $J_i$ and $J_{ij}$ in the $t$th generation. The expectation of $J_i$ is known to be

$$E(J_i) = 1 - (1 - P_2)\lambda_1^t \tag{13}$$

where $\lambda_1 = 1 - 1/2N$ and $P_2 = \sum p_m^2$, in which $p_m$ is the initial frequency of the $m$th allele. On the other hand,

$$E(J_{ij}) = P_2 . \tag{14}$$

The variances and covariances of $J_i$ and $J_{ij}$ can be obtained by the method in Appendix I. They become

$$V(J_i) = 2C(1 - P_2)\lambda_1^t - (1 - P_2)^2 \lambda_1^{2t} + (\tfrac{2}{3} - 2P_2 + \tfrac{4}{3}P_3)\lambda_2^t + \{\tfrac{1}{3} - 2C(1 - P_2) - \tfrac{4}{3}P_3 + (P_2)^2\}\lambda_3^t, \tag{15}$$

$$V(J_{ij}) = P_2(1 - P_2) - 2(P_2 - P_3)\lambda_1^t + \{P_2(1 + P_2) - 2P_3\}\lambda_1^{2t}, \tag{16}$$

$$\mathrm{Cov}(J_i , J_{ij}) = (P_3 - P_2^2)(\lambda_1^t - \lambda_2^t), \tag{17}$$

$$\mathrm{Cov}(J_{ij} , J_{ik}) = (P_3 - P_2^2)(1 - \lambda_1^t), \tag{18}$$

where $P_3 = \Sigma p_m^3$, and

$$C = \frac{2N - 1}{10N - 6},$$

$$\lambda_1 = \left(1 - \frac{1}{2N}\right),$$

$$\lambda_2 = \left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right),$$

$$\lambda_3 = \left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right)\left(1 - \frac{3}{2N}\right).$$

It is noted that the above formulas depend on $N$ and $t$ separately. In practice, however, if $N$ is sufficiently large, say larger than 10, they are essentially a function of $t/2N$. (When the initial frequency of one allele is close to 1, a larger $N$ is required for this statement to be true.) This is because, for a large $N$, $C$, $\lambda_1$, $\lambda_2$, and $\lambda_3$ can be approximated by $\frac{1}{5}$, $e^{-1/2N}$, $e^{-3/2N}$, and $e^{-6/2N}$, respectively. Therefore, when we deal with a large population, the formulas with these approximations may be used. These approximate formulas have the advantage that they depend only on $t/2N$ rather than on $N$ and $t$ separately.

When the initial frequency is more or less the same for all alleles, a further approximation may be made to obtain the value of $E(G_{ST})$. Namely, when $p_i = p_j$ and there are $r$ alleles, $P_2 = 1/r$, $P_3 = 1/r^2$, and $P_3 - P_2^2 = 0$. Therefore, $\mathrm{Cov}(J_i , J_{ij})$ and $\mathrm{Cov}(J_{ij} , J_{ik})$ are both 0 for all generations, and

$$V(J_i) = \left(1 - \frac{1}{r}\right)\left[2C(\lambda_1^t - \lambda_3^t) - \left(1 - \frac{1}{r}\right)\lambda_1^{2t} + \frac{2}{3}\left(1 - \frac{2}{r}\right)\lambda_2^t + \frac{1}{3}\left(1 + \frac{1}{r}\right)\lambda_3^t\right],$$

$$V(J_{ij}) = \frac{1}{r}\left(1 - \frac{1}{r}\right)(1 - \lambda_1^t)^2.$$

In this case it can be shown that the terms $\mathrm{Cov}(H_S, H_T)/E^2(H_T)$ and $E(H_S)\,V(H_T)/E^3(H_T)$ in (6) are very small compared with $E(H_S)/E(H_T)$. Therefore, we have

$$E(G_{ST}) = \frac{(1 - 1/s)(1 - e^{-t/2N})}{1 - (1 - e^{-t/2N})/s} \tag{19}$$

approximately, which is identical with Nei's (1975) formula for $E(G_{ST})$*. As expected, when $s \to \infty$, $E(G_{ST})$ reduces to (3). Of course, when $s = \infty$, $E(G_{ST})$ reduces to (3) for any initial gene frequencies, since in this case the variances and covariance of $H_S$ and $H_T$ are all 0. When $s$ is large, say 50, formula (19) holds approximately if the initial gene frequencies are all less than 0.9.

TABLE I

Theoretical Means and Standard Deviations of $G_{ST}$ for Various Values of Initial Gene Frequencies ($p_i$) and the Number of Subpopulations ($s$)[a]

| Generations | $s = 2$ | | | $s = 10$ | | | $s = 100$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $t/2N$ | $E(G_{ST})$ | $s(G_{ST})$ | $k$ | $E(G_{ST})$ | $s(G_{ST})$ | $k$ | $E(G_{ST})$ | $s(G_{ST})$ | $k$ |
| | | | (1) | $p_1 = p_2 = 0.5$ | | | | | |
| 0.05 | 0.025 | 0.034 | 1.9 | 0.044 | 0.020 | 1.8 | 0.048 | 0.007 | 1.8 |
| 0.10 | 0.050 | 0.067 | 1.8 | 0.086 | 0.037 | 1.7 | 0.094 | 0.012 | 1.6 |
| 0.25 | 0.122 | 0.155 | 1.6 | 0.203 | 0.077 | 1.3 | 0.219 | 0.025 | 1.3 |
| 0.50 | 0.230 | 0.269 | 1.4 | 0.368 | 0.115 | 0.9 | 0.391 | 0.036 | 0.8 |
| 1.00 | 0.410 | 0.391 | 0.9 | 0.606 | 0.134 | 0.4 | 0.630 | 0.040 | 0.4 |
| | | | (2) | $p_1 = 0.2, p_2 = 0.8$ | | | | | |
| 0.05 | 0.024 | 0.035 | 2.1 | 0.044 | 0.020 | 1.9 | 0.048 | 0.007 | 1.9 |
| 0.10 | 0.047 | 0.069 | 2.1 | 0.086 | 0.039 | 1.9 | 0.094 | 0.013 | 1.8 |
| 0.25 | 0.111 | 0.164 | 2.2 | 0.200 | 0.086 | 1.7 | 0.219 | 0.028 | 1.6 |
| 0.50 | 0.202 | 0.297 | 2.2 | 0.358 | 0.135 | 1.3 | 0.390 | 0.043 | 1.2 |
| 1.00 | 0.358 | 0.454 | 1.6 | 0.590 | 0.167 | 0.7 | 0.628 | 0.051 | 0.6 |
| | | | (3) | $p_1 = p_2 = p_3 = p_4 = 0.25$ | | | | | |
| 0.05 | 0.025 | 0.020 | 0.6 | 0.044 | 0.012 | 0.6 | 0.048 | 0.004 | 0.6 |
| 0.10 | 0.050 | 0.039 | 0.6 | 0.086 | 0.022 | 0.6 | 0.094 | 0.007 | 0.6 |
| 0.25 | 0.124 | 0.093 | 0.6 | 0.204 | 0.048 | 0.5 | 0.219 | 0.015 | 0.5 |
| 0.50 | 0.243 | 0.168 | 0.5 | 0.369 | 0.076 | 0.4 | 0.391 | 0.024 | 0.4 |
| 1.00 | 0.456 | 0.261 | 0.3 | 0.607 | 0.096 | 0.2 | 0.630 | 0.029 | 0.2 |

[a] $k$: Lewontin and Krakauer's $k$ parameter. These values were obtained by using formulas (13)–(18) with approximations $C = \frac{1}{5}$, $\lambda_1 = e^{-1/2N}$, $\lambda_2 = e^{-3/2N}$, and $\lambda_3 = e^{-6/2N}$.

The means and variances of $G_{ST}$ for various values of initial gene frequencies and the number of subpopulations are presented in Table I. All these values were obtained by using formulas (13)–(18) with the approximations $C = \frac{1}{5}$,

$\lambda_1 = e^{-1/2N}$, etc., so that they are applicable to populations of any size except extremely small ones. It is seen that $E(G_{ST})$ depends on $s$ as well as on the initial gene frequencies. It is small when $s$ is small. This is because, if $s$ is small, the chance that all subpopulations have similar gene frequencies is high. Particularly when $s = 2$, this chance is considerably high. Note also that if $s$ is small and the initial frequency of an allele is close to 1, this allele may be fixed in all subpopulations with an appreciable probability even in relatively early generations. Then, $G_{ST}$ becomes 0 (see the next section). Thus, the value of $E(G_{ST})$ in the case of $p_1 = 0.2$ and $p_2 = 0.8$ is somewhat smaller than that in the case of $p_1 = p_2 = 0.5$ in all generations. The approximate formula (19) holds better when there are many alleles of similar initial frequencies than when there are only two alleles. For example, the values of $E(G_{ST})$ for case (3) in Table I are identical with the values from (19) up to the fourth significant digit, though they are not presented here. On the other hand, the $E(G_{ST})$ values for case (1) are somewhat deviated from these values.

The formula for the variance of $G_{ST}$ is quite complicated even in the case of $p_i = p_j$, and the property of this quantity is best understood by examining the numerical values in Table I, where the standard deviation ($s(G_{ST})$) rather than the variance is presented. As expected, the standard deviation is very large when $s$ is small. For example, in the case of two subpopulations with two alleles the standard deviation is always larger than the expectation in the early generations. The ratio of $s(G_{ST})$ to $E(G_{ST})$, however, gradually declines as generation proceeds. This ratio is also smaller when there are many alleles segregating than when there are only two alleles.

By using computer simulation, Lewontin and Krakauer (1973) studied the sampling variance of $F_{ST}$ when $s$ independent subpopulations are sampled. They showed that if the gene frequency distribution is binomial with the mean gene frequency being intermediate, the quantity

$$k = \frac{(s-1)\,V(F_{ST})}{E^2(F_{ST})}$$

is approximately 2, whereas if the gene frequency distribution is flatter than the binomial, $k$ is smaller than 2. As argued by Robertson (1975a,b), when a population is split into a large number of subpopulations, the gene frequency distribution is expected to be close to the binomial or normal distribution in the early generations, so that $k$ is expected to be close to 2. This property can be studied by using $E(G_{ST})$ and $V(G_{ST})$ for the case of the number of alleles ($r$) equal to 2.

When $T \equiv t/2N$ is much smaller than 1, $E(G_{ST})$ can be written as

$$E(G_{ST}) = \frac{(s-1)T}{s-T} \tag{20}$$

approximately. Furthermore, if $p_i = p_j$ for all pairs of alleles, $V(G_{ST})$ is given by

$$V(G_{ST}) = \frac{2sT^2}{(r-1)(s-T)^2}\left[1 + \frac{s(1-T)^2}{(s-T)^2} - \frac{2(1-T)}{(s-T)}\right] \tag{21}$$

approximately. Therefore, we have

$$\begin{aligned}\frac{V(G_{ST})}{E^2(G_{ST})} &= \frac{2s}{(r-1)(s-1)^2}\left[1 + \frac{s(1-T)^2}{(s-T)^2} - \frac{2(1-T)}{(s-T)}\right] \\ &\approx \frac{2}{(r-1)(s-1)}.\end{aligned} \tag{22}$$

In the case of two alleles, $r = 2$, so that $k$ becomes approximately 2 in the early generations, as predicted by Robertson. Numerical computations in Table I show that formula (22) approximately holds for $T \leqslant 0.1$ (cases 1 and 3). For a large value of $T$, however, $k$ often becomes smaller than the value given by (22).

TABLE II

Observed Values of the Mean $[E(G_{ST})]$ and Standard Deviation $[s(G_{ST})]$ of $G_{ST}$ in Monte Carlo Simulations[a]

| Generations $t/2N$ | $s = 2$ | | | $s = 10$ | | |
|---|---|---|---|---|---|---|
| | $E(G_{ST})$ | $s(G_{ST})$ | $k$ | $E(G_{ST})$ | $s(G_{ST})$ | $k$ |
| | (1) $p_1 = p_2 = 0.5$ | | | | | |
| 0.05 | 0.030 | 0.044 | 2.2 | 0.045 | 0.021 | 1.9 |
| 0.10 | 0.053 | 0.068 | 1.6 | 0.089 | 0.038 | 1.7 |
| 0.25 | 0.124 | 0.143 | 1.3 | 0.211 | 0.079 | 1.3 |
| 0.50 | 0.228 | 0.234 | 1.1 | 0.376 | 0.121 | 1.0 |
| 1.00 | 0.332 | 0.323 | 1.0 | 0.620 | 0.142 | 0.5 |
| | (2) $p_1 = 0.2, p_2 = 0.8$ | | | | | |
| 0.05 | 0.027 | 0.034 | 1.6 | 0.046 | 0.022 | 2.1 |
| 0.10 | 0.046 | 0.058 | 1.6 | 0.087 | 0.040 | 1.9 |
| 0.25 | 0.096 | 0.111 | 1.3 | 0.208 | 0.077 | 1.2 |
| 0.50 | 0.151 | 0.182 | 1.5 | 0.358 | 0.135 | 1.3 |
| 1.00 | 0.166 | 0.241 | 2.1 | 0.575 | 0.199 | 1.1 |
| | (3) $p_1 = p_2 = p_3 = p_4 = 0.25$ | | | | | |
| 0.05 | 0.024 | 0.020 | 0.7 | 0.046 | 0.012 | 0.6 |
| 0.10 | 0.048 | 0.037 | 0.6 | 0.088 | 0.022 | 0.6 |
| 0.25 | 0.133 | 0.100 | 0.6 | 0.209 | 0.052 | 0.6 |
| 0.50 | 0.245 | 0.180 | 0.5 | 0.376 | 0.076 | 0.4 |
| 1.00 | 0.452 | 0.269 | 0.4 | 0.620 | 0.097 | 0.2 |

[a] $s$: number of subpopulations. $p_i$: the initial gene frequency. $k$: Lewontin and Krakauer's $k$ parameter. The simulations were conducted with $N = 10$. The number of replications used is 400.

This property holds approximately even for the case of $p_i \neq p_j$ (case 2), unless the deviation from equal initial gene frequencies is very large. That is,

$$k \leqslant 2/(r-1) \tag{23}$$

approximately. In the case of $p_1 = 0.2$ and $p_2 = 0.8$ with $s = 2$, $k$ is slightly larger than 2 in the early generations, but this seems to be due to the fact that the accuracy of our approximate formulas declines in this case (see Table II). However, if the deviation from equal initial gene frequencies is very large, the above property no longer holds and $k$ may become substantially larger than $2/(r-1)$, as will be seen later.

## Monte Carlo Simulation

Since our formulas for $E(G_{ST})$ and $V(G_{ST})$ involve approximations, we examined the accuracy of these formulas by computer simulation. At the same time we also studied the distribution of $G_{ST}$. The number of subpopulations ($s$) used in this simulation was 2 and 10, and in all cases the size ($N$) of a subpopulation was 10. We used these small values of $s$ and $N$ because the accuracy of our formulas is poor when these are small. In each generation $2N = 20$ genes were sampled at random for the next generation in each subpopulation. Sampling of genes was done by using pseudorandom numbers of a uniform distribution between 0 and 1. After obtaining gene frequencies, the $G_{ST}$ value was computed for successive generations in each replication. The number of replications used was either 400 or 800, depending on the initial gene frequencies. The observed means and standard deviations of $G_{ST}$ for the case of intermediate initial gene frequencies (corresponding to that of Table I) are presented in Table II.

Before going to the discussion of Table II, it should be mentioned that when the same allele was fixed in all populations we defined $G_{ST} = 0$, though the formal application of formula (2) gives 0/0. This definition is justified, because $G_{ST}$ tends to 0 when the frequency of an allele in the total population approaches 1 (Appendix II). It is also biologically natural, since there is no genetic variability among subpopulations if the same allele is fixed in all subpopulations.

It is clear from Table II that in the early generations the agreement between the theoretical and observed values of $E(G_{ST})$ and $s(G_{ST})$ is satisfactory, particularly when all alleles are initially equally frequent and the number of subpopulations is large. Thus, if $s = 10$ and all alleles are initially equally frequent, the agreement between the theoretical and observed values is very good until $T = t/2N = 1.0$, at which $E(G_{ST})$ is 0.62. In the case of $p_1 = 0.2$ and $p_2 = 0.8$ with $s = 10$ the agreement is satisfactory until $T = 0.5$, at which $E(G_{ST}) = 0.36$. Since the average $G_{ST}$ or $F_{ST}$ values observed in natural populations are generally much smaller than 0.36 (Nei, 1975, p. 175), our formulas for the mean and

variance of $G_{ST}$ seem to hold in most practical cases, as long as $s$ is sufficiently large. We note that the observed value of $E(G_{ST})$ or $E(F_{ST})$ in man is almost always lower than 0.10, even among genetically quite different populations. (In real populations, of course, the effect of migration cannot be neglected; see the accompanying paper.) Note also that in natural populations $N$ is generally much larger than 100.

When $s$ is extremely small and the initial allele frequencies deviate from equality, however, the agreement is not good except in the very early generations. This is particularly so with respect to the $k$ value. This suggests that our approximate formulas should not be used in such a case. Fortunately, the $F_{ST}$ or $G_{ST}$ statistic is rarely used when $s$ is extremely small. In fact, when $s$ is small, a more meaningful study of the gene differentiation among populations is accomplished by pairwise comparisons of populations (see Nei, 1975, Chap. 7).

It is noted that the values in Table I were obtained by the approximate formulas for $E(J_i)$, $V(J_i)$, etc. Numerical computations, however, have shown that there are only small differences between the values obtained by the exact and approximate formulas in the present case ($N = 10$ and initial allele frequencies not close to 1).

TABLE III

Observed Values of the Mean $[E(G_{ST})]$ and Standard Deviation $[s(G_{ST})]$ of $G_{ST}$ in Monte Carlo Simulations[a]

| Generations $t/2N$ | $E(G_{ST})$ | $s(G_{ST})$ | $k$ | $k'$ | $n'$ |
|---|---|---|---|---|---|
| | | (1) $p_1 = 0.1, p_2 = 0.9$ | | | |
| 0.05 | 0.046(0.045) | 0.021(0.021) | 1.9(1.9) | 1.9 | 800 |
| 0.10 | 0.090(0.087) | 0.041(0.041) | 1.8(2.0) | 1.8 | 800 |
| 0.25 | 0.194(0.197) | 0.087(0.100) | 1.8(2.3) | 1.8 | 800 |
| 0.50 | 0.319(0.346) | 0.164(0.172) | 2.4(2.2) | 2.2 | 785 |
| 1.00 | 0.455(0.566) | 0.279(0.223) | 3.4(1.4) | 2.0 | 709 |
| | | (2) $p_1 = 0.05, p_2 = 0.95$ | | | |
| 0.05 | 0.045(0.045) | 0.021(0.021) | 1.9(1.9) | 1.9 | 800 |
| 0.10 | 0.083(0.085) | 0.038(0.045) | 1.9(2.5) | 1.9 | 800 |
| 0.25 | 0.172(0.182) | 0.102(0.123) | 3.1(4.1) | 2.7 | 773 |
| 0.50 | 0.250(0.304) | 0.182(0.227) | 4.8(5.0) | 3.1 | 700 |
| 1.00 | 0.317(0.496) | 0.304(0.306) | 8.3(3.4) | 2.6 | 537 |
| 1.50 | 0.349(0.653) | 0.378(0.288) | 10.5(1.7) | 2.1 | 455 |

[a] The number of subpopulations is 10 with $N = 10$. $p_i$ : initial gene frequencies. $k$: Lewontin and Krakauer's $k$ parameter. $k'$: $k$ value obtained after removal of the cases of $G_{ST} = 0$. The number of replications used is 800. $n'$: number of replications in which $G_{ST} \neq 0$. The figures in parentheses are the theoretical values obtained by using formulas (13)–(18) without approximations.

As we have seen above, the accuracy of our mathematical formulas declines if the initial frequencies of different alleles are extremely dissimilar. To see this point, we conducted another computer simulation with two sets of initial frequencies, i.e., $p_1 = 0.1$; $p_2 = 0.9$ and $p_1 = 0.05$; $p_2 = 0.95$. The results for $s = 10$ with $N = 10$ are given in Table III. In this case the exact formulas for $E(J_i)$, $V(J_i)$, etc., were used. (When the initial frequencies are extremely dissimilar, the exact formulas give a better result than the approximate formulas.) It is clear that when $p_1 = 0.1$ and $p_2 = 0.9$, the agreement between the observed and theoretical values of $E(G_{ST})$ and $s(G_{ST})$ are still reasonably good, particularly in the early generations. There is, however, a substantial disagreement about the $k$ value in the later generations. In the case of $p_1 = 0.05$ and $p_2 = 0.95$ the agreement between the observed and theoretical values is poor except in the very early generations. It should also be noted that the agreement depends considerably on the number of subpopulations. If this number is large, the agreement is better.

One important feature of Table III is the fact that, unlike Lewontin and Krakauer's prediction, the $k$ value increases in the later generations and can exceed 2. One reason for this is that in our computation the value of $G_{ST} = 0$ is included. As mentioned earlier, $G_{ST}$ becomes 0 when the same allele is fixed in all subpopulations, and this occurs quite frequently when $p_1$ or $p_2$ is close to 1. In practice, however, the loci at which the same allele is fixed in all subpopulations would not be included in Lewontin and Krakauer's test, since in this test monomorphic loci are not used. We therefore recomputed the $k$ value, eliminating all cases in which the same allele is fixed in all subpopulations. This value, denoted by $k'$, is presented in Table III. It is clear, however, that even the $k'$ value exceeds 2 substantially in the later generations.

The $k'$ value higher than 2 is caused by the extremely skewed and leptokurtic distributions of gene frequencies in these generations. The gene frequency distribution under pure random genetic drift is close to the binomial distribution in the early generations. Thus, $k$ becomes approximately 2, unless $p_1$ is extremely close to 0 or 1. However, if $p_1$ is about 0.05 (or 0.95), the gene frequency distribution gradually becomes skewed and leptokurtic in the later generations, since the frequencies of gene frequency classes close to 1 (or 0) increase by genetic drift (Kimura, 1955). As shown by Jacquard (1974), if the gene frequency distribution is extremely leptokurtic, the $k$ or $k'$ may become larger than 2.

In this connection, however, it should be noted that Jacquard's approximate formula for $k$ for a pair of alleles generally gives an overestimate. That is, approximating $\bar{x}(1 - \bar{x})$ by $E(x)[1 - E(x)]$ in (1) and assuming $s$ large, he obtained the following formula:

$$k \approx \gamma_2 + 2,$$

where $\gamma_2$ is the kurtosis of the gene frequency distribution (see also Ewens and Feldman, 1976). In the present case $\gamma_2$ can be evaluated by using Robertson's

(1952) formulas for the moments of gene frequencies with $N = 10$. Using this $\gamma_2$, we obtain the following values of $k$:

| $T = t/2N$ | 0.05 | 0.1 | 0.5 | 1.0 | 1.5 |
|---|---|---|---|---|---|
| $p_1 = 0.1$ | 2.3 | 3.0 | 6.3 | 7.3 | 7.4 |
| $p_1 = 0.05$ | 2.8 | 4.7 | 14.0 | 16.9 | 17.4 |

Comparison of these values with those ($k$'s) in Table III indicates that the $k$ value obtained by Jacquard's formula is a gross overestimate. It is noted that the $k$ values obtained by our approximate formulas for $E(G_{ST})$ and $V(G_{ST})$ (the values in parentheses) are much closer to the $k'$ values than those obtained by Jacquard's formula. The poor performance of Jacquard's formula when the initial gene frequency is close to 0 or 1 is mainly due to the positive correlation between $\sigma_x^2$ and $\bar{x}(1 - \bar{x})$, i.e., between $H_S$ and $H_T$ in (6) and (7) (see formula (12)). One might wonder if the relatively small value of $s$ we used in our simulation ($s = 10$) has anything to do with the discrepancy between the observed value of $k$ and Jacquard's $k$. We examined this possibility by conducting another simulation with $p_1 = 0.1$, $p_2 = 0.9$, and $s = 20$. However, the results obtained from 800 replicate runs were not very different from those for the case of the same initial gene frequencies but with $s = 10$.

At any rate, Table III clearly shows that $k$ can be larger than 2 even under pure random genetic drift when the initial gene frequency is close to 0 or 1. Since the gene frequency in natural populations is often close to 0 or 1 at many polymorphic loci, it is quite likely that the above situation occurs in nature. In fact, in the populations of *Thomomys talpoides* studied by Nevo (1973) the pattern of gene frequency distributions at some loci is similar to that considered above.

Figures 1 and 2 give the distributions of $G_{ST}$ for the cases of $p_1 = p_2 = 0.5$ and $p_1 = 0.1$; $p_2 = 0.9$. It is clear that the distribution is highly dependent on the number of subpopulations and the initial gene frequencies. In the case of $s = 2$ the distribution is inverse J-shaped or L-shaped when $T = t/2N$ is 1 or smaller, whereas if $s \geqslant 10$ the distribution becomes close to the bell-shape. The inverse J-shaped distribution of $F_{ST}$ for a small value of $s$ ($= 4$) was also observed by MacCluer (1974), who studied the genetic differentiation of a group of human populations by means of computer simulation. The reason for the inverse J-shaped distribution for a small value of $s$ is that the chance of all subpopulations having similar gene frequencies is appreciably high. Furthermore, in the later generations the same allele may be fixed in all subpopulations. In this case $G_{ST}$ is necessarily 0. Therefore, the mean of $G_{ST}$ is lower than 1 even when $t = \infty$. At $t = \infty$, $G_{ST}$ is either 0 or 1. For example, in the extreme case of $s = 2$, the expected frequencies of $G_{ST} = 0$ and $G_{ST} = 1$ at $t = \infty$ are $\sum p_i^2$ and $1 - \sum p_i^2$, respectively. Thus, the mean of $G_{ST}$ is $1 - \sum p_i^2$, which is equal to the heterozygosity in the initial population.

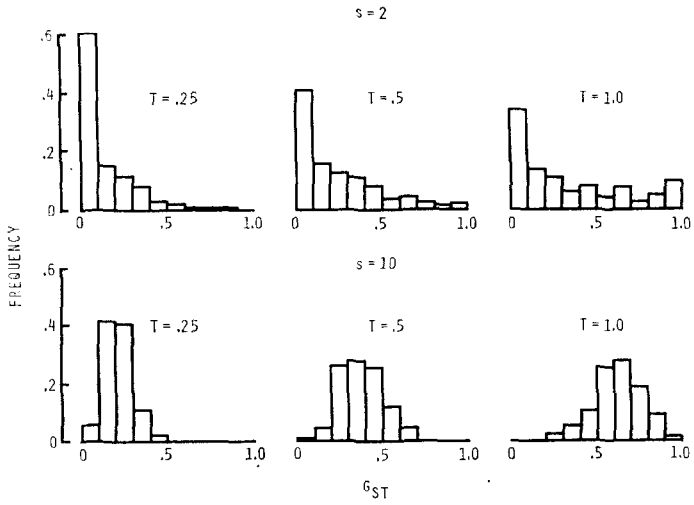


FIG. 1. Frequency distributions of $G_{ST}$ (or $F_{ST}$) obtained by computer simulation. The initial gene frequencies are $p_1 = p_2 = 0.5$. $T = t/2N$ and $s$ is the number of subpopulations.

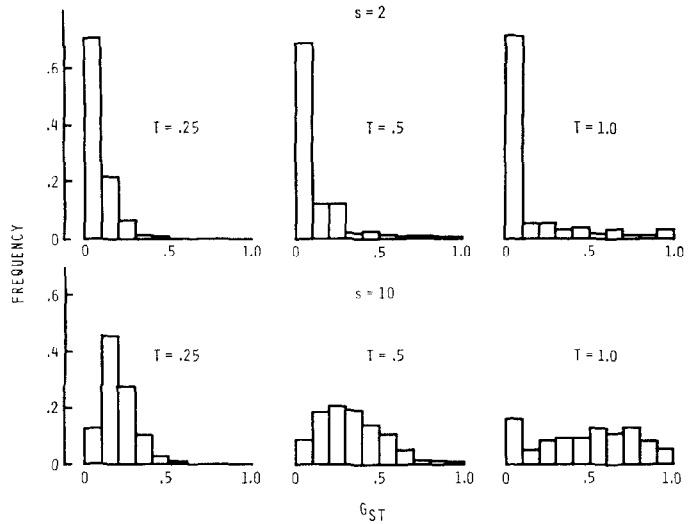


FIG. 2. Frequency distributions of $G_{ST}$ (or $F_{ST}$) obtained by computer simulation. The initial gene frequencies are $p_1 = 0.1$ and $p_2 = 0.9$. $T = t/2N$ and $s$ is the number of subpopulations.

When $s$ is large, the chance of the same allele being fixed in all subpopulations is extremely small. Therefore, $G_{ST}$ almost always becomes 1 at $t = \infty$. In this case the variance of $G_{ST}$ is virtually 0. In the intermediate generations, however, the variance of $G_{ST}$ is appreciably high even if $s$ is as large as 100.

In the present work, we studied the mean and variance of $G_{ST}$ in the probability space defined earlier. For some purposes, however, it is more meaningful to consider the mean and variance of $G_{ST}$ conditional on the event that at least two alleles exist in the total population (i.e., excluding the event that the same allele is fixed in all subpopulations). This conditional mean becomes 1 at $t = \infty$ irrespective of the number of subpopulations. Though this is a good property, it does not appear to be easy to derive the theoretical value of $E(G_{ST})$ at time $t$ in this case. In the presence of migration, however, an approximate formula for this conditional mean at steady state has been worked out (Nei *et al.*, 1977).

## Discussion

In the past three decades the $F_{ST}$ statistic has been used extensively for the study of population differentiation. In many cases $F_{ST}$ was estimated from a single locus. The present study, however, shows that $F_{ST}$ has a large variance due to genetic drift if the number of subpopulations is small. Therefore, in order to study the genetic differentiation of populations, a large number of loci should be used. This is particularly so if $F_{ST}$ is to be used for estimating the divergence time of subpopulations. It should also be noted that when the number of independent populations is small, formula (3) does not give the correct expected value of $F_{ST}$.

In this connection it should be mentioned that for the purpose of estimation of the time after divergence of populations it is better to use the formula $E(G_{ST})^* = 1 - E(H_S)/E(H_T)$ rather than $E(G_{ST}) = 1 - E(H_S/H_T)$ and estimate $E(H_S)$ and $E(H_T)$ by the average values of $H_S$ and $H_T$ for all loci studied, respectively. If the number of loci used is sufficiently large, then the divergence time may be estimated by

$$t = 2N \ln \frac{s + \hat{G} - 1}{(s - 1)(1 - \hat{G})} \tag{24}$$

where $\hat{G} = 1 - \hat{H}_S/\hat{H}_T$, in which $\hat{H}_S$ and $\hat{H}_T$ are the estimates of $E(H_S)$ and $E(H_T)$, respectively. In particular, for $s = 2$, (24) reduces to

$$t = 2N \ln[(2\hat{H}_T - \hat{H}_S)/\hat{H}_S]. \tag{25}$$

The above formulas ignore the effect of mutation, so they are not valid when $t$ is very large. In this case the more rigorous formulas given by Nei (1972, 1976) should be used.

In the present paper we have studied the mean and variance of $F_{ST}$ or $G_{ST}$.

The mathematical formulas developed here are directly applicable to experimental data such as those obtained by Kerr and Wright (1954), Buri (1956), and Sing *et al.* (1973). Application of the formulas to natural populations, however, requires some caution. This is because splitting of natural populations does not generally occur simultaneously as assumed here (except in the case of $s = 2$), and gene migration almost always takes place among subpopulations at least in the early generations. When $F_{ST}$ is small and the distribution of gene frequency in subpopulations is close to the normal distribution, the effect of nonsimultaneous splitting of subpopulations on the variance of $F_{ST}$ can be evaluated by Robertson's (1975b) method, but otherwise it is difficult. Gene migration, which must be quite irregular in natural populations, also complicates the problem considerably. It is not easy to evaluate the general effect of migration, but in the accompanying paper we examine this problem in some detail by using the island model of finite size.

Lewontin and Krakauer (1973) proposed a method of testing the selective neutrality of polymorphic genes on the assumption of $k \leqslant 2$. In practice, however, $k$ can be larger than 2 even under pure genetic drift, as shown in this paper. Furthermore, in natural populations the assumptions of simultaneous splitting of populations as well as of no migration and no mutation are not usually satisfied. Violation of these assumptions further weakens the power of their test, as indicated by Nei and Maruyama (1975), Robertson (1975a,b), and Ewens and Feldman (1976). Therefore, their test cannot be used as a general method. Of course, in experimental populations such as those of Sing *et al.* (1973), all the parameters can be artificially controlled. In this case the Lewontin–Krakauer test may be used. The only problem in artificial populations is the effect of associative overdominance, which may be introduced when the initial population is synthesized.

## APPENDIX I: EXACT VARIANCES AND COVARIANCES OF GENE IDENTITIES $J_i$ AND $J_{ij}$

$J_i$ and $J_{ij}$ are defined as

$$J_i = \sum_{m=1}^{r} x_{im}^2 \qquad \text{and} \qquad J_{ij} = \sum_{m=1}^{r} x_{im}x_{jm} ,$$

respectively. Therefore, the variance of $J_i$ is

$$\begin{aligned} V(J_i) &= E\left(\sum_m x_{im}^2\right)^2 - E^2\left(\sum_m x_{im}^2\right), \\ &= \sum_m E(x_{im}^4) + \sum_{m \neq n} E(x_{im}^2 x_{in}^2) - \left(\sum_m E(x_{im}^2)\right)^2, \end{aligned} \tag{A1}$$

while the variance of $J_{ij}$ is

$$V(J_{ij}) = E\left(\sum_m x_{im}x_{jm}\right)^2 - E^2\left(\sum_m x_{im}x_{jm}\right)$$

$$= \sum_m E(x_{im}^2)\,E(x_{jm}^2) + \sum_{m\neq n} E(x_{im}x_{in})\,E(x_{jm}x_{jn}) \qquad \text{(A2)}$$

$$- \left(\sum_m E(x_{im})E(x_{jm})\right)^2.$$

Similarly, the covariance $\{\mathrm{Cov}(J_i, J_{ij})\}$ of $J_i$ and $J_{ij}$ and the covariance $\{\mathrm{Cov}(J_{ij}, J_{ik})\}$ of $J_{ij}$ and $J_{ik}$ are

$$\mathrm{Cov}(J_i, J_{ij}) = \sum_m E(x_{im}^3)\,E(x_{jm}) + \sum_{m\neq n} E(x_{im}^2x_{in})\,E(x_{jn})$$

$$- \left(\sum_m E(x_{im}^2)\right)\left(\sum_m E(x_{im})\,E(x_{jm})\right). \qquad \text{(A3)}$$

$$\mathrm{Cov}(J_{ij}, J_{ik}) = \sum_m E(x_{im}^2)\,E(x_{jm})\,E(x_{km}) + \sum_{m\neq n} E(x_{im}x_{in})\,E(x_{jm})\,E(x_{kn})$$

$$- \left(\sum_m E(x_{im})E(x_{jm})\right)\left(\sum_n E(x_{in})\,E(x_{kn})\right). \qquad \text{(A4)}$$

Therefore, if we know the moments of gene frequencies of the forms $\mu_r = E(x_{im}^r)$ and $\mu_{rs} = E(x_{im}^r x_{in}^s)$, we can evaluate the variances and covariances of $J_i$ and $J_{ij}$. (Although we are dealing with moments around the origin, we use the notation $\mu$ rather than $\mu'$ for simplicity.) The moments $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$ in the $t$th generation have already been worked out by Robertson (1952) and given by

$$\mu_1^{(t)} = p,$$

$$\mu_2^{(t)} = p - p(1-p)\lambda_1{}^t,$$

$$\mu_3^{(t)} = p - \tfrac{3}{2}p(1-p)\lambda_1{}^t - \tfrac{1}{2}p(1-p)(2p-1)\lambda_2{}^t,$$

$$\mu_4^{(t)} = p - c_1p(1-p)\lambda_1{}^t - p(1-p)(2p-1)\lambda_2{}^t + p(1-p)[p(1-p) - c_2]\lambda_3{}^t,$$

where $p$ is the initial gene frequency, $c_1 = (18N-11)/(10N-6)$, $c_2 = 2 - c_1$, $\lambda_1 = (1-1/n)$, $\lambda_2 = (1-1/n)(1-2/n)$, $\lambda_3 = (1-1/n)(1-2/n)(1-3/n)$, in which $n = 2N$.

The other moments required ($\mu_{11}$, $\mu_{12}$, $\mu_{21}$, $\mu_{22}$) are obtained in the following

way. From the property of multinomial distribution, we can derive the following recurrence equations for these moments:

$$\mu_{11}^{(t+1)} = \left(1 - \frac{1}{n}\right)\mu_{11}^{(t)},$$

$$\mu_{12}^{(t+1)} = \frac{1}{n}\left(1 - \frac{1}{n}\right)\mu_{11}^{(t)} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\mu_{12}^{(t)},$$

$$\mu_{21}^{(t+1)} = \frac{1}{n}\left(1 - \frac{1}{n}\right)\mu_{11}^{(t)} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\mu_{21}^{(t)},$$

$$\mu_{22}^{(t+1)} = \frac{1}{n^2}\left(1 - \frac{1}{n}\right)\mu_{11}^{(t)} + \frac{1}{n}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\{\mu_{12}^{(t)} + \mu_{21}^{(t)}\} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3}{n}\right)\mu_{22}^{(t)},$$

where $\mu^{(t)}$ refers to the moment at the $t$th generation. These simultaneous recurrence equations can easily be solved by using matrix algebra. The solutions are

$$\mu_{11}^{(t)} = \mu_{11}^{(0)}\lambda_1{}^t,$$

$$\mu_{12}^{(t)} = \frac{1}{2}\mu_{11}^{(0)}\lambda_1{}^t + \left\{\mu_{12}^{(0)} - \frac{1}{2}\mu_{11}^{(0)}\right\}\lambda_2{}^t,$$

$$\mu_{21}^{(t)} = \frac{1}{2}\mu_{11}^{(0)}\lambda_1{}^t + \left\{\mu_{21}^{(0)} - \frac{1}{2}\mu_{11}^{(0)}\right\}\lambda_2{}^t,$$

$$\mu_{22}^{(t)} = c_2\mu_{11}^{(0)}\lambda_1{}^t + \frac{1}{3}\{\mu_{12}^{(0)} + \mu_{21}^{(0)} - \mu_{11}^{(0)}\}\lambda_2{}^t + \left[\mu_{22}^{(0)} - c_2\mu_{11}^{(0)} - \frac{1}{3}\{\mu_{12}^{(0)} + \mu_{21}^{(0)} - \mu_{11}^{(0)}\}\right]\lambda_3{}^t.$$

We note that $\mu_{11}^{(0)} = p_m p_n$, $\mu_{12}^{(0)} = p_m p_n{}^2$, $\mu_{21}^{(0)} = p_m{}^2 p_n$, and $\mu_{22}^{(0)} = p_m{}^2 p_n{}^2$ in the present case. Therefore, putting the above moments and $\mu_r$ into (A1) to (A4), we obtain formulas (15) to (18).

## APPENDIX II: VALUE OF $G_{ST}$ WHEN AN ALLELE IS FIXED IN ALL SUBPOPULATIONS

We intend to show that $G_{ST}$ tends to be 0 when the frequency of an allele in the total population approaches 1. For this purpose, we consider a situation where an allele ($A_1$) is just about to be fixed in all subpopulations. That is, we assume that the frequency of this allele is 1 in all subpopulations except one where alleles $A_1$ and $A_2$ exist with frequencies $1 - x$ and $x$, respectively, in which

$x$ is a small quantity. In this case $H_S = 2x(1 - x)/s$ and $H_T = 2(x/s)(1 - x/s)$ from (4a) and (5a), respectively. Therefore,

$$G_{ST} = 1 - \frac{1 - x}{1 - x/s}.$$

Thus, $G_{ST}$ tends to be 0 when $x$ approaches 0.

## ACKNOWLEDGMENTS

We thank Yoshio Tateno for his valuable help in computer programming. This study was supported by grants from the National Institute of Health and the National Science Foundation.

## REFERENCES

BURI, P. 1956. Gene frequency in small populations of mutant *Drosophila*, *Evolution* **10**, 367–402.

EWENS, W. J., AND FELDMAN, M. W. 1976. The theoretical assessment of selective neutrality, *in* "Population Genetics and Ecology" (S. Karlin and E. Nevo, Eds.), pp. 303–337, Academic Press, New York.

JACQUARD, A. 1974. "The Genetic Structure of Populations," Springer-Verlag, New York.

KERR, W. E., AND WRIGHT, S. 1954. Experimental studies of the distribution of gene frequencies in very small populations of *Drosophila melanogaster:* I. Forked, *Evolution* **8**, 172–177.

KIMURA, M. 1955. Solution of a process of random genetic drift with a continuous model, *Proc. Nat. Acad. Sci. USA* **41**, 144–150.

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

MACCLUER, J. W. 1974. Monte Carlo simulation: The effects of migration on some measures of genetic distance, *in* "Genetic Distance" (J. F. Crow and C. Denniston, Eds.), pp. 77–95, Plenum Press, New York.

NEI, M. 1972. Genetic distance between populations. *Amer. Naturalist* **106**, 283–292.

NEI, M. 1973. Analysis of gene diversity in subdivided populations, *Proc. Nat. Acad. Sci. USA* **70**, 3321–3323.

NEI, M. 1975. "Molecular Population Genetics and Evolution," North-Holland, Amsterdam/New York.

NEI, M. 1976. Mathematical models of speciation and genetic distance, *in* "Population Genetics and Ecology" (S. Karlin and E. Nevo, Eds.), pp. 723–765 Academic Press, New York.

NEI, M., CHAKRAVARTI, A., AND TATENO, Y. 1977. Mean and variance of $F_{ST}$ in a finite number of incompletely isolated populations, *Theor. Pop. Biol.* **11**, 291–306.

NEI, M., AND MARUYAMA, T. 1975. Lewontin–Krakauer test for neutral genes, *Genetics* **80**, 395.

NEVO, E. 1973. Test of selection and neutrality in natural populations, *Nature (London)* **244**, 573–575.

ROBERTSON, A. 1952. The effect of inbreeding on the variation due to recessive genes, *Genetics* **37**, 189–207.

ROBERTSON, A. 1975a. Remarks on the Lewontin–Krakauer test, *Genetics* **80**, 396.

ROBERTSON, A. 1975b. Gene frequency distributions as a test of selective neutrality, *Genetics* **81**, 775–785.

SING, C. F., BREWER, G. J., AND THIRTLE, B. 1973. Inherited biochemical variation in *Drosophila melanogaster:* Noise or signal ?, I, Single-locus analyses, *Genetics* **75**, 381–404.

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

WRIGHT, S. 1951. The genetical structure of populations, *Ann. Eugenics* **15**, 323–354.