# A MODEL FOR ANALYSIS OF POPULATION STRUCTURE[1]

EDWARD D. ROTHMAN
*Departments of Statistics and Human Genetics*

CHARLES F. SING
*Department of Human Genetics*

AND

ALAN R. TEMPLETON
*Society of Fellows, University of Michigan, Ann Arbor, Michigan 48104*

Manuscript received November 19, 1973
Revised copy received April 6, 1974

ABSTRACT

Arguments have been presented for the appropriateness of a multinomial Dirichlet distribution for describing single-locus genotypic frequencies in a subdivided population. This distribution is defined as a function of allele frequency, the average (over the entire population) inbreeding coefficient and the correlation between genotypes within a subdivision. Alternative parameterizations and their genetic interpretations are given.—We then show how information from a sample drawn from this subdivided population, in the absence of pedigrees, can be combined with the multinomial Dirichlet model to form a likelihood function. This likelihood function is then used as the basis for estimation and testing hypotheses concerning the genetic parameters of the model. Comparisons of this approach to the alternative procedure of COCKERHAM (1969) and (1973) are made using human data obtained from Tecumseh, Michigan and Monte Carlo simulations.—Finally, implications of these results to statistical inference and to mutation rates are presented.

THE evolution of sexually reproducing organisms is determined in part by the pattern of genetic differentiation among subdivisions of the population. This pattern, in turn, is influenced by the size of the subdivisions, the degree of migration among them and differences in the mode of selection in different parts of the range of the population. Within each subdivision, the mating pattern and differential selection among genotypes contribute to the deviation from Hardy-Weinberg proportions. At any point in time, the effects of the combined operation of these causative forces have been alternatively described by the variance of allele frequencies among subdivisions (WAHLUND 1928), the correlation between alleles or genotypes within or between subdivisions (WRIGHT 1951), and the probability of identity by descent (MALECOT 1948) of alleles in an individual. COCKERHAM (1973) argues that in the absence of pedigrees a realistic treatment of subpopulations must consider the general correlational definitions introduced

[1] This study was supported by the U.S. Atomic Energy Commission, Contract AT(11-1)-1552 to the Department of Human Genetics (CFS), University of Michigan.

by WRIGHT. In this paper we present a theory based on the maximum likelihood principle for making statistical statements about the correlation between alleles within and between individuals and the correlation between genotypes within subdivisions of a population. We have no intent to evaluate the relative roles of alternative forces which may lead to the structure of the population. A likelihood is presented for the distribution of genotype frequencies among subdivisions for a locus with two codominant alleles under the assumptions of no selection, no migration, and subdivision of a population into finite subpopulations at some undetermined time in the past. We will make the necessary assumption that every subdivision contributes equally to the total differentiation of the parental population and that there is no correlation of gene or genotype frequencies between subdivisions.

## THE LIKELIHOOD MODEL

The definitions of means, variances, and covariances computed from a sample coupled with the sampling assumptions inherent in a genetic study are clearly insufficient to completely specify a likelihood. Because the sampling theory approach does not require a likelihood function it has been the method of choice by COCKERHAM (1969, 1973) to estimate correlations between allele frequencies both within individuals and between individuals in a subdivided population. On the other hand, the advantages of using the likelihood principle rather than a sampling approach have been discussed by a number of authors (for a review see EDWARDS 1972). In particular, if the model used to construct a likelihood is a good *approximation* to reality, additional information about the parameters of interest may be extracted from the data. And, although the underlying set of assumptions necessary to derive a likelihood may not be met in practice, experience has shown that a useful distribution need only accurately describe the measurable outcome of the process of interest.

Consider the $h^{\text{th}}$ subpopulation of $N_h$ individuals categorized according to their genotype at a codominant locus with two alleles, $A$ and $a$. Let $T_{h1}$, $T_{h2}$, and $T_{h3}$, respectively, denote the actual number of $AA$, $Aa$, and $aa$ individuals in this subpopulation such that $T_{h1} + T_{h2} + T_{h3} = N_h$. Set the variable

$$Y_i = \begin{cases} 1 & \text{if individual } i \text{ is } AA \\ \frac{1}{2} & \text{if individual } i \text{ is } Aa \\ 0 & \text{if individual } i \text{ is } aa \end{cases}$$

for all $i$, $(i = 1,2, \ldots, N_h)$ of the $h^{\text{th}}$ subpopulation. If the ancestry of each individual is unknown it is reasonable to assume that the joint distribution of the sample of size $n_h$ from deme $h$ (denoted $Y_{i1}, Y_{i2}, \ldots, Y_{in_h}$) is the same as the joint distribution of any other sample from the $h^{th}$ subpopulation (say $Y_{k1}, Y_{k2}, \ldots, Y_{kn_h}$) where $1 \leq n_h \leq N_h$. This implies that the $Y_i$ from subpopulation $h$ are finitely exchangeable. Consequently, (see DEFINETTI 1959) if a sample of

size $n_h$ is drawn from this subpopulation, the probability of the sample may be written

$$Pr[t_{h1}, t_{h2}, t_{h3}] = \sum_{T_{h1}} \sum_{T_{h2}} \sum_{T_{h3}} \frac{\binom{T_{h1}}{t_{h1}}\binom{T_{h2}}{t_{h2}}\binom{T_{h3}}{t_{h3}}}{\binom{N_h}{n_h}} Pr(T_{h1}, T_{h2}, T_{h3}) \qquad (1)$$

where $t_{h1}$, $t_{h2}$, and $t_{h3}$ are the number of $AA$, $Aa$, and $aa$, respectively, where $t_{h1} + t_{h2} + t_{h3} = n_h$, and where the sums are taken over all allowable $T_{hi}$'s in (1). The special case of sampling $n_h$ from $N_n$ without replacement is obtained when $T_{h1}$, $T_{h2}$ and $T_{h3}$ are regarded as constant parameters and not random variables as in (1). In this case (1) reduces to the standard hypergeometric distribution with random variables $t_{h1}$ and $t_{h2}$;

$$Pr(t_{h1}, t_{h2}, t_{h3} | T_{h1}, T_{h2}, T_{h3}) = \frac{\binom{T_{h1}}{t_{h1}}\binom{T_{h2}}{t_{h2}}\binom{T_{h3}}{t_{h3}}}{\binom{N_h}{n_h}} \qquad (2)$$

where $T_{h1} + T_{h2} + T_{h3} = N_h$. In general $T_{h1}$ and $T_{h2}$ are also random variables which reflect the stochastic effects of subdivision of the parental population into subpopulations as well as the random fluctuations of genotype frequencies within a subpopulation over generations since the subdivision.

One of the simplest models that yields a realistic distribution for $T_{h1}$ and $T_{h2}$ involves the following assumptions. First, assume an infinitely large population characterized by the allele frequency for $A$ as $p$. Second, the correlation between alleles within an individual averaged over the entire population is F. The relationship between these two parameters is such that the frequency of $AA$, $Aa$ and $aa$ individuals in this large reference population is

$$\begin{aligned} P_1 &= f(AA) = p^2 + pqF, \\ P_2 &= f(Aa) = 2pq(1 - F), \\ \text{and } P_3 &= f(aa) = q^2 + pqF, \text{ respectively.} \end{aligned}$$

The $h^{\text{th}}$ subdivision is created by drawing $N_h$ individuals at random and independently from one another from this infinitely large population. The resulting distribution of $T_{h1}$ and $T_{h2}$ is the usual multinomial

$$\binom{N_h}{T_{h1}\, T_{h2}\, T_{h3}} P_1^{T_{h1}} P_2^{T_{h2}} P_3^{T_{h3}}, \qquad (3)$$

which is an unrealistic model since, in practice, individuals are not drawn independently from one another to form the $h^{th}$ deme. For example, studies of the South American Indians (NEEL 1967) indicate that villages result from subdivision along familial lines. To construct a likelihood model that incorporates correlation between genotypes within a subdivision we can let the first individual in the $h^{th}$ population be drawn at random such that

$$\begin{aligned} P_{h11} &= p^2 + pqF, \\ P_{h21} &= 2pq(1-F), \\ \text{and } P_{h31} &= q^2 + pqF \end{aligned}$$

for the genotypes $AA$, $Aa$, and $aa$, respectively. In general, for the $i^{th}$ draw the respective probabilities are

$$P_{h1i} = \frac{\rho T_{h1i-1} + (p^2 + pqF)(1-\rho)}{(i-1)\rho + (1-\rho)}$$

$$P_{h2i} = \frac{\rho T_{h2i-1} + (p^2 + (2pq(1-F))(1-\rho)}{(i-1)\rho + (1-\rho)}$$

$$P_{h3i} = 1 - P_{h1i} - P_{h2i} ,$$

given that $i-1$ individuals have already been drawn for the subpopulation and $T_{h1i-1}$ of these were $AA$ while $T_{h2i-1}$ were of the $Aa$ genotype. Under this sampling procedure the correlation between the genotypes of any two individuals in the $h^{th}$ subpopulation is $\rho$. The distribution of $T_{h1}$ and $T_{h2}$ for such a situation takes the form

$$\prod_{K=0}^{T_{h1}-1} \frac{\rho K + P_1(1-\rho)}{K\rho + (1-\rho)} \prod_{L=0}^{T_{h2}-1} \frac{\rho L + P_2(1-\rho)}{L\rho + (1-\rho)} \prod_{M=0}^{N_h - T_{h1} - T_{h2} - 1} \frac{\rho M + P_3(1-\rho)}{M\rho + (1-\rho)} \tag{4}$$

By considering the parameter $\emptyset = (1-\rho)/\rho$, equation (4) can be rewritten as

$$Pr(T_{h1},T_{h2},T_{h3}|P_1,P_2,\emptyset) = \frac{\binom{T_{h1}+P_1\emptyset-1}{T_{h1}}\binom{T_{h2}+P_2\emptyset-1}{T_{h2}}\binom{T_{h3}+P_3\emptyset-1}{T_{h3}}}{\binom{N_h+\emptyset-1}{N_h}} \tag{5}$$

where the corresponding factorials utilized to compute (5) are taken to be the equivalent gamma function. Under this parameterization, $\emptyset$ represents the force of attraction between successive draws of $Y_i$ to form the $h^{\text{th}}$ subpopulation. The multinomial in $P_1$, $P_2$ and $P_3$ may be obtained by letting $\emptyset \to \infty$ (or equivalently $\rho \to 0$) as

$$\lim_{\emptyset \to \infty} \frac{\binom{T_{h1}+P_1\emptyset-1}{T_{h1}}\binom{T_{h2}+P_2\emptyset-1}{T_{h2}}\binom{T_{h3}+P_3\emptyset-1}{T_{h3}}}{\binom{N_h+\emptyset-1}{N_h}} \tag{6}$$

$$= \frac{N_h!}{T_{h1}!T_{h2}!T_{h3}!} (P_1)^{T_{h1}} (P_2)^{T_{h2}} (P_3)^{T_{h3}}$$

in which case successive $Y_i$ for the $h^{\text{th}}$ subpopulation are sampled without correlation and (6) reduces to the multinomial given by (3).

The distribution described by (5) is known as the multinomial Dirichlet (MD) which can be derived by mixing a multinomial with a Dirichlet distribution. (See JOHNSON and KOTZ 1969 for the details on the mixing operation.) This latter fact suggests that an essential difference between the model that yields a multinomial in $T_{h1}$ and $T_{h2}$ and the model that yields a MD is that the probabilities of drawing individuals with certain genotypes are constant for the multi-

nomial but are random variables in the MD case. An alternate way of modeling this process is to assume once again that the distribution of $T_{h1}$ and $T_{h2}$ given $P_1$, $P_2$, and $P_3$ is the multinomial given by (6) but that this distribution is now mixed with the hypergeometric sampling distribution given by (2). We then obtain the multinomial

$$\binom{n_h}{t_{h1}\,t_{h2}\,t_{h3}} P_1^{t_{h1}} P_2^{t_{h2}} P_3^{t_{h3}} \quad . \tag{7}$$

However, here we no longer consider the $P$'s to be constants but instead random variables whose values are determined by a Dirichlet distribution which is influenced by the nature of the subdivision process. A mixture of the multinomial given by (7) with the Dirichlet yields an MD form for equation (1). It is

$$Pr(t_{h1},t_{h2},t_{h3}|P_1,P_2,\phi) = \frac{\binom{t_{h1}+P_1\phi-1}{t_{h1}}\binom{t_{h2}+P_2\phi-1}{t_{h2}}\binom{t_{h3}+P_3\phi-1}{t_{h3}}}{\binom{n_h+\phi-1}{n_h}} \tag{8}$$

where the $P_i$ values now represent the mean of the random variable $P_i$ given in (7). Equation (8) represents the mixture of the three distributions, the hypergeometric sampling distribution, the multinomial for $T_{h1}$, $T_{h2}$ and $T_{h3}$, and the Dirichlet which incorporates the reality of subdivision represented by $\phi$. Equation (8) can also be obtained by substitution of equation (5) into equation (1) directly.

Further insight into the interpretation of the MD model and the parameters of interest using this model can be obtained by considering a more generalized set of conditions which yield the model. Consider a reference population of individuals consisting of $T_{.1}$ $T_{.2}$, and $T_{.3}$ of the $AA$, $Aa$ and $aa$ genotypes, respectively. Suppose this population subdivides into $H$ subpopulations according to the law

$$\begin{aligned} &Pr[T_{11},T_{12},T_{13};T_{21},T_{22},T_{23};\ldots T_{H1},T_{H2},T_{H3} \mid T_{.1},T_{.2},T_{.3}] \\ &= \binom{T_{.1}}{T_{11}} \beta_{11}^{T_{11}} (1-\beta_{11})^{T_{.1}-T_{11}} \binom{T_{.1}-T_{11}}{T_{21}} \beta_{21}^{T_{21}} (1-\beta_{21})^{T_{.1}-T_{11}-T_{21}} \\ &\ldots \binom{T_{.1}-T_{11}-\ldots-T_{H-1,1}}{T_{H,1}} \beta_{H1}^{T_{H1}} (1-\beta_{H1})^{T_{.1}-T_{11}-\ldots-T_{H-1,1}} \\ &\quad \binom{T_{.2}}{T_{12}} \beta_{12}^{T_{12}} (1-\beta_{12})^{T_{.2}-T_{12}} \quad \ldots \text{ etc.} \end{aligned} \tag{9}$$

Different choices of the parameters, $\beta_{hi}$, will give a variety of decompositions of the population which can take into account both the random and nonrandom association of genotypes within a subpopulation as well as associations between genotypes in different subpopulations. An equivalent form of equation (9) is

$$\frac{T_{.1}!}{T_{11}!\,T_{21}!\ldots T_{H1}!} \Delta_{11}^{T_{11}} \Delta_{21}^{T_{21}} \ldots \Delta_{H-1,1}^{T_{H-1,1}} \left(1-\sum_{h=1}^{H} \Delta_{h1}\right)^{T_{H1}}$$

$$\frac{T_{\cdot 2}!}{T_{12}!\,T_{22}!\ldots T_{H2}!}\,\Delta_{12}^{T_{12}}\Delta_{22}^{T_{22}}\ldots\Delta_{H-1,2}^{T_{H-1,2}}\Big(1-\sum_{h=1}^{H}\Delta_{h2}\Big)^{T_{H2}}$$

$$\frac{T_{\cdot 3}!}{T_{13}!\,T_{23}!\ldots T_{H3}!}\,\Delta_{13}^{T_{13}}\Delta_{23}^{T_{23}}\ldots\Delta_{H-1,3}^{T_{H-1,3}}\Big(1-\sum_{h=1}^{H}\Delta_{h3}\Big)^{T_{H3}} \tag{10}$$

where

$$\Delta_{hi}=\beta_{h,i}(1-\beta_{h-1,i})(1-\beta_{h-2,i})\ldots(1-\beta_{H-1,i}), \tag{11}$$
$$h=1,2,\ldots,H\text{-}1,\ i=1,2,3,$$
$$\text{and } \beta_{h,i}=0 \text{ for } h<1.$$

In practice, $T_{\cdot 1}$, $T_{\cdot 2}$, and $T_{\cdot 3}$ are random variables observed at a fixed time, $t$. We take $T_{\cdot i}$ to have independent Poisson distributions over time. This is equivalent to assuming that

$$Pr(T_{\cdot 1},T_{\cdot 2},T_{\cdot 3}\mid N=T_{\cdot 1}+T_{\cdot 2}+T_{\cdot 3})$$

is a multinomial distribution. Then the joint distribution of the genotypes in the $H$ subpopulations may be written

$$Pr(T_{11},T_{12},T_{13};T_{21},T_{22},T_{23};\ldots;T_{H1},T_{H2},T_{H3})$$
$$=\frac{(\lambda_1\Delta_{11})^{T_{11}}e^{-\lambda_1\Delta_{11}}}{T_{11}!}\cdots\frac{(\lambda_1\Delta_{H1})^{T_{H1}}e^{-\lambda_1\Delta_{H1}}}{T_{H1}!}$$
$$\frac{(\lambda_2\Delta_{12})^{T_{12}}e^{-\lambda_2\Delta_{12}}}{T_{12}!}\cdots\frac{(\lambda_2\Delta_{H2})^{T_{H2}}e^{-\lambda_2\Delta_{H2}}}{T_{H2}!}$$
$$\frac{(\lambda_3\Delta_{13})^{T_{13}}e^{-\lambda_3\Delta_{13}}}{T_{13}!}\cdots\frac{(\lambda_3\Delta_{H3})^{T_{H3}}e^{-\lambda_3\Delta_{H3}}}{T_{H3}!}. \tag{12}$$

From this basic distribution the multinomial distribution of $T_{hi}$ into the $H$ subpopulations is obtained by setting

$$\Delta_{1j}=\Delta_{2j}=\Delta_{3j}=\ldots=\Delta_{Hj}=\Delta_j$$

for each $j=1,2,3$ and conditioning on subpopulation size. It is

$$\prod_{h=1}^{H}\frac{N_h!}{T_{h1}!T_{h2}!T_{h3}!}\,\Delta_1^{T_{h1}}\Delta_2^{T_{h2}}\Delta_3^{T_{h3}}. \tag{13}$$

The basic equation (12) takes the form of the MD when either of two assumptions are true. They are (1) the $\Delta_{hi}$; $h=1,2,\ldots,H$ are equal for each $i$ but the genotypes are distributed according to a Dirichlet or (2) $T_{\cdot 1}$, $T_{\cdot 2}$, and $T_{\cdot 3}$ have a negative binomial distribution rather than a Poisson. As previously shown, the parameter $\phi$ of the resulting MD may be interpreted as the attraction of like genotypes into the same subpopulation in the splitting process.

Regardless of the specific process which may give rise to the MD, equation (8) represents a convenient realistic likelihood. First, there is sufficient flexibility in

the parameters of the likelihood to deal with a wide spectrum of situations. Second, the MD likelihood function is consistent with the correlational definitions given by WRIGHT and used by COCKERHAM. To see this, recall that there are three main parameters used in the likelihood model; $P_1$, $P_2$ and $\phi$ or equivalently $p$, $F$ and $\rho$ where

$p$ = $P(A)$ is the frequency of $A$ in the total population,
$F$ = the correlation between alleles in the same individual averaged over the entire population,
$P_1 = p^2 + pqF$,
$P_2 = 2pq(1-F)$,
$P_3 = q^2 + pqF$,
$P_1 + P_2 + P_3 = 1$,
$\rho$ = the correlation between $Y_i$ and $Y_j$, $i \neq j$ irrespective of the subpopulation from which $i$ and $j$ are drawn,
and $\phi = (1-\rho)/\rho$.

We note that, for the special case $Y_i = 1$ if individual is $AA$, ½ if individual is $Aa$, or 0 if individual is $aa$, then $\rho$ in equation (4) is equivalent to WRIGHT's coefficient of relationship and may be written as

$$\frac{2\Theta}{1+F}$$

where $\Theta$ is taken to be the correlation between pairs of alleles drawn from different individuals within a subpopulation irrespective of the subpopulation being considered. As with $F$, $\theta$ is taken to be the expectation over all subpopulations of the population. There are $2N_h$ $(N_h-1)$ such pairs contributing to each $\theta_h$.

For any subpopulation the expected relative frequency of the $AA$ genotype is

$$E[T_{h1}/N_h] = p^2 + pqF$$

and of the $Aa$ genotype is

$$E[T_{h2}/N_h] = 2pq(1-F) .$$

Therefore, the expected value of the gene count estimator of the frequency of the $A$ allele is

$$E[\hat{p}_h] = E\left[\frac{2T_{h1} + T_{h2}}{2N_h}\right] = p .$$

The variance of the estimate is

$$V(\hat{p}_h) = p(1-p)\left[\frac{1+F}{2N_h} + \frac{N_h - 1}{N_h}\Theta\right] \qquad (14)$$

where $F$ and $\Theta$ are the average values defined above.

This result is analogous to that presented for the variance of allele frequencies by COCKERHAM (1969, 1973). Our analysis of $F$ and $\Theta$ allows a somewhat dif-

ferent interpretation than COCKERHAM's of the dispersion of genotype frequencies among subdivisions. The parameter $\rho$ may be defined as either a correlation between $Y_i$, $Y_j$; $i \neq j$, or as a function of difference in variance between the MD and the multinomial. It is, for any $h$, $h = 1, .. H$,

$$\rho_0 = \frac{V(T_{h1}) + V(T_{h2}) + V(T_{h3}) - N_h(P_1Q_1 + P_2Q_2 + P_3Q_3)}{(N_h - 1)\, N_h(P_1Q_1 + P_2Q_2 + P_3Q_3)} \tag{15}$$

where $Q_i = (1-P_i)$.

In equation (15), $\rho_0$ represents an alternate interpretation of the correlation parameter as a function of the subpopulation size and the genotypic variances. This reflects the fact that the genotypes in a subdivision may be correlated due to both population structure *per se* and finiteness of the subpopulations making up the population. Consequently a difficulty arises in interpreting possible differences in $\rho$ as due to differences in population structure or due to a difference in the finite size of the subpopulations in the populations being compared. That is, the size of the subpopulations and population structure are confounded in the genotypic correlation measure $\rho$. However, if we require, as seems natural, that $V(\hat{p}_h)$ of equation (14) is of the order $N_h^{-1}$, then $\rho$ and $\theta$ must be of the same order. Hence an approximate index for comparing the structure of two population which differ in the size of their subpopulations, say $N_1$ for the first population and $N_2$ for the second, should be based on a function which measures the discrepancy between $N_1\rho_1 + (1 - \rho_1)$ and $N_2\rho_2 + (1 - \rho_2)$. That is, one might compute an estimate of $S_j$, $j = 1, 2$

$$S_j = \frac{V(T_{h1}) + V(T_{h2}) + V(T_{h3})}{N_h\,(P_{h1}Q_{h1} + P_{h2}Q_{h2} + P_{h3}Q_{h3})}$$

where $P_{h1}$, $P_{h2}$ and $P_{h3}$ are the frequencies of $AA$, $Aa$ and $aa$ respectively, in the $h^{\text{th}}$ subpopulation of the $j^{\text{th}}$ population with subpopulations of size $N_h$. Then the ratio

$$S_1/S_2$$

gives a measure of relative structure corrected for any effect due solely to subpopulation size. For example, suppose two populations having subpopulations of size $N_h = 100$ and $N_h = 200$, respectively, are found to have $\rho_1 = .02$ and $\rho_2 = .01$, respectively. The question naturally arises whether this difference in $\rho$ between populations is due to the fact that one population consists of subpopulations twice the size of the other or is due to some difference in population structure other than subpopulation size. A ratio $S_1/S_2$ of approximately 1 indicates that both populations have the same level of coancestry due to population structure as measured by $\rho$, and the observed difference, $\rho_1 - \rho_2$, may be attributable to a difference between the populations in the size of their subpopulations.

## STATISTICAL INFERENCE; ESTIMATION AND TESTING

We will consider the application of the MD given by equation (8) to genotype data obtained at one point in time from $H$ subpopulations with no available in-

formation on ancestry or coancestry. The likelihood of a sample representing $H$ subpopulations in terms of the three parameters $P_1$, $P_2$, and $\rho$ is proportional to

$$\prod_{h=1}^{H} \prod_{i=1}^{3} \frac{\binom{t_{hi} + P_i \frac{(1-\rho)}{\rho} - 1}{t_{hi}}}{\binom{n_h + \frac{1-\rho}{\rho} - 1}{n_h}} \tag{16}$$

Ignoring the constant term, the log-likelihood may be shown to equal

$$l(P_1, P_2, \rho) = \sum_{h=1}^{H} \sum_{i=1}^{3} \sum_{L=0}^{t_{hi}-1} \ln[(1-\rho)P_i + \rho L] - \sum_{h=1}^{H} \sum_{L=0}^{n_h} \ln[1-\rho) + \rho L].$$

Recall that $P_i$ are functions of allele frequencies and $F$ while $\rho$ is a function of the allele correlations $F$ and $\Theta$. The estimators of these parameters for this nonlinear form of the MD have not been extensively studied. MOSIMANN (1962) suggests a method of moments type estimator for the frequency parameters, $P_i$, By solving the relationship

$$\hat{P}_i \left(\frac{1-\hat{\rho}}{\hat{\rho}}\right) = \frac{1}{H} \sum_{h=1}^{H} \left(\frac{t_{hi}}{n_h}\right) \quad , \; i = 1,2,3$$

one obtains estimators of $P_i$. Then by equating the determinant of the sample variance-covariance matrix to the determinant of the theoretical variance-covariance matrix (evaluated at $\hat{P}_i$) an estimator of $\rho$ is found. The properties of these estimates have not been studied in general but it can be shown that in the practical situation where $H$ is small and the $n_h$ are large, inconsistent estimators result. The corresponding estimate of the allele frequency. $p$, is the familiar gene counting form

$$\hat{p} = \frac{1}{H} \sum_{h=1}^{H} \left(\frac{2t_{h1} + t_{h2}}{2n_h}\right) .$$

Although this estimate is unbiased it will have large variance

$$V(\hat{p}) = \frac{1}{H} V \left(\frac{2t_{h1} + t_{h2}}{2n_h}\right)$$

unless $H \to \infty$. Consistency of the estimate (mean square sense) is obtained only when the component of correlation between pairs of alleles drawn from different individuals but the same subpopulation, $\Theta$, (and hence $\rho = 2\Theta/1 + F$) is zero.

We now turn to the maximum likelihood estimation procedure. Several alternate methods of evaluating the maximum likelihood estimators are available. One can begin by searching the region of the space suggested by the method of moments. From there a scoring technique (RAO 1965) utilizing the asymptotic variance-covariance matrix based on the inverse of the information matrix evaluated at the estimated values can be used to expedite the search of the surface. In the absence of information about the shape of the entire surface it may be neces-

sary to utilize some type of grid search technique to be certain a secondary maximum was not obtained because of saddlepoints or other inconsistencies of the surface. The matrix of second derivatives is

$$\begin{matrix} \frac{\partial^2 l}{\partial P_1^2} & \frac{\partial^2 l}{\partial P_1 \partial P_2} & \frac{\partial^2 l}{\partial P_1 \partial \rho} \\ \frac{\partial^2 l}{\partial P_1 \partial P_2} & \frac{\partial^2 l}{\partial P_2^2} & \frac{\partial^2 l}{\partial P_2 \partial \rho} \\ \frac{\partial^2 l}{\partial P_1 \partial \rho} & \frac{\partial^2 l}{\partial P_2 \partial \rho} & \frac{\partial^2 l}{\partial \rho^2} \end{matrix}$$

where

$$\frac{\partial^2 l}{\partial P_i^2} = -(1-\rho)^2 \sum_{h=1}^{H} \left\{ \sum_{L=0}^{t_{hi}-1} [(1-\rho)P_i+\rho L]^{-2} + \sum_{L=0}^{t_{h3}-1} [(1-\rho)P_3+\rho L]^{-2} \right\} \quad i = 1,2$$

$$\frac{\delta l}{\delta P_1 \delta P_2} = -(1-\rho)^2 \sum_{h=1}^{H} \sum_{L=0}^{t_{h3}-1} [(1-\rho)P_3+\rho L]^{-2}$$

$$\frac{\partial^2 l}{\partial \rho^2} = \sum_{h=1}^{H} \left\{ \sum_{L=0}^{n_h} \left[ \frac{L-1}{(1-\rho)+\rho L} \right]^2 - \sum_{i=1}^{3} \sum_{L=0}^{t_{hi}-1} \left[ \frac{L-P_i}{(1-\rho)P_i+\rho L} \right]^2 \right\}$$

$$\frac{\partial^2 l}{\partial P_i \partial \rho} = (1-\rho) \sum_{h=1}^{H} \left\{ \sum_{L=0}^{t_{h3}-1} \frac{L-P_3}{[(1-\rho)P_3+\rho L]^2} - \sum_{L=1}^{t_{hi}-1} \frac{L-P_i}{[(1-\rho)P_i+\rho L]^2} \right\}$$

$$+ \sum_{h=1}^{H} \left\{ \sum_{L=0}^{t_{h3}-1} [(1-\rho)P_3+\rho L]^{-1} - \sum_{L=1}^{t_{hi}-1} [(1-\rho)P_i+\rho L]^{-1} \right\} \quad i = 1,2 .$$

The second derivatives converge with probability one to their expected values (KENDALL and STUART 1973, p. 43); consequently the negative of the second derivative matrix converges to the information matrix. By evaluating this matrix at the maximum likelihood estimates of $P_1$, $P_2$ and $\rho$ and inverting the resulting matrix, it is possible to estimate the variance-covariance matrix for the maximum likelihood estimates of $\rho$, $P_1$ and $P_2$. Also, if one desires, the likelihood function can be parameterized in terms of $p$, $F$ and $\Theta$ (or $p$, $F$ and $\rho$) and estimates found for these parameters directly.

One can also put restrictions on the parameters when estimating them and then test the validity of these restrictions. For example, consider the hypothesis $F = 0$. The maximum likelihood estimators of $p$, the gene frequency, and $\rho$ can be found in that subset of the parameter space in which $F = 0$ utilizing either a grid search or a scoring technique. Let these estimators be designated by $\hat{p}'$ and $\hat{\rho}'$, which in general are not equal to the corresponding estimates when $F \neq 0$. Then the quantity

$$2[l(\hat{P}_1,\hat{P}_2,\hat{\rho}) - l(\hat{p}',\hat{\rho}')]$$

is asymptotically distributed as a chi-square with one degree of freedom (RAO 1965). This log likelihood ratio thus provides a test of the hypothesis $F = 0$.

Similarly, one could estimate $p$, the gene frequency (or $P_1 = p^2$) under the hypothesis that $F = 0$ and $\rho = 0$. Letting $\hat{p}''$ designate the gene frequency estimate under this hypothesis, the test criterion for $F = 0$ and $\rho = 0$ (which in our parameterization is equivalent to testing for Hardy-Weinberg equilibrium) is

$$2[l(\hat{P}_1,\hat{P}_2,\hat{\rho}) - l(\hat{p}'')]$$

which is asymptotically distributed as a chi-square with two degrees of freedom.

A similar log-likelihood ratio test could be performed for the hypothesis $\rho = 0$, but as ROTHMAN and WOODROOFE (1973) show, a simple $\chi^2$ test will be quite efficient. Specifically, one computes

$$X^2 = \sum_{h=1}^{H} \sum_{i=1}^{3} \frac{(t_{hi} - n_h \hat{p}_i)^2}{n_h \hat{p}_i}$$

where $\hat{p}_i$ is $\Sigma_{h=1}^{H} t_{hi}/\Sigma_{h=1}^{H} n_h$ and compares this statistic with a $\chi^2$ with $2H-2$ degrees of freedom.

## APPLICATIONS: COMPARISON WITH COCKERHAM'S APPROACH

A comparison of the likelihood approach developed above with the least-squares approach of COCKERHAM (1969) was made in two ways. First, the six largest kindreds defined by SING, CHAMBERLAIN and EGGLESTON (1973) from the Tecumseh Community Health Study were chosen for convenience to represent a random sample of subdivisions. Of course the assumptions of our underlying model do not strictly hold, but the data serve to illustrate the application of the model. Data from codominant loci, each with two alleles, were available on individuals from each kindred. The data set is presented in Table 1. Second, a computer simulation was employed to generate six independent samples of data from a known reference population defined by specific parameter values of $N_h$, $P_1$, $P_2$ and $\rho$. The six samples (subpopulations) of equal size were then subjected to the two analyses. This experiment (and the analyses) was replicated 100 times to obtain the distribution properties of estimates for each of six combinations of parameters. Those six were three values of $N_h$; 20, 50, and 100, each combined with two values of $\rho$; 0.10 and 0.25. The same $P_1$ and $P_2$ were considered in all cases. Each of the six samples of a replication was generated by utilizing a random number generator and a fixed relationship between the parameters $P_1$, $P_2$, and $\rho$. The decision regarding the genotype of the $i^{\text{th}}$ individual was made by generating a random number from a uniform distribution between zero and one and comparing it with the probabilities determined by the parameters under consideration. If $r$ is the random number then the $i^{\text{th}}$ genotype of the sample $i = 1 \ldots N_h$ is assigned according to the rule

$$\begin{aligned} &AA \text{ if } 0 \leq r \leq P_{h1i} \\ &Aa \text{ if } P_{h1i} < r < P_{h1i} + P_{h2i}, \text{ and} \\ &aa \text{ if } P_{h1i} + P_{h2i} < r \leq 1 \end{aligned}$$

TABLE 1

*Data on five codominant loci from six large kindreds available from Tecumseh Community Health Study*

| Phenotype | Kindred 1 | 2 | 3 | 4 | 5 | 6 | Relative frequency |
|---|---|---|---|---|---|---|---|
| MM | 22 | 15 | 6 | 7 | 17 | 21 | .2268 |
| MN | 41 | 40 | 31 | 39 | 27 | 23 | .5180 |
| NN | 34 | 30 | 9 | 10 | 6 | 10 | .2551 |
| N | 97 | 85 | 46 | 56 | 50 | 54 | |
| SS | 4 | 1 | 16 | 2 | 7 | 14 | .1083 |
| Ss | 29 | 23 | 25 | 30 | 29 | 26 | .3990 |
| ss | 65 | 63 | 14 | 25 | 18 | 15 | .4926 |
| N | 98 | 87 | 55 | 57 | 54 | 55 | |
| Hp1–1 | 23 | 11 | 9 | 6 | 18 | 8 | .1847 |
| Hp2–1 | 48 | 38 | 26 | 32 | 29 | 28 | .4950 |
| Hp2–2 | 27 | 38 | 20 | 19 | 7 | 19 | .3201 |
| N | 98 | 87 | 55 | 57 | 54 | 55 | |
| Gc1–1 | 54 | 42 | 41 | 40 | 26 | 24 | .5563 |
| Gc2–1 | 37 | 36 | 14 | 17 | 25 | 24 | .3750 |
| Gc2–2 | 7 | 9 | 0 | 0 | 5 | 7 | .0686 |
| N | 98 | 87 | 55 | 57 | 56 | 55 | |
| Gm(a+) | 18 | 15 | 4 | 0 | 8 | 1 | .1185 |
| Gm(a+b+) | 46 | 38 | 13 | 19 | 26 | 24 | .4278 |
| Gm(b+) | 33 | 32 | 29 | 37 | 16 | 29 | .4536 |
| N | 97 | 85 | 46 | 56 | 50 | 54 | |

where $P_{h1i}$, $P_{h2i}$ and $P_{h3i}$ are the relative probabilities of the genotypes after $i$–1 individuals have been assembled for the $h^{\text{th}}$ subdivision. That is, for the first decision in the $h^{\text{th}}$ subdivision

$$P_{h1i} = p^2 + pqF,$$
$$P_{h2i} = 2pq(1-F), \text{ and}$$
$$P_{h3i} = q^2 + pqF.$$

For the $i^{\text{th}}$ decision of the same subdivision

$$P_{h1i} = \frac{\rho T_{h1i-1} + P_{h11}(1-\rho)}{(i-1)\rho + (1-\rho)}$$

$$P_{h2i} = \frac{\rho T_{h2i-1} + P_{h21}(1-\rho)}{(i-1)\rho + (1-\rho)}, \text{ and}$$

$$P_{h3i} = 1 - P_{h2i} - P_{h3i} \text{ where}$$

$T_{h1i-1}$ and $T_{h2i-1}$ are the number of $AA$ and $Aa$ after $i$–1 decisions have been made. The $P_{h11}$, $P_{h21}$ and $P_{h31}$ values for the simulation of all samples, h = 1. . . H, and every replication, were taken to be .18, .50, and .32, respectively.

Table 2 gives the comparison of the MLE and LS procedures using the six largest kindreds from the Tecumseh study. The parameters $P_1$, $P_2$, and $\rho$ were estimated by maximum likelihood using equation (12). The least-squares (L.S.)

TABLE 2

*A comparison of estimation procedures using the six largest kindreds in the Tecumseh, Michigan data set*

| Parameter | Estimation procedure | *MN* | *Ss* | Locus Haptoglobin | *Gc* | *Gm* | Mean |
|---|---|---|---|---|---|---|---|
| $\rho$ | MLE | .031 | 0.86 | .016 | .028 | .051 | .042 |
| $P_1$ | MLE | .229 | .117 | .184 | .561 | .099 | |
| $P_2$ | MLE | .533 | .436 | .501 | .377 | .429 | |
| | | (—392.62065)* | (—362.88879) | (—414.18962) | (—361.13392) | (—372.5727) | |
| $\rho$ | L.S. | .042 | .203 | .051 | .060 | .093 | .090 |
| $P_1$ | L.S. | .227 | .108 | .187 | .556 | .118 | |
| $P_2$ | L.S. | .518 | .399 | .495 | .375 | .428 | |
| | | (—392.83585)† | (—365.45464) | (—415.62207) | (—361.58550) | (—373.12872) | |
| $\rho_0$ | See text Equation 15 | .030 | .096 | .010 | .015 | .031 | .036 |

* The log-likelihood for the MLE estimates.
† The log-likelihood for the LS estimate.

procedure is that developed by COCKERHAM (1969) using the weighting method for unequal subdivision size given in his 1973 paper. The L.S. analysis is based on correlations among alleles, whereas in Table 2 we are concerned primarily with parameters which define genotype frequencies. COCKERHAM (1969) gives the (unbiased) estimators of $P_1$, $P_2$, and $\rho$ to be

$$P_2 = T_{.2}/N, \text{ and }.$$
$$p_1 = T_{.1}/N,$$
$$P_2 = T_{.2}/N, \text{ and },$$
$$\rho = \frac{2\hat{\Theta}}{1+\hat{F}}$$

where $\hat{\Theta}$ and $\hat{F}$ are the least-squares estimates of the parameters defining correlations between pairs of alleles ignoring subdivision, $F$, and the correlation, $\Theta$, between pairs of alleles within subdivisions averaged over all subdivisions. The assumption is made that the relationship between $\rho$, $\Theta$, and $F$ is affected by sampling only and that all subdivisions are homogeneous with respect to this relationship. For all loci considered the L.S. estimate of $\rho$ is greater than the M.L.E. estimate. On the average, for $\rho$, the least-squares estimator is double the M.L. estimator. Taking twice the standard error of the M.L. efficient variance to estimate the 95% confidence interval for the ML estimate we see that the L.S. estimates of $\rho$ from the *Ss* and Haptoglobin markers appear to be substantially different, whereas the estimates from the other three markers are not. However, no significance can be attached to this difference since a comparison is based on an estimate of the M.L. variance and not that applicable for least-squares estimates. Estimates of $P_1$ and $P_2$ are remarkably similar for all loci. The estimates based on the difference in variance given by $\rho_0$, equation 15, are similar to the M.L. estimates. On the average $\rho_0$ is smaller than $\hat{\rho}_{\text{MLE}}$.

A more meaningful comparison between the L.S. estimator and the M.L.E. is best accomplished on the likelihood axis rather than the parameter axes. Large differences between the two estimates may in fact correspond to a ratio of the likelihood evaluated at the L.S. estimate to the likelihood evaluated at the M.L.E. which is close to 1. This would indicate a rather flat likelihood surface in a neighborhood of the M.L.E. containing the L.S. estimate. This appears to be the case for all loci considered (see Table 2). The implication of this for testing hypotheses about the parameter values is clear. Little loss of power would result if the L.S. estimates are used in place of the M.L.E. in a likelihood ratio test. The loss of power may, however, be considerable with other data sets.

The results of the application of the M.L.E. and L.S. procedures to the output of the computer simulations are given in Table 3. As expected the M.L. estimates are biased more than the corresponding L.S. values. The effect of increased subdivision size to reduce the percent bias is a function of the parameter $\rho$. Regardless of the estimator, for the smaller value of true $\rho$ the bias is affected little by the increase in $N_h$ from 20 to 100. The data suggest that the larger the value of true $\rho$, the greater the *reduction* in bias due to consideration of larger subdivis-

TABLE 3

*Summary statistics on $\hat{\rho}$ computed according to three procedures for four combinations of parameters each replicated 100 times*

| Estimator $N_h$ | Mean | | | Percent bias* | | | Variance × $10^3$ | | | Coefficient of variation | | | Mean squared loss | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 |
| | | | | | | | True $\rho = .1$ | | | | | | | | |
| $\rho_0$ | .081 | 082 | .088 | 19.2 | 18.2 | 11.9 | 2.98 | 1.67 | 1.68 | .68 | .50 | .47 | 0.0033 | 0.0026 | 0.0003 |
| $\rho_{MLE}$ | .084 | .085 | .086 | 16.2 | 15.4 | 14.1 | 3.15 | 1.86 | 1.200 | .67 | .51 | .40 | 0.0034 | 0.0021 | 0.0002 |
| $\rho_{LS}$ | .108 | .093 | .108 | 8.2 | 7.4 | 7.9 | 7.41 | 3.83 | 3.24 | .80 | .67 | .53 | 0.0078 | 0.0038 | 0.0021 |
| | | | | | | | True $\rho = .25$ | | | | | | | | |
| $\rho_0$ | .193 | .217 | .214 | 22.8 | 12.2 | 14.5 | 7.95 | 8.77 | 6.01 | .46 | .43 | .36 | 0.0112 | 0.0099 | 0.0073 |
| $\rho_{MLE}$ | .196 | .217 | .212 | 21.8 | 13.3 | 15.1 | 6.27 | 6.96 | 5.10 | .40 | .38 | .34 | 0.0092 | 0.0081 | 0.0065 |
| $\rho_{LS}$ | .221 | .234 | .256 | 11.4 | 6.4 | 2.3 | 16.66 | 14.94 | 12.70 | .58 | .52 | .44 | 0.0175 | 0.0157 | 0.0127 |

* $\dfrac{\hat{\rho}-\text{True } \rho}{\text{True}}$

ions. For the values of $\rho$ expected in real sexually reproducing populations, say $\rho < 0.1$, it appears that each of the three estimation procedures err badly for small subdivision sizes.

The dispersion of estimates yields a somewhat different conclusion. The effect of varying subdivision size is greatest for the smaller value of true $\rho$. The coefficient of variation reflects a decrease in relative variance for all subdivision sizes as true $\rho$ increases. Regardless of the subdivision size or value of true $\rho$ considered, the variances of least squares estimates are more than double the corresponding variances of M.L. estimates (and $\hat{\rho}_0$ values).

The combined influences of bias and variance are reflected in Table 3 by the mean square loss measure. By this criterion the M.L. and $\rho_0$ estimators are equally good and 2 to 20 times better than the least-squares estimates.

## FURTHER APPLICATIONS

In this section we consider implications of the model to statistical inference. Consider a subpopulation of individuals characterized by their genotype at a given locus. To test whether this subpopulation is in Hardy-Weinberg equilibrium a sample of size $n$ is drawn and a chi-square test (or some modified version as proposed by CANNINGS and EDWARDS 1969) would be used. Because of coancestry or finite size of the subpopulation the variance of the genotypic frequencies will be greater than one would expect if individuals were unrelated. Thus the underlying distribution of the sample may be of the form of a multinomial Dirichlet rather than the usually assumed multinomial. Specifically, suppose we observe $t_1$, $t_2$ and $n - t_1 - t_2$ of genotypes $AA$, $Aa$ and $aa$ respectively. If the subpopulation is in Hardy-Weinberg equilibrium then, whether or not individuals are related,

$$E(t_1) = np^2$$

$$E(t_2)\ 2npq$$

$$E(n-t_2 - t_2) = nq^2$$

where $p$ denotes the frequency of the $A$ allele and $q = 1 - p$ is the frequency of the $a$ allele in the subpopulation. However, since individuals are related it may be shown that

$$\mathrm{Var}(t_1) > np^2(1-p^2)$$
$$\mathrm{Var}(t_2) > 2npq(1-2pq)$$

This result shows that the level of the chi-square test for goodness-of-fit will be affected. In particular the probability of a type 1 error will be larger. Quantification of this effect is discussed in ROTHMAN and WOODROOFE (1973). Clearly the impact on the level of the test will be greatest in highly structured populations. Recognition of this effect has been reported previously by GERSHOWITZ *et al.* (1967) in studies of the Xavante Indians.

Perhaps a more important application along the same line is found in testing for association between some observed pathology and a genetic characteristic.

When pedigree data are unavailable a chi-square test may be used to examine the hypothesis of independence. Here, as in the first application, the level of the test will be affected in a positive manner if $\rho > 0$. Thus one would reject the hypothesis of independence with higher probability when in fact the pathology may be unrelated to this particular genetic characteristic.

We next consider the implications of population structure on estimates of mutation rates. One approach to estimation of a mutation rate in a human population, proposed by NEEL (1973), depends on the expected time to extinction of a mutant allele (given that it goes to extinction). NEEL's results concerning the expected time to extinction based on data from several villages of Yanomama Indians and a simulation study by LI and NEEL (1973) are in disagreement with the theory of KIMURA and OHTA (1969). One reason (others have been proposed) for the lack of agreement may be that KIMURA's result depends on binomial variation of gene frequency in the diffusion equation of the form

$$D_1 = \frac{x(1-x)}{2N_e}$$

where $N_e$ is the effective population size. Note that $D_1$ is on the order of $N_e^{-1}$, which is a necessary assumption for the applicability of the diffusion equations commonly used in genetics. When, in fact, there is population structure, $D$ would take the form of equation (14), that is

$$D_2 = x(1-x)\left(\frac{1+F}{2N} + \frac{N-1}{N}\bar{\theta}\right)$$

where $N$ is the population size. If $F$ is non-zero, $D_2$ is on the order of $N^{-1}$, and consequently a non-zero $F$ can be easily accounted for by an effect of effective population size. However, when $\bar{\theta}$ is also non-zero, $D_2$ is no longer on the order of $N^{-1}$ unless it is further assumed that $\bar{\theta}$ is on the order of $N^{-1}$. When biological conditions are such that this assumption about the magnitude of $\bar{\theta}$ cannot be made, conclusions based on $D_1$ using an $N_e$ would be inappropriate.

## LITERATURE CITED

CANNINGS, E. and A. W. F. EDWARDS, 1969 Expected genotypic frequencies in a small sample: Deviations from Hardy-Weinberg equilibrium. Am. J. Hum. Genet. **21**: 245–247.

COCKERHAM, C. C., 1969 Variance of gene frequencies. Evolution **23**: 72–84. ——, 1973 Analyses of gene frequencies. Genetics **74**: 679–700.

EDWARDS, A. W. F., 1972 *Likelihood.* University Press, Cambridge.

DE FINETTI, B., 1959 La probabilita e la statistica nei rapporti con L'induzione, secondo i diversi punti di vista, *Centro Internazione Matematico Estivo* (C.I.M.E.) Cremonese, Rome.

GERSHOWITZ, H., P. C. JUNQUEIRA, F. M. SALZANO and J. V. NEEL, 1967 Further studies on the Xavante Indians. III. Blood groups and ABH-Le$^a$ secretor types in the Simões Lopes and São Marcos Xavantes. Am. J. Hum. Genet. **19**: 502–513.

JOHNSON, N. L. and S. KOTZ, 1969 *Discrete Distributions.* John Wiley and Sons, Inc., Somerset, N. J.

KENDALL, M. G. and A. STUART, 1973 *The Advanced Theory of Statistics*, Vol. II. V. Hafner, New York.

KIMURA, M. and T. OHTA, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61**: 763–771.

LI, F. H. and J. V. NEEL, 1973 A simulation of the fate of a mutant gene of neutral selective value in a primitive population. In: *Computer Simulation on Human Population Studies.* Edited by B. DYKE and J. MACCLUER. Seminar Press, New York.

MALÉCOT, G., 1948 *Les Mathématiques de l'Hérédité*. Masson, Paris.

MOSIMANN, J. E., 1962 On the compound multinomial distribution, the multivariate beta distribution, and correlations among proportions. Biometrika **49**: 65–82.

NEEL, J. V., 1967 The genetic structure of primitive human populations. Japan J. Hum. Genet. **12**: 1–16. ——, 1973 Private genetic variants and the frequency of mutation. Proc. Nat. Acad. Sci. U.S. (In press.)

RAO, C. R., 1965 *Linear Statistical Inference and Its Applications.* Wiley and Sons, New York.

ROTHMAN, E. and M. WOODROOFE, 1973 Test of Co-ancestry. Tech. Report No. 30, Department of Statistics, Univ. of Michigan, Ann Arbor.

SING, C. F., M. A. CHAMBERLAIN and B. K. EGGLESTON, 1973 An analysis of variance of gene frequencies in a human population. In: *Human Population Structure.* Edited by N. MORTON. Univ. of Hawaii Press, Honolulu.

WAHLUND, S., 1928 Zusammensetzung non Populationen und Korrelationsercheinungen non Standpunkt der Vererbungslehre aus betrachtet. Hereditas **11**: 65–106.

WRIGHT, S., 1951 The genetic structure of populations. Ann. Eugen. **15**: 322–354.

Corresponding editor: R. C. LEWONTIN