DEVIATIONS FROM HARDY-WEINBERG PROPORTIONS: SAMPLING VARIANCES AND USE IN ESTIMATION OF INBREEDING COEFFICIENTS

ALAN ROBERTSON AND WILLIAM G. HILL

Institute of Animal Genetics, University of Edinburgh, Edinburgh EH9 3JN, Scotland

Manuscript received December 6, 1983
Revised copy accepted April 23, 1984

ABSTRACT

An analysis is made of the distribution of deviations from Hardy-Weinberg proportions with k alleles and of estimates of inbreeding coefficients (f) obtained from these deviations.——If f is small, the best estimate of f in large samples is shown to be $2\sum_i(T_{ii}/N_i)/(k - 1)$, where T_{ii} is an unbiased measure of the excess of the ith homozygote and N_i the number of the ith allele in the sample [frequency = $N_i/(2N)$]. No extra information is obtained from the T_{ij}, where these are departures of numbers of heterozygotes from expectation. Alternatively, the best estimator can be computed from the T_{ij}, ignoring the T_{ii}. Also (1) the variance of the estimate of f equals $1/(N(k - 1))$ when all individuals in the sample are unrelated, and the test for $f = 0$ with 1 d.f. is given by the ratio of the estimate to its standard error; (2) the variance is reduced if some alleles are rare; and (3) if the sample consists of full-sib families of size n, the variance is increased by a proportion $(n - 1)/4$ but is not increased by a half-sib relationship.——If f is not small, the structure of the population is of critical importance. (1) If the inbreeding is due to a proportion of inbred matings in an otherwise random-breeding population, f as determined from homozygote excess is the same for all genes and expressions are given for its sampling variance. (2) If the homozygote excess is due to population admixture, f is not the same for all genes. The above estimator is probably close to the best for all f values.

TESTS of departure from Hardy-Weinberg proportions are frequently made to check on random mating in populations, and the deviations from expectation are used to estimate inbreeding coefficients. In this paper we shall investigate some of the sampling properties of the deviations and of the estimates from them.

We first need to clarify some aspects of the current usage of the inbreeding coefficient F. In discussing his early work in his recent volumes, WRIGHT (1969, p. 173) emphasizes that the coefficient relates two populations, one present and one past: "The symbol F is to be interpreted as the correlation between pairs of homologous genes in the uniting gametes that trace in the way indicated by the pedigree to the foundation stock, relative to the array of genes at any neutral locus in that stock," and later: "the relativity referred to above has sometimes been overlooked or misinterpreted."

Later he modified his usage by reinterpreting F as a description of population structure, by defining it as the correlation between uniting gametes relative to the gamete pool of the present population. HALDANE (1954) uses essentially the same approach for "the inbreeding coefficient of a population" by a definition (f) in terms of the excess of homozygotes above Hardy-Weinberg expectations in the population.

Wright extended the second concept in a population separated into a large number of equivalent subpopulations within which mating may not be at random, *e.g.*, a human population, consisting of separate races in each of which there is a proportion of first cousin matings. He defined the following correlations: F_{IT}, between uniting gametes relative to the whole population; F_{IS}, between uniting gametes relative to their own subpopulation; F_{ST}, between random gametes from the same subdivision relative to the whole population; and showed that

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}).$$

We shall restrict the term f to measures of the inbreeding coefficient from the excess of homozygotes in the population and F for that obtained from pedigrees, because these are not necessarily equivalent. We consider two extreme models. The first is a large population in which there is no permanent subdivision but a proportion of matings are between close relatives ($F_{ST} = 0$, $F_{IS} \neq 0$). The population value of f will then be the same for each locus and will equal F_{IS}. The second is a population made up of n separate subpopulations in each of which there is random mating ($F_{ST} \neq 0$, $F_{IS} = 0$). Because the extent of divergence of allele frequency between the subpopulations will differ among alleles at multiallelic loci and among loci, the value of f will not then be the same for each locus and its expected value will be $(n - 1)F_{ST}/n$.

In later sections of this paper, we shall consider estimation of f in both of these cases. Initially, we analyze the sampling properties for populations that are in Hardy-Weinberg equilibrium (HWE), or can be assumed to be as a null hypothesis in tests, where we need make no distinction between the two situations that give rise to departures from HWE.

POPULATIONS IN HARDY-WEINBERG PROPORTIONS

Notation: Let the sample size be N and the number of individuals of genotype A_iA_j be N_{ij}, $i \leq j$, where $i, j = 1, \ldots, k$ alleles. The number of the allele A_i in the sample is $N_i = \sum_{j=1}^{i} N_{ji} + \sum_{j=i}^{k} N_{ij}$, the frequency of A_i in the sample is $N_i/2N$ and the frequency in the population from which the sample was drawn is p_i. The departure, D_{ii} or D_{ij}, of numbers of individuals from Hardy-Weinberg expectation in the sample is given by

$$D_{ii} = N_{ii} - N_i^2/(4N), \qquad D_{ij} = N_{ij} - N_iN_j/(2N) \tag{1}$$

where $2D_{ii} + D_{1i} + \ldots + D_{ik} = 0$. For two alleles

$$D_{11} = D_{22} = -D_{12}/2 = (N_{11}N_{22} - N_{12}^2/4)/N = D, \quad \text{say.}$$

As we shall discuss subsequently, even for a population in HWE, it is well known that $E(D) \neq 0$. Thus, HALDANE (1954) defined a quantity for two alleles,

$$T = (4ND + N_{12})/[4(N - 1)] \tag{2}$$

which generalizes (SMITH 1970) to

$$\begin{aligned} T_{ii} &= (4ND_{ii} + N_i - 2N_{ii})/[4(N - 1)] \\ &= [2(2N - 1)N_{ii} - N_i(N_i - 1)]/[4(N - 1)] \\ \text{and} \quad T_{ij} &= [(2N - 1)N_{ij} - N_iN_j]/[2(N - 1)] \end{aligned} \tag{3}$$

Also, let $\mathbf{N}_i = (N_1, \ldots, N_k)$ and $\mathbf{N}_{ij} = (N_{11}, \ldots, N_{kk})$ be vectors of allele and genotype numbers, respectively, in the sample and $\mathbf{p} = (p_1, \ldots, p_k)$ be the vector of allele frequencies in the population.

Sampling properties of the disequilibria: These can be considered at two levels: first, conditional on the numbers of each allelic type in the sample, N_i, and then unconditionally. Formulas are given for conditional means and variances in APPENDIX (1). In particular, these show

$$\begin{aligned} E(D_{ii} \,|\, \mathbf{N}_i) &= -N_i(2N - N_i)/[4N(2N - 1)] \\ E(D_{ij} \,|\, \mathbf{N}_i) &= N_iN_j/[2N(2N - 1)], \qquad i < j \end{aligned} \tag{4}$$

these values departing from zero because the sampling of genotypes, given allele frequencies, is without replacement. However, $E(T_{ij} \,|\, \mathbf{N}_i) = 0$, for all i and j. When unconditional on the sample numbers, from the multinomial distribution

$$E(D_{ii}) = -p_i(1 - p_i)/2, \qquad E(D_{ij}) = p_ip_j \tag{5}$$

This bias in D_{ii} can also be obtained by noting that the allele frequency $N_i/(2N)$ varies among samples, with variance $p_i(1 - p_i)/(2N)$, so the expected frequency of homozygotes computed from these sample frequencies is $Np_i^2 + p_i(1 - p_i)/2$, compared with the population frequency of Np_i^2.

The conditional variances of D_{ii} and T_{ii} are, from the appendix,

$$\begin{aligned} \operatorname{var}(D_{ii} \,|\, \mathbf{N}_i) &= N_i(N_i - 1)(2N - N_i)(2N - N_i - 1)/[2(2N - 1)^2(2N - 3)] \\ \operatorname{var}(T_{ii} \,|\, \mathbf{N}_i) &= ((2N - 1)^2/[4(N - 1)]^2) \operatorname{var}(D_{ii} \,|\, \mathbf{N}_i) \end{aligned} \tag{6}$$

and the unconditional variances are

$$\begin{aligned} \operatorname{var}(D_{ii}) &= p_i^2(1 - p_i)^2[(4N^2 - 2N - 3)/(4N)] + p_i(1 - p_i)/(8N) \\ \operatorname{var}(T_{ii}) &= [N(2N - 1)/(2(N - 1))]p_i^2(1 - p_i)^2. \end{aligned} \tag{7}$$

Values for $\operatorname{var}(T_{ii})$ and $\operatorname{var}(T_{ij})$ are also given by SMITH (1970). They differ in form from those for the variances of $D_{ij} (i < j)$ because the T_{ij} have mean 0, whereas the D_{ij} do not (equation 5). Note, however, that the bias in D_{ij} is small: for example, from (5) and (7)

$$E(D_{ii})/\text{SE}(D_{ii}) = \tfrac{1}{2}\{N[1 + 1/(8N^2p_i(1 - p_i))]\}^{-1/2},$$

i.e., the bias is of order $1/\sqrt{N}$ of its standard error. The sampling properties of D_{ij} and T_{ij} can also be compared in terms of their mean square error (MSE) to check, following WEIR and COCKERHAM (1984), whether unbiassed estimators of parameters describing inbreeding have minimum MSE. In this case

$$\text{MSE}(D_{ii}) - \text{MSE}(T_{ii}) = p_i(1 - p_i)[1/(8N) - (3/4)p_i(1 - p_i)],$$

and so, unless gene frequencies are very extreme, D_{ii} has the lower MSE. We shall generally use the unbiassed estimator, however. As sample size increases, formulas for the conditional variances, with N_i replaced by $2Np_i$, and the unconditional variances of N_{ij}, D_{ij} and T_{ij} approach the same values. The large sample variances and covariances are shown in Table 1.

The entries in Table 1 may be summarized as follows: (1) genotypes with no alleles in common have positive correlations between their deviations; (2) the deviations of a homozygote and a heterozygote with an allele in common are negatively correlated; (3) the correlations for two heterozygotes with an allele in common are positive if the frequency of the common allele is greater than 0.5 and vice versa.

In the two-allele case, where $D_{11} = D_{22} = -D_{12}/2 = D$, a simple demonstration of these formulas is given by noting that χ^2 with 1 d.f. used to test the significance of the deviation is, for large sample sizes (where p is the frequency of the first allele),

$$\chi^2 = D^2/[Np^2] + 4D^2/[2Np(1 - p)] + D^2/[N(1 - p)^2] = D^2/[Np^2(1 - p)^2].$$

Since χ_1^2 is distributed asymptotically as a standardized normal deviate, it follows that $\text{var}(D) = Np^2(1 - p)^2$. The deviation of the number of heterozygotes from expectation (H), therefore, has variance H^2/N, where $H = 2Np(1 - p)$. This does not apply in the multiple-allele case, however.

In general, the expectation of the usual χ^2 statistic does not equal its degrees of freedom in small samples, both because the D_{ij} do not have zero mean and because the expected numbers of each genotype are based on those in the sample. As we shall discuss subsequently when considering estimation of the inbreeding coefficient, these departures are likely to be greatest if there are rare alleles.

The consequences of family structure in the sample: Equations (5) to (7) and the classical tests for goodness-of-fit to Hardy-Weinberg proportions assume random sampling from a large population, such that members of the sample are not related. This is often unlikely to be true. To what extent is the test then invalid? We consider the case of HWE and use approximations for variances appropriate for a large sample, such that terms of $O(1/N^2)$ can be ignored relative to those of $O(1/N)$.

First, the biases in the D_{ij} (5) are increased by family structure in the sample. If it is contributed equally by F fathers and M mothers, the deficiency of homozygotes is given by

$$E(D_{ii}) = -Np_i(1 - p_i)(1/(8M) + 1/(8F) + 1/(4N)) \tag{8}$$

(ROBERTSON 1965), which reduces to $-p_i(1 - p_i)/2$ when $M = F = N$.

Second, the variances of the D_{ij} are also increased because related individuals are more likely to have the same genotype. We consider the two-allele case for illustration. Expanding D in a Taylor series of terms of N_{11}, N_{12} and N_{22}

TABLE 1

Examples of variances and covariances of departures from Hardy-Weinberg expectation for a locus with four or more alleles

	D_{11}	D_{22}	D_{12}	D_{13}	D_{34}
D_{11}	$p_1^2(1-p_1)^2$	$p_1^2p_2^2$	$-2p_1^2(1-p_1)p_2$	$-2p_1^2(1-p_1)p_3$	$2p_1^2p_3p_4$
D_{22}		$p_2^2(1-p_2)^2$	$-2p_1p_2^2(1-p_2)$	$2p_1p_2^2p_3$	$2p_2^2p_3p_4$
D_{12}			$2p_1p_2[p_1p_2+(1-p_1)(1-p_2)]$	$2p_1(2p_1-1)p_2p_3$	$4p_1p_2p_3p_4$
D_{13}				$2p_1p_3[p_1p_3+(1-p_1)(1-p_3)]$	$2p_1p_3(2p_3-1)p_4$
D_{34}					$2p_3p_4[p_3p_4+(1-p_3)(1-p_4)]$

All entries should be multiplied by N.

and taking expectations,

$$\operatorname{var}(D) = \left(\frac{\partial D}{\partial N_{11}}\right)^2 \operatorname{var}(N_{11}) + 2\left(\frac{\partial D}{\partial N_{11}}\right)\left(\frac{\partial D}{\partial N_{12}}\right)\operatorname{cov}(N_{11}, N_{12}) + \ldots,$$

with higher order terms removed and derivatives evaluated at N_{ij} equal to their expected value with HWE in the population. After rearrangement, this equation reduces to

$$\operatorname{var}(D) = \operatorname{var}[N_{11}(1-p)^2 - N_{12}p(1-p) + N_{22}p^2],$$

which is the variance of the sum of the "scores" of N individuals, when the three genotypes are assigned scores $(1-p)^2$, $-p(1-p)$ and p^2, respectively. Score has mean zero and variance $p^2(1-p)^2$. If the sample consists of S groups of relatives of size n, then the variance may be written as

$$\operatorname{var}(D) = p^2(1-p)^2[Sn + Sn(n-1)r] = Np^2(1-p)^2[1 + (n-1)r],$$

where r is the correlation of score within groups. It is simple to show that r is zero for half-sibs and ¼ for full-sibs. The increase in variance is then dependent only on full-sib relationship and may be written as

$$\operatorname{var}(D) = Np^2(1-p)^2[1 + (n-1)/4]. \qquad (9)$$

Allowing for full-sib families of variable size

$$\operatorname{var}(D) = Np^2(1-p)^2[3/4 + \bar{n}(1 + \sigma_n^2/\bar{n}^2)/4]$$

where $\bar{n}$ and σ_n^2 are the mean and variance of family size, respectively. Using Monte Carlo simulation, we have shown that this expression for the increase in variance is a good approximation even in small samples with few families.

B. S. WEIR (personal communication) has shown that the coefficient of r in (9) is, more generally, twice the two-gene descent measure (COCKERHAM 1971) defined as the probability that, for two individuals, both their paternally and maternally derived genes are identical by descent or the paternal of the first is identical with the maternal of the second and vice versa. Also, the formula applies for multiple alleles where, for allele i, p is replaced by p_i.

The variance of the estimated departure from HWE is thus increased if some of the individuals sampled are full sibs (or otherwise related through both parents) but not if half sibs (or otherwise related through one parent) since the latter does not affect the probability that individuals have identical genotypes. The χ^2 goodness-of-fit test is thus biased with full-sib families; to illustrate this we have assumed that, for two alleles, D is normally distributed with mean zero and variance given by (9) and computed the probability that a "significant" departure from HWE will be obtained using tabulated 5 and 1% significance levels (Table 2). The biases in the Hardy-Weinberg tests are seen to become large rapidly if there are appreciable relationships among the sampled individuals.

Estimation of f *and testing for deviations from equilibrium*: We now consider the estimation of f and properties of the estimates in two- and multiallele cases. The expected genotypic frequencies are

$$E(N_{ii}) = N[p_i^2 + fp_i(1-p_i)] \quad \text{and} \quad E(N_{ij}) = 2Np_ip_j(1-f).$$

TABLE 2

Effect of family structure on bias in Hardy-Weinberg tests

	Probability of rejection of HWE (%)	
Sample of unrelated individuals	5%	1%
Half families of $n = 1$, half of $n = 2$	7.0	1.7
$n = 2$	8.0	2.1
$n = 4$	13.9	5.2
$n = 8$	23.7	12.0

The "obvious" estimator for the two-allele case, where there is only 1 d.f., is from D,

$$\hat{f}_D = 4ND/[N_i(2N - N_i)];$$

but since, for $f = 0$, $E(D \mid \mathbf{N}_i) \neq 0$, $\hat{f}_D$ is biased (4), whereas Haldane's estimator [from (3)]

$$\hat{f}_T = 4NT/[N_i(2N - N_i)]$$

is not. If, however, $f \neq 0$, $\hat{f}_T$ is not unbiased. HALDANE (1954, p. 633) forgot to include the gene frequency terms in the denominator of his formula and thus concluded that $\hat{f}_T$ was always unbiased. The bias is small, of order $f/(2N)$, and is, therefore, not serious. Nevertheless, we shall use $\hat{f}_T$ in subsequent analyses, because its unbiasedness at $f = 0$ is a desirable property and leads to some simplification in the formulas. We shall first consider estimation of f in the multiallelic case before considering the sampling properties in more detail.

When we wish to test for homozygote excess the null hypothesis is of HWE; therefore, it is reasonable to extract 1 d.f. to estimate f and for this estimate to have optimal properties when $f = 0$. We are then able to obtain explicit formulas for the estimator and its variance, whereas, if $f \neq 0$, iterative methods are required and will be discussed subsequently. LI and HORVITZ (1953) have considered alternative estimators but not their sampling properties; YASUDA (1968) gave an expression for the variance of the maximum likelihood estimator but not for the estimator itself.

In view of the complexity of some of the formulas, we shall derive the estimator on the assumption that the sample size is large but investigate its properties more generally. For k alleles there are $k(k + 1)/2$ different estimators of f from T_{ij} given by

$$\hat{f}_{ii} = 4NT_{ii}/[N_i(2N - N_i)], \qquad \hat{f}_{ij} = -2NT_{ij}/(N_iN_j). \tag{10}$$

These are constrained by k equations of the form $2T_{ii} + T_{1i} + \ldots + T_{ik} = 0$, and the large variances for $\hat{f}_{ij}$ are given by (7) and Table 1 after appropriate scaling, for example, as

$$\begin{aligned} \text{var}(\hat{f}_{ii}) &= 1/N \\ \text{cov}(\hat{f}_{ii}, \hat{f}_{jj}) &= p_ip_j/[N(1 - p_i)(1 - p_j)]. \end{aligned} \tag{11}$$

In APPENDIX (2) we show that a set of weights that minimize the variance of a pooled estimator $\hat{f}$ are $(1 - p_i)/(k - 1)$ for $\hat{f}_{ii}$ and 0 for $\hat{f}_{ij}$, *i.e.*, $\hat{f} = \sum_i(1 - p_i)\hat{f}_{ii}/(k - 1)$. This uses only the departures in homozygotes but nevertheless makes full use of the data. Because of the dependencies between the T_{ii} and T_{ij}, the same estimate can also be obtained using only the heterozygotes and giving $\hat{f}_{ij}$ weight $(p_i + p_j)$, *i.e.*, $\hat{f} = \sum\sum_{i \neq j} (p_i + p_j)\hat{f}_{ij}/(k - 1)$. Thus, we have, at the optimum,

$$\hat{f} = \left[\sum_{i=1}^{k} T_{ii}/(Np_i)\right] \Big/ \sum_{i=1}^{k} (1 - p_i). \tag{12}$$

When the p_i are estimated, the estimator in (12) is not the most efficient, but derivation of anything better is not feasible. Equation (12) then becomes

$$\hat{f}_T = 2\left[\sum_{i=1}^{k} (T_{ii}/N_i)\right] \Big/ (k - 1). \tag{13}$$

In large samples, using (7) and (12), it can be shown that

$$\text{var}(\hat{f}_T) = 1/[N(k - 1)], \tag{14}$$

which is also the variance of the maximum likelihood estimator obtained by large-sample theory (YASUDA 1968). For two alleles, this reduces to $1/N$ and a simple illustration is that, if f is viewed as a correlation of genes in uniting gametes, its variance is that of an estimate of a point correlation coefficient for a sample of size N with a true correlation of zero.

For smaller samples, using the APPENDIX, we obtain from (12)

$$\text{var}(\hat{f}_T) = \frac{1}{N(k - 1)}\left[1 - \frac{1}{2N(k - 1)} \sum_{i=1}^{k} \frac{1}{p_i}\right], \tag{15}$$

which, for two alleles, reduces to

$$\text{var}(\hat{f}_T) = (1/N)[1 - 1/(2Np(1 - p))].$$

Thus, we see that, somewhat surprisingly, rare alleles give most information about the inbreeding level, providing it is low. CANNINGS and EDWARDS (1969) noted this using a different formulation. As a consequence, when estimating f, it is very inefficient to combine data on rare alleles: if there are three or more, it is better to combine the frequent than the rare ones. (It will be shown later, however, that when $f \neq 0$, rare alleles lead to high variances of f.)

The estimator (13) was suggested as "the simplest method of estimation in the sense that it involves the least arithmetic labor" in a discussion of several estimators by LI and HORVITZ (1953). However, they did not give sampling variances of the different estimates nor realize that this was also the minimum variance estimator when the true value of f is zero.

WARD and SING (1970) consider the sample sizes required for testing departure from HWE as a function of significance level (α), power (β) and true value of f. For multiple alleles, however, they used the overall χ^2 test with $k(k - 1)/2$ d.f. For example, with $\alpha = 0.05$, $\beta = 0.9$ and $f = 0.05$, they computed required sample sizes of 4205, 2324, 1887 and 1692 for $k = 2$, 4, 6 and 8 alleles, respectively. The power of the test for departure can be increased considerably by using only the single degree of freedom that maximizes information about f. For low values of f, such as this, the required sample size to give the same power with k alleles when a single degree of freedom is extracted will decline in proportion to $k - 1$ as shown by (14). Hence, using the single degree of freedom test, the sample sizes required in the example will be 4205, 1402, 841 and 601 approximately, for $k = 2$, 4, 6 and 8, respectively. These represent a substantial increase in efficiency over the overall χ^2 test, although the numbers required remain large.

ESTIMATION OF f IN INBRED POPULATIONS

Large populations with no subdivision; $F_{ST} = 0$; $F_{IS} \neq 0$: There are two alternatives: (1) all individuals have the same inbreeding coefficient or (2) a proportion of matings are between close relatives, *e.g.*, full cousins, and all others are at random. It is in fact difficult to imagine any practical situation leading to (1) (a cyclical mating scheme as occasionally used in laboratory experiments would be an example) and in practice most analyses are concerned with structure (2). The two differ only in that in (2) samples will differ in the proportion of inbred individuals, leading to a correlation between f values at different loci in the same sample. This proves to be very small and will be ignored.

Many aspects of the problem have been discussed previously (LI and HORVITZ 1953; YASUDA 1968; SMITH 1970; MANTEL and LI 1974; CURIE-COHEN 1982). Analyses of population mixture which would lead to heterogeneity among the proportional deficiencies of different heterozygotes are deferred to the next section.

If f in the sampled population is nonzero, the estimate (13) is not minimum variance. Both f and the gene frequencies have to be estimated simultaneously. For loci with two alleles, LI and HORVITZ (1953) show that the simple gene count, $N_i/2N$, is the maximum likelihood estimator, $\hat{p}_i$, of gene frequency; but for multiple alleles, this is not so, although LI and HORVITZ, quoting SEWALL WRIGHT, argue that it should be taken as such. We do not accept this view for the case we consider here of homogeneous inbreeding of all alleles, for it is clear that, when there is an excess of homozygotes due to inbreeding, they give less information about gene frequency. It turns out, however, that $N_i/(2N)$ is very close to $\hat{p}_i$ unless the population is very highly inbred. We present deviations in terms of D_{ii}(*i.e.*, as $N_{ii} - N\hat{p}_i^2$) for simplicity, although there might be some benefit in using the T_{ii} to remove bias at small f. CURIE-COHEN (1982) also gives a maximum likelihood estimator but takes $\hat{p}_i = N_i/(2N)$.

A suitable iterative solution to the maximum likelihood equations is as follows, which uses information only on the homozygotes, optimal for at least small f. Given a set of estimates $\hat{f}_L$ and $\hat{p}_i$, $i = 1, \ldots, k$, obtain a new estimate, $\hat{f}'_L$ of the inbreeding coefficient as

$$\hat{f}'_L = \frac{\sum_{i=1}^{k} [N_{ii} - N\hat{p}_i^2]/[\hat{p}_i + \hat{f}_L(1 - \hat{p}_i)]}{N \sum_{i=1}^{k} [\hat{p}_i(1 - \hat{p}_i)]/[\hat{p}_i + \hat{f}_L(1 - \hat{p}_i)]}, \tag{16}$$

and, replacing $\hat{f}_L$ by $\hat{f}'_L$, new estimates of the allelic frequencies as solutions to the quadratic equations

$$N(2 - \hat{f}_L)(1 - \hat{f}_L)\hat{p}_i^2 + [N(2 - \hat{f}_L)\hat{f}_L - N_i(1 - \hat{f}_L)]\hat{p}_i - (N_i - N_{ii})\hat{f}_L = 0. \quad (17)$$

Equation (16) is now recomputed using the estimates from (17) and so on, and convergence is very rapid. Starting with $\hat{p}_i = N_i/2N$ and $\hat{f}_L = 0$, (16) reduces to (12), so the latter gives the first iterate for the inbreeding coefficient.

The large sample variance of $\hat{f}_L$ is given by (14) when $f = 0$ (YASUDA 1968), but it has a very involved formula otherwise. For two alleles the variance of T was computed by SMITH (1970), from which that for $\hat{f}_L$ can be deduced. Small sample values were given, but for simplicity consider just large sample sizes where second order terms can be ignored. Then, where $q = 1 - p$

$$\text{var}(D) = Npq[(1 - f^2)pq + f(p - q)^2]$$

and

$$\text{var}(\hat{f}) = [(1 - f^2)pq + f(1 - f)^2(p^4 + q^4 - 2p^2q^2 + 2pq) + f^2(1 - f)]/(Npq)$$

which reduces to

$$\text{var}(\hat{f}) = [1 + f(p - q)^2/(pq)]/N, \quad \text{if } f \text{ is small,}$$

and

$$\text{var}(\hat{f}) = (1 - f^2)/N, \quad \text{if } p = q = 0.5.$$

Large sample formulas for var($\hat{f}$) from (12) for multiple alleles are given by CURIE-COHEN (1982) but not for the maximum likelihood estimator (16).

Likelihood ratio methods can be used to test whether $f = 0$ and whether the data fit the model of homogeneous inbreeding at all alleles.

Rare alleles: If one or more alleles are rare, especially if a homozygote class is missing, the estimating procedures (13), (16) and (17) may not work: either giving negative expected genotype frequencies or, for $\hat{f}_L$, failing to converge. A numerical "hill-climbing" procedure can be used to maximize the likelihood, with constraints imposed to prevent solutions going out of bounds.

An alternative procedure is to combine rare alleles. This loses efficiency if $f = 0$ (14) but increases efficiency if f is large, as the above expressions and those of CURIE-COHEN show. The best recipe is not obvious.

A subdivided population, with random mating within subpopulations; $\text{F}_{\text{ST}} \neq 0$; $\text{F}_{\text{IS}} = 0$: Here, we face problems of a different kind. Suppose that, unknown to the investigator, the population consists of two isolated subpopulations, A and B, descended from the same foundation stock and now equally numerous. We might expect that, for any gene, the gene frequencies p_A and p_B in the subpopulations would differ because of genetic drift. A sample from the population would then be expected to have an excess of homozygotes, equal to $(p_A - p_B)^2/4$. This is the well-known Wahlund effect. Furthermore, if we consider several genes, starting at the same frequency in the foundation stock, we would not expect $p_A - p_B$ to be the same for all. Thus, in this situation we expect the population f value, measured from homozygote excess, $p_{ii} - p_i^2$, to be different for different genes. The same will be true for different alleles at the same locus. Superimposed on the sampling that leads to different f values for different genes, we have the estimation of f for a particular gene by repeated sampling of diploid individuals from the population. For this latter problem alone, the method using (16) and (17) applies. But, in fact, we frequently need to summarize information over many genes and ask more general questions, such as "what is the best estimate of f for this population?" or "does the population of all cows have a higher f than that of all humans?" In such questions, we treat the loci involved as a random sample of such loci and wish to make inferences about the genetic structure and history of the population. The main source of error may then be variation over loci. Therefore, we need to discuss how with this model we should combine information from several alleles at a locus and from several loci. We first need to make the approach with two subpopulations more general. Consider a population consisting of n randomly chosen subpopulations, and a gene with frequency p_{im} in the mth subpopulation. It is easily shown that the excess of homozygotes, $p_{ii} - p_i^2$, equals

$$\sum_m \frac{p_{im}^2}{n} - \frac{\left(\sum_m p_{im}\right)^2}{n^2} = \frac{n-1}{n}\,\frac{\left[\sum_m p_{im}^2 - \left(\sum_m p_{im}\right)^2 \Big/ n\right]}{n-1}. \qquad (18)$$

Now, the second term on the right hand side of (18) is an estimate of the variance, V, of gene frequency between subpopulations, based on $(n - 1)$ d.f. We have then

$$E(f_{ii}) = \frac{n-1}{n}\left(\frac{V}{p_{io}(1 - p_{io})}\right) = (n - 1)F_{ST}/n$$

where p_{io} is the frequency in the foundation stock. The variance of V over loci is $2V^2/(n - 1)$, assuming the inbreeding of the subpopulations, F_{ST}, is sufficiently small that the p_{im} are normally distributed. It follows that the variance of f over genes in the same population is $2E^2(f)/(n - 1)$. This variance may be much larger than that due to repeated sampling of individuals. In a population with $n = 5$, $E(f) = 0.1$ and a sample size of 800, the variance of f over genes is $2 \times (0.1)^2/4 = 0.0050$ and that due to sampling within populations is 0.0012.

Note that, if the pedigree inbreeding coefficient of the subpopulations relative to the foundation stock is F, the expected value of f, as the correlation between uniting gametes relative to the population as a whole, is not F but $F(n - 1)/n$.

How then do we combine information for the deviations of the $k(k + 1)/2$ genotypes at a locus with k alleles? Consider only the variation due to genetic drift and ignore that due to present sampling of individuals. Using f_{ij} for the population value derived from the ijth genotype, we need the drift variances and covariances of the separate f values, noting that f_{ij} is an estimate of $-(n - 1)\text{cov}(p_i, p_j)/(np_{io}p_{jo})$. If f values are small, we can derive these from known formulas for sampling from a normal multivariate distribution, *i.e.*, that in samples of size n from a population in which the covariance between variates x and y is C_{xy}, the sampling covariance between C_{xy} and C_{zw} is $(C_{xz}C_{yw} + C_{xw}C_{yz})/(n - 1)$ (KENDALL and STUART 1968, eq. 41.98). It appears that the variance-covariance matrix of the various genotypic deviations in a subdivided population is proportional to that given in (7), which arises when sampling individuals from a population when $f = 0$. This is perhaps not surprising—the present estimate is an intraclass coefficient of genes in subpopulations arising by drift from the foundation stock, whereas that in (7) is the coefficient between genes in the same individual in a sample of N from a population at equilibrium.

There are three practical implications: (1) In large samples, (13) is the minimum variance estimator in a subdivided population, as it is in a population with $f = 0$. This would suggest its general use. (2) The variance of f estimates, given subdivision, contains the unknown n, the number of subpopulations involved. An empirical estimate of the standard error of the population estimate would have to be obtained from the observed variation over genes. (3) The sampling variance of f estimates is the same for all alleles. The information from a multiallelic locus is proportional to the number of alleles minus one.

Note that, in the conventional χ^2 analysis of deviations from expectation with several alleles at a locus, fitting the best estimate of f to the data will not remove all significance from the remaining deviations—different alleles may and probably will differ in their population f values.

We examined this problem briefly by simulation at a locus with four alleles at equal frequencies. We produced 40 subpopulations by repeated sampling over several "generations" and from each took a final sample of 400 diploid individuals. The samples were combined at random in pairs to give 20 samples of 800 individuals with the required structure, with $n = 2$ and $k = 4$. We then have ten genotypes, constrained by four allele frequencies, giving 6 d.f. for deviations. The expected value of f in such samples was 0.095. We calculated $\hat{f}$ in each of the 20 samples, first using (13) and, second, by the maximum likelihood procedure using (16) and (17), although the latter is not strictly applicable to populations in which alleles may have different f values. The results were as follows: (1) Using (13), $\hat{f}_T$ had a mean of 0.100 and a standard deviation of 0.068. The expected variance of $\hat{f}$ due to genetic drift, $2\hat{f}^2/3$, is 0.0067 and that due to present sampling $[1/(800 \times 3)]$ is 0.0004, giving an expected standard deviation of 0.084. After fitting the best $\hat{f}$ value, 5 d.f. remain for residual deviations. The χ^2_5 was significant at the 5% level in 18 of 20 samples. (2) Using (16) and (17), $\hat{f}_L$ had a mean of 0.122 and a standard deviation of 0.108. Thus, this estimator was both biased and inefficient. The difference between estimator and actual f value was particularly large for populations with high f.

Whatever the population structure, it appears that loci contribute information on $\hat{f}$ proportional to their number of alleles segregating less one. But the actual amount of information available is not calculable *a priori*, although an empirical measure can be obtained from the variation in estimates between loci. This effect of structure parallels closely that discussed between NEI and

MARUYAMA (1975) and ROBERTSON (1975) on the one hand and LEWONTIN and KRAKAUER (1973) on the other on f values calculated from the variance in gene frequency between subpopulations. There it appeared that the expected variance of f over loci, in the absence of selection, depended critically on population structure. This treatment is also relevant to another problem in population genetics. We discussed the present problems in terms of populations made up of isolated subpopulations, all descended from the same foundation with the same inbreeding relative to it. An obvious measure of the extent to which any pair of populations would have separated by genetic drift (the "distance" between the two) is then F_{ST} or some multiple of it.

Suppose we knew the gene frequencies of a given allele in the two subpopulations to be p_{iA} and p_{iB} with a mean of $\bar{p}_i$. Then, the estimate of distance obtained from the allele is $(p_{iA} - p_{iB})^2/[\bar{p}_i(1 - \bar{p}_i)]$ with expectation $2F_{ST}$. For this allele, the f value in the joint population is $(p_{iA} - p_{iB})^2/[4\bar{p}_i(1 - \bar{p}_i)]$. Thus, for large samples, the estimate of distance we should get would be four times the estimate of f.

We may then apply the present theory to the estimation of distance from multiple alleles at a locus. For allele i, we have a distance estimate to which we should give weight $(1 - \bar{p}_i)$ to give a overall estimate $\sum_{i=1}^{k} (p_{iA} - p_{iB})^2/[\bar{p}_i(k - 1)]$. Estimates at different loci would have weight $(k - 1)$ and the overall estimate (as a variance estimate with $\sum(k - 1)$ d.f.) has sampling error equal to itself times $\sqrt{2/\sum(k - 1)}$. This estimator was originally suggested by SANGHVI (1953), and his earlier work on the estimation of distance is reviewed by SMITH (1977). As EDWARDS and CAVALLI-SFORZA (1972) point out, these are equivalent, when f is small, to the angular measures of distance which they have used. These contrast with those of REYNOLDS, WEIR and COCKERHAM (1983) who, apart from other differences, used weights closer to $p_i(1 - p_i)$ and are, therefore, probably less efficient.

LITERATURE CITED

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

COCKERHAM, C. C., 1971 Higher order probability functions of identity of alleles by descent. Genetics **69:** 235–246.

CURIE-COHEN, M., 1982 Estimates of inbreeding in a natural population: a comparison of sampling properties. Genetics **100:** 339–358.

EDWARDS, A. W. F. and L. L. CAVALLI-SFORZA, 1972 Affinity as revealed by differences in gene frequencies. pp. 37–47 In: *The Assessment of Population Affinities in Man,* Edited by J. S. WEINER and J. HUIZINGA, Clarendon Press, Oxford.

EMIGH, T. H., 1980 A comparison of tests for Hardy-Weinberg equilibrium. Biometrics **36:** 627–642.

HALDANE, J. B. S., 1954 An exact test for randomness of mating. J. Genet. **52:** 631–635.

KENDALL, M. G. and A. STUART, 1968 *The Advanced Theory of Statistics,* Vol. III. Charles Griffin and Company, London.

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

LI, C. C. and D. G. HORVITZ, 1953 Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. **5:** 107–117.

MANTEL, N. and C. C. LI, 1974 Estimation and testing of a measure of non-random mating. Ann. Hum. Genet. **37:** 445–454.

NEI, M. and T. MARUYAMA, 1975 Lewontin-Krakauer test for neutral genes. Genetics **80:** 395.

REYNOLDS, J., B. S. WEIR and C. C. COCKERHAM, 1983 Estimation of the coancestry coefficient: basis for short-term genetic distance. Genetics **105:** 767–779.

ROBERTSON, A., 1965 The interpretation of genotypic ratios in domestic animal populations. Anim. Prod. **7:** 319–324.

ROBERTSON, A., 1975 Remarks on the Lewontin-Krakauer test. Genetics **80:** 396.

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

SMITH, C. A. B., 1970 A note on testing the Hardy-Weinberg law. Ann. Hum. Genet. **33:** 377–383.

SMITH, C. A. B., 1977 A note on genetic distance. Ann. Hum. Genet. **40:** 463–479.

WARD, R. H. and C. F. SING, 1970 A consideration of the power of the χ^2 test to detect inbreeding effects in natural populations. Am. Nat. **104:** 355–365.

WEIR, B. S. and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution. In press.

WRIGHT, S., 1969 *Evolution and the Genetics of Populations, Vol. 2: The Theory of Gene Frequencies.* University of Chicago Press, Chicago.

YASUDA, N., 1968 Estimation of the inbreeding coefficient from phenotypic frequencies by a method of maximum likelihood scoring. Biometrics **24:** 915–935.

Corresponding editor: B. S. WEIR

APPENDIX

(1) *Exact sampling formulas for populations in HWE*: An extension of HALDANE'S (1954) formula for two alleles (EMIGH 1980) gives an expression for the distribution of numbers of each genotype in a sample from a population in HWE, conditional on the numbers of each allele in the sample:

$$P(\mathbf{N}_{ij}\,|\,\mathbf{N}_i) = \left[N!\prod_i N_i!\;2^{(N-\Sigma_i N_{ii})}\right]\Big/\left[(2N)!\prod_{i\leq j}\prod N_{ij}!\right] \tag{1A}$$

where $\mathbf{N}_{ij}$ and $\mathbf{N}_i$ are vectors of the numbers of each genotype and allele. Means, variances and covariances for the $\mathbf{N}_{ij}$ conditional on the $\mathbf{N}_i$ are readily derived from (1A). Typical values are as follows:

$$E(N_{11}\,|\,\mathbf{N}_i) = -N_1(2N - N_1)/[4N(2N-1)] + N_1^2/(4N)$$

$$E(N_{12}\,|\,\mathbf{N}_i) = N_1N_2/[2N(2N-1)] + N_1N_2/(2N)$$

$$\text{var}(N_{11}\,|\,\mathbf{N}_i) = \alpha N_1(N_1-1)(2N-N_1)(2N-N_1-1)/2$$

$$\text{var}(N_{12}\,|\,\mathbf{N}_i) = \alpha N_1N_2[(2N-N_1-1)(2N-N_2-1) + (N_1-1)(N_2-1)]$$

$$\text{cov}(N_{11}, N_{22}\,|\,\mathbf{N}_i) = \alpha N_1(N_1-1)N_2(N_2-1)/2$$

$$\text{cov}(N_{11}, N_{12}\,|\,\mathbf{N}_i) = -\alpha N_1(N_1-1)(2N-N_1-1)N_2$$

$$\text{cov}(N_{11}, N_{23}\,|\,\mathbf{N}_i) = \alpha N_1(N_1-1)N_2N_3$$

$$\mathrm{cov}(N_{12}, N_{13} \mid \mathbf{N}_i) = -\alpha N_1(2N - 2N_1 - 1)N_2N_3$$

$$\mathrm{cov}(N_{12}, N_{34} \mid \mathbf{N}_i) = \alpha 2N_1N_2N_3N_4$$

where $\alpha = 1/[(2N - 1)^2(2N - 3)]$ (2A)

$$\text{Because } D_{ii} = N_{ii} - N_i^2/(4N) \quad \text{and} \quad D_{ij} = N_{ij} - N_iN_j/(2N),$$

$$E(D_{ii} \mid \mathbf{N}_i) = -N_i(2N - N_i)/[4N(2N - 1)] \quad \text{and}$$

$$E(D_{ij} \mid \mathbf{N}_i) = N_iN_j/[2N(2N - 1)]$$

and the matrix $\mathrm{var}(D_{ij} \mid \mathbf{N}_i) = \mathrm{var}(N_{ij} \mid \mathbf{N}_i)$; and because

$$T_{ii} = [2(2N - 1)N_{ii} - N_i(N_i - 1)]/[4(N - 1)]$$

$$T_{ij} = [2(2N - 1)N_{ij} - 2N_iN_j]/[2(N - 1)]$$

it follows that $E(T_{ij} \mid \mathbf{N}_i) = 0$ and

$$\mathrm{var}(T_{ij} \mid \mathbf{N}_i) = \frac{(2N - 1)^2}{4(N - 1)^2}\,\mathrm{var}(N_{ij} \mid \mathbf{N}_i) \qquad (3A)$$

i.e., α in (2A) is replaced by $1/[4(N - 1)^2(2N - 3)]$.

(2) *Minimum variance estimator of* **f** *for multiple alleles*: Let **V** be the variance-covariance matrix of estimates of f from each possible pair of k alleles, ordered as follows $\mathbf{x}' = (\hat{f}_{11}, \hat{f}_{22}, \ldots, \hat{f}_{kk}, \hat{f}_{12}, \hat{f}_{13}, \ldots, \hat{f}_{k-1,k})$. Thus, **V** is square and **x** a column vector, each of dimension $l = k(k + 1)/2$. Since the $\hat{f}_{ij}$ are unbiased estimators, we require a set of weights **w** such that the linear estimate $\hat{f} = \mathbf{w}'\mathbf{x}$ has minimum variance. Because the $\hat{f}_{ij}$ are not independent, there is more than one set of weights **w**, but each gives the same value of $\hat{f}$. An optimal solution for **w** must satisfy

$$\sum_{j=1}^{l} v_{ij}w_j = C, \quad \text{an arbitrary constant,} \quad \text{for all } i \qquad (4A)$$

and

$$\sum_{j=1}^{l} w_j = 1, \quad \text{for all } i \qquad (5A)$$

One possible solution is $w_i = (1 - p_i)/(k - 1)$ for $i = 1, \ldots, k$ and $w_i = 0$, otherwise, *i.e.*, $\hat{f} = \sum_{i=1}^{k} (1 - p_i)\hat{f}_{ii}/(k - 1)$, as we now show. Note that, from (10), (11) and Table 1, the relevant large sample variances and covariances are, for example,

$$N\,\mathrm{var}(\hat{f}_{11}) = 1, \qquad N\,\mathrm{cov}(\hat{f}_{11}, \hat{f}_{22}) = p_1p_2/[(1 - p_1)(1 - p_2)]$$

$$N\,\mathrm{cov}(\hat{f}_{11}, \hat{f}_{12}) = 1, \qquad N\,\mathrm{cov}(\hat{f}_{11}, \hat{f}_{23}) = -p_1/(1 - p_1).$$

Hence, for the row 1 of **V** corresponding to $\hat{f}_{11}$, for example,

$$\sum_{j=1}^{l} v_{1j}w_j = \sum_{j=1}^{k} v_{1j}(1 - p_j) = N\left[1 - p_1 + \sum_{j=2}^{k} p_1p_j/(1 - p_1)\right] = N,$$

and for the row $k + 1$ corresponding to $\hat{f}_{12}$,

$$\sum_{j=1}^{k} v_{k+1,j}(1 - p_j) = N\left[1 - p_1 + 1 - p_2 - \sum_{j=3}^{k} p_j\right] = N.$$

Thus, (4A) is satisfied and, since $\sum_{j=1}^{k} (1 - p_j)/(k - 1) = 1$, (5A) is also satisfied.

(3) *Derivation of var*($\hat{\mathrm{f}}_{\mathrm{T}}$) *in (15) for* f = 0: The conditional mean $E(\hat{f}_{Tii} | \mathbf{N}_i) = 0$ for $f = 0$; therefore, for $\hat{f}_T$ given by (12) and using the above formulas for var(T_{ii}) and cov(T_{ii}, T_{jj})

$$\mathrm{var}(\hat{f}_T | \mathbf{N}_i) = \frac{4}{(k-1)^2} \times \frac{1}{8(N-1)^2(2N-3)}$$

$$\times \left\{\sum_i \frac{N_i(N_i - 1)(2N - N_i)(2N - N_i - 1)}{N_i^2} + \sum_{i \neq j}\sum \frac{N_i(N_i - 1)N_j(N_j - 1)}{N_i N_j}\right\}$$

To compute the unconditional expectation, it is convenient to split the term in $\sum_i$ into two parts, *i.e.*, $(N_i - 1)/N_i = 1 - 1/N_i$. Then, taking expectations, we have

$$\mathrm{var}(\hat{f}_T) = \frac{1}{2(k-1)^2(N-1)^2(2N-3)}$$

$$\cdot\left\{2N(2N-1)\left[\sum_i (1 - p_i)^2 + \sum_{i \neq j}\sum p_i p_j\right] - E\sum_i \frac{(2N - N_i)(2N - N_i - 1)}{N_i}\right\}.$$

Dividing up the last term again, and simplifying, we obtain

$$\mathrm{var}(\hat{f}_T) = \frac{N(2N-1)}{(k-1)(N-1)^2(2N-3)}$$

$$- \frac{1}{2(k-1)^2(N-1)^2(2N-3)}\left[2N(2N-1)\sum_i E(1/N_i) - k(4N-1) + 2N\right].$$

Note that $E(1/N_i)$ is not trivial if p_i is small, and as a first approximation, take $E(1/N_i) = 1/(2Np_i)$. Ignoring second order terms, we get

$$\mathrm{var}(\hat{f}_T) = \frac{1}{N(k-1)} - \frac{1}{2N^2(k-1)^2}\sum(1/p_i).$$