ESTIMATES OF INBREEDING IN A NATURAL POPULATION: A COMPARISON OF SAMPLING PROPERTIES

MARTIN CURIE-COHEN

Laboratory of Genetics, University of Wisconsin, Madison Wisconsin 53706

Manuscript received May 18, 1981
Revised copy accepted November 16, 1981

ABSTRACT

The average inbreeding coefficient f of a population can be estimated in several different ways based solely on the genotypic frequencies at a single locus. The means and variances of four different estimates have been compared. While the four estimates are equivalent when there are two alleles, the best estimates when there are three or more alleles are based upon total heterozygosity ($\hat{f}_1 = \frac{x-y}{x}$ where x and y are the expected and observed number of heterozygotes) and the proportion of alleles that are homozygous ($\hat{f}_2 = \frac{1}{k-1}\left[\sum_i \frac{a_{ii}}{\sum_j a_{ij}} - 1\right]$ where $k =$ the number of alleles, $a_{ii} =$ the number of A_iA_i homozygotes, and $2a_{ij} =$ the number of A_iA_j heterozygotes). Both are minimally biased estimates of f and have identical sampling variances when all alleles are equally frequent. However, when alleles have different frequencies, the choice between these two estimates depends on the gene frequencies and the true inbreeding coefficient of a population; $\hat{f}_2$ is the best estimate when the true average inbreeding coefficient is suspected to be low or $f = 0$, while $\hat{f}_1$ is best in populations with large average inbreeding coefficients. Approximate sampling variances of these two estimates are given for any f and any number of alleles with arbitrary gene frequencies; these approximations are accurate for samples as small as $n = 100$. The chi-square and maximum likelihood estimates of f are not as good for realistic sample sizes.

INBREEDING depression can threaten the survival of a small population. Inbreeding depression was recognized early by plant and animal breeders (WRIGHT 1977), but this problem has only recently been recognized in zoo populations (RALLS, BRUGGER and BALLOU 1979; SENNER 1980) and in the management and restocking of endangered populations in the wild. Unfortunately, pedigrees are not usually available for individuals within these populations so that inbreeding cannot be directly detected. However, the average inbreeding coefficient (f) of a population can be measured indirectly from genotypic data.

Inbreeding is expected to increase the number of homozygotes, thus decreasing the number of heterozygotes in a population. A comparison of the number of observed heterozygotes and the number expected under random mating can be used to measure the average inbreeding of a population. However, the expected number of heterozygotes may be determined in several different ways;

for example, it may be calculated from a founding population's gene frequencies (reflecting genetic drift) or from the parental gene frequencies (minimizing the effects of drift). To be pragmatic, a typical field study involves capturing a sample from a wild population and assaying the captured animals for genetic polymorphisms. In this situation the expected number of heterozygotes is calculated from the gene frequencies observed in that sample, and the number of heterozygotes over repeated samples could be used to measure genetic drift.

The interpretation of the average inbreeding coefficient will depend on the way in which it is measured. When inbreeding is measured from a single sample, the estimate of inbreeding ($\hat{f}$) measures the deviation from random pairing of the genes, and it is this case which we will now consider further.

When only a single locus is assayed, an unrealistically large sample size is needed to detect small but significant deviations in heterozygosity (WARD and SING 1970); at a locus with two alleles, the x^2-test can detect an inbreeding coefficient of $f = 0.0001$ (a realistic value for human populations) at the 5% significance level only 50% of the time in a sample as large as 4×10^8; almost twice the population of the U.S. However, it is now feasible to assay many immunogenetic and biochemical polymorphisms simultaneously. The average inbreeding coefficient f can be estimated at each locus, and these estimates then averaged, weighted by the reciprocal of each estimate's sampling variance (KIDD *et al.* 1980). An even more accurate estimate of inbreeding can be obtained from the joint distribution of heterozygosity at many loci. This has been demonstrated for two tightly-linked loci (YASUDA 1968) and the approach can be extended to multiple loci.

Multi-locus measures of average inbreeding are more closely related than single locus measures to the individual inbreeding coefficients measured from pedigrees. Inbreeding measured from genotypic data is conceptually different from inbreeding measured from pedigrees; average inbreeding measured from genotypic frequencies is independent of an individual's pedigree, while inbreeding from pedigrees is independent of an individual's genotype. To illustrate, individuals with the same pedigree inbreeding coefficient may differ as to whether their genes are identical by descent (WEIR, AVERY and HILL 1980). Moreover, since genes which are identical by descent cannot usually be identified, the observed number of homozygotes (or heterozygotes) is affected not only by consanguinity (inbreeding) but also by selection, genetic drift, assortative mating, and other evolutionary forces (COCKERHAM 1973). However, if a large number of loci could be assayed, including linked loci and many loci with rare alleles, it might be possible to estimate an individual's (pedigree) inbreeding coefficient solely from genotypic data.

Nevertheless, as a first step toward an accurate and efficient measure of inbreeding in a small population, it is expedient to initially resolve the best single locus measure of inbreeding and to determine its sampling variance. This paper will compare the sampling means and variances of various measures of inbreeding calculated from the genotypic data at a single locus.

A measure of inbreeding: One locus with two alleles

Assume that a random sample is drawn from a real population. We may wish to compare this sample with what would be expected from an "*idealized inbred population*". An idealized inbred population is defined as an infinitely large population with no mutation, migration, or selection (so that gene frequencies remain constant) and with random mating except for a fixed amount of inbred matings resulting in an average inbreeding coefficient of f. The sampling of a whole population may be considered as a random sample from an infinitely large pool of zygotes. In an idealized inbred population, only the fixed amount of inbreeding (consanguinity) will affect the proportion of heterozygotes. For an autosomal codominant locus having two alleles A_1 and A_2 with frequencies p and q $(p+q=1)$, the proportions of homozygotes are p^2+pqf and q^2+pqf for A_1A_1 and A_2A_2, and the proportion of A_1A_2 heterozygotes is $2pq\ (1-f)$ (CROW and KIMURA 1970). In a random sample of n individuals, let a_{11}, $2a_{12}$, and a_{22} be the numbers of individuals who have genotype A_1A_1, A_1A_2, and A_2A_2 respectively $(a_{11}+2a_{12}+a_{22}=n)$. The observed and expected numbers of each genotype in a sample can be summarized in the following table:

genotype:	A_1A_1	A_1A_2	A_2A_2
observed number:	a_{11}	$2a_{12}$	a_{22}
expected number:	$(p^2+pqf)n$	$2pq(1-f)n$	$(q^2+pqf)n$

Sampling from an idealized population is equivalent to multinomial sampling with probabilities $Q_1=p^2+pqf$, $Q_2=2pq(1-f)$, and $Q_3=q^2+pqf$. The true values of p, q $(=1-p)$ and f are presumably unknown (even if a whole population is sampled), but unbiased estimates of the gene frequencies are

$$\hat{p}=\frac{a_{11}+a_{12}}{n} \text{ and } \hat{q}=\frac{a_{12}+a_{22}}{n}\ :$$

Using these estimates for p and q, f is often estimated from the deviation in heterozygosity from that expected under random mating:

$$\hat{f}=\frac{x-y}{x}=\frac{a_{11}a_{22}-a_{12}^2}{(a_{11}+a_{12})(a_{12}+a_{22})} \qquad (1)$$

where $y=2a_{12}$ is the observed number of heterozygotes and $x=2\hat{p}\hat{q}n$ is the number expected assuming random mating $(f=0)$.

While the estimate $\hat{f}$ is a consistent statistic ($\hat{f}$ approaches f as n gets larger), $\hat{f}$ is biased (the expected value of $\hat{f}$ is not f). The expected value $E(\hat{f})$ is not known precisely, but the expected value of the numerator $(x\text{-}y)$ can be calculated exactly.

In general, for a sample of size n from a multinomial distribution with m distinct classes, it is well known that the number observed in the i^{th} class, a_i ($a_1 + a_2 + \ldots + a_m = n$), has mean and variance

$$E(a_i) = nQ_i \ ,$$
$$V(a_i) = nQ_i(1-Q_i)$$
$$\text{and } \operatorname{cov}(a_i,a_j) = -nQ_iQ_j (i \neq j)$$

where Q_i is the probability of a random individual belonging to the i^{th} class. The higher degree moments and product moments of the multinomial distribution are given by KENDALL and STUART (1977, p. 149).

At a locus with two alleles, the expected number of observed heterozygotes is $E(y) = E(2a_{12}) = 2pq(1-f)n$ from binomial sampling theory. Similarly, the mean number of heterozygotes from an idealized population with random mating has the expected value

$$E(x) = 2npq - pq(1+f) \ .$$

Thus, the expected departure in heterozygosity is

$$E(x-y) = 2pqfn - pq(1+f)$$
$$= -pq[1-(2n-1)f] \ .$$

While only the numerator of Eq. (1) is evaluated here, by the Taylor's series expansion (KENDALL and STUART 1977, p. 246), the expected value of $\hat{f}$ is:

$$E(\hat{f}) = f + O\left(\frac{1}{n}\right)$$

where $O\left(\frac{1}{n}\right)$ are terms of order $\left(\frac{1}{n}\right)$. The expected value of $\hat{f}$ was computed exactly for sample sizes $n = 50$ and $n = 100$. When $f = 0$, $E(\hat{f})$ is independent of the gene frequencies and

$$E(\hat{f}) \simeq f - \frac{1}{2n}.$$

This equation also holds for a wide range of values for f when the two alleles are equally frequent. However, if $f = 0.2$ for example, the bias is approximately $-1/n$ when $p = 0.9$.

The sampling variance of $\hat{f}$ can also be approximated by the Taylor's series expansion (KENDALL and STUART 1977, p. 247),

$$V(\hat{f}) = V\left(\frac{y}{x}\right)$$
$$= \frac{E^2(x)V(y) + E^2(y)V(x) - 2E(x)E(y)\operatorname{cov}(x,y)}{E^4(x)} + O\left(\frac{1}{n^2}\right). \quad (2)$$

From binomial sampling theory,

$$V(y) = 2npq(1-f)[1-2pq(1-f)], \text{ and}$$
$$V(x) = 2npq(p-q)^2(1+f) + 2pq\{(1+f)[pq(1+f) - 2q(p-q)] + (p-q)[1+3f-4pf]\} + \frac{pq}{2n}[1-6pq+(7-36pq-6pqf)f] .$$

In addition, from the product-moments of the multinomial distribution,

$$\text{cov}(x,y) = 2npq(p-q)^2(1-f) - pq(1-f)[(p-q)^2 - 2pq(1+f)] .$$

Thus, in a large sample, the variance of $\hat{f}$ is approximated by

$$V(\hat{f}) = \frac{(2n)^3(1-f)[1-2pq(1-f) - (p-q)^2(1-f)^2]}{pq(2n-1-f)^4} + O\left(\frac{1}{n^2}\right) \simeq \frac{(1-f)[1-2pq(1-f) - (p-q)^2(1-f)^2]}{2pqn} . \quad (3)$$

In Figure 1, the variance of $\hat{f}$, $V(\hat{f})$, times the sample size is plotted against f for

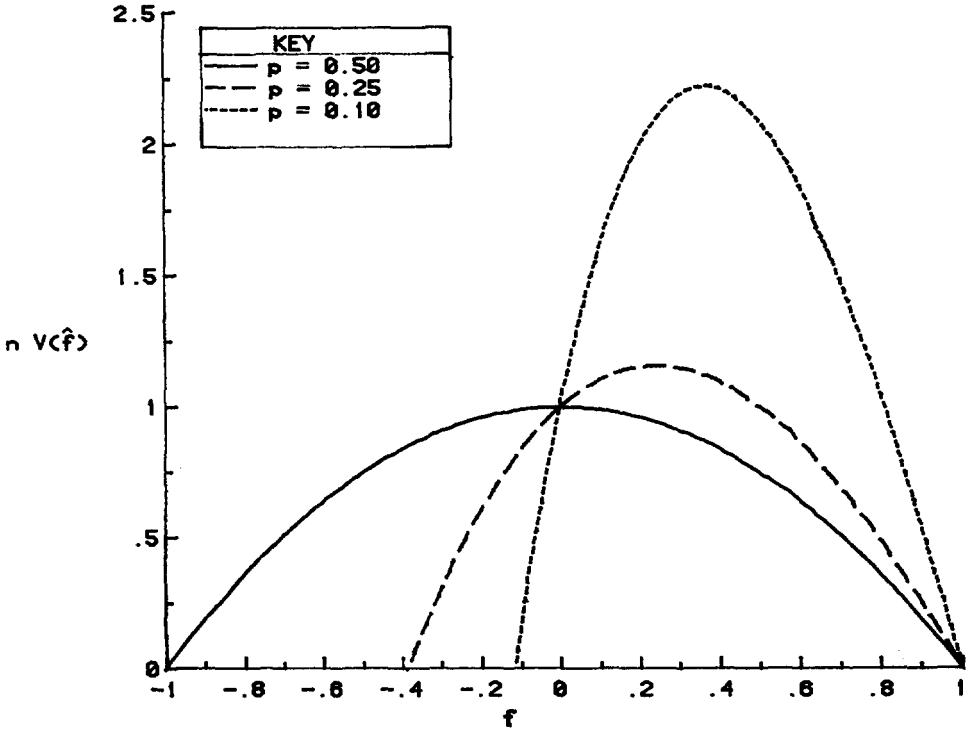

FIGURE 1.—The sampling variance of $\hat{f}$, $V(\hat{f})$, times the sample size (n) as a function of the average inbreeding coefficient f for a population with two alleles at a locus. When $f = 0$, $V(\hat{f}) = \frac{1}{n}$, but when $f \neq 0$, $V(\hat{f})$ depends on the gene frequency, shown here for $p = 0.5$, 0.25 and 0.10.

three different values of the gene frequency, p. The range of f is limited between $-\frac{p}{1-p}$ and 1 (where $p \leqslant q$) since the homozygote frequencies cannot be negative, *i.e.* $p^2 + p(1-p)f \geqslant 0$. In general, the sampling variance of $\hat{f}$ depends both on f and p; however, when $f = 0$, then $V(\hat{f})$ is independent of the gene frequencies and, as shown by YASUDA (1968),

$$V(\hat{f}) = \frac{(2n)^4}{n(2n-1)^4} + O\left(\frac{1}{n^2}\right) \simeq \frac{1}{n} .$$

When $f \neq 0$, then $V(\hat{f})$ can vary dramatically and it greatly increases with small increases in f when one allele is rare.

Measures of inbreeding: Multiple alleles

Assume that locus A has k codominant alleles $(A_1, A_2, \ldots, A_k)$ in an idealized inbred population where the frequency of A_i is p_i $(p_1 + p_2 + \ldots + p_k = 1)$. A random sample from this population will follow a multinomial distribution with $\frac{k(k+1)}{2}$ classes (k homozygote and $\frac{k(k-1)}{2}$ heterozygote types). If n individuals are randomly sampled, the observed and expected numbers of each genotype are

genotype:	A_iA_i	A_jA_j $(i \neq j)$
observed number:	a_{ii}	$2a_{ij}$
expected number:	$[p_i^2 + p_i(1-p_i)f]n$	$2p_jp_j(1-f)n$

An unbiased estimate of the gene frequency p_i $(i = 1, 2, \ldots, k)$ is

$$p_i = \frac{1}{n} \sum_{j=1}^{k} a_{ij}$$

and the sampling variance of this estimate is given by

$$V(\hat{p}_i) = \frac{1}{n^2}\left[\sum_{j=1}^{k} V(a_{ij}) + 2\sum_{j=1}^{k-1}\sum_{l=j+1}^{k} \operatorname{cov}(a_{ij}, a_{il})\right] = \frac{p_i(1-p_i)}{2n}(1+f) .$$

The gene frequency estimate is analogous to the two allele case. However, the average inbreeding coefficient f can be estimated in several different ways, each yielding a different numeric value. LI and HORVITZ (1953) described eight methods of estimating f from the genotypic data in a single sample. While all eight yield estimates with identical values of f for a two-allele locus, they give 12 different values for a three-allele locus (two methods yield three possible values). Since they did not calculate the sampling variances of these estimates, they could not recommend one estimate over the others.

Unfortunately, all eight methods proposed by LI and HORVITZ produce biased estimates of f. HALDANE (1954) suggested a ninth method which he believed unbiased but he was misled by an error in computation (SMITH 1970). No unbiased estimate of inbreeding has yet been demonstrated.

The four most natural measures proposed by LI and HORVITZ are the estimates based on total heterozygosity and on the proportion of alleles that are homozygous, the χ^2-estimate, and the maximum likelihood estimate; the sampling properties of these four estimates will now be compared both theoretically and by computer simulation. Other measures will be considered in the discussion.

The total heterozygosity, $\hat{f}_1$: The number of heterozygotes in a population sample is easily calculated for a locus with codominant alleles. If we assume that the sample is drawn from an idealized inbred population, the total number of heterozygotes is expected to be $(1-f)$ times the number expected under random mating, so that f may be estimated by an extension of Equation (1):

$$\hat{f}_1 = \frac{x-y}{x} \tag{4}$$

where $y = 2\sum_{i<j} a_{ij}$ is the observed number of heterozygotes and $x = 2n\sum_{i<j} \hat{p}_i\hat{p}_j$ is the number expected under random mating.

In a large sample, the variance of this estimate can be approximated as in Equation (2) from the sampling properties of x and y and their covariance (see Appendix 1), so that

$$V(\hat{f}_1) \simeq \frac{(1-f)\left[\sum_{i<j} 2p_ip_j - (1-f)\left(\sum_{i<j} 2p_ip_j\right)^2 - (1-f)^2\sum_{i<j} 2p_ip_j(p_i-p_j)^2\right]}{n\left(\sum_{i<j} 2p_ip_j\right)^2}. \tag{5}$$

In a population with two alleles, $V(\hat{f}_1)$ reduces to Equation (3) as expected. In addition, two other special cases are often encountered and will be useful for comparison, *i.e.*, populations in which

i) all alleles are equally frequent ($p_i = \frac{1}{k}$ for $i = 1, \ldots, k$) so that

$$V(\hat{f}_1) \simeq \frac{(1-f)[1+f(k-1)]}{n(k-1)}, \text{ and}$$

ii) $f = 0$ so that

$$V(\hat{f}_1) \simeq \frac{\sum_{i<j} 2p_ip_j(1-\sum_{i<j} 2p_ip_j) - \sum_{i<j} 2p_ip_j(p_i-p_j)^2}{n\left(\sum_{i<j} 2p_ip_j\right)^2}.$$

The proportion of alleles homozygous, $\hat{f}_2$: Rather than just consider the total number of heterozygotes or homozygotes, the observed numbers of each homozygote class can be compared separately to the number expected under random mating. If we divide the proportion of each homozygote by the gene frequency

of its corresponding allele, the sum of these ratios should equal $1 + f(k-1)$ in an idealized inbred population. This formula can be solved for f to produce a second estimate of a population's average inbreeding coefficient

$$\hat{f}_2 = \frac{1}{k-1}\left[\sum_{i=1}^{k} \frac{\overline{a_{ii}}}{\hat{p}_i} - 1\right] \qquad (6)$$
$$= \frac{1}{k-1}\left[\sum_{i=1}^{k} \frac{a_{ii}}{\sum_{j=1}^{k} a_{ij}} - 1\right] .$$

Alternatively, each homozygote could be weighted by a value w_i, but ROBERTSON and HILL (1981) have shown that (for $f=0$) $\hat{f}_2$ has the smallest variance of all estimates having the form

$$\hat{f} = \sum_i w_i \left(\frac{a_{ii} - n\hat{p}_i^2}{n\hat{p}_i^2}\right)$$

or equivalently

$$\hat{f} = \sum_{i<j} w_{ij} \left(\frac{2a_{ij} - 2n\hat{p}_i\hat{p}_j}{2n\hat{p}_i\hat{p}_j}\right) .$$

This measure essentially estimates f from each homozygote and averages the different estimates, weighted to minimize the total variance.

Again, the variance $V(\hat{f}_2)$ cannot be calculated exactly but can be approximated (see Appendix 2) so that

$$V(\hat{f}_2) \simeq \frac{1-f}{2n(k-1)^2}\left\{2(k-1) - 2(2k-1)f + k^2f^2 + f(2-f)\sum_{i=1}^{k}\frac{1}{p_i}\right\} . \qquad (7)$$

Again, $V(\hat{f}_2)$ reduces to Equation (3) in a population with two alleles. The two other cases of special interest are:

i) when all alleles are equally frequent ($p_i = \frac{1}{k}$ for every i) then

$$V(\hat{f}_2) \simeq \frac{(1-f)[1+f(k-1)]}{n(k-1)} .$$

Thus, the variances $V(\hat{f}_1)$ and $V(\hat{f}_2)$ are the same in large samples from such a population; and

ii) when $f = 0$, then

$$V(\hat{f}_2) \simeq \frac{1}{(k-1)n}$$

so that this variance is independent of gene frequencies when $f=0$; this case agrees with that discussed by ROBERTSON and HILL (1981).

The Chi-Square Estimate, $\hat{f}_3$: The chi-square goodness of fit test has been used in two different methods of estimating the average inbreeding coefficient f. One estimate is the value of f which minimizes the χ^2-value when testing the observed numbers of each genotype against the number expected in an idealized inbred population with average inbreeding coefficient f (WRIGHT; quoted in LI and HORVITZ 1953). For two alleles, a value of f can always be found which exactly fits the data (*i.e.* the χ^2 has zero degrees of freedom) and this estimate of f agrees with the other two-allele estimates. This method uses a maximum likelihood approach but maximizes a less precise probability distribution of the genotype numbers than the maximum likelihood method discussed in the next section and seems, therefore, inferior.

More commonly, f is estimated from the χ^2-value of testing the observed numbers of each genotype against those expected under random mating ($f = 0$). By setting $a_{ii} = [p_i^2 + p_i(1-p_i)f]n$ as in an infinitely large sample, the value of χ^2 becomes

$$\chi^2 = \sum_{i=1}^{k} \frac{(a_{ii}-np_i)^2}{np_i^2} + \sum_{i<j} \frac{(2a_{ij}-2np_ip_j)^2}{2np_ip_j}$$
$$= n(k-1)f^2$$

with $k(k-1)/2$ degrees of freedom. Solving this formula for f, a third estimate of average inbreeding is taken as the positive root

$$\hat{f}_3 = \sqrt{\frac{\chi^2}{n(k-1)}} \,. \tag{8}$$

This estimate is very appealing since the chi-square value can simultaneously provide an estimate of f and a significance test of the hypothesis $f = 0$ (WARD and SING 1970). However the estimate $\hat{f}_3$ differs in interpretation from the previous two estimates. While $\hat{f}_1$ and $\hat{f}_2$ may be positive or negative in value (negative values indicating an excess of heterozygotes) the chi-square estimate $\hat{f}_3$ is always greater than or equal to zero. It is not practical to adjust the sign of this estimate to agree with the other two estimates, since $\hat{f}_1$ and $\hat{f}_2$ may also differ in sign when there is an excess of one heterozygote and a deficiency of another. While $\hat{f}_1$ and $\hat{f}_2$ try to measure a uniform or overall departure in the number of heterozygotes or homozygotes, $\hat{f}_3$ detects any departure from random mating proportions, even if the total number of heterozygotes remains constant.

Since $\hat{f}_3$ is the positive root of a quadratic equation, its sampling properties are much more difficult to determine. However, the squares of these three estimates of f can easily be compared. The variances of $\hat{f}_1^2$ and $\hat{f}_2^2$ are approximately

$$V(\hat{f}_i^2) \simeq 4f^2\, V(\hat{f}_i) \quad (i = 1,2) \,.$$

The statistic χ^2 approximately has a non-central chi-square distribution with

non-centrality parameter $\lambda = nf^2(k-1)$ (WARD and SING 1970; HABER 1980), so that

$$V(\hat{f}_3^2) = [k(k-1) + 4nf^2(k-1)]/n^2(k-1)^2$$
$$= \frac{4f^2}{n(k-1)} + \frac{k}{n^2(k-1)} .$$

However, the chi-square distribution is only approximate and the variance of $\hat{f}_3^2$ can be more precisely approximated by again using the Taylor's series expansion so that

$$V(\hat{f}_3^2) = \frac{2f^2(1-f)}{n(k-1)^2}\left\{2(k-1) - 2(2k-1)f + k^2f^2 + f(2-f)\sum_{i=1}^{k}\frac{1}{p_i}\right\} + O\left(\frac{1}{n^2}\right)$$
$$\simeq 4f^2V(f_2) . \qquad (9)$$

Thus, the sampling variance $V(\hat{f}_3^2)$ is approximately equal to $V(\hat{f}_2^2)$. As a result, we would expect the variance $V(\hat{f}_3)$ to be less than or equal to $V(\hat{f}_2)$ since $\hat{f}_3$ is never less than zero and thus has a narrower range than $\hat{f}_2$. Note that the variance of the noncentral chi-square distribution provides a good approximation of the variance $V(\hat{f}_3^2)$ only when n is large, f is small, and the p_i's are moderate. When f is close to zero, the first term of Equation (9) vanishes and the terms of order $\frac{1}{n^2}$ must be evaluated.

The maximum likelihood estimate, $\hat{\hat{f}}_4$: Maximum likelihood estimates are often preferred because they are sufficient statistics and will attain the minimum variance as the sample size gets infinitely large (*e.g.* FREUND 1962). However, maximum likelihood estimates may be biased and do not necessarily have the minimum variance in samples with a more realistic size. Unfortunately, the maximum likelihood estimate of the average inbreeding coefficient cannot be explicitly written, but must be solved numerically by iteration. If the likelihood of the observed numbers of each genotype are maximized simultaneously for the gene frequencies p_i $(i = 1, \ldots, k)$ and f, the gene frequency estimates are not generally equal to the natural unbiased estimates $\hat{p}_i = \frac{1}{n}\sum_{j=1}^{k} a_{ij}$. However, if the gene frequencies are fixed by these unbiased estimates, then the maximum likelihood estimate $\hat{f}_4$ may be obtained by solving the following equation for Θ:

$$\sum_{i=1}^{k}\frac{a_{ii}(1-\hat{p}_i)}{\hat{p}_i + \Theta} = 2\sum_{i<j} a_{ij} \qquad (10)$$

where $\hat{f}_4 = \frac{\Theta}{1+\Theta}$ (LI and HORVITZ 1953). Since $\hat{f}_4$ cannot be explicitly written, its sampling properties are most difficult to evaluate and will be calculated only empirically.

Simulation results

A computer simulation was used to confirm and extend the comparisons among these four estimates of inbreeding. An idealized inbred population was

sampled and each of the four estimates was calculated. This sampling was repeated 500 times and the mean and variance for each estimate was calculated. The populations were assumed to have three alleles with fixed gene frequencies (either $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$ or $p_1 = p_2 = 0.475$, $p_3 = 0.05$) while the size of the sample or the actual average inbreeding coefficient f was allowed to vary.

Figure 2 shows the average values of the four estimates when the sample sizes varied between 50 and 500, in steps of 50. The data are 500 repeats of a computer simulated sampling from an idealized inbred population with gene frequencies $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$ and with average inbreeding coefficient $f = 0.05$. The average values of the estimates $\hat{f}_1$ and $\hat{f}_2$ are nearly identical and show very little bias, even in samples as small as $n = 100$. The average value of the maximum likelihood estimate $\hat{f}_4$ differs from the first two only when $n = 50$. However, the chi-square estimate $\hat{f}_3$ drastically overestimates f for small samples and even overestimates by 50% for samples as large as $n = 500$. This was expected since $\hat{f}_3$ is always greater than or equal to zero and will detect any deviations from a random assortment of genes. An almost identical curve was produced by sampling an idealized inbred population with gene frequencies $p_1 = p_2 = 0.475$, $p_3 = 0.05$ and

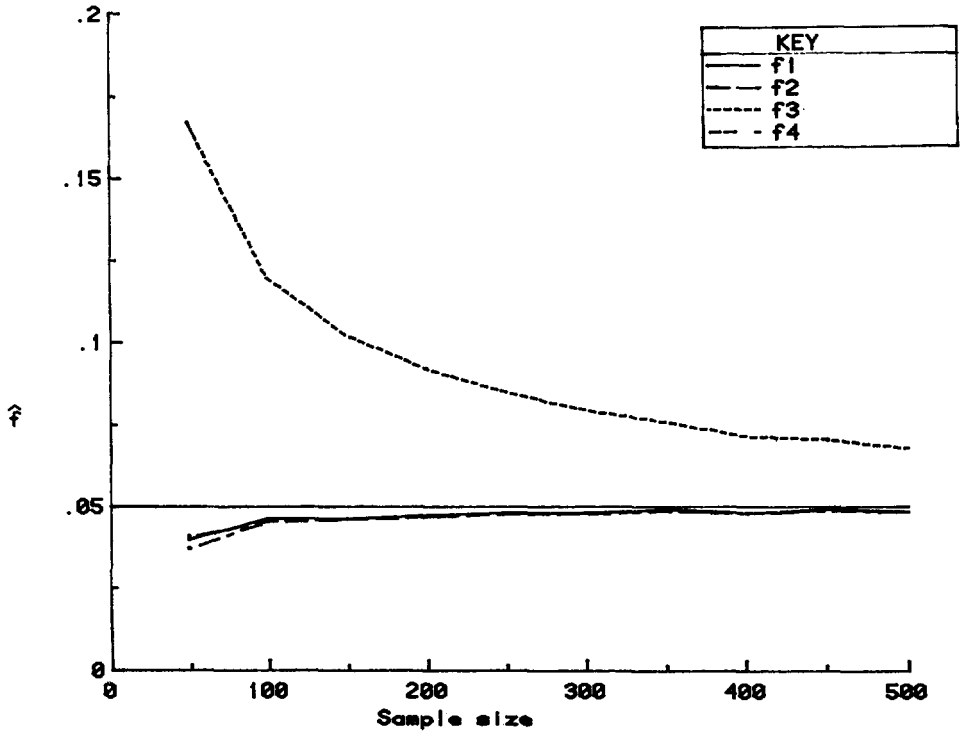

FIGURE 2.—The average values for each of the four estimates of f in a computer simulation where the true value of the average inbreeding coefficient is $f = 0.05$. For each sample size between 50 and 500 (in steps of 50), sampling was repeated 500 times from an idealized inbred population ($p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$, and $f = 0.05$) and the results averaged.

average inbreeding coefficient $f = 0.05$. Similar but less drastic results were observed by setting $f = 0.2$. In these simulations, the sample variances of the four estimates (not shown) were approximately a constant function of $\frac{1}{n}$ in the range $n = 100$ to 500. Thus, the approximations of $V(\hat{f}_i)$, by ignoring the terms of order $\frac{1}{n^2}$, appear valid for sample sizes as small as n = 100.

Simulations were also performed to study the relationships among the four estimates as a function of the true value of f. For these simulations, the size of each sample was fixed at 200, which seemed large enough to minmize bias and allow safe prediction yet small enough to be realistic.

Figure 3 shows the deviation of the average values of each estimate of inbreeding from the true value of f (*i.e.* the average bias) as a function of f. The data are 500 repeats of the sampling of an idealized inbred population with gene frequencies $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$ and the average inbreeding coefficient varying from its minimum value $(-\frac{p_{min}}{1 - p_{min}} = -0.25)$ to 1 in steps of 0.05. The x^2 estimate $\hat{f}_3$ drastically overestimates f when f is small or negative ($\hat{f}_3 \simeq 0.26$ when $f = -0.25$). When f is large, samples rarely produce negative estimates of

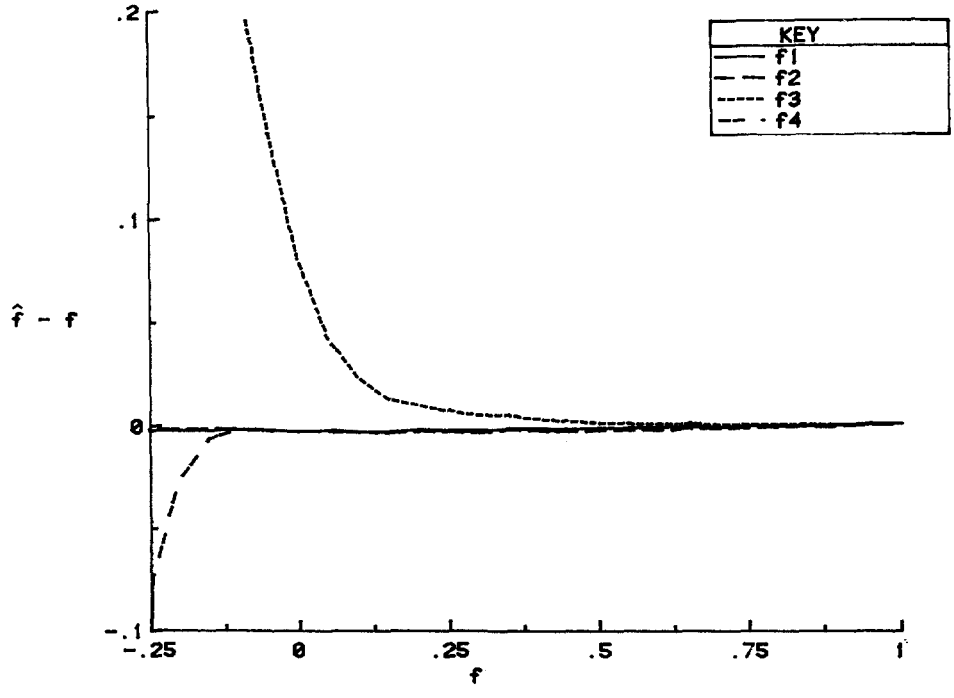

FIGURE 3.—The difference between the average values for each of the four estimates of f and the true value of f ($\hat{f}$—f) as a function of f. For each value of f between —0.25 and 1 (in steps of 0.05), samples of size 200 were repeated 500 times from an idealized inbred population ($p_1 = 0.2$, $p_2 = 0.3$, and $p_3 = 0.5$).

f so that $\hat{f}_3$ more often agrees with the other three estimates. The maximum likelihood estimate $\hat{f}^4$ underestimates f for small negative values of f. The estimate $\hat{f}_1$ and $\hat{f}_2$ (and $\hat{f}_4$ for $f \geqslant -0.1$) have almost identical expected values although they will differ for any given sample; all three on the average, slightly underestimate inbreeding.

Figure 4 shows the product of the variance and the sample size for each estimate since the sampling variances are all approximately proportional to $\frac{1}{n}$. The simulation is the same as that used to generate Figure 3. Estimates $\hat{f}_1$ and $\hat{f}_2$ have almost the same variance although $V(\hat{f}_2) \leqslant V(\hat{f}_1)$ for low values of f. The x^2 estimate $\hat{f}_3$ has the lowest variance, particularly around $f = 0$, but it nearly coincides with $V(\hat{f}_2)$ when f is large enough to effectively eliminate samples with negative values of $\hat{f}_2$. The variance of the maximum likelihood estimate, while high for the smallest negative values of f, nearly coincides with the variance of $\hat{f}_1$ for positive values of f. Figure 5 shows the theoretical values of $V(\hat{f}_1)$ and $V(\hat{f}_2)$ from Equations (5) and (7); They are in excellent agreement with the variances in the computer simulation. In this population, the second estimate $\hat{f}_2$ is better for small values of f and $\hat{f}_1$ is slightly better for large f while $V(\hat{f}_1) = V(\hat{f}_2)$ when $f = 0.207$.

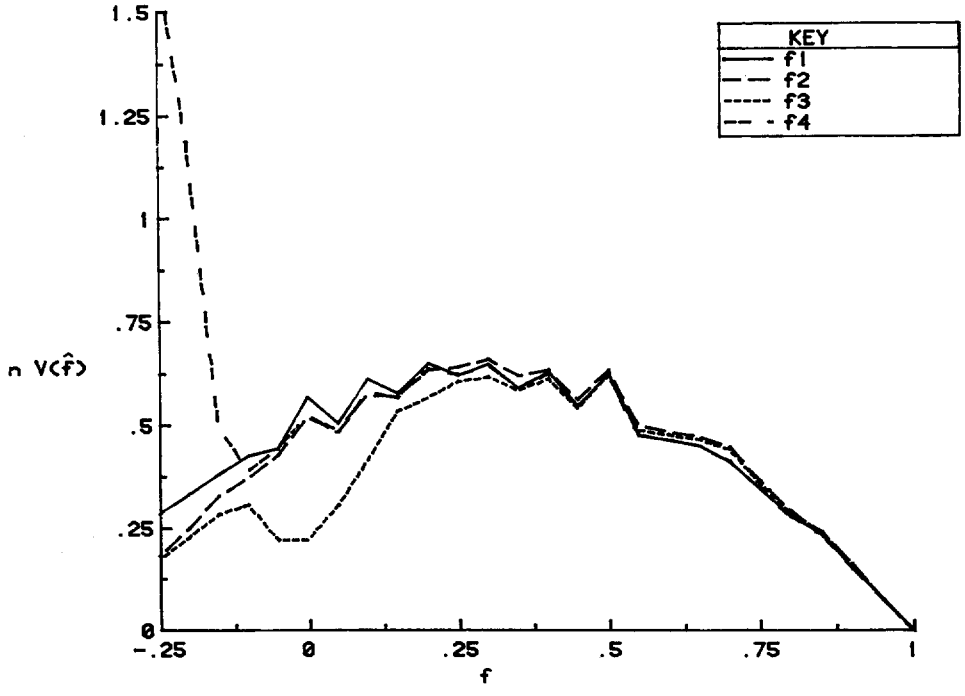

FIGURE 4.—The sampling variances $V(\hat{f})$ times the sample size (n) for each of the four estimates of f from the same computer simulation as in Figure 3.

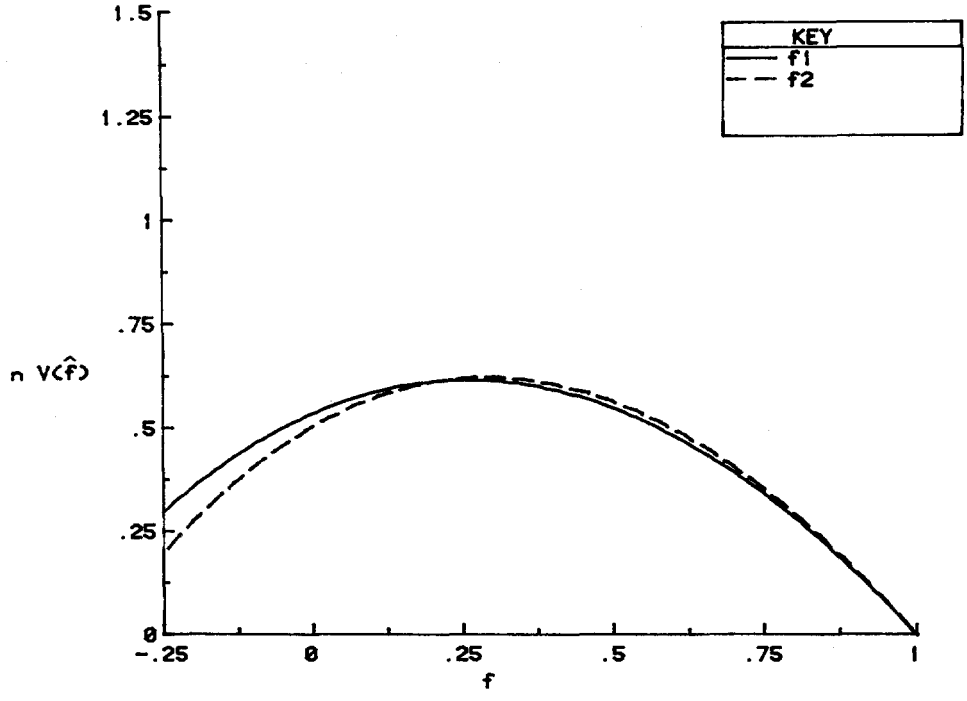

FIGURE 5.—The theoretical sampling variances $V(\hat{f}_1)$ and $V(\hat{f}_2)$ times the sample size (n) as a function of the average inbreeding coefficient f for a population with three alleles ($p_1 = 0.2$, $p_2 = 0.3$, and $p_3 = 0.5$).

The differences between the estimates are more pronounced when one allele is rare. Figures 6 and 7 show the observed and theoretical sampling variances for these estimates when $p_1 = p_2 = 0.475$ and $p_3 = 0.05$. Again, the theoretical variances of $\hat{f}_1$ and $\hat{f}_2$ are in excellent agreement with the variances in the computer simulation. The difference between $V(\hat{f}_1)$ and $V(\hat{f}_2)$ is more pronounced than in the previous population and the two variances are equal for a smaller value of f; $V(\hat{f}_1) = V(\hat{f}_2)$ when $f = 0.077$. The chi-square variance $V(\hat{f}_3)$ is again the lowest for small f and nearly coincides with $V(\hat{f}_2)$ for high f; the maximum likelihood variance $V(\hat{f}_4)$ is highest for small f and nearly coincides with $V(\hat{f}_1)$ for high f.

These four estimates have been compared by computer simulated sampling from an idealized inbred population with two different sets of gene frequencies and varying values of the average inbreeding coefficient f. In nature , however, many factors may affect genotype frequencies so that each heterozygote may depart from random mating proportions in a different way. To test this situation Figure 8 shows the variances for the four estimates of inbreeding as a function of sample size from a population with gene frequencies $p_1 = 0.2$, $p_2 = 0.3$ and $p_3 = 0.5$ but genotype frequencies:

A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_2A_3	A_3A_3
0.04	0.24	0.08	0.03	0.30	0.31

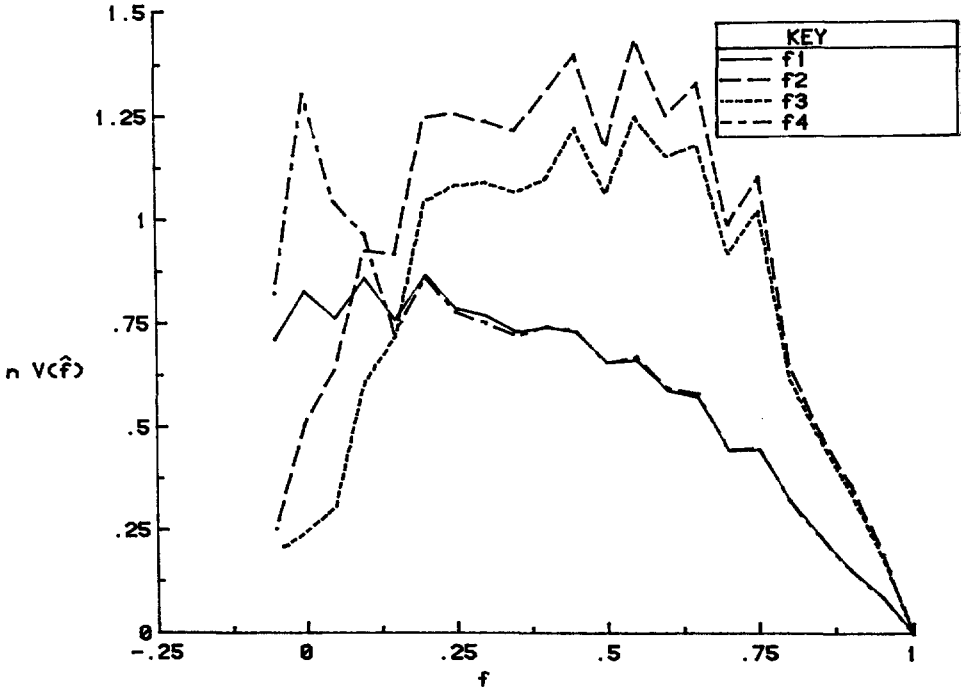

FIGURE 6.—The sampling variances $V(\hat{f})$ times the sample size (n) for each of the four estimates of f as a function of f. For each value of f between $f_{min} = -0.05$ and 1 (in steps of 0.05), samples of size 200 were repeated 500 times from an idealized inbred population ($p_1 = p_2 = 0.475$ and $p_3 = 0.05$).

This population has an excess of some heterozygotes and a deficiency of others, while the total heterozygosity is that expected under random mating. The maximum likelihood estimate $\hat{f}_4$ has the largest variance, while the other three estimates have similar variances which are roughly constant functions of $\frac{1}{n}$ between $n = 50$ and 500. The average values of the estimates (not shown) are also constant between $n = 50$ and 500. As expected $\hat{f}_1 \simeq 0$, but $\hat{f}_2 \simeq \hat{f}_4 \simeq -0.05$ and $\hat{f}_3 \simeq 0.35$.

DISCUSSION

Four estimates of f have been considered here. While a locus with three alleles has been considered, the results presented here are qualitatively the same for larger numbers of alleles. No single estimate is always the best when there are more than two alleles. The two best estimates are those based on total heterozygosity ($\hat{f}_1$) and the proportion of alleles that are homozygous ($\hat{f}_2$). Both are nearly unbiased, even in small samples, and both have identical variances when all alleles are equally frequent. However, when alleles have different frequencies in the population, then $\hat{f}_1$ has a smaller variance for large values of f whereas $\hat{f}_2$ has

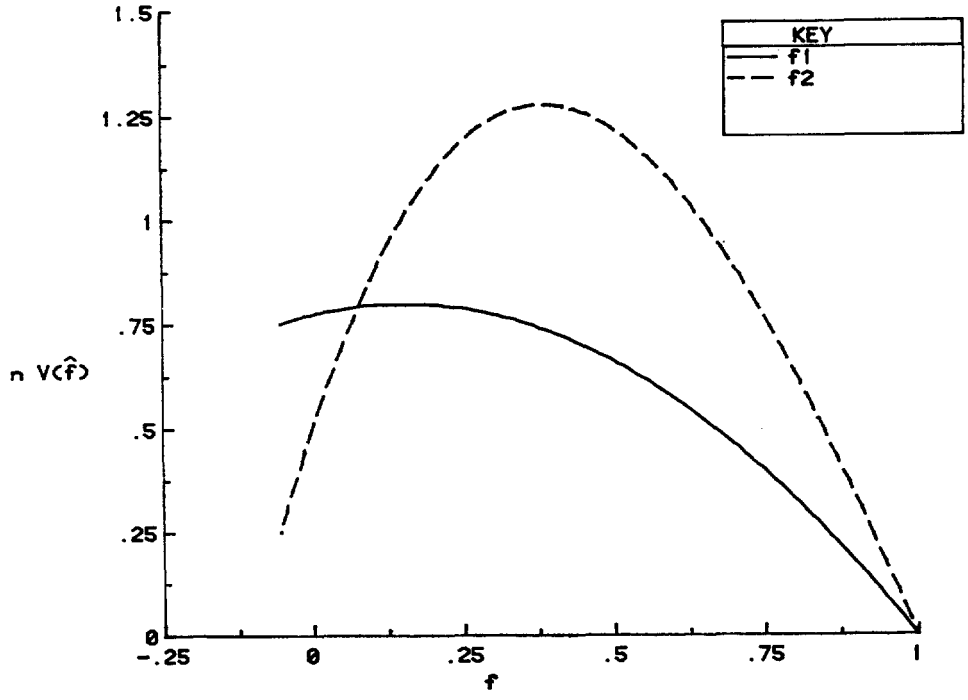

FIGURE 7.—The theoretical sampling variances $V(\hat{f}_1)$ and $V(\hat{f}_2)$ times the sample size (n) as a function of f for a population with three alleles ($p_1 = p_2 = 0.475$ and $p_3 = 0.05$).

a smaller variance for small f including $f = 0$. Moreover, as the gene frequencies become more disparate, the two variances differ more drastically and are equal at a value closer to $f = 0$.

The chi-square estimate ($\hat{f}_3$) has a variance which is less than the variance of $\hat{f}_2$ and nearly equal for large f, as expected (Figures 4 and 6). However, $\hat{f}_3$ is always positive and measures any departure from random mating. Thus, it tends to overestimate the average inbreeding coefficient, particularly for a small f and a small sample size. Consequently, $\hat{f}_3$ is not a good estimate of the average inbreeding coefficient of a population, despite its popular use.

The maximum likelihood estimate ($\hat{f}_4$) is also more biased than $\hat{f}_1$ or $\hat{f}_2$ (Figures 2 and 3) and has the largest variance for small values of f. Otherwise the variance $V(\hat{f}_4)$ is nearly identical to $V(\hat{f}_1)$. This differs from the finding of YASUDA (1968) that the variance of $\hat{f}_4$ approximates $V(\hat{f}_2)$ at $f = 0$.

LI and HORVITZ (1953) proposed three additional measures which have not been discussed here. Their estimate, based on the gametic determinant, actually estimates f^2 (for $k = 3$) and is thus subject to the same problems as the chi-square estimate. The remaining two methods, based on the product-moment correlation and on the reduction to two alleles, do not give unique estimates of f; the first

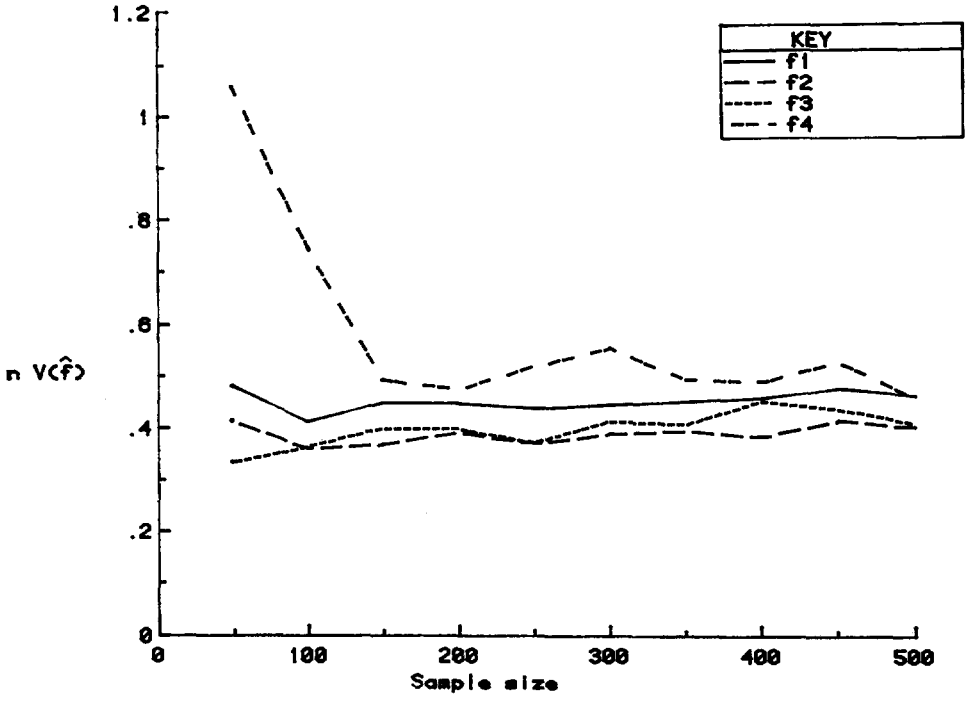

FIGURE 8.—The sampling variances $V(\hat{f})$ times the sample size (n) for a population with three alleles and genotype frequencies $P(A_1A_1) = 0.04$, $P(A_1A_2) = 0.24$, $P(A_1A_3) = 0.08$, $P(A_2A_2) = 0.3$, $P(A_2A_3) = 0.30$ and $P(A_3A_3) = 0.31$. For each sample size between 50 and 500 (in steps of 50), sampling was repeated 500 times.

depends upon the choice of allelic values and the second upon the partition of alleles into two classes. Moreover, reduction to two alleles would appear to increase the variance since the variances of the estimates examined here are inversely related to the number of alleles. Consequently these other measures seem less desirable than the four studied here.

There are, of course, other possible measures of average inbreeding which may prove less biased or have smaller sampling variances. For example, the χ^2 estimate $\hat{f}_3$ might be improved. EMIGH (1980) recently considered 11 different tests for random mating proportions, including χ^2 and six other statistics closely related to the χ^2. His interest was in tests of hypothesis (*i.e.*, $f = 0$) and tests of significance, not in estimation of f. EMIGH compared the power and sensitivity of these statistics extensively for two alleles and suggested any one of nine tests in various situations. For three alleles, he reduced the situation to tests for the three inbreeding coefficients possible by reducing the locus to two alleles. However, all of these statistics are designed as goodness-of-fit tests; they could be used to derive estimates of f but they would not be qualitatively different from that based on the χ^2 test, $\hat{f}_3$. Nevertheless, other measures of average inbreeding in a population have not yet been examined and must remain for future work.

Finally, in any real situation, the true values of f and $p_1, \ldots, p_k$ are unknown. Thus, the choice of $\hat{f}_1$ or $\hat{f}_2$ may be based on prior knowledge of their values or on preliminary calculations. Whichever estimate is used, its sampling variance $V(\hat{f})$ can be approximated by substituting $\hat{f}_1$ or $\hat{f}_2$ for the unknown value of f.

Paper No. 2472 from the Laboratory of Genetics, University of Wisconsin-Madison. This work was supported by Public Health Service grant RR 01216. I am greatly indebted to C. E. KAHN, JR. for developing the computer simulations, to A. ROBERTSON for the initial approach, to the anonymous reviewers for helpful suggestions, and to one reviewer for the exact formula for $V(x)$ for two alleles.

LITERATURE CITED

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

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

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

FREUND, J. E., 1962 *Mathematical Statistics*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

HABER, M., 1980 Detection of inbreeding effects by the χ^2 test on genotypic and phenotypic frequencies. Am. J. Hum. Genet. **32:** 754–760.

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

KENDALL, M. and A. STUART, 1977 *The Advanced Theory of Statistics*, Vol. 1, 4th Ed. MacMillan Publishing Company, New York.

KIDD, K. K., W. H. STONE, C. CRIMELLA, C. CARENZI, M. CASATI and G. ROGNONI, 1980 Immunogenetic and population genetic analyses of Iberian cattle. Anim. Blood Groups Biochem. Genet. **11:** 21–38.

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

RALLS, K., K. BRUGGER and J. BALLOU, 1979 Inbreeding and juvenile mortality in small populations of ungulates. Science **206:** 1101–1103.

ROBERTSON, A. and W. G. HILL, 1981 Sampling variances of deviations from Hardy-Weinberg proportions. Submitted to Ann. Hum. Genet.

SENNER, J. W., 1980 Inbreeding and the survival of zoo populations. pp. 209–223. In: *Conservation Biology*. Edited by M. SOULE and B. WILCOX. Sinauer Assoc., Sunderland, Massachusetts.

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

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., P. J. AVERY and W. G. HILL, 1980 Effect of mating structure on variation in inbreeding. Theoret. Pop. Biol. **18:** 396–429.

WRIGHT, S., 1977 *Evolution and the Genetics of Populations*, Vol. 3. University of Chicago Press, Chicago.

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

Corresponding editor: B. WEIR

APPENDIX 1

Variance of $\hat{f}_1 = \dfrac{x-y}{x}$

The observed number of heterozygotes y follows a binomial distribution with probability $Q = \sum_{i<j} 2p_ip_j(1-f)$. Thus,

$$E(y) = n[\sum_{i<j} 2p_ip_j(1-f)] = nQ$$

and

$$V(y) = nQ(1-Q)$$

The expected value of x is also found easily:

$$\begin{aligned} E(x) &= nE(1-\sum_{i=1}^{k} \hat{p}_i^2) \\ &= n - n\sum_{i=1}^{k} [V(\hat{p}_i) + E^2(\hat{p}_i)] \\ &= (2n-1-f)\sum_{i<j} p_ip_j \end{aligned}$$

The variance $V(x)$ cannot be calculated exactly, but it can be approximated by use of the Taylor's series expansion (KENDALL and STUART 1977). First, notice that

$$V(\hat{p}_i^2) = V(\hat{p}_i)(2p_i)^2 + O\left(\frac{1}{n^2}\right)$$

$$= \frac{2p_i^3(1-p_i)}{n}(1+f) + O\left(\frac{1}{n^2}\right); \text{ also}$$

$$V(\hat{p}_i\hat{p}_j) = p_i^2V(\hat{p}_j) + p_j^2V(\hat{p}_i) + 2p_ip_j\,\mathrm{cov}(\hat{p}_i\hat{p}_j) + O\left(\frac{1}{n^2}\right)$$

and

$$\mathrm{cov}(\hat{p}_i, \hat{p}_j) = -\frac{p_ip_j}{2n}(1+f)$$

so that

$$\begin{aligned} \mathrm{cov}(\hat{p}_i^2, \hat{p}_j^2) &= V(\hat{p}_i\hat{p}_j) + E^2(\hat{p}_i\hat{p}_j) - E(\hat{p}_i^2)E(\hat{p}_j^2) \\ &= -\frac{2}{n}p_i^2p_j^2(1+f) + O\left(\frac{1}{n^2}\right) . \end{aligned}$$

The sampling variance of x is then

$$\begin{aligned} V(x) &= n^2V\left(\sum_{i=1}^{k} \hat{p}_i^2\right) \\ &= n^2\sum_{i=1}^{k} V(\hat{p}_i^2) + 2n^2\sum_{i<j} \mathrm{cov}(\hat{p}_i^2, \hat{p}_j^2) \\ &= n^2\sum_{i=1}^{k} \frac{2p_i^3(1-p_i)}{n}(1+f) - 2n^2\sum_{i<j}\frac{2}{n}p_i^2p_j^2(1+f) + O(1) \\ &= 2n(1+f)\sum_{i<j} p_ip^j(p_i-p_j)^2 + O(1) . \end{aligned}$$

Similarly, cov (x, y) can be calculated as

$$\mathrm{cov}(x, y) = 2n(1-f)\sum_{i<j} p_ip_j(p_i-p_j)^2 + O(1) .$$

Consequently, the sampling variance of the estimate $\hat{f}_1$ is approximately

$$V(\hat{f}_1) = \frac{E^2(x)V(y) + E^2(y)V(x) - 2E(x)E(y)\,\mathrm{cov}(x,y)}{E^4(x)} + O\left(\frac{1}{n^2}\right)$$

$$= \frac{n^3(1-f)\left(\sum_{i<j} 2p_ip_j\right)^2\left[\sum_{i<j} 2p_ip_j - (1-f)\left(\sum_{i<j} 2p_ip_j\right)^2 - (1-f)^2\sum_{i<j} 2p_ip_j(p_i-p_j)^2\right]}{\left[\sum_{i<j}(2np_ip_j - p_ip_j(1+f))\right]^4} + O\left(\frac{1}{n^2}\right)$$

$$\simeq \frac{(1-f)\left[\sum_{i<j} 2p_ip_j - (1-f)\left(\sum_{i<j} 2p_ip_j\right)^2 - (1-f)^2\sum_{i<j} 2p_ip_j(p_i-p_j)^2\right]}{n\left(\sum_{i<j} 2p_ip_j\right)^2}.$$

APPENDIX 2

$$\text{Variance of } \hat{f}_2 = \frac{1}{k-1}\left[\sum_{i=1}^{k} \frac{\frac{a_{ii}}{n}}{\hat{p}_i} - 1\right]$$

Noting first that

$$V\left(\frac{a_{ii}}{\sum_{j=1}^{k} a_{ij}}\right) = \frac{p_i + (1-p_i)f}{2np_i}\{2 + [p_i+(1-p_i)f][p_i+(1-p_i)(f-3)]\} + O\left(\frac{1}{n^2}\right)$$

and $$\mathrm{cov}\left(\frac{a_{ii}}{\sum_{l=1}^{k} a_{il}}, \frac{a_{jj}}{\sum_{l=1}^{k} a_{jl}}\right) = \frac{1-f}{2n}[p_i+(1-p_i)f][p_j+(1-p_j)f] + O\left(\frac{1}{n^2}\right),$$

the sampling variance of the estimate $\hat{f}_2$ can be approximated in a large sample by

$$V(\hat{f}_2) = \left(\frac{1}{k-1}\right)^2 V\left(\sum_{i=1}^{k} \frac{a_{ii}}{\sum_{j=1}^{k} a_{ij}}\right)$$

$$= \left(\frac{1}{k-1}\right)^2\left[\sum_{i=1}^{k} V\left(\frac{a_{ii}}{\sum_{j=1}^{k} a_{ij}}\right) + 2\sum_{i<j}\mathrm{cov}\left(\frac{a_{ii}}{\sum_{l=1}^{k} a_{il}}, \frac{a_{jj}}{\sum_{l=1}^{k} a_{jl}}\right)\right]$$

$$\simeq \frac{1-f}{2n(k-1)^2}\left\{2(k-1) - 2(2k-1)f + k^2f^2 + f(2-f)\sum_{i=1}^{k}\frac{1}{p_i}\right\}.$$