Theoretical and Applied Genetics 48, 171-178 (1976)

Estimation of Inbreeding by Random Walks in Pedigrees

R.D. Cook

School of Statistics, University of Minnesota, Saint Paul, Minnesota (USA)

D.L. Hartl*

Department of Biological Sciences, Purdue University, West Lafayette, Indiana (USA)

Summary. Wright and McPhee (1925) suggested a method of estimating the inbreeding coefficient of an individual based on the probability that a pair of lineages traced randomly, one through the maternal line and one through the paternal line, both contain a common ancestor. (One-half of this probability is an unbiased estimate of the inbreeding coefficient). In their procedure, maternal and paternal lines are chosen in pairs, and comparisons are made only between the lines in a pair. A more efficient procedure is to compare every maternal line with every paternal line, a procedure used by Robertson and Mason (1954). In this paper we provide estimates of the sampling variance of the inbreeding coefficient as estimated by the multiple comparison method, and we examine the relative efficiency of this method and the Wright-McPhee procedure. Formulae are also provided for ascertaining the optimal sampling method for estimating the average inbreeding coefficient of a group or herd.

The inbreeding coefficient F was first defined by Wright in 1922, and since that time the quantity F has been used extensively as a measure of the amount of heterozygosity to be expected in an individual or group. Although the inbreeding coefficient was originally defined as the correlation coefficient between uniting gametes, Cotterman (1940) and Malecot (1948) preferred to interpret it as the probability that a pair of alleles in uniting gametes are identical by descent. For most purposes the definitions are equivalent. In this paper, we will find it convenient to make use of Wright's (1922) expression for the evaluation of the inbreeding coefficient of an individual with an irregular pedigree:

$$F = \sum_{i=1}^{r} (1/2)^{m_i+n_i+1} (1+F_i). \qquad (1)$$

A brief explanation of this expression is in order: consider two lines of ancestry of the individual in question, one line tracing back through the sire (paternal line) and the other tracing back through the dam (maternal line), and suppose that the lines connect for the first time at some common ancestor, say an ancestor denoted "i". Such a pair of lines is referred to as a "path". In expression (1), r is the number of unique paths associated with the individual in question; m_i and n_i are the number of generations from the sire and dam, respectively, to the common ancestor associated with the i^{th} path, and F_i is the inbreeding coefficient of the common ancestor associated with the i^{th} path. The inbreeding coefficient of a group of individuals is defined to be the average of the coefficients defined in (1) of all individuals in the group.

The definition and computation of the inbreeding coefficient has been considered by several authors (e.g. Haldane and Moshinsky 1939, Cotterman 1940 and Kudo 1962), and good detailed discussions are available in standard texts (e.g. Crow and Kimura 1970, Elandt-Johnson 1971 and Jacquard 1974). One point about the inbreeding coefficient that warrants special emphasis is that the inbreeding coefficient is a relative rather than an absolute measure of the homozygosity of an individual or group. It measures the degree of homozygosity relative to what it would be in an individual obtained by breeding randomly among descendants of a specified foundation stock. In the evaluation of an inbreeding coefficient, the foundation stock may be defined (perhaps implicitly) as the stock in existence at the beginning of the herdbook. Alternatively, the pedigrees could be traced back only to some selected date, and the foundation stock would be the stock in existence at the chosen date.

* Work supported by NSF grant GB43209, NIH grant GM21732 and by Research Career Award GM0002301.

The inbreeding coefficient of any individual or group could be computed by tabulating and comparing all possible maternal and paternal lines and using expression (1). In all but the simplest cases, this is an arduous task. Wright and McPhee (1925) point out that the complete pedigree of a modern Shorthorn would require the tabulation of several million names. To overcome this difficulty, Wright and McPhee (1925) suggested estimating the coefficient by comparing randomly chosen maternal and paternal lines. Sequences of sires and dams to be traced back in the herdbook are selected by repeated tosses of a fair coin, and the inbreeding coefficient is estimated as one-half the probability that such a random walk through the maternal and paternal ancestries results in lines that connect at a common ancestor. In addition to showing the usefulness of this deeply insightful sampling scheme and illustrating the estimation procedure, Wright and McPhee point out how the scheme could be used in combination with a complete listing of part of the pedigree.

Using the basic sampling procedure proposed by Wright and McPhee, we suggest an alternative estimation procedure that can be expected to yield more precise estimates of the inbreeding coefficient of an individual or group. Instead of comparing each maternal line with only one paternal line in search of "ties" (lines that contain the same individual), we propose that each maternal line of an individual be compared with every paternal line of that same individual. Essentially this procedure was employed to good effect by Robertson and Mason (1954) in their study of inbreeding in the Red Danish breed of cattle. In this paper we describe the estimation procedure in detail and derive approximations of the variance of the estimates. We assume, without significant loss of generality, that the inbreeding coefficient of each common ancestor is zero (i.e. $F_i = 0$ in expression (1)). The actual estimation procedure can easily be amended according to the discussion in Wright and McPhee (1925) to provide estimates when this assumption is unwarranted.

Estimation of the Inbreeding Coefficient of an Individual

The Wright-McPhee method has been widely used to estimate the average inbreeding coefficient of a group of individuals, but it is also applicable to the estimation of the inbreeding coefficient of a single individual. Wright and McPhee (1925, p. 351) state: "[Our] method may be used to calculate the inbreeding coefficient of an individual, by finding the percentage of ties in a large number of two-column samples from his pedigree." They also state (p. 350) that "a single [two-column] sample of this sort is of practically no value as an indication of the inbreeding of an individual."

Consider n pairs of maternal lines chosen randomly according to the procedure described in the previous section. A pair of lines containing a common ancestor will be said to be "tied". Thus, in a sample of n pairs of lines there are n possible ties. Let P denote the probability of a tie. It is easily shown that $P = 2F$; thus, the probability of a tie between randomly chosen lines is twice the inbreeding coefficient. In the following discussion we consider the estimation of P of a single individual.

In the Wright-McPhee method the distribution of the number of ties is binomial with parameters n and P. The maximum likelihood estimated of P is simply

$$\hat{P} = (\text{number of ties})/n \qquad (2)$$

with variance

$$V_1(\hat{P}) = P(1-P)/n \qquad (3)$$

Our alternative procedure is first to visualize the sample as one of 2n randomly chosen lines (n maternal and n paternal lines) rather than as a sample of n pairs of lines and, second, to estimate the inbreeding coefficient by making all possible comparisons between the maternal and paternal lines. This procedure requires little extra effort in practice, since the most time-consuming part of obtaining an estimate of the inbreeding coefficient is tracing sequences of sires and dams through the herdbook. Once a number of sequences have been tabulated, it is relatively simple to compare them pairwise searching for a common ancestor. Moreover, for the same total number of lines, the alternative procedure can be expected to yield more precise estimates than the one originally proposed by Wright and McPhee. We now consider a more detailed formulation of this procedure.

With n randomly chosen maternal and paternal lines, let

$$X_{ij} = 1 \quad \text{if there is a tie between the ith maternal and jth paternal lines}$$
$$= 0 \quad \text{elsewhere}$$

$i = 1,2,\ldots,n$; $j = 1,2,\ldots,n$. Clearly,

$$EX_{ij} = P$$

and therefore,

$$\hat{P} = \frac{\sum_{ij} X_{ij}}{n^2} = \frac{\text{number of observed ties}}{\text{number of possible ties}} \qquad (4)$$

is an unbiased estimate of P. The variance of this estimate is no longer in the form of a simple binomial variance since the X_{ij}'s are not independent. Also, without simplifying assumptions, it seems impossible to write the likelihood in a tractable form because it depends on the entire configuration of the pedigree. An expression for the variance of P can, however, be easily obtained by first writing

$$V(n^2\hat{P}) = \sum_{ij} V(X_{ij}) + \sum_{\substack{ijkl \\ (i,j)\neq(k,l)}} \operatorname{Cov}(X_{ij}, X_{kl}).$$

Making the reasonable assumptions that

(a) $\operatorname{Cov}(X_{ij}, X_{kl}) = 0$ if $i \neq k$ and $j \neq l$

(b) $\operatorname{Cov}(X_{ij}, X_{il}) = \operatorname{Cov}(X_{kj}, X_{kl})$ if $j \neq l$

and (c) $\operatorname{Cov}(X_{ij}, X_{kj}) = \operatorname{Cov}(X_{il}, X_{kl})$ if $i \neq k$

the previous expression simplifies to

$$V(\hat{P}) = \frac{P(1-P)}{n^2} + \frac{(n-1)}{n^2}[\operatorname{Cov}(X_{11}, X_{12}) + \operatorname{Cov}(X_{11}, X_{21})]. \qquad (5)$$

Again, without simplifying assumptions, the covariance terms cannot be written as functions of P only. An estimate of $V(\hat{P})$ based on the data can be found, however, by noting that

$$\operatorname{Cov}(X_{11}, X_{12}) = \Pr(X_{11} = 1, X_{12} = 1) - P^2 = P_{12} - P^2$$

$$\operatorname{Cov}(X_{11}, X_{21}) = \Pr(X_{11} = 1, X_{21} = 1) - P^2 = P_{21} - P^2$$

and using the following unbiased estimates of P_{12} and P_{21}:

$$\hat{P}_{12} = \frac{2}{n^2(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{j=k+1}^{n} X_{ik}X_{ij}$$
$$= \frac{1}{n^2(n-1)} \sum_{i=1}^{n} X_{i\bullet}(X_{i\bullet} - 1)$$

$$\hat{P}_{21} = \frac{2}{n^2(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{j=k+1}^{n} X_{ki}X_{ji}$$
$$= \frac{1}{n^2(n-1)} \sum_{j=1}^{n} X_{\bullet j}(X_{\bullet j} - 1)$$

where a "dot" in a subscript position indicates summation over the subscript. These expressions essentially count the number of times each maternal (paternal) line ties with two paternal (maternal) lines. Substituting $\hat{P}$, $\hat{P}_{12}$ and $\hat{P}_{21}$ into (5), we obtain the following estimate of $V(\hat{P})$,

$$\hat{V}(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n^2} + \frac{(n-1)}{n^2}(\hat{P}_{12} + \hat{P}_{21} - 2\hat{P}^2). \qquad (6)$$

Using expressions (4) and (6), an estimate of the inbreeding coefficient, F, and its variance are simply obtained as

$$\hat{F} = \hat{P}/2$$

and

$$\hat{V}(\hat{F}) = \hat{V}(\hat{P})/4.$$

Relative Efficiency

It is worthwhile to consider the efficiency of this procedure relative to the one originally proposed by Wright and McPhee. To this end we make the simplifying assumption that any maternal or paternal line can contain at most one common ancestor. Further insights into the problem can also be gained under this assumption. Let $\underline{a}$ denote the number of common ancestors and let $q_m(i)$ [$q_f(i)$] denote the probability that a randomly chosen maternal (paternal) line contains the i^{th} ancestor. With these specifications, we have

$$P = \sum_{i=1}^{a} q_m(i)\, q_f(i)$$

and

$$\hat{P} = \sum_{i=1}^{a} \hat{q}_m(i)\, \hat{q}_f(i) \qquad (7)$$

where $\hat{q}_m(i) = Y_i/n$, $\hat{q}_f(i) = Z_i/n$, and Y_i and Z_i are the number of maternal and paternal lines, respectively, that pass through the i^{th} ancestor. Clearly, $\{Y_i\}$ and $\{Z_i\}$ have independent multinomial distributions with parameters $\{q_m(i)\}$ and $\{q_f(i)\}$, respectively, provided sampling is with replacement. Note that expression (7) is equivalent to expression (4).

To obtain an easily interpretable expression of relative efficiency we would like the variance of $\hat{P}$ to be a function of P only. This is not possible at this point since $V(\hat{P})$ when computed from (7) still depends on the entire configuration of the pedigree through $\{q_m(i)\}$ and $\{q_f(i)\}$. An approximate expression can be obtained by assuming that $q_m(i) \approx q_f(i)$, $i = 1,2,\ldots,a$. One can then calculate how much each common ancestor considered individually contributes to the variance of $\hat{P}$, and these contributions can be summed over all common ancestors. (The authors are grateful to Professor Alan Robertson for pointing this out.) Suppose, then, that there is only one common ancestor in the pedigree. In such a case

$$V(\hat{P}) = V_2(\hat{P}) = \frac{P(1-P)}{n^2} + \frac{2(n-1)}{n^2} P(\sqrt{P} - P). \qquad (8)$$

If there are a common ancestors and all of them contribute equally to the overall estimate, then we may write P = ap and, summing (8) over all common ancestors and neglecting terms of order p^2, we have

$$V(\hat{P}) = \frac{P}{n^2} + \frac{2(n-1)}{n^2}\sqrt{\frac{P^3}{a}}.$$

Professor Robertson (personal communication) has pointed out that a useful definition of the effective value of a would be $(\Sigma p)^3/(\Sigma p^{3/2})^2$. In most populations of domestic animals the effective value of a is not expected to be large. In the Red Danish breed of cattle, the effective value of a turns out to be about 6 (A. Robertson, personal communication).

In any case, expression (8) is clearly a maximum estimate of $V(\hat{P})$. Thus (8) is the conservative estimate to be used in comparing the efficiency of our estimation procedure with that of Wright and McPhee. The efficiency of our procedure relative to the Wright-McPhee procedure is at least as great as

$$\frac{V_2(\hat{P})}{V_1(\hat{P})} = \frac{1}{n} + \frac{2(n-1)}{n^2}\,\frac{\sqrt{P} - P}{1 - P} \qquad (9)$$

where $V_1(\hat{P})$ is from expression (3). This ratio increases monotonically from 1/n to 1 as P goes from 0 to 1. Thus, as one might have anticipated, the proposed estimation procedure can be expected to be considerably better if the inbreeding coefficient is small.

For example, when n = 5 and $\dot{P} = 0.10$, $V_2(\hat{P})/V_1(\hat{P}) = 0.28$, so the proposed procedure is certainly the preferred one. Indeed, when P is small,

$$\frac{V_2(\hat{P})}{V_1(\hat{P})} \approx \frac{1}{n} + \frac{2(n-1)\sqrt{P}}{n^2}.$$

Estimation of the Inbreeding Coefficient of a Population

In this section we follow Wright and McPhee in defining the inbreeding coefficient of a group as the average of the individual inbreeding coefficients of the members of the group. There will almost inevitably be a variance in the individual inbreeding coefficients in a group because different individuals in the group will have somewhat different ancestries. Thus, semi-permanent differences between lines in the group induce variation in F.

Here we are concerned with estimating F_g, the inbreeding coefficient of a group, defined as the average of the individual inbreeding coefficients over all members of the group. An estimate of F_g (or equivalently $P_g = 2F_g$) can be obtained by randomly choosing k members of the group and averaging the estimates of the individual inbreeding coefficients:

$$2\hat{F}_g = \hat{P}_g = \sum_{i=1}^{k} \hat{P}_i/k \qquad (10)$$

where $\hat{P}_i$ is an estimate of $2F_i$ for the ith member of the sample. For the original Wright-McPhee procedure, the variance of such an estimate is

$$V(\hat{P}_g) = \frac{1}{k}\left[\frac{P_g(1 - P_g)}{n} + \left(1 - \frac{1}{n}\right)\sigma_g^2\right] \tag{11}$$

where n is the number of randomly chosen pairs of lines for each individual in the sample and σ_g^2 is the variance of twice the individual inbreeding coefficients in the group. From this expression it can be easily shown that when the total number (nk) of pairs of lines is taken to be fixed, $V(\hat{P}_g)$ is minimized when n = 1; that is, when the original Wright-McPhee procedure is employed, only one random pair of lines per individual should be used. When this is the case (8) reduces to

$$V(\hat{P}_g) = P_g(1 - P_g)/nk$$

which can be estimated by

$$\hat{V}(\hat{P}_g) = \hat{P}_g(1 - \hat{P}_g)/nk\,. \tag{12}$$

When our proposed procedure is used to estimate $P_i = 2F_i$ for each member of the sample, the variance of $\hat{P}_g$ is

$$V(\hat{P}_g) = \frac{P_g(1 - P_g)}{nc} + \frac{n^2 - 1}{nc}\sigma_g^2 + \frac{n - 1}{nc}E\,. \tag{13}$$

In expression (13), n and σ_g^2 are as previously defined, c = nk, and E denotes the expectation of the sum of the covariance terms in (5). When the total number of lines (c) is fixed, it is no longer clear that the optimal choice is n = 1. Inspection of (13) shows that the choice of n should now depend on $P_g(1 - P_g)$, σ_g^2 and E. A little manipulation of (13) reveals that n should be chosen to be 1 if

$$Z = \frac{[P_g(1 - P_g) - \sigma_g^2 - E]}{\sigma_g^2} < 4\,.$$

Otherwise, n should be chosen as the closest integer less than $\sqrt{Z}$. The presence of E makes these conditions difficult to evaluate in practice. However, under the simplifying conditions used previously to obtain expression (8), $E = 2P(\sqrt{P} - P)$. Thus, an approximate expression for Z is

$$Z = [P_g(1 + P_g - 2\sqrt{P_g}) - \sigma_g^2]/\sigma_g^2 \tag{14}$$

and in this form the optimal choice of n can be calculated with relative ease.

Intuition would lead one to expect that, if the variance in the inbreeding coefficients of a group of individuals is large, then (for a fixed expenditure of effort) the optimal procedure for estimating F_g would be to obtain relatively imprecise estimates of the individual inbreeding coefficients of as many individuals as possible and to average these. If, on the other hand, σ_g^2 is small, then you would want to estimate with greater precision the inbreeding coefficients of fewer individuals. Intuition is supported by the above rule that n should be the closest integer less than $\sqrt{Z}$. When P = 0.22, for example (this is roughly the average value of P in Red Danish bulls (see Robertson and Mason 1954), then the optimal choice of n is n = 1 for $\sigma_g^2 > 0.012$. However, the optimal n is 2 if $\sigma_g^2 = 0.01$, 7 if $\sigma_g^2 = 0.001$, and 24 if $\sigma_g^2 =$ = 0.0001.

Complete Tabulation of Part of the Pedigree

Instead of tracing a total of 2n randomly chosen paths, it may be desirable to completely tabulate the first few generations in the pedigree back from the individual in question and then continue to trace the ancestry using the random pedigree method. This procedure was also proposed by Wright and McPhee. It suffers to some extent, however, because it can require considerable effort to achieve a substantial increase in precision: in addition to completely tabulating part of the pedigree, one must keep track of the number of generations back to a common ancestor for part of the computations. Also, the total number of pairs is restricted to being a power of two (i.e., if the pedigree is tabulated completely for 4 generations there will be $2^4 = 16$ random lines continuing back to the foundation stock). On the other hand, in many breeds a substantial part of the variation in the inbreeding coefficient between individuals may be contributed by recent generations; it is thus of value to tabulate these generations completely.

Assume that the pedigree of the individual in question is to be completely tabulated for g generations back of his parents, and that the ancestry after the gth generation will be traced using, for each individual in generation g, a single random path. The inbreeding coefficient may be decomposed as follows

$$P = 2F = \sum_{\alpha_1} \left(\frac{1}{2}\right)^{n_i+m_i} + \sum_{\alpha_2} \left(\frac{1}{2}\right)^{n_i+m_i} + \sum_{\alpha_3} \left(\frac{1}{2}\right)^{n_i+m_i} + \sum_{\alpha_4} \left(\frac{1}{2}\right)^{n_i+m_i} = Q_1 + Q_2 + Q_3 + Q_4$$

where

$$\alpha_1 = \{i \mid \text{both } m_i \text{ and } n_i \leq g\}$$
$$\alpha_2 = \{i \mid n_i \leq g \text{ and } m_i > g\}$$
$$\alpha_3 = \{i \mid n_i > g \text{ and } m_i \leq g\}$$
$$\alpha_4 = \{i \mid \text{both } n_i \text{ and } m_i > g\} .$$

Note that α_1 corresponds to the common ancestors in the complete portion of the pedigree, α_2 and α_3 correspond to the common ancestors in the complete portion of the pedigree on one side and the random portion on the other, and α_4 corresponds to the common ancestors in the random portion of the pedigree.

Because of the nature of the sampling scheme Q_1, Q_2, Q_3, Q_4 are to be estimated separately.

Q_1 is determined directly from the computational formula and is not estimated but will be known with certainty. It may be taken as a lower bound on 2F.

Q_2 is the probability of a tie for the common ancestors in the complete portion of the pedigree on the sire side and the random portion on the dam side. Assuming that there are a_s observed common ancestors in the complete portion of the pedigree on the sire side, Q_2 may be estimated unbiasedly by

$$\hat{Q}_2 = \sum_{i=1}^{a_s} \left(\frac{1}{2}\right)^{n_i+g}$$

where n_i is the number of generations to the i^{th} observed common ancestor on the sire side of the pedigree. To see that $\hat{Q}_2$ is an unbiased estimate of Q_2, consider the following: the probability of detecting the i^{th} common ancestor in both the complete portion of the pedigree on the sire side and the random portion of the pedigree on the dam side is $\left(\frac{1}{2}\right)^{m_i-g}$. Define the random variable W_i by

$$W_i = \begin{cases} 1 & \text{if the } i^{th} \text{ common ancestor is detected both in the complete sire side and in the random dam side of the pedigree} \\ 0 & \text{elsewhere.} \end{cases}$$

Then $EW_i = \left(\frac{1}{2}\right)^{m_i-g}$ and

$$E\Sigma\, W_i \left(\frac{1}{2}\right)^{n_i+g} = \sum_{\alpha_2} \left(\frac{1}{2}\right)^{n_i+m_i} .$$

It follows immediately that $\hat{Q}_2$ is unbiased.

The analogous unbiased estimate of Q_3 is

$$\hat{Q}_3 = \sum_{i=1}^{a_d} \left(\frac{1}{2}\right)^{m_i+g}$$

where m_i is the number of generations to the i^{th} common ancestor in the complete portion of the pedigree on the dam side and a_d is the observed number of common ancestors.

The idea behind estimating Q_4 is basically the same as that discussed previously. However, the lines in the random portion of the pedigree that comprise the pedigrees of the common ancestors used in the calculation of Q_1, $\hat{Q}_2$ and $\hat{Q}_3$ should not be used. Letting $Y_{ij} = 1$ $(i = 1,2\ldots s,\ j = 1,2\ldots d)$ if there is a tie between the i^{th} relevant sire line and the j^{th} relevant dam line and zero elsewhere, we have that

$$\hat{Q}_4 = \frac{\sum_{i=1}^{s} \sum_{j=1}^{d} Y_{ij}}{sd\, 2^{2g}}$$

is an unbiased estimate of Q_4. Note that maximum number of possible ties, sd, between sire and dam lines in the random portion of the pedigree is simply $sd = 2^{2g}$ if there are no common ancestors in the complete portion of the pedigree.

The estimate of $P = 2F$ can now be written simply as

$$\hat{P} = Q_1 + \hat{Q}_2 + \hat{Q}_3 .+ \hat{Q}_4 .$$

The variance of $\hat{P}$ is easily seen to be the sum of the variances of the individual estimates,

$$V(\hat{P}) = V(\hat{Q}_2) + V(\hat{Q}_3) + V(\hat{Q}_4) .$$

$V(\hat{Q}_4)$ can be determined following the procedure previously outlined:

$$V(\hat{Q}_4) = \frac{Q_4 2^{2g}(1 - Q_4 2^{2g})}{sd\, 2^{2g}} + \frac{d-1}{sd} \operatorname{Cov}(Y_{11}, Y_{12}) + \frac{s-1}{sd} \operatorname{Cov}(Y_{11}, Y_{21}) .$$

An estimate of this variance can be found by setting $Q_4 = \hat{Q}_4$ and estimating the covariance terms in the same way the covariance terms in expression (5) were estimated. A first approximation to $V(\hat{Q}_2)$ and $V(\hat{Q}_3)$ can be easily found by assuming that each pedigree of the individuals in generation g can contain at most one common ancestor. Under this assumption we have

$$V_1(\hat{Q}_2) = \sum \left(\frac{1}{2}\right)^{2n_i+2g} V(W_i)$$

$$= \sum \left(\frac{1}{2}\right)^{2n_i+2g} \left(\frac{1}{2}\right)^{m_i-g} \left(1 - \left(\frac{1}{2}\right)^{m_i-g}\right) . \quad (15)$$

Similarly,

$$V_1(\hat{Q}_3) = \sum \left(\frac{1}{2}\right)^{2m_i+2g} \left(\frac{1}{2}\right)^{n_i-g} \left[1 - \left(\frac{1}{2}\right)^{n_i-g}\right] .$$

Estimates of these variances can be found by evaluating the sums over the set of observed common ancestors.

If some pedigrees initiating from the individuals in generation g contain more than one common ancestor the random variables (W_i) associated with these pedigrees will be correlated. Indexing the random variables so that $W_i^{(l)}$ are associated with the common ancestors in the l^{th} pedigree on the dam side ($i = 1,2,\ldots R_l$ and $l = 1,2,\ldots L$, say) we have

$$\operatorname{Cov}\left(W_i^{(l)}, W_j^{(l')}\right) = 0 \qquad l \neq l'$$

and

$$\operatorname{Cov}\left(W_i^{(l)}, W_j^{(l)}\right) = - \left(\frac{1}{2}\right)^{m_i^{(l)}+m_j^{(l)}-2g} \qquad i \neq j$$

where $m_i^{(l)}$ denotes the number of generations to the common ancestor associated with $W_i^{(l)}$. The latter result follows from the observation that for each l the distribution of $\{W_i^{(l)}\}$ is multinomial. With these results a little algebra will verify that

$$V(\hat{Q}_2) = V_1(\hat{Q}_2) - \sum_l \left[\sum_i \sum_{\substack{j \\ i \neq j}} \left(\frac{1}{2}\right)^{m_i^{(l)}+m_j^{(l)}+n_i^{(l)}+n_j^{(l)}} \right] \quad (16)$$

where $V_1(\hat{Q}_2)$ is from expression (15). If each pedigree initiating from the individuals in generation g contains only one common ancestor, then $R_l = 1$ for all l and this expression reduces immediately to (15). Moreover, under the present sampling scheme (one random data per individual in generation g) the only estimate available for the covariance portion of (16) is zero and we are thus led back to (15) for the purpose of estimating $V(\hat{Q}_2)$. Analogous results for $V(\hat{Q}_3)$ can be obtained by interchanging n_i and m_i.

Discussion

The procedure of estimating the inbreeding coefficient by making all possible pairwise comparisons between a set of n maternal and n paternal ancestries has been used previously by Robertson and Mason (1954 -- see also Robertson and Asker 1951). Robertson and Mason sampled each pedigree by tracing one line back at random on both sides of the pedigree and at each step in the procedure writing down both the male and female parent, even though the line was carried back through only one of the parents. They thereby generated two maternal and two paternal lines. The counting of ties in this method must be modified somewhat to take into account the nonindependence of both the maternal lines and the paternal lines. Specifically, when a tie appears in the randomly chosen line, then the sire and dam must appear twice in the preceding generation and the resulting ties must not be counted. Moreover, ties lying behind an initial tie are to be counted only when the random lines from the animal constituting the initial tie passed one to the sire and one to the dam. With these modifications, the method of estimation proceeds as before.

It is worth emphasizing that the suggested method of estimating the inbreeding coefficient by multiple comparison of sire and dam lines provides a much more efficient use of data than does comparing each sire line with only one dam line. As mentioned previously, a large part of the effort in estimating inbreeding by the Wright-McPhee method is tracing ancestries through the herdbook. Once this has been done, there seems little reason to extract from the ancestries less than the maximum amount of information they contain.

Acknowledgement

Our thanks to Professor Alan Robertson for his detailed and helpful comments on the manuscript.

Literature

Cotterman, C.W.: A Calculus for Statistico-genetics. Unpublished thesis, Ohio State University, Columbus, Ohio (1940)

Crow, J.F.; Kimura, M.: An introduction to population genetics theory. New York: Harper and Row 1970

Elandt-Johnson, E.: Probability models and statistical methods in genetics. New York: John Wiley 1971

Haldane, J.B.S.; Moshinsky, P.: Inbreeding in Mendelian populations with special reference to human cousin marriage. Ann. Eugen. 9, 321-340 (1939)

Jacquard, A.: The genetic structure of populations, tr, by D. and B. Charlesworth. New York: Springer-Verlag 1974

Kudo, A.: A method for calculating the inbreeding coefficient. Am. J. Hum. Genet. 14, 426-432 (1962)

Malécot, G.: Les mathématiques de l'hérédité. Paris: Masson 1948

Robertson, A.; Asker, A.A.: The genetic history and breed-structure of British Friesian cattle. Empire J. Expt. Agric. 19, 113-130 (1951)

Robertson, A.; Mason, I.L.: A genetic analysis of the Red Danish breed of cattle. Acta Agriculturae Scandinavica 4, 257-265 (1954)

Wright, S.: Coefficients of inbreeding and relationship. Amer. Natur. 56, 330-338 (1922)

Wright, S.; McPhee, H.C.: An approximate method of calculating coefficients of inbreeding and relationship. J. Agric. Res. 31, 377-383 (1925)

Received April 4, 1975/
February 25, 1976
Communicated by A. Robertson

R. Dennis Cook
School of Statistics
University of Minnesota
Saint Paul, Minnesota 55101 (USA)

Daniel L. Hartl
Department of Biological Sciences
Purdue University
West Lafayette, Indiana 47907 (USA)