PROCEDURES FOR CALCULATING ESTIMATES OF THE COEFFICIENT OF INBREEDING OF AN INDIVIDUAL

by

KEITH LAVERNE HOFFMAN

B. S., Kansas State University, 1965

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:

H. C. Fryer

Major Professor

TABLE OF CONTENTS

INTRODUCTION

The problem of computing the coefficient of inbreeding, even in the most complicated pedigrees, is simply computing the amount of heterozygosis probably lost because of inbreeding; inbreeding being defined as the mating together of individuals that are related by ancestry. As a result of inbreeding the zygotic proportions within a population are altered in such a way as to increase the amount of homozygosis, thus decreasing the amount of heterozygosis. Hence, for various degrees of inbreeding the zygotic proportions become for dominants (AA), heterozygotes (Aa) and recessives (aa), respectively $p^2 + Fpq$, $2pq(1-F)$ and $q^2 + Fpq$, where p and q are gene frequencies of A and a.

The symbol F in the above proportions refers to the coefficient of inbreeding. Wright (1922) defines F as the correlation coefficient between uniting gametes; whereas, Malécot (1948) defines F as the probability that the two genes at any locus in an individual are identical by descent. These definitions are equivalent, the difference being in the approach to the problem of computing the coefficient of inbreeding.

The concept of the coefficient of inbreeding had its beginning in the early 1920's with the work of Sewall Wright. His procedures consisted primarily of tracing lines of descent on a pedigree chart by the use of path coefficients. In 1925 Wright and H. C. McPhee combined efforts to condense Wright's original procedure.

During the late 1940's and the early 1950's other methods for estimating the coefficient of inbreeding were developed. These methods attempted to simplify S. Wright's original method of path coefficients. A Frenchman, Gustave Malécot, approached the problem of coefficients of inbreeding by making use of probabilities. Other procedures formulated at this time made use of the work of Li (1953,1955), Horvitz (1953), Emik (1949), Terrill (1949), Cruden (1949), Plum (1954), Hazel (1950) and Lush (1950). While examining information on this topic, one becomes increasingly aware that there was a trend of heightened interest concerning the importance of estimating the coefficient of inbreeding during this time period of late 1940's and early 1950's; that there has been a decline of interest on this topic in the past decade.

With this background information in mind, one may begin to discuss the coefficient of inbreeding of an individual.

PROCEDURE INVOLVING PATH COEFFICIENTS

In order to examine path coefficients some elementary concepts of statistics need to be reviewed. This review is necessary because path coefficients involve statistical concepts. The statistics included in this review are the correlation coefficient, the sum of variables, multiplying independent variables and partial correlation (Wright, 1921,1934; Li, 1955; Kempthorne, 1957).

The correlation coefficient assumes a linear relation between two variables, say A and B; i.e., a given change in A will always involve a certain constant change in the corresponding average value of B. Let $\overline{A}$ and $\overline{B}$ be the mean values of A and B respectively, then the correlation coefficient between A and B is defined as

$$r_{AB} = \frac{\sum (A-\overline{A})\ (B-\overline{B})}{\sqrt{[\sum (A-\overline{A})^2 \sum (B-\overline{B})^2]}} = \frac{\sigma_{AB}}{\sigma_A\ \sigma_B} \qquad (1)$$

In connection with the idea of regression, when the variances of A and B are equal the following is obtained

$$r_{AB} = \frac{\sigma_{AB}}{\sigma_A^2} = b_{AB} = \frac{\sigma_{AB}}{\sigma_B^2} = b_{BA} \qquad (2)$$

Also, by definition $r_{AA} = 1$, and if A and B are independent, $r_{AB} = 0$. These concepts may be extended to N pairs of values of A and B.

Let $X = A+B$, then the variance of the summed variable is

$$\sigma_X^2 = \sigma_{A+B}^2 = \sigma_A^2 + \sigma_B^2 + 2\,\sigma_A\,\sigma_B\ r_{AB} \qquad (3)$$

When A and B are independent, or uncorrelated, $r_{AB} = 0$ and (3) reduces to

$$\sigma_X^2 = \sigma_A^2 + \sigma_B^2 \qquad (4)$$

This may be extended to any number of factors.

Let $X = AB$ and assume $r_{AB} = 0$, then the variance of the product is

$$\sigma_X^2 = \bar{B}^2 \sigma_A^2 + \bar{A}^2 \sigma_B^2 + \frac{1}{N} \sum (A-\bar{A})^2 (B-\bar{B})^2 \qquad (5)$$

Generally speaking the last term in (5) is much smaller than either of the first two terms, and (5) becomes approximately

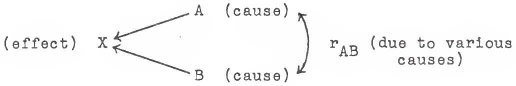
$$\sigma_X^2 = \bar{B}^2 \sigma_A^2 + \bar{A}^2 \sigma_B^2 \qquad (6)$$

Suppose there are three correlated variables: A, B, and C. The partial correlated coefficient between A and B when C is kept constant is

$$r_{AB.C} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{\left[(1-r_{AC}^2)\ (1-r_{BC}^2)\right]}} \qquad (7)$$

This may also be extended to any number of variables.

In addition to the degree of relationship furnished by the coefficients of correlation, some knowledge of the nature of the relationship between the variables must be taken into account. It is not necessary to know what constitutes "cause" and "effect" (Wright, 1921,1923a,1934; Tukey, 1954; Li, 1955); one needs only to be aware that there are many cases in which certain factors are direct causes of variation in others or that other pairs are related as effects of a common cause. In a system of related variables, "causes" and "effects" are connected by arrows as in the following diagram.



The arrows connecting causes and effect in the above diagram are referred to as "paths".

Let σ_X be the total standard deviation of X and $\sigma_{X.A}$ denote the standard deviation of X due to the influence of A, while all other causes (except A) remain constant. The path coefficient, $p_{X.A}$, (Wright, 1921,1923a,1934; Tukey, 1954; Li, 1955; Kempthorne, 1957) is defined as the ratio of the standard deviation of X due to A to the total standard deviation.

$$p_{X.A} = \frac{\sigma_{X.A}}{\sigma_X} \qquad (8)$$

The path coefficient, $p_{X.A}$, is an absolute number without any physical unit. In this respect it is similar to correlation coefficient. However, the path coefficient has a direction (from A to B); in this respect being similar to regression coefficients. Thus, one may state that path coefficients are standardized linear regression coefficients.

Another property of path coefficients is the determination of X by cause A. The coefficient of determination, $d_{X.A}$, is defined as the square of the path coefficient.

A process preliminary to calculating the total correlation between two variables is tracing connecting paths (Wright, 1922, 1923b; Li, 1955); because this correlation is the sum of all paths connecting these two variables in a causal scheme. There are certain rules that must be followed in tracing these connecting paths.

1. No "first-forward-then-backward" motion in tracing any connecting paths.
2. The correct way of tracing a connecting path is a "first-backward-then-forward" motion.
3. For chains of variables one may continue to trace backward (no change in direction) for as many steps as are available, then forward for as many steps as are available, without any change in direction.

It is necessary to add that these rules apply only in cases of independent causes, not in cases where the causes are dependent.

The coefficient of inbreeding (Wright, 1922,1923b,1934; Li, 1955; Kempthorne, 1957) is obtained by a summation of path coefficients for every line of descent by which the parents are connected, each line tracing back from the sire to a common ancestor and hence forward to the dam, and passing through no individual more than once. The same common ancestor may, of course, be involved in more than one line.

The path coefficient for the path, sire (X) to offspring

(O), is given by the formula

$$p_{O.X} = \frac{1}{2} \sqrt{\frac{(1+f_X)}{(1+f_O)}} \quad , \qquad (9)$$

where f_X and f_O are the coefficients of inbreeding for sire and offspring respectively.

In the case of the grand sire (G) and offspring (O), the path coefficient is

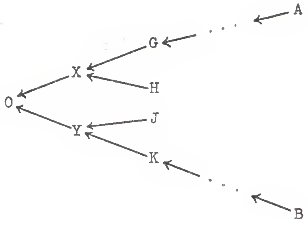
$$p_{O.G} = p_{O.S}\, p_{S.G} = \frac{1}{4} \sqrt{\frac{(1+f_G)}{(1+f_O)}} \quad , \qquad (10)$$

and for any ancestor (A) one has for the coefficient pertaining to a given line of descent

$$p_{O.A} = (\frac{1}{2})^n \sqrt{\frac{(1+f_A)}{(1+f_O)}} \quad , \qquad (11)$$

where n is the number of generations between individuals (O) and the ancestor (A) in this line.

The following path diagram will aid in understanding (9), (10) and (11).



The correlation between two individuals (r_{XY}) is obtained by a summation of the coefficients for all connecting paths.

Thus,

$$r_{XY} = \sum p_{X.A}\, p_{Y.A}$$

$$= \sum \left(\frac{1}{2}\right)^{m+n} \frac{1 + f_A}{\sqrt{(1+f_X)\,(1+f_Y)}} \qquad (12)$$

where m and n are the number of generations in the paths from A to X and from A to Y, respectively.

The correlation between uniting gametes, the coefficient of inbreeding, is

$$f_0 = \frac{1}{2}\, r_{XY} \sqrt{(1+f_X)(1+f_Y)} \qquad (13)$$

where r_{XY} is the correlation between sire and dam and f_X and

f_Y are coefficients of inbreeding of sire and dam. Substituting the value of r_{XY} results in

$$f_0 = \sum \left[\left(\frac{1}{2}\right)^{m+n+1} (1+f_A) \right] . \qquad (14)$$

AN APPROXIMATING PROCEDURE

The preceeding material formulated by Sewall Wright has presented not only the most widely cited procedure for calculating the coefficient of inbreeding, but also was one of the first procedures created.

Later attempts to formulate methods for calculating this coefficient condensed the number of time-consuming computations.

One of these later methods (Wright and McPhee, 1925) made a definite attempt for condensation. To understand this approximating method for calculating the coefficient of inbreeding one must refer back to (14). In (14), $(\frac{1}{2})^{m+n+1}$ $(1+f_A)$ refers to the contribution of a particular tie between the pedigrees of sire and dam.

This approximate procedure rests on the tabulation of random samples of the pedigrees of sire and dam. The reliability of the results can be tested by the ordinary theory of sampling. It is necessary that the sample lines be chosen wholly at random.

The simplest possible sample which can show a connection between sire and dam is obtained by tracing back two ancestral lines, one on the sire's side and one on the dam's side. Random sequences of S's and D's are then written in columns below each parent, extending sufficiently to include the foundation stock. The line of ancestry is then traced back in the pedigree, the sire being looked up where S occurs in the column and the dam for each D in the column. Although a second sample will probably not show the same sequence of sires and dams, a single sample is of practically no value as an indicator of the inbreeding of the individual. However, the average obtained from a large number of such samples should not differ appreciably from the true value.

The following explanation will be concerned with the two-column samples which show an ancestral connection, since those which do not show ancestral connection have a coefficient of zero, as far as the sample indicates. In the former cases a contribution of $(\frac{1}{2})^{m+n+1}$ $(1+f_A)$ is indicated if the common ancestor A is m generations back of the sire and n back of the dam. The sire has 2^m ancestors in the m^{th} generation and the dam 2^n possible pairs going back as far as the common ancestor. If the single pair of lines is a fair sample of the total, its contribution must be multiplied by 2^{m+n} to obtain an estimate of the inbreeding of the whole pedigree. On carrying out this multiplication, m and n disappear and the coefficient takes the simple form $\frac{1}{2}(1+f_A)$. Thus, in calculating the inbreeding indicated by a two-column pedigree, it is not necessary to count the generations to the closest common ancestor; it is merely necessary to note whether there is a tie between the pedigree sire and dam, and what animal is responsible for it.

It should be noted that increased accuracy in the approximating procedure may be obtained by combining the approximating procedure with the previously explained complete procedure using path coefficients.

PROBABILITY APPROACH

The probability approach (Malécot, 1948; Kempthorne, 1957) is unique in its computation of the coefficient of inbreeding. Malécot uses the term "coefficient de parente" which is equivalent to Wright's term, coefficient of inbreeding. Malécot's procedure involves the relationship between two individuals; henceforth, let "coefficient de parente" be referred to as the "coefficient of parentage".

Each individual I has two parents, four grandparents, . . . , 2^n ancestors of the order n. One gene of I has the probability of $\frac{1}{2}$ of originating from the father, $\frac{1}{2}$ from the mother, $\frac{1}{4}$ from each of the grandparents, . . . , $\frac{1}{2}^n$ of originating from a given ancestor of the order n along a determined chain of ascendance. (An ancestor of I can be connected to him by several chains of ascendance.)

Let the coefficient of parentage f_{IL} of two individuals, I and L, be the probability that two genes at a locus taken, one on I and the other on L, are identical, that is to say they descend from the same locus. The complimentary probability

$(1-f_{IL})$ represents the probability that these two genes are a result of ancestors with no relationship, in other words, are stochastically independent (because then the knowledge of the gene which occupies one gives no information on the gene which occupies the other; these two genes can be identical or different but their probabilities are independent).

Call the coefficient of consanguity f_m of an individual M the probability that its two genes at a locus are identical by descent. As one originates from its father and the other from its mother, f_m is the coefficient of parentage between the two parents.

The coefficient of parentage f_{IL} of two individuals I and L is greater than zero only if I and L have one or several common ancestors A_1, A_2, etc. Assume at first that there is only one ancestor A of the order m of I and of the order n of L by chains of unique ascendance whose combination constitutes a chain of relationship connecting I and L.

The probability that one gene of I and one homologous gene of L originate from A is $(\frac{1}{2})^{m+n}$; but in this eventuality they have a probability of $\frac{1}{2}$ of originating from the same locus A, and a probability of $\frac{1}{2}$ of originating from different loci in which case they are only identical with the probability f_A. Hence, $f_{IL} = (\frac{1}{2})^{m+n} \dfrac{1+f_A}{2}$. In particular the coefficient of parentage of one individual with a common ancestor of the order of m corresponds to n = 0; the coefficient of parentage of one individual with himself corresponds to $m = n = 0$.

One may now deal with the general case where I and L are connected by any number of chains of relationship, each chain being the union of two chains of ascendance coming from I and L to a common ancestor A_i and having no common point other than A_i; two chains of relationship are regarded as distinct even if they have a common part, provided that they differ by at least one link. As the transmission of identical genes along a determined chain of relationship excludes their transmission along all others, the principle of total probabilities gives:

$$f_{IL} = \sum (\tfrac{1}{2})^{m_i+n_i} \frac{(1+f_{A_i})}{2} \quad . \qquad (15)$$

The sum being extended to all distinct chains of relationship connecting I and L, the i^{th} is comprised of m_i+n_i links and coming to the common ancestor A_i with coefficient of consanguity f_{A_i}.

Thus, one finds that the final formula reached by Malécot using the probability approach is equivalent to Wright's original formula (14).

USE OF MODELS BASED ON A PANMICTIC POPULATION

Proportion of Heterozygotes

In order to understand the next seven methods of calculation of the coefficient of inbreeding, it is necessary to review some relevant definitions.

In a large panmictic (random mated) population in which the frequency of the allele A_i is q_i , the proportions of the various genotypes in an equilibrium condition are given by the coefficients of the A's in the expression

$$\left[\sum_i q_i A_i\right]^2 = \sum_i q_i^2 A_i A_i + 2\sum_{i<j} q_i q_j A_i A_j \qquad \text{(I)}$$

where $q_i = 1(i = 1,2,. . . ,k)$. This population will be referred to as Model I.

As compared to Model I, there will be relatively more homozygous individuals in the population when the gametes are not uniting entirely at random, but are correlated. When the population is not mating at random, the genotypic frequencies will be:

$$(1-F)\left[\sum_i q_i A_i\right]^2 + F\sum_i q_i A_i A_i$$

$$= \sum_i \left[(1-F)q_i^2 + Fq_i\right] A_i A_i + 2(1-F)\sum_{i<j} q_i q_j A_i A_j \qquad \text{(II)}$$

This population will be referred to as Model II. When F (coefficient of inbreeding) = 0, Model II becomes Model I.

One of the simplest methods (Li and Horvitz, 1953; Li, 1955) of estimating F is based upon the total proportion of heterozygotes in a sample. Let this proportion be H. Assign the terms H_0 and H_F to denote the total proportions of heterozygotes in Models I and II, respectively.

Thus,

$$H_0 = 2\sum q_i q_j \ , \ (i < j)$$

$$= 1 - \sum q_i^2$$

and $$H_F = 2(1-F)\sum q_i q_j$$

Therefore,

$$H_F = (1-F)\ H_0$$

or $$F = \frac{H_0 - H_F}{H_0}$$

regardless of the number of alleles involved. Substitution of the observed H (= $2\sum a_{ij}/N$, from Table 2) for H_F and calculation of the value of H_0 taking $q_i = n_i/N$, one may estimate F. Hence, using f to denote the sample estimate of F, one has

$$f = 1 - \frac{H}{H_0} = 1 - \frac{N\sum a_{ij}}{\sum n_i n_j} \ , \ (i < j) \ . \qquad (16)$$

Product-Moment Correlation

The fact that the product-moment correlation coefficient (Li and Horvitz, 1953) between the gametes of the following table is F, may be varified by assigning any arbitrary numerical values to alleles $A_1, A_2, \ . \ . \ . \ , A_k$.

Table I. Gametic Correlation of Model II

	A_1	A_2	· · ·	A_k	
A_1	$(1-F)q_1^2 + Fq_1$	$(1-F)q_1q_2$	· · ·	$(1-F)q_1q_k$	q_1
A_2	$(1-F)q_2q_1$	$(1-F)q_2^2 + Fq_2$	· · ·	$(1-F)q_2q_k$	q_2
· ·	· ·	· ·		· ·	· ·
A_k	$(1-F)q_kq_1$	$(1-F)q_kq_2$	· · ·	$(1-F)q_k^2 + Fq_k$	q_k
	q_1	q_2	· · ·	q_k	1

The order of the arrangement of the alleles is immaterial. This is, however, not the case with actual sample numbers. For instance, when $k = 3$, there are three different ways of arranging the sample data and thus three different correlation values could be obtained. It is convenient to assign the values 1, 0, -1 to A_1, A_2, A_3, respectively, in which case the correlation coefficient (estimate of F) is given by

$$f = \frac{N(a_{11} - 2a_{13} + a_{33}) - (n_1 - n_3)^2}{N(n_1 + n_3) - (n_1 - n_3)^2} \qquad (17)$$

Two other similar expressions may be derived by interchanging the subscripts 2 and 3, and 1 and 2. If the sample data are consistent with Model II, the values of the three correlations should not differ to any great extent. Although, each of them is a consistent estimate of F, it is desirable to devise some methods of estimation which are independent of the order of arrangement of the sample numbers and yield a unique estimate. The preceding method and the following five methods satisfy these conditions.

Determinant of the Gametic Correlation Matrix

If one arranges the zygotic proportions of Model II in the form of a gametic correlation as in Table 1, the determinant of the matrix formed by the elements is a function of F (Li and Horvitz, 1953; Kempthorne, 1957). One must then remove the factors in q common to all the elements of each row, add each of the columns to the first, all of whose elements are equal to unity, and subtract the first row from each of the remaining rows. The result is:

$$q_1 \cdot \cdot \cdot q_k \begin{vmatrix} (1-F)q_1+F & (1-F)q_2 & \cdot\cdot\cdot & (1-F)q_k \\ (1-F)q_1 & (1-F)q_2+F & \cdot\cdot\cdot & (1-F)q_k \\ \cdot & \cdot & & \cdot \\ (1-F)q_1 & (1-F)q_2 & \cdot\cdot\cdot & (1-F)q_k+F \end{vmatrix}$$

$$= q_1 \cdot \cdot \cdot q_k \begin{vmatrix} 1 & (1-F)q_2 & \cdot\cdot\cdot & (1-F)q_k \\ 1 & (1-F)q_2+F & \cdot\cdot\cdot & (1-F)q_k \\ \cdot & \cdot & & \cdot \\ 1 & (1-F)q_2 & \cdot\cdot\cdot & (1-F)q_k+F \end{vmatrix}$$

$$= q_1 \cdot \cdot \cdot q_k \begin{vmatrix} 1 & (1-F)q_2 & (1-F)q_3 & \cdot \cdot \cdot & (1-F)q_k \\ 0 & F & 0 & \cdot \cdot \cdot & 0 \\ 0 & 0 & F & \cdot \cdot \cdot & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdot \cdot \cdot & F \end{vmatrix}$$

$$= q_1 \cdot \cdot \cdot q_k F^{k-1} \tag{18}$$

Table 2. Observed Numbers of Individuals

	A_1	A_2	. . .	A_k	
A_1	a_{11}	a_{12}	. . .	a_{1k}	n_1
A_2	a_{21}	a_{22}	. . .	a_{2k}	n_2
.	.	.		.	.
A_k	a_{k1}	a_{k2}	. . .	a_{kk}	n_k
	n_1	n_2	. . .	n_k	N

Therefore, the determinant of the observed numbers in Table 2 divided by the product of its marginal totals will yield an estimate of the $(k-1)^{th}$ power of F; thus,

$$f^{k-1} = \frac{\begin{vmatrix} a_{11} & \cdot & \cdot & \cdot & a_{1k} \\ \cdot & & & & \cdot \\ a_{k1} & \cdot & \cdot & \cdot & a_{kk} \end{vmatrix}}{n_1 \times n_2 \times \cdot \cdot \cdot \times n_k} \qquad (19)$$

Chi-Square

Using the proportions of Model I as the "expected" and those of Model II as "observed" numbers, the difference between the zygotic proportions of (I) and (II), as caused by the existence of F, may be measured by the value of Chi-square (Li and Horvitz, 1953; Li, 1955).

$$\begin{aligned} \chi^2 &= \sum_{i} \frac{\left[NFq_i(1-q_i) \right]^2}{Nq_i^2} + \sum_{i<j} \frac{\left[2NFq_iq_j \right]^2}{2Nq_iq_j} \\ &= NF^2 \left\{ \sum_{i} (1-2q_i + q_i^2) + 2\sum_{i<j} q_iq_j \right\} \\ &= NF^2(k - 2 + 1) \\ &= NF^2(k - 1) \qquad (20) \end{aligned}$$

Let $q_i = n_i/N$ and calculate the zygotic proportions (I) on the assumption of panmixia, the value of Chi-square obtained on comparing them with the observed will give one an estimate of F; viz.,

$$f^2 = \frac{\chi^2}{N(k-1)} \text{, with } \frac{k(k+1)}{2} - k = \frac{k(k-1)}{2} \text{ d.f.} \quad (21)$$

One may obtain a sampling distribution of f by a transformation of that of Chi-square. The advantage of using this method is that the test of significance of f is equivalent to testing the significance of χ^2.

Proportions of Alleles in Homozygous Condition

The method of estimating the coefficient of inbreeding that involves the least amount of arithmetic labor follows (Li and Horvitz, 1953). Let $z_{ii} = (1-F)q_i^2 + Fq_i$ denote the proportion of A_iA_i in Model II whose frequency of allele A_i is q_i. Hence, the proportion of A_i's in the population is z_{ii}/q_i . The sum of such proportions over all alleles is:

$$\frac{z_{11}}{q_1} + \frac{z_{22}}{q_2} + \ldots \frac{z_{kk}}{q_k} = 1 + F(k-1)$$

In Model I (F=0), the sum of such proportions is unity. From this consideration the sample estimate of F is obviously

$$f = \frac{1}{k-1} \left\{ \sum_i \frac{a_{ii}}{n_i} - 1 \right\} . \quad (22)$$

Maximum Likelihood

In this method (Li and Horvitz, 1953) let the number of alleles be two, i.e., k =2. Let the observed numbers of A_1A_1, A_1A_2, A_2A_2 in the sample be a, 2b, c, respectively, where a + 2b + c = N. To avoid subscripts let p be the frequency of the allele A_1 and q be that of A_2, where p + q = 1. Then the likelihood function is

$$L = \frac{N!}{a!(2b)!c!} \left[(1-F)p^2 + Fp \right]^a \left[2(1-F)pq \right]^{2b}$$
$$\left[(1-F)q^2 + Fq \right]^c$$

and the logarithm of the likelihood function is, ignoring constant terms,

$$\log L = a \times \log\left[(1-F)p^2 + Fp \right] + 2b \times \log\left[2(1-F)pq \right]$$
$$+ c \times \log\left[(1-F)q^2 + Fq \right]$$

Setting $\partial \log L/\partial p = 0$ and $\partial \log L/\partial F = 0$, these two equations upon simplification become

$$\frac{a(2p+\theta)}{p(p+\theta)} + \frac{2b(1-2p)}{pq} - \frac{c(2q-\theta)}{q(q+\theta)} = 0$$

$$\frac{aq}{p+\theta} - 2b + \frac{cp}{q+\theta} = 0 \qquad (23)$$

where $\theta = F/(1-F)$. On eliminating their middle terms by multiplying the second equation of (23) by (1-2p)/pq and then adding the two equations together, we obtain upon simplification the relation $aq/(p+\theta) = cp/(q+\theta)$. Hence, (23) may be written in the much more simplified form of two linear equations:

$$\frac{aq}{p+\theta} = b \quad ; \quad \frac{cp}{q+\theta} = b \,. \tag{24}$$

The simultaneous solution of these equations gives

$$p = \frac{a+b}{N} \;, \quad \theta = \frac{ac-b^2}{Nb} \;, \tag{25}$$

the expression for θ here being equivalent to those for f in previous sections.

For $k \geq 3$, it seems best to accept the observed gene frequencies ($q_i = n_i/N$) and estimate the value of F under this set of conditions in order to give F the biological meaning attached to it. Hence, with given values of q_i, one has only to solve the equation $\partial \log L / \partial F = 0$, i.e.,

$$\sum_i \frac{a_{ii}(1-q_i)}{q_i + \theta} = 2 \sum_{i<j} a_{ij} \tag{26}$$

where $\theta = F/(1-F)$ and $2 \sum a_{ij}$ is the total number of heterozygotes in the sample. Note that when $k = 2$, (26) reduces to the second equation of (23). To solve (26) for θ, an initial trial value may be obtained from one of the previous methods of this section, and a more accurate solution for θ may be obtained by iteration.

Reducing the Number of Alleles

Instead of solving (26) as a whole, one may break it into component parts and just solve the following equation (Li and Horvitz, 1953):

$$\frac{a_{ii}(1-q_i)}{q_i + \theta} = \sum_j a_{ij}(i \neq j) = n_i - a_{ii} \; . \qquad (27)$$

That is, one may set the fraction on the left side equal to half the number of heterozygotes containing the allele A_i. This equation is analogous to (24) for the case of two alleles. Solving (27) for θ , we have, still taking $n_i = Nq_i$,

$$\theta = \frac{a_{ii} - n_i q_i}{n_i - a_{ii}} = \frac{a_{ii} - Nq_i^2}{Nq_i - a_{ii}} \qquad (28)$$

or,

$$f = \frac{a_{ii} - Nq_i^2}{Nq_i(1-q_i)} = \frac{Na_{ii} - n_i^2}{n_i(N-n_i)}$$

It should be noted that this approximation method is equivalent to pooling all the non-A_i alleles together as one allele, thus reducing the original k x k gametic correlation table into a 2 x 2 table involving only A_i and $\overline{A}_i$ as shown in Table 3 (the symbol $\overline{A}_i$ denotes non-A_i alleles). Applying the method of estimating F for k = 2, we obtain the solution (28), which is the maximum likelihood estimate as far as the data in Table 3 are concerned. Note that this pooling method is quite different from the first method of estimation in this section. There are k ways of doing this kind of reduction. In practice, however, one may choose the allele with the highest frequency to be A_i and pooling the remaining k-1 alleles together. Similarly, one may reduce the k alleles to any number smaller than k. This procedure, though an approximate method, is perhaps advisable when some of the alleles have very low frequencies.

Table 3. Reduced 2 x 2 Gametic Correlation Table

	A_i	$\bar{A}_i$	Total
A_i	a_{ii}	$n_i - a_{ii}$	n_i
$\bar{A}_i$	$n_i - a_{ii}$	$N-2n_i + a_{ii}$	$N - n_i$
Total	n_i	$N - n_i$	N

SYSTEMATIC PROCEDURES

Sire-Ancestor Procedure

In some systems of mating it is not always possible to have a regular system of inbreeding. However, some measure of inbreeding is essential, because the degree of inbreeding could vary widely. This method (Emik and Terrill, 1949) is an attempt to condense Sewall Wright's original formula for F (14).

The relationship of the parents (X and Y), for the purpose of calculating the inbreeding coefficient of the offspring, would be

$$R_{XY} = \sum \left[\left(\frac{1}{2}\right)^{m+n} (1 + f_A) \right] \qquad (29)$$

which is the genetic covariance, but will be called the numerator relationship since it is the numerator of the true relationship.

$$R = \frac{\sum \left[\left(\frac{1}{2}\right)^{m+n} (1 + f_A) \right]}{\sqrt{(1 + f_X)\ (1 + f_Y)}} \qquad (12)$$

The numerator relationship for any pair of parents is twice the value of the inbreeding coefficient of the offspring.

From (14) and (29) the numerator relationship of parent (X) to offspring (O) becomes:

$$R_{XO} = f_O + \frac{1}{2}(1 + f_X) \qquad (30)$$

and the numerator relationship of an animal (O) to itself is $1+f_O$.

To avoid tracing out each line of descent on each pedigree as necessary in (14), one may use methods of combining the numerator relationships. One method involves the determination of the numerator relationship of a sire to each of his ancestors through which he may be related to any dam. These numerator relationships would be arranged in a table with the appropriate derivatives in columns designating the number of generations that the common ancestor may be removed from the dam. Derivatives are then added for each ancestor in the dam's pedigree; the result being divided by two to give the inbreeding coefficient of the offspring.

Advantages of this procedure include a) the ability to calculate the inbreeding coefficient of offspring resulting from crossing related inbred lines; and b) the ability to determine the degree of inbreeding in a herd or flock at certain intervals of time where it is not practical to calculate coefficients continuously. The sire-ancestor procedure is very useful when the number of females is large and the number of males small. It is also used for breeding plans that attempt to avoid inbreeding to determine if that requirement has been met.

Numerator Relationship Coefficient Charts

The preparation of numerator relationship charts for all the animals in an inbred line is necessary for this procedure (Emik and Terrill, 1949). These charts are an attempt to simplify the calculation of inbreeding coefficients.

The charts may be initiated by calculating the numerator relationships of the foundation animals to each other by ordinary pedigree analysis. The sire-ancestor procedure just described is useful for this purpose. Then numerator relationships may be computed by use of the formula:

$$R_{XY} = 0.5(R_{SY} + R_{DY}) \qquad (31)$$

To reduce the number of numerator relationships to be calculated to a minimum one may obtain only the relationship of each generation group with the preceding and succeeding generation groups. (It is impossible to follow this plan exactly if the generations are irregular.)

One must be aware that precautions are necessary in developing the charts. The numerator relationships which are used to obtain the relationship of one animal to another should always be the younger animal to the older animal. This is essential if the two animals are in direct lines of descent. All work should be independently checked. Errors in recording the numbers of the sire and dam or in calculating or recording the inbreeding as relationship coefficients may be carried on indefinitely as they are not apt to be detected in later work.

When inbred lines are first started, the calculation of inbreeding coefficients by pedigree inspection or by the sire-ancestor method may be more rapid than the development of numerator relationship charts. However, after five to ten generations of inbreeding this will not be true.

When relatively small numbers of females are involved within a line and when inbreeding is to be continued for many generations, the numerator relationship charts method proves particularly useful. These charts are also more efficient when an inbreeding coefficient is needed for each offspring from the line.

Procedure for Closed Population

The computation of inbreeding coefficients for isolates of limited size is often a laborious procedure. A method is available which permits the accumulation of data, so that the inbreeding coefficients for any generation may be directly determined from those obtained for preceding ones (Cruden, 1949). Thus eliminating the preparation and examination of long pedigree charts. (A second advantage is to be found in the speed of computation since the data obtained for any one generation furnish the basis of calculation for each succeeding generation.)

The method requires the computation of inbreeding coefficients of all possible matings and some hypothetical matings for a single (hereafter referred to as the base) generation early in the history of the line. The coefficients for later generations are then constructed as simple functions of the coefficients of the base generation.

This method yields the same result without requiring that any paths be traced. It is based on the fact that the inbreeding coefficients of the offspring of two parents is equal to the average inbreeding coefficients of offspring, perhaps hypothetical, from any one of the following examples of matings:

1. Paternal parent mated with each of the two maternal grandparents;
2. Maternal parent mated with each of the two paternal grandparents;

3. Each of the two maternal grandparents with each of the paternal grandparents;
4. Each of the two maternal grandparents with each of the four paternal grandparents.

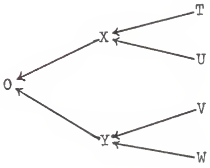


Fig. 1. Sample Pedigree.

For example, from Fig. 1, if we designate the coefficient of inbreeding which would have been obtained for the progeny of any two animals 1 and 2 by f_{12}, f_0 is equal to any of the following expressions:

$$\frac{f_{XV} + f_{XW}}{2}$$

$$\frac{f_{YT} + f_{YU}}{2}$$

$$\frac{f_{TV} + f_{TW} + f_{UV} + f_{UW}}{4} \qquad (32)$$

It may be noted in the formulation presented that the actual sexes of the hypothetical parents need not be taken into account in using the coefficient f_{12}. In fact, self-fertilization, represented by f_{11}, may be assumed without prejudice to the technique.

It must be emphasized that a coefficient which represented, hypothetically, a self-fertilization cannot be expressed as an average of other coefficients.

Covariance Charts

Wright's original formula (14) shows that the coefficient of inbreeding of an individual is simply ½ of the genic covariance between the individual's sire and dam.

In many cases inbreeding coefficients may be computed quite rapidly from covariance charts involving only: 1) the mates to the females in the direct female line of ancestry, often referred to as the bottom of the pedigree; and 2) the females that have female descendants represented in the population and at the same time are ancestors of one of the mates (Plum, 1954). The number of these females is often very small.

The inbreeding of the individual 0 in Fig. 2 is (cov B_0A_0)/2, but according to the procedure preceding this one

$$\text{cov } B_0A_0 = (\text{cov } B_0B_1)/2 + (\text{cov } B_0A_1)/2$$

and proceeding with this expansion, we arrive at the general formula which is

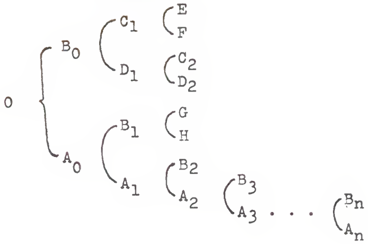


Fig. 2. The Pedigree of Individual 0 Represented in the Conventional Form.

$$\text{cov } B_0A_0 = (\text{cov } B_0B_1)/2 + (\text{cov } B_0B_2)/4 + \cdots + (\text{cov } B_0B_n)/2^n + (\text{cov } B_0A_n)/2^n \qquad (33)$$

where B_0, B_1, . . . , B_n are the mates to the females in the direct female line at the bottom of the pedigree (A_0, A_1, . . . , A_n).

Formula (33) indicates that only the direct female line together with their mates needs to be traced in order to compute the inbreeding coefficient of individual 0. Since all calculations of inbreeding coefficients are relative to some base date, the direct female line needs only to be traced back to this base date after which the term (cov B_0A_n)/2^n may be dropped from the formula in most cases.

There is one limitation: whenever one of the A-animals is also an ancestor of one of the B-animals, the computation cannot be carried back of the particular A-animal in question. For example, if A_3 is a common ancestor to both A_2 and B_2 the computation cannot be carried back to A_3. The genic covariance between B_0 and A_3 must be computed and the complete formula for the genic covariance between A_0 and B_0 will be:

$$\text{cov } B_0A_0 = (\text{cov } B_0B_1)/2 + (\text{cov } B_0B_2)/4 + (\text{cov } B_0B_3)/8 + (\text{cov } B_0A_3)/8 \; . \qquad (34)$$

In most cases the last term of this formula may be computed by going back of the individual which is not a common ancestor (B_0) because cov B_0A_3 = (cov A_3C_1)/2 + (cov A_3D_1)/2 and this formula may be further expanded according to the principle of formula (33).

When applying this procedure, the first step is to tabulate the direct female ancestry together with their mates for each animal whose inbreeding coefficient is to be computed. Once this is done the actual covariance chart will be limited to the males appearing in these female ("family") pedigrees. If individuals which are (female) ancestors of some of the males appear in any of the "families", these "foundation" females should also be included in the covariance chart. When a covariance chart has been computed, the inbreeding of any individual may be computed by means of formula (33) or (34).

The covariances between the males and the foundation females may be computed by the use of punched cards, in accordance with the following procedure. If however, the males are bred within the population under study, it may be simpler to start with the foundation males and females and work forward according to the principle outlined by the previous procedure using either formula (33) or (34).

From Punched Cards

The calculation of the coefficient of inbreeding may be reduced from a complex operation to a routine procedure with the use of punched cards (Hazel and Lush, 1950). Although it becomes primarily mechanical and clerical in operation, the numerical results of this procedure are identical with those of Wright's original formula (14). The steps of this procedure will not be discussed in this report; for further information one may refer to the source cited.

ACKNOWLEDGEMENTS

The writer wishes to express his appreciation to Dr. H. C. Fryer and Dr. A. D. Dayton, of the Department of Statistics, Kansas State University, for their helpful suggestions and advice during the preparation of this report.

The writer would also like to express his gratitude to Dr. R. F. Nassar, of the Department of Statistics, and Dr. K. A. Huston, of the Department of Dairy Science, for their assistance in the preparation of this report.

REFERENCES

Cruden, D. 1949. The Computation of Inbreeding Coefficients for Closed Populations. *J. of Heredity*, 40: 248-251.

Emik, L. O. and C. E. Terrill. 1949. Systematic Procedures for Calculating Inbreeding Coefficients. *J. of Heredity*, 40: 51-55.

Hazel, L. N. and J. L. Lush. 1950. Computing Inbreeding and Relationship Coefficients from Punched Cards. *J. of Heredity*, 41: 301-306.

Kempthorne, O. 1957. *An Introduction to Genetic Statistics*. John Wiley and Sons, Inc., New York.

Li, C. C. and D. G. Horvitz. 1953. Some Methods of Estimating the Inbreeding Coefficient. *Am. J. of Human Genetics*, 5: 107-117.

Li, C. C. 1955. *Population Genetics*. The University of Chicago Press, Chicago. Fifth Impression, 1966.

Malécot, G. 1948. *Les Mathématiques de L'Hérédité*. Masson Et C^{ie}, Editeurs, Paris.

Plum, M. 1954. Computation on Inbreeding and Relationship Coefficients. *J. of Heredity*, 45: 92-94.

Tukey, J. W. 1954. Causation, Regression, and Path Analysis. *Statistics and Mathematics in Biology*, O. Kempthorne, T. Bancroft, J. Gowen and J. Lush, editors. The Iowa State College Press, Ames.

Wright, S. 1921. Correlation and Causation. *J. of Agric. Research*, 20: 557-585.

Wright, S. 1922. Coefficients of Inbreeding and Relationship. *Am. Naturalist*, 56: 330-338.

Wright, S. 1923a. The Theory of Path Coefficients--A Reply to Nile's Criticism. *Genetics*, 8: 239-255.

Wright, S. 1923b. Mendelian Analysis of the Pure Breeds of Livestock. I. The Measurement of Inbreeding and Relationship. *J. of Heredity*, 24: 339-348.

Wright, S. and H. C. McPhee. 1925. An Approximate Method of Calculating Coefficients of Inbreeding and Relationship. *J. of Agric. Research*, 31: 377-383.

Wright, S. 1934. The Method of Path Coefficients. *Ann. of Math. Stat.*, 5: 161-215.

PROCEDURES FOR CALCULATING ESTIMATES OF THE COEFFICIENT OF INBREEDING OF AN INDIVIDUAL

by

KEITH LAVERNE HOFFMAN

B. S., Kansas State University, 1965

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

This report discusses the procedures involved in calculating the estimates of F (coefficient or inbreeding) of an individual. The earliest procedures were as follows:

A. Wright's original formula using path coefficients, and

B. approximating procedure involving random sampling.

Malécot approached the problem of calculating the coefficient of inbreeding by considering probabilities.

When models based on a panmictic population are reviewed, one may arrive at these procedures:

A. method involving proportions of heterozygotes;

B. product-moment correlation method using a table of gametic correlations:

C. use of the determinant of the matrix of the previous table of gametic correlations:

D. method involving a value of Chi-square;

E. method considering the proportions of alleles in homozygous condition;

F. maximum likelihood method of estimation; and

G. method involving reduction of the number of alleles.

The final set of procedures considered in this report may be classified as systematic. These include:

A. sire-ancestor method for irregular systems of inbreeding;

B. preparation of numerator relationship charts;

C. procedure for a closed population of limited size;

D. application of covariance charts; and

E. procedure using punched cards.

This report outlines the individual steps in each procedure that are necessary to arrive at an estimate of the coefficient of inbreeding. In some instances, procedures are compared as to their effectiveness under given circumstances; and, advantages and disadvantages are stated.