

Paternity Testing: I. Calculation of Paternity Indexes

L. Kaiser and G.A.F. Seber

Division of Statistics, University of California at Davis, and Department of Mathematics and Statistics, Auckland University, New Zealand

An algorithm, readily adaptable to microcomputers, is given for computing paternity indexes. A closed-form expression based only on gene frequencies and phenotype structures is derived for the paternity index for a given mother/child/alleged father trio and any blood group system. This above work is applied to the problem of estimating gene frequencies from sample data.

Key words: paternity testing, paternity index, gene frequencies, gene count method

INTRODUCTION

With the growing interest in blood group typing, as applied to paternity cases and forensic medicine [cf Dodd and Lincoln, 1978; Salmon and Salmon, 1980], a need has arisen for the automatic calculation of various probabilities and indexes. A particularly useful index, commonly called the Paternity Index, is the ratio of two probabilities, $P(C|M,AF)/P(C|M)$. For a given blood group system, $P(C|M,AF)$ is the conditional probability of obtaining the child's phenotype (C), given the phenotypes of the mother (M) and alleged father (AF), and $P(C|M)$ is the conditional probability of obtaining the child's phenotype, given the mother and a random father (RF). Once this ratio is calculated for each system ABO, Rh, etc, an overall index is obtained by multiplying all the ratios together.

Lee [1980] proposed a method of computing the individual and overall paternity indexes based on storing a table of indexes for each system and, given a particular entry for C, M, and AF, searching each table for the appropriate index. However the calculation of the tables is complicated, particularly for the Rh system, and the search procedure computationally inefficient. More recently [eg, Chastang, 1976; Minakata et al, 1980] interest has centered on methods for calculating paternity indexes directly without the need for searching tables. We give below such an algorithm for calculating a paternity

Received for publication November 2, 1982; revision received January 24, 1983.

Address reprint requests to Dr. G.A.F. Seber, Department of Mathematics and Statistics, Auckland University, Private Bag, Auckland, New Zealand.

index for any child, mother, and alleged father combination, which is particularly adaptable to microcomputers. The algorithm is demonstrated using an example from the ABO system.

EXAMPLE OF THE ALGORITHM

Consider the case when the child has phenotype A_1 , the mother phenotype A_2 , and the alleged father phenotype A_1 in the ABO system. Then the paternity index is

$$\frac{P(C = A_1 | M = A_2, AF = A_1)}{P(C = A_1 | M = A_2)} \quad (1)$$

We begin with the denominator. The phenotype A_1 is any one of the genotypes A_1A_1 , A_1A_2 , and A_1O so that

$$\begin{aligned} P(C = A_1 | M = A_2) &= P(A_1A_1 | M = A_2) + P(A_1A_2 | M = A_2) + P(A_1O | M = A_2) \\ &= 0 + P(M \text{ transmits } A_2 \text{ gamete} | M = A_2)P(RF \text{ transmits } A_1 \text{ gamete}) \\ &\quad + P(M \text{ transmits } O \text{ gamete} | M = A_2)P(RF \text{ transmits } A_1 \text{ gamete}). \end{aligned}$$

The conditional probabilities that the mother transmits a particular gamete are listed in Table I: the row position gives the gamete, and the column position the phenotype of the mother. The letters a_1 , a_2 , etc are the gene probabilities (relative frequencies) for the population. Thus, using matrices and the columns of Table I:

$$\begin{aligned} P(C = A_1 | M = A_2) &= \frac{a_2 + \theta}{a_2 + 2\theta} \cdot a_1 + \frac{\theta}{a_2 + 2\theta} \cdot a_1 \\ &= (\text{column } A_2)' \mathbf{H}(\text{column RF}). \end{aligned} \quad (2)$$

TABLE I. Entry in the (j,k)th Position is the Probability That a Person Produces Gamete j, Given Their Phenotype is k, for the ABO Group.

Gamete	RF ^a	Phenotype					
		A_1 (1)	A_2 (2)	B (3)	O (4)	A_1B (5)	A_2B (6)
A_1 (1)	a_1	$\frac{a_1 + a_2 + \theta}{a_1 + 2a_2 + 2\theta}$	ϕ	ϕ	ϕ	$\frac{1}{2}$	ϕ
A_2 (2)	a_2	$\frac{a_2}{a_1 + 2a_2 + 2\theta}$	$\frac{a_2 + \theta}{a_2 + 2\theta}$	ϕ	ϕ	ϕ	$\frac{1}{2}$
B (3)	b	ϕ	ϕ	$\frac{b + \theta}{b + 2\theta}$	ϕ	$\frac{1}{2}$	$\frac{1}{2}$
O (4)	θ	$\frac{\theta}{a_1 + 2a_2 + 2\theta}$	$\frac{\theta}{a_2 + 2\theta}$	$\frac{\theta}{b + 2\theta}$	1	ϕ	ϕ

^aColumn RF gives the gene probabilities for a random man: $\phi = 0$ and denotes an impossible combination.

The matrix **H** is given by

$$\mathbf{H} = \begin{matrix} & \begin{matrix} A_1 & A_2 & B & O \end{matrix} \\ \begin{matrix} A_1 \\ A_2 \\ B \\ O \end{matrix} & \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{matrix} \quad (3)$$

and it is determined by the phenotype of the child. The rows and columns are both labeled A_1 , A_2 , B , and O , in that order, and a 1 is entered in the matrix if the combination of genes yields the phenotype of the child, otherwise there is zero. Since $C = A_1$ there is 1 in the A_1A_1 , A_1A_2 , A_2A_1 , A_1O , and OA_1 positions; and zero in the OO , A_1B , etc positions.

Similar arguments apply to the numerator of (1). We have

$$\begin{aligned}
 P(C = A_1 | M = A_2, AF = A_1) &= P(M \text{ transmits } A_2 | M = A_2)P(AF \text{ transmits } A_1 | AF = A_1) \\
 &\quad + P(M \text{ transmits } O | M = A_2)P(AF \text{ transmits } A_1 | AF = A_1) \\
 &= \frac{a_2 + \theta}{a_2 + 2\theta} \cdot \frac{a_1 + a_2 + \theta}{a_1 + 2a_2 + 2\theta} + \frac{\theta}{a_2 + 2\theta} \cdot \frac{a_1 + a_2 + \theta}{a_1 + 2a_2 + 2\theta} \\
 &= (\text{column } A_2)' \mathbf{H} (\text{column } A_1). \quad (4)
 \end{aligned}$$

which is simply equation (2) with RF replaced by AF. The computation of (2) and (4), therefore, involves the columns of Table I and the symmetric **H** matrix corresponding to the child's phenotype. For example the **H** matrices corresponding to $C = O$ and $C = A_2$ are

$$O: \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad A_2: \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

In general, if we have a trio (C,M,AF) with respective phenotypes (X,Y,Z), then

$$P(C = X | M = Y) = (\text{column for } Y)' (\text{matrix for } X) (\text{column for } RF)$$

and

$$P(C = X | M = Y, AF = Z) = (\text{column for } Y)' (\text{matrix for } X) (\text{column for } Z)$$

$$= \mathbf{y}'\mathbf{H}\mathbf{xz}, \text{ say.} \tag{5}$$

$$= \sum_{i=1}^4 \sum_{j=1}^4 y(i)h_{ij}z(j), \tag{6}$$

where $y(i)$ is the i th element of \mathbf{y} etc.

In devising a suitable algorithm it would appear that we have to store the expressions in Table I and an \mathbf{H} matrix for each phenotype in the computer. However, we shall see in the next section that Table I itself may be generated from the gene frequencies and the \mathbf{H} matrices. Further, we now show that only the nonzero elements of each \mathbf{H} matrix need to be stored.

Let (i,j) denote the position of an element of \mathbf{H} in the intersection of the i th row and j th column of \mathbf{H} . If \mathbf{H} is given by (3), then we have five positions where the element is unity, namely $(1,1)$, $(1,2)$, $(2,1)$, $(1,3)$, and $(3,1)$. If we label these $k = 1, 2, 3, 4, 5$, the five positions are $(i[k],j[k])$, where $i[1] = 1, j[1] = 1; i[2] = 1, j[2] = 2; i[3] = 2, j[3] = 1; i[4] = 1, j[4] = 3; \text{ and } i[5] = 3, j[5] = 1$. Then, from (6),

$$\mathbf{y}'\mathbf{H}\mathbf{xz} = \sum_{k=1}^5 y(i[k])z(j[k]),$$

and this can be computed, given the five values of the pair $(i[k],j[k])$.

The above algorithm can be applied to any blood group system but it is most useful when there are a large number of phenotypes as for example in the ABO, MNSs, and, in particular, the Rh and HLA systems.

In practice the \mathbf{H} matrices do not need to be stored separately: in fact only one matrix is needed. If we place in the (i,j) th position of the following matrix the phenotype arising from a combination of gene i and gene j , we get

$$\begin{matrix} & A_1 & A_2 & B & O \\ \begin{matrix} A_1 \\ A_2 \\ B \\ O \end{matrix} & \begin{bmatrix} A_1 & A_1 & A_1B & A_1 \\ A_1 & A_2 & A_2B & A_2 \\ A_1B & A_2B & B & B \\ A_1 & A_2 & B & O \end{bmatrix} \end{matrix}$$

With the phenotypes suitably coded, this is the matrix referred to by Minakata et al [1980] as the translational matrix. To obtain the \mathbf{H} matrix for $C = A_1$ we set $A_1 = 1$ and all other entries equal to zero in the above matrix.

FORMULA FOR PATERNITY INDEX

To allow application to any blood group system it is convenient to have a change in notation. Consider a particular system with r gametes and s phenotypes, gene probabilities $\mathbf{g}' = (g_1, g_2, \dots, g_r)$, and \mathbf{H} matrices $\mathbf{H}(k) = [(h_{ij}(k)], k = 1, 2, \dots, s$. Since

$$P(C = c | M = m) = \sum_{i=1}^r \sum_{j=1}^r h_{ij}(c)P(M \text{ transmits } i | M = m)P(RF \text{ transmits } j),$$

we see that (2) holds in general. A similar expression shows that (4) also holds in general. As a first step we now derive an expression for the conditional probability of observing a particular gamete from a person with known phenotype.

Let (i,j) denote the genotype of a person arising from gametes i and j . Then, for a randomly chosen person from the population and a randomly chosen gamete from this person we have

$$\begin{aligned} & P(\text{gamete is } j \mid \text{phenotype is } k) \\ &= \sum_{i=1}^r P(\text{gamete } j \mid \text{genotype } (i,j), \text{phenotype } k)P(\text{genotype } (i,j) \mid \text{phenotype } k) \\ &= \sum_{i=1}^r P(\text{gamete } j \mid \text{genotype } (i,j))P(\text{genotype } (i,j) \mid \text{phenotype } k), \end{aligned} \tag{7}$$

since the second probability in (7) is zero if phenotype k is inconsistent with genotype (i,j) . Now

$$\begin{aligned} & P(\text{genotype } (i,j) \mid \text{phenotype } k) \\ &= \frac{P(\text{phenotype } k \mid \text{genotype } (i,j))P(\text{genotype } (i,j))}{P(\text{phenotype } k)} \\ &= \begin{cases} \frac{h_{ij}(k)2g_i g_j}{\mathbf{g}'\mathbf{H}(k)\mathbf{g}}, & i \neq j \\ \frac{h_{ii}(k)g_i^2}{\mathbf{g}'\mathbf{H}(k)\mathbf{g}}, & i = j \end{cases} \end{aligned}$$

and

$$P(\text{gamete } j \mid \text{genotype } (i,j)) = \begin{cases} \frac{1}{2}, & i \neq j \\ 1, & i = j. \end{cases}$$

Let \mathbf{I}_r be the $r \times r$ identity matrix with columns $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_r$. Then, substituting in (7),

$$\begin{aligned} P(\text{gamete } j \mid \text{phenotype } k) &= \{1 \cdot h_{jj}(k)g_j^2 + \sum_{\substack{i=1 \\ i \neq j}}^r \frac{1}{2}h_{ij}(k)2g_i g_j\} / \mathbf{g}'\mathbf{H}(k)\mathbf{g} \\ &= g_j \sum_{i=1}^r h_{ij}(k)g_i / \mathbf{g}'\mathbf{H}(k)\mathbf{g} \\ &= \mathbf{g}'\mathbf{e}_j\mathbf{e}_j'\mathbf{H}(k)\mathbf{g} / \mathbf{g}'\mathbf{H}(k)\mathbf{g}, \end{aligned} \tag{8}$$

and this is the (j,k) th element of a matrix, \mathbf{F}_g say (cf Table I for the ABO system). Let $\mathbf{G} = \text{diag}(g_1, g_2, \dots, g_r)$, then the column of \mathbf{F}_g corresponding to phenotype k is the column vector $\mathbf{GH}(k)\mathbf{g} / \mathbf{g}'\mathbf{H}(k)\mathbf{g}$, and the column in Table I corresponding to RF is \mathbf{g} . Therefore the paternity index for $C = c, M = m$, and $AF = a$ is, from (5),

$$\frac{(\text{col. for m})' \mathbf{H}(\text{c})(\text{col. for a})}{(\text{col. for m})' \mathbf{H}(\text{c})(\text{col. for RF})} = \frac{\mathbf{g}' \mathbf{H}(\text{m}) \mathbf{G} \mathbf{H}(\text{c}) \mathbf{G} \mathbf{H}(\text{a}) \mathbf{g}}{\mathbf{g}' \mathbf{H}(\text{m}) \mathbf{G} \mathbf{H}(\text{c}) \mathbf{g} \cdot \mathbf{g}' \mathbf{H}(\text{a}) \mathbf{g}}$$

This formula can be used to compute a paternity index directly and forms the basis of several likelihood ratio tests for paternity to be considered in a further paper.

ESTIMATION OF GENE PROBABILITIES

Given a random sample of size n from a population, we can estimate the probability that an individual has a given phenotype in a particular blood group system by the relative frequency of that phenotype in the sample. For example, in the ABO system, if $P(A_1)$ is the probability of phenotype A_1 , then it can be estimated by $n(A_1)/n$, where $n(A_1)$ is the number in the sample with phenotype A_1 . From these phenotype probability estimates we can then obtain estimates of the gene probabilities using a method called gene counts which is equivalent to maximum likelihood [Elandt-Johnson, 1971]. For example, in the ABO system we can set up the following equations for the gene probabilities a_1 , a_2 , b , and θ :

$$\begin{aligned} a_1 &= \sum_{\text{phenotypes}} P(\text{RM transmits } a_1 \mid \text{phenotype}) P(\text{phenotype}) \\ &= \frac{a_1 + a_2 + \theta}{a_1 + 2a_2 + 2\theta} P(A_1) + \frac{1}{2} P(A_1B), \end{aligned}$$

$$a_2 = \frac{a_2}{a_1 + 2a_2 + 2\theta} P(A_1) + \frac{a_2 + \theta}{a_2 + 2\theta} P(A_2) + \frac{1}{2} P(A_2B),$$

$$b = \frac{b + \theta}{b + 2\theta} P(B) + \frac{1}{2} P(A_1B) + \frac{1}{2} P(A_2B),$$

and

$$\theta = 1 - a_1 - a_2 - b.$$

Using the order in Table I, if $\mathbf{g}' = (a_1, a_2, b, \theta)$,

$$\mathbf{P}' = [P(A_1), P(A_2), P(B), P(O), P(A_1B), P(A_2B)],$$

and \mathbf{F}_g (a function of \mathbf{g}) is the matrix given by Table I with column RF omitted, then the above equations can be expressed in the compact matrix form

$$\mathbf{g} = \mathbf{F}_g \mathbf{P}.$$

To estimate \mathbf{g} we replace each phenotype probability $P(\cdot)$ by its estimate $\hat{P}(\cdot) = n(\cdot)/n$, form the column of estimates $\hat{\mathbf{P}}$, and solve the nonlinear equations $\mathbf{g} = \mathbf{F}_g \hat{\mathbf{P}}$ for \mathbf{g} . The simplest way of doing this is to begin with an initial estimate \mathbf{g}_1 of \mathbf{g} and then solve for \mathbf{g} by iterating the cycle $\mathbf{g}_{i+1} = \mathbf{F}_{\mathbf{g}_i} \hat{\mathbf{P}}$ until \mathbf{g}_{i+1} stabilizes.

We conclude that the gene count method can readily be incorporated into the above algorithm for computing paternity indexes. The last values of \mathbf{g}_i and $\mathbf{F}_{\mathbf{g}_i}$ give us an estimate of Table I, which is then used for paternity index calculations. The only further information

required is an initial estimate of \mathbf{g} [cf Elandt-Johnson, 1971]. Using the gene count method in tandem with the above algorithm is particularly appropriate when there is a paucity of phenotype data as in some racial groups. Every new item of data then needs to be used and the estimate of \mathbf{g} continuously updated by adding into $\hat{\mathbf{P}}$. However, if an accurate estimate of \mathbf{g} is available, then $\mathbf{F}_{\mathbf{g}}$ does not need to be derived algebraically and stored, and its elements can be computed as part of the algorithm using (8).

ACKNOWLEDGMENTS

We would like to thank Dr. Graeme Woodfield, Director of the Auckland Blood Transfusion Center, for introducing us to this subject.

REFERENCES

- Chastang C: Pater (1976): a program for parenthood diagnosis. *Comput Programs Biomed* 76:251-258.
- Dodd BE, Lincoln PJ (1978): An analysis of 1,556 cases of doubtful paternity submitted for blood group investigation. *Med Sci Law* 18:185-190.
- Elandt-Johnson RC (1971): "Probability Models and Statistical Methods in Genetics." New York: Wiley.
- Lee CL (1980): Numerical expression of paternity test results using predetermined indexes. *Am J Clin Pathol* 73:522-536.
- Minakata K, Asano M, Hattori H (1980): A new approach to the computation of indices of paternity. *Comput Programs Biomed* 80:191-200.
- Salmon D, Salmon S (1980): Blood groups and genetic markers polymorphism and probability of paternity. *Transfusion* 20:684-694.

Edited by John M. Opitz