Ann. Hum. Genet., Lond. (1975), **39**, 173
Printed in Great Britain

The estimation of pairwise relationships

BY E. A. THOMPSON*
The Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge

I. INTRODUCTION

Pairwise relationships between individuals have been extensively studied from the point of view of conditional prediction or genetic counselling. Cotterman (1940) first introduced the '*k-coefficients*' which are a sufficient specification of the relationship between any two non-inbred individuals;

$$\boldsymbol{k} = (k_0, k_1, k_2), \quad k_i \geqslant 0, \quad k_0 + 2k_1 + k_2 = 1, \tag{1}$$

where k_0, $2k_1$ and k_2 are (respectively) the probabilities that the two individuals have 0, 1 and 2 genes in common (see, for example, Crow & Kimura (1970)). In this paper it will be more convenient to use the coefficients

$$\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2), \quad \text{where} \quad \kappa_0 = k_0, \quad \kappa_1 = 2k_1 \quad \text{and} \quad \kappa_2 = k_2, \tag{1'}$$

κ_i thus being the probability of i genes in common.

Cotterman's work has been extended by, amongst many others, Malécot (1948) and Li & Sacks (1954), who first gave the pairwise genotype distributions for two related non-inbred individuals. The specification of relationships and the derivation of joint genotype distributions has been extended by Jacquard (1972) to cases of two possibly inbred individuals, and by Thompson (1974*a*) to cases of several individuals.

Edwards (1967) suggested that the relationships between the members of a population should be studied from the point of view of inference rather than of prediction. Given sufficient data on individual genotypes, and perhaps also the ages and sexes of individuals, it should be possible to reconstruct the detailed genealogy of any population. It may be shown that in practice sufficient data may well be available to enable useful inferences regarding population structure to be made. Turnbull (1972) has detailed some of the problems of anthropologists in obtaining accurate information of relationships in some small-scale societies, and it has been estimated that often only 80% may be correct. With currently obtainable data it is possible to reconstruct the genealogy of a small population with at least this accuracy (Thompson, 1974*b*). Methods of genealogy reconstruction may also be of practical use in animal populations, but human populations provide a useful starting point, particularly since the most suitable genetic data as yet available are those on the human blood-group polymorphisms.

However, a study of structure or relationship inferences has a further purpose. Even where inference as to detailed population structure is not yet a practical proposition, a study of relationships from the point of view of inference rather than prediction can greatly clarify the relation between genealogical and genetic structure.

To consider the complete structure of a population it is clearly necessary to consider joint relationships between an arbitrary number of individuals, and it is towards this end that

* Address for academic year 1974/5: Department of Genetics, Stanford, California 94305.

Thompson (1974*a*) has considered the specification of multiple genetic relationships. In this paper we consider a problem which introduces ideas on which more general population structure inferences may be based. We wish to infer the pairwise relationship between two individuals, given their genotypes at several loci, and to do so we consider the likelihood of the Cotterman k-coefficients of equation (1) (or $\boldsymbol{\kappa}$ of equation (1′).

II. THE LIKELIHOOD FOR PAIRWISE RELATIONSHIPS

Consider first the problem of inferring the relationship $\boldsymbol{\kappa}(A, B)$ between two individuals A and B who are assumed to be non-inbred. We assume that data are available on the genotypes $G(A)$ and $G(B)$ of A and B at s unlinked autosomal loci, for which the allele frequencies in the general population are known. These restrictions are discussed in Section VI. For pairwise relationships we may consider the two individuals to be either ordered or unordered, provided the convention is not varied with the hypothesis of relationship considered. Relationship coefficients are symmetrical with respect to just two individuals; $\boldsymbol{\kappa}(A, B) = \boldsymbol{\kappa}(B, A)$. Although there is clearly a distinction between father and son in age, or mother and son in sex, in a large population obeying the Hardy–Weinberg law the relationships offspring and parent, etc., are identical as far as autosomal genotype probabilities are concerned. However, for joint relationships between several individuals there may be alternative hypotheses which do not consider all individuals equally – for example, a hypothesis of joint sibship against an alternative where two individuals are sibs but the specified third individual is unrelated to them. The individuals are then necessarily ordered, and for consistency we thus consider always ordered sets of genotypes.

Consider first a single t-allele locus, with alleles a_i having frequencies p_i

$$\left(1 \leqslant i \leqslant t, \sum_{i=1} p_i = 1\right).$$

Suppose that A and B have genotypes G_1 and G_2 respectively and let $r(A, B)$ be the number of genes common to A and B at the locus in question. Then the likelihood for relationship $\boldsymbol{\kappa}$ between A and B is

$$\begin{aligned} L(\boldsymbol{\kappa}) &= \Pr[G(A) = G_1,\ G(B) = G_2 | \boldsymbol{\kappa}(A, B) = \boldsymbol{\kappa}] \\ &= \sum_{i=0}^{2} \Pr[G(A) = G_1,\ G(B) = G_2 | r(A, B) = i] \Pr[r(A, B) = i | \boldsymbol{\kappa}(A, B) = \boldsymbol{\kappa}], \end{aligned}$$

or

$$L(\boldsymbol{\kappa}) = \sum_{i=0}^{2} \kappa_i P_i(G_1, G_2), \tag{2}$$

where

$$P_i(G_1, G_2) = \Pr[G(A) = G_1,\ G(B) = G_2 | r(A, B) = i]. \tag{3}$$

Thus provided $P_i(G_1, G_2)$ may be simply computed for any ordered genotype pair (G_1, G_2), the likelihood of any given genetic relationship $\boldsymbol{\kappa}$ may be found. Table 1 gives the required probabilities for each type of genotype pair. Equation (2) is equivalent to that given by Li & Sacks (1954), while Jacquard (1972) gives an extension of this equation which covers cases in which A and/or B may be inbred.

Suppose now that at locus j the genotypes of A and B are $G_1^{(j)}$ and $G_2^{(j)}$ respectively $(1 \leqslant j \leqslant s)$ and that

$$\Pr[G(A) = G_1^{(j)},\ G(B) = G_2^{(j)} | r(A, B) = i] = P_i^{(j)}(G_1^{(j)}, G_2^{(j)}), \tag{3′}$$

Table 1. *Values of $P_i(G_1, G_2)$ for a locus with t alleles a_j with frequencies p_j $(j = 1, \ldots, t)$*

(Note that $\sum_{z \in \mathbf{Z}} n(z;t) = (t(t+1)/2)^2$ and $\sum_{(G_1, G_2) \in \mathbf{G}} P_i(G_1, G_2) = 1$ for $i = 0, 1, 2$ as is required.)

Type of genotype pair; $z \in \mathbf{Z}$	Ordered genotype pair $(G_1, G_2) \in \mathbf{G}$: G_1	G_2	No. of different ordered genotype pairs of type z; $n(z;t)$	$P_h(G_1, G_2)$: $h = 0$	$h = 1$	$h = 2$
1	a_ia_i	a_ia_i	t	p_i^4	p_i^3	p_i^2
$2a$	a_ia_i	a_ia_j	$t(t-1)$	$2p_i^3p_j$	$p_i^3p_j$	0
$2b$	a_ia_j	a_ia_i	$t(t-1)$			
3	a_ia_i	a_ja_j	$t(t-1)$	$p_i^2p_j^2$	0	0
4	a_ia_j	a_ia_j	$t(t-1)/2$	$4p_i^2p_j^2$	$p_ip_j(p_i+p_j)$	$2p_ip_j$
$5a$	a_ia_i	a_ja_l	$t(t-1)(t-2)/2$	$2p_i^2p_jp_l$	0	0
$5b$	a_ja_l	a_ia_i	$t(t-1)(t-2)/2$			
6	a_ia_j	a_ia_l	$t(t-1)(t-2)$	$4p_i^2p_jp_l$	$p_ip_jp_l$	0
7	a_ia_j	a_la_m	$t(t-1)(t-2)(t-3)/4$	$4p_ip_jp_lp_m$	0	0

Table 2. *$\boldsymbol{\varkappa}$-values for some standard genealogical relationships, in the absence of inbreeding (other relationships have $\kappa_1 \leqslant \frac{1}{2}$ and $\kappa_2 \leqslant \frac{1}{16}$)*

Relationship of A to B	Code letter	$\boldsymbol{\varkappa}(A, B)$: κ_0	κ_1	κ_2
Unrelated	U	1	0	0
Offspring, parent	Q	0	1	0
Sib	B	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Self, identical twin	R	0	0	1
Niece, nephew, uncle, aunt; Grandparent, grandchild; Half-sib	N	$\frac{1}{2}$	$\frac{1}{2}$	0
First cousin; Parent's half-sib, half-sib's child	C	$\frac{3}{4}$	$\frac{1}{4}$	0
Double first cousin	D	$\frac{9}{16}$	$\frac{6}{16}$	$\frac{1}{16}$
Half-sibs whose non-identical parents are				
(i) sibs or parent–offspring	NB, NQ	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$
(ii) half-sibs	NN	$\frac{7}{16}$	$\frac{1}{2}$	$\frac{1}{16}$

where a superscript (j) is attached to the function P_i since it is dependent upon the allele frequencies at locus j. The likelihood is then multiplicative over the s independent (since assumed unlinked) loci.

$$L_s(\boldsymbol{\varkappa}) = \prod_{j=1}^{s} \left(\sum_{i=0}^{2} \kappa_i P_i^{(j)}(G_1^{(j)}, G_2^{(j)}) \right) \tag{2'}$$

and the support, or log-likelihood, is

$$S_s(\boldsymbol{\varkappa}) = \log_e L_s(\boldsymbol{\varkappa}) = \sum_{j=1}^{s} \log_e \left(\sum_{i=0}^{2} \kappa_i P_i^{(j)}(G_1^{(j)}, G_2^{(j)}) \right). \tag{4}$$

Several different genealogical relationships have the same $\boldsymbol{\varkappa}$; clearly these cannot be distinguished on the basis of genotype data alone. For example, data on the relative ages of individuals is required to distinguish a grandparent from a half-sib. Table 2 gives the $\boldsymbol{\varkappa}$ values for some standard genealogical relationships in the absence of inbreeding of common ancestors; we shall identify these genetic relationships by the given code letters. Only these nine relation-

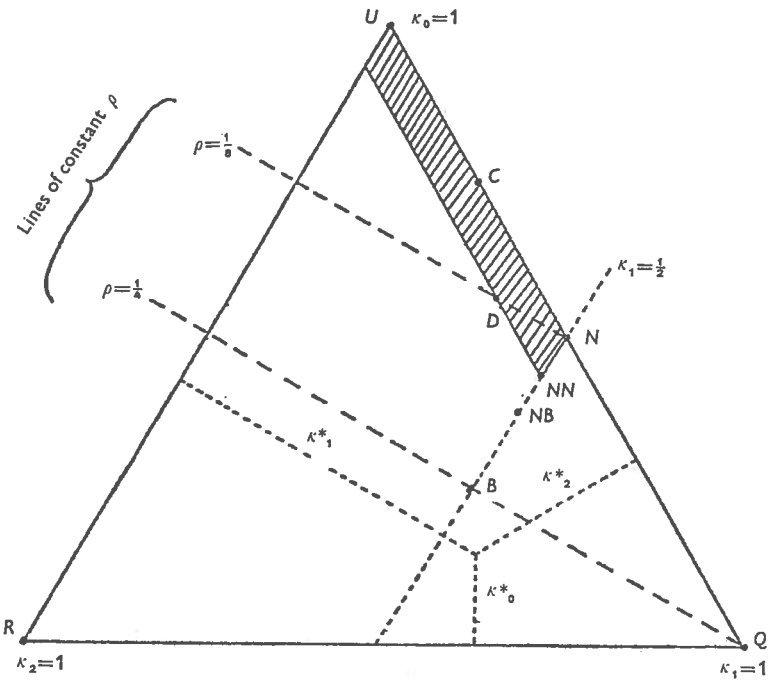

Fig. 1. The space K of pairwise genetic relationships, showing the location of the genealogical relationships of Table 2. $K = \{\boldsymbol{\kappa}; \kappa_0+\kappa_1+\kappa_2 = 1, \kappa_i \geqslant 0\}$. Any point $\boldsymbol{\kappa}^*$ of K may be represented by a point in an equilaterial triangle as shown. The shaded area indicates the region in which the $\boldsymbol{\kappa}$ values for other genealogical relationships are located. The lines of constant kinship (Malécot, 1948) are also shown; $\rho = (\kappa_1+2\kappa_2)/4$.

ships have $\kappa_2 \geqslant \frac{1}{16}$ or $\kappa_2 \geqslant \frac{1}{2}$. Thus $\boldsymbol{\kappa}$ values corresponding to genealogical relationships are discrete points concentrated mainly in a small region of the space

$$K = \{\boldsymbol{\kappa}, \kappa_0+\kappa_1+\kappa_2 = 1, \kappa_i \geqslant 0\} \quad \text{(see figure 1)}.$$

For convenience, however, $S(\boldsymbol{\kappa})$ may be considered as a continuous function over the whole triangle; if common ancestors of A and B are inbred the relationship may still be specified by some $\boldsymbol{\kappa}$ in K, but the $\boldsymbol{\kappa}$ values corresponding to the usual relationships will be slightly modified, and in this case many more $\boldsymbol{\kappa}$ values are theoretically possible. With currently available data such relationships cannot usually be distinguished from the same relationship between A and B in the absence of inbreeding of their common ancestor, but we shall see that with sufficient data any $\boldsymbol{\kappa}$ must be accurately determined.

For a single locus $L(\boldsymbol{\kappa})$ is linear in $\boldsymbol{\kappa}$ and the maximum occurs

at U if (G_1, G_2) is of type 3, 5, or 7 (Table 1); A and B have no alleles in common;
or of type 2 with $p_i > \frac{1}{2}$, or of type 6 with $p_i > \frac{1}{4}$,

at Q if (G_1, G_2) is of type 2 with $p_i < \frac{1}{2}$, or of type 6 with $p_i < \frac{1}{4}$,

and at R (or at B if R is excluded from consideration),

if (G_1, G_2) is of type 1 or 4; A and B have identical genotypes.

These results correspond with intuitive estimates of the relative merits of U, Q and B in that sibs are the individuals who most frequently have identical genotypes, and that rare alleles are more likely to be identical by descent.

Now, over s loci the likelihood is the product of linear functions with positive coefficients, and if $x \geqslant 0$, $y \geqslant 0$ and $0 \leqslant \lambda \leqslant 1$ then

$$\lambda x + \{1-\lambda\} y \geqslant x^{\lambda} y^{[1-\lambda]}.$$

Hence
$$L_s(\lambda \boldsymbol{\varkappa}^{(1)} + \{1-\lambda\}\boldsymbol{\varkappa}^{(2)}) \geqslant \{L_s(\boldsymbol{\varkappa}^{(1)})\}^{\lambda} \{L_s(\boldsymbol{\varkappa}^{(2)})\}^{(1-\lambda)}, \tag{5}$$

for any $\boldsymbol{\varkappa}^{(1)}$ and $\boldsymbol{\varkappa}^{(2)}$ in K, and $S_s(\boldsymbol{\varkappa}) = \log_e(L_s(\boldsymbol{\varkappa}))$ is therefore concave, $(s \geqslant 1)$. Further, K is a closed convex subspace of Euclidean three-dimensional space, and provided $S_s(\boldsymbol{\varkappa})$ is strictly concave it has a unique local maximum in K, possibly on the boundary. For a given pair of individuals the maximum likelihood estimate of $\boldsymbol{\varkappa}(A, B)$ may be found by some standard 'hill-climbing' technique.

However, few points in K correspond to genealogical relationships, and in many cases it is more relevant simply to evaluate $S_s(\boldsymbol{\varkappa})$ at the points of interest. To evaluate the support efficiently it is necessary to rapidly identify the type of genotype pair. A computer routine LHOOD accepts the four alleles of the two individuals at any given locus, and their population frequencies. Binary tests of alikeness of alleles are then made between relevant allele pairs, and result in a binary number between 0 and 15 specifying the type of genotype pair. $P_i(G_1, G_2)$ $(i = 0, 1, 2)$ are then rapidly computed by function routines indexed by the correct binary number.

In situations of dominance only the phenotypes $\phi(A)$ and $\phi(B)$ of A and B are observable. Suppose that at locus j $\phi(A) = \phi_1^{(j)}$ and $\phi(B) = \phi_2^{(j)}$, and let $H_i^{(j)}$ denote the set of genotypes corresponding to $\phi_i^{(j)}$ $(i = 1, 2; j = 1, \ldots, s)$. Then

$$\begin{aligned} S_s(\boldsymbol{\varkappa}) &= \sum_{j=1}^{s} \log_e P^{(j)}(\phi(A) = \phi_1^{(j)},\ \phi(B) = \phi_2^{(j)} | \boldsymbol{\varkappa}(A, B) = \boldsymbol{\varkappa}) \\ &= \sum_{j=1}^{s} \log_e \left(\sum_{i=0}^{2} \kappa_i P^{(j)}(\phi(A) = \phi_1^{(j)},\ \phi(B) = \phi_2^{(j)} | r(A, B) = i) \right) \\ &= \sum_{j=1}^{s} \log_e \left(\sum_{i=0}^{2} \kappa_i P_i^{(j)}(\phi_1^{(j)}, \phi_2^{(j)}) \right) \quad \text{(say)}, \end{aligned} \tag{4'}$$

where
$$P_i^{(j)}(\phi_1^{(j)}, \phi_2^{(j)}) = \sum_{G_1^{(j)}} \ldots \sum_{\in H_i^{(j)}} P_i^{(j)}(G_1^{(j)}, G_2^{(j)}). \tag{6}$$

Thus apart from computational problems in the determination and specification of $H_i^{(j)}$ for different dominance systems, there is no difficulty in extending the pairwise estimation of relationship to situations of dominance. As before the support function $S_s(\boldsymbol{\varkappa})$ is concave.

III. THE THEORY OF THE ESTIMATION OF PAIRWISE RELATIONSHIP

Only relationships with different $\boldsymbol{\varkappa}$ values can be distinguished on the basis of genotype data alone; in this section we consider first whether different $\boldsymbol{\varkappa}$ values are necessarily distinguishable.

We shall say that '$\boldsymbol{\varkappa}$ is *identifiable*' with respect to a given locus if

$$P(\phi_1, \phi_2 | \boldsymbol{\varkappa}^{(1)}) = P(\phi_1, \phi_2 | \boldsymbol{\varkappa}^{(2)})$$

for *all* ordered phenotype pairs (ϕ_1, ϕ_2) implies that $\boldsymbol{\varkappa}^{(1)} = \boldsymbol{\varkappa}^{(2)}$. (7)

We see that $\boldsymbol{\varkappa}$ is not identifiable for loci for which there is only one phenotype. In this case all individuals are alike and clearly pairwise relationships cannot be determined; $L(\boldsymbol{\varkappa}) = 1$ for all $\boldsymbol{\varkappa}$. Conversely we shall show that if $\boldsymbol{\varkappa}$ is identifiable with respect to a given locus we can determine $\boldsymbol{\varkappa}$ from the phenotypic characteristics of two individuals, provided we have genetic data for

a sufficient number of such loci. Identifiability is equivalent to strict concavity of the support function over a sufficient number of loci; the support over several loci is strictly concave unless the same lines of constant support obtain for all loci.

Now, at a single locus

$$L(\boldsymbol{\kappa}) = \sum_{i=0}^{2} \kappa_i P_i(\phi_1, \phi_2)$$
$$= \kappa_0(P_0(\phi_1,\phi_2) - P_2(\phi_1,\phi_2)) + \kappa_1(P_1(\phi_1,\phi_2) - P_2(\phi_1,\phi_2)) + P_2(\phi_1,\phi_2)$$
$$\text{since} \quad \kappa_0+\kappa_1+\kappa_2 = 1.$$

Thus $P(\phi_1,\phi_2|\boldsymbol{\kappa}^{(1)}) = P(\phi_1,\phi_2|\boldsymbol{\kappa}^{(2)})$ for all (ϕ_1,ϕ_2) if and only if either $\boldsymbol{\kappa}^{(1)} = \boldsymbol{\kappa}^{(2)}$ *or*

$$F(\phi_1,\phi_2) \text{ is constant over all ordered pairs } (\phi_1,\phi_2), \tag{8}$$

where
$$F(\phi_1,\phi_2) = [P_1(\phi_1,\phi_2) - P_2(\phi_1,\phi_2)]/[P_0(\phi_1,\phi_2) - P_2(\phi_1,\phi_2)]. \tag{9}$$

If $F(\phi_1,\phi_2) = f$ for all (ϕ_1,ϕ_2), $L(\boldsymbol{\kappa})$ is necessarily a function only of $\kappa_0 + f\kappa_1$. That equation (8) should *not* hold is a necessary and sufficient condition for identifiability. (F is well defined unless there is only one phenotype).

There are many different possible dominance systems; a complete enumeration is given by Bennett (1957). However, we shall restrict attention to systems in which different homozygotes have different phenotypes, and a heterozygote has a phenotype either uniquely distinguishable or identical to that of either one of the two relevant homozygotes. The human blood-group loci, and many others, fall within this class. Consider a locus with t alleles.

Case 1. If there exist (i) an allele a_1 not dominant to any other, and (ii) an allele a_2 such that a_1 and a_2 are co-dominant, then $\boldsymbol{\kappa}$ is identifiable.

Proof. Let $\phi(G)$ denote the phenotype of genotype G, and $\phi_1 = \phi(a_1a_1)$, $\phi_2 = \phi(a_2a_2)$. Then $H_1 = \{a_1a_1\}$ and H_2 contains only genotypes without a_1 genes, where, as before, H_i is the set of genotypes G with $\phi(G) = \phi_i$. From Table 1 and equation (9):

$$F(\phi_1,\phi_1) = (p_1^3 - p_1^2)/(p_1^4 - p_1^2) = 1/(1+p_1) \quad \text{and} \quad F(\phi_1,\phi_2) = 0 + F(\phi,\phi_1).$$

Case 2. If $\boldsymbol{\kappa}$ is identifiable over any subset of a genetic system it is identifiable overall. If condition (i) above does not hold there must be some subset of cyclic dominance (a_1 dominant to a_2 dominant to ... dominant to a_1). Systems of cyclic dominance do not occur amongst the human blood-group loci, but for completeness it may be shown that $\boldsymbol{\kappa}$ is *not* identifiable only in the special case $t = 3$ and $p_1 = p_2 = p_3 = \frac{1}{3}$. In this case $F(\phi_1,\phi_2) = \frac{2}{3}$ for all (ϕ_1,ϕ_2) and $L(\boldsymbol{\kappa})$ is necessarily constant along lines of constant

$$3\kappa_0 + 2\kappa_1 \quad \text{or} \quad \kappa_0 - 2\kappa_2.$$

Case 3. Finally if condition (i) holds but (ii) does not, a_1 is recessive to all other a_j ($j = 2, \ldots, t$). Unless there is some allele a_2 not dominant to any allele other than a_1, we again have some subsystem of cyclic dominance.

If a_1 is recessive to all other alleles and a_2 is dominant only to a_1, $\boldsymbol{\kappa}$ is not identifiable if and only if $p_1 + p_2 = 1 (t = 2)$.

Proof. Let $\phi_1 = \phi(a_1a_1)$; $H_1 = \{a_1a_1\}$. Let $\phi_2 = \phi(a_2a_2)$; $H_2 = \{a_1a_2, a_2a_2\}$. Then

$$F(\phi_1,\phi_1) = 1/(1+p_1)$$

as before, and

$$F(\phi_1,\phi_2) = [p_1^2p_2 - 0]/[p_1^2(p_2^2 + 2p_1p_2) - 0] = 1/(p_2 + 2p_1)$$

(see Table 1 and equations (6) and (9)). These are equal if and only if $(p_1+p_2)=1$; that is, if and only if $t=2$. In this case

$$F(\phi_2,\phi_2) = [(p_2^3+2p_2^2p_1+p_1p_2(p_1+p_2))-(p_2^2+2p_1p_2)]/[(p_2^2+2p_1p_2)^2-(p_2^2+2p_1p_2)]$$
$$= 1/(1+p_1).$$

Thus $F(\phi_2,\phi_2)=F(\phi_1,\phi_2)=F(\phi_2,\phi_1)=F(\phi_1,\phi_1)$ and $\boldsymbol{\kappa}$ is not identifiable, since these are the only possible ordered phenotype pairs. $\boldsymbol{\kappa}$ values having equal $\kappa_0+\kappa_1/(1+p_1)$, or equivalently equal $p_1\kappa_0-\kappa_2$, (10)

have equal support, whatever the phenotypes of the individuals.

Combining the above results we have that

the only general exception to identifiability of $\boldsymbol{\kappa}$ *at a single locus within the class of dominance systems considered is two-allele simple dominance* (an 'SD' locus).

The problems arising in the use of SD loci are theoretical rather than practical. The lines of non-identifiability (10) depend on the frequency p_1 of the recessive allele. Over a pair of such loci with different p_1 values we do have identifiability of $\boldsymbol{\kappa}$. Further for loci of equal p_1 $(0<p_1<1)$ the net support is strictly concave in the direction orthogonal to the lines given by (10). We can thus, with data from a sufficient number of such loci, accurately estimate $p_1\kappa_0-\kappa_2$, and this may be sufficient to determine the correct relationship amongst a discrete set of alternatives.

We shall say that a locus with respect to which $\boldsymbol{\kappa}$ is identifiable is

informative, (11)

and shall now show that identifiability is equivalent to strict concavity of the mean support surface, and asymptotic consistency of the maximum likelihood estimate of $\boldsymbol{\kappa}$. Since we do not have identical gene loci, consistency is not immediate; the usual arguments must be modified. Let $\boldsymbol{\kappa}^*$ be the true value of $\boldsymbol{\kappa}(A,B)$ and consider any fixed $\boldsymbol{\kappa}\neq\boldsymbol{\kappa}^*$. As before, let $S_s(\boldsymbol{\kappa})$ be the support for $\boldsymbol{\kappa}$ over s loci, and $\phi_1^{(j)}$, $\phi_2^{(j)}$ be the phenotypes of A and B respectively at locus j. Then

$$S_s(\boldsymbol{\kappa})-S_s(\boldsymbol{\kappa}^*)=\sum_{j=1}^{s}Z_j(\boldsymbol{\kappa},\boldsymbol{\kappa}^*), \tag{12}$$

where
$$Z_j(\boldsymbol{\kappa},\boldsymbol{\kappa}^*)=\log_e(P^{(j)}(\phi_1^{(j)},\phi_2^{(j)}|\boldsymbol{\kappa}(A,B)=\boldsymbol{\kappa})/P^{(j)}(\phi_1^{(j)},\phi_2^{(j)}|\boldsymbol{\kappa}(A,B)=\boldsymbol{\kappa}^*)). \tag{13}$$

The random variables Z_j are independent but not identically distributed. Any given Z_j is distributed over a finite set of points, but the set is not uniformly bounded above and may, if $\kappa_0=0$ and $\kappa_0^*\neq 0$, include $-\infty$. Since the divisor in (13) is the true probability of the data at locus j, the expectation

$$E(Z_j(\boldsymbol{\kappa},\boldsymbol{\kappa}^*)|\boldsymbol{\kappa}(A,B)=\boldsymbol{\kappa}^*)\leqslant 0, \tag{14}$$

with equality if and only if

$$P^{(j)}(\phi_1^{(j)},\phi_2^{(j)}|\boldsymbol{\kappa}(A,B)=\boldsymbol{\kappa})=P^{(j)}(\phi_1^{(j)},\phi_2^{(j)}|\boldsymbol{\kappa}(A,B)=\boldsymbol{\kappa}^*)$$

for all $(\phi_1^{(j)},\phi_2^{(j)})$; in this case Z_j is zero with probability 1; $P(Z_j(\boldsymbol{\kappa},\boldsymbol{\kappa}^*)=0)=1$.

A locus is not informative if and only if there exist some $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}^*$ with $\Pr[Z_j(\boldsymbol{\kappa},\boldsymbol{\kappa}^*)=0]=1$. However, even for informative loci there is no strictly negative upper bound to the mean (14). We *define* a locus to be

'(ϵ,δ)-*informative between* $\boldsymbol{\kappa}$ *and* $\boldsymbol{\kappa}^*$' if $\Pr[|Z_j(\boldsymbol{\kappa},\boldsymbol{\kappa}^*)|>\epsilon]>\delta$. (15)

Any informative locus is (ϵ,δ)-informative between $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}^*$ for some strictly positive ϵ and δ (dependent upon $\boldsymbol{\kappa}$, $\boldsymbol{\kappa}^*$ and the locus in question).

Now, suppose that all but a fixed number of loci are (ϵ, δ)-informative for some fixed, arbitrarily small, but strictly positive ϵ and δ. In this case we may consider random variables

$$Y_j(\boldsymbol{\kappa}, \boldsymbol{\kappa}^*) = \max(-2, Z_j(\boldsymbol{\kappa}, \boldsymbol{\kappa}^*)).$$

Then $Z_j \leqslant Y_j$, and the Y_j have a uniformly bounded variance, and a mean bounded above by $-\mu$, for some strictly positive μ dependent upon ϵ, δ, $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}^*$ (Thompson, 1974*b*). Then by the Strong Law of large numbers for non-identically distributed random variables (see, for example, Feller (1968), p. 259); for any chosen (positive) C,

$$\Pr\left[\sum_{j=1}^{s} Y_j(\boldsymbol{\kappa}, \boldsymbol{\kappa}^*) < -C \quad \text{for all} \quad s > s_0\right] \to 1 \quad \text{as} \quad s_0 \to \infty.$$

Thus, since $Y_j \geqslant Z_j$,

$$\Pr\left[\sum_{j=1}^{s} Z_j(\boldsymbol{\kappa}, \boldsymbol{\kappa}^*) < -C \text{ for all } s > s_0\right] \to 1 \quad \text{as} \quad s_0 \to \infty,$$

or

$$\Pr[S_s(\boldsymbol{\kappa}) - S_s(\boldsymbol{\kappa}^*) < -C \text{ for all } s > s_0] \to 1 \quad \text{as} \quad s_0 \to \infty, \tag{16}$$

and we have (strong) consistency in the estimation of $\boldsymbol{\kappa}$ over a large number of gene loci.

This theorem is not strictly applicable. We do not in fact have an indefinite number of loci, and those which we have are not independent since they lie in linkage groups; the information in the genome is finite. Hence (16) does not guarantee that any relationship can ever be determined with absolute reliability. However, besides showing that with unlimited information relationships must be correctly determined, consistency also provides that even with finite data the correct relationship has on average maximal support. As we shall see below, in practice the information available is amply sufficient to ensure that a correct estimate of any near relationship could be obtained from a reasonable number of markers. The condition used in the proof of (16) of a uniform lower bound on the information provided by each locus has some genetic significance. If loci become progressively less informative (for example, progressively more monomorphic), the total information available may be bounded by some finite limit regardless of the number of loci.

IV. THE PRACTICE OF THE ESTIMATION OF PAIRWISE RELATIONSHIP

Although we have consistency we have still to consider the reliability of inferences made in any given situation. To consider the practical possibility of estimating $\boldsymbol{\kappa}(A, B)$ with currently available data we compute the mean increase in support difference

$$E_{\boldsymbol{\kappa}^*}(Z_j(\boldsymbol{\kappa}, \boldsymbol{\kappa}^*)), \tag{17}$$

due to a t-allele locus with allele frequencies $p_1, \ldots, p_t$ when the true relationship is $\boldsymbol{\kappa}^*$. (In this and the following sections a subscript on expectation (E) or probability (P) will be used to denote the true relationship, and the code letters for the relationships of Table 2 will be used interchangeably with their $\boldsymbol{\kappa}$ values.) The expression (17) measures the information contained (on average) in the locus, with regard to distinguishing relationships $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}^*$. In the absence of dominance, loci with equifrequent alleles provide on average the most information. Loci with rare alleles can occasionally be very informative; for example, on those rare occasions when two individuals have this rare allele.

The mean support difference provided by a locus with t equifrequent alleles and no dominance may be readily computed; some results for relationships B, Q and U are given in Table 3. It may be seen that individual loci usually provide little information; many loci are required to determine a relationship with any reliability. Mean support differences are additive over unlinked loci, but

Table 3. *Mean support differences between the given relationships for a locus with t equifrequent alleles and no dominance, when the true relationship is as shown (mean support differences are additive over unlinked loci)*

Number of alleles (t)	Mean support differences					
	$E_U(S(U)-S(Q))$	$E_U(S(U)-S(B))$	$E_Q(S(Q)-S(U))$	$E_Q(S(Q)-S(B))$	$E_B(S(B)-S(U))$	$E_B(S(B)-S(Q))$
2	∞	0·16	0·20	0·06	0·15	∞
3	∞	0·28	0·33	0·11	0·27	∞
4	∞	0·38	0·48	0·16	0·40	∞
5	∞	0·47	0·61	0·22	0·50	∞
8	∞	0·63	0·94	0·32	0·76	∞
10	∞	0·70	1·14	0·36	0·90	∞
15	∞	0·84	1·46	0·44	1·18	∞
20	∞	0·92	1·71	0·49	1·40	∞
Large t	∞	$2\log_e 2$	$(\log_e t - 2\log_e 2)$	$\log_e 2$	$(\log_e t - (\frac{5}{2})\log_e 2)$	∞

the total mean support difference is not a complete measure of the reliability of inferences. In estimating $\varkappa(A,B)$ it is more important to have data for many loci than that each locus should provide large support differences; such a locus enables $r(A,B)$ for that given locus to be determined with greater certainty, but $\kappa_i(A,B)$ is the proportion of loci with $r(A,B)=i$.

For any given $\varkappa^*$ not equal to Q (or R) $P_{\varkappa^*}(S(Q)=-\infty)>0$, and thus mean support differences are infinite (Table 3). In this case it is more relevant to consider

$$P_{\varkappa^*}(S(Q)=-\infty)=\kappa_0^* P_U(S(Q)=-\infty). \tag{18}$$

$P_U(S(Q)=-\infty)$ is the probability that a parent–offspring relationship between unrelated individuals is positively excluded. For relationships $\varkappa^*$ other than U, Q, B and R the mean support differences $E_{\varkappa^*}(S(\varkappa^*)-S(U))$ and $E_{\varkappa^*}(S(\varkappa^*)-S(B))$ are usually very small; such relationships cannot be readily distinguished from U and/or B relationships.

To obtain practical results, sample genotypes of individuals in a large genealogy were obtained by repeated simulation of gene flow. Data for 20 loci each with between 2 and 5 alleles with varying alleles frequencies were obtained, and a large number of pairs of individuals in relationships U, Q, B, R, N, C and D were considered. Even with this amount of genetic data, which is as much as we could normally have in practice, it is found that true relationships N, C and D are often incorrectly estimated. Even though for unrelated pairs relationship Q is excluded in over 96 % of cases, averaging over true relationships B, N, C, and D, Q is excluded in only 75 % of cases. Particularly in the case of no dominance, Q, if not excluded, often has maximal support (Thompson, 1974*b*). Estimation, on the basis of genetic information alone (excluding information of age, etc.) of relationships U, Q and B has 85 % accuracy, or 92 % if only these three relationships are considered. Even where the true relationship has maximal support (amongst the possibilities U, Q, B, R, C, N and D) there may often be others within two units of support.

We thus have good determination of U, Q, B and R, but these are the only relationships which can be reliably distinguished with currently available data. However, these are the only relationships that it is normally necessary to consider in inferring a genealogy (Thompson, 1974*b*). For these four true relationships typical examples of the shape of the complete support surface are given in Fig. 2, again using genetic data for pairs of individuals for the above 20 loci. In each case

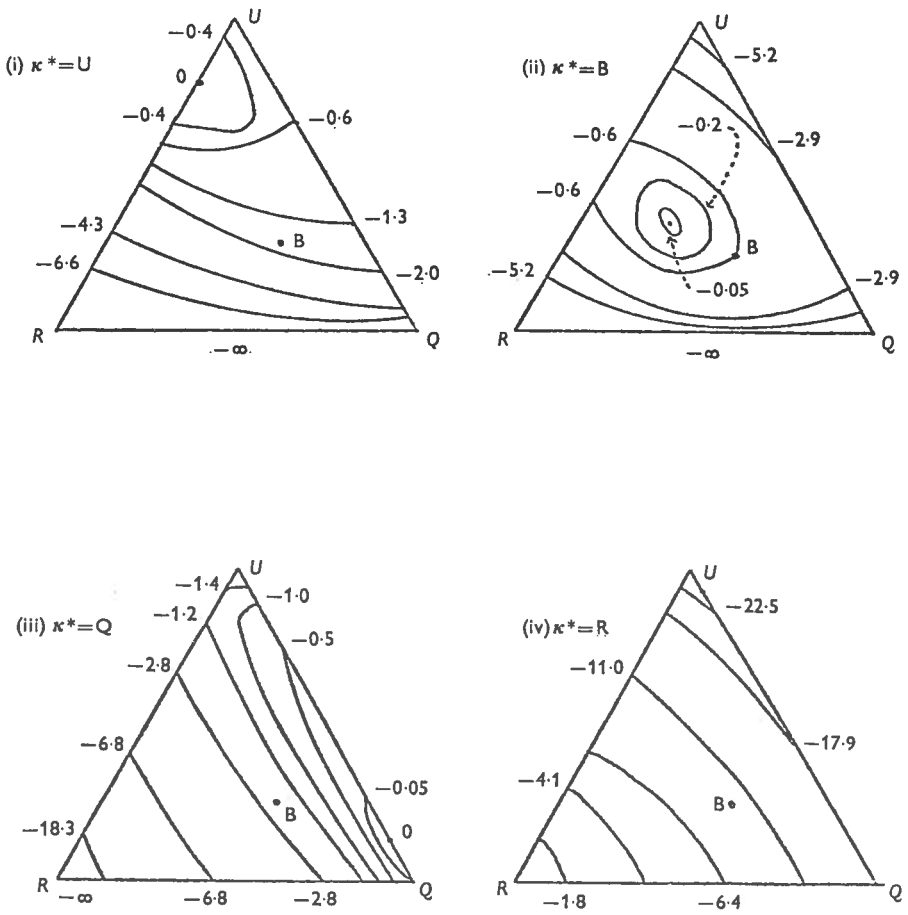

Fig. 2. The forms of the support surface for pairwise relationship for four different cases of true relationship. The support values are given relative to the maximum in each case. (i) True relationship, U; $\kappa^* = (1, 0, 0)$. Maximizing relationship $\hat{\kappa} = (0{\cdot}8, 0{\cdot}0, 0{\cdot}2)$, and $S(\hat{\kappa}) - S(\kappa^*) = 0{\cdot}53$. (ii) True relationship, B; $\kappa^* = (0{\cdot}25, 0{\cdot}5, 0{\cdot}25)$. Maximizing relationship $\hat{\kappa} = (0{\cdot}35, 0{\cdot}25, 0{\cdot}4)$ and $S(\hat{\kappa}) - S(\kappa^*) = 0{\cdot}6$. (iii) True relationship, Q; $\kappa^* = (0, 1, 0)$. Maximizing relationship $\hat{\kappa} = (0{\cdot}15, 0{\cdot}85, 0{\cdot}0)$ and $S(\hat{\kappa}) - S(\kappa^*) = 0{\cdot}05$. (iv) True relationship, R; $\kappa^* = (0, 0, 1)$. $\hat{\kappa} = R = (0, 0, 1)$.

the true relationship κ^* has higher support than any of the other three, the support surface is unimodal, and the maximizing point $\hat{\kappa}$ is nearer the true relationship than to any other, and κ^* has support very little less than $\hat{\kappa}$. However, usually only when $\kappa^* = R$ does $\hat{\kappa} = \kappa^*$, and even with 20 loci the support surface may vary little over the whole space K. Relationship B almost always gives a support surface with internal maximum, while U and Q normally give maxima on the boundary in the neighbourhood of the correct vertex of K.

The considerations of this section may be extended to cases in which dominance occurs. A locus exhibiting dominance necessarily provides less information than if the genotypes were distinguishable, but there are no essential differences in the computation of mean support differences (17) or exclusion probabilities (18). Again relationships U, Q, B and R are relatively readily distinguished, but other support differences are small.

Table 4. *Phenotypes of the six Tristan da Cunha individuals of Fig. 3 at six blood-group loci*

	A_1A_2BO	Rhesus	MNS	P(P_1)	Duffy (Fy^a)	Kell (K)
I.1	A_2B	R_1r	Ms	+	−	−
I.2	A_2B	R_1r	MNS	+	+	+
I.3	A_1	R_0r	MNS	+	+	−
II.1	B	rr	MNs	+	−	−
II.2	A_2	rr	Ms	+	−	−
II.3	B	R_1r	NS	−	+	−

Table 5. *Tristan da Cunha population allele frequencies*

Locus	Alleles	Frequencies (respectively)
ABO	A_1, A_2, B, O	0·18, 0·07, 0·10, 0·65
Rhesus	R_0, R_1, R_2, R′, r	0·03, 0·40, 0·17, 0·03, 0·37
MNS	MS, Ms, NS, Ns	0·18, 0·40, 0·05, 0·37
P	P_1, P_2	0·57, 0·43
Duffy	Fy^a, Fy^b	0·35, 0·65
Kell	K, k	0·05, 0·95

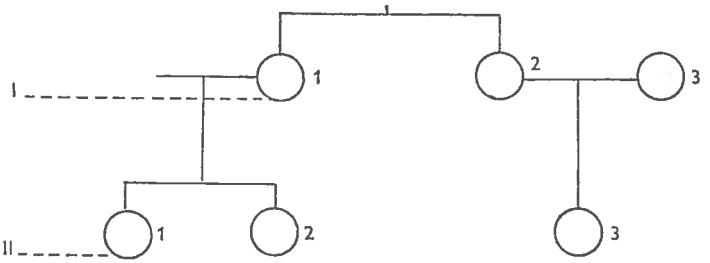

Fig. 3. Genealogy of six Tristan da Cunha individuals.

V. EXAMPLE

As an example consider six individuals from the Tristan da Cunha population, whose correct genealogy is as shown in Fig. 3 and whose phenotypes at six loci are given in Table 4. The approximate Tristan da Cunha population allele frequencies for these loci are given in Table 5.* Since we have a situation with dominance, inferences on the pairwise relationships may not be an accurate reflexion of the estimates that would be obtained by considering all the individuals jointly, but pairwise estimation is a useful preliminary in any genealogy reconstruction. Note that the ABO phenotypes alone indicate that we have a family; in fact, all the comparatively rare A_2 and B genes result from single genes in the parents of I. 1 and I. 2. All individuals differ in some way, and no other conclusions are immediate from Table 4.

Supports for relationships are computed using equations (4′), (6) and Table 1. For example, consider individuals I. 1 and I. 2 at the Rhesus locus. The genotypes are known and identical ($G_1 = G_2 = R_1r$), and the relevant allele frequencies are 0.40 and 0.37. From Table 1 we have directly that (G_1, G_2) is of type 4 and hence

$$P_0(G_1, G_2) = (2 \times 0.40 \times 0{\cdot}37)^2, \quad P_1(G_1, G_2) = 0{\cdot}40 \times 0{\cdot}37 \times (0{\cdot}40 + 0{\cdot}37)$$

and

$$P_2(G_1, G_2) = 2 \times 0{\cdot}40 \times 0{\cdot}37.$$

* See acknowledgements.

For individuals I. 3 and II. 1 at the ABO locus we have the complication of dominance:

$$\phi_1 = A_1; \quad G_1 = A_1A_1, \quad A_1A_2 \quad \text{or} \quad A_1O, \quad \text{and} \quad \phi_2 = B; \quad G_2 = BB \quad \text{or} \quad BO.$$

$P_0(\phi_1, \phi_2)$ is simply the product of the two phenotype probabilities:

$$((0{\cdot}18)^2 + (2 \times 0{\cdot}18 \times 0{\cdot}07) + (2 \times 0{\cdot}18 \times 0{\cdot}65))\,((0{\cdot}10)^2 + (2 \times 0{\cdot}10 \times 0{\cdot}65)).$$

If the individuals have a gene in common only $G_1 = A_1O$ and $G_2 = BO$ are possible; all other combinations have zero probability. (G_1, G_2) is of type 6 and from Table 1,

$$P_1(\phi_1, \phi_2) = P_1(A_1O, BO) = (0{\cdot}18 \times 0{\cdot}10 \times 0{\cdot}65).$$

Finally the two individuals cannot have two genes in common at the ABO locus; $P_2(\phi_1, \phi_2) = 0$.

The complete results of these computations are given in Table 6; in each case the support relative to that for the true relationship is given. A positive value thus indicates that the maximum likelihood relationship is not the true one. In fact, maximum-likelihood estimates are surprisingly accurate considering the available information; all those not involving individual II. 3 are correct. It is seen that the support values can reveal relationships that are far from immediate from the initial data. The success of Table 6 is, however, over-emphasized by the fact that the true relationships are already known. The figures of Table 6 alone allow few firm conclusions to be drawn. In no case do *all* other relationships have even one unit of support less than the true one, and many of the support differences are virtually negligible. The relationships $\varkappa$(I. 1, I. 3) = U, $\varkappa$(I. 1, II. 1) = $\varkappa$(I. 1, II. 2) = Q, and perhaps also $\varkappa$(I. 3, II. 1) = $\varkappa$(I. 3, II. 2) = U and $\varkappa$(I. 2, II. 3) = Q, are the only inferences that can be made with any confidence at all. However, that even this is obtainable from data from six loci, all exhibiting dominance and three having only two alleles, is an indication that these methods may have useful practical applications.

VI. LINKAGE, GENE FREQUENCIES AND INBREEDING

In this section we briefly consider some miscellaneous aspects of the problem. First, we have restricted attention to unlinked autosomal loci. Clearly X-linked loci could also be used to estimate relationship; the single-locus pairwise genotype distributions are given by Li & Sacks (1954). However, their use necessitates a distinction between maternal and paternal relatives and corresponding modification of the $\varkappa$-coefficients. Equally there is no fundamental restriction against the use of linked loci; in this case the genotypes at the linked loci must be considered jointly. Provided the recombination fraction is known, two linked loci jointly provide more information than either separately, but less than were they independent. However, in addition to the increased complexity of the likelihood function, recombination fractions may not be accurately known, and the use of linked loci is best avoided where possible.

We have assumed the absence of mutation and of errors of determination of phenotype; these phenomena are only likely to cause problems when estimation amongst a large number of individuals is considered. Population allele frequencies have been assumed known; in practice they are estimated from samples taken from the population of which we consider both the individuals to be representative members. Formulae for the bias in the $\varkappa$ value maximizing the mean support function for given true and assumed population allele frequencies are given by Thompson (1974*b*). Since the different loci will provide biases in different directions, the net effect over several loci is likely to be small.

In practice estimation between the four relationships U, Q, B and R is insensitive to changes in gene frequency. Using the same simulated data as above an 'error' of gene frequency was intro-

Table 6. *Support differences between the given relationship and the true relationship between each pair of individuals*

(A positive value thus indicates that a given relationship has greater likelihood than the true relationship. **** denotes the true relationship.)

	Relationship (and $\varkappa$ value)					
Pair of individuals	U (1, 0, 0)	Q (0, 1, 0)	B ($\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$)	N ($\frac{1}{2}$, $\frac{1}{2}$, 0)	C ($\frac{3}{4}$, $\frac{1}{4}$, 0)	D ($\frac{9}{16}$, $\frac{6}{16}$, $\frac{1}{16}$)
I. 1, I. 2	−1·46	−1·82	****	−0·97	−0·98	−0·13
I. 1, I. 3	****	−3·26	−2·84	−1·21	−0·54	−1·16
I. 1, II. 1	−2·07	****	−1·19	−0·87	−1·41	−1·20
I. 1, II. 2	−3·08	****	−0·67	−1·20	−2·00	−1·38
I. 1, II. 3	+0·29	−∞	−1·12	****	+0·26	+0·00
I. 2, I. 3	****	−1·35	−0·85	−0·40	−0·15	−0·26
I. 2, II. 1	−0·09	−0·40	−1·48	****	+0·02	−0·29
I. 2, II. 2	−0·10	−0·92	−1·58	****	−0·06	−0·25
I. 2, II. 3	−1·20	****	−0·84	−0·34	−0·68	−0·56
I. 3, II. 1	****	−1·83	−2·06	−0·65	−0·29	−0·74
I. 3, II. 2	****	−1·89	−2·07	−0·67	−0·29	−0·75
I. 3, II. 3	+0·99	****	−0·42	+0·78	+0·94	+0·64
II. 1, II. 2	−0·84	−0·16	****	−0·22	−0·46	−0·16
II. 1, II. 3	−0·25	−0·07	−0·67	+0·08	****	−0·03
II. 2, II. 3	+0·55	−∞	−2·86	−0·73	****	−0·73

duced at each of the 20 loci. At every locus the change for some allele was 0·1, some errors thus being 100 %. The likelihoods given by these modified frequencies were significantly less than when the true simulation frequencies were used, but support differences were very little affected. In no case was a maximizing relationship changed from its true value, and in only a few cases, all of true relationship U, was the support difference substantially reduced. The maximizing points $\hat{\varkappa}$ and the shape of the support contours were also virtually unchanged.

Even relative to a recent gene pool there is a possibility that one or both of the individuals may be inbred. The genotype distribution then depends on the probabilities of the nine genetically distinct identity states given by Jacquard (1972). Although the problems of identifiability and consistency become more complex, there is in theory no difficulty in extending the estimation to these eight identity coefficients (nine probabilities which sum to one). The situation is covered by the case of multiple relationships considered in section VII. In practice however there are normally insufficient data to estimate eight relationship coefficients; we therefore consider the effect of assumed absence of inbreeding. The support function is then maximized in a subspace K (see equation (1)) of the complete space of identity coefficients, in which only the coefficients $(\kappa_0, \kappa_1, \kappa_2)$ corresponding to non-inbred gene identity states are non-zero. The remaining six coefficients are assumed to be zero, but for most actually occurring genealogical relationships the total probability of these six states is small; the true relationship is located close to the restricted subspace. The fact that the full hypothesis set is not considered does not prevent a correct inference provided the true hypothesis falls within the class considered, and otherwise results in that inference within the restricted class that is closest to the true one in terms of the probability distributions involved.

For example, for sibs who are the offspring of first cousins, probably the most frequently occurring situation of interest, the identity states corresponding to $(\kappa_0, \kappa_1, \kappa_2)$ have probabilities

$(\frac{12}{64}, \frac{30}{64}, \frac{15}{64})$, the remaining six states having total probability $\frac{7}{64}$. Maximizing amongst alternatives in K we will usually find that the maximizing relationship is B ($\boldsymbol{\kappa} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$). Thus although the full true relationship cannot be inferred, the predominant relationship may be estimated.

VII. THE ESTIMATION OF JOINT RELATIONSHIPS

Thompson (1974*a*) has shown how the joint relationship W between n relatives may be specified by a distribution $\left\{q_i,\ 1 \leqslant i \leqslant D_n,\ q_i \geqslant 0,\ \sum_{i=1}^{D_n} q_i = 1\right\}$ over the D_n genetically distinct gene identity states $\{V_i,\ 1 \leqslant i \leqslant D_n\}$. The theory of the estimation of relationship extends, essentially without modification, to these general joint relationships. Again several different genealogical relationships may give the same distribution: only estimation of this distribution can be considered. Again we would normally in practice restrict attention to a few alternative hypotheses of interest, rather than attempt a general estimation.

Many of the general properties of Sections II and III remain true. In place of a two-dimensional subspace K of three-dimensional Euclidean space, we now have points in a $(D_n - 1)$-dimensional subspace of D_n-dimensional space. For a single locus the likelihood is a linear function, and hence constant over relationships in spaces of dimension $(D_n - 2)$. Over several loci it is the product of such linear functions and the support is concave, although over s loci the spaces of constant support will be of dimension at least $(D_n - s - 1)$. The space of maximal support (subject to $\Sigma_i q_i = 1, q_i \geqslant 0$) can consist of points with at most s non-zero q_i; but most true relationships have relatively few non-zero q_i.

A more general representation of identifiability is required in this multidimensional space. Consider first identifiability at a single locus j. Let $W = \Sigma_i q_i V_i$ and $W^* = \Sigma_i q_i^* V_i$, where W^* is the true joint relationship between the n individuals, having, at this locus, phenotypes $(\phi_1, \ldots, \phi_n)$. As in Section III,

$$E(Z_j(W, W^*)) = E(\log_e(P^{(j)}(\phi_1, \ldots, \phi_n | W)/P^{(j)}(\phi_1, \ldots, \phi_n | W^*))) \leqslant 0 \tag{19}$$

with strict inequality unless

$$P^{(j)}(\phi_1, \ldots, \phi_n | W) = P^{(j)}(\phi_1, \ldots, \phi_n | W^*) \quad \text{for all possible} \quad (\phi_1, \ldots, \phi_n),$$

or

$$\Sigma_i (q_i - q_i^*) P^{(j)}(\phi_1, \ldots, \phi_n | V_i) = 0 \quad \text{for all} \quad (\phi_1, \ldots, \phi_n). \tag{20}$$

Let $\mathbf{1}$ denote the D_n-dimensional vector of ones, and $\mathbf{P}^{(j)}(\phi_1, \ldots, \phi_n)$ the D_n-dimensional '$P^{(j)}$-*vector*':

$$(P^{(j)}(\phi_1, \ldots, \phi_n | V_i),\ 1 \leqslant i \leqslant D_n). \tag{21}$$

Let $d_i = (q_i - q_i^*)$ and $\mathbf{d} = (d_i,\ 1 \leqslant i \leqslant D_n)$; then (20) and $\Sigma_i q_i = \Sigma_i q_i^*$ may be written

$$\left.\begin{aligned} &\mathbf{d} \,.\, \mathbf{1} = 0 \\ \text{and} \quad &\mathbf{d} \,.\, \mathbf{P}^{(j)}(\phi_1, \ldots, \phi_n) = 0 \quad \text{for all} \quad (\phi_1, \ldots, \phi_n). \end{aligned}\right\} \tag{22}$$

For identifiability of relationship it is necessary and sufficient that (22) implies $\mathbf{d} = \mathbf{0}$, or that the vector $\mathbf{1}$ and the set of all possible $\mathbf{P}^{(j)}$-vectors together span D_n-dimensional space.

If there are h possible phenotypes there are at most h^n distinct $P^{(j)}$-vectors, and a necessary condition is

$$1 + h^n \geqslant D_n \quad \text{or} \quad h \geqslant (D_n - 1)^{1/n}. \tag{23}$$

The required value of h increases with n; for $n = 2, 3, 6$ we require $h \geqslant 3, 5$ and 8 respectively (see Thompson (1974*a*) for the relevant values of D_n). (23) is not a sufficient condition; for $n = 2$

and absence of inbreeding we found in Section III that the vector **1** and the three distinct phenotype-pair **P**-vectors given by an *SD* locus together span only a space of dimension two and so are insufficient to give identifiability of the three parameters $(\kappa_0, \kappa_1, \kappa_2)$.

Thus a single locus does not normally provide identifiability of relationship. For most loci there will be spaces over which the support is necessarily always constant (cf. equation (10)) and relationships in the same space cannot be distinguished by any number of loci of this type. However, over a set of loci of *different* types we may have identifiability, in which case the mean support surface is strictly concave, with maximum at the true relationship. Under a condition of a uniform lower bound to the information provided by each such set of loci, similar to that of equation (15), we will have consistency of the distribution $\{q_i\}$, or relationship $W = \Sigma_i q_i V_i$, over a large number of such sets of loci.

Thus information on the complete genome would result in a correct inference of the full joint relationship between a large number of individuals, but in practice we have the problem, not only of the unreliability of estimation as in Section IV, but of the fact that over the loci available some relationships may necessarily always have equal support, and thus be indistinguishable. In practice we find that specific alternative hypotheses of interest between a few individuals are seldom indistinguishable, and that inferences between a few such alternatives can usually be made, but support differences are small and estimates often inaccurate.

SUMMARY

Relationships between the individuals of a population have been previously studied from the point of view of prediction. Edwards (1967) suggested that the problem of detailed population structure could also be studied from the point of view of inference. Even where inferences of practical applicability cannot be made, such an approach can increase understanding of the relation between genealogical and genetic structure. In this paper we consider a specific problem which provides an introduction to the ideas and methods of genealogical inference. This is the problem of estimating the pairwise relationship between two individuals on the basis of their phenotypes at several loci. There is no theoretical problem in the extension from pairwise to joint relationships.

This research was supported by a research studentship from the Science Research Council (B/71/652), and subsequently by a Sims Scholarship from the University of Cambridge. I am also grateful for the hospitality of the Matematisk Institut, University of Aarhus, where some of this research was carried out, and to Dr A. W. F. Edwards for many helpful discussions. The Tristan da Cunha data of Section V are from a file of unpublished material compiled by the late Dr H. E. Lewis and other members of the Medical Research Council Tristan da Cunha Working Party.

REFERENCES

Bennett, J. H. (1957). The enumeration of genotype–phenotype correspondences. *Heredity* **11**, 403–409.

Cotterman, C. W. (1940). A calculus for statico-genetics. Ph.D. thesis, Ohio State University. Published in *Genetics and Social Structure* (ed. Paul Ballonoff). Benchmark Papers in Genetics; Dowden, Hutchinson and Ross (1974).

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper and Row.

Edwards, A. W. F. (1967). Automatic construction of genealogies from phenotypic information (Autokin). *Bulletin of the European Society for Human Genetics* **1**, 42–43.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, 3rd ed., vol. 1. New York: Wiley.

Jacquard, A. (1972). Genetic information given by a relative. *Biometrics*, **28**, 1101–1114

LI, C. C. & SACKS, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10**, 347–360

MALÉCOT, G. (1948). *Les Mathématiques de l'hérédité*. Paris: Masson et Cie.

THOMPSON, E. A. (1974*a*). Gene identities and multiple relationships. *Biometrics* **30**, 667–680.

THOMPSON, E. A. (1974*b*). Mathematical analysis of human evolution and population structure. Ph.D. thesis, University of Cambridge.

TURNBULL, C. M. (1972). Demography of small-scale societies. In *The Structure of Human Populations* (ed. G. A. Harrison and A. J. Boyce). Oxford: Clarendon Press.