Ann. Rev. Biophys. Bioeng. 1978. 7:253–86

STATISTICAL MODELING AND ANALYSIS IN HUMAN GENETICS

◆9113

R. C. Elston

Department of Biostatistics and the Genetics Curriculum, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514

D. C. Rao

Population Genetics Laboratory, University of Hawaii, Honolulu, Hawaii 96822

1 INTRODUCTION

Soon after the rediscovery of Mendel's work at the turn of the century, a rift developed between two opposing schools of thought. On the one hand were those who stressed the qualitative nature of genetic inheritance, as had been demonstrated by Mendel; on the other hand were the "biometricians," who noted that most human variation is quantitative, rather than qualitative, and therefore amenable to correlation and regression analysis. In 1918, Fisher (39) demonstrated that the rift was more apparent than real; by supposing that the genetic contribution to any quantitative trait is made up of the sum of many small independent additive effects, each controlled by a Mendelian factor, Fisher theoretically derived exactly what the biometricians were finding empirically. This unification of statistical and genetic findings led, in the half century to follow, to a period in which the development of statistical theory and knowledge in human genetics advanced hand in hand. Throughout this period, however, a practical rift remained in human genetics: statistical methods for the analysis of qualitative traits were largely developed in terms of one- and two-gene models, whereas those for the analysis of quantitative traits were for the most part based on the assumption of a model in which many genes act additively. As late as 1960 it was necessary for Edwards (24) to point out that consideration should be given to polygenic models for qualitative traits; and as late as 1970 it was suggested that to fit a one-gene model to a quantitative trait, the trait should first be changed to a simple dichotomy, with each individual classified according to whether his trait value is above or below some threshold value (105).

The past decade has seen a burst of activity in the development of statistical

0084–6589/78/0615–0253$1.00

models for human genetic analysis. This has resulted both from the widespread availability of the computer technology necessary for advanced methods of analysis, and from the increasing relative importance of genetic diseases in man, now that environmental agents such as bacteria and viruses are coming more under control. Furthermore, the human gene map is no longer largely a desert; we now know which specific chromosome carries each of over 200 loci (69). In this review we describe the various models, developed mostly over the last decade, for the purposes of genetic counseling and analysis. We do not consider the many models that have been developed to study population or evolutionary genetics.

Genetic Counseling vs Analysis

Genetic counseling requires several kinds of expertise, but here we are concerned only with the statistical modeling aspects. The question to be answered is this: given that we know the genetic mechanism that underlies the transmission of a disease, together with appropriate values for all the parameters involved, what is the probability that a certain individual, on the basis of what we know about this individual and/or his relatives, should have the disease? More generally, we can ask what the probability is that he should have a particular phenotype. (The individual for whom this question is asked may be as yet unborn, as when a couple wishes to know the probability of their having a child with a particular disease.) This problem can be thought of as the mirror image of the problem posed in genetic analysis: given that we know the phenotypes of a set of individuals with respect to some trait, what is the genetic mechanism that underlies that trait, and how do we estimate the parameters involved? It is far easier to answer the genetic counseling question than the analysis question; in fact it can be shown that in the absence of certain kinds of data, the analysis question is virtually unanswerable. Therefore, we start in Section 2 by answering the counseling question for certain models, restricting ourselves to the situations in which at most two individuals are involved; this enables us to introduce some basic genetic mechanisms before considering the problem of analysis.

Terminology and Symbols

TERMINOLOGY The terminology we use is that mostly used by human geneticists. The normal chromosomal complement in man comprises 22 pairs of homologous autosomes, or autosomal chromosomes, and two sex chromosomes—XX in females and XY in males. Genes occur at loci in linear sequence along the chromosomes. At each autosomal locus two genes occur, one on each of the homologous chromosomes. Different genes that occur at the same locus are termed alleles. Each parent transmits one of his two genes at any given locus independently to each offspring. (This, essentially, is Mendel's first law, or the "law of segregation.") However, the X and Y chromosome in males do not constitute a homologous pair: males transmit their X chromosome to each of their daughters and their Y chromosome to each of their sons. Loci on the X chromosome are called X-linked.

The two genes that an individual has at a given locus comprise his genotype at that locus. An observed characteristic on an individual is his phenotype or phenotypic value; the phenotype may be qualitative (e.g. color of eyes) or quantitative (e.g. height).

If the genes that cause differences in a phenotypic trait are all at the same locus, the trait is called monogenic. Since there must be at least two different genes involved for genetic differences to occur, it would be more logical to call such a trait unilocal; the term monogenic, however, is used far more commonly. Although any one individual can have at most two different genes at a locus, more than two different genes can occur at that locus in the population at large; if this occurs, the locus is termed multiallelic and such a genetic system is still monogenic.

If the phenotype is controlled by the segregation of genes at many loci in any one family, it is termed polygenic—though the term multilocal would be more logical. Later we see that the usual model for polygenic inheritance is a very special case and, as is often implicitly assumed in the literature, this special case is taken to be the definition of polygenic inheritance.

The term multifactorial is often used as a synonym for polygenic, but should more properly have a broader connotation. A monogenic trait is multifactorial if environmental influences are also involved in determining the phenotype. Furthermore, although any trait in which many loci are involved is multifactorial, the term polygenic is usually restricted to models in which normality is assumed, as described in Sections 2 and 4.

SYMBOLS Within each of the major sections of this review a consistent symbolism is used; each symbol has only one meaning. However, to help the reader go back to the original literature, some of the original symbolism has been kept; for this reason the symbolism is not exactly the same in each section. Throughout the paper, however, we consistently use the following abbreviations: $P(A)$, probability of the event A; $E(x)$, expectation or mean of x; $V(x) = E[x - E(x)]^2$, variance of x; $\mathrm{Cov}\ (x,y) = E\{[x - E(x)]\ [y - E(y)]\}$, covariance of x and y; $\phi(z,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp[-\frac{1}{2}(z/\sigma)^2]$, ordinate at z of a normal density function with mean 0 and variance σ^2; $\Phi(x) = \int_{-\infty}^{x}\phi(u,1)du$, cumulative standardized normal distribution; $N(\mu, \sigma^2)$, normally distributed with mean μ and variance σ^2; and δ_{ij}, Kronecker delta: $\delta_{ij} = \begin{cases} 1 \text{ if } i=j \\ 0 \text{ if } i \neq j \end{cases}$.

2 DICHOTOMOUS TRAITS—RANDOM INDIVIDUALS AND RELATIVES OF AFFECTED INDIVIDUALS

Statistical models in human genetics can be classified according to various criteria: the genetic mechanism (e.g. monogenic, polygenic); the kind of phenotype (e.g. qualitative or quantitative, univariate or multivariate); the kinds of individuals sampled (e.g. unrelated individuals, related pairs, nuclear families, large pedigrees); and the purpose of the model (genetic counseling or genetic analysis). In principle we can consider every possible cell in such a multiple classification, but in practice not all of these possibilities have been explored in depth. The models we consider in this section assume that, at most, pairs of related individuals are involved. We also restrict ourselves to a dichotomous phenotype (which we can take to be affected vs unaffected), and in such a situation there are only limited possibilities for genetic analysis, e.g. for determining whether the mode of inheritance is monogenic or polygenic (53, 58). For this reason we consider only the genetic counseling problem, asking for each genetic model what the probability is that a random member of the population is affected, i.e. what is the prevalence of the disease in question, and what the probability is that a particular relative of an affected individual is affected.

Monogenic and Oligogenic Models

BASIC AUTOSOMAL LOCUS For simplicity we consider a locus with just two alleles, A and a, random mating, and no selection or mutation. If the frequency of the allele A in the population is p, and that of a is $q = 1 - p$, then the genotypic frequencies are p^2 for AA, $2pq$ for Aa, and q^2 for aa. If f_i is the probability that a random individual of genotype i ($i = AA, Aa, aa$) is affected, then the answer to the first question, namely the prevalence of the disease in the population, is simply

$$\eta = p^2 f_{AA} + 2pq f_{Aa} + q^2 f_{aa}. \qquad 1.$$

To answer the second question, we use the stochastic matrices developed independently by Geppert & Koller (40a) and by Li & Sacks (66), but with the notation introduced by Campbell & Elston (5). These matrices are

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{T} = \begin{bmatrix} p & q & 0 \\ \tfrac{1}{2}p & \tfrac{1}{2} & \tfrac{1}{2}q \\ 0 & p & q \end{bmatrix}, \mathbf{U} = \begin{bmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{bmatrix}.$$

Let Y and Z be two individuals, and order their possible genotypes $1 = AA$, $2 = Aa$, and $3 = aa$. Then in each matrix the element in the ith row and jth column is the probability that Y has the jth genotype given that Z has the ith genotype, where Y and Z have the appropriate relationship. The matrix **I** gives these probabilities when Y and Z are monozygotic twins, or more generally, conditional on both of Y's genes being identical by descent (i.b.d.) with Z's genes: **T** gives these probabilities when Y is a parent or child of Z, or more generally, conditional upon Y and Z sharing just one gene i.b.d.; and **U** gives these probabilities when Y and Z are unrelated, i.e. conditional upon their sharing no genes i.b.d.

For any particular relationship between Y and Z, let c_T be the probability that they share both genes i.b.d., c_T the probability that they share one gene i.b.d., and c_U the probability that they share no genes i.b.d. at an autosomal locus. It follows that the appropriate matrix for this relationship is $\mathbf{R} = c_I\,\mathbf{I} + c_T\,\mathbf{T} + c_U\,\mathbf{U}$. The coefficients can be derived from Mendel's first law. For the grandparent-grandchild relationship, for example, $c_I = 0$, $c_T = \tfrac{1}{2}$, and $c_U = \tfrac{1}{2}$, so that $\mathbf{R} = \tfrac{1}{2}\mathbf{T} + \tfrac{1}{2}\mathbf{U}$.

Premultiplying **R** by a diagonal matrix D whose ith diagonal element is the frequency of the ith genotype in the population, we obtain the joint genotypic distribution for Y and Z. Thus, if we let **f** be a column vector with transpose $\mathbf{f}'$ whose ith element is f_i, the joint probability that both Y and Z are affected is $\mathbf{f}'\mathbf{DRf}$. It follows that the answer to our second question, the probability that Y is affected given that his relative Z is affected, is $\mathbf{f}'\mathbf{DRf}/\eta$, where **R** is appropriate for the relationship between Y and Z.

We have assumed in this development that the same **f** is applicable to both Y and Z, but only trivial changes are required if this is not so. For example, **f** may be age or sex dependent. By using Y and Z as subscripts to denote for which individual the parameters should be appropriate, the more general expression for the probability that Y is affected given Z is affected is $\mathbf{f}_Z'\mathbf{D}_Z\mathbf{R}_{YZ}\mathbf{f}_Y/\eta_Z$. The extension to a multiallelic locus is also quite simple (40, 66).

BASIC X-LINKED LOCUS Again we consider just two alleles, A and a, random mating, and no selection or mutation; we further assume that the gene frequency is the same in both sexes. For females, just as in the autosomal case, the prevalence of the disease is as given in equation 1. For males, who only have one X chromosome, it is convenient to represent the Y chromosome by a dot. Thus the two genotypes that are possible are $A.$ and $a.$, with frequencies in the population p and q, respectively. We need to define $f_{A.}$ and $f_{a.}$ analogously as before, and then the prevalence of the disease among males is

$$\eta_{\male} = pf_{A.} + qf_{a.}. \qquad 2.$$

If $f_{a.} = f_{aa} = 1$ and $f_{A.} = f_{AA} = f_{Aa} = 0$, the disease is caused by a simple recessive X-linked gene; in this situation we see from equations 1 and 2 that the prevalence of the disease in females is the square of the prevalence in males. Conversely, if $f_{a.} = f_{aa} = 0$ and $f_{A.} = f_{AA} = f_{Aa} = 1$, the disease is caused by a simple dominant X-linked gene; in this situation, in the limit as $p \rightarrow 0$, the prevalence in females is twice the prevalence in males. Thus a sex difference in the prevalence of a disease may suggest X-linked inheritance as a possible cause for that disease.

The probability that Y is affected given that Z is affected can again be expressed as $\mathbf{f}_Z'\mathbf{D}_Z\mathbf{R}_{YZ}\mathbf{f}_Y/\eta_Z$, but with appropriate redefinitions (40a, 66).

LESS RESTRICTED MONOGENIC MODELS Multiallelic loci can easily be allowed for under the general procedures just given. Allowing for nonrandom mating, mutation, and selection, however, poses problems. Much is known about the equilibrium genotypic frequencies, for both autosomal and X-linked loci, under various systems of assortative mating and inbreeding, and under mutation and selection pressures (e.g. 15, 65). With this information, and the appropriate **f**, it is easy to express the prevalence of a disease as a function of the various genetic parameters. Very little general theory, however, has been developed for determining the probability that the relative of an affected individual is affected under these less restrictive conditions.

In an infinitely large population undergoing random mating, consanguineous matings (i.e. matings between relatives) do not occur; this has been tacitly assumed to be the case in our development so far. In a finite population, however, consanguineous matings do occur, even under random mating. Since the presence of consanguineous matings can have a large effect on the probability that the relative of an affected individual is affected, this situation is important for genetic counseling, and a general theory for it is available. Consider four objects that may be pairwise identical or not identical, with the relation "identical" being symmetric and transitive. There are 15 possibilities or identity states (42, 80). If these four objects are the genes of two individuals at an autosomal locus, and the relation is i.b.d., it is found that 9 of these 15 states are genetically distinct; and in 6 of these 9 genetically distinct states the two genes of one or both of the two individuals are i.b.d. Thus only three genetically distinct states exist if we assume an individual cannot have two genes i.b.d., and these states correspond to **I**, **T**, and **U**.

To allow for consanguineous matings it is necessary to consider nine, rather than three, distinct states for an autosomal locus. Furthermore, it is found during the development that **R** cannot be expressed as a linear combination of nine matrices with scalar coefficients, since some of the coefficients depend upon the row of the matrix. Thus each genotype of Z is considered separately, and then the genotype distribution of Y conditional on Z's genotype is obtained as a linear combination of row vectors. Details of the method are given by Jacquard (52) for a multiallelic autosomal locus, but this paper contains a small error (31). The coefficients in the linear combinations can be found, for any relationship whatsoever, by an algorithm developed by Nadot & Vaysseix (80).

OLIGOGENIC MODELS The case where a few loci are involved (which would more logically be termed paucilocal models) can be handled, when the genes at different loci segregate independently, by the use of Kronecker products of the matrices **I**, **T**, and **U** defined above (5). However, genes at different loci on the same chromosome tend to be transmitted together, rather than independently, provided the loci are not too far apart on the chromosome. This phenomenon is termed genetic linkage, and extends Mendel's second law, the "law of independent assortment," which holds good only for genes on nonhomologous chromosomes or genes at loci on homologous chromosomes that are far apart. The derivation of **R** for two linked autosomal loci is dealt with by several authors (5, 20).

Polygenic Model

The basic polygenic model for dichotomous traits is described here under the following restrictive assumptions: all the genes involved are autosomal, with additive effects; there is random mating and no mutation or selection; and all environmental effects are completely independent—i.e. the familial nature of the disease is due to genetic causes alone. Some of these assumptions are relaxed in Section 3, when we discuss quantitative traits. The model has been developed by two different approaches that, although seemingly different, are mathematically equivalent. Furthermore, it is possible to parametrize the model in two different ways since without loss of generality one of the parameters can arbitrarily be set equal to a constant. We give both approaches to the model, but restrict ourselves to the parametrization (but not the notation) used by Curnow (16, 17).

THE RISK FUNCTION APPROACH Consider a locus at which two types of genes occur, which we call 0-genes and 1-genes. In the polygenic model we suppose that there are many such loci, and that the risk to an individual of being affected depends only on the proportion of his genes at these loci that are 1-genes. In a population of randomly mating individuals, as the number of these loci that are independently segregating increases, this proportion will tend to be normally distributed. Thus under this model the dependence of risk on genotype can be replaced by the dependence of risk on a normally distributed random variable, G say, called genetic liability. Since we never actually measure G, we assume without loss of generality it is $N(0,1)$. We now assume that the risk function takes the form of a cumulative normal with mean θ and variance σ^2, i.e. the probability that an individual with genetic liability G is affected is $\Phi[(G-\theta)/\sigma]$. Thus, analogous to equation 1, but integrating over the continuous variable G rather than summing over a finite number of genotypes, the prevalence of the disease in the population is

$$\eta = \int_{-\infty}^{\infty} \phi\,(G,1)\; \Phi[(G-\theta)/\sigma]\; dG = \Phi[-\theta/(1+\sigma^2)^{1/2}]. \qquad 3.$$

It should be noted that the assumption that G is normally distributed over the population is not a strong one; provided it is continuous, it can always be transformed to be normally distributed. Similarly, the assumed form of the risk function is not, by itself, very restrictive. Although individually not important, when taken together these assumptions, for which there is little biological justification, become very strong. A further assumption now needed to apply this model to pairs of relatives is that the distribution of G among such pairs over the population is bivariate normal; the correlation ρ in this distribution is the expected proportion of genes the two relatives share i.b.d., i.e.

$$\rho = c_I + \tfrac{1}{2} c_T.$$

Denote this standardized bivariate normal density with correlation ρ, for a particular pair of relatives Y and Z, $\phi(G_Y, G_Z, \rho)$; this corresponds to the matrix **DR** in the monogenic case. Then the joint probability that both Y and Z are affected, analogous to **f′DRf** (which is a double summation over the joint distribution), is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi[(G_Z-\theta)/\sigma]\phi(G_Y, G_Z, \rho)\Phi[(G_Y-\theta)/\sigma]dG_Y dG_Z. \qquad 4.$$

This quantity, divided by the prevalence η given in equation 3, is then the probability that Y is affected given that Z is affected. We have assumed the same risk function for Y and Z in this development, but as before for **f**, there is no difficulty in allowing the two risk functions to be different. Figure 1 gives a pictorial representation of this model: two risk functions are shown, with differing values of θ: θ_1 and θ_2.

Curnow (16) has shown that expression 4 is equal to

$$\int_{-\infty}^{\infty}\phi(u,\ 1)\{\Phi[-(\theta+\rho^{*1/2}u)/(1+\sigma^2-\rho^*)]\}^2 du,$$

where $\rho^* = \rho(1+\sigma^2)$, which is an easier form to evaluate computationally. Approximations to this expression, of varying degrees of accuracy, also exist (36, 70, 99).

Finally it should be noted in passing that the same model can be used for counseling without the necessity of assuming that the familial nature of the disease is due to genetic causes only, by allowing G to contain an environmental component (16, 18, 100); but then there is no way to derive ρ from genetic principles alone.

THE THRESHOLD APPROACH The original approach to the polygenic model for a dichotomous trait was quite different (13, 36, 115). Assume the existence of a normally distributed random variable L, which we call total liability. (This variable

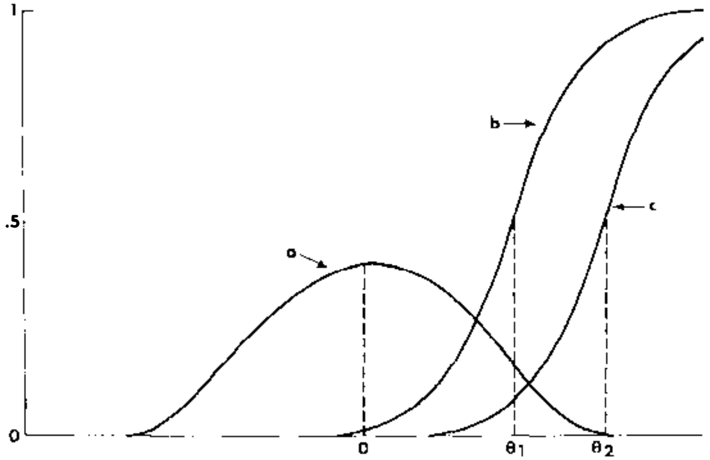

Figure 1 Pictorial representation of the risk function approach to the polygenic model for a dichotamous trait; the abscissa is genetic liability G. (*a*) Density function of G, $N(0,1)$; (*b*) risk function $\Phi[(G-\theta_1)/\sigma]$; (*c*) risk function $\Phi[(G-\theta_2)/\sigma]$. The risk function is the probability that an individual with genetic liability G is affected.

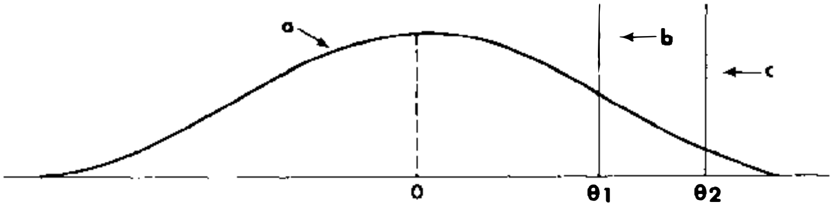

Figure 2 Pictorial representation of the threshold approach to the polygenic model for a dichotomous trait; the abscissa is total liability, $L = G + E$. (*a*) Density function of total liability, $N(0,\ 1+\sigma^2)$; (*b*) threshold θ_1; (*c*) threshold θ_2. Individuals whose total liability is greater than the threshold are affected.

is usually simply called liability.) An individual's total liability is the sum of two uncorrelated components: a genetic liability, the same G as before, and an environmental liability, E,
affected if and only if his total liability is greater than θ, in this approach called the threshold. Finally we assume that any correlation in L between pairs of related individuals is due solely to a correlation in G, i.e. the pairs of values of E are uncorrelated. As before, G is assumed to follow a standardized bivariate normal density among pairs of relatives over the whole population, with correlation $\rho = c_I + \frac{1}{2} c_T$. It can be shown that this model results in the same expression as given in equation 3 and expression 4 above. Figure 2 illustrates this approach, with two different thresholds being shown. Figures 1 and 2 are two different ways of depicting the same model.

Whichever approach is taken, $V(G)$ is arbitrary, in the sense that the model depends on $V(G)$ and $V(E)$ only through the ratio $V(G)/V(E)$, or alternatively through the ratio $V(G)/[V(E) + V(G)]$. As is seen from its definition given below, this latter ratio is the heritability of the (total) liability to the disease, though it is often mistakenly called the heritability of the disease itself (25, 28).

3 PATH ANALYSIS AND VARIANCE COMPONENTS

In this section we deal with the analysis of quantitative traits under the assumption of multifactorial inheritance. The purpose of the methods described in this section is not so much to distinguish between various modes of genetic inheritance as to resolve genetic and environmental effects. There are basically two approaches to this problem, the methods of path analysis and variance components. Path analysis, whose primary purpose is to explain the interrelationships among variables, was originally developed for the analysis of correlations by Wright (114). The method of variance components, on the other hand, was developed as a natural extension of the analysis of variance, i.e. the analysis of the variance of a quantitative trait into component parts. In large samples, both these methods should give essentially identical results under similar assumptions; for the same model it is largely a matter of taste as to which method is chosen. Cavalli-Sforza & Feldman (10, 38) have developed a model of cultural inheritance in which the phenotype of a child is determined by the parental phenotypes as well as by the child's own genotype. Although it has not yet been applied to any body of real data, the path analysis models have been shown to be capable of extracting the main features of such a model (90).

Originally the variance component models included just additive gene effects, intralocus gene interactions (dominance), and random environmental effects (39). The biometrical geneticists in Birmingham, led by Jinks & Eaves (54), have extended the model to allow a distinction between intrafamilial and interfamilial environmental effects; in addition, models that incorporate interlocus gene interactions (epistasis) have been developed for twin studies (11, 47, 81). The method of path analysis was originally developed for linear additive systems (114), i.e. linear models with no statistical interaction effects, but it has recently been extended to allow the approximate treatment of dominance, epistasis, and genotype-environment interactions (116). Models that incorporate indices of familial environment and a logical treatment of causal relationships between generations have been developed by Morton (74) and Rao et al (94, 95), making path analysis a powerful tool for the resolution of genetic and environmental inheritance (85). Among current models those developed for use with path analysis are more realistic, for traits in which the environment is important, than those developed for use with variance components; for this reason we only briefly summarize the method of variance components.

Basic Model

Throughout this section variables (causes and effects) are denoted by capital letters, and parameters are denoted by lower case letters, Greek or Roman. We restrict our attention to the following linear additive model:

$$P = G + C + R, \qquad 5.$$

where P is the phenotype, value of the quantitative trait; G is the genotype, assumed polygenic; C is the controllable environment, called common or family environment; and R is the random environment, unique to each individual, with $V(P) = \sigma_P^2$, $V(G) = \sigma_G^2$, $V(C) = \sigma_C^2$, $V(R) = \sigma_R^2$ and $\mathrm{COV}(G,C) = \sigma_{GC}$, $\mathrm{COV}(G,R) = \mathrm{COV}(C,R) = 0$.

Thus the total phenotypic variance is given by

$$\sigma_P^2 = \sigma_G^2 + \sigma_C^2 + 2\sigma_{GC} + \sigma_R^2. \qquad 6.$$

Important underlying assumptions are that (*a*) a linear additive model exists for the quantitative trait, which assumes no dominance, epistasis, or genotype-environment interactions, and (*b*) genotype-environment covariance is in equilibrium. Furthermore, when twins and adopted children are included in the data it is usual to assume that (*c*) the phenotypic similarity of twins due to common prenatal and postnatal environment, irrespective of zygosity, is no greater or less than for ordinary siblings, (*d*) adoptions are random, with no regard to genetic or environmental variables, and (*e*) true parents are assumed to exert no influence on the children either prior to or after their adoption.

The terms in equation 6 are the commonly estimated variance components. In path analysis they are obtained, relative to the total phenotypic variance, as functions of path coefficients.

HERITABILITY One of the fundamental parameters of multifactorial inheritance is heritability, which we denote h^2; it is defined as the proportion of total phenotypic variance due to genetic factors: $h^2 = \sigma_G^2/\sigma_P^2$. This definition holds regardless of any genotype-environment covariance (51). In the presence of dominance and epistasis the genetic variance is split into components, called additive genetic variance (σ_A^2), dominance variance, and epistatic variance, and accordingly two types of heritability are defined: heritability in the narrow sense, $h_n^2 = \sigma_A^2/\sigma_P^2$; and heritability in the broad sense, $h_b^2 = \sigma_G^2/\sigma_P^2$. When gene interactions are absent, $h_n^2 = h_b^2 = h^2$.

Path Analysis

In path analysis the underlying model is depicted as a path diagram, which expresses as paths the relationships among the variables involved. These paths are either causal paths from causes to effects, indicated by single-headed arrows, or correlational paths, indicated by double-headed arrows. Associated with each path is a path coefficient, formally defined as a standardized partial regression coefficient. Given a path diagram a simple calculus exists for deriving, as functions of the path coefficients, the correlation between any pair of variables in the diagram (64, 116). Effects that are used as imperfect measures of causes are called indices. Thus, two kinds of observable variables exist, effects and indices (*I*), and we indicate how data on these can be used to test hypotheses about the various paths in a diagram. First, we give details of a general model developed for path analysis. Following a recent convention (92), causes are denoted by ellipses and effects (including indices) are denoted by rectangles.

NUCLEAR FAMILIES A general model that incorporates specific maternal effects, shown in Figure 3, has been proposed (D. C. Rao, N. E. Morton, C. L. Gulbrandsen,

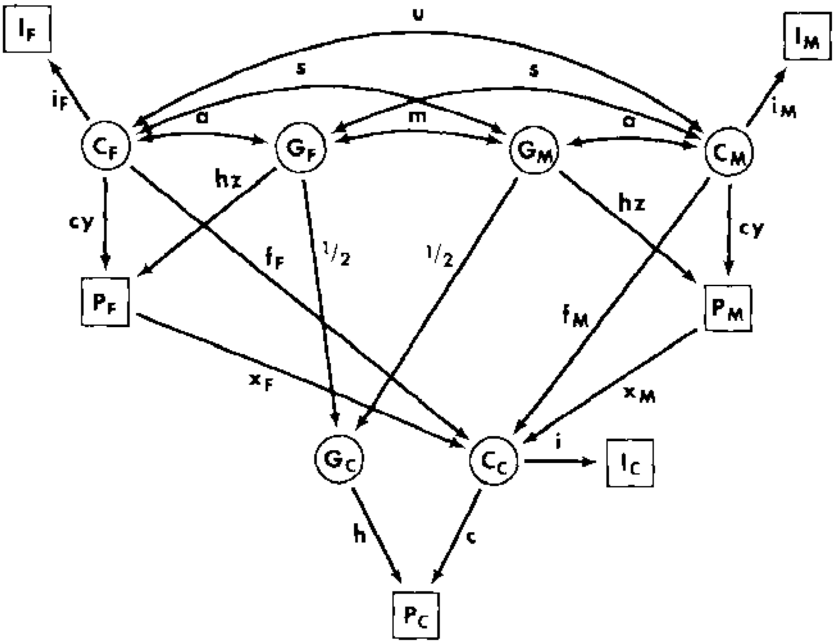

Figure 3 Marital and parent-offspring path diagrams. The subscripts *F, M,* and *C* denote father, mother, and child, respectively. *G* is genotype, *P* is phenotype, and *C* is common environment with index *I*. Each effect (rectangle) has another path from an independent cause; these residual paths are not shown, since they do not contribute to the observed correlations among effects.

Table 1 Path coefficients of the general model (Figure 3)

	Symbol	Definition
Marital		
	m	correlation between parental genotypes.
	u	correlation between common environments of spouses.
	s	correlation between common environment of an adult and spouse's genotype.
Environmental		
	c	effect of common environment on child's phenotype.
	y	ratio of the effects of common environment on adult's phenotype and on child's phenotype.
	f_F	effect of father's common environment on child's common environment.
	f_M	effect of mother's common environment on child's common environment.
	x_F	effect of father's (adult) phenotype on child's common environment.
	x_M	effect of mother's (adult) phenotype on child's common environment.
Genetic		
	h	effect of genotype on child's phenotype (square root of heritability).
	z	ratio of the effects of genotype on adult's phenotype and on child's phenotype.
Indices		
	i	effect of child's common environment on child's index (a measure of adequacy of the index).
	i_F	effect of father's common environment on father's index.
	i_M	effect of mother's common environment on mother's index.
Derived		
	a	correlation between individual's genotype and common environment $= [hz(1 + m)(x_F + x_M) + s(f_F + f_M + cyx_F + cyx_M)]/[2 - (f_F + f_M + cyx_F + cyx_M)]$.[a]

[a] In the absence of maternal effects, $x_F = x_M = x$, $f_F = f_M = f$, and $a = [hzx(1 + m) + s(f + cyx)]/[1 - (f + cyx)]$.

G. G. Rhoads, and A. Kagan, submitted for publication) for the analysis of data on nuclear families (parents and their children). The model contains 14 functionally independent parameters, defined in Table 1; as indicated at the bottom of the table, parameter a, the correlation between an individual's genotype and common environment, is functionally dependent on the other parameters. We distinguish genetic and environmental effects in children and adults: h^2 is the heritability in children, whereas it is h^2z^2 in adults; c^2 is the proportion of the phenotypic variance due to common environment in children, whereas it is c^2y^2 in adults. Specific maternal effects are included through different effects of parental common environments, and phenotypes, on the common environment they provide to their children (f_F, f_M, x_F, x_M). Given phenotypes and environmental indices for both parents and children, nuclear families generate 16 correlations in all, distinguishing paternal and maternal ones: the two parental phenotypes, two parental indices, and child's phenotype and index generate $\binom{6}{2} = 15$ correlations, and in addition there is the sibling correlation derived from the children. Methods for calculating these 16 correlations are known (41, 102), as well as the expected correlations derived from Figure 3 (88). Indices may be created by regressing the phenotype on relevant variables that are not themselves products of the genotype. The general model in terms of 14 parameters is overdeterminate in nuclear families, which leaves at least two degrees of freedom for testing the goodness of fit of the model.

ADOPTED CHILDREN In the basic model $P = G + C + R$, G and C are not correlated for adopted children, since G comes from the true parents whereas C comes from the adoptive parents. Thus the phenotypic variance of such children is less than that for true children. Let the phenotypic variance for adopted children be $\sigma^2_{P^*}$, so that, from equation 6, $\sigma^2_P = \sigma^2_{P^*} + 2\sigma_{GC}$. Then, correlations that involve adopted children are multiplied by θ or θ^2, depending on whether one or both the individuals involved are adopted, where $\theta = \sigma_P/\sigma_{P^*}$ (94).

OTHER RELATIONSHIPS Rao et al (95) presented models for a variety of other biological and social relationships including twins, half-sibs, foster children, uncle-niece, and first cousin. These models did not incorporate maternal effects nor different indices for children and adults; however, it is easy to introduce both these extensions. Data on such relationships can be added to data from nuclear families for tests of consistency as well as for increased power.

ASSORTATIVE MATING As is seen in Figure 3, path analysis easily allows for assortative mating. Fisher (39) gave a comprehensive theory for assortative mating, which remains obscure to many in spite of several attempts to explain it (14, 71, 109, 110, 111, 112). Wright (116) gives an elegant treatment of the general theory, which is briefly explained here. There are four basic types of assortative mating.

Type 1: genetic assortative mating The first type postulates that the only cause of marital correlation is genetic; related individuals, e.g. first cousins, marry. This is the special case of $u = s = 0$ in Figure 3. This corresponds to inbreeding or consanguinity.

Type 2: environmental assortative mating Under type 2, environmental similarity of spouses is the only cause of marital correlation. This is the special case of $m = s = 0$ in Figure 3. This type of assortative mating may be appropriate for metabolic traits.

Type 3: assortative mating based on a common social homogamy According to type 3, assortative mating for status, tastes, contacts, and other aspects of group

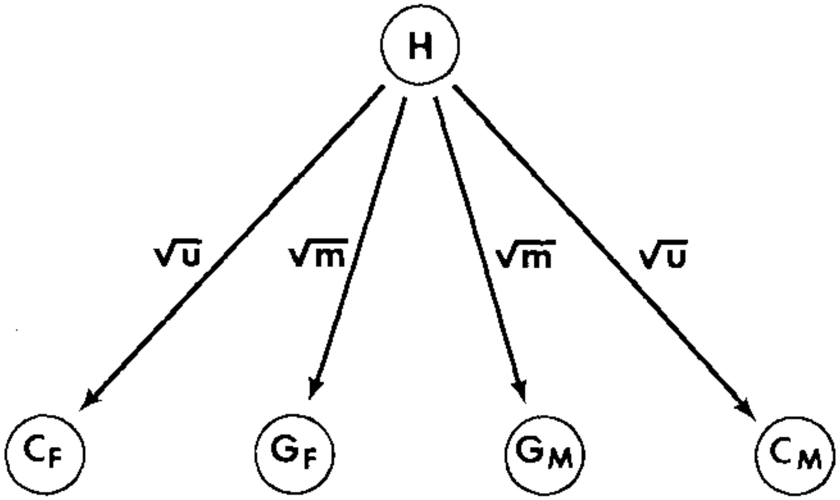

Figure 4 Path diagram for assortative mating type 3. The subscripts F and M denote father and mother, respectively. H is social homogamy, G is genotype, and C is common environment. Residual paths are not shown.

membership leads secondarily to a marital correlation (77, 78). Figure 4 presents the marital variables, which include social homogamy (H), which has a path coefficient $\sqrt{m}$ to genotype and $\sqrt{u}$ to common environment of the mates. This generates the following marital correlations: $m =$ correlation between genotypes of mates; $u =$ correlation between common environments of mates; and $\sqrt{mu} =$ correlation between genotype and spouse's common environment. This corresponds to the special case of $s = \sqrt{mu}$ in Figure 3. This type of assortative mating is most likely for behavioral traits such as I.Q. and is a special case of a model introduced by Wright (116).

Type 4: phenotypic assortative mating Type 4 assumes that potential mates assort on the basis of their phenotypes alone; it is also called direct homogamy. This model is perhaps relevant to certain physical traits such as height, but it is unlikely that even for such traits assortment takes place only in terms of the phenotype. It has been developed in detail by Wright (116) and also by Li (65). When dominance is introduced (116), this case corresponds to Fisher's treatment (39), except for the complications due to common environment and the approximation involved in treating dominance by path analysis.

MAXIMUM LIKELIHOOD ANALYSIS Morton (74) and Rao et al (94) have proposed that analyses be based on Fisher's z-transformation of the correlation coefficients calculated between pairs of observed variables, since this has been shown to approach normality faster than the correlation coefficient (40). Let $z_1, z_2 \ldots z_m$ be m z-transforms of observed and bias-corrected correlations (see 89, 94). If the covariance between z_i and z_j is denoted by σ_{ij}, and the covariance matrix is denoted by Σ, the log-likelihood of the observed data can be written, assuming multivariate normality for all the z's, as

$$\ln L \approx -\chi^2/2 + \text{constant}, \qquad 7.$$

where $\chi^2 = [\mathbf{z} - E(\mathbf{z})]'\Sigma^{-1}[\mathbf{z} - E(\mathbf{z})]$ and $\mathbf{z}$ is a column vector of the observed z transforms. The approximation in equation 7 arises from the fact that a power of $|\Sigma|$ has been incorporated into the constant; empirically it is found that $|\Sigma|$ hardly changes for different models. The elements of $E(\mathbf{z})$ are expressed as functions of the path coefficients, estimated by maximizing $\ln L$ or, equivalently, minimizing χ^2. To test hypotheses, let $\chi^2_{m-\kappa-\omega}$ be the minimum value of χ^2, with $m-\kappa-\omega$ degrees of freedom (df), when $\kappa + \omega$ parameters are estimated, and let $\chi^2_{m-\kappa}$ be the minimum value, with $m-\kappa$ df when only κ of the parameters are estimated—with the ω other parameters fixed at values that correspond to a null hypothesis. Then $\chi^2_\omega = \chi^2_{m-\kappa} - \chi^2_{m-\kappa-\omega}$ is asymptotically distributed as chi square with ω df under the null hypothesis, and thus it provides the likelihood ratio test of the null hypothesis.

Specification of the covariance matrix Σ in equation 7 depends on how the data are obtained. If $z_1, z_2 \ldots z_m$ are independent (estimated from different samples), $\sigma_{ij} = 0$ $(i \neq j)$, and in this situation it is known that the asymptotic properties hold good from even small to moderate sample sizes. For this case, the χ^2 expression of equation 7 simplifies to

$$\chi^2 = \sum_{i=1}^{m} [z_i - E(z_i)]^2 / V(z_i). \qquad 8.$$

If some or all correlations are estimated from the same sample, not all the covariances σ_{ij} are zero; in that case Σ should incorporate the asymptotic correlations between correlations (27). The variances are given by

$$V(z_i) = \sigma_{ii} = \begin{cases} 1/(n_i - 3), \text{ if } z_i \text{ is the transform of an interclass correlation.} \\ 1/(n_i - 1.5), \text{ if } z_i \text{ is the transform of an intraclass correlation.} \end{cases}$$

When some correlations are estimated from independent samples and some are estimated from the same sample, the appropriate log-likelihood may be written as a sum of two components given by equations 7 and 8.

Variance Components

We now turn to the estimation of the variance components in equation 6. This has traditionally been done by equating observed and expected mean squares, and we illustrate this method by considering the analysis of twin data. Haseman & Elston (47) presented a comprehensive genetic model and theory for the estimation of variance components solely from twin data. Their treatment is reparametrized here to conform to the general model in equation 5, which ignores dominance and epistasis. For two monozygotic (MZ) twins reared together, the phenotypes are written as $P_1 = G_1 + C_1 + R_1$, $P_2 = G_1 + C_1 + R_2$, so that P_1 and P_2 differ only with respect to random environments. Similarly, the phenotypes for two dizygotic (DZ) twins are $P_1 = G_1 + C_1 + R_1$, $P_2 = G_2 + C_1 + R_2$, in which the genotypes are different, as the random environments. Table 2 presents the analysis of variance table for *n* MZ and *m* DZ pairs. The phenotypic variances and covariances are recoverable from the expected mean squares, as indicated at the bottom of the table. Not all the four parameters (σ_G^2, σ_C^2, σ_R^2, σ_{GC}) are estimable solely from twin data. Assume $\sigma_{GC} = 0$. The least squares estimates of the three remaining parameters are then obtained by minimizing the sum of the four squared differences between the mean squares and their expected values, which results in the unbiased estimates $\hat{\sigma}_G^2 = \text{AMZ} - \text{WMZ} - \text{ADZ} + \text{WDZ}$, $\hat{\sigma}_C = (\text{WMZ} - \text{AMZ} + 2\ \text{ADZ} - 2\ \text{WDZ})/2$, and $\hat{\sigma}_R = (\text{ADZ} + \text{WDZ} + 3\ \text{WMZ} - \text{AMZ})/4$. Haseman & Elston (47) also gave a weighted least squares solution and a likelihood analysis, assuming the phenotypes of a twin pair follow a bivariate normal distribution.

Any analysis based on twin data alone is suspect (29). Eaves and his colleagues (21–23) have elaborated the biometrical genetical approach for many different types of relationships, confining themselves to the weighted least squares method of estimating variance components from pairs of relatives. On the other hand, Lange et al (63) have developed a method of obtaining maximum likelihood estimates of the variance components and testing hypotheses about them, from pedigrees of arbitrary structure. The phenotypes of the members of the pedigree are assumed to follow a multivariate normal distribution whose covariance matrix is expressed as a function of an additive genetic variance, a dominance variance, and an environmental variance. Two advantages of this approach are that the variance components are estimable even from one large pedigree, and they adopt the most efficient method of analysis through maximum likelihood. If this method could be extended to include in the model assortative mating and the environmental paths of Figure 1, not only would there be the advantage of utilizing pedigree data, but also dominance would be accurately allowed for. It must be pointed out, however, that by any method of analysis dominance is largely confounded

Table 2 Analysis of variance for *n* pairs of *MZ* and *m* pairs of *DZ* twins[a]

Type of twins	Source of variation	df	Mean squares	Expected mean squares
MZ	among pairs	$n-1$	*AMZ*	$2\sigma_G^2 + 2\sigma_C^2 + 4\sigma_{GC} + \sigma_R^2$
	within pairs	n	*WMZ*	σ_R^2
DZ	among pairs	$m-1$	*ADZ*	$\frac{3}{2}\sigma_G^2 + 2\sigma_C^2 + 4\sigma_{GC} + \sigma_R^2$
	within pairs	m	*WDZ*	$\frac{1}{2}\sigma_G^2 + \sigma_R^2$

[a] $V(P) = [E(AMZ) + E(WMZ)]/2 = [E(ADZ) + E(WDZ)]/2$.

$$\text{Cov}\,(P_1, P_2) = \begin{cases} [E(AMZ) - E(WMZ)]/2 \text{ for } MZ \\ [E(ADZ) - E(WDZ)]/2 \text{ for } DZ \end{cases}$$

with environmental effects unless special relationships, such as MZ twins and adopted children, occur in the data.

4 SEGREGATION ANALYSIS—NUCLEAR FAMILIES

In the previous section we described models, and corresponding methods of hypothesis testing, for analyzing a quantitative phenotypic trait into genetic and environmental components. Those models, which are essentially based on the correlations or covariances between pairs of relations, can analyze the genotype variance into additive and other components; however, they do not permit a detailed analysis of the mode of inheritance of the phenotype. We use the term genetic analysis to indicate that the genetic mechanism is being analyzed, and such an analysis becomes more feasible once we utilize in the analysis the complete structure of nuclear families or larger pedigrees. The genetic analysis of quantitative traits has recently taken a leap forward, largely as a result of two developments in methodology: in one (76), the traditional approach of basing tests on the distribution of the offspring phenotypes conditional on the parental phenotypes is followed, whereas in the other, developed with large pedigree structures in mind (34), tests are based on the unconditional likelihood of all the phenotypes. In this section we only consider nuclear families, but to answer the genetic counseling question and to simplify the extension to larger pedigrees in Section 5, unconditional likelihoods are considered first.

Let us briefly consider the genetic counseling problem. Once again, we wish to know the probability that an individual Y is (or will be) affected given that we know the affectation status of his relatives Z_1, Z_2 . . . , in the same nuclear family. Let $y=1$ represent the event Y is affected, let $y=0$ represent the event that Y is not affected, and let $\mathbf{z}$ represent the affectation status of the relatives. Then the probability we want is $P(y=1, \mathbf{z})/P(\mathbf{z})$. But note that $P(\mathbf{z})=P(y=1, \mathbf{z})+P(y=0, \mathbf{z})$. Thus if we calculate the probability that all the family members have their respective phenotypes, including (*a*) Y assumed to be affected and (*b*) Y assumed to be unaffected, then the probability we want is the first of these two divided by their sum. More generally, we can consider the likelihood that Y has any particular phenotype, which may be a probability or a density function. We now build up the likelihoods for a nuclear family under some general models, for the most part following Elston & Stewart (34), assuming we have a random family (of fixed size) from the population. These can be used as just indicated for genetic counseling, so long as all the necessary parameters are known, without invoking Bayes' theorem as is usually done (79). We then discuss how these likelihoods can be used for hypothesis testing and parameter estimation, and finally we take up the problem of nonrandom sampling. Unless otherwise stated we assume random mating.

General Likelihood for a Nuclear Family

OLIGOGENIC MODEL Denote the phenotypes of the father and mother z_f and z_m, and those of their n offspring z_j $(j=1, 2 \dots n)$. Let k be the number of different genotypes that affect the phenotype. (If one or more X-linked loci are involved, k is dependent on the sex of the individual; for ease of exposition we do not explicitly allow for this here.) The genotypes can be arranged in some specific order, and so we can talk of the ith genotype, $i = 1, 2 \dots k$. Let the probability density function of z conditional on the ith genotype be $g_i(z)$.

For reasons that become apparent in Section 5, we index the genotypes of the two parents s_0 and t_0, and those of the children s_1: each of these indices runs from 1 to k. Let $p_{s_0t_0s_1}$ be the probability that a child has genotype s_1, given that his parents' genotypes are s_0 and t_0. If only one locus is involved, $p_{s_0t_0s_1}$ must take on one of the values 0, ¼, ½, or 1; if w independent loci are involved, it must be the product of w factors each of which is 0, ¼, ½, or 1. Given that the parents' genotypes are s_0 and t_0, the likelihood for the jth offspring is $\Sigma_{s_1=1}^{k} p_{s_0t_0s_1} g_{s_1}(z_j)$. Now conditional on the genotypes of the parents, the genotypes of the offspring are independent of one another. We further assume that, conditional on their own respective genotypes, the phenotypes of the offspring are independent of one another; then the likelihood for the whole sibship, given that the parents' genotypes are s_0 and t_0, is

$$\Pi_{j=1}^{n}\Sigma_{s_1=1}^{k} p_{s_0t_0s_1} g_{s_1}(z_j). \qquad 9.$$

It is to be understood that the summation over s_1 in this expression is performed separately for each of the offspring, and the sums then are multiplied together. From now on we simply write symbols such as Π_j and Σ_{s_1}, since the limits will be clear.

Let ψ_i be the probability that a random individual from the population has the ith genotype, i.e. ψ_i, is the frequency of genotype i. Then the likelihood for the two parents can be written as

$$\Sigma_{s_0}\psi_{s_0}g_{s_0}(z_f)\Sigma_{t_0}\psi_{t_0}g_{t_0}(z_m), \qquad 10.$$

where again we assume that, conditional on their respective genotypes, the phenotypes of the two parents are independent. This is the sum of k^2 terms, each term corresponding to a particular pair of genotypes s_0 and t_0 for the parents. Multiplying each term in this sum by the likelihood for the sibship conditional on s_0 and t_0 (expression 9), we arrive at the total likelihood for the whole nuclear family:

$$\Sigma_{s_0}\psi_{s_0}g_{s_0}(z_f)\Sigma_{t_0}\psi_{t_0}g_{t_0}(z_m)\Pi_j\Sigma_{s_1}p_{s_0t_0s_1}g_{s_1}(z_j). \qquad 11.$$

This expression assumes that, conditional on their genotypes, the phenotypes of all the family members are mutually independent.

POLYGENIC MODEL Assuming a completely additive polygenic model, the likelihood for a nuclear family can be expressed exactly as in expression 11, with the following changes. The genotypes are replaced by random variables and summations are replaced by integrations; the random variables are taken to be $N(0, \sigma_G^2)$ in the population, where σ_G^2 is the additive genetic variance of the phenotypic trait. Denote the random variables for the parents a_0 and b_0, replacing s_0 and t_0, so that ψ_{s_0} becomes $\phi(a_0, \sigma_G^2)$ and ψ_{t_0} becomes $\phi(b_0, \sigma_G^2)$. Similarly denote the random variable for each offspring a_1, replacing s_1, and then $p_{s_0t_0s_1}$ becomes $\phi[a_1 - \frac{1}{2}(a_0 + b_0), \frac{1}{2}\sigma_G^2]$. The likelihood for a nuclear family thus becomes, analogous to expression 11,

$$\int_{-\infty}^{\infty}\phi(a_0, \sigma_G^2)g_{a_0}(z_f)\int_{-\infty}^{\infty}\phi(b_0, \sigma_G^2)g_{b_0}(z_m)\Pi_j\int_{-\infty}^{\infty}\phi[a_1 - \tfrac{1}{2}(a_0 + b_0), \tfrac{1}{2}\sigma_G^2]g_{a_1}(z_j)da_1db_0da_0, \qquad 12.$$

in which there is a separate integration over a_1 for each of the offspring.

MIXED MODELS The term mixed model has been used (76) to denote a model that incorporates both oligogenic and polygenic effects. The likelihood of a nuclear family under such a model, assuming the effects of the oligogenic and polygenic loci are additive, is developed quite easily by combining expressions 11 and 12 to obtain:

$$\Sigma_{s_0}\psi_{s_0}\int_{-\infty}^{\infty}\phi(a_0, \sigma_G^2)g_{s_0a_0}(z_f)\Sigma_{t_0}\psi_{t_0}\int_{-\infty}^{\infty}\phi(b_0, \sigma_G^2)g_{t_0b_0}(z_m)\cdot$$
$$\Pi_j\Sigma_{s_1}p_{s_0t_0s_1}\int_{-\infty}^{\infty}\phi[a_1 - \tfrac{1}{2}(a_0 + b_0), \tfrac{1}{2}\sigma_G^2]g_{s_1a_1}(z_j)da_1db_0da_0. \qquad 13.$$

where $g_{iG}(z)$ is the probability density function of z conditional on "oligogenotype" i and "polygenotype" G, i.e. the ith genotype at the oligogenic loci and the value G of the random variable that represents polygenic genotype.

The likelihood expressions 11, 12, and 13 all assume that, conditional on their genotypes, the phenotypes of all the individuals in the family are mutually independent. We can allow for non-independence by assuming that z_m, z_f, and z_i, conditional on the genotypes of all the family members, follow a multivariate distribution. If we allow for too many arbitrary correlations, however, the models will lead to no meaningful analysis in practical situations. As a compromise, Morton & MacLean (76) have considered a mixed model in which just one extra parameter is added, to allow for a common environmental component among all members of the same sibship. Let C be a $N(0, \sigma_C^2)$ random variable that takes on the same value for each member of a sibship. Then the likelihood under this more general model is the same as in expression 13, except that the factor in the second line is modified:

$$\Sigma_{s_0}\psi_{s_0}\int_{-\infty}^{\infty}\phi(a_0, \sigma_G^2)g_{s_0a_0}(z_f)\Sigma_{t_0}\psi_{t_0}\int_{-\infty}^{\infty}\phi(b_0, \sigma_G^2)g_{t_0b_0}(z_m)\cdot$$
$$\int_{-\infty}^{\infty}\phi(C, \sigma_C^2)\Pi_j\ \Sigma_{s_1}p_{s_0t_0s_1}\int_{-\infty}^{\infty}\phi[a_1 - \tfrac{1}{2}(a_0 + b_0), \tfrac{1}{2}\sigma_G^2]g_{s_1a_1C}(z_j)da_1dCdb_0da_0, \qquad 14.$$

where $g_{s_1a_1C}(z_j)$ is the probability density function of z_j conditional on the oligogenotype s_1, the polygenotype a_1, and the common environment C. Suppose $g_{s_1a_1C}(z) = \phi(\mu + C - z, \sigma_R^2)$, where μ depends upon s_1 and a_1. Now $\int_{-\infty}^{\infty}\phi(C, \sigma_C^2)$ $\prod_{j=1}\phi(\mu + C - z_j, \sigma_R^2)dC$ is equal to the ordinate of an n-variate multinormal distribution with a variance matrix whose diagonal elements are all $\sigma_C^2 + \sigma_R^2$, and whose off-diagonal elements are all σ_C^2. It follows that the likelihood expression 14 is identical to the one that would be obtained if expression 13 were modified to allow the joint phenotypic distribution of the children, conditional on their genotypes, to be multinormal with this same variance matrix, i.e. in this special case (76) the model allows for a common sibling environmental correlation equal to $\sigma_C^2/(\sigma_C^2 + \sigma_R^2)$. Thus the parameter σ_C^2 can meaningfully take on negative values, provided it is greater than $-\sigma_R^2/2$.

SPECIFICATION OF THE PHENOTYPIC DISTRIBUTIONS When the phenotypic trait z is quantitative, it is reasonable to assume that after transformation, if necessary, $g(z)$ is a normal distribution. Thus we can put in expression 11, $g_i(z) = \phi(\mu_i - z, \sigma_R^2)$; in expression 12, $g_G(z) = \phi(G + \mu - z, \sigma_R^2)$; in expression 13, $g_{iG}(z) = \phi(G + \mu_i - z, \sigma_R^2)$; and in expression 14, $g_{iGC}(z) = \phi(G + C - z, \sigma_R^2)$. Thus, the parameters of $g(z)$ are μ_i, the mean of the ith oligogenotype, or the overall mean μ for the pure polygenic model, and the environmental variance σ_R^2. Of course it is also possible to let the environmental variance σ_R^2 depend upon genotype, but this increases the number of parameters. Similarly there is no theoretical difficulty in letting the means and/or variances depend upon other characteristics of the individual, e.g. sex, age, parity, or specific environment.

When z is qualitative, $g(z)$ is a multinomial distribution. In the simplest case, a monogenic model of two alleles A and a at an autosomal locus with z a dichotomy (0 or 1), $g(z)$ takes on one of six values, dependent on just three penetrance parameters. If $g_{Aa}(z) = g_{AA}(z)$, we say that A is dominant to a, or equivalently, that a is recessive to A. If $g_{Aa}(1) = g_{AA}(1) = 1$, where $z = 1$ indicates having a disease, we say the disease is due to a fully penetrant dominant gene; then if $g_{aa}(1) = 0$ there are no sporadic cases. Conversely, if $g_{aa}(1) = 1$ and $g_{AA}(1) = g_{Aa}(1) = 0$, the disease is due to a fully penetrant recessive gene with no sporadic cases. For diseases that have a variable age of onset, models in which $g(z)$ is dependent on age have been proposed (35).

For the purely polygenic model, we can let $g(z)$ be defined by a set of risk functions, as described in Section 2. For a dichotomy, we let $g_G(1) = \phi[(G - \theta)/\sigma]$, $g_G(0) = 1 - g_G(1)$; for a trichotomy (96), we let $z = 0$, 1, 2 and

$$g_G(2) = \phi[(G - \theta_2)/\sigma],\quad g_G(1) = \phi[(G - \theta_1)/\sigma] - \phi[(G - \theta_2)/\sigma],$$
$$g_G(0) = 1 - \phi[(G - \theta_1)/\sigma], \qquad 15.$$

which is depicted in Figures 1 and 2; the extension to any polychotomy is clear. Note that, although in principle both θ and σ may depend on z, if the classes of z correspond to differences in disease severity it is probably only meaningful to let θ do so, since the curves $\Phi[(G - \theta_1)/\sigma_1]$ and $\Phi[(G - \sigma_2)/\sigma_2]$, as functions of G, cross if $\sigma_1 \neq \sigma_2$. Sex and/or age, however, could well be related to σ.

In the case of the mixed model we let θ (and/or σ) depend upon the oligogenotype. A simple way of doing this is to define parameters μ_i for the oligogenotypes and let, for a dichotomy,

$$\left.\begin{array}{l}\text{in expression 13, } g_{iG}(1) = \Phi[(G + \mu_i - \theta)/\sigma] \\ \text{in expression 14, } g_{iG}(1) = \Phi[(G + \mu_i + C - \theta)/\sigma]\end{array}\right\} g_{iG}(0) = 1 - g_{iG}(1), \qquad 16.$$

with the extension to a polychotomy entailing subscripting θ, as in equation 15. But when z is solely qualitative, this leads to a redundancy in the parametrization:

without loss of generality we can assume, for example, $\Sigma_i \mu_i = 0$ or $\Sigma_z \theta_z = 0$. Also if z is solely qualitative, when replacing $g(z)$ in expressions 12, 13, and 14 by the above risk functions we can, as in Section 2, without loss of generality set $\sigma_G^2 = 1$. Suppose, however, that z is quantitative for some individuals but qualitative for others. For example, in a study of diabetes we may have the quantitative result of a glucose tolerance test on some individuals, and on others we may only know whether or not they are clinically affected. In this situation we may use a normal distribution for $g(z)$ when z is quantitative, and the risk function 16 when z is qualitative, and then σ_G^2 should be left in the model as a parameter. This is the same as assuming there is a total liability, distributed as a mixture of k normal distributions each with variance $\sigma_G^2 + \sigma^2$, and an individual is clinically affected if and only if his total liability is greater than θ; this liability has variance $\tau^2 + \sigma^2$, where $\tau^2 = \sigma_G^2 + \sigma_C^2 + \Sigma_i \psi_i \mu_i^2 - (\Sigma_i \psi_i \mu_i)^2$ and is linearly related to the quantitative measures z with correlation $\tau^2/\sqrt{(\tau^2 + \sigma^2)(\tau^2 + \sigma_R^2)}$ (76).

NONRANDOM MATING As is seen in Section 5, we can allow for consanguinity between the parents by considering the nuclear family to be part of a larger pedigree structure. Models for assortative mating, on the other hand, have not been well developed for the analysis of nuclear families. In the case of oligogenic models, assortative mating can be allowed for by making the genotypic frequencies ψ_{t_0} dependent on s_0 in expression 11; the difficulty is how best to do this without introducing many more parameters. One possibility, for which some equilibrium theory has been developed in the monogenic case (60), is to assume that the probability of each of the k^2 mating types $s \times t$ is given by $\psi_s\ \psi_t(1 + \alpha\delta_{st})$. Other models for assortative mating in the monogenic case have also been considered (56,98,113). In the case of the polygenic models, it should be possible to incorporate the results of Section 3, but this has not been done; the model shown in Figure 3 is the general model that can allow for a wide variety of possibilities.

Testing Hypotheses and Parameter Estimation

If we have data on a random sample of nuclear families from some defined population, we can use the overall likelihood, i.e. the product of the likelihood for each family, to test any specified hypothesis by means of the asymptotic properties of the likelihood ratio criterion, analogous to the tests indicated in Section 3. Then, when we have found a parsimonious hypothesis that fits the data, the likelihood maximized under that hypothesis gives maximum likelihood estimates of all the necessary parameters. We do not dwell here on statistical and computational details of these procedures, but rather on the more fundamental questions of which likelihood should be used and what are the relevant hypotheses to test.

CONDITIONAL LIKELIHOODS Although we should naturally wish to start with the most general model possible, and hence use the likelihood 14, for practical reasons we are currently limited to either oligogenic models (35) or a mixed model in which the major gene part is monogenic (76). As indicated previously, however, a choice has to be made between basing tests and parameter estimation on the complete likelihoods as developed above, or on the likelihood of the siblings' phenotypes conditional on the parents' phenotypes. Traditionally tests of genetic hypotheses in the simpler cases have been largely based on this latter distribution, though not by means of the likelihood ratio criterion (101). For oligogenic models this conditional likelihood for a family is equal to expression 11 divided by expression 10; for the other models it can similarly be obtained by dividing each of the likelihoods 12, 13, or 14 by the corresponding likelihood for the two parents. Morton & MacLean (76) developed this conditional likelihood directly, rather than as the ratio of two likelihoods.

Go et al (43) compared these two likelihoods in a simulation study of monogenic models and found that parameter estimation is appreciably more efficient when the unconditional likelihood is used. Both likelihoods have been simulated to study the robustness of models to detect major genes (43, 68), but the differences found between the two studies depend more on the precise hypotheses tested, rather than on whether a conditional or an unconditional likelihood was used.

CHOICE OF HYPOTHESES FOR TESTING It is an unfortunate fact that any sample of data from a continuous distribution can be monotonically transformed such that it could reasonably have come from a normal distribution. Nevertheless, when a set of family data clearly fits a bimodal distribution, one is immediately led to think that the bimodality is probably due to an underlying dichotomy, whether genetic or environmental. Therefore it is profitable to consider, for oligogenic models, the different ways of dividing the genotypes into two genetically meaningful sets.

Consider two alleles at an autosomal locus, *A* and *a*. There are three genotypic dichotomies possible, corresponding to a phenotypic equivalence of *AA* and *Aa*, i.e. $g_{AA}(z) = g_{Aa}(z)$, a phenotypic equivalence of *aa* and *Aa*, or a phenotypic equivalence of *AA* and *aa*. The first two of these imply dominance and are hypotheses one would naturally test; however, they represent essentially the same genetic mechanism in that the difference between them is merely one of relabeling the alleles *a* and *A* instead of *A* and *a*. The third dichotomy is quite different and is unlikely to occur except for polymorphic traits that are correlated with Darwinian fitness. The fact that there are two distinct genetic mechanisms is summarized by saying that in diploid organisms two phenotypes and two alleles at one locus lead to two phenograms (12).

Two phenotypes and two alleles at each of two loci lead to 50 phenograms (46), of which five are considered by Defrise-Gussenhoven (19) to be hypotheses that deserve special attention. There is little doubt that the rigorous testing of two-locus hypotheses under more general models deserves more attention.

We consider in a little more detail the mixed model of one two-allele locus plus a polygenic component, since with it one can attempt to answer what possibly is the single most important question that concerns the inheritance of any character: is most of the genetic variation due to one locus, or are many gene loci necessarily involved (34)? Assume $g_i(z) = \phi(\mu_i - z, \sigma_R^2)$, the presence of a sibling environmental correlation, and that ψ_{AA}, ψ_{Aa}, and ψ_{aa} depend on a single parameter, q, the gene frequency. The model then depends upon seven parameters: q, μ_{AA}, μ_{Aa}, μ_{aa}, σ_G^2, σ_R^2, and σ_R^2. It follows that the null hypothesis $q = 0$ and $\mu_{AA} = \mu_{Aa} = \mu_{aa}$ is true only if there is no major gene effect, and rejection of this null hypothesis has been used to test for the presence of such a major gene effect (41, 91). Under this hypothesis, however, the model requires the data to come from a normal distribution, and so this test is sensitive to non-normality, and in particular to skewness (68). Although this problem can be alleviated by the use of a power transformation (67), it can be avoided (33, 43) at the expense of adding three extra parameters, transmission probabilities, to the models: $\tau_{AA\ A} = P$ (an AA individual transmits A to offspring); $\tau_{Aa\ A} = P$ (an Aa individual transmits A to offspring); $\tau_{aa\ A} = P$ (an aa individual transmits A to offspring). Functions of these three parameters replace $p_{s_0 t_0 s_1}$ in the likelihoods, as follows: $p_{s\,t\,AA} = \tau_{s\ A}\ \tau_{s\ A}$, $p_{s\,t\,Aa} = \tau_{s\ A}\ (1 - \tau_{t\ A}) + \tau_{t\ A}\ (1 - \tau_{s\ A})$, and $p_{s\,t\,aa} = (1 - \tau_{t\ A})(1 - \tau_{s\ A})$. Then, under the unrestricted model that allows the transmission probabilities to be anywhere from 0 to 1, we test the null hypothesis $\tau_{AA\ A} = 1$, $\tau_{Aa\ A} = ½$, $\tau_{aa\ A} = 0$. The advantage of this unrestricted model is that it includes the possibility that the transmission probabilities are all equal, τ say, corresponding to no genetic transmission from parents to offspring. In fact it is recommended to test the null hypothesis that the transmission probabilities are equal, though this also implies that the offspring are distributed in the proportions τ^2 AA:$2\tau(1 - \tau)Aa$:$(1 - \tau)^2$ aa; but this is no restriction if only two distinct phenotypic distributions exist, i.e. $g_{Aa}(z) = g_{AA}(z)$ or $g_{Aa}(z) = g_{aa}(z)$.

Nonrandom Sampling

In the genetic study of rare diseases it is inefficient to sample families randomly from the population; most such families will contain only unaffected individuals, yielding no information at all about the kind of genetic segregation that underlies the disease. For this reason families are sampled for study, or ascertained, via probands, affected individuals who bring the family to the notice of the investigator; every family in the sample contains at least one proband. This corresponds to sampling from a subset of the original sampling frame (the whole population), and so the likelihood needs to be modified accordingly.

Suppose all the probands are parents. If we base our analyses on the traditional likelihood conditional on the parents' phenotypes, no modification is necessary, since this likelihood automatically refers to the appropriate subset. But this likelihood does need to be modified when the families are ascertained through the offspring: we need the likelihood of the sibship conditional on at least one child being a proband.

Let $L(z_1 \ldots z_n)$ be the likelihood of the sibship conditional on the parents' phenotypes, and let $\pi(z_i)$ be the probability that the ith child is a proband, i.e. is ascertained, given that his phenotype is z_i. Then, assuming all children are independently ascertained, the likelihood we want is

$$\frac{L(z_1 \ldots z_n)\, P(\text{at least one proband} \mid z_1, z_2 \ldots z_n)}{P(\text{at least one proband})} =$$

$$\frac{L(z_1 \ldots z_n)\,\{1 - \Pi_i[1 - \pi(z_i)]\}}{1 - \Sigma_{z_1} \ldots \Sigma_{z_n} L(z_1 \ldots z_n)\,\{\Pi_i[1 - \pi(z_i)]\}}, \qquad 17.$$

with the summations replaced by integrations if z is qualitative. This is the traditional way of allowing for ascertainment via probands (73) and assumes that (*a*) the likelihood used is conditional on the parental phenotypes and (*b*) the values of $\pi(z_i)$ are known. If the values of $\pi(z_i)$ are unknown, but are functions of one or more parameters that need to be jointly estimated along with the other parameters of the model, then the appropriate likelihood would be the joint likelihood of the sibship phenotypes and the proband status of each child, conditional on the parents' phenotypes and on at least one child being a proband. The effect of this difference is to replace $1 - \Pi_i[1 - \pi(z_i)]$ in the numerator of equation 17 by $\Pi_i[\pi(z_i)]^{b_i}[1 - \pi(z_i)]^{1-b_i}$, where $b_i = 1$ if the ith individual is a proband, 0 otherwise. This is the approach taken by Elston & Yelverton (26, 35), except that they do not condition the likelihood on the parents' phenotypes.

The interpretation of π requires caution (106). Historically, the special case where every affected child has the same probability π of being a proband, especially as $\pi \rightarrow 0$ and $\pi = 1$, has received much attention: $\pi \rightarrow 0$ corresponds to the case where the probability that a family is ascertained is proportional to the number of affected children in the family; $\pi = 1$ corresponds to the case where the sample consists of every family in the population with at least one affected child, or a random sample of such families. Thus, for the likelihood 17 to be relevant, it is not necessary for π to be the probability that an affected individual brings his family into the sample for study; π is the probability that an affected individual brings his family into the (possibly conceptual) sampling frame from which a random sample of families is drawn for study.

The case where the unconditional likelihood is used, and either parents or offspring can be probands, can be considered as a special case of the general pedigree likelihood to which we now turn.

5 SEGREGATION ANALYSIS—PEDIGREES

If many nuclear families are pooled together for analysis, genetic heterogeneity from family to family may obscure the mode of inheritance. For this reason the study of single large pedigrees, which are more likely to be homogeneous, has for a long time been used by geneticists for detecting simple Mendelian diseases. It is only recently, however, that the rigor and generality of segregation analysis, as developed for nuclear families, has been introduced for pedigrees. This has come about from the development of a general method of expressing the likelihood of a pedigree (34), which can then be used for counseling and analysis in the same manner as described in Section 4. In this situation there is nothing comparable to the likelihood conditional on the parental phenotypes, since a large pedigree can contain individuals who are at the same time parents and offspring of other individuals in the pedigree. Therefore all tests should be based on unconditional likelihoods, though with modifications if necessary for nonrandom sampling. We limit our discussion in this section to construction of the likelihood and considerations of sampling.

Likelihood for a Random Set of Pedigrees

We shall now develop the general likelihood, under an oligogenic model, of any number of simple pedigrees, each of which contains no loops and starts with a single pair of original parents. In such pedigrees there are two types of individuals, and it is convenient to denote their phenotypes by two different letters. Persons related to someone in a previous generation have their phenotypes denoted x, and unrelated persons "marrying into" the pedigree have their phenotypes denoted y. In the case of the original parents of the pedigree, we arbitrarily use x for one of them and y for the other.

We use subscripts on subscripts to denote the generation, starting with 0 for the original generation. Let the phenotypes of the original parents of the i_0th pedigree be x_{i_0} and y_{i_0}; let the phenotype of their i_1th child be $x_{i_0 i_1}$, and let his or her spouse's phenotype be $y_{i_0 i_1}$; similarly let the phenotype of the i_2th child of this i_1th child be $x_{i_0 i_1 i_2}$, and that of his or her spouse be $y_{i_0 i_1 i_2}$, and so on (Figure 5).

Now let us rewrite expression 11, for the i_0th set of parents, by using this new notation. The result is

$$\Sigma_{s_0}\psi_{s_0}g_{s_0}(x_{i_0})\Sigma_{t_0}\psi_{t_0}g_{t_0}(y_{i_0})\Pi_{i_1}\Sigma_{s_1}p_{s_0 t_0 s_1}g_{s_1}(x_{i_0 i_1}), \qquad 18.$$

and if this is producted over i_0 the result is the likelihood of a set of nuclear families. Suppose now the offspring all have spouses; the likelihood of the set of nuclear families, together with the spouses of the children, is then

$$\Pi_{i_0}\Sigma_{s_0}\psi_{s_0}g_{s_0}(x_{i_0})\Sigma_{t_0}\psi_{t_0}g_{t_0}(y_{i_0})\Pi_{i_1}\Sigma_{s_1}p_{s_0 t_0 s_1}g_{s_1}(x_{i_0 i_1})\Sigma_{t_1}\psi_{t_1}g_{t_1}(y_{i_0 i_1}). \qquad 19.$$

If we now define the operator

$$\Gamma_j = \Pi_{i_j}\Sigma_{s_j}p_{s_{j-1}t_{j-1}s_j}g_{s_j}(x_{i_0 i_1}\ \ldots\ {}_{i_j})\Sigma_{t_j}\psi_{t_j}g_{t_j}(y_{i_0 i_1}\ \ldots\ {}_{i_j}),$$

where $P_{s_{j-1}t_{j-1}s_j} = \psi_{s_0}$ when $j = 0$, then expression 19 is identical to $\Gamma_0(\Gamma_1)$; and the likelihood for a set of simple pedigrees of any number of generations can be simply written as the sequence of operations $\Gamma_0(\Gamma_1(\Gamma_2(\Gamma_3. \ . \ . \ . \ )))$.

In the same manner, we can define Γ_j for the polygenic model (see expression 12) as

$$\Pi_{i_j}\int_{a_j}\phi[a_j - \tfrac{1}{2}(a_{j-1} + b_{j-1}), \tfrac{1}{2}\sigma_G^2]g_{a_j}(x_{i_0 i_1}\cdot\ .\ \cdot {}_{i_j})\int_{b_j}\phi(b_j, \sigma_G^2)g_{b_j}(y_{i_1 i_2}\ \ldots\ {}_{i_j}), \qquad 20.$$

where the symbol $\int_a$ indicates integration of everything following it with respect to a from minus infinity to plus infinity. The likelihood under the polygenic model is then also given by the sequence of operations Γ_j, except that now, when $j = 0$, $\phi[a_j - \tfrac{1}{2}(a_{j-1} + b_{j-1}), \tfrac{1}{2}\sigma_G^2]$ is replaced by $\phi(a_0, \sigma_G^2)$.

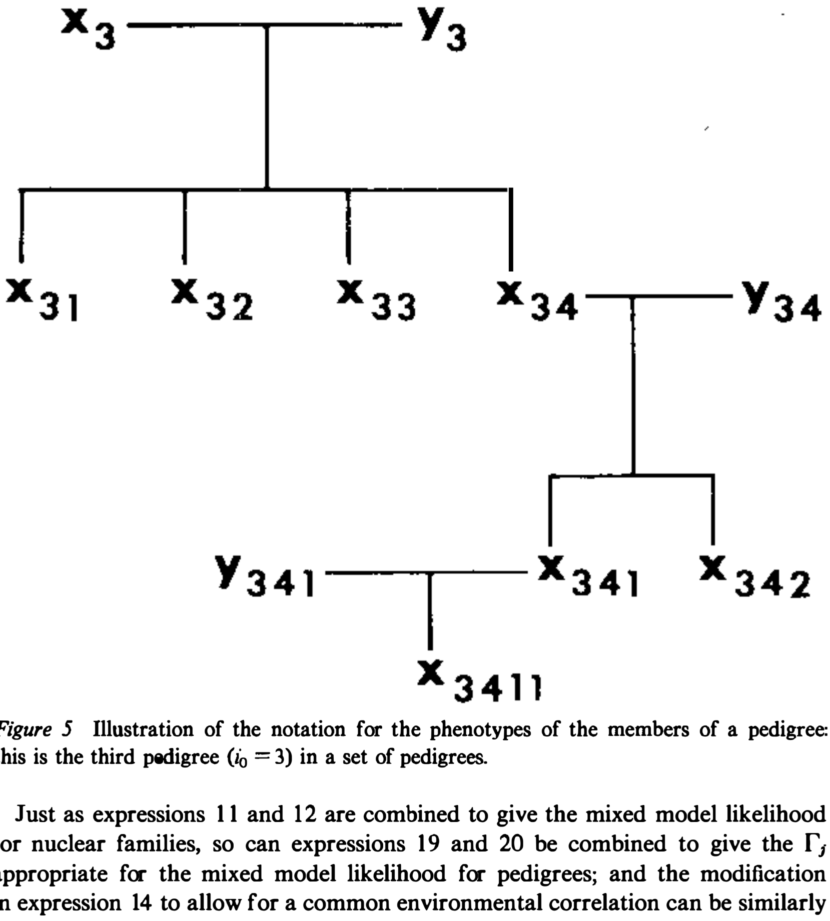

Figure 5 Illustration of the notation for the phenotypes of the members of a pedigree: this is the third pedigree ($i_0 = 3$) in a set of pedigrees.

Just as expressions 11 and 12 are combined to give the mixed model likelihood for nuclear families, so can expressions 19 and 20 be combined to give the Γ_j appropriate for the mixed model likelihood for pedigrees; and the modification in expression 14 to allow for a common environmental correlation can be similarly incorporated.

EXTENSIONS FOR PEDIGREES OF ARBITRARY STRUCTURE The likelihoods just developed can be easily extended to allow for twins and half-sibships (35). To allow for an arbitrary pedigree structure, however, it is simplest to define an algorithm by which the likelihood can be calculated; this has been done for the oligogenic model by Lange & Elston (61). Although the approach is different, the model that Lange et al (63) have used to estimate variance components from pedigrees is a generalization of the polygenic model, for arbitrary pedigrees, when the phenotype is normally distributed. Algorithms for the likelihood under other models, for completely arbitrary pedigrees, remain to be determined.

It should be noted that in all the likelihoods considered here and in Section 4, we can allow for missing phenotypes by setting $g(z) = 1$ if no value is available for individual Z. It follows that if we have a nuclear family in which we know

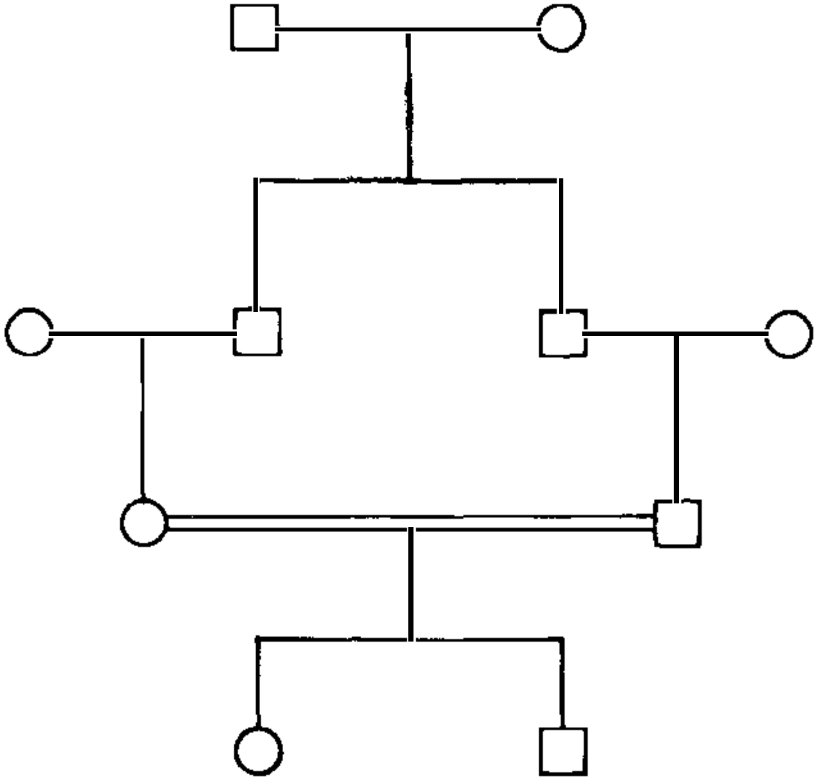

Figure 6 Pedigree containing a loop, in this case involving a mating between first cousins indicated by the double line.

the parents are related, this can be considered a special case of a larger pedigree. If the parents are first cousins, for example, the nuclear family can be considered a four-generation pedigree with a loop, but one in which no information is available on members of the first two generations (Figure 6). Hence consanguineous matings can be allowed for by including the necessary pedigree structure.

Nonrandom Sampling of Pedigrees

As indicated at the beginning of this section, there are advantages to the study of individual large pedigrees. If we have a single pedigree with m members, we can look upon the data as a sample of size one from an m-variate population; however, because under our model the $m \times m$ variance matrix is structured, depending on relatively few parameters, if m is large enough the asymptotic properties of the likelihood ratio criterion can still be used for testing hypotheses (33).

If the pedigree is ascertained via a single proband, the ascertainment probability π cannot be simultaneously estimated. One could base the analysis on the likelihood of the pedigree conditional on the proband's phenotype, but this is hardly necessary. Little bias will result if the pedigree is analyzed as a random pedigree, provided caution is used in interpreting the results. Thus, if a monogenic model is found to fit in a pedigree ascertained via a proband with a rare disorder, in a large enough pedigree the only biases will be in the estimates of the gene frequencies and other functions dependent on them; this has been demonstrated in the analysis of a 195-member pedigree ascertained via as many as four probands (33).

If pedigrees are pooled for analysis, with each pedigree containing at least one proband, we can (and unless the pedigrees are very large we should) estimate the ascertainment function. The solution to this problem given by Elston (26) is as follows. As before, let $\pi(z_1)$ be the probability that an individual with phenotype z_i is a proband, and let $b_i = 1$ if this individual is a proband, 0 otherwise. Suppose there are n individuals in the pedigree, with likelihood $L(z_1, \ldots z_n)$ appropriate for a random pedigree from the population. With this definition, and assuming all the ascertainments are independent, the joint likelihood is given by equation 17. This treatment, however, should be refined whenever possible to take account of the particular sampling scheme used to collect the data.

TWO-STAGE ASCERTAINMENT So far, we have assumed that a pedigree is included in the sampling frame if and only if it contains at least one proband. Frequently, however, a pedigree is analyzed because a trait appears to be familial in that pedigree. Here again, provided the pedigree is large enough, a genetic analysis that ignores that fact can still be meaningful, if carefully interpreted. For small pedigrees, however, it is essential to follow a precise sampling scheme and to allow for that method of sampling in the analysis (7). One possibility is to use a second stage of ascertainment, based on a definite criterion, to subsample from the sampling frame obtained after the initial ascertainment. Expressions for appropriate likelihoods in such cases have been derived (4).

6 LINKAGE ANALYSIS

The ultimate goal of all genetic analysis, whether statistical or biochemical, is to identify individual genes and what they do. In the absence of biochemical methods, we can go no further than the identification of individual genes and their positioning on the chromosomes. Although it is quite possible in certain situations for environmental causes to simulate segregation, it is difficult to conceive of environmental factors that simulate genetic linkage between two phenotypes—especially when one of the phenotypes is known to be monogenic; thus linkage analysis may provide the ultimate statistical proof of a gene's existence. In this section we briefly review the classical problem of linkage between two loci and the building of linkage maps.

Linkage Between Two Monogenic Phenotypes

As indicated in Section 2, linkage is the phenomenon whereby two genes at different loci on the same chromosome tend to be transmitted together, rather than independently. Thus the appropriate model is an oligogenic one with transmission probabilities that depend upon an individual's genotype at two loci. Consider two loci on the same pair of autosomal chromosomes, locus A with alleles A_i and locus B with alleles B_j ($i, j = 1, 2, \ldots$). The possible pairs of genes, one from each locus, are thus A_iB_j; these gametes, one from each parent, pair to form the genotypes of the offspring. Denote the genotype of an individual who received the gamete A_iB_j from one parent, and the gamete A_kB_ℓ from the other, A_iB_j/A_kB_ℓ; and let $\tau_{A_iB_j/A_kB_\ell\,A_mB_n}$ be the probability that an individual with this genotype transmits the gamete A_mB_n to his offspring. Then for two linked loci, each segregating in a Mendelian manner, the transmission probabilities are

$$\tau_{A_iB_j/A_kB_\ell\,A_mB_n} = \tfrac{1}{2}(1-\theta)(\delta_{im}\delta_{jn} + \delta_{km}\delta_{\ell n}) + \tfrac{1}{2}\theta(\delta_{im}\delta_{\ell n} + \delta_{km}\delta_{jn}), \qquad 21.$$

where θ, the recombination fraction, is the probability of an odd number of crossover events between the two loci at the time the gametes are formed (1). These transmission probabilities are multiplied together, one for each parent, and summed as necessary to obtain the elements $p_{s_{j-1}t_{j-1}s_j}$ appropriate for the oligogenic model likelihoods developed in Section 3. In general, θ is dependent upon the sex of the individual with genotype A_iB_j/A_kB_ℓ in equation 21; it can also be made dependent on the age (32) or any other characteristic of that individual.

As noted by Elston & Stewart (34), if this oligogenic model is used together with univariate phenotypic distributions $g(z)$, expressions 11 and 19 are relevant for a single trait genetically determined by the action of genes at two linked loci. The more usual linkage problem, however, entails two linked loci, each of which determines independent traits. In this situation the phenotypic distributions are bivariate, but can be expressed as the product of two univariate distributions, each dependent on the genotype at one locus only. It is usually also possible to express each probability ψ as the product of two probabilities, again with each dependent on the genotype at one locus only; if this is not so, linkage disequilibrium occurs. Ott (82) has illustrated the calculation of a pedigree likelihood in detail for the usual linkage problem. Although other methods have been proposed (97), the use of likelihood methods for the detection and estimation of linkage between two known monogenic phenotypes has a long history (45, 72) and is now generally accepted. It should be noted that it is not necessary for the phenotypic distributions to be qualitative for linkage analysis (62, 84).

Robust Methods of Detecting Linkage

Brief mention is made of several methods that share a certain degree of robustness and are useful for testing for linkage between two traits when the genetic mechanism that underlies one of the traits is unknown. All these methods capitalize upon the fact that linkage between two traits leads to an intrafamilial, and especially intrasibship, correlation between the two traits.

In Penrose's sib-pair method for two dichotomous traits (86), each sib-pair is classified into a 2 X 2 table according as to whether it is concordant or discordant for each of the two traits. If there is linkage we expect an excess of the pairs to be concordant or discordant for both traits, and a dearth of pairs concordant for just one trait. Penrose (87) also developed a test for the case where one trait is quantitative, but it appears to be erroneous (2).

Haseman & Elston (48) also developed a model for a quantitative trait linked to a known monogenic marker, with the possibility of utilizing parental information on the marker. Essentially, the test consists of testing for a correlation between the proportion of genes sib-pairs have i.b.d. at the marker locus and the absolute difference (or the square of the difference) between the sib-pairs' values for the quantitative trait. If there is linkage this correlation is expected to be negative; if no linkage exists it is expected to be zero. The extension of this model to sibships of size three, and the power of both tests, have been investigated; the evidence so far would suggest that it is valid to apply the sib-pair test to sibships of arbitrary size m, assuming the $m(m-1)/2$ sib pairs to which it gives rise are independent (2). This shows that a great increase in power is to be expected from this type of analysis with increased sibship size. An alternative approach to the same problem has been developed by Hill (50), and a nonparametric method of detecting linkage with data on up to three generations has been developed by Smith (103).

It is clear that these tests will have little power to detect linkage if the locus that controls the quantitative trait contributes little to the variance of the trait in the sample. Since it is a general principle that selecting the sample on the basis of one trait alone cannot invalidate a linkage analysis, selection of an appropriate sample is to be encouraged. Thus Elston et al (30) used the Haseman-Elston test on dizygotic twin pairs in which at least one twin was affected with a rare disease. By giving affected individuals a phenotypic value of unity, and unaffected individuals a value of zero, the Haseman-Elston test reduces to testing whether or not the proportion of genes a twin pair have i.b.d. at the marker locus is the same for concordant as for discordant pairs. If there is linkage the proportion is expected to be less than a half for discordant pairs, and greater than a half for concordant pairs.

Chromosome Mapping

Once linkage is detected between several pairs of loci, the next problem is to construct a map that specifies the locations of these loci. Such a map, inferred from recombination frequencies between the pairs of loci, is called a genetic map. The distance between two loci on the genetic map is the map distance (d), defined as the mean number of crossover events between them. Map distances have the property of additivity whereas recombination fractions do not; therefore we need a mapping function that specifies the relationship between d and θ. Different assumptions give rise to different mapping functions, but the criterion for judging which is better must depend on which fits the data better. Most mapping functions proposed so far differ in the extent to which they allow for interference, i.e. the extent to which a crossover event tends to suppress the occurrence of other crossover events in its vicinity (1). Rao et al (93) have introduced the general mapping function

$$d=[p(2p-1)(1-4p)\ln(1-2\theta)+16p(p-1)(2p-1)\tan^{-1}2\theta + 2p(1-p)(8p+2)\tanh^{-1}2\theta+6(1-p)(1-2p)(1-4p)\theta]/6, \qquad 22.$$

which involves one unknown parameter p, estimable under certain assumptions; $p=1$ yields Haldane's function (44), $p=½$ yields Kosambi's (57), $p=¼$ yields that of Carter & Falconer (9), and $p=0$ corresponds to complete interference ($d=\theta$). From the overall data on male meiosis, p was estimated as $\hat{p}=0.351$ with an approximate standard error of 0.007 (93). Assuming positive interference within a chromosome arm, one obligatory crossover event, and no interference across the centromere, Sturt (107) proposed the mapping function

$$\theta=½\{1-(1-d/L)\exp[-d(2L-1)/L]\}, \qquad 23.$$

where L is the total genetic length of the chromosome arm; this tends to Haldane's function (44) for large L. The underlying assumption that there is no interference across the centromere is doubtful (59), as is also the supposition of an obligatory crossover per chromosome arm. Also, although Carter & Falconer (9) have argued that the interference in Kosambi's function (57) is much too low, equation 23 yields even less interference.

For a given set of families with recombination data on a pair of loci, the joint likelihood function $L(\theta)$ is traditionally converted into the lod score (72), $z(\theta)=\log[L(\theta)/L(½)]=\log\{L[\theta(d)]/L(½)\}$, displaying d as a parameter that can be estimated by maximizing z. The problem of mapping n loci whose map locations (unknown) are $w_1, w_2 \ldots w_n$ can be solved as follows (75). Choose a set of starting values for $w_1 \ldots w_n$ and then for every observed pair of loci, convert the map distance $d=|w_i-w_j|$ into θ by using an appropriate mapping function and calculate the lod score for this value of θ. Let the total of all such lod scores, summed over all observed pairs of the n loci, be $z(w_1 \ldots w_n)$; similarly, a small increment (Δ) in w_i would give rise to $z(w_1 \ldots w_i+\Delta \ldots w_n)$. Thus by maximizing the total lod score over all pairs of loci, we obtain the maximum likelihood estimates of the n map locations relative to each other. Morton (75) also indicates a treatment when θ is dependent on sex.

7 CONCLUSION

At the outset of this review, basic genetic models have been described that can be used for deriving probabilities relevant for genetic counseling, for those situations in which the genetic mechanism that underlies a phenotypic trait is known. This has been followed by more general models, under which specific environmental and genetic hypotheses can be tested and relevant parameters can be estimated. We have concentrated on the model and on deriving the appropriate likelihood in each situation, since likelihood theory provides a firm basis, albeit asymptotic, for testing hypotheses and estimation.

We have for the most part avoided discussing algorithms and computational methods, whether for calculating the probabilities relevant for genetic counseling or for maximizing likelihoods over sets of parameters. This has been an active area of research (e.g. 3, 6, 8, 49, 55, 83, 104, 108) and will undoubtedly continue to be so in the future. For example, we need an efficient algorithm for calculating the likelihood of a large pedigree under the mixed model, and we need an efficient method of applying to pedigrees the models that have been developed for path analysis, discussed in Section 3.

Further research into the power and robustness of the various models, when used for the genetic analysis of data, is necessary. We also need to know more about how pedigree structure affects power; although sampling theory is a well-developed discipline, very little rigorous research has gone into determining the best way to sample pedigrees for the different kinds of genetic analysis.

At present the art of modeling and the theory of statistical analysis in human genetics are, as they should be, in advance of our computational capabilities. We can speculate that the development of new models and methods of analysis for multivariate traits will further help in the detection and identification of individual genes (110a). The most useful advances, however, can be expected to arise as a result of deficiencies noted in the application of current models in the analysis of bodies of real data, rather than as a result of model building divorced from data analysis.

ACKNOWLEDGMENTS

RCE was supported by a US Public Health Service Research Scientist Award (MH31732) and research grant (MH26721) from the National Institute of Mental Health, by a research grant (GM 16697) from the National Institute of General Medical Sciences, and by a contract (NHLBI–71–2243) from the National Heart, Lung and Blood Institute. DCR was supported by a US Public Health Service grant (GM17173) from the National Institute of General Medical Sciences.

We thank K. Lange for his helpful comments on the draft of this paper.

Literature Cited

1. Bailey, N. T. J. 1961. *Introduction to the Mathematical Theory of Genetic Linkage*. London: Oxford Univ. Press. 298 pp.
2. Blackwelder, W. K. 1977. *Inst. Stat. Mimeo Ser. No. 1114*. Chapel Hill, N.C.: Univ. N.C. 128 pp.
3. Bolling, D. R., Chase, G. A., Murphy, E. A. 1976. *Ann. Hum. Genet.* 40:25–36
4. Bucher, K. D. 1977. *Inst. Stat. Mimeo Ser. No. 1141*. Chapel Hill: Univ. N.C. 185 pp.
5. Campbell, M. A., Elston, R. C. 1971. *Ann. Hum. Genet.* 35:225–36
6. Cannings, C., Skolnick, M. H., deNevers, K., Stridharan, R. 1976. *Comp. Biomed. Res.* 9:393–407
7. Cannings, C., Thompson, E. A. 1977. *Clin. Genet.* In press
8. Cannings, C., Thompson, E. A., Skolnick, M. H. 1977. *Adv. Appl. Prob.* In press
9. Carter, T. C., Falconer, D. S. 1951. *J. Genet.* 50:307–23
10. Cavalli-Sforza, L. L., Feldman, M. W. 1973. *Am. J. Hum. Genet.* 25:618–37
11. Christian, J. C., Kang, K. W., Norton, J. A. Jr. 1974. *Am. J. Hum. Genet.* 26:154–61
12. Cotterman, C. W. 1953. *Am. J. Hum. Genet.* 5:193–235
13. Crittendon, L. B. 1961. *Ann. NY Acad. Sci.* 91:769–80
14. Crow, J. F., Felsenstein, J. 1968. *Eugen. Q.* 15:85–97
15. Crow, J. F., Kimura, M. 1970. *An Introduction to Population Genetic Theory*. New York: Harper & Row. 591 pp.

16. Curnow, R. N. 1972. *Biometrics* 28:931–46
17. Curnow, R. N. 1974. *Biometrics* 30:655–65
18. Curnow, R. N., Smith, C. 1975. *J. R. Stat. Soc. A* 138:131–69
19. Defrise-Gussenhoven, E. 1962. *Acta Genet. Stat. Med.* 12:65–96
20. Denniston, C. 1975. *Ann. Hum. Genet.* 39:89–104
21. Eaves, L. J. 1973. *Heredity* 30:199–210
22. Eaves, L. J. 1977. *J. R. Stat. Soc. A* 140:In press
23. Eaves, L. J., Last, K. A., Martin, N. G., Jinks, J. L. 1977. *Br. J. Math. Stat. Psychol.* 30:1–42
24. Edwards, J. H. 1960. *Acta Genet. Stat. Med.* 10:63–70
25. Elston, R. C. 1973. *Soc. Biol.* 20:276–79
26. Elston, R. C. 1973. *Hum. Hered.* 23:105–12
27. Elston, R. C. 1975. *Biometrika* 62: 133–48
28. Elston, R. C. 1977. *Biometrics* 33:231–33
29. Elston, R. C., Boklage, C. E. 1978. *Int. Twin Conf., 2nd, 1977, Washington, D.C.* New York: Alan Liss. In press
30. Elston, R. C., Kringlen, E., Namboodiri, K. K. 1973. *Behav. Genet.* 3: 101–6
31. Elston, R. C., Lange, K. 1976. *Ann. Hum. Genet.* 39:493–96
32. Elston, R. C., Lange, K., Namboodiri, K. K. 1976. *Am. J. Hum. Genet.* 28:69–76
33. Elston, R. C., Namboodiri, K. K., Glueck, C. J., Fallat, R., Tsang, R., Leuba, V. 1975. *Ann. Hum. Genet.* 39:67–87
34. Elston, R. C., Stewart, J. 1971. *Hum. Hered.* 21:523–42
35. Elston, R. C., Yelverton, K. C. 1975. *Am. J. Hum. Genet.* 27:31–45
36. Falconer, D. S. 1965. *Ann. Hum. Genet.* 29:51–76
37. Deleted in proof
38. Feldman, M. W., Cavalli-Sforza, L. L. 1975. *Ann. Hum. Biol.* 2:215–26
39. Fisher, R. A. 1918. *Trans. R. Soc. Edinburgh* 52:399–433
40. Fisher, R. A. 1921. *Metron* 1:1–32
40a. Geppert, S., Koller, S. 1938. *Erbmathematik; Theorie der Vererbung in Bevölkerung und Sippe.* Leipzig: Quelle und Meyer. 228 pp.
41. Gerrard, J. W., Rao, D. C., Morton, N. E. 1977. *Am. J. Hum. Genet.* In press
42. Gillois, M. 1964. *La Relation d'Identité en Génétique.* PhD thesis, Université de Paris. 294 pp.
43. Go, R. C. P., Elston, R. C., Kaplan, E. B. 1977. *Am. J. Hum. Genet.* In press
44. Haldane, J. B. S. 1919. *J. Genet.* 8:299–309
45. Haldane, J. B. S., Smith, C. A. B. 1947. *Ann. Eugen.* 14:10–31
46. Hartl, D. L., Maruyama, T. 1968. *J. Theoret. Biol.* 20:129–63
47. Haseman, J. K., Elston, R. C. 1970. *Behav. Genet.* 1:11–19
48. Haseman, J. K., Elston, R. C. 1972. *Behav. Genet.* 2:3–19
49. Heuch, I., Li, F. H. F. 1972. *Clin. Genet.* 3:501–4
50. Hill, A. P. 1975. *Ann. Hum. Genet.* 38:439–49
51. Holroyd, R. G. 1975. *Ann. Hum. Genet.* 38:379
52. Jacquard, A. 1972. *Biometrics* 28: 1101–14
53. James, J. W. 1971. *Ann. Hum. Genet.* 35:47–49
54. Jinks, J. L., Eaves, L. J. 1974. *Nature* 248:287–89
55. Kaplan, E. B., Elston, R. C. 1972. *Inst. Stat. Mimeo Ser. No. 823.* Chapel Hill: Univ. N.C. 83 pp.
56. Karlin, S., Scudo, F. M. 1969. *Genetics* 63:499–510
57. Kosambi, D. D. 1944. *Ann. Eugen.* 12:172–75
58. Krüger, J. 1973. *Humangenetik* 17: 181–252
59. Lalouel, J. M. 1977. *Heredity* 38:61–77
60. Lange, K. 1976. *Math Biosci.* 29:49–57
61. Lange, K., Elston, R. C. 1975. *Hum. Hered.* 25:95–105
62. Lange, K., Spence, M. A., Frank, M. B. 1976. *Am. J. Hum. Genet.* 28:167–73
63. Lange, K., Westlake, J., Spence, M. A. 1976. *Ann. Hum. Genet.* 39:485–91
64. Li, C. C. 1974. *Path Analysis—a Primer.* Pacific Grove, Calif.: Boxwood Press. 346 pp.
65. Li, C. C. 1976. *A First Course in Population Genetics.* Pacific Grove, Calif.: Boxwood Press. 631 pp.
66. Li, C. C., Sacks, L. 1954. *Biometrics* 10:347–60
67. MacLean, C. J., Morton, N. E., Elston, R. C., Yee, S. 1976. *Biometrics* 32:695–99
68. MacLean, C. J., Morton, N. E., Lew, R. 1975. *Am. J. Hum. Genet.* 27:365–84

69. McKusick, V. A., Ruddle, F. H. 1977. *Science* 196:390–405
70. Mendell, N. R., Elston, R. C. 1974. *Biometrics* 30:41–57
71. Moran, P. A. P., Smith, C. A. B. 1966. *Eugenics Laboratory Memoirs XLI.* London: Cambridge Univ. Press. 62 pp.
72. Morton, N. E. 1955. *Am. J. Hum. Genet.* 7:277–318
73. Morton, N. E. 1969. *Computer Applications in Genetics,* ed. N. E. Morton, pp. 129–39. Honolulu: Univ. Ha. Press. 167 pp.
74. Morton, N. E. 1974. *Am. J. Hum. Genet.* 26:318–30
75. Morton, N. E. 1978. *Human Gene Mapping 4.* Basel: Karger. In press
76. Morton, N. E., MacLean, C. J. 1974. *Am. J. Hum. Genet.* 26:489–503
77. Morton, N. E., Rao, D. C. 1977. *Genetic Epidemiology,* ed. N. E. Morton, C. S. Chung. New York: Academic. In press
78. Morton, N. E., Rao, D. C. 1978. *Yearbook of Physical Anthropology.* Philadelphia: The Wistar Inst. Press. In press
79. Murphy, E. A., Chase, G. A. 1975. *Principles of Genetic Counseling,* p. 70. Chicago: Year Book Med. Publ. 391 pp.
80. Nadot, R., Vaysseix, G. 1973. *Biometrics* 29:347–59
81. Nance, W. E., Corey, L. A. 1976. *Genetics* 83:811–26
82. Ott, J. 1974. *Am. J. Hum. Genet.* 26:588–97
83. Ott, J. 1977. *Ann. Hum. Genet.* 40:443–54
84. Ott, J. 1977. *Clin. Genet.* 11:In press
85. Palovino, J. 1976. *Book of Abstracts,* No. 507. Vol. V. Mexico City: Int. Congr. Human Genet.
86. Penrose, L. S. 1935. *Ann. Eugen.* 6:133–38
87. Penrose, L. S. 1938. *Ann. Eugen.* 8:233–37
88. Rao, D. C. 1978. *J. Ind. Soc. Agric. Stat.* In press
89. Rao, D. C., MacLean, C. J., Morton, N. E., Yee, S. 1975. *Am. J. Hum. Genet.* 27:509–20
90. Rao, D. C., Morton, N. E. 1974. *Am. J. Hum. Genet.* 26:767–72
91. Rao, D. C., Morton, N. E. 1977. *Hum. Genet.* 36:317–20
92. Rao, D. C., Morton, N. E., Elston, R. C., Yee, S. 1977. *Behav. Genet.* 7:147–59
93. Rao, D. C., Morton, N. E., Lindsten, J., Hulten, M., Yee, S. 1977. *Hum. Hered.* 27:99–104
94. Rao, D. C., Morton, N. E., Yee, S. 1974. *Am. J. Hum. Genet.* 26:331–59
95. Rao, D. C., Morton, N. E., Yee, S. 1976. *Am. J. Hum. Genet.* 28:228–42
96. Reich, T., James, J. W., Morris, C. A. 1972. *Ann. Hum. Genet.* 36:163–84
97. Renwick, J. H. 1971. *Ann. Rev. Genet.* 5:81–120
98. Scudo, F. M., Karlin, S. 1969. *Genetics* 63:479–98
99. Smith, C. 1971. *Am. J. Hum. Genet.* 23:578–88
100. Smith, C., Mendell, N. R. 1974. *Ann. Hum. Genet.* 37:275–86
101. Smith, C. A. B. 1956. *Ann. Hum. Genet.* 20:257–65
102. Smith, C. A. B. 1957. *Ann. Hum. Genet.* 21:363–73
103. Smith, C. A. B. 1975. *Ann. Hum. Genet.* 38:451–60
104. Smith, C. A. B. 1976. *Ann. Hum. Genet.* 40:37–54
105. Steinberg, A. G., Rushforth, N. B., Bennett, P. H., Burch, T. A., Miller, M. 1970. *Nobel Symp.* 13:237–64
106. Stene, J. 1977. *Biometrics* 33:523–27
107. Sturt, E. 1976. *Ann. Hum. Genet.* 40:147–63
108. Suarez, B. K., Fishman, P. M., Reich, T. 1978. *Ann. Hum. Genet.* In press
109. Vetta, A. 1974. *Nature* 263:316–17
110. Vetta, A., Smith, C. A. B. 1974. *Ann. Hum. Genet.* 38:243–48
110a. Weiss, V. 1976. *Gegenbaurs Morphol. Jahrb.* 122:875–81
111. Wilson, S. R. 1973. *Ann. Hum. Genet.* 37:189–204
112. Wilson, S. R. 1973. *Ann. Hum. Genet.* 37:205–16
113. Wilson, S. R. 1976. *Ann. Hum. Genet.* 40:225–29
114. Wright, S. 1921. *J. Agric. Res.* 20:557–85
115. Wright, S. 1934. *Genetics* 19:506–36
116. Wright, S. 1978. *Evolution and the Genetics of Populations,* Vol. 4. Chicago: The Univ. Chicago Press. In press