*The printing of this publication was supported by NIH General Medical Sciences Grant 5-R01-GM16697-08.

STATISTICAL METHODS FOR DETECTING GENETIC LINKAGE FROM SIBSHIP DATA*

by

William Cudd Blackwelder

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1114

April 1977

STATISTICAL METHODS FOR DETECTING GENETIC LINKAGE FROM SIBSHIP DATA

by

William Cudd Blackwelder

A Dissertation submitted to the faculty of the University of North Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics

Chapel Hill

1977

Approved by:

_____
Adviser

_____
Reader

_____
Reader

WILLIAM CUDD BLACKWELDER. Statistical Methods for Detecting Genetic Linkage from Sibship Data. (Under the Direction of ROBERT C. ELSTON.)

A method for detecting linkage between a locus for a quantitative trait and a marker locus with known inheritance, using data from independent sib pairs, was published by Haseman and Elston in 1972. Using asymptotic theory, a formula for the number of sib pairs required for the test to have specified power is developed; the formula is verified by computer simulations. It is found that, except for cases of very high heritability due to the hypothesized trait locus, the power of the test is quite low.

The sib-pair test is extended to the case of sibships of size three. The power of several statistics appropriate for this case is studied; included among the statistics is the sib-pair statistic on all possible sib pairs within the sibships of size three. It is found that these sib-trio statistics are comparable to each other in power and all are much more powerful than the sib-pair statistic for the same total number of sibs in independent sib pairs.

Based mainly on the results for sib trios, it is suggested that the sib-pair statistic on all possible sib pairs be used for cases of sibships of size four or greater and sibships of mixed sizes.

ACKNOWLEDGMENTS

I am grateful to Dr. Robert Elston for suggestion the topic of this dissertation and for his patient and helpful guidance throughout the course of this investigation. Thanks for their assistance are due also to the other members of my committee: Dr. P. K. Sen, Dr. James Grizzle, Dr. David Kleinbaum, and Dr. Michael Swift.

To my friends in the Biometrics Research Branch of the National Heart, Lung, and Blood Institute, I express my appreciation for their contributions. In particular, Dr. Max Halperin and Dr. James Ware have been quite helpful in both administrative and technical aspects of completing the dissertation, and Ms. Barbara Dyson has been most patient and cooperative in the arduous task of typing the manuscript.

Finally, I want to thank my parents, Samuel and Laura Will Blackwelder, and my sister, Frances Koon, for their encouragement and support of my academic interests throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION AND REVIEW OF LITERATURE

1.1. Introduction

Genetic linkage exists when genes at two different loci on the same chromosome remain together more often than would be expected by chance alone, assuming there is free recombination, when the gametes (sex cells) are formed. In the absence of linkage the recombination fraction, or proportion of recombinants (gametes in which the two genes on the same parental chromosome are separated), is 1/2; if there is linkage, the recombination fraction is less than 1/2. For a discussion of linkage in humans, a text in human genetics, such as McKusick's (1969), is recommended.

1.2. Phase of Linkage

It has been pointed out by many authors that if a mating is to give any information about genetic linkage between two loci, at least one of the partners must be heterozygous at both loci. For example, a mating of the type SsTt x sstt can be informative, but a mating such as SsTT x sstt gives no information about linkage. (Here S and s are the possible alleles, or distinct forms of the gene, at one locus, T and t are the possible alleles at another locus, and the capital letter indicates a dominant allele; if there is dominance, an individual with genetic composition (genotype) Ss, for example, has the same observable expression

of the gene (phenotype) as an individual with genotype SS.) The term "phase of linkage" refers to the arrangement of the alleles on the two homologous chromosomes (that is, members of the same pair) in the double heterozygote. The arrangement such that the alleles S and T are on one chromosome, while s and t are on the other, is arbitrarily referred to as the "coupling phase"; if the arrangement is St on one chromosome and sT on the other, the chromosomes are in the "repulsion phase." If two loci are linked, the probabilities of the various genotypes in the offspring of a double heterozygote depend on the recombination fraction and the phase of linkage.

Analysis of linkage data in humans has often involved observations on only two generations, or perhaps only one, if Penrose's sib-pair method was used. Unfortunately, with data from only two generations, one can never be certain of the phase of linkage in the parents, and hence of the correct probabilities of the various genotypes in the offspring. In large sibships one can sometimes be reasonably sure of the phase of linkage, according to the phenotypes of offspring; for example, if there are six offspring, such that either (i) five are recombinants and one is not or (ii) one is a recombinant and five are not, then almost certainly (ii) is the case if there is linkage. Another way to obtain information on the phase of linkage is through knowlege of grandparents' genotypes; that is, to have data on three generations. In many situations the information available for the first generation allows one to deduce the phase of linkage in the second, and hence the correct probabilities (as functions of the recombination fraction) for different individuals in the third generation.

1.3. Linkage on the X Chromosome

As indicated above, it can be quite advantageous in linkage studies to have information on pedigrees of at least three generations. This is especially true in the case of genes on the X chromosome, as in the work of Haldane and Smith (1947) on color-blindness and hemophilia.

Some researchers have employed a particular "three-generation method" for studying linkage relationships on the X chromosome. In this method the ratio of number of recombinants to total sample size in the third generation is used to test for linkage. Fraser (1968) discussed this method and gave references to the work of others who had employed it.

A related and quite important problem is the detection of sex linkage for a trait. Strictly speaking, sex linkage is not linkage in the sense of two genetic loci occurring close together on a chromosome; rather, the term refers to the situation in which a single gene is located on a "sex" chromosome, so that expression of the trait is associated with sex. In present usage the term "sex-linked" is essentially synonymous with "X-linked"; that is, located on the X chromosome.

An early attempt to demonstrate sex linkage appeared in a paper by Finney (1939), who used analysis of variance and correlation techniques to study possible sex linkage of a gene for height. In a recent paper Bock and Kolakowski (1973) gave evidence to support the hypothesis that an X-linked gene influences ability to visualize spatial relations. They corrected test scores for age and computed correlations for all possible pairings of parent and offspring by sex. The result was an excess of father-daughter correlation over that of father-son and an excess of mother-son correlation over that of mother-daughter; this

result is consistent with X-linkage but is difficult to explain from any simple environmental model.

Lasker and Wanke have suggested an interesting strategy in studying sex linkage from data on large numbers of relatives. They would classify pairs of individuals by degree of relationship and by type (depending on sexes of related antecedents) within degree. For example, male first cousins could be offspring of (i) two brothers, (ii) two sisters, or (iii) brother and sister; their probabilities of sharing an X chromosome or Y chromosome differ according to the type of relationship. Then, for example, if a quantitative trait is modified by a Y-linked gene, male first cousins of type (i) should on the average differ less in their trait values than other pairs of male first cousins.

1.4. Linkage Analysis for Qualitative Traits

1.4.1. Early Work

The problem of linkage analysis in humans has two parts, detection of linkage between two genetic loci and estimation of the recombination fraction between the two loci. Early attempts to detect linkage in humans consisted mainly of inspecting family data. The beginning of statistically valid methods of attacking the problem was apparently the work of Bernstein (1931), who analyzed data on offsping from matings of the type SsTt x sstt. Bernstein proposed the statistic $y = (a+d)(b+c)$ for detecting linkage, where a,b,c,d are observed numbers of offspring with phenotypes ST, St, sT, and st, respectively; using the normal approximation, one can compare y with its expected value under the hypothesis of no linkage.

Wiener (1932) developed a method similar to Bernstein's in which the proportion of crossovers (recombinants) in a group of offspring was estimated and compared with its expected value for arbitrary values of the recombination fraction. Both Wiener and Bernstein analyzed data on the MN and ABO blood groups, concluding there is no linkage between the two loci.

Hogben (1934) pointed out an algebraic error in Bernstein's derivation of an expected value formula, which invalidated the formula except for the special case of no linkage. Hogben extended Bernstein's method to include more mating types, and he pointed out the possibility of estimating the recombination fraction as well as detecting linkage.

Haldane (1934) extended the method to include many more mating types, setting down equations for estimating the recombination fraction as well as using statistics analogous to Bernstein's y for detecting linkage. He noted that this type of statistic is better (in the sense of efficiency or minimum variance) for recombination fractions 0 or 1/2 than for other values.

1.4.2. Fisher's u-Scores

It appears that not many researchers applied these early methods, at least not as originally formulated. There are several methods which have received more attention in the literature. One of the earlies of these was due to Fisher (1935a, 1935b), who proposed a system of u-scores to be used for various mating types. Fisher's u-scores are functions of the observed numbers a, b, c, d of offspring with phenotypes ST, St, sT, and st which Bernstein used. He proposed a set of three functions, one of them appropriate for any particular mating type and genetic assumption;

for example, assuming dominance and the mating

$$SsTt \times sstt,$$

the score assigned was

$$u_{11} = (a-b-c+d)^2 - (a+b+c+d).$$

Fisher's scores have the property that, under the conditions for which they were proposed, all have expected value zero under the hypothesis of no linkage. To test the hypothesis, Fisher proposed that the sum U of the u-scores for all mating types in a sample be compared with its expected value. The validity of the test depends on the assumption of normality, which is clearly not met in small samples. Fisher demonstrated that his statistic is (asymptotically) efficient in the sense of minimum variance for the case of no linkage, but not otherwise. He pointed out the usefulness of u-scores to estimate recombination fractions as well as to detect linkage.

In a series of papers Finney (1940, 1941a, 1941b, 1942a, 1942b, 1942c, 1943) expanded on Fisher's work and demonstrated the use of u-scores in a number of examples; Finney used the term "efficient scores." He derived the u-scores by approximating an individual's probability of having a certain phenotype with a linear function in $\theta = (1-2\lambda)^2$, where λ is the recombination fraction. Finney showed that even if the parental mating type is unknown, the efficient score is a weighted sum of the three distinct u-scores developed by Fisher, where the weights are probabilities of various possible matings.

Haldane (1947) derived moments of Fisher's u-scores beyond the second and suggested approximations which might be used as alternatives to the normal approximation.

Bailey (1951a) suggested computational simplifications in using u-scores. In other papers Bailey (1950, 1951b) discussed the application of u-scores when one of the traits is a partially manifesting abnormality; for example, in the case of dominance an individual with genotype Tt would be expected to have the abnormality, but in some cases he may not.

1.4.3. Penrose's Sib-Pair Method

Penrose (1935) proposed a test for detecting linkage between two genes for qualitative characteristics; his method uses sib-pair data, with no account taken of any knowledge of parental genotypes. He proposed the following 2x2 table, where "like" indicates the two sibs have the same phenotype for a characteristic and "unlike" indicates they have different phenotypes.

		Gene S: Like	Gene S: Unlike
Gene T	Like	n_{11}	n_{12}
Gene T	Unlike	n_{21}	n_{22}

The entries n_{11}, n_{12}, n_{21}, and n_{22} are observed numbers of sib pairs in the sample. This method can be applied for any number of alleles. To test for linkage, standard statistical methods for 2x2 tables are applicable, since the hypothesis of no linkage implies no association between "likes" and "unlikes." An estimate of the recombination fraction is available also. Penrose stated that in families with more than two sibs,

all possible pairs of sibs can be used, but he did not take into account the covariances between observations on different pairs in the same sibship.

In a later paper Penrose (1938) proposed a sib-pair method for detecting linkage when one or both of the characteristics is "graded"; that is, at least ordinal and possibly "continuous." This method will be discussed later.

In several later papers Penrose (1947, 1950, 1951, 1953) extended his original work and demonstrated the use of his sib-pair method for qualitative traits. He compared the sib-pair method with the u-scores of Fisher and Finney, indicating that, except for a rare recessive trait, the use of u-scores to detect linkage is more efficient if there is knowledge of parental genotypes.

Penrose's sib-pair method is relatively simple to apply, and it may be useful as a suggestive tool. It does have deficiencies; as previously stated, Penrose did not account for covariances between observations on different pairs within the same sibship, and he ignored any information on the genotypes of parents.

1.4.4. The Backward Odds Approach

Haldane and Smith (1947) modified a method of linkage analysis which had been proposed by Bell and Haldane (1937). Haldane and Smith employed a backward odds or probability ratio approach, in which they calculated the ratio $P(\lambda)/P(1/2)$, where $P(\lambda)$ is the probability of an observed set of phenotypes assuming a recombination fraction λ. They posed several possibilities for detecting and estimating linkage: an "inverse" probability or Bayesian approach, based on some *a priori* distribution for λ; maximum likelihood, if the curve of $P(\lambda)/P(1/2)$ as

a function of λ is sufficiently close to a normal density curve; or a "confidence interval" approach suggested especially for small samples, based on the observation that $\{P(\lambda)/P(1/2) \geq M\} \leq 1/M$ for any positive M. Haldane and Smith applied their methods to confirm the earlier conclusion of Bell and Haldane that genes for color-blindness and hemophilia, both located on the X chromosome, are linked.

The backward odds approach has been amplified in a review article by Smith (1953). He suggested this approach as an alternative to the efficient scores of Fisher and Finney, since use of efficient scores requires large samples to detect even moderately loose linkage, and the assumption of normality for the test statistic may not be warranted in small samples.

1.4.5. Lod Scores and Sequential Testing

Morton (1955) suggested the combination of samples via the sequential probability ratio test of Wald (1947). He devised a system of scores which he termed "lod" scores, or log odds. That is, if $P(\lambda)$ denotes the probability of a sample for recombination fraction λ, the analysis is based on log $(P(\lambda)/P(1/2))$. Morton showed that sequential testing required in some cases many fewer observations than either Fisher's u-scores or the backward odds test of Haldane and Smith. Employing the sequential method, Morton (1956) concluded there was linkage between a gene for elliptocytosis and the Rh blood type locus. Steinberg and Morton (1956) and Morton (1957) extended the method of lod scores to more complicated genetic situations than had been considered previously, for example multiple allelism.

1.4.6. A Bayesian Analysis for Linkage

Smith (1959) noted that Morton's sequential methods have advantages of simplicity and efficiency. However, he questioned in principle the idea of treating the problem as one of decision making; that is sampling sequentially until linkage, or its absence, has been sufficiently demonstrated. Smith suggested a Bayesian approach, combining Morton's lod scores with a prior distribution for the recombination fraction λ. His idea for a reasonable prior distribution was that $\lambda = 1/2$ with probability 21/22, since humans have 22 pairs of autosomes (non-sex chromosomes), and that otherwise λ is distributed uniformly between 0 and 1/2. This approach is quite similar to the backward odds method of Haldane and Smith (1947). In a later paper Smith (1968) suggested corrections to lod scores applicable in some situations, for example in the case of three-generational data.

The idea of a suitable prior distribution for λ has been refined somewhat. Renwick (1969) assumed that genetic loci are distributed uniformly throughout the genome; that is, the number of loci on a chromosome is proportional to the length of the chromosome (at the metaphase stage of cell division). For example, the prior probability that two autosomal loci are both on chromosome 1 is taken as $(1/11)^2$, since that chromosome's length is about 1/11 of the total length of the 22 autosomes. Then the prior probability that two loci are both on some unspecified autosome is about 1/18.5, rather than 1/22 as assumed by Smith. It should be noted that in this discussion all loci have been assumed to be autosomal; that is, not on the X or Y chromosomes, the "sex" chromosomes.

An interesting application of this Bayesian type of analysis appeared in a paper by Renwick (1971). He discussed the detection of linkage between a genetic locus and a chromosomal inversion, which was treated as a mutant allele at a single locus. Renwick calculated a posterior probability of .55 that the locus for erythrocytic acid phosphatase is on chromosome 2 in humans.

1.5. Linkage Analysis for Quantitative Traits

1.5.1. Motivation

Until recently, the sib-pair test of Penrose (1938) was the only method of linkage analysis proposed for the case when one or both traits are quantitative. Considerable experimental work has been done with plants and animals (see, for example, papers by Lowry and Schultz (1959) and by Niemann-Sorensen and Robertson (1961)), but relatively little attention has been paid to this area of linkage analysis in humans.

Thoday (1967) motivated the search for linkages between genes for quantitative or "continuous" traits and known marker loci by an appeal to the "major-gene" theory. That is, even though there may be several genes contributing to variation in a continuous trait, possibly there is one which is responsible for a large part of the variation; demonstration of linkages with known markers is one means by which such genes might be detected. Thoday did not formulate his ideas in precise mathematical terms, but rather argued from a hypothetical model for IQ and the MN blood group. He made the following assumptions: variation in the quantitative trait is largely accounted for by a single two-allele locus; a heterozygote for the quantitative trait tends to have a trait value intermediate to the average values for the two homozygotes; the trait

locus is closely linked to the marker locus: heterozygotes for the marker locus are distinguishable from homozygotes. Consider progeny from matings of two heterozygotes for the marker; Thoday asserted that there should be more within-family variation and less between-family variation in trait values among the progeny heterozygous for the marker than among those homozygous for the marker. The reason for this assertion, as Thoday demonstrated in a table of gametes and progeny, is that heterozygous progeny within a family should have more genotypic variation at the trait locus than homozygous progeny, whereas the genotypic distribution for the trait should vary more from family to family for homozygotes than for heterozygotes.

Hill (1974 and 1975) has developed an analysis of variance technique for detecting linkage which uses some of Thoday's ideas. Smith (1975) has suggested a method similar to Hill's.

1.5.2. Penrose's Method for Quantitative Traits

As mentioned previously, Penrose (1938) proposed a sib-pair method applicable for linkage analysis when one or both of the traits is quantitative. In Penrose's notation, the statistic used to test the hypothesis of no linkage is

$$\phi = \frac{n\,\Sigma(g_1-g_2)^2(h_1-h_2)^2}{\Sigma(g_1-g_2)^2\Sigma(h_1-h_2)^2} - 1,$$

where n is the number of sib pairs, $(g_1-g_2)^2$ and $(h_1-h_2)^2$ are the squared differences in observed values of the respective traits for a pair of sibs, and the summations are over all sib pairs. Under the hypothesis of no linkage, ϕ has expectation zero and, for large samples, estimated variance (according to Penrose) $\frac{V_\phi}{n-1}$,

where

$$V_{\phi} = \frac{n^2 \Sigma(g_1-g_2)^4 \Sigma(h_1-h_2)^4}{[\Sigma(g_1-g_2)^2 \Sigma(h_1-h_2)^2]^2} - 1.$$

It is unclear what assumptions Penrose made and how he derived this result, for the expression for the variance of ϕ appears to be in error. An asymptotically consistent estimate of the variance of ϕ is

$$\frac{n\{\Sigma(g_1-g_2)^4 - \frac{1}{n}[\Sigma(g_1-g_2)^2]^2\}\{\Sigma(h_1-h_2)^4 - \frac{1}{n}[\Sigma(h_1-h_2)^2]^2\}}{\{[\Sigma(g_1-g_2)^2]^2 - \Sigma(g_1-g_2)^4\}\{[\Sigma(h_1-h_2)^2]^2 - \Sigma(h_1-h_2)^4\}}$$

Penrose's methods, sometimes with modifications such as adjusting for observed correlations between trait values, have been applied in several studies. Kloepfer (1946) searched for linkages among 19 traits (171 pairs of traits). Brues (1970) studied possible linkages between body build and several other characteristics. Howells and Slowey (1956) analyzed 37 traits (666 pairs of traits) for linkage; they used data on only 75 pairs of sibs, whereas Penrose had recommended 100 sib pairs as a minimum number for study. Linkages were not conclusively demonstrated in any of the above studies, although there were some suggestive results. One difficulty is that one expects by chance alone some significant results from a large number of statistical tests.

1.5.3. Jayakar's Methods

Jayakar (1970) proposed methods of linkage analysis for a quantitative trait and a qualitative marker. He assumed that in the vicinity of

a two-allele marker locus there is a single two-allele locus which affects the quantitative trait. Jayakar derived his methods by considerations of (i) variances of trait values and (ii) covariances between trait values for sibs. He assumed that an individual's trait value, given his genotype at the trait locus, is distributed normally with variance σ^2, where σ^2 includes variance due to the environment and any variance due to the remainder of the genome.

Jayakar admitted that his methods have drawbacks, and he considered his paper a first step in attacking the problem. One drawback in applying his results to human data is that the suggested method for detecting linkage requires at least two individuals in a family with one marker genotype and one individual with another genotype. He does claim, by citing results of Monte Carlo trials, that the method is useful in detecting linkage. Another difficulty is that the variance σ^2 is not estimable directly; hence the recombination fraction, which Jayakar expresses as a function of several variances (including σ^2), is not estimable unless σ^2 is assumed known. Even in this case the resulting variance of the recombination fraction estimate tends to be quite large.

1.5.4. General Pedigree Analysis

Elston and Stewart (1971) formulated a general expression for the likelihood of a set of pedigree data, which might typically cover four or more generations. They allowed for various genetic models, including linkage between two autosomal loci; the traits considered can be qualitative or quantitative. They proposed methods for using the likelihood to test genetic hypotheses. In another paper Elston and Stewart (1973) presented likelihood formulations appropriate for analyzing the genetics

of traits in experimental animals; for this case they assumed that the pedigrees were of a particular type.

It is interesting to note that the likelihood expressions are practically identical for the cases of (i) two linked loci which together determine a single trait and (ii) two linked loci, each of which determines a separate trait. In (i) a univariate distribution for the single trait value is involved; in (ii) the univariate distribution must be replaced with a bivariate distribution.

Haseman and Elston (1972), in their maximum likelihood procedure for estimating linkage, applied this type of pedigree analysis for the special case in which each pedigree consists of a sib pair, with parental information on a marker locus taken into account.

1.5.5. Regression and Maximum Likelihood Methods of Haseman and Elston

Haseman and Elston (1972) considered the problems of detecting and estimating linkage between a two-allele locus for a quantitative trait and an m-allele marker locus. They developed a method of analysis for sib-pair data which, unlike Penrose's sib-pair method, takes knowledge of parental genotypes into account. The methods they proposed are particularly applicable when little is known about the genetics affecting the quantitative trait. They assumed random mating, linkage equilibrium (that is, no association in the population between phenotypic frequencies for the linked loci), and an additive model for an individual's trait value with genetic and environmental components. It might be noted that these assumptions are no more restrictive than those of most authors; in many papers the genetic assumptions have not been stated explicitly.

The detection of linkage is based on the squared difference of the trait values for the two sibs in a family. This method will be discussed extensively and extended to cases of sibships larger than two in the succeeding chapters of this dissertation. For estimation of the recombination fraction, Haseman and Elston suggested a maximum likelihood method using the absolute values of differences, rather than the squared differences, and assuming normality for the trait values.

1.6. Heterogeneity of Recombination Fractions

Until recently, in most studies a common recombination fraction for all individuals has been assumed. However, it is now well established that there is a sex difference, and other differences in recombination fraction can perhaps also occur, due to factors such as age, race, and geographic region. Some of the statistical methods for linkage analysis are sufficiently general to allow for heterogeneity of recombination fractions, and some researchers have apparently detected such heterogeneity in specific cases. Morton (1956) concluded from his data on elliptocytosis and Rh blood type that there were different recombination fractions in different families. Smith (1963) proposed a method for detecting heterogeneity by comparing likelihoods; applying his method to Morton's data, Smith also concluded that there were different recombination fractions. Renwick and Schulze (1965) analyzed data on linkage between the loci for the nail-patella syndrome and the ABO blood group; they discussed sex and age differences in recombination fractions. Weitkamp (1973) reviewed the phenomenon of heterogeneity of recombination fractions and summarized the findings from a number of specific studies.

CHAPTER II

THE SIB-PAIR LINKAGE TEST OF HASEMAN AND ELSTON AND ITS POWER

2.1. Model

Haseman and Elston (1972) proposed a method of detecting linkage between a locus for a quantitative trait and a marker locus for which the mode of inheritance is known, such as the ABO blood group locus. The development they give is for the case of a single two-allele trait locus, although under their conditions the method is valid for k two-allele loci, all linked to the same marker. The method is based on data for independent sib pairs and takes into account information on parental phenotypes for the marker, if available. The assumptions they make include random mating, linkage equilibrium and no selection (which result in no association in the population between phenotypic frequencies for the two linked loci), no epistasis (that is, no interaction between the trait and marker loci), and an additive model as follows.

Suppose we have a random sample of n sib pairs, and let x_{ij} ($i = 1, 2$; $j = 1, \ldots, n$) denote the value of the quantitative trait for the ith member of the jth sib pair. The model is of the form

$$x_{ij} = \mu + g_{ij} + e_{ij}, \tag{2.1}$$

where μ is an overall mean and g_{ij} and e_{ij} are "genetic" and "environmental" effects, respectively. (The "environmental" effect includes any contribution to x_{ij} not due to the single two-allele trait locus under

consideration, and may in fact include a genetic component due to loci unlinked to the marker.)

Denote the alleles at the trait locus by T and t, with respective gene frequencies p and q = 1-p; T is the dominant allele if there is complete dominance. Then the genetic effect g_{ij} is as follows.

$$\begin{aligned} g_{ij} &= a \text{ for a TT individual,} \\ &= d \text{ for a Tt individual,} \\ &= -a \text{ for a tt individual} \end{aligned}$$

For the case of no dominance, d = 0; the case d = a is referred to as "complete dominance." Under random mating, the genetic variance σ_g^2 (variance of the genetic effect over the entire population) is given by

$$\left.\begin{aligned} \sigma_g^2 &= \sigma_a^2 + \sigma_d^2, \\ \text{where} \quad \sigma_a^2 &= 2pq[a-d(p-q)]^2 \\ \text{and} \quad \sigma_d^2 &= 4p^2q^2d^2. \end{aligned}\right\} \qquad (2.2)$$

Hence the genetic variance σ_g^2 is the sum of an additive component σ_a^2 and a dominance component σ_d^2 (see, for example, Li (1955)).

Let $e_j = e_{1j} - e_{2j}$. It is assumed that e_j is distributed as a normal random variable with mean 0 and variance σ_e^2.

Knowledge of the sibs' and (possibly) their parents' phenotypes at the marker locus is used to estimate π_{mj}, the true proportion of genes identical by descent (i.b.d.) at the marker locus, for the jth sib pair. (In general, two genes are i.b.d. if each is a copy of the same gene in a common ancestor. For the model considered here, sibs have genes i.b.d. only through their parents; at a particular locus they have 0, 1, or 2 genes i.b.d.) Haseman and Elston estimate π_{mj} by

$$\hat{\pi}_j = f_{2j} + f_{1j}/2, \tag{2.3}$$

where f_{ij} $(i = 0, 1, 2)$ is the probability that the jth sib-pair, given their and their parents' phenotypes, have exactly i genes i.b.d. at the marker locus. We denote the true proportion of genes i.b.d. at the trait locus by π_{tj}. Since $f_{0j} + f_{1j} + f_{2j} = 1$, we see that the information on sibs' and parents' phenotypes is summarized in $\hat{\pi}_j$ and f_{1j}.

Let λ denote the recombination frequency between the quantitative trait locus and the marker locus (λ is assumed constant for a particular population) and define $\psi = \lambda^2 + (1 - \lambda)^2$. Also, define $Y_j = (x_{ij} - x_{2j})^2$, the squared difference between the trait values for the jth sib pair.

Now the conditional density of Y_j, given $\hat{\pi}_j$ and f_{1j}, can be written

$$P(Y_j|\hat{\pi}_j, f_{1j}) = \sum_{\pi_{tj}} P(Y_j|\pi_{tj}) \sum_{\pi_{mj}} P(\pi_{tj}|\pi_{mj}) P(\pi_{mj}|\hat{\pi}_j, f_{1j}),$$

or
$$P(Y_j|\hat{\pi}_j, f_{1j}) = \sum_{h=0,1,2} P(Y_j|\pi_{tj} = h/2) R_{hj},$$

where

$$R_{hj} = \sum_{k=0,1,2} P(\pi_{tj} = h/2|\pi_{mj} = k/2) P(\pi_{mj} = k/2|\hat{\pi}_j, f_{1j}), \qquad h = 0,1,2. \tag{2.4}$$

From (2.4) it follows that, for a general function $g(Y_j)$,

$$E(g(Y_j)|\hat{\pi}_j, f_{1j}) = \sum_{h=0,1,2} E(g(Y_j)|\pi_{tj} = h/2) R_{hj}, \tag{2.5}$$

where R_{hj} is defined as in (2.4). (Haseman and Elston derive (2.5) for the case $g(Y_j) = Y_j$.)

From Table 2.1, which is an augmented version of Table I of Haseman and Elston (1972), we can derive $P(Y_j|\pi_{tj})$ or $E(g(Y_j)|\pi_{tj})$, for $\pi_{tj} = 0, 1/2, 1$. (Assuming normality of e_j, the distribution of Y_j/σ_e^2, where σ_e^2 is the variance of e_j, for a particular type of sib pair is noncentral χ^2 with one degree of freedom; the noncentralities are given in the last column of Table 2.1.)

Table 2.2, which is Table IV of Haseman and Elston, gives the joint distribution of π_{mj} and π_{tj}. From it we get the probabilities $P(\pi_{tj} = h/2|\pi_{mj} = k/2)$ for h, k = 0, 1, 2. Noting that

$$P(\pi_{mj} = k/2|\hat{\pi}_j, f_{1j}) = f_{kj}$$

by the definition of f_{kj}, we have from (2.3) and Table 2.2 that the R_{hj} in (2.4) are given by

$$\left.\begin{aligned} R_{0j} &= \psi^2 + (1 - 2\psi)\hat{\pi}_j - 1/2(1 - 2\psi)^2 f_{1j} \\ R_{1j} &= 2\psi(1 - \psi) + (1 - 2\psi)^2 f_{1j} \\ R_{2j} &= (1 - \psi)^2 - (1 - 2\psi)\hat{\pi}_j - 1/2(1 - 2\psi)^2 f_{1j}. \end{aligned}\right\} \qquad (2.6)$$

We see from (2.6) and Table 2.1 that (2.5) can be rewritten in general as

$$E(g(Y_j)|\hat{\pi}_j, f_{1j}) = \alpha + \beta\hat{\pi}_j + \gamma f_{1j}, \qquad (2.7)$$

where α, β, γ are functions of the parameters σ_e^2, a, d, p, and ψ but do not depend on the values of $\hat{\pi}_j$ or f_{1j}. Letting $g(Y_j)$ be the ℓth moment about the origin, we write

$$E(Y_j^{\ell}|\hat{\pi}_j, f_{1j}) = \alpha_{\ell} + \beta_{\ell}\hat{\pi}_j + \gamma_{\ell} f_{1j}. \qquad (2.8)$$

TABLE 2.1

DISTRIBUTION OF Y_j GIVEN π_{tj}, AND NONCENTRALITY PARAMETERS

p = gene frequency of allele T; q = 1 - p

Sib Pair (ordered)	Y_j	$P(\text{sib pair}\mid\pi_{tj})$ $\pi_{tj}=0$	$\pi_{tj}=1/2$	$\pi_{tj}=1$	Noncentrality
TT-TT	e_j^2	p^4	p^3	p^2	0
tt-tt	e_j^2	q^4	q^3	q^2	0
Tt-Tt	e_j^2	$4p^2q^2$	pq	$2pq$	0
TT-Tt	$(a-d+e_j)^2$	$2p^3q$	p^2q	0	$(a-d)^2/\sigma_e^2$
Tt-TT	$(-a+d+e_j)^2$	$2p^3q$	p^2q	0	$(a-d)^2/\sigma_e^2$
Tt-tt	$(a+d+e_j)^2$	$2pq^3$	pq^2	0	$(a+d)^2/\sigma_e^2$
tt-tt	$(-a-d+e_j)^2$	$2pq^3$	pq^2	0	$(a+d)^2/\sigma_e^2$
TT-tt	$(2a+e_j)^2$	p^2q^2	0	0	$4a^2/\sigma_e^2$
tt-TT	$(-2a+e_j)^2$	p^2q^2	0	0	$4a^2/\sigma_e^2$

TABLE 2.2

JOINT DISTRIBUTION OF π_{mj} AND π_{tj}

	$P(\pi_{tj},\pi_{mj})$			
π_{tj} \ π_{mj}	0	1/2	1	Total
0	$\psi^2/4$	$\psi(1-\psi)/2$	$(1-\psi)^2/4$	1/4
1/2	$\psi(1-\psi)/2$	$(1-2\psi+2\psi^2)/2$	$\psi(1-\psi)/2$	1/2
1	$(1-\psi)^2/4$	$\psi(1-\psi)/2$	$\psi^2/4$	1/4
Total	1/4	1/2	1/4	1

Using (2.6), Table 2.1, and the assumption of normality with mean 0 for e_j, it follows from straightforward algebra that for the first two moments the parameters in (2.8) are

$$\left.\begin{aligned}
\alpha_1 &= \sigma_e^2 + 2\psi\sigma_g^2 + 2\psi(1-\psi)\sigma_d^2 \\
\beta_1 &= 2(1-2\psi)\sigma_g^2 \\
\gamma_1 &= (1-2\psi)^2\sigma_d^2
\end{aligned}\right\}$$

and

$$\left.\begin{aligned}
\alpha_2 &= 3\sigma_e^4 + 12\psi\sigma_e^2\sigma_g^2 + 12\psi(1-\psi)\sigma_e^2\sigma_d^2 \\
&\qquad + 4\psi pq[p(a-d)^4 + q(a+d)^4] + 8p^2q^2\psi^2(3a^4-6a^2d^2-d^4) \\
\beta_2 &= (1-2\psi)[12\sigma_e^2\sigma_g^2 + 4pq(p(a-d)^4 + q(a+d)^4) + 8p^2q^2(3a^4-6a^2d^2-d^4)] \\
\gamma_2 &= (1-2\psi)^2[6\sigma_e^2\sigma_d^2 - 4p^2q^2(3a^4-6a^2d^2-d^4)].
\end{aligned}\right\} \tag{2.9}$$

(Haseman (1970, Chapter VIII) has derived α_1, β_1, and γ_1.) From (2.8) we see that

$$\begin{aligned}
V(Y_j|\hat{\pi}_j, f_{1j}) = (\alpha_2-\alpha_1^2) + (\beta_2-2\alpha_1\beta_1)\hat{\pi}_j + (\gamma_2-2\alpha_1\gamma_1)f_{1j} \\
- \beta_1^2\hat{\pi}_j^2 - 2\beta_1\gamma_1\hat{\pi}_j f_{1j} - \gamma_1^2 f_{1j}^2.
\end{aligned} \tag{2.10}$$

If there is no dominance at the trait locus, then $\sigma_d^2 = 0$ and the regression of Y_j on $\hat{\pi}_j$ and f_{1j} is a function of $\hat{\pi}_j$ alone. Haseman and Elston base their linkage test on the simple regression of Y_j on $\hat{\pi}_j$, regardless of the assumption about dominance at the trait locus. (It will be seen in a later section that consideration of the regression of Y_j on both $\hat{\pi}_j$ and f_{1j} appears to contribute little extra to the

detection of linkage.) The test for linkage is to reject the null hypothesis of no linkage $H_0:\lambda = 1/2$ (equivalently, $\psi = 1/2$ or $\beta_1 = 0$), in favor of the one-sided alternative $H_A:\lambda < 1/2$ (equivalently, $\beta_1 < 0$), at the 100α% level, if $b/s_b < -z_{1-\alpha}$, where

$$\left.\begin{aligned} b &= \frac{\sum_{j=1}^{n} Y_j(\hat{\pi}_j - \bar{\hat{\pi}})}{\sum_{j=1}^{n} (\hat{\pi}_j - \bar{\hat{\pi}})^2} \\ s_b &= \left(\frac{\sum_{j=1}^{n} (Y_j - \bar{Y})^2 - b^2 \sum_{j=1}^{n} (\hat{\pi}_j - \bar{\hat{\pi}})^2}{(n-2)\sum_{j=1}^{n} (\hat{\pi}_j - \bar{\hat{\pi}})^2} \right)^{1/2} \\ \bar{\hat{\pi}} &= \frac{1}{n}\sum_{j=1}^{n} \hat{\pi}_j \ , \end{aligned}\right\} \qquad (2.11)$$

and $z_{1-\alpha}$ is the 100(1-α)% point of the standard normal distribution. (That is, a standard normal deviate will exceed $z_{1-\alpha}$ with probability α.) The test is then essentially the usual normal theory one-sided test for a zero regression coefficient. For convenience the normal distribution, rather than Student's t-distribution as proposed by Haseman and Elston, is used in determining the critical value of the test statistic; for reasonably large samples the difference is trivial.

2.2. Power-Asymptotic Theory

It is obvious from the definition of Y_j as $(x_{1j}-x_{2j})^2$ that Y_j has a non-normal distribution. As shown later in section 3.1, if e_j is normally distributed then the exact distribution of Y_j, given $\hat{\pi}_j$ and f_{1j}, is that of a mixture of noncentral χ^2 variates each with one degree

of freedom. However, it is shown below that the estimated regression coefficient b has a normal distribution asymptotically, given the sets of $\hat{\pi}_j$ and f_{1j}. Using this result, approximate expressions will be derived for power and for the sample size required to attain a certain power.

We now prove the asymptotic normality (conditional on the $\hat{\pi}_j$ and f_{1j}) of

$$\sum_{j=1}^{n} Y_j$$

and of the estimated regression coefficient b, defined in (2.11). All moments of Y_j, given $\hat{\pi}_j$ and f_{1j}, are finite. Then for any $\delta > 0$ we can write that $E[|Y_j-EY_j|^{2+\delta}|\hat{\pi}_j,f_{1j}] \leq L$ for some $L < \infty$. Then

$$w_n = \sum_{j=1}^{n} E[(|Y_j-EY_j|^{2+\delta})|\hat{\pi}_j,f_{1j}] \leq nL.$$

Similarly,

$$v_n = [\sum_{j=1}^{n} V(Y_j|\hat{\pi}_j,f_{1j})]^{(2+\delta)/2} \geq (Mn)^{(2+\delta)/2}$$

for some $M > 0$. Let $\rho_n = w_n/v_n$; then $\rho_n \leq nL/(nM)^{(2+\delta)/2}$, and we have $\lim_{n\to\infty} \rho_n = 0$. Hence it follows from Liapounoff's central limit theorem (see, for example, Feller (1971)) that

$$\sum_{j=1}^{n} Y_j$$

is asymptotically normal, conditional on the $\hat{\pi}_j$ and f_{1j}. (We can argue similarly that the unconditional distribution of

$$\sum_{j=1}^{n} Y_j$$

is asymptotically normal also.) A simple extension of the above argument,

replacing Y_j by $Y_j(\hat{\pi}_j - \bar{\hat{\pi}})$, shows that the numerator of b in (2.11) is asymptotically normal, assuming the $\hat{\pi}_j$ are not all equal. Then, since the denominator of b, given the $\hat{\pi}_j$, is a constant, we see that the conditional distribution of b is normal asymptotically.

Using the asymptotic normality of b, we can derive an approximate result for the power of the linkage test. Let $\underline{\hat{\pi}}$ and $\underline{f}_1$ be nx1 vectors of values of $\hat{\pi}_j$ and f_{1j}, respectively, and let H_A represent a specific alternative hypothesis about the value of λ. The power is given by

$$\Pr\{b/s_b < -z_{1-\alpha} | \underline{\hat{\pi}}, \underline{f}_1, H_A\},$$

or

$$\Pr\left\{ \frac{b - E(b|\underline{\hat{\pi}}, \underline{f}_1, H_A)}{\sqrt{V(b|\underline{\hat{\pi}}, \underline{f}_1, H_A)}} < \frac{-z_{1-\alpha}s_b - E(b|\underline{\hat{\pi}}, \underline{f}_1, H_A)}{\sqrt{V(b|\underline{\hat{\pi}}, \underline{f}_1, H_A)}} \,\middle|\, \underline{\hat{\pi}}, \underline{f}_1, H_A \right\}.$$

Given the alternative hypothesis, the expression on the left above is asymptotically distributed N(0,1), so we have the large-sample result

$$\text{Power} \cong \Pr\left\{ z < \frac{-z_{1-\alpha}s_b - E(b|\underline{\hat{\pi}}, \underline{f}_1, H_A)}{\sqrt{V(b|\underline{\hat{\pi}}, \underline{f}_1, H_A)}} \right\},$$

where z denotes a standard normal variate. Letting $1-\tau$ denote power, we have the equivalent expression

$$z_{1-\tau} \cong \frac{-z_{1-\alpha}s_b - E(b|\underline{\hat{\pi}}, \underline{f}_1, H_A)}{\sqrt{V(b|\underline{\hat{\pi}}, \underline{f}_1, H_A)}}. \qquad (2.12)$$

2.3. Derivation of Sample Size Formula

To determine approximate sample size required for a certain power, we replace the terms of (2.12) with appropriate asymptotic results, first in terms of the Y_j and then in terms of the $\hat{\pi}_j$ and f_{1j}. Throughout this section we will use u_n to denote a random variable, based on n observations, which is normally distributed with zero mean and finite variance; it should be clear from the context whether u_n is a function of the Y_j, given the $\hat{\pi}_j$ and f_{1j}, or u_n is a function of the $\hat{\pi}_j$ and f_{1j}. The notation $o_p(n^{\xi})$ will denote the nth member of a sequence of random variables $\{x_n\}$ such that x_n/n^{ξ} converges to zero in probability as $n\to\infty$ (see Pratt (1959)). We note that much ordinary algebraic manipulation holds for $o_p(n^{\xi})$; for example, we will use the facts that

$$n^{\xi_1} o_p(n^{\xi_2}) = o_p(n^{\xi_1+\xi_2}) \text{ for all real } \xi_1 \text{ and } \xi_2,$$

$$o_p(n^{\xi_1}) o_p(n^{\xi_2}) = o_p(n^{\xi_1+\xi_2}) \text{ for all real } \xi_1 \text{ and } \xi_2,$$

and

$$[1+o_p(n^{\xi})]^m = 1 + o_p(n^{\xi}) \text{ for } \xi < 0.$$

We can use the same type of argument used in the previous section for

$$\sum_{j=1}^{n} Y_j$$

to show that in general

$$\sum_{j=1}^{n} Y_j^k$$

is asumptotically normal, given the $\hat{\pi}_j$ and f_{1j}, for any real k. Then we can write

$$\sum_{j=1}^{n} Y_j^k = \sum_{j=1}^{n} E(Y_j^k \,|\hat{\pi}_j, f_{1j}) + n^{1/2} u_n + o_p(n^{1/2}). \tag{2.13}$$

From (2.13) we have

$$\sum_{j=1}^{n} (Y_j - \bar{Y})^2 = \sum_{j=1}^{n} E(Y_j^2 | \hat{\pi}_j, f_{1j}) - [\sum_{j=1}^{n} E(Y_j | \hat{\pi}_j, f_{1j})]^2/n \tag{2.14}$$

$$+ n^{1/2} u_n + o_p(n^{1/2}).$$

From (2.11), (2.13), and (2.14) it follows that

$$s_b^2 = \{1/[n^{1/2} \sum_{j=1}^{n} (\hat{\pi}_j - \bar{\hat{\pi}})^2]^2\} \Big(\sum_{j=1}^{n} (\hat{\pi}_j - \bar{\hat{\pi}})^2 \{ \sum_{j=1}^{n} E(Y_j^2 | \hat{\pi}_j, f_{1j})$$

$$- [\sum_{j=1}^{n} E(Y_j | \hat{\pi}_j, f_{1j})]^2/n + n^{1/2} u_n + o_p(n^{1/2}) \} \tag{2.15}$$

$$- [\sum_{j=1}^{n} E(Y_j | \hat{\pi}_j, f_{1j})(\hat{\pi}_j - \bar{\hat{\pi}}) + n^{1/2} u_n + o_p(n^{1/2})]^2 \Big).$$

From (2.8) and the definition of b in (2.11) we have

$$E(b | \underline{\hat{\pi}}, \underline{f}_1, H_A) = \beta_1 + \gamma_1 \sum_{j=1}^{n} f_{1j} (\hat{\pi}_j - \bar{\hat{\pi}}) / \sum_{j=1}^{n} (\hat{\pi}_j - \bar{\hat{\pi}})^2, \tag{2.16}$$

where β_1 and γ_1 are defined in (2.9). (We see from the definition of γ_1 that under H_A b is a biased estimate of β_1 unless $\sigma_d^2 = 0$.) Similarly, it follows from (2.11) that

$$V(b|\underline{\hat{\pi}},\underline{f}_1,H_A) = \sum_{j=1}^{n} V(Y_j|\hat{\pi}_j,f_{1j})(\hat{\pi}_j-\bar{\hat{\pi}})^2/[\sum_{j=1}^{n}(\hat{\pi}_j-\bar{\hat{\pi}})^2]^2, \qquad (2.17)$$

where $V(Y_j|\hat{\pi}_j,f_{1j})$ is given by (2.9) and (2.10).

We now require asymptotic results for the functions of $\underline{\hat{\pi}}$ and $\underline{f}_1$ in (2.15), (2.16), and (2.17). We will use the following general result.

Lemma 2.1

Let (x_j,y_j), $j = 1, \ldots, n$, be independent and identically distributed (IID) random variables with all moments finite, and let

$$S = \sum_{j=1}^{n} x_j^a y_j^b (x_i-\bar{x})^2,$$

where

$$\bar{x} = \sum_{j=1}^{n} x_j/n,$$

for some real a and b. Then

$$S = n\{E(x^{a+2}y^b) - 2E(x)E(x^{a+1}y^b) + [E(x)]^2E(x^ay^b)\} + n^{1/2}u_n + o_p(n^{1/2}).$$

Proof. We can write S as

$$\sum_{j=1}^{n} x_j^{a+2}y_j^b - 2\sum_{j=1}^{n} x_j^{a+1}y_j^b(\sum_{j=1}^{n} x_j/n) + \sum_{j=1}^{n} x_j^a y_j^b(\sum_{j=1}^{n} x_j/n)^2.$$

Since the (x_j,y_j) are IID, all the above sums are asymptotically normal; for example,

$$\sum_{j=1}^{n} x_j^a y_j^b = nE(x^a y^b) + n^{1/2} u_n + o_p(n^{1/2}).$$

Replacing all the above sums by similar expressions, we have

$$S = nE(x^{a+2} y^b) + n^{1/2} u_n + o_p(n^{1/2}) - 2[nE(x^{a+1} y^b)$$

$$+ n^{1/2} u_n + o_p(n^{1/2})][E(x) + n^{-1/2} u_n + o_p(n^{-1/2})]$$

$$+ [nE(x^a y^b) + n^{1/2} u_n + o_p(n^{1/2})][E(x) + n^{-1/2} u_n$$

$$+ o_p(n^{-1/2})]^2,$$

or

$$S = n\{E(x^{a+2} y^b) - 2E(x^{a+1} y^b)E(x) + E(x^a y^b)[E(x)]^2\} + n^{1/2} u_n + o_p(n^{1/2}).$$

This proves the lemma.

Applying Lemma 2.1 to the relevant sums of the form

$$\sum_{j=1}^{n} \hat{\pi}_j^a f_{1j}^b (\hat{\pi}_j - \bar{\hat{\pi}})^2,$$

we have the following:

$$\left.\begin{aligned}
&\sum_{j=1}^{n}(\hat{\pi}_j-\bar{\hat{\pi}})^2 = nV(\hat{\pi}) + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\hat{\pi}_j(\hat{\pi}_j-\bar{\hat{\pi}})^2 = n\{E(\hat{\pi}^3) - 2E(\hat{\pi})E(\hat{\pi}^2) + [E(\hat{\pi})]^3\} + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\hat{\pi}_j^2(\hat{\pi}_j-\bar{\hat{\pi}})^2 = n\{E(\hat{\pi}^4) - 2E(\hat{\pi})E(\hat{\pi}^3) + [E(\hat{\pi})]^2E(\hat{\pi}^2)\} \\
&\qquad\qquad + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum\hat{\pi}_j f_{1j}(\hat{\pi}_j-\bar{\hat{\pi}})^2 = n\{E(\hat{\pi}^3 f_1) - 2E(\hat{\pi})E(\hat{\pi}^2 f_1) + [E(\hat{\pi})]^2E(\hat{\pi}f_1)\} \\
&\qquad\qquad + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum f_{1j}(\hat{\pi}_j-\bar{\hat{\pi}})^2 = n\{E(\hat{\pi}^2 f_1) - 2E(\hat{\pi})E(\hat{\pi}f_1) + [E(\hat{\pi})]^2E(f_1)\} \\
&\qquad\qquad + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum f_{1j}^2(\hat{\pi}_j-\bar{\hat{\pi}})^2 = n\{E(\hat{\pi}^2 f_1^2) - 2E(\hat{\pi})\,E(\hat{\pi}f_1^2) + [E(\hat{\pi})]^2E(f_1^2)\} \\
&\qquad\qquad + n^{1/2}u_n + o_p(n^{1/2}).
\end{aligned}\right\} \quad (2.18)$$

By the same methods we can show that

$$\sum_{j=1}^{n} f_{1j}(\hat{\pi}_j-\bar{\hat{\pi}}) = n\,\mathrm{cov}(\hat{\pi},f_1) + n^{1/2}u_n + o_p(n^{1/2}). \qquad (2.19)$$

Substitution from (2.8), (2.9), (2.10), (2.18), and (2.19) in (2.15), (2.16), and (2.17) yields the following:

$$
\left.
\begin{aligned}
s_b^2 &= [nU + n^{1/2}u_n + o_p(n^{1/2})]/[\sum_{j=1}^{n}(\hat{\pi}_j-\bar{\hat{\pi}})^2]^2 \\
E(b|\underline{\pi},\underline{f}_1,H_A) &= \{n[\beta_1 V(\hat{\pi}) + \gamma_1 \mathrm{cov}(\hat{\pi},f_1)] + n^{1/2}u_n \\
&\qquad + o_p(n^{1/2})\}/\sum_{j=1}^{n}(\hat{\pi}_j-\bar{\hat{\pi}})^2 \\
V(b|\underline{\hat{\pi}},\underline{f}_1,H_A) &= [nW + n^{1/2}u_n + o_p(n^{1/2})]/[\sum_{j=1}^{n}(\hat{\pi}_j-\bar{\hat{\pi}})^2]^2,
\end{aligned}
\right\} \quad (2.20)
$$

where

$$
\left.
\begin{aligned}
U &= V(\hat{\pi})\{(\alpha_2-\alpha_1^2) + (\beta_2-2\alpha_1\beta_1)E(\hat{\pi}) + (\gamma_2-2\alpha_1\gamma_1)E(f_1) - \beta_1^2 E(\hat{\pi}^2) \\
&\qquad - 2\beta_1\gamma_1 E(\hat{\pi}f_1) - \gamma_1^2[E(f_1)]^2\} - \gamma_1^2[\mathrm{cov}(\hat{\pi},f_1)]^2 \\
W &= (\alpha_2-\alpha_1^2)V(\hat{\pi}) + (\beta_2-2\alpha_1\beta_1)\{E(\hat{\pi}^3) - 2E(\hat{\pi})E(\hat{\pi}^2) + [E(\hat{\pi})]^3\} \\
&\qquad + (\gamma_2-2\alpha_1\gamma_1)\{E(\hat{\pi}^2 f_1) - 2E(\hat{\pi})E(\hat{\pi}f_1) + [E(\hat{\pi})]^2 E(f_1)\} \\
&\qquad - \beta_1^2\{E(\hat{\pi}^4) - 2E(\hat{\pi})E(\hat{\pi}^3) + [E(\hat{\pi})]^2 E(\hat{\pi}^2)\} \\
&\qquad - 2\beta_1\gamma_1\{E(\hat{\pi}^3 f_1) - 2E(\hat{\pi})E(\hat{\pi}^2 f_1) + [E(\hat{\pi})]^2 E(\hat{\pi}f_1)\} \\
&\qquad - \gamma_1^2\{E(\hat{\pi}^2 f_1^2) - 2E(\hat{\pi})E(\hat{\pi}f_1^2) + [E(\hat{\pi})]^2 E(f_1^2)\}
\end{aligned}
\right\} \quad (2.21)
$$

and the α_k, β_k, γ_k are given in (2.9). Substituting from (2.20) into (2.12) and performing simple indicated algebra, we have

$$z_{1-\tau}W^{1/2} + z_{1-\alpha}U^{1/2} = -\sqrt{n}\,[\beta_1 V(\hat{\pi}) + \gamma_1 \mathrm{cov}(\hat{\pi},f_1)] + u_n + o_p(1), \quad \text{where}$$

U and W are given in (2.21). For n sufficiently large, the term $o_p(1)$ is negligible and we have

$$\sqrt{n} = -[z_{1-\tau}W^{1/2} + z_{1-\alpha}U^{1/2} + u_n]/[\beta_1 V(\hat{\pi}) + \gamma_1 \mathrm{cov}(\hat{\pi}, f_1)]. \qquad (2.22)$$

Finally, we approximate the required sample size by taking the expected value of the right side of (2.22) (note that u_n in this case is a function of the $\hat{\pi}_j$ and f_{1j}) and then squaring both sides, obtaining

$$n = [z_{1-\tau}W^{1/2} + z_{1-\alpha}U^{1/2}]^2/[\beta_1 V(\hat{\pi}) + \gamma_1 \mathrm{cov}(\hat{\pi}, f_1)]^2, \qquad (2.23)$$

where U and W are defined in (2.21) and the α_k, β_k, and γ_k are defined in (2.9).

From (2.23) we can determine the approximate sample size (i.e., number of sib pairs required) for the sib-pair test at the $100\alpha\%$ level to have power $1-\tau$, given the parameters σ_e^2, a, d, and p at the trait locus and the joint distribution of the $(\hat{\pi}_j, f_{1i})$. Let us assume no dominance and complete information on parental phenotypes for a two-allele marker locus. Let M and m denote the marker alleles, and denote the respective gene frequencies of M and m by u and $v = 1-u$. Table 2.3, which is essentially Table 6.2 from Haseman (1970, Chapter VI) for the case of two alleles, gives the joint probabilities of sibs' and parents' genotypes, as well as the f_i and $\hat{\pi}$, for all the possible combinations of sibs and parents. From Table 2.3 we can easily derive the joint distribution of $\hat{\pi}$ and f_1, which is given in Table 2.4.

TABLE 2.3

PROBABILITIES OF SIBS AND PARENTS, AND f_i ($i = 0,1,2$) AND $\hat{\pi}$ FOR A TWO-ALLELE LOCUS WITH ALLELES M,m AND RESPECTIVE GENE FREQUENCIES u AND v, ASSUMING RANDOM MATING

Mating	Sib Pair (Unordered)	Joint Probability of Sibs and Parents	f_0	f_1	f_2	$\hat{\pi}$
MM x MM	MM - MM	u^4	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
MM x Mm	MM - MM	u^3v	0	$\frac{1}{2}$	$\frac{1}{2}$	3/4
	MM - Mm	$2u^3v$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{4}$
	Mm - Mm	u^3v	0	$\frac{1}{2}$	$\frac{1}{2}$	3/4
MM x mm	Mm - Mm	$2u^2v^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
Mm x Mm	MM - MM	$u^2v^2/4$	0	0	1	1
	MM - MM	u^2v^2	0	1	0	$\frac{1}{2}$
	MM - mm	$u^2v^2/2$	1	0	0	0
	Mm - Mm	u^2v^2	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$
	Mm - mm	u^2v^2	0	1	0	$\frac{1}{2}$
	mm - mm	$u^2v^2/4$	0	0	1	1
Mm x mm	Mm - Mm	uv^3	0	$\frac{1}{2}$	$\frac{1}{2}$	3/4
	Mm - mm	$2uv^3$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{4}$
	mm - mm	uv^3	0	$\frac{1}{2}$	$\frac{1}{2}$	3/4
mm x mm	mm - mm	v^4	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$

TABLE 2.4

JOINT DISTRIBUTION OF $\hat{\pi}$ AND f_1, ASSUMING NO DOMINANCE AND COMPLETE PARENTAL INFORMATION

		$\hat{\pi}$					
		0	1/4	1/2	3/4	1	Total
f_1	0	$u^2v^2/2$	0	u^2v^2	0	$\frac{u^2v^2}{2}$	$2u^2v^2$
	½	0	$2u^3v+2uv^3$	$u^4+2u^2v^2+v^4$	$2u^3v+2uv^3$	0	$u^4+4u^3v+2u^2v^2+4uv^3+v^4$
	1	0	0	$2u^2v^2$	0	0	$2u^2v^2$
Total		$\frac{u^2v^2}{2}$	$2u^3v+2uv^3$	$u^4+5u^2v^2+v^4$	$2u^3v+2uv^3$	$\frac{u^2v^2}{2}$	1

From Table 2.4 we can derive moments for the joint distribution of $\hat{\pi}$ and f_1 for the case of no dominance and complete parental information at the marker locus, as follows.

$$
\left.
\begin{aligned}
&E(\hat{\pi}) = 1/2 \\
&E(\hat{\pi}^2) = [1 + uv(1-uv)]/4 \\
&V(\hat{\pi}) = uv(1-uv)/4 \\
&E(\hat{\pi}^3) = [1 + 3uv(1-uv)]/8 \\
&E(\hat{\pi}^4) = 1/16 + 25uv/64 - 11u^2v^2/32 \\
&E(f_1) = 1/2 \\
&E(f_1^2) = (1+4u^2v^2)/4 \\
&V(f_1) = u^2v^2 \\
&E(\hat{\pi}f_1) = 1/4 \\
&Cov(\hat{\pi},f_1) = 0 \\
&E(\hat{\pi}^2f_1) = (1+u^3v+uv^3)/8 \\
&E(\hat{\pi}f_1^2) = (1+3u^3v+3uv^3)/16 \\
&E(\hat{\pi}^3f_1) = (1+4u^2v^2)/8 \\
&E(\hat{\pi}^2f_1^2) = [1 + uv(1+2uv)]/16.
\end{aligned}
\right\} \qquad (2.24)
$$

Substituting from (2.24) into (2.23), we can calculate approximate sample sizes for given values of α, τ, and the parameters at the trait locus.

2.4. Sample Size Calculations

The sample size calculated from (2.23) depends on the genetic parameters σ_e^2, p, a, d at the trait locus only through the parameters α_1, β_1, γ_1, α_2, β_2, γ_2 defined in (2.9). From (2.23) and the definitions in (2.2) and (2.9) it can be shown that multiplication of σ_e^2, a^2, and

d^2 by a common factor leaves the sample size in (2.23) unchanged; that is, the sample size depends on a^2/σ_e^2 and d^2/σ_e^2 rather than on all of σ_e^2, a, and d separately. Hence we can, without loss of generality, let $\sigma_e^2 = 1$.

Futher, the sample size for given p, u, and λ is nearly a monotonic function of heritability, where heritability is defined by $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2/2)$. (This definition corresponds to the usual definition of heritability in the broad sense (Lush (1949)) if e_{1j} and e_{2j} in (2.1) are uncorrelated and represent environmental effects only.) Note that heritability depends on σ_e^2 only through a^2/σ_e^2 and d^2/σ_e^2. If d, p, u, and λ are all fixed, then sample size is exactly a monotonically decreasing function of h^2.

Table 2.5 gives sample sizes from (2.23) for the case u = .5, no dominance and complete parental information at the marker locus, and $\alpha = .05$, power $1-\tau = .90$, $\lambda = 0$, $h^2 = .1(.1).9$, and

(i) p = .5 for no dominance at the trait locus,

(ii) p = .1(.2).9 for complete dominance at the trait locus.

Also given are simple rules so that approximate sample sizes can be easily obtained for d = 0 and d = a, p = .1(.2).9, u = .1(.2).9, $h^2 = .1(.1).9$, and $\lambda = 0(.1).2$. The approximation gives results for the case d = 0 which differ from those calculated directly from (2.23) by less than 4%; for sample sizes less than 2000, the approximation gives results higher than those from (2.23), but not more than 2% higher. For the case d = a, the approximation gives results within 13% of the results of (2.23) (within 5% for sample sizes less than 2000).

TABLE 2.5

HASEMAN-ELSTON SIB-PAIR LINKAGE TEST:

SAMPLE SIZE (NUMBER OF SIB PAIRS) REQUIRED FOR 90% POWER AT $\alpha = .05$, ASSUMING NO DOMINANCE AND COMPLETE PARENTAL INFORMATION AT MARKER LOCUS

$n = n(h^2, \lambda, p, u)$ = number of sib pairs; $\lambda = 0$, $u = .5$.

	No Dominance at Trait Locus (d=0)	Complete Dominance at Trait Locus (d=a)				
	p=.5	p=.1	p=.3	p=.5	p=.7	p=.9
h^2	n	n	n	n	n	n
.1	33010	33210	33280	33570	34260	40180
.2	7421	7585	7550	7723	8249	13950
.3	2953	3104	3037	3169	3641	9267
.4	1481	1625	1541	1653	2099	7688
.5	841	981	888	987	1417	6984
.6	516	653	554	645	1064	6617
.7	334	469	365	450	862	6404
.8	224	358	250	331	737	6271
.9	155	287	177	255	656	6185

To approximate $n(h^2, \lambda, p, u)$ for $\lambda = 0(.1).2$, $u = .1(.2).9$:

I. For d = 0

1. To approximate $n(h^2, 0, p, .5)$, add to $n(h^2, 0, .5, .5)$ (first column of table):

$$\begin{cases} 18 \text{ for } p=.3 \text{ or } .7 \\ 163 \text{ for } p=.1 \text{ or } .9 \end{cases}$$

2. To approximate $n(h^2, 0, p, u)$, multiply $n(h^2, 0, p, .5)$ obtained in step 1 by

$$\begin{cases} 1.13 \text{ for } u = .3 \text{ or } .7 \\ 2.29 \text{ for } u = .1 \text{ or } .9 \end{cases}$$

3. To approximate $n(h^2, \lambda, p, u)$, multiply $n(h^2, 0, p, u)$ obtained in step 2 by

$$\begin{cases} 2.49 \text{ for } \lambda = .1 \\ 7.96 \text{ for } \lambda = .2. \end{cases}$$

II. For d = a

1. To approximate $n(h^2, 0, p, u)$, multiply $n(h^2, 0, p, .5)$ (last five columns of the table) by

$$\begin{cases} 1.14 \text{ for } u = .3 \text{ or } .7 \\ 2.39 \text{ for } u = .1 \text{ or } .9 \end{cases}$$

2. To approximate $n(h^2, \lambda, p, u)$, multiply $n(h^2, 0, p, u)$ obtained in step 1 by

$$\begin{cases} 2.58 \text{ for } \lambda = .1 \\ 8.35 \text{ for } \lambda = .2. \end{cases}$$

Effect of gene frequency at the marker locus

The rules given under Table 2.5 reflect the fact that the ratio of sample sizes for any two values of u is nearly independent of h^2, λ, p, and d. The moments in (2.24) are functions of uv; hence for this case of no dominance at the marker locus, the sample size is symmetric about u=.5. The sample size appears to be minimized for fixed (h^2, λ, p) at u=.5.

Effect of gene frequency at the trait locus

The relationship of sample size to p is more complicated. For $d = 0$ and fixed u and h^2, the sample size is symmetric about $p = .5$ and decreases as p approaches .5; this is because the dominance variance component σ_d^2 is then 0 and σ_a^2 depends on p only through pq, as we see from (2.2). For $d \neq 0$, the relationship is changed; for $d = a$, $\lambda = 0$, and fixed u and h^2, the sample size decreases with increasing p for small p and then for larger p increases as p increases. It is not clear from Table 2.5 at what value of p the sample size is minimized for the case $d = a$; further computation indicates that for $\lambda = 0$ and $u = .5$, the minimum is around $p = .25$ for $h^2 = .3$ to .7, lower for lower h^2, and higher for higher h^2.

Effect of dominance at the trait locus

The dependence of sample size on d is not clear. The sample size for $d = a$, relative to that for $d = 0$, is a function of h^2 and p; for some h^2 and p the sample size for $d = a$ is larger than for $d = 0$, and for other h^2 and p it is larger for $d = 0$. For $p > .5$, the sample size for $d = a$ is often quite large compared to that for $d = 0$, with the percentage difference increasing with h^2.

Effect of recombination fraction

Sample size must of course increase with λ, since $\lambda = 0$ represents the case of "complete linkage" and $\lambda = .5$ is the case of no linkage (i.e., the null hypothesis). The rules given in Table 2.5 indicate that sample size increases quite rapidly with increasing λ.

Effects of α and τ

Other obvious relationships are that the sample size in (2.23) decreases with increasing α and/or with decreasing power $(1-\tau)$. At least for the case of no dominance and complete parental information at the marker locus and no dominance at the trait locus, the values of U and W in (2.21) are approximately equal, so that the ratio of the sample size from (2.23) for α and τ' to that for α and τ is approximately

$$\left(\frac{z_{1-\tau'} + z_{1-\alpha'}}{z_{1-\tau} + z_{1-\alpha}}\right)^2.$$

Table 2.6 gives values of this ratio, expressed as a percentage of the sample size for $\alpha = .05$ and $1-\tau = .90$, for selected values of α' and $1-\tau'$. Direct computation from (2.23) yields percentages which are within one percentage point of those in Table 2.6, for all combinations of the following genetic parameter values: $\sigma_e^2 = 1$; $\lambda = 0, .2$; $p = .1, .5$; $h^2 = .1, .5, .9$; $u = .1, .5$.

TABLE 2.6

HASEMAN-ELSTON SIB-PAIR LINKAGE TEST: APPROXIMATE SAMPLE SIZE REQUIRED FOR SELECTED VALUES OF α AND POWER, ASSUMING NO DOMINANCE AND COMPLETE PARENTAL INFORMATION AT THE MARKER LOCUS AND NO DOMINANCE AT THE TRAIT LOCUS, EXPRESSED AS A PERCENTAGE OF THE SAMPLE SIZE REQUIRED FOR $\alpha = .05$, POWER = .90

α	Power	Percentage Sample Size
.10	.90	77
.05	.80	72
.10	.80	53
.05	.70	55
.05	.50	33

Whereas the sample sizes required for $\alpha = .05$ and 90% power are in many cases prohibitive, by relaxing the requirements on α and/or τ one can obtain reasonable sample sizes. The price for this is of course an increased chance of concluding there is linkage when there is no linkage and/or a decreased chance of detecting linkage when it exists.

Effects of Information and Dominance at the Marker Locus

We have thus far considered only the case of no dominance and complete parental information at the marker locus; for the most part we will continue to base our investigation into the behavior of the Haseman-Elston test on this case. It is of interest, however, to compare sample sizes obtained under different assumptions. Sample sizes have been calculated from (2,23) for the following cases at a two-allele marker locus:

(1) No dominance and complete parental information

(2) No dominance and no parental information

(3) Dominance and complete parental information

(4) Dominance and no parental information.

Values taken for the genetic parameters were $\sigma_e^2 = 1$, $d = 0$, $h^2 = .1(.1).9$, $\lambda = 0(.1).2$, $p = .1(.2).9$, and $u = .1(.2).5$; $\alpha = .05$. The results are roughly what one would expect. For all values of the genetic parameters used, sample sizes are smallest for case (1) and largest for case (4); that is, the test has greatest power for case (1) and least power for case (4). Expressed as a multiple of the sample size for case (1), the ranges of sample sizes from (2.23) for the other cases are as follows:

case (2), 1.45 to 1.83;

case (3), 1.12 to 2.38;

case (4), 1.56 to 3.44.

For the cases of dominance at the marker locus ((3) and (4)), the multipliers depend mostly on u and relatively little on the other parameters. For case (2) the range of values is much smaller; the multiplier increases with increasing h^2, increasing u, and decreasing λ, but changes in p have little effect on it besides that due to changes in h^2.

CHAPTER III

SIMULATION STUDIES OF POWER AND ROBUSTNESS OF THE HASEMAN-ELSTON TEST

3.1. Distribution of Y_j

The question remains whether the sample sizes calculated in the previous chapter are adequate to attain the nominal powers for which they were calculated. Using the actual distributions of the Y_j and the $(\hat{\pi}_j, f_{1j})$ (assuming no dominance and complete parental information at the marker locus), samples of a specified size have been drawn via computer simulation, the Haseman-Elston test done, and the observed power calculated. As shown in the following section, the results are quite good, even for relatively small sample sizes; that is, the observed power is quite close to the theoretical power obtained from asymptotic considerations.

The conditional density of Y_j, given $\hat{\pi}_j$ and f_{1j}, has been given in general form by (2.4) in section 2.1. It follows from the definition of Y_j and the assumption of normality that the distribution of Y_j/σ_e^2, given the true proportion of genes i.b.d. at the trait locus, is a mixture of noncentral χ^2 variates each with one degree of freedom; the noncentralities are given in Table 2.1 in section 2.1.

Now let $P_i (i = 1,2,3,4)$ denote the density $\sigma_e^2 \chi_1^2(\delta_i^2)$, where the noncentrality δ_i^2 is as follows:

$$\left.\begin{aligned} \delta_1^2 &= 0 \\ \delta_2^2 &= (a-d)^2/\sigma_e^2 \\ \delta_3^2 &= (a+d)^2/\sigma_e^2 \\ \delta_4^2 &= 4a^2/\sigma_e^2. \end{aligned}\right\} \qquad (3.1)$$

From (2.4), (2.23), and Table 2.1 we can write the density of Y_j/σ_e^2, given $\hat{\pi}_j$ and f_{1j}, as

$$\left.\begin{aligned} P(Y_j\sigma_e^2|\hat{\pi}_j,f_{1j}) &= c_{1j}P_1 + c_{2j}P_2 + c_{3j}P_3 + c_{4j}P_4, \\ \text{where} \\ c_{1j} &= (p^4+q^4+4p^2q^2)R_{0j} + (p^3+q^3+pq)R_{1j} + R_{2j} \\ c_{2j} &= 4p^3qR_{0j} + 2p^2qR_{1j} \\ c_{3j} &= 4pq^3R_{0j} + 2pq^2R_{1j} \\ c_{4j} &= 2p^2q^2R_{0j} \end{aligned}\right\} \qquad (3.2)$$

and the R_{hj} are given in (2.6) in section 2.1.

3.2. Verification of Sample Size Formula

The computer simulations in this section were based on the conditional distribution of Y_j/σ_e^2 just described, together with the distribution of $(\hat{\pi}_j,f_{1j})$ from Table 2.4 in section 2.3. The required uniform pseudorandom numbers were generated using methods described by MacLaren and Marsaglia (1965); an algorithm described by Kronmal (1964) was used to generate

normal pseudorandom numbers. The simulations were as follows. For the jth sib pair, a uniform pseudorandom number was used to pick values of $\hat{\pi}_j$ and f_{1j} with appropriate probability. The mixture coefficients c_{ij} in (3.2) were calculated; a second uniform pseudorandom number was then used, along with the mixture coefficients, to pick one of the four possible distributions (that is, to choose one of the noncentralities in (3.1) with appropriate probability). Finally, a standardized normal pseudorandom number was generated and added to the square root of the noncentrality, and the result was squared, yielding the simulated squared difference Y_j. (The "environmental" variance σ_e^2 was set equal to 1; hence no multiplication by σ_e^2 was necessary to obtain Y_j.) After the required sample size was reached, the sib-pair test statistic was calculated. The entire process was repeated 100 times and the observed power (or proportion of tests in which the null hypothesis $H_o:\lambda = 1/2$ was rejected) calculated.

The first five rows of Table 3.1 show, for $\alpha = .05$, $\sigma_e^2 = 1$, and selected values of p, u, λ, a, d, and sample size, the theoretical power (from (2.23) in section 2.3) and the observed power after 100 simulations. The examples were chosen to include one extremely small sample size as well as cases both of no dominance and of complete dominance at the trait locus and cases of both zero and nonzero λ. For all these cases the agreement between theoretical and observed power is quite good. The last five rows give the results from another set of simulations with the same (a,d,p,u) but with $\lambda = .5$, the null hypothesis value. The observed and theoretical powers agree quite well for these simulations also; hence the nominal α-level appears to be a good approximation to the true Type I error, at least for the cases tabulated. Note that the results are conservative

(fewer falsely significant results than expected) for the two cases in which the observed α-level differs most from the nominal .05.

TABLE 3.1

THEORETICAL AND OBSERVED POWER FROM COMPUTER SIMULATIONS OF THE HASEMAN-ELSTON SIB-PAIR TEST, ASSUMING NO DOMINANCE AND COMPLETE PARENTAL INFORMATION AT THE MARKER LOCUS

For each sample size (number of sib pairs), 100 samples were drawn; $\alpha = .05$, $\sigma_e^2 = 1$.

p	u	λ	a	d	h^2	Sample Size	Theoretical Power	Observed Power
.3	.3	0	3	0	.88	205	.90	.95
.3	.5	0	4	4	.97	142	.90	.88
.5	.5	.1	2	0	.80	553	.90	.90
.3	.5	0	5	5	.98	46	.50	.44
.1	.5	0	3	3	.92	90	.50	.52
.3	.3	.5	3	0	.88	205	.05	.06
.3	.5	.5	4	4	.97	142	.05	.03
.5	.5	.5	2	0	.80	553	.05	.02
.3	.5	.5	5	5	.98	46	.05	.05
.1	.5	.5	3	3	.92	90	.05	.06

Since some fairly small sample sizes were included in Table 3.1, it seems reasonable to infer that the nominal α-level and power associated with sample sizes calculated from (2.23) will generally be good approximations to the true values. We have assumed normality of the trait values throughout this development; in the next section the power and significance

level of the test statistic will be investigated for small sample sizes for the case of normality and also for other distributions of the trait values.

It has been remarked earlier that the regression of Y_j, given $\hat{\pi}_j$ and f_{1j}, is a function of both $\hat{\pi}_j$ and f_{1j} if there is dominance at the trait locus ($d \neq 0$). For the three simulations in Table 3.1 with $d \neq 0$ and theoretical power .9 or .5, the multiple regressions indicated in (2.8) in section 2.1 have been done and the usual F-statistic for testing $H_o: \beta_1 = 0, \gamma_1 = 0$ calculated. In each case, the observed power (for $\alpha = .05$) was less than the power observed for the test of regression on $\hat{\pi}_j$ alone; the power of the F-test was also less than the power of a two-sided test for regression on $\hat{\pi}_j$. These results are not surprising, since we know from (2.9) that $\beta_1 \leq 0$ and $\gamma_1 \geq 0$, and the F-statistic tests H_o against the general alternative $H_A: \beta_1 \neq 0, \gamma_1 \neq 0$. Hence it seems that consideration of the regression on both $\hat{\pi}_j$ and f_{1j} does not increase power, even when the full model (2.8) is theoretically more appropriate.

Let us consider the regression of Y_j on f_{1j} alone, ignoring $\hat{\pi}_j$. The sample size required, based on a formula like (2.23) in section 2.3, is much larger than that from (2.23) for the regression on $\hat{\pi}_j$ alone. Further, even for the simulations with $d \neq 0$ in Table 3.1, tests of regression on f_{1j} alone have almost no power (observed power $< .10$).

In summary, examination of the case of no dominance and complete parental information at the marker locus indicates that the Haseman-Elston test is more powerful than either (i) a similar regression test on f_1 alone or (ii) a multiple regression test on both $\hat{\pi}$ and f_1. Also, the nominal α-level and power associated with sample sizes from (2.23) appear to be reasonably good approximations to the true values.

3.3. Simulations of Power and Robustness for Small Sample Sizes

For the cases $\lambda = 0$ (complete linkage) and $\lambda = .5$ (null hypothesis), 100 samples of sib pairs were drawn via computer simulation for various distributions of the original trait values and various combinations of the genetic parameters, assuming complete parental information and no dominance at the marker locus. For most of the simulations, the sample size was 20 and it was assumed there was no dominance at the trait locus (d=0). Some additional situations were simulated for the case of normally distributed trait values. For each sample the sib-pair test was done; the number of significant results at the nominal .05 level was calculated for each set of 100 samples.

The simulations were done for the following distributions of $e_j = e_{1j} - e_{2j}$, where e_{ij} and e_{2j} are the "environmental" effects in (2.1): normal; normal, with the natural logarithm of the squared difference used instead of the squared difference itself; difference of independent exponential variates, each with location parameter 1; difference of independent chi-square variates, each with one degree of freedom; and difference of independent beta variates, each with parameters 3 and 0.5. In each case e_j was scaled to have variance 1. The exponential and chi-square distributions for individual e_{ij} are skewed to the right, and the beta distribution with parameters 3 and 0.5 is skewed to the left.

Pseudorandom numbers were generated for the normal distribution as described in the previous section. For the exponential distribution, pseudorandom numbers x were obtained by setting x equal to the inverse of the distribution function F(x), where a uniform pseudorandom number was drawn as described in the previous section and taken as the value of F(x); that is,

$x = -\ell n(1-F(x))$, where $F(x) = 1 - e^{-x}$. For the chi-square distribution with one degree of freedom, normal pseudorandom numbers were generated and squared. Pseudorandom numbers were generated for the beta distribution by the method of Johnk (1964), which has been described also by Phillips and Beightler (1972).

Table 3.2 summarizes the results for power of the simulations described above; in this case, $\lambda = 0$. For each distribution shown, 8100 samples of the indicated sample size were generated, 100 for each combination of h^2, p, and u; the sample size was 20 except for one set of simulations with sample size 50, which were done to be compared with the results in the first column of the table. The results for a particular value of h^2 were averaged over the indicated values of p and u; similarly, the results for p and u were averaged over h^2 and u and over h^2 and p, respectively. The observed power was in general an increasing function of h^2, an increasing function of p, and an increasing function of u. There were no apparent interactions among the three parameters in determining power. As we would expect from the sample sizes generated for 90% power in Table 2.5, section 2.4, the power for the sample sizes shown was quite low. The observed power was higher for the normal case with sample size 50 than with sample size 20, as would be expected, but only for heritabilities about .4 or greater. (It should be emphasized that the quantity we are calling "heritability" is heritability due entirely to the two-allele trait locus under study, and does not include heritability due to other (unlinked) loci.) The various distributions simulated all resulted in about the same observed power on the average. Taking the logarithms of squared differences of trait values, which one might expect

to result in more normally distributed values of the dependent variables in the regressions, actually decreased the average observed power.

TABLE 3.2

AVERAGE OBSERVED POWER FROM COMPUTER SIMULATIONS OF THE HASFMAN-ELSTON TEST, ASSUMING COMPLETE LINKAGE ($\lambda = 0$)

Also assumed are no dominance and complete parental information at the marker locus, and no dominance at the trait locus (d=0). Unless otherwise specified, size of each sample is 20 sib pairs; $\alpha = .05$, $\sigma_e^2 = 1$.

		Distribution of "Environmental" Effect					
		Normal†	Normal, with Log Transform of Squared Differences†	Exponential	Chi-Square	Beta	Normal (Sample size 50)
h^2	.1	.060	.066	.071	.083	.063	.067
	.2	.084	.070	.091	.066	.076	.076
	.3	.087	.067	.080	.080	.093	.080
	.4	.072	.071	.100	.097	.097	.109
	.5	.112	.078	.111	.117	.104	.149
	.6	.121	.101	.121	.137	.146	.187
	.7	.142	.099	.150	.146	.149	.237
	.8	.174	.151	.169	.174	.179	.309
	.9	.226	.196	.214	.207	.216	.369
p	.1	.103	.083	.113	.108	.112	.156
	.3	.134	.104	.122	.131	.128	.185
	.5	.123	.113	.134	.130	.134	.187
u	.1	.106	.084	.109	.113	.114	.133
	.3	.123	.107	.126	.120	.124	.195
	.5	.131	.109	.134	.136	.135	.199
All Samples		.120	.100	.123	.123	.125	.176

†These results are based on the same set of simulations.

The results for robustness (that is, simulations under the null hypothesis) are given in the middle column of Table 3.3. For all cases tabulated except the beta distribution, the results shown were averaged over 8100 simulations, 100 for each combination of $h^2 = .1(.1).9$, $p = .1(.2).5$, and $u = .1(.2).5$; for the beta distribution the values used for h^2 were $.1(.2).9$. The observed significance levels did not appear to depend on h^2, p, or u. Considering the small sample sizes, the observed significance levels seem quite satisfactory; hence the sib-pair test appears reasonably robust against non-normality of the squared differences of trait values, at least for those distributions simulated.

An important characteristic of the normal distribution is symmetry, or lack of skewness, and under normal theory the residuals from a fitted linear regression are approximately normally distributed. It was felt that a transformation to eliminate most of the skewness in the residuals might bring the observed significance level closer to the nominal .05 level. For each of the simulations in Table 3.3, the squared difference Y was transformed by taking Y^s such that the skewness of the residuals was minimized, where s was allowed to range from .1 to 1 in intervals of .1. The skewness of the residuals was calculated as

$$\sum_{i=1}^{n} \frac{(r_i-\bar{r})^3}{n-1} \Bigg/ \left[\sum_{i=1}^{n} \frac{(r_i-\bar{r})^2}{n-1} \right]^{3/2},$$

where $r_i (i=1, \ldots, n)$ is the ith residual and $\bar{r} =$

$$\sum_{i=1}^{n} r_i/n.$$

As shown in the table, the average observed significance level after the transformation was actually higher than before; this occurred even though

the skewness of the residuals was usually reduced substantially. Another interesting result (not shown) is that for both sets of tests, the average skewness of the residuals was lower when the tests were significant than when they were not; this indicates that, at least for the cases simulated, we are unlikely to get spurious significance for the sib-pair test merely because of skewness in the residuals.

TABLE 3.3

AVERAGE OBSERVED SIGNIFICANCE LEVEL FROM COMPUTER SIMULATIONS OF THE HASEMAN-ELSTON TEST, ASSUMING NOMINAL .05 LEVEL

Also assumed are no dominance and complete parental information at the marker locus. Unless otherwise specified, size of each sample is 20 sib pairs and no dominance (d=0) is assumed at the trait locus; $\sigma_e^2 = 1$.

Distribution of "Environmental" Effects	Average Significance Level	
	No Skewness Transformation	After Transformation to Reduce Skewness of Residuals
Normal†	.058	.068
Normal, with Log Transform of Squared Differences†	.058	Not Done
Normal, with d = a	.064	.072
Exponential	.061	.066
Chi-Square	.060	.068
Beta	.065	.071
Normal (Sample Size 30)	.059	Not Done
Normal (Sample Size 30)	.053	.056
Exponential (Sample Size 30)	.054	.057

†These results are based on the same set of simulations.

To summarize, the results of the simulations in this section suggest that for small samples, (i) the power of the Haseman-Elston test is quite low, (ii) the significance level of the test is satisfactory for a number of continuous distributions of trait values, and (iii) transforming the squared differences of trait values to reduce the skewness of the residuals does not in general improve the significance level.

In the next two chapters, extensions of the Haseman-Elston test for sibships of size three will be studied, and it will be shown that higher power can be attained for that case than for the same total number of sibs in independent sib pairs.

CHAPTER IV

EXTENSION OF THE HASEMAN-ELSTON TEST TO SIBSHIPS OF SIZE THREE

4.1. Model

In this chapter we consider tests to detect linkage in sibships of size three. The notation is similar to that introduced in Chapter II, with some modifications. We now have three observed trait values x_1, x_2, and x_3, where the ordering is arbitrary. There are now three sib pairs within a sibship; we arbitrarily define the three squared differences of trait values by

$$\left.\begin{aligned} Y_1 &= (x_1-x_2)^2 \\ Y_2 &= (x_1-x_3)^2 \\ Y_3 &= (x_2-x_3)^2. \end{aligned}\right\} \qquad (4.1)$$

Similarly, we have $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$, the estimated proportions of genes identical by descent at the marker locus; $\hat{\pi}_1$ is the estimated proportion for sibs 1 and 2, $\hat{\pi}_2$ is the proportion for sibs 1 and 3, and $\hat{\pi}_3$ is the proportion for sibs 2 and 3. We define in an analogous manner other quantities which in Chapter II were defined for a single sib pair; for example, we now have f_{11}, f_{12}, and f_{13}, rather than f_1.

We now let e_1^*, e_2^*, and e_3^* denote the "environmental" effects for the three sibs and let e_1, e_2, and e_3 be the corresponding differences; that is,

$$\left.\begin{aligned} e_1 &= e_1^* - e_2^* \\ e_2 &= e_1^* - e_3^* \\ e_3 &= e_2^* - e_3^* \,. \end{aligned}\right\} \tag{4.2}$$

We assume the joint distribution of e_1^*, e_2^*, and e_3^* is multivariate normal with zero means, common variance for e_1^*, e_2^*, and e_3^*, and common covariance for all pairs e_j^*, e_k^* $(j \neq k)$. Letting σ_e^2 be the variance of e_1, it follows that the joint distribution of e_1, e_2, and e_3 is multivariate normal with zero means and covariance matrix

$$\sigma_e^2 \begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}.$$

In addition, the following moments, which will be used later in the derivation of $E(Y_1Y_2|$genotypes of sibs at trait locus), can be derived easily from the moment generating function for this multivariate normal distribution:

$$\left.\begin{aligned} E(e_1^2 e_2) &= E(e_1 e_2^2) = 0 \\ E(e_1^2 e_2^2) &= 3\sigma_e^4/2 \,. \end{aligned}\right\} \tag{4.3}$$

Note that the results in this paragraph do not require that e_1^*, e_2^*, and e_3^* be uncorrelated.

For the case of sib trios we introduce a new set of probabilities of true numbers of genes identical by descent at the marker locus. Let $h_{i\ell s}$ denote the probability that sib pair 1 has i genes i.b.d., sib pair

2 has ℓ genes i.b.d., and sib pair 3 has s genes i.b.d. at the marker locus, where i, ℓ, and s range from 0 to 2. (We assume throughout our study of sib trios that the sib pairs are numbered as indicated in (4.1) and (4.2); also, this use of h should not be confused with the use of h^2 to indicate heritability.) We note by inspection that $h_{i\ell s} = 0$ for many combinations of i, ℓ, and s, since only 4 distinct alleles can occur in a sibship. For example, if i = 0 and ℓ = 2, we can represent the genotypes of the ordered sibs by M_1M_2, M_3M_4, and M_1M_2, respectively, where the subscripts denote distinct parental alleles; it then follows that s must be 0, so that $h_{021} = 0$ and $h_{022} = 0$. Similar examination for all possible cases shows that there are only 10 $h_{i\ell s}$ which can have nonzero values.

We now consider the six distinct unordered mating types for a two-allele locus. Denoting the alleles by M and m, these mating types can be represented as MMxMM, MMxMm, MMxmm, MmxMm, Mmxmm, and mmxmm. For three of the mating types (MMxMM, MMxmm, and mmxmm) there is only one possible genotype for an offspring; for two (MMxMm and Mmxmm) there are two possible offspring genotypes; and for one (MmxMm) there are three possible offspring genotypes. Thus there are $3(1^3) + 2(2^3) + 1(3^3) = 46$ possible combinations of parents' and sibs' genotypes for an unordered mating and ordered sibship of size three. Enumeration of these 46 cases shows that for a two-allele locus there are 19 distinct sets of $h_{i\ell s}$, which are given in Table 4.1a, along with the corresponding matings and sibships.

The entries in Table 4.1a are obtained as in the following example. Again using subscripts to denote distinct parental alleles, consider the mating $M_1M_2 \times M_3m_4$ and ordered sibs MM, Mm, and Mm. The MM sib can be

either M_1M_3 or M_2M_3, and each of the Mm sibs can be either M_1m_4 or M_2m_4. Given the mating and sibship, there are 8 equally probable cases when we consider the distinct parental genes. These 8 cases and the corresponding numbers of genes i.b.d. for sib pairs 1,2, and 3, respectively, are as follows:

M_1M_3	M_1m_4	M_1m_4	1	1	2
M_1M_3	M_1m_4	M_2m_4	1	0	1
M_1M_3	M_2m_4	M_1m_4	0	1	1
M_1M_3	M_2m_4	M_2m_4	0	0	2
M_2M_3	M_1m_4	M_1m_4	0	0	2
M_2M_3	M_1m_4	M_2m_4	0	1	1
M_2M_3	M_2m_4	M_1m_4	1	0	1
M_2M_3	M_2m_4	M_2m_4	1	1	2

Then given the mating and ordered sibship, $h_{002} = h_{112} = h_{101} = h_{011} = 1/4$, and all other $h_{i\ell s}$ are 0. This result is indicated in the penultimate line of Table 4.1a. Assuming random mating, the probability of the unordered mating and ordered sibship is $u^3v/2$, where u and v = 1-u are the gene frequencies of the alleles M and m, respectively. Similarly, the probabilities of the other three matings and ordered sibships which have the same $h_{i\ell s}$ (that is, the other matings and sibships shown in the right half of the penultimate line of Table 4.1a) are $u^3v/2$, $uv^3/2$, and $uv^3/2$, respectively. Thus the total probabibility of having a mating and ordered sibship with the indicated values for the $h_{i\ell s}$ is $u^3v + uv^3$, as shown in the last column and penultimate line of Table 4.1b.

The 19 distinct sets of $h_{i\ell s}$ at a two-allele marker locus, along with their corresponding probabilities for the case of no dominance and complete

parental information, are given in Table 4.1b. The 19 probabilities are the probabilities of the corresponding sets of matings and ordered sibships from Table 4.1a; we note that they sum to

$$u^4 + 4u^3v + 6u^2v^2 + 4uv^3 + v^4 = 1.$$

Also given in Table 4.1b are the values of f_{11}, f_{12}, f_{13} and of $\hat{\pi}_1$, $\hat{\pi}_2$, $\hat{\pi}_3$, where

$$\hat{\pi}_k = f_{2k} + f_{1k}/2 \tag{4.4}$$

and the f_{1k} and f_{2k} are calculated from the h's as follows:

$$\left.\begin{aligned} f_{11} &= h_{101} + h_{110} + h_{121} + h_{112} \\ f_{12} &= h_{011} + h_{110} + h_{211} + h_{112} \\ f_{13} &= h_{011} + h_{101} + h_{211} + h_{121} \end{aligned}\right\} \tag{4.5}$$

and

$$\left.\begin{aligned} f_{21} &= h_{200} + h_{211} + h_{222} \\ f_{22} &= h_{020} + h_{121} + h_{222} \\ f_{23} &= h_{002} + h_{112} + h_{222}. \end{aligned}\right\}$$

TABLE 4.1a

PROBABILITIES h_{ijs} OF NUMBERS OF GENES IDENTICAL BY DESCENT AT A TWO-ALLELE LOCUS WITH ALLELES M AND m, GIVEN PARENTS' AND SIBS' GENOTYPES, AND CORRESPONDING SETS OF UNORDERED MATINGS AND ORDERED SIBSHIPS

The probabilities h_{ijs} are 0 unless otherwise specified.

h_{002}	h_{020}	h_{200}	h_{110}	h_{101}	h_{011}	h_{112}	h_{121}	h_{211}	h_{222}	Matings and Sibships
									1	MmxMm(MM,MM,MM;mm,mm,mm)
								1		MmxMm(MM,MM,Mm;mm,mm,Mm)
							1			MmxMm(MM,Mm,MM;mm,mm,Mm)
						1				MmxMm(Mm,MM,MM;Mm,mm,mm)
		1								MmxMm(MM,MM,mm;mm,mm,MM)
	1									MmxMm(MM,mm,MM;mm,MM,mm)
1										MmxMm(mm,MM,MM;MM,mm,mm)
			1							MmxMm(Mm,MM,mm;Mm,mm,MM)
				1						MmxMm(MM,Mm,mm;mm,Mm,MM)
					1					MmxMm(MM,mm,Mm;mm,MM,Mm)
			1/2			1/2				MmxMm(MM,Mm,Mm;mm,Mm,Mm)
				1/2			1/2			MmxMm(Mm,MM,Mm;Mm,mm,Mm)
					1/2			1/2		MmxMm(Mm,Mm,MM;Mm,Mm,mm)
1/4	1/4	1/4							1/4	MmxMm(Mm,Mm,Mm)
						1/4	1/4	1/4	1/4	MMxMm(MM,MM,MM;Mm,Mm,Mm): mmxMm(mm,mm,mm;Mm,Mm,Mm)
		1/4	1/4	1/4				1/4		MMxMm(MM,MM,Mm;Mm,Mm,MM): mmxMm(mm,mm,Mm;Mm,Mm,mm)
	1/4		1/4		1/4		1/4			MMxMm(MM,Mm,MM;Mm,MM,Mm); mmxMm(mm,Mm,mm;Mm,mm,Mm)
1/4				1/4	1/4	1/4				MMxMm(Mm,MM,MM;MM,Mm,Mm); mmxMm(Mm,mm,mm;mm,Mm,Mm)
1/16	1/16	1/16	1/8	1/8	1/8	1/8	1/8	1/8	1/16	MMxMM(MM,MM,MM): mmxmm(mm,mm,mm): MMxmm(Mm,Mm,Mm)

TABLE 4.1b

PROBABILITIES h_{ijs} OF NUMBERS OF GENES IDENTICAL BY DESCENT, ASSUMING NO DOMINANCE AND COMPLETE PARENTAL INFORMATION AT A TWO-ALLELE MARKER LOCUS

Also shown are the resulting f_{1k} and $\hat{\pi}_k$, and the probabilities for each of the 19 distinct sets of h_{ijs} (that is, probabilities of the corresponding sets of matings and ordered sibships). The probabilities h_{ijs} are 0 unless otherwise specified; u = probability of either allele, and v = 1-u.

h_{002}	h_{020}	h_{200}	h_{110}	h_{101}	h_{011}	h_{112}	h_{121}	h_{211}	h_{222}	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	f_{11}	f_{12}	f_{13}	Probability
									1	1	1	1	0	0	0	$u^2v^2/8$
								1		1	1/2	1/2	0	1	1	$u^2v^2/4$
							1			1/2	1	1/2	1	0	1	$u^2v^2/4$
						1				1/2	1/2	1	1	1	0	$u^2v^2/4$
		1								1	0	0	0	0	0	$u^2v^2/8$
	1									0	1	0	0	0	0	$u^2v^2/8$
1										0	0	1	0	0	0	$u^2v^2/8$
			1							1/2	1/2	0	1	1	0	$u^2v^2/4$
				1						1/2	0	1/2	1	0	1	$u^2v^2/4$
					1					0	1/2	1/2	0	1	1	$u^2v^2/4$
			1/2			1/2				1/2	1/2	1/2	1	1	0	$u^2v^2/2$
				1/2			1/2			1/2	1/2	1/2	1	0	1	$u^2v^2/2$
					1/2			1/2		1/2	1/2	1/2	0	1	1	$u^2v^2/2$
1/4	1/4	1/4							1/4	1/2	1/2	1/2	0	0	0	$u^2v^2/2$
						1/4	1/4	1/4	1/4	3/4	3/4	3/4	1/2	1/2	1/2	u^3v+uv^3
		1/4	1/4	1/4				1/4		3/4	1/4	1/4	1/2	1/2	1/2	u^3v+uv^3
	1/4		1/4		1/4		1/4			1/4	3/4	1/4	1/2	1/2	1/2	u^3v+uv^3
1/4				1/4	1/4	1/4				1/4	1/4	3/4	1/2	1/2	1/2	u^3v+uv^3
1/16	1/16	1/16	1/8	1/8	1/8	1/8	1/8	1/8	1/16	1/2	1/2	1/2	1/2	1/2	1/2	$u^4+v^4+2u^2v^2$

Expected values of functions of the h's, $\hat{\pi}$'s, and f_1's can be calculated directly from Table 4.1b; these will be given explicitly as needed later in the chapter.

Making the same genetic assumptions as in section 2.1, we have for the kth sib pair (k=1,2,3), as in (2.8) and (2.10) of section 2.1,

$$\left.\begin{aligned} E(Y_k^{\ell}|I_m) &= \alpha_{\ell} + \beta_{\ell}\hat{\pi}_k + \gamma_{\ell}f_{1k} \\ \text{and} \\ V(Y_k|I_m) &= (\alpha_2-\alpha_1)^2 + (\beta_2-2\alpha_1\beta_1)\hat{\pi}_k + (\gamma_2-2\alpha_1\gamma_1)f_{1k} \\ &\quad - (\beta_1\hat{\pi}_k+\gamma_1 f_{1k})^2, \end{aligned}\right\} \qquad (4.6)$$

where $\alpha_1,\beta_1,\gamma_1$ and $\alpha_2,\beta_2,\gamma_2$ are defined in (2.9), section 2.1, and I_m denotes the available information on parents' and sibs' phenotypes at the marker locus. (We will continue to use I_m throughout the remainder of this dissertation; in the present instance the information in I_m is summarized by the h's.)

Now, however, we no longer have mutually independent Y's, since the Y_k within a sibship are correlated (for a sibship of size three, any two distinct sib pairs must have one sib in common). For $k\neq k'$, we have

$$E(Y_kY_{k'}|I_m) = \sum_{\underline{\pi}_t} E(Y_kY_{k'}|\underline{\pi}_t) \sum_{\underline{\pi}_m} P(\underline{\pi}_t|\underline{\pi}_m)h_{2\underline{\pi}_m}, \qquad (4.7)$$

where $\underline{\pi}_t$ and $\underline{\pi}_m$ are column vectors of true proportions of genes i.b.d. at trait and marker loci, respectively, for the three sib pairs. As indicated by the notation, $h_{2\underline{\pi}_m}$ is the probability that the true proportions of genes i.b.d. at the marker locus are given by $\underline{\pi}_m$ (that is, doubling the

elements of $\underline{\pi}_m$ gives the subscripts for h). Hence both $\underline{\pi}_t$ and $\underline{\pi}_m$ have 10 possible values, corresponding to the 10 possible values of $h_{i\ell s}$, and the summations in (4.7) are over these 10 values. In (4.7) we have a special case of the extension to sibships of size three of the relation

$$E(g(Y)|I_m) = \sum_{\pi_t} E(g(Y)|\pi_t) \sum_{\pi_m} P(\pi_t|\pi_m)P(\pi_m|I_m),$$

which was given in slightly different notation in (2.5), section 2.1.

We now derive $E(Y_1Y_2|I_m)$. Table 4.2 gives Y_1, Y_2, and the expectation of Y_1Y_2, conditional on the genotypes at the trait locus for the three sibs. The expectations follow from the joint distribution of e_1 and e_2; in particular, the moments in (4.3) are used.

Table 4.3 gives the probabilities of the 27 possible ordered genotypes at a two-allele trait locus for sib trios, conditional on the set of true probabilities $\underline{\pi}_t$ of genes i.b.d. As there are 10 possible nonzero $h_{i\ell s}$, there are 10 possible values of $\underline{\pi}_t$. The analogous probabilities for sib pairs, as well as the values of the squared difference Y, were given in Table 2.1.

TABLE 4.2

VALUES OF Y_1 AND Y_2 AND EXPECTATION OF Y_1Y_2, GIVEN GENOTYPES AT TRAIT LOCUS

Sib Trio (Ordered) 1st	2nd	3rd	$Y_1=(x_1-x_2)^2$	$Y_2=(x_1-x_3)^2$	$E(Y_1Y_2 \mid \text{Sib Trio})$
TT	TT	TT	e_1^2	e_2^2	$3\sigma_e^4/2$
		Tt	e_1^2	$(a-d+e_2)^2$	$(a-d)^2\sigma_e^2+3\sigma_e^4/2$
		tt	e_1^2	$(2a+e_2)^2$	$4a^2\sigma_e^2+3\sigma_e^4/2$
TT	Tt	TT	$(a-d+e_1)^2$	e_2^2	$(a-d)^2\sigma_e^2+3\sigma_e^4/2$
		Tt	$(a-d+e_1)^2$	$(a-d+e_2)^2$	$(a-d)^4+4(a-d)^2\sigma_e^2+3\sigma_e^4/2$
		tt	$(a-d+e_1)^2$	$(2a+e_2)^2$	$4a^2(a-d)^2+(3a-d)^2\sigma_e^2+3\sigma_e^4/2$
TT	tt	TT	$(2a+e_1)^2$	e_2^2	$4a^2\sigma_e^2+3\sigma_e^4/2$
		Tt	$(2a+e_1)^2$	$(a-d+e_2)^2$	$4a^2(a-d)^2+(3a-d)^2\sigma_e^2+3\sigma_e^4/2$
		tt	$(2a+e_1)^2$	$(2a+e_2)^2$	$16a^4+16a^2\sigma_e^2+3\sigma_e^4/2$
Tt	TT	TT	$(-a+d+e_1)^2$	$(-a+d+e_2)^2$	$(a-d)^4+4(a-d)^2\sigma_e^2+3\sigma_e^4/2$
		Tt	$(-a+d+e_1)^2$	e_2^2	$(a-d)^2\sigma_e^2+3\sigma_e^4/2$
		tt	$(-a+d+e_1)^2$	$(a+d+e_2)^2$	$(a^2-d^2)+4d^2\sigma_e^2+3\sigma_e^4/2$
Tt	Tt	TT	e_1^2	$(-a+d+e_2)^2$	$(a-d)^2\sigma_e^2+3\sigma_e^4/2$
		Tt	e_1^2	e_2^2	$3\sigma_e^4/2$
		tt	e_1^2	$(a+d+e_2)^2$	$(a+d)^2\sigma_e^2+3\sigma_e^4/2$
Tt	tt	TT	$(a+d+e_1)^2$	$(-a+d+e_2)^2$	$(a^2-d^2)+4d^2\sigma_e^2+3\sigma_e^4/2$
		Tt	$(a+d+e_1)^2$	e_2^2	$(a+d)^2\sigma_e^2+3\sigma_e^4/2$
		tt	$(a+d+e_1)^2$	$(a+d+e_2)^2$	$(a+d)^4+4(a+d)^2\sigma_e^2+3\sigma_e^4/2$
tt	TT	TT	$(-2a+e_1)^2$	$(-2a+e_2)^2$	$16a^4+16a^2\sigma_e^2+3\sigma_e^4/2$
		Tt	$(-2a+e_1)^2$	$(-a-d+e_2)^2$	$4a^2(a+d)^2+(3a+d)^2\sigma_e^2+3\sigma_e^4/2$
		tt	$(-2a+e_1)^2$	e_2^2	$4a^2\sigma_e^2+3\sigma_e^4/2$
tt	Tt	TT	$(-a-d+e_1)^2$	$(-2a+e_2)^2$	$4a^2(a+d)^2+(3a+d)^2\sigma_e^2+3\sigma_e^4/2$
		Tt	$(-a-d+e_1)^2$	$(-a-d+e_2)^2$	$(a+d)^4+4(a+d)^2\sigma_e^2+3\sigma_e^4/2$
		tt	$(-a-d+e_1)^2$	e_2^2	$(a+d)^2\sigma_e^2+3\sigma_e^4/2$
tt	tt	TT	e_1^2	$(-2a+e_2)^2$	$4a^2\sigma_e^2+3\sigma_e^4/2$
		Tt	e_1^2	$(-a-d+e_2)^2$	$(a+d)^2\sigma_e^2+3\sigma_e^4/2$
		tt	e_1^2	e_2^2	$3\sigma_e^4/2$

TABLE 4.3

PROBABILITIES $c_{r\ell}$ OF SIB TRIOS, GIVEN VECTOR $\underline{\pi}_t$ OF TRUE PROPORTIONS OF GENES IDENTICAL BY DESCENT AT TRAIT LOCUS

Probabilities are 0 unless otherwise specified; p = gene frequency of allele T, and q = 1-p.

r	Sib Trio (Ordered)			ℓ and corresponding $\underline{\pi}_t' = (\pi_1\pi_2\pi_3)$									
	1st	2nd	3rd	1 (001)	2 (010)	3 (100)	4 ($\frac{1}{2}\frac{1}{2}$0)	5 ($\frac{1}{2}$0$\frac{1}{2}$)	6 (0$\frac{1}{2}\frac{1}{2}$)	7 ($\frac{1}{2}\frac{1}{2}$1)	8 ($\frac{1}{2}1\frac{1}{2}$)	9 ($1\frac{1}{2}\frac{1}{2}$)	10 (111)
1			TT	p^4	p^4	p^4	p^4	p^4	p^4	p^3	p^3	p^3	p^2
2	TT	TT	Tt			$2p^3q$	p^3q	p^3q				p^2q	
3			tt			p^2q^2							
4			TT		$2p^3q$		p^3q		p^3q		p^2q		
5	TT	Tt	Tt	$2p^3q$			p^2q^2	p^3q	p^3q	p^2q			
6			tt					p^2q^2					
7			TT		p^2q^2								
8	TT	tt	Tt						p^2q^2				
9			tt	p^2q^2									
10			TT	$2p^3q$				p^3q	p^3q	p^2q			
11	Tt	TT	Tt		$2p^3q$		p^3q	p^2q^2	p^3q		p^2q		
12			tt				p^2q^2						
13			TT			$2p^3q$	p^3q	p^3q	p^2q^2			p^2q	
14	Tt	Tt	Tt	$4p^2q^2$	$4p^2q^2$	$4p^2q^2$	$2p^2q^2$	$2p^2q^2$	$2p^2q^2$	pq	pq	pq	2pq
15			tt			$2pq^3$	pq^3	pq^3	p^2q^2			pq^2	
16			TT				p^2q^2						
17	Tt	tt	Tt		$2pq^3$		pq^3	p^2q^2	pq^3		pq^2		
18			tt	$2pq^3$				pq^3	pq^3	pq^2			
19			TT	p^2q^2									
20	tt	TT	Tt						p^2q^2				
21			tt		p^2q^2								
22			TT					p^2q^2					
23	tt	Tt	Tt	$2pq^3$			p^2q^2	pq^3	pq^3	pq^2			
24			tt		$2pq^3$		pq^3		pq^3		pq^2		
25			TT			p^2q^2							
26	tt	tt	Tt			$2pq^3$	pq^3	pq^3				pq^2	
27			tt	q^4	q^4	q^4	q^4	q^4	q^4	q^3	q^3	q^3	q^2

From Table 4.2 and Table 4.3 we can get the conditional expectations $E(Y_1Y_2|\underline{\pi}_t)$ for the 10 possible values of $\underline{\pi}_t$. These expected values are given in Table 4.4.

TABLE 4.4

CONDITIONAL EXPECTATION OF Y_1Y_2, GIVEN VECTOR $\underline{\pi}_t$ OF TRUE PROPORTIONS OF GENES IDENTICAL BY DESCENT AT TRAIT LOCUS

$\underline{\pi}_t'$	$E(Y_1Y_2\|\underline{\pi}_t)$
(001)	$3\sigma_e^4/2+4p^3q[(a-d)^4+4(a-d)^2\sigma_e^2]+4pq^3[(a+d)^4+4(a+d)^2\sigma_e^2]$ $+ 32p^2q^2(a^4+a^2\sigma_e^2)$
(010)	$3\sigma_e^4/2+4p^3q(a-d)^2\sigma_e^2+4pq^3(a+d)^2\sigma_e^2+8p^2q^2a^2\sigma_e^2$
(100)	$3\sigma_e^4/2+4p^3q(a-d)^2\sigma_e^2+4pq^3(a+d)^2\sigma_e^2+8p^2q^2a^2\sigma_e^2$
($\frac{1}{2}\frac{1}{2}$0)	$3\sigma_e^4/2+4p^3q(a-d)^2\sigma_e^2+4pq^3(a+d)^2\sigma_e^2+4p^2q^2[(a^2+d^2)^2+(2a^2+4d^2)\sigma_e^2]$
($\frac{1}{2}$0$\frac{1}{2}$)	$3\sigma_e^4/2+2p^3q[(a-d)^4+5(a-d)^2\sigma_e^2]+2pq^3[(a+d)^4+5(a+d)^2\sigma_e^2]$ $+ p^2q^2[4a^2(a-d)^2+4a^2(a+d)^2+(3a-d)^2\sigma_e^2+(a-d)^2\sigma_e^2$ $+ (3a+d)^2\sigma_e^2+(a+d)^2\sigma_e^2]$
(0$\frac{1}{2}\frac{1}{2}$)	$3\sigma_e^4/2+2p^3q[(a-d)^4+5(a-d^2\sigma_e^2+2pq^3[(a+d)^4+5(a+d)^2\sigma_e^2]$ $+ p^2q^2[4a^2(a-d)^2+4a^2(a+d)^2+(3a-d)^2\sigma_e^2+(a-d)^2\sigma_e^2$ $+ (3a+d)^2\sigma_e^2+(a+d)^2\sigma_e^2]$
($\frac{1}{2}\frac{1}{2}$1)	$3\sigma_e^4/2+2p^2q[(a-d)^4+4(a-d)^2\sigma_e^2]+2pq^2[(a+d)^4+4(a+d)^2\sigma_e^2]$
($\frac{1}{2}$1$\frac{1}{2}$)	$3\sigma_e^4/2+2p^2q(a-d)^2\sigma_e^2+2pq^2(a+d)^2\sigma_e^2$
(1$\frac{1}{2}\frac{1}{2}$)	$3\sigma_e^4/2+2p^2q(a-d)^2\sigma_e^2+2pq^2(a+d)^2\sigma_e^2$
(111)	$3\sigma_e^4/2$

We now evaluate the terms

$$\sum_{\underline{\pi}_m} P(\underline{\pi}_t | \underline{\pi}_m) h_{2\underline{\pi}_m'}$$

in (4.7). We obtain the joint distribution of $\underline{\pi}_t$ and $\underline{\pi}_m$ as follows. Let the mating be represented by

$$\frac{A_1B_1}{A_2B_2} \times \frac{A_3B_3}{A_4B_4},$$

where A and B represent the trait and marker loci, respectively, and the subscripts denote the four distinct parental chromosomes. Suppose the three ordered offspring are

$$\frac{A_1B_1}{A_3B_3}, \frac{A_1B_2}{A_4B_3}, \frac{A_1B_2}{A_3B_3};$$

the probabilities of obtaining these offspring are $\frac{1}{4}(1-\lambda)^2$, $\frac{1}{4}\lambda^2$, and $\frac{1}{4}\lambda(1-\lambda)$, respectively. Then the probability of the sibship is $\frac{1}{64}\lambda^3(1-\lambda)^3$, and $\underline{\pi}_t' = (\frac{1}{2}\ 1\ \frac{1}{2})$, $\underline{\pi}_m' = (\frac{1}{2}\ \frac{1}{2}\ 1)$. Enumerating all possibilities for the three ordered offspring, we get the joint distribution of $\underline{\pi}_t$ and $\underline{\pi}_m$, which is given in Table 4.5. (Remember that $\psi = \lambda^2+(1-\lambda)^2$.)

TABLE 4.5

JOINT DISTRIBUTION OF π_t AND π_m

$P(\pi_t, \pi_m)$

π_t \ π_m	(001)	(010)	(100)	$(\frac{1}{2}\frac{1}{2}0)$	$(\frac{1}{2}0\frac{1}{2})$	$(0\frac{1}{2}\frac{1}{2})$	$(\frac{1}{2}\frac{1}{2}1)$	$(\frac{1}{2}1\frac{1}{2})$	$(1\frac{1}{2}\frac{1}{2})$	(111)	Total
(001)	$\frac{(3\psi-1)^2}{64}$	$\frac{(1-\psi)^2}{64}$	$\frac{(1-\psi)^2}{64}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)^2}{64}$	$\frac{1}{16}$
(010)	$\frac{(1-\psi)^2}{64}$	$\frac{(3\psi-1)^2}{64}$	$\frac{(1-\psi)^2}{64}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)^2}{64}$	$\frac{1}{16}$
(100)	$\frac{(1-\psi)^2}{64}$	$\frac{(1-\psi)^2}{64}$	$\frac{(3\psi-1)^2}{64}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)^2}{64}$	$\frac{1}{16}$
$(\frac{1}{2}\frac{1}{2}0)$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{1-4\psi+5\psi^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)^2}{32}$	$\frac{1}{8}$
$(\frac{1}{2}0\frac{1}{2})$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\frac{1}{2})^2}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{\psi(1-\psi)}{16}$	$\frac{1-4\psi+5\psi^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)^2}{32}$	$\frac{1}{8}$
$(0\frac{1}{2}\frac{1}{2})$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\frac{1}{2})(3\psi-1)}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{1-4\psi+5\psi^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)^2}{16}$	$\frac{(1-\psi)^2}{32}$	$\frac{1}{8}$
$(\frac{1}{2}\frac{1}{2}1)$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\frac{1}{2})^2}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{1-4\psi+5\psi^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{1}{8}$
$(\frac{1}{2}1\frac{1}{2})$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\frac{1}{2})(3\psi-1)}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{1-4\psi+5\psi^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{1}{8}$
$(1\frac{1}{2}\frac{1}{2})$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{(1-\psi)^2}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{\psi(1-\psi)}{16}$	$\frac{1-4\psi+5\psi^2}{16}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{1}{8}$
(111)	$\frac{(1-\psi)^2}{64}$	$\frac{(1-\psi)^2}{64}$	$\frac{(1-\psi)^2}{64}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1\cdot\psi)^2}{32}$	$\frac{(1-\psi)^2}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(1-\psi)(3\psi-1)}{32}$	$\frac{(3\psi-1)^2}{64}$	$\frac{1}{16}$
Total	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	1

We obtain the conditional probabilities $P(\underline{\pi}_t|\underline{\pi}_m) = P(\underline{\pi}_t,\underline{\pi}_m)/P(\underline{\pi}_m)$ from Table 4.5. Combining each conditional probability with the proper $h_{i\ell s}$ gives us the $P(\underline{\pi}_t|\underline{\pi}_m)P(\underline{\pi}_m|I_m)$, or

$$P(\underline{\pi}_t|\underline{\pi}_m)h_{2\underline{\pi}_m}.$$

A little algebra, and use of the fact that the sum of the 10 nonzero $h_{i\ell s}$ is 1, yields the results in Table 4.6.

TABLE 4.6

DISTRIBUTION OF VECTOR $\underline{\pi}_t$ OF TRUE PROPORTIONS OF GENES IDENTICAL BY DESCENT AT TRAIT LOCUS, GIVEN FAMILY INFORMATION I_m AT MARKER LOCUS

ℓ	$\underline{\pi}_t'$	$w_\ell = P(\underline{\pi}_t \mid I_m) = \sum_{\underline{\pi}_m} P(\underline{\pi}_t \mid \underline{\pi}_m) P(\underline{\pi}_m \mid I_m) = \sum_{\underline{\pi}_m} P(\underline{\pi}_t \mid I_m) h_{2\underline{\pi}_m'}$
1	(001)	$\frac{(1-\psi)^2}{4} - \psi(1-2\psi)h_{002} - \frac{(1-\psi)(1-2\psi)}{2}(h_{101}+h_{011}+h_{112})$
2	(010)	$\frac{(1-\psi)^2}{4} - \psi(1-2\psi)h_{020} - \frac{(1-\psi)(1-2\psi)}{2}(h_{110}+h_{011}+h_{121})$
3	(100)	$\frac{(1-\psi)^2}{4} - \psi(1-2\psi)h_{200} - \frac{(1-\psi)(1-2\psi)}{2}(h_{110}+h_{101}+h_{211})$
4	$(\frac{1}{2}\frac{1}{2}0)$	$\frac{(1-\psi)^2}{2} - \psi(1-2\psi)h_{110} - \frac{(1-\psi)(1-2\psi)}{2}(h_{101}+h_{011}+h_{121}+h_{211}+2h_{020}+2h_{200})$
5	$(\frac{1}{2}0\frac{1}{2})$	$\frac{(1-\psi)^2}{2} - \psi(1-2\psi)h_{101} - \frac{(1-\psi)(1-2\psi)}{2}(h_{110}+h_{011}+h_{112}+h_{211}+2h_{002}+2h_{200})$
6	$(0\frac{1}{2}\frac{1}{2})$	$\frac{(1-\psi)^2}{2} - \psi(1-2\psi)h_{011} - \frac{(1-\psi)(1-2\psi)}{2}(h_{110}+h_{101}+h_{112}+h_{121}+2h_{002}+2h_{020})$
7	$(\frac{1}{2}\frac{1}{2}1)$	$\frac{(1-\psi)^2}{2} - \psi(1-2\psi)h_{112} - \frac{(1-\psi)(1-2\psi)}{2}(h_{101}+h_{011}+h_{121}+h_{211}+2h_{002}+2h_{222})$
8	$(\frac{1}{2}1\frac{1}{2})$	$\frac{(1-\psi)^2}{2} - \psi(1-2\psi)h_{121} - \frac{(1-\psi)(1-2\psi)}{2}(h_{110}+h_{011}+h_{112}+h_{211}+2h_{020}+2h_{222})$
9	$(1\frac{1}{2}\frac{1}{2})$	$\frac{(1-\psi)^2}{2} - \psi(1-2\psi)h_{211} - \frac{(1-\psi)(1-2\psi)}{2}(h_{110}+h_{101}+h_{112}+h_{121}+2h_{200}+2h_{222})$
10	(111)	$\frac{(1-\psi)^2}{4} - \psi(1-2\psi)h_{222} - \frac{(1-\psi)(1-2\psi)}{2}(h_{112}+h_{121}+h_{211})$

Finally, by using the results of Tables 4.4 and 4.6 and performing simple but tedious algebraic manipulations, we have the following result:

$$
\begin{aligned}
E(Y_1Y_2|I_m) = {} & 3\sigma_e^4/2+6\psi\sigma_e^2\sigma_g^2+6\psi(1-\psi)\sigma_e^2\sigma_d^2+2\psi pq[p(a-d)^4+q(a+d)^4] \\
& + 4\psi^2p^2q^2(3a^4-6a^2d^2-d^4) \\
& +(1-2\psi)\sum_{k=1}^{3}\hat{\pi}_k\{2\sigma_e^2\sigma_g^2+(2pq/3)[p(a-d)^4+q(a+d)^4] \\
& \qquad + (4p^2q^2/3)(3a^4-6a^2d^2-d^4)\} \\
& + (1-2\psi)^2\sum_{k=1}^{3}f_{1k}[\sigma_e^2\sigma_d^2-(2p^2q^2/3)(3a^4-6a^2d^2-d^4)] \\
& + (1-2\psi)[(\sum_{k=1}^{3}\hat{\pi}_k/3) - \hat{\pi}_3]\{6\sigma_e^2\sigma_g^2+4pq[p(a-d)^4+q(a+d)^4] \\
& \qquad + 4p^2q^2(3a^4-9a^2d^2-2d^4)\} \\
& + (1-2\psi)^2\{[(\sum_{k=1}^{3}f_{1k}/3)-f_{13}](3\sigma_e^2\sigma_d^2+4p^2q^2d^4) \\
& \qquad + p^2q^2a^4(4\sum_{k=1}^{3}\hat{\pi}_k-12\hat{\pi}_3+4\sum_{k=1}^{3}f_{1k}-8+24h_{002}+8h_{222}) \\
& \qquad + p^2q^2a^2d^2(-4\sum_{k=1}^{3}\hat{\pi}_k+12\hat{\pi}_3-12f_{13}+8-24h_{002}-8h_{222})\}.
\end{aligned}
\tag{4.8}
$$

We obtain $E(Y_1Y_3|I_m)$ from (4.8) by replacing $\hat{\pi}_3$, f_{13}, and h_{002} with $\hat{\pi}_2$, f_{12}, and h_{020}; similarly, we get $E(Y_2Y_3|I_m)$ by substituting $\hat{\pi}_1$, f_{11}, and h_{200}.

The unconditional expectation $E(Y_1Y_2)$ can be derived from Table 4.4 and the marginal probabilities for possible values of $\underline{\pi}_t$ in Table 4.5. The result is

$$E(Y_1Y_2) = 3\sigma_e^4/2 + 3\sigma_e^2\sigma_g^2 + 3\sigma_e^2\sigma_d^2/2 + p^2q(a-d)^4 + pq^2(a+d)^4 + p^2q^2(3a^4 - 6a^2d^2 - d^4). \tag{4.9}$$

We note that, unconditionally, $E(Y_1Y_2) = E(Y_1Y_3) = E(Y_2Y_3)$, since the numbering of sib pairs is arbitrary.

We note also from (4.8) that for $\psi = 1/2$ (equivalently, the recombination fraction λ is 1/2), we have the same result as (4.9); that is, $E(Y_1Y_2|I_m, \lambda = 1/2)$ is the same as the unconditional expectation $E(Y_1Y_2)$. Similarly, $E(Y_k|I_m, \lambda = 1/2) = E(Y_k)$ and $V(Y_k|I_m, \lambda = 1/2) = V(Y_k)$ for $k = 1,2,3$; also, $E(Y_1) = E(Y_2) = E(Y_3)$ and $V(Y_1) = V(Y_2) = V(Y_3)$. Then under the null hypothesis $H_o: \lambda = 1/2$, the covariance matrix of $\underline{Y}' = (Y_1Y_2Y_3)$ is of the form

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

where σ^2 and $\rho\sigma^2$ are the unconditional variance and covariance, respectively; that is,

$$\sigma^2 = V(Y_1) = V(Y_2) = V(Y_3)$$

and

$$\rho\sigma^2 = \text{cov}(Y_1,Y_2) = \text{cov}(Y_1,Y_3) = \text{cov}(Y_2,Y_3).$$

Explicitly,

$$\left.\begin{aligned}\sigma^2 &= 2\sigma_e^4+4\sigma_e^2\sigma_g^2+2\sigma_e^2\sigma_d^2-(\sigma_g^2+\sigma_d^2/2)^2 \\ &\quad + 2[p^2q(a-d)^4+pq^2(a+d)^4] \\ &\quad + 2p^2q^2(3a^4-6a^2d^2-d^4)\end{aligned}\right\} \qquad (4.10)$$

and

$$\left.\begin{aligned}\rho\sigma^2 &= \sigma_e^4/2+\sigma_e^2\sigma_g^2+\sigma_e^2\sigma_d^2/2-(\sigma_g^2+\sigma_d^2/2)^2 \\ &\quad + p^2q(a-d)^4+pq^2(a+d)^4 \\ &\quad + p^2q^2(3a^4-6a^2d^2-d^4).\end{aligned}\right\}$$

We note that (4.10) implies, from the definitions of α_1 and α_2 in (2.9), section 2.1, that for the case $\psi = 1/2, \sigma^2 = \alpha_2-\alpha_1^2$ and $\rho\sigma^2 = \alpha_2/2 - \alpha_1^2$. It is easily seen from (4.10) that ρ is of the form $(A-B+C)/(4A-B+2C)$, where

$$A = \sigma_e^4/2 + \sigma_e^2\sigma_g^2 + \sigma_e^2\sigma_d^2/2$$

$$B = (\sigma_g^2+\sigma_d^2/2)^2$$

$$C = p^2q(a-d)^4+pq^2(a+d)^4+p^2q^2(3a^4-6a^2d^2-d^4).$$

Since $A \geq 0$, $B \geq 0$, and $4A-B+2C > 0$ (the last must be true since $4A-B+2C$ is a variance), it follows that $\rho \leq 1/2$. We will see later that ρ is between $1/4$ and $1/3$ for a large range of parameter values of interest; also, $\rho \to 1/4$ as $\sigma_e^2 \to \infty$ for fixed a and d. However, it is possible for ρ to be less than $1/4$; for example, if $\sigma_e^2 = 0$, $a = 0$, and $p = .5$, then $A = 0$, $C = 4B/3$, and $\rho = .2$ for any $d > 0$.

4.2. Possible Regression Tests

We wish to test $H_0:\lambda = 1/2$, given a random sample of n sibships of size three, under the same genetic assumptions as in Chapter II. Within

each sibship, the expectation and variance of Y_k, conditional on the family information I_m at the marker locus, are given in (4.6), for k = 1,2,3; for $k \neq k'$, we have $\text{cov}(Y_k\ Y_{k'}|I_m) = E(Y_kY_{k'}|I_m) - E(Y_k|I_m)E(Y_{k'}|I_m)$, where $E(Y_kY_{k'}|I_m)$ is obtained from (4.8). From the previous section we know also that under H_o the covariance matrix of $\underline{Y}' = (Y_1Y_2Y_3)$ is of the form

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

where σ^2 and $\rho\sigma^2$ are given by (4.10); for the present we assume that σ^2 and ρ are known. We make the further simplifying assumption that $E(Y_k|I_m) = \alpha_1 + \beta_1\hat{\pi}_k$ (that is, we ignore γ_1, as in the Haseman-Elston test). An estimator for β_1 can then be obtained using ordinary least squares theory.

We introduce a subscript in the Y's, $\hat{\pi}$'s, and f_1's to indicate sibship; thus Y_{kj} is the squared difference for the kth sib pair in the jth sibship, and similarly we write $\hat{\pi}_{kj}$ and f_{1kj}. Suppose we have a random sample of n sibships each of size three. In matrix notation we are assuming $E(\underline{Y}|I_m) = X\underline{\theta}$ and $V(\underline{Y}|I_m) = \mathcal{V}$, where

$$
\begin{aligned}
\underset{1x3n}{\underline{Y}'} &= (Y_{11}Y_{21}Y_{31} \quad Y_{12}Y_{22}Y_{32} \quad \cdots \quad Y_{1n}Y_{2n}Y_{3n}) \\
\underset{2x3n}{X'} &= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 \\ \hat{\pi}_{11}\hat{\pi}_{21}\hat{\pi}_{31} & & & \hat{\pi}_{12}\hat{\pi}_{22}\hat{\pi}_{32} & & & \cdots & \hat{\pi}_{1n}\hat{\pi}_{2n}\hat{\pi}_{3n} & & \end{pmatrix} \\
\underset{1x2}{\underline{\theta}'} &= (\alpha_1 \ \beta_1) \\
\underset{3nx3n}{\Sigma} &= \begin{pmatrix} \Sigma & & 0 \\ & \Sigma & \\ 0 & & \Sigma \end{pmatrix}
\end{aligned}
\qquad (4.11)
$$

and

$$
\underset{3x3}{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.
$$

We want to estimate α_1 and β_1 so as to minimize $S = (\underline{Y}-\underline{Y}^*)'\Sigma^{-1}(\underline{Y}-\underline{Y}^*)$, where $\underline{Y}^*$ is the vector of estimated squared differences; that is, $Y^*_{kj} = \alpha^*_1+\beta^*_1\hat{\pi}_{kj}$ for $k = 1,2,3$ and $j = 1, \ldots, n$, where α^*_1 and β^*_1 are the least squares estimators for α_1 and β_1. We know that, assuming Σ is known, the resulting estimator $\underline{\theta}^*$ and its covariance matrix are given by

$$
\underline{\theta}^* = \begin{pmatrix} \alpha^*_1 \\ \beta^*_1 \end{pmatrix} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\underline{Y} \qquad (4.12)
$$

and

$$
V(\underline{\theta}^* | I_m) = (X'\Sigma^{-1}X)^{-1}.
$$

We define the following sums of squares and cross-products which will be useful in the remainder of this chapter and in Chapter V:

$$
\begin{aligned}
S_T(Y,Y) &= \sum_{j=1}^{n}\sum_{k=1}^{3}(Y_{kj}-\bar{\bar{Y}})^2 \\
S_W(Y,Y) &= \sum_{j=1}^{n}\sum_{k=1}^{3}(Y_{kj}-\bar{Y}_j)^2 \\
S_B(Y,Y) &= S_T(Y,Y)-S_W(Y,Y) \\
S_T(Y,\hat{\pi}) &= \sum_{j=1}^{n}\sum_{k=1}^{3}(Y_{kj}-\bar{\bar{Y}})(\hat{\pi}_{kj}-\bar{\bar{\hat{\pi}}}) \\
S_W(Y,\hat{\pi}) &= \sum_{j=1}^{n}\sum_{k=1}^{3}(Y_{kj}-\bar{Y}_j)(\hat{\pi}_{kj}-\bar{\hat{\pi}}_j) \\
S_B(Y,\hat{\pi}) &= S_T(Y,\hat{\pi})-S_W(Y,\hat{\pi}) \\
S_T(\hat{\pi},\hat{\pi}) &= \sum_{j=1}^{n}\sum_{k=1}^{3}(\hat{\pi}_{kj}-\bar{\bar{\hat{\pi}}})^2 \\
S_W(\hat{\pi},\hat{\pi}) &= \sum_{j=1}^{n}\sum_{k=1}^{3}(\hat{\pi}_{kj}-\bar{\hat{\pi}}_j)^2 \\
S_B(\hat{\pi},\hat{\pi}) &= S_T(\hat{\pi},\hat{\pi})-S_W(\hat{\pi},\hat{\pi}) \\
S_T(\hat{\pi},f_1) &= \sum_{j=1}^{n}\sum_{k=1}^{3}(\hat{\pi}_{kj}-\bar{\bar{\hat{\pi}}})(f_{1kj}-\bar{\bar{f}}_1) \\
S_W(\hat{\pi},f_1) &= \sum_{j=1}^{n}\sum_{k=1}^{3}(\hat{\pi}_{kj}-\bar{\hat{\pi}}_j)(f_{1kj}-\bar{f}_{1j}),
\end{aligned}
\qquad (4.13)
$$

where

$$
\bar{\bar{Y}} = \frac{1}{3n}\sum_{j=1}^{n}\sum_{k=1}^{3}Y_{kj},\quad \bar{Y}_j = \frac{1}{3}\sum_{k=1}^{3}Y_{kj},\quad \bar{\bar{\hat{\pi}}} = \frac{1}{3n}\sum_{j=1}^{n}\sum_{k=1}^{3}\hat{\pi}_{kj},
$$

$$
\bar{\hat{\pi}}_j = \frac{1}{3}\sum_{k=1}^{3}\hat{\pi}_{kj},\quad \bar{\bar{f}}_1 = \frac{1}{3n}\sum_{j=1}^{n}\sum_{k=1}^{3}f_{1kj},\quad \text{and}\quad \bar{f}_{1j} = \frac{1}{3}\sum_{k=1}^{3}f_{1kj}.
$$

It follows from (4.11) and (4.13) that

$$\Sigma^{-1} = \frac{1}{\sigma^2(1-\rho)(1+2\rho)} \begin{pmatrix} 1+\rho & -\rho & -\rho \\ -\rho & 1+\rho & -\rho \\ -\rho & -\rho & 1+\rho \end{pmatrix},$$

$$(x'\Sigma^{-1}x)^{-1} = \frac{\sigma^2(1+2\rho)}{3n[(1-\rho)S_T(\hat{\pi},\hat{\pi}) + 3\rho S_W(\hat{\pi},\hat{\pi})]} \begin{pmatrix} (1-\rho)S_T(\hat{\pi},\hat{\pi})+3\rho S_W(\hat{\pi},\hat{\pi})+3n(1-\rho)\bar{\hat{\pi}}^2, & -3n(1-\rho)\bar{\hat{\pi}} \\ -3n(1-\rho)\bar{\hat{\pi}} & 3n(1-\rho) \end{pmatrix},$$

and

$$x'\Sigma^{-1}\underline{Y} = \frac{1}{\sigma^2(1-\rho)(1+2\rho)} \begin{pmatrix} 3n(1-\rho)\bar{Y} \\ (1-\rho)S_T(Y,\hat{\pi})+3\rho S_W(Y,\hat{\pi})+3n(1-\rho)\bar{Y}\,\bar{\hat{\pi}} \end{pmatrix}.$$

From the above and (4.12) we obtain the estimator

$$\underset{\sim}{\theta}^* = \begin{pmatrix} \alpha_1^* \\ \beta_1^* \end{pmatrix} = \frac{\begin{pmatrix} \bar{\bar{Y}}[(1-\rho)S_T(\hat{\pi},\hat{\pi})+3\rho S_W(\hat{\pi},\hat{\pi})] - \bar{\hat{\pi}}[(1-\rho)S_T(Y,\hat{\pi})+3\rho S_W(Y,\hat{\pi})] \\ (1-\rho)S_T(Y,\hat{\pi}) + 3\rho S_W(Y,\hat{\pi}) \end{pmatrix}}{(1-\rho)S_T(\hat{\pi},\hat{\pi}) + 3\rho S_W(\hat{\pi},\hat{\pi})} . \quad (4.14)$$

Also,

$$V(\beta_1^* | I_m) = \frac{\sigma^2(1-\rho)(1+2\rho)}{(1-\rho)S_T(\hat{\pi},\hat{\pi}) + 3\rho S_W(\hat{\pi},\hat{\pi})} . \quad (4.15)$$

Then, even though the Y_{kj} are not normally distributed, for large samples we can treat the statistic

$$t = \beta_1^* \Big/ \sqrt{V(\beta_1^* | I_m)} = \frac{(1-\rho)S_T(Y,\hat{\pi}) + 3\rho S_W(Y,\hat{\pi})}{\{\sigma^2(1-\rho)(1+2\rho)[(1-\rho)S_T(\hat{\pi},\hat{\pi}) + 3\rho S_W(\hat{\pi},\hat{\pi})]\}^{1/2}} \quad (4.16)$$

as a normal deviate and hence test the null hypothesis. (Remember we have assumed σ^2 and ρ are known.) If the Y_{kj} were normally distributed, then (4.14) would give maximum likelihood estimators for α_1 and β_1.

In order to use (4.16) as a test statistic we need to have estimates for required functions of ρ and σ^2. Let

$$
\left.
\begin{aligned}
V_1 &= \{S_W(Y,Y) - [S_W(Y,\hat{\pi})]^2/S_W(\hat{\pi},\hat{\pi})\}/(2n-1) \\
V_2 &= \{S_W(Y,Y) - \sum_{j=1}^{n} [S_j(Y,\hat{\pi})]^2/S_j(\hat{\pi},\hat{\pi})\}/(n_1+2n_2) \\
V_3 &= \{S_B(Y,Y) - [S_B(Y,\hat{\pi})]^2/S_B(\hat{\pi},\hat{\pi})\}/(n-2) \\
V_4 &= \{S_T(Y,Y) - [S_T(Y,\hat{\pi})]^2/S_T(\hat{\pi},\hat{\pi})\}/(3n-2),
\end{aligned}
\right\} \qquad (4.17)
$$

where

$$
\left.
\begin{aligned}
S_j(Y,\hat{\pi}) &= \sum_{k=1}^{3} (Y_{kj}-\bar{Y}_j)(\hat{\pi}_{kj}-\bar{\hat{\pi}}_j), \\
S_j(\hat{\pi},\hat{\pi}) &= \sum_{k=1}^{3} (\hat{\pi}_{kj}-\bar{\hat{\pi}}_j)^2,
\end{aligned}
\right\}
$$

and n_1 and n_2 are the numbers of sibships with $S_j(\hat{\pi},\hat{\pi})\neq 0$ and $S_j(\hat{\pi},\hat{\pi}) = 0$, respectively ($n_1+n_2 = n$). Then under the null hypothesis, since $V(\underline{Y}|I_m) = \Sigma$ (that is, the variances and covariances of the Y_{kj} do not depend on the family information I_m), we have

$$
\left.
\begin{aligned}
E(V_1|I_m) &= \sigma^2(1-\rho) \\
E(V_2|I_m) &= \sigma^2(1-\rho) \\
E(V_3|I_m) &= \sigma^2(1+2\rho) \\
E(V_4|I_m) &= \sigma^2\{1-\rho\,[4-3S_W(\hat{\pi},\hat{\pi})/S_T(\hat{\pi},\hat{\pi})]/(3n-2)\}.
\end{aligned}
\right\} \qquad (4.18)
$$

We can use V_1, V_2, or V_3 to estimate the appropriate expectation from (4.18), but we still need an estimate of ρ to use the test statistic (4.16). Table 4.7 gives the values of ρ for a wide range of values of heritability h^2 and of gene frequency p at a two-allele trait locus, for the two cases $d = 0$ and $d = a$; given h^2, p, and the ratio of d to a, the

correlation ρ does not depend on σ_e^2 or the actual value of a. For most values tabulated, ρ is between 1/4 and 1/3; values of ρ greater than 1/3 are even more frequent for p very close to 0 or 1 (not shown).

TABLE 4.7

UNCONDITIONAL CORRELATION ρ BETWEEN Y_1 AND Y_2

	d=0*			d=a (p= frequency of the dominant allele)				
h^2	p=.1	p=.3	p=.5	p=.1	p=.3	p=.5	p=.7	p=.9
.1	.251	.250	.250	.251	.250	.251	.254	.288
.2	.256	.251	.251	.255	.251	.253	.265	.359
.3	.265	.253	.252	.263	.253	.258	.282	.412
.4	.278	.257	.254	.274	.256	.264	.307	.445
.5	.296	.262	.257	.290	.260	.274	.333	.464
.6	.318	.269	.261	.310	.266	.287	.360	.475
.7	.343	.278	.267	.333	.274	.302	.385	.482
.8	.369	.291	.275	.357	.284	.319	.408	.487
.9	.394	.307	.286	.382	.297	.339	.426	.491

*For d=0, ρ is the same for p and (1-p).

From (4.18) we see that under the null hypothesis we can get unbiased estimators for ρ and σ^2 by taking

$$\left.\begin{aligned} \hat{\sigma}^2 &= (V_3+2V_1)/3 \\ \text{and} \\ \hat{\rho} &= (V_3-V_1)/(V_3+2V_1). \end{aligned}\right\} \qquad (4.19)$$

(V_1 is arbitrarily chosen in (4.19) rather than V_2.) Using these estimators for ρ and σ^2, we estimate β_1 and its variance by

$$\hat{\beta}_1 = \frac{V_1 S_T(Y,\hat{\pi}) + (V_3-V_1) S_W(Y,\hat{\pi})}{V_1 S_T(\hat{\pi},\hat{\pi}) + (V_3-V_1) S_W(\hat{\pi},\hat{\pi})} \tag{4.20}$$

and

$$\widehat{V(\hat{\beta}_1)} = \frac{V_1 V_3}{V_1 S_T(\hat{\pi},\pi) + (V_3-V_1) S_W(\hat{\pi},\hat{\pi})} \tag{4.21}$$

The next section describes computer simulations of the power of the statistic in (4.16) for selected sample sizes and selected values of the genetic parameters when ρ is arbitrarily taken to be 0, 1/4, 1/3, or its "true" value; each of the functions V_1,V_2, and V_3 was used in turn in estimating the variance. The case ρ=0 was simulated also with V_4 used to estimate the variance; this test is the Haseman-Elston test for the 3n possible sib pairs. Also included in the simulations were the statistic $\hat{\beta}_1/\sqrt{\widehat{V(\hat{\beta}_1)}}$ from (4.20) and (4.21) and two statistics based on the estimator

$$\tilde{\beta}_1 = S_W(Y,\hat{\pi})/S_W(\hat{\pi},\hat{\pi}), \tag{4.22}$$

which has conditional variance

$$V(\tilde{\beta}_1 | I_m) = \sigma^2(1-\rho)/S_W(\hat{\pi},\hat{\pi}). \tag{4.23}$$

(We know from the work of Halperin (1971) that the least squares estimator of β_1, using data from a single sibship, is the same as the usual estimator for the case of independent observations; $\tilde{\beta}_1$ is a weighted average of these separate estimators.)

4.3. Preliminary Simulations

Computer simulations were done

(1) to compare observed power of various three-sib tests with the power of the Haseman-Elston sib-pair test for comparable sample size;

(2) to compare observed and nominal α-levels;

(3) to compare power and α-level for various arbitrary values for ρ in (4.16), including the "true"value (that is, the value calculated from (4.10), given the genetic parameters); and

(4) to compare power and α-level when ρ is estimated from (4.19) with the results for arbitrary values for ρ in (4.16).

The method used in the simulations is as follows, where the pseudo-random numbers were generated as described in section 3.2.

(1) Given the gene frequency u at the marker locus and a uniform pseudorandom number, sample from the joint distribution of the h's and thereby obtain the $\hat{\pi}$'s and f_1's. This has been done only for the case of no dominance and complete parental information at the marker locus.

(2) Given a second uniform pseudorandom number, pick a specific ordered sib trio. The probability d_r of the rth ordered sib trio (r=1, ..., 27), given the gene frequency p at the trait locus, the recombination fraction λ, and the family information I_m, can be calculated from

$$d_r = \sum_{\ell=1}^{10} c_{r\ell} w_\ell,$$

where $c_{r\ell}$ is the probability of the rth ordered sib trio, given p and the

ℓth set of true proportions of genes identical by descent at the trait locus (ℓ=1, ..., 10), and w_ℓ is the probability of the ℓth set of true proportions, given λ and I_m. The $c_{r\ell}$, which are analogous to the probabilities for sib pairs in Table 2.1, section 2.1, were given in Table 4.3; the w_ℓ, given in Table 4.6, correspond to the R_{hj} in (2.6), section 2.1. Choosing the sib trio is equivalent to choosing the three genetic effects g_k (k=1,2,3), where g_k is defined as in (2.1), section 2.1.

(3) From three uncorrelated normal pseudorandom numbers, each with variance $\sigma_e^2/2$, determine the random effects e_k (k=1,2,3) for the three sibs. This gives, for example, $V(e_1-e_2) = \sigma_e^2$ and $\text{cov}(e_1-e_2, e_1-e_3) = \sigma_e^2/2$.

(4) Calculate $x_k = g_k+e_k$ (k=1,2,3).

(5) Calculate $Y_1 = (x_1-x_2)^2$

$Y_2 = (x_1-x_3)^2$

$Y_3 = (x_2-x_3)^2$.

(6) From the Y's and $\hat{\pi}$'s for a sample of sib trios, estimate the regression of Y on $\hat{\pi}$ using the estimators in (4.14), (4.20), and (4.22), and determine whether or not the estimated β_1 is significantly different from 0 via an ordinary t-statistic (ratio of estimated β_1 to its estimated standard error), tested as a normal deviate.

(7) Repeat the above process for a total of 100 samples with the selected genetic parameters. Determine the observed powers of the various regression tests used.

Table 4.8a gives the sets of genetic parameters and sample sizes used in the simulations, as well as expected and (where done for Table 3.1, section 3.2) observed power of the sib-pair test for approximately the same total number of sibs (never differing in total sibs by more than 2). Table 4.8b gives the various combinations of estimated value of ρ, estimator for β_1 (β_1^* from (4.14), $\hat{\beta}_1$ from (4.20), or $\tilde{\beta}_1$ from (4.22)), and estimator for $\sigma^2(1-\rho)$ or $\sigma^2(1-\rho)(1+2\rho)$ which were used. Table 4.8c gives the results (observed power) for the various simulations.

TABLE 4.8a

THREE-SIB SIMULATION PROBLEMS:

SETS OF GENETIC PARAMETERS USED, SAMPLE SIZES, AND POWER OF CORRESPONDING SIB-PAIR TEST

For each sample size (number of sibships of size three), 100 samples were drawn; $\alpha=.05$, $\sigma_e^2=1$.

	p	u	λ	a	d	h^2	Sample Size	Expected Power of Sib-Pair Test	Observed Power of Sib-Pair Test
(1)	.1	.5	0	3	3	.92	60	.50	.52
(2)	.3	.3	0	3	0	.88	136	.90	.95
(3)	.3	.5	0	4	4	.97	94	.90	.88
(4)	.3	.5	0	5	5	.98	30	.50	.44
(5)	.3	.5	0	3.27	0	.90	62	.70	Not done
(6)	.3	.5	0	5	0	.95	30	.50	Not done
(7)	.5	.5	.1	5	0	.96	65	.50	Not done
(8)	.3	.3	0	3	0	.88	42	.50	Not done
(9)	.7	.3	0	3	0	.88	42	.50	Not done

TABLE 4.8b

COMBINATIONS OF ESTIMATOR FOR β_1, ASSUMED OR ESTIMATED VALUE FOR ρ, AND ESTIMATOR FOR VARIANCE IN TEST STATISTICS FOR THREE-SIB SIMULATIONS

Test Statistic	Estimator for β_1	Value Used for ρ	Estimator for $\sigma^2(1-\rho)$ or $\sigma^2(1-\rho)(1+2\rho)$, As Needed
1	$\tilde{\beta}_1$	Not needed	V_1
2	$\tilde{\beta}_1$	Not needed	V_2
3	β_1^*	1/4	$V_1(1+2\rho)$
4	β_1^*	1/4	$V_2(1+2\rho)$
5	β_1^*	1/4	$V_3(1-\rho)$
6	β_1^*	1/3	$V_1(1+2\rho)$
7	β_1^*	1/3	$V_2(1+2\rho)$
8	β_1^*	1/3	$V_3(1-\rho)$
9	β_1^*	0	V_1
10	β_1^*	0	V_2
11	β_1^*	0	V_3
12	β_1^*	"true" value	$V_1(1+2\rho)$
13	β_1^*	"true" value	$V_2(1+2\rho)$
14	β_1^*	"true" value	$V_3(1-\rho)$
15	β_1^*	0	V_4
16	$\hat{\beta}_1$	Estimated from (4.19)	Functions of V_1 and V_3 from (4.19)

Note: $\tilde{\beta}_1$, β_1^*, and $\hat{\beta}_1$ are defined in (4.22), (4.14), and (4.20), respectively.

TABLE 4.8c

OBSERVED POWER FOR THREE-SIB SIMULATIONS

Set of Genetic Parameters	Test Statistic																"True" Value of ρ
			ρ=1/4			ρ=1/3			ρ=0			ρ="True" Value			ρ=0**		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
(1)	.72	.76	.84	87	.77	.83	.85	.78	.89	.91	.59	.82	.84	.81	.82	.82	.387
(2)	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00	1.00	1.00	1.00	1.00	.304
(3)	.96	.96	.97	.98	.97	.97	.97	.97	.99	.99	.90	.97	.97	.97	.98	.97	.308
(4)	.58	.62	.69	.72	.64	.67	.69	.67	.73	.74	.50	.67	.70	.67	.66	.68	.310
(5)	.85	.86	.92	.92	.92	.90	.91	.93	.95	.95	.80	.91	.91	.92	.92	.91	.307
(6)	.71	.70	.73	.73	.69	.74	.73	.74	.79	.77	.55	.74	.73	.73	.66	.74	.317
(7)	.73	.78	.82	.82	.78	.79	.80	.81	.84	.83	.61	.80	.82	.79	.78	.80	.294
(8)	.68	.68	.75	.76	.70	.74	.76	.78	.76	.78	.53	.75	.76	.76	.72	.75	.304
(9)	.81	.79	.82	.82	.80	.83	.82	.84	.85	.84	.64	.82	.82	.82	.77	.83	.304
(1)*	.09	.09	.07	.09	.07	.08	.08	.07	.10	.11	.04	.08	.08	.08	.07	.04	.387
(2)*	.07	.07	.06	.07	.05	.05	.06	.07	.09	.09	.00	.06	.06	.06	.05	.05	.304
(3)*	.05	.06	.06	.06	.05	.05	.07	.06	.08	.09	.03	.06	.07	.05	.07	.05	.308
(4)*	.04	.04	.03	.04	.03	.03	.03	.03	.05	.05	.01	.03	.04	.03	.03	.04	.310
(5)*	.05	.06	.04	.05	.03	.03	.04	.05	.06	.06	.02	.03	.04	.05	.04	.07	.307
(6)*	.08	.09	.07	.07	.04	.06	.07	.06	.09	.10	.02	.06	.07	.06	.06	.08	.317
(7)*	.03	.04	.04	.05	.05	.04	.05	.05	.08	.08	.04	.04	.05	.05	.08	.05	.294
(8)*	.06	.08	.04	.06	.04	.04	.06	.06	.07	.08	.03	.04	.06	.04	.04	.06	.304
(9)*	.08	.08	.06	.08	.05	.07	.08	.06	.07	.07	.04	.07	.08	.06	.05	.03	.304

*λ=.5

**Haseman-Elston test for independent sib pairs

The test statistics numbered 12-14 in Table 4.8 were included for comparison purposes only; they could not be used to analyze real data since the genetic parameters will always be unknown. For the test statistics 1-11, 15, and 16, comparisons were based on average levels attained for power and significance level. The differences in these measures for the various statistics were sometimes quite small, so that the choice of the best statistics was somewhat arbitrary. The main conclusions drawn from Table 4.8c are as follows.

1. All statistics have greater power than the sib-pair test for the same total number of sibs.
2. Statistic 11 has lower power than the others; statistics 1 and 2 also have relatively low power.
3. The power is greatest for statistics 9 and 10, but these statistics also give significance levels substantially higher than the nominal .05 value.
4. The statistics using V_2 to estimate variance (2,4,7,10) tend to be very slightly higher in power than the statistics using V_1 (1,3,6,9), but the significance levels of the former are too high.
5. Statistics 3, 5, 6, 8, and 16 appear to have reasonable power, as well as about the right significance level. Of these, statistic 5 is somewhat lower on the average than the others in observed power, and statistic 8 is on the average slightly farther from the nominal .05 significance level than are statistics 3, 6, and 16.
6. The power of statistic 15, the Haseman-Elston test statistic, is only a little lower on the average than that for statistics 3, 6, 8, and 16; this result is surprising, since the Haseman-Elston statistic

ignores the correlation between squared differences for sib pairs in the same sibship. The significance level for the Haseman-Elston statistic appears reasonably close to the nominal .05, as we would expect.

Based on these preliminary simulations, reasonable choices for a three-sib test seem to be statistics 3 or 6, which estimate β_1 from (4.14), use V_1 to estimate $\sigma^2(1-\rho)$, and set $\rho=1/4$ and $1/3$, respectively. Using $\rho=1/4$ has the attraction that it is equivalent to taking

$$\beta_1^* = \frac{S_T(Y,\hat{\pi}) + S_W(Y,\hat{\pi})}{S_T(\hat{\pi},\hat{\pi}) + S_W(\hat{\pi},\hat{\pi})},$$

where the sums of squares and cross-products are as in (4.13). Statistic 16, which estimates ρ and σ^2 from the data, appears equivalent to these two but presents greater difficulty in deriving an expression for power analytically. The Haseman-Elston statistic appears to be almost as powerful as statistics 3, 6 and 16.

In the next chapter a sample size formula will be derived for a three-sib test in a manner analogous to the derivation of (2.23) in section 2.3, using the estimator β_1^* in (4.14), and the power of this and other test statistics will be examined further.

CHAPTER V

POWER OF SIB-TRIO TESTS

5.1. Sample Size Formula

In this section a sample size formula analogous to (2.23) in section 2.3 is derived. The notation will be as in Chapters II and IV if not defined explicitly in this section. We use V_1 from (4.17) as an estimator for $\sigma^2(1-\rho)$. The assumed value for the null hypothesis correlation between two different Y's will be left variable and will be denoted by ρ_0; hence the sample size required for a certain power can be studied as a function of ρ_0:

Define

$$\left.\begin{aligned} b &= \frac{(1-\rho_0)S_T(Y,\hat{\pi}) + 3\rho_0 S_W(Y,\hat{\pi})}{(1-\rho_0)S_T(\hat{\pi},\hat{\pi}) + 3\rho_0 S_W(\hat{\pi},\hat{\pi})} \\ \text{and} \quad s_b^2 &= \frac{V_1(1+2\rho_0)}{(1-\rho_0)S_T(\hat{\pi},\hat{\pi}) + 3\rho_0 S_W(\hat{\pi},\hat{\pi})} . \end{aligned}\right\} \qquad (5.1)$$

Then b and s_b^2 correspond to β_1^* in (4.14) and $V(\beta_1^*|I_m)$ in (4.15), respectively, with ρ replaced by its assumed or gussed value ρ_0 and $\sigma^2(1-\rho)$ replaced by

$$V_1 = \{S_W(Y,Y) - [S_W(Y,\hat{\pi})]^2/S_W(\hat{\pi},\hat{\pi})\}/(2n-1),$$

previously defined in (4.17). Throughout this chapter, sums of squares and cross-products of the form S(a,b) are defined as in (4.13). As in Chapter II,

the statistic b/s_b is used to test the null hypothesis $H_o:\lambda = 1/2$ at the 100 α% level by rejecting H_o if $b/s_b < -z_{1-\alpha}$, where $z_{1-\alpha}$ is the $100(1-\alpha)$% point of the standard normal distribution. Proceeding as in section 2.2, we can write an expression for the power of the test for a given alternative hypothesis H_A concerning the value of λ; rewriting the resulting expression as in (2.12), we obtain

$$z_{1-\tau} \cong \frac{-z_{1-\alpha}s_b - E(b|I_m,H_A)}{\sqrt{V(b|I_m,H_A)}}, \tag{5.2}$$

where $1-\tau$ is the power. Note that the result is conditional on the family information I_m as well as on H_A.

The asymptotic theory for functions of the Y_j in Chapter II holds now for functions of the Y_{kj} $(k=1,2,3)$, conditional on the family information I_m. For example, we can prove the asymptotic normality of

$$\sum_{j=1}^{n}\sum_{k=1}^{3} Y_{kj}$$

by letting $T_j = Y_{1j} + Y_{2j} + Y_{3j}$ and arguing as in section 2.2, replacing Y_j with T_j and replacing $\hat{\pi}_j$ and f_{1j} with the more general I_m.

Now let $U_j = Y_{1j}^2 + Y_{2j}^2 + Y_{3j}^2$, and define T_j as above. Then

$$S_W(Y,Y) = \sum_{j=1}^{n} U_j - \sum_{j=1}^{n} T_j^2/3.$$

From (4.6) we have

$$E(Y_{kj}^2|I_m) = \alpha_2 + \beta_2\hat{\pi}_{kj} + \gamma_2 f_{1kj},$$

where α_2, β_2, and γ_2 are defined in (2.9), so that

$$E(U_j|I_m) = 3\alpha_2 + 3\beta_2\bar{\hat{\pi}}_j + 3\gamma_2\bar{f}_{1j}.$$

Also, we have from (4.8) and (2.9) that

$$\sum_{k\neq k'} E(Y_{kj}Y_{k'j}|I_m) = 3\alpha_2 + 3\beta_2\bar{\hat{\pi}}_j + 3\gamma_2\bar{f}_{1j},$$

and so

$$E(T_j^2|I_m) = 6\alpha_2 + 6\beta_2\bar{\hat{\pi}}_j + 6\gamma_2\bar{f}_{1j}.$$

Then, appealing to the asymptotic normality of

$$\sum_{j=1}^{n} T_j^2$$

and

$$\sum_{j=1}^{n} U_j,$$

we have

$$S_W(Y,Y) = \sum_{j=1}^{n} E(U_j|I_m) - \left[\sum_{j=1}^{n} E(T_j^2|I_m)/3\right] + n^{1/2}u_n + o_p(n^{1/2})$$

$$= n(\alpha_2+\beta_2\bar{\bar{\pi}}+\gamma_2\bar{\bar{f}}_1) + n^{1/2}u_n + o_p(n^{1/2}),$$

where u_n and $o_p(n^{\xi})$ are as in Chapter II. In a similar manner we can show that

$$S_W(Y,\hat{\pi}) = \beta_1 S_W(\hat{\pi},\hat{\pi}) + \gamma_1 S_W(\hat{\pi},f_1) + n^{1/2}u_n + o_p(n^{1/2}),$$

where β_1 and γ_1 are as in (2.9). Then, combining the above results for $S_W(Y,Y)$ and $S_W(Y,\hat{\pi})$ and multiplying numerator and denominator of s_b^2 in (5.1) by the denominator, we have

$$s_b^2 = \frac{(1+2\rho_o)}{2} \frac{\{\alpha_2 + \beta_2\bar{\hat{\pi}} + \gamma_2\bar{\bar{f}}_1 - \frac{1}{n}[\beta_1 S_W(\hat{\pi},\hat{\pi}) + \gamma_1 S_W(\hat{\pi},f_1)]^2/S_W(\hat{\pi},\hat{\pi})\}}{[(1-\rho_o)S_T(\hat{\pi},\hat{\pi}) + 3\rho_o S_W(\hat{\pi},\hat{\pi})]^2} \times$$

$$[1-\rho_o)S_T(\hat{\pi},\hat{\pi}) + 3\rho_o S_W(\hat{\pi},\hat{\pi})][1 + n^{-1/2}u_n + o_p(n^{-1/2})]. \tag{5.3}$$

From (4.6) we have $E(Y_{kj}|I_m) = \alpha_1 + \beta_1\hat{\pi}_{kj} + f_{1kj}$, so that

$$E(b|I_m,H_A) = \beta_1 + \gamma_1 \frac{(1-\rho_o)S_T(\hat{\pi},f_1) + 3\rho_o S_W(\hat{\pi},f_1)}{(1-\rho_o)S_T(\hat{\pi},\hat{\pi}) + 3\rho_o S_W(\hat{\pi},\hat{\pi})}. \tag{5.4}$$

Now let

$$\ell_{kj} = (1-\rho_o)(\hat{\pi}_{kj} - \bar{\hat{\pi}}) + 3\rho_o(\hat{\pi}_{kj} - \bar{\hat{\pi}}_j). \tag{5.5}$$

It follows from (5.1) and (4.13) that

$$b = \sum_{j=1}^{n} \sum_{k=1}^{3} Y_{kj}\ell_{kj} / [(1-\rho_o)S_T(\hat{\pi},\hat{\pi}) + 3\rho_o S_W(\hat{\pi},\hat{\pi})]. \tag{5.6}$$

It follows easily that

$$V(b|I_m,H_A) = \frac{\{\sum_{j=1}^{n} \sum_{k=1}^{3} \sum_{k'=1}^{3} E(Y_{kj}Y_{k'j}|I_m)\ell_{kj}\ell_{k'j} - \sum_{j=1}^{n} [\sum_{k=1}^{3} E(Y_{kj}|I_m)\ell_{kj}]^2\}}{[(1-\rho_o)S_T(\hat{\pi},\hat{\pi}) + 3\rho_o S_W(\hat{\pi},\hat{\pi})]^2} \cdot \tag{5.7}$$

Substituting from (4.6) for $E(Y_{kj}|I_m)$ and $E(Y_{kj}^2|I_m)$ and from (4.8) for $E(Y_{kj}Y_{k'j}|I_m)$, using the definition of α_2 in (2.9), and rearranging terms, we obtain the following expression for the numerator of (5.7):

$$\sum_{j=1}^{n}\sum_{k=1}^{3}\ell_{kj}^2(\alpha_2-\alpha_1^2) + \sum_{j=1}^{n}\sum_{k\neq k'}\ell_{kj}\ell_{k'j}[(\alpha_2/2) - \alpha_1^2 - (1-2\psi)^2(8p^2q^2)(a^4-a^2d^2)] + \sum_{j=1}^{n}\sum_{k=1}^{3}\hat{\pi}_{kj}\ell_{kj}^2(\beta_2-2\alpha_1\beta_1)$$

$$+ \sum_{j=1}^{n}\sum_{k\neq k'}\hat{\pi}_{kj}\ell_{kj}\ell_{k'j}\{-2\alpha_1\beta_1 + (1-2\psi)[8\sigma_e^2\sigma_g^2 + 4(p^2q(a-d)^4 + pq^2(a+d)^4) + p^2q^2(16a^4-40a^2d^2-8d^4)]$$
$$+ (1-2\psi)^2(8p^2q^2)(a^4-a^2d^2)\}$$

$$+ \sum_{j=1}^{n}\sum_{k=1}^{3}f_{1kj}\ell_{kj}^2(\gamma_2-2\alpha_1\gamma_1) + \sum_{j=1}^{n}\sum_{k\neq k'}f_{1kj}\ell_{kj}\ell_{k'j}\{-2\alpha_1\gamma_1 + (1-2\psi)^2[4\sigma_e^2\sigma_d^2 + 4p^2q^2(a^2+d^2)^2]\}$$

$$- \beta_1^2\sum_{j=1}^{n}\sum_{k=1}^{3}\hat{\pi}_{kj}^2\ell_{kj}^2 \quad - \gamma_1^2\sum_{j=1}^{n}\sum_{k=1}^{3}f_{1kj}^2\ell_{kj}^2 \quad - 2\beta_1\gamma_1\sum_{j=1}^{n}\sum_{k=1}^{3}\hat{\pi}_{kj}f_{1kj}\ell_{kj}^2 \qquad (5.8)$$

$$- \beta_1^2\sum_{j=1}^{n}\sum_{k\neq k'}\hat{\pi}_{kj}\hat{\pi}_{k'j}\ell_{kj}\ell_{k'j} - \gamma_1^2\sum_{j=1}^{n}\sum_{k\neq k'}f_{1kj}f_{1k'j}\ell_{kj}\ell_{k'j} - 2\beta_1\gamma_1\sum_{j=1}^{n}\sum_{k\neq k'}\hat{\pi}_{kj}f_{1k'j}\ell_{kj}\ell_{k'j}$$

$$+ \sum_{j=1}^{n}\sum_{k\neq k'\neq k''}\ell_{kj}\ell_{k'j}\hat{\pi}_{k''j}\{(1-2\psi)[-2\sigma_e^2\sigma_g^2 - 2(p^2q(a-d)^4 + pq^2(a+d)^4) + 4p^2q^2(-a^4+4a^2d^2+d^4)]$$
$$+ (1-2\psi)^2(8p^2q^2)(-a^4+a^2d^2)\}$$

$$+ \sum_{j=1}^{n}\sum_{k\neq k'\neq k''}\ell_{kj}\ell_{k'j}f_{1k''j}(1-2\psi)^2[-\sigma_e^2\sigma_d^2 + 2p^2q^2(a^4-4a^2d^2-d^4)] + \sum_{j=1}^{n}\sum_{k\neq k'}\ell_{kj}\ell_{k'j}h_{222,j}(1-2\psi)^2(8p^2q^2)(a^4-a^2d^2)$$

$$+ \sum_{j=1}^{n}(\ell_{1j}\ell_{2j}h_{002,j} + \ell_{1j}\ell_{3j}h_{020,j} + \ell_{2j}\ell_{3j}h_{200,j})(1-2\psi)^2(48p^2q^2)(a^4-a^2d^2).$$

As in section 2.3, we derive asymptotic representations for relevant functions of the $\hat{\pi}$'s and f_1's in (5.3), (5.4), and (5.8) and then substitute into (5.2) to obtain a formula for sample size. First we note that, since the numbering of sib pairs within a sibship is arbitrary, there is symmetry in expected values of $\hat{\pi}$'s and f_1's with different subscripts; for example,

$$E(\hat{\pi}_1) = E(\hat{\pi}_2) = E(\hat{\pi}_3),$$

$$E(\hat{f}_{11}) = E(\hat{f}_{12}) = E(\hat{f}_{13}),$$

$$E(\hat{\pi}_1\hat{\pi}_2) = E(\hat{\pi}_1\hat{\pi}_3) = E(\hat{\pi}_2\hat{\pi}_3).$$

This symmetry is easily demonstrated from Table 4.1b for the case of no dominance and complete parental information at the marker locus. (The subscript j, indicating the jth sibship, is unnecessary in the expectations, since we assume that the h's , $\hat{\pi}$'s, and f_1's have the same distribution for all sibships).

Proceeding as in the proof of Lemma 2.1, the following results are easily shown:

$$\left.\begin{aligned}
\bar{\hat{\pi}} &= E(\hat{\pi}_1) + n^{-1/2}u_n + o_p(n^{-1/2}) \\
\bar{f}_1 &= E(f_1) + n^{-1/2}u_n + o_p(n^{-1/2}) \\
S_T(\hat{\pi},\hat{\pi}) &= 3n\{E(\hat{\pi}_1^2) - [E(\hat{\pi}_1)]^2\} + n^{1/2}u_n + o_p(n^{1/2}) \\
S_W(\hat{\pi},\hat{\pi}) &= 2n[E(\hat{\pi}_1^2) - E(\hat{\pi}_1\hat{\pi}_2)] + n^{1/2}u_n + o_p(n^{1/2}) \\
S_T(\hat{\pi},f_1) &= 3n[E(\hat{\pi}_1 f_{11}) - E(\hat{\pi}_1)E(f_{11})] + n^{1/2}u_n + o_p(n^{1/2}) \\
S_W(\hat{\pi},f_1) &= 2n[E(\hat{\pi}_1 f_{11}) - E(\hat{\pi}_1 f_{12})] + n^{1/2}u_n + o_p(n^{1/2}).
\end{aligned}\right\} \qquad (5.9)$$

Similarly, we get the following results for the summations in (5.8):

$$\left.\begin{aligned}
&\sum_{j=1}^{n}\sum_{k=1}^{3} \ell_{kj}^2 = nw_1 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k\neq k'} \ell_{kj}\ell_{k'j} = nw_2 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k=1}^{3} \hat{\pi}_{kj}\ell_{kj}^2 = nw_3 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k\neq k'} \hat{\pi}_{kj}\ell_{kj}\ell_{k'j} = nw_4 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k=1}^{3} f_{1kj}\ell_{kj}^2 = nw_5 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k\neq k'} f_{1kj}\ell_{kj}\ell_{k'j} = nw_6 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k=1}^{3} \hat{\pi}_{kj}^2\ell_{kj}^2 = nw_7 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k\neq k'} \hat{\pi}_{kj}\hat{\pi}_{k'j}\ell_{kj}\ell_{k'j} = nw_8 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k=1}^{3} f_{1kj}^2\ell_{kj}^2 = nw_9 + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k\neq k'} f_{1kj}f_{1k'j}\ell_{kj}\ell_{k'j} = nw_{10} + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k=1}^{3} \hat{\pi}_{kj}f_{1kj}\ell_{kj}^2 = nw_{11} + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k\neq k'} \hat{\pi}_{kj}f_{1k'j}\ell_{kj}\ell_{k'j} = nw_{12} + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k\neq k'\neq k''} \ell_{kj}\ell_{k'j}\hat{\pi}_{k''j} = nw_{13} + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}\sum_{k\neq k'\neq k''} \ell_{kj}\ell_{k'j}f_{1k''j} = nw_{14} + n^{1/2}u_n + o_p(n^{1/2})
\end{aligned}\right\} \qquad (5.10)$$

$$\left.\begin{aligned}
&\sum_{j=1}^{n}\sum_{k\neq k'}\ell_{kj}\ell_{k'j}h_{222,j} = nw_{15} + n^{1/2}u_n + o_p(n^{1/2}) \\
&\sum_{j=1}^{n}(\ell_{1j}\ell_{2j}h_{002,j} + \ell_{1j}\ell_{3j}h_{020,j} + \ell_{2j}\ell_{3j}h_{200,j}) = nw_{16} + n^{1/2}u_n + o_p(n^{1/2}),
\end{aligned}\right\} \quad (5.10)\ (\text{cont.})$$

where

$$\left.\begin{aligned}
w_1 &= 3\{(1+2\rho_o+3\rho_o^2)E(\hat{\pi}_1^2) - (1-\rho_o)^2[E(\hat{\pi}_1)]^2 - 2\rho_o(2+\rho_o)E(\hat{\pi}_1\hat{\pi}_2)\} \\
w_2 &= 6\{-\rho_o(2+\rho_o)E(\hat{\pi}_1^2) - (1-\rho_o)^2[E(\hat{\pi}_1)]^2 + (1+2\rho_o^2)E(\hat{\pi}_1\hat{\pi}_2)\} \\
w_3 &= 3\{(1+\rho_o)^2E(\hat{\pi}_1^3) - 2(1-\rho_o)(1+\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1^2) + (1-\rho_o)^2[E(\hat{\pi}_1)]^3 \\
&\quad - 2\rho_o(2+\rho_o)E(\hat{\pi}_1^2\hat{\pi}_2) + 4\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1\hat{\pi}_2) + 2\rho_o^2E(\hat{\pi}_1\hat{\pi}_2\hat{\pi}_3)\} \\
w_4 &= 6\{-\rho_o(1+\rho_o)E(\hat{\pi}_1^3) - (1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1^2) + (1-\rho_o)^2[E(\hat{\pi}_1)]^3 \\
&\quad + (1+2\rho_o^2)E(\hat{\pi}_1^2\hat{\pi}_2) - (1-\rho_o)(1-2\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1\hat{\pi}_2) - \rho_oE(\hat{\pi}_1\hat{\pi}_2\hat{\pi}_3)\} \\
w_5 &= 3\{(1+\rho_o)^2E(\hat{\pi}_1^2f_{11}) - 2(1-\rho_o)(1+\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{11}) + (1-\rho_o)^2[E(\hat{\pi}_1)]^2E(f_{11}) \\
&\quad - 4\rho_o(1+\rho_o)E(\hat{\pi}_1f_{11}\hat{\pi}_2) + 2\rho_o^2E(\hat{\pi}_1^2f_{12}) + 4\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{12}) \\
&\quad + 2\rho_o^2E(\hat{\pi}_1\hat{\pi}_2f_{13})\} \\
w_6 &= 6\{-\rho_o(1+\rho_o)E(\hat{\pi}_1^2f_{11}) - (1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{11}) + (1-\rho_o)^2[E(\hat{\pi}_1)]^2E(f_{11}) \\
&\quad + (1+\rho_o+2\rho_o^2)E(\hat{\pi}_1f_{11}\hat{\pi}_2) - \rho_oE(\hat{\pi}_1^2f_{12}) - (1-\rho_o)(1-2\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{12}) \\
&\quad - \rho_oE(\hat{\pi}_1\hat{\pi}_2f_{13})\} \\
w_7 &= 3\{(1+\rho_o)^2E(\hat{\pi}_1^4) - 2(1-\rho_o)(1+\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1^3) + (1-\rho_o)^2[E(\hat{\pi}_1)]^2E(\hat{\pi}_1^2) \\
&\quad - 4\rho_o(1+\rho_o)E(\hat{\pi}_1^3\hat{\pi}_2) + 2\rho_o^2E(\hat{\pi}_1^2\hat{\pi}_2^2) + 4\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1^2\hat{\pi}_2) \\
&\quad + 2\rho_o^2E(\hat{\pi}_1^2\hat{\pi}_2\hat{\pi}_3)\}
\end{aligned}\right\} \quad (5.11)$$

$$w_8 = 6\{-2\rho_o(1+\rho_o)E(\hat{\pi}_1^3\hat{\pi}_2) + (1+2\rho_o+2\rho_o^2)E(\hat{\pi}_1^2\hat{\pi}_2^2) - 2(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1^2\hat{\pi}_2)$$
$$+ (1-\rho_o)^2[E(\hat{\pi}_1)]^2E(\hat{\pi}_1\hat{\pi}_2) - \rho_o(2-\rho_o)E(\hat{\pi}_1^2\hat{\pi}_2\hat{\pi}_3)$$
$$+ 2\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1\hat{\pi}_2\hat{\pi}_3)\}$$

$$w_9 = 3\{(1+\rho_o)^2E(\hat{\pi}_1^2f_{11}^2) - 2(1-\rho_o)(1+\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{11}^2) + (1-\rho_o)^2E(\hat{\pi}_1)]^2E(f_{11}^2)$$
$$- 4\rho_o(1+\rho_o)E(\hat{\pi}_1f_{11}^2\hat{\pi}_2) + 2\rho_o^2E(\hat{\pi}_1^2f_{12}^2) + 4\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{12}^2)$$
$$+ 2\rho_o^2E(\hat{\pi}_1\hat{\pi}_2f_{13}^2)\}$$

$$w_{10} = 6\{-2\rho_o(1+\rho_o)E(\hat{\pi}_1^2f_{11}f_{12}) - 2(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{11}f_{12})$$
$$+ (1-\rho_o)^2[E(\hat{\pi}_1)]^2E(f_{11}f_{12}) + (1+2\rho_o+2\rho_o^2)E(\hat{\pi}_1f_{11}\hat{\pi}_2f_{12})$$
$$- 2\rho_oE(\hat{\pi}_1f_{11}\hat{\pi}_2f_{13}) + \rho_o^2E(\hat{\pi}_1^2f_{12}f_{13}) + 2\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{12}f_{13})\}$$

$$w_{11} = 3\{(1+\rho_o)^2E(\hat{\pi}_1^3f_{11}) - 2(1-\rho_o)(1+\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1^2f_{11})$$
$$+ (1-\rho_o)^2[E(\hat{\pi}_1)]^2E(\hat{\pi}_1f_{11}) - 4\rho_o(1+\rho_o)E(\hat{\pi}_1^2f_{11}\hat{\pi}_2)$$
$$+ 4\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{11}\hat{\pi}_2) + 2\rho_o^2E(\hat{\pi}_1^2\hat{\pi}_2f_{12}) + 2\rho_o^2E(\hat{\pi}_1f_{11}\hat{\pi}_2\hat{\pi}_3)\}$$

$$w_{12} = 6\{-\rho_o(1+\rho_o)E(\hat{\pi}_1^2f_{11}\hat{\pi}_2) - \rho_o(1+\rho_o)E(\hat{\pi}_1^3f_{12}) - (1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1^2f_{12})$$
$$+ (1-\rho_o)^2[E(\hat{\pi}_1)]^2E(\hat{\pi}_1f_{12}) + (1+2\rho_o+2\rho_o^2)E(\hat{\pi}_1^2\hat{\pi}_2f_{12})$$
$$- (1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1f_{11}\hat{\pi}_2) - \rho_oE(\hat{\pi}_1f_{11}\hat{\pi}_2\hat{\pi}_3) - \rho_o(1-\rho_o)E(\hat{\pi}_1^2\hat{\pi}_2f_{13})$$
$$+ 2\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1\hat{\pi}_2f_{13})\}$$

$$w_{13} = 6\{\rho_o^2E(\hat{\pi}_1^3) + 2\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1^2) + (1-\rho_o)^2[E(\hat{\pi}_1)]^3 - 2\rho_o(2+\rho_o)E(\hat{\pi}_1^2\hat{\pi}_2)$$
$$- 2(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1\hat{\pi}_2) + (1+2\rho_o+2\rho_o^2)E(\hat{\pi}_1\hat{\pi}_2\hat{\pi}_3)\}$$

(5.11) (cont.)

$$
\left.
\begin{aligned}
w_{14} &= 6\{\rho_o^2 E(\hat{\pi}_1^2 f_{11}) + 2\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1 f_{11}) + (1-\rho_o)^2[E(\hat{\pi}_1)]^2 E(f_{11}) \\
&\quad - 2\rho_o E(\hat{\pi}_1 f_{11}\hat{\pi}_2) - 2\rho_o(1+\rho_o)E(\hat{\pi}_1^2 f_{12}) - 2(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1 f_{12}) \\
&\quad + (1+2\rho_o+2\rho_o^2)E(\hat{\pi}_1\hat{\pi}_2 f_{13})\} \\
w_{15} &= 6\{-\rho_o(2+\rho_o)E(\hat{\pi}_1^2 h_{222}) - 2(1-\rho_o)^2 E(\hat{\pi}_1)E(\hat{\pi}_1 h_{222}) \\
&\quad + (1-\rho_o)^2[E(\hat{\pi}_1)]^2 E(h_{222}) + (1+2\rho_o^2)E(\hat{\pi}_1\hat{\pi}_2 h_{222})\} \\
w_{16} &= 3\{\rho_o^2 E(\hat{\pi}_1^2 h_{200}) + 2\rho_o(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1 h_{200}) \\
&\quad + (1-\rho_o)^2[E(\hat{\pi}_1)]^2 E(h_{200}) - 2\rho_o(1+\rho_o)E(\hat{\pi}_1^2 h_{020}) \\
&\quad - 2(1-\rho_o)E(\hat{\pi}_1)E(\hat{\pi}_1 h_{020}) - 2\rho_o E(\hat{\pi}_1\hat{\pi}_2 h_{200}) \\
&\quad + (1+2\rho_o+2\rho_o^2)E(\hat{\pi}_1\hat{\pi}_2 h_{002})\}.
\end{aligned}
\right\} \quad (5.11)\ (\text{cont.})
$$

Substituting from (5.9) and (5.10) into (5.3), (5.4), (5.7), and (5.8), we have

$$
\left.
\begin{aligned}
s_b^2 &= nU[1+n^{-1/2}u_n + o_p(n^{-1/2})]/D^2 \\
E(b|I_m,H_A) &= nQ[1+n^{-1/2}u_n + o_p(n^{-1/2})]/D \\
V(b|I_m,H_A) &= nW[1+n^{-1/2}u_n + o_p(n^{-1/2})]/D^2,
\end{aligned}
\right\} \quad (5.12)
$$

where

$$D = (1-\rho_o)S_T(\hat{\pi},\hat{\pi}) + 3\rho_o S_W(\hat{\pi},\hat{\pi})$$

$$U = 3(1+2\rho_o)[(1+\rho_o)V(\hat{\pi}_1) - 2\rho_o \mathrm{cov}(\hat{\pi}_1,\hat{\pi}_2)]/2$$

$$\times\Big[\alpha_2+\beta_2 E(\hat{\pi}_1) + \gamma_2 E(f_{11})$$

$$- \frac{2\{\beta_1[V(\hat{\pi}_1) - \mathrm{cov}(\hat{\pi}_1,\hat{\pi}_2)] + \gamma_1[\mathrm{cov}(\hat{\pi}_1,f_{11}) - \mathrm{cov}(\hat{\pi}_1,f_{12})]\}^2}{[V(\hat{\pi}_1) - \mathrm{cov}(\hat{\pi}_1,\hat{\pi}_2)]}\Big]$$

$$Q = 3\{\beta_1[(1+\rho_o)V(\hat{\pi}_1) - 2\rho_o \mathrm{cov}(\hat{\pi}_1,\hat{\pi}_2)] + \gamma_1[(1+\rho_o)\mathrm{cov}(\hat{\pi}_1,f_{11}) - 2\rho_o \mathrm{cov}(\hat{\pi}_1,f_{12})]\}$$

$$\begin{aligned} W = {} & w_1(\alpha_2-\alpha_1^2) + w_2[(\alpha_2/2)-\alpha_1^2-(1-2\psi)^2(8p^2q^2)(a^4-a^2d^2)] \\ & + w_3(\beta_2-2\alpha_1\beta_1) + w_4\{-2\alpha_1\beta_1 + (1-2\psi)[8\sigma_e^2\sigma_g^2+4(p^2q(a-d)^4 \\ & + pq^2(a+d)^4) + p^2q^2(16a^4-40a^2d^2-8d^4)] \\ & + (1-2\psi)^2(8p^2q^2)(a^4-a^2d^2)\} \\ & + w_5(\gamma_2-2\alpha_1\gamma_1) + w_6\{-2\alpha_1\gamma_1 + (1-2\psi)^2[4\sigma_e^2\sigma_d^2 + 4p^2q^2(a^2+d^2)^2]\} \\ & - (w_7+w_8)\beta_1^2 - (w_9+w_{10})\gamma_1^2 - (w_{11}+w_{12})(2\beta_1\gamma_1) \\ & + w_{13}\{(1-2\psi)[-2\sigma_e^2\sigma_g^2-2(p^2q(a-d)^4 + pq^2(a+d)^4) \\ & + p^2q^2(-4a^4+16a^2d^2+4d^4)] + (1-2\psi)^2(8p^2q^2)(-a^4+a^2d^2)\} \\ & + w_{14}(1-2\psi)^2[-\sigma_e^2\sigma_d^2+p^2q^2(2a^4-8a^2d^2-2d^4)] \\ & + w_{15}(1-2\psi)^2(8p^2q^2)(a^4-a^2d^2) + w_{16}(1-2\psi)^2(48p^2q^2)(a^4-a^2d^2) \end{aligned} \qquad (5.13)$$

and the w_i's ($i=1, \ldots, 16$) are defined in (5.11). Substituting from (5.12) into (5.2) and performing some obvious algebraic manipulation

yields

$$\sqrt{n} = -(z_{1-\tau}W^{1/2}+z_{1-\alpha}U^{1/2}+u_n)[1+n^{-1/2}u_n+o_p(n^{-1/2})]/Q.$$

For sufficiently large n we can ignore the terms $n^{-1/2}u_n$ and $o_p(n^{-1/2})$; taking the expectation of the resulting expression and solving for n, we have

$$n = [z_{1-\tau}W^{1/2}+z_{1-\alpha}U^{1/2}]^2/Q^2. \tag{5.14}$$

The sample size formula (5.14) is the analogue of (2.23) for sibships of size three. That is, (5.14) gives the approximate number of independent sib trios required for the test statistic b/s_b from (5.1) to have power $(1-\tau)$ when the null hypothesis $H_o:\lambda = 1/2$ is tested at level α, given the assumed value ρ_o, the distribution of the probabilities h_{ijk} of genes identical by descent at the marker locus, and the parameters σ_e^2,a,d, and p at the trait locus. We now compare some results from (5.14) with the results for sib pairs in section 2.4.

5.2. Asymptotic Power-Results

We assume no dominance and complete parental information at a two-allele marker locus. From Table 4.1 we derive the moments required to determine sample size, as follows.

$$\left.\begin{aligned} E(\hat{\pi}_1) &= 1/2 \\ E(\hat{\pi}_1^2) &= [1+uv(1-uv)]/4 \\ V(\hat{\pi}_1) &= uv(1-uv)/4 \end{aligned}\right\} \tag{5.15}$$

$$E(\hat{\pi}_1^3) = [1+3uv(1-uv)]/8$$

$$E(\hat{\pi}_1^4) = 1/16 + 25uv/64 - 11u^2v^2/32$$

$$E(f_{11}) = 1/2$$

$$E(f_{11}^2) = (1+4u^2v^2)/4$$

$$V(f_{11}) = u^2v^2$$

$$E(\hat{\pi}_1 f_{11}) = 1/4$$

$$Cov(\hat{\pi}_1, f_{11}) = 0$$

$$E(\hat{\pi}_1^2 f_{11}) = [1+uv(1-2uv)]/8$$

$$E(\hat{\pi}_1^3 f_{11}) = [1+3uv(1-2uv)]/16$$

$$E(\hat{\pi}_1 f_{11}^2) = (1+4u^2v^2)/8$$

$$E(\hat{\pi}_1^2 f_{11}^2) = [1+uv(1+2uv)]/16$$

$$E(\hat{\pi}_1 \hat{\pi}_2) = 1/4$$

$$E(\hat{\pi}_1^2 \hat{\pi}_2) = [1+uv(1-uv)]/8$$

$$E(\hat{\pi}_1^3 \hat{\pi}_2) = [1+3uv(1-uv)]/16$$

$$E(\hat{\pi}_1^2 \hat{\pi}_2^2) = 1/16 + 9uv/64 - u^2v^2/8$$

$$E(f_{11} f_{12}) = 1/4$$

$$E(\hat{\pi}_1 f_{12}) = 1/4$$

$$E(\hat{\pi}_1^2 f_{12}) = [1+uv(1-uv)]/8$$

$$E(\hat{\pi}_1^3 f_{12}) = [1+3uv(1-uv)]/16$$

$$E(\hat{\pi}_1 f_{12}^2) = (1+4u^2v^2)/8$$

$$E(\hat{\pi}_1^2 f_{12}^2) = [1+uv(1+4uv)]/16$$

(5.15) (cont.)

$$E(\hat{\pi}_1\hat{\pi}_2 f_{11}) = 1/8$$

$$E(\hat{\pi}_1^2\hat{\pi}_2 f_{11}) = [1+uv(1-2uv)]/16$$

$$E(\hat{\pi}_1^3\hat{\pi}_2 f_{11}) = [1+3uv(1-2uv)]/32$$

$$E(\hat{\pi}_1\hat{\pi}_2 f_{11}^2) = (1+4u^2v^2)/16$$

$$E(\hat{\pi}_1^2\hat{\pi}_2 f_{12}) = [1+uv(1-uv)]/16$$

$$E(\hat{\pi}_1 f_{11} f_{12}) = 1/8$$

$$E(\hat{\pi}_1^2 f_{11} f_{12}) = [1+uv(1-2uv)]/16$$

$$E(\hat{\pi}_1\hat{\pi}_2 f_{11} f_{12}) = 1/16$$

$$E(\hat{\pi}_1\hat{\pi}_2\hat{\pi}_3) = 1/8 + uv(1-uv)/16$$

$$E(\hat{\pi}_1^2\hat{\pi}_2\hat{\pi}_3) = [1+2uv(1-uv)]/16$$

$$E(\hat{\pi}_1\hat{\pi}_2 f_{13}) = 1/8$$

$$E(\hat{\pi}_1^2\hat{\pi}_2 f_{13}) = [1+uv(1-uv)]/16$$

$$E(\hat{\pi}_1\hat{\pi}_2 f_{13}^2) = (1+4u^2v^2)/16$$

$$E(\hat{\pi}_1 f_{12} f_{13}) = 1/8$$

$$E(\hat{\pi}_1^2 f_{12} f_{13}) = (1+uv)/16$$

$$E(\hat{\pi}_1\hat{\pi}_2\hat{\pi}_3 f_{11}) = 1/16 + uv(1-2uv)/32$$

$$E(\hat{\pi}_1\hat{\pi}_2 f_{11} f_{13}) = 1/16$$

$$E(h_{222}) = 1/16$$

$$E(\hat{\pi}_1 h_{222}) = [1+2uv(1-uv)]/32$$

$$E(\hat{\pi}_1^2 h_{222}) = [1+uv(5-4uv)]/64$$

$$E(\hat{\pi}_1\hat{\pi}_2 h_{222}) = [1+uv(5-4uv)]/64$$

(5.15) (cont.)

$$
\left.
\begin{aligned}
E(h_{200}) &= 1/16 \\
E(\hat{\pi}_1 h_{200}) &= [1+2uv(1-uv)]/32 \\
E(\hat{\pi}_1^2 h_{200}) &= [1+uv(5-4uv)]/64 \\
E(\hat{\pi}_1 h_{020}) &= [1-2uv(1-uv)]/32 \\
E(\hat{\pi}_1^2 h_{020}) &= [1-uv(3-4uv)]/64 \\
E(\hat{\pi}_1 \hat{\pi}_2 h_{200}) &= (1-uv)/64 \\
E(\hat{\pi}_1 \hat{\pi}_2 h_{002}) &= [1-uv(3-4uv)]/64.
\end{aligned}
\right\} \quad (5.15) \text{ (cont.)}
$$

Sample sizes calculated from (5.14) are given in Table 5.1, using the moments in (5.15), for $\alpha = .05$, power $= .90$, $\rho_o = .25$, $\lambda = 0$, $u = .5$, , $h^2 = .1(.1).9$,

(i) $p = .5$ for no dominance at the trait locus,

(ii) $p = .1(.2).9$ for complete dominance at the trait locus.

As was done in Table 2.5, simple rules are given for obtaining approximate sample sizes for $d = 0$ and $d = a$, $p = .1(.2).9$, $u = .1(.2).9$, $h^2 = .1(.1).9$, and $\lambda = 0(.1).2$; approximate sample sizes calculated using these rules agree with sample sizes calculated from (5.14) to an extent similar to that reported for the calculating rules in Table 2.5.

TABLE 5.1

SIB-TRIO LINKAGE TEST: SAMPLE SIZE (NUMBER OF SIB TRIOS) REQUIRED FOR 90% POWER AT α=.05, ASSUMING NO DOMINANCE AND COMPLETE PARENTAL INFORMATION AT MARKER LOCUS

$n = n(h^2,\lambda,\rho,u)$ = number of sib trios; $\lambda = 0, u = .5, \rho_o = .25$.

	No Dominance at Trait Locus (d=0)	Complete Dominance at Trait Locus (d=a)				
	p=.5	p=.1	p=.3	p=.5	p=.7	p=.9
h^2	n	n	n	n	n	n
.1	10053	10116	10135	10226	10431	12135
.2	2297	2347	2336	2389	2546	4170
.3	930	975	955	996	1135	2730
.4	475	518	493	527	658	2237
.5	275	317	289	319	445	2015
.6	172	213	183	211	333	1897
.7	113	154	123	149	268	1828
.8	78	117	86	110	228	1785
.9	55	94	62	85	201	1756

To approximate $n(h^2,\lambda,p,u)$ for $\lambda = 0(.1).2$, $u = .1(.2).9$, $h^2 = .1(.1).9$:

I. For d = 0

1. To approximate $n(h^2,0,p,.5)$, add to $n(h^2,0,.5,.5)$ (first column of table):

$$\begin{cases} 5 \text{ for } p = .3 \text{ or } .7 \\ 49 \text{ for } p = .1 \text{ or } .9 \end{cases}$$

2. To approximate $n(h^2,0,p,u)$, multiply $n(h^2,0,p,.5)$ obtained in step 1 by

$$\begin{cases} 1.13 \text{ for } u = .3 \text{ or } .7 \\ 2.29 \text{ for } u = .1 \text{ or } .9 \end{cases}$$

3. To approximate $n(h^2,\lambda,p,u)$, multiply $n(h^2,0,p,u)$ obtained in step 2 by

$$\begin{cases} (2.43-.1h^2) \text{ for } \lambda = .1 \\ (7.71-.7h^2) \text{ for } \lambda = .2. \end{cases}$$

II. For d = a

1. To approximate $n(h^2,0,p,u)$, multiply $n(h^2,0,p,.5)$ (last five columns of the table) by

$$\begin{cases} 1.14 \text{ for } u = .3 \text{ or } .7 \\ 2.34 \text{ for } u = .1 \text{ or } .9 \end{cases}$$

2. To approximate $n(h^2,\lambda,p,u)$, multiply $n(h^2,0,p,u)$ obtained in step 1 by

$$\begin{cases} 2.43-.1h^2 \text{ for } \lambda = .1 \\ 7.71-.7h^2 \text{ for } \lambda = .2. \end{cases}$$

We have previously seen from the simulations in Chapter IV that various sib-trio linkage tests are considerably more powerful than the Haseman-Elston sib-pair test for the same total number of sibs. From Tables 5.1 and 2.5 (in general, from (5.14) and (2.23)) we can compare, for the particular sib-trio test chosen, the approximate sample sizes required for it and the sib-pair test to have the same power. Ratios of numbers of sib trios in Table 5.1 to the corresponding numbers of sib pairs in Table 2.5 are on the order of one third, ranging from .28 to .35 (.32 to .35 for numbers of sib trios less than 500 in Table 5.1). Thus the total number of sibs required for the sib-trio test used in

Table 5.1 is about half the number required for the sib-pair test, under the assumptions of Tables 5.1 and 2.5. Sample sizes have also been calculated from (5.14) for the combinations of α and power indicated in Table 2.6; the results, expressed as percentages of the sample size for $\alpha = .05$ and power = .90, are quite similar to those for Table 2.6. Hence the total number of sibs required for the sib-trio test is on the order of half the number of sibs required for the Haseman-Elston test on independent sib pairs for a range of values of α and power.

Calculations for $\rho_o = 0, .25$, and .33 indicate that sample size calculated from (5.14) is a monotonically increasing function of ρ_o, at least for ρ_o ranging from 0 to .33. However, the dependence on ρ_o is not great; sample sizes calculated for $\rho_o = .33$ are slightly larger and those calculated for $\rho_o = 0$ are slightly smaller than those for $\rho_o = .25$. For example, (5.14) yields sample sizes 51, 55, and 56 for $\rho_o = 0, .25$, and .33, respectively, for the case $\lambda = 0$, $p = .5$, $u = .5$, $h^2 = .9$. It has been noted previously in section 4.3 that the use of $\rho_o = 0$ tends to result in a higher significance level that the nominal value.

As would be expected from comparing Tables 5.1 and 2.5 (and the approximating rules included with them) and from the above discussion, the effects of changes in p,u,d, and λ on sample sizes calculated from (5.14) are similar to those reported in section 2.3 for samples of sib pairs calculated from (2.23). Sample sizes have not been calculated from (5.14) for cases other than that of no dominance and complete parental information at the marker locus; however, as with samples of sib pairs, sample sizes for corresponding values of $p, u, h^2, \lambda, \alpha$, and

power for other assumptions about dominance and information at the marker locus must be larger than those shown in Table 5.1.

In addition to the simulations described in section 4.3, on which the choice of a particular sib-trio test was based, further simulations have been done to check the power and significance level of several sib-trio tests, using sample sizes calculated from (5.14). Table 5.2, which is analogous to Table 3.1 for sib pairs, gives the results of these simulations for selected examples. The four test statistics shown correspond to statistics 3,6,16,and 15, respectively, in Table 4.8b; the first statistic, using $\rho_o = .25$, is the one for which the sample sizes were calculated. The examples were chosen to include the smallest sample sizes which were calculated from (5.14) for $\sigma_e^2 = 1$, $\alpha = .05$, and power = .50 and .90, as well as one larger sample size. (The ranges of other parameters used in the calculations are $p = .1(.2).9$, $u = .1(.2).5$, $\lambda = 0(.1).2$, and $h^2 = .1(.1).9$.) The first nine rows of Table 5.2 indicate good agreement between theoretical and observed power for all the statistics shown; the power of the Haseman-Elston test on all possible sib pairs is only slightly less on the average than the power for the other three. These results, along with the first nine rows of Table 4.8c, indicate that the sample size required for a certain power is about the same for all four statistics. The last nine rows of Table 5.2 indicate that the significance level is fairly close to the nominal level for each of the four statistics; the observed significance levels averaged .05 to .06, compared to a nominal level of .05. These results also are consistent with those in Table 4.8c. It should be noted that the assumption of normality of individual sibs' trait values was made in both sets of simulations.

TABLE 5.2

THEORETICAL AND OBSERVED POWER FROM COMPUTER SIMULATIONS OF A LINKAGE TEST ON SIB TRIOS, ASSUMING NO DOMINANCE AND COMPLETE PARENTAL INFORMATION AT THE MARKER LOCUS

For each sample size (number of sibships of size three), 100 samples were drawn; $\alpha = .05$, $\sigma_e^2 = 1$.

p	u	λ	a	d	h^2	ρ (True Value)	Sample Size	Theoretical Power*	Observed Power			
									$\rho_o=.25$*	$\rho_o=.33$**	ρ,σ^2 est.***	H-E†
.5	.5	0	3	0	.90	.286	55	.90	.88	.87	.87	.86
.7	.3	0	3.27	0	.90	.307	67	.90	.97	.97	.97	.95
.7	.5	0	3.27	0	.90	.307	60	.90	.92	.91	.91	.90
.5	.3	0	3	0	.90	.286	61	.90	.84	.84	.84	.85
.5	.5	.1	3	0	.90	.286	128	.90	.87	.87	.88	.85
.3	.5	0	2.12	2.12	.90	.297	62	.90	.83	.81	.83	.84
.3	.5	0	3.27	0	.90	.307	15	.50	.62	.57	.60	.60
.5	.3	0	3	0	.90	.286	16	.50	.46	.44	.46	.43
.5	.5	0	3	0	.90	.286	14	.50	.49	.49	.52	.41
.5	.5	.5	3	0	.90	.286	55	.05	.07	.05	.06	.05
.7	.3	.5	3.27	0	.90	.307	67	.05	.09	.09	.09	.06
.7	.5	.5	3.27	0	.90	.307	60	.05	.09	.09	.09	.10
.5	.3	.5	3	0	.90	.286	61	.05	.04	.03	.04	.03
.5	.5	.5	3	0	.90	.286	128	.05	.03	.03	.03	.04
.3	.5	.5	2.12	2.12	.90	.297	62	.05	.03	.03	.03	.02
.3	.5	.5	3.27	0	.90	.307	15	.05	.07	.06	.06	.06
.5	.3	.5	3	0	.90	.286	16	.05	.05	.05	.05	.05
.5	.5	.5	3	0	.90	.286	14	.05	.06	.06	.06	.03

*Sib-trio test with assumed value $\rho_o=.25$

**Sib-trio test with assumed value $\rho_o=.33$

***Sib-trio test with ρ and σ^2 estimated from the data

†Haseman-Elston test on all possible sib pairs

No simulations have been done to study the power and robustness of sib-trio statistics for small samples, as was done in section 3.3 for the case of sib pairs. It seems reasonable to expect that the results for sib trios would be similar to those in section 3.3, with the sib-trio statistics having greater power than the sib-pair statistic for the same total number of sibs.

To summarize, the simulations in Chapter IV and the results in this chapter indicate that (i) various sib-trio test statistics, in particular the four statistics in Table 5.2, have much greater power to detect linkage than the Haseman-Elston statistic on independent sib pairs; and (ii) the sib-trio test statistics in Table 5.2 have satisfactory significance levels, compared with a nominal .05 level, even for some rather small sample sizes.

CHAPTER VI

LINKAGE TESTS FOR LARGER SIBSHIPS AND FOR SIBSHIPS OF MIXED SIZES

In this chapter we consider briefly tests to detect linkage in sibships of size four or larger and in samples consisting of sibships of various sizes. First we prove the following result.

Lemma 6.1

Given a sibship of size 4, ordered with respect to sibs' genotypes at a two-allele trait locus, with corresponding trait values $x_i = \mu_i + g_i + e_i^*$ (i=1, ..., 4), and making the same assumptions as in previous chapters on the above linear model for x_i, let $Y_1 = (x_1-x_2)^2$ and $Y_2 = (x_3-x_4)^2$. Then, unconditionally (that is, averaging over the distribution of sibships, assuming random mating), $E(Y_1Y_2) = E(Y_1)E(Y_2)$, so that $\text{cov}(Y_1,Y_2) = 0$.

Proof.

We can enumerate all 81 possibilities for ordered sibships of size 4, along with the corresponding probabilities and values of $Y_1, Y_2, E(Y_1|\text{sibship})$, and $E(Y_2|\text{sibship})$, and then derive $E(Y_1Y_2)$ directly. Alternatively, we argue as follows.

(i) Assuming, as we have throughout, that genotypes are independent within a sibship, for any ordered sibship we have

P(sibship) = P(sibs 1 and 2) P(sibs 3 and 4).

(ii) Let

$$e_1 = e_1^* - e_2^* \text{ and } e_2 = e_3^* - e_4^*.$$

Then

$$E(e_1) = E(e_2) = 0 \text{ and } E(e_1^2) = E(e_2^2) = \sigma_e^2.$$

From the multivariate normal distribution we have e_1 and e_2 independent, so that

$$E(e_1e_2) = E(e_1^2e_2) = E(e_1e_2^2) = 0$$

and

$$E(e_1^2e_2^2) = \sigma_e^4.$$

(iii) The squared differences Y_1 and Y_2 can be written as $(w_1+e_1)^2$ and $(w_2+e_2)^2$, respectively, where w_1 and w_2 take the values 0, ± (a-d), ± 2a, depending on the sibs' genotypes. From (ii) it follows that

$$\begin{aligned} E(Y_1Y_2|\text{sibship}) &= (w_1^2+\sigma_e^2)(w_2^2+\sigma_e^2) \\ &= E(Y_1|\text{sibship})E(Y_2|\text{sibship}) \\ &= E(Y_1|\text{sibs 1 and 2})E(Y_2|\text{sibs 3 and 4}). \end{aligned}$$

(iv) Then we have

$$\begin{aligned} E(Y_1Y_2) &= \sum_{\text{sibships}} E(Y_1Y_2|\text{sibship})P(\text{sibship}) \\ &= \sum_{\text{sibships}} E(Y_1Y_2|\text{sibship})P(\text{sibs 1 and 2})P(\text{sibs 3 and 4}) \\ &= \sum_{\substack{\text{sibs 1} \\ \text{and 2}}} E(Y_1|\text{sibs 1 and 2})P(\text{sibs 1 and 2}) \\ &\quad \times \sum_{\substack{\text{sibs 3} \\ \text{and 4}}} E(Y_2|\text{sibs 3 and 4})P(\text{sibs 3 and 4}) \\ &= E(Y_1)E(Y_2), \text{ and the lemma is proved.} \end{aligned}$$

It is clear that the above lemma holds also for sibships larger than four, as long as the squared differences Y_1 and Y_2 correspond to sib pairs with no member in common. It is clear also that the same result holds for $E(Y_1Y_2|I_m,\lambda=1/2)$, where, as before, I_m represents the information on parents' and sibs' genotypes at the marker locus and λ is the frequency of recombination between the marker and trait loci. The latter is true because, in the absence of linkage, information about the marker locus tells us nothing about the trait values.

Considering (i) the lack of correlation (in the absence of linkage) between squared differences of trait values for sib pairs with no common member and (ii) the results of the simulations in Chapter IV and Chapter V for sib trios, it seems quite reasonable to suggest that the Haseman-Elston test be used on all possible sib pairs as a test for linkage, regardless of size of sibship. Thus sibships of size 4 would contribute 6 sib pairs, sibships of size 5 would contribute 10 sib pairs, and in general sibships of size n would contribute n(n-1)/2 sib pairs.

CHAPTER VII

SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

7.1. Summary

We have presented the model for the test derived by Haseman and Elston (1972) to detect genetic linkage between a locus for a quantitative trait and a marker locus with known pattern of inheritance. The test uses data on independent sib pairs and incorporates available information on parents' phenotypes at the marker locus. The power of the Haseman-Elston test has been studied by means of both asymptotic theory and computer simulations. It was found that the sample sizes required for high power are impractically large except for some situations of tight linkage and high heritability of the quantitative trait values. One of the assumptions required for the test to be valid for small samples is normality of the squared differences of sibs' trait values; computer simulations indicated also that the test is robust against various specific types of departures from this assumption, in all of which the trait values themselves have continuous distributions.

Various statistics similar to the Haseman-Elston statistic were considered for a linkage test on independent sib trios. Several appeared from computer simulations to be comparable to one another in power. For one of them, asymptotic theory was used to derive a sample size formula, as in the case of sib pairs. The resulting formula, as well as computer simulations, indicated that in most cases

of interest these sib-trio tests require only about half the total number of sibs to have the same power to detect linkage as the sib-pair test on independent sib pairs. However, the Haseman-Elston statistic on all possible sib pairs within the sib trios is almost as powerful as other sib-trio statistics.

Linkage tests for sibships of size four or greater and for sibships of mixed sizes were discussed briefly. It was suggested that in these cases the Haseman-Elston statistic for all possible sib pairs be used.

7.2. Suggestions for Further Research

Further investigation might be made of linkage tests for the cases of sibships of size four or greater and of sibships of mixed sizes. The power of any such tests should be studied, as well as the power of the Haseman-Elston test on all possible sib pairs. A speculation at this point is that the Haseman-Elston test increases substantially in power with size of sibship for samples of fixed total number of sibs. (If the $n(n-1)/2$ squared differences of sibs' trait values for a sibship of size n were independent, then (n-1) times as many individual sibs in sibships of size two would be required to attain the same power as the test in sibships of size n.)

The least squares approach of Chapter IV could be used to derive a test statistic for sibships of any particular size. However, the direct methods used in this dissertation to study power for the cases of independent sib pairs and sib trios are quite impractical for larger sibships, so that either other analytical methods (perhaps making more use of arguments by analogy and symmetry) or further Monte Carlo studies would be needed.

BIBLIOGRAPHY

Bailey, N. T. J. 1950. The influence of partial manifestation on the detection of linkage. Heredity 4: 327-336.

Bailey, N. T. J. 1951a. On simplifying the use of Fisher's u-statistics in the detection of linkage in man. Annals of Eugenics 16: 26-32.

Bailey, N. T. J. 1951b. The detection of linkage for partially manifesting rare "dominant" and recessive abnormalities in man. Annals of Eugenics 16: 33-44.

Bell, J., and Haldane, J. B. S. 1937. The linkage between the genes for colour-blindness and hemophilia in man. Proceedings of the Royal Society B 123: 119-150.

Bernstein, F. 1931. Zur Grundlegung der Chromosomentheorie der Vererbung beim Menschen mit besondere Berücksichtigung Blutgruppen. Zeitsschrift für Induktive Abstammungsund Vererbungslehre 57: 113-138.

Bock, R. D., and Kolakowski, D. 1973. Further evidence of sex-linked major-gene influence on human spatial visualizing ability. American Journal of Human Genetics 25: 1-14.

Brues, A. M. 1950. Linkage of body build with sex, eye color, and freckling. American Journal of Human Genetics 2: 215-239.

Elston, R. C., and Stewart, J. 1971. A general model for the genetic analysis of pedigree data. Human Heredity 21: 523-542.

Elston, R. C., and Stewart, J. 1973. The analysis of quantitative traits for simple genetic models from parental, F1 and backcross data. Genetics 73: 695-711.

Feller, W. 1971. An Introduction to Probability Theory and Its Applications, Volume II (Second Edition). New York: Wiley (p.286).

Finney, D. J. 1939. Tests for sex linkage in a quantitative character. Annals of Eugenics 9: 203-207.

Finney, D. J. 1940. The detection of linkage. Annals of Eugenics 10: 171-214.

Finney, D. J. 1941a. The detection of linkage II. Further mating types: Scoring of Boyd's data. Annals of Eugenics 11: 10-30.

Finney, D. J. 1941b. The detection of linkage III. Incomplete parental testing. Annals of Eugenics 11: 115-135.

Finney, D. J. 1942a. The detection of linkage IV. Lacking of parental records and the use of empirical estimates of information. Journal of Heredity 33: 157-160.

Finney, D. J. 1942b. The detection of linkage V. Supplementary tables. Annals of Eugenics 11: 224-232.

Finney, D. J. 1942c. The detection of linkage VI. The loss of information from incompleteness of parental records. Annals of Eugenics 11: 233-244.

Finney, D. J. 1943. The detection of linkage VII. Combination of data from matings of known and unknown phase. Annals of Eugenics 12: 31-43.

Fisher, R. A. 1935a. The detection of linkage with dominant abnormalities. Annals of Eugenics 6: 187-201.

Fisher, R. A. 1935b. The detection of linkage with recessive abnormalities. Annals of Eugenics 6: 339-351.

Fraser, G. R. 1968. Some complications of the use of the three-generation method in the estimation of linkage relationships on the X chromosome in man. Annals of Human Genetics 32: 65-79.

Hájek, J., and Sidák, Z. 1967. Theory of Rank Tests. New York: Academic Press.

Haldane, J. B. S. 1934. Methods for the detection of autosomal linkage in man. Annals of Eugenics 6: 26-65.

Haldane, J. B. S., and Smith, C. A. B. 1947. A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. Annals of Eugenics 14: 10-31.

Halperin, M. 1971. Normal regression theory in the presence of intra-class correlation. Annals of Mathematical Statistics 22: 573-580.

Haseman, J. K. 1970. The genetic analysis of quantitative traits using twin and sib data. Unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill.

Haseman, J. K., and Elston, R. C. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2: 3-19.

Hill, A. P. 1974. Quantitative linkage: a statistical method for its detection and estimation. Unpublished Ph.D. dissertation, University of California at Los Angeles.

Hill, A. P. 1975. Quantitative linkage: a statistical procedure for its detection and estimation. Annals of Human Genetics 38: 439-449.

Hogben, L. T. 1934. The detection of linkage in human families. Proceedings of the Royal Society B 114: 340-363.

Howells, W. W., and Slowey, A. P. 1956. "Linkage studies" in morphological traits. American Journal of Human Genetics 8: 154-161.

Jayakar, S. D. 1970. On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. Biometrics 26: 451-464.

Johnk, M. D. 1964. Erzung von Beta Verteiler und Gammaverteiler Zufallzahlen. Metrika Zeitschrift fur Theoretische und Angewandte Statistik 8: 5-15.

Kloepfer, H. W. 1946. An investigation of 171 possible linkage relationships in man. Annals of Eugenics 13: 35-71.

Kronmal, R. 1964. Evaluation of a pseudorandom norm.l number generator. Journal of the Association for Computing Machinery 11: 357-363.

Lasker, G. W., and Wanke, W. D. A method for the study of sex chromosome influences in normal populations. Unpublished paper.

Li, C. C. 1955. Population Genetics. Chicago: University of Chicago Press.

Lowry, D. C., and Shultz, F. T. 1959. Testing association in metric traits and marker genes. Annals of Human Genetics 23: 83-90.

Lush, J. L. 1949. Heritability of quantitative characters in farm animals. Proceedings of the Eighth International Congress of Genetics, Stockholm, 1948 (Supplement to Hereditas). Edited by G. Bonnier and R. Larsson. Berlingska Boktryckeriet, Lund. Pages 356-375.

MacLaren, M. D., and Marsaglia, G. 1965. Uniform random number generators. Journal of the Association for Computing Machinery 12: 83-89.

McKusick, V. A. 1969. Human Genetics. Englewood Cliffs, N. J.: Prentice-Hall.

Morton, N. E. 1955. Sequential tests for the detection of linkage. American Jourual of Human Genetics 7: 277-318.

Morton, N. E. 1956. The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. American Journal of Human Genetics 8: 80-96.

Morton, N. E. 1957. Further scoring types in sequential linkage tests, with a critical review of autosomal and partial sex linkage in man. American Journal of Human Genetics 9: 55-75.

Niemann-Sorensen, A., and Robertson, A. 1961. The association blood group and several production characteristics in three Danish cattle breeds. Acta Agriculturae Scandinavica 11: 163-196.

Penrose, L. S. 1935. The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Annals of Eugenics 6: 133-138.

Penrose, L. S. 1938. Genetic linkage in graded human characters. Annals of Eugenics 8: 233-237.

Penrose, L. S. 1947. A further note on the sib-pair linkage method. Annals of Eugenics 13: 25-29.

Penrose, L. S. 1950. Data for the study of linkage in man: red hair and the ABO locus. Annals of Eugenics 15: 243-247.

Penrose, L. S. 1951. Data for the study of linkage in man: phenylketonuria and the ABO and MN loci. Annals of Eugenics 16: 241-248.

Penrose, L. S. 1953. The general purpose sib-pair linkage test. Annals of Eugenics 18: 120-124.

Phillips, D. T., and Beightler, C. S. 1972. Procedure for generating gamma variates with non-integer parameter sets. Journal of Statistical Computing and Simulation 1: 197-208.

Pratt, J. W. 1959. On a general concept of "in probability." Annals of Mathematical Statistics 30: 549-558.

Renwick, J. H. 1969. Progress in mapping human autosomes. British Medical Bulletin 25: 65-73.

Renwick, J. H. 1971. Assignment and map-positioning of human loci using chromosomal variation. Annals of Human Genetics 35: 79-97.

Renwick, J. H., and Schulze, J. 1965. Male and female recombination fractions for the nail-patella: ABO locus in man. Annals of Human Genetics 28: 379-392.

Smith, C. A. B. 1953. The detection of linkage in human genetics. Journal of the Royal Statistical Society B 15: 153-192.

Smith, C. A. B. 1959. Some comments on the statistical methods used in linkage investigations. American Journal of Human Genetics 11: 289-304.

Smith, C. A. B. 1963. Testing for heterogeneity of recombination fractions in human genetics. Annals of Human Genetics 27: 175-182.

Smith, C. A. B. 1968. Linkage scores and correction in simple two and three generation families. Annals of Human Genetics 32: 127-150.

Smith, C. A. B. 1975. A non-parametric test for linkage with a quantitative character. Annals of Human Genetics 38: 451-460.

Steinberg, A. G., and Morton, N. E. 1956. Sequential test for linkage between cystic fibrosis of the pancreas and the MNS locus. American Journal of Human Genetics 8: 177-189.

Thoday, J. M. 1967. New insights into continuous variation. Proceedings of the Third International Congress of Human Genetics Edited by J. F. Crow and J. V. Neel. Baltimore: Johns Hopkins Press (pp. 339-350).

Wald, A. 1947. Sequential Analysis. New York: Wiley.

Weitkamp, L. R. 1973. Human autosomal linkage groups. Proceedings of the Fourth International Congress of Human Genetics, Paris, 1971. Birth Defects: Original Article Series, 8, no. 7.

Wiener, A. S. 1932. Method of measuring linkage in human genetics: with special reference to blood groups. Genetics 17: 335-350.