THEORETICAL ASPECTS OF PEDIGREE ANALYSIS

E. Ginsburg, I. Malkin, R.C. Elston

THEORETICAL ASPECTS OF PEDIGREE ANALYSIS

RAMOT PUBLISHING HOUSE – TEL-AVIV UNIVERSITY, ISRAEL

E. Ginsburg, I. Malkin, R.C. Elston

THEORETICAL ASPECTS OF PEDIGREE ANALYSIS

RAMOT PUBLISHING HOUSE - TEL AVIV UNIVERSITY, ISRAEL

Ramot Publishing - Tel Aviv University

ISBN 965-274-424-7

Published in Israel, 2006

PREFACE

It is a great pity that I have to write these few paragraphs on the last book by Emil Ginsburg when he, as first author, is no longer with us. Emil considered science and the opportunity to do science as one of the major values and pleasures one may have in life. He was a very critical person and very demanding of others, especially of his friends; they repaid him with love. His inability to make the compromises necessary for the routine of scientific life in the Soviet was a challenge to him, requiring a lot of spirit and courage and making him a standard of integrity.

He was one of the first investigators in the former USSR who understood the theoretical and practical limitations of the classical biometrical genetics approach to quantitative traits, built on the statistical concept of heritability, and considered Mendelian analysis to be a promising alternative. For many years, we could hardly find anyone else to support either side of our opposing views when we discussed quantitative genetics together. Unlike me, he considered segregation analysis to be an important tool for unraveling the genetic basis of quantitative variation, whereas I was more biased toward marker analysis as a major tool.

Emil's contribution to pedigree analysis is very impressive. He developed a unified approach to genetic analysis, allowing for the simultaneous treatment of data on qualitative and quantitative characters, for either controlled crosses or samples from populations with arbitrary family structures. His method allowed for linkage, viability disturbances, competition, differential penetrances, etc., with the possibility of testing hypotheses about gene action (major gene, combination of major gene and polygenes). Some of his results established the basis for the original

algorithms implemented in a software package for pedigree analysis, MAN.

One of Emil's major interests was the genetics of quantitative variation. The analysis of quantitative traits was for many decades considered one of the most difficult fields of genetics. Basically, the situation has not much changed with the evolving genome paradigm and genomic technologies. The subject of this book is the theoretical foundation of pedigree analysis, with complex (quantitative) traits being its major focus, especially in the context of human genetics. It can be considered as an attempt to discuss in a comprehensive fashion the basic concepts and notions of pedigree analysis, a thorough inspection of its applicability and limitations, and the corresponding statistical methodologies and challenges. The subjects reviewed and discussed include pedigree sampling, likelihoods on pedigrees, parameterization problems and model comparisons, various formulations of complex trait analysis on pedigrees (continuous, binary, longitudinal), model-based vs. model-free linkage analysis, perspectives on and limitations of the genetic dissection of complex traits, etc. Emil made valuable contributions to many of these problems, which are only partly reflected in this book. I believe that this book will be very helpful to those engaged in both research and teaching of quantitative genetics.

Abraham Korol
Professor of Genetics
Institute of Evolution, University of Haifa

A fairly complete draft of this book was written by Emil Ginsburg before his death, our contribution being one of rephrasing, clarifying and updating what were essentially his ideas. Nevertheless, we have tried hard to keep the spirit of Emil's writing. This has been no easy task. Emil was not a native English speaker and his desire to explain each idea as precisely and comprehensively as possible sometimes resulted in lengthy phrases and sentences that were difficult to understand. On the one hand, we did not want to completely change everything; and on the other hand, we wanted to clarify as much as possible the phrases and terminology Emil used. We hope we have succeeded in making this body of theoretical work, especially the whole new way of viewing pedigree ascertainment (a topic that was first investigated by Weinberg almost a hundred years ago), including the selective inclusion of pedigree data in the dataset that is eventually analyzed, accessible to a wider audience. In this book, those of us who have studied and contributed to the theory of pedigree analysis will find much to ponder over. We always enjoyed discussing pedigree analysis with Emil; and we fervently hope that our work on this book will help keep alive his incisive contributions to the topic.

Robert C. Elston. Ida Malkin.

Cleveland, Ohio, August 2005. Tel Aviv, September 2005

Contents

INTRODUCTION

Let us define the *inheritance* of a trait as the mechanism by which the joint phenotypic distribution of that trait in members of any pedigree can be explicitly described. If the inheritance of a trait is known, it is possible to make a probabilistic prediction about the joint co-variation and co-segregation of its phenotypic characteristics in pedigree members in future generations. **The study of trait inheritance is called *pedigree analysis* because the inheritance of the trait being studied is made explicit by collecting and studying a sample of pedigrees. We assume that the interdependence between the phenotypic characteristics of the members in each sampled pedigree implicitly reflects the real inheritance of the trait being studied.**

Most widely used is ***genetic pedigree analysis,*** **in which a description of the trait inheritance is made assuming that the main factors underlying this inheritance are *genes*** – the DNA segments positioned on the chromosomes and transmitted from parent to offspring in accordance with Mendelian laws.

When describing the trait inheritance, we suppose that the genes involved in the control of this trait can be unambiguously identified, and that their separate and joint phenotypic distributions can be predicted for each environmental condition. This knowledge can be further used in practice to predict later disease development in a patient, to find the optimal treatment he/she should be given, etc. In other words, the trait inheritance is defined in such a way that, using phenotypic data on other pedigree members, we can determine the physiological status expected in an individual to be completely characterized by his/her specific genetic

make-up and environment. At present we are far from such knowledge, although various techniques (biochemical, molecular, mathematical, etc.) are being actively used to obtain it.

A widely used approach to study the inheritance of a particular function of an organism is to describe it with the use of mathematical-genetic models. The initial stage of this modeling requires the choice of one or several *traits* describing the biological function to be studied. Then mathematical-genetic models for the inheritance of these traits are formulated and tested, using *pedigree samples* collected from a particular population. **The *pedigree* comes to the attention of the researcher as a structured subset of its members, this structure being determined by the relationships among the pedigree members. As already stated, we suppose that the pedigree implicitly contains details of the trait's mode of inheritance through the co-variation and co-segregation of the trait characteristics among the pedigree members.** Using specially constructed *statistical tests*, a formal model of trait inheritance is tested and either rejected or accepted as a satisfactory description of the mode of inheritance. If accepted, the model thus constructed is used to solve practical problems that are directly suggested by the model's ability to make predictions. The basic factors of the trait inheritance formulated by such a model are the same as they are in any other technique: the genes controlling the trait inheritance, their phenotypic distribution in different environmental conditions, etc.

Being only one of many approaches, pedigree analysis results in a detailed description of the trait inheritance when combined with other, less formal, methods of study - such as physiological, cytogenetic and molecular-genetic methods. Only by combining the information obtained

by different techniques can we obtain an adequate and (relatively) complete description of the inheritance, which can then be later used in various applications.

Until recently, a distinction has been made between two versions of pedigree analysis, segregation analysis and linkage analysis. In the former, the genetic model of the trait inheritance is constructed using the pedigree sample, which by assumption contains the necessary information about the true trait inheritance, expressed implicitly through the joint phenotypic co-segregation in the collected pedigree members. In the latter, linkage analysis, the purpose has been to localize in specific chromosomal segments the putative gene(s) established in the segregation analysis. This distinction in purpose of these two types of analysis was accompanied by differences in the sampling design used to collect the pedigree sample and in the statistical techniques used.

Recent advances in pedigree analysis have caused us to reconsider this comparatively simple scheme. First of all, we have come to appreciate that segregation analysis is a procedure with very limited capabilities. Although a genetic model for the trait inheritance can be formulated and statistically tested using a pedigree sample, and a determination made whether or not to accept what is found as an adequate mathematical-genetic description of the trait inheritance, the result can be trusted only if the trait is under relatively simple genetic and environmental control – or, to be more correct, if the function phenotypically characterized by the trait under study is controlled by genes and environmental factors in a relatively simple manner. However, most human characteristics, especially those describing the human organism as a whole and directly related to an individual's physiological status and his/her health, are not inherited in a

simple fashion. Their genetic and environmental control is so complicated that it is hardly conceivable that a simple genetic model can adequately describe their inheritance.

The second advance that has substantially changed the classical scheme of pedigree analysis is the successful completion of the Human Genome Project, providing a large number of DNA marker loci all along the human genome. Accordingly, the direction of pedigree analysis has shifted. Using the possibilities presented by the results of the Human Genome Project, the main productive technique of pedigree analysis has become linkage analysis. As was noted by Rao (1998, p.2), "…whereas linkage analysis has been used in the past merely to map genes that were already known to exist, linkage analysis of complex traits serves a dual purpose: first, to prove the very existence of a trait gene, and then to map it".

This change in the formulation of the pedigree analysis problem has been accompanied by a corresponding change in the design of pedigree samples and in formulating genetic models, and in the statistical methods used. Elston (1998) distinguished *model-based* and *model-free* pedigree analyses. In the first, we formulate models of the trait inheritance and the sampling procedure that are used, while in the second the analysis proceeds without such explicit models. The complementary nature of these two methods of pedigree analysis, and how they relate to each other, is a problem that has to be clearly formulated and solved. Moreover, the very construction of a genetic model describing the inheritance of a multifactorial trait under study becomes different from that used in earlier segregation analysis. Usually, the analysis results in the construction of a complex *compound* genetic model, which, rather than describing the trait

inheritance as a whole, results in a set of component genetic models each describing the phenotypic effects of the different genes that control trait being studied. These component models are then combined, taking into account the possible pleiotropic gene effects and common environmental factors modifying their phenotypic distributions. This is a much more complicated task, but it appears to be the only way to describe in an adequate fashion the multifactorial traits that are currently being studied.

Thus model-free and model-based analyses are used to solve different problems: the first is directed towards establishing the very existence of genes taking part in the trait control and identifying them by localizing their positions on chromosomal segments (with whatever accuracy the set of marker loci used can provide), while the second is directed towards constructing a mathematical-genetic model describing the inheritance of the trait (the joint distribution of phenotypes in the pedigrees analyzed) and towards testing that model on the pedigree sample. Currently, attempts are being made to develop a practical strategy to unite these two kinds of pedigree analysis in a complementary way, for example by using the results of the model-free approach to improve, i.e., to make more effective, the prognostic capabilities of the mathematical-genetic model for the inheritance of the trait under study.

In what follows, we consider the theoretical aspects of pedigree analysis, concentrating mostly on defining its basic concepts, on a detailed description of the situations when these concepts can be properly used, and on ways of formulating pedigree analysis problems.

1. BASIC DEFINITIONS

1.1. Population

The process of determining the inherited characteristics of individuals in subsequent generations takes place, and therefore can be studied, on large groups of individuals more or less separated from one another.

Let us define a *population* as an "inwardly connected" and "outwardly isolated" set of individuals sharing a common range of environmental conditions. The inward connectivity means that there is a non-zero probability that any pair of individuals of opposite sex will have descendents in future generations of that population. The outward isolation means that this probability is substantially less for an individual from one population and any other individual from a different population. Each population is characterized by its *relationship structure,* meaning that a specific relationship "connection" exists for each pair of individuals (parents and offspring, siblings, cousins of different lineages, and so on up to unrelated or individuals with unknown relationship, if there are any such that exist).

The characteristics of each population, its size, its relationship structure, its particular mating structure - in short, all the population characteristics that distinguish it from other such populations - are usually fully determined by its origin and later history, including all demographic, social, cultural and other processes.

1.2. Trait - phenotype

The *trait* is defined as any characteristic describing a certain biological function of the persons in the study population. How to characterize this function is usually determined by the instruments available to the investigator. In some cases, the method of characterizing the function is almost uniquely determined (height, weight, etc). In other cases, the same function can be described by different traits. The instrument that is adopted usually has no direct connection with the biological, in particular the genetic, nature of the function being studied. This is why its characterization by a particular trait is also termed *phenotypic* (from the Greek word meaning "appearance"), and any individual characteristic is called a *phenotype*.

Depending on the manner in which the characteristic is measured (whether a categorical description or a continuous measurement) the trait can be *qualitative*, in particular binary, *quantitative* (discrete or continuous), or *complex*. In the first case, the function is described by a set of categories or types, while in the second a numerical measure of the biological function is introduced. Both qualitative and quantitative traits can be observed or measured as a single value, or be represented by a vector of results, numerical or categorical, of several observations and/or measurements. A complex trait is represented by an array of qualitative and/or quantitative traits, each obtained as a single observation (measurement) or as a function of several observations.

1.3. Gene, genotype

Inheritance can be defined as a set of correlations between the degree of two individuals' relationship and their level of phenotypic similarity (Galton, 1888; Pearson, 1920).

In genetics, the concept most often used to describe inheritance is the *gene*. Presently, molecular, biochemical and cytogenetic studies have produced such a complicated and multi-semantic construction called "gene" that it is doubtful whether a complete and exhaustive definition of this term can be reasonably proposed. In the limited context of our present considerations, the following definition can be considered satisfactory:

A gene is a hereditary unit that has the following three necessary properties:

(i) The gene is represented by a certain chromosomal segment (from this originates the second, almost equivalent term, *genetic locus*).

(ii) The gene is transmitted from parents to offspring according to Mendelian rules of segregation (nowadays what is transmitted is often termed an *allele*, to distinguish it from the *locus*; and

(iii) The gene takes part in the control of the trait's inheritance.

For the sake of convention, we shall call the chromosomal segment that takes part in the trait's phenotypic control a gene, and its chromosomal position a locus - often known on the basis of a nearby marker locus.

Thus the gene is defined here as a *Mendelian factor* controlling the inheritance of a trait. From an information point of view, the gene can be considered to be an instruction (written in a special four-letter alphabet) by which an organism constructs a certain phenotype during its ontogenesis. There can be more than one version of the same instruction and these

versions are called the *alleles* of the gene. **We can think of genes as factors in the sense that heat is a factor, and alleles are the levels of such a Mendelian factor in the sense that hot and cold are levels of the factor heat.** Alleles, that represent different instructions not manifested at the phenotypic level (i.e. that are unobservable), are called *isoalleles*. Note that whether or not we can detect an allele is directly determined by the manner in which we phenotypically characterize the function we are studying. For example, two alleles may be distinguished on an electrophoregram but, at the same time, could be considered as isoalleles when compared at the level of the whole organism.

The specific function of each gene is determined by its set of alleles. **Each individual has his/her *genotype* regarding the trait under consideration, which is defined as the sub-set of allele pairs (for diploid organisms) at each locus taking part in the control of the trait.**

Let us assume by definition that ***a gene takes part in the control of the trait's inheritance*** **if, and only if, at least one substitution of the gene's alleles in the diploid genotype leads to a discernible change in the trait's distribution (a different set of trait value for a quantitative trait, or different phenotypic variants for a qualitative trait).** By this definition, monomorphic genes having only one allele (or any number of alleles isoallelic to one another) do not take a part in the control of the trait, even if it is known that products of these genes are necessary links in the biochemical chain of synthesis that results in the phenotype.

To make the above definitions correspond to the molecular organization of a real gene, it is helpful to present the latter as follows. The gene can be described as the structured sequence of two antiparallel nucleotide strands in a configuration that forms a double helix. A

homologous pair of nucleotides forms the basic DNA unit – the base pair (bp). Each gene contains the specific bp sequences that determine its beginning and end. Between the beginning and the end are two types of bp sequences alternating with each other, exons and introns that determine the specific gene function. The exon is the portion of the bp sequence that is translated into protein, while the intron is not a translated sequence and is usually removed by splicing. The exon is often considered as determining the gene's function, but it has lately been found that intronic sequences can also determine function.

The length of the Human genome is about 3,200 megabases (Mb; 1 Mb = 1,000 bp). It contains approximately 30,000 named and unnamed genes and many intergenic DNA sequences. A fraction of about 1.5% performs the coding functions that are somehow manifested in the gene's intracell or intercell activity, or expressed at the phenotypic level. Noncoding sequences of genes include control regions, such as promoters, operators, and terminators, as well as intron sequences.

Genes differ from one another in their size. The average size of a gene is 27 kb. The small β-Globin gene is 1.6 kb in length and contains 3 exons each with an average length of 150 bp, and introns with an average length of 490 bp. For the large Distrophin gene, these parameters are: 2,400 kb, 79 exons of on average 180 bp length, and introns each approximately 30,000 bp long.

Certain regions of intergenic DNA are highly polymorphic; they contain a number of tandem repeats, short sequences repeated one after another multiple times. Also in the gene sequence are single nucleotide polymorphisms (SNPs) that occur as often as about every 200 bp in the human genome, i.e. approximately 27kb/200bp = 135 per

gene of average length. Taking into account that only about 1% of its DNA sequence affects the function of the gene, the polymorphic gene contents may or may not result in a change of phenotype. Different molecular contents of the gene sequences are expressed as *alleles* of the gene. If they are not recognizable phenotypically, they should be considered as isoalleles.

There are specific loci called *marker*s used in linkage analysis to establish the position on the chromosomes of trait genes of interest. These markers can be defined as genes often themselves having no clear phenotypic expression, but with known chromosomal positions (determined in a previous study). Currently, mostly single nucleotide polymorphisms (SNPs) are used as marker loci. However, the presence or absence of other specific DNA sequences found by molecular techniques are also widely used, for example, restriction fragment length polymorphisms (RFLPs) and variations in the number of short tandem repeat polymorphisms (STRPs) in specific chromosomal regions, also known as microsatellites. The accuracy of any gene mapping study clearly depends on the particular set of marker loci used, including their chromosomal positions. In this connection, it should be noted that the length of a trait gene is usually much larger than that of marker loci, including RFLPs, STRPs and even shared haplotypes (see section 10.4). This means that "fine-scale" gene mapping can be performed in linkage analysis, if the molecular technique used provides us with the possibility of identifying marker loci within the limits of the trait gene being studied.

Thus, from the above very schematic description of molecular gene structure, it follows that the way we have defined a gene – the Mendelian factor, the compact discrete unit successively positioned with other such

units along a chromosome and responsible for the hereditary process – represents a certain idealization of the term. However, this idealized definition serves our purpose quite satisfactorily when we consider pedigree analysis problems.

1.4. Genotype-phenotype correspondence

We have given above an operational definition of genes that take part in the control of a trait: if an allele substitution in the diploid genotype results in a recognizable change in the trait distribution (by the method of phenotypic characterization used), then alleles of this gene should be included in the trait genotype. However, the question of how (through what molecular, cytological and physiological processes) this genotypic control is realized has not been raised. At this stage, any generalization of ontogenetic regularity that could be used in a formal description of the genotype-trait correspondence does not yet appear to be possible. Without such knowledge, the correspondence can be defined only phenomenologically. This means that, for each genotype, we have to introduce a probability distribution of possible phenotypic manifestations of this genotype (trait values for a quantitative trait or phenotypic variants for a qualitative trait). This distribution gives only a more or less adequate approximation of the results of the whole highly complicated ontogenetic process that leads from the genotype to the observed trait.

Let $\mathbf{X} = \{x\}$ be a set of possible phenotypic characteristics (binary, qualitative, or quantitative, discrete or continuous). Denote by $X_n = \{x_1,...,x_n\}$, $x_i \in \mathbf{X}$, the set of phenotypes observed on the n members of a pedigree, and by $f(X_n|G_n)$ the joint probability (density) of the set of phenotypes X_n, i.e. on the n members of the pedigree, conditional on their

given set of genotypes $G_n = \{g_1,...,g_n\}$, $g_i \in \mathbf{G}$, where **G** is the set of genotypes that are involved in controlling the trait being studied. This probability describes the genotype-phenotype correspondence. If there is no other common factor (besides these genes) causing the trait co-variation among relatives, then $f(X_n|G_n) = \prod f(x_i|g_i)$ where $f(x|g)$ is the probability that phenotype x is manifested by an individual with genotype g; for a quantitative continuous trait, this probability describes the *residual* distribution about a genotype-specific mean of the trait x in individuals with the given g, caused by all factors influencing the trait except the genotype. Note again that $f(x|g)$ and $f(X_n|G_n)$ are introduced only as an approximate phenomenological description of the genotype-phenotype correspondence, and no molecular or cytological mechanism is taken into account in their definition.

1.5. Genetic model of inheritance

Define a *genetic model* for the inheritance of a trait as the following three distributions determined on the two sets defined above:

$$\theta = \{ p(g_1, g_2), P(g \mid g_1, g_2), f(X_n \mid G_n) \mid \mathbf{X},\mathbf{G}\}, \qquad (1.1)$$

where $p(g_1,g_2)$ is the joint population distribution of genotypes in spouse pairs determined by the population mating structure; $P(g|g_1,g_2)$ is the conditional probability that an offspring receives genotype g from parents having genotypes g_1 and g_2 - the core of the genetic inheritance; and $f(X_n|G_n)$ is the joint distribution of the phenotypes of the n pedigree members given their set of genotypes G_n.

Each of these three distributions is determined, first by its type (the particular mathematical form, e.g. binomial, Poisson, Gaussian, n-variable normal, etc.) and second by an array of parameters specific to this type of

distribution. These parameters are called genetic model parameters or *genetic parameters*.

The model (1.1) represents the most complete formal description of the mode of trait inheritance. Its particular form is justified as follows. First, this model is intended to describe the trait inheritance in the particular population, given its specific way of forming spouse pairs. Second, this model defines the process of transmission of the parental alleles to offspring – the core of genetic inheritance. Last, this model defines how the genotypes jointly manifest the observed (measured) traits. The three distributions of the model represent three physically different and complementary aspects of the inheritance process.

Given the model θ, the joint distribution of phenotypes in members of any collected pedigree can be unambiguously expressed as:

$$f(X_n \mid \theta) = \sum_{G_n} P(G_n \mid \theta) f(X_n \mid G_n, \theta),$$

where $P(G_n|\theta)$ is the joint probability of the subset of genotypes, $G_n = \{g_1,...,g_n\}$, in the n pedigree members determined by the population distribution $p(g_1,g_2)$ and the *transition probabilities* $P(g|g_1,g_2)$. Elston and Stewart (1971) called these latter probabilities transition probabilities because they have the Markov property of not depending on the genotypes of ancestors in previous generations, which we can think of as previous "states" visited in an evolutionary process. The only difference from the usual Markov process is that each individual's genotype depends on two, rather than one, genotypes in the previous generation. They reserved the term *transmission probability* for the probability that an individual with a given genotype transmits a particular allele (or, more generally, haplotype) to an offspring. Later authors have unfortunately used the term

transmission probability for these two quite different probability distributions.

As we can see, the main concept, underlying model (1.1), is the specifically structured set **G** of discrete objects, genotypes that determine the trait inheritance. This is why the model is called genetic.

1.6. Genetic analysis

The mode of inheritance of a trait can be determined if, and only if, we can define an algorithm by which it is possible to predict, with a certain degree of accuracy, the phenotypes of offspring on the basis of the phenotypes of their relatives in previous generations and also, perhaps, of certain environmental conditions in which the inherited potentials of these offspring will be manifested. We do not specify here the manner in which the algorithm should be constructed - in other words, no limitation is introduced on the construction of the models describing the inheritance process.

A *phenomenological* description of the trait inheritance combines all those algorithms whose construction includes no assumption about a particular inheritance mechanism. This way of describing inheritance is the simplest ideologically and was the one first proposed. Galton (1889) introduced a linear statistical approximation for the dependence of the trait values of an offspring on those of his/her parents. Further development of the linear statistical methods led to some more complicated versions of this description (e.g., in the form of a joint n-variable normal distribution of the trait values in members of a pedigree) without changing its main characteristic – the total absence of any particulars of the hereditary mechanism being included in the algorithm. The construction, mutual

comparison and selection of (in a certain sense) the best phenomenological algorithm can be defined as a particular case of *inheritance analysis*.

Genetic analysis is another particular case of inheritance analysis, but in which the dependence between the trait manifestations (values) of relatives are formulated in terms of factorial genetics, including concepts of the gene, genotype, Mendelian rules of gene transmission across generations and the phenomenologically introduced genotype–phenotype correspondence.

Formally, genetic analysis can be defined as follows (Ginsburg, 1984; Ginsburg and Livshits, 1999). Let $\theta = \{\theta_i\}$ be a set of *a priori* constructed hypothetical genetic models that differ from one another in their genotypic set, G, and in the types of the three component model distributions and/or the values of the genetic parameters. Introduce an operator Ω that establishes a particular *order of preference* for the all models in θ with respect to the collected pedigree data $\{X_n\}$:

$$\Omega(\theta|\{X_n\}) = \tilde{\theta} = \{\theta_1, \theta_2, \theta_3, ...\}\ , \qquad (1.2)$$

where $\tilde{\theta}$ is the re-ordered set of models in which θ_i is at least preferable to θ_j (in a certain sense, determined by an explicit form of the Ω operator) for any $i< j$.

Based on this ordering $\tilde{\theta}$, genetic analysis is defined as the choice of the first ranked model as supposedly the best mathematical-genetic descriptor of the inheritance of the trait being studied.

1.7. Pedigree analysis

The starting point of any genetic analysis is the empirical data that implicitly contain information about the genetic nature of the trait analyzed. These data can be represented by two different types of datasets.

The first represents the results of a hybridization experiment that has the following main features. Initially, two so-called pure lines that are phenotypically different are crossed. The genetic "purity" of the crossed lines is not always unambiguously defined. In some experiments, each line is obtained by a prolonged close inbreeding accompanied by casting away individuals phenotypically different from a certain standard. In this case, we hope that the resultant line consists of individuals having the same homozygous genotypes, at least for genes in any way connected with the trait being studied. In other cases, the hybridized "lines" have not been obtained by this inbreeding technique, but represent two heterogeneous sub-populations that (i) differ from one another by some subset of traits, and (ii) exhibit phenotypic variability of these traits that is negligible within each sub-population. It is again assumed that each of the crossed sub-populations is homozygous for genes controlling the trait subset (i.e. that they are non-segregating lines with respect to the trait).

The lines are hybridized to produce first-generation hybrids. In different hybridization schemes, individuals from the latter can be crossed with one another to produce second-generation hybrids, or with individuals from one of the parental lines (backcrosses). Several other generations can be obtained according to the chosen hybridization plan. The data observed (measured) on individuals from these generations are then the basis for genetic hybrid analysis.

The second type of empirical data subjected to genetic analysis comprises *pedigree data*. The information about the specific genetic control of the trait is implicitly represented by the joint phenotypic distribution of the pedigree members' phenotypes, the subset of observed individuals having certain relationships with one another and segregating their phenotypes across generations in a dependent manner.

For a pedigree, the joint distribution and transmission of phenotypes across generations, together with the usual supposition that the genotypes of the pedigree founders are a representative sample of genotypes from the same population, provides the empirical description of the inheritance of the trait under study.

The empirical basis of *pedigree genetic analysis* is a sample of pedigrees.

1.8. A note on phenotypic characterization

It is necessary to clarify one of the problems of genetic analysis, namely, the manner in which the phenotypes of individuals are characterized - the trait definition. The genetic analysis *of qualitative* (*binary*) *traits* is distinguished from the genetic analysis *of quantitative traits*. While this sub-division is convenient methodologically because a different model formulation is needed for the different trait types, it is not critical from a genetic point of view.

In the mid-forties, there was the widespread belief among geneticists that the genetic nature of qualitative traits differs substantially from that of quantitative traits. First, this difference was defined as the difference in the number of genes taking part in the trait control and in the magnitudes of their effects. In particular, it was stated that quantitative traits are controlled by a large number of genes distributed independently

and each having a small additive effect on the trait value. This idea was derived from the following seemingly unconnected set of facts. First, Fisher (1918), in his demonstration that no contradiction exists between the discrete nature of genes (Mendelian factors) and the continuous variation of measured traits, introduced the concept of the average allele effect and of additive genetic variance. Second, Wright (1968) had successfully used a linear statistical technique to describe trait co-variation among relatives. Third, it is well known that the binomial distribution can be well approximated by the normal distribution, given a large number of possible combinations. Thus, assuming that a large number of minor genes are responsible for the co-variation of trait values in relatives, an additive polygenic model was constructed for quantitative trait inheritance using such parameters as heritability, genetic correlations etc. (see, for example, Mather and Jinks, 1982). Ginsburg and Nikoro (1982) made a detailed analysis of the basic concepts and methods of this theory and showed that 1) the assumptions underlying the additive polygenic model are too restrictive with respect to our current level of knowledge of the mechanisms underlying the genetic control of phenotypic variation, and 2) these assumptions are in principle untestable on the basis of usual pedigree samples.

Next, statements about the specific nature of genes controlling quantitative traits appeared. It was assumed that these genes are not structural, like those controlling the qualitative traits, but are of some specific nature, mostly polygene modifiers taking various forms at the different stages of genetic development. For example, the discovery of heterochromatin provoked the assumption that the polygenes controlling quantitative traits were located in heterochromatin chromosomal segments.

Successful studies of the molecular regulation of gene activity in some microorganisms led to another hypothesis, that the quantitative trait genes are regulatory genes. Next, it was assumed that the polygenes are represented by special DNA repetitions, or that they can be in the form of disperse mobile DNA elements, or of special gene-enhancers, etc. A special abbreviation has even been accepted for them, QTL – quantitative trait loci (Geldermann, 1975). As was to be expected, each new hypothesis about the specific polygenic nature of the QTL appeared at the very beginning of studying the new phenomenon and, with progress, the corresponding hypothesis has not been disproved, but rather tacitly buried.

It is clear that confusing two different classifying factors has caused the appearance of this list of hypotheses. The first factor is the method of phenotypic characterization, which was chosen and then accepted regardless of the particular nature of the genes controlling the trait under study. The second factor is the genetic control of the trait, including both the number of genes and the particular type of DNA involved. It was wrongly assumed that if the trait is descriptive and typologically defined, then it should be determined by simple structural genes; while if the phenotypic characterization was made by some measuring instrument, then the control of the corresponding quantitative trait becomes too complicated to be attributed to the effect of structural genes.

It is now well known that there are qualitative, in particular binary, traits that have complicated inheritance. The widely used binary characterization “affected-unaffected” applied to some diseases is typologically determined, based on a specific set of symptoms. For diseases with a complex etiology, the special term “multifactorial” was

introduced, meaning that (i) there is a number of environmental factors affecting the development and further manifestation of the disease (the so-called risk factors), and (ii) the genetic potential for the disease is expected to be determined by several different genes because, under the present set of methods used for the analysis of trait inheritance, it was found impossible to reduce their genetic control to be the effects of one or two genes. On the other hand, there are quantitative traits, such as the activity of some enzymes that are controlled by one or two well-known structural genes.

2. SAMPLE SPACE

2.1. True pedigrees

Any particular pedigree study implicitly defines the population under study as a set of discrete units, *true pedigrees*. **We define a true pedigree to be such that the relationship connecting any pair of its members can be determined, and there is no other individual whose relationship with any of these pedigree members can be established.** Some of the true pedigree members may not be available for observation but, as we shall show, their existence can be established. Let τ denote both the structure (i.e. the relationships) and the phenotypes (trait values, discrete or continuous) associated with the members of a true pedigree.

The set of all pedigrees defined in this way constitutes what we shall call the population under study, in contrast to the real population from which these true pedigrees are formed. This basic concept of pedigree analysis has still not received a clear and proper explanation (but see an attempt to clarify this problem in Thompson, 1986) and so we discuss this definition here in some detail.

To start with, let us stress that it is quite natural, given that the sampled objects are pedigrees, to consider the set of true pedigrees $\{\tau\}$ to be a population from which the pedigrees are sampled. However, we call the pedigrees in this population under study "true" pedigrees to distinguish them from the pedigree structures that occur in the actual sample, which represent the sampled parts of the true pedigrees.

Let us now define the correspondence between the real population from which the pedigree samples are collected and the population of true pedigrees under study. This correspondence is

determined by two kinds of factors. The first is determined by the specific population, its origin and its history, including demographic, social, political and other processes. These are factors that are usually uncontrolled and, what is important, unknown - and, therefore, cannot be properly documented. Said mathematically, the real population is mapped into a set of disjoint true pedigrees whose structures are limited by a set of factors, usually unknown, that makes it impossible to establish and document all the relationships that connect members of the real population.

The second kind of factors determining the population of true pedigrees is more subjective. **Pedigree analysis begins with the definition of a *sampling design* introduced by the investigator when the study is at the planning stage. This design determines which pedigrees are to be sampled from the set of true pedigrees, and how. Practical execution of this design requires that we use an instrument, with which we can learn the structure of the true pedigree from which a sampled pedigree comes. We shall call the instrument used to do this a *questionnaire*.**

Thus, for any real population, these two sets of factors determine the population of true pedigrees $\{\tau\}$ from which the pedigrees for analysis are sampled. Depending on the factors determining the structures of the true pedigrees and on details of the sampling design (i.e. how the questionnaire is defined), different sets $\{\tau\}$ can be formed from the same real population. Note that the *implicit* way in which $\{\tau\}$ is formed means that no attempt is to be made to draw up a complete list of true pedigrees. This is not only unnecessary, but also, in practice, impossible for a large population. This set can nevertheless be defined in principle; at least, the

structure of the true pedigree from which each particular sampled pedigree comes, and hence is included in the pedigree sample analyzed, can be learned by using the questionnaire.

Consider now the results of a pedigree analysis. Because the pedigrees are sampled from the set $\{\tau\}$, the analysis result, i.e., the genetic model providing the most accurate description of the trait inheritance, is applicable and relevant to only this set $\{\tau\}$. It may or may not be relevant to the real population because, as follows from the above considerations, there is no one-to-one correspondence between the real population and the set of true pedigrees under study. This means that we study the trait inheritance in the population $\{\tau\}$ rather than in the real population. It is usually assumed that $\{\tau\}$ *represents* the real population in which the investigator performs the study, so that the analysis results may also be applicable to this real population, if $\{\tau\}$ represents it adequately. However, the adequacy of such a representation is not in general guaranteed. Thus, although the results of any pedigree analysis should be applicable to the population $\{\tau\}$ by definition, applicability of the results to the real population is not the problem that is solved.

Consider the following two examples. In the first, a more or less isolated population is under study. After some thorough questioning and rechecking of answers (including, if necessary, paternity testing and other techniques), it is possible to reconstruct the relationships among all (or a large number of) the living members of this population, forming in such a way a single true pedigree that adequately represents the population. Many such attempts to reconstruct true pedigrees have been reported, for example by Bonné et al. (1970) for the Habbanite isolate of Jews, and by Neel (1978) for the Yanomama Indians of the Amazon rain forest. Clearly,

any analysis result (segregation and/or linkage) that is obtained on pedigree samples collected from such true pedigrees, adequately reconstructed, is equally applicable to both the {τ} population and the real population.

In the second example, the sampling design dictates that only sibships are to be sampled. In this case, all relationships outside each sibship are consciously neglected. The {τ} population consists of the sibships that can be formed from members of the real population using the questionnaire and, therefore, each analyzed sample consists of sibships. In this case, we may doubt whether the analysis results (e.g., the genetic model of inheritance chosen), applicable by definition to this particular {τ} population, can be applied to the real population. Between these two extremes there can be a large number of intermediate sampling designs that determine how the {τ} population is formed to "represent" the real population, with different degrees of completeness and, therefore, allowing different interpretations of how the analysis results relate to the real population.

To sum up these considerations, we note that the real population, in which the study is performed, gives rise to a population of true pedigrees {τ}, and the definition of this new population depends both on properties of the real population being studied and on the sampling design introduced by the investigator. By definition, the trait inheritance is studied in the population {τ} from which the pedigree samples are collected. The analysis results are also applicable to this {τ} population. Extrapolating these results to the real population is not always justifiable, but would be justifiable in many particular cases. Using the questionnaire, we can determine the structure of any true pedigree from which a sampled

pedigree is collected. Without any questionnaire, no pedigree can be sampled. We therefore propose here that, when the questionnaire is used to determine the availability of pedigree members for study so that a pedigree sample can be collected, it is designed to collect at the same time the information, indicated below, necessary to correct the pedigree likelihood for the sampling procedures used.

2.2. Measures of model similarity

The sampling design the investigator established at the outset determines which pedigrees should be sampled and how. The design determines the *sampling procedures* used in the process of pedigree collection.

The sampling procedures S generate a *sample space* of pedigrees that can be in principle sampled from the $\{\tau\}$ population. Evidently, for each sampled pedigree (X,C), having phenotypic content X and the structure C, there exists at least one such true pedigree τ that satisfies $(X,C) \subseteq \tau$. Different procedures generate different sample spaces for the same population $\{\tau\}$. A more detailed formulation of the sampling procedure and how it is modeled will be given below.

The pedigree sample, collected from the particular sample space, is used to distinguish the set of genetic models preliminarily formed in the set θ from which the "best" is chosen as the pedigree analysis result. The only operational way to introduce a *measure of similarity* between models is to use their phenotypic expression, namely, their accuracy in describing the pedigree distribution. Let $\Phi(\tau|\theta)$ be the distribution of the true pedigrees $\{\tau\}$ generated by the particular genetic model θ. Thus, the information measure

$$I_{ij} = \sum_{\tau \in \{\tau\}} \ln[\Phi(\tau \mid \theta_i) / \Phi(\tau \mid \theta_j)]\Phi(\tau \mid \theta_0)$$

defines the relative similarity of models θ_i and θ_j to θ_0 in the population $\{\tau\}$, where θ_0 is the true mode of trait inheritance

Given the particular sampling procedures *S* used in the process of pedigree collection, each model θ_i generates its specific distribution $P^S(X, C \mid \theta_i)$ of *derivative* pedigrees from the sample space defined by *S* and, therefore,

$$I_{ij}^S = \sum_{(X,C) \in \{(X,C)\}} \ln[P^S(X, C \mid \theta_i) / P^S(X, C \mid \theta_j)]P^S(X, C \mid \theta_0)$$

measures the similarity of θ_i to θ_0 in comparison with θ_j in the sample space generated by *S*.

Each of these information measures, I_{ij} and I_{ij}^S, is positive if θ_i is more similar to θ_0, negative if it is less similar, and equal to zero if θ_i and θ_j describe the pedigree (true or derivative) distribution with exactly the same accuracy. Different definitions of the sample space and different probability distributions for the derivative pedigrees could result in different signs for the information measures I_{ij} and I_{ij}^S. It is assumed here that, for each pair of probability models, θ_i and θ_j are absolutely continuous with regard to one another in the sense defined by Kullback (1959, ch. 2, section 2).

Thus, the pedigree analysis result is defined to be adequate if the model of trait inheritance having the first rank in θ provides the most accurate description of the pedigree distribution, even if the formulation of this model differs from the true one, and this difference may cause inconsistent estimators of some model parameters.

This definition seems reasonable because of the following. Sawyer (1990) showed that parameters of a genetic model of trait inheritance, as well as the nuisance parameters of a sampling model, can be consistently estimated only if these models are formulated correctly, meaning that they describe at least the most important features of the trait inheritance and sampling processes. If this is not the case, the parameter estimates are doomed to be more or less biased asymptotically. This bias is in practice indeterminable for most cases, which substantially limits the possibility of adequately interpreting any analysis results. The relation that these results of Sawyer bear to the above definition of pedigree analysis can be discussed as follows.

1) In practice, the set θ of trait inheritance models that are to be compared with one another in the process of pedigree analysis is quite limited. The main limiting factor is that of model complexity, expressed partially in the number of parameters that need to be estimated from pedigree data with usually limited information. This is why most often monogenic (major gene) models of trait inheritance are currently used in segregation and linkage analyses. This is also why it seems hardly reasonable to expect to find that the true model of trait inheritance, θ_0, is included in set θ - because the true genetic and environment control of the trait inheritance is usually (much) more complicated than what can be formulated as a testable model. This means that all the models in the set θ differ from θ_0 and all of them are doomed to produce inconsistent estimators of their parameters. The model providing the “most accurate” description of the pedigree (true or derivative) phenotypic distribution is usually not the true one, and is chosen among others that also differ from θ_0. It is important to stress that, for each particular set θ that is formed, it is

in practice impossible to evaluate how similar to θ_0 the chosen model is, whether in its construction (the three distributions defined on the given set of genotypes, **G**), or in its accuracy in providing a formal description of the mode of inheritance.

2) It is well known that, for finite-size samples, the maximum likelihood estimators of model parameters are usually biased, more or less. This is true for any particular model of trait inheritance, including the true one, θ_0. Moreover, for finite-size samples, consistent estimators are not necessarily less biased than inconsistent ones - at least, as yet the opposite has not been proved. This holds until the sample size increases up to a level that the asymptotic properties considered by Sawyer become true, this level being in practice indeterminable. Thus, the bias in estimators of the trait inheritance model should be considered an unavoidable annoyance of pedigree analysis. It seems reasonable to become reconciled with it and to agree that the particular estimates of each of the model parameters bear only a secondary importance as soon as the choice of the "best" model from the given set θ is correctly made. Being genetic parameters, they keep their genetic relevance and the estimates can be interpreted in terms of the genetic model only after the model is determined, i.e., accepted as the analysis result.

Here and below, a pedigree analysis result will be called *correct* if the model chosen from θ for describing the trait inheritance (the analysis result) is the one that is most similar to the true inheritance model, even if the chosen model cannot be characterized by consistent parameter estimators. As follows from Sawyer (1990), the analysis result would be not only correct, but also *consistent,* if the true model of the trait

inheritance, θ_0, is included in the set of tested models (which we hardly expect to occur in practice).

2.3. Sampling procedure and pedigree subsets

Let us define the sampling procedure as a combination, $S = (\alpha, \varepsilon, \psi)$, of three separate sub-procedures, the pedigree ascertainment α, the pedigree extension ε, and the pedigree inclusion ψ. Below, each of these sub-procedures is considered in some detail.

Here we will consider only ascertainment through *probands* (a term first introduced in the English language by Fisher, 1934). **A proband is defined here as an individual who, because of his/her characteristics, is ascertained (not necessary independently of other such individuals) and becomes the cause of his/her pedigree entering the sample for study. (It should be carefully noted that this definition differs from the usual one that requires independent ascertainments as part of the definition of a proband).** For each true pedigree, let us define a subset of its members who could *potentially* become probands by reason of how, independently of their phenotypes, the probands are defined. Typically, a geographic area or a catchment area of one or more hospitals defines this subset. Let τ_p be the subset of potential probands in the true pedigree τ, both their relationships determining the subset structure $C(\tau_p)$ and its phenotypic content $X(\tau_p)$. Using the terminology due to Elston and Sobel (1979), this subset is called the *proband sampling frame* (PSF) of the pedigree. We shall use the term *proband combination* (PC) to denote a set of persons who simultaneously cause the pedigree to enter the sample for study – without any implication that they do so independently. For a pedigree with $n_p = |\tau_p|$ PSF members, $2^{n_p} - 1$ is the maximum number of

actual PCs that can be formed and ascertained in the sampling process. The initially ascertained part of a pedigree, its PC, is then extended by incorporating other members of the true pedigree into the sample. The PC constitutes an obligate part of any sampled pedigree, but the pedigree can also contain other non-proband members, who may differ for the different PCs. The set $\{\tau_p\}$ of subsets of potential probands in the true pedigrees is the second characteristic of the studied population that determines the ascertainment design.

Let C and C_1 be the structures of a sampled pedigree and its initially ascertained part, the PC, and X and X_1, respectively, be sets of phenotypes observed on members of these two structures: $C \supseteq C_1$ and $X \supseteq X_1$. By $C_2 = C \backslash C_1$ we shall denote the complementary structure of the subpedigree collected in the extension process. We assume that the pedigree extension (intrafamilial sampling) does not distinguish the τ_p members of the true pedigree from other τ members. By definition, $(X,C) \subseteq \tau$, and $(X_1,C_1) \in \tau_p$. Further, let $(X_p,C_p) = (X,C) \cap \tau_p$ be the subset of sampled potential probands (the sampled part of the pedigree PSF), denoting both their relationships to each other and their phenotypes. It is always assumed that the PC is an obligate part of this subset. By definition, $(X_1,C_1) \subseteq (X_p,C_p) \subseteq \tau_p$. Fig. 1 illustrates these definitions.

Thus, any particular sampling event initiated by the pedigree ascertainment divides all members of each sampled pedigree, (X,C), into three distinctly different categories (Fig. 1): $(X,C) = (X_1,C_1) + (X_p,C_p) \backslash (X_1,C_1) + (X,C) \backslash (X_p,C_p)$, where the first is the PC – the initially ascertained unit; and the two others, sampled in the process of pedigree extension, are: the sampled potential probands in $(X_p,C_p) \backslash (X_1,C_1)$ that have not realized their potentials (have not become probands), and the

individuals in $(X,C)\backslash(X_p,C_p)$ who just cannot become probands under the given ascertainment scheme.

Assume further that there is selective inclusion of the ascertained, and further extended, pedigrees in the sample that is subjected to pedigree analysis. The inclusion of each pedigree in the sample analyzed could be determined by requiring it to meet a condition in the following way. In each pedigree, define a substructure $C^a \subseteq C$ in such a way that some characteristics of its members affect the probability of including this pedigree in the sample that is analyzed (Fig. 1A). These characteristics may have nothing in common with the studied phenotype, or they may include the phenotypes (they may contain, for example the marker data collected for linkage analysis – see below). One possible version of this condition is to analyze only pedigrees having parents not exceeding a certain age, or living only inside a prescribed district. In this case, the phenotype data do not define the inclusion process. Another condition could be that at least one parent should be "affected". (The fact that a parent is affected would not be connected with the ascertainment process if the ascertainment is performed via offspring probands). In both these examples, the additional substructure C^a contains all pairs of spouses in the sampled pedigree. If the inclusion condition is somehow related to the number of affected members in the sampled pedigree, then, evidently, $C^a \equiv C$. Denote by X^a the set of phenotypes observed on the members of this substructure, C^a (this set can be empty). We assume that, for a given pedigree, there is only one substructure C^a uniquely determined by the sampled pedigree structure C, although the converse is not true: different pedigree structures can contain the same substructure C^a. Note that this is not the case for the sampled PSF structure: different structures C_p may be

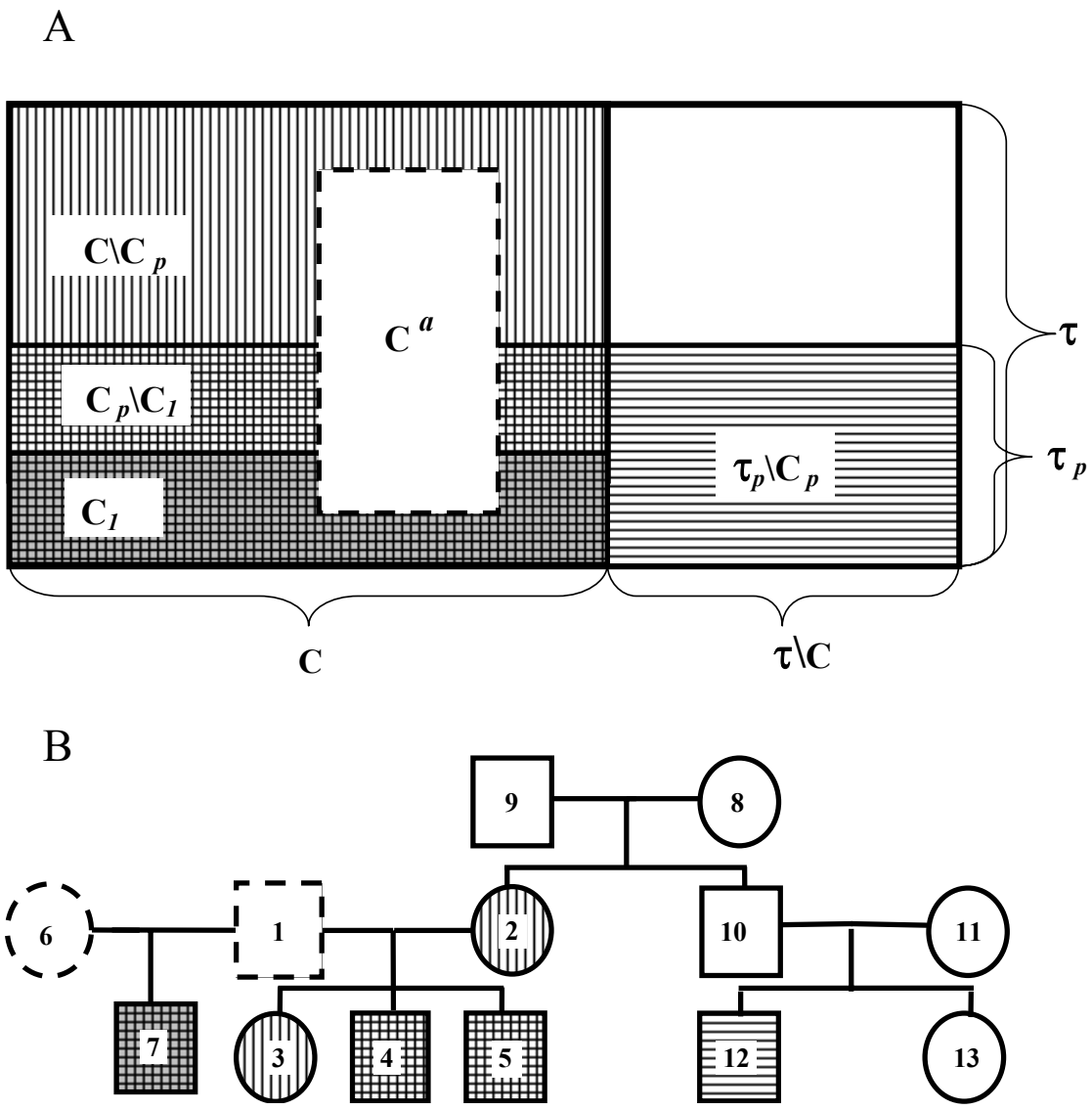


Figure 1.

A. Schematic representation of the true and sampled pedigree substructures. τ - the true pedigree; τ_p – the true pedigree PSF; C – the sampled pedigree structure; C_1 – the initially ascertained PC (proband combination); C_2 – the complementary part sampled in the process of pedigree extension; C_p – the sampled part of the pedigree PSF; C^a – the pedigree substructure responsible for inclusion of the pedigree in the sample analyzed. See details in sections 2.3.

B. Example illustrating the schematic representation. The members included in the pedigree are marked by the same hatching as used in Fig.1.A.

associated with the same pedigree C. Clearly, C^a can overlap with C_p (and even with C_1), so that there can be some pedigree members whose phenotypes affect the pedigree ascertainment and at the same time determine whether the pedigree is to be included in the sample that is analyzed.

Thus we assume that, to be sampled (more accurately, to be included in the sample subjected to pedigree analysis), the pedigree should first be ascertained in accordance with a specified proband ascertainment scheme; then it should be extended according to a given extension rule; and then it is censored according to whether the phenotypes in its predefined substructure C^a are compatible with the specified condition. Operationally, it is not necessary for the second and third stages of this sampling process to be separated from one another. Sometimes, the specified condition can be checked in parallel with the extension process. In this case, the intrafamilial sampling stops as soon as the condition is not fulfilled, e.g., a nuclear family is encountered that contains less than 2 affected members. However, note that this can lead to a substantial difference in the sampling result. If the condition (e.g., at least two affected members in each nuclear family) is included in the extension model, then the pedigree extension is stopped but the collected pedigree is left in the sample to be analyzed. If, on the other hand, this condition is a part of the inclusion model, then the collected pedigree is *excluded* from the sample analyzed. That is why these two procedures, which yield different sampling results, should be considered separately.

2.4. Example

To make the above definition more understandable, consider the following example. Let there be a set of true pedigrees each consisting of

13 individuals, with relationships determined by the graph drawn in Fig. 1B. Assume that only men under a certain age can be probands, and individuals 4, 5, 7 and 12 satisfy this condition. This means that the true pedigree is represented by the set $\tau = \{1,2,...,13\}$, while the ascertainment procedure defines its PSF as the subset $\tau_p = \{4,5,7,12\}$. There are 2^4-1 = 15 different PCs that could cause the ascertainment event.

The sampled pedigree structure C is determined by the initially ascertained PC and the extension rule determining what relatives of these PC members are to be observed, and in what order. Assume that the extension model is formulated as follows. Observe all first-degree relatives of members of the ascertained PC. Consider each member of the observed part of the pedigree as a *pointer*, if his/her still unobserved first-degree relatives are available for observation (Lalouel and Morton, 1981). If the pointer's phenotype happens to be of a certain *a priori* defined type, then his/her first-degree relatives are to be observed, including in this way one additional nuclear family into the sampled pedigree. Otherwise, relatives of this pointer are not to be observed. Let us denote this "extension" type as x^+, and the alternative type as x^-. If a quantitative continuous trait is considered, the extension is made if the trait value in the pointer belongs to the trait interval x^+, otherwise no extension is made through this pointer. Assume that, by this definition, individuals 1, 2 and 10 can be pointers as soon as they are sampled.

Fig. 2 shows some of the pedigrees that can be sampled if the PC contains either proband 4 or proband 5, or both of them. In each case, the extended pedigree part, C_2, is determined by the phenotypes of the pedigree pointers: whether each of them belongs to the "extension" type (denoted by +), or not (–).

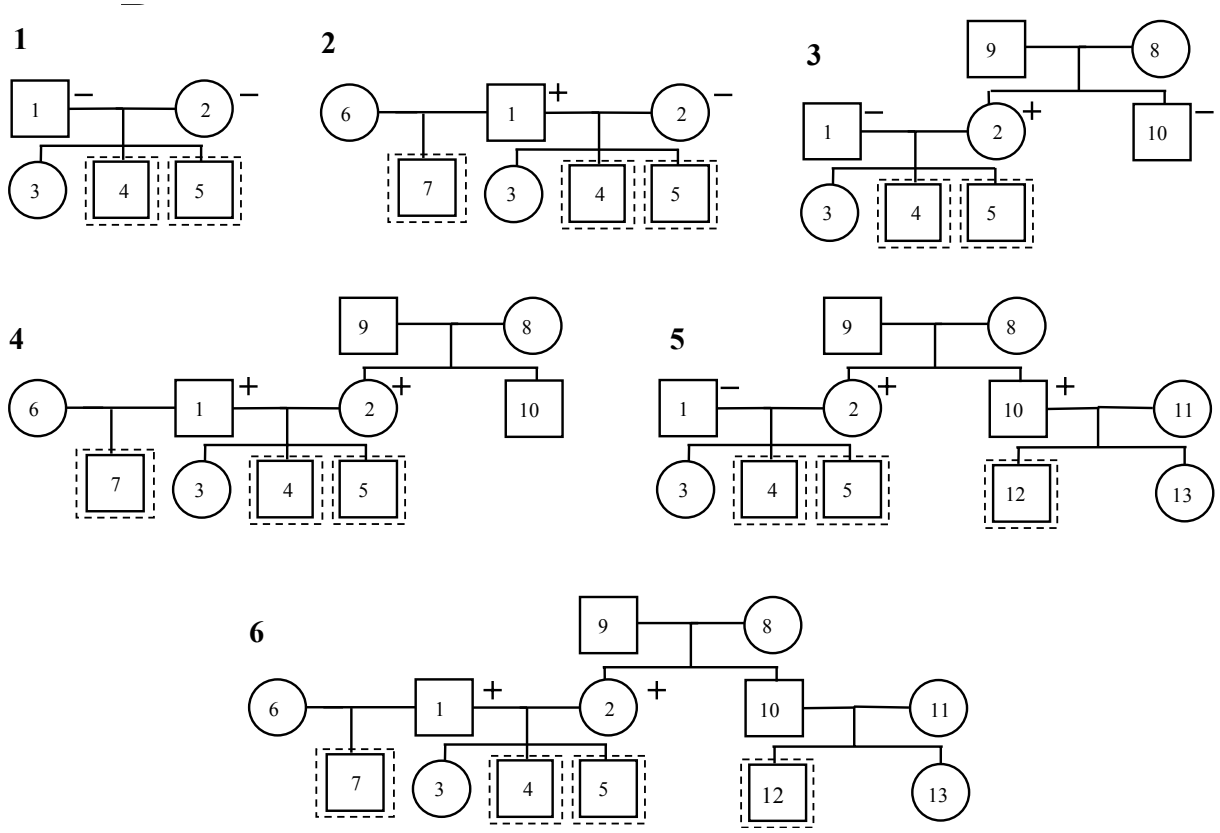


Figure 2. Structures of the 6 pedigrees that can be sampled from the true pedigree presented in Fig.1.B through a proband combination consisting either of only the 4th member, or of only the 5th, or of both of them. The sampled structures depend on the pointers' phenotypes. See details in sections 2.4.

Finally, let the inclusion rule be as follows. The sampled pedigree is included in the analysis only if each nuclear family collected at the first extension stage, i.e., the nuclear family containing the actual proband(s), has at least one parent with a phenotype of the "inclusion" type z^+ and it is not included otherwise. In this case, each of the sampled pedigrees in Fig. 2 has the same subset $C^a = \{1,2\}$, and would be included in the analyzed sample only if either pedigree member 1 or 2, or both of them, have the "inclusion" phenotype z^+. As we can see, the sampling procedure

considered here: (i) is proband dependent (PD – a term defined by Vieland and Hodge 1995), because the sampled pedigree structure depends on the ascertained PC, (ii) contains a trait-dependent pedigree extension, and (iii) contains trait-dependent pedigree censoring.

2.5. Planned and employed procedures

The sample space, i.e., the set of pedigrees that can be sampled from the set of true pedigrees $\{\tau\}$, is defined by two conditions. An objective condition determines the population of true pedigrees, while a subjective condition is defined by the initially introduced sampling design that determines the set of pedigrees that can be collected for the analysis.

In this connection, the following should be noted. There is a difference between the planned sampling procedures and the ones actually employed. In practice, there are a number of various social-demographic factors in the population under study that affect the exact fulfillment of the planned sampling design. Usually, it is impossible to make a complete list of these factors because they are unknown. Thus, for each initially planned sampling design, there is a set of different possible sampling procedures that could in practice be actually employed. Which one happens to be employed in each particular case is unknown and, therefore, cannot be explicitly defined when the sampling procedures are formulated according to the sampling design introduced above. The only possible way to describe the sampling process in formal terms is to formulate the sampling design as initially planned, neglecting the (unknown) procedures of pedigree collection that are actually employed in the particular realization of the pedigree collection.

Thus, the formulation of the sampling procedures defined above (the pedigree ascertainment, extension and censoring) can be made using the initially planned sampling design and sampling procedures, but not using the unknown procedure that are in fact used in practice.

2.6. Adequate sampling

Here we introduce the definition of an adequate sampling procedure – one of the basic items underlying the theory of pedigree analysis.

Among the possible sampling procedures, let us single out one that is special or, to be more accurate, degenerate, $S_0 = (\alpha_0,\varepsilon_0,\psi_0)$, which we will call here the *zero-sampling* procedure and define as follows. Let the initially ascertained PC be not related, either directly or indirectly, with the trait studied, so that the probability of ascertaining any part of the pedigree, $P(asc \mid X_1, C_1, \tau_p, \alpha_0)$, does not depend on the trait content of this part or on the phenotypes of any other sampled or true pedigree members. It terms of the ascertainment probability, this means in turn that, without loss of generality, we can put: $P(asc \mid X_1, C_1, \tau_p, \alpha_0) \equiv 1$. Let the pedigree extension model, ε_0, also be trait-independent, so that any initially sampled sub-pedigree is further extended incorporating all relatives available for observation regardless of their phenotypes and with a random (trait-independent) stopping rule. Lastly, let no trait-dependent censoring occur in the degenerate inclusion procedure ψ_0, which means that we can assume, without any loss of generality, $P(incl \mid X^a, C^a, \psi_0) \equiv 1$. In sum, under S_0 any *(X,C)* is randomly selected from those true pedigrees that contain this sub-pedigree. Thus, the sample space contains

any sub-pedigrees that can be carved out from the given set of true pedigrees, including all the true pedigrees themselves. The pedigree distribution generated on this sample space is defined as:

$$P^{S_0}(X,C\,|\,\theta_0) = \sum_{\tau\in\{\tau\}} \Phi_i(\tau\,|\,X,C,\theta_0)\,.$$

Let us assume that, *for each particular set*, $I_{ij}^{S_0} > 0$ *if, and only if,* $I_{ij} > 0$. This means that the ranks of the models in θ are the same whether obtained by using I_{ij} on the true pedigree set $\{\tau\}$, if this set is obtainable for observation, or by using $I_{ij}^{S_0}$ on the pedigree set generated by the zero-sampling model S_0. In turn, this means that the pedigree analysis performed on the samples obtained without any special sampling procedure (ascertainment, extension and inclusion) provides correct results; i.e., from the *a priori* formed set θ it always chooses the model that is most similar to the true model θ_0.

The sample space generated by any other, non-zero, sampling procedure differs from that generated by S_0. We see from the above description of the sampling sub-procedures, ascertainment α, extension ε, and inclusion ψ, only *S*-specific sub-pedigrees are carved out from the true pedigrees $\{\tau\}$. This is why the following definition of adequate sampling is now introduced:

The sampling procedure S is adequate if, for any pair of models in θ, *it provides* $I_{ij}^{S} > 0$ *if, and only if,* $I_{ij}^{S_0} > 0$, which, taking into account the previous assumption, means *if, and only if,* $I_{ij} > 0$.

This means that only those *S* are adequate, i.e., are permitted to be used in pedigree analysis, that provide exactly the same ranking order of

the genetic models in the given θ as would be obtained for the zero-sampling procedure or for the set of true pedigrees. Thus adequacy means that the genetic model chosen under sampling procedure *S* (the pedigree analysis result) is expected to be the same (up to the accuracy of the statistical operations performed) as that which would be chosen on the basis of the true pedigrees, were they available for observation. This way of formulating the condition of adequacy is sufficient, but not necessary because, strictly speaking, we are only interested in the model most similar to θ_0 being ranked first in order to provide the correct result of a pedigree analysis; ranks of other models in θ are of no interest to us.

Even after introducing the concept of an adequate sampling procedure, we are still unable to list the necessary conditions that need to be fulfilled in order to formulate such a procedure. However, we can point out some obvious conditions that make the sampling procedure inadequate. First among these is any sampling procedure that is explicitly based on a model of how the trait is inherited. Except for the zero-sampling procedure, the component sub-procedures are usually determined by the phenotypic contents of certain pedigree subsets - the pedigree PSF, the set of pointers determining the pedigree extension, and/or the phenotypes of the members responsible for inclusion of the pedigree in the sample analyzed. However, the dependence of these sub-procedures on the phenotypic content of these subsets in no way implies their dependence on the trait inheritance model. Up to now, we are lucky in that we have not yet met a sampling procedure that depends on the trait inheritance model, e.g., on the major gene frequency or on genotype penetrance. It is clear that such a model would induce a bias in ranking the inheritance models

tested and, therefore, the analysis result would also be expected to be biased.

On the other hand, even if the sampling procedure does not depend on the mode of inheritance, it could still provide inadequate results if it is in some sense not coordinated with the mode of inheritance. Consider, for example, the heterogeneity that occurs when there is polygenic control of low and intermediate trait values, while a major gene mutation causes large trait values. Such a mixed model of trait control could occur quite often. In this case, if it is mostly probands who have large trait values that cause the pedigree ascertainment, then we can expect major gene models to be chosen more frequently as the result of any analysis. Although such a result would not be incorrect, it would not provide a completely satisfactory description of the mode of trait inheritance.

These two examples show that a complete definition of an adequate sampling procedure (and its modeling) is hardly possible; but this, of course, does not make our definition unnecessary.

3. PEDIGREE LIKELIHOOD

Pedigree analysis is performed on sampled pedigrees collected from the set of true pedigrees $\{\tau\}$. The subset of pedigrees that in principle can be sampled is defined by the sampling design planned at the outset of the study by the investigator. This subset is called the *sample space*. It is generated by the sampling procedures defined by the sampling design (the pedigree ascertainment, the intrafamilial extension, and the selective inclusion in the sample analyzed). We now derive a probability model for a pedigree that can be sampled.

3.1. Component probabilities

The ascertainment process is modeled by $P(asc \mid X_1, C_1, \tau_p, \alpha)$, the probability (mass or density) of the pedigree being ascertained given its PC and the particular model of ascertainment α. We assume that this probability does not depend on the model of trait inheritance. However, it is determined by the subset τ_p of potential probands in the true pedigree from which the ascertainment takes place. Accordingly, $1 - P(asc \mid X_1, C_1, \tau_p, \alpha)$ is the probability that the particular PC, (X_1, C_1), formed from the given τ_p would not cause the pedigree to be ascertained.

Let $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$ be the probability of collecting the pedigree complement, (X_2, C_2), of the initially ascertained PC. In addition to the probabilistic correspondence between the complementary structure and phenotypes of its members, determined by the trait model of inheritance $\theta \subseteq \boldsymbol{\theta}$, this probability is determined by the model ε (and its

parameter(s), if specified) of the pedigree extension. Some modeling details of this extension will be considered below. Accordingly,

$$P(X,C \mid \theta,\varepsilon) = P(X_2,C_2 \mid X_1,C_1,\theta,\varepsilon)P(X_1,C_1 \mid \theta)$$

is the probability of a sampled pedigree having the structure C and phenotypic content X. Here, $P(X_1,C_1 \mid \theta)$ is the joint probability of phenotypes in the PC members given the model θ. As we can see from these definitions, the probability $P(X,C \mid \theta,\varepsilon)$ is defined only if the initially ascertained substructure C_1 is defined for each C.

Finally, let $P(incl \mid X^a,C^a,\psi)$ be the probability that the pedigree is included in the sample subjected to analysis if it contains the particular subset of phenotypes X^a in the pedigree substructure C^a. This probability can depend on some parameter(s), ψ, modeling the inclusion procedure. The inclusion probability can then be determined as: $P(incl \mid X^a,C^a,\psi)$ = 1 if at least one C^a member is of the "inclusion" type, and = 0 otherwise. This probability could also be defined in another way. For example, following the approximation proposed by Stene (1977, 1978) the probability can be formulated as $P(incl \mid X^a,C^a,\psi) = const \cdot k^{\psi}$, where k is the number of C^a members that are of the "inclusion" type, and ψ (≥ 0) is a parameter that determines how the probability of pedigree inclusion depends on this number. Let us consider some special cases of this inclusion scheme. For example, the case $\psi = 0$ could be called "complete" inclusion when all pedigrees having a substructure C^a containing at least one member of the inclusion type have the same probability of being included in the sample; and $\psi = 1$ could be called "single" inclusion when the probability of inclusion is proportional to the number of members in C^a who are of the inclusion type.

3.2. General form of the pedigree likelihood

The probability (mass or density) that the pedigree is of type (*X,C*) with its particular PC (X_1,C_1), conditional on having been sampled, is called the pedigree likelihood and can be expressed as:

$$P(X,C,smpl \mid \theta,\varepsilon,\alpha,\psi) =$$

$$\frac{P(X,C \mid \theta,\varepsilon)P(asc \mid X_1,C_1,\tau_p,\alpha)P(incl \mid X^a,C^a,\psi)}{P(smpl \mid \theta,\varepsilon,\alpha,\psi)}, \quad (3.1)$$

where the normalizing coefficient

$$P(smpl \mid \theta,\varepsilon,\alpha,\psi) =$$

$$\sum_{(X,C)} P(X,C \mid \theta,\varepsilon) \sum_{C_1 \subseteq C_p} P(asc \mid X_1,C_1,\tau_p,\alpha)P(incl \mid X^a,C^a,\psi)$$

defines the sample space for the given sampling procedures; it is the probability that at least one pedigree is sampled, i.e., ascertained under the given ascertainment model α, further extended in accordance with the extension model ε, and then censored according to the model ψ.

Note that $\sum_X P(X,C,smpl \mid \theta,\varepsilon,\alpha,\psi) = P(C \mid \theta,\varepsilon,\alpha,\psi)$, which means that, in general, the distribution of the sampled pedigree structures depends on details of the sampling procedure (and the way it is modeled) and, if the execution of this procedure is determined by phenotype values in the sampled pedigrees, this distribution may depend on the model of inheritance being studied.

Thus, on the given sample space generated by sampling procedure *S,* we express the likelihood of a pedigree (*X,C*) as the probability that it can be collected from the set of true pedigrees $\{\tau\}$; and a specific likelihood is determined for each model of trait inheritance tested:

$$P^S(X,C \mid \theta_i) = P(X,C,smpl \mid \theta_i,\varepsilon,\alpha,\psi).$$

3.3. Sample likelihood

Most generally, pedigree analysis is performed on a *sample* of pedigrees $\{(X_k,C_k)\}$ each of which is ascertained, extended and censored in accordance with its own specific sampling procedure S_k. As a particular case, all the pedigrees in the sample could be sampled using the same sampling scheme ($S_k = S$), but this is not necessary. Some of the pedigrees may be collected purely randomly, some may be ascertained through probands and extended regardless of the phenotypes of their members, and others may be extended using some phenotype-dependent procedure. If, as is usually the case, the pedigrees are sampled independently of one another, then the sample likelihood is simply the product of the likelihoods for the pedigrees included in the analysis (or equivalently the sum of their log-likelihoods):

$$L(\{(X_k,C_k)\}\,|\,\theta_i,\varepsilon,\alpha,\psi) = \prod_k [P^{S_k}(X_k,C_k\,|\,\theta,\varepsilon)]^{N(X_k,C_k)} =$$

$$= \prod_k \left[\frac{P(X_k,C_k\,|\,\theta,\varepsilon)P(asc\,|\,X_{1k},C_{1k},\tau_{pk},\alpha)P(incl\,|\,X^a_k,C^a_k,\psi)}{P_k(smpl\,|\,\theta,\varepsilon,\alpha,\psi)} \right]^{N(X_k,C_k)},$$

where $N(X_kC_k)$ is the number of sampled pedigrees having the same (X_k,C_k), i.e., the same PC, (X_{1k},C_{1k}), the same subset of collected potential probands (X_{pk},C_{pk}), the same subset censoring the pedigree inclusion (X^a_k,C^a_k), and the same subset of observed members who cannot be probands and do not affect the inclusion procedure, $(X_k,C_k)\backslash[(X_{pk},C_{pk})\cup(X^a_k,C^a_k)]$. For a quantitative continuous trait, it is most probable that $N(X_k,C_k) = 1$ for each sampled pedigree. Each pedigree likelihood should be conditioned on the specific sampling procedure used to collect it. From a formal point of view, the choice of a particular

condition for the pedigree likelihood is equivalent to the choice of the specific sample space on which the probability of this sampled pedigree is defined. As we see, pedigrees whose likelihoods are defined on different sample spaces can be present in the same sample (Note, in this connection, the statement made to the contrary by Hodge and Vieland, 1996). The only common requirement for all the sampled pedigrees is that the model of the trait inheritance be the same.

4. GENETIC MODELS FOR QUANTITATIVE TRAITS

The general definition of a genetic model of trait inheritance given above is given in detail by explicit formulation of its component distributions defined in (1.1), $p(g_1,g_2)$, $P(g|g_1,g_2)$, and $f(X_n|G_n)$. Clearly, models describing the inheritance of different traits should be formulated differently. In this chapter and the next, we review and compare the traditional ways of formulating the genetic model for quantitative traits, whether continuous or discrete, and for qualitative (binary, affected-unaffected) and complex traits.

In what follows, in order to simplify our considerations of the main operations involved in pedigree analysis, namely, the formulation of a mathematical-genetic model, the probabilistic description of the pedigree sample to be analyzed (the sample likelihood), and the formation of the set of models to be tested, we will use the simplest version of the genetic model, which we call the *major gene* (MG) model. This model explicitly includes two kinds of effects, the effect of a diallelic gene called MG, which forms a three-component genotypic set **G**, and all the other effects involved in the trait control that determine the joint phenotypic distribution among members of the sampled pedigree. As will be clear from the discussion below, this use of the simplest model in developing the theory in no way leads to a loss of generality of the conclusions we come to.

4.1. Population characteristics

The first distribution determining the genetic model (1.1) is the genotypic distribution of pairs of spouses. This is a characteristic of the population in which the inheritance of the trait is being studied. This does not mean that other model characteristics cannot be considered as being

relevant to the population structure. The population is determined by its specific gene pool, which implies population specificity of the set **G**. Particular environmental conditions, as well as the population's history, determine the trait distribution.

The distribution $p(g_1,g_2)$, where g_1 and g_2 are the genotypes of two spouses, is determined by the genotype frequencies and by the type of assortative mating occurring with respect to the trait under consideration. Under panmixia, $p(g_1,g_2) = p_{g_1} p_{g_2}$, where p_g is the population frequency of genotype g. For a single gene with two alleles, A_1 and A_2, the set **G** contains only three possible genotypes, A_1A_1, A_1A_2 and A_2A_2, which we can number g = 1, 2 and 3, respectively. The Hardy-Weinberg distribution of these genotypes is given by: $p_g = p^2$, $2p(1\text{-}p)$ or $(1\text{-}p)^2$ for g = 1, 2 or 3, respectively, where p is the population frequency of allele A_1. For any other genotype set (not MG) that follows an equilibrium distribution, each genotype frequency is proportional to the product of frequencies of the alleles constituting the genotype.

Assortative mating is a very complex process governed by a number of physical and social characteristics of the mates. Here, we define only that part of the general assortative mating effect that can be ascribed to the joint distribution of the spouse genotypes, with regard to a given genetic model describing the trait inheritance.

Formally, **any assortative mating effect is expressed in genetic model terms as the inequality:** $p(g_1,g_2) \neq p_{g_1} p_{g_2}$. A particular formulation and parameterization of the assortative mating effect for quantitative traits can be proposed as follows (Ginsburg et al., 1986). Assume that the probability of mating between a pair of individuals with genotypes g_i and g_j (g_i , g_j $\in$ **G**) is proportional to a factor

$q(g_i)q(g_j)\Psi(g_i,g_j)$, where $q(g_i) = p_{g_i}$ if the i-th spouse has no parents included in the analyzed pedigree (founder), and $q(g_i) = P(g_i|g_m,g_f)$ if g_m and g_f are genotypes of the parents of the i-th member. The proposed assortative mating factor is of the form: $\Psi(g_i,g_j) = exp[\eta(\mu_{g_i} - \mu_{g_j})^2/\sigma^2_\mu]$ where η is the coefficient of non-random mating; μ_g is the genotypic value of genotype g, and $\sigma^2_\mu = \sum p_g \mu^2_g - [\sum p_g \mu_g]^2$ is the genotypic variance. The normalizing factor for a pedigree likelihood is calculated for the whole pedigree instead of for each pair of spouses included in the pedigree.

Hasstedt (1995) proposed another parameter measuring the assortative mating effect, ϕ, - the correlation between the genotypic values in spouses. In nuclear pedigrees, for a given p and μ_g, there is one-to-one correspondence between the coefficient η and this correlation, ϕ, as expressed by the equation:

$$\phi = [\sum_{g_i g_j} p(g_i,g_j)\mu_{g_i}\mu_{g_j} - \mu^2]/[\sum_{g_i}\mu^2_{g_i}\sum_{g_j} p(g_i,g_j) - \mu^2],$$

where $\mu = \sum \mu_{g_i} p(g_i,g_j)$; $g_i, g_j = 1, 2, 3$.

However, for a pedigree of more complex structure that includes several nuclear pedigrees, the correlation ϕ (but not η) depends on the number of "inner" spouses (i.e. not founders) and on their particular positions in the pedigree. Thus, the correspondence between η and ϕ becomes ambiguous, depending on the structure of pedigrees under consideration. This is the penalty for not being able to formulate $p(g_1, g_2)$ independently of the pedigree structure.

4.2. Transmission probabilities

Each offspring genotype *g* is formed from two parental haplotypes (ξ,η) **or** (η,ξ)**, where** $\xi,\eta \in \{h\}$ **– the set of haplotypes from which offspring genotypes can be formed. We now introduce the *transmission probability*,** $P(\xi \mid g)$**, the probability that an individual having genotype** $g \in G$ **produces the haplotype** $\xi \in \{h\}$ **in the formation of his/her offspring genotypes** (Elston and Stewart, 1971). By definition, $\sum_{\xi \in \{h\}} P(\xi \mid g) = 1$ for any possible *g*.

Using these transmission probabilities, we can express the second distribution defining the genetic model (1.1), i.e., the distribution of offspring genotypes given the genotypes of their parents, or transition probabilities, $P(g|g_1g_2)$, as follows. For each $g = (\xi,\eta) = (\eta,\xi)$

$$P(g \mid g_1, g_2) =$$

$$= \frac{\upsilon(\xi,\eta)}{F}[P(\xi \mid g_1)P(\eta \mid g_2) + (1-\delta_{\xi\eta})P(\xi \mid g_2)P(\eta \mid g_1)], \quad (4.1)$$

where: 1) δ_{ab} is the Kronecker product symbol, used here to take into account the fact that homozygous genotypes transmit only one possible haplotype to their offspring; 2) $\upsilon(\xi,\eta)$ is a coefficient intended to take into account any departure from random association in the formation of the offspring genotype from the gametes obtained from the parents; and 3) the normalizing factor *F* is defined for the given pair of parental genotypes (g_1g_2).

This formulation of trait inheritance, the transmission of the genes from parents to offspring, can hardly be proposed in a general form because of (usually unknown) infringement of the classical Mendelian rules of gene transmission during gametogenesis and formation of the

offspring genotypes. Clearly, distribution (4.1) is determined by parameters that include both those determining transmission probabilities and parameters for the association of parental gametes in the offspring genotype. The most common assumption, used as an approximation, is that we have random mating and Hardy-Weinberg genotypic proportions.

4.3. Continuous quantitative trait

For a continuous quantitative trait, the conditional distribution of trait values among individuals having the same genotype g, $f(x|g)$, is usually assumed (after transformation, if necessary) to be normal with expectation μ_g (genotypic value) and residual variance σ^2_g. This residual phenotypic variation is caused by all the factors (genetic and environmental), other than the genes defined in the model **G**, which is the main factor responsible for the trait inheritance, that somehow affect the value of the trait under study. The residual co-segregation of the trait values in relatives is caused by two factors. First, potential minor-genes that are involved in the trait control, but have not been identified in the model explicitly as main genes controlling the trait variation, produce an interdependence of trait values among relatives sharing the same alleles of these minor genes. Second, common familial environmental (household) factors that influence the trait produce correlated modifications of the trait value in members of the same pedigree.

Usually, the n-variable normal approximation is used as the joint distribution of trait residuals $f(X_n|G_n)$ of the n members of a pedigree, the residual of the i-th individual being defined as $x_i - \mu_{g_i}$, where x_i is his/her trait value and μ_{g_i} is the genotypic value of his/her genotype $g_i \in \mathbf{G}$. This

distribution is determined by an $n \times n$ symmetric covariance matrix $\{\sigma_{g_i}\sigma_{g_j}r_{ij}\}$, where $\sigma^2_{g_i}$, $\sigma^2_{g_j}$ are the residual variances of genotypes g_i, $g_j \in \mathbf{G}$ and r_{ij} is the pairwise correlation coefficient between the residuals of relatives i and j . It should be stressed that the normal form of the distribution is only a more or less suitable approximation to describe the joint residual co-variation. After an appropriate transformation of the trait value it is often good enough, being justified by the central limit theorem of probability theory, and it is good operationally because during many decades statisticians have used it for various applications. However, the assumption of normality without a transformation can be unsatisfactory for a particular continuous quantitative trait. For this reason, allowing for a transformation such as the Box and Cox (1964) power transformation should be included as part of the model and tested.

The model parameterization of the n-variable normal distribution can be made in various ways, as follows.

First, the pairwise correlation matrix R can be parameterized in terms of variance components as has been incorporated in the program package PAP (Hasstedt, 2002) with the model parameters being polygenic heritability h^2 and shared environmental factors c_i^2. The number of shared environmental components is unrestricted. Each corresponds to a group of individuals who share a particular environmental effect. For example, all the members of nuclear family they may share a household effect, and in addition there may be sharing of spouses-specific and/or sibling-specific environmental effects. The resulting correlation matrix has in general non-zero elements for each pair of pedigree members. “Exact computation of the likelihood when the model includes correlations between pedigree

members requires summing over the probabilities of all combinations of genotypes for pedigree members; exact calculation requires too much computer time for pedigrees sizes exceeding about ten members" (Hasstedt, 2002).

In a second way of parameterization, which is also available in PAP, the pairwise correlations determining the R-matrix are taken to be the model parameters, with the following assumptions (Hasstedt, 2002, p. 30):

1) Only three types of correlations are introduced, namely, between spouses, *r*, between parents and offspring, *b*, and between sibs, *e*.

2) The residual correlation between any given pair of relatives (spouses, siblings, etc.) is the same regardless of either the particular position of this pair in the pedigree under consideration or of the pedigree structure.

3) The correlation between any pair of pedigree members not belonging to the same nuclear family equals zero.

However, for a sample of pedigrees with substantially different structures, it was shown (Ginsburg, 1997) that the last two conditions do not allow for minor-gene residual co-variability among relatives and lead, in particular, to non-interpretable values of *partial correlations* between the trait residuals of some pairs of relatives. *Partial correlation* between the trait residuals of two pedigree members means here the correlation after "partialling out" their correlations with the residuals of all other pedigree members. Just as the pairwise correlation coefficient between residuals of individuals *i* and *j* describes their general joint behavior, the partial correlation describes their joint behavior when the residuals of all other pedigree members, other than *i* and *j,* are held fixed.

Assume, for example, that independently distributed genes with additive effect are responsible for the non-major-gene residual co-variation between pairs of relatives. For a panmictic population, under pure polygenic inheritance the three pairwise correlations are expected to be: $r = 0.0$, $b = 0.5$ and $e = 0.5$ (see, e.g., Thompson, 1986, Chapter 6). The corresponding partial correlations, ρ between spouses, β between parents and offspring and ε between sibs, when calculated for a nuclear pedigree with 3 offspring are, respectively, -0.600, 0.447 and 0.00. As can be seen, while ε corresponds exactly to its expected value (no correlation between sibs when the parental genotypes are fixed), the large negative partial correlation between spouses that would be required if a polygenic component is not included in the model can hardly be interpreted. Consider, for example, the simplest extension of the nuclear pedigree - including grandparents, the parents of one of the two parents, into the pedigree - thus forming a 7-member pedigree. The above three residual correlations and a zero correlation between pedigree members who do not belong to the same nuclear pedigree (e.g. grandchild-grandparent) would form a *negatively-determined correlation matrix* and, therefore, no 7-variable normal distribution exists with these parameter values. The same is true for more complicated pedigree structures. These examples show that any approximate model for the residual co-variation should be formulated with care, to provide the possibility of a reasonable interpretation for the genetic model parameters.

Ginsburg (1997) proposed an other (third) way of parameterization in which the *partial correlations* between pedigree members are used as model parameters. Three assumptions, analogous to that of Hasstedt, are applied here to the partial correlation coefficients instead of to the pairwise

correlation coefficients. Correspondingly, three partial correlations are introduced as the model parameters: ρ between spouses, β between parents and offspring and ε between sibs. Conditions 2 and 3, but formulated in terms of the *partial correlations*, allow for the possibility of minor-gene residual co-variation among relatives. In particular, if the minor-gene genotype of an individual is fixed, no genotypic partial correlation is expected between residuals among his/her offspring or between his/her parents and siblings. The same is true for any partial minor-gene correlation between siblings, for fixed genotypes of their parents. The pairwise correlations between residuals of spouses, parent and offspring, or siblings, as well as those of any other pair of relatives, are expressed as functions of ρ, β and ε and depend, additionally, on the pedigree structure and on the position within the pedigree of each pair of relatives. In particular, the pairwise correlation between any pair of relatives does not necessarily equal zero, even if these relatives belong to different nuclear pedigrees. So conditions 2 and 3 (now formulated in terms of the partial correlations) seem more justified if there is minor-gene residual co-variation among relatives, if familial environmental factors cause the residual co-variation between relatives, then these conditions seem justifiable in this case as well.

When pedigrees of the same structure are under consideration, the two last parameterizations are equivalent, because one correlation triplet (r, b, e) has a one-to-one correspondence with a correspond triplet (ρ, β, ε), which follows from the formulae:

$$\rho_{ij} = -\frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}}; \rho_{ii} = 1; \text{ and } r_{ij} = \frac{\overline{P}_{ij}}{\sqrt{\overline{P}_{ii}\overline{P}_{jj}}}; r_{ii} = 1; \qquad (4.2)$$

where ρ_{ij} and r_{ij} are, respectively, the partial and pairwise correlations between residuals in the *i*-th and *j*-th pedigree members. The correlation matrix *R* was determined above; the matrix $\overline{P}$ has non-diagonal (*ij*) elements equal to - ρ_{ij}, while its diagonal elements equal 1. However, when pedigrees of different structure form the sample being analyzed, the difference in the choice of the parameter triplet for the model parameters can be important, at least in order to interpret the genetic model.

The program package S.A.G.E. (2004) includes the regressive models due to Bonney (1984) which, when the pedigrees are nuclear families, can include as a special case residual correlations as expected under polygenic inheritance. For more extended pedigree structures, these models do so approximately. In particular, the class D regressive model assumes that, conditional on the major genotypes and the phenotypes of the two parents, the residual correlation between the phenotypes of any pedigree member and a previous ancestor is zero. The grandparent-grandchild residual correlation, for example, is then the square of the parent-offspring residual correlation, in contrast to half the parent-offspring correlation (which is what is expected under polygenic inheritance).

4.4. Trait covariates

The model formulated above describes the joint variability of quantitative trait values among members of a pedigree without any reference to trait covariates, i.e., other characteristics of the pedigree members that, if observed, give additional information about the transmission of the trait across generations. Two groups of covariates should be distinguished. The first group contains those traits that have

common genes pleiotropically responsible for variation in another phenotype but also, indirectly, for the variation of the trait being studied. Their effects can be tested by a specially designed bivariate analysis - see below, section 5.3). The second group includes covariates that are not genetically transmitted or, at least, that share no genes in common with the trait being studied. Among these latter, an individual's sex and age are the most important. Many quantitative traits, such as body measurements, plasma concentrations of lipids or hormones, etc., show significant sex differences and change substantially with age and aging. Age-dependency may occur in quite different ways, this being a function of the nature of the trait and of the age interval studied. For example, bone traits related to aging (e.g., bone mineral density, or the metacarpal cortical index) hardly change between adolescence and the mid-forties, but then gradually decrease until death. Body height, on the other hand, increases until early adulthood and then remains virtually the same, although there is thereafter a gradual decrease with age. The serum concentration of many hormones, for instance parathyroid hormone, increases until adulthood and then gradually decreases with age.

Some models of inheritance of anthropometric traits incorporating various genotype-sex-specific interactions with age have been well described (e.g. Pérusse et al., 1991; Comuzzie et al 1995; Cheng et al, 1998). It is of special interest to mention the genotype-dependent effect of various covariates (age, sex, hormones, etc), also included as regression parameters in the regressive models of Bonney (1984) (the models were called regressive models because they include regression on the same phenotypes of previous relatives, just as autoregressive models include as parameters the regression on previous values of the same trait). These

models were successfully used in several studies of quantitative traits, including anthropometrics (e.g. Borecki et al, 1993; Mahaney et al, 1995; Lecomte et al., 1997). These publications can be referred to for details.

The simplest way to account for age and sex effects on inter-individual trait variation is the regression adjustment of the trait values for age and sex effects made prior to any pedigree analysis. The better way is to incorporate the age and sex effects explicitly into a penetrance function determining the MG effect, or any other genetic model, and this permits us to account for genotype-specific effects. In this approach, age and sex determine the trait inheritance model together with the other genetic parameters. Below, we consider a MG model that explicitly formulates genotype-sex-age interaction.

Let x_{gst} be the trait value in an individual having MG genotype g, sex s (m - male, f - female) and age t ($t > 0$) . The following linear model is assumed for this trait:

$$x_{gst} = \mu_{gst} + \xi = \mu_{gs} + \varphi_{gs}(t) + \xi,$$

where μ_{gst} is the expected trait value of individuals having the same g, s and t; ξ is the trait residual not affected by the MG, sex or age; μ_{gs} is the expected trait value of individuals having genotype g and sex s, and $\varphi_{gs}(t)$ is a function describing the genotype-sex specific age dependence of the trait value of these individuals; $\sum_t \nu_t \varphi_{gs}(t) = 0$, where ν_t is the frequency of individuals having age t.

Ginsburg (1997) proposed several genotype-sex specific functions of age. Here, we consider one particular formulation of $\varphi_{gs}(t)$ that is of special anthropological interest and can be used to approximate age dependent changes in bone anatomy:

$$\varphi_{gs}(t) = a_{gs}[t(1 - \delta(T_{gs})) + T_{gs}\,\delta(T_{gs}) - \bar{t}_{gs}], \qquad (4.3)$$

where a_{gs} is a slope coefficient measuring the rate of change in the trait per year; T_{gs} is a genotype-sex specific threshold, introduced in such a way that the trait increases (or decreases, if $a_{gs} < 0$) linearly with age after the latter exceeds this threshold, while it stays constant, with the value $a_{gs}(T_{gs} - \bar{t}_{gs})$, at early ages; $\delta(T_{gs}) = 0$ if $t \geq T_{gs}$ and = 1 otherwise, and $\bar{t}_{gs} = \sum_{t=0}^{T_{gs}} t\nu_t + T_{gs}(1 - \sum_{t=0}^{T_{gs}} \nu_t)$, which follows from the condition $\sum_t \nu_t \varphi_{gs}(t) = 0$.

4.5. Example

The above formulation of a covariate effect on the outcome of an inherited quantitative trait can be illustrated as follows. The metacarpal cortical index (CI) is the ratio of the combined cortical thickness to the total diameter of the bone. It serves as an indirect measure of bone mass and can be used in the prediction of osteoporosis. We tested whether genetic control of CI variation in large samples of pedigrees from Chuvashia, Russia (Karasik et al., 2000) can be satisfactorily described by a MG model of inheritance that includes (relatively) large gene effects and effects of other secondary genes and environmental factors. The trait shows quite pronounced age and sex dependence of the individual's genotype in determining phenotype. Thus, according to the above formulation of the genetic model, three main pleiotropic effects of the major gene were formulated. The first was major gene control of the baseline level of the CI, the second was the age at onset of involutive bone changes (i.e. the inflection point), and the third was the rate of decrease of

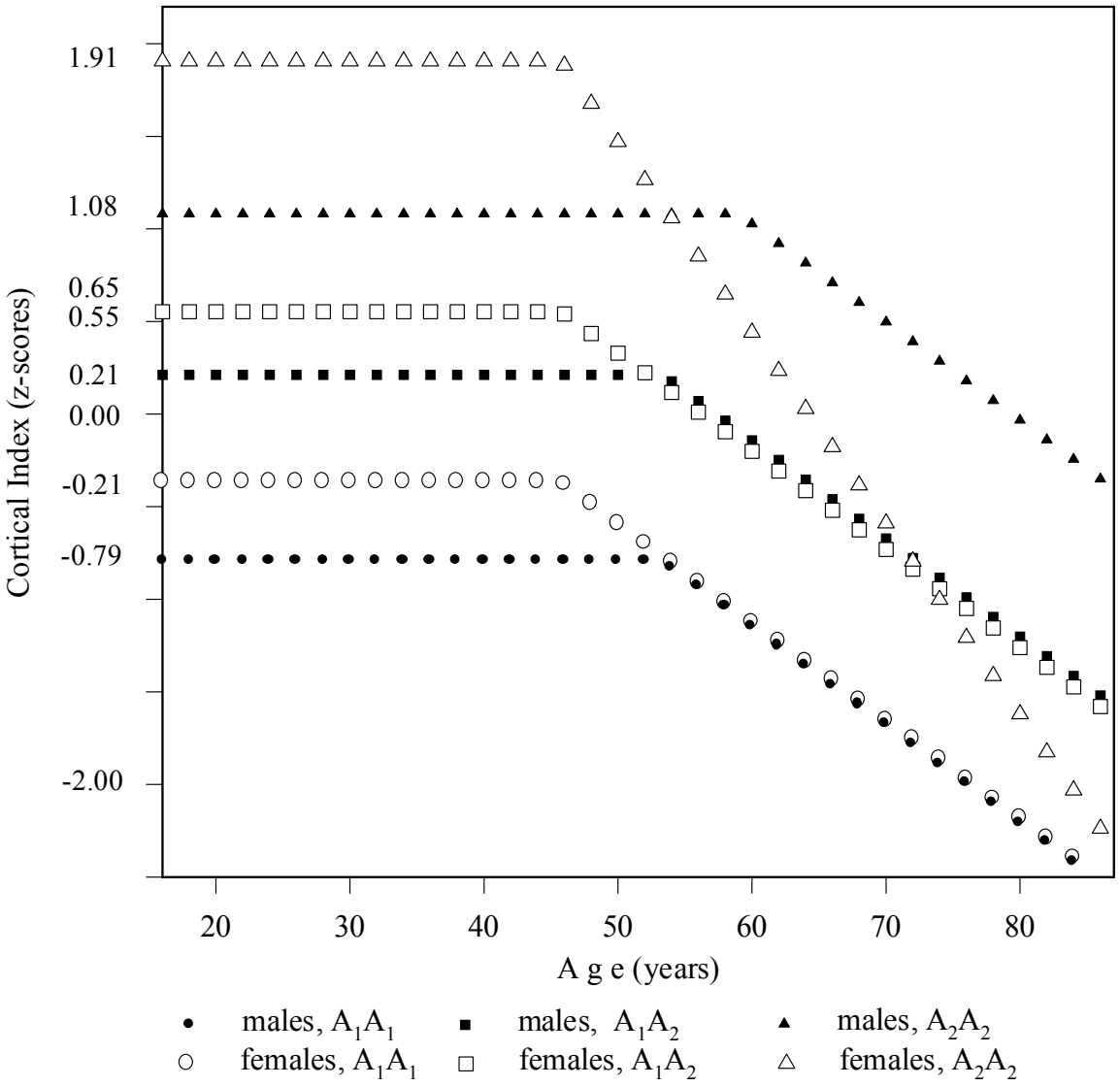


Figure 3. Dependence of the metacarpal cortical index (CI) on genotype, sex and age (from Karasik et al., 2000).

For each genotype, A_1A_1, A_1A_2 and A_2A_2, separately for males and females, the dependence of the genotypic value on age as given by formula (4.3) is shown. The CI values are shown in standardized units .See details in section 4.5.

the CI with age (i.e. the slope coefficient). The analysis showed that this complex model of CI inheritance can be statistically accepted. Non-major-gene effects shared by pedigree members (residual familial correlations) were found to be statistically non-significant. About 73% of the inter-individual variation in CI was attributable to the effects explicitly included in the model.

Figure 3 shows schematically the results obtained. Without dwelling on particulars of the analysis performed and any numerical results, we can summarize the main findings as follows. All the effects included in the model were found to be statistically significant (except the residual correlations). A difference was found in the inflection points between the different genotypes in males and females. A similar difference was found between the slope coefficients. Because this model was accepted in not just a single study, but also confirmed in other studies as a satisfactory model of CI inheritance, this result can be quite reasonably interpreted in biological terms and be of practical (predictive) interest.

4.6. Quantitative discrete trait

It is well known that individuals differ from one another in the thickness of their hair, and that this difference is not infrequently inherited. Consider this as a quantitative discrete trait expressed as the number of hair follicles per square unit of homogenous skin area. The distribution of this trait can be quite satisfactorily approximated by a Poisson distribution, i.e., the probability of observing exactly x follicles is given by $P(x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, where λ is the expected number of follicles per square unit of homogenous skin studied.

If there is no residual correlation between the number of hair follicles of relatives, the joint distribution of phenotypes in members of the pedigree, given their genotypes, is expressed as $f(X_n \mid G_n, \lambda) = \prod_{i=1}^{n} P(x_i \mid \lambda_i) = \prod_{i=1}^{n} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}$, where λ_i is the expectation of the follicle number in the *i*-th pedigree member. However, to take account of a possible correlation between relatives in the density of the follicles, the model should be formulated in a more sophisticated form. In particular, this can be done as follows. Assume that the expected follicle density in pedigree members, λ_i, has an *n*-normal distribution with variance-covariance matrix defined as in the previous section through partial correlations between spouses, parent-offspring and sibs. In this case, the joint distribution of phenotypes in pedigree members, given their particular genotypes, can be expressed as

$$f(X_n \mid G_n, \rho, \beta, \varepsilon) =$$

$$= \int [\prod_{i=1}^{n} P(x_i \mid \lambda_i)] N(\lambda_1, ..., \lambda_n \mid g_1, g_2, ..., g_n; \rho, \beta, \varepsilon) d\lambda_1 ... d\lambda_n \ ,$$

where the integration goes over all possible follicle numbers in each pedigree member.

4.7. Parameterization problems

In formulating the diallelic MG model above, we have attempted to make the model as general as possible, but keeping unaltered the structure of the model definition (1.5). Assuming that only one gene is responsible for the genetic control of the trait, other factors affecting the major genotype manifestation were described in terms of the joint residual phenotypic distribution among members of the pedigree. The residual

variation was assumed to be influenced by many factors that are not major genotypes, and that are not specified in detail. The following explicit formal assumptions were made: 1) We allowed each major genotype to manifest a (normal) trait distribution with a specific expectation, μ_g, and variance, σ^2_g. 2) We described each genotype by a genotypic mean value, μ_g, that depended on sex and age - the genotype-sex specific slope coefficient, a_{gs}, measuring the rate of change in the trait per year - and by a genotype-sex specific threshold, T_{gs}, assuming that the genotypic value linearly increases (or decreases) with age after age exceeds this threshold. 3) We introduced assortative mating described by the parameter η. 4) We modeled the common household specific conditions causing co-variation of the trait residuals in pairs of relatives by the partial correlations between residuals in spouses, ρ, in parent and offspring, β, and in siblings, ε.

This does not make the assumption that no genes other than the putative major gene take part in the control of this trait. We assumed only that the effects of other genes are such that they can be adequately described by an *n*-variable normal distribution of residuals. Such a formulation of the model limits our ability to describe the trait inheritance to a genetic model that has a reasonable number of genetic parameters. It would in principle be possible, for example, to construct a MG model with, say, three alleles, or a two-gene two-allele model. The genetic trait variation would then be controlled by 6 monogenic genotypes in the first case, or by 10 digenic genotypes in the second case (we do not assume that cis- and trans- double heterozygous genotypes have the same phenotypic distributions). Accordingly, keeping the model as general as possible, the number of genotype-dependent model parameters (genotypic values, slope coefficients, etc) would have to be substantially increased.

Each model of trait inheritance is tested on a particular pedigree sample of given size and structure. Clearly, the information this sample implicitly contains about the trait inheritance (expressed through the joint phenotypic distribution of its members) is limited. At the same time, a necessary part of pedigree analysis is estimation of the model parameters, which is difficult to achieve with any accuracy if the number of parameters is too large. If this were the case all the other statistical operations of the analysis would also be performed with insufficient accuracy, affecting the result of the analysis at the final stage.

Because we have a most general formulation for the genetic model, it seems justifiable to introduce some constraints on the model parameters that would result in a simpler model, with the number of parameters to be estimated being in line with the information available in the sample being analyzed. However, in general, this model simplification is not a procedure that can be unambiguously defined. An approach to this problem that can be statistically justified is to use the likelihood ratio to test the null hypothesis that the simpler model is not significantly worse in describing the trait inheritance (see more details in section 8.3). However, such a test is expected to have low power, if only because of the large number of parameters that need to be estimated under the general model. At the same time, other non-statistical ways of simplifying the model can hardly be justified in practice. The assumptions implied by reducing the number of model parameters are most often made, without any serious justification, in order to simplify the model and, therefore, the calculation of the pedigree likelihood.

Thus, formulation of the genetic model, and in particular its parameterization, should be coordinated with the pedigree sample on

which the model is tested. This is why, for reasonably informative pedigree samples, for the most part simple versions of (for example, MG) genetic models are usually considered for testing. However, some theoretical investigations point to the need to include more complicated models in pedigree analysis in order to increase the power of linkage tests (see, for example, Risch, 1990a, b, c; Dizier et al., 1993; Schork et al. 1993; Dizier et al., 1996), even at the cost of requiring a substantially larger sample size.

5. THE VARIETY OF TRAITS

The wide diversity of phenotypes that can characterize the biological function under study (collected with questionnaires) necessarily leads to the formulation of a corresponding variety of genetic models for the inheritance of the traits.

5.1 Binary trait

A binary trait is defined when each individual can be more or less accurately classified into one of only two phenotypic classes, expressed, for example, as affected or unaffected. In section 1.8, we stressed the independence of two different classifying factors, namely, the method used to characterize the phenotypes of individuals and the mode of inheritance of the trait. The genetic model for a binary trait would appear to be very simple. But the way an individual is characterized phenotypically is in no way determined by the true genetic, ontogenetic and environmental effects that lead to the binary phenotype. It is determined only by the available instrumentation - the questionnaire used. A binary trait can be controlled by a single gene more or less separate in its action from that of other individual genes (a MG model), by several genes each with relatively large effect on the phenotype (an oligogenic model), by polygenes (a term that needs precise definition), by epigenes, and/or by non-genetic environmental factors. Thus, formulation of a genetic model for the inheritance of a binary trait is not a simpler, but in fact a more difficult, task than for a quantitative trait.

By definition, the first two distributions that define the genetic model (1.1), namely $p(g_1,g_2)$ and $P(g|g_1,g_2)$, are formulated with no direct connection to the type of trait being studied, so the formulation of these

distributions made in sections 4.1 is applicable for binary traits also. What is specific to binary traits starts when we introduce a correspondence between the genotypes of the pedigree members and their phenotypic distributions.

The simplest model for binary trait inheritance assumes that each individual phenotypically manifests its genotype independently of the phenotypes of his/her relatives, though not necessarily uniquely. If the model does not include clearly formulated non-genetic factors that affect these distributions, the genotype–phenotype correspondence is expressed by the *genotype penetrance*: the probability w that the given genotype will have the affected phenotype. Accordingly, $1 - w$ is the probability that this genotype will have the unaffected phenotype. In this case, the joint distribution of phenotypes of the pedigree members, conditional on their subset of genotypes, is represented by the product: $f(X_n|G_n) = \prod_{i=1}^{n} f(x_i \mid g_i)$, where $f(x_i|g_i) = w(g_i)$ if the phenotype of the i-th pedigree member, x_i, is specified as affected, and $f(x_i|g_i) = 1 - w(g_i)$ otherwise.

In most cases, the genotype–phenotype correspondence for a binary trait would not be adequately formulated in this simple form. The inheritance of a trait observed under a dichotomous classification of individuals can be very complicated and, therefore, sometimes needs a very complicated, even sophisticated, formulation of the model.

In the last few decades, these complicated genetic models have often been constructed using an underlying quantitative trait (the so-called liability) model. It is assumed that during ontogenesis, under the control of the genotype, the specific environment and the specific influence of the closest relatives, each individual forms an underlying quantitative trait, or liability. This genotype expression cannot be accurately measured for

reasons that are unrelated to the nature of the trait. Instead, a *threshold* is added as part of the model: if the underlying quantitative liability of an individual, formed as a result of multifactorial effects, exceeds a certain given threshold t, this individual is specified as affected; and otherwise he/she is considered to be unaffected (Mendel and Elston, 1974). Ginsburg and Axenovich (1986) considered the case where the threshold t is not given (i.e. not previously found, for example in some preliminary statistical analysis of pedigree or population data) but is rather included as an additional parameter of the genetic model, t ($\in\theta$).

This approach to constructing complicated models for binary trait inheritance has some technical aspects in its practical application that are still not solved. Let X_n be a set of phenotypes on the n members of a pedigree, and $f(X_n|\theta)$ be the pedigree likelihood generated by model θ. Assume that the pedigree members have formed (according to the formulated model for quantitative trait inheritance) their underlying quantitative traits Y_n. In this case, the pedigree likelihood for the binary trait studied would be expressed as:

$$f(X_n \mid \theta,\varepsilon) =$$

$$= \sum_{G_n} P(G_n \mid \theta,\varepsilon) \int_{y_1=c_1}^{d_1} \ldots \int_{y_n=c_n}^{d_n} f(y_1,\ldots,y_n \mid g_1,\ldots,g_n,\theta).dy_1\ldots dy_n\,, \quad (5.1)$$

where the integral limits are defined as follows: $c_i = t$ and $d_i = \infty$ for the case when the i-th pedigree member is classified as affected, and $c_i = -\infty$ and $d_i = t$ when the i-th pedigree member is classified as unaffected.

The underlying model can be formulated in such a way that the genotype–phenotype correspondence, $f(Y_n|G_n,\theta)$, is given as an n–variable normal conditional on the given subset of genotypes in the pedigree

members. One of the methods of approximating the multivariate normal integral (5.1) was considered by Rice et al, (1979) in an application to multifactorial quantitative traits. However, the construction of genetic models for the multifactorial control of a binary trait in this manner still needs further development. The same is true for any qualitative trait where the individuals in the population under study are divided into more than two types. In the simple monogenic model, more than two penetrances need to be defined for each genotype, while for the multifactorial model we need more than one threshold. Karunaratne and Elston (1998) proposed a multivariate logistic model to allow for correlations in a binary trait across family members.

5.2. Complex traits

Quite expectedly, difficulties arise in formulating the geneticmodel when individuals are identified in a complex manner, i.e., each is characterized by an array of differently observed characteristics (according to the instrumentation employed), such as binary, qualitative or quantitative, discrete or continuous. Reasonable genetic modeling of these complex traits is hardly possible in most cases. This is why two approaches are used. The first is to analyze each of the component traits separately, find the “best” model describing its inheritance, and then try to combine these models into a single *compound* model describing the inheritance of the complex trait. The main problem in doing this is how to take account of the fact that the putative major genes found in these separate analyses of the component traits may be not only linked with one another, but some of them could also be the same genes that pleiotropically control a subgroup of these component traits. Mostly, when such complex traits describe multifactorial diseases, this approach seems

to be the only one that can be used for segregation analysis, even though it appears to be very difficult, tiresome and cannot guarantee results that can be used to solve practical problems - in particular to make predictions. The alternative approach for the genetic study of such complex traits is to use linkage analysis, in an attempt to find the chromosomal locations of the genes that take part in the control of each component trait, identify these genes, and only then construct a compound genetic model for the complex trait.

In practice, however, it is possible to use an approach that we shall call “preliminary reduction” to study the genetics of such complex traits. Livshits et al. (1998) studied the genetic control of human adiposity, in particular the inheritance of a set of 22 adiposity measures on each pedigree member. The pedigree samples were collected from the 3 ethnically and geographically different populations of Kirghizstan, Turkmenia and Chuvasha. Among the set of adiposity traits that were directly measured were skinfold thicknesses and circumferences, and also additionally some constructed indices. Because they are characteristics of the same biological subsystem, these traits are expected to be under the pleiotropic control of genes somehow responsible for the development of adiposity, together with the additional involvement of environmental influences. This is why, to simplify the analysis, a phenomenological analysis of these traits was initially performed. Using the matrix of genetic correlations (obtained under a polygenic model as described by Hazel and Lush, 1942) the 22 traits were subjected to principal component analysis. In each sample that was studied, 4 principal components were retained that showed negligible genetic correlations with one another. This transformation to principal components was interpreted as forming new

quantitative traits of adiposity that are controlled by non-overlapping groups of genes. This reduction further permitted a separate analysis of each of the 4 new traits and, for each of them, the acceptance of a most parsimonious MG model (see details in section 8.3). Combining these results, a four-gene genetic model of adiposity was proposed (of course, in the sense that such a complicated construct as adiposity was defined in the study).

To justify this reduction approach, it seems reasonable to note the similarity of the results obtained in the three ethnically and geographically different populations of Kirghizstan, Turkmenia and Chuvasha. In each of the 3 samples, 4 genetic factors (principal components) were found that had an almost identical biological interpretation, namely, the amount of subcutaneous fat, the total body obesity, the pattern of distribution of subcutaneous fat and the distribution of central adiposity.

Let us note that this example of constructing a compound oligogenic model of trait inheritance uses the so-called additive-polygene theory of quantitative trait inheritance that was so fashionable in the early stages of genetic analysis. The term “genetic correlation” between traits means the correlation between the first trait in a parent and the second in an offspring, and vice versa. Interpretation of this in genetic terms is valid only under strict conditions (Ginsburg and Nikoro, 1973a, 1973b, 1982) that are hardly justified for quantitative traits at the current level of our genetic knowledge. Thus, strictly speaking, the compound model was constructed using a preliminary transformation of the traits, phenomenological in nature, that has very little in common with the joint genetic control of the traits. This does not mean, of course, that other, more genetically based methods, cannot be constructed.

5.3. Bivariate models

A particular case of a complex trait is a bivariate trait, in which each observed pedigree member is described by two traits each of which can be either qualitative or quantitative.

In defining the genetic model, let the set of possible trait phenotypes **X** be represented by the two sets **Y** and **Z**, for each of the two traits, respectively. Next, to make the genetic model as general as possible with no limitation on the number of genetic parameters, assume that there exist two major genes each controlling its own specific trait. Thus, the distributions $p(g_1,g_2)$ and $P(g|g_1,g_2)$ are to be parameterized by 1) two population frequencies p and q of alleles A_1 and B_1 of the two genes and by the disequilibrium parameter D determining the joint population distribution of their alleles, and 2) the recombination fraction ρ between these major loci.

We can formulate the genotype-phenotype correspondence as follows. Let each trait be described by a MG model in which $f(Y_n|G_{1n})$ and $f(Z_n|G_{2n})$ are given in one of the forms considered above, where G_{1n} and G_{2n} are the sets of major genotypes on the n pedigree members for the first and second trait, respectively. Then, $f(Y_n,Z_n|G_n)$ defines the distribution of the $2n$ phenotypes observed on the n pedigree members given their particular set of two-locus genotypes G_n and can be expressed, for example, by a $2n$-variable normally distributed with $2n\times2n$ matrix of partial correlations having the form:

R_Y	W
W	R_Z

where R_Y and R_Z are the matrices of partial correlations between relatives defined for each of the traits and W is a diagonal $n\times n$ matrix with the same

diagonal elements ω - the partial correlation between residuals of the two traits in each pedigree member. The variance-covariance matrix for this 2*n*-variable normal distribution can be easily expressed using the formulas given in the previous chapter.

5.4 Longitudinal model

It is not uncommon to study the effect of sex and age on the ontogenetic development of human characteristics by performing a longitudinal study, i.e. to use sequentially repeated trait observations made on the same individual, in order to learn about the specific dynamics of, for example, a certain disease; or, pooling the individuals into groups (by sex, by living condition, etc.) in order to learn about the effects of these group on these dynamics. Usually, such longitudinal studies have been performed statistically by including the genetic effects on phenotype development only indirectly, for example, when the groups that are compared are sampled from ethnically different populations.

Let us consider how a genetic model of inheritance can be formulated and estimated on a pedigree sample, explicitly accounting for the specific genotype-sex effect on the development of the trait with age. Let each pedigree member be observed not once, but several times (not necessarily at the same times for the different pedigree members), and in this way be described by a subset of phenotypes $\{x_i(t_i)\}$ where t_i denotes the particular age at which the trait is observed on the i-th pedigree member. Let us denote by $X(T)_n$ the set of these subsets for a pedigree with n members. Because only one trait is under study, the most parsiomonious genetic model is the MG model, described for a quantitative trait in the

previous chapter. Thus, the only problem is to formulate properly the genotype–phenotype correspondence, $f[X(T)_n|G_n]$.

Before formulating this, consider a special situation in which the person collecting the pedigree data observes the pedigree sample for a second time after a certain interval of time. In this case, each pedigree member is described by two phenotypes observed with the same time interval between them. Considering these two phenotypes as observations of two different traits, the bivariate MG model described in the previous section should be quite applicable in this situation also. The partial correlation between residuals of the two traits in each pedigree member, ω, represents here the partial correlation between residuals of two sequentially observed phenotypes of the same trait.

In the general case, this partial correlation is expected to depend on the time interval between the successive observations: we express it as $\omega(|t_2 - t_1|)$, where t_1 and t_2 are at the ages at the time of the two observations. In the same way, we should represent the partial correlations between spouses by $\rho(|t_2 - t_1|)$, between parent and offspring by $\beta(|t_2 - t_1|)$, and between siblings by $\varepsilon(|t_2 - t_1|)$. It is reasonable to assume that all of these correlations are decreasing functions of the age interval and, to introduce a minimal number of new parameters, they can each be formulated, for example, as a linear function, $a_v(1–b_v|t_2 - t_1|)$, or an exponential function, $a_v exp(-b_v|t_2 - t_1|)$. Here, $a_v = 1$, ρ, β or ε (defined for the observations on relatives obtained at the same age), and the b_v are coefficients determining how the partial correlations decrease with increasing time interval between the pair of trait observations; $v = 1, 2, 3$ and 4 for the functions $\omega(|t_2 - t_1|)$, $\rho(|t_2 - t_1|)$, $\beta(|t_2 - t_1|)$, and $\varepsilon(|t_2 - t_1|)$, respectively. In this particular (phenomenological) formulation of the

residual distribution conditional on the given set of genotypes in the pedigree members, $f[X(T)_n|G_n]$, only four parameters b_v, $v = 1,\ldots,4$, have been added, to be estimated together with all the other model parameters as defined, for example, in the previous chapter for a quantitative trait.

If k_i denotes the number of successive observations on the i-th pedigree member, the n-member pedigree is described by $K = \sum_{i=1}^{n} k_i$ phenotypes. For a quantitative trait, the joint distribution of their residuals could be approximated by a K-variable normal with a $K \times K$ matrix of partial correlations constructed using the above functions defined on the age intervals.

Table 5.1. Matrix of partial correlations between the 7 phenotypic longitudinal observations made on a 4-member family.

	1	2	3	4	5	6	7
1	1.0	ω(5)	ω(10)	ρ(5)	β(15)	β(0)	β(15)
2		1.0	ω(5)	ρ(0)	β(20)	β(5)	β(20)
3			1.0	ρ(5)	β(25)	β(10)	β(25)
4				1.0	β(20)	β(5)	β(20)
5					1.0	ω(15)	ε(0)
6						1.0	ε(15)
7							1.0

Notes: The age difference between the observations is given in parentheses. See explanation is in the text.

To illustrate, consider as an example a four-member pedigree (two parents numbered 1 and 2 and two offspring numbered 3 and 4) and assume that the following longitudinal trait observations were made:

$x_1(30)$, $x_1(35)$, $x_1(40)$, $x_2(35)$, $x_3(15)$, $x_3(30)$, $x_4(15)$; a total of 7 phenotypes ordered as they are written here, for the 4-member family. As can be seen, parent 1 was observed 3 times, at ages 30, 35 and 40, parent 2 and sibling 4 were observed only once, while offspring 3 was observed twice. Table 5.1 shows the partial correlations between all these observations. The variance-covariance matrix of the 7-variable normal distribution can be obtained from the matrix of partial correlations given in Table 5.1.

5.5. Other formulations

The way of constructing genetic models considered above is specifically intended to describe in a formal fashion differently defined traits, but is not the only one possible, of course. Different formulations and parameterizations of genetic models were proposed by Bonney together with his co-authors and followers (Bonney, 1984, 1988; Bonney et al., 1988; Demenais, 1991; Demenais et al., 1992). Initially, Bonney's formulation was intended to produce models for binary traits with non-genetic correlations between the phenotypes of relatives. Instead of the three independent component distributions clearly defined and separately formulated, as given in the genetic model definition (1.1), the regressive models for binary traits tried to introduce a set of formally equivalent parameters that determine the mode of inheritance. Each parameter, the allelic frequency, genotypic value, correlation between residuals, transmission probability and so on, was included in the model as a logistic factor. However, when compared with the traditional model, one of the versions of which was described above in chapter 4, it was discovered that, although intended initially to have the same biological interpretation, the parameters determining regressive models differ in meaning from the traditional parameters. Then, Demenais (1991) made an attempt to

reconcile these two ways of model formulation and showed that this is possible if the three component distributions of the genetic model (1.1) are defined and formulated separately to take into account particular details of the population genotype distribution, the specific mode of transmission and the specific genotype–phenotype correspondences. The multivariate logistic model developed by Karunaratne and Elston (1998) is another formulation, incorporated into the program package S.A.G.E (2004).

5.6. Control of heterogeneity

Let us define *genetic heterogeneity* of the trait inheritance as the case where not all the phenotypes from the set **X** have the same genetic control (Whittemore and Halpern (2001). Although observed in the same way (measured by the same instrumentation) these phenotypes can be the result of different genes. These genes could be involved in the trait control because of a special environmental condition, because of other minor genes involved in the trait control, etc. As a result, different individuals who manifest the same phenotypically described trait would be doing so as a result of the effects of different genes, i.e., their traits would be differently inherited, and our formulation of the model of inheritance should explicitly account for this fact.

The easiest way to do it is to sort the pedigree data phenotypically prior to their analysis, i.e., to form homogenous pedigree sub-samples, assuming that the true genetic control of the trait is the same within each sub-sample of pedigree members being analyzed. At the same time, different sub-samples are assumed to have different genetic control of the trait being studied. In this way, performing an analysis of the pedigree data in all sub-samples separately, and then combining the results, there is hope

that the heterogeneity of trait inheritance would be adequately described by not one, but several, genetic models.

Unfortunately, this simple way of studying the heterogeneity of genetic control is rarely possible in practice, especially at the early stages of studying the trait inheritance. At further developed stages, after some information about the heterogeneity has been obtained, it is sometimes possible to distinguish pedigrees in which the trait is controlled by two different large effect genes, pedigrees where the trait is controlled by polygenes, or by phenocopies, or pedigrees where no genes are involved in the control – there is pure environmental control of the trait variation. Clearly, this phenotypic distinction between the sub-samples implicitly assumes that there are *different traits,* differently inherited but phenotypically defined in the same manner, when using the same typological or measuring technique. It is reasonable to expect that these traits would be further distinguished by more adequate phenotypic characterization. In his review, Rao (1998) among others has discussed this way of dissecting multifactorial traits.

This real heterogeneity of the genetic control of a trait should not be confused with complicated, but not heterogeneous, genetic control. The simplest example of this can be formulated as follows. Let the inheritance model of the trait include a large effect gene and this gene mutation (allele) causes extremely large trait values in individuals who have it, while low and moderate trait values are shown by individual genotypes not having this particular allele. In this case, members of the same pedigree can simultaneously manifest large trait values in those having genotypes with this specific allele and low or moderate values in those whose genotypes do not have it. It is evident that the unambiguous construction

of sub-samples to be separately analyzed is then hardly possible, even if a number of special stipulations are made. We need to construct a rather complicated model for this trait inheritance, explicitly formulating the distribution $f(X_n|G_n)$ of different trait values that can be manifested by different individual genotypes.

5.7. On the genotype-phenotype formulation

In the last two chapters, examples of genetic model formulation were considered for differently defined traits, i.e., for the different phenotypic descriptions of the biological function being studied.

Once the phenotypic characteristics of this function have been established, the core of each genetic model is the specifically constructed set of genotypes, **G**, that determine the trait genetic control through their particular population distribution, their transmission from parents to offspring, and their phenotypic distributions.

The formulation of the first two component distributions of (1.1), i.e., the distribution of genotypes in spouses and the transmission of genotypes from generation to generation, could be made quite straightforwardly as soon as the set **G** is determined. This formulation is the same for any phenotypic description of the trait(s) being studied. However, a natural ambiguity in the model formulation arises as soon as the third component of model (1.1) is introduced, namely the joint distribution of phenotypes on pedigree members given their genotype combination, $f(X_n|G_n)$. In the last two chapters, examples of how to formulate this distribution were given for variously defined traits, quantitative, continuous and discrete, qualitative (binary), and complex. In each case, the mathematical model includes a set of parameters each

having a clear genetic interpretation. The formulated genetic model allows us to write down the joint probability of phenotypes on the members of each given pedigree.

Clearly, other formulations of the distribution $f(X_n|G_n)$ are quite possible. In some cases, the justification of the particular chosen form of $f(X_n|G_n)$ could be made traditionally, as, for example, the n-variable normal distribution for a quantitative continuous trait, representing the result of the central limit theorem of probability theory. In other cases, this choice could be justified by results of some preliminary investigations of the trait under study. However, in each case the third component distribution in the genetic model formulation should be considered as an approximation to the form of the genotype-phenotype correspondence, the level of approximation being unknown. Taking into account the evident ambiguity in this formulation, it seems reasonable to consider a set of genetic models having the same first two component distributions, $p(g_1g_2)$ and $P(g|\ g_1g_2)$, and differently formulated $f(X_n|G_n)$ covering a possible range of genetic models.

6. THE CORRECTED PEDIGREE LIKELIHOOD

In statistics, the likelihood technique is the one that is most developed asymptotically. Accordingly, this technique is the one most widely used to solve statistical problems when complex empirical data are analyzed. In contrast to the pure mathematical accuracy of the asymptotic results that have been obtained, practical rules indicating how to use this technique for solving complex statistical problems on *finite-size* samples are still far from satisfactory, and need substantiation and further investigation in general, as well as in each particular case.

The basic concept of this technique is the *likelihood* for the data to be analyzed. In the general case, the likelihood is the probability of the data defined on a particular sample space, whether using some explicitly formulated model or not. In pedigree analysis, this is the pedigree likelihood constructed under a previously formulated genetic model. Because it is the probability (mass or density) of the pedigree data to be analyzed, the likelihood must be defined to be positive and normalized on the sample space determined by the sampling design that was initially introduced. The problem of defining the sample space was considered in chapter 2. Here, we consider the methods of adequately, but not necessarily uniquely (for the given sample space), defining the pedigree likelihood.

6.1. Likelihood calculability

Sawyer (1990) considered pedigrees as discrete sampling units, neglecting their inner substructures. He showed that estimators of the genetic model obtained from a pedigree sample are asymptotically unbiased if the model corresponds exactly to the true mode of inheritance

of the trait being studied, including the procedures used in the process of sample collection. This correspondence can be described in more detail by the following conditions that are necessary and sufficient to make the pedigree likelihood (3.1) adequate and, in particular, calculable:

C_1 Each model of inheritance is formulated explicitly and the formulation corresponds to the true mode of inheritance. The only possible unknowns are the model parameters θ_i.

C_2 The two pedigree substructures C_1 and C_2 are identified. The model of pedigree extension ε is known and provides an explicit and correct expression for $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$; only some parameters determining the extension model are possibly unknown.

C_3 The pedigree PSF, i.e., the subset of potential probands in the true pedigree from which the sampling was performed, is identified. The ascertainment model is given in a form that provides an explicit and correct expression for the ascertainment probability $P(asc \mid X_1, C_1, \tau_p, \alpha)$. The only possible unknowns are the ascertainment parameters α.

C_4 The population distribution of the PSFs, τ_p, is known.

C_5 The pedigree subset $X^a C^a$ that determines the inclusion of the pedigree in the sample analyzed is identified. The model for this inclusion, ψ, is known in enough detail to formulate the inclusion probability $P(incl \mid X^a, C^a, \psi)$, but parameter(s) of this expression might be unknown.

As we can see, the first condition means that the genetic model should correspond exactly to the true mode of inheritance of the trait being studied, including the genes controlling the trait, their phenotypic effects

manifested in different environmental conditions, etc. Only the parameters of the formulated model are to be estimated from the pedigree sample. This estimation of the genetic model is the main goal of the analysis. The other conditions, C_2 - C_5, define the sampling procedures used to collect the pedigree sample. These sampling models are a nuisance; they are estimated, together with the genetic model of inheritance, and are used only to account for the particular sampling process. If the formulated conditions are true, the estimator of the model being tested is asymptotically unbiased.

At the same time, if the genetic model is formulated differently from the true mode of trait inheritance, or if the sampling procedures (ascertainment, extension and/or inclusion) do not accurately reflect the real sampling process, the analysis result is doomed to be asymptotically biased, which means there will be asymptotically biased estimation of the genetic description of the trait inheritance. The important point is that the bias, its magnitude and direction, depends on the particular models (the genetic model and the sampling model) used in the analysis. Thus, incorrect ranking of the inheritance models in θ is to be expected, and we will not be able to interpret unambiguously the results of the analysis.

To the extent that strict fulfillment of these conditions does not occur in practice, likelihood (3.1) will be of limited use. To understand the main difficulties in fulfilling these conditions, we now consider some details of modeling the sampling procedures.

6.2. Pedigree extension

The model of pedigree extension determines how the pedigree complement, $(X_2,C_2) = (X,C) \setminus (X_1,C_1)$, is to be sampled, given the initially

ascertained PC. Collecting for observation any available relatives from the true pedigree, regardless of their characteristics or other factors, represents the special degenerate extension model, ε_0. The only version of a directed pedigree extension that is usually considered is *sequential intrafamilial sampling* (Cannings and Thompson, 1977; Thompson and Cannings, 1979). This assumes that, at each stage of data collection, the decision about what additional part of the true pedigree should be observed is made using the data already collected up to the current stage and some limitecd information about the true pedigree, e.g., the existence and availability for observation of the closest relatives of already observed individuals (The pointer extension considered in section 2.4 represents a particular case of this sequential procedure; the very existence of these still unobserved relatives cannot be established without the questionnaire that was introduced above).

Let us consider some general principles of how the extension process can be formulated. Denote the part of the pedigree collected at all n previous stages $[X_{(n)},C_{(n)}]$, and the part of the pedigree collected at the n-th extension stage $[X^{(n)},C^{(n)}]$. Here, $[X_{(1)},C_{(1)}] \equiv (X_1,C_1)$ and $[X_{(n+1)},C_{(n+1)}] = [X_{(n)},C_{(n)}] \cup [X^{(n+1)},C^{(n+1)}]$. Let us assume that, for each particular true pedigree τ, at the n-th extension stage the extension model ε allows only one possible decision about extending the pedigree data already collected. In other words, for the given true pedigree, there is only one unique part of the pedigree on which further data are to be collected, $[X^{(n+1)},C^{(n+1)}]$ (including as a special case no part). In principle, it is possible to define a *randomized* extension model where, at each extension stage, we permit several extensions of the already collected sub-pedigree, and we make the choice among them using a random number generator. Here, we only

consider *deterministic* procedures ε. We represent the extended sub-pedigree as the union of the extensions made during all the extension stages: $(X_2,C_2) = (X,C)\backslash(X_1,C_1) = \bigcup_{n=2}[X^{(n)},C^{(n)}]$.

Now consider a subset of true pedigrees, all having the same sub-pedigree $[X_{(n)},C_{(n)}]$, that has been already collected under the extension model ε, i.e., the same PC and the same extension $[X_{(n)},C_{(n)}] \setminus (X_1,C_1)$. Depending on its particular information content, the *k*-th of these true pedigrees proposes its next specific extension $[X_k^{(n+1)},C_k^{(n+1)}]$. On the universe of all such extensions, the conditional probability $P(X^{(n+1)},C^{(n+1)} \mid X_{(n)},C_{(n)},\theta,\varepsilon)$ is defined as the proportion of the particular extensions $[X^{(n+1)},C^{(n+1)}]$ that are acceptable under model ε. By definition,

$$\sum_{[X^{(n+1)},C^{(n+1)}]} P(X^{(n+1)},C^{(n+1)} \mid X_{(n)},C_{(n)},\theta,\varepsilon) = 1,$$

where the sum goes over all extensions of $[X_{(n)},C_{(n)}]$ in the population possible under the given model ε.

Consider next the subset of true pedigrees from which it is possible to collect the extension $(X_2,C_2) = (X,C)\backslash(X_1,C_1)$. It seems reasonable to assume that the same extension can be collected, in different ways and at different stages, from different true pedigrees. For example, structure 5 in Fig. 2 can be collected in different ways. Let the PC contain two probands, $C_1 = \{4,12\}$. Then the pedigree structure having $C_2 = \{1,2,3,5,8,9,10,11,13\}$ can be sampled through different extension stages when either pointer 2 or pointer 10, or both, belong to the extension type x^+.

Let w denotes a particular sequence by which the given pedigree extension $(X_2,C_2) = (X,C)\backslash(X_1,C_1)$ can be collected. In this case,

$$P(X_2,C_2 \mid X_1,C_1,\theta,\varepsilon,w) =$$

$$= \prod_{n_w=1} P(X^{(n_w+1)},C^{(n_w+1)} \mid X_{(n_w)},C_{(n_w)},\theta,\varepsilon)$$

is the probability of collecting extension X_2C_2, given the extension model ε and the particular extension sequence w. Accordingly,

$$P(X_2,C_2 \mid X_1,C_1,\theta,\varepsilon) = \sum_w P(w \mid X_2,C_2) P(X_2,C_2 \mid X_1,C_1,\theta,\varepsilon,w),$$

where $P(w \mid X_2,C_2)$ is the population relative frequency of that particular sequence w by which the extension (X_2,C_2) can be collected. This probability is determined by the population distribution of true pedigrees. Evidently, $\sum_{(X_2,C_2)} P(X_2,C_2 \mid X_1,C_1,\theta,\varepsilon) = 1$, where the sum goes over all extensions of the initially ascertained (X_1,C_1) that can be made in the given population $\{\tau\}$, including all the ways in which the extension is made under the given extension model ε.

Let us introduce two kinds of pedigree extension: *trait-independent* and *trait-dependent.* In the former, at each extension stage the decision about how to extend the sub-pedigree $[X_{(n)},C_{(n)}]$ collected so far is made independently of the phenotypic content $X_{(n)}$ of this sub-pedigree, as well as of the phenotypes of all other members of the true pedigree. In this case, the structure of the pedigree extension C_2 does not depend on the trait model and the following factorization is correct:

$$P(X^{(n+1)},C^{(n+1)} \mid X_{(n)},C_{(n)},\theta,\varepsilon) =$$

$$P(X^{(n+1)} \mid X_{(n)},C_{(n+1)},\theta)\ P(C^{(n+1)} \mid C_{(n)},\varepsilon), \quad (6.1)$$

where the first factor on the right hand side is the joint probability of phenotypes in the pedigree extension given the pedigree structure (both that previously collected and the extension). This probability is fully determined by the trait model. The second factor is the probability of the particular extension structure $C^{(n+1)}$, which, as we have assumed, does not depend on the phenotypes of the pedigree members already collected (nor, therefore, on the trait model), but does depend on other particulars of the extension procedure (relationships with the individuals already collected, availability etc). In this case, $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon) \propto_\theta P(X_2 \mid X_1, C, \theta)$, i.e., the probability of collecting the pedigree complement X_2C_2 is proportional to the joint probability of the phenotypes in the pedigree extension X_2, given the extended structure C_2 and the pedigree PC. This proportionality means that the left and right hand sides differ from one another by a factor that is independent of θ. Otherwise, when the extension model is such that the structure of a newly collected pedigree part depends on the phenotypes of already observed pedigree members, the pedigree extension is trait-dependent and expression (6.1) is incorrect (Hodge and Boehnke, 1984).

It should be noted that mixed extension procedures, including both trait-independent and trait-dependent stages, are possible. In the above example (section 2.4), at the first stage all available first-degree relatives of the ascertained probands are to be collected, while next collected should be the first-degree relatives of only extension-type pointers.

We have made the above probabilistic description of the pedigree extension process in a somewhat general form in order to show that the correct detailed formulation of this process is not a simple task. With an explicitly defined extension model, sometimes the probability

$P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$ can be expressed correctly using only the observed data (X_2,C_2) and (X_1,C_1). In the example considered in section 2.4, all the intermediate steps in the formulation of $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$ were skipped as unneeded: the pedigree is extended if it contains a pointer who is of the extension type. In other cases, more detailed knowledge is needed about the true pedigree from which the sampling is performed, and/or about the distribution of some parts of the true pedigrees. Up to now, we have only considered the simplest versions of the extension model in pedigree analysis, either in its classical form or in sequential pedigree collection (Cannings and Thompson, 1977; Thompson and Cannings, 1979; Lalouel and Morton, 1981). Below, considering the various possibilities for the sampling correction of the pedigree likelihood, we shall assume that the pedigree extension model is given in such detail that the probability $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$ can be correctly formulated using the available data (see condition C_2 in section 6.1).

6.3. Models of proband ascertainment

After Weinberg's (1912) publication, a number of quite thorough investigations were performed in this field. The ascertainment-model-based (AMB) pedigree likelihood (mostly for sibships) was formulated by introducing various explicit expressions for $P(asc \mid X_1, C_1, \tau_p, \alpha)$. Here we consider the main possibilities for ascertainment modeling.

Each ascertainment event divides the set of potential probands of the true pedigree, its PSF, into three component parts (Fig. 1): $\tau_p = (X_1,C_1) + (X_p,C_p)\backslash(X_1,C_1) + \tau_p\backslash(X_p,C_p)$, where, as defined above, (X_1,C_1) is the PC identified in the sampled pedigree; $(X_p,C_p)\backslash(X_1,C_1)$ is the subset of sampled potential probands who did not become probands; and $\tau_p\backslash(X_p,C_p)$ is the

subset of unsampled potential probands about which nothing is known except that they did not become part of the PC. The first two components are represented in the observed data, while the third comprises unobserved data, not being part of the sampled pedigree. In the proband ascertainment model, we assume that *all the members* of τ_p take part in the ascertainment procedure. It so happens that, by chance and because of their phenotypes, some of them (those forming the sub-structure C_1) became the PC, while the others did not.

Thus*, the probability of ascertainment,* $P(asc \mid X_1, C_1, \tau_p, \alpha)$, is the joint probability that, given the set of potential probands or PSF*,* (X_1,C_1) *became* the PC, while the other potential probands of the true pedigree, $\tau_p \backslash (X_1,C_1)$, including both those sampled, $(X_p,C_p)\backslash(X_1,C_1)$, and those not sampled, $\tau_p \backslash (X_p,C_p)$, *did not become part of the* PC:

$$P(asc \mid X_1, C_1, \tau_p, \alpha) =$$

$$= \Pr[(X_1,C_1) \cap \overline{(X_p,C_p) \setminus (X_1,C_1)} \cap \overline{\tau_p \setminus (X_p,C_p)} \mid \alpha)], \quad (6.2)$$

where $\overline{A}$ denotes that no potential proband from the subset A becomes part of the PC. This probability is defined for each given τ_p with its specific structure $C(\tau_p)$, and phenotypic content $X(\tau_p)$. By definition:

$$\sum_{C_p \subseteq C(\tau_p)} \sum_{C_1 \subseteq C_p} \Pr[(X_1,C_1) \cap \overline{(X_p,C_p) \setminus (X_1,C_1)} \cap \overline{\tau_p \setminus (X_p,C_p)}] = 1,$$

where the second sum goes over all C_1, including the empty one, while by definition the pedigree is ascertained only if C_1 is not empty. This equality should hold for any given phenotypic content $X(\tau_p)$ of τ_p.

In general, any particular formulation of probability (6.2) should take explicit account of how members of the pedigree PSF jointly give rise to the pedigree PC. Without dwelling upon the possible formulations of the

probability (6.2), we note that, traditionally, it is often formulated in its more or less simplified form, in terms of the so-called π-model of ascertainment, the most general formulation of which was given by Elston and Sobel (1979). Let $\pi_i = \pi(x_i,\beta_i)$ denote the probability that the *i*-th member of the true pedigree becomes a proband, this member having phenotype x_i and some additional characteristics β_i, such as sex, age, place and duration of residence, etc. Clearly, $\pi_i = 0$ for any pedigree member not belonging to the pedigree PSF. Assume that π_i does not depend on the individual's position in the structure τ_p, nor on the phenotypes and proband statuses of other members of τ_p (i.e. there is no "interaction" between members of the given PSF during the ascertainment process – this is the main independence assumption of the π-ascertainment scheme). However, π_i does depend on the individual's phenotype, and this probability may be different for males and females, it may also change with an individual's age if people of a certain age range enter the catchment area more frequently than others, etc. Importantly, given the potential proband's phenotype, the probability that he/she becomes an actual proband does not depend on the trait model θ. Thus, assuming independent ascertainments (single if $|C_1| = 1$, or multiple if $|C_1| > 1$), probability (6.2) can be expressed as:

$$P(asc \mid X_1, C_1, \tau_p, \alpha) =$$

$$= \prod_{i \in C_1} \pi_i \prod_{j \in C(\tau_p) \setminus C_1} (1-\pi_j) = \prod_{i \in C_1} \pi_i \prod_{j \in C_p \setminus C_1} (1-\pi_j) \prod_{k \in C(\tau_p) \setminus C_p} (1-\pi_k), \quad (6.3)$$

where the first two factor products on the right hand side are determined by the sampled pedigree data, while the factor $\prod_{k \in C(\tau_p) \setminus C_p} (1-\pi_k)$ is

determined by the unsampled subset of potential probands in the true pedigree. This last factor does not depend on the trait model, but it does depend on the result of sampling, both the ascertainment and the extension, because it is different for different structures of the sampled part of the PSF C_p; and, of course, it is determined by the true set τ_p (which is usually unknown, but which could be learnt from the questionnaire).

In the classical π-scheme for a binary trait, it is assumed that $\pi_i \equiv \pi$ for each affected member of τ_p and $\pi_i = 0$ otherwise. In this case, definition (6.3) reduces to $\pi^b(1-\pi)^{k-b}$, where b is the number of probands and the other k affected PSF members, if the pedigree PSF is not sampled in its entirety, comprise k_1 who are observed in the process of pedigree sampling and $k_2 = k - k_1$ who are unobserved. Initially, this simplified form of ascertainment probability was proposed for sibship data and it was later applied (correctly or not, depending on the assumed and actually performed ascertainment procedure) to complex pedigrees characterized by a binary trait. Note, in addition, that this binomial expression is not multiplied by a combinatorial coefficient because it is the probability of ascertaining this particular pedigree, not any pedigree having the same b and k. When the probands have not been identified, the approximate ascertainment probability $1-(1-\pi)^k$ was proposed for all pedigrees having at least one proband among k affected PSF members (see a note on this approximation in Ewens and Shute, 1986).

Independence of becoming probands by members of the pedigree PSF seems too idealized a situation and, in practice, more complicated multiplex forms of ascertainment take place. Hodge and Vieland (1996) proposed a proband *configuration* as the predefined subset of relatives

(e.g., two siblings, father and his son etc.) that jointly cause the pedigree ascertainment (although independently of other such configurations in the pedigree). Let $\pi_{ij} = \pi(x_i\beta_i, x_j\beta_j)$ be the probability that the (ij)-th configuration causes the pedigree ascertainment. In this case, the multiplicative form (6.3) of the ascertainment probability is assumed to hold true not for separate members of the PSF, but for the predefined configurations that can be formed by members of the given τ_p (or of the whole true pedigree when $\tau_p \equiv \tau$).

6.4. Pedigree likelihood – sample space

For the given population of true pedigrees $\{\tau\}$, any particular sampling procedure $S = (\alpha,\varepsilon,\psi)$ determines the sample space – the universe of pedigrees that in principle could be sampled and, therefore, analyzed. The pedigree likelihood is defined on this space as the probability (density) of sampling this pedigree under the given model of trait inheritance. It is evident that there is not just one way to define this probability. The core of any likelihood definition is, of course, the joint distribution of phenotypes on the pedigree members, which is fully determined by the analysis model of trait inheritance. However, for this likelihood to adequately reflect not only the genetic model, but also the specific mode of sampling, and in order to keep the probabilistic nature of the likelihood, the latter should contain a formal description of the sampling procedure and should be normalized. Traditionally, this normalization for the sampling procedure is called *likelihood correction* (e.g. for ascertainment). The normalizing denominator in the likelihood definition (3.1) is the probability of the sample space on which the explicit

form of the likelihood correction is defined for the sampling procedure *S* that is used.

6.5. Ascertainment correction

As defined above, the pedigree sampling includes three distinct sub-procedures, the pedigree ascertainment, its extension and censoring. We have assumed that all particulars of the extension procedure are already taken into account when the probability of the extended pedigree part, $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$, is explicitly formulated. If this is not so, then all other considerations become meaningless. Thus, we consider below the methods of likelihood correction for the other two specific sub-procedures of pedigree ascertainment and censoring.

For the moment, only the ascertainment correction will be considered, assuming there is no selective inclusion, in the sample to be analyzed, of the pedigrees that have been already ascertained and extended. In other words, we shall assume that the subset (X^a, C^a) is empty and, therefore, the pedigree likelihood in this case can be expressed as in (3.1) with $P(incl \mid X^a, C^a, \psi) \equiv 1$.

Bearing in mind the above details of the ascertainment procedure, it becomes clear that likelihood (3.1) cannot usually be calculated, because conditions C_3 and C_4 (see section 6.1) are not fulfilled. Indeed, if the pedigree PSF is not sampled in its entirety (C_3), it is impossible to calculate $P(asc \mid X_1, C_1, \tau_p, \alpha)$. Moreover, even if the pedigree PSF is known, the denominator of (3.1) cannot be found without knowing the population distribution of PSFs (C_4). Therefore it is justifiable to talk about the insolvability (intractability is the term used by Vieland and Hodge, 1995, who first noted this) of the ascertainment problem. Additionally, it

should be noted that this statement has no connection with the inclusion procedures. The latter can be very diverse in their formulation, while pedigree ascertainment through probands can be at least accurately formulated in the form of (6.2) or (6.3).

6.6. Conditioning on the PSF structure

Assume that, in the process of pedigree sampling (or even before this process starts), a survey is conducted of the true pedigree population, using the questionnaire defined above, inquiring about pedigree structures. This would not be exceptionally unusual in practice. If certain characteristics by which potential probands are defined (e.g., belonging to the catchment area) are determined by this questionnaire, then the structure $C(\tau_p)\backslash C_p$ of the unobserved members of the pedigree PSF can be learnt for each sampled pedigree; also, if needed, we can determine the structure relating to those pedigree members who are not potential probands, but who nevertheless provide information about the relationships between the observed pedigree members and the unobserved PSF members. This means that the sampled pedigree is now extended and is represented by the structure $C \cup C(\tau_p)$, which includes the previously sampled structure and the substructure $C(\tau_p)\backslash C_p$ identified by the questionnaire. By doing this, the structure of the true pedigree PSF, $C(\tau_p)$, becomes known regardless of what particular PC was ascertained and how it was further extended.

In this case, it seems natural to formulate the pedigree likelihood conditional on the given PSF structure as follows. Consider a dummy pedigree $(X,C) \cup \tau_p$, which contains the sampled (ascertained and extended) pedigree, (X,C), and, additionally, the PSF of the true pedigree from which the sampling was performed but that has not been included in

the sampling process. In this extended pedigree, the structure of the added PSF part, $C(\tau_p)\backslash C_p$, has been identified by using the questionnaire, while its phenotypic content, $X(\tau_p)\backslash X_p$, is unknown. The probability of this dummy pedigree, $P[(X,C)\cup\tau_p\,|\,\theta,\varepsilon]$, can only be found if the population distribution of τ_p is known, which is hardly expected to occur in practice.

Let $P[(X,C)\cup\tau_p\,|\,C(\tau_p),\theta,\varepsilon]$ be the probability of observing this dummy pedigree given its PSF structure and the trait inheritance and extension models. Then the probability that the pedigree is (X,C) and is sampled, given the PSF structure, can be expressed as

$$P[X,C,smpl\,|\,C(\tau_p),\theta,\varepsilon,\alpha]=$$

$$=\sum_{X(\tau_p)\backslash X_p} P[(X,C)\cup\tau_p\,|\,C(\tau_p),\theta,\varepsilon]P(asc\,|\,X_1,C_1,\tau_p,\alpha)=$$

$$=P(X,C\,|\,\theta,\varepsilon)\sum_{X(\tau_p)\backslash X_p} P[X(\tau_p)\backslash X_p\,|\,C(\tau_p)\backslash C_p,X,C,\theta]\times$$

$$\times P(asc\,|\,X_1,C_1,\tau_p,\alpha)\,. \qquad (6.4)$$

In the particular case that the ascertainment probability can be represented in its multiplicative form (6.3), (6.4) can be expressed as

$$P(X,C\,|\,\theta,\varepsilon)P(asc\,|\,X_1,C_1,X_p,C_p,\alpha)R(X,C,\theta,\alpha)\,, \qquad (6.5)$$

where $P(asc\,|\,X_1,C_1,X_p,C_p,\alpha)=\prod_{i\in C_1}\pi_i\prod_{j\in C_p\backslash C_1}(1-\pi_j)$ is the probability that only the substructure C_1 contains probands while the sampled complement $C_p\backslash C_1$ does not, and

$$R(X,C,\theta,\alpha)=\sum_{X(\tau_p)\backslash X_p} P[X(\tau_p)\backslash X_p\,|\,C(\tau_p)\backslash C_p,X,C,\theta]\times$$

$$\times\prod_{k\in C(\tau_p)\backslash C_p}(1-\pi_k)$$

is the probability that the added PSF structure $C_p \backslash C_1$ contains no proband given the sampled pedigree data, X.

If we have an explicit model for the ascertainment probability (6.2) or (6.3), both expressions (6.4) and (6.5) are calculable, because the conditional probability of phenotypes on members of the added part of the PSF, $P[X(\tau_p) \setminus X_p \mid C(\tau_p) \setminus C_p, X, C, \theta]$ can be calculated – it is determined solely by the trait inheritance model. Note that expressions (6.4) and (6.5) are results of the usual way of handling missing pedigree data, i.e., summing over all possible phenotypes for the members of the unobserved part of the PSF under the given model of trait inheritance.

The denominator of the likelihood is the probability of sampling at least one pedigree from the true ones having a PSF of the given structure. It can be expressed as:

$$P[smpl \mid C(\tau_p), \theta, \varepsilon, \alpha] =$$

$$= \sum_{C_1 \subseteq C(\tau_p)} \sum_{[X, C(C_1)] \cup X(\tau_p)} P[(X, C) \cup \tau_p \mid C(\tau_p), \theta, \varepsilon] \times$$

$$\times P(asc \mid X_1, C_1, \tau_p, \alpha) = \quad (6.6)$$

$$= \sum_{C_1 \subseteq C(\tau_p)} \sum_{[X_p, C_p(C_1)] \cup X(\tau_p)} P[(X_p, C_p) \cup X(\tau_p) \mid C(\tau_p), \theta, \varepsilon] \times$$

$$\times P(asc \mid X_1, C_1, \tau_p, \alpha) =$$

$$= \sum_{X(\tau_p)} P[X(\tau_p) \mid C(\tau_p), \theta] \sum_{C_1 \subseteq C(\tau_p)} P(asc \mid X_1, C_1, \tau_p, \alpha) \times$$

$$\times \sum_{C_p(C_1) \subseteq C(\tau_p)} P[C_p \mid C(\tau_p), \theta, \varepsilon], \quad (6.7)$$

where:

1) the first sum in (6.6) goes over those PCs that can be ascertained from the given PSF, and the second sum goes over all those *(X,C)* that can be extended from the given PC, which is denoted by $C(C_1)$, and over possible phenotypes in the added structure $C(\tau_p)\backslash C_p$;

2) when obtaining (6.7), we took into account that by definition $X_p \cup X(\tau_p) \equiv X(\tau_p)$ for any $C_p \subseteq C(\tau_p)$ and, also by definition, $P(asc \mid X_1, C_1, \tau_p, \alpha)$ is not dependent on the particular sampled C_p. Note, now, that $\sum_{C_p(C_1)\subseteq C(\tau_p)} P[C_p \mid C(\tau_p), \theta, \varepsilon] = 1$ for any given extension model ε, which follows from the normalization of the probability $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$ - see section 6.2. In this equality only those C_p are considered that include the particular given C_1.

Thus, (6.7) can be finally expressed as:

$$= \sum_{X(\tau_p)} P[X(\tau_p) \mid C(\tau_p), \theta] \sum_{C_1 \subseteq C(\tau_p)} P(asc \mid X_1, C_1, \tau_p, \alpha) \equiv$$
$$\equiv P[asc \mid C(\tau_p), \theta, \alpha], \qquad (6.8)$$

which means that the denominator is the probability of ascertaining at least one PC from the PSF having the given structure $C(\tau_p)$, regardless of whether or not, and how, this PC is further extended. As can be seen, the likelihood denominator (6.8) is quite calculable because, as expected, it does not depend on the extension model.

Using (6.4) and (6.8) as the numerator and denominator, the pedigree likelihood conditional on the PSF structure can be expressed as:

$$P[X, C \mid smpl, C(\tau_p), \theta, \varepsilon, \alpha] = \frac{P[X, C, smpl \mid C(\tau_p), \theta, \varepsilon, \alpha]}{P[smpl \mid C(\tau_p), \theta, \alpha]} =$$

$$= \frac{P(X,C \mid \theta,\varepsilon) \sum_{X(\tau_p)\backslash X_p} P[X(\tau_p)\backslash X_p \mid C(\tau_p),X,C,\theta] P(asc \mid X_1,C_1,\tau_p,\alpha)}{\sum_{X(\tau_p)} P[X(\tau_p) \mid C(\tau_p),\theta] \sum_{C_1 \in C(\tau_p)} P(asc \mid X_1,C_1,\tau_p,\alpha)}. \quad (6.9)$$

This likelihood explicitly uses a particular model for the pedigree ascertainment and is called an ascertainment-model-based (AMB) likelihood. Expression (6.9) is calculable because the joint probability of phenotypes in members of any given pedigree sub-structure is determined solely by the trait inheritance model and does not depend on particulars of the extension model. The numerator of likelihood (6.9) uses the usual way of handling missing pedigree data, $X(\tau_p)\backslash X_p$, by summing over all phenotypes possible on members of the unobserved part of the PSF under the given model of trait inheritance (condition C_3). Additionally, conditioning on the PSF structure allows us to avoid the need to use the population distribution of pedigree PSFs (condition C_4): the sample space consists only of pedigrees whose PSF has the same structure $C(\tau_p)$. Moreover, the denominator of (6.9) does not depend on the extension model ε.

This likelihood is expected to yield asymptotically the same estimator of the trait inheritance model as likelihood (3.1) because it is obtained from it (to be more correct, from the (3.1) version for the particular case $C^a = \varnothing$) by conditioning, both the numerator and denominator, on the same PSF structure, which is assumed to be independent of the trait inheritance model.

Let us remember again that (6.9) was obtained by using the following two assumptions: 1) the definition of the potential probands is

made independently of the trait under study (e.g., individuals belonging to a certain catchment area), and 2) at each stage of the intrafamilial extension, no distinction is made between those true pedigree members who could become probands but did not do so and those who could not be probands by the very definition of probands.

We have assumed here that the ascertainment procedure that is actually employed is such that the probability $P(asc \mid X_1, C_1, \tau_p, \alpha)$ can depend on the whole pedigree PSF, its structure and its phenotypic content. At the same time, as follows from (6.9), the modeled distribution of the sampled pedigrees, $P^S(X, C \mid \theta)$, uses only the sampled part of the pedigree PSF, (X_p,C_p), and the structure (but not the phenotypes) of its unsampled part, $C(\tau_p)\backslash C_p$. This permits us to construct a calculable likelihood from which a correct estimator of the trait inheritance model can be obtained even if the two conditions C_3 and C_4, which are in practice most unlikely to hold, are not fulfilled.

6.7. Conditioning on the sampled pedigree structure

Vieland and Hodge (1996, p. 1073) stated that "in practice the likelihoods used in both linkage analysis and segregation analysis are always conditioned on the sampled pedigree structure". However, this is not the conditioning discussed here. A likelihood can only be calculated if the sampled pedigree structure is known, but this technical information has nothing to do with the ascertainment conditioning of the likelihood. The latter is intended to solve a quite different problem: to take into account the fact that the pedigrees are not sampled at random, but rather according to some previously established sampling design. We believe Vieland and Hodge have confused the technical problem of calculation with the

likelihood correction that should account for the specially designed sampling procedures.

Vieland and Hodge (1995, 1996) considered the case where all members of the true pedigree are potential probands, in other words, $\tau_p \equiv \tau$ for each τ. In this case, the AMB pedigree likelihood can be correctly expressed as conditional on the true (not the sampled) pedigree structure – the same expression as (6.9) with substitution of τ for τ_p in both the numerator and denominator, and with summation over all phenotypes possible for the pedigree members in the substructure $C(\tau)\backslash C$. This expression is calculable only if, through the use of the questionnaire, we know the true pedigree structure $C(\tau)$. Our ability to know this is directly related to the meaning of the term "true" pedigree. Vieland and Hodge (1995, p. 42) assumed that "we are all (probably) related to one another (or, at any rate, to say so is a better approximation to the truth than to divide us arbitrarily into small family units)". Although we do not disagree with this statement, we stress that in *pedigree analysis* the subdivision of the real population into family units is a procedure that is far from arbitrary. The sampled objects (pedigrees) are determined by the sampling design (the questionnaire), and different sampling designs determine different $\{\tau\}$ populations from the same real population.

Let us note the difference between the true pedigree structure $C(\tau)$ established through the use of a questionnaire and the structure C of the sampled (ascertained and further extended) pedigree. $C(\tau)$ is determined by the population $\{\tau\}$ and by the study design, not by the ascertained PC and the model used to extend the PC. On the other hand, the sampled pedigree structure C is directly determined by the sampling procedure used. From a true pedigree with a given $C(\tau)$, it is possible to sample

different sub-pedigrees whose structures are different both in the initially ascertained parts and in the resulting extension. Vieland and Hodge (1995) stated that ascertainment correction results in a consistent model estimator only if the ascertainment is single or, if not, it is proband independent (PI); this means that, in the context of the tractability of the ascertainment problem, it seems unhelpful to make a distinction between proband independent (PI) and proband dependent (PD) sampling, as proposed by them. The true pedigree structure $C(\tau)$ is PI by definition, while the sampled pedigree structure C can be PI or PD, depending on the ascertainment and extension procedures used. However, this fact does not affect the possibility of constructing a corrected likelihood. In other words, PI or PD *sampling* is not relevant to the possibility of ascertainment correction. In the analogue of (6.9) for the case $\tau_p \equiv \tau$, the conditioning should be performed on the structure $C(\tau)$ and not on the *sampled* pedigree structure, which can depend on the particular PC ascertained.

The same is true if, as has been assumed here, the PSF constitutes only part of the true pedigree ($\tau_p \subset \tau$): likelihood (6.9) is conditioned on the structure $C(\tau_p)$ but not on the sampled structure C_p. Clearly, $C(\tau_p)$ does not depend on either the ascertained PC or on the extension procedure that is employed. Although the *sampled* pedigree structure is determined by the sampling process and can be PI or PD, this fact does not affect the possibility of constructing the ascertainment correction for the pedigree likelihood.

The likelihood conditioned on the sampled pedigree structure can be especially useful when pedigrees are extended in a trait-independent manner, i.e., when the structure of the extended part of the pedigree, C_2, is independent of the trait model (e.g., when the intrafamilial sampling

includes all the members of a true pedigree available for observation). This lets us use the joint probability of the phenotypes of the extended part of the pedigree $P(X_2 \mid X_1, C, \theta, \varepsilon_0)$, which is determined only by the trait model, instead of the more complicated probability $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$. However, if proband status is still determined by individual phenotypes, the likelihood conditioned on the sampled pedigree structure, even for the particular case of trait-independent pedigree extension, can provide an adequate trait model estimator only under the conditions given above.

6.8. Conditioning on the PSF data

Likelihood (6.9) explicitly uses a model of pedigree ascertainment. However, if particulars of the actual sampling procedures are unknown, then this model could be incorrectly formulated, causing in this way an asymptotically biased estimator of the trait model of inheritance (Sawyer, 1990). To avoid the risk of such bias, which in practice cannot be determined, Ewens and Shute (1986) and Shute and Ewens (1988a,b) proposed the ascertainment-assumption-free (AAF) method of likelihood correction for the sampling procedures. Here, we prefer to call it ascertainment-model-free (AMF) because it may contain specific assumptions about the sampling procedures, but is free from the need of an explicit model formulation of the ascertainment procedure.

Assume that the probability of ascertaining a pedigree is fully and completely determined by the PSF data in the true pedigree from which the sampling is performed, regardless of the particular PC causing the ascertainment. This means that, instead of the explicitly formulated function (6.2) or (6.3) determined by a limited set of parameters α, the

ascertainment probability is defined as $P(asc \mid \tau_p)$ and is the same regardless of the particular PC initiating the pedigree sampling. Following Ewens and Shute, consider this probability as a parameter and find its maximum likelihood estimate $\hat{P}(asc \mid \tau_p)$ based on the pedigree sample analyzed. Substituting this expression $\hat{P}(asc \mid \tau_p)$ into the sample likelihood, it is possible to obtain after some transformation, for the extended pedigree $XC \cup \tau_p$ that contains not only the ascertained and extended parts of the pedigree but also all the data "relevant to sampling", the following expression for the AMF likelihood:

$$P[(X,C) \cup \tau_p \mid smpl, \theta, \varepsilon] \propto_{\theta,\varepsilon} \frac{P[(X,C) \cup \tau_p \mid \theta, \varepsilon]}{P(\tau_p)} =$$

$$= \frac{P[(X,C) \cup X(\tau_p) \mid C(\tau_p), \theta, \varepsilon]}{P[X(\tau_p) \mid C(\tau_p), \theta]}, \quad (6.10)$$

i.e., the likelihood is found to be equivalent to (produces the same estimators of the models θ and ε as) the probability of the pedigree data $(X,C) \cup \tau_p$ conditioned on the given PSF, both its structure and its phenotypic content. The right hand side of (6.10) is found by dividing both the numerator and denominator of this expression by the probability of the τ_p structure, $P[C(\tau_p)]$, which makes this expression calculable because the probability $P[X(\tau_p) \setminus X_p \mid C(\tau_p) \setminus C_p, X, \theta]$ is determined by only the trait inheritance model. The same result can be obtained using the technique of Hodge (1988, see also Kalbfleisch and Sprott, 1970, and the note of Sawyer, 1990, p 355).

Note that likelihood (6.10) is defined on a parameter space that has a larger (sometimes, substantially larger) number of dimensions.

Moreover, the ascertainment probability is defined independently of the particular PC that is ascertained from the pedigree PSF. This last assumption is not in full agreement with, for example, the usually accepted binomial π-model of ascertainment (Morton, 1959; Elston and Sobel, 1979). However, ideologically it is similar to the specific form of the π-model described as "at least one proband" and formulated as $1 - (1-\pi)^k$, where k is the number of affected PSF members (see, for example, Weinberg, 1927; Haldane 1938; Bailey, 1951). Shute and Ewens (1988) pointed out that this latter formulation of the ascertainment probability is a "far less efficient approach" (up to dozens of times!) for estimating parameters than the usual binomial form (6.2). This conditioning on the whole τ_p data substantially decreases the pedigree information. In particular, when all the true pedigree members are "relevant to ascertainment", i.e., when $\tau_p \equiv \tau$ and therefore $C_p = C$, the AMF likelihood degenerates to 1, which makes the very process of estimating the trait inheritance model useless.

Taking this into account, (6.10) produces ML estimates of θ different from those obtained by the AMB likelihood - at least their standard errors are expected to be larger. However, if 1) there is insufficient knowledge about the actual ascertainment procedure used for sampling the pedigree, i.e., there is a risk of incorrectly modeling the ascertainment procedure and, therefore, a risk of non-testable bias in the results of the analysis, and if 2) the pedigree PSF, both its structure and phenotypic content, is known in its entirety, then it is possible to use the AMF likelihood (6.10), which provides an asymptotically unbiased estimator of the trait inheritance model. In addition, note that this is an

excellent example of robust likelihood formulation: it provides unbiased results under minimum assumptions about the sampling procedure.

Vieland and Hodge (1996) were the first to note that likelihood (6.10) is calculable and provides correct estimation of the trait inheritance model only if the pedigree PSF is observed in its entirety, both its structure and phenotypic content. If this is not the case, i.e., part of the PSF happens to be unobserved (e.g., members of the substructure, $C(\tau_p)\backslash C_p$, which can be added if we use a questionnaire, or some other members of the sampled PSF C_p), then the conditional probability (6.10) turns out to be undefined: the likelihood correction cannot be constructed even if the PSF structure is accurately established. This need to have all the PSF members observed substantially limits the practical applicability of the AMF likelihood (6.10). Thus, formulating the AMF likelihood makes it possible to avoid condition C_4 (see section 6.1) as in the AMB formulation. However, contrary to what happens in the case of AMB formulation, condition C_3 becomes more critical with regard to the observed data. There is no need to formulate the explicit ascertainment model, but there is need to observe the whole pedigree PSF. If this is not done, the AMF likelihood is undefined; however, sometimes it is possible to salvage this robust AMF likelihood using an approximate formulation.

The above note, that the AMF likelihood cannot be accurately defined whenever at least one member of the pedigree PSF is not observed, does not mean that the idea of the AMF approach cannot be salvaged approximately. This idea, to avoid explicit formulation of the pedigree ascertainment, could be very useful in many cases. Below, the problem of approximate formulation of the pedigree likelihood will be

considered in more detail. Here, we consider only one simple and practically applicable approximate AMF likelihood, of the form:

$$P[X,C \mid smpl, C(\tau_p), \theta, \varepsilon] \approx \frac{P[(X,C) \cup \overline{X}(\tau_p) \mid C(\tau_p), \theta, \varepsilon]}{P[(X_p,C_p) \cup \overline{X}(\tau_p) \mid C(\tau_p), \theta, \varepsilon]},$$

where $\overline{X}(\tau_p)$ denotes the modified set of phenotypes in the PSF members: it contains the trait values on those PSF members who were observed, and the trait expectation for those who were not observed. As soon as the PSF structure is known, this likelihood can be easily calculated by replacing each missing phenotype on the PSF members by the sample mean.

6.9. Special case of convergent sampling

Consider a method of likelihood correction for a special way of sampling that has been noted several times (Cannings and Thompson, 1977; Vieland and Hodge, 1995) as an example of it being impossible to construct the correct likelihood, not only in the AMB but also in the AMF form. Suppose that two PCs, (X_{11},C_{11}) and (X_{12},C_{12}) have been independently ascertained from different parts of the same true pedigree (more precisely, from different parts of the same pedigree PSF). Next, suppose both of them have been extended according to some extension model. It is possible for these two extension processes to join up, or converge, to the same part of this true pedigree – the same individual or the same nuclear pedigree - and then the pedigree extension further proceeds from this point on. In general, for this case of convergent sampling, it is unknown how to jointly formulate both of these extension processes in the pedigree likelihood.

Strictly speaking, in terms of proband ascertainment, any pedigree sampling is convergent as soon as the pedigree PC is formed from more than one proband, $|C_1| > 1$. Indeed, in any "registry" scheme, probands are ascertained independently of one another, and the discovery that some of them are relatives belonging to the same true pedigree is usually made only after the extension process begins and then proceeds up to a level that is different for different pedigrees. This "convergence" has no effect if the extension procedure is simple, e.g., any available pedigree member is to be observed. However, in the case of trait-dependent sequential extension, formulation of the AMB likelihood for this convergent sampling appears to be difficult, if not impossible, in many cases; so the AMF likelihood considered above then seems to be the only possibility for likelihood correction.

Let us condition the pedigree likelihood on the whole pedigree sub-structure X_0C_0 collected in all subsequent extension stages up to the converging one: $(X_0,C_0) \supseteq (X_{11},C_{11})$ and $(X_0,C_0) \supseteq (X_{12},C_{12})$. If $(X_0,C_0) \subseteq \tau_p$, i.e., this sampled substructure is a part of the PSF, then the SMF likelihood (6.10) is adequate for this convergent sampling. If (X_0,C_0) represents a more extended substructure, then the pedigree likelihood could be conditioned on the combined sub-pedigree $(X_0,C_0) \cup \tau_p$ instead of on τ_p. This likelihood is calculable provided that all the information necessary to write down the conditional probability $P[(X,C) \cup \tau_p \,|\, (X_0,C_0),\theta,\varepsilon]$ is given, regardless of how the converging (X_0,C_0) has been formed. However, it should be noted again that the problem of missing data in the pedigree subset (X_0,C_0) is exactly the same as was mentioned above. Conditioned in this way the likelihood is

undefined whenver at least one member of the subset C_0 has not been observed.

Summarizing what has been said in the last two sections, the following should be stressed once more. There is no formal limitation for conditioning the pedigree likelihood on any part of the sampled data, e.g., on the sampled pedigree structure or on the part of the pedigree collected in the case of convergent sampling. However, if we need to obtain analysis results free of sampling bias, then the pedigree likelihood defined on the given sample space should properly account for the sampling procedure, regardless of whether or not there is any additional conditioning.

6.10. Likelihood correction of Cannings and Thompson

Let us consider now a very special case that occurs when the ascertainment likelihood correction is exactly as proposed by Cannings and Thompson, (1977; Thompson and Cannings, 1979). They stated that, if all the sampled pedigree data are included in the analyzed sample, then the likelihood $P(X_2,C_2 \mid X_1,C_1,\theta,\varepsilon)$ provides an asymptotically unbiased estimator of the genetic model.

Let us assume that:

1) $(X_1,C_1) = (X_pC_p)$, which means that *no* PSF member is collected in the process of pedigree extension, i.e., the sampled part of the PSF is represented by the initially ascertained subset of probands - in this case, $P(X_2,C_2 \mid X_p,C_p,\theta,\varepsilon) \equiv P(X_2,C_2 \mid X_1,C_1,\theta,\varepsilon)$;

2) given the sampled part of the PSF, (X_1,C_1) (or (X_p,C_p), which is the same in this case), the two trait subsets $X_2 = X\backslash X_1$ and $X(\tau_p)\backslash X_1$ are distributed independently of one another, which means that

$$P[X(\tau_p)\backslash X_1 \mid C(\tau_p)\backslash C_1,X,\theta] \equiv P[X(\tau_p)\backslash X_1 \mid C(\tau_p)\backslash C_1,X_1,\theta].$$

In this case, the pedigree likelihood can be defined as conditional on its PC. Summing (6.6) over possible extensions X_2C_2, the denominator can be expressed as:

$$\sum_{(X_2,C_2)} P[X,C,smpl \mid C(\tau_p),\theta,\varepsilon,\alpha] =$$

$$= P[X_1,C_1,smpl \mid C(\tau_p),\theta,\varepsilon,\alpha] =$$

$$= P(X_1,C_1 \mid \theta,\varepsilon)\times$$

$$\times \sum_{X(\tau_p)\setminus X_1} P[X(\tau_p)\setminus X_1 \mid C(\tau_p)\setminus C_1, X_1, \theta]P(asc \mid X_1,C_1,\tau_p,\alpha) =$$

$$= P(X_1,C_1 \mid \theta,\varepsilon)\times$$

$$\times \sum_{X(\tau_p)\setminus X_1} P[X(\tau_p) \mid C(\tau_p), X_1, \theta]P(asc \mid X_1,C_1,\tau_p,\alpha) . \quad (6.11)$$

Dividing the numerator (6.6) and the denominator (6.11) by

$$\sum_{X(\tau_p)\setminus X_1} P[X(\tau_p)\setminus X_1 \mid C(\tau_p)\setminus C_1, X_1, \theta]P(asc \mid X_1C_1,\tau_p,\alpha) \equiv$$

$$\equiv \sum_{X(\tau_p)\setminus X_1} P[X(\tau_p) \mid C(\tau_p), X_1, \theta]P(asc \mid X_1,C_1,\tau_p,\alpha) ,$$

the pedigree likelihood can be expressed in the following simple form:

$$P(X,C \mid smpl, X_1C_1,\theta,\varepsilon) \propto_{\theta,\varepsilon} \frac{P(X,C \mid \theta,\varepsilon)}{P(X_1,C_1 \mid \theta)} ,$$

which is what Cannings and Thompson (1977) proposed and repeatedly used (Thompson and Cannings, 1979; Thompson, 1986). It is important to stress that this equivalence does not hold whenever either of the two conditions formulated above is not fulfilled. Thus, the simple ascertainment correction of conditioning the pedigree likelihood on the sampled PC can provide consistent estimation of the trait inheritance model only in very special situations - the occurrence of which in practice would appear to be rather doubtful.

6.11. Censoring pedigrees

Consider now the general case of a sampling sub-procedure where not all the sampled pedigrees are included in the sample that is analyzed, but only those having the predefined phenotypic content X^a in a certain pedigree substructure C^a. There can be various reasons for the selective inclusion of pedigrees in the sample subjected to analysis. For example we may find, when studying a complex trait, that no simple model satisfactorily describes its inheritance; we may believe that this is caused by heterogeneity of the sampled data (whether real or caused by a badly defined phenotype). In this case, it is natural to form sub-samples for separate analysis - for example, we may select out for separate analysis pedigrees containing members with some special phenotype.

To provide a correct estimate of the trait inheritance model, the SMB (sampling model based) likelihood of the pedigree that has been ascertained, extended *and* included in the sample that is analyzed can be formulated as follows. Its numerator should be the same as in (3.1), explicitly including not only the ascertainment, but also the inclusion procedure. Its denominator, i.e., the probability of the sample space on which the likelihood of the sampled pedigree is defined, depends on the particular formulation of the inclusion procedure.

It was assumed above that all sampled pedigrees having the same structure C contain the same substructure C^a. The inclusion condition can be, for example, the number of affected members in the sampled pedigree, or the existence of at least one spouse pair having previously defined phenotypes, or special phenotypes in members of those component nuclear pedigrees that contain probands, etc. Therefore, to estimate a trait inheritance model that is free of asymptotic bias caused by this sampling

sub-procedure, it is sufficient (but not necessary) to use the pedigree likelihood conditioned on both structures, the pedigree structure C and the structure of the true pedigree PSF:

$$P[X,C \mid smpl,C,C(\tau_p),\theta,\varepsilon,\alpha] = \frac{P[X,C,smpl \mid C(\tau_p),\theta,\varepsilon,\alpha]}{P[C,smpl \mid C(\tau_p),\theta,\varepsilon,\alpha]} \quad (6.12)$$

$$= \frac{P(X,C \mid \theta,\varepsilon) \sum_{X(\tau_p)\setminus X_p} P[X(\tau_p)\setminus X_p \mid C(\tau_p),X,C,\theta] Q(X,C,\tau_p)}{\sum_{X \cup X(\tau_p)} P[(X,C) \cup X(\tau_p) \mid C(\tau_p),\theta,\varepsilon] Q(X,C,\tau_p)},$$

where $Q(X,C,\tau_p) = P(asc \mid X_1,C_1,\tau_p,\alpha)P(incl \mid X^a,C^a,\psi)$.

This SMB likelihood is calculable whenever the models of the trait inheritance and sampling procedures (ascertainment, extension and inclusion) are given and the PSF structure is known, i.e., (6.12) can be used in the same cases as (6.9). Note in addition that conditioning on the sampled structure C necessarily presupposes conditioning on the initially ascertained substructure C_1. Otherwise, the probability $P(XC \mid \theta,\varepsilon)$ cannot be calculated.

If the substructure C^a is defined differently, the likelihood conditioning can be made less rigorous. For example, let the pedigree PSF consist of only children under a certain age, the parents of each ascertained proband are obligatorily observed, and the pedigree is included in the sample that is analyzed if at least one such parent is affected. In this sampling scheme, the substructure C^a relevant to inclusion is uniquely determined by the sampled PC and does not depend on the pedigree structure collected outside C^a. Thus the pedigree likelihood can be conditioned on only the structure C^a (together, of course, with the PSF structure), i.e., the denominator of (6.12) can be expressed as

$$\sum_{X^a \cup X(\tau_p)} P[(X^a, C^a) \cup X(\tau_p) \mid C(\tau_p), \theta, \varepsilon] \times$$

$$\times P(asc \mid X_1, C_1, \tau_p, \alpha) P(incl \mid X^a, C^a, \psi),$$

where $P[(X^a, C^a) \cup X(\tau_p) \mid C(\tau_p), \theta, \varepsilon] =$

$$= \sum_{(X,C) \setminus (X^a, C^a)} P[(X, C) \cup X(\tau_p) \mid C(\tau_p), \theta, \varepsilon].$$

To construct the sampling model free (SMF) likelihood, let us introduce the subset of members "relevant to sampling" (RS) as those who are either "relevant to ascertainment" or "relevant to inclusion". This subset RS has structure $C_p^a(\tau) = C^a \cup C(\tau_p)$ and phenotypic content $X_p^a(\tau) = X^a \cup X(\tau_p)$.

Reiterating now the arguments that led to the AMF likelihood (Section 6.8), it is possible to show that a correct (consistent if $\theta_0 \subset \theta$) estimator of the trait inheritance model can be obtained using the SMF likelihood conditional on the RS data $[X_p^a(\tau), C_p^a(\tau)] = (X^a, C^a) \cup \tau_p$. This can be expressed in the form (6.10) with the following replacements: $C_p^a(\tau)$ instead of $C(\tau_p)$, and $X_p^a(\tau)$ instead of $X(\tau_p)$. Let us stress once again that the problem of missing data on members of $C_p^a(\tau)$ is the same as was mentioned for likelihood (6.10): these missing data make undefined the sample space on which the probabilistic measure, the pedigree likelihood, should be mathematically defined.

6.12. Bivariate analysis

In the above considerations, no special distinction was made between the variously defined traits under study. The results obtained for

the likelihood formulation hold true for any set **X**, including qualitative, quantitative and complex traits. They are applicable, in particular, to a bivariate trait, for which the general model of joint inheritance of two traits, *Y* and *Z*, is given in section 5.3. Using this model and explicitly formulated sampling models, we can construct each of the pedigree likelihood functions considered in this chapter.

However, taking into account that not just one, but two, phenotypes are observed on the pedigree members, some specific changes in the formulation of the sampling procedures α, ε and ψ might be needed if each of the traits determines these procedures differently. Assume, for example, that the pedigree ascertainment is fully determined by the *Y*-content of the true pedigree PSF and not by its *Z* observations. The ascertainment probability is then defined as $P(asc \mid Y_1, Z_1, C_1, \tau_p, \alpha) \equiv P(asc \mid Y_1, C_1, \tau_p, \alpha)$, where τ_p is represented by three sets, $Y(\tau_p)$, $Z(\tau_p)$ and $C(\tau_p)$. Assume further that the pedigree extension (up to its stopping condition) is also determined by the Y-observations of the pedigree members already collected. Lastly, assume that the selective inclusion of the pedigree in the analyzed sample is determined by predefined *Z*-phenotypes of the C^a members, or by a combination of both phenotypes. As we can see, in this example the pedigree ascertainment and (sequential) extension is governed by one trait, while the inclusion of the collected pedigree in the analyzed sample is determined by the second trait also. This not infrequently occurs in practice, and the example shows that an adequate formulation of the component probabilities in the SMB and SMF likelihoods presented above makes them fully adequate for bivariate analysis. However, the identification of the subset RS, or at least of its

structure $C_p^a(\tau)$ (for SMB likelihood), is a necessary condition for formulation of the pedigree likelihood.

Once the pedigree likelihood has been formulated, the main goal of a bivariate pedigree analysis is to test the hypothesis that the joint inheritance of both traits can be described by pleiotropic control of two separately observed phenotypes manifested by the same major gene.

6.13. Illustration

The above theory of correcting the pedigree likelihood for the sampling procedures employed can be illustrated as follows. If the tested genetic model exactly models all the particulars of the true inheritance of the studied trait (the genetic model of inheritance and the sampling procedures, except, as usual, for the model parameters that are to be estimated from the particular pedigree sample), then all the proposed pedigree likelihoods (AMB and SMB, AMF and SMF) result in a consistent model estimator. This means that these likelihoods provide adequate correction for the sampling procedures determined by the previously introduced sampling design. No asymptotic bias of the model parameters is expected. This fact justified the above theoretical construction. At the same time, if the tested model of trait inheritance differs from the true one, because the latter is unknown, then the genetic model estimator is always more or less biased. The amount of bias depends on particulars of how the tested genetic model is formulated.

Without dwelling upon numerical details, consider once more the example of section 2.4. Pedigree samples were simulated from the sample space determined by the sampling design described in 2.4. and for each of them the AMB and AMF corrected likelihoods were used to estimate the

genetic model parameters. The results obtained can be presented as follows.

The simulated data were for MG control of a quantitative trait with the same ascertainment probability π for all PSF members having trait values exceeding a previously established threshold. For each simulated sample of size *n*, ML estimates of the trait model parameters were found. Using 10,000 replications, the expectation $E(\hat{\theta})$ and standard deviation $\sigma(\hat{\theta})$ were empirically found for each of these estimators. Fig. 4 shows the dependence of the bias, $\Delta\hat{\theta} = E(\hat{\theta})-\theta$, and standard deviation of the estimated allele frequency on sample size. (We use as our example the allele frequency because this is the population parameter most affected by the bias caused by the sampling procedure that is used).

As we can see, for finite-size samples all likelihoods produced biased estimators of the parameters, as is usual for the maximum likelihood technique. However, note what in practice is the important fact that the estimation bias is substantially less than the corresponding standard deviation of the parameter estimate, and this is especially so for small sample sizes. In large samples, the AMB likelihood provides unbiased estimators, as expected.

In this example, it was impossible to use the AMF corrected likelihood because some of the pedigree PSF members were not observed in their entirety and, therefore, the sample space was undefined. However, we tried to salvage the AMF approach in order to correct the pedigree likelihood without explicitly formulating the ascertainment model (an excellent example of a robust approach that provides an asymptotically unbiased genetic model estimator with minimum assumptions about the sampling procedure). We used an approximate AMF likelihood of the

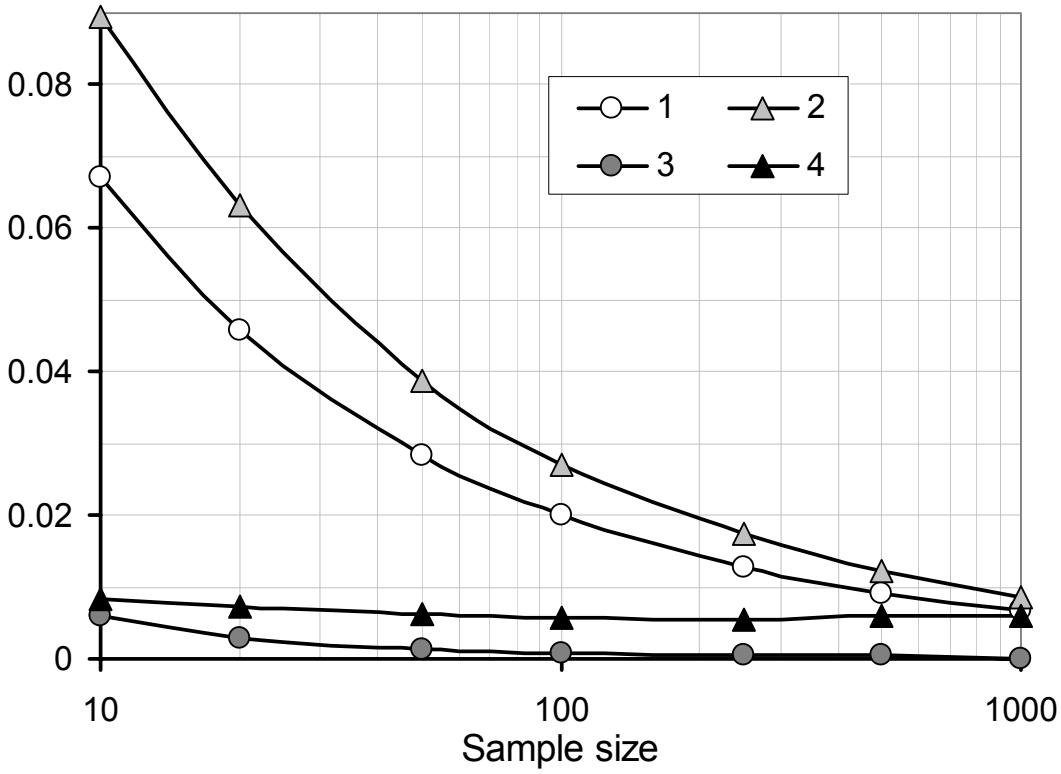


Figure 4. Dependence on the sample size of the standard deviation (1, 2) and bias (3, 4) of the estimator of allele frequency.

Pedigree likelihood conditioned on the PSF structure - 1 and 3; approximate_AMF likelihood - 2 and 4. See details in section 6.13.

form (6.8) with the following modifications. The pedigree data were taken to be the phenotype for each pedigree member who was observed and the trait mean for the pedigree members whose phenotypes were not observed. Provided the structure of the PSF is known, this approximate AMF likelihood can be easily calculated by replacing each missing phenotype on the PSF members by the sample mean. As can be seen in Fig.4, this approximate AMF likelihood provided negligible large-sample bias of the

genotypic value estimator: the bias is always smaller than the corresponding standard deviation.

6.14. SMB and SMF formulations

The above considerations were intended to show how it is possible to define a probabilistic measure in the form of the pedigree likelihood on the given sample space. It is possible in the AMB, SMB, AMF and SMF formulations. In each case, given enough knowledge about the genetic model and the sampling procedures determining the particular sample space, the proposed likelihood formulations are mathematically correct and, therefore, can be used in pedigree analysis.

The comparative usefulness of the two approaches considered above (SMB and SMF) cannot be established unequivocally for all possible situations. On the one hand, to venture upon an explicit formulation of the ascertainment (sampling) model is reasonable only if the real procedure used in the sampling process is known in sufficient detail. Any inadequate formulation of the ascertainment model can bring about untestable bias in the estimates of the trait inheritance model (including in the case of linkage). At the same time, it is well known that the field trial conditions of pedigree sampling hardly allow the possibility of strictly adhering to any previously established rules. Thus these rules can either only be roughly formulated with a relatively simplistic model, or an accurate formulation of them may be even impossible, if they are very complicated or incidentally violated. On the other hand, the SMF approach, although based on being independent of the details of the actual sampling procedure, automatically causes a loss (sometimes substantial) of pedigree information. As seen from previous sections, the SMF approach only *appears* to be free of the sampling model. In reality, as mentioned

above, the model is implicitly formulated. Moreover, in general, the SMF approach is applicable only under a rather strict condition: the pedigree PSF should be known in its entirety, both structurally and phenotypically, and we might doubt whether this can actually occur in many situations.

However, all the likelihood constructions considered above were explicitly based on the main assumption that the collected pedigree data are informative. We found it necessary to identify the subset of members in the pedigree PSF responsible for the pedigree to be sampled. Otherwise, it is impossible to define accurately the sample space to which the sampled pedigree belongs and to define the corresponding pedigree likelihood. We also found it necessary to identify the subset of pedigree members responsible for selective inclusion of the already collected (ascertained and extended) pedigrees in the sample that is analyzed. Otherwise, the sample space and the pedigree likelihood cannot be accurately defined mathematically, at least not in the form of the above formulations. In other words, only if we can accurately define the pedigree subsets responsible for the particular sampling procedure employed can the above likelihood constructions have any probabilistic sense.

We assumed above that the subset responsible for inclusion of the pedigree in the sample analyzed is fully and uniquely determined by the sampled pedigree structure $C^a \subseteq C$. This means that, given an explicitly formulated condition of what phenotypic content of this subset should determine each particular inclusion, an accurate identification of this subset seems to be a quite solvable problem. This is also the case when the pedigree PSF is identified using the questionnaire introduced above. The PSF needs to be known for all the above likelihood constructions.

Practical construction and use of such a questionnaire is not an easily solved problem. In some cases, for some pedigrees, it would be possible to identify the PSF structure quite accurately; while in others, which might be more numerous and more important, such accuracy may not be achievable and, therefore, the basic conditions for the above likelihood construction cannot be fulfilled. An accurate likelihood construction in those cases becomes impossible, and the only way to define the likelihood measure on the given sample space is “approximate”. Below, we shall consider in some detail the problems of this approximate likelihood construction and its impact on the results of a pedigree analysis.

7. SAMPLING CORRECTION IN LINKAGE ANALYSIS

7.1. Linkage problems

Two main approaches are recognized in linkage analysis, model-free and model-based (Elston, 1998). In the former, which often uses pedigree data specially designed for this type of analysis, the transmission of alleles at a marker locus from generation to generation is directly compared with the transmission of the trait phenotype being studied. If these two transmission patterns are found to be significantly "associated" with one another, then, taking into account the previously established design of the study, the hypothesis of no linkage between the marker locus and the trait is rejected; otherwise it is accepted. In the latter case, a genetic model of joint inheritance of the trait and marker phenotypes is used to formulate the likelihood explicitly for each sampled pedigree, and the (parametric) null hypothesis H_0: $\rho = 0.5$ of no linkage between the marker locus and the trait-controlling locus is tested against the alternative: H_1: $\rho < 0.5$, where ρ denotes the recombination fraction between the trait and marker loci. If H_0 is rejected, then a second stage of the linkage analysis is carried out, namely, point and interval estimation of the recombination fraction(s).

Linkage analysis is performed on pedigrees that are usually not randomly sampled, but rather sampled according to a certain scheme that is determined by a sampling design previously defined by the investigator planning the analysis. For model-based analysis, this fact is taken into account through a special correction of the pedigree likelihood obtained by conditioning on the particulars of the sampling procedure. If this procedure cannot be adequately formulated, then a likelihood correction that has been

proposed is to condition on the sampled trait data (Risch, 1984; Clerget-Darpoux et al., 1986; Elston, 1989; Greenberg, 1989; Clerget-Darpoux and Bonaïti-Pellié, 1992; Hodge and Elston, 1994; Wang et al., 2000). Vieland and Hodge (1996, p.1073) noted that the problem of ascertainment correction in linkage analysis is "fundamentally intractable" "not only for lod scores per se, but for any likelihood-based method for linkage or joint linkage/segregation analysis".

7.2. Basic notation

In previous chapters, we considered the general theoretical possibility of adequately correcting the likelihood for the sampling procedures. It was shown that this correction can be made in some practical cases of pedigree analysis. This possibility was shown in a quite general form: no special assumptions were made about the ascertainment probability; no special assumption was made about the manner of intrafamilial pedigree extension (except the possibility of explicitly formulating the probability of the extended pedigree $P(X_2, C_2 \mid X_1, C_1, \theta, \varepsilon)$ - the necessary condition for all other sampling corrections), and no special conditions were put on determining the selective inclusion of the pedigree in the sample analyzed. From this, we might expect that, contrary to the statement by Vieland and Hodge (1996), the problem of correctly accounting for the sampling procedures in linkage analysis can be solved. We shall discuss below in what cases and how it is possible to do this.

In linkage analysis, each pedigree member is characterized by two functionally different traits, the trait describing the biological function being studied and the specific so-called marker locus, which usually does

not have its own phenotypic manifestation of interest, but for which the chromosomal position is known (perhaps from a previous study). Suppose that, after analyzing the pedigree sample, we find a statistically significant association between the transmission across generations of the phenotypes of the trait under study and the marker genotypes. Then linkage between a putative locus controlling the trait and the marker locus will have been established and we can find the chromosomal position of this putative locus by point and interval estimation of the recombination fraction between the two loci.

7.2.1. Joint trait-marker model of inheritance

Suppose the trait inheritance can be described by a diallelic monogenic model determined by a set of parameters θ_t, and let A_1 and A_2 be the two alleles at the trait locus (the assumption of a single diallelic trait locus is made for simplicity only, without loss of generality). Similarly, suppose the marker phenotype inheritance is described by a monogenic model, but not necessarily diallelic, with parameters θ_l (population frequencies of the marker locus alleles, penetrances of the marker genotypes etc.), and let M_m be the m-th allele at the marker locus, $m = 1,\ldots,M$. Besides θ_t and θ_l, the joint inheritance of the trait and marker is described by additional parameters: the recombination fraction(s) ρ, determining linkage between the two loci, and the disequilibrium parameter(s) $D = \{D_{am}\}$, where $D_{am} = \Pr(A_a M_m) - p_a q_m$ is the difference between the population frequency of the haplotype $A_a M_m$ and the frequency expected when the alleles at the two loci are distributed independently; $a = 1,2$; $p_1 = \Pr(A_1)$ and $p_2 = 1 - p_1$ are frequencies of the trait alleles; and $q_m = \Pr(M_m)$ is the frequency of m-th marker allele, $\sum_m q_m$

= 1. There are only $M - 1$ independent values of D_{am} because of the constraints $D_{1m} = -D_{2m}$ for any m, and $\sum_m D_{am} = 0$ for any a.

7.2.2. Pedigree data

The rest of the notation is similar to that used before: the set $\{\tau\}$ determines the population of true pedigrees under study and the set $\{\tau_p\}$ of true pedigree PSFs determines the ascertainment design of the study. The sampled pedigree data are specified as follows: $X = \{x_i\}$ and $Y = \{y_i\}$; $i = 1,\ldots,n$, are the sets of trait and marker data, respectively, observed on the n pedigree members whose relationships form the pedigree structure C. By definition, $(X,Y,C) \subseteq \tau$, where τ is the true pedigree from which the sampling was performed. Here we use a triple notation for the sampled pedigree to characterize its structure C and its trait X and marker Y phenotypic contents. As was defined in section 6.3, we divide these n pedigree members into three groups forming three pedigree substructures (see Fig. 1):

- C_1 is the proband combination (PC), the substructure of probands who caused the pedigree ascertainment - this is the initially ascertained part of the pedigree.

- C_p ($\subseteq C$) is the substructure of sampled pedigree members who could in principle be probands because of their characteristics such as age, duration of residence, etc. – the sampled part of the pedigree PSF: $C_p \subseteq \tau_p$. This subset necessarily includes the actual probands: $C_1 \subseteq C_p$. By chance, members from the complementary substructure $C_p \setminus C_1$ did not realize their proband potentials in this particular ascertainment event.

- $C \setminus C_p$ is the subset of pedigree members who could not be probands because of the way the probands are defined for this particular pedigree

ascertainment design. The trait and marker sets, X and Y, are each divided into the three corresponding subsets: $X = X_1+(X_p\backslash X_1)+(X\backslash X_p)$ and $Y = Y_1+(Y_p\backslash Y_1)+(Y\backslash Y_p)$.

We assume additionally that not all the sampled (ascertained and extended) pedigrees are included in the sample that is analyzed. Inclusion in the sample is determined by an additional condition: a substructure $C^a \subseteq C$ is defined in such a way that the trait and/or marker data of its members affect the probability of including this pedigree in the sample analyzed. For example, it is known that a pedigree is not informative for linkage if it contains no doubly heterozygous parent in at least one of its component nuclear families. In this case, C^a could be defined as the subset of all parent pairs in the pedigree, and the condition for inclusion would be that the ascertained and extended pedigree must have among its C^a members some who have trait and marker phenotypes compatible with a doubly heterozygous genotype. Denote by X^a and Y^a the sets of trait and marker phenotypes, respectively, on members of this substructure C^a.

We consider below the possibility of adequately correcting the likelihood in linkage analysis to account for three specific sampling design procedures: the pedigree ascertainment, extension and censoring.

7.3. Component probabilities

Let us define the basic probabilities that describe the joint trait-marker inheritance and the pedigree sampling process.

Let $P(X,Y,C \mid \theta_t,\theta_l,\varepsilon,D,\rho)$ be the joint probability of the pedigree having the structure C and the trait and marker data X and Y, respectively. This probability is constructed using an explicitly formulated model of joint trait-marker inheritance and a model of pedigree extension.

It is evident that the sampled structure C is specified in such a way that the initially ascertained pedigree PC, C_1, is also identified. Otherwise, the probability of the extended part of the pedigree, $(X,Y,C)\backslash(X_1,Y_1,C_1)$, cannot be defined. Assuming that the marker phenotypes of the pedigree members have no connection with the intrafamilial pedigree extension, the extension model can be described as follows. The structure of each newly incorporated pedigree part is determined by the structure and trait phenotypes of the pedigree members already sampled and, which is not infrequently the case, by some knowledge about the true pedigree from which the sampling is performed, e.g., the fact that certain of its members are available for observation. This information can be obtained by using the questionnaire defined above as the necessary instrument for sampling the pedigree in the first place. In general, this sequential extension results in a *trait-dependent* pedigree structure. If incorporation of the next pedigree members occurs independently of the phenotypes of the previously observed members, e.g., all members available for observation are included, then the sampled pedigree structure is *trait-independent*. In this case,

$$P(X,Y \mid C,\theta_t,\theta_l,D,\rho) \propto_{\theta_t,\theta_l,D,\rho} P(X,Y,C \mid \theta_t,\theta_l,D,\rho),$$

by which we mean that the joint probability of the observed data (X,Y) conditional on the pedigree structure is equivalent to that not conditioned in this manner (regardless of the particular trait-independent scheme of pedigree extension actually used), in the sense that both sides achieve their maximum at the same values of the parameters θ_t, θ_l, D and ρ.

In terms of the proband ascertainment scheme, the probability $P(asc \mid X_1,C_1,\tau_p,\alpha)$ of ascertaining a pedigree having its particular PC is determined by the whole set of potential probands (τ_p) in the true pedigree

from which the ascertainment takes place. This is the probability that the PSF members forming the structure C_1 simultaneously become probands and together cause the pedigree ascertainment, while the other PSF members do not realize their proband potentials. Usually, only part of this pedigree PSF is sampled. The questionnaire helps us reconstruct the unsampled substructure $C(\tau_p)\backslash C_p$, i.e., to identify its members - but not necessarily their phenotypes. If necessary, other unsampled pedigree members, providing the relationship connections between C and $C(\tau_p)\backslash C_p$, are also identified.

Let $P(incl \mid X^a, Y^a, C^a, \psi)$ be the probability that the pedigree is included in the sample subjected to linkage analysis when it contains the particular subsets of trait and marker phenotypes X^a and Y^a, respectively, in the pedigree substructure C^a. Accordingly, $P(\overline{incl} \mid X^a, Y^a, C^a, \psi) = 1 - P(incl \mid X^a, Y^a, C^a, \psi)$ is the probability that a pedigree having the subset (X^a, Y^a, C^a) is not included in the sample analyzed. The inclusion probability can depend on some specific parameter(s) ψ modeling this procedure.

Without dwelling upon a possible definition of the inclusion condition, i.e., upon the particular formulation of the inclusion probability $P(incl \mid X^a, Y^a, C^a, \psi)$, let us recall that Stene (1977, 1978) proposed an approximate formulation for the ascertainment probability. He assumed that this was proportional to the number of affected members in the pedigree PSF. Similar to this, the inclusion probability could be approximated as: $P(incl \mid X^a, Y^a, C^a, \psi) = 1 - (1 - \Psi)^k$, where k is the number of those pairs of spouses that have at least one member whose trait and marker phenotypes are compatible with a doubly heterozygous

genotype, and $0 \leq \psi \leq 1$ is a parameter that determines how the probability of pedigree inclusion depends on the number of such spouse pairs. Here, the case $\psi \rightarrow 1$ can be called "complete" inclusion, when all pedigrees of the given structure that have at least one doubly heterozygous parent are included in the sample for analysis; and $\psi \rightarrow 0$ can be called "single" inclusion, when the probability of inclusion is proportional to the number of component nuclear families that have at least one doubly heterozygous parent.

7.4. General form of the linkage likelihood

Our goal is to consider in more detail the conditions under which the pedigree likelihood yields asymptotically unbiased estimates of the parameter ρ, the main purpose of linkage analysis. To do this, it is necessary to distinguish between two versions of linkage analysis, pure linkage analysis and what is often called joint segregation-linkage (JSL) analysis, but which in reality is linkage analysis while jointly estimating the inheritance model parameters (i.e., there is no segregation *analysis* in the sense of determining whether or not there is major gene segregation, but there *is* estimation of some segregation model parameters). In the first, it is assumed that the trait inheritance model θ_t is given together with its parameters. The same is assumed about the models determining ascertainment α, and extension ε. Linkage analysis is performed purely to estimate ρ, while the parameters D, θ_l and ψ are considered known. In the second version, JSL analysis, the form of the genetic model of trait inheritance (usually, but not necessarily, monogenic) is also assumed to be known - otherwise, it is not clear what linkage with the given marker locus is being estimated - but the parameters θ_t are assumed to be unknown, and

are to be estimated together with α, ε, D, θ_l, ψ and ρ. Making this distinction, the following can be said about the possibility of obtaining a consistent estimator of ρ.

In general, the pedigree likelihood is defined as the probability of the particular pedigree data, (X,Y,C), conditional on the pedigree having been sampled, i.e., ascertained, extended and included in the sample analyzed:

$$P(X,Y,C \mid smpl,\theta_t,\theta_l,\varepsilon,D,\alpha,\psi,\rho) =$$

$$= \frac{P(X,Y,C \mid \theta_t,\theta_l,\varepsilon,D,\rho)P(asc \mid X_1,C_1,\tau_p,\alpha)P(incl \mid X^a,Y^a,C^a,\psi)}{P(smpl \mid \theta_t,\theta_l,\varepsilon,D,\alpha,\psi,\rho)}, \quad (7.1)$$

where the denominator is the probability that the pedigree is sampled, and is expressed as the sum of the numerator taken over all pedigree data (X,Y,C) possible under the sampling procedure employed.

In this general form, the likelihood cannot always be calculated using only the sampled data. In particular, the second factor in the numerator cannot be calculated because it contains unobserved data in $C(\tau_p)\backslash C_p$, and the denominator cannot be found if the population distribution of possible PSFs, τ_p, is unknown. This caused Vieland and Hodge (1996) to note that it is impossible to construct an adequate likelihood correction "not only for lod scores per se, but for any likelihood-based method for linkage or joint linkage/segregation analysis". As follows from the previous chapters, the real problem lies in not having complete knowledge of the ascertainment procedure. There is no additional linkage-specific cause of intractability of this likelihood.

7.5. SMB likelihood for linkage

In the previous chapter, it was noted that the SMB pedigree likelihood provides a consistent estimator of the trait inheritance model if it is conditioned on the whole pedigree substructure, which necessarily includes the structure of the subset RS. Its structure is $C_p^a(\tau) = C(\tau_p) \cup C^a$. This holds true for linkage analysis also.

The structure "relevant to ascertainment", $C(\tau_p)$, is not known if the pedigree PSF is not sampled in its entirety. Only part of this structure, $C_p \subseteq C(\tau_p)$, is identified in the sampled pedigree and the members of this part are phenotypically observed (except possibly for missing observations). However, we assume that the questionnaire is used to reconstruct (but not to observe phenotypically) the substructure $C(\tau_p) \backslash C_p$ of the PSF members not included in the sampled pedigree. In this way, the structure $C(\tau_p)$ of the true pedigree PSF becomes known.

The second component of the structure RS, the structure of the subset "relevant to inclusion", can be different for differently sampled pedigrees. However, as was assumed above that, for each given sampled structure C, the substructure C^a is uniquely determined whenever the procedure for pedigree inclusion is defined.

Thus, to obtain a pedigree likelihood that yields a consistent estimator of the joint trait-marker inheritance model, it is sufficient (but not necessary; see the example below) to condition it on the joint structure $C \cup C(\tau_p)$ of the sampled pedigree structure and the structure learnt from the questionnaire:

$$P[X, Y, C \mid smpl, C \cup C(\tau_p), \theta_t, \theta_l, \varepsilon, D, \alpha, \psi, \rho] =$$

$$= \frac{P[X,Y,C,smpl \mid C(\tau_p),\theta_t,\theta_l,\varepsilon,D,\alpha,\psi,\rho]}{P[C,smpl \mid C(\tau_p),\theta_t,\theta_l,\varepsilon,D,\alpha,\psi,\rho]}, \qquad (7.2)$$

where the numerator is the joint probability that the pedigree is (*X,Y,C)* and is sampled (ascertained, extended and included in the analysis), conditional on the PSF structure of the true pedigree from which the sampling was performed. The denominator is the probability of having the sample of structure *C* collected from a true pedigree having the given PSF structure $C(\tau_p)$. The sample space on which likelihood (7.2) is defined is the set of all pedigrees having the same structure $C \cup C(\tau_p)$. Let us find explicit expressions for the numerator and denominator, respectively, and discuss whether they can be calculated.

Let the sampled pedigree be (*X,Y,C*) and let the PSF of the true pedigree from which the sampling was performed be τ_p, with structure $C(\tau_p)$ and trait content $X(\tau_p)$. The non-empty part of the PSF having structure C_p has been sampled ($C_p \subseteq C$) and the trait and marker phenotypes of its members have been observed, while the other part, $C(\tau_p)\backslash C_p$, has been reconstructed (but not phenotypically observed) using the questionnaire, so that the data $X(\tau_p)\backslash X_p$ and $Y(\tau_p)\backslash Y_p$ are missing.

The joint probability that the pedigree is (*X,Y,C*) and is sampled, given the PSF structure, can be expressed as:

$$P[X,Y,C,smpl \mid C(\tau_p),\theta_t,\theta_l,\varepsilon,D,\alpha,\psi,\rho] =$$

$$= P(X,Y,C \mid \theta_t,\theta_l,\varepsilon,D,\rho)P(incl \mid X^a,Y^a,C^a,\psi) \times$$

$$\times \sum_{X(\tau_p)\backslash X_p} P[X(\tau_p)\backslash X_p \mid C(\tau_p)\backslash C_p, X,Y,\theta_t,\theta_l,D,\rho] \times$$

$$\times P(asc \mid X_1,C_1,\tau_p,\alpha), \qquad (7.3)$$

where the sum over possible $X(\tau_p)\backslash X_p$ is taken in accordance with the usual practice of handling missing data: the members included in the structure $C(\tau_p)\backslash C_p$ were not phenotypically observed. This expression is calculable because each of its components can be calculated: $P(X,Y,C\,|\,\theta_t,\theta_l,\varepsilon,D,\rho)$, $P(incl\,|\,X^a,Y^a,C^a,\psi)$ and $P(asc\,|\,X_1,C_1,\tau_p,\alpha)$ are explicitly given in each model-based linkage analysis. The conditional probability

$$P[X(\tau_p)\setminus X_p\,|\,C(\tau_p)\setminus C_p,X,Y,\theta_l,\theta_t,D,\rho]$$

depends not only on the trait inheritance model but in general also on joint marker-trait model parameters. If the marker and trait loci are linked or are in allelic disequilibrium, the marker data Y determine this conditional probability together with X.

The sample space on which the pedigree likelihood is defined should be chosen depending on the particular extension and inclusion procedures used.

As we have assumed, in each particular case the structure C^a of the pedigree members determining the inclusion of the pedigree in the analysis is uniquely determined by the structure of the sampled pedigree. This means that conditioning on the sampled pedigree structure, C, automatically implies conditioning at the same time on the structure C^a. Thus, the sample space can be determined by the following probability (the denominator of the likelihood):

$$P[C,smpl\,|\,C(\tau_p),\theta_t,\theta_l,\varepsilon,D,\alpha,\psi,\rho]=$$

$$=\sum_{X,Y}P[X,Y,C,smpl\,|\,C(\tau_p),\theta_t,\theta_l,\varepsilon,D,\alpha,\psi,\rho]=$$

$$= \sum_{X,Y} P(X,Y,C \mid \theta_t,\theta_l,\varepsilon,D,\rho) P(incl \mid X^a,Y^a,C^a,\psi) \times$$

$$\times \sum_{X(\tau_p)\setminus X_p} P[X(\tau_p)\setminus X_p \mid C(\tau_p)\setminus C_p, X, Y, \theta_t, \theta_l, D, \rho] \times$$

$$\times P(asc \mid X_1, C_1, \tau_p, \alpha), \quad (7.4)$$

where the sum goes over all trait and marker data that are possible for the given sampled pedigree structure, C, and the structure $C(\tau_p)\setminus C_p$ reconstructed using the questionnaire. Note that given C implies given C_1, the particular PC substructure from which C was extended in accordance with model ε. Expression (7.4) is calculable if the model of trait inheritance and the models of pedigree sampling (ascertainment, extension and inclusion) are given.

In general, likelihood (7.2) depends on all the parameters introduced above that determine the joint trait-marker inheritance and the sampling procedures. This means that, provided it can be calculated, this likelihood can be used in linkage analysis when only the parameters D and ρ are to be estimated, or, in JSL analysis, when nuisance parameters of the models of the trait and marker inheritance, θ_t and θ_l, and the parameters determining the sampling procedures, α, ε and ψ, are to be estimated together with D and ρ. Clearly, these estimators are not the same in linkage and JSL analyses. For any one particular pedigree sample, they are differently biased and have different sampling errors because the likelihoods are defined on two different parameter spaces. Likelihood (7.2) uses only the sampled data (with the addition of the structure $C(\tau_p)\setminus C_p$); and the ML estimators of the parameters, in particular those of the recombination fraction(s) ρ, are asymptotically unbiased and most efficient. It is necessary to stress that, should the calculability of (7.2) be in

doubt because of a very complicated sampling procedure, this technical problem of calculation is completely separate from the problem considered here, i.e. whether or not it is in principle possible to formulate an adequate sampling correction for the pedigree likelihood.

7.6. Marker-independent sampling

It was assumed above that the pedigree ascertainment and extension are in no way determined by the marker data. Assume now that the inclusion procedure also does not depend on the marker data, i.e., $P(incl \mid X^a, Y^a, C^a, \psi) \equiv P(incl \mid X^a, C^a, \psi)$. For example, the ascertained and extended pedigree enters the sample analyzed depending only on the number of "affected" pedigree members it contains. In this case, the denominator (7.4) of the likelihood (7.2) does not depend on the parameters determining the marker inheritance and the joint trait-marker distribution, θ_l, D and ρ, because of the obvious equality: $\sum_Y P(X, Y, C \mid \theta_t, \theta_l, \varepsilon, D, \rho) = P(X, C \mid \theta_t, \varepsilon)$.

The numerator (7.3) can now be rewritten as follows

$$P[X, Y, C, smpl \mid C(\tau_p), \theta_t, \theta_l, \varepsilon, D, \alpha, \psi, \rho] =$$

$$= P(X, Y, C \mid \theta_t, \theta_l, \varepsilon, D, \rho) P(incl \mid X^a, C^a, \psi) R(X, Y, \theta_t, \theta_l, D, \rho),$$

where $R(X, Y, \theta_t, \theta_l, D, \rho, \alpha) =$

$$= \sum_{X(\tau_p) \setminus X_p} P[X(\tau_p) \setminus X_p \mid C(\tau_p) \setminus C_p, X, Y, \theta_t, \theta_l, D, \rho] \times$$

$$\times P(asc \mid X_1, C_1, \tau_p, \alpha).$$

Let us consider two special conditions: 1) all members of the pedigree PSF have measured trait values, $X(\tau_p) \setminus X_p$ is empty, and 2) the ascertainment procedure is single in the sense that

$P(asc \mid X_1, C_1, \tau_p, \alpha) = P(asc \mid X_1, C_1, \alpha)$. If condition 1) or 2) is true, then $R(X, Y, \theta_t, \theta_l, D, \rho) = R(X, \theta_t)$ does not depend on θ_l, D, ρ or Y. In this case only one component of likelihood (7.2), namely, $P(X, Y, C \mid \theta_t, \theta_l, \varepsilon, D, \rho)$ depends on details of the joint trait-marker inheritance. The other components in the numerator and denominator do not depend on the parameters θ_l, D or ρ. Thus, we can then write the following equivalences:

$$P(X, Y, C \mid smpl, \theta_t, \theta_l, \varepsilon, D, \alpha, \psi, \rho) \propto_{\theta_l, D, \rho}$$

$$\propto_{\theta_l, D, \rho} P(X, Y, C \mid \theta_t, \theta_l, \varepsilon, D, \rho) \quad (7.5)$$

$$\propto_{\theta_l, D, \rho} \frac{P(X, Y, C \mid \theta_t, \theta_l, \varepsilon, D, \rho)}{P(X, C \mid \theta_t, \varepsilon)} = P(Y \mid X, C, \theta_t, \theta_l, \varepsilon, D, \rho), \quad (7.6)$$

where (7.6) is true because the probability of the condition $\sum_Y P(X, Y, C \mid \theta_t, \theta_l, \varepsilon, D, \rho) = P(X, C \mid \theta_t, \varepsilon)$ does not depend on the parameters θ_l, D or ρ. This means that, regardless of the procedures employed for ascertainment and inclusion (but provided that they are marker-independent and condition 1) or 2) is satisfied), the parameters θ_l, D and ρ determining the joint trait-marker inheritance can be estimated using the joint probability of the sampled data (7.5), or the conditional probability of the marker data given the sampled pedigree structure and the trait data (7.6). Neither of these likelihoods needs any explicit formulation of the ascertainment and inclusion procedures and, therefore, they avoid any bias in linkage analysis results that could be caused by incorrect formulation of these procedures.

We have assumed that the models of trait inheritance θ_t, and pedigree extension ε, are known; for example, these models could have

been estimated in a previous segregation analysis. In other words, (7.5) can be used only for standard linkage analysis. When either θ_t or α, or both, need to be estimated from pedigree data (JSL analysis), then this likelihood is expected to produce asymptotically biased estimators of ρ. Because it contains no sampling correction, not only the trait model parameters θ_t are expected to be biased, but so also are all the other parameters estimated from the same likelihood together with θ_t, including the estimator of ρ.

Likelihood (7.6), however, can also be used in JSL analysis in the case that sampling is marker independent and $R(X,Y,\theta_t,\theta_l,D,\rho)$ does not depend on θ_l,D,ρ or Y, because it then follows from conditioning (7.2) on the entire trait data. If the sampling is marker-independent, but condition 1) or 2) is not true, likelihoods (7.5) and (7.6) produce in general asymptotically biased estimators of ρ, although the value of the bias can be almost trivial.

The pedigree likelihood (7.6) conditional on the pedigree trait data has been recommended in a number of publications, although not all of them clearly stipulated the necessary condition of having marker-independent sampling (Risch, 1984; Clerget-Darpoux et al., 1986; Elston, 1989; Greenberg, 1989; Clerget-Darpoux and Bonaïti-Pellié, 1992; Hodge and Elston, 1994; Wang et al., 2000). Vieland and Hodge (1996) noted that this likelihood correction follows from the AAF approach of Ewens and Shute (1988) if 1) the pedigree sampling is really marker-independent, and 2) the pedigree PSF is completely known.

7.7. Marker-dependent sampling, SMF likelihood

To avoid inconsistency of the parameter estimators due to incorrect formulation of the ascertainment procedure, Ewens and Shute (1986) proposed the ascertainment-assumption-free (AAF) method of correcting the pedigree likelihood as used in segregation analysis. They showed that the pedigree likelihood conditioned on that part of the pedigree data "relevant to ascertainment" provides consistent estimators of the trait model parameters, regardless of the particular ascertainment scheme that has been used. However, it should be noted that these authors did not consider the case where we do not observe all the pedigree members relevant to ascertainment. The same approach can be applied to the more general sampling procedure that includes pedigree ascertainment, its extension, and also selective inclusion of the pedigree in the sample analyzed. In this case, it seems reasonable to call the method *sampling-model-free* (SMF).

Let us re-formulate likelihood (7.1) by substituting, for the product of the explicitly formulated ascertainment and inclusion probabilities

$$P(asc \mid X_1, C_1, \tau_p, \alpha) P(incl \mid X^a, Y^a, C^a, \psi),$$

the joint probability $P[X_p^a(\tau), Y^a, C_p^a(\tau)]$ of the pedigree being ascertained and included in the sample analyzed given its subset RS, $[X_p^a(\tau), Y^a, C_p^a(\tau)] = \tau_p \cup (X^a, Y^a, C^a)$, its structure and trait and marker contents. Assume now that the subset RS fully and uniquely determines this probability, although no explicit model of either ascertainment or inclusion is introduced. Note that this assumption means that the sampling probability is the same for the different PCs that provide the same PSF subset. Then, using the Ewens and Shute (1986) technique, i.e.,

considering this probability as a parameter and replacing it by its ML estimator in the likelihood expression, it is possible to obtain the following SMF likelihood:

$$P(X,Y,C \mid smpl,\theta_t,\theta_l,\varepsilon,D,\rho) \propto_{\theta_t,\theta_l,\varepsilon,D,\rho}$$

$$\propto_{\theta_t,\theta_l,\varepsilon,D,\rho} \frac{P(X,Y,C \mid \theta_t,\theta_l,\varepsilon,D,\rho)}{P[X_p^a(\tau),Y^a,C_p^a(\tau) \mid \theta_t,\theta_l,\varepsilon,D,\rho]} \ . \quad (7.7)$$

This is the probability of the sampled pedigree conditional on its subset RS, combining both the pedigree PSF and the "inclusion" data (X^a,C^a). This likelihood provides asymptotically unbiased estimators of all the parameters determining the joint trait-marker inheritance, θ_t,θ_l,D and ρ, and of the extension parameter ε. Usually (see Ewens and Shute, 1986; Shute and Ewens, 1988; Hodge, 1988; Sawyer, 1990), the ML estimators of parameters obtained from likelihood (7.7) are less efficient (sometimes substantially so) than those obtained from the SMB likelihood. This likelihood formulation is really robust, providing consistent estimation of the genetic model with minimal assumptions about the sampling procedure. The sample space on which (7.7) is defined consists of pedigrees having the same given content RS, both structurally and phenotypically. However, if the subset RS contains some unobserved data, the very space on which this conditional probability is to be defined becomes undefined and theoretically the SMF likelihood correction cannot be constructed.

In the case of trait-independent pedigree extension, the simpler likelihood can be used instead of (7.7): the conditional probabilities of the

sampled data (X,Y) given C, and of the data $(X_p^a(\tau), Y^a)$ given $C_p^a(\tau)$, can be used instead of the corresponding probabilities in (7.7).

7.8. Example

7.8.1. The pedigree data

The following example illustrates the above statements. Suppose the population consists of nuclear pedigrees each having at least 2 offspring. The phenotypes of the pedigree members are described by a binary trait (affected-unaffected) and a marker genotype. The trait is controlled by a diallelic locus in such a way that genotype A_1A_1 always has an affected phenotype, denoted 1, while genotypes A_1A_2 and A_2A_2 have an unaffected phenotype, denoted 0, where A_1 and A_2 are two alleles with population frequencies p and $1 - p$, respectively. The codominant marker locus has two alleles, M_1 and M_2, with frequencies q and $1 - q$, respectively. The joint distribution of the trait and marker alleles is determined by the disequilibrium parameter $D = \Pr(A_1M_1) - pq$, and by the recombination fraction ρ between the two loci.

The sampling procedure is defined as follows. The pedigree PSF consists of two members, the father and the oldest offspring. If affected, each of them becomes a proband with the same probability π independently of one another. Thus, the pedigree PC (proband combination) can be represented by only the father, by only the oldest offspring, or by both. The PC is further extended to include the mother and just one as yet unobserved offspring. There are two additional conditions for including the pedigree in the sample: 1) the marker genotypes in the parents should be M_1M_2 in the mother and M_1M_1 in the father (back-

cross), and 2) the mother should be unaffected, i.e., her phenotypic observation is always $0M_1M_2$ (otherwise, the pedigree is not informative for linkage). Thus, having in mind the scheme shown in Fig. 1, the pedigree PSF is represented by the trait phenotypes on the father and the oldest offspring. Here, this PSF is always sampled in its entirety, so: $(X_p,C_p) \equiv \tau_p$; $X_p^a \equiv X_p^a(\tau)$ and $C_p^a \equiv C_p^a(\tau)$. The subset controlling inclusion, (X^a,Y^a,C^a), contains the mother's trait phenotype and the marker genotypes of both parents. The complementary pedigree subset $(X,Y,C)\backslash(X_p^a,Y^a,C_p^a)$ is represented by the phenotype of at most one member – the younger offspring, if sampled – and by the marker genotypes of both offspring.

Table 7.1 presents the 28 different nuclear pedigrees that are possible under the sampling scheme just described, their pedigree structures, the phenotypes and marker genotypes of their members, and the pedigree probands (denoted by bold **1**). The mother's observation is not shown because it is always $0M_1M_2$. The first 4 three-member pedigrees can be collected when only the father becomes a proband, but not his oldest offspring. The latter in this case is observed in the process of pedigree extension and can be either affected or unaffected. The other pedigrees are ascertained through the oldest offspring proband and, in some cases, the father-proband combination, and all of them contain two offspring.

From the sampling scheme considered we can conclude that, in pedigrees 1-20, the father's genotype is unequivocally $A_1A_1M_1M_1$, while in the other 8 pedigrees it is $A_1A_2M_1M_1$, because $A_2A_2M_1M_1$ would provide no offspring proband. The mother's genotype can only

Table 7.1. List of 28 pedigrees that can be sampled from the population of nuclear families defined in the example considered.

N^0	F	Offspring	N^0	F	Offspring
1	**1**	0 M_1M_1	15	1	**1** M_1M_2 +1 M_1M_1
2	**1**	0 M_1M_2	16	1	**1** M_1M_2 +1 M_1M_2
3	**1**	1 M_1M_1	17	**1**	**1** M_1M_1 +1 M_1M_1
4	**1**	1 M_1M_2	18	**1**	**1** M_1M_1 +1 M_1M_2
5	1	**1** M_1M_1+0 M_1M_1	19	**1**	**1** M_1M_2 +1 M_1M_1
6	1	**1** M_1M_1+0 M_1M_2	20	**1**	**1** M_1M_2 +1 M_1M_2
7	1	**1** M_1M_2+0 M_1M_1	21	0	**1** M_1M_1+0 M_1M_1
8	1	**1** M_1M_2+0 M_1M_2	22	0	**1** M_1M_1+0 M_1M_2
9	**1**	**1** M_1M_1+0 M_1M_1	23	0	**1** M_1M_2+0 M_1M_1
10	**1**	**1** M_1M_1+0 M_1M_2	24	0	**1** M_1M_2+0 M_1M_2
11	**1**	**1** M_1M_2+0 M_1M_1	25	0	**1** M_1M_1+1 M_1M_1
12	**1**	**1** M_1M_2+0 M_1M_2	26	0	**1** M_1M_1+1 M_1M_2
13	1	**1** M_1M_1 +1 M_1M_1	27	0	**1** M_1M_2+1 M_1M_1
14	1	**1** M_1M_1 +1 M_1M_2	28	0	**1** M_1M_2+1 M_1M_2

*Notes: For each pedigree, its structure, the trait phenotypes of its members, and their marker genotypes are shown. F denotes the father phenotype (*1 – *affected,* 0 – *unaffected,* **1** – *the father is a proband).*

be A_1M_1/A_2M_2 or A_1M_2/A_2M_1 in all pedigrees except the first, where the genotype A_2M_1/A_2M_2 is also possible.

Fig. 5A shows how the bias of the estimator of ρ, $\Delta\hat{\rho}_n = E(\hat{\rho}_n) - \rho$, and its standard deviation, $\sigma(\hat{\rho}_n) = \sqrt{E[\hat{\rho}_n - E(\hat{\rho}_n)]^2}$, depend on the sample size, where $\hat{\rho}_n$ denotes the estimator found for a

sample of size n, and $E(x)$ denotes the expectation of x found by averaging the estimate obtained from 10,000 simulation replicates. These characteristics were obtained in a JSL analysis, when the nuisance parameters p, q, D (and π, for the SMB likelihood) were estimated together with ρ. Three likelihoods corrected for the sampling procedures were used: the SMB likelihood (7.5) conditional on the sampled pedigree structure; the SMB likelihood conditional on the structure RS (the mother's trait phenotype and the marker genotypes of both parents), and the SMF likelihood (7.7). This was possible because the subset RS was observed in its entirety. As we see, all three likelihoods produce asymptotically unbiased estimators of ρ.

As expected in the likelihood technique, the estimators of ρ obtained from these likelihoods were biased for finite sample sizes. However, for each n, the bias of the estimator of ρ was smaller than the corresponding standard deviation. Also as expected, the estimator of ρ was more efficient when the SMB likelihood was conditioned on only the structure RS than when it was conditioned on the whole pedigree structure. At the same time, contrary to what was expected, the SMF estimator of ρ turned out to be more efficient, i.e., its standard deviation was smaller, than that of the SMB estimator (Fig. 5A). However, this contradiction is only apparent. For the other parameters, the standard deviation was larger for the SMF method. This means that, when conditioning on more data, the general result holds that the norm of the inverse matrix of second derivations is larger (Sawyer, 1990), but this does not mean that *each* parameter estimator has a larger standard deviation.

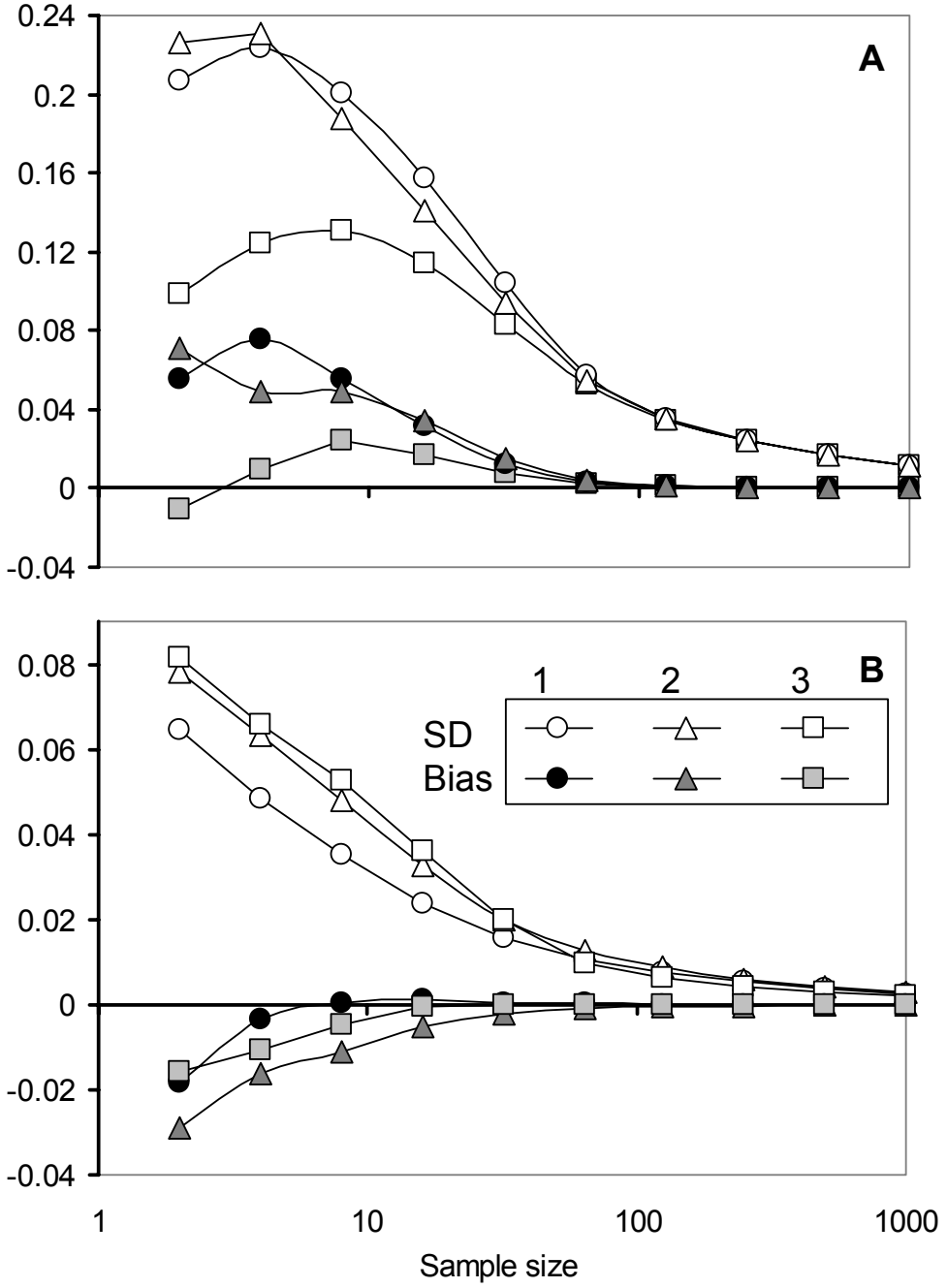


Figure 5. Dependence on the sample size of the bias and standard deviation of the estimators of the recombination fraction (A) and the disequilibrium parameter (B).

The estimates were obtained using: 1) the SMB likelihood conditioned on the pedigree structure; 2), the SMB likelihood conditioned on the RS structure, and 3) the SMF likelihood. See details in section 7.8.

7.9. Correction of the linkage likelihood

The theory developed above to correct the likelihood for the sampling procedures used in pedigree collection is quite applicable for linkage analysis: the same SMB and SMF likelihood forms can be used to obtain asymptotically unbiased estimators of the recombination fraction. This follows from the fact that the theory of adequate sampling correction for the pedigree likelihood was proved in the previous chapter quite generally. The trait studied can be quantitative or qualitative. Nothing special was assumed about the proband definition or how the ascertainment probability is formulated. The only condition required for the model of pedigree extension is that it should permit an explicit expression for the probability of the pedigree data collected in the process of intrafamilial pedigree extension. That is why this theory is perfectly applicable to linkage analysis, the main goal of which is to obtain accurate (in particular, asymptotically unbiased) estimation of the recombination fraction between the locus controlling the trait studied and the marker. If the pedigree samples on which this estimation is performed are collected using the previously planned sampling procedures (ascertainment, intrafamilial extension and censoring), the pedigree likelihood used in linkage analysis should be corrected for these sampling procedures in the way described by the above theory. This includes unambiguous identification of the subset of the pedigree members RS, modeling the sampling procedures (ascertainment and inclusion) reflecting the true ones, at least in their main details (Sawyer, 1990), and, of course, the procedure used for pedigree extension, without which it is impossible even to write down the joint probability of the collected pedigree data.

If these three conditions are met, the pedigree likelihood can accurately account for the sampling procedures and provide a consistent estimator of the trait inheritance, including the recombination fraction ρ. Concerning the SMF method of sampling correction, it should be once again noted that likelihood (7.7), which has the excellent property of robustness, is simply not defined if the subset RS contains missing data.

Note that, although the sampling correction should directly address the subset RS, it is always permissible, if deemed desirable, to condition the pedigree likelihood on other parts of the sampled data. The important conclusion that follows from the above considerations is that this additional conditioning should be made *together with* the sampling condition, not *instead of* it. For example, conditioning on the pedigree structure must include conditioning on the substructure of the subset RS.

In section 6.10, it was explained why the condition of using all the sampled data was formulated by Cannings and Thompson (1977). Although they said nothing against selective inclusion in other cases, avoiding it has been widely accepted since then (see Vieland and Hodge, 1995, 1996). On the one hand, accepting this seems understandable because, contrary to the more or less clear formulation of a proband ascertainment scheme, the inclusion procedures could be very diverse and, therefore, hardly amenable to a general formulation. On the other hand, selective inclusion of pedigrees in the sample analyzed seems to be widely practiced (and not only in linkage analysis), although this fact is not always explicitly stated. Here, we have considered selection of the pedigree for analysis as a special part of the sampling procedure which, together with the ascertainment, should be accounted for in order to obtain consistent estimators of the recombination fraction and other parameters

that determine joint trait-marker inheritance. We considered one characteristic of the members of C^a, as an example: the trait and marker data should be compatible with non-zero linkage information in the pedigree. However, this is equally applicable for any formulation of the probability $P(incl \mid X^a, Y^a, C^a, \psi)$.

7.10. Linkage test

The LRT statistics used to test the null hypothesis of no linkage, H_0: $\rho = 0.5$ (or $\rho_m = \rho_f = 0.5$), can be formulated as follows:

$$\lambda = 2 \ln[P(X, Y \mid \hat{\theta}, \hat{\rho}) / P(X, Y \mid \hat{\theta}, 0.5)], \qquad (7.8)$$

where $\hat{\rho}$ is the recombination fraction estimated together with other parameters, $\hat{\theta}$, of the joint distribution in pedigree members of the two phenotypes, the trait and marker. Asymptotically, the test statistic (7.8) is distributed as a central χ^2 with df = 1 (provided $\hat{\theta}$ is not restricted to being ≤ 0.5).

8. THE SET OF TESTED GENETIC MODELS

Pedigree analysis was defined in sections 1.6 and 1.7 as the formation of a set θ of genetic models, the introduction of an operator Ω ranking them in order of preference and, as the analysis result, the choice of the model providing the most accurate description of the inheritance of a trait. Here we consider these problems in more detail.

8.1. Likelihood ratio

The statistical analogue of the information measure I_{ij}^{S} of similarity between two inheritance models introduced in section 2.2 can be presented for *finite-size* samples in the form of the logarithm of a likelihood ratio (LR):

$$\lambda_{ij} = \sum_k \lambda_{ij}^{S_k} = \sum_k \ln[P^{S_k}(X_k, C_k \mid \hat{\theta}_i) / P^{S_k}(X_k, C_k \mid \hat{\theta}_j)], \quad (8.1)$$

where the sum goes over all sampled pedigrees; $P^{S_k}(X_k, C_k \mid \hat{\theta}_i)$ is the maximum likelihood for pedigree (X_k,C_k), and $\hat{\theta}_i$ denotes the ML estimate(s) of θ_i yielding this maximum. In (8.1), we explicitly take into account the fact that each pedigree (X_k,C_k) can be sampled with its own specific sampling procedure S_k . Provided they are sampled independently of one another, the sample log-likelihood is the sum of the log-likelihoods for the pedigrees included in the analyzed sample.

As we can see, (8.1) is the statistical realization of the general operator Ω that ranks the previously formed set of genetic models θ, for finite-size samples. Thus, if the sample distribution of this LR is known, the basic statistical problem of comparing the genetic models in θ can be

solved and characterized in terms of the traditional statistical "significance" of the model comparisons.

8.2. Transmission probability tests

Let us consider now what models are to be included in the set θ. First of all, they should be mathematical-genetic models. Second, this set should cover as many types of trait inheritance models as possible, to provide an analysis result that describes the mode of inheritance with maximum accuracy. Third, this set is limited by the complexity of the models to be tested. Any particular pedigree sample contains limited genetic information, and therefore we usually cannot distinguish complex multiparametric models from one another. Thus, at present mostly MG models are considered for pedigree samples of a reasonable (practically achievable) size, and additionally some types of two-locus models containing parameter constraints that may not be testable. It seems worthwhile to add the following about the particular formulation of these models.

Elston and Stewart (1971) introduced a transmission probability model under which to test hypotheses of inheritance in pedigree segregation analysis. To test for a MG model, where the trait is under the control of a single locus with two alleles, A_1 and A_2, their statistical model can be outlined as follows. Introduce the transmission probabilities $\tau_g = \Pr(A_1|g)$ that a parent with genotype g (g = 1, 2 and 3 for genotypes A_1A_1, A_1A_2 and A_2A_2, respectively) transmits allele A_1 to his/her offspring. Under this statistical model, the parameters τ_g are estimated from the pedigree sample to have arbitrary values, together with other parameters specified by the model θ_i. Now two specific genetic hypotheses can be

tested under this model. If the transmission is Mendelian, then τ_g = 1.0, 0.5 and 0.0 for g = 1, 2 and 3, respectively, and this forms the first hypothesis of interest. For the second hypothesis, we assume that all three τ_g are equal to the same value, $\bar{\tau}$ (which means that the offspring genotype is independent of his/her parental genotypes), estimated together with the other model parameters. Accordingly, two LR tests are introduced ("transmission probability tests"):

$$\lambda_{i1} = 2\sum_k \ln[P(X_k, C_k \mid \hat{\theta}_i, \hat{\tau}_g) / P(X_k, C_k \mid \hat{\theta}_i, \tau_M)]$$

and

$$\lambda_{i2} = 2\sum_k \ln[P(X_k, C_k \mid \hat{\theta}_i, \hat{\tau}_g) / P(X_k, C_k \mid \hat{\theta}_i, \bar{\tau})].$$

where τ_M denotes the Mendelian transmission probabilities, and $\hat{\tau}_g$ and $\bar{\tau}$ are estimates of the transmission probabilities found together with other model parameters.

For the pedigree sample $\{(X_k, C_k)\}$, the null hypothesis, H_0: $\tau_g = \tau_M$, that the trait is really controlled by one Mendelian diallelic locus, can be tested by the LRT statistic which, if the estimate $\hat{\tau}_g$ is unconstrained, is distributed asymptotically as the central χ^2 with df = 3 (Kendall and Stuart, 1970).

The distribution of the second test statistic was assumed asymptotically to be χ^2 with 2 df, if the offspring genotype is really independent of the parental genotypes. Using these tests, the monogenic model θ_i is included in the set $\boldsymbol{\theta}$ if it is not rejected by the test λ_{i1} and, at the same time, $\theta_i(\bar{\tau})$ is rejected by the test λ_{i2}.

Note that the use of such a statistical model to test hypotheses nested within it raises a specific problem. While each of the tests λ_{i1} or

λ_{i2}, provided their distribution are correctly established, can be characterized by the traditional statistical type I and II errors, the characteristics of the combined test, when $\theta_i(\hat{\tau})$ is not rejected and at the same time $\theta_i(\bar{\tau})$ is rejected, cannot be derived from any asymptotic assumptions (Ginsburg, 1984; Ginsburg and Livshits, 1999). They are expected to be different for different situations.

It should be stressed that the models $\theta_i(\hat{\tau})$ and $\theta_i(\bar{\tau})$ were not intended to be used for a description of the trait inheritance, but rather to allow us to test whether a simple monogenic model can be used to describe the mode of inheritance. Accordingly, neither of these auxiliary models is included in the set $\boldsymbol{\theta}$ from which the "best" (in the sense defined above) is to be chosen as the analysis result. Thus, the transmission probability tests provide the first limitation of the genetic models that are to be included in the set $\boldsymbol{\theta}$. Only those models are included whose genetic content is compatible with genetically determined MG inheritance (the genotypic set **G** and the three component distributions defining the model).

Based on a particular sample of pedigrees, the transmission probability tests result in either acceptance or rejection of the null hypothesis H_0 that the tested genetic model can be included in the set $\boldsymbol{\theta}$. H_0 is accepted in two cases. In the first, the model formulation exactly corresponds to the conditions formulated in section 6.1; in other words, the model describes the real MG genetic control, except for the model parameters, which are to be estimated from the particular pedigree sample. In the second, the transmission probability tests λ_{i1} and λ_{i2} are not sufficiently powerful to reject H_0 based on the given pedigree sample with its limited information. In this case, conditions $C_1 - C_5$ (at least not all of them, see section 6.1) do not strictly correspond to of the particular mode

of inheritance of the trait being studied. It is natural to assume that a corresponding increase in the analyzed sample size, i.e., an increase in the test's power, would result in rejection of H_0, in exclusion of the tested model from the set θ, and in the need to re-formulate the model - making corresponding changes to the components of the genetic model defined in section 1.5.

Unfortunately, it is in practice impossible to extend this idea of transmission probability tests, developed for the MG models, to more complicated genetic models. The number of probabilities that describe the transmission of parental alleles to offspring (gametes, if more than one gene is involved in the genotypic control) increases sharply, making it impossible to estimate them with reasonable accuracy from finite-size pedigree samples. Skipping the proof, let us mention only two examples. For the three-allele monogenic model the number of transmission probabilities that should be estimated from the pedigree sample equals 20, instead of 3, for the MG model. For the digenic model, each gene having 2 alleles, this number is already 30.

From what has been said it follows that formation of the set θ is usually made only approximately. In practice it is highly improbable that the formulated genetic model would correspond exactly to the real mode of inheritance and to the sampling procedures employed (conditions C_1 – C_5). Thus, inclusion of the formulated models in the set θ is determined mostly by the sample that is analyzed because this sample is not informative enough to reject the model from this set.

8.3. Most parsimonious models

Any accepted parameter constraint leads to a particular simplified model of inheritance (additive trait control, equal residual variances for the three major genotypes, etc). By making the description of the mode of inheritance more "economical", such a constraint clearly decreases the maximum pedigree likelihood, $P^{S_k}(X_k, C_k \mid \hat{\theta}_i)$, in comparison with that of a model not having that constraint. If this decrease is found to be statistically non-significant, the simpler model would be preferred. This does not mean that such a simplification provides the same accuracy in describing the trait inheritance. In all probability, we might expect that the simpler model would lose some details in describing the trait inheritance. However, if the loss in this description is not substantial, and its statistical acceptance seems to confirm this fact, then the simplified model that has been obtained under constraints that are not statistically rejected can be used as an approximate description of the mode of trait inheritance.

The *most parsimonious* (MP) model is defined as the one for which any further parameter constraint is rejected statistically (with predefined type I error and power that depends on the amount of information in the sample). This idea of using the MP model provides further simplification of the set of genetic models from which to obtain the pedigree analysis result. The initially formed set θ is divided into groups, each group containing a "general" genetic model and the genetic models that can be derived from this general model by various parameter constraints.

If model θ_j is a special case of θ_i, i.e., can be obtained from the latter by some parameter constraints, then the sampling distribution of $2\lambda_{ij} = 2\ln[P(X, C \mid \hat{\theta}_i) / P(X, C \mid \hat{\theta}_j)]$ can be approximated asymptotically by

a central χ^2 with df equal to the number of independent constraints, under the assumption that θ_i and θ_j do not differ from one another in their ability to describe the pedigree distribution. Otherwise, the distribution could be approximated by a non-central χ^2 (Wald, 1943). Using these approximations, we can test the null hypothesis H_0 that the particular parameter constraint is acceptable, i.e., does not result in a statistically significant change in the genetic description of the trait inheritance. If H_0 is accepted, the genetic model can be formulated in its simplified form. Otherwise, this constraint is rejected.

Using this technique for each group of genetic models within which these nested relations can be established, it is possible to replace this group operationally by only one (usually) most parsimonious model, thus forming the reduced set $\{\hat{\theta}_i^{MP}\}$.

The two procedures reducing the set of genetic models that should be compared with one another for the choice of the one providing the most accurate description of the trait inheritance should ideally be performed in a certain order. The transmission probability test should be used first, and only then should the MP models be constructed. Attempts to use the transmission probability test on MP models can lead to biased results, because each MP model represents an approximation of the formulated genetic model, for which the level of approximation bias, and even its direction, is usually unknown.

There may be ambiguity regarding the construction of the most parsimonious model for each group. Assume for example that, at a certain stage of the process of testing constraints for a MG model, the two following constraints are statistically accepted: 1) additive major gene control with unconstrained residual variances, and 2) equal residual

variances with unconstrained genotypic values. It may be the case that both of these are found to be models that are most parsimonious, in the sense that any further parameter constraint would be significantly rejected in either of them. Statistical comparison of two such MP models can be made by a simulation test as described below in 8.4.

Thus, we set limits when we form the set of genetic models to be compared for their ability to describe the trait inheritance. The set is reduced, first by including only those models that are genetically formulated (using the transmission probability test) and second, by including only the MP models.

8.4. Comparison of differently formulated models

Let us rank the models from $\{\hat{\theta}_i^{MP}\}$ by their likelihoods. Now it is necessary to find out whether the difference between the first ranked and second ranked model is statistically significant (given the pedigree sample with its information about the trait inheritance and the predefined type I error of the statistical decision). The distribution of the LRT is defined unambiguously if the second ranked model is obtained from the first ranked by parameter constraints. In this case, the LRT distribution is approximated asymptotically by the central χ^2 with df equal to the number of the constraints placed, if these constraints are acceptable, or by a non-central χ^2, if not.

Consider now the case where the first and second ranked models from $\{\hat{\theta}_i^{MP}\}$ are not nested in this way one within the other. In this case, a simulation test can be used to test whether there is a statistically significant difference between these two genetic models in their description of the

trait inheritance. The test can be constructed as follows. Let $P(X, C \mid \hat{\theta}_1)$ and $P(X, C \mid \hat{\theta}_2)$ be the likelihoods, obtained on the same pedigree sample, for two differently formulated genetic models (different sets of genotypes, different control of phenotypes etc.; for example model θ_1 is digenic while θ_2 is the three-allele monogenic model). Assume that $P(X, C \mid \hat{\theta}_1) > P(X, C \mid \hat{\theta}_2)$ and the LR statistic for these two models is LRT = $\ln[P(X, C \mid \hat{\theta}_1) / P(X, C \mid \hat{\theta}_2)]$.

Assume that model $\hat{\theta}_2$ is true. Using the known parameter estimates $\hat{\theta}_2$, simulate a pedigree sample of the given size and structure and, estimating $\hat{\theta}_1$ and $\hat{\theta}_2$, find the new LR statistic. Repeating this simulation many times, it is possible to find the empirical distribution of the LR statistic for these two genetically different models. For a predefined type I error, α, and given sample size, n, the simulated critical value $c_{\alpha n}$ can be found as the upper 100α- percentile of this distribution. Then, to compare θ_1 and θ_2, the statistical decision is made as follows: H_0 ($\hat{\theta}_2$ is true) is rejected if the LRT = $\ln[P(X, C \mid \hat{\theta}_1) / P(X, C \mid \hat{\theta}_2)]$ found for the compared models exceeds the critical value $c_{\alpha n}$. This means that there is a statistically significant difference in the description of the trait inheritance: model $\hat{\theta}_2$ provides a less accurate description of the trait inheritance than does $\hat{\theta}_1$.

Repeating this simulation testing for any pair of models from the set $\{\hat{\theta}_i^{MP}\}$, we can obtain the resulting multiple comparisons. The set of models should then be divided into several groups such that all the models belonging to the same group do not differ statistically from one another.

Then it is possible to rank these groups by their accuracy in describing the trait inheritance: models from the first group provide the most accurate description, and this fact is statistically significant on using the simulation test. Models in the second group describe the trait inheritance significantly better than the models in the third group etc.

Thus, the problem of comparing and ranking the genetic models in $\{\hat{\theta}_i^{MP}\}$ can be solved statistically using the standard likelihood technique.

8.5. Planned and employed sampling models

Explicit formulation of the sampling procedures is a necessary part of constructing the genetic model. Above, we several times referred to the sampling procedure "employed" when the sample for pedigree analysis is collected. However, there is quite a difference between the conceived sampling design carefully planned initially by an investigator and the ascertainment, extension and inclusion procedures actually employed in practice. The deviations from the planned design could be negligible or substantial, even in some important aspects of it. Unfortunately, it is usually impossible to identify and to document all the factors that happened to cause these deviations and, therefore, it is impossible to unambiguously describe the sampling procedure employed in practice. Note in this connection that the sample analyzed could contain pedigrees sampled, i.e., ascertained, extended and/or censored, differently from one another. Thus, the factors causing the difference between the planned and employed procedures should be documented (if possible) separately for each sampled pedigree, because each is described by its specific probability (likelihood) defined on its specific sample sub-space.

Thus, because it is impossible to describe unambiguously the employed sampling procedure, the sampling models can only be formulated using the initially planned sampling design.

The obvious conclusion from what has been said is that we need to formulate not just one sampling model, but a reasonable set of models $\{S_H\}$, all formulated explicitly using the initially planned sampling design and, hopefully, covering the range of possible sampling procedures that could have occurred while collecting the pedigree sample analyzed. This means that each of the set of competing genetic models from $\{\hat{\theta}_i^{MP}\}$ should be compared with one another when the sample likelihood explicitly includes each sampling model from $\{S_H\}$. This increases the number of MP models that should be compared, but this increase is quite justified when we take into account the more complete coverage it affords of the unknown, but actually employed, sampling procedure. In this case, the first-rank model that provides the most correct model of the trait inheritance simultaneously indicates the most parsimonious sampling model.

8.6. Statistical equivalence of the models compared

The problems of pedigree analysis considered up to now were mostly theoretical. Note carefully that the construction of sampling-corrected pedigree likelihoods (chapters 6 and 7) was made in a quite general form. No explicit specification of the ascertainment or inclusion probabilities was used, only general formulas were considered. Accordingly, our theoretical studies have examined the *general* possibility and method of constructing pedigree likelihoods corrected for the sampling procedures.

At the same time, when performing pedigree analysis, it is not enough to prove that it is in principle possible to construct a pedigree likelihood that accurately corrects for the sampling procedures. There are a number of usually unknown social-demographic factors in each population that can substantially modify how we use the theoretical results. In each particular population, for a particular initially planned sampling design, we must take into account the socio-demographic factors affecting the exact realization of the planned sampling procedures, and so we should consider in more detail how we can in practice distinguish the genetic models in the set $\{\hat{\theta}_i^{MP}\}$.

The definition of pedigree analysis (1.2) implies two conditions for being able to compare models. First, for any two genetically nonequivalent models of trait inheritance there must always exist specific pedigree data (with respect to both structure and phenotypic content) with which these models can be distinguished. Second, the operator Ω introduced above must provide an absolutely correct ranking of the set of models of trait inheritance. If the necessary data are included in the sample analyzed, then Ω necessarily distinguishes two genetically different models, i.e., one of them would be specified as preferable to the other.

In practice, these two conditions will not necessarily be satisfied. First of all, the distinguishing data could be unobtainable. The sample space, i.e., the set of pedigrees that can in principle be sampled from the true pedigree set $\{\tau\}$ defined above is constructed without any connection to the model set θ and, therefore, the sample space formed does not necessarily contain the whole conceivable range of pedigree data. If this sample space does not contain such distinguishing pedigree data, then the genetically different models would not be distinguished (even

asymptotically). Second, even if such distinguishing data can be sampled, the power of the statistical test used for the model comparison (the statistical version of the operator **Ω)** could be insufficient to make an unambiguous decision in favor of one of the two models being compared.

This means that different genetic models of trait inheritance could be considered as *statistically equivalent,* being indistinguishable even asymptotically, on the given particular sample space using the given statistical technique. In practice, the solvability of pedigree analysis, i.e. the ranking of genetic models one of which is to be chosen as the best descriptor of the trait inheritance, is determined by both the sampled pedigrees analyzed and the statistical test used.

As usual, it is in practice impossible to list the conditions necessary to distinguish all models of the preformed set θ in a particular study. Except for a few very trivial studies, the practical solvability of pedigree analysis is considered in the same way as for any other analysis, mostly *a posteriori*.

9. ON APPROXIMATE SAMPLING CORRECTIONS

9.1. Once more about the genotype-phenotype correspondence

Among the three component distributions defining the genetic model (section 1.5) the most uncertain is the one establishing how the set of genotypes in pedigree members display their phenotypes. This genotype-phenotype correspondence, $f(X_n,C_n|G_n)$, was introduced phenomenologically simply because our current knowledge about the ontogenetic process determining any trait that we may study is such that it cannot be formulated with any certainty on the basis of a genetic mechanism. This means that the distribution $f(X_n,C_n|G_n)$ can be formulated only as an approximation and, as such, in different forms. In turn, this means that the genetic model definition (1.1) is also given simply as an approximation. In each particular formulation of the genetic model, the only condition that the distribution $f(X_n,C_n|G_n)$ must satisfy is its ability to describe the joint distribution of the trait residuals – the joint distribution of phenotypes in the pedigree members caused by all genetic and environmental factors other than the effects of the genotypes included in the genotypic set **G**.

From this point of view, we must re-formulate the genetic model given above in (1.1). It can only be satisfactorily defined if there is some definite preliminary knowledge about the distribution of the residuals. In other cases, in reality, not one but a set of genetic models should be introduced as versions of (1.1) that differ from one another in the particular formulation of the trait residual distributions.

9.2. Accurate and approximate formulations

Consider one more problem of practical importance in pedigree analysis. When noting the "inherent intractability" of the ascertainment problem in pedigree analysis, Vieland and Hodge (1995) recommended "to pursue robust *approximate* approaches that will provide numerically acceptable results". Indeed, the rather strict (and in practice hardly likely to be met) conditions needed to make the likelihoods formulated in the two previous sections adequate, point directly to a need of some "approximate" approaches for the sampling correction or, to be more accurate, the need of an approximate formulation of the pedigree likelihood in probability terms - provided, of course, the level of approximation is acceptable. However, we should note certain things about the term "approximate" in this context.

Usually, "accurate" formulation means the introduction of explicit probabilistic models for the pedigree ascertainment, extension and inclusion, with their parameters having a clear probabilistic interpretation, e.g., $\pi(x_i,\beta_i)$ is the probability that a potential proband who has the phenotype(s) x_i and additional characteristic(s) β_i actually becomes a proband. Contrary to that, an "approximate" formulation assumes the introduction of the ascertainment probability as a phenomenological function of some pedigree characteristics such as, for example, the number of affected members in the pedigree PSF (Stene, 1977, 1978). The distinction between such accurate and approximate formulations is not always clear-cut. On the one hand, most of the accurate formulations are made using some assumptions that are in principle untestable (the sampling procedure that is in practice employed is always much more complicated than that modeled and used in constructing the likelihood). On the other hand, it is sometimes possible to introduce a clearly

formulated probabilistic model for some phenomenological formulations (Ginsburg and Axenovich, 1992).

Strictly speaking, any (either accurate or approximate) model for the pedigree ascertainment, extension or inclusion used in the likelihood formulation represents only an approximate description of the real sampling procedure employed, just because it is only a model. This means that 1) it is hardly probable that any initially defined sampling procedure will be strictly followed in practice, and 2) the procedure S_E employed in practice is expected to be (much) more complicated than the testable model (hypothesis) S_H that is formulated and used to construct the pedigree likelihood.

It follows from the above consideration that we should distinguish between at least two types of approximation that occur in pedigree analysis. The first, in which we do not accurately identify the pedigree subset responsible for one or another sampling sub-procedure, and the second, in which this subset is accurately identified but formulation of the corresponding probability (ascertainment and/or inclusion) is made only approximately, with insufficient correspondence to the procedure that was actually employed.

The sampling models (hypotheses) $\{S_H\}$ that can be formulated using the initially planned sampling design differ from the particular procedure S_E that was actually employed when the pedigree was collected and included in the analyzed sample. The difference is caused by (usually unknown) factors specific to the particular population under study. Different SMB and SMF likelihoods use different ascertainment models, but both require an accurate and complete identification of the pedigree PSF. If this condition is fulfilled, adequacy of the likelihood formulation

(in particular, correct estimation of the trait inheritance parameters) depends, in turn, on the correctness of the extension and inclusion models, i.e., on their similarity to the sub-procedures actually employed. These different sampling models represent different formal approximations of the same real sampling procedure $S = (\alpha,\varepsilon,\psi)$. We showed that these formulations of the pedigree likelihood provide correct (consistent) analysis results whenever they adequately describe the most decisive factors of the sampling procedure used, namely, the PSF and C^a structures of the true pedigree and either the form (but not parameter values) of the ascertainment and inclusion probability (SMB), or the assumed independence of it from the particular ascertained PC (SMF). We noted that the last formulation (SMF) is possible only when the phenotypes of all the pedigree members involved in the PSF are known. It is evident that other formulations are possible, including approximate ones. Some of them have been already proposed and become widely accepted and, therefore, deserve closer inspection. Most approximations proposed up to now have been connected with the classical π-scheme of ascertainment applied to a binary trait. Below, we consider some of them.

9.3. At least one proband

Haldane (1938), Bailey (1951), Morton (1959), Elston and Sobel (1979) and many others defined the ascertainment probability as the probability that *at least one* of the PSF members *becomes a proband*: $P(asc \mid X_1 C_1, \tau_p, \alpha) = 1 - \Pr(\overline{\tau_p} \mid \alpha)$. This means that the ascertainment probability is modeled the same for all pedigrees sampled from the same true pedigree, regardless of the particular PC that caused the ascertainment. This formulation can be considered as an approximation of

the proband ascertainment probability to be used when the actual pedigree probands have not been registered, but the pedigree PSF *has been*. This formulation of the ascertainment probability is not inconsistent with the usual binomial version of the multiplicative form of ascertainment probability (6.3), but it leads to a decrease in the pedigree information. Shute and Ewens (1988a) reported that this formulation of the ascertainment probability is "far less effective" (up to 50 times so!) for parameter estimation than the multiplicative form (6.3).

Note that the authors cited did not consider the important cases where the pedigree PSF, or at least its structure, is not known. In this case, the probability of no proband in a pedigree PSF, $\Pr(\overline{\tau_p} \mid \alpha)$, cannot be accurately found.

9.4. Single ascertainment

In the classical binomial ascertainment scheme, the special case of *single ascertainment* was distinguished as occurring when the probability π of any potential proband actualizing his/her potential is negligibly small. In this case, the binomial formulation of the ascertainment probability considered in section 6.3 can be approximately replaced in the likelihood expression by a very simple factor – the number of pedigree probands. This approximation decreases the number of parameters that need to be estimated from the given pedigree sample. Sometimes, this procedure can even increase the power of the tests used, but it is clearly only possible if π really is negligible.

Hodge and Vieland (1996) introduced a "generalized single ascertainment … not just through a single proband, but through only one *type* of proband configuration", a subset of relatives who become probands

and are ascertained simultaneously as one ascertainment unit, e.g., an affected sib pair, an affected parent-child pair etc. They defined *single ascertainment*, without a direct reference to the value of π, as that having the probability of finding more than one proband configuration (one proband if the configuration includes only one person) per pedigree exactly equal to zero. Because this ascertainment form shows excellent properties, the authors argued that this definition "corresponds both to single selection under the classical π-model (in the limit as $\pi \to 0$) and to the case considered by Cannings and Thompson (1977)". (These authors used the term "selection" instead of "ascertainment" as used here). This means that (6.3) is redefined as $P(asc \mid X_1, C_1, \alpha)$ regardless of the number, relationships and phenotypes of the pedigree PSF members; in other words, regardless of the structure and phenotypic content of either $(X_p, C_p) \setminus (X_1, C_1)$ or $\tau_p \setminus (X_p, C_p)$. In the sampling context, this expression can be interpreted, in particular, as a total suppression of the potentials of all other τ_p members after those in the configuration C_1 have somehow become actual probands.

Using this ascertainment probability, the pedigree likelihood for single ascertainment can be simplified to:

$$P(X, C \mid smpl, \theta, \varepsilon, \alpha) = \frac{P(X, C \mid \theta, \varepsilon)\Pr(X_1, C_1 \mid \alpha)}{\psi(\theta, \alpha)}, \qquad (9.1)$$

where $\psi(\theta, \alpha) = \sum_{(X,C)} P(X, C \mid \theta, \varepsilon)\Pr(X_1, C_1 \mid \alpha)$ is the population probability of ascertaining the pre-established proband configuration (individual, pair of siblings, parent-child pair, etc). As we can see, the pedigree likelihood in the form (9.1) is quite calculable if the above

conditions C_0, C_1 and C_3 are fulfilled and condition C_4 is differently formulated, namely as C'_4: the population distribution of the given proband configuration should be known.

Note that it is only for the particular π-model of ascertainment, in which all $\pi_i \equiv \pi$, that the denominator in (9.1) is proportional to the population prevalence of the proband configuration, $\psi(\theta,\alpha) \propto \psi(\theta)$, i.e., only for this particular case is the "fundamental definition" of Hodge and Vieland (1996) of single ascertainment true. If, as generalized by Elston and Sobel (1979), this probability of becoming a proband differs for different members of τ_p, in particular if it depends on the individual's specific phenotype and/or auxiliary characteristics, $\pi_i = \pi(x_i,\beta_i)$, then the proportionality $\psi(\theta,\alpha) \propto \psi(\theta)$ does not hold - even when all the other conditions for "single" ascertainment are satisfied.

We can discuss the adequacy of such a single ascertainment model as follows. If the ascertainment sub-procedure used is similar to that modeled, i.e., if there really has been the suppression of the potentials of all the other τ_p members, then likelihood (9.1) provides a correct (consistent) estimator of the inheritance model. However, such a suppression seems hardly likely to occur in practice unless special precautions are taken when collecting the pedigree sample (Elston, 1995). That is why, in terms of the proband ascertainment scheme, it seems more justifiable to consider single ascertainment (formulated with any definition of the sampling configuration) in the traditional way, i.e., as an approximation of the ascertainment probability when π (or π_i) approaches zero ($1 - \pi \approx 1$), i.e., when the potential probands actualize their potentials

very rarely. This assumption can, at least, be confirmed or refuted on the basis of a preliminary population study.

9.5. Phenomenological formulation

For a binary trait, a corollary result from classical single ascertainment can be presented by the approximate proportionality: $P(asc \mid X_1 C_1, \tau_p, \alpha) \propto_\theta r$, where r is the number of affected members in the pedigree PSF, the proportionality coefficient being independent of the trait model. This approximation was generalized by Stene (1977, 1978) in the form:

$$P(asc \mid X_1, C_1, \tau_p, \alpha) \propto_\theta br^c.$$

In this representation, $c = 1$ corresponds to classical single ascertainment; $c = 0$ to *complete* ascertainment, when the pedigrees having at least one proband are ascertained with probability 1, and values of c outside the [0,1] interval are also permitted. For example, $c = 2$, the "quadratic" ascertainment case, has been considered by Haldane (1938), Elston and Bonney (1984) and Ewens and Shute (1986). Ginsburg and Axenovich (1992) proposed a quite simple probabilistic scheme explaining this approximation in proband ascertainment terms: this definition of the ascertainment probability corresponds to the case where the probability π that the affected individual becomes a proband is different for different sizes of τ_p.

This phenomenological approximation can be generalized to the cases where the probability of becoming a proband, $\pi_i = \pi(x_i,\beta_i)$, is different for different PSF members. However, it can be used to construct a calculable pedigree likelihood only under the same conditions as the SMB likelihood formulated above: the total number of affected PSF

members, r, or the set of characteristics $\{x_i,\beta_i\}$ in all the τ_p members, should be known.

9.6. Adequacy of the approximate proposition

It seems hardly possible to formulate a theory or, at least, a list of reasonable recommendations of how to construct an approximate description of pedigree sampling when either the pedigree subsets responsible for the sampling procedures employed cannot be accurately identified, or the probabilities of pedigree ascertainment and inclusion cannot be accurately formulated, or both. Usually, some particular forms of such approximations are proposed simply because their authors consider the results of their analysis as promising and sufficiently accurate, which, from their point of view, apparently justifies the proposition.

Consider, for example, the approximate ascertainment correction for complex pedigrees proposed by Bonney (1998). His proposition was based on a number of practically untestable assumptions and can be schematically outlined as follows. Contrary to what we have said above about the ascertainment, extension and inclusion procedures (section 2.3), Bonney assumed that a sampled complex pedigree can be represented as a union of "family units" – nuclear pedigrees that are ascertained *independently* of one another. For each unit, the ascertainment correction is made *separately,* using the approximation "at least one proband" and assuming that each (affected) member of the unit can potentially be a proband. The likelihood correction for the complex pedigree is made by *multiplying* the correction coefficients of each constituent nuclear family. Unfortunately, Bonney did not present either proof of adequacy of this approximation or any illustration of it.

Let us stress once more that if the formulated sampling model does reflect some important features of the actually performed ascertainment procedure, then it is possible to expect at least approximate adequacy of the ascertainment correction. However, if this is not so, i.e., the degree of similarity between the really performed procedure S_E and its model S_H (as in the case proposed by Bonney) is unknown, the analysis result simply cannot be reasonably interpreted. It can be shown in addition that Bonney's recommendations for the ascertainment correction could lead to rather doubtful results in some cases ("uniform proband status ascertainment correction" in his terms).

Summing up the above examples of constructing an "approximate" ascertainment (and inclusion) correction, and taking into account the necessarily approximate character of any SMB likelihood - even if it appears to be "not very approximate" because it is based on an explicitly formulated ascertainment model - the following should be noted. The inherent approximate character of practically every ascertainment (and extension and inclusion) model should be taken into account in the likelihood formulation. This can be done by giving either a set of conditions (explicitly formulated and practically testable), under which the particular approximation is justified, or a specific algorithm that provides robust analysis results (see below). Up to now, the first approach is quite undeveloped. Bonney's (1998) proposition has been made, but it was not accompanied by any description of the conditions under which this proposition would provide adequate analysis results

9.7. On robust algorithms

In robust algorithms for pedigree analysis, we would consider the analysis results conclusive if the genetic model of the trait inheritance that

is accepted is the same (at least, in its main details) for different models of the unknown (or insufficiently known) sampling procedure. It is very doubtful that a more or less complete list of these different hypothetical models $\{S_H\}$, which should provide robustness of the analysis results, could be given in each particular situation. However, this is the very basis for constructing robust algorithms. Incidentally, the following heuristic algorithm for making a genetic decision about the trait inheritance model has been proposed, widely used, but not as yet studied systematically: if the accepted model of trait inheritance is the same, up to sufficient detail (e.g., not including estimates of only some parameters), for both single ascertainment ($\pi \to 0$) and complete ascertainment ($\pi = 1$), then this model is considered reliable. Later on it turned out that these two special ascertainment models do not cover the range of possible ascertainment procedures (Stene, 1977, 1978), but the very idea of how to construct a robust algorithm had been proposed. This idea has still not been sufficiently investigated. We mentioned above that some problems of forming the set of hypothetical sampling models $\{S_H\}$ are not yet solved, and as yet there is no proposal of how to solve them, even in relatively simple situations. Also still unsolved are the problems connected with the statistical properties of this decision procedure about the genetic model. For example, it is unknown how to find the probabilities of type I and II errors of this decision.

9.8. Sample space and likelihood formulation

Although justified theoretically, and quite alluring from an application point of view, the idea "to pursue robust *approximate* approaches that will provide numerically acceptable results" does not

appear to be easy to do and, obviously, needs some additional study to define the very principles of how to construct such approaches.

To describe accurately the pedigree distribution in probabilistic terms, i.e., to define the accurate pedigree likelihood on the given sample space, we need quite detailed information about the sampling sub-procedures defined on the set of true pedigrees. However, in practice this could be unavailable. We considered above some widely accepted methods of approximate likelihood formulation. Most of them were proposed many years ago, when the inheritance of mostly qualitative (binary) traits was under intensive study. It was shown that each particular approximation has its own area of use, consistent with the theory developed above. However, in practice an accurate description of this area is not always possible.

It is important to formulate clearly exactly what it means to say that these methods are approximate. Until pedigree analysis explicitly uses a pedigree likelihood, any approximation is formally expressed in a different but quite unique manner of defining a probabilistic measure (pedigree likelihood) on the given sample space. This space may not be completely defined, or it could be biased because of the limitations imposed by the sampling design on the possibility of its strict fulfillment in the practical process of collecting the pedigree data; and because of the technical (social, demographic etc) availability of the information needed about the individuals' relationships, etc. However, for each specific problem the sample space is either determined and, therefore, it is in principle possible to define on it a probabilistic measure (the pedigree likelihood), or the sample space cannot be defined unambiguously by the set of true pedigrees and the questionnaire used and so, therefore, no pedigree likelihood can be defined. All the accurate and approximate methods of

pedigree analysis are based on the probabilistic distribution of pedigrees that can be sampled from the given sample space (described in more or less detail, correctly or incorrectly). On each sample space, it is possible to define more than one probabilistic measure. Differently formulated likelihoods (SMB, SMF, approximate or not) produce different analysis results that are more effective the more complete the information that is available about the sample space.

10. MODEL - FREE PEDIGREE ANALYSIS

Up to now, we have considered methods of pedigree analysis that are explicitly based on constructing mathematical-genetic models of the trait inheritance and models of the sampling procedures, ascertainment, extension and inclusion. These models were explicitly formulated using parameters determined by the three basic distributions that define the genetic model (1.1): $p(g_1, g_2)$, $P(g \mid g_1, g_2)$, and $f(X_n \mid G_n)$.

As we have seen from the above considerations, to produce results that are reliable and, therefore, interpretable in terms of the genetic model of the trait inheritance, especially when the trait is multifactorial with complicated inheritance, the practical execution of such an analysis requires a highly informative pedigree sample, an adequate formulation of the tested models of trait inheritance and sampling procedures, and very complicated calculations that can only be performed by means of specially designed computer programs.

In their place, *model-free* methods of genetically studying such traits have been proposed and extensively exploited over the past several years. They are called model – free because they are constructed without explicit formulation of a genetic model for the trait inheritance. This kind of analysis is especially appropriate at the early stages of a genetic study of such traits as, for example, a disease susceptibility having a very complicated genetic control and, therefore, hardly lending itself to constructing a reasonably adequate genetic model for its inheritance. Initially, these methods dealt with rather simply structured pedigree data and used comparatively simple statistical tests. Their further development

has resulted in using all the sampled pedigree data, which substantially increases the power of the method.

Owing to the successful completion of the Human Genome Project, which resulted in being able to identify many DNA markers distributed along all the human chromosomes, the most widely used model–free methods are directed at testing linkage between marker loci and the trait being studied. Dependent transmission across generations of the marker genotype and the phenotype of the trait can be caused by the gene that takes part in the trait control being positioned on the same chromosome in the vicinity of the marker locus (or loci). Thus, linkage analysis itself changes its classical purpose. As was formulated by Rao (1998), linkage analysis of complex traits should first prove the very existence of a gene involved in the control of the trait being studied, and only then map it.

Below, we give a description of four types of model-free methods. Two of them are widely used and are described in some detail, while the other two are only briefly outlined.

10.1. The Haseman–Elston method

To demonstrate the existence of genetic control of a quantitative trait, Haseman and Elston (1972) proposed a method of establishing a linkage relationship between the trait and a marker locus, assuming that *environmental factors cannot simulate* the effect of genetic linkage.

Operationally, their proposed test of linkage between the quantitative trait being studied (to be more correct, between a locus putatively taking part in the control of that trait) and a marker locus was as follows. Consider a sample of nuclear pedigrees each having two offspring. Let x_{i1} and x_{i2} be the trait values measured (and adjusted for age,

sex and any other relevant covariates) in the offspring of the *i*-th pedigree, and $d_i^2 = (x_{i1} - x_{i2})^2$ be the squared difference between them.

First let us define the proportion π_i of marker alleles the sibs of the *i*-th pedigree have identical by descent (i.b.d.). Two relatives share an allele i.b.d. (as opposed to identical in state) at a particular locus if one of the alleles they each have there are copies of the same ancestral allele. Thus at any one locus sibling pairs can share 0, 1 or 2, i.e. 0, 0.5 or all, of their two alleles at a marker locus i.b.d. Sometimes we can observe this fraction, but at other times, for example if we cannot observe the marker genotypes of the parents, it must be estimated on the basis of the marker phenotypes available. Thus the proportion π_i is estimated by the estimator $\hat{\pi}_i$, defined as the probability, conditional on all the marker data available, that the *i*-th pair share 2 alleles i.b.d. plus half the probability they share 1 allele i.b.d. This is the expectation of π_i found by taking into account the population frequencies of the marker alleles and the marker genotype penetrances (it is possible to calculate this expectation even if some marker data on the pedigree members are missing). Haseman and Elston (1972) gave formulas for estimating π_i for various situations.

Consider the regression of the squared sib-pair trait differences d_i^2 on the i.b.d. estimator $\hat{\pi}_i$ and estimate its parameters on the given sample of nuclear families by the usual least mean square method, i.e. by minimizing

$$Q = \sum_i (d_i^2 - \alpha - \beta\hat{\pi}_i)^2 , \qquad (10.1)$$

where α and β are the regression coefficients to be estimated.

The null hypothesis tested is H_0: $\beta = 0$ against the alternative H_1: $\beta < 0$, which formulation is explained as follows. Haseman and Elston

showed that, for the special case of a diallelic marker locus, random mating, no dominance, no epistasis, no disequilibrium in the joint distribution of the trait and marker loci, and complete parental information, the expectation of the squared difference of the sibling traits conditional on the i.b.d. estimator is: $E(d_i^2 \mid \hat{\pi}_i) = \alpha + \beta\hat{\pi}_i$ with $\beta = -2(1-2\rho)^2\sigma_a^2$, where ρ is the recombination fraction between the trait and marker loci and σ_a^2 is the additive component of the trait locus genotypic variance. Thus, acceptance of H_0 means that $\rho = 0.5$ or $\sigma_a^2 = 0$, i.e., no linkage between the trait and marker loci is detected in the given pedigree sample. The rejection of H_0 suggests the existence of linkage (with corresponding type I and type II errors) and, therefore, the existence of genetic control of the quantitative trait being studied. Accepting H_1, it is possible, under certain conditions, to obtain the maximum likelihood estimates of both the effect of the gene taking part in the trait control and the recombination fraction between it and the marker locus.

Proposed in 1972, this method initially used mostly serological and biochemical chromosomal markers. Later on, stimulated by a sharp increase in available DNA markers, this method has been substantially developed, extending the main idea to not only sibships (more than one pair in a pedigree) but also to half-sibs (Risch, 1990 a, b), and then to any type of noninbred relative pair, which allows the use of information from extended pedigrees (Amos and Elston, 1989; Olson and Wijsman, 1993; Schaid et al, 2000); and to form i.b.d. matrices for multipoint linkage analysis (Markianos et al, 2001). These new versions, used under rather mild conditions, resulted in a substantial increase of power (see, for example Holmans, 2001). Elston et al. (2000) and Elston et al. (2005)

discussed the possibility of using this test for binary traits. It should be especially noted that this test is robust, providing comparatively reliable information about linkage without any assumptions about the model of trait inheritance, and with minimal assumptions about the sampling procedures (Allison et al., 2000; Schaid and Rowland, 2000).

Currently, this method has been further improved (Shete et. al. 2003) and is widely used, in particular in genome screening to find chromosomal segments having marker loci linked to the trait being studied.

10.2 Transmission disequilibrium test

The transmission disequilibrium test (TDT) was formulated by Spielman et al. (1993) to test linkage between a marker locus and the trait being studied, provided that there is linkage disequilibrium between the marker locus and the gene taking part in the control of the trait. Initially, the test was used for a binary trait (affected-unaffected). Operationally, the test was constructed as follows.

Suppose n pairs of spouses with at least one affected offspring are sampled from the population, and only one affected offspring of each parent pair is randomly selected. Let M and m be the two alleles of a diallelic marker locus. Then the $2n$ sampled parents can each transmit or not transmit a particular marker allele (M or m) to an affected offspring, as shown in Table 10.1, taken from Ewens and Spielman (1995).

In terms of the quantities given in this Table, the contingency statistic can be written:

$$T_A = 4n(b-c)^2 /[(2a+b+c)(b+c+2d)], \qquad (10.2)$$

whose distribution, if the hypothesis about no association between the marker locus and the trait studied is true, can be approximated under certain conditions by the central χ^2 with df = 1. Thus, the statistic T_A can be used to test the presence of an association between the trait and the marker locus. It should be especially stressed that, contrary to the case-control samples that could induce an association caused, for example, by population stratification, this family-control data does not depend on the population structure. However, if the population from which the affected offspring are sampled has arisen as a result of admixture, there must have been at least two generations of random mating, with no further admixture, before the sample is taken.

Table 10.1. Combinations of transmitted and nontransmitted marker alleles M and m among 2*n* parents of affected children (Ewens and Spielman, 1995)

	Nontransmitted alleles		
Transmitted alleles	M	m	Total
M	*a*	*b*	*a*+*b*
m	*c*	*d*	*c*+*d*
Total	*a*+*c*	*b*+*d*	*2n*

Using only the heterozygote parents, Mm, it is possible to construct the statistic to test the null hypothesis H_0: $\rho = 0.5$ against the alternative H_1: $\rho < 0.5$, where ρ is the recombination fraction, i.e., to test linkage between the marker locus and a gene taking part in the trait control. This TDT is given in the form:

$$T_L = (b-c)^2/(b+c) \qquad (10.3)$$

The genetic basis of these tests can be explained using the following simple ideas. Let the trait inheritance model be determined as follows. 1) The trait is controlled by one gene with two alleles, A_1 and A_2, having frequencies, p and $1 - p$, respectively. 2) The genotypic values are not constrained (any degree of dominance). 3) The diallelic marker locus is codominant, with allele frequency $P(M) = q$ and $P(m) = 1 - q$. The joint distribution of the alleles at the two loci is determined additionally by the disequilibrium parameter $D = P(A_1M_1) - pq$ and by the recombination fraction ρ. In this case, it is possible to prove that the overall association between the two loci (this is exactly what the data in Table 10.1 describe) is measured by the parameter $\delta = D(1-2\rho)$ [Spielman et al., 1993, 1994]. This means that the test for association (10.2) is correctly defined only if the two loci are linked, and the linkage test (10.3) only if there exists disequilibrium in the joint distribution of alleles at the trait and marker loci. Thus, the TDT can be used either to test association between transmission of the studied trait and marker alleles to the next generation under the condition that there is linkage ($\rho < 0.5$), or to test linkage between the trait and marker loci under the condition that their alleles are not distributed independently ($D \neq 0$). For the former test to be valid only one child per family can be included in the data; for the latter test, because the transmissions are independent under the null hypothesis of no linkage, it is permissible to include more than one affected child per family when constructing Table 10.1.

Because of the relative simplicity of collecting data of the type shown in Table 10.1, and because of the simple form of tests (10.2) and (10.3), this model-free method has been rapidly developed further by its authors and by other investigators. Schaid and Sommer (1994) showed that

the power of the TDT could be increased if the trait inheritance model is known. However, Ewens and Spielman (1995) noted that this increase is only modest and, if the true mode of trait inheritance is additive, the TDT is the most powerful test. Next, Cleves at al. (1997) extended the TDT test to the case of multiallelic markers. Then, the situation was considered where marker genotypes in the parents are missing, but marker information on unaffected siblings can be used to determine the transmitted and untransmitted marker alleles (Spielman and Ewens, 1998).

Without dwelling further upon very interesting developments of the TDT directed mostly to its use in non-standard situations, let us just note that Allison [1997] proposed the TDT for use with *quantitative traits* on the same family-based sample data (parents with at least one of them heterozygous at the marker locus and one of their offspring). The data analyzed in this case are represented by two sub-samples, one having the transmitted marker allele, and the other having the untransmitted marker allele.

Allison proposed the TDT for an *extreme-threshold* (ET) sampling design defined as follows. Let u and v be upper and lower thresholds such that trios having offspring trait values between these thresholds are excluded from the analysis. In the particular case $u = v$, all the trios are used. Allison showed that this ET design *increases the power* of the test but did not dwell in detail on how to choose the optimal thresholds. Clearly, they should be chosen differently for different trait distributions. For example, suppose the trait is under the control of a single major gene with a dominant allele effect, and the trait distribution is bimodal. Then the optimal values of u and v are on either side of the point of the trait density that is a local minimum, even if that point differs substantially from the

median. The trios included in the analyzed sub-samples are only those with offspring trait values belonging to the tails of the distribution. The power of the test is expected to increase as the proportion of excluded intermediate trait values increases, up to the point where this exclusion sufficiently reduces the number of trios in the analysis that the power starts to decrease. This limiting point is determined by the total sample size, as well as by the mode of inheritance of the trait being studied, and by details of the marker locus, such as its allele frequencies, its distance from the trait locus, and the level of linkage disequilibrium. This means that, for each particular trait-marker pair, there is a specific optimal proportion of trait values that should be excluded to provide the maximum power for the test.

Malkin et al. (2002), keeping intact the rationale behind the construction of the TDT, proposed an *extreme-offspring* (EO) sampling design that removes the problem of establishing this optimal proportion of excluded trios, and that substantially increases the power of the TDT. For each parent pair, to form the trio one selects the offspring having *the most extreme trait value* among those siblings whose trait values belong to the predefined tails of the distribution. Of course, there is no offspring selection for parent pairs who have only one offspring with a trait value outside the intermediate trait range – this single offspring is included in the trio. Thus, for each given proportion of excluded offspring, we have the same number of trios for both EO and ET testing.

Malkin et al. (2002) showed that the maximal effect of the ET design is smaller than the effect of the EO design. The difference depends on the number of offspring from which the one with extreme trait value is

selected and on the true mode of trait inheritance, including the trait heritability and the level of dominance.

It is important to stress that, as for the Haseman-Elston method, the TDT is also *robust*, providing information about linkage of the genes taking part in the trait control without any explicit formulation of the genetic model of inheritance, and using relatively easily collected pedigree data. Currently, it seems to be used mostly for screening markers that have been previously found to be linked with the trait in particular chromosomal regions (so-called "candidate genes"), but in the future it may be used for whole genome-wide association testing.

The requirement of statistical independence of the parent-offspring trios restricts the amount of information that can be used when the available sample contains extended pedigrees. Thus recent development of the TDT technique has been directed towards using the maximal amount of information that one can obtain from a sample of extended pedigrees. These attempts have been made for both binary traits (Martin et al, 2000) and quantitative traits (George et al, 1999, Abecasis et al, 2000).

10.3 Test of disequilibrium for pedigrees

Let us consider how the association/disequilibrium can be used to find a more accurate localization for a trait gene.

Consider the model of trait inheritance in which it is assumed that the trait genotypes on pedigree members are exactly their marker genotypes. Let $P(X_n, C_n \mid \mu)$ be the pedigree likelihood of this model when it is assumed that all the marker genotypes have the same genotypic value μ for the trait in each pedigree member. Further, let $P(X_n, C_n \mid \mu_g)$ be the same pedigree likelihood defined by the MG model for which the

three genotypic values μ_g that determine the trait control (g = 1, 2 and 3 for genotypes A_1A_1, A_1A_2 and A_2A_2, respectively) are distinguished.

Introduce the test in the form:

$$\text{LRT} = 2\ln[P(X_n, C_n \mid \hat{\mu}_g) / P(X_n, C_n \mid \bar{\mu})],$$

where $\hat{\mu}_g$ are the maximum likelihood estimates of the μ_g model parameters, and $\bar{\mu}$ is the parameter maximizing the null hypothesis model $P(X_n, C_n \mid \bar{\mu})$. We expect that, if this null hypothesis is true, the LRT is distributed asymptotically as χ^2 with 2 df (the difference between the numbers of estimated parameters in the two likelihoods). The null hypothesis is rejected if the LRT value exceeds the critical value c_α corresponding to the pre-established probability of type I error α. Any association found could be caused by different population effects such as either population stratification or a recently introduced trait gene mutation.

Let us assume that a linkage between the trait and marker loci was accepted in some previously performed analysis using, for example, the Haseman-Elston method. In this case, there is an interval on the chromosome in the neighborhood of the selected marker where some polymorphic markers demonstrate significant linkage with the putative trait gene. For the Haseman-Elston method, the length of this interval may easily be greater than 10cM. However, for the purpose of localizing a disease gene, a much narrower region is required. This can be found using a pedigree likelihood that includes marker loci that are in disequilibrium with the trait gene. Explicitly including in the pedigree likelihood a disequilibrium parameter that is not negligibly small substantially increases the power of linkage analysis. Recent evaluations show that in a stabilized population the disequilibrium parameter for a pair of loci has a

significant value only for distances less then 0.1cM. Then, using the above test for those marker loci that are in disequilibrium with the trait gene, the initially relatively wide chromosomal interval in which the trait gene demonstrates linkage with the set of marker loci can be substantially narrowed, providing a more precise evaluation of the trait gene position.

The test described in this section is very sensitive to disequilibrium, because it uses the whole pedigree data. This test can be used for fine scale gene mapping for distances where the level of linkage disequilibrium is significant. (Note, however, that in a stratified population an association can also be found in the absence of linkage).

10.4. Method of haplotype sharing

This method, named decay of haplotype sharing (DHS), was proposed by McPeek and Strahs (1999).

Suppose the trait mutation originated some generations ago. Define a set of tightly linked marker loci in the vicinity of the mutation. It is assumed that the mutation originated in a haplotype that can be constructed from alleles of this set of marker loci.

Genotyping multiple tightly linked markers enables one to use the pedigree structure to reconstruct sequential haplotypes in a given pedigree member (Stephens et al., 2001; Markianos et al. 2001) either exactly or, if different haplotype combinations are possible in the particular pedigree member, with corresponding probabilities, Because of the recombination process that has occurred across generations since the mutation originated, the haplotypes identified in pedigree members would be different from the initial haplotype in which the mutation originated and they would differ from one another; the difference is larger, the larger the number of generations since the mutation was introduced. From this, we might expect

to find shared fragments in pedigree members representing regions of the initial haplotype that were preserved during the successive generations, their lengths being smaller the more recombination that has occurred from generation to generation.

The location of the mutation among the marker loci is usually unknown. It is considered as a parameter and is evaluated simultaneously with the other two parameters, namely, the initial marker haplotype in which the mutation originated and the number of generations since that time. The estimation of these parameters is made by maximizing the likelihood, which is formed as the product of the likelihoods found separately for each pedigree member phenotypically displaying the mutation. The model of McPeek and Strahs (1999) assumes that the parts of a haplotype shared by these pedigree members are fragments of the initial haplotype in which the mutation originated, saved during the recombination process. The expected length of these fragments decrease exponentially with increase in the number of intervening generations.

Suppose the age of the mutation is infinite, i.e., it came into existence so long ago that the recombination process has destroyed any dependence in the distribution of alleles of the trait and marker loci. Rejection of this hypothesis means that the mutation is not too ancient, i.e., the marker alleles and the trait gene alleles (the mutation is one of them) are distributed dependently with one another. The dependence is measured by the mutation age parameter, which results in a mean (non-zero) length of shared haplotypes among pedigree members containing the mutation. This means that the trait gene and the marker loci are linked, which is exactly what is tested in this method.

10.5. Characteristics of the model-free methods

Linked loci are often (but certainly not always) in allelic disequilibrium with each other. Thus, two types of linkage tests should be distinguished, the test of linkage itself, the result of which depends on the distance between the trait and marker loci, and the fine-scale mapping test that explicitly uses the disequilibrium association between the two loci.

The Haseman-Elston algorithm (and, based on the same idea, variance-component linkage analysis, see Amos, 1994) is the linkage test. Its power depends on the recombination fraction between the trait and marker loci and on the additive component of genetic variance of the linked quantitative trait gene. Its dependence on any allelic disequilibrium between the trait and marker loci is not too strong, even if the loci are tightly linked, so this method is used for long distance genome scanning.

The association tests explicitly use the allelic disequilibrium. They test significance of the disequilibrium parameters and are more powerful than linkage tests when the trait and marker loci are extremely close together. For a stabilized population, this distance should usually be less than 1cM. With larger distances the disequilibrium coefficient tends to zero exponentially, together with the power of any association test. These tests are effective for examining candidate genes and also for fine scale genome mapping.

We have mentioned here three types of association/disequilibrium tests. The TDT has a specific design that allows us to avoid the effects of population stratification. Only disequilibrium in offspring of heterozygous marker individuals is tested, which is always equal to zero if the loci are not linked.

The pedigree disequilibrium test is applicable to the whole sample and has significantly greater power. But unless a TDT type TSgt is performed (see e.g. George et al., 1999) a significant result can then indicate not just linkage disequilibrium but also the effects of population stratification, which is only formally expressed as nonzero *D*. For mapping the trait gene, this test is effective for distances of about 0.1cM or less, for which the linkage disequilibrium should have a significant value.

The DHS method is based on another disequilibrium measure, which is applicable only to a chromosomal interval containing tightly linked marker loci. This is the average length of an ancestral haplotype that all the affected offspring have inherited together with the mutant allele. This length is determined by the number of generations of recombination since the mutation originated. The method deals with the "finest scale" mapping of the trait gene investigated because, among the set of tightly linked markers, often those are chosen that are located inside the limits of the trait gene (see section 1.3).

The tests considered here represent only examples of the model-free technique. Because of intensive current development of this technique, other new methods can be expected to be proposed in the near future. They use different statistical techniques and test different manifestations associated with linkage. Their use is justified by various assumptions about the history of the trait and marker loci, the distance between them and their joint allele distribution.

10.6 Limitations of model-based linkage results

Concerning these model-free methods of pedigree analysis, it is widely accepted that, being robust, they are to be used to provide only preliminary information about the linkage between the genes taking part in

the control of a trait under study and the chromosomal markers (DNA or candidate genes). Their use is assumed to be especially appropriate for a disease susceptibility having complicated genetic and environment multifactorial control and, therefore, hardly lending itself to constructing a reasonably simple genetic model for its inheritance. Initially, these methods dealt with rather simply structured pedigree data, while currently they use the whole sampled pedigree. However, technically (in particular, statistically) they are easier to perform than the model-based methods.

If the model-free analysis results in rejection of the hypothesis of no linkage, this result is considered as preliminary, justifying all further efforts to collect a sufficiently informative pedigree sample, and analyzing it explicitly formulating the inheritance and sampling models, *adequately* reflecting at least their most important features, using appropriate program packages. In other words, the linkage test used in model-based pedigree analysis is expected to be (much) more powerful than that used in model-free methods. Goldin and Weeks (1993) and Greenberg et al (1996) attempted to justify this expectation in detail, considering different arguments for and against it.

However, these arguments are not unambiguously convincing. Any model-based pedigree analysis result, i.e., the genetic model estimated on the given pedigree sample as the best descriptor of the trait inheritance, does not always make sense. If this model reflects at least some important features of the inheritance of the trait under study, which can hardly be guaranteed in practice, then its acceptance, interpretation and further application are justified. However, if the trait is controlled by genes and environmental factors in so complicated a fashion that it is simply impossible to construct a reasonable genetic model for its inheritance (in

particular, with a reasonable number of parameters that are to be estimated on a pedigree sample of practically achievable size), then even the best among all tested models could be absurd (see Beaty, 1997).

Any *a priori* set of genetic models θ or, in its reduced version, $\{\hat{\theta}_i^{MP}\}$, is limited by both their content and their number. Formulation of the genetic models is limited by the possibilities of estimating their parameters from finite-size samples. It is inconceivable that one can formulate such a general model that it can describe the inheritance of any complicatedly controlled multifactorial trait. If the genetic model tested is substantially simpler than the true inheritance of the trait being studied, then the model estimator would be biased (Sawyer, 1990). It is usually impossible to evaluate this bias and, therefore, to give any reasonable interpretation of the analysis results - regardless of what they are, e.g., the tested linkage is accepted or rejected - because in the general case the false positive and false negative results are indistinguishable. The proportion of such false results can be, and has been, studied only for the cases of a trait with relatively simple inheritance, described by a similarly simply formulated genetic model. In turn, this means that subjecting any trait to pedigree analysis, i.e., trying to describe its inheritance in terms of relatively simply formulated genetic models, should hardly be considered justifiable, and any interpretation of its results would definitely be ambiguous. This is especially true for traits that are at the initial stages of their study. Later on, at other stages, we might learn that the complicatedly controlled trait can be dissected and presented by its “component” traits, each having simpler inheritance that can be reasonably modeled and, therefore, subjected to pedigree analysis. One of the methods is to use

auxiliary traits that biologically correlate with the trait being studied (Ott, 1995).

10.7. Genetic dissection of multifactorial traits

Regarding model-free pedigree analysis, up to now we have placed no limitations on the complexity of the inheritance of the trait. This problem has not even been formulated. This justifies the use of model-free pedigree analysis for any trait, including those controlled in a complicated multifactorial manner, until these limitations are found and explicitly formulated. The results obtained, e.g., finding linkage with a certain chromosomal marker, could be used in further study of the trait inheritance. Lander and Kruglyak (1995), Rao (1998) and many other authors, have discussed methods of *dissecting* complicatedly inherited multifactorial traits using genomic scans, where model-free analysis is used to establish the existence of the genes taking part in the control of the trait, and then to localize them in chromosomal segments with the accuracy that the method can provide.

Consider now what the term “dissecting” means. The most that can be obtained using the model-free technique in a genome scan is a (relatively) complete list of the genes taking part in the trait control, through rejection of null hypotheses of no linkage. The method is not intended to localize these genes by estimating the recombination fraction. Further details of the chromosomal positions of these genes can be obtained only through confirming such testing of the null hypothesis with the newer more dense sets of DNA markers (SNPs, RFLPs, STRPs etc.) located in small genomic fragments. Thus, the dissection of multifactorial traits with model-free linkage analysis have been made possible by the

molecular technique of identifying marker loci that can statistically establish the genes involved in the trait control, using the pedigree sample.

10.8. Phenotypic dissection of multifactorial traits

We have already mentioned several times the limitations of genetic modeling (it is intended for traits inherited in a relatively simple fashion) to investigate a biological function that is complicatedly controlled, both genetically and environmentally, and that we can replace the initially considered multifactorial trait by another phenotypic description of the function to be studied. It has been proposed that the function described by a complex trait be replaced by "component" traits, formed by using, for example, biological markers - intermediate phenotypes - as discussed by Ott (1995).

Let us assume that the component traits of the new complex phenotypic description of the biological function under study are such that:

1) these component traits describe the same function that is characterized phenotypically by the multifactorial trait;

2) they are controlled in a relatively simple manner by genetic and environmental factors, regardless of how they are defined phenotypically (qualitative or quantitative);

3) the genes taking part in the control of each component trait (it is assumed that these genes can be identified because of the assumed simplicity of their genetic control) are the same (or almost the same) as were found in a model-free analysis to be taking part in the multifactorial trait control.

In this case, the studied function initially described by a multifactorial binary trait can be described by a complex trait. This replacement of the description of the function would be adequate, if we

can assume that the components of this complex trait together describe the same (or almost the same) characteristics of the function under study, but expressed as different phenotypes.

In this new description, the problem of completeness arises. The component traits describe the biological function with varying degrees of completeness for the different multifactorial traits that might be under study, and for different sets of available biological markers often used as prognostic markers. This new description of the function under study could be considered to be a *phenotypic dissection* of the initially considered multifactorial trait.

If, as we assume, the inheritance of each of the component traits is described by a genetic model that can be adequately estimated and tested using a pedigree sample, then, instead of the model of inheritance of the multifactorial phenotype, which in practice it is hardly possible to construct, we can construct several comparatively simple models constituting together a *compound* genetic model for this complex trait, and then use them in various applications. Further efforts in this direction would increase the adequacy of such a compound model, taking into account, for example, pleiotropic effects of some genes on several component traits and the joint environmental effect on the phenotypic distributions arising from these genes. This seems to be a quite natural step-by-step way of studying the inheritance of a complicatedly controlled trait. The result expected from this study is the mathematical-genetic model of the complex trait inheritance, which, if achieved with satisfactory completeness, can be used in application problems. Ideally, this description of the inheritance of the biological function would be constructed with such completeness as to provide adequate and accurate

prediction of this function in each individual whose genetic and physiological characteristics have been determined in previously performed tests.

CONCLUSION

Thus, pedigree analysis is defined as a method to formulate the mathematical-genetic description of the inheritance of a particular biological function using sampled pedigree data. The sample of pedigrees represents the basis on which the pedigree analysis is performed, i.e., on which the genetic models of the studied function are formulated and statistically tested. The goal of the analysis is to construct the model that provides the most accurate and maximally complete description of the inheritance of the function. If the pedigree analysis results in such a model, the latter can be used in various applications. In particular, performing some previously established set of special tests (clinical, biochemical, molecular), it may be possible to identify an individual's genotype, the dependence of the phenotypic distribution on the given range of environmental conditions, etc. These tests would provide the detailed information about the genetic specificity of the individual that allows us to predict his/her physiological reaction to a particular medical treatment. This, clearly, is the ultimate goal of any study performed on the inheritance of a biological function.

We simplified our consideration of the main operations of pedigree analysis, namely, the formulation of the mathematical-genetic model, the probabilistic characterization of the analyzed pedigree sample (sample likelihood), and the formation of a set of models to be tested, mostly by using the MG model. The use of this model can be explained as follows.

First of all, there are probably traits whose phenotypic distribution is determined by genotypes of only one diallelic gene.

Second, the possibility of finding a reasonably accurate estimator of a genetic model from a pedigree sample is determined directly by a correspondence between the model complexity - in particular, the number of parameters that need to be estimated - and the information available in the sample analyzed. Because the latter is always limited, mostly MG models are formulated, statistically tested and used in order to have an (approximate) description of the trait inheritance.

However, the main reason for using the MG model was to remove the need to give the details of formulating the various possible genetic models, which are *irrelevant* once the *general theoretical* results have been proved for the various aspects of pedigree analysis.

Indeed, the formulation and parameterization of the mathematical-genetic model for different traits characterizing the biological function studied (qualitative, binary, or quantitative - discrete or continuous) are considered without any direct specification of the genotypic set **G.** All these considerations hold true for more complicated multifactorial models as well, were such models to be formulated. At the same time, the specificity of different model formulations for different characterizations of the biological function is more clearly described using the MG model because it is then not encumbered by a need to consider irrelevant effects.

A similar argument holds when the second basic problem is considered, namely, the pedigree likelihood correction for the sampling procedures used. As we saw from the considerations in chapters 6 and 7, no direct connection was specified between the corrected forms of the pedigree likelihood (SMB or SMF) and the structure of the genotype set, **G**. The likelihood corrections hold for any **G** and, therefore, it is quite natural to consider them for the simplest MG model.

The same is true for forming the set of most parsimonious models $\{\theta_i^{MP}\}$ - it is clear that the structure of **G** has no connection with this process.

Thus, it is quite justifiable to use the simple MG models to analyze how the genetic-mathematical model of trait inheritance, and the sampling procedures determining the process of collecting pedigrees, are formulated and parameterized.

At the same time, the following should be noted in connection with the definition, given in section 1.5, of a gene involved in controlling the trait inheritance. Wolf (1995) considered the theoretical possibility of an unambiguous genotype-phenotype correspondence and illustrated his conclusions with an example. Referring directly to the biological system of the ontogenesis process, he pointed out that "genes are a necessary but not sufficient component of it. The structures already present, gradients, threshold values, position relationships, and conditions of the internal milieu, are equally essential. ... even monofactorial traits can be considered to be of multifactorial causation" (ibid, p. 127). Thus, we should not expect each trait mutation to have a consistent phenotypic outcome and, therefore, the genotype-phenotype relationship may be irregular.

Dipple and McCabe (2000) further reviewed some attempts of empirical methods to establish the genotype-phenotype relationship and showed that they are still far from being able to produce unambiguous results. Studying relatively simple Mendelian disorders, they showed that many of them were found to be multifactorial traits, in the sense that the ontogenetic process resulting in the phenotypic manifestation of each mutation constituting an individual genotype is too complicated to be unambiguously expressed in some direct linear genotype-phenotype

correspondence. The ontogenetic process was in many cases found to be multifactorial, in the sense that the phenotypic distribution of the genotype is influenced by many genetic, epigenetic and environmental factors, including not only the specific mutation studied, but also the effects of other genes involved in the biochemical realization of the trait, the effects of other genes having maybe no direct connection with the trait studied (modifiers), and the effects of very complicated and not easily recognized environmental conditions of the ontogenetic process. In other words, up to now, attempts to formally establish the genotype-phenotype correspondence explicitly using all the ontogenetic mechanisms involved in the genotype manifestation (or, at least, some of them) are still hardly possible.

In our considerations, we used only the operational definition (see 1.5) of the genes taking part in the trait control. The multifactorial control discussed by Wolf (1995) and Dipple and McCabe (2000) means that the ontogenetic process involves many genes as necessary links in the biochemical chain that results in the observed phenotype. However, only those allele substitutions that lead to phenotypic change (established by the trait characterization employed) form the set of genotypes, **G**, that determines the genetic model.

Consider now a connection between the two versions of pedigree analysis, model-free and model-based. If, using the former, a set of genes is established that take part in the trait control, it is possible, in principle, to formulate a model-based genetic model that describes the inheritance of the trait explicitly, thus formalizing this set. The problem to be solved in this case is *purely technical,* whether or not it is possible to collect an informative pedigree sample on which the formulated multifactorial model

of trait inheritance could be analyzed, i.e., its parameters estimated and statistical tests of its fit performed. It may be that further development of pedigree analysis methods will provide such a possibility.

However, up to now, this possibility can be considered only in *principle*. In the current practice of pedigree analysis, using pedigree samples of practically achievable size, it is hardly conceivable to formulate and statistically analyze such multifactorial models of trait inheritance. Up to now, the only common thread that we can establish between the model-based and model-free methods is that they are both just pedigree analyses, which means that the basis on which the analysis results are obtained in both cases is represented by pedigree samples. Moreover, we still do not know how to combine the qualitatively different results obtained by these two methods of analysis. For example, we do not know whether it is possible, and if so how, to use the information about genes established as taking part in the trait control to obtain a more accurate description of the trait inheritance, which is the main goal of genetic model construction.

To complete our discussion of pedigree analysis, let us stress once more that it was not our intention to write a manual describing the various practical methods of performing the analysis, given the particular trait that characterizes the biological function to be studied, the size and structure of the pedigree sample on which the analysis is to be performed, and the pre-established sampling design (this last determining the sampling procedures that are to be formulated and used to correct the sample likelihood). At the same time, this has not been a strictly consistent development of the mathematical theory of pedigree analysis. The current development of pedigree analysis is still far from affording us such a possibility, notwithstanding the fact that many aspects of the analysis can now be

formalized and the problems explicitly solved using the mathematical apparatus available.

We have only considered basic concepts and made theoretical statements that provide the correct formulation of, and solution to, pedigree analysis and, what we regard as especially important, we have discussed in detail the conditions under which these concepts and statements hold true.

REFERENCES

Abecasis GR, Cardon LR and Cookson CWO (2000) A General Test of Association for Quantitative Traits in Nuclear Families. Am J Hum Genet 66: 279-292.

Abecasis GR, Cookson CWO and Cardon LR (2000) Pedigree tests of transmission disequilibrium. Eur J Hum Genet 8: 545-51.

Allison DB (1997) Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 60: 676-690.

Allison DB, Fernandez JR, Heo M, Beasley TM (2000) Testing the robustness of the new Haseman-Elston quantitative-trait loci-mapping procedure. Am J Hum Genet 67: 249-252.

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535-543.

Amos CI, Elston RC (1989) Robust methods for detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol 6: 349 -360.

Bailey NTJ (1951) The estimation of the frequencies of recessives with incomplete multiple selection. Ann Eugenics 16: 215-222.

Beaty TH (1995) Evolving methods in genetic epidemiology. I. Analysis of genetic and environmental factors in family studies. Epidemiologic Reviews 19: 14-23.

Bonne B, Ashbel S, Modai M, Godber MJ, Mourant AE, Tills D, Woodhead BG (1970) The Habbanite isolate. I. Genetic markers in the blood. Human Heredity 20: 609-622.

Bonney GE (1984) On the statistical determination of major gene mechanisms in continuous human traits: regressive models. Am J Med Genet 18: 731-749.

Bonney GE (1988) Regressive logistic models for familial disease and other binary traits. Biometrics 42: 611-625.

Bonney GE (1998) Ascertainment correction based on smaller family units. Am J Hum Genet 63: 1202-1215.

Bonney GE, Lathrop GM, Lalouet J-M (1988) Combined linkage and segregation analysis using regressive models. Am J Hum Genet 43: 029-037.

Borecki IB, Bonney GE, Rice T, Bouchard C, Rao DC (1993) Influence of genotype-dependent effects of covariates on the outcome of segregation analysis of the body mass index. Am J Hum Genet 53: 676-87.

Box GEP and Cox DR (1964) An analysis of transformations. J Roy Stat Soc [B] 26: 211-252.

Burton PR Palmer LJ, Jacobs J, Keen KJ, Olson JM, Elston RC (2001) Ascertainmemnt adjustment: where does it take us? Am J Hum Genet 67: 1505-1514.

Cannings C, Thompson EA (1977) Ascertainment in sequential sampling of pedigrees. Clin Genet 12: 208-212.

Cheng LS, Livshits G, Carmelli D, Wahrendorf J, Brunner D (1998) Segregation analysis reveals a major gene effect controlling systolic blood pressure and BMI in an Israeli population. Hum Biol 70: 59-75.

Cleaves MA, Olson JM, Jacobs KB (1997) Exact transmission - disequilibrium tests with multiallelic markers. Genet Epidemiol 14: 337-347.

Clerget-Darpoux F, Bonaïti-Pellié C (1992) Strategies on marker information for the study of human disease. Ann Hum Genet 46: 145-153.

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42: 393-399.

Comuzzie AG, Blangero J, Mahaney MC, Mitchell BD, Hixson JE, Samollow PB, Stern MP, MacCluer JW (1995) Major gene with sex-specific effects influences fat mass in Mexican Americans. Genet Epidemiol1 2: 475-88.

Demenais FM (1991) Regressive logistic models for familial diseases: A formulation assuming an underlying liability model. Am J Hum Genet 49: 773-785.

Demenais FM, Elston RC (1981) A general transmission probability model for pedigree data. Hum Hered 31: 93-99.

Demenais FM, Laing AE, Bonney GE (1992) Numerical comparison of two formulation of the logistic regressive models with the mixed model in segregation analysis of discrete traits. Genet Epidemiol 9: 419-435.

Dipple KM, McCabe ERB (2000) Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and system dynamics. Am J Hum Genet 66: 1729-1735.

Dizier M-H, Babron M-C, Clerget-Darpoux F (1994) Interactive effect of two candidate genes in a disease: extension of the marker-association-segregation $\chi 2$ method. Am J Hum Genet 55: 1042-1049.

Dizier M-H, Babron M-C, Clerget-Darpoux F (1996) Conclusion of LOD-Score analysis for family data generated under two-locus models. Am J Hum Genet 58: 1338-1346.

Dizier M-H, Bonaiti-Pellie C, Clerget-Darpoux F (1993) Conclusion of segregation analysis for family data under two-locus models. Am J Hum Genet 53: 1338-1346.

Elson RC, Sobel E (1979) Sampling considerations in gathering and analysis of pedigree data. Am J Hum Genet 31: 62-69.

Elston RC (1989) Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. Am J Med Genet 34: 487-488.

Elston RC (1995) Invited editorial: Twixt cup and lip: How intractable is the ascertainment problem? Am J Hum Genet 56:15-17.

Elston RC (1998) Methods of linkage analysis – and the assumptions underlying them. Am J Hum Genet 63: 931-934.

Elston RC, Bonney GE (1984) Sampling consideration in the design and analysis of family studies. In: Rao DC, Elston RC, Kuller LH, Feinlieb M, Carter C, Havlik R. (eds) Genetic epidemiology of coronary heart disease: past, present and future. Alan R Liss, New York, p. 349-371.

Elston RC, Bonney GE (1986) Sampling via probands in the analysis of family studies. Proc XIII-th Internat Biometric Conf, University of Washington, Seattle, July 27 - August 1, 1986.

Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. Genet Epidemiol 19: 1-17.

Elston RC, Song D, Iyengar SK (2005) Mathematical assumptions versus biological reality: Myths in affected sib-pair linkage analysis. Am J Hum Genet 76:152-156.

Elston RC, Stewart J (1971) A general model for genetic analysis of pedigree data. Hum Hered 21: 523-542.

Ewens WJ, Shute NCE (1986) A resolution of the ascertainment sampling problem. I Theory. Theor Popul Biol 30: 388-412.

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision and admixture. Am J Hum Genet 57: 455-564.

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Thans Roy Soc: Edinburgh, p 399-433.

Fisher RA (1934) The effect of methods of ascertainment upon the estimation of frequencies. Ann Eugen 6:13-25.

Galton F (1888). Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society 45: 135-145.

Galton F (1889) Natural inheritance. London, 259 p.

Geldermann H (1975) Investigations on inheritance of quantitative characters in animals by gene markers. I. Meth Theoret Appl Genet 46: 319-330.

George V, Tiwari HK, Zhu X, Elston RC (1999) A test of transmission / disequilibrium for quantitative traits in pedigree data, by multiple regression. Am J Hum Genet 65: 236-245.

George VT, Elston RC (1991) Ascertainment: an overview of the classical segregation analysis model for independent sibships. Biomed J 33: 741-753.

Ginsburg E, Axenovich T (1992) A cooperative binomial ascertainment model. Am J Hum Genet 51: 1156-1160.

Ginsburg E, Livshits G, Yakovenko K, Kobyliansky E (1999) Genetics of human body size and shape: Evidence for oligogenic control of adiposity. Ann Hum Biol 26: 79-87.

Ginsburg E, Livshits G. (1999) Segregation analysis of quantitative traits. Ann Hum Biol, 26: 103-129.

Ginsburg E, Malkin I, Elston RC (2003) Sampling correction in pedigree analysis. Statistical Applications in Genetics and Molecular Biology 2: (1) Article 2, http://www.bepress.com/sagmb/vol2/iss1/art2.

Ginsburg E, Malkin I, Elston RC. (2004) Sampling correction in linkage analysis. Genet Epidemiol 27: 87-96.

Ginsburg EKh (1974) On formulation and formalization of two problems of interpopulation breeding. Biom Zeitschr 16: 511-517.

Ginsburg EKh (1975) On a method of formalization of recombination process in polygenic models. Biom Zeitschr 17: 41-47.

Ginsburg EKh (1984) Description of inheritance of quantitative traits. Novosibirsk, “Nauka”, 247 p.

Ginsburg EKh (1997) Program Package for Mendelian Analysis of Pedigree Data (MAN). Version 4. Technical Report. Department of

Anatomy and Anthropology, Sackler Faculty of Medicine, Tel Aviv University.

Ginsburg EKh, Axenovich TI (1986) Testing of monogenic hypotheses on the basis of arbitrary structured pedigrees sampled via proband. III. Quantitative trait. Genetika 22: 599-607.

Ginsburg EKh, Fedotov AM, Chepkasov IL (1986) Program package for testing of major - gene models of quantitative trait MAN-1. Novosibirsk, Inst of Cytology and Genetic Press, 42 p.

Ginsburg EKh, Nikoro ZS (1973a) On genetic correlations. I. Pleiotropy and disequilibrium. Genetika 9: 45-54.

Ginsburg EKh, Nikoro ZS (1973b) On genetic correlations. II. Methods of estimation. Genetika 9: 148-155.

Ginsburg EKh, Nikoro ZS (1982) Analysis of variance and breeding problems. Novosibirsk, "Nauka", 168 p.

Goldin LR, Weeks DE (1993) Two-locus models of disease: Comparison of likelihood and nonparametric linkage methods. Am J Hum Genet, 53: 908-915.

Greenberg DA (1989) Inferring mode of inheritance by comparison of lod scores. Am J Med Genet 35: 480-486.

Greenberg DA, Hodge SE, Vieland VJ, Spence MA (1996) Affected-only linkage methods are not a panacea. Am J Hum Genet 58: 892-895.

Haldane JBS (1938) The estimation of the frequencies of recessive conditions in man. Ann Eugenics 8: 255-262.

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and marker locus. Behav Genet 2: 3-19.

Hasstedt SJ (1995) Phenotypic assortative mating in segregation analysis. Genet Epidemiol 12: 109-127.

Hasstedt SJ (2002) Pedigree Analysis Package. Version 5.0. Department of Human Genetics University of Utah, Salt Lake City.

Hazel LN, Lush JL (1942) The efficiency of three methods of selection. J Hered 33: 393-399.

Hodge SE (1988) Conditioning on subsets of the data: applications to ascertainment and other genetic problems. Am J Hum Genet 43: 364-373.

Hodge SE, Boehnke M (1984) A note on Cannings and Tompson's sequential sampling scheme for pedigrees. Am J Hum Genet 39: 274-281.

Hodge SE, Elston RC (1994) Lods, wrods and mods: the interpretation of lod scores calculated under different models. Genet Epidemiol 11: 329-342.

Hodge SE, Vieland VJ (1996) The essence of single ascertainment. Genetics 144: 1215-1223.

Holmans P (2001) Likelihood ratio affected sib-pair tests applied to multiply affected sibships: issues of power and type I error rate. Genet Epidemiol 20: 44-56.

Kalbfleish JD, Sprott DA (1970) Application of likelihood methods to models involving large numbers of parameters. J R Stat Soc 32: 175-208.

Karasik D, Ginsburg E, Livshits G, Pavlovsky O, Kobyliansky E (2000) Evidence of major gene control of cortical bone loss in Humans. Genet Epidemiol 19: 410-421.

Karunaratne PM, Elston RC (1998) A multivariate logistic model (MLM) for analyzing binary family data. Am J Med Genet 76: 428-437.

Kendall MC, Stuart A (1970) The Advanced Theory of Statistics. Vol 2. Inference and Relationship. 2nd Ed. Charles Griffin & Co. London.

Kruglyak L, Lander ES (1995) Complete Multipoint Sib-Pair Analysis of Qualitative and Quantitative Traits, Am J of Hum Genet 57: 439-454.

Kruglyak L, Lander ES (1995) High-resolution genetic mapping of complex traits.Am J Hum Genet 56: 1212-23.

Kullback S (1959) Information theory and statistics. New York – John Wiley&Sons Inc. London – Chapman&Hall, Limited.

Lalouel JH, Morton NE (1981) Complex segregation analysis with pointers. Hum Hered 31: 312-321.

Lecomte E, Herbeth B, Nicaud V, Rakotovao R, Artur Y, Tiret L (1997) Segregation analysis of fat mass and fat-free mass with age- and sex-dependent effects: the Stanislas Family Study. Genet Epidemiol 14: 51-62.

Livshits G, Yakovenko K, Ginsburg E, Kobyliansky E (1998) Genetics of human body size and shape: Pleiotropic and independent genetic determinants of adiposity. Ann Hum Biol 25: 221-236.

Mahaney MC, Jaquish CE, Comuzzie AG (1995) Statistical genetics of normal variation in family data for oligogenic diseases. Genet Epidemiol 12: 783-7.

Malkin I, Ginsburg E and Elston RC (2002) Increase in Power of Transmission-Disequilibrium Tests for Quantitative Traits. Genet Epidemiol 23: 234–244.

Markianos K, Daly MJ, Kruglyak L (2001) Efficient Multipoint Linkage Analysis through Reduction of Inheritance Space, Am J Hum Genet 68: 963–977.

Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: The pedigree disequilibrium test. Am J Hum Genet 67: 146-154.

Mather K, Jinks JL (1982) Biometrical genetics. 3rd ed. London, New York, Chapman and Hall.

McPeek MS, Strahs A (1999) Assessment of linkage disequilibrium by decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65: 858-875.

Mendel NR, Elston RC (1974) Multivariate quantitative traits: genetic analysis and prediction of recurence risks. Biometrics 30: 41-57.

Morton NE (1959) Genetic tests under incomplete ascertainment. Am J Hum Genet 11: 1-16.

Neel JV (1978) The population structure of an Amerindian Tribe, the Yanomama. Ann Rev Genet 12: 365-413.

Nikoro ZS, Stakan GA, Charitonova SN, Vasileva LA, Ginsburg EKh, Reshetnikova NF (1968) Theoretical grounds of animal breeding. Moscow, “Kolos”, 440 p.

Olson JM, Wijsman EM (1993) Linkage between quantitative trait and marker loci: methods using all relative pairs. Genet Epidemiol 10: 87-102.

Ott J (1995) Linkage analysis with biological markers. Hum Hered 45: 169-174.

Pearson K (1904) On generalized theory of alternative inheritance with special reference to Mendel’s laws. Phil Trans Roy Soc. A. N203: 53-86.

Pearson K (1920) Notes on the history of correlation. Biometrika 13: 25-45.

Pérusse L, Moll PP, Sing CF (1991) Evidence that a single gene with gender- and age-dependent effects influences systolic blood pressure determination in a population-based sample. Am J Hum Genet 49: 94-105.

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69: 1-14.

Rao DC (1998) CAT scans, PET scans, and genomic scans. Genet Epidemiol 15: 1-18.

Rice J, Reich T, Cloninger CR (1979) An approximation to the multivariate normal integral: Its application to multifactorial quantitative traits. Biometrics 35: 451-459.

Risch N (1990a) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46: 222-228.

Risch N (1990b) Linkage strategies for genetically complex traits. II. The power of relative pairs. Am J Hum Genet 46: 229-241.

Risch N (1990c) Linkage Strategies for Genetically Complex Traits. III. The effect of Marker Polymorphism on Analyses of Affected Relative Pairs. Am J Hum Genet 46: 242-253.

Risch N. (1984) Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 36: 363-386.

S.A.G.E. (2004) Statistical Analysis for Genetic Epidemiology, S.A.G.E. 5.0 0 http://darwin.cwru.edu/sage/

Sawyer S (1990) Maximum likelihood estimators for incorrect models, with an application to ascertainment bias for continuous characters. Theor Popul Biol 38: 351-366.

Schaid DJ, Elston RC, Tran L, Wilson AF (2000) Model-free sib-pair linkage analysis: Combining full-sib and half-sib pairs. Genet Epidemiol 19: 30-51.

Schaid DJ, Rowland CM (2000) Robust transmission regression models for linkage and association. Genet Epidemiol 19(Suppl): S78-S84.

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate gene association studies using cases and parents. Am J Hum Genet 55: 402-409.

Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53: 1127-36.

Shete S, Jacobs KB, Elston RC (2003) Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: Weighting sums and differences. Hum Hered 55 :79-85.

Shute NCE, Ewens WJ (1988a) A resolution of the ascertainment sampling problem. II. Generalization and numerical results. Am J Hum Genet 43: 374-386.

Shute NCE, Ewens WJ (1988b) A resolution of the ascertainment sampling problem. III. Pedigrees. Am J Hum Genet 43: 387-395.

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62: 450-458.

Spielman RS, McGinnes RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the Insulin gene region and insulin-dependent Diabetes Mellitus (IDDM). Am J Hum Genet, 52: 120-132.

Spielman RS, McGinnis RE, Ewens WJ (1994) The transmission / disequilibrium test detects cosegregation and linkage. Am J Hum Genet 54: 559-560.

Stene J (1977) Assumption for different ascertainment models in human genetics. Biometrics 33: 523-527.

Stene J (1978) Choice of ascertainment model. I. Discrimination between single-proband models by means of birth order data. Ann Hum Genet 42: 219-229.

Stephens M, Smith NJ, Donnelly P A (2001) New Statistical Method for Haplotype Reconstruction from Population Data. Am J Hum Genet 68: 978–989.

Thompson E (1986) Pedigree analysis in human genetics. The John Hopkins University Press, London.

Thompson E (1987) Likelihoods in pedigree analysis under sequential sampling. Am J Hum Genet 41: 687-689.

Thompson EA, Cannings C (1979) Sampling schemes and ascertainment. In: Sing CF and Skolnick MH (Eds) Genetic Analysis of Common Diseases: Application to Predictive Factors in Coronary Disease, Alan R. Liss, New York, pp. 363-382.

Vieland VJ (1998) Bayesian linkage analysis, or: how I learned to stop worrying and love the posterior probability of linkage. Am J Hum Genet 63: 947-954.

Vieland VJ, Hodge SE (1995) Inherent intractability of the ascertainment problem for pedigree data: A general likelihood framework. Am J Hum Genet 56: 33-43.

Vieland VJ, Hodge SE (1996) The problem of ascertainment for linkage analysis. Am J Hum Genet, 58: 1072-1084.

Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Thans Amer Math. Soc 54: 426.

Wang K, Huang J, Vieland VJ. (2000) The consistency of the posterior probability of linkage. Am J Hum Genet 64: 533-553.

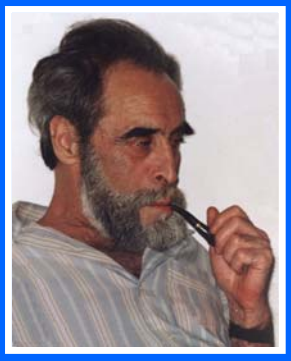
Weinberg W (1912) Weitere Beitrage sur Theory der Vererbung. IV. Uber Methode und Fehlerquellen der Untersuchung auf Mendelsche Zahlen beim Menschen. Arch Rassen Gesellschaftbiol 9: 165-174.

Whittemore AS, Halpern J (2001) Problems of the definition, interpretation and evaluation of genetic heterogeneity. Am J Hum Genet 68: 457-465.

Wolf U (1995) The genetic contribution to the phenotype. Hum Genet 95: 127-148.

Wright S (1968) Evolution and genetics of population. Vol 1, Genetic and biometric foundations. Univ of Chicago Press, Chicago – London.

THEORETICAL ASPECTS OF PEDIGREE ANALYSIS



Emil Ginsburg – Doctor of Biological Sciences, Professor, Novosibirsk State University (Russia) and Sackler Faculty of Medicine, Tel-Aviv University (Israel); author of 4 books and over 100 articles in Population Biology, Genetic Epidemiology and Biometrics.

Ida Malkin – PhD in Physics and Mathematics, Senior Researcher, Sackler Faculty of Medicine Tel-Aviv University (Israel); author of 20 articles in Population Biology and Genetic Epidemiology.

Robert C. Elston – Director, Division of Genetic and Molecular Epidemiology, and Professor, Department of Epidemiology and Biostatistics, Case Western Reserve University (Cleveland, USA); author of 6 books and over 500 articles in Human Genetics, Genetic Epidemiology, Biometrics and Statistics.

Theoretical Aspects of Pedigree Analysis

by E. Ginsburg, I. Malkin, R.C. Elston

Errata and clarifications

Page 14, line 2: "genes that control **the** trait"
Page 18, line 4: no comma after "Alleles"
Page 32, line 5: remove "_"
Page 53: The notation on this page would be clearer if, in the left hand side of equation (3.1) and the equation 12 lines from the bottom, and in the right hand side of the equation at the bottom of the page, *smpl* came to the right of the conditioning symbol |
Page 42, figure legend: "C_2 = **C\C$_1$** – the complementary part…"
Page 47, line 10 from bottom: "**In** terms of"
Page 61, line 5 from bottom: remove "**they**"
Page 63, line 4 from bottom: "**another**"
Page 73, line 7: "If this **is** the case**,** all"
Page 78, line 13: "genetic model"
Page 82, line 12: "**groups**"
Page 116, line 10 from bottom: **SMF** – sampling model free (not defined until page 121)
Page 125: it is stated that the "approximate AMF likelihood cn be easily calculated by replacing each missing phenotype on the PSF members by the sample mean" This mean was calculated from all the non-missing phenotypes in the sample, both those of PSF members and those of non-PSF members.
Pages 149-151: n is the number of pedigrees sampled; the parameter values used for the simulation are - $\rho = 0.05$; $p = 0.2$; $q = 0.7$; $\pi = 0.2$, and $D = 0.0$
Page 154: in equation (7.8) the estimate of θ is not the same in both the numerator and the denominator, as incorrectly suggested by the notation. At the bottom of the page: "provided $\hat{\rho}$ is not restricted"
Page 195, line 2: "a TDT type **test**"
Page 198, last line: "analysis **has** been made"
Page 208, line 3 from bottom: "(**1997)**"
Page 209, line 1: "**Bonné**"
Page 209, line 12 from bottom: "**Lalouel**"
Page 210, line 7: "**Cleves**"
Page 210, line 10 from bottom and page 211, line 5 from bottom: "**Bonaïti-Pellié**"
Page 210, line 4 from bottom: "Genet **Epidemiol 12**"
Page 213, line 6: "**Trans** Roy Soc"
Page 213, line 2 from bottom: "**Biometrical** J"
Page 214, line 9 from bottom: "Genet Epidemiol: **24: 1-10"**
Page 216, line 8 from bottom: "**Thompson's**"
Page 222, line 10 from bottom: "**Assumptions**"
Page 223, line 7 from bottom: "**Trans** Amer Math Soc"