# Contents

1 **Introduction** ............................................. 5  
   1.1 Chromosomes and Genes ............................... 5  
   1.2 Inheritance of Genes ................................ 6  
   1.3 Determining Genetic Mechanisms and Gene Positions .................. 9  

2 **Probability Theory** .................................. 13  
   2.1 Random Models and Probabilities ..................... 13  
   2.2 Random Variables and Distributions .................. 20  
   2.3 Expectation, Variance and Covariance ............... 34  
   2.4 Exercises ........................................... 43  

3 **Inference Theory** .................................... 45  
   3.1 Statistical Models and Point Estimators ............ 45  
   3.2 Hypothesis Testing .................................. 49  
   3.3 Linear regression ................................... 54  
   3.4 Exercises ........................................... 57  

4 **Parametric Linkage Analysis** ......................... 59  
   4.1 Introduction .......................................... 59  
   4.2 Two-Point Linkage Analysis ......................... 59  
      4.2.1 Analytical likelihood and lod score calculations .......... 59  
      4.2.2 The pedigree likelihood .......................... 67  
      4.2.3 Missing marker data .............................. 71  
      4.2.4 Uninformativeness ............................... 75  
      4.2.5 Other genetic models ............................ 76  
   4.3 General pedigrees ................................... 83  
   4.4 Multi-Point Linkage Analysis ....................... 84  
   4.5 Power and simulation ................................ 85  
   4.6 Software ........................................... 86
CONTENTS

5 Nonparametric Linkage Analysis 89
  5.1 Affected sib pairs ........................................ 89
    5.1.1 The Maximum Lod Score (MLS) .................. 92
    5.1.2 The NPL Score .................................... 94
    5.1.3 Incomplete marker information ................. 97
    5.1.4 Power and p-values ................................ 99
  5.2 General pedigrees ...................................... 102
  5.3 Extensions ............................................. 105
  5.4 Exercises .............................................. 107

6 Quantitative Trait Loci 109
  6.1 Properties of a Single Locus ......................... 111
    6.1.1 Characterizing the influence of a locus on the phenotype . 111
    6.1.2 Decomposition of the genotypic value, (Fisher 1918) . . 112
    6.1.3 Partitioning the genetic variance ............... 115
    6.1.4 Additive effects, average excesses, and breeding values . 116
    6.1.5 Extensions for multiple alleles ................. 118
  6.2 Genetic Variation for Multilocus Traits ............ 120
    6.2.1 An extension of the least-squares model for genetic effects . 121
    6.2.2 Some notes on Environmental Variation .......... 125
  6.3 Resemblance between relatives ...................... 126
    6.3.1 Genetic covariance between relatives .......... 126
  6.4 Linkage methods for quantitative traits ........... 133
    6.4.1 Analysis of sib-pairs: The Haseman-Elston Method . 136
    6.4.2 Linkage Analysis in General Pedigrees: Variance Component Analysis .................. 140
    6.4.3 Software for quantitative trait linkage analysis .... 144
  6.5 Exercises .............................................. 145

7 Association Analysis 147
  7.1 Family-based association methods .................. 152
    7.1.1 The Transmission/Disequilibrium Test, TDT ...... 152
    7.1.2 Tests using a multiallelic molecular marker .... 158
    7.1.3 No parental information available ............. 159
    7.1.4 An association test for extended pedigrees, the PDT .... 163

8 Answers to Exercises 167

A The Greek alphabet 171
Preface

These lecture notes are intended to give an overview of statistical methods employed for localization of genes that are involved in the causal pathway of human diseases. Thus the applications are mainly within human genetics and various experimental design techniques employed in animal genetics are not discussed.

The statistical foundations to gene mapping were laid already until the 1950s. Several simple Mendelian traits have been mapped since then and the responsible gene cloned. However, much remains to be known about the genetic components of more complex diseases (such as schizophrenia, adult diabetes and hypertension). Since the early 1980s, a large set of genetic markers has been discovered, e.g. restriction fragment length polymorphisms (RFLPs) and single nucleotide polymorphisms (SNPs). In conjunction with algorithmic advances, this has enabled more sophisticated mapping techniques, which can handle many markers and/or large pedigrees. Further, the completion of the human genome project (Lander et al. 2001), (Venter et al. 2001) has made it possible with automated genotyping. All these facts together imply that 'statistics in gene mapping’ is still a very active research field.

Our main focus is on statistical techniques, whereas the biological and genetics material is less detailed. For readers who wish to get a broader understanding of the subject, we refer to a textbook covering gene mapping of (human) diseases, e.g. Haines and Pericak-Vance (1998). Further, more details on statistical aspects of linkage and association analysis can be found in Ott (1999), Sham (1998), Lynch and Walsh (1997) and Terwillinger and Göring (2000).

Some prior knowledge of probability or inference theory is useful when reading the lecture notes. Although all statistics concepts used are defined ‘from scratch’, some familiarity with basic calculus is helpful to get a deeper understanding of the material.

The lecture notes are organized as follows: A very comprehensive introduction to genetics is given in Chapter 1. In Chapters 2 and 3 basic concepts from probability and inference theory are introduced, and illustrated with genetic examples. The following four chapters show in more detail how statistical techniques are applied to various areas of genetics; linkage analysis, quantitative trait loci methods, and association analysis.
Chapter 1

Introduction

1.1 Chromosomes and Genes

The genetic information of an individual is contained in 23 pairs of chromosomes in the cell nucleus; 22 paired autosomes and two sex chromosomes.

The chemical structure of the chromosomes is deoxyribonucleic acid (DNA). One single strand of DNA consists of so called nucleotides bond together. There are four types of nucleotide bases, called adenine (A), guanine (G), cytosine (C) and thymine (T). The sequence of DNA bases constitutes a code for synthesizing proteins, and they are arranged in groups of three, so called codons, e.g. ACA, TTG and CCA. The basis of the genetic code is that the $4^3 = 64$ possible codons specify 20 different amino acids.

Watson and Crick correctly hypothesized in 1953 the double-helical structure of the chromosomes, with two bands of DNA strands attached together. Each base in one strand is attached to a base in the other strand by means of a hydrogen bond. This is done in a complementary way (A is bonded to T and G to C), and thus the two strands carry the same genetic information. The total number of base pairs along all 23 chromosomes is about $3 \times 10^9$.

It was Mendel who 1865 first proposed that discrete entities, now called genes, form the basis of inheritance. The genes are located along the chromosomes, and it is currently believed that the total number of genes for humans is about 30 000. A gene is as a segment of the DNA within the chromosome which specifies uniquely an amino acid sequence, which in turn specifies the structure and function of a subunit in a protein. More details on how the protein synthesis is achieved can be found in e.g. Haines and Pericak-Vance (1998). There can be different variants of a gene, called alleles. For instance, the normal allele a might have been mutated into a disease allele A. A locus is a well-defined position along a chromosome, and a genotype consists of a pair of alleles at the same locus, one inherited from the father and one from the mother. For instance, three genotypes (aa), (Aa) and (AA) are possible for a
biallelic gene with possible alleles $a$ and $A$. A person is \textit{homozygous} if both alleles of
the genotype are the same (e.g. (aa) and (AA)) and \textit{heterozygous} if they are different
(e.g. (Aa)). A sequence of alleles from different loci received from the same parent is
called a \textit{haplotype}.

1.2 Inheritance of Genes

Among the 46 chromosomes in each cell, there are 23 inherited from the mother
and 23 from the father. Each maternal chromosome consists of segments from both
the (maternal) grandfather and the grandmother. The positions where the DNA seg-
ments switch are called \textit{crossovers}. Thus, when an egg is formed, only half of the nu-
cleotides from the mother are passed over. This process of mixing grandpaternal and
grandmaternal segments is called \textit{meiosis}. In the same way, meiosis takes place du-
dring formation of each sperm cell, with the (paternal) grandfather and grandmother
DNA segments being mixed. A simplified picture of meiosis (for one chromosome)
is shown\footnote{Figure 1.1 is simplified, since in reality two pairs of chromosomes mix, where the chromosomes
within each pair are identical. Cf. e.g. Ott (1999) for more details.} in Figure 1.1.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{meiosis.png}
\caption{A simplified picture of meiosis when one chromosome of the mother
or father is formed. The dark and light segments correspond the grandfather’s and
grandmother’s DNA strands respectively. In this picture, two crossovers occur.}
\end{figure}
1.2. INHERITANCE OF GENES

Crossovers occur randomly along each chromosome. Two loci are located one Morgan (or 100 centiMorgans, cM) from each other when the expected number of crossovers between them is one per meiosis \(^2\). This is a unit of (genetic) map length, which is different for males and females and also different than the physical distance (measured in units of 1000 base pairs, kb, or million base pairs, Mb), cf. Table 1.1. The total map length of the 22 autosomes is 28.5 Morgans for males and 43 Morgans for females. Often one simplifies matters and uses the same map distance for males and females by sex-averaging the two map lengths of each chromosome. This gives an total map length of approximately 36 Morgans for all autosomes. The lengths of the chromosomes vary a lot, but the average map length of an autosome is about \(\frac{36}{22} = 1.6\) Morgans.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Map length Male</th>
<th>Map length Female</th>
<th>Ph length</th>
<th>Chr</th>
<th>Map length Male</th>
<th>Map length Female</th>
<th>Ph length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>221</td>
<td>263</td>
<td>13</td>
<td>13</td>
<td>107</td>
<td>157</td>
<td>144</td>
</tr>
<tr>
<td>2</td>
<td>193</td>
<td>255</td>
<td>14</td>
<td>14</td>
<td>106</td>
<td>151</td>
<td>109</td>
</tr>
<tr>
<td>3</td>
<td>186</td>
<td>214</td>
<td>15</td>
<td>15</td>
<td>84</td>
<td>149</td>
<td>106</td>
</tr>
<tr>
<td>4</td>
<td>157</td>
<td>203</td>
<td>16</td>
<td>16</td>
<td>110</td>
<td>152</td>
<td>98</td>
</tr>
<tr>
<td>5</td>
<td>149</td>
<td>194</td>
<td>17</td>
<td>17</td>
<td>108</td>
<td>152</td>
<td>92</td>
</tr>
<tr>
<td>6</td>
<td>142</td>
<td>183</td>
<td>18</td>
<td>18</td>
<td>111</td>
<td>149</td>
<td>85</td>
</tr>
<tr>
<td>7</td>
<td>144</td>
<td>171</td>
<td>19</td>
<td>19</td>
<td>113</td>
<td>121</td>
<td>67</td>
</tr>
<tr>
<td>8</td>
<td>135</td>
<td>155</td>
<td>20</td>
<td>20</td>
<td>104</td>
<td>120</td>
<td>72</td>
</tr>
<tr>
<td>9</td>
<td>130</td>
<td>145</td>
<td>21</td>
<td>21</td>
<td>66</td>
<td>77</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>144</td>
<td>144</td>
<td>22</td>
<td>22</td>
<td>78</td>
<td>89</td>
<td>56</td>
</tr>
<tr>
<td>11</td>
<td>125</td>
<td>143</td>
<td>X</td>
<td>-</td>
<td>193</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>136</td>
<td>232</td>
<td>Autos</td>
<td>2849</td>
<td>4301</td>
<td>3093</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: Chromosome lengths for males and females, measured in units of map length (cM) and physical length (Mb) respectively. The table is taken from Collins et al. (1996), cf. also Ott (1999). Autos refers to the sum over all the 22 autosomes.

Consider two loci on the same chromosome, with possible alleles \(A, a\) at the first locus and \(B, b\) at the second one. Suppose an individual has inherited a haplotype \(AB\) from the father and \(ab\) from the mother respectively (so that the genotypes of the two loci are \((Aa)\) and \((Bb)\)). If a gamete (egg or sperm cell) receives a haplotype \(Ab\) during meiosis, it is said to be recombinant, meaning that the two alleles come from different parents. The haplotype \(aB\) is also recombinant, whereas \(AB\) and \(ab\) are non-recombinant. Thus two loci are recombinant or non-recombinant if an odd

\(^2\)This simply means that in a large set of meioses, there will be on the average one crossover per meiosis.
or even number of crossovers occur between them. The recombination fraction $\theta$ is the probability that two loci become recombinant during meiosis. Obviously, the recombination fraction must be a function of the map distance $x$ between the two loci, since it is less likely that two nearby loci become recombinant.

There exist many probabilistic models for the occurrence of crossovers. The simplest (and most often used) one is due to Haldane (1919). By assuming that crossovers occur randomly along the chromosome according to a so called Poisson process, one can show that the recombination fraction is given, as a function of map distance, by

$$\theta(x) = 0.5(1 - \exp(-0.02x)),$$

when the map distance $x$ is measured in cM. Equation (1.1) is referred to as Haldane’s map function, and depicted in Figure 1.2. For small $x$ we have $\theta \approx 0.01x$, whereas for large $x$ the recombination fraction has increased to $\theta \approx 0.5$. Two loci are called linked when $\theta < 0.5$ and this is always the case when they belong to the same chromosome. Loci on different chromosomes are unlinked, meaning that $\theta = 0.5$. Formally, we may say that the map distance between loci on different chromosomes is infinite, meaning that inheritance at the two loci are independent events.

![Figure 1.2: Recombination fraction $\theta$ as a function of map distance $x$ according to Haldane’s map function. The map distance is measured in cM.](image-url)
1.3 Determining Genetic Mechanisms and Gene Positions

In general, the genotypes cannot be determined unambiguously. The phenotype is the observable expression of a genotype that is being used in a study. For instance, a phenotype can be binary (affected/nonaffected) or quantitative (adult length, body weight, body mass index, insulin concentration, ...). The way in which genes and environment jointly affect the phenotype is described by means of a genetic model, as schematically depicted in Figure 1.3.

Figure 1.3: Schematic description of a genetic model, showing how the genotype at a susceptibility locus and environmental factors give rise to a phenotype.

The genetic model might involve one or several genes. For a monogenic disease, only one gene increases susceptibility to the disease. This is the case for Huntington's disease, cf. Gusella et al. (1983). When the genetic component of the disease has contributions from many genes, we have a complex or polygenic disease. For instance, type 2 diabetes is likely to be of this form, cf. e.g. Horikawa et al. (2000). Further, it might happen that different gene(s) are responsible for the disease in different sub-populations. In that case, we speak of a heterogenic disease. Hereditary breast cancer is of this kind, where two genes responsible for the disease in different populations have been found so far, cf. Hall et al. (1990) and Wooster et al. (1995).

The way in which the gene(s) affect the phenotypes (or how 'the genetic component penetrates') is described by a number of penetrance parameters. For instance, for a monogenic disease the penetrance parameters reveal if the disease is dominant (one disease allele of the disease genotype is sufficient for becoming affected), or recessive (both disease alleles are needed).

The objective of segregation analysis is to determine the penetrance and environmental parameters of the genetic models, using phenotype data from a number of
Two typical pedigrees with dominant (a) and recessive (b) modes of inheritance are shown in Figure 1.4. Individuals without ancestors in the pedigree are called founders, whereas the remaining ones are called nonfounders. Males are depicted with squares and females by circles, respectively. The phenotype is binary with black and white indicating affected and unaffected individuals, respectively. In family b), one of the founders have a disease allele, which has then been segregated down through three generations. Because of the recessive nature of the trait, the disease allele is hidden until the third generation. Then two of five offspring from a cousin marriage become affected, by getting one disease allele from the father and one from the mother.

Another pedigree is shown in Figure 1.5. All individuals have been genotyped at two loci, as indicated. When it is known which two alleles come from the father and mother respectively, we say that the phase of the individual is known, meaning that the paternal and maternal haplotypes can be determined. This is indicated by vertical lines in Figure 1.5. The phase of all founders is typically unknown, unless previous family history is available. Sometimes the phase can be determined by pure inspection. For instance, the male in the second generation must have inherited the A- and B-alleles from the father and the a- and b-alleles from the mother. Since he is doubly heterozygous, his phase is known, with paternal and maternal haplotypes AB and ab respectively. The male of the third generation has known phase too. Moreover, we know that his paternal haplotype is recombinant, since the A and b-alleles must come from different grandparents.

Another important task is to locate the locus (loci) of the gene(s) in the genetic model. For this genotypes from a number of markers are needed. These are loci (not necessarily genes) with known positions along the chromosomes with at least two possible alleles. By typing, i.e. observing the marker genotypes of as many individuals as possible in the pedigrees, one can trace the inheritance pattern. In linkage analysis regions are sought for where the inheritance pattern obtained from the markers are
1.3. **DETERMINING GENETIC MECHANISMS AND GENE POSITIONS**

Figure 1.5: A pedigree with binary phenotypes and alleles from two loci shown. The first locus has alleles $A$, $a$, and the second one alleles $B$, $b$. Cf. Figure 1.1 in Ott (1999).

Highly correlated with the inheritance pattern observed from the phenotypes. The rationale for this is that nearby loci must have correlated inheritance patterns, because crossovers occur between the two loci with low probability. In *association analysis*, one uses the fact that markers in close vicinity of a disease locus might be in *linkage disequilibrium* with the disease locus. This means that some marker alleles are over represented among affected individuals. One reason for this is that the haplotype of an ancient disease founder is left intact through many generations in a chromosomal region surrounding the disease locus.
2.1 Random Models and Probabilities

A model is a simplified map of reality. Often, the model is aimed at solving a particular practical problem. To this end, we need to register a number observable quantities from the model, i.e. perform a so called experiment. In a deterministic model, these quantities can just attain one value (which is still unknown before we observe it), whereas for a random model, the outcome of the observed quantities might differ if we repeat the experiment.

Example 1 (A randomly picked gene.) Suppose we pick at random a person from a population, and wish to register the genotype at a certain locus. If the locus is monoallelic with allele $A$, only one genotype, $(AA)$, is possible. Then the model is deterministic. If, on the other hand, two alleles $A$ and $a$ are possible and at least two of the three corresponding genotypes $(AA)$, $(Aa)$, and $(aa)$ occur in the population, the outcome (and hence the model) is random.

Let $\omega$ be the outcome of the experiment in a random model and $\Omega$ be the set of all possible values that $\omega$ can attain, the so called sample space. A subset $B \subset \Omega$ of the sample space is referred as an event. A probability function is a function which to each event $B$ assigns a number $P(\omega \in B)$ between 0 and 1 (‘the probability of $\omega$ falling into $B$’). Sometimes we write just $P(B)$ (‘the probability of $B$’), when it is clear from the context what $\omega$ is.

Example 2 (Binary (dichotomous) phenotypes.) Consider a certain disease for which individuals are classified as either affected or unaffected. Thus the sample space is $\Omega = \{\text{unaffected, affected}\}$. The prevalence $K_p$ of the disease is the proportion of affected individuals in the population. Using probability functions we write this as

$$K_p = P(\omega = \text{affected}),$$  \hspace{1cm} (2.1)
i.e. the probability of the event $B = \{ \text{affected} \}$. 

Figure 2.1: Graphical illustration of the intersection between two events $B$ and $C$ which are not disjoint (a) and disjoint (b) respectively.

Since events are subsets of the sample space, we can form set theoretic operations such as intersections, unions and complements with them, see Figures 2.1 and 2.2. This we write as

$$B \cup C = \text{ \textquoteleft at least one of } B \text{ and } C \text{ occur\textquoteleft}$$

$$B \cap C = \text{ \textquoteleft both } B \text{ and } C \text{ occur\textquoteleft}$$

$$B^\complement = \text{ \textquoteleft } B \text{ does not occur\textquoteleft}. $$

Figure 2.2: Illustration of (a) an event $B$ and (b) its complement $B^\complement$.

**Example 3 (Full and disjoint events.)** Notice that $\Omega$ is a subset of itself, and thus an event (the so called 'full event'). The complement $\Omega^\complement$ of the full event is $\emptyset$, the
empty set. Two events $B$ and $C$ are disjoint if $B \cap C = \emptyset$. In Example 2, \{affected\} and \{unaffected\} are disjoint, since a person cannot be both affected and unaffected.

Any probability function must obey some intuitively very plausible rules, given in the following axioms.

**Definition 1 (Kolmogorov’s axiom system.)** Any probability function, $P$, must satisfy the following three rules:

- $(i) : P(\Omega) = 1.$
- $(ii) :$ If $B$ and $C$ are disjoint, then $P(B \cup C) = P(B) + P(C)$.
- $(iii) :$ For any event $B$, $0 \leq P(B) \leq 1$.

**Example 4 (Probability of set complements.)** Suppose the prevalence of a certain disease is 0.1. What is the probability that a randomly picked individual is not affected? Obviously, this must be $0.9 = 1 - 0.1$. Formally, we can deduce this from Kolmogorov’s axiom system. Let $B = \text{affected in Example 2}$. Then $B^* = \text{unaffected}$. Since $B$ and $B^*$ are disjoint and $B \cup B^* = \Omega$, it follows from $(i)$, $(ii)$ in Kolmogorov’s axiom system that

\[ 1 = P(\Omega) = P(B \cup B^*) = P(B) + P(B^*) \]

\[ \iff P(B^*) = 1 - P(B) = 1 - 0.1 = 0.9. \]

A very important concept in probability theory is conditional probability. Given two events $B$ and $C$, we refer to $P(B|C)$ as the conditional probability of $B$ given $C$. It is the probability of $B$ given that (or conditioning on the fact that) $C$ has occurred. Formally it is defined as follows:

**Definition 2 (Conditional probability.)** Suppose $C$ is an event with $P(C) > 0$. Then the conditional probability of $B$ given $C$ is defined as

\[ P(B|C) = \frac{P(B \cap C)}{P(C)}. \]
Example 5 (Sibling relative risk.) Given a sib pair, let \( B \) and \( C \) denote the events that the first and second sibling is affected by a disease respectively. Then

\[ K_s = P(C|B) \]

is defined as the sibling prevalence of the disease. Whereas the prevalence \( K_p \) in (2.1) was the probability that a randomly chosen individual was affected, \( K_s \) is the probability of being affected given the extra information that the sibling is affected. For a disease with genetic component(s), we must obviously have \( K_s > K_p \). The extent to which the risk increases when the sibling is known to be affected, is quantified by means of the the relative risk for siblings,

\[ \lambda_s = \frac{K_s}{K_p}. \quad (2.3) \]

The more \( \lambda_s \) exceeds one, the larger is the genetic component of the disease.

Example 6 (Penetrances of a binary disease.) Suppose we have an inheritable monogenic disease, i.e. the susceptibility to the disease depends on the genotype at one certain locus. Suppose there are two possible alleles \( A \) and \( a \) at this locus. Usually \( A \) denotes the disease susceptibility allele and \( a \) the normal allele, respectively. We know from Example 2 that the prevalence is the overall probability that an individual is affected. However, with extra information concerning the disease genotype of the individual, this probability changes. The penetrance of the disease is the conditional probability that an individual is affected given the genotype. Thus we introduce

\[
\begin{align*}
f_0 &= P('\text{affected}'|(aa)), \\
f_1 &= P('\text{affected}'|(Aa)), \\
f_2 &= P('\text{affected}'|(AA)),
\end{align*}
\]

the three penetrance parameters of the genetic model. For instance, if it is known that a proportion 0.1 of the individuals in the population are \( AA \)-homozygotes, and that a fraction 0.08 are affected and \( AA \)-homozygotes. Then

\[ f_2 = \frac{P('\text{affected and } (AA)')}{P((AA))} = \frac{0.08}{0.1} = 0.8. \]

In other words, for an homozygote \( AA \) the conditional probability is 0.8 of having the disease.

Normally the probability of being affected increases with the number of disease alleles in the genotype, i.e. \( f_0 \leq f_1 \leq f_2 \). If \( f_0 > 0 \), there are phenocopies in the population, meaning that not only the gene (but also environmental factors and other genes) may be responsible for the disease. A fully penetrant autosomal dominant
2.1. RANDOM MODELS AND PROBABILITIES

disease has \( f_1 = f_2 = 1 \), i.e. one disease allele is sufficient to cause the disease with certainty. However, apart from some genetic traits that are manifest at birth, it is usually the case that \( 0 < f_0; f_1; f_2 < 1 \). The disease is dominant if \( f_1 = f_2 \) and recessive if \( f_0 = f_1 \).

Even though the penetrance parameters \( f_0, f_1 \) and \( f_2 \) model a monogenic disease very well, a drawback with them is that they are more difficult to estimate from data than e.g. the relative risk \( \lambda \) for siblings. □

Two events \( B \) and \( C \) are independent if the occurrence of \( B \) does not affect the conditional probability of \( C \) and vice versa. In formulas, this is written

\[
P(C) = P(C|B) = \frac{P(B \cap C)}{P(B)} \iff P(B \cap C) = P(B)P(C). \tag{2.5}\]

Thus we have an intuitive multiplication principle regarding independent events. The probability that both of them occur equals the product of the probabilities that each one of them occur.

**Example 7 (Hardy-Weinberg equilibrium.)** Suppose the proportion of the disease allele \( A \) in a population is \( p = P(A) \). Usually, \( p \) is a small number, like 0.0001, 0.001, 0.01 or 0.1. If the locus is two-allelic, a randomly picked allele is either \( A \) or \( a \). The events of picking \( a \) and \( A \) are therefore complements of each other. By Example 4, the probability of the normal allele is

\[
q = P(a) = 1 - P(A) = 1 - p.
\]

The probability of a randomly chosen individual’s genotype is in general a complicated function of the family history of the population as well as the mating structure. (For instance, is it more probable that a homozygote (aa) mates with another (aa) than with a heterozygote (Aa)?). The simplest assumption is to postulate that the paternal allele is independent of the maternal allele. Under this assumption, and with the acronyms ‘pa = paternal allele’ and ‘ma = maternal allele’, the probability that a randomly chosen genotype is (Aa) is

\[
P((Aa)) = P(\{’pa=A’ and ’ma=a’\} \cup \{’pa=a’ and ’ma=A’\})
\]

\[
= P(’pa=A’ and ’ma=a’) + P(’pa=a’ and ’ma=A’)
\]

\[
= P(’pa=A’)P(’ma=a’) + P(’pa=a’)P(’ma=A’)
\]

\[
= pq + qp = 2pq.
\]

In the second equality we used (ii) in Kolmogorov’s axiom system, since the events \( \{’pa=A’ and ’ma=a’\} \) and \( \{’pa=a’ and ’ma=A’\} \) are disjoint (both of them cannot happen simultaneously). In the third equality we used the independence between the
events \{ \text{pa=A} \} \text{ and } \{ \text{ma=a} \} \text{ on one hand and between } \{ \text{pa=a} \} \text{ and } \{ \text{ma=A} \} \text{ on the other hand. Similar calculations yield}

\begin{align*}
P((AA)) &= p^2, \\
P((aa)) &= q^2.
\end{align*}

(2.6)

If the genotype probabilities are given by the above three formulas, we have Hardy-Weinberg equilibrium. If for instance \( p = 0.1 \), we get \( P((AA)) = 0.01, \) \( P((Aa)) = 0.18 \) and \( P((aa)) = 0.81 \) under HW equilibrium.

Independence of more than two events can be defined analogously. If \( B_1, B_2, \ldots, B_n \) are independent, it follows that

\[ P(B_1 \cap B_2 \cap \ldots \cap B_n) = P(B_1) \cdot P(B_2) \cdot \ldots \cdot P(B_n) = \prod_{i=1}^{n} P(B_i). \]

In many cases, we wish to compute the probability of an event \( B \) when the conditional probability of \( B \) given a number of other events are given. For instance, the proportion of males having (registered) a certain type of cancer in a country can be found weighting the known proportions for different regions of the country. The formula for this is given in the following theorem:

![Diagram](image)

Figure 2.3: The law of total probability. \( C_1, C_2, \ldots, C_8 \) are disjoint subsets of the sample space and therefore we have that \( P(B) = \sum_{i=1}^{8} P(B \cap C_i) = \sum_{i=1}^{8} P(B|C_i)P(C_i) \). The diagram shows that \( P(B|C_1) = P(B|C_2) = P(B|C_4) = 0. \)
Theorem 1 (Law of total probability.) Let \( C_1, \ldots, C_k \) be a disjoint decomposition of the sample space\(^1\). Then, for any event \( B \),
\[
P(B) = \sum_{i=1}^{k} P(B|C_i)P(C_i). \tag{2.7}
\]

Example 8 (Prevalence under HW equilibrium.) What is the prevalence \( K_p \) of a monogenic disease for a population in Hardy-Weinberg equilibrium when the disease allele frequency is \( p = 0.02 \) and the penetrance parameters are \( f_0 = 0.03, f_1 = 0.3 \) and \( f_2 = 0.9 \)? We apply Theorem 1 with \( B = \) 'affected', and \( C_1, C_2 \) and \( C_3 \) the events that a randomly picked individual has genotype \((aa)\), \((Aa)\) and \((AA)\) respectively at the disease locus. Clearly \( C_1, C_2, C_3 \) form a disjoint decomposition of the sample space, since an individual has exactly one of the three genotypes \((aa)\), \((Aa)\) and \((AA)\). The probabilities of the \( C_i \)-events can be deduced from Example 7, and so
\[
K_p = P(B) = P(B|C_1)P(C_1) + P(B|C_2)P(C_2) + P(B|C_3)P(C_3)
\]
\[
= f_0 \cdot (1 - p)^2 + f_1 \cdot 2p(1 - p) + f_2 \cdot p^2
\]
\[
= 0.03 \cdot (1 - 0.02)^2 + 0.3 \cdot 2 \cdot 0.02 \cdot (1 - 0.02) + 0.9 \cdot 0.02^2
\]
\[
= 0.0409. \tag{2.8}
\]

The next theorem is very useful in many applications when the conditional probabilities are given 'in wrong order':

Theorem 2 (Bayes' Theorem.) Let \( B, C_1, \ldots, C_n \) be as given in Theorem 1. Then, for any \( i = 1, \ldots, n \),
\[
P(C_i|B) = \frac{P(B|C_i)P(C_i)}{P(B)} = \frac{P(C_i|B)P(C_i)}{\sum_{j=1}^{n} P(B|C_j)P(C_j)}. \tag{2.9}
\]

In the second equality of (2.9), we used the Law of Total Probability to equate the two denominators.

Example 9 (Probability of \((aa)\) for an affected.) In Example 8, what is the probability that an affected individual is a homozygote \((aa)\)? Using the same notation as in that example, we seek the conditional probability \( P(C_1|B) \). Since \( P(B) \) has already been calculated in (2.8), we apply Bayes’ Theorem to get
\[
P(C_1|B) = \frac{P(B|C_1)P(C_1)}{P(B)} = \frac{0.03 \cdot (1 - 0.02)^2}{0.0409} = 0.7037.
\]

\(^1\)This means that \( C_i \cap C_j = \emptyset \) when \( i \neq j \) and \( C_1 \cup \ldots \cup C_k = \Omega \).
The relative high proportion 70% of affecteds that are homozygotes \((aa)\) is explained by the fact that the phenocopy rate \(f_0\) is larger than the disease allele frequency \(p\). Thus the genetic component of the disease is rather weak.

We end this section with another application of the Law of Total Probability.

**Example 10 (Heterozygosity of a marker.)** In linkage analysis inheritance information from a number of markers with known positions along the chromosomes is used, cf. Chapters 4 and 5. The term **polymorphism** denotes the fact that a locus can have several possible allelic forms. The more polymorphic a marker is, the easier it is to trace the inheritance of that marker in a pedigree, and hence the more useful is the marker for linkage analysis. This is illustrated in Figure 2.4, where inheritance of two markers is shown for the same pedigree.

The degree of polymorphism of a marker depends on the number of allelic forms, but also on the allele frequencies. The **heterozygosity** \(H\) of a marker is defined as the probability that two independently picked marker alleles are different. It is frequently used for quantifying the degree of polymorphism.

In order to derive an explicit expression for \(H\), we assume that the marker has \(k\) allelic forms with allele frequencies \(p_1, \ldots, p_k\). We will apply the law of total probability (2.7), with \(B = \) ‘the two alleles are of the same type’ and \(C_i = \) ‘allele 1 is of type \(i\)’. Then, by the definition of allele frequency \(P(C_i) = p_i\). Further, given that \(C_i\) has occurred, the event \(B\) is the same thing as ‘allele 2 is of type \(i\)’. Therefore, since the two alleles are picked independently,

\[
P(B|C_i) = P(\text{‘allele 2 is of type } i\text{’}|C_i) = P(\text{‘allele 2 is of type } i\text{’}) = p_i.
\]

Finally, we get from (2.7);

\[
H = P(B^c) = 1 - P(B) = 1 - \sum_{i=1}^{k} P(B|C_i)P(C_i) = 1 - \sum_{i=1}^{k} p_i^2.
\]

The closer to 1 \(H\) is, the more polymorphic is the marker. For instance, a biallelic marker with \(p_1 = p_2 = 0.5\) has \(H = 1 - 0.5^2 - 0.5^2 = 0.5\). This is considered as a low degree of polymorphism. A marker with five possible alleles and equal allele frequencies \(p_1 = \ldots = p_5 = 0.2\) is more polymorphic, and has \(H = 1 - 5 \cdot 0.2^2 = 0.8\).

---

### 2.2 Random Variables and Distributions

A **random variable** (r.v.) \(X = X(\omega)\) is defined as a function of the outcome \(\omega\) in a random experiment. For instance, \(X\) may represent that part of the outcome which we can observe or the part we are currently interested in.
2.2. RANDOM VARIABLES AND DISTRIBUTIONS

Figure 2.4: Inheritance of markers at two different loci for a pedigree with three founders with known phases. The six founder alleles are a) all different b) all equal. In a) the inheritance pattern of the pedigree can be determined unambiguously from the marker information and in b) the markers give no information at all about inheritance.

A random variable \((r.v.) \, X\) is discrete if the set of possible values is countable, i.e. can be arranged in a sequence (this is always the case if there are finitely many values that \(X\) can attain). The random variation of \(X\) can be summarized by the following function:

**Definition 3 (Probability function.)** Suppose \(X\) is a discrete random variable. The probability function is then defined by

\[
x \rightarrow P(X = x),
\]

with \(x\) ranging over the countable set of values which \(X\) can attain\(^2\).

**Example 11 (Two-point distribution.)** It is common to code the possible values of a discrete random variable as integers. For instance, if the phenotype \(Y\) is binary, we let \(Y = 0\) and \(Y = 1\) correspond to ’unaffected’ and ’affected’ respectively. Then, the probability function of \(Y\) is given by

\[
P(Y = 0) = 1 - K_p, \quad P(Y = 1) = K_p,
\]

where \(K_p\) is the prevalence of the disease. \(\square\)

\(^2\)Usually, the symbol \(p_x(x) = P(X = x)\) is used for the probability function. In order to avoid too much notation, we will avoid that symbol here.
Example 12 (Binomial distribution and IBD sharing.) A sequence of \( n \) random experiments are conducted. Each experiment is successful with probability \( p \), \( 0 < p < 1 \). Let \( N \) denote the number of successful experiments. Then \( N \) is a discrete random variable with possible values \( 0, 1, \ldots, n \). It can be shown that the probability function is

\[
P(N = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n,
\]

where \( q = 1 - p \) and \( \binom{n}{k} = \frac{n!}{k!(n-k)!} \) is a binomial coefficient (read as '\( n \) choose \( k \)'). The short-hand notation is \( N \in \text{Bin}(n, p) \). The probability function of four different \( \text{Bin}(n, p) \)-distributions are depicted in Figure 2.5.

Two individuals share an allele identical by descent (IBD) if there is a founder in the pedigree that has passed on one of its two alleles to both individuals. Consider a pedigree without inbreeding loops where all founders are unrelated. If the pedigree contains a sib pair, each of the two sibs gets one allele from the father. The probability is 0.5 that these two alleles are IBD, i.e. that they both come from the paternal grandfather or the paternal grandmother respectively. Similarly, the probability is 0.5 that the two alleles passed on to the sibs from the mother are IBD. Let \( N \) be the total number of alleles shared IBD by the sibs. Then \( N \) can be viewed as the number of successes in two experiments, where success means that the parent passes on two
alleles IBD to the sib pair. Since the probability of success is 0.5, it follows that
\( N \sim \text{Bin}(2, 0.5) \), and hence (2.10) implies
\[
\begin{align*}
P(N = 0) &= (1 - 0.5)^2 = 0.25, \\
P(N = 1) &= \binom{2}{1} 0.5(1 - 0.5) = 0.5, \\
P(N = 2) &= 0.5^2 = 0.25.
\end{align*}
\]

IBD sharing of related individuals is the basis of nonparametric linkage analysis, which will be dealt with in Chapter 5.

A random variable \( X \) which can attain all values in an interval such as \([0, 1]\) is not discrete. The reason is that there are uncountably many values that \( X \) can attain along \([0, 1]\). It can be shown that it is impossible to assign positive probability to all outcomes. However, it is possible to define a so called 

**probability density function** instead of a probability function:

**Definition 4 (Probability density functions.)** A random variable is said to be continuous, if there exists a function \( x \to f_X(x) \) such that
\[
P(b < X \leq c) = \int_b^c f_X(x) dx
\]
for all real numbers \( b < c \). The function \( f_X \) is referred to as the probability density function (or just the density function) of \( X \).

By letting \( c \to b \) in (2.12), we find that \( P(X = b) = 0 \) for any number \( b \). This seems as a contradiction. However, there are so many values which \( X \) can attain, so each single number must be assigned zero probability. Only intervals are given positive probabilities. An intuitive characterization of the density function is obtained by noticing that
\[
P(b < X \leq b + h) \approx f_X(b)h
\]
if \( h \) is small. Thus the probability of a small interval around \( b \) is approximately that interval’s length times the density function of \( X \) evaluated at \( b \). We sometimes write \( f \) instead of \( f_X \), when it is clear that the density function of \( X \) is referred to.

**Example 13 (Uniform distribution.)** Let \( b < c \) be two arbitrary numbers. A continuous random variable \( X \) is said to have a uniform distribution on the interval \([b, c]\) if the density function is given by
\[
f_X(x) = \begin{cases} 
 0, & x < b, \\
\frac{1}{c - b}, & b \leq x \leq c, \\
0, & x > c.
\end{cases}
\]
(2.13)
Thus the density function is constant over \([b, c]\) and zero outside, cf. Figure 2.6. The short-hand notation is \( X \sim U(b, c) \).
CHAPTER 2. PROBABILITY THEORY

Example 14 (Normal distribution.) A continuous random variable \( X \) has a normal (Gaussian) distribution if there are real numbers \( \mu \) and \( \sigma > 0 \) such that

\[
    f_X(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right), \quad -\infty < x < \infty. \tag{2.14}
\]

Figure 2.6: Density functions of two different uniform distributions. The dotted vertical lines are shown just to emphasize the discontinuities of the density functions at these points.

Notice that \( f_X \) is symmetric around \( \mu \) and the width of the function around \( \mu \) depends on \( \sigma \). We will find in the next section that \( \mu \) and \( \sigma \) represent the mean value (expected value) and standard deviation of \( X \) respectively. The short-hand notation is \( X \sim N(\mu, \sigma^2) \). The case \( \mu = 0 \) and \( \sigma = 1 \) is referred to as a standard normal distribution \( N(0, 1) \). Figure 2.7 shows the density function of two different normal distributions.

The normal distribution is perhaps the most important distribution in probability theory. One reason for this is that quantities which are sums of many small (independent) contributions, each of which has small individual effect, can be shown to be approximately normally distributed\(^3\).

In genetics, quantitative phenotypes such as blood pressure, body mass index and body weight are often modelled as being normal random variables.

\(^3\)This is a consequence of the so called Central Limit Theorem.
2.2. RANDOM VARIABLES AND DISTRIBUTIONS

Figure 2.7: Density function of two different normal distributions; $N(0, 1)$ and $N(2, 0.5^2)$.

**Example 15 ($\chi^2$-distribution.)** A continuous random variable $X$ is said to have a chi-square distribution with $n$ degrees of freedom, $n = 1, 2, 3, \ldots$, if

$$f_X(x) = \frac{1}{2^{n/2} \Gamma(n/2)} x^{n/2-1} \exp(-x/2), \quad x > 0,$$

where $\Gamma$ is the Gamma function. For positive integers $n$ we have $\Gamma(n) = (n - 1)!$. The short-hand notation is $X \in \chi^2(n)$. Four different chi-square densities are shown in Figure 2.8.

The $\chi^2$-distribution is used in hypothesis testing theory for computing $p$-values and significance levels, cf. Section 3.2 and Chapter 4.

A slight disadvantage of the exposition so far is that discrete and continuous random variables must be treated separately, with either probability functions or density functions being defined. The distribution function on the other hand can be attributed to any random variable:

**Definition 5 (Distribution functions.)** The (cumulative) distribution function (cdf) of any random variable $X$ is defined as

$$F_X(x) = P(X \leq x), \quad -\infty < x < \infty.$$
Figure 2.8: Density functions of four different $\chi^2$-distributions $\chi^2(n)$.

The distribution function $x \rightarrow F_X(x)$ is always non-decreasing, with limits 0 and 1 as $x$ tends to $-\infty$ and $\infty$, respectively. For a continuous random variable, it can be shown that $F_X$ is continuous and differentiable, with derivative $f_X$. For a discrete random variable, $F_X$ is piecewise constant and makes vertical jumps $P(X = x)$ at all points $x$ which $X$ can attain, cf. Figure 2.9.

The basic properties of the cdf can be summarized in the following theorem:

**Theorem 3 (Properties of cdfs.)** The cdf of a random variable $X$ satisfies

\[
F_X(x) \rightarrow 0 \text{ as } x \rightarrow -\infty, \\
F_X(x) \rightarrow 1 \text{ as } x \rightarrow \infty, \\
P(X = x) = \text{vertical jump size of } F_X \text{ at } x.
\]

Further,

\[
F_X(x) = \begin{cases} 
\sum_{y \leq x} P(X = y), & \text{if } X \text{ is a discrete r.v.} \\
\int_{-\infty}^{x} f_X(y) dy, & \text{if } X \text{ is a continuous r.v.,}
\end{cases}
\]

where, in the discrete case, $y$ ranges over the countable set of values that $X$ can attain which are not larger than $x$.

**Example 16 (The cdf of a standard normal distribution.)** Suppose $X$ has a standard normal distribution, i.e. $X \in \mathcal{N}(0, 1)$. Its cumulative distribution function $F_X$
2.2. RANDOM VARIABLES AND DISTRIBUTIONS

Figure 2.9: Cumulative distribution functions for the same four binomial distributions $\text{Bin}(n, p)$ as in Figure 2.5, where the probability functions were plotted instead.

occurs so often in applications that it has been given a special symbol $\Phi$. Thus, by combining (2.14) (with $\mu = 0$ and $\sigma = 1$) and (2.15) we get

\[
\Phi(x) = \int_{-\infty}^{x} f_{X}(y) \, dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2}\right) \, dy. \tag{2.16}
\]

Figure 2.10 shows the cdf of the standard and one other normal distribution. 

Quantiles can conveniently be defined in terms of the cdf. For instance, the median of (the distribution of $X$) is that value $x$ which satisfies $F_X(x) = 0.5$, meaning that the probability is 0.5 that $X$ does not exceed $x$. More generally, we have the following definition:

**Definition 6 (Quantiles.)** Let $0 < \alpha < 1$ be a given number. The the $\alpha$-quantile of (the distribution of) the random variable $X$ is defined as that number $x$ which satisfies

\[
F_X(x) = \alpha, \tag{4}
\]

We tacitly assume that there exists such an $x$. This is not always the case, although it holds for e.g. normal distributions and other continuous random variables with a strictly positive density. A more general definition of quantiles, which covers all kinds of random variables, can be given. This is however beyond the scope of the present monograph.
i.e. the probability is $\alpha$ that $X$ does not exceed $x$.

Figure 2.11 illustrates two quantiles for the standard normal distribution. The choice $\alpha = 0.5$ corresponds to the median of $X$, as noted above. Further, $\alpha = 0.25$ and $0.75$ give the lower and upper quartiles of $X$, respectively.

Often we wish to find the distribution of a random variable $Y$ given the fact that we have observed another random variable $X$. This brings us to the important concept of conditional probability and density functions:

**Definition 7 (Conditional probability and density functions.)** Suppose we have two random variables $X$ and $Y$, of which $X = x$ is observed. If $Y$ is discrete, we define

$$y \rightarrow P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

as the conditional probability function of $Y$ given $X = x$, with $y$ ranging over

---

5 The definition is in fact only strict if $P(X = x) > 0$. Otherwise, we refer to an advanced textbook in probability theory.

6 For the interested reader: We are actually using conditional probabilities for events here. Since $Y = Y(\omega)$ and $X = X(\omega)$ are functions of the outcome $\omega$, (2.17) corresponds to formula (2.5), with events $C = \{ \omega : Y(\omega) = y \}$ and $B = \{ \omega : X(\omega) = x \}$.
the countable sequence of values which \( Y \) can attain\(^7\). If \( Y \) is continuous and has a continuous distribution given \( X = x \) as well, we define the conditional density function \( y \rightarrow f_{Y|X}(y|x) \) of \( Y \) given \( X = x \) through

\[
P(b < Y \leq c|X = x) = \frac{P(b < Y \leq c|X = x)}{P(X = x)} = \int_b^c f_{Y|X}(y|x)dy, \tag{2.18}
\]

which holds for all \( b < c \).

We usually speak of the conditional distribution of \( Y|X = x \), as given by either (2.17) in the discrete case or (2.18) in the continuous case.

**Example 17 (Affected sib pairs, contd.)** Consider a sib pair with both siblings affected by some disease. Given this knowledge, is the distribution of \( N \), the number of alleles shared IBD by the sibs at the disease locus, changed? Without conditioning, the distribution of \( N \) is given by (2.11). However, it is intuitively clear that an affected sib pair is more likely to have at least one allele IBD than a randomly picked sib pair. For instance, for a rare recessive disease, it is probably so that both parents are heterozygous \((Aa)\) whereas both children are \((AA)\)-homozygotes. In that case, both \(A\)-alleles must have been passed on IBD, giving \( N = 2 \).

\(^7\)The usual notation is \( P_{Y|X}(y|x) = P(Y = y|X = x) \) to denote conditional probability functions.
We may formalize this reasoning as follows: If \( Y_1 \) and \( Y_2 \) indicate the disease status of the two sibs (with '0=unaffected' and '1=affected'), our information is that \( Y_1 Y_2 = 1 \). Thus we wish to compute the conditional probability function of \( N \) given that \( Y_1 Y_2 = 1 \). Let us use the acronym ASP (Affected Sib Pair) for \( Y_1 Y_2 = 1 \). Then the sought probabilities are written as

\[
egin{align*}
  z_0 &= P(N = 0 | ASP), \\
  z_1 &= P(N = 1 | ASP), \\
  z_2 &= P(N = 2 | ASP).
\end{align*}
\]

(2.19)

Suarez et al. (1978) have obtained expressions for how \( z_0, z_1, \) and \( z_2 \) depend on the disease allele frequency and penetrance parameters for a monogenic disease. Some examples are given in Table 2.1. As mentioned above, for a fully penetrant recessive model \( (f_0 = f_1 = 0 \) and \( f_2 = 1) \) with a very rare disease allele, it is very likely that an affected sib pair has \( N = 2 \), i.e. that the corresponding probability \( z_2 \) is close to one, as indicated in the second row of Table 2.1.

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
p & f_0 & f_1 & f_2 & z_0 & z_1 & z_2 & E(N|ASP) \\
\hline
0.001 & 0 & 1 & 1 & 0.001 & 0.500 & 0.499 & 251 & 1.498 \\
0.001 & 0 & 0 & 1 & 0.000 & 0.002 & 0.998 & 2.5 \cdot 10^3 & 1.998 \\
0.001 & 0.2 & 0.5 & 0.8 & 0.249 & 0.500 & 0.251 & 1.002 & 1.001 \\
0.1 & 0 & 1 & 1 & 0.081 & 0.491 & 0.428 & 3.08 & 1.346 \\
0.1 & 0 & 0 & 1 & 0.083 & 0.165 & 0.826 & 30 & 1.818 \\
0.1 & 0.2 & 0.5 & 0.8 & 0.223 & 0.500 & 0.277 & 1.12 & 1.054 \\
\hline
\end{array}
\]

Table 2.1: Values of conditional IBD-probabilities \( z_0, z_1, \) and \( z_2 \) in (2.19) and expected number of alleles shared IBD for an affected sib pair. The genetic model corresponds to a monogenic disease with allele frequency \( p \) and penetrance parameters \( f_0, f_1, \) and \( f_2 \). The sibling relative risk \( \lambda_s \) can be computed from \( \lambda_s = 0.25/z_0 \), cf. Risch (1987) and Exercise 2.8.

**Example 18 (Phenotypes conditional on genotypes; quantitative traits.)** For quantitative traits such as body weight or body mass index, it is common to assume that the phenotype varies according to a normal distribution given the genotype. This can be written

\[
Y | G = (aa) \in N(\mu_0, \sigma^2), \quad Y | G = (Aa) \in N(\mu_1, \sigma^2), \quad Y | G = (AA) \in N(\mu_2, \sigma^2).
\]
2.2. RANDOM VARIABLES AND DISTRIBUTIONS

More precisely, this means that the conditional density function is given by

$$f_{Y|G}(y|(aa)) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{1}{2} \left( \frac{y - \mu_0}{\sigma} \right)^2 \right)$$

when $G = (aa)$ and similarly in the other two cases. Thus $\mu_0$, $\mu_1$, and $\mu_2$ represent the mean values of the trait given that the individual has 0, 1 or 2 disease alleles. The remaining random variation can be thought of as being environmentally caused and having standard deviation $\sigma$. Thus $\mu_0$, $\mu_1$, $\mu_2$ are genetically caused penetrance parameters, whereas $\sigma$ is an environmental parameter. The dominant case corresponds to $\mu_1 = \mu_2$ (one disease allele is sufficient to increase the mean level of the phenotype) and $\mu_0 = \mu_1$ (both disease alleles are needed). Figure 2.12 shows the total and conditional densities in the additive case when $\mu_1$ equals the average of $\mu_0$ and $\mu_2$.

Two discrete random variables $Y$ and $X$ are independent if the distribution of $Y$ is unaffected if we observe $X = x$. By (2.17), this means that

$$P(Y = y) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

$$P(Y = y, X = x) = P(X = x)P(Y = y)$$

for all $x$ and $y$.

We will now give an example which involves both independent random variables and random variables that are independent given the fact that we observe some other random variables:

**Example 19 (Marker genotype probabilities.)** Consider a biallelic marker (e.g. a *single nucleotide polymorphism*, SNP) $M$ with possible alleles 1 and 2. The genotype at the marker is thus (11), (12) or (22). Let $p = P(\text{marker allele} = 1)$. Under Hardy-Weinberg equilibrium, the genotype probabilities can be computed exactly as for a disease susceptibility gene, cf. Example 7. Thus

$$P((11)) = p^2, \quad P((12)) = 2p(1-p), \quad P((22)) = (1-p)^2.$$  \hfill (2.21)

Consider the pedigree in Figure 2.13. It has four individuals; two parents and two offspring. Further, all of them are genotyped for the marker, so we can register the genotypes of all pedigree members\(^9\) and put them into a vector $G = (G_1, \ldots, G_4)$. The

---

\(^8\)If this environmental variation is the sum of many small contributions, it is reasonable with a normal distribution.

\(^9\)This is in contrast with disease susceptibility genes, or more generally genes with unknown location on the chromosome. Then only phenotypes can be registered, and usually the genotypes cannot be determined unambiguously from the phenotypes.
Figure 2.12: Density function of $Y$ in Example 18, when the disease allele frequency $p$ equals 0.2 and further $\mu_0 = 0$, $\mu_l = 2$, $\mu_2 = 4$ and $\sigma = 1$ (solid line). Shown in dash-dotted lines are also the three conditional densities of $Y|G = (aa)$ (equals $N(0, 1)$), $Y|G = (Aa)$ (equals $N(2, 1)$) and $Y|G = (AA)$ (equals $N(4, 1)$). These are scaled so that the areas under the curves correspond to the HW proportions $(1 - p)^2 = 0.64$, $2p(1 - p) = 0.32$ and $p^2 = 0.04$.

two parents are founders and the two siblings nonfounders. If we assume that the two parents are listed first, what is the probability of observing $g = ((11), (12), (11), (12))$ under HW equilibrium when $p = 0.4$?

![Pedigree diagram]

$G_1 = (11)$ \hspace{1cm} $G_2 = (12)$

$G_3 = (11)$ \hspace{1cm} $G_4 = (12)$

Figure 2.13: Segregation of a biallelic marker in a family with two parents and two offspring. The probability for this pedigree to have the displayed marker genotypes is calculated in Example 19.
2.2. RANDOM VARIABLES AND DISTRIBUTIONS

We start by writing

\[ P(G = g) = P(G_1 = (11), G_2 = (12)) \cdot P(G_3 = (11), G_4 = (12)|G_1 = (11), G_2 = (12)), \]  

(2.22)

i.e. we condition on the value of the parents' genotypes. Assuming that the two founder genotypes are independent random variables, it follows from (2.20) and (2.21) that

\[ P(G_1 = (11), G_2 = (12)) = P(G_1 = (11))P(G_2 = (12)) = p^2 \cdot 2p(1-p) = 2 \cdot 0.4^3 \cdot 0.6 = 0.0768. \]  

(2.23)

Assume further that the sibling genotype probabilities are determined via Mendelian segregation. We condition on the genotypes of both parents. Given the genotypes of the parents, the genotypes of the two siblings are independent random variables, corresponding to two independent sets of meioses. Thus

\[ P(G_3 = (11), G_4 = (12)|G_1 = (11), G_2 = (12)) = P(G_3 = (11)|G_1 = (11), G_2 = (12))P(G_4 = (12)|G_1 = (11), G_2 = (12)) = 0.5 \cdot 0.5 = 0.25, \]

where the two segregation probabilities \( P(G_3 = (11)|G_1 = (11), G_2 = (12)) = 0.5 \) and \( P(G_4 = (11)|G_1 = (11), G_2 = (12)) = 0.5 \) are obtained as follows: The father (with genotype (11)) always passes on allele 1, whereas the mother can pass on both 1 and 2 with equal probabilities 0.5. Combining the last three displayed equations, we arrive at

\[ P(G = g) = 0.0768 \cdot 0.25 = 0.0192. \]

More generally, if \( n \) discrete random variables \( X_1, \ldots, X_n \) are independent, it follows that

\[ P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n) \]  

(2.24)

for any sequence \( x_1, \ldots, x_n \) of observed values.

For independent continuous random variables \( X_1, X_2, \ldots, X_n \), we must use probabilities instead of intervals. If \( b_1, \ldots, b_n \) are small positive numbers, then

\[ P(X_1 \in [x_1, x_1 + b_1], X_2 = [x_2, x_2 + b_2], \ldots, X_n = [x_n, x_n + b_n]) \approx f_{x_1}(x_1)f_{x_2}(x_2) \cdots f_{x_n}(x_n) b_1 b_2 \cdots b_n, \]  

(2.25)

\[ ^{10} \text{To be precise: we apply (2.20) with } Y = G_1, y = (11), X = G_2 \text{ and } x = (12). \]

\[ ^{11} \text{To be strict, we now generalize (2.20), where only independence of random variables are discussed without any conditioning.} \]

\[ ^{12} \text{The exact definition is actually obtained by replacing the right-hand side of (2.25) by } \int_{x_1}^{x_1+b_1} f_{x_1}(x) dx \cdots \int_{x_n}^{x_n+b_n} f_{x_n}(x) dx \]
i.e. the probability of the vector \((X_1, \ldots, X_n)\) falling into a small box with side lengths \(h_1, \ldots, h_n\) and one corner at \(x = (x_1, \ldots, x_n)\), is approximately equal to the product of the side lengths times the product of the density functions of \(X_1, \ldots, X_n\) evaluated at the points \(x_1, \ldots, x_n\).

### 2.3 Expectation, Variance and Covariance

How do we define an expected value \(E(X)\) of a random variable \(X\)? Intuitively, it is the value obtained 'on average' when we observe \(X\). We can formalize this by repeating the experiment that lead to \(X\) independently many times; \(X_1, \ldots, X_n\). It turns out that by the Law of Large Numbers, the mean value

\[
\frac{X_1 + X_2 + \ldots + X_n}{n}
\]

(2.26)
tends to a well-defined limit as \(n\) grows over all bounds. This limit \(E(X)\) can in fact be computed directly from the probability or density function of \(X\) (cf. Definitions 3 and 4), without needing the sequence \(X_1, X_2, \ldots\):

**Definition 8 (Expected value of a random variable.)** The expected value of a random variable \(X\) is defined as

\[
E(X) = \begin{cases} 
\sum_{x} x P(X = x), & \text{if } X \text{ is a discrete r.v.}, \\
\int_{-\infty}^{\infty} x f_X(x) dx, & \text{if } X \text{ is a continuous r.v.} 
\end{cases}
\]

(2.27)

with \(x\) ranging over the sequence of values that \(X\) can attain in the discrete case.

**Example 20 (Dice throwing.)** A dice is thrown once, resulting in a face with \(X\) eyes. Assuming that all values 1, \ldots, 6 have equal probability, the expected value is

\[
E(X) = \sum_{x=1}^{6} x P(X = x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5.
\]

Figure 2.14 shows that the mean values in (2.26) approach the limit 3.5 as the number of throws \(n\) grows.

**Example 21 (Uniform \((0, 1)\)-distribution.)** Let \(X \in U(0, 1)\) have a uniform distribution on the interval \([0, 1]\). By putting \(b = 0\) and \(c = 1\) in (2.13), it is seen that \(f_X(x) = 1\) when \(x \in [0, 1]\) and \(f_X(x) = 0\) when \(x \notin [0, 1]\) respectively. Thus, it follows from Definition 8, that

\[
E(X) = \int_{0}^{1} x f_X(x) dx = \int_{0}^{1} x \cdot 1 dx = \left[ \frac{x^2}{2} \right]_{0}^{1} = \frac{1^2}{2} - \frac{0^2}{2} = 0.5.
\]

The intuitive result is that \(E(X)\) equals the midpoint of the interval \([0, 1]\).
2.3. EXPECTATION, VARIANCE AND COVARIANCE

Figure 2.14: Mean value \((X_1 + \ldots + X_n)/n\) as a function of \(n\) for 500 consecutive dice throws.

Example 22 (Mean of normal distribution.) If \(X \in \mathcal{N}(\mu, \sigma^2)\) has a normal distribution, then, according to (2.14),

\[
E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma \sqrt{2 \pi}} \exp \left( -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right) \, dx = \ldots = \mu,
\]

where in the last step, we skipped some calculations\(^{13}\) to arrive at what we previously have remarked: \(\mu\) is the expected value of \(X\). This is not surprising, since the density function of \(X\) is symmetric around the point \(\mu\).

It is of interest to know not only the expected value of a random variable, but also a quantity relating to how spread out the distribution of \(X\) is around \(E(X)\). Two such measures are defined as follows:

Definition 9 (Standard deviation and variance.) The variance of a random vari-

\(^{13}\)For the interested reader, we remark \(E(X)\) can be written as \(\int (x - \mu)f_X(x) \, dx + \mu \int f_X(x) \, dx = 0 + \mu \cdot 1 = \mu\), since the integrand \((x - \mu)f_X(x)\) is skew-symmetric around \(\mu\) and therefore must integrate to 0, whereas a density function \(f_X(x)\) always integrates to 1.
CHAPTER 2. PROBABILITY THEORY

A random variable is defined as

\[ V(X) = E[(X - E(X))^2] = \begin{cases} \sum_{x \in \mathbb{X}} (x - E(X))^2 P(X = x), & \text{if } X \text{ is a discrete r.v.} \\ \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x) dx, & \text{if } X \text{ is a continuous r.v.} \end{cases} \]

(2.29)

with \( x \) ranging over the sequence of values that \( X \) can attain in the discrete case. The standard deviation of \( X \) is defined as the square root of the variance, i.e.

\[ D(X) = \sqrt{V(X)}. \]

Notice that \( x - E(X) \) is the deviation of an observed value \( X = x \) from the expected value \( E(X) \). Thus \( V(X) \) can be interpreted as the average (expected value of the) observed squared deviation \((x - E(X))^2\). Since the squared deviation is non-negative for each \( x \), it must also be non-negative on average, i.e. \( V(X) \geq 0 \). Notice however that \( V(X) \) has a different dimension\(^{14}\) than equals the square of \( X \). To get a measure of spread with the same dimension as \( X \), we take the square root of \( V(X) \) and get \( D(X) \).

Example 23 (Variance of a uniform distribution.) The expected value, variance and standard deviation of some distributions are given in Table 2.2. Let us calculate the variance and standard deviation in one particular case; the uniform distribution on \([0, 1] \): We already found in Example 21, that \( E(X) = 0.5 \) when \( X \sim U(0, 1) \). Thus the variance becomes

\[ V(X) = \int_0^1 (x - \frac{1}{2})^2 f_X(x) dx = \int_0^1 (x - \frac{1}{2})^2 dx = \int_0^1 (x^2 - x + \frac{1}{4}) dx = \left[ \frac{x^3}{3} - \frac{x^2}{2} + \frac{x}{4} \right]_0^1 = (1^3/3 - 1^2/2 + 1/4) - (0^3/3 - 0^2/2 + 0/4) = 1/12, \]

and the standard deviation is given by

\[ D(X) = \sqrt{V(X)} = 1/\sqrt{12} = 0.289. \]

Some basic scaling properties of the expected value, variance and standard deviation is given in the following theorem:

**Theorem 4 (Scaling properties of \( E(X) \), \( V(X) \) and \( D(X) \).)** Let \( X \) be a random variable and \( b \) and \( c \) constants. Then

\[ E(bX + c) = bE(X) + c, \]
\[ V(bX + c) = b^2 V(X), \]
\[ D(bX + c) = |b| D(X). \]

\(^{14}\)If for instance \( X \) in measured in cm, then so is \( E(X) \), whereas \( V(X) \) is given in cm\(^2\).
2.3. EXPECTATION, VARIANCE AND COVARIANCE

Table 2.2: Expected value, variance and standard deviation of some distributions.

<table>
<thead>
<tr>
<th>Distribution of $X$</th>
<th>$E(X)$</th>
<th>$V(X)$</th>
<th>$D(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin($n, p$)</td>
<td>$np$</td>
<td>$np(1-p)$</td>
<td>$\sqrt{np(1-p)}$</td>
</tr>
<tr>
<td>$U(b, c)$</td>
<td>$(b+c)/2$</td>
<td>$(c-b)^2/12$</td>
<td>$(c-b)/\sqrt{12}$</td>
</tr>
<tr>
<td>$N(\mu, \sigma^2)$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>$\chi^2(n)$</td>
<td>$n$</td>
<td>$2n$</td>
<td>$\sqrt{2n}$</td>
</tr>
</tbody>
</table>

If a fixed constant, say 50, is added to a random variable $X$ (corresponding to $b = 1$ and $c = 50$ above), it is clear that the expected value of $X + 50$ will increase by 50, whereas the standard deviation of $X + 50$ is the same as that for $X$, since the spread remains unchanged when we add a constant. On the other hand, if we change units of a measurement from meters to centimeters, then $X$ is replaced by $100X$, corresponding to $b = 100$ and $c = 0$ above. It is natural that both the expected value and the standard deviation get multiplied by the same factor 100. The variance on the other hand quantifies squared deviations from the mean and gets multiplied by a factor $100^2 = 10^4$.

**Example 24 (Standardizing a random variable.)** Let $X$ be a random variable with $D(X) > 0$. Then

$$Z = \frac{X - E(X)}{D(X)} \quad (2.30)$$

is referred to as the *standardized random variable* corresponding to $X$. It measures the deviation of $X$ from its expected value on a scale determined by the standard deviation $D(X)$. Observe that

$$E(Z) = D(X)^{-1}E(X) - D(X)^{-1}E(X) = 0,$$
$$D(Z) = D(X)^{-1}D(X) = 1,$$

where we applied Theorem 4 with constants $b = D(X)^{-1}$ and $c = -D(X)^{-1}E(X)$. The canonical example of a standardized random variable is $Z \sim N(0, 1)$, which can be obtained by standardizing any normally distributed random variable $X$ according to (2.30).

In order to check how two random variables $X$ and $Y$ depend on each other, one can compute the conditional distribution of $Y$ given $X = x$ (cf. Definition 7) or analogously the conditional distribution of $X$ given $Y = y$. However, sometimes a single number is preferable as a quantifier of dependence:

**Definition 10 (Covariance and correlation coefficient.)** Given two random variables $X$ and $Y$, the covariance between $X$ and $Y$ is given by

$$C(X, Y) = E [(X - E(X))(Y - E(Y))],$$
whereas the correlation coefficient between $X$ and $Y$ is defined as

$$
\rho(X, Y) = \frac{C(X, Y)}{D(X)D(Y)}.
$$

Figure 2.15: Plots of 100 pairs $(X, Y)$, when both $X$ and $Y$ have standard normal distributions $N(0, 1)$ and the correlation coefficient $\rho = \rho(X, Y)$ varies. Notice that $D(X) = D(Y) = 1$, and hence $C(X, Y) = \rho(X, Y)$ in all four subfigures.

Figure 2.15 shows plots of 100 pairs $(X, Y)$ for four different values of the correlation coefficient $\rho(X, Y)$. It can be seen from these figures that when $\rho(X, Y) > 0$ (and hence also $C(X, Y) > 0$), most pairs $(X, Y)$ tend to get large and small simultaneously. On the other hand, if $\rho(X, Y) < 0$, a large value of $X$ is more often accompanied by a small value of $Y$ and vice versa. If $X$ and $Y$ are independent random variables, there is no preference of $X$ to get large or small when a value of $Y$ is observed (and vice versa). Thus the following result is reasonable:

**Theorem 5 (Independence and correlation.)** Let $X$ and $Y$ be two random variables. If $X$ and $Y$ are independent random variables, then $C(X, Y) = \rho(X, Y) = 0$, but the converse is not true.

Two random variables $X$ and $Y$ are said to be uncorrelated if $C(X, Y) = 0$. Theorem 5 implies that 'non-correlation' is a weaker requirement than independence.
2.3. EXPECTATION, VARIANCE AND COVARIANCE

A disadvantage of the covariance is that $C(X, Y)$ changes when we change units. If for instance $X$ and $Y$ are measured in centimeters instead of meters, the dependency structure between $X$ and $Y$ has not been affected, only the magnitude of the values. However, $C(X, Y)$ gets multiplied by a factor $100 \times 100 = 10^4$, which is not totally satisfactory. Notice however that the product $D(X)D(Y)$, which appears in the denominator of the definition of $\rho(X, Y)$, also gets increased by a factor $10^3 \cdot 10^2$ when we turn to centimeters. Thus the correlation coefficient $\rho(X, Y)$ is a normalized version of $C(X, Y)$ which is unaltered by change of units. The following theorem gives some basic scaling properties of the covariance and the correlation coefficient:

**Theorem 6 (Scaling properties of covariance and correlation.)** Let $X$ and $Y$ be two random variables and $b$, $c$, $d$ and $e$ four given constants. Then

$$C(bX + c, dY + e) = bdC(X, Y)$$

$$\rho(bX + c, dY + e) = \rho(X, Y) \text{ if } b, c > 0.$$  

Finally, it always holds that

$$-1 \leq \rho(X, Y) \leq 1,$$

with $\rho(X, Y) = 1$ if and only if $Y = bX + c$ for some $b > 0$ and $\rho(X, Y) = -1$ if and only if $Y = bX + c$ for some $b < 0$.

The covariance and correlation coefficient measure the degree of linear dependency between random variables $X$ and $Y$. The maximal degree of linear dependency ($\rho = \pm 1$) is attained when $Y$ is a linear function of $X$ and vice versa.

Often, the expected value, variance or standard deviation of sums of random variables is of interest. The following theorem shows how these can be computed:

**Theorem 7 (Expected value and variance for sums of r.v.'s.)** Let $X$, $Y$, $Z$ and $W$ be given random variables. Then

$$E(X + Y) = E(X) + E(Y),$$

$$V(X + Y) = V(X) + V(Y) + 2C(X, Y),$$

$$D(X + Y) = \sqrt{V(X) + V(Y) + 2C(X, Y)},$$


(2.31)

In particular, if $X$ and $Y$ are uncorrelated (e.g. if they are independent), then

$$V(X + Y) = V(X) + V(Y),$$

$$D(X + Y) = \sqrt{V(X) + V(Y)},$$

(2.32)

Notice that the calculation rule for $V(X + Y)$ is analogous to the algebraic addition rule $(x + y)^2 = x^2 + y^2 + 2xy$ whereas $C(X + Y, Z + W)$ corresponds to the
rule \((x + y)(z + w) = xz + xw + yz + yw\). Theorem 7 can easily be extended to cover sums with more than two terms\(^{15}\), although the formulas look a bit more technical.

**Example 25 (Heritability of a disease.)** Continuing Example 18, we might alternatively write the phenotype

\[ Y = X + e \]

as a sum of a genetic component \((X)\) and environmental variation \((e)\). Here \(X\) is a discrete random variable taking values \(\mu_0, \mu_1, \mu_2\) depending on whether the genotype at the disease locus is \((aa)\), \((Aa)\) or \((AA)\). Under HW equilibrium, the probabilities for these three genotypes are \((1 - p)^2\), \(2p(1 - p)\) and \(p^2\), where \(p\) is the disease allele frequency. The environmental variation \(e \in N(0, \sigma^2)\) is assumed normal with mean 0 and variance \(\sigma^2\).

Assuming that \(X\) and \(e\) are independent random variables, it follows from (2.32) that the total phenotype variation is

\[ V(Y) = V(X) + V(e) = V(X) + \sigma^2, \]

where \(V(X)\) can be expressed in terms of \(\mu_0, \mu_1, \mu_2\) and \(p\). The fraction

\[ H = \frac{V(X)}{V(Y)} = \frac{V(X)}{V(X) + \sigma^2} \]

of the total phenotype variance caused by genetic variation is referred to as the *heritability* of the disease. The closer to one \(H\) is, the stronger is the genetic component.

For instance, referring to Figure 2.12, assume that \(p = 0.2\), \(\mu_0 = 0\), \(\mu_1 = 2\), \(\mu_2 = 4\) and \(\sigma^2 = 1\). Then the genotype probabilities under HW equilibrium are \(p^2 = 0.04\), \(2p(1 - p) = 0.32\) and \((1 - p)^2 = 0.64\). Hence, using the definitions of \(E(X)\) and \(V(X)\) in (2.27) and (2.29) we get

\[
\begin{align*}
E(X) &= 0 \cdot 0.64 + 2 \cdot 0.32 + 4 \cdot 0.04 = 0.8, \\
V(X) &= (0 - 0.8)^2 \cdot 0.64 + (2 - 0.8)^2 \cdot 0.32 + (4 - 0.8)^2 \cdot 0.64 = 1.28.
\end{align*}
\]

Finally, the heritability is

\[ H = \frac{1.28}{1.28 + 1} = 0.561. \]

More details on variance decomposition of quantitative traits will be given in Chapter 6. \(\square\)

\(^{15}\)For instance, the sum and variance of a sum of \(n\) random variables is given by \(E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)\) and \(V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} C(X_i, Y_j)\). For pairwise uncorrelated random variables, the latter formula simplifies to \(V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i)\).
The following two formulas are sometimes useful when calculating the variance and covariance:

**Theorem 8 (Two useful calculation rules for variance and covariance.)** Given any two random variables \( X \) and \( Y \), it holds that

\[
\begin{align*}
V(X) &= E(X^2) - E(X)^2, \\
C(X, Y) &= E(XY) - E(X)E(Y).
\end{align*}
\]

**(2.34)**

**Example 26 (Heritability of a disease, contd.)** Continuing Example 25, let us compute the variance \( V(X) \) of the genetic component by means of formula (2.34). Notice first that

\[
E(X^2) = 0^2 \cdot P(X = 0) + 2^2 \cdot P(X = 2) + 4^2 \cdot P(X = 4) = 0^2 \cdot 0.04 + 2^2 \cdot 0.32 + 4^2 \cdot 0.64 = 11.52.
\]

Since \( E(X) = 3.2 \) has already been calculated in Example 25, formula (2.34) implies

\[
V(X) = E(X^2) - E(X)^2 = 11.52 - 3.2^2 = 1.28,
\]

in agreement with (2.33) □

Sometimes, we are just interested in the average behavior of \( Y \) given \( X = x \). This can be achieved by computing the expected value of the conditional distribution in

**Definition 11 (Conditional expectation.)** Suppose we have two random variables \( X \) and \( Y \) of which \( X = x \) is observed. Then, the conditional expectation of \( Y \) given \( X = x \) is defined as

\[
E(Y|X = x) = \begin{cases} 
\sum_y y P(Y = y|X = x), & \text{if } Y|X = x \text{ is a discrete r.v.}, \\
\int_{-\infty}^{\infty} y f_Y|X(y|x) \, dy, & \text{if } Y|X = x \text{ is a continuous r.v.},
\end{cases}
\]

and the summation over \( y \) ranges over the countable set of values that \( Y|X = x \) can attain in the discrete case. □

**Example 27 (Expected number of alleles IBD.)** It was shown in Example 12 that \( N \), the number of alleles shared IBD by a randomly picked sib pair, had a binomial distribution. It follows from (2.11) that the expected number of IBD alleles is one, since\(^{16}\)

\[
E(N) = 0 \cdot P(N = 0) + 1 \cdot P(N = 1) + 2 \cdot P(N = 2) = 0 \cdot 0.25 + 1 \cdot 0.5 + 2 \cdot 0.25 = 1.
\]

\(^{16}\)Alternatively, since \( N \in \text{Bin}(2, 0.5) \), we just look at Table 2.2 to find that \( E(N) = 2 \cdot 0.5 = 1.\)
CHAPTER 2. PROBABILITY THEORY

What is then the expected number of alleles shared IBD by an affected sib pair? The IBD distribution for an affected sib pair (ASP) was formulated as a conditional distribution in (2.19), and thus from Definition 11 we get

$$E(N|\text{ASP}) = 0 \cdot P(N = 0|\text{ASP}) + 1 \cdot P(N = 1|\text{ASP}) + 2 \cdot P(N = 2|\text{ASP}).$$

Values of this conditional expectation are given in the last column of Table 2.1 for different genetic models. The stronger the genetic component is, the closer to 2 is the expected number of alleles IBD for an affected sib pair.

Recall that the Law of total probability (2.7) was used for calculating probabilities of events when the conditional probabilities given a number of other events were given beforehand. In the same way, it is often the case that the expected value of a random variable $Y$ is easier to calculate if first the conditional expectation given some other random variable $X$ is computed. This is described in the following theorem:

**Theorem 9 (Expected value via conditional expectation.)** The expected value of a random variable $Y$ can be computed by conditioning on the outcome of another random variable $X$ according to

$$E(Y) = \begin{cases} \sum_{x \in X} E(Y|X = x)P(X = x), & \text{if } X \text{ is a discrete r.v.}, \\ \int_{-\infty}^{\infty} E(Y|X = x)f_X(x)dx, & \text{if } X \text{ is a continuous r.v.}, \end{cases}$$

where the summation ranges over the countable set of values that $X$ can attain in the discrete case.

We illustrate Theorem 9 by computing the expected value of a quantitative phenotype.

**Example 28 (Expectation of a quantitative phenotype.)** Consider the quantitative phenotype $Y = X + e$ of Example 25 for a randomly chosen individual in a population. Assuming the same model parameters as in Figure 2.12, the genetic component $X$ equals $\mu_0 = 0$, $\mu_1 = 2$ and $\mu_2 = 4$ for an individual with genotype (aa), (Aa) and (AA) respectively. Under Hardy-Weinberg equilibrium, and if the disease allele frequency $p$ is 0.2, the expected value of $Y$ can be obtained from (2.35) by means of

$$E(Y) = E(Y|X = 0) \cdot P(X = 0) + E(Y|X = 2) \cdot P(X = 2)$$

$$+ E(Y|X = 4) \cdot P(X = 4)$$

$$= 0 \cdot (1-p)^2 + 2 \cdot 2p(1-p) + 4 \cdot p^2$$

$$= 0 \cdot 0.8^2 + 2 \cdot (2 \cdot 0.2 \cdot 0.8) + 4 \cdot 0.2^2$$

$$= 0.8.$$

For the conditional expectations, we reasoned as follows: The conditional distribution of $Y$ given $X = 4$ is $N(4, \sigma^2) = N(4, 1)$, and hence $E(Y|X = 4) = 4$. Similarly one has $E(Y|X = 0) = 0$ and $E(Y|X = 2) = 2$. \qed
2.4 Exercises

2.1. The probability of the union of two events $B$ and $C$ can be derived by means of

$$ P(B \cup C) = P(B) + P(C) - P(B \cap C). $$

The rationale for this formula can be seen from Figure 2.1 a). When summing $P(B)$ and $P(C)$ the area $P(B \cap C)$ is counted twice, and this must be compensated for by subtracting $P(B \cap C)$. Suppose $P(B) = 0.4$ and $P(C) = 0.5$. Compute $P(B \cup C)$ if

(a) $B$ and $C$ are disjoint.
(b) $B$ and $C$ are independent.

2.2. In Exercise 2.1, compute

(a) $P(B*)$
(b) $P(B^* \cup C)$ if $B$ and $C$ are independent. (Hint: If $B$ and $C$ are independent, so are $B^*$ and $C$.)

2.3. A proportion 0.7 of the individuals in a population are homozygotes ($aa$), i.e. have no disease allele. Further, a fraction 0.1 are homozygotes ($aa$) and affected. Compute the phenocopy rate $f_0^*$ in equation (2.4).

2.4. Compute the probability of a heterozygote ($Aa$) under HW-equilibrium if the disease allele frequency is 0.05.

2.5. Consider a monogenic disease with disease allele frequency $p = 0.05$ and penetrance probabilities $f_0 = 0.08$, $f_1 = 0.6$ and $f_2 = 0.9$ in (2.4). Compute, under HW-equilibrium,

(a) the probability that a randomly chosen individual is affected and has $i$ disease alleles, $i = 0, 1, 2$,
(b) the conditional probability that an affected individual is a heterozygote ($Aa$).

2.6. A random variable $N$ has distribution Bin(2, 0.4). Compute $P(N = 1)$.

2.7. A continuous random variable $X$ has density function

$$ f(x) = \begin{cases} 
0, & x < 0, \\
2x, & 0 \leq x \leq 1, \\
0, & x > 1.
\end{cases} $$

Plot the density function and evaluate $P(X < 0.6)$.  

2.8. Consider Example 17. We will find a formula for \( z_0 = P(N = 0|\text{ASP}) \) in terms of the sibling relative risk \( \lambda_s \).

(a) Compute \( P(\text{ASP}) \) in terms of \( \lambda_s \) and the prevalence \( K_p \).
(b) Compute \( P(N = 0, \text{ASP}) \) in terms of \( K_p \). (Hint: \( P(N = 0, \text{ASP}) = P(N = 0)P(\text{ASP}|N = 0) \)).
(c) Give an expression for \( z_0 \) in terms of \( \lambda_s \).

2.9. A dice is thrown twice. Let \( X_1 \) and \( X_2 \) be the outcomes of the two throws and \( Y = \max(X_1, X_2) \). Assume that the two throws are independent and compute

(a) the probability distribution for \( Y \). (Hint: There are 36 possible outcomes \((X_1, X_2)\). Check how many of these that give \( Y = 1, \ldots, 6 \).
(b) \( E(Y) \),
(c) the probability function for \( Y|X_1 = 5 \),
(d) \( E(Y|X_1 = 5) \).

2.10. Compute the expected value, variance and standard deviation of the random variable \( X \) in Exercise 2.7.

2.11. (Before doing this exercise, read through Example 25.) Consider a sib pair with two alleles IBD. The values of a certain quantitative trait for the sibs are \( Y_1 = X + e_1 \) and \( Y_2 = X + e_2 \), where the genetic component \( X \) is the same for both sibs and \( e_1 \) and \( e_2 \) are independent environmental components. Assume that \( V(e_1) = V(e_2) = 4 \) and that the heritability \( H = 0.3 \). Compute

(a) \( V(Y_1) = V(Y_2) \). (Hint: Use \( H = V(X)/V(Y_1) \) and the fact that \( V(Y_1) = V(X) + V(e_1) \)).
(b) \( C(Y_1, Y_2) \). (Hint: Use formula (2.31) for the covariance, and then Theorem 5.)
(c) \( \rho(Y_1, Y_2) \).
Chapter 3

Inference Theory

3.1 Statistical Models and Point Estimators

Statistical inference theory uses probability models to describe observed variation in data from real world phenomena. In general, any conclusions drawn are only valid within the framework of the assumptions used when formulating the mathematical model.

This is formalized using a statistical model: The observed data is typically a sequence of numbers, say \( x = (x_1, \ldots, x_n) \). We assume that \( x_i \) is an observation of a random variable \( X_i \), \( i = 1, \ldots, n \). The distribution of \( X = (X_1, \ldots, X_n) \) depends on an unknown parameter \( \psi \in \Psi \), where \( \Psi \) is the parameter space, i.e. the set of possible values of the parameter. The parameter \( \psi \) represents the information that we wish to extract from the experiment.

**Example 29 (Coin tossing.)** Suppose we flip a coin 100 times, resulting in 61 heads and 39 tails. Let \( \psi \) be the (unknown) probability of head. For a symmetric coin, we would put \( \psi = 0.5 \). Suppose instead that \( \psi \in [0, 1] \) is an unknown number between 0 and 1. We put \( n = 100 \) and let \( x_i \) be the result of the \( i \)th throw, with \( x_i = 1 \) if head occurs and \( x_i = 0 \) if tail does. Then \( x_i \) is an observation of \( X_i \), having a two point distribution, with probability function

\[
P(X_i = 0) = 1 - \psi, \quad P(X_i = 1) = \psi.
\]

A convenient way to analyze an experiment is to compute the likelihood function \( \psi \rightarrow L(\psi) \), where \( L(\psi) \) quantifies how likely the observed sequence of data is. It is defined a bit differently for discrete and continuous random variables:
**Definition 12 (Likelihood function, independent data.)** Suppose \( X_1, \ldots, X_n \) are independent random variables. Then, the likelihood function is defined as

\[
L(\phi) = \begin{cases} 
\prod_{i=1}^n P(X_i = x_i), & \text{if } X_1, \ldots, X_n \text{ are discrete r.v.'s} \\
\prod_{i=1}^n f_X(x_i), & \text{if } X_1, \ldots, X_n \text{ are continuous r.v.'s.}
\end{cases}
\]  

(3.1)

In the discrete case, it follows from (2.24) that \( L(\phi) \) is the probability \( P(X = x) \), i.e. the probability of observing the whole sequence \( x_1, \ldots, x_n \). This value depends on \( \phi \), which is unknown\(^1\). Therefore, one usually plots the function \( \phi \rightarrow L(\phi) \) to see which parameter values that are more or less likely to correspond to the observed data set. In the continuous case, it follows similarly from (2.25) that \( L(\phi) \) is proportional to having the observed value of \( X \) in a small surrounding of \( x = (x_1, \ldots, x_n) \).

**Example 30 (Coin tossing, contd.)** The likelihood function for the coin tossing experiment of Example 29 can be computed as \( L(\phi) = (1 - \phi)^{39} \phi^{61} \), since there are 39 factors \( P(X_i = 0) = (1 - \phi) \) and 61 factors \( P(X_i = 1) = \phi \). A more formal way of deriving this is

\[
L(\phi) = \prod_{i=1}^{100} P(X_i = x_i) = \prod_{i=1}^{100} (1 - \phi)^{1-x_i} \phi^{x_i} = (1 - \phi)^{100-x_i} \phi^{x_i},
\]  

(3.2)

since \( \sum_{i=1}^{100} x_i = 61 \) is the total number of heads.

A point estimator \( \hat{\phi} = \hat{\phi}(x) \) is a function of the data set which represents our ‘best guess’ of \( \phi \), given the information we have from data and assuming the statistical model to hold. A very intuitive choice of \( \hat{\phi} \) is to use the parameter value which maximizes the likelihood function, i.e. the \( \phi \) that most likely would generate the observed data vector:

**Definition 13 (Maximum likelihood estimator.)** The maximum likelihood (ML) estimator is defined as

\[
\hat{\phi} = \arg \max_{\phi \in \Phi} L(\phi),
\]

meaning that \( \hat{\phi} \) is the parameter value which maximizes \( L \).

If \( L \) is differentiable, a natural procedure to find the ML estimator would be to check where the derivative \( L' \) of \( L \) w.r.t. \( \phi \) equals zero. Notice however that if \( \hat{\phi} \) maximizes \( L \) it also maximizes the log likelihood function \( \ln L \) and it is often more convenient to differentiate \( \ln L \), as the following example shows:

\(^1\)Often, one writes \( P(X = x \mid \phi) \), to highlight that the probability of observing the data set at hand depends on \( \phi \). This can be interpreted as conditioning on \( \phi \), i.e. the probability of \( X = x \) given that \( \phi \) is the true parameter value. This should not be confused with (2.2), where we condition on random events.
Example 31 (ML-estimator for coin tossing.) If we take the logarithm of (3.2) we get
\[
\ln L(\psi) = 39 \ln(1 - \psi) + 61 \ln \psi.
\]
This function is shown in Figure 3.1. Differentiating this w.r.t. and putting the derivative to zero we get
\[
0 = \frac{d \ln L'(\psi)}{d \psi} \bigg|_{\psi = \hat{\psi}} = \frac{61}{\hat{\psi}} - \frac{39}{1 - \hat{\psi}} \iff \hat{\psi} = \frac{61}{100}.
\]
The ML-estimator of \(\psi\) is thus very reasonable; the relative proportion of heads obtained during the throws.

Figure 3.1: Log likelihood function \(\ln L\) for the coin tossing problem of Example 31. The ML-estimator is indicated with a vertical dotted line.

Estimation of disease allele frequencies and penetrance parameters is the subject of segregation analysis. It can be done by maximum likelihood, although the likelihood functions are quite involved. For instance, for a monogenic disease with binary responses (as in Example 6), the parameter vector to estimate is \(\psi = (p, f_0, f_1, f_2)\), where \(p\) is the disease allele frequency and \(f_0, f_1, f_2\) the penetrances.

In contrast, at markers, only the allele frequencies need to be estimated, since the genotypes are observed directly, not indirectly via phenotypes. Estimation of marker allele frequencies is important, since it is used both in parametric and nonparametric linkage analysis.
Example 32 (Estimating marker allele probabilities.) Consider a data set with 100 pedigrees. We wish to estimate the allele frequency \( p = P(\text{‘allele 1’}) \) for a biallelic marker with possible alleles 1 and 2. We assume that all founders are being typed for the marker and that the total number of founders with genotypes (11), (12) and (22) are 181, 392 and 240 respectively.

In order to write an expression for the likelihood function \( p \to L(p) \) (\( p \) is the unknown parameter), we introduce some notation: Let \( G_i \) denote the collection of genotypes for the \( i \)-th pedigree, and \( G_i^{\text{founder}} \) and \( G_i^{\text{nonfounder}} \) the corresponding subsets for the founders and non-founders. Then, assuming that the genotypes of different pedigrees are independent, we have

\[
L(p) = \prod_{i=1}^{100} P(G_i) \\
= \prod_{i=1}^{100} P(G_i^{\text{founder}} \cdot P(G_i^{\text{nonfounder}} | G_i^{\text{founder}}) \\
= \prod_{i=1}^{100} P(G_i^{\text{founder}}) \cdot \prod_{i=1}^{100} P(G_i^{\text{nonfounder}} | G_i^{\text{founder}}),
\]

where for each pedigree we conditioned on the genotypes of the founders and divided \( P(G_i) \) into two factors, as in (2.22). In the last equality, we simply rearranged the order of the factors. Each \( P(G_i^{\text{nonfounder}} | G_i^{\text{founder}}) \) only depends on Mendelian segregation, not on the allele frequency. Thus we can regard \( C = \prod_{i=1}^{100} P(G_i^{\text{nonfounder}} | G_i^{\text{founder}}) \) as a constant, independent of \( p \). As in (2.23), we further assume that all founder genotypes in a pedigree are independent. Under Hardy-Weinberg equilibrium, this means that each \( P(G_i^{\text{founder}}) \) is a product of genotype probabilities (2.21). The total number of founder genotype probabilities of the kind \( P((11)) = p^2 \) is 181, and similarly for 392 and 240 of the kind \( P((12)) = 2p(1 - p) \) and \( P((22)) = (1 - p)^2 \) respectively. Thus

\[
L(p) = C \prod_{i=1}^{100} P(G_i^{\text{founder}}) \\
= C(p^2)^{181}(2p(1 - p))^{392}(1 - p)^{240} \\
= 2^{392} \cdot C \cdot p^{754}(1 - p)^{872},
\]

where 754 and 872 is the total number of founder marker alleles of type 1 and 2 respectively. Since \( 2^{392} \cdot C \) is a constant not depending on \( p \), we can drop it when maximizing \( L(p) \). Then, comparing with (3.2), we have a coin tossing problem with 754 heads and 872 tails. Thus, the ML estimator of the allele frequency is the ‘relative proportion of heads’ i.e.

\[
\hat{p} = \frac{754}{754 + 872} = 0.4637.
\]

Our example is a bit over-simplified in that we required all founder genotypes to be known. This is obviously not realistic for large pedigrees with many generations.

\[\text{\footnote{A more strict notation would be } P(G_i = g_i), \text{ where } g_i \text{ is the observed set of genotypes for the } i \text{-th pedigree.}}\]
3.2. HYPOTHESIS TESTING

Still, one can estimate marker allele frequencies by means of relative allele frequencies among the genotyped founders. This is no longer the ML-estimator though if there are untyped founders, since we do not make use of all data (we can extract some information about an untyped founder genotype from the non-founders in the same pedigree).

The advantage of the ML estimator is its great generality; it can be defined as soon as a likelihood function exists. It also has good properties for most models when the model is specified correctly. However, a disadvantage of ML-estimation is that misspecification of the model may result in poor estimates.

3.2 Hypothesis Testing

Hypothesis testing refers to testing the value of the parameter \( \psi \) in a statistical model given data. For instance, in the coin tossing Example 29, we might ask whether or not the coin is symmetric. This corresponds to testing a null hypothesis \( H_0 \) (the coin is symmetric, \( \psi = 0.5 \)) against an alternative hypothesis \( H_1 \) (the coin is not symmetric, i.e. \( 0 < \psi < 1 \) but \( \psi \neq 0.5 \)). More generally we formulate the testing problem as

\[
H_0 : \psi \in \Psi_0, \\
H_1 : \psi \in \Psi_1 = \Psi \setminus \Psi_0,
\]

where \( \Psi_0 \subset \Psi \) is a subset of the parameter space and \( \Psi_1 = \Psi \setminus \Psi_0 \) consists of all parameters in \( \Psi \) but not in \( \Psi_0 \). If \( \Psi_0 \) consists on one single parameter (as in the coin tossing problem), we have a simple null hypothesis. Otherwise, we speak of a composite null hypothesis.

How do we, based on data, decide whether or not to reject \( H_0 \)? In the coin tossing problem we could check if the proportion of heads is sufficiently close to 0.5. In order to specify what ‘sufficiently close’ means, we need to construct a well-defined rule when to reject \( H_0 \). In general this can be done by defining a test statistic \( T = T(X) \), which is a function of the data vector \( X \). The test statistic is then compared to a fixed threshold \( t \), and \( H_0 \) is rejected for values of the test statistic exceeding \( t \), i.e.

\[
T(X) \geq t \quad \Rightarrow \quad \text{reject } H_0, \\
T(X) < t \quad \Rightarrow \quad \text{do not reject } H_0.
\]

We will now give an example of a test for allelic association between a marker and a trait locus. A more detailed treatment of association analysis is given in Chapter 7.

Example 33 (The Transmission Disequilibrium Test.) Consider segregation of a certain biallelic marker with alleles 1 and 2. To this end, we have a number of trios consisting of two parents and one affected offspring where all pedigree members have been genotyped. Among all heterozygous parents we register how many
times allele 1 has been transmitted to the offspring. We may then test allelic association between the disease and marker locus by checking if the fraction of transmitted 1-alleles significantly deviates from 0.5.

For instance, suppose there are 100 heterozygous parents, and let \( \psi \) denote the probability that allele 1 is transmitted from the parent to the affected child. It can be shown that the hypotheses \( H_0 : \text{ 'no allelic association' } \) versus \( H_1 : \text{ 'an allelic association is present' } \) can be formulated as

\[
H_0 : \psi = 0.5, \\
H_1 : \psi \neq 0.5.
\]

Let \( N \) be the number of times marker allele 1 is transmitted. The transmission disequilibrium test (TDT) was introduced by Spielman et al. (1993). It corresponds to using a test statistic

\[
T = |N - 50|,
\]

and large values of \( T \) result in rejection of \( H_0 \). With threshold \( t = 10 \) we reject \( H_0 \) when

\[
T \geq 10 \iff N \leq 40 \text{ or } N \geq 60.
\]

Now \( N \) has a binomial distribution \( \text{Bin}(100, \psi) \), since it counts the number of successes (= allele 1 being transmitted) in 100 consecutive independent experiments, with the probability of success being \( \psi \). In fact, the hypothesis testing problem is identical to registering a coin that is tossed 100 times and testing whether or not the coin is symmetric (with \( \psi = \text{ 'probability of heads' } \)). The probability of rejecting the null hypothesis even though it is true is referred to as the significance level \( \alpha \) of the test. Since \( H_0 \) corresponds to \( \psi = 0.5 \), we have \( N \in \text{Bin}(100, 0.5) \) under \( H_0 \). Therefore the significance level

\[
\alpha = P(N \leq 40|H_0) + P(N \geq 60) = 0.0569,
\]

a value that can be obtained from a standard computer package. The set of outcomes which correspond to rejection of \( H_0 \) are drawn with black bars in Figure 3.2.

Obviously, we can control the significance level \( \alpha \) by our choice of threshold. For instance, if \( t \) is increased from 10 to 15 in the coin tossing problem, the significance level drops down to \( \alpha = 0.0035 \). A lower significance level corresponds to a safer test, since more evidence is required to reject \( H_0 \). There is, however, never a free lunch, so a safer test implies, on the other hand, that it is more difficult to detect \( H_1 \) when it is actually true. This is reflected in the power function.

---

\( ^3 \)The most frequently used test statistic of the TDT is actually a monotone transformation of \( T \), cf. equation (3.6) below.

\( ^4 \)This is achieved by summing over probabilities \( P(N = x) \) in (2.10), with \( n = 100 \) and \( p = q = 0.5 \).
3.2. **HYPOTHESIS TESTING**

Figure 3.2: Probability function of \( N \) under \( H_0 \) (\( N \in \text{Bin}(100, 0.5) \)). The black bars correspond to rejection of \( H_0 \) and their total area give the significance level 0.0569.

**Definition 14 (Significance level and power.)** Consider an hypothesis test of the form (3.3). Then the **significance level** of the test is defined as

\[
\alpha = P(T \geq t | H_0),
\]

(3.4)

provided the distribution of \( T \) is independent of which particular \( \psi \in \Psi_0 \) applies\(^5\). The **power function** is a function of the parameter \( \psi \) and is defined as

\[
\beta(\psi) = P(T \geq t | \psi),
\]

i.e. the probability of rejecting \( H_0 \) given that \( \psi \) is the true parameter value. \( \square \)

Figure 3.3 shows the power function for the binomial experiment in Example 33 for two different thresholds. As seen from the figure, the lower threshold \( t = 10 \) gives a higher significance level (\( \beta(0.5) = 0.0569 \)) but also a higher power for all \( \psi \neq 0.5 \).

Significance levels often used in practice are, depending on the application, 0.05, 0.01 and 0.001. The outcome of a test is referred to as **statistically significant** at the level \( \alpha \) when \( H_0 \) is rejected.

An alternative to specifying a significance level in advance is to compute a **p-value**. This can be thought of as the significance level achieved by data, i.e. the probability, under \( H_0 \), of observing a test statistic at least as large as the one we actually observed.

---

\(^5\)This condition is always satisfied for a simple null hypothesis, since then \( \Psi_0 \) contains just one parameter value.
Figure 3.3: Power of TDT for 100 parent-child pairs as a function the probability \( \psi \) that allele 1 is transmitted. The null hypothesis \( H_0 : \psi = 0.5 \) is rejected when \( |N - 50| \geq t \), with \( N \) the number of transmitted 1-alleles.

**Definition 15 (p-value.)** Let \( \mathbf{x} \) be the observed value of the data vector for a test of the form (3.3). Then the \( p \)-value is defined as

\[
\alpha(\mathbf{x}) = P(T(\mathbf{X}) \geq T(\mathbf{x})|H_0),
\]

i.e. the significance level that would be obtained with a threshold \( t = T(\mathbf{x}) \).

**Example 34 (TDT, contd.)** Suppose that marker allele 1 is transmitted 38 times among the 100 parent-child pairs of Example 33. This corresponds to a threshold \( t = |38 - 50| = 12 \). The \( p \)-value is

\[
P(T \geq 12|H_0) = P(N \leq 38|H_0) + P(N \geq 62|H_0) = 0.021.
\]

The probability is thus 2.1\% that by chance the number of transmitted 1-alleles deviates from 50 with at least 12 when no allelic association is present.

Just as the likelihood function (3.1) was a convenient tool for defining the ML-estimator in the previous section, we will now demonstrate how it can be used in hypothesis testing:
3.2. HYPOTHESIS TESTING

Definition 16 (Likelihood ratio tests.) Consider a statistical model, with a likelihood function \( L(\phi) \) defined as in (3.1). Then the likelihood ratio (LR) test uses a test statistic \( T = \text{LR} \), where

\[
\text{LR} = \frac{\max_{\phi \in \Psi} L(\phi)}{\max_{\phi \in \Psi_0} L(\phi)}.
\]

is the ratio of the likelihood for the most likely parameter divided by the likelihood for the most likely parameter under \( H_0 \).

It can be shown that the TDT is in fact a likelihood ratio test.

A test is said to be powerful when its power to detect an alternative \( \phi \in \Psi_1 \) is high, given a certain restriction on the significance level\(^6\). The power of a test increases when we collect more data, but it also depends on how well the test utilizes the information present in data. The LR test often has very good properties in terms of power. Just as for the ML-estimator, its performance can drastically decrease though when the statistical model is not correctly specified.

In order to compute significance levels or \( p \)-values, we must know the distribution of the test statistic under \( H_0 \), cf. (3.4) and (3.5). In statistical genetics applications the test statistic is often so complicated that computer simulation is needed for this. Sometimes asymptotic methods, which are valid for large data sets, can be used instead. For instance, suppose the parameter vector \( \psi \) has \( f \) 'free components' when it varies over \( \Psi_1 \), the alternative hypothesis. Then it can be shown, under certain regularity conditions, that twice the log likelihood ratio

\[
\Lambda = 2 \ln \text{LR},
\]

has approximately a \( \chi^2(f) \)-distribution (cf. Example 15)\(^7\). In linkage analysis (Chapters 4 and 5), it is common practice to replace \( \Lambda \) by the so called lod score log LR = \( \Lambda/(2 \ln(10)) \). Here log refers to the base 10 logarithm.

Example 35 (TDT, contd.) It is more common to replace the test statistic \( T = |N - 50| \) by\(^8\) TDT = \((N - 50)^2/25\). More generally, with a sample of \( n \) parent-affected child pairs, where all the parents are marker heterozygotes, one has

\[
\text{TDT} = \frac{(N - 0.5n)^2}{0.25n},
\]  

(3.6)

---

\(^6\)We recall that the power can always be increased by increasing the significance level. To get an objective performance measure of the test, we must keep the significance level fixed.

\(^7\)Since \( \Lambda \) is a monotone transformation of LR, the two test statistics give equivalent test if we transform the threshold \( t \) by the same transformation.

\(^8\)Again, the two tests statistics are related through a monotone transformation and yield equivalent tests.
and it can be shown that TDT has approximately a $\chi^2(1)$-distribution under $H_0$ when $n$ is large. For instance, with $N = 38$ observed, as in Example 34, an approximate $p$-value based on the $\chi^2(1)$-approximation is\footnote{Here we replace 38 by 38.5 in the formula since $N$ has a discrete distribution but not $\chi^2(1)$. This is a so called half-correction, cf. a textbook in probability theory for details.}

$$P(\chi^2(1) \geq (38.5 - 50)^2/25) = 0.0214.$$ Comparing this with the exact $p$-value in Example 34, we find that the approximation is good. 

The TDT given above is an example of a method used for association analysis. Also in linkage analysis, hypothesis testing is central. The objective is then to test whether or not the disease locus is located on a certain chromosome ($H_1$) or not ($H_0$). If $H_0$ is rejected, the next step is to estimate the position of the disease locus as well as possible. In parametric linkage analysis, it is usually assumed that disease allele frequencies and penetrance parameters have been estimated beforehand by means of methods from segregation analysis. The test is then carried out using (the base 10 logarithm of) a likelihood ratio, i.e. the lodscore. In contrast, nonparametric linkage analysis uses test statistics based on excess allele sharing identical by descent among affected individuals in the same pedigree. No genetic model needs to be specified, and this is an advantage for complex diseases such as inheritable diabetes and psychiatric disorders.

### 3.3 Linear regression

Depending on the causal connections between two variables, $X$ and $Y$, their true relationship may be linear or nonlinear. In any case, a linear model can always be used as a first approximation to the true pattern of association. Assume that the conditional mean of $Y$, i.e. the expected value of $Y$ given $X = x$, $E(Y|x)$, is a linear function of $x$,

$$Y = E(Y|x) + (Y - E(Y|x)) = E(Y|x) + e = \alpha + \beta x + e$$

(3.7)

where $\alpha$ is the $y$-intercept, $\beta$ is the slope of the line, the regression coefficient, and $e$ is the residual error. The mean of $e$ is zero by construction. The residual error is the deviation of $Y$ from the regression line, $\alpha + \beta x$. Even if the relationship between $X$ and $Y$ is truly linear, deviations from this straight-line relation can nevertheless be observed in data due to e.g. measurement error. Furthermore, the true values of $\alpha$ and $\beta$ are generally not known in this situation and have to be estimated from sampled...
3.3. LINEAR REGRESSION

data. In least squares linear regression estimation of $\alpha$ and $\beta$ is based on minimization of the sum of squared residuals,

$$\sum_{i} e_{i}^{2} = \sum_{i} (y_{i} - \hat{y}_{i})^{2};$$

where summation is over observed pairs of data, $(x_{i}, y_{i})$, and $\hat{y}_{i}$ is the predicted response or fitted value,

$$\hat{y}_{i} = \hat{\alpha} + \hat{\beta}x_{i}.$$

That is, the least squares solution yields estimates, $\hat{\alpha}$ and $\hat{\beta}$, that minimizes the average value of the squared vertical deviations of the observed $y$'s from the values predicted by the regression line, cf. Figure 3.4. It has the useful property of maximizing the

![Figure 3.4: Least-squares linear regression of $y$ on $x$. Fitted line and residual deviations. The open circles correspond to the true values of the response $y$.](image)

amount of variance in $y$ that can be explained by a linear dependence on $x$ and is given by,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (3.8)$$

and

$$\hat{\beta} = \frac{C(x, y)}{V(x)}, \quad (3.9)$$

where $\bar{x} = (\sum x_{i})/n$ is the mean of $x$ and $\bar{y} = (\sum y_{i})/n$ is the mean of $y$. 
Some important properties of least squares regression

1. The regression line passes through the means of both $x$ and $y$, i.e. for $x = \bar{x}$ we have $\hat{y} = \bar{y}$.

2. The average value of the residuals is zero meaning that $\sum \hat{e}_i = 0$.

3. The residual errors are uncorrelated with the predictor variable, $x$, and therefore also uncorrelated with the predicted values, $\hat{y}$, i.e. $C(x, \hat{e}) = C(\hat{y}, \hat{e}) = 0$.

4. The least squares solution, $\hat{\alpha}$ and $\hat{\beta}$, maximizes the amount of variation in $y$ that can be explained by a linear regression on $x$.

It follows that the total variation in $y$ can be split in two parts or components of variance: variation due to regression on $x$ plus residual variation,

$$V(y) = V(\hat{y} + \hat{e}) = V(\hat{y}) + V(\hat{e})$$  \hspace{1cm} (3.10)

where $V(\hat{y}) = \rho^2(x, y)V(y)$ and, consequently, $V(\hat{e}) = (1 - \rho^2(x, y))V(y)$. Here the squared correlation coefficient of $x$ and $y$, $\rho^2(x, y)$, can be interpreted as the proportion of the total variance in $y$ that is explained by linear dependence on $x$.

The decomposition of $V(y)$ in (3.10) is of central importance for the modelling of genetic effects in quantitative traits, see Chapter 6.

**Example 36 (Regression on number of disease alleles.)** In Example 18 the mean of $y$ varies as a function of the number of disease alleles at a biallelic locus. Whether or not this relation is truly linear we can always, as a first approximation at the least, fit a straight line to describe the dependence. In Chapter 6 we will see that this particular regression leads to a subdivision of the total genetic variance due to the disease locus into two component parts: the additive and dominant components of variance, respectively.

In order to be able to make statements in terms of statistical significance concerning the regression parameters we need to impose assumptions on the probability distribution of the data. For example, within the framework of the linear model we might be interested in testing whether $y$ and $x$ are in fact unrelated, $H_0 : \beta = 0$. The most commonly used set of assumptions, which of course have to be validated in each separate application of the model, considers the $y$’s to be uncorrelated and normally distributed with constant variance that is independent of $x$. It can be shown that under these assumptions the maximum-likelihood estimates of $\alpha$ and $\beta$ coincide with the least-squares estimates in (3.8) and (3.9).
3.4. EXERCISES

In many situations it is necessary to consider the joint impact of several predictor variables $x_1, x_2, \ldots, x_p$ on the outcome or response $y$. The model in (3.7) can be extended to allow for more than one covariate $x$ in a straightforward way:

$$
y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e. \tag{3.11}$$

The expression in (3.11) is called a multivariate linear regression and again it is easy to calculate the least-squares estimates of the regression coefficients $\alpha, \beta_1, \beta_2, \ldots, \beta_p$. Similar properties to the case with a single predictor apply for the multivariate least-squares solution. For example, the mean of the residuals is again 0, the residuals are uncorrelated with the fitted values, $\hat{y}$, and the amount of variation in $y$ that can be explained by a linear regression on $x_1, x_2, \ldots, x_p$ is maximized.

3.4 Exercises

3.1. A coin is tossed 60 times with 23 heads and 37 tails.

(a) Compute the likelihood function $L(\phi)$, where $\phi \in (0, 1)$ is the probability of heads.

(b) Compute $\ln L(\phi)$.

(c) Find the ML-estimator $\hat{\phi}$.

3.2. A coin is tossed 100 times with $N$ heads and $(100 - N)$ tails.

(a) Find the ML-estimator of $\phi$, the probability of heads.

(b) Assume that the coin is symmetric, i.e. $\phi = 0.5$. Which binomial distribution does 100 $\hat{\phi}$ have?

(c) Using Table 2.2 and Theorem 4, compute $E(\hat{\phi}), V(\hat{\phi})$ and $D(\phi)$ when $\phi = 0.5$.

(d) Can you generalize the results in c) to an arbitrary number of throws $n$?

3.3. Consider a data set of 200 parent-offspring pairs with all parents heterozygous (12) at a certain biallelic marker locus. If allele 1 is transmitted 131 times, compute the $p$-value for the null hypothesis that $\phi$, the probability that allele 1 is transmitted, is 0.5.
Chapter 4

Parametric Linkage Analysis

4.1 Introduction

Linkage analysis is a statistical technique used to find the approximate chromosomal locations of e.g. disease genes relative to a map of other genes with known locations. The idea is to look for evidence of co-segregation between the disease and genes (markers) whose locations are already known. With co-segregation, we mean a tendency for two or more genes to be inherited together, and hence for related individuals with the disease phenotype to share alleles at some nearby marker locus.

The two most widely used methods for linkage analysis are relative pair methods and lod score (likelihood based) methods. The former methods will be discussed in chapter 5 whereas this chapter is devoted to the latter.

4.2 Two-Point Linkage Analysis

We use the term two-point linkage analysis for analysis of linkage between two genes, usually, but not necessarily, a disease gene and a marker gene. The parameter of interest is the recombination fraction $\theta$. Two genes perfectly linked to each other will always be transmitted together during meiosis, corresponding to $\theta = 0$, whereas unlinked genes, e.g. genes located on different chromosomes, are transmitted independently, corresponding to $\theta = 0.5$. The co-segregation of disease- and marker alleles in a pedigree can be summarized in the likelihood function which measures the support, given by the data, for different $\theta$-values.

4.2.1 Analytical likelihood and lod score calculations

In this section we define the basis for parametric linkage analysis, and to keep things mathematically tractable we deliberately make assumptions that may not always be
that realistic. Linkage data encountered in human mapping studies, where crosses cannot be planned, will usually exhibit complications efficiently prohibiting every attempt to find closed form expressions for pedigree likelihoods. Hence, few attempts were made to analyze linkage data from extended pedigrees until an efficient recursive algorithm was introduced by Elston and Stewart (1971). An elegant example from the pre-Elston-Stewart era, showing the potential complexity of these likelihood expressions, was presented by one of the founders of linkage analysis, Newton E. Morton (1956). He derived a closed form expression for the likelihood corresponding to a 5-generation pedigree which turned out to be a polynomial of degree 20 (in $\theta$) with 60 terms! Using standard software available now, but of course not at that time, it is simple to check the result of his heroic effort. The results differ, but not dramatically. His main conclusion, significant linkage between a gene for elliptocytosis and the Rh blood type at an estimated recombination fraction of 0.05, seems to be correct.

Most linkage problems are impossible to solve analytically, but to fully appreciate and understand linkage analysis it is important to be familiar with the basic mathematics behind. Let us therefore start with simple problems that can be treated analytically.

**Example 37 (Direct counting)** If all meioses in a pedigree can be classified as recombinants or nonrecombinants, the likelihood function is simply the binomial probability function. Let $n$ denote the total number of observed meioses, $r$ the number of recombinants, and consequently $n - r$ the number of nonrecombinants. Then the likelihood is given by

$$L(\theta) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}. \quad (4.1)$$

The value of $\theta$ that maximizes this function, or in other words, the value that best fits the observed data, is the relative frequency of recombinants. But, in linkage analysis theta values above 0.5 (free recombination between the two loci) are usually considered biologically irrelevant, so let us instead define the estimated recombination fraction

$$\hat{\theta} = \begin{cases} \frac{r}{n} & \text{if } \frac{r}{n} \leq \frac{1}{2} \\ 0.5 & \text{if } \frac{r}{n} > \frac{1}{2} \end{cases}.$$

The shape of the likelihood function (4.1) for a few choices of $n$ and $r$ is shown in Figure 4.1. □

Once $\theta$ has been estimated, we proceed by testing the null hypothesis of no linkage ($\theta = 0.5$), that is whether the observed deviation from 50% recombination is statistically significant at some predefined level. For this purpose we could use the
4.2. TWO-POINT LINKAGE ANALYSIS

Figure 4.1: The binomial likelihood (4.1) for a few choices of \((n, r)\) where \(n\) denotes the total number of observed meioses and \(r\) the number of recombinant meioses

likelihood ratio test statistic

\[
\Lambda(\hat{\theta}) = 2 \ln \frac{L(\hat{\theta})}{L(0.5)};
\]

introduced in section 3.2, but in linkage analysis it is common practice to use the logarithm (base 10) of the likelihood ratio instead and compare it to a predefined threshold. This statistic

\[
Z(\theta) = \log \frac{L(\theta)}{L(0.5)};
\]

known as the lod score (for log-odds), is the single most important concept in linkage analysis. It was introduced for human studies by Haldane and Smith (1947). Positive lod scores indicate evidence in favor of linkage, whereas negative indicate evidence against linkage. Morton (1955) used the theory of sequential test procedures to define critical values (thresholds for rejection of the null hypothesis) corresponding to this test statistic. A lod score above 3 is generally accepted as significant evidence of linkage whereas a lod score below -2 is deemed sufficient to 'accept' the null hypothesis of free recombination.

A lod score of 3 at \(\theta = \hat{\theta}\) means that the observed data is 1000 times \((10^3)\) more likely when \(\theta = \hat{\theta}\) than under the null hypothesis \(\theta = 0.5\), or equivalently that the odds for linkage is 1000:1. This threshold corresponds to a \(p\)-value of 0.0001. The
usual argument for the extremely low significance level in linkage studies is that of multiple testing\(^1\). For linkage to a locus on the X chromosome, a lod score of 2 is usually considered sufficient for significant linkage.

**Example 38 (Direct counting, contd.)** The lod score corresponding to the likelihood (4.1) in Example 37 is given by:

\[
Z(\theta) = \log(\binom{n}{r} \theta^r (1-\theta)^{n-r}) \\
= \log(\theta^r (1-\theta)^{n-r}) \\
= \log \theta^r + \log (1 - \theta)^{n-r} - \log 0.5^n \\
= r \log \theta + (n - r) \log(1 - \theta) + n \log 2.
\]

Lod score curves corresponding to the likelihood curves presented in Figure 4.1 are shown in Figure 4.2.

![Lod score curves](image)

**Figure 4.2:** Lod score curves corresponding to the likelihoods in Figure 4.1

A nice property of the lod score function is that it is additive over independent families. The total lod score at a fixed \(\theta\) for a set of pedigrees is thus obtained by

\(^1\)A lod score is significant at the 5% level if the probability of observing a lod score greater than or equal to the lod score actually observed is 5% or less if the null hypothesis of no linkage is true. Twenty independent tests at the 5% level will thus, under the null hypothesis, on average produce one significant result. A high threshold, corresponding to a very low *pointwise* p-value, is therefore usually used when studying linkage to a large set of DNA markers.
summing the family-wise lod scores. This principle is illustrated in Figure 4.3 where the left panel shows the family-wise lod score functions and the right panel the total (cumulative) lod score functions for one, two, and three families (top to bottom).

![Figure 4.3: Family-wise and cumulative lod scores for three small pedigrees. The labels (n, r) denote number of observed meioses (n) and number of observed recombinants (r), respectively.](image)

Note that the support for \( \theta = 0 \) disappears (the lod score plunges to minus infinity) as soon as the first recombinant gamete is observed. This is intuitively OK since an observed recombination between the marker locus and the disease locus implies that the distance between the two loci is greater than zero.

So far we have assumed that all meioses in a pedigree can be scored unambiguously as recombinant or nonrecombinant with regard to a disease locus and a marker locus. This is, however, seldom the case. Complicating factors such as unknown phase, marker and/or disease locus homozygosity, phenocopies, incomplete penetrance, diagnostic uncertainty, unknown mode of inheritance, unequal male and female recombination fraction, and missing marker or phenotype data will often blur the picture. These problems will be defined and discussed below, but let us start with a simple situation.

Consider the nuclear family in Figure 4.4. This family has four members, labeled 1 to 4. The mother (2) and the second daughter (4) are affected with a certain disease (filled symbols) whereas the other two family members are unaffected (open symbols). Let us assume that the disease is dominantly inherited (one disease allele
is sufficient to become affected), that it is fully penetrant (all disease allele carriers are affected), that no phenocopies exist (aa-carriers can not be affected), and that all members of the family have been successfully typed at one informative marker locus. Let us further assume that the disease allele (A) is rare in the population (e.g. relative frequency p=0.0001). Then we can confidently assume that affected individuals are heterozygous (Aa) at the disease locus. The genotypes at the disease locus and the marker locus are shown below each individual. Four meioses can be observed in this pedigree. It is obvious that the mother has transmitted the haplotype (a5) to her unaffected daughter (3) and (A3) to her affected daughter (4), but in this case, and in fact in all two-generation pedigrees, it is impossible to deduce whether a haplotype is recombinant or not. The reason is that we do not know the phase of the mother. If her phase is

\[ P_1 : (A3|a5) \]

then the daughters are nonrecombinant, if it is

\[ P_2 : (A5|a3) \]

then both daughters are recombinant. The likelihood of this pedigree can now be calculated using the law of total probability, see Theorem 1 in Chapter 2. The disjoint decomposition of the sample space is in this example given by the two possible phases of the mother, so

\[
L(\theta) = \sum_{i=1}^{2} L(\theta|P_i)P(P_i) = 0.5(L(\theta|P_1) + L(\theta|P_2)) = 0.5((1 - \theta)^2 + \theta^2)
\]
The constant 0.5 comes from the fact that both phases have the same a priori probability. This constant is multiplied by the probability of observing either two non-recombinants \((1 - \theta)^2\) or two recombinants \(\theta^2\). The likelihood reaches its maximum (over the parameter space \(\theta \in [0, 0.5]\)) at \(\theta = 0\), leading to the lod score

\[
Z(\theta = 0) = \log \frac{L(\theta=0)}{L(\theta=0.5)}
= \log \frac{0.5(1^2+0^2)}{0.5(0.5^2+0.5^2)}
= \log(2)
= 0.301.
\]

This small, but optimally informative, nuclear family gave a lod score of about 0.3. Using the additivity principle, it is thus clear that a significant lod score (> 3) can be reached in a study of ten small families of this kind.

The number of observable meioses in pedigree 1 is four, but only the two maternal meioses were used in the likelihood calculations. The reason for that is that the unaffected father is not, or at least was not assumed to be, a disease gene carrier. He is thus homozygous \((aa)\) at the disease locus and therefore it is impossible to score the paternal meioses as recombinant or nonrecombinant. It might seem like a waste of resources to type the father, and in fact it was for this particular marker in this example, but this is not true in general. Genotyping information from the unaffected husband would have been very important if no sample was available from the affected mother.

In pedigree 1, we see one affected and one unaffected child. Since these children differ not only at the disease locus but also at the marker locus (they inherited different marker alleles from their mother) they support the hypothesis of close linkage between the marker locus and the disease locus. The same would have been true for two affected children having inherited the same marker allele from their affected mother. In fact, the symmetry implies that also two unaffected children having inherited the same marker allele from the mother give a lod score of 0.3, but families selected for linkage studies usually have more than one affected individual.

Let us now consider pedigree 2 in Figure 4.5 which is identical to pedigree 1 except for the affection status of the first daughter. When two affected siblings have inherited different marker alleles from their mother, one of them must be a recombinant and the other a nonrecombinant. It is thus not surprising that the maximum likelihood estimate of the recombination fraction turns out to be 0.5 in this case. To see that consider the likelihood

\[
L(\theta) = 0.5(\theta(1 - \theta) + (1 - \theta)\theta)
= \theta(1 - \theta).
\]

It is an increasing function of \(\theta\) over the interval \([0, 0.5]\). Thus, the lod score \(Z(\theta)\) is negative for all \(\theta < 0.5\) and 0 for \(\theta = 0.5\).
The ideas outlined above apply also to larger nuclear families. The three children in pedigree 3, shown in Figure 4.6, are either all recombinant or all nonrecombinant leading to a likelihood proportional to

$$\theta^3 + (1 - \theta)^3.$$ 

The multiplicative constant was left out because it will cancel out in the likelihood ratio anyway. The maximum likelihood estimate \( \hat{\theta} = 0 \) and the lod score \( Z(0) = \log(4) = 0.6 \). In general, the lod score will increase by 0.3 for each child supporting the hypothesis of tight linkage. Thus, it is hypothetically possible to reach a lod score of 3.0 in one large family with 11 children. The lod score functions for pedigree 1, 2, and 3 and the total lod score for the pedigrees is shown in Figure 4.7.
If we change the affection status of the son in pedigree 3 (from affected to unaffected) the likelihood will be proportional to

$$\theta^2 (1 - \theta) + (1 - \theta)^2 \theta = \theta (1 - \theta)$$

which reaches its maximum at $\theta = 0.5$, corresponding to a maximum lod score of 0.

The pedigree likelihoods discussed so far do not take marker and disease allele frequencies into account. That is OK since lod scores do not depend on these parameters when all members of a pedigree have been genotyped and the genetic model is autosomal dominant with full penetrance and no phenocopies. The pedigree likelihoods above should in fact have been functions of the allele frequencies, but we choose not to show that explicitly because the factors involving these parameters are identical under the two hypotheses (linkage and no linkage) and will thus cancel out in likelihood ratios. This is, however, not true in general, so before we take a look at more complicated scenarios, let us move a few steps backwards and define the pedigree likelihood more rigorously.

### 4.2.2 The pedigree likelihood

The likelihood for a pedigree with $n$ individuals is defined as the probability of observing the phenotypes $y = (y_1, y_2, \ldots, y_n)$ given the model parameter $\theta$ and the
penetrance parameters $f_0, f_1,$ and $f_2^2$. These intra-pedigree phenotypes are typically dependent for genetic and/or environmental reasons, but let us assume that the dependency is purely genetic and that it can be completely accounted for by our model for shared genotypes. Thus, we assume that individuals’ phenotypes $y_i$ are independent conditional on their joint marker-disease genotypes $g_i = (m_i, d_i)$, i.e.,

$$P(y | g) = \prod_{i=1}^{n} P(y_i | g_i),$$

(4.2)

where $m_i = (m_{i1}, m_{i2})$ is the marker genotype and $d_i = (d_{i1}, d_{i2})$ is the disease genotype. The unconditional probability of the phenotypes $y = (y_1, y_2, \ldots, y_n)$ is the pedigree likelihood:

$$L(\theta) = P(y | \theta) = \sum_{g} P(y, g | \theta) = \sum_{g} P(y | g) P(g | \theta),$$

(4.3)

where the summation is taken over all joint marker-disease genotypes (including phase) compatible with the observed data. The first factor in the summand, $P(y | g)$, depends on the penetrance parameters $f = (f_0, f_1, f_2)$ whereas the second factor in the summand, $P(g | \theta)$, depends on the recombination fraction, the disease allele frequency $p$, and the marker allele frequencies $p_M$. Let us first note that

$$P(g) = P(g_1 \cap g_2 \cap \ldots \cap g_n)$$

(4.4)

which can be expressed as a product of conditional probabilities:

$$P(g) = P(g_1) \cdot P(g_2 | g_1) \cdot P(g_3 | g_1, g_2) \cdot \ldots \cdot P(g_n | g_1, \ldots, g_{n-1}).$$

(4.5)

This expression can be simplified further because for non-founders (pedigree members with parents in the pedigree) genotypes are independent conditional on the genotypes of the parents. For founders, genotypes are assumed independent and hence depend only on the population allele frequencies. Let us therefore divide the $n$ pedigree members into two groups – founders ($F$) and non-founders ($NF$). Now (4.5) can be written

$$P(g) = \prod_{i \in F} P(g_i) \prod_{j \in NF} P(g_j | g_{Fj}, g_{Mj}),$$

---

2The penetrance parameters $f_0, f_1,$ and $f_2$ were introduced in Chapter 2, Example 6.

3This multiplication rule follows directly from the definition of conditional probability in Chapter 2. By replacing B with $g_2$ and C with $g_1$ in (2.2) we see that

$$P(g_1 \cap g_2) = P(g_1) \cdot P(g_2 | g_1).$$
where indexes $F_j$ and $M_j$ denote father and mother of non-founder number $j$, respectively. The first product depends on the marker allele frequencies $p_M$ and the disease allele frequency $p$, whereas the second depends on the recombination fraction $\theta$. For founders, it is common practice to assume that all pairs of alleles at the two genes (disease and marker) are in linkage equilibrium$^4$. The probability of a joint marker-disease genotype can under this assumption be written

$$P(g_i) = P(m_i; d_i; p_M, p) = P(m_i; p_M) \cdot P(d_i; p),$$

where $m_i$ and $d_i$ represent the pair of marker and disease genotypes for individual $i$, i.e. $g_i = (m_i, d_i) = (m_{i_1}, m_{i_2}, d_{i_1}, d_{i_2})$. Furthermore we assume that each of the two genes is in Hardy-Weinberg equilibrium.

For non-founders, we have

$$P(g_i|g_{Fj}, g_{Mj}) = P(m_i; d_i; m_{Fj}, m_{Mj}, d_{Fj}, d_{Mj}; \theta).$$

Putting it all together one obtains the pedigree likelihood:

$$L(\theta) = \sum_g \prod_{i=1}^n P(y_i|g_i; f) \prod_{i\in F} P(g_i) \prod_{j\in NF} P(g_j|g_{Fj}, g_{Mj}). \tag{4.6}$$

**Example 39 (Pedigree 1 revisited)** The joint marker-disease genotypes (including phase) for the four individuals in pedigree 1 are:

- Father (1) $a1|a2$
- Mother (2) $A3|a5$ or $A5|a3$
- First daughter (3) $a1|a5$
- Second daughter (4) $a1|A3$

The father and the first daughter (3) are both homozygous at the disease locus so their joint marker-disease genotypes are unambiguously known. The mother, on the other hand, who is doubly heterozygous has two possible genotypes. Finally, the second daughter (4) is also doubly heterozygous, but her phase is known (from the genotypes of her parents) so she has just one possible genotype. Thus, only two joint marker-disease genotypes $g$ are compatible with the observed data:

$$G_1 : (a1|a2, A3|a5, a1|a5, a1|A3)$$

and

$$G_2 : (a1|a2, A5|a3, a1|a5, a1|A3).$$

$^4$The marker allele $M_j$ and the a disease allele $D_j$ are in linkage equilibrium if the proportion of haplotypes $(M_j,D_j)$ in the population is equal to the proportions of $M_j$-alleles at the marker locus times the proportion of $D_j$-alleles at the disease locus.
We have assumed a dominant model without phenocopies, corresponding to penetrance parameters \( f = (f_0, f_1, f_2) = (0, 1, 1) \). Under this model, phenotypes will be completely determined by genotypes, so \( P(y_i|g_i) = 1 \) for all members of the pedigree. The pedigree likelihood will thus be a sum of two products, one for each joint marker-disease genotype. Let us denote the marker allele frequencies\(^5\)

\[ p_j = P(m = j) \quad j = 1, 2, \ldots \]

and the frequency of the normal allele at the disease locus

\[ q = 1 - p \]

then the contribution to the likelihood from the father will be\(^6\)

\[ q^2 \times 2p_1p_2 \]

and that from the mother

\[ 2pq \times 2p_3p_5. \]

These expressions are products of Hardy-Weinberg equilibrium probabilities. The contribution from each daughter is

\[ \frac{1}{2} \times \frac{(1 - \theta)}{2} \]

if the joint marker-disease genotype is \( G_1 \), and it is

\[ \frac{1}{2} \times \frac{\theta}{2} \]

if it is \( G_2 \). The first factor \( \frac{1}{2} \) is the probability of receiving \((a1)\) from the father conditional on his joint marker-disease genotype\(^7\), whereas the second factor reflects the transmission from the mother to a daughter. Let us for example take a look at the transmission from the mother to her unaffected daughter \(3\) when the genotype is \( G_1 \). We know that the daughter received the haplotype \((a5)\) from her mother. The probability of this haplotype conditional on the phase of the mother is:

\[ P(a5|G_1) = P(a|G_1)P(5|a, G_1) = \frac{1}{2} \times (1 - \theta). \]

\(^5\)The notation \( p_M \) was introduced above for the marker allele frequencies, but, for more compact notation, we drop the index \( M \) when denoting specific marker alleles.

\(^6\)Note that we assume no association between the two loci, i.e. the probability of observing a disease allele at the disease locus is the same no matter what pair of alleles we observed at the marker locus.

\(^7\)He will transmit either \((a1)\) or \((a2)\) each with probability \( \frac{1}{2} \).
Putting it all together we get

\[
L(\theta) = q^2 \times 2pq \times 2p_1 p_2 \times 2pq \times 2p_3 p_5 \times \frac{1}{2} \times \frac{(1-\theta)}{2} \times \frac{1}{2} \times \frac{(1-\theta)}{2} + \\
q^2 \times 2pq \times 2p_1 p_2 \times 2pq \times 2p_3 p_5 \times \frac{1}{2} \times \frac{q}{2} \times \frac{1}{2} \times \frac{q}{2} \\
= C \times ((1-\theta)^2 + \theta^2),
\]

where

\[
C = \frac{1}{2} q^3 p_1 p_2 p_3 p_5.
\]

Since \(L(\theta = 0) = C\) and \(L(\theta = 0.5) = \frac{C}{2}\) the lod score at \(\theta = 0\) is

\[
Z(0) = \log\left(\frac{L(0)}{L(0.5)}\right) = \log(2) = 0.3.
\]

As we saw earlier in Section 4.2.1 it is simple to calculate lod scores in this situation without considering the full pedigree likelihood, but in general there are no shortcuts when the scenarios get more realistic and complicated.

### 4.2.3 Missing marker data

Consider the pedigree in Figure 4.8. Let us once again assume that the genetic model is autosomal dominant with full penetrance and no phenocopies. The father in this family was for some reason not available for genotyping, but the homozygosity of one of his daughters (3) tells us that he must have at least one 1-allele. The other paternal marker allele is, however, impossible to infer in this example, and as a consequence, it is possible to score only one of the maternal meioses unambiguously (as recombinant

---

Figure 4.8: Pedigree 4 - Father not typed.
or non-recombinant) conditional on the phase of the mother. The second paternal allele could be 1, 2, or an allele \(w\) that was not observed in this pedigree. In missing-data situations like this, it is common practice to condition not only on the phase of the mother, but also on the genotype of the father. The number of possible joint marker-disease genotypes conditional on the observed pedigree data and the genetic model is eight in this example. We have three possible configurations for the father: \((1a|1a)\), \((1a|2a)\), or \((1a|wa)\), and two possible configurations (phases) for the mother: \((1A|2a)\) or \((1a|2A)\). The genotypes of the daughters are \((1a|1a)\) and \((1a|2A)\) if the father has no 2-allele, but if he carries a 2-allele, we have two possible configurations for the affected daughter \((4)\): \((1a|2A)\) and \((1A|2a)\). The possible joint marker-disease genotypes are listed below:

\[
G_1 : (1a|1a, 1A|2a, 1a|1a, 1a|2A) \\
G_2 : (1a|2a, 1A|2a, 1a|1a, 1a|2A) \\
G_3 : (1a|2a, 1A|2a, 1a|1a, 1A|2a) \\
G_4 : (1a|wa, 1A|2a, 1a|1a, 1a|2A) \\
G_5 : (1a|1a, 1a|2A, 1a|1a, 1a|2A) \\
G_6 : (1a|2a, 1a|2A, 1a|1a, 1a|2A) \\
G_7 : (1a|2a, 1a|2A, 1a|1a, 1A|2a) \\
G_8 : (1a|wa, 1a|2A, 1a|1a, 1a|2A)
\]

The pedigree likelihood will thus be a sum of eight products where each product includes two founder probabilities and two non-founder probabilities. The founder probabilities are, using the same notation as in the previous section:

\[
\text{Father (1)} : p_1^2 \times q^2 \quad \text{for } G_1 \text{ and } G_5 \\
= 2p_1p_2 \times q^2 \quad \text{for } G_2, G_3, G_6 \text{ and } G_7 \\
= 2p_1p_w \times q^2 \quad \text{for } G_4 \text{ and } G_8 \\
\text{Mother (2)} : 2p_1p_2 \times 2pq \quad \text{for } G_1, G_2, \ldots, G_8
\]

The unaffected daughter \((3)\) will with probability 1 receive the haplotype \((1a)\) from her father conditional on his joint marker-disease genotype under \(G_1\) and \(G_5\), whereas the corresponding probability is 0.5 if the father is heterozygous at the marker locus. The probability that she receives \((1a)\) from her mother is

\[
P(1a) = P(a)P(1|a) = \frac{1}{2} \times \theta
\]

\(8\)It is common practice to pool all unobserved alleles at a marker locus into a pseudo-allele, so-called lumping. It will reduce computational time without affecting the likelihoods.
4.2. TWO-POINT LINKAGE ANALYSIS

if the mothers phase is \((1A|2a)\), i.e. for \(G_1, G_2, G_3,\) and \(G_4,\) and it is

\[
P(1a) = P(a)P(1|a) = \frac{1}{2} \times (1 - \theta)
\]

if the mothers phase is \((1a|2A)\). Following this recipe it is fairly straightforward to find the non-founder probabilities:

| Daughter (3) | 1 \times \frac{\theta}{2} | for \(G_1\) |
| :           | \frac{1}{2} \times \frac{\theta}{2} | for \(G_2, G_3,\) and \(G_4\) |
| :           | 1 \times \frac{(1-\theta)}{2} | for \(G_5\) |
| :           | \frac{1}{2} \times \frac{(1-\theta)}{2} | for \(G_6, G_7,\) and \(G_8\) |

| Daughter (4) | 1 \times \frac{\theta}{2} | for \(G_1\) |
| :           | \frac{1}{2} \times \frac{\theta}{2} | for \(G_2, G_4,\) and \(G_7\) |
| :           | \frac{1}{2} \times \frac{(1-\theta)}{2} | for \(G_3, G_6,\) and \(G_8\) |
| :           | 1 \times \frac{(1-\theta)}{2} | for \(G_5\) |

Before we write down the pedigree likelihood we note that all factors from the father include \(p_1q^2\) and that all factors from the mother are identical. Hence these factors will cancel out in the likelihood ratio, so we drop them already at this stage and note that

\[
L(\theta) \propto p_1^2 \frac{\theta^2}{4} + 2p_2(\frac{\theta^2}{16} + \frac{\theta(1-\theta)}{16}) + 2p_w \frac{\theta^2}{16} + p_1 \frac{(1-\theta)^2}{4} + 2p_2 \frac{(1-\theta)^2}{16} + 2p_w \frac{(1-\theta)^2}{16} + \frac{p_w}{2} (\theta^2 + (1-\theta)^2) + \frac{p_w}{2}.
\]

Most textbooks on this topic use the term pedigree likelihood for everything that is proportional to the ‘real’ pedigree likelihood. We follow that tradition and redefine

\[
L(\theta) = (p_1 + \frac{p_w}{2})(\theta^2 + (1-\theta)^2) + \frac{p_w}{2}.
\]

The corresponding lod score at \(\theta = 0\) is

\[
Z(\theta = 0) = \frac{L(0)}{L(0.5)} = \log \left( \frac{p_1 + \frac{(p_1 + p_w)}{2}}{\frac{p_1 + p_w}{2} + \frac{p_w}{4}} \right). \tag{4.7}
\]

Table 4.1 shows the value of this function (4.8) for some allele frequency combinations. We see that the lod score function reaches its maximum 0.301 when the relative frequency of the 2-allele at the marker locus is 0, just as one would expect, because in this situation we know the phase of second daughter (4). We are thus back in the phase known situation discussed in the previous section. The other extreme situation is when the relative frequency of the 2-allele is close to 1.0. Then the family is uninformative for linkage corresponding to a lod score close to 0.
We saw in this example that the lod score for the pedigree depends on the marker allele frequencies. This is true in general for pedigrees with untyped founders. It is thus important to use 'good' estimates of the allele frequencies, especially in situations where genotype data is missing for many founders.

Estimated marker allele frequencies can be found e.g. at the Marshfield web site http://research.marshfieldclinic.org/genetics/Freq/FreqInfo.htm, but such estimates should be used with caution. They are usually based on a small number of chromosomes, and furthermore, those chromosomes might have a completely different origin compared to the population under study. A 'quick and dirty' alternative would be to use the observed allele frequencies instead. This might be reasonable, especially when a large number of families are studied, but the allele frequency estimates will be biased. Consider e.g. two nuclear families with heterozygous parental genotypes (1,2), (3,4) in family one and (5,6), (7,8) in family two. Assume that the first family has one child with genotype (1,4) whereas the second family has three children with genotypes (6,7), (6,7), and (5,7). The observed relative frequency of the 6-allele is 3/16=0.1875, but is that really a good estimate? No, a better alternative would be to use only founder alleles, resulting in equal estimated allele frequencies 1/8=0.125 in this example. Another potential cause of bias is that ascertained families with a rare disease might share a marker allele close to the disease locus inherited from the same ancient ancestor. This allele might be very common in the families but very rare in the population leading to biased allele frequency estimates and also to biased lod scores if some founder genotypes are missing for this marker.

An alternative way of avoiding bias is to type e.g. 50 unrelated blood donors from the same genetic population and use their 100 chromosomes for allele frequency estimation, and an even better idea would be to use them in addition to the genotyped founders in the pedigrees. Also genotypes from non-founders can be used in the estimation step without introducing bias, cf. Terwilliger and Ott (1994).

### Table 4.1: The lod-score function (4.8) evaluated at different marker allele frequencies.

<table>
<thead>
<tr>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_w$</th>
<th>$Z(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0.00</td>
<td>0.50</td>
<td>0.301</td>
</tr>
<tr>
<td>0.45</td>
<td>0.10</td>
<td>0.45</td>
<td>0.272</td>
</tr>
<tr>
<td>0.30</td>
<td>0.40</td>
<td>0.30</td>
<td>0.185</td>
</tr>
<tr>
<td>0.10</td>
<td>0.80</td>
<td>0.10</td>
<td>0.064</td>
</tr>
<tr>
<td>0.01</td>
<td>0.99</td>
<td>0.00</td>
<td>0.004</td>
</tr>
<tr>
<td>0.50</td>
<td>0.50</td>
<td>0.00</td>
<td>0.176</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
<td>0.155</td>
</tr>
<tr>
<td>0.10</td>
<td>0.50</td>
<td>0.40</td>
<td>0.138</td>
</tr>
</tbody>
</table>
Another frequently used approach is to assume equal allele frequencies, but in general that is not a good idea. It might, in unfortunate situations, lead to strong evidence for linkage even if no linkage exists. This problem has been extensively studied, see e.g. Freimer (1993), Ott (1992), and Ott (1999).

The bottom line is that good allele frequency estimates are necessary in order to get reliable lod scores, especially when genotype data is missing for a large proportion of the founders. It is also worth noting that missing data for founders is a very common problem, especially in studies involving multi-generational extended pedigrees where founders at the top are usually deceased long time ago and hence not available for genotyping.

4.2.4 Uninformativeness

A nuclear family must have at least two children to be informative for linkage. This is easily verified using a simplified version of pedigree 1, consisting of father (1), mother (2), and the second daughter (4). The likelihood $L(\theta) = 0.5(\theta + (1 - \theta)) = 0.5$ for all theta, corresponding to a lod score of 0. One child is thus not sufficient to give us phase information necessary to score the meioses in a nuclear family, but each additional child might, as we have seen above, add 0.3 to the lod score. The first child is used to establish the linkage phase.

Marker homozygosity

The mother in pedigree 1 is heterozygous at the marker locus. That is very important, because it is impossible to distinguish between the two linkage phases for homozygous markers. This used to be a problem in the early days of linkage analysis when the number of markers and their heterozygosity\(^\text{10}\) was low, but nowadays it is not. Now, thousands of markers with high heterozygosity (say $>0.70$) are available, and if one marker at a potentially interesting locus turns out to be non-informative it is an easy task to type one or a few nearby markers. The optimal marker from a mathematical point of view has an infinite number of alleles, each with a population allele frequency close to zero, but highly polymorphic markers with 30-40 alleles are seldom used because the fact that so many different alleles has evolved over time might indicate that mutations occur frequently at the marker locus. Such mutations might lead to Mendelian inconsistencies or even worse, to less clear marker-disease segregation. Another aspect on the use of highly polymorphic markers is that of computational

---

\(^9\)The full pedigree likelihood is proportional to this likelihood, but henceforth we call everything that is proportional to the pedigree likelihood a likelihood.

\(^{10}\)The term heterozygosity was introduced in Chapter 2, Example 10.
CHAPTER 4. PARAMETRIC LINKAGE ANALYSIS

time - especially in multipoint analysis. The optimal marker, when considering all these aspects might be a marker with about 10 equally frequent alleles.

Disease locus homozygosity

We saw in pedigree 1 that the meioses from the father, who is homozygous (aa) at the disease locus, to his daughters were impossible to score as recombinant or non-recombinant. Only doubly heterozygous individuals, like the mother in this pedigree, can provide meioses that are possible to score. Disease locus homozygosity of the other type (AA) is more problematic, but fortunately rare. A family with one affected parent, assumed to be (Aa), and two affected children, also assumed to be (Aa), will, as we have seen above, give a lod score of 0.3 if both meioses are either recombinant or nonrecombinant.\footnote{Once again we assume a dominant model with full penetrance and no phenocopies} This lod score is a false positive finding if the mother is homozygous (AA) at the disease locus. The probability that an affected individual is AA-homozygous at the disease locus is negligible if the disease allele is rare in the population, but it is always a good idea to study the pattern of disease transmission in each pedigree. If all children in a large nuclear family are affected, the reason might be AA-homozygosity. A marker allele inherited identical by descent by all children in such a family might give a high lod score, even though no linkage exists.

4.2.5 Other genetic models

The genetic model we have studied so far is characterized by:

- Autosomal dominant inheritance
- Full penetrance
- No phenocopies
- Rare disease allele

The first three of these assumptions are related to the penetrance parameters $f_0, f_1,$ and $f_2$ introduced in Chapter 2, Example 6. The model above corresponds to $f_0 = 0$ and $f_1 = f_2 = 1$. This model is convenient to work with, at least from a mathematical point of view, but these assumptions are often far from realistic, even for diseases with a seemingly typical dominant inheritance pattern. Hereditary breast cancer will be used to illustrate this.

Breast cancer is the most common cancer among females living in the western part of the world. About one in nine will develop the disease during their life time. The disease has a genetic component, but about 90% of the cases are so-called sporadics who develop the disease for other nongenetic (environmental) reasons. Linkage
studies designed to identify breast cancer genes will thus inevitably enroll phenocopies (sporadic cases), a fact that is usually taken into account by letting the penetrance parameter $f_0$ take some positive, possibly age-dependent, value less than 1.0.

To complicate things even further, some disease gene carriers might be unaffected. In this situation, the disease gene is said to have incomplete or reduced penetrance. A good example of this phenomenon in hereditary cancer syndromes is given by Knudson’s two-hit hypothesis, Knudson (1971), stating that one working copy of a tumor suppressor gene is sufficient for a specific cell regulation mechanism to work properly. Individuals with a germline mutation in one of the two copies of a tumor suppressor gene will thus be at higher risk of developing the disease than those born with two working copies since a single somatic mutation of the gene is sufficient for the individual to lose the protection provided by the working gene. Some individuals with a germline mutation will live all their life with one working copy of the tumor suppressor gene in each cell whereas other germline mutation carriers experience a second hit towards the gene, initiating tumor growth emanating from the cell which has lost both copies of the tumor suppressor gene.

The lifetime penetrance for the two breast cancer genes BRCA1 and BRCA2 is about 80%, see e.g. Ford et al. (1998). This reduced penetrance is usually accounted for in the genetic model by age dependent penetrances $f_1(\text{age})$ and $f_2(\text{age})$.

Consider a 20-year-old phenotypically unaffected daughter to a woman carrying a mutated breast cancer gene (BRCA1 or BRCA2). She might have inherited the normal copy of the gene from her mother, but that is far from sure. The hereditary form of the disease is characterized by early age at onset, but symptoms before thirty years of age are rare. The probability that she is a disease gene carrier is thus about 50% and she is therefore uninformative for linkage. Assume that she has two affected sisters and that all three sisters share a marker allele at a specific locus identical by descent from their affected mother. A parametric linkage analysis of this family under an autosomal dominant model with full penetrance will give a maximum lod score of 0 at $\theta = 0.5$ for this marker whereas age dependent penetrances with $f_1(20) = f_2(20) = 0$ will lead to a lod score of 0.3 effectively ignoring the young unaffected sister in the analysis.

**Reduced penetrance**

To see how reduced penetrance affects the pedigree likelihood we reanalyze pedigree 1 under a dominant model without phenocopies, assuming that $0 < f_1 = f_2 < 1$. We assume the same marker data, i.e. father (12) mother (35), unaffected daughter (15), and affected daughter (13), but this time, all we know about the alleles at the disease locus is that the two affected individuals carry at least one disease allele (A). The number of joint marker-disease genotypes (including phase) to sum over to get

---

12 A germline mutation is an inherited mutation present in all diploid cells of the body
the pedigree likelihood will thus be quite large. For each individual we have the following possibilities:

- Father (1) \( a_1 | a_2, A_1 | a_2, A_2 | a_1, \) or \( A_1 | A_2 \)
- Mother (2) \( A_3 | a_5, \) or \( A_5 | a_3, \) or \( A_3 | A_5 \)
- First daughter (3) \( a_1 | a_5, A_1 | a_5, A_5 | a_1, \) or \( A_1 | A_5 \)
- Second daughter (4) \( A_1 | a_3, A_3 | a_1, \) or \( A_1 | A_3 \)

The number of ways to combine these genotypes is \( 4 \times 3 \times 4 \times 3 = 144 \), but all combinations will not follow the Mendelian laws of segregation, so the actual number of terms in the sum is somewhat smaller. It is not that tricky to go through every case, but it is self-torture\(^{13}\). Therefore, from here on, we rely on computer programs when calculating pedigree likelihoods and lod scores. The lod scores in this example will depend on the disease allele frequency \( p \), the penetrance\(^{14}\) \( f \) and the recombination fraction \( \theta \). Lod scores at \( \theta = 0 \) for a few choices of the other two parameters are shown in Table 4.2. The lod scores at \( \theta = 0 \) are close to 0.3 (the maximal lod score

<table>
<thead>
<tr>
<th>( p )</th>
<th>( f )</th>
<th>( Z(0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0001</td>
<td>0.99</td>
<td>0.297</td>
</tr>
<tr>
<td>0.01</td>
<td>0.99</td>
<td>0.297</td>
</tr>
<tr>
<td>0.01</td>
<td>0.80</td>
<td>0.221</td>
</tr>
<tr>
<td>0.10</td>
<td>0.80</td>
<td>0.210</td>
</tr>
<tr>
<td>0.01</td>
<td>0.50</td>
<td>0.123</td>
</tr>
<tr>
<td>0.10</td>
<td>0.50</td>
<td>0.107</td>
</tr>
<tr>
<td>0.10</td>
<td>0.20</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Table 4.2: Lod scores at \( \theta = 0 \) depend on the disease allele frequency \( p \) and the penetrance \( f \) when the penetrance is reduced and phenocopies are not allowed for.

in this family) when the penetrance is only slightly reduced, but the picture will be more and more blurred the lower the penetrance. To locate low-penetrant disease genes is therefore a tricky business.

**Phenocopies and genetic heterogeneity**

If individuals can develop the disorder even though they do not carry a copy of the disease allele, we must allow for this additional complexity by letting the penetrance

\(^{13}\)A worked example for a nuclear family assuming a rare disease allele can be found in Terwilliger and Ott (1994) p. 40-42.

\(^{14}\)We used \( f \) to denote the vector of penetrance parameters \( f = (f_0, f_1, f_2) \) before, but here we use \( f \) for the only penetrance parameter, i.e. \( f = f_1 = f_2 \).
4.2. TWO-POINT LINKAGE ANALYSIS

parameter $f_0$ take positive values. Let us once again turn back to pedigree 1, but this time we assume that the penetrance parameters are $f = (f_0, 1, 1)$, i.e. a dominant model with full penetrance for disease gene carriers and penetrance $f_0 > 0$ for aa-carriers. If this model is correct, the mother and/or the affected daughter might be phenocopies, and the higher the probability that they actually are, the lower the lod scores at $\theta = 0$, see Table 4.3.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$f_0$</th>
<th>$Z(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.01</td>
<td>0.296</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.231</td>
</tr>
<tr>
<td>0.01</td>
<td>0.10</td>
<td>0.111</td>
</tr>
</tbody>
</table>

Table 4.3: Lod scores at $\theta = 0$ depend on the disease allele frequency $p$ and the penetrance parameter $f_0$ when phenocopies are allowed for.

During the hunt for the first breast cancer gene in the late eighties and early nineties, two different types of phenocopies were making life hard for the scientists: sporadic cases and cases of different genetic origin (genetic heterogeneity). Now, when we know that two major breast cancer genes (BRCA1 and BRCA2) exist, it is clear that not only sporadic cases but also BRCA2 cases can be regarded as phenocopies when studying BRCA1-related breast cancer, and consequently that BRCA1 cases can be regarded as phenocopies when studying BRCA2-related breast cancer. This problem has been extensively studied and it is common practice to test for genetic heterogeneity using e.g. the admixture test (Smith, 1963). The idea is to introduce a parameter $\alpha$ representing the probability that a family is linked to a specific disease locus. The likelihood for family $i$ can then be written as

$$L_i(\alpha, \theta) = \alpha L_i(\theta) + (1 - \alpha) L_i(\theta = 0.5)$$

and the total likelihood for $n$ families as

$$L(\alpha, \theta) = \prod_{i=1}^{n} L_i(\alpha, \theta).$$

The null hypothesis $\alpha = 1$ can now easily be tested against the alternative $\alpha < 1$ using a one-sided likelihood ratio test.

The ideal way of dealing with incomplete penetrance and phenocopies is to look for clinical features that might be useful for stratification of the pedigrees into groups with 'similar characteristics'. One example is stratification on the number of cases among first degree relatives, another stratification on median age at onset. It might also be possible to use gene expression profiles from cDNA microarrays for pedigree stratification, an idea suggested by Hedenfalk et al. (2001) who showed that the expression profiles of BRCA1 and BRCA2 carriers are very different.
Varying penetrances

Many genetically caused disorders are not phenotypically visible at the time of birth, but develop at some time in life. Symptoms might be visible in childhood as is the case for e.g. Duchennes muscular dystrophy, or later in life (e.g. Huntingtons disease). To assume the same constant penetrance for individuals of different age is therefore often not appropriate. The usual approach is to define so-called liability classes in such a way that the penetrance parameters are the same for all individuals belonging to the same class. The liability classes are not necessarily age classes, they might also represent e.g. sex or subtypes or severity of the disease phenotype. An example of successful use of liability classes is the CASH model (Easton et al., 1993) used for identification of the breast cancer susceptibility genes (BRCA1 and BRCA2). Data from a large population based study, the Cancer and Steroid Hormone study, was used, first by Claus et al., (1991) and later by Easton et al., to estimate the penetrance of breast cancer (familial and sporadic) in seven age classes each (< 30, 30-39, 40-49, 50-59, 60-69, 70-79, and ≥ 80). In Easton et al. (1993) all male breast cancer cases and all ovarian cancer cases were assigned to the 'female affected before age 30' liability class ignoring the actual age at onset. Thus, they decided that these syndromes are more likely to be caused by a non-working, at that time unknown, breast cancer gene than female breast cancer. Now, we know that ovarian cancer is part of the BRCA1-syndrome whereas male breast cancer is part of the BRCA2-syndrome.

Recessive mode of inheritance

The typical sign of a recessive mode of inheritance is affected children to unaffected parents, but such a pattern is no guarantee that the disease is recessive. A dominantly inherited disease with reduced penetrance might also lead to this disease pattern if the parent carrying the disease allele never developed the disease. Another family pattern indicating recessive inheritance is consanguineous matings, i.e. matings between relatives, e.g. cousins or second cousins. A single disease allele in an unaffected founder can in this scenario be found in two copies in a child to a pair of unaffected carrier parents both related to the founder. The proportion of affected children in a sibship is 25% under the recessive model (if the parents are heterozygous (Aa)) compared to 50% in the autosomal dominant situation (assuming one aa-parent and one Aa-parent). The pedigree likelihood calculations for recessive models are analogous to those in the autosomal dominant case.

Example 40 (Linkage under a recessive model) Assume that the genetic model is autosomal recessive with penetrance parameters \( f = (0, 0, 1) \) corresponding to 100% penetrance and no phenocopies. Assume further that pedigree 5 in Figure 4.9 has been typed for a highly polymorphic marker with observed alleles shown below each
4.2. TWO-POINT LINKAGE ANALYSIS

individual. All four meioses can be scored as recombinant or non-recombinant, con-

3 4
A A
2 4
A A

Figure 4.9: Pedigree 5 - Recessive mode of inheritance.

ditional on phase. The four possible joint marker disease genotypes are:

\[ G_1 : (1A|2a, 3A|4a) \]

\[ G_2 : (1A|2a, 3a|4A) \]

\[ G_3 : (1a|2A, 3A|4a) \]

\[ G_4 : (1a|2A, 3a|4A) \]

each with probability 0.25. The genotypes of the affected daughters are both un-

ambiguously known to be \((2A|4A)\) under the four parental phase combinations. All

the four meioses are recombinant under \(G_1\) and non-recombinant under \(G_4\), whereas

two are recombinant and two non-recombinant under \(G_2\) and \(G_3\). The likelihood is

\[ L(\theta) = 0.25(\theta^4 + 2\theta^2(1 - \theta)^2 + (1 - \theta)^4) \]

and the lod score

\[ Z(\theta) = \log(\theta^4 + 2\theta^2(1 - \theta)^2 + (1 - \theta)^4) - \log(0.5^4 + 2 \times 0.5^4 + 0.5^4). \]

The maximum lod score is \(Z(0) = \log(4) = 0.6. \)

Unaffected siblings add very little information if the mode of inheritance is au-

tosomal recessive. To see that we add a third child (unaffected) to pedigree 5. The
possible joint marker-disease genotypes of this child, conditional on those of the parents are:

\[
\begin{align*}
G_1^{(5)} &: (1A|3a) \\
G_2^{(5)} &: (1A|4a) \\
G_3^{(5)} &: (1a|3a) \\
G_4^{(5)} &: (1a|3A) \\
G_5^{(5)} &: (1a|4a) \\
G_6^{(5)} &: (1a|4A) \\
G_7^{(5)} &: (2A|3a) \\
G_8^{(5)} &: (2A|4a) \\
G_9^{(5)} &: (2a|3a) \\
G_{10}^{(5)} &: (2a|3A) \\
G_{11}^{(5)} &: (2a|4a) \\
G_{12}^{(5)} &: (2a|4A)
\end{align*}
\]

but only the four combinations of two parental marker alleles can be distinguished. The maximum lod score for the family for each of these four genotype combinations are shown in Table 4.4: The lod score will increase about 20% (from 0.606 to 0.727) unless the additional unaffected sibling has exactly the same marker genotype as the affected siblings. On the other hand, each additional affected sibling with the same marker alleles as the other affected siblings would increase the lod score by 0.6. Hence, a significant lod score can, theoretically, be reached in a family with six affected children if the mode of inheritance is autosomal recessive.

<table>
<thead>
<tr>
<th>alleles</th>
<th>$Z(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,3)</td>
<td>0.727</td>
</tr>
<tr>
<td>(1,4)</td>
<td>0.727</td>
</tr>
<tr>
<td>(2,3)</td>
<td>0.727</td>
</tr>
<tr>
<td>(2,4)</td>
<td>$-\infty$</td>
</tr>
</tbody>
</table>

Table 4.4: Lod scores at $\theta = 0$ for a nuclear family with two affected and one unaffected child. The affected children share the marker alleles 2 and 4. The total lod score when adding the unaffected child will depend on its alleles at the marker locus.
4.3 General pedigrees

We saw in section 4.2.2 that the pedigree likelihood can be written

\[ L(\theta) = P(y|\theta) = \sum_g P(y, g|\theta) = \sum_g P(y|g)P(g|\theta). \]

This sum can also be written

\[ L(\theta) = \sum_{g_1} \sum_{g_2} \ldots \sum_{g_n} P(y|g)P(g|\theta), \]

where \( g_i, i = 1, \ldots, n \) is the set of possible joint marker-disease genotypes for individual number \( i \) in the pedigree. The number of terms in the sum grows exponentially with the size of the pedigree, \( n \), but fortunately it is possible to carry out the calculations sequentially in a way that the amount of computing rises only linearly with pedigree size. The idea is to break down the pedigree into nuclear families and peel the result from each nuclear-family calculation onto the individual linking that particular nuclear family to the rest of the pedigree. The procedure is known as the Elston-Stewart algorithm (Elston et al., 1971).

**Example 41 (The Elston-Stewart algorithm)** Consider pedigree 6 in Figure 4.10. To keep the notation readable, we carry out the calculations without specifying marker alleles and affection status for the members of the pedigree. The key individuals in this pedigree are number 4, who links the nuclear family including 3, 4, 7, and 8 to the rest of the pedigree, and number 5 who links the family including 5, 6, 9, and 10 to the rest of the pedigree. Let us first calculate the likelihood of \( (y_3, y_7, y_8) \)
conditional on the genotype of the linking individual (4):

\[
P(y_3, y_7, y_8 | g_4) = \sum_{g_1} \sum_{g_2} \sum_{g_3} P(y_3, y_7, y_8, g_3, g_7, g_8 | g_4)
= \sum_{g_1} \sum_{g_2} \sum_{g_3} (P(y_1 | g_1) P(y_7 | g_7) P(g_5 | g_7, g_4)
\times P(y_8 | g_8) P(g_8 | g_3, g_4) P(g_3)).
\]

The likelihood of \((y_6, y_9, y_{10})\) conditional on the genotype of the linking individual (5) is calculated analogously as:

\[
P(y_6, y_9, y_{10} | g_5) = \sum_{g_1} \sum_{g_2} \sum_{g_3} \sum_{g_{10}} P(y_6, y_9, y_{10}, g_6, g_9, g_{10} | g_5)
= \sum_{g_1} \sum_{g_2} \sum_{g_3} \sum_{g_{10}} (P(y_6 | g_6) P(y_9 | g_9) P(g_9 | g_5, g_6)
\times P(y_{10} | g_{10}) P(g_{10} | g_5, g_6) P(g_6)).
\]

Now we use the conditional independence of the two nuclear families to calculate the likelihood of \((y_3, \ldots, y_{10})\) conditional on the genotypes of the individuals at the top of the pedigree.

\[
P(y_3, \ldots, y_{10} | g_1, g_2) = \sum_{g_1} P(y_3, y_7, y_8 | g_4) P(y_4 | g_4) P(g_4 | g_1, g_2)
= \sum_{g_1} (P(y_6, y_9, y_{10} | g_5) P(y_5 | g_5) P(g_5 | g_1, g_2))
\]

Finally we sum over \(g_1\) and \(g_2\) to get the full pedigree likelihood

\[
P(y_1, \ldots, y_{10}) = \sum_{g_1} \sum_{g_2} P(y_1 | g_1) P(y_2 | g_2) P(y_3, \ldots, y_{10} | g_1, g_2) P(g_1) P(g_2).
\]

Thanks to this algorithm it is now computationally feasible to analyze large extended pedigrees, at least using two-point analysis, but things get worse when we move on to the multipoint situation.

### 4.4 Multi-Point Linkage Analysis

Parametric two-point linkage analysis is often used as the first approach for analysis of genotype data, at least when mapping disease genes for Mendelian disorders. The natural extension is to use multiple markers simultaneously in order to extract as much information as possible from the data. The idea is to regard the marker positions as fixed and then vary the location \(x\) of a new marker across the fixed map. The new marker might be a newly detected short tandem repeat (STR) marker whose exact location is unknown, but let us think of it as a disease locus that we want to map. For each tentative disease locus position \(x\), we calculate a multilocus likelihood and compare it to a multilocus likelihood at an unlinked position, for details see Terwilliger and Ott (1994). Significant linkage to the map is, just as in the two-point
4.5. POWER AND SIMULATION

Case, defined as 1000:1 odds for a specific location \( x \) relative to a position off the map. The logarithm (base 10) of this likelihood ratio is known as the multipoint lod score. Some computer packages report the location score, which is defined as two times the natural logarithm of the likelihood ratio, instead of the lod score. The reason for that is that the location score asymptotically follows a chi-square distribution with 1 degree of freedom. To convert a location score to a lod score just divide by \( 2 \ln(10) \approx 4.6 \).

Multipoint linkage analysis is very computer intensive. The computational complexity of the parametric approach grows linearly with the number of individuals but exponentially with the number of markers included in the calculations. Theoretically, it is possible to calculate parametric multipoint lod scores for a dense grid of locations \( x \) over a map with many fixed markers, but in practice sliding n-point analysis\(^{15}\) is often carried out. The least demanding multipoint analysis is the sliding three-point analysis where lod scores are computed over all fixed sub-maps of two adjacent markers, but more loci should of course be utilized simultaneously if possible.

One advantage of multipoint analysis compared to two-point analysis is that the problem of marker homozygosity is not that devastating. Nearby markers provide information that is not available in the two-point situation. Another advantage is that we usually get information on crossovers on both sides of the disease locus in multipoint analysis. A disadvantage of the multipoint method is that it is much more sensitive to misspecification of the disease model (Risch and Giuffra, 1992).

The standard software for this type of analysis has been LINKMAP which is part of the LINKAGE package, but nowadays a faster implementation, VITESSE, is often used. References for these software packages will be given in Section 4.6.

4.5 Power and simulation

Assume that families with a well defined disease phenotype have been collected and that the disease model is known. At this stage of planning it might be tempting to order a set of markers and start the time consuming and expensive genotyping phase of the project, but first it is important to find out if the effort is worthwhile. This step is usually carried out using simulation software like e.g. SLINK of the LINKAGE package. Genotypes are simulated for all founders in the pedigrees, usually under the assumption of a polymorphic marker with five to ten equally frequent alleles. These founder genotypes are then transmitted to the non-founders according to an assumed recombination fraction. Simulated genotype data from individuals not available

\(^{15}\)The location of a test locus is varied over window covering \( n - 1 \) consecutive markers with known locations, and a n-locus likelihood is calculated for each position. This procedure is repeated for all windows covering \( n - 1 \) consecutive markers. The window is thus sliding from one end of the marker map to the other.
for genotyping is then thrown away before lod scores are calculated. After having repeated this procedure a large number of times, it is simple to calculate expected lod scores (ELODS) at the assumed $\theta$, expected maximum lod scores (EMLODS), and finally the probability to find a lod score above a fixed threshold. This simulation procedure is explained in detail in e.g. Haines and Pericak-Vance (1998).

The simulation is often performed under more than one scenario. It might e.g. be interesting to study a broad and a narrow definition of the phenotype, and also to see how the parameters defining the genetic model affect the power of the study. The perfectly linked marker ($\theta = 0$) is usually simulated in order to find the upper limit of the lod score, and the unlinked marker ($\theta = 0.5$) in order to see the distribution of the lod scores expected at unlinked positions. A third useful alternative is to simulate a marker tightly linked to the disease locus (e.g. $\theta = 0.1$).

For data consisting of fully informative gametes, finding analytical expressions for ELOD, EMLOD and power is straightforward, see e.g. Sham (1998), p. 134-138.

### 4.6 Software

A lot of software has been developed for linkage analysis, and short descriptions and links to most of it can be found at Jurg Ott’s excellent web site http://linkage.rockefeller.edu/soft/. Much of the software is freeware, but commercially available packages like e.g. S.A.G.E. (Statistical Analysis of Genetic Epidemiology; http://darwin.cwru.edu/pub/sage.html) are also included in the alphabetical list of genetic analysis software. The most commonly used software for parametric linkage analysis is probably the LINKAGE package (Lathrop et al., 1984) based on the Elston-Stewart algorithm. This package is a set of Fortran programs useful for different linkage analysis settings. Some of the most useful routines are MLINK (two-point analysis for fixed values of the recombination fraction $\theta$), ILINK (iterative search for the $\theta$ that maximizes the two-point lod score), LINKMAP (the program for parametric multipoint analysis), LCP (the linkage control program), LRP (the linkage report program), SLINK (a program for power simulations), and HOMOG (tests for locus heterogeneity). A very practically oriented introduction to parametric linkage analysis in general, and the LINKAGE programs in particular is ‘Handbook of human genetic linkage’, by Terwilliger and Ott (1994).

A few improvements to the original LINKAGE programs are worth mentioning. Firstly VITESSE (O’Connell and Weeks, 1995) which uses more efficient algorithms for multipoint analysis than LINKMAP in the LINKAGE package (e.g. fuzzy inheritance, and pooling of unobserved alleles), and secondly FASTLINK (Cottingham et al., 1993) which is a C-implementation of the linkage package which runs much faster than the original package. Another advantage of FASTLINK, compared to LINKAGE, is that some of the hard coded constants, like e.g. the maximum num-
4.6. SOFTWARE

...ber of liability classes, are fixed at considerably higher values in FASTLINK. For an introduction, see http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html.

Parametric linkage analysis can also be carried out using GENEHUNTER (Kruglyak et al., 1996), but in this package, it is not possible to calculate two-point lod scores for $\theta \neq 0$. Multipoint lod scores are, however, much more easily calculated using GENEHUNTER than using LINKMAP or VITESSE, but the multipoint analysis in GENEHUNTER uses only affected individuals in the analysis and that may be suboptimal. We have seen earlier in this chapter that unaffected individuals can be as informative as affected individuals, so the recommendation is to calculate parametric multipoint lod scores using an algorithm that takes advantage of the linkage information in all informative meioses of the pedigrees.
CHAPTER 4. PARAMETRIC LINKAGE ANALYSIS
Chapter 5

Nonparametric Linkage Analysis

The linkage analysis methods of the preceding chapter require knowledge of disease allele frequency/frequencies, as well as penetrance parameters. For complex disorders (polygenic traits), whose manifestation depends on the joint action of multiple genes in addition to environmental agents, it is much harder to specify a genetic model. Therefore, alternative nonparametric linkage (NPL) methods have been developed. The basic idea is the following: Consider a pedigree with several affected individuals\(^1\). It is likely that the affected individuals share the same disease alleles from one or a few founders. Such allele sharing identical by descent (IBD) was introduced already in Chapter 2, Examples 12 and 17. Using information from markers in the vicinity of a particular test locus, we can actually estimate the inheritance pattern at that locus and test how likely it is that the IBD pattern of the affected individuals occurred just by chance according to Mendelian segregation.

5.1 Affected sib pairs

Affected sib pairs are very often used in NPL analysis. One reason for this is that they are fairly easy to collect. Further, it is easier to first introduce the statistical NPL methods for a small pedigree consisting of two parents and two (affected) offspring.

Two alleles are said to be identical by state (IBS) if they are of the same kind, no matter what the ancestral origin is. It is important to distinguish IBD and IBS sharing from each other. We give an example which illustrates the difference between the two concepts:

Example 42 (IBD and IBS sharing.) Consider the sib pair family of Figure 5.1. It illustrates how a marker with three possible alleles 1, 2, and 3 is transmitted. Both the mother and the father have the third allele 3, denoted as 3\(_1\) and 3\(_2\) respectively.

\(^1\)As in Chapter 4, we will throughout assume that the phenotype is binary, i.e. unaffected/affected.
The children in the family both have a 3-allele. However, they come from different parents and are not IBD. Thus the two siblings share 0 alleles IBD. On the other hand, the $3_1$ and $3_2$-alleles in Figure 5.1 are IBS. This means that the siblings share one allele IBS.

Figure 5.1: Segregation of a marker with three possible alleles in a nuclear family with two offspring.

IBD sharing is the important concept in linkage analysis. IBS is a weaker concept, yielding less powerful statistical methods. The reason is that two alleles can be IBS without originating from the same ancestral founder allele.

Consider now a fixed locus $x$ on the genome. We wish to test if there is a disease locus linked to it ($H_1$) or not ($H_0$). Let $N = N(x)$ be the number of alleles shared IBD by an affected sib pair at locus $x$. Further, let $z_0 = z_0(x)$, $z_1 = z_1(x)$, and $z_2 = z_2(x)$ be the probabilities that $N$ equals 0, 1, or 2 respectively. These probabilities were introduced in Example 17, Chapter 2, for the case when $x$ coincides with the disease locus of a monogenic disease. Then, as shown in Table 2.1, the probabilities $z_j$ will depend on the genetic model, i.e. the disease allele frequency and the penetrance parameters. However, $(z_0, z_1, z_2)$ can also be defined for other loci $x$, linked or unlinked to the disease locus. Then $(z_0, z_1, z_2)$ depends not only on the genetic model, but also on how closely linked $x$ is to the disease locus, cf. e.g. Dudoit and Speed (1999) and Figure 5.2.

If $x$ is unlinked to the disease locus ($H_0$), the fact that the sib pair is affected gives no extra information about the distribution of $N$. In that case $z_j$ equals the binomial probabilities derived in Example 12, Chapter 2, i.e. $z_0 = 0.25$, $z_1 = 0.5$, and $z_2 = 0.25$. Thus, the hypothesis testing problem at locus $x$ can be formulated as

$$H_0 : (z_0, z_1, z_2) = (0.25, 0.5, 0.25) \text{ at locus } x,$$
$$H_1 : (z_0, z_1, z_2) \neq (0.25, 0.5, 0.25) \text{ at locus } x.$$  \hspace{1cm} (5.1)

Throughout this chapter, we implicitly understand that the number of alleles shared IBD by the sib pair is conditioned on the fact that both siblings are affected. For ease of notation, we don't indicate this conditioning on 'ASP', as was done in Chapter 2, Example 17.
5.1. AFFECTED SIB PAIRS

Figure 5.2: The probabilities $z_1$ and $z_2$ that an affected sib pair share 1 or 2 alleles IBD for different loci along a chromosome of length 150 cM. At the disease locus, positioned at 75 cM, $z_1 = 0.15$ and $z_2 = 0.8$. The more distant $x$ is from the disease locus, the closer are $z_1$ and $z_2$ to the $H_0$-values 0.5 and 0.25.

We will use

$$\phi = (z_0, z_1, z_2)$$

(5.2)

as our vector of genetic model parameters. The advantage of this is that $\phi$ can be used both for monogenic, polygenic and heterogenic diseases, and thus is suitable to use for complex diseases\(^3\).

Given a data set with $n$ independent affected sib pairs, we wish to design a test which checks if the relative proportions of sib pairs with 0, 1 or 2 alleles IBD significantly deviate from the $H_0$-proportions. This can be done by using either likelihood based methods (Section 5.1.1) or methods based on excess average IBD sharing (Section 5.1.2).

If $H_1$ is true, the power of the test at locus $x$ will depend on how much $(z_0, z_1, z_2)$ deviates from $(0.25, 0.5, 0.25)$. This in turn depends both on the genetic model and the recombination fraction between $x$ and the disease locus (loci). Four obvious

\(^3\)As mentioned in Chapter 2, Example 17 and Table 2.1, $(z_0, z_1, z_2)$ is a function of the disease allele frequency and the penetrance parameters $f_0$, $f_1$ and $f_2$ for a monogenic disease when $x$ coincides with the disease locus. Similarly, $(z_0, z_1, z_2)$ can be written as a function of the larger number of parameters needed to describe a heterogenic and polygenic disease.
constraints for \((z_0, z_1, z_2)\) are \(z_0 + z_1 + z_2 = 1\), \(z_0 \geq 0\), \(z_1 \geq 0\), and \(z_2 \geq 0\). Since \(z_0\) can be computed once \(z_1\) and \(z_2\) are known, it suffices to restrict ourselves to \((z_1, z_2)\). Holman (1993) and Faraway (1994) showed exactly which IBD probabilities \((z_1, z_2)\) that are possible under Hardy-Weinberg equilibrium. In addition to the constraints given above there are two more, giving the 'possible triangle' of Figure 5.3. This triangle gives us important extra information, and thus enables us to increase the power of the test, as will be seen in the next section. Sib pair triangle constraints under more general assumptions have been considered by Dudoit and Speed (1999).

![Figure 5.3: Holman's possible triangle for the probabilities \(z_1\) and \(z_2\) that an affected sib pair share one or two alleles IBD at the disease locus. The upper line is \(z_1 + z_2 = 1\) and the lower line \(3z_1 + 2z_2 = 2\).](image)

### 5.1.1 The Maximum Lod Score (MLS)

We wish to test if a specific locus \(x\) is linked to a locus affecting the trait studied. To this end, we have a data set consisting of \(n\) affected sib pairs. For simplicity, let us assume that we have perfect marker information at locus \(x\) for all sib pairs. (Incomplete marker information will be treated in Subsection 5.1.3.) The assumed perfect marker information means that we can observe \(N_i = N_i(x)\), the number alleles shared IBD at locus \(x\) by the \(i\):th sib pair, unambiguously from marker data. Let \(j\) be the observed value of \(N_i\). When testing \(H_0\) versus a fixed alternative vector \((z_0, z_1, z_2)\) in (5.1), the likelihood ratio at locus \(x\) for the \(i\):th sib pair becomes (cf.
5.1. AFFECTED SIB PAIRS

Definition 16, Chapter 3)

\[
\text{LR}_i(x; \psi) = \frac{P(N_i = j|H_1)}{P(N_i = j|H_0)} = \begin{cases} 
  z_0/0.25 = 4z_0, & j = 0, \\
  z_1/0.5 = 2z_1, & j = 1, \\
  z_2/0.25 = 4z_2, & j = 2, 
\end{cases}
\]

(5.3)

using the IBD probabilities under \(H_0\) as given in (5.1). Notice that the observed IBD count \(j\) in (5.3) as well as the probabilities \((z_0, z_1, z_2)\) depend on the locus \(x\), as illustrated in Figure 5.2.

The total likelihood ratio for all sib pairs is formed by multiplying the familywise likelihood ratios in (5.3);

\[
\text{LR}(x; \psi) = \prod_{i=1}^{n} \text{LR}_i(x; \psi) = (4z_0)^{n_0}(2z_1)^{n_1}(4z_2)^{n_2},
\]

where \(n_0 = n_0(x)\) is the number of sib pairs that share 0 alleles IBD at \(x\), and similarly for \(n_1 = n_1(x)\) and \(n_2 = n_2(x)\). This is analogous to the coin tossing likelihood (3.2). The difference is that there are now three possible outcomes (IBD = 0, 1 or 2) for each family rather than two, as in the coin tossing. The locus \(x\) is the parameter of main interest in the linkage analysis, since it is the one we vary along the genome in order to find regions that may be linked to the disease.

For a fixed \(\psi \in H_1\), the LOD score is computed as

\[
Z(x; \psi) = \log \text{LR}(x; \psi) = n_0 \log(4z_0) + n_1 \log(2z_1) + n_2 \log(4z_2).
\]

(5.4)

Notice that the vector \(\psi = \psi(x)\) varies with the locus \(x\), as shown in Figure 5.2. It depends on the genetic model, which is assumed known in parametric linkage analysis and unknown in nonparametric linkage analysis. Further, \(\psi\) depends on the recombination fraction between \(x\) and the disease susceptibility locus (loci), which is always unknown. Thus, \(\psi\) must be estimated at each \(x\), regardless of whether we know the genetic model or not. The resulting method is referred to as the maximum lod score (MLS) statistic. It was introduced by Risch (1990). The MLS score at locus \(x\) is defined as

\[
\hat{Z}(x) = \max_{\hat{\psi}} Z(x; \hat{\psi})
\]

(5.5)

\[
= \max_{(\hat{z}_0, \hat{z}_1, \hat{z}_2)} \left( n_0 \log(4\hat{z}_0) + n_1 \log(2\hat{z}_1) + n_2 \log(4\hat{z}_2) \right)
\]

where \(\hat{\psi} = (\hat{z}_0, \hat{z}_1, \hat{z}_2)\) is the ML-estimator\(^4\) of \(\psi\) at locus \(x\). We can use \(\hat{Z}(x)\) as test statistic for testing pointwise (locuswise) \(H_0\) against \(H_1\) in (5.1). In order to combine all such pointwise MLS scores into one score we define the test statistic

\[
\hat{Z}_{\text{max}} = \max_x \hat{Z}(x)
\]

(5.6)

\(^4\)In fact, \(\hat{\psi}\) is that parameter \(\hat{\psi}\) which maximizes the lod score \(Z(x; \hat{\psi})\). It can be shown that \(\hat{\psi}\) also maximizes the likelihood of the observed data and thus it is the ML-estimator.
where \( x \) varies over the chromosomal region(s) of interest. This is the (maximum) MLS score. The null hypothesis of no linkage to any locus in the tested region is rejected when \( \hat{Z}_{\text{max}} \) exceeds a given threshold\(^5\). More details on choosing the threshold will be given in Subsection 5.1.4.

If no a priori constraints are put on \( \phi \), the ML-estimator in (5.5) equals the relative frequencies of the number of families with 0, 1 or 2 alleles IBD, i.e.

\[
\begin{align*}
\hat{z}_0 &= n_0/n, \\
\hat{z}_1 &= m_1/n, \\
\hat{z}_2 &= n_2/n.
\end{align*}
\] (5.7)

This is analogous to the coin tossing in Example 31, Chapter 3, where the ML-estimator of the probability of heads was the relative proportion of heads during the throws. It is important to notice that the ML-estimator (5.7) is recomputed at each locus \( x \) that is being tested in the linkage analysis. This is so, since \( n_0, n_1 \) and \( n_2 \) refer to the number of families with 0, 1 or 2 alleles IBD at locus \( x \).

In practice, the ML-estimator is more involved than the relative frequencies in (5.7) for two reasons. First, the power of the test is increased if the unconstrained maximization in (5.5) is replaced by maximization over Holman’s possible triangle in Figure 5.3. This means that the ML-estimator differs from (5.7) if the vector \((\hat{z}_1, \hat{z}_2)\) of relative proportions of sib pairs with one and two IBD alleles is located outside Holman’s possible triangle. Secondly, if the marker information is incomplete, the MLS score gets more complicated than in (5.5). This affects computation of the ML-estimator as well, as described in Subsection 5.1.3.

MLS sib pair analysis has been implemented in the MAPMAKER/SIBS program, cf. Kruglyak and Lander (1995).

### 5.1.2 The NPL Score

In this section, we will formulate a testing procedure which is an alternative to the MLS score and easier to generalize to arbitrary pedigrees.

As for the MLS score, we assume perfect marker information. Further, there are \( n \) affected sib pairs, with \( N_i = N_i(x) \) the number of alleles shared IBD at locus \( x \) by the \( i \):th sib pair. When \( x \) is unlinked with the disease locus, the distribution of \( N_i \) was derived in Chapter 2, Example 12, and it follows that\(^6\) \( E(N_i) = 1 \) and \( V(N_i) = 0.5 \).

---

\(^5\)Let us write the null hypothesis in (5.1) as \( H_0(x) \), to highlight its dependence on the locus \( x \). Then, when investigating a whole region of loci, we are actually testing the null hypothesis \( H_0 = \cap_x H_0(x) \) against the alternative that some locus in this region is linked to the disease.

\(^6\)The formulas for \( E(N_i) \) and \( V(N_i) \) can be derived by direct calculation, using \( P(N_i = 0) = P(N_i = 2) = 0.25 \) and \( P(N_i = 1) = 0.5 \). Alternatively, we might use Table 2.2 for binomial distributions, since \( N_i \in \text{Bin}(2, 0.5) \).
5.1. AFFECTED SIB PAIRS

If we standardize \( N_i \) by its mean and standard deviation for unlinked loci (cf. Chapter 2, Example 24) we obtain

\[
Z_i = \sqrt{2}(N_i - 1),
\]

which is referred to as the \( i \):th NPL family score at locus \( x \). This means that \( Z_i = Z_i(x) \) attains the values \(-\sqrt{2}, 0, \) and \( \sqrt{2} \) when the \( i \):th sib pair has 0, 1, or 2 alleles IBD at locus \( x \).

The probabilities are \( z_0 = z_0(x), z_1 = z_1(x), \) and \( z_2 = z_2(x) \) that \( Z_i \) equals \(-\sqrt{2}, 0, \) and \( \sqrt{2} \), respectively. Thus

\[
E(Z_i) = \sqrt{2}(z_2 - z_0),
\]

\[
V(Z_i) = 2(z_2 + z_0) - 2(z_2 - z_0)^2.
\]

Clearly, \( E(Z_i) = 0 \) and \( V(Z_i) = 1 \) under the null hypothesis \((z_0, z_1, z_2) = (0.25, 0.5, 0.25)\) in (5.1). Further, it can be shown that\(^7\) \( E(Z_i) > 0 \) whenever \((z_0, z_1, z_2)\) belongs to the alternative hypothesis in (5.1). Hence, the hypothesis testing problem can be rewritten as

\[
H_0 : E(Z_i) = 0 \text{ at locus } x, \\
H_1 : E(Z_i) > 0 \text{ at locus } x.
\]

for all \( i = 1, \ldots, n \).

The total NPL score at locus \( x \) is then defined by summing the family scores and normalizing by a factor \( 1/\sqrt{n} \);

\[
Z(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(x) = \sqrt{\frac{2}{n}} (n_2(x) - n_0(x)),
\]

where \( n_0 = n_0(x) \) and \( n_2 = n_2(x) \) is the number of ASPs with 0 and 2 alleles IBD, respectively. The normalization makes \( E(Z(x)) = 0 \) and \( V(Z(x)) = 1 \) under \( H_0 \).\(^9\)

The final (maximum) NPL score is computed by maximizing the locuswise NPL score over the chromosomal region(s) of interest;

\[
Z_{\text{max}} = \max_x Z(x).
\]

The null hypothesis of no disease locus linked to the region(s) is rejected if \( Z_{\text{max}} \) exceeds a predefined threshold.

\( ^7 \)These formulas are derived as follows: \( E(Z_i) = -\sqrt{2} \cdot z_0 + 0 \cdot z_1 + \sqrt{2} \cdot z_2 = \sqrt{2}(z_2 - z_0) \) and, using Theorem 7 for the variance, \( V(Z_i) = E(Z_i^2) - E(Z_i)^2 = \ldots = 2(z_2 + z_0) - 2(z_2 - z_0)^2 \).

\( ^8 \)This is a consequence of Holman’s triangle restriction in Figure 5.3.

\( ^9 \)The formulas for \( E(Z(x)) \) and \( V(Z(x)) \) under \( H_0 \) can be obtained from by combining (5.9) with the algebraic rules for expected values and variances in Theorem 4 Chapter 2, yielding \( E(Z(x)) = \sqrt{n}E(Z_i(x)) \) and \( V(Z(x)) = V(Z_i(x)) \).
CHAPTER 5. NONPARAMETRIC LINKAGE ANALYSIS

Notice that large NPL scores lead to rejection of the null hypothesis of no linkage. This is a consequence of (5.10). The expected value of the pointwise NPL score at locus $x$ is

$$E(Z(x)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E(Z_i(x)) = \sqrt{n}E(Z_1(x)) = \sqrt{2n(z_2(x) - z_0(x))} > 0$$

under $H_1$. Thus the NPL score at locus $x$ will on the average be positive when a disease locus is linked to it. Quite naturally, the expected value also increases with the sample size, indicating that the power to detect linkage then increases.

Figure 5.4: The NPL score for 100 and 500 sib pair families along one chromosome. The marker information is perfect and the chromosome has length 150 cM, with a trait locus positioned at 75 cM. The genetic model is recessive, with disease allele frequency $p = 0.1$ and penetrance probabilities $(f_0, f_1, f_2) = (0.1, 0.1, 0.9)$ in a) and b), (uppermost figures). In c) and d), (bottom figures), the phenocopy rate $f_0$ is changed from 0.1 to 0. The dash-dotted curves are the expected NPL scores under the given disease model, locus position and sample size scenarios.

Figure 5.4 shows the NPL score along a chromosome for two different sample sizes and two different recessive models. It can be seen from the figure that the presence of phenocopies dramatically decreases the NPL score.

NPL score analysis has been implemented (for arbitrary pedigrees) in the Genehunter (Kruglyak et al. (1996)) and Allegro (Gudbjartsson et. al. (2000)) programs.
5.1. AFFECTED SIB PAIRS

5.1.3 Incomplete marker information

Assume that the marker information for the $i$:th sib pair is incomplete at locus $x$. Then we can no longer observe the number of alleles $N_i = N_i(x)$ shared IBD by the sibs unambiguously. Let us introduce the probabilities

$$
\begin{align*}
\pi_0 &= P(N_i = 0 | MD_i), \\
\pi_1 &= P(N_i = 1 | MD_i), \\
\pi_2 &= P(N_i = 2 | MD_i),
\end{align*}
$$

where $MD_i$ is an acronym for 'marker data' for the $i$:th sib pair. Notice that we only utilize information from the marker data in the conditioning, not whether or not $x$ is linked to the disease.

The marker data is informative if one of the probabilities $\pi_0$, $\pi_1$ and $\pi_2$ is close to one\(^{10}\). The information content is a more quantitative measure of marker informativeness. It was introduced by Kruglyak et al. (1996) for pedigrees of general form. The information content $I_E = I_E(x)$ at locus $x$ varies between zero and one\(^{11}\). A completely informative set of markers corresponds to $I_E = 1$ and a completely uninformative one to $I_E = 0$.

In single point analysis only one marker is used. Then the information content at $x$ will depend on the recombination fraction between the marker and $x$, how many of the parents (in addition to the siblings) that are being genotyped for the marker and the number of different marker alleles in the genotyped family members. For instance, if all the genotyped persons happen to be homozygotes with the same marker allele, the information content becomes 0 for that family.

In multipoint analysis, a number of markers on the same chromosome as $x$ are being used. It is then more involved to compute the marker probabilities $\pi_0$, $\pi_1$ and $\pi_2$, as well as the information content. Lander and Green (1987) showed how Hidden Markov models (HMM) can be used for devising an algorithm\(^{12}\). In general, the information content depends on several of the markers used; both their positions along the chromosome and their informativity\(^{13}\). Figure 5.5 shows the information content along the chromosomes for different scenarios:

\(^{10}\)Note that the probabilities ($\pi_0$, $\pi_1$, $\pi_2$) are family- and locus-dependent. A more precise notation is therefore $\pi_0 = \pi_0(x)$, $\pi_1 = \pi_1(x)$, and $\pi_2 = \pi_2(x)$.

\(^{11}\)For one sib pair, one has $I_E = (4 + m_0 \log_2(p_0) + m_1 \log_2(p_1) + p_1 \log_2(p_1) + p_2 \log_2(p_2)) / 4$, with $p_0$, $p_1$, $p_2$ the conditional probability that, given the marker data, the sib pair shares one allele IBD and that this allele is passed on from the father (mother). Thus $\pi_1 = \pi_1f + \pi_1m$. Further, $\log_2$ is the base 2 logarithm. For a collection of sib pairs, $I_E$ is the average of the familywise information contents.

\(^{12}\)The Lander and Green-algorithm is based on Haldane's map function (1.1).

\(^{13}\)To be precise, if no single marker is fully informative, then all markers will contribute to $I_E$ at $x$. On the other hand, if all pedigree members are being genotyped and there are two perfectly informative markers on either side of $x$, then the more distant markers will have no effect on $I_E$ at $x$. 
Figure 5.5: The information content $I_E = I_E(x)$, when $x$ varies along a chromosome of length 150 cM. The data set consists of 100 sib pairs. The markers are positioned on an equally spaced grid of size $\delta$ cM and have $M$ equally probable alleles (heterozygosity $H = 1 - 1/M$ in Example 10, Chapter 2). In a) all four family members are genotyped and in b) only the sib pair. The definition of $I_E$ for the whole data set is the average of the 100 familywise information contents.

It is important to distinguish between the probabilities $\pi_j$ in (5.13) and $z_j$ in (2.19). The numbers $z_0$, $z_1$ and $z_2$ correspond to the relative proportion of affected sib pairs with 0, 1 or 2 alleles IBD in a large population (whether or not we can observe these IBD numbers). They will depend on the genetic model and how closely linked the locus of interest is to the disease locus. The probabilities $\pi_0$, $\pi_1$, and $\pi_2$ on the other hand refer to the information that the marker data gives about the IBD sharing. They also depend on the position of the locus. However, as opposed to $(z_0, z_1, z_2)$, they vary between families, since the quality of the marker data is typically family-dependent.

Let us now describe how computation of the MLS and NPL scores are affected when the marker information is incomplete. First, the likelihood ratio for the $i$:th sib
5.1. AFFECTED SIB PAIRS

The total lod and MLS score for $n$ sib pairs is then obtained by summing the logarithm of (5.14) over all sib pairs and then maximizing w.r.t. $\psi$, as in Subsection 5.1.1.

Even though the ML-estimator $\hat{\psi}$ gets more complicated for incomplete marker data it can still be rapidly computed by means of the so called EM-algorithm.

The NPL score can be generalized to handle incomplete data in several ways. We will describe the method introduced by Kruglyak et al. (1996). In Subsection 5.1.2, we defined $Z_i$ as the value of the standardized score function (5.8) for the $i$:th family at locus $x$. For incomplete markers, the conditional probability given MD is $\pi_0$, $\pi_1$ and $\pi_2$ that $Z_i$ equals $-\sqrt{2}$, 0, and $\sqrt{2}$, respectively. By taking the conditional expectation

$$\tilde{Z}_i = E(Z_i|MD) = -\sqrt{2} \cdot \pi_0 + 0 \cdot \pi_1 + \sqrt{2} \cdot \pi_2 = \sqrt{2}(\pi_2 - \pi_0),$$  

we obtain the $i$:th family score for incomplete marker data. These are then replacing $Z_i(x)$ in (5.11) to obtain the total NPL score. Figure 5.6 shows how the quality of the marker data affects the NPL score for a collection of ASP families.

An alternative likelihood based method for NPL analysis with incomplete data was introduced by Kong and Cox (1997). It is slightly more complicated to formulate but often more powerful in regions between markers.

5.1.4 Power and $p$-values

The power of the MLS-method (5.6) and the NPL-method (5.12) depend on the underlying (and unknown) genetic model, the chromosomal region of interest, the number of families in the data set, and the informativity of the markers. Which of the two methods that is most powerful depends on the genetic model. It can be shown that the NPL score is slightly more powerful for additive models (with $z_1$ fixed to 0.5), whereas the MLS score is preferable when there is no a priori information of additivity (and thus ($z_1, z_2$) can be located anywhere in Holman’s triangle). In this way, one might say the the MLS score is more nonparametric in spirit, since its performance is not optimized for a particular set of parameters.

---

14 This equality can be deduced as follows: Expand the numerator of the likelihood ratio as $P(MD_i|H_1) = \sum_{j=0}^{2} P(MD_i|N_i = j)P(N_i = j|H_1) = \sum_{j=0}^{2} P(MD_i|N_i = j)z_j$ (using Theorem 1). By Bayes’ Theorem 2, the factors $P(MD_i|N_i = j)$ can be written as $P(MD_i|N_i = j) = P(N_i = j|MD_i)P(MD_i)/P(N_i = j) = \pi_j P(MD_i)/P(N_i = j)$. Finally $P(MD_i) = P(MD_i|H_0)$, since the probability of observing the marker data without conditioning on disease status is the same thing as conditioning on $H_0$. 

---
Figure 5.6: The NPL score for 100 ASP families and the same scenario as in Figure 5.4 c) when the marker information is imperfect. The markers are positioned on a grid of size $\delta$ cM with the two middle markers at equal distance $\delta/2$ cM from the disease locus. The markers have $M$ equally frequent alleles, corresponding to heterozygosity $H = 1 - 1/M$.

When calculating $p$-values (cf. (3.5)), it is important to distinguish between pointwise and regionwise $p$-values. If $z(x)$ is the observed NPL score at locus $x$, the pointwise $p$-value of the NPL score at locus $x$ is

$$P(Z(x) \geq z(x) | H_0)$$

and the regionwise $p$-value is

$$P(Z_{\text{max}} \geq z_{\text{max}} | H_0),$$

with $Z_{\text{max}}$ the NPL score in (5.12) and $z_{\text{max}} = \max_x z(x)$ its observed value. The $p$-value formulas for the MLS score are analogous, replacing $Z(x)$ and $Z_{\text{max}}$ by $\hat{Z}(x)$ and $\hat{Z}_{\text{max}}$ respectively.

When investigating linkage to a certain region, it is incorrect to report the pointwise $p$-value at the locus with maximal pointwise NPL (MLS) score. This is so,
since the pointwise \( p \)-value ignores the fact that we perform tests at several loci. The regionwise \( p \)-value takes this multiple testing into account and is larger than the pointwise \( p \)-value at the locus with maximal NPL- or MLS-score.

Exact calculation of \( p \)-values is sometimes complicated, but accurate approximations can be obtained by simulation. Alternatively, approximations by normal distributions give simple-to-use \( p \)-value formulas for the NPL score. These may or may not be accurate depending on e.g. the size of the data set, the value of the observed NPL score and the informativity of the marker data. Such approximative pointwise and genomewide\(^{15}\) \( p \)-values are reported in Table 5.1. These numbers are usually conservative for incomplete marker data, meaning that the true \( p \)-values are then smaller.

It is more difficult to find simple approximations for the MLS-score \( p \)-values. It can be shown, however, that the 'additive MLS score', obtained by restricting the maximization in (5.5) to the line \( z_1 = 0.5 \) in Holman's triangle, is roughly equivalent to a certain transformation of the NPL score\(^{16}\), cf. the second column of Table 5.1. The \( p \)-value of the full model MLS score is larger than that for the additive MLS score.

<table>
<thead>
<tr>
<th>NPL score</th>
<th>add MLS score</th>
<th>pointw. ( p )-value</th>
<th>genomew. ( p )-value</th>
<th>false positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.86</td>
<td>0.023</td>
<td>1.00</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>1.3 \cdot 10^{-3}</td>
<td>0.80</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>3.5</td>
<td>3.2 \cdot 10^{-5}</td>
<td>0.065</td>
<td>0.067</td>
</tr>
<tr>
<td>5</td>
<td>5.4</td>
<td>2.9 \cdot 10^{-7}</td>
<td>9.5 \cdot 10^{-4}</td>
<td>9.5 \cdot 10^{-4}</td>
</tr>
<tr>
<td>6</td>
<td>7.8</td>
<td>9.9 \cdot 10^{-10}</td>
<td>4.7 \cdot 10^{-6}</td>
<td>4.7 \cdot 10^{-6}</td>
</tr>
</tbody>
</table>

Table 5.1: Pointwise and genomewide \( p \)-values for observed NPL and additive MLS scores under perfect marker information and normal approximation. For the NPL score, this means that \( Z(x) \) in (5.16) is assumed to have a \( N(0, 1) \)-distribution under \( H_0 \) at each locus \( x \). Thus the pointwise \( p \)-value is reported as \( 1 - \Phi(z) \), where \( \Phi \) is the cdf (2.16) of the \( N(0, 1) \)-distribution and \( z \) the observed NPL score. For the genomewide \( p \)-values, we use the normal approximation formulas defined by Lander and Kruglyak (1995). The same method is used for the last column, containing the \( H_0 \)-expected number of false positives along the entire genome exceeding the given threshold.

Lander and Kruglyak (1995) introduced the terms suggestive and significant linkage. They are defined as those NPL scores when the expected number of false positives is less than or equal to a given threshold.
positives along the entire genome is 1 (NPL score 3.1) and 0.05 (NPL score 4.1) respectively under the null hypothesis of no disease locus linked to any part of the genome, cf. the last column of Table 5.1.

5.2 General pedigrees

Even though the MLS score is an excellent nonparametric linkage method for sib pairs, it is difficult to generalize to pedigrees of arbitrary form. On the other hand, the NPL score can be generalized to arbitrary pedigrees, as shown by Kruglyak et al. (1996). When doing so, the inheritance vector of a pedigree is an extremely useful tool. It is a binary vector (i.e. a vector with zeros and ones as entries) describing the inheritance pattern of a pedigree.

For an affected sib pair family, the inheritance vector at a certain locus is written as

\[ v = (p_1, m_1, p_2, m_2), \]

where \( p_i = 0 \) or 1 according to whether a grandpaternal or grandmaternal allele was transmitted in the paternal meiosis giving rise to the \( i \):th sibling. In the same way, \( m_i \) is defined in for the maternal meiosis giving rise to the \( i \):th sibling. An example is given in Figure 5.7. Since each component of \( v \) can take on two values, \( v \) itself can have \( 2^4 = 16 \) different values. The number of alleles \( N \) shared IBD by the sibs is a function of the inheritance vector, i.e. \( N = N(v) \). This is illustrated in Table 5.2.

<table>
<thead>
<tr>
<th>( v )</th>
<th>( N(v) )</th>
<th>( v )</th>
<th>( N(v) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0,0)</td>
<td>2</td>
<td>(1,0,0,0)</td>
<td>1</td>
</tr>
<tr>
<td>(0,0,0,1)</td>
<td>1</td>
<td>(1,0,0,1)</td>
<td>0</td>
</tr>
<tr>
<td>(0,0,1,0)</td>
<td>1</td>
<td>(1,0,1,0)</td>
<td>2</td>
</tr>
<tr>
<td>(0,0,1,1)</td>
<td>0</td>
<td>(1,0,1,1)</td>
<td>1</td>
</tr>
<tr>
<td>(0,1,0,0)</td>
<td>1</td>
<td>(1,1,0,0)</td>
<td>0</td>
</tr>
<tr>
<td>(0,1,0,1)</td>
<td>2</td>
<td>(1,1,0,1)</td>
<td>1</td>
</tr>
<tr>
<td>(0,1,1,0)</td>
<td>0</td>
<td>(1,1,1,0)</td>
<td>1</td>
</tr>
<tr>
<td>(0,1,1,1)</td>
<td>1</td>
<td>(1,1,1,1)</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.2: Values of the score function \( N \), the number of alleles shared by a sib pair, for different inheritance vectors \( v \).

Under \( H_0 \), the disease is not linked to the locus of interest. Thus, the fact that both siblings are affected gives no extra information about the inheritance pattern. Instead, inheritance is solely determined by the Mendelian segregation laws: For each meiosis, the probability is 1/2 that either the paternal or maternal allele
5.2. GENERAL PEDIGREES

Figure 5.7: Inheritance vector at a certain marker locus, and the corresponding transmission of alleles. The phase of both parents is assumed to be known. Since further, the parents’ four alleles are all different, the inheritance vector \( v = (0, 1, 1, 0) \) can be determined unambiguously. If the phase of both parents are unknown, all four inheritance vectors \( (0, 1, 1, 0), (1, 1, 0, 0), (0, 0, 1, 1) \) and \( (1, 0, 0, 1) \) are possible.

is transmitted. Since further, the four meioses in the ASP pedigree are independent, all 16 inheritance vectors have the same probability \( (1/2)^4 = 1/16 \) under \( H_0 \). Since there are 4 of the 16 inheritance vectors with \( N = 0 \), it follows that \( z_0 = P(N = 0) = 4/16 = 1/4 \) under \( H_0 \), in agreement with (5.1). The probabilities \( z_1 = 1/2 \) and \( z_2 = 1/4 \) are derived in the same way, since there are 8 and 4 inheritance vectors with \( N = 1 \) and \( N = 2 \), respectively.

Consider now a general pedigree with \( f \) founders and \( k \) nonfounders. If both parents of each nonfounder is present in the pedigree, there is a total of \( 2k \) meioses present in the pedigree. These can be gathered into the inheritance vector

\[
v = (p_1, m_1, p_2, m_2, \ldots, p_k, m_k),
\]

(5.19)

with \( p_i = 0 \) or 1 depending on whether the paternal meiosis which resulted in the \( i \)th nonfounder had a grandpaternal or grandmaternal origin. The definition of \( m_i \) is similar.

The usefulness of the inheritance vector comes from the fact that a score function \( S \) used in linkage analysis for general pedigrees can be written as a function

\[
S = S(v).
\]

(5.20)

This is a generalization of the score function \( N = N(v) \) for sib pairs. Several such score functions (5.20) have been proposed in the literature. They all give large values
when the inheritance vector is such that the affected individuals share a few founder alleles. The basis for this reasoning is the following: If the disease allele frequency is small and the phenocopy rate not too high, it is likely that the affected individuals have disease alleles from just one (or a few) common founder(s).

The most natural generalization of the sib pair score function is to consider all pairs of affected individuals in the pedigree, and then sum the number of alleles shared IBD for all such pairs. This score function, $S_{\text{pair}}$, was defined by Whittemore and Halpern (1994), who also defined another score function; $S_{\text{all}}$. In contrast to $S_{\text{pair}}$, $S_{\text{all}}$ considers allele sharing of all affected individuals simultaneously, not just pairwise.\(^\text{17}\)

Which score function is best to use depends on the underlying (and unknown) genetic model. Recent simulation studies by McPeek (1999) and Sengul et al. (2001) have shown that $S_{\text{all}}$ is quite powerful over a large range of genetic models and also robust, i.e. it never or seldom performs badly. Therefore, $S_{\text{all}}$ should be recommended before $S_{\text{pair}}$ if there is small a priori knowledge of the genetic model. Other score functions that also perform well over a wide range of genetic models are $S_{\text{robdom}}$ (McPeek, 1999) and Sobel and Lange’s $C$-statistic (Sobel and Lange, 1996).

Consider now a data set with $n$ pedigrees, which typically are of different form. Let $S_i = S_i(x)$ be the value of the score function (5.20) evaluated for the $i$:th family at locus $x$. Then, in analogy to (5.8), introduce the normalized score function $Z_i = Z_i(x)$ according to

$$Z_i = (S_i - \mu_i)/\sigma_i;$$  \hspace{1cm} (5.21)

where $\mu_i = E(S_i|H_0)$ and $\sigma^2_i = V(S_i|H_0)$ is the mean and variance of $S_i$ under the null hypothesis that $x$ is unlinked to the disease.\(^\text{18}\) The standardization (5.21) is made so that $E(Z_i) = 0$ and $V(Z_i) = 1$.

We wish to test if the normalized family scores $Z_i$ are significantly larger than zero. To this end, we define the total NPL score $Z(x)$ at locus $x$ as in (5.11), and the total regionwise NPL-score as in (5.12).

The expected value of $Z(x)$ can be computed both under $H_0$ and $H_1$ that $x$ is unlinked and linked to the disease respectively. In the former case, $Z(x)$ has zero mean and unit variance because of the standardization made for each family in (5.21). Under $H_1$, we need to condition on the affection status (‘affected’, ‘unaffected’ or

\(^{17}\) $S_{\text{all}}$ is defined as the sum, over all possible sets consisting of one allele from each affected individual, of the number of permutations leaving the founder alleles intact. Cf. Whittemore and Halpern (1994) for the exact mathematical definition.

\(^{18}\) Let $v_i = v_i(x)$ be the inheritance vector for the $i$:th pedigree at locus $x$. When there is no a priori information about the inheritance we have $P(v_i = w) = 2^{-2k}$ for all $2^k$ possible inheritance vectors $w$. With $S_i = S(v_i)$, this yields $\mu_i = E(S(v_i)|H_0) = \sum_w S(w)P(v_i = w) = 2^{-2k}\sum_w S(w)$ and, using (2.34), $\sigma^2_i = V(S(v_i)|H_0) = E(S^2(v_i)|H_0) - E(S(v_i)|H_0)^2 = 2^{-2k}\sum_w S^2(w) - \mu^2_i$. Notice in particular that for a collection of pedigrees of arbitrary form, the constants $\mu_i$ and $\sigma^2_i$ will depend on $i$. For a sib pair collection, on the other hand, $\mu_i = 1$ and $\sigma^2_i = 1/2$ for all $i$. 
5.3. Extensions

of all individuals in the pedigrees, and the expected NPL score will depend on the genetic model, how closely linked \( x \) is to the disease, the number of pedigrees, the graphical structure of the pedigrees, and the affection status of the pedigree members. A ‘good’ score function has large positive values of \( E(Z(x)) \) under \( H_1 \) for a wide range of genetic models and pedigree types.

The approximate \( p \)-values of Table 5.1 are based on the normal approximation \( Z(x) \in N(0, 1) \) under \( H_0 \) and the assumption of fully informative markers. These \( p \)-value formulas are sometimes accurate even for general pedigrees. For instance, the normal approximation requires that \( n \) is fairly large and that no single large pedigree dominates the whole data set. Otherwise, it is better to compute the \( p \)-values by simulation.

For incomplete marker data, the NPL family score for the \( i \):th pedigree can be computed similarly as in Section 5.1.3. Let

\[
P_{\text{marker}}(w) = P(v = w | \text{MD}_i)
\]

be the probability distribution of the inheritance vector \( v \) for the \( i \):th pedigree at locus \( x \), given the observed marker data (MD\(_i\)). Then, the \( i \):th family NPL score at locus \( x \) is defined as

\[
\bar{Z}_i(x) = E(Z_i(x) | \text{MD}_i) = \sum_{w} Z_i(x; w) P_{\text{marker}}(w),
\]

with \( Z \) the normalized score function (5.21) corresponding to the \( i \):th pedigree.

Even the information content \( I_E \) can be generalized to arbitrary pedigrees\(^{19}\). The interpretation is still the same: \( I_E \) ranges between 0 (no informativity) and 1 (full informativity) depending on the quality of the markers for the locus of interest.

5.3 Extensions

A possible extension of the total NPL score in (5.11) is to put different weights \( \gamma_i \) to the family scores;

\[
Z(x) = \sum_{i=1}^{n} \gamma_i \bar{Z}_i(x).
\]

As mentioned in Section 5.2, each family score \( \bar{Z}_i(x) \) has zero mean and unit variance under the null hypothesis that the trait is not linked to the disease, and if the marker

\(^{19}\) The exact definition of the information content for one family at locus \( x \) is \( I_E(x) = (2k + \sum_{w} P_{\text{marker}}(w) \log_2(P_{\text{marker}}(w))) / (2k) \), with \( P_{\text{marker}} \) the inheritance distribution of \( v \) at locus \( x \). The information content for the whole data set is computed by taking the average of the familywise \( I_E \)-values.
data information is perfect. By repeated use of Theorem 7 this implies \( E(Z(x)) = 0 \) under \( H_0 \). If we add the constraint

\[
\sum_{i=1}^{n} y_i^2 = 1
\]

on the weights, the total NPL score will also satisfy \( V(Z(x)) = 1 \) under \( H_0 \) when the marker data is perfect.

The rationale behind (5.22) is to assign larger weights to ‘more informative pedigrees’. For instance, a large pedigree with many affected individuals should be given a larger weight than a small ASP family. However, exactly which pedigrees that are informative depends on the genetic model. For instance, consider a pedigree with two parents and many children, of which all have the disease. For a dominant trait, it is likely that one of the parents is homozygous for the disease allele, and therefore the segregation is random from that parent. On the other hand, for a recessive disease, it is most likely that at least one of the parents is heterozygous for the disease allele and then the same allele is passed on to all sibs from that parent, making the segregation highly non-uniform and the family much more informative. Since the genetic model is more or less unknown in NPL analysis, the optimal weighting scheme in (5.22) is unknown as well. One possibility is to use weights that are fairly powerful over a large class of genetic models, cf. e.g. Sham et al. (1997) and Nilsson (2001) for details.

Most diseases of interest for NPL analysis are complex, meaning that several loci interact and jointly increase susceptibility to the disease. Consider a trait for which a least two (unknown) loci contribute to the disease and a data set of ASP families. Then proceed with conditional NPL analysis as follows: Compute first an unweighted NPL score (5.11). If a peak is found at some locus \( x_0 \), recompute a weighted NPL score function (5.22), where families with large (small) family scores \( \tilde{Z}_i(x_0) \) at \( x_0 \) are given large (small) weights \( y_i \). The main idea is that another peak of the unweighted NPL score at a second locus \( x_1 \) will be magnified when we choose non-uniform weights. This is possible if the NPL family scores at \( x_0 \) and \( x_1 \) are positively correlated, indicating so called epistasis between the two loci. Conditional NPL analysis has been applied to increase power in linkage analysis for non-insulin-dependent diabetes (NIDDM), cf. e.g. Cox et al. (1999) and Ångquist (2001).

For a heterogeneous disease, it suffices to have a disease causing allele at one of the trait loci in order get the disease. Then, the NPL family scores are expected to be negatively correlated at the trait loci, and a large (small) value of \( \tilde{Z}_i(x_0) \) should imply a small (large) weight \( y_i \).
5.4 *EXERCISES*

5.1. Consider the family of Figure 5.8. Determine the number alleles shared IBS and IBD by the sibs.

5.2. A data set of 100 ASP families is given. Assume that the marker information is perfect at a certain locus \( x \), and that the number of sib pairs with 0, 1 and 2 alleles IBD at \( x \) is 14, 45 and 41 respectively. Compute the MLS score \( \hat{Z}(x) \).

5.3. Consider the data set of Exercise 5.2.
   
   (a) Compute the NPL score \( Z(x) \).
   
   (b) Using the normal approximation \( Z(x) \in N(0, 1) \) under \( H_0 \) of \( x \) being unlinked to the disease, compute the pointwise \( p \)-value at \( x \).
   
   (c) Is the result in b) trustworthy?

5.4. In this exercise, we will compute the marker data probabilities (5.13) for a simple single point analysis example with one marker. The marker is located at recombination fraction \( \theta = 0.1 \) from a certain locus \( x \). If the IBD sharing can be observed perfectly at the marker, the marker data MD for one ASP is the number of alleles shared IBD at the marker. Let further \( N \) be the number of alleles shared IBD by the sib pair at locus \( x \). Assume \( MD = 2 \). This means that both parents transmit the same granparental alleles to both siblings at the marker.

   (a) What is the probability \( P \) that one parent transmits the same grandparental allele to both of its offspring at \( x \)?

   (b) Compute the probabilities \( \pi_j = P(N = j | MD) \), \( j = 0, 1, 2 \), that the sib pair shares zero, one or two alleles IBD at \( x \).
5.5. Consider the pedigree of Figure 5.9, for which the components of the inheritance vector are shown. Which is the number of alleles shared IBD by the two affected first cousins? If more than zero, determine the founder allele(s) corresponding to the IBD sharing.

Figure 5.9: Components of the inheritance vector for a two-generation family with two affected first cousins.
Chapter 6

Quantitative Trait Loci

A locus at which alleles determine the average level of a quantitative trait phenotype is called a QTL (quantitative trait locus). Typically the word ‘quantitative’ is used in connection with continuously varying characters, as opposed to the dichotomous subdivision of individuals for qualitative traits, e.g. based on affection status. Normal human trait variation and common diseases involve many genetic and environmental components and their interaction. The genetic analysis of such complex phenotypes requires new statistical approaches for the localization and evaluation of the relative importance of specific quantitative trait loci. One approach for identifying genetic factors in the aetiology of complex diseases studies quantitative phenotypes that are, in turn, risk factors for the disease. The underlying quantitative phenotypes that predispose to disease development may be aetiologically more homogeneous than the diseases themselves and may therefore provide more precise information for genetic linkage studies. Furthermore, some qualitative phenotypes, such as hypertension, type II diabetes and obesity, occur once an individual has exceeded a threshold for susceptibility. In such a case, studying the binary phenotype is not as informative as studying the actual phenotypic measurement, cf. Figure 6.1.

Sections 7.1 and 7.2 give information on terminology and concepts used within the area of quantitative genetics. Of central importance is the decomposition of the genotypic value, i.e. the mean phenotypic value conditional on the genotype, based on least-squares linear regression. This leads in subsection 7.1.3 to the definition of the additive and dominance components of variance associated with the effect of a single quantitative trait locus. Section 7.2 considers multilocus traits and introduces the concept of epistasis, i.e. gene interaction between different unlinked loci. The impact of environmental trait determinants is touched upon in subsection 7.2.2. Trait resemblance between relatives is discussed in Section 7.3. It is found that the genetic source of resemblance is expressed through sharing of alleles identical by

\[1 \text{The science and study of the causes, origins and reasons of diseases and their mode of operation.} \]
Figure 6.1: Pedigree of a three-generation family with three affected individuals. The affected individuals are colored black and the non-affected individuals are colored white. However, the phenotype studied is not binary (qualitative), but rather multinomial or continuous (quantitative). The level of phenotypic expression for each individual is represented by a thermometer showing the degree of affection on a scale from 0 to 100%. The degree of affection for the individuals are, 90%, 0%, 60%, 0%, 30% and 70%, respectively. Here we have chosen 50% to represent the threshold where we say an individual is affected by a trait.

descent at the trait locus (loci) involved. In Section 7.4 two frequently used so-called 'model-free' methods of quantitative trait linkage analysis, the Haseman-Elston regression and the variance component analysis, are introduced. Both methods use the covariance structure of data and rely on estimates of the inheritance pattern in genomic regions inferred from genotyped polymorphic markers. The material in Sections 7.1 to 7.3 is based on the exposition in Lynch and Walsh (1998). Several recently published review articles are relevant for the linkage discussion in Section 7.4, see e.g. Amos and de Andrade (2001), Blangero et al. (2000) and Feingold (2001).

In Chapter 2, Example 25, the phenotypic value, $Y$, of an individual was partitioned into a genetic component, the genotypic value $X$, and an environmental deviation $e$,

$$Y = X + e.$$ 

For a given genotype, $X$ is the expected, or population average, phenotypic value resulting from the joint expression of all the genes underlying the trait. The environmental variation is modelled as a deviation from the mean and hence has zero expectation. For a multilocus trait, $X$ is a potentially complicated function. We will however first consider the simplest case with a single autosomal biallelic locus affecting the trait. In this case there are (at most) three different expected phenoty-
pic values corresponding to the three different genotypes at the locus, cf. Chapter 2, Figure 2.12.

6.1 Properties of a Single Locus

6.1.1 Characterizing the influence of a locus on the phenotype

Denote the two alleles of a biallelic autosomal trait locus $A_1$ and $A_2$. Let $2a$ be the difference between the mean phenotypes of $A_2A_2$- and $A_1A_1$-individuals and let $(1 + k)a$ denote the difference in mean phenotypic values between heterozygotes ($A_1A_2$) and $A_1A_1$ homozygotes. Linear transformations of the measurement scale employed are immaterial as far as differences are concerned and we may arbitrarily fix the genotypic value of $A_1A_1$-homozygotes to 0, Figure 6.2.

\[
\begin{array}{c|c|c|c}
\text{Genotype} & A_1A_1 & A_1A_2 & A_2A_2 \\
\hline
\text{0} & (1 + k)a & 2a \\
\end{array}
\]

Genotypic value

Figure 6.2: Representation of genotypic values (mean phenotypic values) for a biallelic trait locus.

Different values of $a$ and $k$ correspond to different genetic models:

- $a$ is referred to as the homozygous effect and is a measure of additivity of alleles.
- $k$ measures departure from additivity, i.e. dominance, and is referred to as the dominance coefficient:
  1. Alleles $A_1$ and $A_2$ behave in a completely additive fashion when $k = 0$.
  2. $k = 1$ corresponds to complete dominance of the $A_2$ allele.
  3. $k = -1$ corresponds to complete dominance of the $A_1$ allele.
  4. $k > 1$ if the locus exhibits overdominance and, analogously, $k < -1$ corresponds to underdominance.

Example 43 (The pygmy gene in mouse.) The pygmy gene, denoted pg, in the mouse greatly reduces body size. An experiment reported the following means for
body weight in grams: \( X_{++} = 14, X_{+pg} = 12 \) and \( X_{pgpg} = 6 \). We will take these to be the expected phenotypic values, (i.e. the genotypic values). We have \( 2a = 14 - 6 = 8 \), hence \( a = 4 \) and \( (1 + k)a = 12 - 6 = 6 \) implying \( k = 0.5 \). These data thus suggests recessivity (although not complete) of the pygmy gene, (from Falconer and Mackay (1996)).

### 6.1.2 Decomposition of the genotypic value, (Fisher 1918)

The number of copies of a particular allele in a genotype (0,1, or 2) is referred to as the gene content. Unless this allele interacts additively with all other alleles, there will be a nonlinear relation between the gene content and the genotypic value, cf. Figure 6.2. We will consider the best linear approximation to this relation, since this leads to a partitioning of the genotypic value into an 'expected value' based on additivity (\( \hat{X} \)) and a deviation resulting from dominance (\( \delta \)), cf. Figure 6.3.

![Figure 6.3: Linear least-squares regression of the genotypic value, \( X \), of a single biallelic locus on the number of \( A_2 \)-alleles, \( N_2 \), in the genotype of an individual. From left to right the points correspond to the \( A_1A_1 \), \( A_1A_2 \), and \( A_2A_2 \) genotypes. Circles represent the true genotypic values, \( X \), while squares are the fitted or 'predicted' values, \( \hat{X} \), based on the linear regression. The deviation, \( \delta = X - \hat{X} \) between \( X \) and \( \hat{X} \) is the dominance deviation.](image)

Formally, by least-squares regression of the genotypic value on the number of \( A_1 \) and \( A_2 \) alleles in the genotype, \( N_1 \) and \( N_2 \), respectively,

\[
X_{ij} = \hat{X}_{ij} + \delta_{ij} = \mu_X + \alpha + \gamma + \delta_{ij} = \mu_X + \alpha N_1 + \alpha_2 N_2 + \delta_{ij}, \tag{6.1}
\]
where \( X_{ij} \) is the genotypic value of \( A_iA_j \)-individuals, \( \mu_X \) is the mean genotypic value in the population (assuming as before \( X_{11} = 0 \)), \( \alpha_i \) and \( \alpha_2 \), the slopes of the regression, \( N_1 \) and \( N_2 \), the predictors and \( \delta_{ij} \) the residual deviation. \( \alpha_i \) is termed the additive effect of allele \( A_i \), \( i = 1, 2 \).

The genotypic values predicted by the regression are

\[
\hat{X}_{ij} = \mu_X + \alpha_i + \alpha_j = \begin{cases} 
\mu_X + 2\alpha_1 & \text{for } A_1A_1, \\
\mu_X + \alpha_1 + \alpha_2 & \text{for } A_1A_2, \\
\mu_X + 2\alpha_2 & \text{for } A_2A_2.
\end{cases}
\]

(6.2)

Thus the predicted values \( \hat{X}_{ij} \) correspond to a strictly additive genetic model (\( k = 0 \)) with \( a = \alpha_2 - \alpha_1 \). The additive allelic effects, \( \alpha_i \) are defined as deviations from the population mean, \( \mu_X \), and hence must have population expectation equal to zero.

To see this, let \( P_{ij} \) be the population relative frequency of genotype \( A_iA_j \) and let \( p_i \) denote the population relative frequency of allele \( A_i \), \( i = 1, 2 \). Consider picking an allele at random from the population by first drawing an individual at random and then with equal probability choosing one of the two alleles present in the genotype of the individual. The chosen allele is certain to be \( A_i \) if the individual is \( A_iA_i \) and is \( A_j \) with probability 0.5 if the individual is \( A_iA_j \). Since these are the only possibilities for picking an \( A_i \)-allele we have, irrespective of mating behavior,

\[
p_i = P_{ii} + P_{ij} = 2P_{ii} + P_{ij}, \quad i = 1, 2.
\]

From properties of least-squares regression (see e.g. Chapter 3, Section 3.3), the residual, \( \delta_{ij} \), has zero mean and hence \( E(\hat{X}_{ij}) = \mu_X \). It follows from (6.2) that

\[
0 = E(\hat{X}_{ij}) - \mu_X = 2\alpha_1P_{11} + (\alpha_1 + \alpha_2)P_{12} + 2\alpha_2P_{22} = 2\alpha_1p_1 + 2\alpha_2p_2,
\]

i.e. the expected additive effect of a randomly drawn allele from the population, \( \alpha_1p_1 + \alpha_2p_2 \), is zero. (Note that under random mating each \( N_i \in \text{Bin}(2, p_i) \)).

Since two predictors, \( N_1 \) and \( N_2 \), appear in 6.1 the equation is a multiple regression. However in the biallelic case we can rewrite the model, noting that for any individual, \( N_1 = 2 - N_2 \) so that

\[
X_{ij} = \mu_X + \alpha_1(2 - N_2) + \alpha_2N_2 + \delta_{ij} = \hat{\mu}_X + (\alpha_2 - \alpha_1)N_2 + \delta_{ij}
\]

where \( \hat{\mu}_X = \mu_X + 2\alpha_1(= \hat{X}_{11}) \) is the new intercept. We denote the slope of this regression by

\[
\alpha = \alpha_2 - \alpha_1.
\]

Since \( p_1 + p_2 = 1 \), it follows that

\[
\alpha_1 = -p_2\alpha \quad \alpha_2 = p_1\alpha.
\]
Recall from Section 3.3 that the slope of a univariate regression is simply the covariance between response and predictor divided by the variance of the predictor. Thus

$$\alpha = \frac{C(X, N_2)}{V(N_2)}.$$  

Here $C(X, N_2)$ and $V(N_2)$ are functions of the gene effects, i.e $a$ and $k$, and the population allele frequencies, $p_i = P(A_i), i = 1, 2$. Assuming that mating is random we have, (remember $X_{11} = 0$ by assumption),

$$E(X)(= \mu_X) = a(1 + k) \cdot 2p_1p_2 + 2a \cdot p_2^2$$  

$$= 2ap_2(1 + p_1k)$$  

$$E(N_2) = 1 \cdot 2p_1p_2 + 2 \cdot p_2^2$$  

$$= 2p_2$$  

$$E(N_2^2) = 1 \cdot 2p_1p_2 + 4 \cdot p_2^2$$  

$$= 2p_2(1 + p_2)$$  

$$E(X \cdot N_2) = a(1 + k) \cdot 1 \cdot 2p_1p_2 + 2a \cdot 2 \cdot p_2^2$$  

$$= 2ap_2(2p_2 + p_1(1 + k))$$  

$$C(X, N_2) = E(X \cdot N_2) - E(X)E(N_2)$$  

$$= 2p_1p_2a(1 + k(p_1 - p_2))$$  

$$V(N_2) = E(N_2^2) - (E(N_2))^2$$  

$$= 2p_1p_2$$  

which implies

$$\alpha = \frac{C(X, N_2)}{V(N_2)} = a(1 + k(p_1 - p_2)).$$

Under the assumption of random mating, $\alpha$ is known as the average effect of allelic substitution. It represents the expected change in genotypic value that results when an $A_2$ allele is randomly substituted for an $A_1$ allele:

$$\alpha = a(1 + k(p_1 - p_2)) = a(1 + k)p_1 + [2a - a(1 + k)]p_2$$  

$$= (X_{12} - X_{11})p_1 + (X_{22} - X_{12})p_2.$$

For the purely additive case ($k = 0$), $\alpha$ is simply equal to $a$. In general, however, $\alpha$ is also a function of $k$ and of the allele frequencies in the population, Figure 6.4.

With dominance present, the phenotypic effect of a gene substitution depends on the status of the unsubstituted allele. If $A_2$ is dominant ($k > 0$), then $\alpha$ will be inflated relative to the case of additivity if $A_2$ is rare ($p_1 > p_2$), but diminished if $A_2$ is
6.1. PROPERTIES OF A SINGLE LOCUS

Figure 6.4: The slope $\alpha$ of the linear least-squares regression of genotypic value on gene content as a function of allele frequency, $p_2$, and degree of dominance, $k$. Note that, except for the case of complete additivity ($k = 0$), the regressions differ with different allele frequencies.

common ($p_1 < p_2$). Thus, except in the case of additivity, the average effect of allelic substitution is not simply a function of inherent physiological properties of the allele. It can only be defined in the context of the population.

6.1.3 Partitioning the genetic variance

Let $\sigma_X^2$ denote the total genetic variance, i.e. the variance of the genotypic value $X$. Then

$$\sigma_X^2 = V(X) = V(\hat{X} + \delta) = V(\hat{X}) + V(\delta) + 2C(\hat{X}, \delta) = V(\hat{X}) + V(\delta) = \sigma_A^2 + \sigma_D^2,$$

where the fourth equality ($C(\hat{X}, \delta) = 0$) follows from properties of least-squares regression (see Chapter 3, Section 3.3). Statistically speaking, $\sigma_A^2$ is the amount of the variance of $X$ that is explained by the regression on $N_2$, whereas $\sigma_D^2$ is the residual variance. Biologically, $\sigma_A^2$ is the genetic variance associated with the average additive effects of alleles (the additive genetic variance), and $\sigma_D^2$ is the additional genetic variance associated with dominance effects (the dominance genetic variance).
Further, straightforward but somewhat tedious calculations show that

\[ \sigma_A^2 = 2(p_1 \alpha_1^2 + p_2 \alpha_2^2) = 2p_1p_2 \alpha^2 = 2p_1p_2 a^2(1 + k(p_1 - p_2))^2 \]

\[ \sigma_D^2 = (2p_1p_2 a k)^2. \]

Notice that the additive genetic variance is twice the population variance of the additive allelic effect, \( p_1 \alpha_1^2 + p_2 \alpha_2^2 \) and thus, since mating is assumed random, equals the variance of the sum of the two additive allelic effects of a randomly drawn individual from the population. Both components of variance depend upon the allele frequencies, the dominance coefficient \( k \), and the homozygous effect \( a \).

A common misconception is that the relative magnitudes of additive and dominance genetic variance provide information on the additivity of gene action. However, through its influence on \( \alpha \), dominance contributes to the additive genetic variance, and for certain allele frequencies, can cause \( \sigma_A^2 \) to reach much higher levels than in the purely additive case. Even in the case of complete dominance, \( \sigma_D^2 \) is unlikely to greatly exceed \( \sigma_A^2 \), and it is often substantially smaller, Figure 6.5.

### 6.1.4 Additive effects, average excesses, and breeding values

In randomly mating diploid species, a parent transmits only one allele per locus to each of its offspring. The transmitted allele exhibits its additive effect when randomly combined with a gene from another parent. The dominance deviation of a parent, which is a function of the interaction between the two parental genes, is eliminated when gametes are produced. One way to think of \( \hat{X} \) and \( \delta \) is thus as the heritable and nonheritable components of an individual’s genotypic value, respectively.

A somewhat different measure of the effect of an allele is called the average excess, \( \alpha^* \). It has a simpler biological interpretation and can, for randomly mating populations, be shown to be equivalent to the additive effect, \( \alpha \). Suppose that the maternal (or paternal) allele of an individual is \( A_2 \) and consider the status of the paternal (maternal) allele for the same individual. The average excess of allele \( A_2 \) is the difference between the conditional expected genotypic value of the individual and the mean genotypic value of a randomly drawn individual from the entire population, i.e.

\[ \alpha^*_2 = E(X|ma = A_2) - \mu_X \]

\[ = X_{12} P(pa = A_1|ma = A_2) + X_{22} P(pa = A_2|ma = A_2) - \mu_X, \]

where ‘pa’ and ‘ma’ is short for paternal and maternal allele, respectively. However, under random mating, the status of the paternal allele is independent of the corresponding maternal allele and hence, in this case

\[ \alpha^*_2 = X_{12} p_1 + X_{22} p_2 - \mu_X. \]
6.1. PROPERTIES OF A SINGLE LOCUS

Figure 6.5: The dependence of the genetic variance components at a locus on the degree of dominance, $k$, and the relative frequency of the $A_2$ allele, $p_2$. Solid lines denote total genetic variance, a dashed line the additive genetic variance and a dotted line the dominance variance.

Plugging in the values of $X_{12}(=a(1+k))$ and $X_{22}(=2a)$ gives, after some simplification, $\sigma_2^2 = p_1 \alpha = \alpha_2$.

An individual’s breeding value, $\Lambda_{ij}$, is the sum of the additive effects of its genes, i.e. $\Lambda_{ij} = \alpha_i + \alpha_j$ if the genotype is $A_iA_j$. Under random mating the breeding value of an individual can be shown to equal twice the expected deviation of its offspring mean phenotype from the population mean. Thus, in experimental settings, it is possible to estimate the breeding value of an individual by mating it to many randomly chosen individuals from the population and taking twice the deviation of its offspring mean from the population mean.

Example 44 (The measured-genotype approach.) Litter size in the Merino sheep of Australia is determined largely by a single polymorphic locus. Consequences of the Booroola gene ($B$) in two hypothetical random-mating populations with gene frequencies of 0.5 and 0.1 are shown in Table 6.1. We assume that phenotypic me-
### Table 6.1: Consequences of the Booroola gene, \((B)\), in two hypothetical random-mating populations with allele frequencies of 0.5 and 0.1. Table from Lynch and Walsh (1998).

<table>
<thead>
<tr>
<th></th>
<th>(p_B = 0.5)</th>
<th></th>
<th></th>
<th>(p_B = 0.1)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(bb)</td>
<td>(Bb)</td>
<td>(BB)</td>
<td>(bb)</td>
<td>(Bb)</td>
<td>(BB)</td>
</tr>
<tr>
<td>Genotypic value ((X_{ij}))</td>
<td>1.48</td>
<td>2.17</td>
<td>2.66</td>
<td>1.48</td>
<td>2.17</td>
<td>2.66</td>
</tr>
<tr>
<td>Frequency ((P_{ij}))</td>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
<td>0.81</td>
<td>0.18</td>
<td>0.01</td>
</tr>
<tr>
<td>Mean genotypic value (\mu_X = \sum P_{ij} X_{ij})</td>
<td>2.120</td>
<td></td>
<td></td>
<td>1.616</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additive effects (\alpha_B = p_B X_{Bb} + p_B X_{BB} - \mu_X)</td>
<td>0.295</td>
<td></td>
<td></td>
<td>0.603</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\alpha_b = p_b X_{bb} + p_B X_{Bb} - \mu_X)</td>
<td>-0.295</td>
<td></td>
<td></td>
<td>-0.067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breeding values (\Lambda_{ij} = \alpha_i + \alpha_j)</td>
<td>-0.59</td>
<td>0.00</td>
<td>0.59</td>
<td>-0.134</td>
<td>0.536</td>
<td>1.206</td>
</tr>
<tr>
<td>Dominance deviations (\delta_{ij} = X_{ij} - (\mu_X + \Lambda_{ij}))</td>
<td>-0.05</td>
<td>0.05</td>
<td>-0.05</td>
<td>-0.002</td>
<td>0.018</td>
<td>-0.162</td>
</tr>
<tr>
<td>Variance components (\sigma_A^2 = \sum P_{ij} \Lambda_{ij}^2)</td>
<td>0.1740</td>
<td></td>
<td></td>
<td>0.0808</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\sigma_D^2 = \sum P_{ij} \delta_{ij}^2)</td>
<td>0.0012</td>
<td></td>
<td></td>
<td>0.0003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\sigma_X^2 = \sigma_A^2 + \sigma_D^2)</td>
<td>0.1752</td>
<td></td>
<td></td>
<td>0.0811</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When \(n\) alleles are present at a locus, the average excess, \(\alpha_i^*\), for any allele \(A_i\) is given by

\[ \alpha_i^* = \frac{1}{n} \sum_{j=1}^{n} \frac{X_{ij}}{P_{ij}} \]

where \(X_{ij}\) is the genotypic value of allele \(A_i\) in genotypic class \(ij\) and \(P_{ij}\) is the frequency of that class. The variance component due to dominance deviations is then

\[ \sigma_D^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{X_{ij}^2}{P_{ij}} - \frac{1}{n} \frac{\sum_{i,j=1}^{n} X_{ij}^2}{\sum_{i,j=1}^{n} P_{ij}} \]

### 6.1.5 Extensions for multiple alleles

**Average Excesses**

When \(n\) alleles are present at a locus, the average excess, \(\alpha_i^*\), for any allele \(A_i\) is given by

\[ \alpha_i^* = \frac{1}{n} \sum_{j=1}^{n} \frac{X_{ij}}{P_{ij}} \]
by,

\[ \alpha_i^* = \sum_{j=1}^{n} X_{ij} P(a_2 = A_j | a_1 = A_i) - \mu_X, \]

where \( a_1 \) and \( a_2 \) denote the paternal and maternal allele, respectively (or vice versa should you prefer). Under random mating this reduces to

\[ \alpha_i^* = \sum_{j=1}^{n} X_{ij} P(a_2 = A_j) - \mu_X = \sum_{j=1}^{n} X_{ij} p_j - \mu_X \]

where \( p_j \) is the frequency of the \( j \)th allele, \( A_j \).

**Additive Effects**

Additive effects are defined to be the set of \( \alpha_i \) that minimizes residual variance, obtained from the least-squares solution for the (multivariate) linear regression

\[ X = \mu_X + \sum_{i=1}^{n} \alpha_i N_i + \delta, \]

where \( N_i \) is the number of copies of allele \( A_i \) carried by an individual. Under random mating we have \( \alpha_i = \alpha_i^* \) so that, in this case, the additive effects can be characterized by

\[ \alpha_i = \sum_{j=1}^{n} X_{ij} p_j - \mu_X. \]

It can further be shown that when mating is random

\[ \alpha_i = \alpha_i^* = \frac{C(X,N_i)}{E(N_i)} = \frac{C(X,N_i)}{2p_i}, \]

where \( C(X,N_i) \) is the covariance between genotypic value and the number of copies of the \( A_i \) allele.

**Additive Genetic Variance**

The additive genetic variance is the variance of the breeding values of individuals in the population. Since the breeding value of an individual is defined as the sum of the additive effects of the two alleles and since under random mating these effects are independent we have

\[ \sigma_A^2 = 2 \sum_{i=1}^{n} p_i \alpha_i^2. \]

**In Summary**
• The homozygous effect \( a \), and the dominance coefficient \( k \) are intrinsic properties of allelic products. That is, they are not functions of allele frequencies but may vary with the genetic background.

• The additive effect \( \alpha \), and the average excess \( \bar{\alpha} \) are properties of alleles in a particular population. They are functions of homozygous effects, dominance coefficients and genotype frequencies.

• The breeding value is a property of a particular individual in reference to a particular population. It is the sum of the additive effects of an individual’s alleles.

• The additive genetic variance, \( \sigma^2_A \), is a property of a particular population. It is the variance of the breeding values of individuals in the population.

6.2 Genetic Variation for Multilocus Traits

Several questions and new concepts need to be considered when trying to explain the phenotypic variation of traits whose genetic component is governed by multiple loci: Do the genotypic effects associated with single loci combine additively or do there exist nonlinear interactions between different loci or epistasis? Are the inheritance and distribution of genes at one locus independent of those at other loci? Several sources of environmental variance influence the expression of polygenic traits, and this raises questions as to whether gene expression varies with the environmental context, gene \( \times \) environment interaction, and whether specific genotypes are associated with particular environments, introducing covariance of genotypic values and environmental effects.

Although the presence of the many different sources of variation will make dissection of the genetic contributions to trait expression a difficult task to accomplish, we might still hope to be able to characterize populations with respect to the relative magnitudes of different sources of phenotypic variance. In Section 7.1 we saw that the genetic variance associated with a single locus can be partitioned into additive and dominance components. This approach can be generalized to account for all of the loci contributing to the expression of a quantitative trait, as well as to allow for variance arising from gene interaction among loci, i.e. epistasis. The dominance effect at a locus is defined to be the deviation of the observed genotypic value from the expectation based on additive effects. Hence, dominance is a measure of non-additivity of allelic effects within loci. Analogously, epistasis describes the non-additivity of effects between loci.

Example 45 (Artificial teosinte.) As a numerical example of epistasis consider the genotypic values associated with two biallelic loci in an artificially constructed po-
6.2. GENETIC VARIATION FOR MULTILocus TRAITS

pulation of teosinte, the presumed wild progenitor of cultivated maize, Table 6.2. Estimates of homozygous-effect coefficients \((a)\) and dominance coefficients \((k)\), conditional on the genotypic states of the alternate locus, are given in the last two columns and rows of Table 6.2, assuming that observed mean phenotypes are accurate estimates of genotypic values. The fact that both \(a\) and \(k\) vary dramatically with genetic background provides strong evidence of epistatic interaction between the genes associated with the two markers, \(UMC107\) and \(BV302\). From Lynch and Walsh (1998).

<table>
<thead>
<tr>
<th>BV302</th>
<th>UMM</th>
<th>UMT</th>
<th>UIT</th>
<th>a</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMM</td>
<td>18.0</td>
<td>40.9</td>
<td>61.1</td>
<td>27.0</td>
<td>0.33</td>
</tr>
<tr>
<td>BMT</td>
<td>54.6</td>
<td>47.6</td>
<td>66.5</td>
<td>6.0</td>
<td>-2.17</td>
</tr>
<tr>
<td>BT</td>
<td>47.8</td>
<td>83.6</td>
<td>101.7</td>
<td>21.6</td>
<td>0.06</td>
</tr>
</tbody>
</table>

\(a\): 14.9 21.4 20.3

\(k\): 1.46 -0.69 -0.73

Table 6.2: Average length of vegetative internodes in the lateral branch (in mm) mainly determined by the joint genotype of two biallelic loci. Data from an artificially constructed population of teosinte. Table from Lynch and Walsh (1998).

### 6.2.1 An extension of the least-squares model for genetic effects

We will confine our attention to two loci since the extension to three or more loci will be quite obvious, and we furthermore assume that the two loci are unlinked, i.e., they reside on different chromosomes. In addition, a random mating population is assumed throughout. Consider an individual with alleles \(A_i\) and \(A_j\) at one locus and \(B_k\) and \(B_l\) at another. The corresponding genotypic value, \(X_{ij,kl}\), can be written as the sum of the effects within loci and a deviation \(\varepsilon\) due to interaction between loci,

\[
X_{ij,kl} = \mu + (\alpha_i^{(1)} + \alpha_j^{(1)} + \delta_{ij}^{(1)}) + (\alpha_k^{(2)} + \alpha_l^{(2)} + \delta_{kl}^{(2)}) + \varepsilon_{ij,kl},
\]

where the superscripts, (1) and (2), correspond to the two loci involved. The epistatic interactions between loci can arise in three different ways: additive \(\times\) additive (\(\alpha\alpha\)), additive \(\times\) dominance (\(\alpha\delta\)), and dominance \(\times\) dominance (\(\delta\delta\)). The number of different types of epistasis grows steadily with an increasing number of loci considered jointly.

In the previous section, the additive effect of an allele was, under random mating, found to be equal to the expected phenotypic deviation of members of the
population with the allele (at a fixed ‘position’) from the population mean phenotype. Specifically, for each separate locus consider the paternally inherited allele to be the first allele and the maternally derived allele the second allele, (or vice versa). Let \( a_i^{(r)} \) denote the \( r \)th allele at locus \( r \), \( i = 1, 2, r = 1, 2 \) and let \( E[X|a_1^{(1)} = A_i] \) be the conditional mean phenotype of individuals with \( a_1^{(1)} \) equal to \( A_i \) without regard to the other allele at the locus or to the genotype at the second locus. Then, using that mating is random and that the loci are assumed unlinked,

\[
E \left[ X|a_1^{(1)} = A_i \right] = \sum_{j,k,l} X_{ij,kl} p_{ij,l}^A \left[ d_2^{(1)} = A_j, a_1^{(2)} = B_k, d_2^{(2)} = B_l|a_1^{(1)} = A_i \right]
\]

where \( p_{ij,l}^A \) is the relative frequency of the \( s \)th allele at the \( r \)th locus, \( r = 1, 2 \) and we define

\[
a_i^{(1)} = E \left[ X|a_1^{(1)} = A_i \right] - \mu_X.
\]

Analogously \( a_k^{(2)} = E[X|a_1^{(2)} = B_k] - \mu_X \), where \( E[X|a_1^{(2)} = B_k] = \sum_{j,k,l} X_{ij,kl} p_{ij,l}^B p_{ij,k}^{(1)} p_{ij,l}^{(2)} \). Within each locus, the mean value of the average effects is equal to zero, i.e.

\[
\sum_i a_i^{(1)} p_i^{(1)} = \sum_i a_i^{(2)} p_i^{(2)} = 0.
\]

The dominance effects are defined by considering the conditional mean phenotype of individuals with \( a_1^{(1)} = A_i \) and \( a_2^{(1)} = A_j \) without regard to the alleles at the second locus. We have \( E[X|a_1^{(1)} = A_i, a_2^{(1)} = A_j] = \sum_{k,l} X_{ij,kl} p_{kl}^{(2)} p_{ij,l}^{(2)} \) and

\[
\delta_{ij}^{(1)} = \delta_{ij}^{(1)} = E \left[ X|a_1^{(1)} = A_i, a_2^{(1)} = A_j \right] - \mu_X - a_i^{(1)} - a_j^{(1)}
\]

and

\[
\delta_{kl}^{(2)} = \delta_{lk}^{(2)} = E \left[ X|a_1^{(2)} = B_k, a_2^{(2)} = B_l \right] - \mu_X - a_k^{(2)} - a_l^{(2)}
\]

Here, we subtracted the mean genotypic value and the additive effects, leaving the dominance effect as the only unexplained portion of the conditional mean at the locus. The mean dominance deviation at each locus is equal to zero, \( \sum_i \delta_{ij}^{(1)} p_i^{(1)} p_j^{(1)} = \sum_{kl} \delta_{kl}^{(2)} p_k^{(2)} p_l^{(2)} = 0 \).

The definition of epistatic effects proceeds in a similar fashion. Considering the mean phenotype of individuals with \( a_1^{(1)} = A_i \) and \( a_1^{(2)} = B_k \) without regard to the other two alleles, i.e. \( E[X|a_1^{(1)} = A_i, a_1^{(2)} = B_k] = \sum_{j,k,l} X_{ij,kl} p_{ij,l}^{(1)} p_{ij,k}^{(2)} \), the additive x additive effects are

\[
(\alpha \alpha)_{i,k} = E[X|a_1^{(1)} = A_i, a_1^{(2)} = B_k] - \mu_X - a_i^{(1)} - a_k^{(2)}
\]
i.e. the deviation of the conditional mean from the expectation based on the population mean and the additive effects, $\alpha_i^{(1)}$ and $\alpha_k^{(2)}$.

An additive×dominance effect measures the interaction between an allele at one locus with a particular genotype of another locus. It is defined as the deviation of the conditional mean $E[X|a_1^{(1)} = A_i, a_1^{(2)} = B_k, a_2^{(2)} = B_l]$ from the expectation based on all lower-order effects: the three additive effects, one dominance effect, and two additive × additive effects,

$$(\alpha \delta)_{i,kl} = (\alpha \delta)_{i,lk} = E \left[ X|a_1^{(1)} = A_i, a_1^{(2)} = B_k, a_2^{(2)} = B_l \right]$$

$$- \mu_X - \alpha_i^{(1)} - \alpha_k^{(2)} - \alpha_l^{(2)} - \delta_{kl} - (\alpha \alpha)_{i,k} - (\alpha \alpha)_{i,l},$$

and analogously

$$(\alpha \delta)_{j,k} = (\alpha \delta)_{ji,k} = E \left[ X|a_1^{(1)} = A_i, a_1^{(2)} = A_j, a_1^{(2)} = B_k \right]$$

$$- \mu_X - \alpha_i^{(1)} - \alpha_j^{(1)} - \alpha_k^{(2)} - \delta_{ij} - (\alpha \alpha)_{i,k} - (\alpha \alpha)_{j,k},$$

Finally for a dominance × dominance effect,

$$(\delta \delta)_{ij,kl} = X_{ij,kl} - \mu_X - \alpha_i^{(1)} - \alpha_j^{(1)} - \alpha_k^{(2)} - \alpha_l^{(2)} - \delta_{ij} - \delta_{kl} - (\alpha \alpha)_{i,k} - (\alpha \alpha)_{i,l} - (\alpha \alpha)_{j,k} - (\alpha \alpha)_{j,l} - (\delta \delta)_{ij,kl}.$$ 

Note: $(\delta \delta)_{ij,kl} = (\delta \delta)_{ij,kl} = (\delta \delta)_{ji,kl} = (\delta \delta)_{ji,kl}$. In summary: we started with the lowest-order effects, the additive effects of alleles. They account for as much of the variance in genotypic values as possible. We then progressively defined higher-order effects, each time accounting for as much of residual variation as possible. Summing up terms we have the following partitioning of the genotypic value,

$$X_{ij,kl} = \mu_X + [\alpha_i^{(1)} + \alpha_j^{(1)} + \alpha_k^{(2)} + \alpha_l^{(2)} + [\delta_{ij}^{(1)} + \delta_{kl}^{(2)}]$$

$$+ [(\alpha \alpha)_{i,k} + (\alpha \alpha)_{i,l} + (\alpha \alpha)_{j,k} + (\alpha \alpha)_{j,l}]$$

$$+ [(\alpha \delta)_{i,kl} + (\alpha \delta)_{i,kl} + (\alpha \delta)_{ji,kl} + (\alpha \delta)_{ji,kl}] + (\delta \delta)_{ij,kl}.$$ 

(6.3)

The different effects in (6.3) depend on allele frequencies in the population. However, the mean value of each type of effect is always zero.

Now, since we assumed random mating and independently segregating loci, the total genetic variance is simply the sum of the variances of the individual effects,

$$\sigma_X^2 = V(X) = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2$$ 

(6.4)
where

\[
\begin{align*}
\sigma_A^2 &= \sigma_{A,1}^2 + \sigma_{A,2}^2 \\
&= 2 \sum_i (\alpha_i^{(1)})^2 p_i^{(1)} + 2 \sum_k (\alpha_k^{(2)})^2 p_k^{(2)} \\
\sigma_D^2 &= \sigma_{D,1}^2 + \sigma_{D,2}^2 \\
&= \sum_{i,j} (\delta_{ij}^{(1)})^2 p_i^{(1)} p_j^{(1)} + \sum_{k,l} (\delta_{kl}^{(2)})^2 p_k^{(2)} p_l^{(2)} \\
\sigma_{AA}^2 &= 4 \sum_{i,j,k} ((\alpha\delta)_{i,j,k}^{(1)})^2 p_i^{(1)} p_j^{(1)} p_k^{(2)} \\
\sigma_{AD}^2 &= \sigma_{AD,12}^2 + \sigma_{AD,21}^2 \\
&= 2 \sum_{i,j,k} ((\alpha\delta)_{i,j,k}^{(1)})^2 p_i^{(1)} p_k^{(2)} p_j^{(1)} p_l^{(2)} + 2 \sum_{i,j,k} (\alpha\delta)_{i,j,k}^{(2)} p_i^{(1)} p_j^{(1)} p_k^{(2)} \\
\sigma_{DD}^2 &= \sum_{i,j,k,l} ((\delta\delta)_{i,j,k,l}^{(1)})^2 p_i^{(1)} p_j^{(1)} p_k^{(2)} p_l^{(2)}.
\end{align*}
\] (6.5)

Because of the hierarchical way in which genetic effects are defined, in general the magnitude of genetic variance components become progressively smaller at higher stages in the hierarchy. This, however, is not a valid argument for routinely ignoring epistatic effects altogether. Unless information on gene frequencies is available, the relative magnitude of variance components provide only limited insight into the physiological mode of gene action. Depending on the gene frequencies, epistatic interactions can greatly inflate the additive and/or dominance components of genetic variance. For example relatively small epistatic components of variance are not incompatible with the existence of strong epistatic gene action. This raises a serious issue for quantitative genetics since, from the standpoint of statistical power and experimental design, the many different types of epistatic variance components will be hard to track down. In summary, epistatic interactions are likely to be important in the expression of many quantitative traits as well as in the determining of levels of additive genetic variance.

In the above calculations we have been treating the transmission of genes at different loci as independent events. Such independence is generally true for genes located on different chromosomes, but when loci are physically linked on the same chromosome, a dependence can exist between the genes incorporated into gametes. Genes that lie on the same chromosome tend to be inherited as a group, a tendency that declines with increasing distance between the loci. As a result, haplotype frequencies may deviate from expectations based allele frequencies, a phenomenon known as linkage disequilibrium or, more general, gametic phase disequilibrium.

We still have little knowledge of the loci underlying most quantitative traits. However, theoretical arguments suggest that the aggregate effects of gametic phase disequilibrium might be extensive for quantitative traits whose expression is based on large numbers of loci, even if the average level of disequilibrium between pairs of
loci is relatively small. Gametic phase disequilibrium can cause either inflation or depression of both the additive and dominance genetic variance. When epistatic interactions exist between loci in gametic phase disequilibrium, the picture becomes extremely complicated.

### 6.2.2 Some notes on Environmental Variation

The expression of most quantitative traits is not completely governed by genetic determinants. Environmental effects are often subtle, causing simple amplifications or reductions in sizes of parts, numbers of progeny, physiological performance etc. As with genetic variance, sources of environmental variation can be partitioned in different ways. General environmental effects refer to influential factors that are shared by groups of individuals whereas special environmental effects are residual deviations from the expected phenotype based on genotype and general environmental effects. Ideally, the genotypic value and the different environmental effects behave in an additive fashion, with the phenotype of an individual simply being the sum of the genetic and environmental effects, \( y = X + E \). However, the picture is complicated by the fact that, in some instances, different genotypes respond to environmental change in nonparallel ways causing \( \text{genotype} \times \text{environment interaction} \).

Let \( E \) and \( e \) denote contributions of general and specific environmental effects to the phenotypic expression of a trait, and let \( I \) denote the genotype \( \times \) environment interaction effect. Then the phenotype of the \( k \)th individual of the \( i \)th genotype exposed to the \( j \)th level of the general environment \( E \) can be written as a linear function of four components,

\[
Y_{ijk} = X_i + E_j + I_{ij} + e_{ijk}.
\]

Each of these components may be further subdivided. For example we have already seen that the genotypic value, \( X_i \), is a potentially complicated function of population mean phenotype, additive allelic effects, dominance deviations and epistasis. The terms \( I_{ij}, \) \( E_j \) and \( e_{ijk} \) are defined in a least-squares sense as deviations from lower-order expectations, implying that their population mean values are equal to zero. The genotypic value, \( X_i \), is the mean phenotypic value of the particular genotype \( i \) averaged over all environmental conditions, whereas \( E_j \) is the deviation from the population mean caused by general environment \( j \). The quantity \( X_i + E_j + I_{ij} \) is the expected phenotype of genotype \( i \) in environment \( j \). Hence, \( I_{ij} \) is the residual deviation left after assuming that genotypic and environmental values act in an additive way.

By construction \( I \) and \( e \) are uncorrelated with the other variables (and each other). Hence the total genotypic variance, \( \sigma^2_X \), can be written

\[
\sigma^2_Y = \sigma^2_X + \sigma^2_E + \sigma^2_I + 2C(X, E) + \sigma^2_e.
\]
Genotype-environment covariance, $C(X, E)$, is a measure of the physical association of particular genotypes with particular environments. If individuals are randomly distributed with respect to environments, then $C(X, E)$ is zero.

### 6.3 Resemblance between relatives

In the previous sections it was seen that the phenotypic variance of a trait can be partitioned into a number of genetic and environmental components. From a practical point of view we need methods to estimate the magnitude of these components. The key to this matter is the fact that various genetic and environmental sources of variance contribute differentially to the resemblance between different types of relatives. For simplicity, we will concentrate on the genetic covariance between relatives, i.e. environmental causes of resemblance will not be considered here. Hence our basic model for the phenotypic values will be $Y = X + e$, where $e$ is the mean zero environmental deviation, which is assumed to be uncorrelated with the genotypic value $X$. Furthermore, because the environmental effects in this model are random residual deviations, they are uncorrelated among individuals and do not contribute to the resemblance between relatives. Let $Y_{i1} = X_{i1} + e_{i1}$ and $Y_{i2} = X_{i2} + e_{i2}$ be the phenotypic values of two members, $i_1$ and $i_2$ say, of a particular relationship. Under the assumed model, the phenotypic covariance between relatives $i_1$ and $i_2$ equals their genetic covariance, i.e.

$$C(Y_{i1}, Y_{i2}) = C(X_{i1} + e_{i1}, X_{i2} + e_{i2}) = C(X_{i1}, X_{i2}).$$

The genetic covariance, $C(X_{i1}, X_{i2})$, is a consequence of relatives inheriting copies of the same genes. As with genetic variance, the genetic covariance can be partitioned into components attributable to additive, dominance and epistatic effects. Each term consists of one of the components of variance, cf. Section 7.2, weighted by a coefficient that is determined by the distribution of the number of alleles identical by descent at the loci involved. We will only consider the ideal situation in which mating is random and trait loci are unlinked and in gametic phase equilibrium.

#### 6.3.1 Genetic covariance between relatives

Technically speaking all members of a population are related to some degree. To fix a frame of reference, we will consider the founders of a given pedigree, i.e. individuals whose parents are not members of the pedigree, to be unrelated. Two often used measures of relatedness are the kinship coefficient and the fraternity coefficient of two related individuals. For a single autosomal locus there are four ways in which we can choose one gene from each relative. Picking one allele at random from each relative, the kinship coefficient, $\Theta$, is defined as the probability that the two chosen
alleles are identical by descent (IBD). The fraternity coefficient, $\Delta$, is defined as the probability that the relative pair share both alleles IBD. These measures can be rather tricky to calculate if inbreeding is possible (individuals that contain pairs of alleles at a locus that are IBD are said to be inbred). However, if we only regard outbred pedigrees things are more straightforward.

**Example 46 (Calculation of coefficients of kinship and fraternity.)** A parent and its offspring always have one allele IBD. Hence if one allele is picked at random from both parent and child, the probability that the two chosen alleles are IBD, i.e. the coefficient of kinship, equals $0.5^2 = 0.25$. Obviously, the coefficient of fraternity is 0 for a parent and its offspring.

Similarly, two full sibs have 0,1, or 2 alleles IBD with probabilities 0.25, 0.50, and 0.25, respectively. If the sibs have one allele IBD, the probability that two randomly chosen alleles (one from each sib) are IBD is 0.25 (cf. the parent-offspring relation). If on the other hand the siblings have two alleles IBD, the corresponding probability equals 0.5. Since these are the only possibilities of picking a pair of alleles IBD from the two sibs, the coefficient of kinship equals $0.25 \times 0.50 + 0.50 \times 0.25 = 0.25$. The coefficient of fraternity is just the probability that the sibs share two alleles IBD, i.e. 0.25.

Relationship coefficients for other types of relationships can be examined in a similar way, cf. Table 6.3.

There is an alternative interpretation of the kinship coefficient which is of relevance for the derivation of linkage methods for quantitative trait loci, see Section 6.4. Consider the IBD-distribution of a relative pair at a given locus, i.e. the distribution of the number of alleles shared IBD at the locus. For an outbred pedigree, let $N_{i_1i_2}$ denote the number of alleles shared identical by descent between two relatives, $i_1$ and $i_2$. Let $C$ represent the event that two randomly drawn alleles, one from each relative, are IBD at the considered locus. Hence $P(C) = \Theta_{i_1i_2}$, the coefficient of kinship for the two relatives. If $p_C(i)$ denotes the conditional probability of the event $C$ given that $N_{i_1i_2} = i, i = 0, 1, 2$, a moments thought shows that $p_C(0) = 0, p_C(1) = 1/4$ and $p_C(2) = 1/2$. In summary, $P(C|N_{i_1i_2}) = N_{i_1i_2}/4$ and averaging with respect to the distribution of IBD-values shows that $\Theta_{i_1i_2}$ equals half the expected proportion of alleles shared identical by descent at the locus, i.e. $2\Theta_{i_1i_2} = E(N_{i_1i_2})/2$. Of course $\Delta_{i_1i_2} = P(N_{i_1i_2} = 2)$, by definition.
Assume first that a single locus is responsible for the genetic contribution to the trait variance. Let $\Lambda^j$ and $\tilde{\delta}^j$ denote the breeding value and dominance deviation, respectively, for the two relatives, $j = 1, 2$. Since $E(\Lambda^j) = E(\tilde{\delta}^j) = C(\Lambda^j, \tilde{\delta}^j) = 0$, we have

$$C(X_{i_1}, X_{i_2}) = C(\Lambda^{i_1} + \tilde{\delta}^{i_1}, \Lambda^{i_2} + \tilde{\delta}^{i_2})$$

$$= C(\Lambda^{i_1}, \Lambda^{i_2}) + C(\tilde{\delta}^{i_1}, \tilde{\delta}^{i_2})$$

$$= E(\Lambda^{i_1}\Lambda^{i_2}) + E(\tilde{\delta}^{i_1}\tilde{\delta}^{i_2})$$

(6.6)

In order to evaluate the two expectations in the last line of (6.6) we first condition on the number of alleles shared identical by descent by the two relatives at the locus. The breeding value of an individual, $\Lambda^j$, is given by

$$\Lambda^j = Z_1^{(j)} + Z_2^{(j)},$$

(6.7)

where the $Z$-terms correspond to the two additive allelic effects determined by the genotype of the individual. If mating is random, $Z_1^{(j)}$ and $Z_2^{(j)}$ are independent with mean 0 and variance $\sigma_A^2/2 = p_1 \tilde{\delta}_1^2 + p_2 \tilde{\delta}_2^2$. If the two relatives have no alleles IBD the resulting four $Z$-variables (additive allelic effects) are independent and equally distributed; if the pair have one allele IBD, two of the $Z$'s must be completely identical and there are only three independent $Z$'s involved in the covariance computation;

<table>
<thead>
<tr>
<th>Relationship</th>
<th>$\Theta$</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent-offspring</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
<tr>
<td>Grandparent-grandchild</td>
<td>$\frac{1}{8}$</td>
<td>0</td>
</tr>
<tr>
<td>Great grandparent-great grandchild</td>
<td>$\frac{1}{16}$</td>
<td>0</td>
</tr>
<tr>
<td>Half sibs</td>
<td>$\frac{1}{8}$</td>
<td>0</td>
</tr>
<tr>
<td>Full sibs, dizygotic twins</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>Uncle(aunt)-nephew(niece)</td>
<td>$\frac{1}{8}$</td>
<td>0</td>
</tr>
<tr>
<td>First cousins</td>
<td>$\frac{1}{16}$</td>
<td>0</td>
</tr>
<tr>
<td>Double first cousins</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{1}{16}$</td>
</tr>
<tr>
<td>Second cousins</td>
<td>$\frac{1}{64}$</td>
<td>0</td>
</tr>
<tr>
<td>Monozygotic twins</td>
<td>$\frac{1}{2}$</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.3: Coefficients of kinship and fraternity under the assumption of no inbreeding.
6.3. RESEMBLANCE BETWEEN RELATIVES

finally, if the relatives have two alleles IBD there are two pairs of identical Z’s among
the four additive allelic effects. From this it is relatively straightforward to show that

\[ E[A^i A^i | N_{i2} = n] = \begin{cases} 
0, & n = 0, \\
\sigma_A^2 / 2, & n = 1, \\
\sigma_A^2, & n = 2.
\end{cases} \]  

(6.8)

Similar reasoning for the product of dominance deviations gives

\[ E[\delta^i \delta^i | N_{i2} = n] = \begin{cases} 
0, & n = 0, \\
0, & n = 1, \\
\sigma_D^2, & n = 2.
\end{cases} \]  

(6.9)

(The case \( n = 1 \) in (6.9) relies on the fact that, in addition to \( E(\delta) = \sum_k \delta_k p_k \)
being equal to zero, we also have \( \sum_k \delta_k p_k = \sum_i \delta_i p_i = 0 \).) In summary,

\[ E[A^i A^i | N_{i2}] = \sigma_A^2 N_{i2} / 2 \]

(6.10)

\[ E[\delta^i \delta^i | N_{i2}] = \sigma_D^2 I\{N_{i2} = 2\} \]

where \( I\{N_{i2} = 2\} \) equals 1 or 0 depending on whether \( N_{i2} \) equals 2 or not.

Taking expectations with respect to the distribution of IBD-values, recalling that
\( E(N_{i2})/2 = 2 \Theta_{i2} \) and noting that \( E(I\{N_{i2} = 2\}) \) is just the probability of sharing
two alleles IBD, i.e. the fraternity coefficient \( \Delta_{i2} \), we finally have,

\[ C(X_i, X_i) = 2 \Theta_{i2} \sigma_A^2 + \Delta_{i2} \sigma_D^2. \]  

(6.11)

Next we consider the genetic covariance when two unlinked loci are governing
the genotypic value of an individual. First, conditioning on the IBD-status at the two
loci, \( N_{i2}^{(1)} \) and \( N_{i2}^{(2)} \), using notation from (6.5) we can show that

\[ C(X_i, X_i | N_{i2}^{(1)}, N_{i2}^{(2)}) = \]

\[
\sigma_{A,1}^2 N_{i2}^{(1)} / 2 + \sigma_{A,2}^2 N_{i2}^{(2)} / 2 \\
+ \sigma_{D,1}^2 I\{N_{i2}^{(1)} = 2\} + \sigma_{D,2}^2 I\{N_{i2}^{(2)} = 2\} \\
+ \sigma_{A}\sigma_{D} N_{i2}^{(1)} N_{i2}^{(2)} / 4 \\
+ \sigma_{AD,12}^2 N_{i2}^{(1)} I\{N_{i2}^{(2)} = 2\} / 2 + \sigma_{AD,21}^2 I\{N_{i2}^{(1)} = 2\} N_{i2}^{(2)} / 2 \\
+ \sigma_{DD}^2 I\{N_{i2}^{(1)} = 2\} I\{N_{i2}^{(2)} = 2\} 
\]

(6.12)

so that averaging with respect to the joint distribution of \( N_{i2}^{(1)} \) and \( N_{i2}^{(2)} \) (which is just
the product of the marginal distributions due to independence) we have

\[ C(X_i, X_i) = 2 \Theta_{i2} \sigma_A^2 + \Delta_{i2} \sigma_D^2 + (2 \Theta_{i2} \sigma_A^2 \sigma_D^2 + 2 \Theta_{i2} \Delta_{i2} \sigma_{AD}^2 + (\Delta_{i2})^2 \sigma_{DD}^2. \]  

(6.13)
It is straightforward to extend formulas (6.12) and (6.13) to an arbitrary number of loci. For example, in the special case of \( n \) contributing unlinked loci, excluding any interaction effects between loci, we simply have an additive contribution from each separate locus,

\[
C(X_{i1}, X_{i2} | N_{i1i2}; l = 1, \ldots, n) = \sum_{l=1}^{n} \left( \sigma^2_{A,l} N^{(l)}_{i1i2} / 2 + \sigma^2_{D,l} I\{N^{(l)}_{i1i2} = 2\} \right) \tag{6.14}
\]

and hence, by taking expectations with respect to the joint distribution of \( N^{(l)}_{i1i2}; l = 1, \ldots, n \)

\[
C(X_{i1}, X_{i2}) = 2 \Theta_{i1i2} \sigma^2_A + \Delta_{i1i2} \sigma^2_D \tag{6.15}
\]

where

\[
\sigma^2_A = \sum_{l=1}^{n} \sigma^2_{A,l} \quad \text{and} \quad \sigma^2_D = \sum_{l=1}^{n} \sigma^2_{D,l}
\]

is the total additive and dominance variance, respectively, summing contributions from each involved locus. In Figures 6.6 and 6.7 the correlation between sibling phenotypic values is illustrated by their joint phenotypic distribution for a simple model with a single biallelic QTL determining the genotypic values and an uncorrelated normally distributed environmental deviation.

Using the resemblance coefficients given in Table 6.3, explicit expressions for the genetic covariances of common types of relatives are given in Table 6.4 for the case with two contributing loci. Remember that the additive genetic variance is a function of all higher-order types of gene action and thus the relative magnitude of the coefficients in Table 6.4 does not imply that resemblance between relatives is influenced only slightly by dominance and epistatic gene action.

The coefficients in Table 6.4 permit the estimation of different variance components from linear combinations of different observed genetic covariances between relatives. For example, ignoring higher-order epistasis and environmental sources of covariance, \( 8 \times [\text{parent-offspring covariance} - (2 \times \text{half-sib covariance})] \) gives an estimate of \( 8[(\sigma^2_A/2 + \sigma^2_{AA}/4) - 2(\sigma^2_A/4 + \sigma^2_{AA}/16)] = \sigma^2_{AA}. \) Similarly, \( 2 \times [(4 \times \text{half-sib covariance}) - (\text{parent-offspring covariance})] \) is an estimate of \( \sigma^2_A. \) Hence, in principle, the analysis of a series of relationships provides a basis for partitioning the phenotypic variance into its elementary components. In practice, however, there are limitations that will prevent us from obtaining precise estimates of the variance components. Part of the variance, such as that caused by higher-order epistatic interactions, is essentially beyond reach in a statistical sense. Most practical applications of quantitative genetics have been concentrated on the additive genetic component of the phenotypic variance, with the remaining components being treated as noise.
6.3. RESEMBLANCE BETWEEN RELATIVES

Figure 6.6: Joint probability distribution of sibling phenotypic values for a trait influenced by a single biallelic locus with alleles $A_1$ and $A_2$ and a residual environmental deviation. The dominance coefficient, $k$, is put equal to 1 corresponding to a dominant genetic model with possible genotypic values either 0 or $2a$. The relative frequency used for the $A_2$ allele is 0.3. The conditional joint distribution given the siblings genotypic values is described by independent normal random variables each with variance $\sigma^2 = 0.64$ corresponding to a (broad sense) heritability for $a = 1$ of 61%. The phenotypic correlation between siblings is 0.46.

The ratio of the (total) additive variance to the total trait variance, $h^2 = \sigma_A^2 / \sigma_Y^2$, is known as the *heritability* of a trait, or more precisely the narrow-sense heritability. (Heritability in the broad sense is defined as the total genetic contribution to the trait variance, $H^2 = \sigma_X^2 / \sigma_Y^2$, cf. Example 25.) One reason for using the additive variance in the definition of the heritability concept is the desire for a parameter that describes the genetic resemblance between parents and offspring. It is however clear that we can use relationships other than parents and their offspring to approximately estimate the heritability of a trait. The first term in any genetic covariance expression is $2\Theta_{hi:z}\sigma_A^2$, cf. (6.13). Thus, under the assumption that the additive genetic variance is the major source of phenotypic covariance,

$$h^2 \approx \frac{C(Y_{hi}, Y_{hi})}{2\Theta_{hi:z}\sigma_Y^2}$$
Figure 6.7: Joint probability distribution of sibling phenotypic values for a trait influenced by a single biallelic locus with alleles $A_1$ and $A_2$ and a residual environmental deviation. The dominance coefficient, $k$, is put equal to 0.3 giving possible genotypic values either 0, 1.3$a$ or 2$a$. The relative frequency used for the $A_2$ allele is 0.5. The conditional joint distribution given the siblings genotypic values is described by independent normal random variables each with variance $\sigma^2 = 0.04$ corresponding to a (broad sense) heritability for $a = 1$ of 93%. The phenotypic correlation between siblings is 0.49.

This gives an approximation to the heritability. However, when the assumption of an ideal additive model doesn’t hold, heritability will be over-estimated on average. The possible bias can be evaluated when estimates of phenotypic covariance are available for more than one type of relatives.

We will next discuss methods used to track down the chromosomal locations of susceptibility loci underlying a continuous trait. The locus-specific (narrow sense) heritability of the trait defined as the part of $h^2$ attributable to a specific locus, say $l$, (i.e. $\sigma^2_{A_l}/\sigma^2_Y$) is seen to be one of the key parameters for successful applications of these procedures.
6.4. **LINKAGE METHODS FOR QUANTITATIVE TRAITS**

Table 6.4: Coefficients for the components of genetic covariance between different types of relatives.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>$\sigma_{AA}^2$</th>
<th>$\sigma_{AD}^2$</th>
<th>$\sigma_{AD}^2$</th>
<th>$\sigma_{DD}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent-offspring</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{4}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grandparent-grandchild</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{16}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Great grandparent-great grandchild</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{1}{64}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half sibs</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{16}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full sibs, dizygotic twins</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{8}$</td>
</tr>
<tr>
<td>Uncle(aunt)-nephew(niece)</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{16}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First cousins</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{1}{64}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double first cousins</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{16}$</td>
<td>$\frac{1}{16}$</td>
<td>$\frac{1}{64}$</td>
</tr>
<tr>
<td>Second cousins</td>
<td>$\frac{1}{32}$</td>
<td>$\frac{1}{1024}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monozygotic twins</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

6.4 **Linkage methods for quantitative traits**

Quantitative traits are often influenced by several, possibly interacting, genetic loci. Hence a fully parametric linkage approach requires accurate knowledge of the number of trait loci, the number of alleles at each trait locus as well as the distributional form of the trait values, conditional upon genotypes, and the parameters describing inheritance of marker and trait alleles. However, specifying the number of trait affecting loci, and the number of alleles at each locus is difficult or impossible unless the specific causative loci have been identified. Also, most segregation analytic methods underlying such an approach only model a single locus so that the correct genetic model cannot be obtained in these cases. If the trait model is incorrectly specified, estimates will generally be biased to some extent. To circumvent these issues in modelling, several model-free tests for genetic linkage have been proposed. The term ‘model-free’ is short for ‘do not require knowledge of the underlying genetic model’. The best known of these methods are the relative-pair regression approach of Hase-\_man and Elston (1972), together with its different modifications and extensions, and techniques for general pedigrees using variance component analysis, see e.g. Blangero et al. (2000). In the sequel we will restrict attention to these model-free approaches.

Model-free methods for quantitative traits evaluate the similarity among pairs of
individuals for both the marker and trait phenotypes. These methods estimate genetic similarities among individuals in a first step. Since this part of the analysis is independent of the particular phenotype measured, the same estimation techniques as for binary data can be applied. The next step is then to evaluate the evidence for a trait-influencing locus at specified locations. The model-free approaches typically involve parameters that partition the inter-individual trait variability among subjects into components that are due to a major locus linked to a marker locus versus components due to residual polygenic effects from unlinked loci. In single marker analysis, the recombination fraction, and hence the location of the disease locus, is often difficult to estimate due to confounding with the linked component of variance, i.e. they cannot be separately estimated from data. However, when multiple markers are used, a so called 'interval-mapping strategy' can be effectively employed to identify regions that show greatest evidence for linkage. The most tightly linked area gives the highest proportion of variance attributable to linked genetic factors.

Both the Haseman-Elston method (HE) and the variance component (VC) approach are based upon identity-by-descent sharing, conditional on pedigree marker information. With non-perfect marker information, conditional probabilities that pairs of individuals share 0, 1, or 2 alleles IBD can be estimated using available family members and population genotype frequencies, cf. subsection 5.1.3. Similarity in IBD sharing is then used to evaluate trait similarity by using either linear regression, as in HE, or variance component analysis. Both methods require specification of a model that partitions variances and covariances among pairs of relatives into components reflecting genetic and environmental factors. Thus, the strategy depends only upon observable quantities, unlike 'model-dependent' strategies.

Suppose that we have phenotypic pedigree data on an assumed polygenic continuous trait, $Y$, and assume further that the following simple relationship hold with respect to genetic and environmental influences,

$$ Y = X + e = \mu + \sum_{l=1}^{n} X^{(l)} + e $$

(6.16)

where $X$, as before, denotes the total genetic contribution, $\mu$ is the overall mean phenotypic value, $X^{(l)} = \alpha^{(l)} + \beta^{(l)}$ is the effect of the $l$th QTL, $n$ is the (unknown) number of influential QTLs and $e$ represents environmental deviation. We further assume that the $n$ loci are unlinked and that $X^{(l)}$ and $e$ are uncorrelated with expectation zero. The additivity of QTL effects assumed in (6.16) corresponds to a genetic model excluding epistatic interaction effects between influential loci. This basic model can be extended in several ways. It is possible to include interaction effects, e.g. by modelling additive $\times$ additive epistasis, cf. subsection 6.4.2. Other extensions involve shared environmental effects and genotype $\times$ environment interaction. However, we will not elaborate on these further model extensions here. Some methods,
6.4. **LINKAGE METHODS FOR QUANTITATIVE TRAITS**

...e.g. those based on variance components analysis, allow modelling of the mean of \( Y \) in terms of measured or known covariate information. For example, mean phenotype levels might vary with gender and/or age and is often influenced by different environmental exposures such as smoking behavior etc. In order to explain as much as possible of the variability in \( Y \) it is then desirable to include such information when the model is estimated. This will in addition increase the power to detect linkage to a QTL. When covariate information is included, the model (6.16) can be written as

\[
Y = \mu + \sum_{i=1}^{p} \beta_i z_i + \sum_{l=1}^{n} X^{(l)} + \epsilon 
\]

(6.17)

where the \( \beta_i \)'s are regression coefficients for the corresponding \( p \) predictors or covariate values \( z_i \). A model of this type, involving both so called fixed effects of known or measured covariates together with random effects of unknown quantities, such as the individual genotypic values, is said to be mixed.

Both the HE method and approaches using VC analysis involve the expected covariance between pairs of relatives under the assumed model, e.g. (6.17), and given the inheritance information from genotyped markers. For a pair of relatives, \( i_1 \) and \( i_2 \), let \( N^{(l)}_{i_1 i_2} / 2 \) be the proportion of alleles identical by descent at the \( l \)th trait locus and let \( I\{N^{(l)}_{i_1 i_2} = 2\} \) equal 1 or 0 depending on whether the pair shares two alleles IBD at the \( l \)th locus or not. These two quantities are measures of genetic similarity at the involved trait loci. Recall that the genetic covariance of two relatives in a pedigree is a function of IBD-status at the involved trait loci,

\[
C(X_{i_1}, X_{i_2} | N^{(l)}_{i_1 i_2} ; l = 1, \ldots, n) = \\
\sum_{l=1}^{n} \left( \sigma^2_{A,l} N^{(l)}_{i_1 i_2} / 2 + \sigma^2_{D,l} I\{N^{(l)}_{i_1 i_2} = 2\} \right).
\]

(6.18)

This formula applies for both models (6.16) and (6.17). Suppose now that we want to test for the presence of an effect of a QTL at a specified point in a given chromosomal region using, either single or multipoint, marker genotype information in the given area. Let \( X^{(l)} \) denote the genetic effect of a putative QTL in the studied region. Information about IBD-status at a specific location, \( x \), can be obtained from the genotypes of polymorphic genetic markers at or around that location. When information is complete, i.e. in case of perfect marker information, it will be known whether the pair of relatives share 0, 1, or 2 alleles IBD at \( x \). It follows that both \( N^{(l)}_{i_1 i_2} / 2 \) and \( I\{N^{(l)}_{i_1 i_2} = 2\} \) are known, should \( L \) be situated at \( x \). E.g. \( N^{(l)}_{i_1 i_2} / 2 \) would equal either 0, 0.5, or 1. If IBD-information is incomplete at the location for the trait locus \( L \), information can be summarized by the conditional probabilities of IBD-sharing, given marker genotype data. Using similar notation as in (5.13) let \( \pi^{(l)}_{k_1 i_2} = P(N^{(l)}_{i_1 i_2} = \)
appearance of the kinship coefficient, the expected value of the squared phenotype difference assuming model (6.16), the expected proportion alleles shared IBD at locus (6.19), is due to the fact that marker data in the specified chromosomal region provide no information on IBD-status at other unlinked locus

Let the quantitative trait values of a relative-pair be \( Y_i \), and \( Y_j \). The Haseman-Elston method is, in its original formulation, simply to regress the squared phenotype differences, \((Y_i - Y_j)^2\), from relative pairs of the same type on the expected proportion of alleles IBD, i.e. \( n_{11}^{(L)} \), at the test locus, see e.g. Sham (1998). The most common design uses trait data from sib-pairs. A regression coefficient significantly less than 0 is considered as evidence for linkage. To see that this procedure makes sense, consider the expected value of the squared phenotype difference assuming model (6.16),

\[
E \left[ (Y_i - Y_j)^2 \right] = V(Y_i - Y_j) \\
= V(Y_i) + V(Y_j) - 2C(Y_i, Y_j) \\
= 2\sigma_Y^2 - 2C(X_i, X_j)
\]

where \( \sigma_Y^2 \) is the total phenotypic variance and \( X \) refers to genotypic value. The total variance is a population parameter, independent of IBD-status, whereas in view of observed marker data in a chromosomal region, the covariance may be replaced by the expression in (6.19). Summing up terms, the conditional mean squared trait difference is a linear function of \( n_{11}^{(L)} \) and \( n_{21}^{(L)} \):

\[
\hat{C}(X_i, X_j | \text{MD}) = \\
\sum_{l=1}^\infty \left( \sigma_{A,l}^2 E[N_{i,i2}^{(l)} | \text{MD}] / 2 + \sigma_{D,l}^2 P\{N_{i,i2}^{(l)} = 2 | \text{MD}\} \right) \\
= n_{11}^{(L)} \sigma_{A,L}^2 + n_{21}^{(L)} \sigma_{D,L}^2 + 2\Theta_{n,i2} \sigma_{A,R} + \Delta_{n,i2} \sigma_{D,R}
\]

where

\[
\sigma_{A,R} = \sum_{l \neq L} \sigma_{A,l}^2
\]

\[
\sigma_{D,R} = \sum_{l \neq L} \sigma_{D,l}^2
\]

denote the residual additive and dominance genetic variances, respectively. The appearance of the kinship coefficient, \( \Theta_{n,i2} \), and the coefficient of fraternity, \( \Delta_{n,i2} \), in (6.19), is due to the fact that marker data in the specified chromosomal region provide no information on IBD-status at other unlinked trait loci. For example, for an unlinked locus \( l \neq L \), \( E[N_{i,i2}^{(l)} | \text{MD}] / 2 = E[N_{i,i2}^{(1)}] / 2 = 2\Theta_{n,i2} \).

### 6.4.1 Analysis of sib-pairs: The Haseman-Elston Method

Let the quantitative trait values of a relative-pair be \( Y_i \) and \( Y_j \). The Haseman-Elston method is, in its original formulation, simply to regress the squared phenotype differences, \((Y_i - Y_j)^2\), from relative pairs of the same type on the expected proportion of alleles IBD, i.e. \( n_{11}^{(L)} \), at the test locus, see e.g. Sham (1998). The most common design uses trait data from sib-pairs. A regression coefficient significantly less than 0 is considered as evidence for linkage. To see that this procedure makes sense, consider the expected value of the squared phenotype difference assuming model (6.16),

\[
E \left[ (Y_i - Y_j)^2 \right] = V(Y_i - Y_j) \\
= V(Y_i) + V(Y_j) - 2C(Y_i, Y_j) \\
= 2\sigma_Y^2 - 2C(X_i, X_j)
\]
\[ \hat{E} [(Y_i - Y_{i2})^2|\text{MD}] = \alpha + \beta_1 \pi_{i1i2}^{(L)} + \beta_2 \pi_{i2i2}^{(L)} \]

with

\[
\begin{align*}
\alpha &= 2 \left( \sigma_Y^2 - 2 \Theta_{i12} \sigma_{A,R}^2 - \Delta_{i12} \sigma_{D,R}^2 \right) \\
\beta_1 &= -2 \sigma_{A,L}^2 \\
\beta_2 &= -2 \sigma_{D,L}^2.
\end{align*}
\]

If we neglect the dominance contribution assuming strictly additive gene action at each locus, i.e. putting \( \sigma_{D,L} = \sigma_{D,R} = \sigma_D^2 = 0 \), the linear regression is simple using only one predictor: the expected proportion alleles IBD at the putative trait locus \( L \), \( \pi_{i1i2}^{(L)} \),

\[ \hat{E} [(Y_i - Y_{i2})^2|\text{MD}] = 2 \left( \sigma_Y^2 - 2 \Theta_{i12} \sigma_{A,R}^2 \right) - 2 \sigma_{A,L}^2 \pi_{i1i2}^{(L)}. \]

In a single-marker (two-point) analysis with recombination fraction \( \theta \) between marker and trait locus, the expected proportion of alleles IBD at the trait locus can be written as a linear function, involving \( \theta \), of the corresponding expected proportion alleles IBD at the marker locus. It follows that the regression can be formulated, again neglecting dominance, as

\[ \hat{E} [(Y_i - Y_{i2})^2|\text{MD}] = \tilde{\alpha} + \tilde{\beta} \pi_{i1i2}^{(M)} \]  

(6.20)

where \( \pi_{i1i2}^{(M)} = E[N_{i1i2}^{(M)}|\text{MD}]/2 \) is the expected proportion of marker alleles IBD. In this case the value of the regression coefficient, \( \tilde{\beta} \), depends on the type of relation between the two relatives. In the most common design, i.e. for (full) sibs it can be shown that

\[ E \left[ N_{i1i2}^{(L)}|N_{i1i2}^{(M)} \right] = 4 \theta (1 - \theta) + (1 - 2 \theta)^2 N_{i1i2}^{(M)} \]

and hence,

\[ \pi_{i1i2}^{(L)} = 2 \theta (1 - \theta) + (1 - 2 \theta)^2 \pi_{i1i2}^{(M)}. \]

It follows that in a sib-pair single marker analysis, the regression coefficient, \( \tilde{\beta} \), in (6.20) equals \(-2(1 - 2 \theta)^2 \sigma_{A,L}^2 \). In summary, in a single-marker analysis using \( n \) pairs of the same type of relatives, one regresses the squared phenotype differences of the pairs on the estimated fraction of alleles IBD at the marker locus, cf. Figure 6.8.

A significant negative slope indicates linkage to a QTL. This is a one-sided test, as the null hypothesis (no linkage) is \( H_0 : \tilde{\beta} = 0 \ (\sigma_{A,L}^2 = 0) \) versus the alternative \( H_1 : \tilde{\beta} < 0 \ (\sigma_{A,L}^2 > 0) \).

There are however several caveats with this approach. First, different types of relatives cannot easily be mixed in one test. Second, parents and their offspring share exactly one allele IBD and therefore cannot be used to estimate this regression,
since there is no variability in the predictor. Further, from a statistical point of view, the distributional requirements necessary for maximum-likelihood estimation, i.e. normally distributed, uncorrelated errors with constant variance, are likely not to be fulfilled, even in the case of perfect marker information. Finally, QTL position ($\theta$) and effect size ($\sigma^2_{\lambda,e}$) are confounded and cannot be separately estimated. Thus, in its simplest form, the HE method is a detection test rather than an estimation procedure. It can be shown that ignorance of the effects of dominance in the regression equation does not bias the estimate of $\beta$ in a sib-pair analysis.

The general conclusion from power studies is that the HE test has poor power in many settings. For example, in Figure 6.9 it can be seen that rather substantial random fluctuations of the slope estimates persist for reasonably sized samples even for heritabilities as large as 50%. This has lead to the development of various modifications and extensions of the procedure. The original one-marker analysis has been

![Graph](image-url)
6.4. LINKAGE METHODS FOR QUANTITATIVE TRAITS

Figure 6.9: Haseman-Elston regression based on simulated data from \( n = 200 \) sibpairs. The assumed genetic model is strictly additive with an additive genetic variance \( s_i^2_A = 0.5 \). Marker information is assumed to be perfect, and hence the true slope of the regression line is \(-1\) in each case. The three subplots in each row use a common value of the heritability, \( h^2 \): upper row subplots have \( h^2 = 0.50 \), middle row subplots have \( h^2 = 0.33 \), and lower row subplots have \( h^2 = 0.25 \). Bars represent 95\% pointwise confidence intervals, estimated from data, for the true values of the regression line.

extended in several ways by incorporating genetic information from two or more linked markers. In addition to gains in statistical power, estimation of QTL position, e.g. via a so called interval mapping approach is then possible. In the multipoint method of Kruglyak and Lander (1995), all marker information is used to provide a maximum-likelihood estimate at each point along the chromosomal segment of the actual distribution of IBD-values for each relative pair. A so called EM method (for references see e.g. Kruglyak and Lander (1995)) is then used to compute the regression using this distribution. Regressions are computed at each point along a chromosome, generating a LOD-score plot as a function of putative QTL position.

Perhaps the most serious drawback with the Haseman-Elston procedure is its inability, being based on relative pairs of the same type only, to account for all the
genetic information present in extended pedigrees. This is however not the case for the methods based on variance component analysis which consider all pairs of relatives in a pedigree simultaneously. However, in recent years several more powerful extensions of the original HE method have been suggested in the literature. For a recent review see Feingold (2002).

6.4.2 Linkage Analysis in General Pedigrees: Variance Component Analysis

In view of the comparatively low power to detect linkage to a QTL using relative-pair methods, such as the different variants of the Haseman-Elston approach, some pessimism has arisen concerning the possibility to dissect the genetic components underlying complex traits using linkage methods. However, methods based on variance component (VC) analysis are more powerful tools for linkage analysis of quantitative traits, compared to the Haseman-Elston type regression techniques, Williams and Blangero (1999). The reason for their superiority in this respect is mainly that they allow simultaneous consideration of all pedigree members, and not just relative pairs of fixed type. As in the HE method, the central idea is to identify loci making a significant contribution to the population variance of a trait by use of IBD probabilities estimated from genotyped marker loci. VC analysis is an old and well-established statistical technique used to separate the total variance of a quantitative variable into components due to various sources. The earliest versions of this method, in connection with linkage studies, used only one or two markers at a time. In recent years the method has been extended to a more powerful multipoint analysis, both with exact calculation of the conditional IBD-distribution in the studied chromosomal region for small to moderately sized pedigrees and using an approximate method for larger pedigree sizes. The expressions for the expected genetic covariance between relatives, which constitute the basis for the decomposition of the trait variability, are exactly the same as those underlying the HE method. In a single marker, two-point, analysis the formulas involve the recombination fraction between the marker and the trait locus together with the estimated distribution of IBD-values at the marker locus. When multiple relative pair types are considered, the recombination fraction can be estimated. However, usually an interval-mapping approach is used in which evidence for a genetic effect at a locus is based on multiple markers, and the strongest evidence is provided at the point with maximal LOD score. In the sequel we will assume that we have genotype information from multiple markers in a chromosomal region of interest.

The power to detect linkage using the VC method (and related methods) for quantitative trait linkage analysis is almost solely a function of the QTL-specific heritability. The power is an increasing function of the proportion of total trait variance attributed to the QTL. Heritabilities below 10% will in general lead to unrealistic
6.4. LINKAGE METHODS FOR QUANTITATIVE TRAITS

sample sizes. Theoretical considerations concerning the evolutionary history of mankind predict most QTLs to have a heritability below this limit. The hope is, from a linkage perspective, that most traits are influenced by at least one major locus with an effect large enough.

Testing effects of a single locus

Suppose that we focus interest on detection of the (marginal) effect of a single major locus whose unknown position is varied within the genomic region of interest. Hence the possible effects of other unlinked trait loci are lumped together in a residual polygenic effect which can be, and usually is, incorporated in the model as well. Assume that phenotypic values in a sampled pedigree can be described according to a (with respect to the different loci) strictly additive genetic model as in (6.16), or if we want to take covariate information into consideration, by (6.17). In any case, the first step of the VC approach is identical to the procedure used with the HE method: genotyped markers in the region for a putative QTL is used to estimate the inheritance pattern (the IBD-values), or its distribution, at a given test locus. The information acquired gives an estimate of the proportion of alleles IBD at the test locus for each relative pair in the studied pedigree through its conditional expectation given marker data, i.e. \( \pi_{niz}^{(L)} \). Similarly, for each relative pair, the probability of sharing both alleles IBD at the test locus in view of the marker data equals \( \pi_{niz}^{(L)} \). Hence, conditional on genotyped marker data, the phenotypic covariance of pedigree relative pairs is modelled by (6.19), exactly as for the HE method.

By assuming that the joint distribution of phenotypic values within pedigrees is multivariate normal, the likelihood of data can easily be written and numerical procedures used to estimate the variance components, \( \sigma_{A,L}^2 \), \( \sigma_{D,L}^2 \), \( \sigma_{A,R}^2 \), \( \sigma_{D,R}^2 \), and \( \sigma_e^2 \). This is a consequence of the multivariate normal distribution being fully specified by the mean value and covariance alone\(^2\). We can then test the null hypothesis that the genetic variance resulting from the \( L \)th QTL equals 0 (no linkage) by comparing the likelihood of this restricted model, i.e. the null model \( \sigma_{A,L}^2 = \sigma_{D,L}^2 = 0 \), with that of a model in which the variance from the \( L \)th QTL is estimated. The difference between the \( \log_{10} \)-likelihoods gives a LOD score that is the equivalent of the classical LOD score of linkage analysis. Twice the difference in ln-likelihoods of these two models yields a test statistic whose asymptotic distribution is a mixture of different

\(^2\)The multivariate normal ln likelihood for a pedigree with \( n \) individuals corresponding to model (6.17) is given by

\[
-\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln \det(\Sigma) - \frac{1}{2} r^T \Sigma^{-1} r
\]

where \( \Sigma^{-1} \) is the inverse of the covariance matrix with elements given by (6.19), \( \det(\Sigma) \) is the determinant of \( \Sigma \), \( r \) is a column vector of residuals with \( k \)th component \( Y_k - E(Y_k|z_k) = Y_k - \mu - \sum_j \beta_j z_{kj} \) and \( z_k \) the covariate values for the \( k \)th individual.
chi-square distributions. The simplest case results when only the additive variance component is being modelled, i.e. $\sigma^2_{D,L}$ is assumed equal to zero. In that case the test statistic is approximately distributed as a fifty-fifty mixture of a chi-square variable with 1 degree of freedom and a point mass at 0 when the null hypothesis is true.

Although the multipoint VC method is the currently most powerful technique for quantitative trait linkage analysis, a number of issues concerning the validity of the approach have been raised. Perhaps the biggest assumption underlying VC analysis is that of multivariate normality of the phenotype within pedigrees. It has been shown by Allison et al. (1999) that rather extreme deviations from normality may result in inflated Type I errors, i.e. too many false positive findings. One possible solution in these situations is to transform the phenotypic values such that the distribution of the transformed observations is closer to normality. Then the results of the analysis may on the other hand be harder to interpret since they refer to a transformed measurement scale. In any case, different findings with respect to the validity of the method are somewhat contradictory and there is a need for development of diagnostic tools.

Another point of concern in connection with the VC method is the selection principles applied when pedigrees are sampled, i.e. the ascertainment scheme used. VC analysis considers the phenotypic distribution given IBD information (and not vice versa). If pedigrees are not randomly sampled but instead selected on the basis of e.g. extreme phenotypic values of a proband and this ascertainment is not corrected for in the analysis, bias may result. Again opinions vary in the literature about the seriousness of the introduced bias in different situations. Two different proposed methods for ascertainment correction can be found in Hopper and Matthews (1982) and Elston and Sobel (1979).

It is very often stated that an additional benefit in VC analysis is the possibility to estimate the proportion of total variance attributable to a detected locus, the QTL-specific heritability. However, recently this has been strongly questioned, at least in the context of genomewide scans, Göring et al. (2001). In a pointwise approach, i.e. for a fixed genomic location $x$, it is correct that an essentially unbiased estimate of the locus-specific heritability is produced when the VC model is fitted. (At least this is true if we disregard possible ascertainment bias and consider effects of deviations from normality to be negligible). The situation is however radically different with genomewide scans and the authors show that the locus-specific effect size at genomewide LOD score peaks tend to be grossly inflated and can even be virtually independent of the true effect size. The reason for the bias is to be found in the high correlation between the observed statistical significance and the effect-size estimate. When the LOD score is maximized over the many pointwise tests being conducted throughout the genome, the locus-specific effect-size estimate is maximized as well. To get an impression of the possible magnitude of this bias the authors provide simulation results with varying number of QTLs each with a heritability of 10%. The simulations were
based on 1,000 randomly ascertained nuclear families with two offspring each. The mean estimated heritability at the position of the maximum LOD score was however approximately 25% and virtually independent of the number of QTLs. The situation was even worse when all LOD score peaks above 3 were considered (mean heritability estimate approximately 30%). It is true that the magnitude of the genomewide bias decreases both with sample size and the true value of the locus-specific heritability. However, it is argued in Göring et al. (2001) that most current data sets for mapping of complex human traits have nowhere near the required size to make the bias negligible. Further, the authors argue that the findings have wide-ranging implications, as they apply to all statistical methods of gene localization.

Joint consideration of several loci with or without epistasis

Most linkage analysis of oligogenic traits, influenced by several genetic loci, consider only one locus at a time. In principle, using VC analysis, it is possible to model the joint impact of several trait loci simultaneously. The effect of each QTL is assessed through QTL-specific additive and dominance variance components. Epistasis variance components can be included as well, although the experience of applications of epistatic models is still very limited. Based on theoretical arguments, a simultaneous analysis should be more powerful than a single-locus analysis. To some extent the different steps in the joint procedure resemble the approach of a conditional NPL analysis, (Section 5.3).

Suppose that we want to model the joint impact of two unlinked loci, $L_1$ and $L_2$, residing in two different chromosomal regions. To keep things reasonably simple we assume a model containing no marginal or epistatic dominance effects whatsoever. On the other hand we allow for additive $\times$ additive epistatic interaction between ‘the model loci’ $L_1$ and $L_2$ in addition to an additive polygenic effect accounting for a residual genetic contribution. Using (6.12) the genetic covariance of two relatives conditional on known IBD-status at each loci is given by

$$
C \left( X_{i_1}, X_{i_2} | N_{i_1 i_2}^{(l)}; l = 1, \ldots, n \right) = \sum_{l=1}^{n} \sigma_{AA}^{2} N_{i_1 i_2}^{(l)} / 2 + \sigma_{AA}^{2} N_{i_1 i_2}^{(l_1)} N_{i_1 i_2}^{(l_2)} / 4
$$

(6.21)

where the variance component $\sigma_{AA}^{2} = \sigma_{AA, L_1 L_2}^{2}$ corresponds to additive $\times$ additive epistatic interaction between loci $L_1$ and $L_2$. Usually information is incomplete concerning the inheritance patterns at loci $L_1$ and $L_2$. Further we do not include inheritance information at all with respect to the unlinked loci in the residual term of the model. Let $MD_1$ and $MD_2$ denote marker information in the two test regions for loci $L_1$ and $L_2$ respectively. Taking expectation of the expression in (6.21), conditional on marker data $MD_1$ and $MD_2$ we get
CHAPTER 6. QUANTITATIVE TRAIT LOCI

\[
\hat{C}(X_{i1}, X_{i2} | MD_1, MD_2) = \pi_{i1}^{(L_1)} \sigma_{A,L_1}^2 + \pi_{i2}^{(L_2)} \sigma_{A,L_2}^2 + \pi_{i1}^{(L_1)} \pi_{i2}^{(L_2)} \sigma_{A,A}^2 + 2 \theta_{i1i2} \sigma_{A,R}^2
\]  

(6.22)

where \( \sigma_{A,R}^2 = \sum_{l \neq L_1, L_2} \sigma_{A,l}^2 \) denote the residual additive variance.

To use the joint model in a genomewide scan we start by considering single-locus models at different locations throughout the genome. For a locus with a LOD score above some predefined threshold value we may fit the joint two-locus model by conditional inclusion of a second unlinked locus while keeping the first locus in the model. This can, for each (unlinked) location of the second locus, be done in two steps. First we might consider a model without the interaction term, i.e. putting \( \sigma_{A,A}^2 \) equal to zero. The conditional LOD score for inclusion of a marginal effect from the second locus when the first locus is already in the model is given by the difference between the log_{10}-likelihoods of the two models with and without the marginal effect of the second locus (both however containing an effect of the first, conditioning, locus). In a second step we might proceed by a similar comparison of a model with the two marginal effects to a larger model including the epistatic interaction between the two loci considered. For an example of joint modelling of marginal effects of trait loci affecting unesterified cholesterol concentrations in ten large Mexican American pedigrees see Almasy et al. (1999).

6.4.3 Software for quantitative trait linkage analysis

A very flexible program for genetic variance component analysis is the SOLAR package, Southwest foundation for Biomedical Research. The authors have implemented a multipoint IBD method for estimation of IBD sharing at arbitrary points along a chromosome for each relative pair. The method is approximative and is based on regression on IBD values at marker loci. The multipoint algorithm can handle large and complex pedigrees. It is for instance possible to model multiple trait loci, dominance, and epistasis. The main reference to the program is Almasy and Blangero (1998). More information can be found at http://www.sfbr.org/sfbr/public/software/solar/index.html.

Haseman-Elston regression and (a somewhat less flexible variant of) variance component analysis have been implemented in the GENEHUNTER (v. 2.1) software. The multipoint algorithm used is exact and can be applied to pedigrees of moderate size. For more information see Pratt et al. (2000) and for online documentation (concerning v. 2.0) http://linkage.rockefeller.edu/soft/gh/.
6.5 Exercises

6.1. The genotypic values of a quantitative trait are determined by a single biallelic locus with alleles $A_1$ and $A_2$. The mean phenotypic levels for the three different genotypes are: $m_{11}$ for $A_1$-homozygous individuals, $m_{12}$ for heterozygous individuals and $m_{22}$ for $A_2$-homozygous individuals.

(a) Characterize the genetic model in terms of $m_{11}$, $m_{12}$ and $m_{22}$ if
i. the locus has no effect on the trait;
ii. the two alleles act in a completely additive way.
iii. the $A_2$ allele is completely dominant;
(b) Calculate the values of $a$ and $k$ corresponding to $m_{11} = 10$, $m_{12} = 12$, $m_{22} = 16$.

6.2. For a randomly mating population and a single biallelic trait locus the average effect of allelic substitution is equal to the difference between the additive allelic effects, $\alpha = \alpha_2 - \alpha_1$. It was shown that the expected additive effect of a randomly drawn allele from the population is 0. Show that

(a) $\alpha_1 = -p_2 \alpha$
(b) $\alpha_2 = p_1 \alpha$

6.3. For a randomly mating population and a single biallelic trait locus the number of $A_2$-alleles, $N_2$, has a binomial distribution with parameters 2 and $p_2$. Use this to show that the additive genetic variance $\sigma_A^2$ is given by $2p_1p_2\alpha^2$.

6.4. For a randomly mating population and a single biallelic trait locus where the two alleles are equally frequent, is it possible for the dominance variance to be larger than the additive variance?

6.5. For a randomly mating population and a single biallelic trait locus, assume that $p_2 = 0.2$ and that the mean phenotypic values given the genotypes are: 1 ($A_1A_1$), 3 ($A_1A_2$), and 8 ($A_2A_2$).

(a) Calculate the mean phenotypic value in the population, the additive allelic effects, the dominance deviations, the additive variance and the dominance variance.
(b) Calculate the genetic correlation between
i. Monozygotic twins;
ii. Full sibs;
iii. Half sibs;
iv. First cousins;
v. Two unrelated individuals.
Chapter 7

Association Analysis

Genetic linkage is the tendency of short chromosomal segments to be inherited intact from parents to offspring. As a result, some combinations of alleles, i.e., haplotypes, on these short segments may be preserved over a large number of generations. This co-segregation of alleles is more pronounced the shorter the genetic distance is between the corresponding loci. The excessive co-occurrence of certain haplotypes, because of tight linkage or for other reasons, is known as allelic association. Linkage analysis can be used to perform a genome-wide search for the existence of trait loci using a relatively small number of markers. Association analysis, on the other hand, is often used in an attempt to confirm the involvement of a suspected allele thought to be of importance for a trait of interest, or of an associated allele at a closely linked locus. It is thus an important tool for mapping of genetic loci, which is complementary to linkage analysis. Historically, association studies between diseases and polymorphisms such as ABO blood groups and HLA antigens have resulted in many consistent findings. Association analysis promises to become an even more powerful tool in the near future and it has been suggested (Risch and Merikangas 1996) that it may become possible to test every locus in the genome for its involvement in a disease trait, using association analysis methods.

Consider two loci A and B with alleles $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ occurring at relative frequencies $p_1, p_2, \ldots, p_m$ and $q_1, q_2, \ldots, q_n$ in the population. There are a total of $mn$ possible haplotypes, which can be denoted as $A_1B_1, A_1B_2, \ldots, A_mB_n$, with corresponding relative frequencies $h_{11}, h_{12}, \ldots, h_{mn}$. If the occurrence of allele $A_i$ and the occurrence of allele $B_j$ in a haplotype are independent events, then the relative frequency of the joint occurrence of alleles $A_i$ and $B_j$ in a gamete is equal to the product of the marginal frequencies,

$$h_{ij} = p_i q_j.$$

If this equality does not hold, the alleles are said to be associated.
CHAPTER 7. ASSOCIATION ANALYSIS

Let \( \theta \) be the recombination fraction between the two loci \( A \) and \( B \) and let \( h_{ij0} \) denote the relative frequency of haplotype \( A_iB_j \) in the current generation. What will be the frequency of the same haplotype in the next generation under the assumption of random mating? Each haplotype in the next generation is either a recombinant (probability \( \theta \)) or a non-recombinant (probability \( 1 - \theta \)), with respect to loci \( A \) and \( B \). When the haplotype is non-recombinant it has probability \( h_{ij0} \) of being \( A_iB_j \). When it is recombinant, the probability of the haplotype being \( A_iB_j \) is simply \( p_iq_j \) under the assumption of random mating. Therefore the probability that a haplotype transmitted to the next generation is \( A_iB_j \) equals

\[
h_{ij1} = (1 - \theta)h_{ij0} + \theta p_i q_j.
\]

(7.1)

The change in haplotype relative frequency from generation 0 to generation 1 is thus

\[
h_{ij1} - h_{ij0} = \theta (p_i q_j - h_{ij0}).
\]

Hence the haplotype frequency will not change if there is no allelic association in the current generation, and if there is an association, the change is proportional to \( \theta \). If there are changes in haplotype frequencies between generations the two loci are said to be in gametic phase disequilibrium. The rate with which a randomly mating population approaches gametic phase equilibrium depends on the recombination fraction \( \theta \). Rewriting (7.1) as

\[
h_{ij1} - p_i q_j = (1 - \theta) (h_{ij0} - p_i q_j)
\]

we see that the distance between the haplotype frequency and its equilibrium value is diminished by a factor of \( (1 - \theta) \) per generation so that after \( k \) generations

\[
h_{ijk} - p_i q_j = (1 - \theta)^k (h_{ij0} - p_i q_j).
\]

(7.2)

The difference between a haplotype relative frequency and its equilibrium value is sometimes used as a measure of the magnitude of association between alleles. In Figure 7.1 the decline of gametic phase disequilibrium over generations is plotted for different values of the recombination fraction.

**Example 47** Suppose the recombination fraction between two loci equals 0.01. How many generations would it take to halve the magnitude of allelic associations as measured by the discrepancies \( h_{ij} - p_i q_j \), if we assume a large population in random mating? After \( k \) generations the magnitude of allelic associations is diminished by a factor of \( (1 - \theta)^k \). Hence equating \( (1 - \theta)^k \) with 0.5 gives the solution

\[
k = \ln(0.5) / \ln(1 - \theta)
\]

Substituting \( \theta = 0.01 \) gives \( k = 69 \) generations. \( \square \)
When a disease mutation first occurs, it does so on a particular chromosome and so is associated with all the other alleles at nearby loci on that chromosome. This association breaks down over the generations due to recombination, cf. (7.2). Most association studies use markers that are believed to be tightly linked, and hence associated, with disease loci. There are however multiple possible causes underlying an inferred allelic association.

How is gametic phase disequilibrium and allelic association generated in the first place? It can be generated by several different mechanisms such as random genetic drift, mutation, selection, and population admixture and stratification. It is useful to have an intuitive understanding of these processes in order to appreciate the strength and limitations of allelic association studies for gene mapping.

Random genetic drift is a term used to describe random changes in allele or haplotype distributions from one generation to the next in a finite population. The gene pool of one generation can be viewed as a random sample from the gene pool of the previous generation. Allele and haplotype relative frequencies are therefore subject to sampling variation. In addition, mutations can alter the allele and haplotype distribution from one generation to the next. The smaller the population the larger are the effects of mutation and sampling variation. The expected magnitude of gametic phase disequilibrium between two loci in a stable population is thus a function of
CHAPTER 7. ASSOCIATION ANALYSIS

<table>
<thead>
<tr>
<th>$N$</th>
<th>$A_i$</th>
<th>$B_j$</th>
<th>$A_iB_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.3</td>
<td>0.5</td>
<td>0.15</td>
</tr>
<tr>
<td>2000</td>
<td>0.2</td>
<td>0.4</td>
<td>0.08</td>
</tr>
<tr>
<td>10000</td>
<td>0.05</td>
<td>0.1</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 7.1: Allele and haplotype relative frequencies in three subpopulations

population size, recombination fraction, and mutation rate.

There are other mechanisms that can generate gametic phase disequilibrium regardless of the size of the population. One such mechanism is selection, which occurs when an individual’s genotype has an influence on reproductive fitness. Another important mechanism is population admixture and stratification. When a population consists of two or more subgroups which, for cultural or geographical reasons, have evolved more or less separately for many generations, two loci that are in gametic phase equilibrium in every subgroup may be in disequilibrium for the population as a whole. For this phenomenon to arise the population subgroups must show some variability in allele frequencies of the two loci.

**Example 48 (Spurious associations.)** Consider three populations that have reached gametic phase equilibrium with respect to alleles $A_i$ and $B_j$. Suppose that the population size ($N$), the relative frequency of allele $A_i$, the relative frequency of allele $B_j$, and the relative frequency of the haplotype $A_iB_j$ are as in Table 7.1. If these three subpopulations are merged what will be the allele and haplotype relative frequencies before any interbreeding takes place? We have

\[
P(A_i) = \frac{[0.3 \times 1000 + 0.2 \times 2000 + 0.05 \times 10000]}{13000} = 0.0923
\]

\[
P(B_j) = \frac{[0.5 \times 1000 + 0.4 \times 2000 + 0.1 \times 10000]}{13000} = 0.1770
\]

\[
P(A_iB_j) = \frac{[0.15 \times 1000 + 0.08 \times 2000 + 0.005 \times 10000]}{13000} = 0.0277
\]

The equilibrium relative frequency of $A_iB_j$ is $0.0923 \times 0.1770 = 0.0163$ which is distinct from 0.0277. Alleles $A_i$ and $B_j$ are therefore associated in the merged population.

As we have seen association between alleles at two loci occurs when the distribution of alleles at one of the loci is dependent of the allele present at the other locus. Linkage of a marker locus and a disease susceptibility locus always leads to an association, but for most pairs of loci that association is solely intra familial, i.e. there is no association at the population level. Allelic association (on the population level), on the other hand occurs when particular marker alleles appear more frequently in individuals with the disease than in individuals without the disease. The association
may or may not be due to linkage. The term ‘linkage disequilibrium’ is often used as a synonym for allelic association, but this is somewhat misleading in view of the many different possible underlying causes of allelic associations. A possibly more accurate term to describe the phenomenon is ‘gametic phase disequilibrium’. Linkage disequilibrium should properly refer only to allelic association that is due to linkage, i.e. that has not yet been broken up by recombination. Allelic association that is not caused by linkage disequilibrium can not be used for mapping loci.

Linkage analysis is a powerful tool for detecting the presence of a disease locus in a chromosomal region. However, it is not very efficient for fine mapping, since the discrimination between small differences in recombination fraction requires data on a large number of informative gametes. For example, observing no recombinant in 50 fully informative gametes suggests a recombination fraction of 0, but it is also not incompatible with a recombination fraction of 5%. (The probability of observing 50 non-recombinants if the true recombination fraction is 5% equals $0.95^{50} \approx 0.077$).

However, linkage analysis can be followed up by association analysis methods for finer mapping of disease loci. The rational is that, for most human populations, allelic associations due to tight linkage are only expected to exist between loci with recombination fractions of less than 1%. Many association methods can either be used to test for association in the presence of linkage or be used as linkage tests in the presence of association. Huge sample sizes are expected to fine map loci by linkage in the absence of allelic association, whereas feasible sample sizes often suffice when there is linkage disequilibrium.

The detection of allelic associations between two loci suggests that they are in tight linkage. However, as we already have seen, allelic associations can occur between loosely linked or even unlinked loci in the presence of e.g. population admixture and stratification. This must be taken into consideration in the design, analysis and interpretation of allelic association studies.

The simplest and oldest association analysis method is the case-control study. Two random samples are collected, one of persons with a particular disease (cases), the other of persons without that disease (controls). We can then test for whether a particular marker allele is more common among the cases than the controls. If the disease arose as a mutation on a chromosome bearing that particular marker allele, then a (statistically) significant association could be due to linkage disequilibrium between the marker and disease loci. But if the sample comes from a heterogeneous population made up of two or more strata (i.e. subpopulations), and the strata differ with respect to their joint disease-marker distribution, this can by itself cause an overall disease-marker association, even if there is no such association in any of the separate strata, cf. Example 48. Although this type of association is of no biological interest, it is a true population association, caused merely by heterogeneity in the population.

There are three main ways to avoid an association due merely to heterogeneity.
The first is to sample from a homogeneous population, but this may be difficult to achieve in practice. The second way is to include appropriate covariates, such as e.g. ethnicity, in the analysis, but this is not possible if the appropriate covariates, whether genetic or environmental, are not known. The third way is to use matched controls. Matching for ethnicity is necessary if other genetic factors could be causing an association, and one way to do this is to use family-based controls. The most commonly used family-based association method is the transmission/disequilibrium test (TDT). Many other recently proposed association analysis methods are generalizations and extensions of the TDT.

For a recent and comprehensive review of the many different family-based association methods proposed in the literature see Zhao (2000).

7.1 Family-based association methods

Family-based association designs offer a compromise between traditional linkage studies and case-control association studies. All methods proposed in the literature have the common feature of comparing alleles transmitted from the parents to alleles not transmitted to the affected offspring or alleles transmitted to the unaffected offspring. We will start by considering the original formulation of the TDT test, Spielman et al (1993). It was intended as a test for linkage with a marker located near a candidate gene, in cases where association between the marker and disease status had already been found.

7.1.1 The Transmission/Disequilibrium Test, TDT

The TDT method differs from IBD methods and parametric linkage methods in that the TDT evaluates departures from random assortment of alleles across families, whereas the other methods evaluate departures from random assortment of alleles within families. In other words, the TDT focuses on linkage between a specific marker allele and the disease allele, whereas the other linkage tests focus on linkage between a specific marker locus and the disease locus.

Consider a family with one affected child and marker allele information available for the child and both parents, i.e. a family trio. Assume that there is no segregation distortion, i.e. the particular allele passed on to a child is chosen randomly from the two alleles of each parent. In addition, assume that there are, effectively, only two alleles at the disease locus, with allele $D_1$ the disease allele and allele $D_2$ the normal allele. $D_1$ and $D_2$ may be groups of alleles rather than single alleles. The differential transmission of the parents’ alleles at the marker locus to the affected child provides evidence of both linkage and allelic associations in the population. We will first discuss the special case of only two alleles at the marker locus.
7.1. FAMILY-BASED ASSOCIATION METHODS

<table>
<thead>
<tr>
<th>Transmitted allele</th>
<th>Non-transmitted allele</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_1$</td>
</tr>
<tr>
<td>$M_1$</td>
<td>$a$</td>
</tr>
<tr>
<td>$M_2$</td>
<td>$c$</td>
</tr>
<tr>
<td>Total</td>
<td>$a + c$</td>
</tr>
</tbody>
</table>

Table 7.2: Numbers $a$, $b$, $c$, and $d$ of transmitted and non-transmitted marker alleles $M_1$ and $M_2$ among $2n$ parents of $n$ affected children.

Let the two alleles at the marker be denoted $M_1$ and $M_2$. Table 7.2 summarizes the number of alleles transmitted and not transmitted to the $n$ affected children of $2n$ parents. It is quite intuitive that only parents which are marker heterozygotes with genotype $M_1M_2$ can provide any information about the recombination fraction $\theta$. The irrelevance of the $M_1M_1$- and $M_2M_2$-parents leads to the TDT statistic of Spielman et al. (1993),

$$ TDT = \frac{(b - c)^2}{(b + c)}. \quad (7.3) $$

The TDT statistic uses an approximate $\chi^2$ test or an exact binomial test to compare the number of times that heterozygous parents transmits the alleles $M_1$ and $M_2$ to an affected child. If we think of the transmitted marker alleles as 'cases' and the non-transmitted alleles as 'controls' we see that they are perfectly matched for ethnicity (among other things), since they are from the very same persons. It can be shown that the TDT tests the joint null hypothesis that there is no linkage or there is no allelic association (the cause of any such association usually being assumed to be linkage disequilibrium). A significant result, if not due to chance, must be due to the presence of both linkage and association, or to any other reason that a particular allele is preferentially transmitted to offspring with disease. Thus the test assumes for its validity absence of what is referred to as meiotic drive or selection. Provided that association is present, the TDT often has more power than conventional linkage tests, but, since it uses within-family comparisons only, it is not affected by aspects of population structure that can lead to associations in the absence of linkage, (Ewens and Spielman 1995). For the interested reader, a more detailed motivation of the TDT test is given next.

The expected values of the entries in Table 7.2 need to take account of the selection of the family through the affected child. Let $p$ denote the relative frequency of the disease allele, $D_1$, and let $q$ be

---

1This material follows closely the computations in Curnow et al. (1998) and it is possibly somewhat demanding from a probabilistic point of view. It can be optionally skipped.
Here the unconditional probabilities, \( P(T_{ij}) \) are given by
\[
P(T_{ij}) = P(M_i M_j) P(F_1 \rightarrow M_i | M_i M_j) = \begin{cases} q^2 & \text{for } i = j = 1, \\ q(1-q) & \text{for } (i,j) = (1,2) \text{ or } (2,1), \\ (1-q)^2 & \text{for } i = j = 2 \end{cases}
\]
and by conditioning on the child’s disease genotype
\[
P(A) = P(A | D_1 D_1) p^2 + P(A | D_1 D_2) 2p(1-p) + P(A | D_2 D_2) (1-p)^2
= f_{11} p^2 + 2f_{12} p(1-p) + f_{22} (1-p)^2.
\]
Further, again by conditioning on the child’s disease genotype,
\[
P(A | T_{ij}) = \frac{P(A | C = D_1 D_1, T_{ij}) P(C = D_1 D_1 | T_{ij})}{P(C = D_1 D_1 | T_{ij})} + \frac{P(A | C = D_1 D_2, T_{ij}) P(C = D_1 D_2 | T_{ij})}{P(C = D_1 D_2 | T_{ij})} + \frac{P(A | C = D_2 D_2, T_{ij}) P(C = D_2 D_2 | T_{ij})}{P(C = D_2 D_2 | T_{ij})}
= f_{11} P(C = D_1 D_1 | T_{ij}) + f_{12} P(C = D_1 D_2 | T_{ij}) + f_{22} P(C = D_2 D_2 | T_{ij}).
\]
7.1. FAMILY-BASED ASSOCIATION METHODS

Here, with \( \alpha_j \) denoting the probability that \( F_1 \) transmits a disease allele conditional on the event \( T_{ij} \),

\[
P(C = D_1 D_1 | T_{ij}) = P(F_2 \rightarrow D_1) P(F_1 \rightarrow D_1 | T_{ij})
\]

\[
= p \alpha_j,
\]

\[
P(C = D_1 D_2 | T_{ij}) = P(F_2 \rightarrow D_2) P(F_1 \rightarrow D_1 | T_{ij})
\]

\[
+ P(F_2 \rightarrow D_1) P(F_1 \rightarrow D_2 | T_{ij})
\]

\[
= (1 - p) \alpha_j + p (1 - \alpha_j),
\]

\[
P(C = D_2 D_2 | T_{ij}) = P(F_2 \rightarrow D_2) P(F_1 \rightarrow D_2 | T_{ij})
\]

\[
= (1 - p)(1 - \alpha_j).
\]

From (7.8) and (7.9)

\[
P(A | T_{ij}) = f_1 \alpha_j + f_2 [(1 - p) \alpha_j + p (1 - \alpha_j)] + f_{22} (1 - p)(1 - \alpha_j)
\]

\[
= \alpha_j f_1 + f_2 (1 - p) f_{22}
\]

with \( Z \) shorthand for \( p f_1 - f_{12} + (1 - p)(f_{12} - f_{22}) \). Note that \( p f_{12} + (1 - p)f_{22} \) is the conditional probability that the child is affected given that \( F_1 \) transmits a normal \( D_2 \) allele, i.e., \( P(A | F_1 \rightarrow D_2) \). Let \( 1 - B \) denote the ratio of the probability that the chromosome of an affected child has a normal allele at the disease locus to the same probability for a random chromosome in the population. That is

\[
1 - B = \frac{P(F_1 \rightarrow D_2 | A)}{P(F_1 \rightarrow D_2)} = \frac{P(A | F_1 \rightarrow D_2)}{P(A)}
\]

(7.11)

where the last equality follows from Bayes theorem. It is now easy to show that

\[
B = \frac{PZ}{P(A)} = \frac{p [p (f_{11} - f_{12}) + (1 - p)(f_{12} - f_{22})]}{p^2 f_{11} + 2p (1 - p) f_{12} + (1 - p)^2 f_{22}}
\]

and hence from (7.5) and (7.10)

\[
P(T_{ij} | A) = P(T_{ij}) [1 + (\alpha_j - p) B / p].
\]

(7.12)

It remains to evaluate the conditional probabilities \( \alpha_j \) that \( F_1 \) transmits the disease allele given that \( F_1 \) has marker genotype \( M_i M_j \) and transmits an \( M_i \)-allele. Denote by \( R \) the event that the transmitted gamete is recombinant with respect to marker and disease locus and let \( \bar{R} \) denote the complementary event that no such recombination has occurred. Then

\[
\alpha_j P(T_{ij}) = P(F_1 \rightarrow D_1, \bar{R}) + P(F_1 \rightarrow D_1, T_{ij}, R)
\]

\[
= P(F_1 = M_i D_1 / M_i D_1, F_1 = M_i M_j, \bar{R})
\]

\[
+ P(F_1 = M_i D_1 / M_i D_1, F_1 = M_i M_i, R)
\]

\[
= P(F_1 = M_i D_1 / M_i D_1, F_1 = M_i D_1, \bar{R})
\]

\[
+ P(F_1 = M_i D_1 / M_i D_1, F_1 = M_i D_1, R)
\]

\[
+ P(F_1 = M_i D_2 / M_i D_1, F_1 = M_i D_1, \bar{R})
\]

\[
+ P(F_1 = M_i D_2 / M_i D_1, F_1 = M_i D_1, R)
\]

\[
= P(F_1 = M_i D_1 / M_i D_1, F_1 = M_i D_1)
\]

\[
+ \left[ P(M_i D_1 / M_i D_2)(1 - \theta) + P(M_i D_2 / M_i D_1)\theta \right] / 2
\]

\[
= (1 + I(i = j)) P(M_i D_1 / M_i D_1) + P(M_i D_1 / M_i D_2) / 2
\]

\[
+ \left[ \theta [P(M_i D_2 / M_i D_1) - P(M_i D_1 / M_i D_2)] / 2
\]

\[
+ [1 \theta (P(M_i D_2 / M_i D_1) - P(M_i D_1 / M_i D_2))] / 2
\]
Note from the last line of (7.13) that the recombination fraction $\theta$ only enter the expression for $\alpha_{ij}$ if $(i,j) = (1, 2)$ or $(i,j) = (2, 1)$. From the assumption of random mating and using (7.13) and (7.4) we find

$$\alpha_{ij} = \begin{cases} 
\frac{(pq + \delta)q}{q} & \text{if } i = j = 1, \\
\frac{(pq + \delta)q}{q - \theta \delta}q(1 - q) & \text{if } (i,j) = (1, 2), \\
\frac{(p(1 - q) - \delta)(1 - q) + \theta \delta}{q(1 - q)} & \text{if } (i,j) = (2, 1), \\
\frac{(p(1 - q) - \delta)}{q(1 - q)} & \text{if } i = j = 2
\end{cases}$$

Combining (7.6), (7.12), and (7.14) finally gives the probabilities in Table 7.3. It follows that the difference in expected values of $b$ and $c$ in Table 7.2 is

$$E(b - c) = 2nB\delta(1 - 2\theta)/p$$

Since $B = 0$ only if the three penetrances are equal and so the disease locus has no effect on the occurrence of the disease, the expected value of $b - c$ will only be 0 if either $\delta = 0$, no allelic association, or $\theta = 0$, no linkage. Thus the TDT statistic can only test the null hypothesis of no association, $\delta = 0$, if there is linkage, $\theta < \frac{1}{2}$, or the null hypothesis of no linkage, $\theta = \frac{1}{2}$, if there is association, $\delta \neq 0$. The cause of the association is not important.

The derivation of the TDT test assumes that the contributions from the two parents of an affected child are independent. This is true under the null hypothesis of no association, $\delta = 0$, since the selection of the parents depends on the disease alleles, and with no association, there is no correlation in the occurrences of disease alleles and marker alleles in the parents. The independence of parental transmissions is however not obvious when there is no linkage but there is allelic association. In fact, it can be shown that the independence assumption holds true in this situation only if the two marker alleles are equally frequent or the penetrances are multiplicative, i.e.

$$f_1^2 = f_1 f_{22}.$$  

However, it has been shown (Whittaker et al. 1998) that the mean and variance of $b - c$ are unaffected by linkage. Hence the $\chi^2$-test is still valid for large sample sizes when both parents are included in the analysis. Furthermore, there do exist an extension of the TDT which do not presuppose independence between parental contributions, see Schaid and Sommer (1993); Knapp et al. (1995). This more general test will have less power than the TDT when the independence assumption holds true, but will be more powerful when there is substantial deviation from the condition $f_1^2 = f_1 f_{22}$.

In the TDT the transmitted and the non-transmitted alleles contributed by a parent is regarded as a paired observation. It is also possible to analyze data from independent family trios by regarding the total collections of transmitted and non-transmitted alleles as two independent case-control samples. This gives the haplotype-based haplotype relative risk test (HHRR) introduced in Terwilliger and Ott (1992). Using the notation of Table 7.2 the test statistic is given by

$$\text{HHRR} = \frac{(b - c)^2}{(2a + b + c)(b + c + 2d)/4n}$$

which is asymptotically distributed as a $\chi^2$ random variable with one degree of freedom under the null hypothesis of no association. Unlike the TDT, the HHRR can
not be used as a test for linkage but it is a powerful test of allelic association when the recombination fraction is near 0. The test is, unlike the TDT, however not protected against population stratification.

**Validity of the TDT test**

The TDT is a test either for linkage in the presence of association, or for association in the presence of linkage. As a test for linkage it is valid for any number of affected children in the nuclear families, i.e. the parental alleles can be counted once for each affected offspring in this case. In fact, the TDT is a valid test of linkage in all situations, e.g. using extended pedigree data. The reason is that under the null hypothesis of no linkage, the Mendelian inheritance (random assortment) implies that the transmission or non-transmission to each offspring occurs independently. On the other hand, as a test for association either the parental alleles are counted only once (however many affected children there are), or the dependence among offspring must be allowed for in the analysis. The reason is that the transmissions from a parent to its affected children are correlated if there is linkage, even if there is no association (the null hypothesis in this case), Spielman and Ewens (1996).

Martin et al. (1997) provide a test statistic, $T_{sp}$, that employs the information on transmissions to both members of an affected sib pair and that is valid as a test of both linkage and association. The $T_{sp}$ is similar to the TDT and comparison of the two statistics gives some insight as to why the TDT is invalid as a test for association with affected sib pair data, cf. Wicks (2000). For parents with heterozygous marker genotype $M_1M_2$, let $n_{11}$ be the number who transmit $M_1$ to both of their (affected) children, let $n_{22}$ be the number who transmit $M_2$ to both of their children, and let $n_{12}$ be the number who transmit $M_1$ to one child and transmit $M_2$ to the other child. Then

$$T_{sp} = \frac{(n_{11} - n_{22})^2}{n_{11} + n_{22}}$$

and

$$TDT = \frac{(n_{11} - n_{22})^2}{(n_{11} + n_{22} + n_{12})/2}.$$ 

The TDT is a more powerful test of linkage for affected sib pair data than is $T_{sp}$. Wicks (2000) argue that the reason is due to the fact that the TDT applied to data from affected sib pairs utilizes excess sharing in identity-by-descent transmissions. That is, the tendency for $n_{11} + n_{22}$ to exceed $n_{12}$ in the presence of linkage. To see this note that

$$TDT = T_{sp} \times \frac{n_{11} + n_{22}}{(n_{11} + n_{22} + n_{12})/2}.$$ 

Hence the TDT when applied to affected sib pair data can be written as a product of $T_{sp}$ and a factor that is a measure of excess sharing. The presence of linkage
alone, without association, results in a tendency towards excess sharing. Therefore, positive test results for the TDT will sometimes be attributable to the presence of excess sharing when $T_{sp}$ alone is not large enough to provide significant evidence for the presence of association in addition to linkage.

7.1.2 Tests using a multiallelic molecular marker

If the marker employed has more than two different allelic variants, several different marker alleles might be associated with the disease allele. For a marker with $m$ different alleles, let $t_{ij}$ denote the number of parents with marker alleles $M_i$ and $M_j$ that transmit marker allele $M_i$ to the affected child. A rather obvious generalization of the biallelic TDT statistic is the multiallelic TDT (Bickemöller and Clerget-Darpoux, 1995)

$$TDT_a = \sum_{i=1}^{m} \sum_{j<i} (t_{ij} - t_{ji})^2 / (t_{ij} + t_{ji}),$$

(7.17)

that compares the number of $M_iM_j$-parents who transmit $M_i$ with the number who transmit $M_j$, summing over all heterozygous parental marker genotypes. The $TDT_a$ statistic evaluates evidence for asymmetry in a contingency table of transmitted and non-transmitted marker alleles and its value is compared with a $\chi^2$-distribution with $m(m-1)/2$ degrees of freedom. Since there is one degree of freedom associated with each different heterozygous marker genotype, corresponding to $m(m-1)/2$ possibly different association parameters, the test may lack power versus simpler patterns of association, e.g. when only one of the alleles is associated with disease.

To concentrate on the associations of individual marker alleles, a biallelic TDT statistic can be calculated, and tested, for each of the $m$ marker alleles combining all other marker alleles as a single allele. Thus for allele $M_i$ we have

$$TDT_i = \left[ \sum_{j\neq i} (t_{ij} - t_{ji}) \right]^2 / \sum_{j\neq i} (t_{ij} + t_{ji}).$$

(7.18)

The statistics $TDT_i$ are not independently distributed and so a correction for the multiple testing involved must rely on a Bonferroni type correction. However, it is possible to use a randomization procedure to evaluate the significance of e.g. the largest of $TDT_i$-statistics, the $TDT_{\text{MAX}}$. Morris et al. (1997) suggested that randomized data sets are generated by deciding, randomly and independently for each parent, whether or not to exchange the transmitted and untransmitted allele. For each randomized data set the $TDT_{\text{MAX}}$ statistic can be calculated. The value of the $TDT_{\text{MAX}}$ actually observed from data is then compared to the distribution of $TDT_{\text{MAX}}$-values calculated from the randomized data sets in order to assess statistical significance.
7.1. FAMILY-BASED ASSOCIATION METHODS

Specified parametric models have been used to derive other tests for multiallelic associations. Based on a generalization to more than two marker alleles of the calculations underlying the probabilities in Table 7.3, Sham and Curtis (1995) derived the ETDT, the extended TDT-test, from a logistic model for the probability that a parent of an affected child has marker alleles $M_i$ and $M_j$ and transmits $M_i$ to the child. The multiallelic test proposed in Clayton and Jones (1999) is derived from the conditional distribution of the marker genotype of the affected child given marker genotypes of the parents. A specific model, the generalized haplotype relative risk model, is used to formulate the likelihood of data, and the proposed test is a so called score test derived from the log likelihood.

7.1.3 No parental information available

The TDT uses data from families in which marker genotypes are known for the father, the mother, and the affected offspring, but only parents who are marker heterozygotes are considered. Since the TDT tests for unequal transmission of alleles from the parents to affected offspring, it cannot be performed if genotypic data for the parents are not available.

When diseases with onset in adulthood or in old age are studied, it may be impossible to obtain genotypes for markers in parents of the affected offspring. Instead several methods have been proposed for this situation that compare the marker genotypes in affected and unaffected offspring. Curtis (1997) has introduced a discordant-sibship test for association that compares the allele frequencies of sib pairs sampled from discordant sibships according to the following procedure: for each discordant sibship, randomly choose one affected sibling (the case) and then choose (randomly, if necessary) an unaffected sibling (the control) whose genotype is maximally different from that of the case. The sampling of maximally discordant sib pairs avoids the introduction of correlation terms arising from the use of multiple sibs. The procedure may lead to a loss of some information, especially when there are several affected siblings. To calculate the test statistic each marker allele in the affected individual is compared with each marker allele in the unaffected sibling. The approach is unbiased and can be extended to markers with multiple alleles by way of a likelihood model similar to that of Sham and Curtis (1995). Another approach, focused on markers with multiple alleles, were proposed by Boehnke and Langefeld (1998). The discordant-alleles test (DAT) is based on a homogeneity statistic for a $2 \times m$ contingency table, where $m$ is the number of marker alleles.

The sib TDT, (S-TDT)

The sib TDT (S-TDT) of Spielman and Ewens (1998) generalizes to sibships that contain more than single affected and single unaffected siblings. However, when
using these larger sibships, the test is valid only as a test of linkage, (similar to the TDT). For sib pairs it is identical to the test suggested by Curtis (1997). The S-TDT does not reconstruct parental genotypes and does not depend on estimates of allele frequencies. For situations where some families have parental genotypes available, other families have genotypes of unaffected sibs but not the parents available, and still others with both kinds of data available the S-TDT statistic can easily be combined with the TDT statistic into one overall test.

The sibships used have to meet two requirements: (1) there must be at least one affected and one unaffected sibling; and (2) the members of the sibship must not all have the same genotype. The ‘minimal configuration’ possible then consists of a discordant sib pair with different marker genotypes. The S-TDT determines whether the marker allele frequencies among affected offspring differ significantly from their unaffected sibs. As for the TDT, the S-TDT is protected against spurious association due to population admixture and stratification. Two procedures for evaluating statistical significance were proposed in Spielman and Ewens (1998), one Monte Carlo method based on permutation of affection status within each family and one large sample approach based on the normal distribution. The latter procedure is preferred in connection with the overall test combining the S-TDT with the TDT. Hypothetical data for a biallelic marker are shown in Table 7.4. With data of this type, it is perhaps tempting to use an ordinary $\chi^2$ test to look for departures from the null hypothesis. However this is not a valid approach with these aggregated data because of the dependence of the observations on sibs from the same family. Below we briefly outline the Monte Carlo permutation procedure suggested by Spielman and Ewens (1998) to evaluate statistical significance.

Consider a family with $a$ affected and $u$ unaffected sibs, each with known marker genotype. To determine what differences between affected and unaffected sibs would be produced by chance, permute the observed genotypes within each sibship as follows. Ignoring actual affection status, randomly choose $a$ of the sibs and assign them to the ‘affected’ category. The remaining $u$ sibs are assigned to the ‘unaffected’ category. Since the permutation is carried out within families, potential problems resulting from population structure are eliminated (as is true with the original TDT). The resulting number of different alleles in ‘affected’ and ‘unaffected’ sibs are then totaled over families. For a biallelic marker the simulation result is of the same form as the example in Table 7.4. This procedure is

<table>
<thead>
<tr>
<th>Sib status</th>
<th>No. of alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_1$</td>
</tr>
<tr>
<td>Affected</td>
<td>8</td>
</tr>
<tr>
<td>Unaffected</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 7.4: Total number of marker alleles in affected and unaffected members of sibships.
repeated a large number of times and in each replicate results in a data table analogous to Table 7.4. Together these tables provide the 'null' distribution for a test of linkage.

If there are only two marker alleles, \( M_1 \) and \( M_2 \), or if one marker (e.g. \( M_1 \)) is of particular interest and all other markers are grouped together as \( M_2 \), we proceed as follows: The number of \( M_1 \) alleles among individuals randomly chosen as 'affected' is used to test for linkage. A p-value is then calculated as the proportion of randomly generated tables in which this number is equal to, or more extreme than the observed value in the actual data. The precision in the p-value calculation can be made arbitrarily high by increasing the number of randomly generated tables.

Another approach for evaluation of statistical significance uses a normal approximation valid for large samples. In a sibship with \( a \) affected and \( u \) unaffected members, the total number of sibs is \( t = a + u \). For a biallelic marker, suppose that the number of sibs who are \( M_1M_1 \) is \( r \) and the number of sibs with genotype \( M_1M_2 \) is \( s \). Then, assuming the null hypothesis is true, the mean and variance of the number of \( M_1 \) alleles among affected sibs, \( Y \), conditional on \( a, u, r \) and \( s \) can be calculated by means of the hypergeometric distribution. The overall mean \( A \) and variance \( V \) under the null hypothesis of the number of \( M_1 \) alleles among affected sibs are given by summation over all families in the sample:

\[
A = \sum (2r + s)a/t
\]

and

\[
V = \sum au[4r(t - r - s) + s(t - s)]/[t^2(t - 1)].
\]

The induced test statistic \( z = (Y - A)/\sqrt{V} \) is approximately distributed according to a standard normal random variable when the null hypothesis holds true. This so called \( z \) score is well-suited for use when combining results from the S-TDT and the (original) TDT; see Spielman and Ewens (1998) for details.

The S-TDT is a test of linkage between marker and disease. However, it can be used to test for association when all sibships have the 'minimal configuration', i.e. precisely one affected and one unaffected sibling with different marker genotypes in each family. If there is no association between marker and disease, the two possible genotypic assignments for the affected and the unaffected sib are equally likely in these sibships. This property is what is simulated by the permutation procedure and hence the S-TDT is valid as a test for association for families of this type. For sibships that do not have the minimal configuration, the S-TDT is not valid as a test for association.

Curtis and Sham (1995) have shown that bias can arise in the original TDT if the genotype of one parent is missing. This is the case, even if it is clear which marker allele the available (heterozygous) parent transmitted to an affected child. For these families there might be marker information on unaffected sibs and, if so, the S-TDT can be used instead.
The procedure outlined above becomes more complex when a marker with multiple alleles, say \( m \), is investigated. Spielman and Ewens (1998) suggest the calculation of a \( z_{\text{MAX}} \) score analogously to the computation of the TDT\(_{\text{MAX}} \) statistic. That is, a \( z \) score can be calculated for each separate marker allele treating all other alleles as one. The \( z_{\text{MAX}} \) score is chosen as the largest absolute \( z \) score. Approximate significance points for this statistic can be found by simulation.

The Sibship Disequilibrium Test, (SDT)

Horvath and Laird (1998) introduced a discordant-sibship test, the sibship disequilibrium test (SDT), that uses data from all the affected and all the unaffected siblings. It can be used to detect both linkage in the presence of association and association in the presence of linkage.

Let \( M_1 \) and \( M_2 \) denote the alleles of a biallelic marker. For each sibship let \( m_A \) (\( m_U \)) be the mean number of \( M_1 \) alleles among the affected (unaffected) siblings and let \( d \) denote the difference \( m_A - m_U \). The SDT is a sign test based on these differences. Let \( d_+ \) be the number of sibships for which \( d > 0 \) and let \( d_- \) be the number of sibships with \( d < 0 \). The test statistic is defined as

\[
\text{SDT} = \frac{(d_+ - d_-)^2}{(d_+ + d_-)}.
\]

Significance can be evaluated by exact calculation using the binomial distribution or approximately via a \( \chi^2 \) distribution with one degree of freedom.

For a marker with \( m \) alleles the SDT is defined as a multivariate sign test based on differences \( d^i = m^i_A - m^i_U \), where \( m^i_A \) (\( m^i_U \)) is the average number of \( M_i \) alleles in the affected (unaffected) members of the sibship. Note that since \( \sum_j m^i_A = \sum_j m^i_U = 2 \) we have \( d^m = -\sum_{i=1}^{m-1} d^i \). Therefore \( d^m \) can be dropped without loss of information. There are several multivariate sign tests, the one suggested for use by Horvath and Laird (1998) is as follows: Let \( S^i = (s^1, s^2, \ldots, s^{m-1})^t \), where \( s^j = \sum_i \text{sgn}(d^j_i) \), \( d^j_i \) denotes the difference for the \( i \)th sibship, and \( \text{sgn}(d) \) equals 1, 0, or -1 depending on whether \( d > 0 \), \( d = 0 \), or \( d < 0 \). The test rejects the null hypothesis for large values of the statistic

\[
\text{SDT} = S^t V^{-1} S,
\]

where the matrix \( V \) has elements \( V_{jk} = \sum_i \text{sgn}(d^j_i)\text{sgn}(d^k_i) \). Under the null hypothesis of no linkage or no association, the multiallelic SDT asymptotically has a \( \chi^2 \) distribution with \( m - 1 \) degrees of freedom.

It is straightforward to combine TDT and SDT when data consist of a mixture of families with and without parental information for a biallelic marker. For families with marker genotypes of both parents available let \( b \) and \( c \) denote the total number of heterozygous parents that transmits an \( M_1 \) allele and \( M_2 \) allele to an affected child.
respectively. Calculate $d_+$ and $d_-$ for families with discordant sibships and missing parental genotype information. Define $B = b + d_+$ and $C = c + d_-$. The statistic

$$Z^2 = \frac{(B - C)^2}{B + C}$$

has a $\chi^2$ distribution with 1 degree of freedom under the null hypothesis of no linkage. If the families with parental genotype information consist of trios, this holds true also for the null hypothesis of no association.

7.1.4 An association test for extended pedigrees, the PDT

All tests discussed so far are concerned with unrelated nuclear families and/or sibships. A limitation of these tests is that, although they remain valid tests of linkage, they are not valid tests of association if related family entities from larger pedigrees are used. Martin et al. (2000) and Martin et al. (2001) have developed the pedigree disequilibrium test (PDT) to use with data from related nuclear families and/or discordant sibships from extended pedigrees. Like the original TDT it is valid even when there is population stratification.

The problem with testing for association with related families, is that genotypes of related individuals are correlated if there is linkage, even if there is no allelic association in the population. The strategy to overcome this difficulty proposed in Martin et al. (2000) is to base a test on a random quantity measuring association for the entire pedigree, rather than treating related nuclear families or sibships as if they were independent. A measure of association is defined for each triad and each discordant sib pair within a pedigree, and the average of these quantities is the measure of association for the pedigree. The contributions from different pedigrees are considered independent.

There are two types of families that may give information about association. Informative nuclear families consist of at least one affected child, both parents genotyped at the marker and at least one parent heterozygous. Informative discordant sibships have at least one affected and one unaffected sibling with different marker genotypes and may or may not have parental genotype data.

Consider a biallelic marker locus with alleles $M_1$ and $M_2$. For each triad within an informative nuclear family define $X_T$ to be the difference between the number of transmitted and non-transmitted $M_1$ alleles. A heterozygous parent contributes with either +1 or -1 to $X_T$. Similarly, define for each discordant sib pair within an informative discordant sibship $X_S$ to be the difference between the number of $M_1$ alleles in the affected and the unaffected child. This difference will be either -2, -1, 0, 1, or 2. For each pedigree with at least one informative nuclear family and/or at
least one informative discordant sibship define a summary measure\(^2\)

\[
D = \sum_j X_{Tj} + \sum_j X_{Sj}
\]

Under the null hypothesis of no association the expected values of both \(X_T\) and \(X_S\) are 0 for any triad and any discordant sib pair. Hence \(E(D) = 0\) for any pedigree. If \(N\) is the total number of unrelated pedigrees with at least one informative nuclear family or informative discordant sibship in the sample and \(D_i\) is the summary measure for the \(i\)th pedigree, then under the null hypothesis of no association,

\[
E \left( \sum_{i=1}^N D_i \right) = 0
\]

and

\[
V \left( \sum_{i=1}^N D_i \right) = \sum_{i=1}^N V(D_i) = E \left( \sum_{i=1}^N D_i^2 \right).
\]

This suggests the following test statistic for the PDT,

\[
PDT = \frac{\left( \sum_{i=1}^N D_i \right)^2}{\sum_{i=1}^N D_i^2}
\]

which is approximately \(\chi^2\)-distributed with 1 degree of freedom for large sample sizes.

Suppose that the data consist only of independent family trios. In this case the TDT can be used as a test for association. The TDT differs from the PDT in that it treats the contribution from heterozygous parents as independent. For the PDT, the trios are the independent units. How do the two tests compare in a sample of independent family trios?

For a biallelic marker locus, define a random variable for each heterozygous parent of an affected child, \(Y_i\) equal to the difference between the number of transmitted and non-transmitted \(M_1\) alleles, \(i = 1, \ldots, h\), with \(h\) denoting the total number of heterozygous parents in the sample. The TDT statistic is then given by

\[
TDT = \frac{\left( \sum_{i=1}^h Y_i \right)^2}{\sum_{i=1}^h Y_i^2}
\]

The numerators of the two statistics, TDT and PDT, are the same, but the variance estimates in the denominators differ. Families with a single heterozygous parent contribute equally to both statistics, but variances are estimated differently for families

\(^2\)Two other, slightly different, summary measures are discussed in Martin et al. (2000) and Martin et al. (2001)
with two heterozygous parents. We have
\[ \sum_{i=1}^{b} Y_i^2 = h \]
\[ \sum_{i=1}^{b} D_i^2 = h + 2(n_c - n_d) \]  

(7.21)

where \( n_c \) is the number of times that two heterozygous parents in a family transmit the same allele to the affected child, i.e. the number of concordant transmissions, and \( n_d \) is the number of times that two heterozygous parents in a triad transmit different alleles to the affected child, i.e. the number of discordant transmissions. It follows directly that
\[ \frac{TDT}{PDT} = 1 + \frac{2(n_c - n_d)}{h}. \]

Under the null hypothesis of no linkage or no association, \( E(n_c - n_d) = 0 \) and thus the two tests are asymptotically equivalent under the null hypothesis. Under the alternative hypothesis, when there is both linkage and association, the two tests are however not necessarily equivalent. Martin et al. (2000) refer to different genetic model examples in which each test is more powerful than the other. However, the authors conclude that under realistic assumptions there is likely to be little difference between the outcomes of the two tests. When the sample consists only of independent discordant sib pairs the PDT is the same as the sib TDT and the test of Curtis for a marker locus with two alleles, (Spielman and Ewens 1998; Curtis 1997).

Martin et al. (2000) showed by simulation that, when extended-pedigree data are available, substantial gains in power can be attained by using the PDT rather than other methods such as the TDT, the \( T_p \), the sib TDT or the SDT, that can only use a subset of the data in order to remain valid tests of association in the presence of linkage.

The discussion above has only concerned biallelic markers. One possible extension of the PDT for use with multiallelic markers, that was suggested by Martin et al. (2000), is to consider each allele versus all of the others and calculate a value for the PDT statistic for each allele. The multiple testing issue has to be addressed when evaluating statistical significance.
Chapter 8

Answers to Exercises

2.1. a) 0.9 b) 0.7
2.2. a) 0.6 b) 0.8
2.3. 0.1/0.7 = 0.143
2.4. 0.095
2.5. a)

\[ P(\text{affected and } i \text{ disease alleles}) = \begin{cases} 
(1 - p)^2 \cdot f_0 = 0.0722, & i = 0, \\
2p(1 - p) \cdot f_1 = 0.0570, & i = 1, \\
p^2 \cdot f_2 = 0.00225, & i = 2.
\end{cases} \]

b) 0.0570/(0.0722+0.0570+0.00225) = 0.4336

2.6. \( P(N = 1) = 2 \cdot 0.4 \cdot 0.6 = 0.48. \)

2.7. \( P(X < 0.6) = \int_0^{0.6} f(x)dx = [x^2]_0^{0.6} = 0.6^2 - 0^2 = 0.36. \)

2.8. a) \( P(\text{ASP}) = P(Y_1 = 1) P(Y_2 = 1|Y_1 = 1) = K_p^2 \lambda, \) with \( Y_1 \) and \( Y_2 \) as in Example 17.

b) Given \( N = 0, Y_1 \) and \( Y_2 \) are independent. Thus \( P(N = 0, \text{ASP}) = P(N = 0)P(Y_1 = 1)P(Y_2 = 1) = 0.25K_p^2. \)

c) \( z_0 = P(N = 0, \text{ASP})/P(\text{ASP}) = 0.25/\lambda, \)

2.9. a) There are \((2k - 1)\) outcomes \((X_1, X_2)\) with \( Y = k. \) Thus \( P(Y = k) = (2k - 1)/36, k = 1, \ldots, 6. \)

b) \( E(Y) = 1 \cdot P(Y = 1) + \ldots 6 \cdot P(Y = 6) = 2 \cdot 1/36 + \ldots + 6 \cdot 11/36 = 161/36 = 4.47. \)

c) \( P(Y = 5|X_1 = 5) = 5/6, \ P(Y = 6|X_1 = 5) = 1/6. \)
d) \( E(Y | X_1 = 5) = 5 \cdot P(Y = 5 | X_1 = 5) + 6 \cdot P(Y = 6 | X_1 = 5) = 5 \cdot 5/6 + 6 \cdot 1/6 = 31/6 = 5.17 \).

2.10. 
\[
E(X) = \int_0^1 x \cdot f(x) \, dx = \int_0^1 2x^2 \, dx = 2/3, \\
V(X) = E(X^2) - E(X)^2 = \int_0^1 x^2 \cdot f(x) \, dx - (2/3)^2 \\
\quad = \int_0^1 2x^2 \, dx - (2/3)^2 = 1/2 - (2/3)^2 = 1/18, \\
D(X) = 1/\sqrt{18} \approx 0.236.
\]

2.11. a) \( H = V(X)/V(Y_1) = (V(Y_1) - V(e_1))/V(Y_1) \iff V(Y_1) = V(e_1)/(1 - H) = 4/(1 - 0.3) = 5.71 \).
b) \( C(Y_1, Y_2) = C(X + e_1, X + e_2) = C(X, X) + C(X, e_2) + C(e_1, X) + C(e_1, e_2) = V(X) + 0 + 0 + 0 = V(X) = H \cdot V(Y_1) = 0.3 \cdot 5.71 = 1.72. \)
c) \( \rho(Y_1, Y_2) = C(Y_1, Y_2)/(D(Y_1)D(Y_2)) = HV(Y_1)/(\sqrt{V(Y_1)V(Y_2)}) = H = 0.3. \)

3.1. a) \( L(\psi) = \psi^{23}(1 - \psi)^{37} \)
b) \( \ln L(\psi) = 23 \ln(\psi) + 37 \ln(1 - \psi) \)
c) \( \hat{\psi} = 23/60 \approx 0.383 \), i.e. the relative proportion of heads.

3.2. a) \( \hat{\psi} = N/100 \)
b) \( 100N \in \text{Bin}(100, 0.5) \)
c) \( E(\hat{\psi}) = E(N)/100 = (100 \cdot 0.5)/100 = 0.5, \quad V(\hat{\psi}) = V(N)/100^2 = 100 \cdot 0.5 \cdot (1 - 0.5)/100^2 = 0.0025. \quad D(\hat{\psi}) = \sqrt{V(\hat{\psi})} = 0.05. \)
d) \( E(\hat{\psi}) = 0.5, \quad V(\hat{\psi}) = 0.25/n, \quad \text{and} \quad D(\hat{\psi}) = 0.5/\sqrt{n}. \) Thus there is no systematic error in \( \hat{\psi} \) and the precision of the estimator increases with the number of throws \( n. \)

3.3. Under \( H_0 \) the number of transmitted 1-alleles \( N \) has as \( \text{Bin}(200, 0.5) \)-distribution. Thus the \( p \)-value is \( P(|N - 100| \geq |131 - 100||H_0) = 1.39 \cdot 10^{-5}. \)

6.1. The number of alleles shared IBD (IBS) is 0 (1).

6.2. The ML-estimates \( \hat{z}_0 = 0.14, \hat{z}_1 = 0.45 \) and \( \hat{z}_2 = 0.41 \) in (5.7) falls within Holman’s triangle (\( \hat{z}_1 \leq 0.5, \hat{z}_1 + \hat{z}_2 \leq 1 \) and \( 3\hat{z}_1 + 2\hat{z}_2 \geq 2). \) Thus \( \hat{Z}(x) = 100\hat{z}_0 \log(4\hat{z}_0) + 100\hat{z}_1 \log(2\hat{z}_1) + 100\hat{z}_2 \log(4\hat{z}_2) = 3.224. \)

6.3. a) Using (5.11), the NPL score \( Z(x) = \sqrt{2/100(41 - 14)} = 3.818. \)
b) The \( p \)-value is \( 1 - \Phi(3.818) = 6.72 \cdot 10^{-5}. \)
c) No. Without normal approximation, \( Z(x) \) has a transformed binomial distribution under \( H_0. \) The exact pointwise \( p \)-value becomes 8.209 \cdot 10^{-5}. Further, the genomewide \( p \)-value corresponding to an NPL score of 3.818 is even larger. With the approximative method of Table 5.1, it is 0.123.
6.4. a) The probability that the marker locus and \( x \) are non-recombinant for I) both offspring and II) no offspring is \( \theta^2 \) and \( (1 - \theta)^2 \) respectively. Then \( P = \theta^2 + (1 - \theta)^2 = 0.82 \).

b) \( N|MD \in \text{Bin}(2, P) \), since both parents transmit the same grandparental allele at \( x \) with pr. \( P \). Thus \( \pi_0 = (1 - P)^2 = 0.0324 \), \( \pi_1 = 2P(1 - P) = 0.295 \) and \( \pi_2 = P^2 = 0.672 \).

6.5. The cousins share one allele IBD, which comes from the two siblings paternal grandfather.

7.1. (a) 
\[
m_{11} = m_{12} = m_{22} \quad \text{(i)};
\]
\[
m_{12} = (m_{11} + m_{22})/2 \quad \text{(ii)};
\]
\[
m_{12} = m_{22} \quad \text{(iii)}.
\]

(b) 
\[ a = 3 \] and \( k = -1/3 \), partly recessive model.

7.2. (a) 
\[
0 = \alpha_1 p_1 + \alpha_2 p_2 = \alpha_1 p_1 + (\alpha + \alpha_1)p_2 = \alpha_1(p_1 + p_2) + \alpha p_2 = \alpha_1 + \alpha p_2,
\]
since \( p_1 + p_2 = 1 \). Thus \( \alpha_1 = -p_2 \alpha \).

(b) 
From (a), \( \alpha_2 = \alpha + \alpha_1 = \alpha - p_2 \alpha = (1 - p_2) \alpha = p_1 \alpha \).

7.3. \( \sigma_A^2 = V(\hat{\mu}_X) = V(\hat{\mu}_X + \alpha N_2) = \alpha^2 V(N_2) \). Since \( N_2 \) is binomial with parameters 2 and \( p_2 \) it follows that \( \sigma_A^2 = 2p_1p_2 \alpha^2 \).

7.4. \( p_1 = p_2 \) implies \( \alpha = \alpha \) so the dominance variance is larger than the additive variance if \( (ak/2)^2 > a^2/2 \) i.e. \( k^2 > 2 \). Hence, the answer is yes, if \( |k| > \sqrt{2} \).

7.5. (a) 
\[
\mu_X = 1.92, \; \alpha_1 = -0.52, \; \alpha_2 = 2.08, \; \delta_{11} = 0.12, \; \delta_{12} = \delta_{21} = -0.48, \; \delta_{22} = 1.92, \; \sigma_A^2 = 2.1632, \; \sigma_D^2 = 0.2304.
\]

(b) 
(i) 1; (ii) 0.48; (iii) 0.23; (iv) 0.11; (v) 0.
Appendix A

The Greek alphabet

\[ \alpha, A \text{ alpha} \]
\[ \beta, B \text{ beta} \]
\[ \gamma, \Gamma \text{ gamma} \]
\[ \delta, \Delta \text{ delta} \]
\[ \varepsilon, E \text{ epsilon} \]
\[ \zeta, Z \text{ zeta} \]
\[ \eta, H \text{ eta} \]
\[ \theta, \Theta \text{ theta (also } \vartheta) \]
\[ \iota, I \text{ iota} \]
\[ \kappa, K \text{ kappa} \]
\[ \lambda, \Lambda \text{ lambda} \]
\[ \mu, M \text{ mu} \]
\[ \nu, N \text{ nu} \]
\[ \xi, \Xi \text{ xi} \]
\[ \pi, \Pi \text{ pi} \]
\[ \rho, P \text{ rho} \]
\[ \sigma, \Sigma \text{ sigma} \]
\[ \tau, T \text{ tau} \]
\[ \upsilon, \Upsilon \text{ upsilon} \]
\[ \varphi, \Phi \text{ phi} \]
\[ \chi, \chi \text{ chi} \]
\[ \psi, \Psi \text{ psi} \]
\[ \omega, \Omega \text{ omega} \]

Table A.1: The Greek alphabet
References


Index

Φ, see standard normal distribution
χ²-distribution, 24
p-value, 51, 99

a priori probability, 65
additive effect, 113
additive genetic variance, 115
affected sib pairs, 89
allele, 5
allelic association, 147
alternative hypothesis, see hypothesis
amino acid, 5
ASP, 30
Association Analysis, 147
association analysis, 11
autosomes, 5
average effect of allelic substitution, 114
average excess, 116

Bayes' rule, see Bayes’ theorem
Bayes’ theorem, 19
biallelic, 111
binomial distribution, 21
breeding value, 117
broad-sense heritability, 131
cdf, see distribution function
centiMorgan, see genetic map length
Central Limit Theorem, 24
chromosomes, 5
cM, see genetic map length
codon, 5
complement, see set complement
components of variance, 56
conditional density function, 28
conditional distribution, 29
conditional expectation, 41, 42
conditional NPL, 106
conditional probability, 15
conditional probability function, 28
continuous random variable, 23
correlation, 37, 38
covariance, 37
crossover, 6
cumulative distribution function, see distribution function
density function, see probability density function
deoxyribonucleic acid, 5
deterministic model, 13
discrete random variable, 21
disease locus homozygosity, 76
disease susceptibility allele, 16
disjoint events, 14, 15
distribution function, 25
DNA, see deoxyribonucleic acid
dominance coefficient, 111
dominance genetic variance, 115
dominant, 9
dominant disease, 17

EM-algorithm, 99
epistasis, 106, 109, 120
event, 13
expected value, 34, 42
fitted value, see predicted response
**INDEX**

<table>
<thead>
<tr>
<th>term</th>
<th>page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed effects</td>
<td>135</td>
</tr>
<tr>
<td>founders</td>
<td>10</td>
</tr>
<tr>
<td>fraternity coefficient</td>
<td>126</td>
</tr>
<tr>
<td>fully penetrant</td>
<td>16</td>
</tr>
<tr>
<td>gametic phase disequilibrium</td>
<td>124, 148</td>
</tr>
<tr>
<td>Gaussian distribution, see normal distribuition</td>
<td></td>
</tr>
<tr>
<td>gene</td>
<td>5</td>
</tr>
<tr>
<td>gene content</td>
<td>112</td>
</tr>
<tr>
<td>genetic covariance</td>
<td>126</td>
</tr>
<tr>
<td>genetic map length</td>
<td>7</td>
</tr>
<tr>
<td>genetic model</td>
<td>9</td>
</tr>
<tr>
<td>genotype</td>
<td>5, 17</td>
</tr>
<tr>
<td>genotypic value</td>
<td>110</td>
</tr>
<tr>
<td>Greek alphabet</td>
<td>171</td>
</tr>
<tr>
<td>Haldane's map function</td>
<td>8</td>
</tr>
<tr>
<td>haplotype</td>
<td>6, 64</td>
</tr>
<tr>
<td>haplotype-based haplotype relative risk</td>
<td>156</td>
</tr>
<tr>
<td>Hardy-Weinberg equilibrium</td>
<td>17–19</td>
</tr>
<tr>
<td>Haseman-Elston Method</td>
<td>136</td>
</tr>
<tr>
<td>heritability</td>
<td>40, 131</td>
</tr>
<tr>
<td>heterogeneity</td>
<td>78</td>
</tr>
<tr>
<td>heterogenic disease</td>
<td>9</td>
</tr>
<tr>
<td>heterozygosity</td>
<td>20</td>
</tr>
<tr>
<td>heterozygous</td>
<td>6</td>
</tr>
<tr>
<td>HHRR, see haplotype-based haplotype relative risk</td>
<td></td>
</tr>
<tr>
<td>Hidden Markov models</td>
<td>97</td>
</tr>
<tr>
<td>Holman's possible triangle</td>
<td>92</td>
</tr>
<tr>
<td>homozygous</td>
<td>6</td>
</tr>
<tr>
<td>homozygous effect</td>
<td>111</td>
</tr>
<tr>
<td>hypothesis</td>
<td>49</td>
</tr>
<tr>
<td>hypothesis testing</td>
<td>49</td>
</tr>
<tr>
<td>IBD, see identical by descent, see identical by descent</td>
<td></td>
</tr>
<tr>
<td>IBS, see identical by state</td>
<td></td>
</tr>
<tr>
<td>identical by descent</td>
<td>22, 89</td>
</tr>
<tr>
<td>identical by state</td>
<td>89</td>
</tr>
<tr>
<td>incomplete marker information</td>
<td>97</td>
</tr>
<tr>
<td>incomplete penetrance</td>
<td>77</td>
</tr>
<tr>
<td>independence</td>
<td>38</td>
</tr>
<tr>
<td>independent events</td>
<td>17</td>
</tr>
<tr>
<td>independent random variables</td>
<td>31</td>
</tr>
<tr>
<td>information content</td>
<td>97</td>
</tr>
<tr>
<td>inheritance vector</td>
<td>102</td>
</tr>
<tr>
<td>kinship coefficient</td>
<td>126</td>
</tr>
<tr>
<td>Kolmogorov's axiom system</td>
<td>15</td>
</tr>
<tr>
<td>Law of Large Numbers</td>
<td>34</td>
</tr>
<tr>
<td>Law of total probability</td>
<td>18</td>
</tr>
<tr>
<td>least squares linear regression</td>
<td>55</td>
</tr>
<tr>
<td>likelihood function</td>
<td>45</td>
</tr>
<tr>
<td>likelihood ratio test</td>
<td>52</td>
</tr>
<tr>
<td>linear regression</td>
<td>54</td>
</tr>
<tr>
<td>linkage analysis</td>
<td>10</td>
</tr>
<tr>
<td>linkage disequilibrium</td>
<td>11, 124</td>
</tr>
<tr>
<td>linked loci</td>
<td>8</td>
</tr>
<tr>
<td>location score</td>
<td>85</td>
</tr>
<tr>
<td>locus</td>
<td>5</td>
</tr>
<tr>
<td>lod score</td>
<td>53, 61</td>
</tr>
<tr>
<td>log-odds, see lod score</td>
<td></td>
</tr>
<tr>
<td>marker homozygosity</td>
<td>75</td>
</tr>
<tr>
<td>markers</td>
<td>10</td>
</tr>
<tr>
<td>maximum likelihood estimator</td>
<td>46</td>
</tr>
<tr>
<td>maximum lod score</td>
<td>92, 93</td>
</tr>
<tr>
<td>maximum-likelihood estimation</td>
<td>138</td>
</tr>
<tr>
<td>meiosis</td>
<td>6</td>
</tr>
<tr>
<td>missing marker data</td>
<td>71</td>
</tr>
<tr>
<td>mixed genotypic values</td>
<td>135</td>
</tr>
<tr>
<td>ML-estimator, see maximum likelihood esti</td>
<td></td>
</tr>
<tr>
<td>MLS, see maximum lod score model</td>
<td>13</td>
</tr>
<tr>
<td>monogenic disease</td>
<td>9</td>
</tr>
<tr>
<td>Morgan, see genetic map length</td>
<td></td>
</tr>
<tr>
<td>multi-point linkage analysis</td>
<td>84</td>
</tr>
<tr>
<td>narrow-sense heritability</td>
<td>131</td>
</tr>
</tbody>
</table>
INDEX

Nonparametric Linkage Analysis, 89
nonparametric linkage score, 94
normal allele, 16
normal distribution, 24, 35
NPL, see Nonparametric Linkage Analysis
NPL score, see nonparametric linkage score
nucleotides, 5
null hypothesis, see hypothesis
overdominance, 111

Parametric Linkage Analysis, 59
parametric linkage analysis, 54
pedigree likelihood, 67
penetrance, 16
penetrance parameters, 9
phase, 10
phenocopies, 16
phenocopy, 78
phenotypic value, 110
point estimator, 46
pointwise \( p \)-values, 100
polygenic disease, 9
polymorphism, 20
power, 50, 85, 99
power function, 50, 51
powerful test, 53
predicted response, 55
prevalence, 13
probability density function, 23
probability function, 21

QTL, see quantitative trait loci
quantiles, 27
quantitative trait loci, 109
quartiles, 28

random effects, 135
random model, 13
random variable, 20

recessive, 9
recessive disease, 17
recessive mode, 80
recombinant, 7, 10
recombination fraction, 8
reduced penetrance, 77
regionwise \( p \)-values, 100
regression coefficient, 54
relative risk for siblings, 16
residual error, 54

S-TDT, see sib TDT
sample space, 13
scaling properties, 36, 39
SDT, see Sibship Disequilibrium Test
segregation analysis, 9
set complement, 15
sib TDT, 159
sibling prevalence, 15, 16
Sibship Disequilibrium Test, 162
significance level, 50, 51
significant, 51
significant linkage, 101
simulation, 85
single locus, 111
single nucleotide polymorphism, 31
SNP, 31
spurious association, 150
squared deviation, 36
standard deviation, 35
standard normal distribution, see normal distribution, 26
standardized random variable, 37
statistical model, 45
suggestive linkage, 101

TDT, see Transmission Disequilibrium Test, see Transmission Disequilibrium Test
test statistic, 49
Transmission Disequilibrium Test, 49, 152
two-point linkage analysis, 59
uncorrelated random variables, 38
underdominance, 111
uniform distribution, 23, 34
uninformativeness, 75
unlinked loci, 121
unlinked loci, 8

variance, 35
variance component analysis, 140
variance of a uniform distribution, 36
VC analysis, see variance component analysis