The Statistical Theory Underlying Human Genetic Linkage Analysis Based on Quantitative Data from Extended Families

Ushma Galal

A thesis submitted in fulfillment of the requirements for the degree of Magister Scientiae in the Department of Statistics, University of the Western Cape

Supervisors: Professor Lize van der Merwe and Professor Renette Blignaut

November 2010

Keywords

Fixed effects

Variance-components

Random effects

Mixed-models

Genetics

Inherited traits

Family studies

Extended pedigrees

Statistical genetics

Linkage analysis

Abstract

The Statistical Theory Underlying Human Genetic Linkage Analysis Based on Quantitative Data from Extended Families

U. Galal
MSc Thesis, Department of Statistics, University of the Western Cape

Background

Traditionally in human genetic linkage analysis, extended families were only used in the analysis of dichotomous traits, such as Disease/No Disease. For quantitative traits, analyses initially focused on data from family trios (for example, mother, father, and child) or sib-pairs. Recently however, there have been two very important developments in genetics: It became clear that if the disease status of several generations of a family is known and their genetic information is obtained, researchers can pinpoint which pieces of genetic material are linked to the disease or trait. It also became evident that if a trait is quantitative (numerical), as blood pressure or viral loads are, rather than dichotomous, one has much more power for the same sample size. This led to the development of statistical mixed models which could incorporate all the features of the data, including the degree of relationship between each pair of family members. This is necessary because a parent-child pair definitely shares half their genetic material, whereas a pair of cousins share, on average, only an eighth. The statistical methods involved here have however been developed by geneticists, for their specific studies, so there does not seem to be a unified and general description of the theory underlying the methods.
The aim of this dissertation is to explain in a unified and statistically comprehensive manner, the theory involved in the analysis of quantitative trait genetic data from extended families. The focus is on linkage analysis: what it is and what it aims to do. There is a step-by-step build up to it, starting with an introduction to genetic epidemiology. This includes an explanation of the relevant genetic terminology. There is also an application section where an appropriate human genetic family dataset is analysed, illustrating the methods explained in the theory sections.

November 2010

Declaration

I declare that *The Statistical Theory Underlying Human Genetic Linkage Analysis Based on Quantitative Data From Extended Families* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Ushma Galal

February 2011

Signed: .....

Acknowledgments

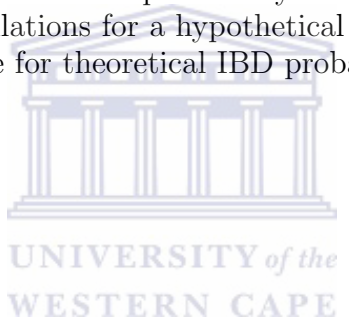
Thank you to the Biostatistics Unit of The South African Medical Research Council, Tygerberg, Cape Town for their support and encouragement.

Thank you also to Professor Johanna C. Moolman-Smook from the MRC Centre for Molecular and Cellular Biology, University of Stellenbosch Health Sciences Faculty, for allowing me to use their data.

Finally, thank you to Professors Lize van der Merwe and Renette Blignaut, my supervisors, for their help, support and dedication to this study.

Contents

List of Figures



List of Tables

List of Symbols

Symbol	Meaning
	Greek letters represent fixed effects
	Latin symbols denote random effects
Italics	Italicised words indicate definitions
Slanted	Slanted words represent variable names
b	Scalars- represented by lower-case letters
$\underline{b}$	An underlined lower-case letter denotes a vector
$\mathcal{X}$	Matrix- represented by upper-case letters
$\mathbf{b}, \underline{\mathbf{b}}, \mathcal{X}$	Boldface font indicates random variables, vectors and matrices (or command-line instructions)
$\underline{\mathbf{b}}^T$	Superscript 'T' denotes the transpose of a column vector or matrix
$E(\cdot)$	Mathematical expectation, expected value or mean of a random variable, vector or matrix
$var(\cdot)$	Variance of a random variable
$cov(\cdot, \cdot)$	Covariance of two random variables
$cov(\cdot)$	Covariance matrix of two random vectors
$\underline{1}_r$	$(r \times 1)$ Vector of 1's,
$\underline{0}_r$	$(r \times 1)$ Vector of 0's,
$\mathcal{I}_{\mathbf{n_i}}$	$(n_i \times n_i)$ Identity matrix
$\mathcal{J}_{\mathbf{n_i}}$	$(n_i \times n_i)$ Unity matrix (matrix of 1's)
$\mathcal{O}_{\mathbf{n_i}}$	$(n_i \times n_i)$ Matrix of 0's,
φ_{ijk}	Kinship coefficient between individuals j and k in family i
$\pi_{ijk}(a)$	Probability that individuals j and k in family i share a alleles at a single marker locus
$\sim$	Indicates the distribution of the random variable to the left of it
$\mathcal{N}(\mu, \sigma^2)$	A (univariate) normal distribution with mean μ and variance σ^2
$\mathcal{N}_r(\mathcal{X}\underline{\beta}, \mathbf{\Omega})$	An r-variate normal distribution with mean vector $\mathcal{X}\underline{\beta}$ and covariance matrix $\mathbf{\Omega}$
$F(t-1, r-t)$	Indicates the F-distribution with (t-1) and (r-t) degrees of freedom.

List of Acronyms

(In alphabetical order)

Acronym	**Meaning**
ANOVA	Analysis of variance
DNA	Deoxyribonucleic acid
HCM	Hypertrophic cardiomyopathy
HE	Haseman-Elston
HWE	Hardy-Weinberg equilibrium
HWL	Hardy-Weinberg law
IBD	Identical by descent
IBS	Identical by state
ICC	Intraclass correlation coefficient
LR	Likelihood ratio
lods	LR test calculated using $log_{10}(\cdot)$ rather than the natural logarithm
MLE	Maximum likelihood estimate/estimation
MME	Method of moment estimate/estimation
MS	Mean square
MS_b	Between-group mean square
MS_e	Mean square error
MS_t	Mean square treatment
QTL	Quantitative trait loci
REML	Restricted maximum likelihood estimation
SD	Standard deviation
SNP	Single nucleotide polymorphism
SS_b	Between-group sum of squares
SS_{error}	Error/within-group sum of squares
SS_{total}	Total sum of squares
$SS_{treatment}$	Treatment sum of squares
URL	Uniform resource locator

1 Introduction and Objectives

> STATISTICAL MODELS: A statistical model is a formal representation of the way in which data are thought to arise, and the features of the model dictate how questions of interest may be stated unambiguously and how the data should be manipulated and interpreted to address the questions. Different models embody different assumptions about how the data arise; thus, the extent to which valid conclusions may be drawn from a particular model rests on how relevant its assumptions are to the situation at hand ... Formally, a statistical model uses *probability distributions* to describe the mechanism believed to generate the data. That is, responses are represented by a [sic] *random variables* whose probability distributions are used to describe the chances that a response takes on different values. How responses arise may involve many factors; thus, how one "builds" a statistical model and decides which probability distributions are relevant requires careful consideration of the features of the situation (Davidian, 2005:13).

Genetic studies are generally carried out to isolate and identify the genetic factor(s) responsible for the trait under investigation. These traits are usually a disease or some characteristic which indicates disease severity, such as blood pressure or weight. Due to recent technological advances, genetic sequencing has become more viable and, as a result of this, there is also an increasing need for advanced statistical techniques which can be used to analyse genetic data.

It has been found that, genetically, two random unrelated individuals are 99.9% identical. The 0.1% difference observed is responsible for the variation between individuals. As with any statistical analysis, studies undertaking the analysis of genetic data need to account for the sources of variation observed in the data. This is just one of the factors to consider when building an appropriate statistical model for genetic data. In the area of genetic research, there are two broad types of studies which are carried out on humans: population-based studies where unrelated individuals are recruited, and family-

based studies where relatives are recruited. The latter is the focus of our study.

Family-based genetic studies are different to population-based studies as well as standard epidemiological studies. There are several reasons for this, which combine to make family-based studies both unique and interesting. These are:
1. considering the correlation between trait values on related individuals;
2. considering the correlation between genes (pieces of genetic material that may code for a biological function) that are close together on the genome; and
3. accounting for the fact that each individual has two independent genetic observations for each gene under observation– one of which comes from their mother and the other from their father– resulting in a correlation between the genetic material of related individuals.

The first point above draws attention to a fundamental difference between most research studies and studies in human family genetics. In the former, researchers assume subjects are randomly selected from a population. Therefore, observations on them are assumed to be independent and the corresponding statistical methods can be used. In family genetic studies, affected individuals are first selected, then their entire family is also recruited for the study. These individuals are called *probands*.

When statisticians use the word 'sample', they are referring to a group of observational units selected from the population of interest. However, when geneticists talk about a sample, they are referring to a piece of a person's genetic material. To avoid confusion, we will use 'study group' to refer to the group under observation. In addition, when referring to the way in which the study group was selected, we will use the word 'recruited', rather than the word 'sampled', to again avoid confusion. Geneticists generally refer to the recruitment of the study group as 'ascertainment'.

Families are recruited to investigate whether or not trait values for the individuals in the study are correlated. Correspondingly, our statistical analysis should account for the fact that family members are genetically related and thus expected to be similar for inherited traits, such as eye colour and height. However, they will also be similar for traits which

are due to sharing a common environment, such as radiation poisoning caused by the family home being near a faulty nuclear power plant. The first step in a family genetic study of quantitative (numerical) traits, is to distinguish between these two sources of correlation because the former is genetic and the latter, called *environmental*, is not. We can determine if traits are actually inherited or if familial similarity is due to shared environment by including in analyses the degree to which individuals are related to one another. If a trait is inherited, then the more closely related two people are, the more similar they are expected to be with respect to the trait. Thus the degree of correlation between family pairs will differ according to the degree of relation. Conversely, for a trait which is caused by environmental factors, all the family members are expected to be similar with respect to the trait, regardless of the degree of relation between them. Thus, all family pairs should have the same correlation for these trait values.

An obvious method for analysing this data is cluster analysis. However, we cannot use traditional cluster analysis for family-based genetic studies; we must modify it. Since there are different degrees of relation between different family pairs, cluster analysis does not correctly capture the statistical variation which exists in family genetic data. In addition, individuals that marry into a family are only linked to the family through their children. Therefore, including a single per-family random effect ("cluster") in a statistical model is not appropriate. In this study we will illustrate how to include the degree of relation between family pairs, using a matrix of relationships known as a *kinship* matrix, together with the per-family random effect.

Pedigree or family studies require information for as many family members as is feasible. It is not always possible to obtain genetic information for all family members, as they may be deceased or unreachable. However, their relationship to the rest of the family is useful for analysis because they inherit genetic material from their predecessors and transmit this same genetic material to their descendants. As a result, they are included in datasets even when no other information is available for them.

For the second point, we need to consider the phenomenon called *linkage*, which occurs when two genes are inherited together in such a way that one gene acts as though it is joined or linked to the other gene. This is investigated using genes called *markers*, which have known locations. There are many markers on the genome and many of them are non-functional genes. If two linked markers are both found to be associated with the trait of interest, then the causal gene (the gene responsible for the trait) is likely to either lie between these two markers or close to one of them. In this case, we say that the markers are informative since their locations narrow-down the area in which the causal gene can be sought.

The third point highlights a unique and important aspect of genetic data. For every part of a gene in an individual, there are two genetic observations, one from each parent. It is not possible to determine which part originates from which parent, unless parental genetic information is available and distinguishable. Also, each of the two pieces of genetic information is considered to be independent of the other piece, except if the parents are related. If the two pieces are different, then it is normally assumed that only one of them is correlated to the trait of interest and that the piece and trait value are inherited from the same parent. We are essentially searching for this piece, but we have to count each person and his trait twice. Thus, the pairs of genetic data inside each person are independent, but the people involved are not, especially if we are considering an extended family. In a statistical analysis, we somehow need to account for this phenomenon. In this study, we will illustrate how to do this through a family-specific random effect and a matrix depicting the genetic similarities between family pairs, for a particular marker.

These three points bring to light the unique aspects of family-based genetic studies, as well as their complexities. If an extended family, which is the subject of this study, is particularly large, then calculating a kinship matrix, for instance, becomes more computationally intensive. In addition, no standard statistical packages are equipped to carry out these calculations.

In terms of modeling extended pedigrees, the obvious choice is a linear mixed-effects model, which can be specified in most standard statistical packages. Some basic assumptions of such a model are that the random effect is the same for each family, the correlation between family pairs can be modeled through a known correlation structure and that family sizes are the same. However, all of these assumptions fail for the type of model we are considering here. Therefore, the linear mixed-effects model requires modification. Our solution is to use the matrices of genetic relationships to specify coefficients for the variances, thus creating potentially different covariances for each family pair. As a result, only the variance-components are actually estimated. Building up and explaining these models is the main focus of this thesis.

Another challenge we face is that the modifications suggested here are impossible to specify in most standard statistical packages. Fortunately, there are now several programs available which do allow the specifications we require.

1.1 Data

Individuals in a family study are categorised as *founders*- people who have no parents in the pedigree- such as the spouses of family members ('marry-ins') and the top generation in a pedigree, and *non-founders*-those people who have both parents in the pedigree. Founders are related to their children but it is usually assumed that their genetic information is independently obtained from the population.

Historically, genetic family studies were only carried out on sibling pairs, nuclear families, or family trios. Family trios consist of parents and one (usually affected) child. While the proband is always the main focus of interest, the parents are considered because they are the source of the child's genes. Nuclear families are families which consist of two generations of individuals: parents and at least one child. They are ideal for studies of major gene (single-gene) disorders, such as cystic fibrosis. However, they are less appropriate for complex (many-gene) diseases, such as cancer, heart disease, diabetes and

obesity. One reason is because disease susceptibility genes can be difficult to identify if, in different families, different genes have a significant effect on the same disease (Ellsworth and Manolio, 1999). For such studies, extended pedigrees are more informative. Another reason for nuclear families being inappropriate is that there may be many genes having undetectable effects, but together they cause a disease.

Due to recent developments in the area of human genetics, there is a need for more complex statistical methods because:
1. if the disease status of several generations of a family is known (as opposed to just two or three family members) and their genetic information is obtained, then researchers can pinpoint with greater accuracy which genes are linked to the trait of interest;
2. until recently, the analysis of quantitative genetic data was not computationally viable. However, with improvements in modern computers, this is now possible.

It is important to be able to analyse quantitative data because, for a study group of the same size (and all else being equal), statistical power is much greater for quantitative rather than dichotomous data. The advantage of using quantitative data was recognised in terms of genetic analysis many years ago by Douglas Falconer when he wrote:

> The genetic principles underlying the inheritance of metric [quantitative] characters are basically those of *population genetics* ... But since the segregation of the genes concerned cannot be followed individually, new methods of study are needed and new concepts have to be introduced. The branch of genetics concerned with metric characters is called *quantitative genetics* or *biometrical genetics*. The importance of this branch of genetics need hardly be stressed ... It is therefore in this branch that genetics has its most important application to practical problems and also its most direct bearing on evolutionary theory (Falconer, 1989:104).

This led to the subsequent development of statistical models for quantitative measures. One such model is the modified linear mixed-effects model mentioned earlier, which can

incorporate all the features of the data.

Programs which can be used to analyse family genetic data require the data in a specific format for analysis. This is to ensure the inclusion of the unique information required for a family study. For example, family data must show how each individual is related to every other individual in the family. One way to do this is to create, for each individual, a list of names of related individuals, as well as a description of their relationship. However, this is cumbersome. A simpler, more elegant method, lists the parents of each individual. All the other relationships can be inferred once the parents are identified: if they share parents, they are siblings; if their parents share parents, they are first cousins; and so on. This is the method used to set up the data for family studies. Table 1 is an example of the first few entries of a genetic dataset. It shows the information for two families; the first is an extended family while the second is a family trio.

Table 1: Example data

Family ID	***Person ID***	***Dad ID***	***Mom ID***	***Sex***	***Affection***	***Quantitative Trait***	***Genetic Marker***
F100	1	0	0	1	0	X	0 0
F100	2	0	0	2	2	X	2 1
F100	3	1	2	2	2	140.5	2 1
F100	4	1	2	2	1	122.8	2 2
F100	5	0	0	1	1	150.7	0 0
F100	6	5	4	2	1	145.0	2 1
F100	7	0	0	1	1	161.0	2 2
F100	8	7	3	1	1	158.9	2 2
F100	9	7	3	1	2	156.2	2 1
F100	10	7	3	2	1	152.4	2 2
F105	1	0	0	1	1	X	1 2
F105	2	0	0	2	2	145.5	2 2
F105	3	1	2	1	2	166.3	2 1

Datasets for genetic family studies contain, for each individual and at the very least, variables identifying: the ID of the family; the individual's ID within the family; the IDs of the individual's parents; the sex of the individual; information for the trait(s) of interest; and information for the gene(s) of interest. As such, the first four columns in Table 1 give the pedigree information. They identify all the individuals in each family

and give their relationships to each other. Column 1 gives the family ID, while column 2 gives the ID for each individual in the respective families. Columns 3 and 4 give the parental IDs for each individual. For example, in family F100, the parents of founder individuals 1 and 2 are unknown, as indicated by the zeros, but 1 and 2 are the respective father and mother of individuals 3 and 4. Similarly, in family F105, individual 3's father and mother are individuals 1 and 2, respectively.

Column 5 gives the gender of each individual, where 1=male and 2=female. These first five columns appear in every genetic pedigree dataset, in this exact order.

We can infer the relationship between other family pairs through the parental IDs. So for family F100, since individuals 3 and 4 share the same parents, they must be siblings. In the same way, we see that person 7 has no parents in the family, and must thus be a founder, specifically a married-in. This is substantiated when we move further down the table and find that individuals 7 and 3 are the parents of individuals 8, 9 and 10, thus implying that individual 7 is married to individual 3. This also implies that persons 1 and 2 are the grandparents of individuals 8, 9 and 10, and also of 6, who is the offspring of individuals 4 and 5, where 5 is also a married-in. In this way a family tree, known as a *pedigree*, can be built for each family in the study.

Column 6 of Table 1 gives the affection status of the individual, for the trait, where 1=unaffected, 2=affected and 0=unknown. This column is omitted if the trait of interest is not dichotomous. Column 7 gives the quantitative trait value (height (cm), say) for each person, where X denotes missing values. Finally, column 8 contains the genetic information. Each cell of this column consists of either a pair of numbers (for example, 2 1) or symbols (for example, a A), where 0=unknown. In datasets, these are separated by either a forward slash (2/1), a space (2 1) or no separator (21). One of the symbols represents the part of a gene that comes from the individual's mother, while the other comes from the individual's father. In practice it is usually not known which part comes from which parent as this is something that cannot be determined in the laboratory.

As some of the components shown in Table 1 are unique to genetic family studies, the methods involved in the analysis of such data have traditionally been developed by geneticists for their specific studies. As a result, it is difficult for statisticians without sufficient knowledge of genetics to understand the methodology. In addition, a unified and general theory of the statistical methodology involved here, does not appear to exist.

The aim of this study is to formulate, in unambiguous mathematical notation, a statistical model that has as special cases some of the models which are currently used to investigate the existence of linkage in human family genetics. The methodology will be written up for statisticians, under the assumption that they have little to no knowledge about genetics. As a result, we will begin our study by introducing some important and necessary genetic definitions and concepts (Chapter 2). This is because, as stated in Burton et al. (2005:941):

> ... knowledge about the underlying biology, coupled with the inferential tools of modern epidemiology and biostatistics, allows important aetiological questions to be answered in ways that are more rigorous, and often more powerful, than approaches that fail to make the best use of both the epidemiology and the genetics.

The next step will be to illustrate how to explore genetic data (Chapter 3). In Chapter 4, we introduce the statistical methodology that will be the focal point of this study, namely variance-components methods. In Chapters 5 and 6 we describe systematically and in statistical language, the methodology involved in a genetic family analysis. By systematic we mean that we will start by explaining methods for familial aggregation (Section 5.1), which is the simplest model. We will then consecutively add specific random effects, leading to models enabling inference on segregation analysis and broad-sense heritability (Section 5.3) and lastly, linkage analysis and narrow-sense heritability (Chapter 6). These are the first steps in any genetic study investigating the heritability of disease in humans.

Familial aggregation aims to determine if a trait runs in families by checking if trait

values are correlated inside families. If there is evidence of aggregation, we investigate further to see if it is due to inherited genes or to shared environment. Heritability allows us to determine what proportion of the trait variance is due to heritable genetic factors. Segregation analysis involves using family relatedness to fit specific genetic models to trait data. These models assist in determining whether correlations are higher if relatives are more closely related. For all the models, hypothesis tests are used to determine if the traits are in fact inherited.

Linkage analysis aims to determine if a particular trait in an individual is transmitted through families together with a specific gene. If there is evidence of this occurring, the location of this gene is sought. Linkage analysis tests a marker for linkage with a hypothetical causal gene. If linkage is found to exist, the conclusion is that the true causal gene is in linkage with, and hence physically close to, the marker. This thus reduces the area in which the true disease gene should be searched for.

Even though the approach to a family genetic analysis is presented in a systematic way here, this is not necessarily the way it is done in practice. For instance, if there is other evidence that familial inheritance exists, then the first step may be skipped. However, all the steps are presented here as they illustrate how standard statistical methods can seldom be used to analyse human genetic data. As a result, they also illustrate the limitations of standard statistical packages for such analyses.

The model that is developed will be explained and built up by considering the unique properties of genetic data, that were discussed previously. The models that are used to investigate linkage already exist and are those that underly, among others, the QTDT (Quantitative Transmission Disequilibrium Tests) software (Abecasis et al., 2000a, 2000b). However, these have not been presented in the literature using one set of notation, or in a language that is familiar to statisticians.

1.2 Practical example

To demonstrate the statistical methods used in practice and explained in the theoretical sections of this study, we will present the results of the analyses of data obtained from research carried out in the area of hereditary cardiovascular disease in extended pedigrees. The data, which has already been analysed and published (Revera et al., 2007, Revera et al., 2008, Van der Merwe et al., 2008 and Heradien et al., 2009), consists of information for 22 families obtained from research conducted in South Africa on hypertrophic cardiomyopathy (HCM). These families together contain 507 individuals, but genetic information is only available for at most 329 individuals, depending on the marker under observation. The remaining people are included because they provide information on the way in which those family members with available information, are related to one another. For example, if my grandparents are not included in our pedigree, it will not be apparent that my first cousins and I are related.

HCM is a frequently inherited cardiac muscle disease, characterised by thickening of the left ventricular wall of the heart. Hypertrophic means 'excessive thickening' while cardiomyopathy implies a disease of the heart muscles. It is an inherited disorder which is known to cause sudden death in young people (under the age of 35). It is caused by (known) mutations in the genes that encode the protein components of the cardiac sarcomere. These proteins are responsible for heart contraction. Ventricular thickening is highly variable and the variability is due to both genetic factors and non-genetic factors such as age and sex (Revera et al., 2007).

The quantitative traits used to illustrate the data analysis are both measures of ventricular thickness: *LVMecho* is left ventricular mass as measured by echocardiography; and *cwtscore* is a composite measure of ventricular thickness.

We will explore the data, then demonstrate familial aggregation and heritability, and investigate whether the markers are linked. This example will be referred to as the

Heartdata example from this point onward.

Analysis of the data will be carried out using statistical software packages which are designed specifically for analysing genetic data. The data analysis and results are discussed in Chapter 7. Before we can get to that however, we need to understand the context of this study and the type of data we are interested in. To begin, Chapter 2 introduces the genetic background necessary to this study.

2 Genetics: Introduction and Terminology

2.1 Molecular genetics

Human genetic material is composed of deoxyribonucleic acid (DNA) which is divided into chromosomes. Each chromosome consists of two strands which are joined together by the base pairs A/T or C/G, and twisted to form a double-helix structure. The human genome consists of 23 pairs of chromosomes: 22 autosomal pairs and 1 sex-determining pair. Of these 23, one member of each pair is inherited from the mother and one from the father. This bi-parental inheritance is the most important property of the genetic data statisticians must analyse. The members of each chromosomal pair (except the sex-determining chromosomes in males) are known as *homologous chromosomes* because they contain identical gene locations, called gene loci (sing. locus) along their lengths. The loci have identical genetic potential, implying that the gene pairs influence the same characteristics in the individual. However, the genetic sequence (sequence of base pairs) in each chromosomal pair need not be identical, since each locus is defined by the particular sequence of bases found there (Burton et al., 2005). When the DNA sequence at a particular gene locus varies between the chromosomes of different individuals in the population, each version of the sequence is known as an *allele*. Sometimes alleles are not a sequence but have only two possible forms, for example the bases G or T, as seen with single nucleotide polymorphisms (SNPs). Whatever the form, alleles are the hereditary unit factors responsible for transmitting genetic information from one generation to another.

Humans are diallelic, meaning that we have two alleles at each genetic locus, one from our mothers and one from our fathers. The transmission of genetic information from parents to children occurs during *meiosis*. It is the process of cell division which occurs during reproduction, after fertilisation has occurred. During meiosis, the developing child's *gametes* (sperm and ova) form. The cells divide and replicate, and the DNA partitions to create the gametes. Since the gametes are involved in sexual reproduction through fertilization, they contain only one set of chromosomes. When the child grows up and in turn

mates, the one set of chromosomes will then be passed on to the offspring at conception. The offspring's second set of chromosomes will come from the gamete of the other parent.

Alleles that occur frequently in the population are called *wild-type* or *normal* alleles, while those that contain modified genetic information are known as *disease susceptibility* or *mutant* alleles. Mutation occurs during meiosis. In practice, it is not always clear which alleles are wild-type and which are mutant.

2.2 Mendelian (or transmission) genetics

In the mid-1800s, after experimenting with hybridization in the pea plant, Gregor Mendel set down several hypotheses/postulates which formed the cornerstones of studies on genetic inheritance (Klug & Cummings, 2000):

1. Genetic characteristics in individuals are controlled by (paired) unit factors (alleles).
2. The paired alleles (One maternal and one paternal) *segregate* or separate randomly during meiosis, such that each gamete receives one allele or the other with equal probability (Known as Mendel's First Law: Segregation).
3. This law was later disproved for alleles that are linked. When gametes form, segregating pairs of alleles assort independently of each other, leading to extensive genetic variation (Known as Mendel's Second Law: Independent assortment).

The physical expression of some clinical outcome (usually a disease) or inherited characteristic, such as eye colour, blood pressure or height, is known as the individual's *phenotype*. The word 'phenotype' is interchangeable with the word 'trait' and the latter is the one we will continue to use in this study. An individual's genetic status at a single gene locus, often represented by paired alleles (for example $1/2, 1/1, 2/2$), is called his *genotype*. If a locus contains identical alleles, for example $1/1$ or $2/2$, the individual is said to be *homozygous* at that locus, while he is *heterozygous* at that locus for the $1/2$ or $2/1$ alleles. An individual with heterozygous alleles is called a *heterozygote* while one with homozygous alleles is known as a *homozygote*.

Dominance and recessiveness describe the relationships between traits and alleles and thus aid in determining an individual's trait from his genotype. For a dichotomous trait at a diallelic locus, allele 1 is completely dominant with respect to disease susceptibility if an individual needs only one copy of this allele to be affected. On the other hand, allele 1 would be recessive for the same disease if an individual needed two copies to be affected (Burton et al., 2005). Therefore, dominance and recessiveness are two sides of the same coin. For example, suppose that for some disease caused by a diallelic locus, allele 1 is the dominant allele. The penetrance function of the disease, $Pr(disease|G, E)$, is defined as the conditional probability of observing the disease of interest, given the individual's genotype (G) and considering environmental factors that could influence the risk of him having the disease (E). It is a density function if the trait is quantitative and a mass function when the trait is discrete. Although complete penetrance is rare, sickle-cell anemia is an example of a disease where the condition is caused only by genetic factors (Elston, 2004).

For a dichotomous trait (disease/no disease), in the dominance case, the penetrance function shows that an individual with one copy of the disease susceptibility allele, 1, has the same risk of disease as someone with two copies, i.e. $P(disease|1/1) = P(disease|1/2)$. Conversely, if allele 1 is recessive with respect to a disease, then an individual with only one copy of this disease susceptibility allele has the same risk as an individual with no copies of it, i.e. $P(disease|1/2) = P(disease|2/2)$ (Terwilliger, 2005). Therefore, if allele 1 is completely dominant for a disease, then allele 2 is recessive protective against the same disease.

Finally, *codominance* occurs when each of the three genotypes have different effects on the trait, i.e. $P(disease|1/1) \neq P(disease|1/2) \neq P(disease|2/2)$.

Figures 1 and 2 describe the concepts of dominance and recessiveness visually.
Let us first consider the dichotomous trait in the two plots of Figure 1. Here, allele 1 is first completely dominant (left plot) with respect to disease susceptibility, then completely

recessive (right plot). Thus, for a disease that is inherited dominantly, we see that having at least one copy of allele 1 is sufficient to cause illness. However, if the disease is inherited recessively, then two copies of allele 1 are required to cause the disease. In this case, allele 2 is recessive-protective against the disease since one copy of allele 2 will 'protect' against the disease. In other words, in the latter case, allele 2 is dominant with respect to disease protection. However, these examples are of an extreme case. In reality, the situation is not as extreme and complete dominance rarely occurs. Therefore, disease susceptibility usually lies somewhere between these two extremes.

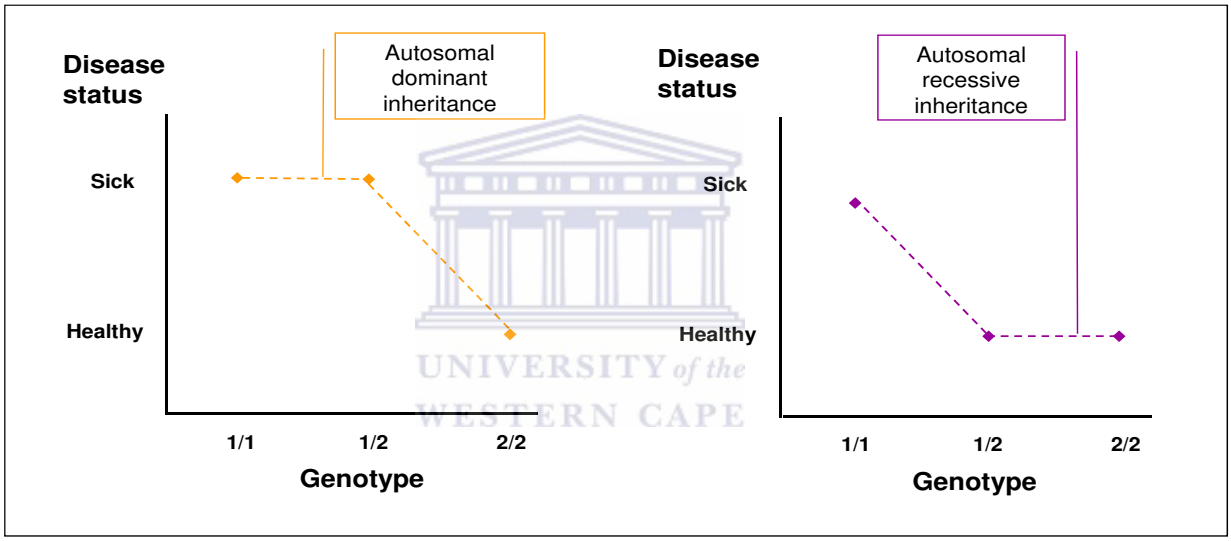


Figure 1: Plot illustrating dominance and recessiveness in a dichotomous trait, where allele 1 is either dominant or recessive with respect to disease susceptibility

For a quantitative trait such as height, shown in Figure 2, we have a similar situation to the dichotomous case in Figure 1. The difference is that now, the trait values have density functions where the mean value of the trait will depend on the type of inheritance. Therefore, if allele 1 is dominant for tall stature and height is inherited dominantly (top left plot), then one copy of allele 1 will be sufficient to result in a person being taller than the average. If height is inherited recessively (top right plot), one copy of allele 2 will result in a person who is shorter than average. So here allele 2 is dominant with respect to short stature (or recessive-protective against tall stature).

For quantitative traits, we are interested in a special type of codominance called *additive* inheritance. It occurs when the expected trait value in a heterozygous allele pair lies exactly half-way between the expected trait values of the two corresponding homozygous pairs. In Figure 2, additive inheritance is depicted in the bottom plot. Here, the average height of someone who is heterozygous lies between the heights of the homozygotes. Thus, if allele 1 is dominant with respect to tall stature, the 1/1 genotype will have a height distribution with a higher mean than the 1/2 and 2/2 genotypes.

The bars going through each point in the graph indicate the spread about each of those means (the density functions).

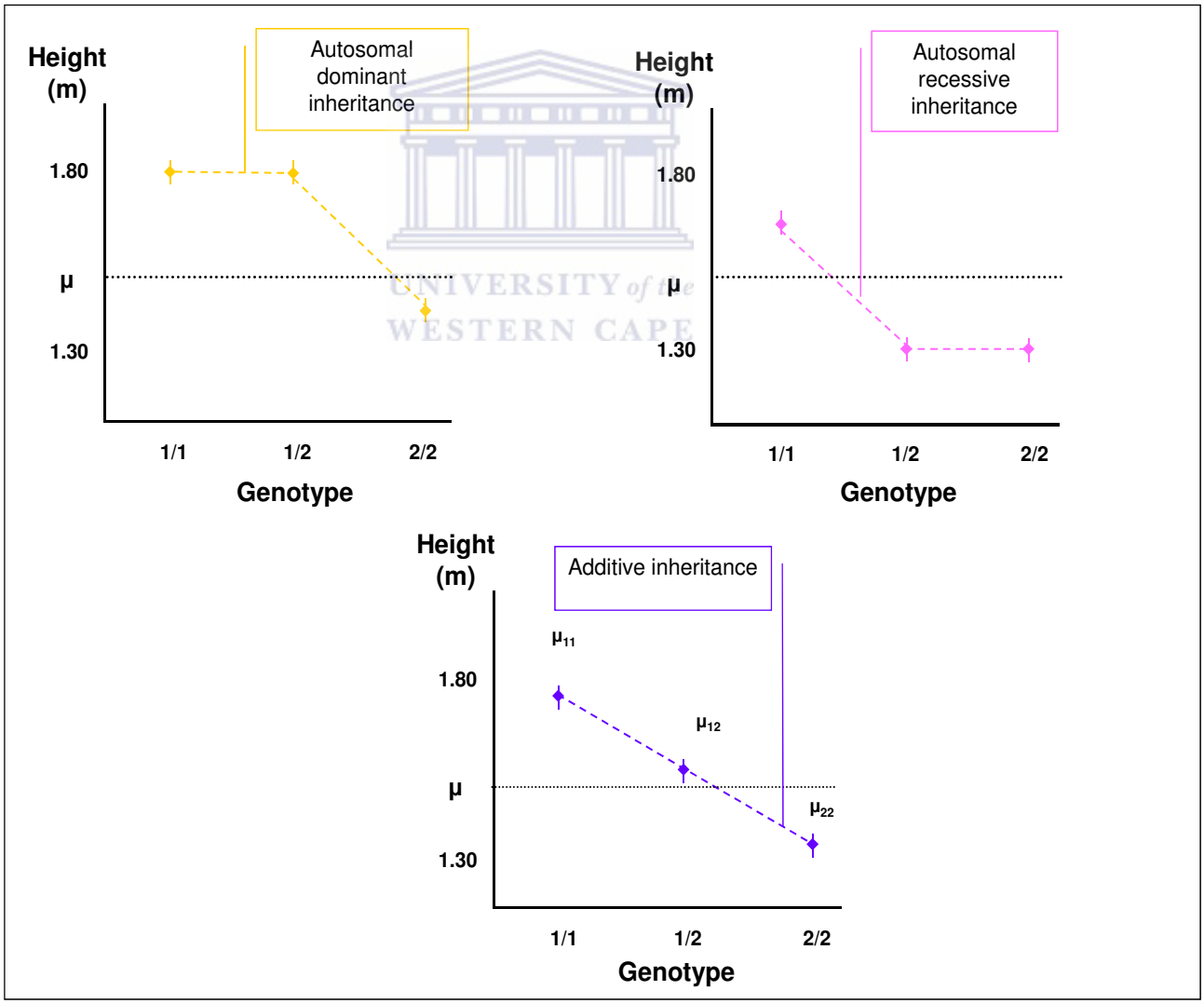


Figure 2: Plot illustrating dominance and recessiveness in a quantitative trait, where allele 1 is respectively dominant, recessive or additive with respect to height

Figures 1 and 2 are two examples of graphs which can be produced to understand genetic data. However, as with any study, all the data needs to be explored and understood before any analysis can be carried out. We look at data exploration in more detail in the next chapter.

3 Exploratory Data Analysis

When analysing any data, the first step is to explore and understand the data so that any possible errors can be found and the appropriate statistical analyses, if any, may be identified and carried out. A data exploration starts by plotting the data to understand it, identify the distributions of the various factors under consideration and detect anomalous values. Histograms and box-and-whisker plots are produced for exploring quantitative data while categorical and ordinal data are graphed with either bar charts or pie charts. Pairwise plotting, such as scatterplots for quantitative data, allows the statistician to identify patterns or trends, thus identifying possible associations between pairs of variables, if they exist. Thereafter, appropriate summary statistics are calculated.

Data exploration helps us to understand the data by telling us what it looks like, what the distribution of various variables may be, whether there are any anomalous observations, whether or not the data needs transformation, and it gives us an idea of how to proceed with the analysis. However, genetic studies require additional, more specialised exploring due to the nature of the information. In particular, family studies involve studying the inheritance patterns of individuals in families, given their traits, genotypes and their relationships to each other. As mentioned before, family members are related and are thus similar with respect to inherited traits, so understanding their relationship is vital to analysing family data. The families used in these studies can be (simple) nuclear families or (complex) multi-generational extended families. Exploring family data, regardless of the family-size, begins with a pedigree such as the one in Figure 3, which depicts a nuclear family. In a pedigree, circles represent females and squares represent males.

At the top of the pedigree are the parents, who are called Dad and Mom. The line joining them represents a mating, which results in two daughters, Sue and Jane. The shaded circle representing Mom implies that she is affected with the trait of interest. Since the circle representing Sue is also shaded, Sue is also affected. Thus, it is possible that Mom passed on the causal allele(s) to Sue. Since Dad and Jane are represented by unshaded

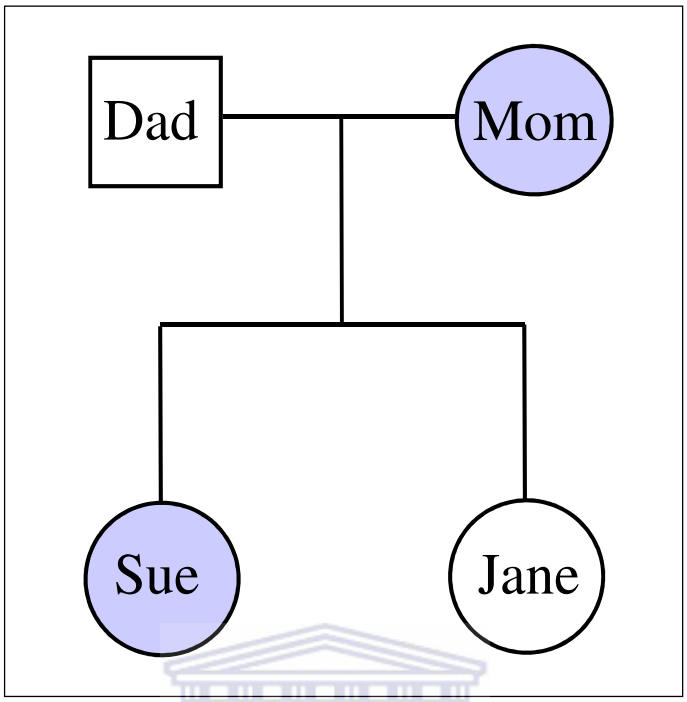


Figure 3: Pedigree of nuclear family

symbols, they are either not affected or their affection status is unknown. This family is one that shows signs of a trait that may be passed on from parent to child, so further inquiry is justified. If we had information for the extended family, there could be more clarity regarding the possible inheritance of the trait. For example, if we had information for the great-grandparents and their great-grandchildren, then we could see if cousins far removed share the trait, and thus possibly the causal alleles. Pedigree diagrams such as the one above can be produced and used to visually assess data for extended families.

Trait vs. genotype plots, such as those shown in Figures 1 and 2 should be made to explore any possible inheritance patterns in the trait under investigation; is it dominant, recessive or additive? In practice though, inheritance patterns are rarely simple.

In Figure 4, the plot on the left summarises the number of members in the various Heartdata families. There is one very large family, with over 90 members, as indicated by the bar on the extreme right-hand side of the plot. The plot on the right illustrates the number of generations in the families. It shows that the Heartdata contains at least two

generations in each family, and at most five. There are three nuclear families, which have two generations, and most of the remaining families have either three or four generations of members.

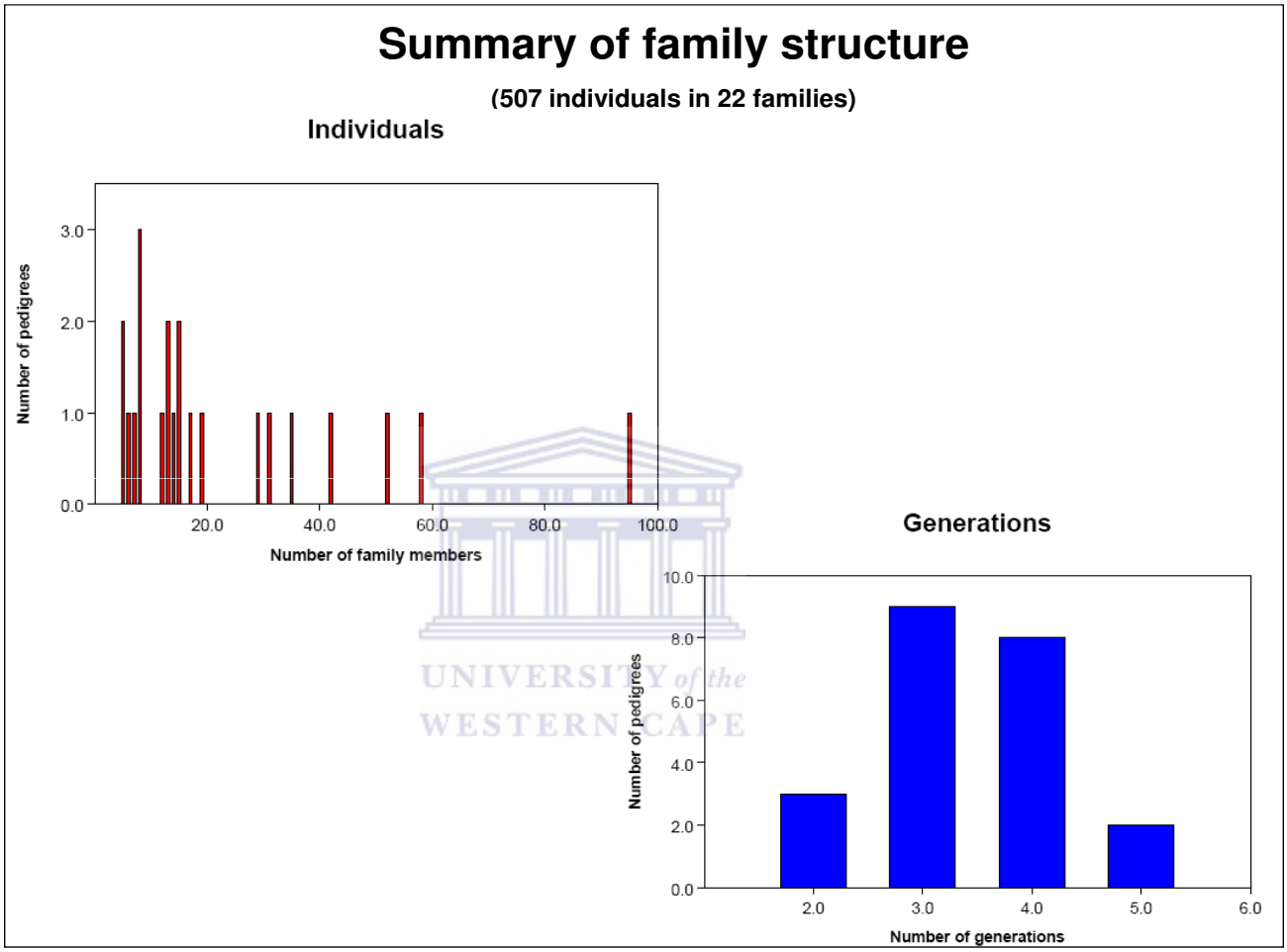


Figure 4: Summary of family structure

As a result of considering extended families with information for several generations, it is vital to assess the family and genotype data for pedigree and genotype errors. Checking for pedigree error involves validating that the relationships between the family members have been correctly captured. For example, a pedigree error would occur if a child was recorded as the parent of his father.

Assessing genotype errors involves ensuring that the individual and family genotypes are correctly typed and adhere to Mendelian inheritance. For example, a genotype error in a

family would occur when a child has an allele that comes from neither parent. Inconsistent Mendelian inheritance, and thus genotype errors, are easier to detect in large pedigrees than in affected sibling pair studies where no other family data is available (Teare and Barrett, 2005). For example, say a child shares no alleles in common with his father, but shares an allele in common with his brother and paternal grandmother. Without the grandmother's genotype, it would be impossible to determine that it is the father's genotype that was incorrectly captured in the dataset. We might then have wrongly concluded that the father is not the child's father, when in reality he is.

One way of checking for genotype errors is using the Hardy-Weinberg Law (HWL), which tests the statistical independence of two alleles at a locus. It is based on binomial probabilities, assuming there is random mating and no mutation, migration or natural selection. Random mating here implies that the probability that two individuals mate is independent of their genotype and ethnic group (Thomas, 2004). Under these assumptions, consider alleles 1 and 2 at a particular diallelic locus, such that they occur in the proportions p and $(1-p) = q$ respectively in the population. Then, assuming random mating, the genotype frequencies for genotypes 1/1, 1/2 and 2/2 occur in the proportions $p^2, 2pq$ and q^2 respectively. This is Hardy-Weinberg Equilibrium (HWE) and it occurs after just a single generation of random mating, regardless of the initial genotype frequencies. Therefore, according to HWE, genotype and allele frequencies in a large, randomly-mating population remain stable over generations.

In population genetics, the HWL is a fundamental law describing the relationship between allele and genotype frequencies in a randomly-mating population. It is used to check the quality of the data and for testing genetic associations.

Deviations from HWE may have several causes, among which are non-random mating and genotyping errors, as discussed in Wiggington et al. (2005) and Foulkes (2009). In their paper, Wiggington et al. state that the former generally leads to false homozygosity, which implies a shortage of heterozygous allele pairs in the population. The latter

however, leads to an excess of heterozygotes, and this can be tested for if genotyping error is suspected. Therefore, testing for compatibility with HWE is a routine check for genotyping errors. Nevertheless, compatibility with HWE does not necessarily imply that there are no genotype errors.

Since HWE is defined for a randomly-mating population, testing it on related individuals would be invalid. In pedigree studies, even though observations are not independent, HWE can still be used as a data-quality check. It is tested on a selected set of unrelated individuals from each of the families, as was done above. This requires identifying and then selecting all the members of each family who are genotyped but are not related to the parents in that family. This usually refers to those people who have married into the family. In addition, the chosen individuals must be genotyped, implying that there are usually many more founders than there are selected unrelated individuals.

As this process is tedious and time-consuming in practice, particularly if the family is very large, specialised software is used to test HWE.

For the Heartdata, the HWE plot in Figure 5 is produced for *Marker 11*, for the unrelated individuals in the data.

In Figure 5, the graph on the left shows the distribution of the genotypes in the study group. The values in each cell indicate the following observed genotype counts:
For the 1/1 genotype, the observed genotype count is 12.
For the 1/3 genotype, the observed genotype count is 15.
For the 3/3 genotype, the observed genotype count is 7.
Therefore, there are $7 + 12 = 19$ homozygotes and 15 heterozygotes. The shading of each cell indicates the expected genotype counts under HWE, such that the darker blocks indicate genotypes with high expected counts, while the lighter blocks indicate genotypes with lower expected counts. Here, allele 3 is the minor allele as the 3/3 genotype has a lower count than the 1/1 genotype.

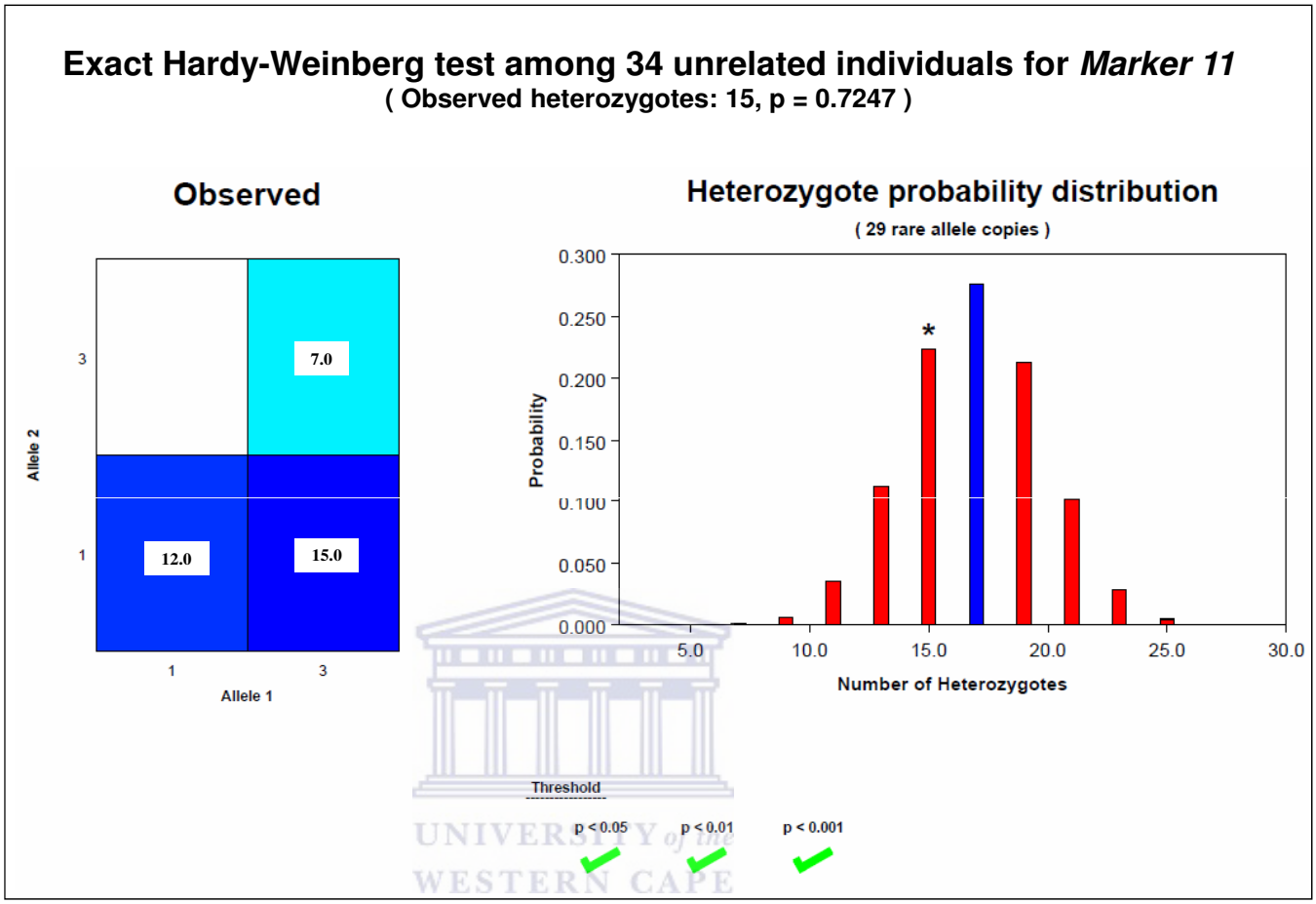


Figure 5: Hardy-Weinberg test on unrelated individuals, for *Marker 11*

The graph on the right shows the exact heterozygote probability distribution derived by Wigginton et al. (2005), for the number of heterozygotes, conditional on the number of rare allele (minor allele) copies. As indicated in the text above the graph, there are 29 rare allele copies: 15 from the heterozygote and 2x7 = 14 from the seven 3/3 homozygotes. The null hypothesis tested is that of HWE. The exact test is analogous to Fisher's exact test for contingency tables. If the number of observed heterozygotes (indicated by a star in the probability distribution graph) falls too far to the left or right on the graph, then the null hypothesis is rejected as this indicates, respectively, either a deficit or excess of heterozygotes. On the graph, the bars in red fall in the area of the distribution which is used to calculate the p-value. Wigginton et al. (2005) and Foulkes (2009) recommend using the exact test, as it is better than the commonly used chi-squared goodness-of-fit

test which is based on asymptotic theory. The chi-squared test can give very large or small p-values and is sometimes extremely anti-conservative, meaning that the type I error rate can exceed the nominal significance level. This is because the expected frequency of rare alleles is not usually high enough for the chi-squared test to be appropriate. On the other hand, the exact test never exceeds the nominal significance level, even when there are only a small number of minor alleles in the sample.

In Figure 5, there are three significant levels at which HWE is tested: 0.05, 0.01 and 0.001. The results are indicated underneath the graph on the right. Since the p-value for the test is 0.7274, as indicated in the sub-heading of the figure, there is sufficient evidence for HWE at all three significant levels. This is indicated by the green tick symbol. If there is no HWE, the green tick is replaced by a red cross.

Another component of exploring family data involves carrying out pairwise age checks on all the relative pairs in each of the families. This ensures that the pedigree makes sense and that the information is entered correctly into the dataset. For example, we can identify discrepancies in the ages of family members if, say, a grandmother is younger than her grandchild or a mother is only 5 years older than one of her children.

For the Heartdata, the scatterplot in Figure 6 shows the age relationship for parent-child pairs. The circled points identify cases where parent-child age anomalies occur. When the age check was carried out, the parents and children causing the anomaly were identified. For the points circled in red, individual 3 in family F100 was identified as being 61.0 years old. This person had 4 children, aged 57.0 (individual 5), 55.0 (individual 6) and 51.0 (individuals 7 and 8), which is clearly impossible.

The age discrepancies found here were reported to the researchers and it was explained that these were not in fact the ages of the individuals in the dataset, but rather the age at which their hearts were assessed, explaining the apparent anomaly. Had an age check not been done, we would have assumed the ages in the dataset were in fact the person-ages rather than age at assessment and this would have affected the interpretation

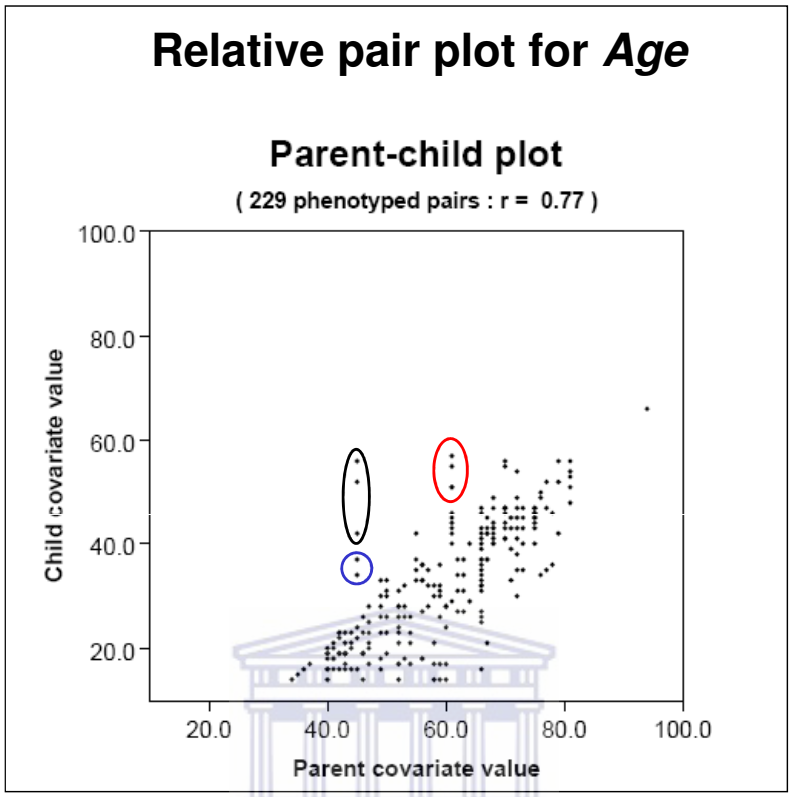


Figure 6: Relative pair plot for *Age*

of any analysis involving age. This example illustrates the importance of data checking and follow-up of discrepancies.

As part of the exploratory analysis, it is also important to report summary information for any dichotomous or quantitative traits, and for covariate data. These reports include information for correlations between sibling and other relative pairs. Various graphs can again be produced to help identify patterns in the data. For quantitative data, histograms or box-and-whisker plots help identify outliers and influential observations. In addition to this, graphs allow the visual detection of departures from normality. This is an important consideration when analysing quantitative data as transformations may be necessary before any statistical analysis will be valid.

Let us consider the quantitative trait, *cwtscore*, from the Heartdata. Figure 7 shows a histogram of this trait.

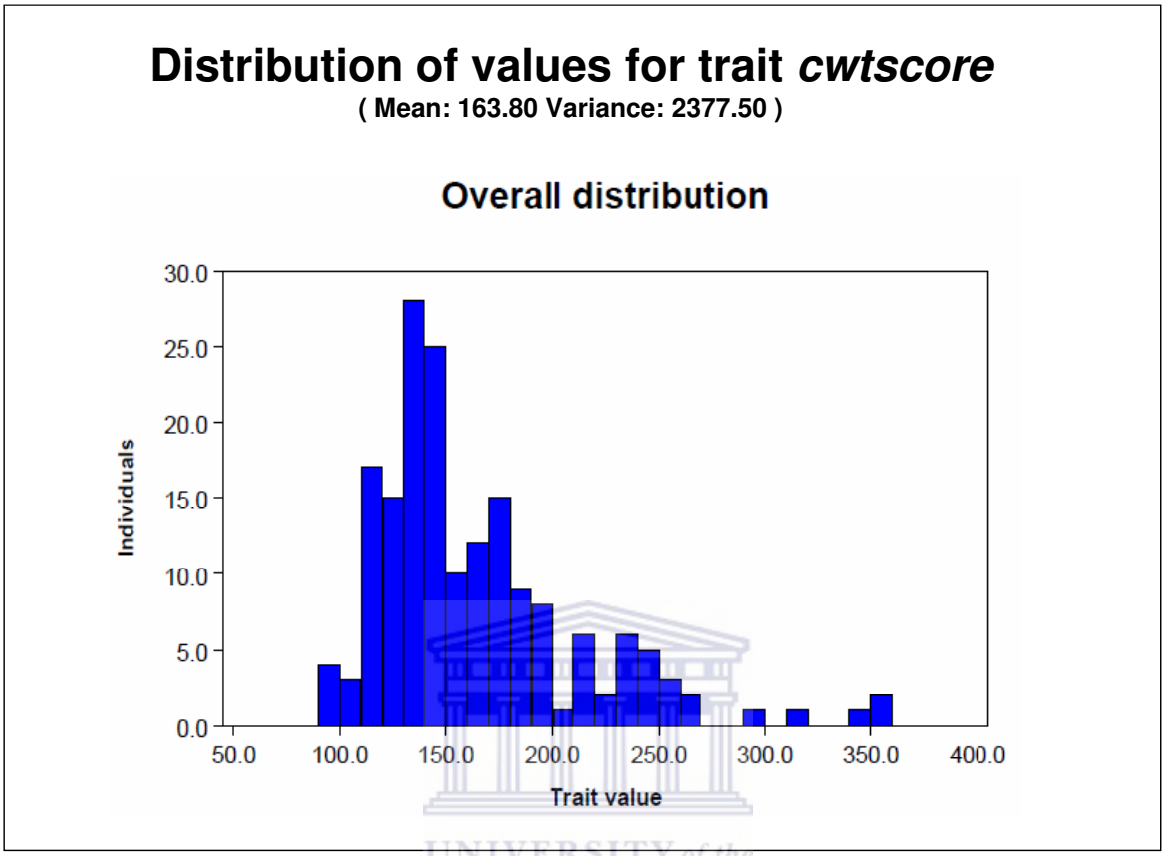


Figure 7: Summary plot for quantitative trait *cwtscore*

From Figure 7 we see that the distribution of *cwtscore* is skewed to the right, with mean 163.80 and variance 2377.50. Before any analysis is carried out, this data may need to be transformed to achieve approximate normality. For the Heartdata traits, we used a method of transformation called *quantile normalisation*; it was chosen as the most appropriate transformation in the original analysis of the data. It is a method which makes the distribution of a variable as close to normal as possible, without changing the order of the observations of that variable. It works as follows: firstly, all the observations are ranked; then 0.5 is subtracted from each rank and the answer is divided by the number of observations. This value is considered as a standard normal cumulative distribution value, and the corresponding standard normal quantile is found.

For *cwtscore*, the quantile normalised data is called *Qcwtscore* and its distribution is plotted in Figure 8.

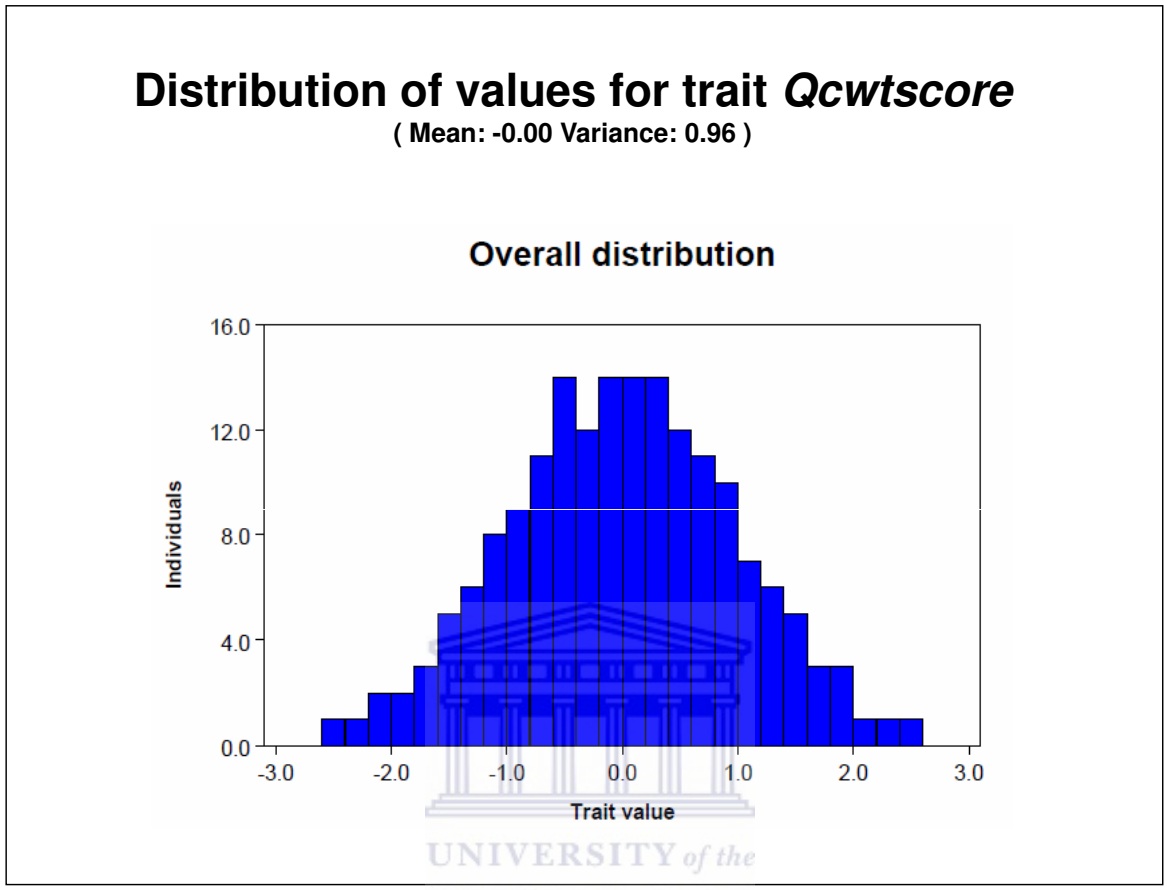


Figure 8: Summary plot for quantitative trait *Qcwtscore*

We see that *Qcwtscore* looks much more symmetrically distributed than *cwtscore*. As expected, the mean is now zero and the variance is 1.

As with any statistical analysis, it is important that the results are carefully and correctly interpreted by reporting estimated effect sizes in practically understandable units of measurement. In the case of quantile normalisation, the effect sizes cannot be transformed back into the original unit of measurement. As a result, they cannot be interpreted in terms of specific estimated effect sizes. To produce effect sizes which can be interpreted, the untranformed data is also analysed and the results are used solely for interpretational purposes.

As mentioned before, if a trait is shared by individuals because it is inherited, then the correlation between pairs of these individuals will get stronger the more closely related

they are. If however, the trait is shared because the individuals share a common environment, then the correlation between all sharing pairs will be the same. Finally, if the trait is independently distributed, as with random strangers, then no correlation is expected.

Bearing this in mind, quantitative traits should be explored by producing scatterplots of the trait values for family pairs, such as siblings or parent-child pairs. These plots could illustrate patterns in trait values, which are due to family relatedness. For example, for an inherited trait such as height, we expect the scatterplot for sibling pairs to show a high correlation, while the same plot for two random strangers should show a completely random scatter because unrelated individuals are not expected to be genetically similar. As such, observations on them should be uncorrelated. For family members that are less closely related, for example uncle-nephew or cousin-cousin pairs, some correlation is expected, but not as much as that for sibling pairs.

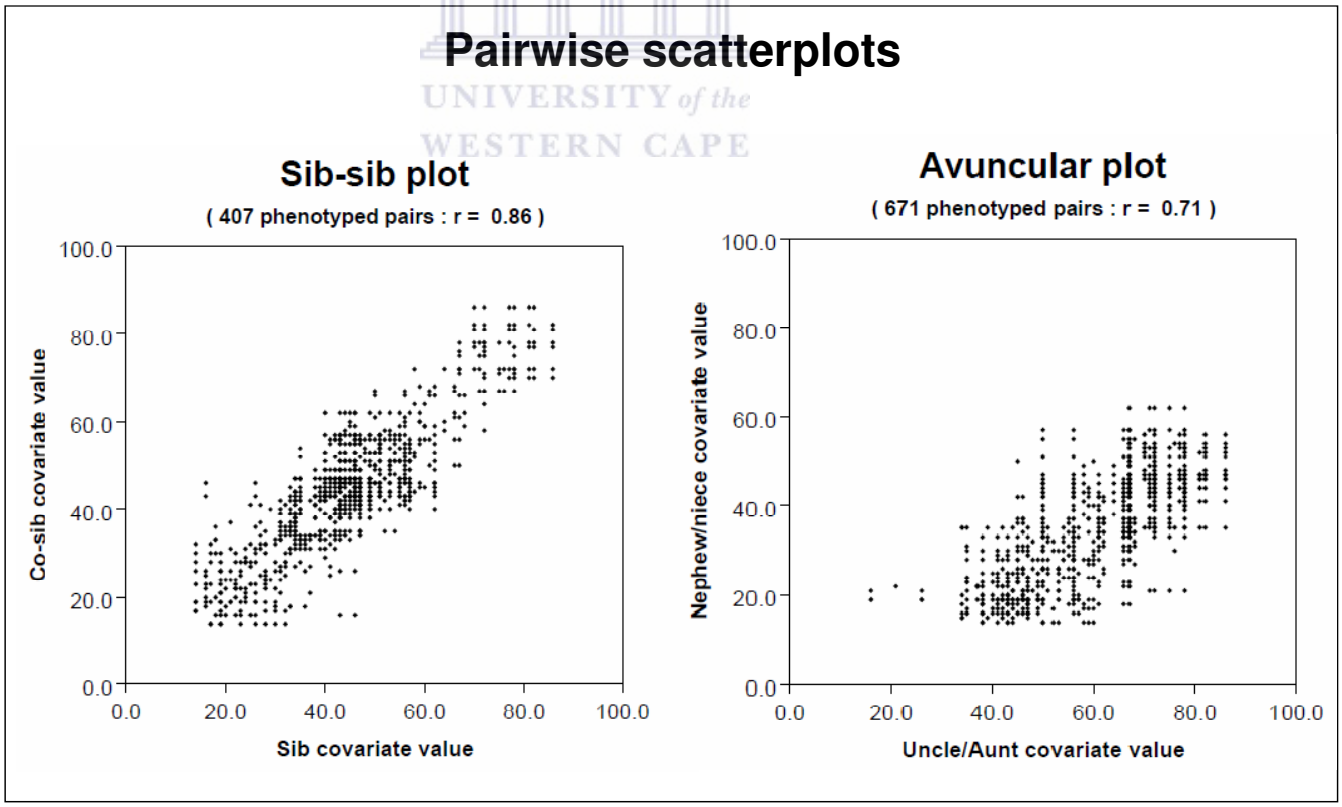


Figure 9: Pairwise scatterplots

Figure 9 illustrates scatterplots for two different family pairs. The graph on the left is

the sib-sib plot for a trait and shows the distribution of the trait among sibling pairs. It indicates that a good positive correlation exists between the sibling trait values, as would be expected for a trait clustering in families. The plot on the right shows the distribution of the same trait for avuncular-pairs, which are pairs made up of an uncle/aunt with a niece/nephew. We see here that the scatter is more random, and thus less correlated, than that for the siblings. This indicates an inherited trait, as avuncular pairs are third degree relatives while siblings are first degree relatives and are thus much more similar genetically.

The histogram in Figure 10 shows the distribution of the quantitative covariate Systolic blood pressure. Here, we again see that the distribution is slightly skewed to the right

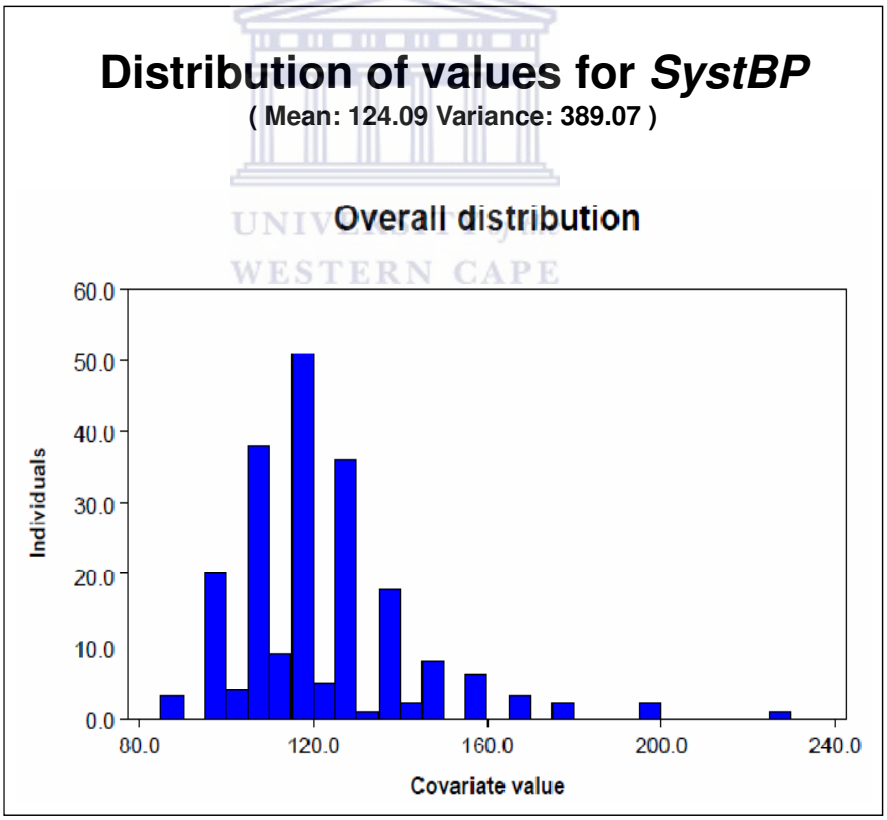


Figure 10: Summary plot for systolic blood pressure

and thus this data may also need to be transformed to achieve approximate normality, before any analysis is carried out.

It is not always the case that a set of data will only contain quantitative covariates. Often, there are also dichotomous covariates, such as ethnicity or gender, as in the Heartdata. One of the other dichotomous variables found in the Heartdata is called *Mutation*. It indicates whether or not an individual is affected with the mutation which causes HCM, where 1 =No and 2 =Yes.

Finally, the genetic marker data can be plotted and explored, as shown for *Marker 14*, in Figure 11. The distribution plot on the left shows that, for this marker, 295 out of 507 individuals are genotyped and allele 2 occurs more frequently than allele 1 (60% vs. 40% respectively).

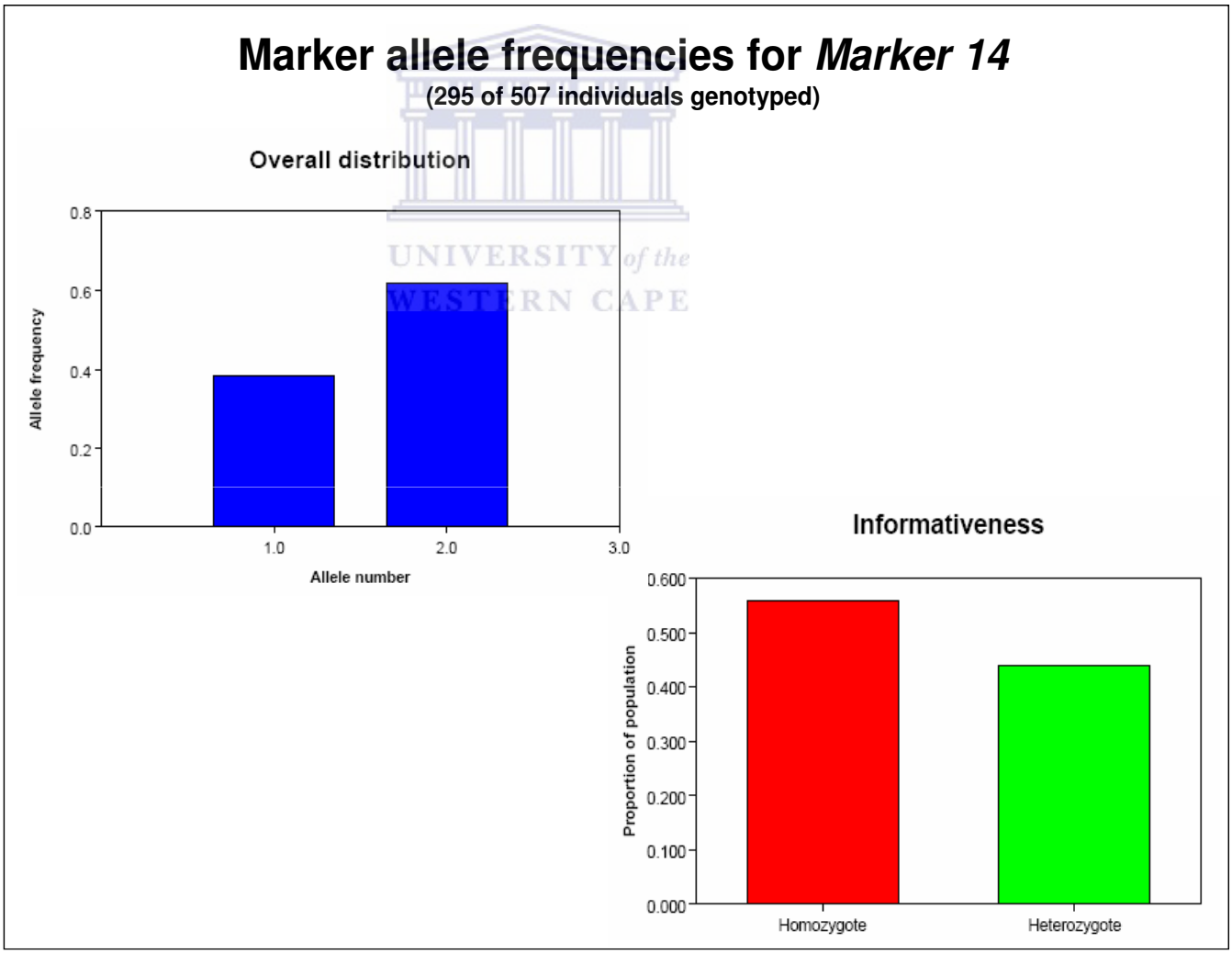


Figure 11: Marker allele frequency for *Marker 14*

The plot of informativeness shows the proportion of genotyped individuals that are homozygous and heterozygous. Marker *informativity* is determined by the proportion of heterozygous individuals in a population and is important because it determines how useful a genetic marker is. For homozygous parents, it cannot be determined which alleles are transmitted to offspring. Therefore, those meioses are not informative for linkage as it is unknown whether or not recombination occurs. This will be explained further in Chapter 6. For *Marker 14*, about 44% of the genotyped individuals are heterozygous and thus informative. The remainder are not informative as they are homozygous.

In the next section, we present linear mixed-effects methodology. This is the focus of our study since mixed-effects models form the basis for one way in which family genetic data can be modeled.

4 Linear Mixed-effects Models

> The notion that haphazard variation may arise from a number of sources and that it may be valuable to identify these sources and measure their impact has a long history and many applications and implications. Indeed, it is only in very simple situations that it is likely to be satisfactory to represent haphazard variation by independent identically distributed random variables or by the essentially equivalent notion of random sampling from a hypothetical infinite population (Cox & Solomon, 2003:ix).

The haphazard (or *random*) variation and the other sources of variation referred to above can be accounted for in the particular statistical models that are used to analyse data. This is done by including random effects in models. Variance-components methodology is used for the analysis of random effects, where different sources of random variation in the data are accounted for in the statistical model. Cox & Solomon (2003) describe how patterns of variation are partly systematic and partly haphazard. The systematic variation is usually the treatment variation or the between-unit variation, which is explained through the dependency on explanatory features. The most common source of haphazard variation is known as either natural variation, measurement error, residual error or sampling error. It expresses the natural variability that exists between similar individuals or experimental units.

We will explain variance-components methodology by describing the simplest model then building it up step by step. This simplest, smallest model consists of one fixed and one random effect. The fixed effect is the mean, μ, of the observed data. It is one source of the systematic variation which Cox & Solomon (2003) refer to. Their haphazard variation is represented by the variance of the random effect, $\mathbf{e_i}$. We call this variance the random error or residual variance, and denote it by σ_e^2. In the simple model, σ_e^2 accounts for all the haphazard or random variation in the data. Therefore, the total variance in the data is just σ_e^2.

We will build up the more complex models that we require by adding fixed and random effects to the simple model. In general, the mean and residual variance are not counted when referring to the number of fixed and random effects added to a model, as every model must contain at least these two elements. For example, when we say a model contains one fixed and two random effects, we assume that this is over and above the mean and residual variance.

The model we ultimately develop here is the specific type of mixed-effects model for variances, which is required for the analysis of human genetic family data. In addition, aside from the simplest models, we will not discuss estimation or testing because, for the complex models, these are research topics on their own.

To explain the models we want to build up, we use a trivial example. This is because the models used in the analysis of family-based genetics have additional levels of complexity which will be added in the next chapter. Our aim therefore, is to explain the necessary statistical techniques here, in a comprehensive manner and using a simple example, then extending these techniques to include the genetic data.

4.1 Simplest case: Fixed mean model

Suppose that we take a random sample of r red apples. We want to estimate the shelf-life, μ, of a red apple. Let $\mathbf{y} \sim \mathcal{N}(\mu, \sigma_e^2)$ be the random shelf-life of an apple. Then, $\underline{\mathbf{y}}(r \times 1) = (\mathbf{y_1}, \ldots, \mathbf{y_r})^T$ is the random vector of the shelf lives of the r apples. It has a multivariate normal distribution, which is an extension of the normal distribution for a single random variable. Through its form it naturally accounts for the correlation among the elements of $\underline{\mathbf{y}}$. Therefore, since $\underline{\mathbf{y}}$ is a random vector consisting of r elements that are all normally distributed, it has a r-variate multivariate normal distribution with a mean and covariance matrix that we have to determine.

In this very simple case, the model for the shelf lives of the apples is:

$$\mathbf{y_i} = \mu + \mathbf{e_i}, \text{ for } i = 1, \ldots, r. \tag{1}$$

Here, $\mathbf{e_i}$ is the random effect or residual of apple i such that the $\mathbf{e_i}$ are independent and identically distributed (i.i.d.) $\mathcal{N}(0, \sigma_e^2)$. It follows that the vector of residuals for the r apples, $\underline{\mathbf{e}}(r \times 1) = (\mathbf{e_1}, \ldots, \mathbf{e_r})^T \sim \mathcal{N}_r(\underline{0}, \sigma_e^2 \mathcal{I}_\mathbf{r})$.

To find the corresponding model for the vector of all apples $\underline{\mathbf{y}}$, we expand Model (1) for each of the r apples:

$$\begin{pmatrix} \mathbf{y_1} = \mu + \mathbf{e_1} \\ \vdots \\ \mathbf{y_r} = \mu + \mathbf{e_r} \end{pmatrix},$$

which implies

$$\begin{pmatrix} \mathbf{y_1} \\ \vdots \\ \mathbf{y_r} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} \mathbf{e_1} \\ \vdots \\ \mathbf{e_r} \end{pmatrix}.$$

Thus

$$\underline{\mathbf{y}} = \underline{1}_r \mu + \underline{\mathbf{e}}.$$

In standard matrix notation, this is

$$\underline{\mathbf{y}} = \mathcal{X}\underline{\beta} + \underline{\mathbf{e}},$$

where $\mathcal{X}(r \times 1) = \underline{1}_r$ is the design matrix of fixed effects and $\underline{\beta}(1 \times 1) = \mu$ represents the vector of regression coefficients.

From this, the mean and covariance matrix are, respectively,

$$\begin{aligned} E(\underline{\mathbf{y}}) &= E(\underline{1}_r \mu + \underline{\mathbf{e}}) \\ &= \underline{1}_r \mu, \text{ since } E(\underline{\mathbf{e}}) = 0 \\ cov(\underline{\mathbf{y}}) &= E[(\underline{\mathbf{y}} - E(\underline{\mathbf{y}}))(\underline{\mathbf{y}} - E(\underline{\mathbf{y}})^T)] \\ &= E[\underline{\mathbf{e}}\underline{\mathbf{e}}^T] \\ &= \sigma_e^2 \mathcal{I}_\mathbf{r}, \text{ because } \mathbf{e_{ij}} \sim \text{ i.i.d.} \end{aligned}$$

Therefore, $\underline{\mathbf{y}} \sim \mathcal{N}_r(\underline{1}_r \mu, \sigma_e^2 \mathcal{I}_\mathbf{r})$.

The parameters μ and σ_e^2 can be jointly estimated via maximum likelihood estimation (MLE) or method of moment estimation (MME), by first estimating μ then using this estimate to estimate σ_e^2. Both MLE and MME give the same unbiased estimates of μ but MLE gives biased estimates of σ_e^2. Fortunately, we can write $\hat{\sigma}_e^2$ as a function of s_e^2, the unbiased sample estimate.

To obtain the MLEs of model parameters, we first need to define a likelihood function: If $\mathbf{y} \sim \mathcal{N}(\mu, \sigma_e^2)$ is the random shelf-life of an apple, then $f(\mathbf{y})$ is the density function of $\mathbf{y}$, such that

$$f(\mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma_e} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma_e})^2}.$$

The likelihood function of $\mathbf{y}$, denoted by $\mathcal{L}(\cdot)$, in terms of μ and σ_e^2, is the joint density of the independently observed sample values. So,

$$\begin{aligned} \mathcal{L}(\mu, \sigma_e^2) &= \prod_{i=1}^{r} f(\mathbf{y}, \mu, \sigma_e^2) \\ &= \prod_{i=1}^{r} \frac{1}{\sqrt{2\pi}\sigma_e} e^{-\frac{1}{2}\left(\frac{\mathbf{y_i}-\mu}{\sigma_e}\right)^2} \end{aligned}$$

To obtain the MLEs of μ and σ_e^2, we simultaneously maximise the natural log of the likelihood with respect to μ and σ_e^2, to get

$$\begin{aligned} \hat{\mu} &= \bar{\mathbf{y}} \\ &= \frac{1}{r}\sum_{i=1}^{r} \mathbf{y_i} \\ &= \frac{1}{r}\underline{1}_r^T \underline{\mathbf{y}} \end{aligned}$$

$$\begin{aligned} (\frac{r}{r-1})\hat{\sigma_e}^2 &= s_e^2 \\ &= \frac{1}{r-1}\sum_{i=1}^{r}(\mathbf{y_i} - \bar{\mathbf{y}})^2 \\ &= \frac{1}{r-1}(\underline{\mathbf{y}} - \underline{1}_r\bar{\mathbf{y}})^T(\underline{\mathbf{y}} - \underline{1}_r\bar{\mathbf{y}}), \end{aligned}$$

where it can be shown that $\hat{\mu} \sim \mathcal{N}(\mu, \frac{1}{r}\sigma_e^2)$ and $\frac{(r-1)s_e^2}{\sigma_e^2} \sim \chi_{r-1}^2$ (non-central), independently of each other.

These estimates can be used to derive confidence intervals and test hypotheses about the mean (for example $H_0 : \mu = \mu_0$) and/or residual variance (for example $H_0 : \sigma_e^2 = \sigma_0^2$) using likelihood ratios (LR), which are based on the previously defined likelihood function. These tests compare different possible values for the unknown parameter being estimated. Let θ denote the unknown parameter, such that $\theta \in$ S, the parameter space. Then the likelihood of the parameter value under the simple null hypothesis, $H_0 : \theta = \theta_0$, is compared to the likelihood of the value under the (simple) alternate hypothesis, $H_1 : \theta = \theta_1$. The likelihood ratio test statistic is then

$$\lambda = \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_1)}.$$

If λ is large, θ_0 is more likely, but if it is small, θ_1 is more likely.

In our applications, we will always test a simple null hypothesis against a composite alternative hypothesis, which is the sample space. Thus, $H_1 : \theta \in S$. Here, the likelihood ratio is

$$\lambda = \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta \in S)}.$$

Since $\mathcal{L}(\theta \in S)$ is evaluated at $\hat{\theta}$, the MLE of θ, λ will always be less than 1. If λ is close to 1, there is sufficient evidence in support of the null hypothesis. However if λ is significantly smaller than 1, the LR test rejects the null hypothesis (Hogg & Tanis, 2006). The hypothesis test is based on $2ln(\lambda)$ having an asymptotic chi-squared distribution, with degrees of freedom equal to the difference in the number of parameters estimated between the null and full (alternate) models.

In Section 4.1, we have introduced a very simple model and explained how to estimate its parameters. Now suppose that we look more closely at our shelf-life data and realise that perhaps μ does not encompass the full mean of the data. This implies that there is another source of systematic variation in our data and it needs to be accounted for. To do this, we need more fixed effects in our model as it is the fixed effects which account for the systematic variation in data.

4.2 Fixed effects model

Suppose that on closer inspection of the data, we found that there was actually a preservative that was tested on the apples, so that the shelf-life of the treated apples ($t = 1$) could be compared to those that were untreated ($t = 2$). Assume that the treatment was randomly assigned to the apples, such that r_1 were treated with the preservative and r_2 were left untreated. This means that there is now a treatment effect, the additional source of systematic variation, which we need to account for in our model, so that the new model will differ from Model (1) in Section 4.1. Including the treatment implies that the total sample size $N = r_1 + r_2 = r$.

Let μ be the fixed mean shelf-life of the untreated apples, and let τ_t denote the fixed effect of preservative t on the shelf-life of the i^{th} apple. Let $\mathbf{y_{ti}}$ denote the random shelf-life of apple i treated with preservative t. Suppose the data is arranged so that the r_1 treated apples are first and the r_2 untreated ones follow. Then a model for this data, looks as follows:

$$\mathbf{y_{ti}} = \mu + \tau_t + \mathbf{e_{ti}}, \text{ for } t = 1, 2;\ i = 1, \ldots, r_t, \tag{2}$$

where $\mathbf{e_{ti}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$ is the residual for the i^{th} apple. Then the vector of residuals $\underline{\mathbf{e}} \sim \mathcal{N}_r(\underline{0}, \sigma_e^2 \mathcal{I}_\mathbf{r})$.

In linear Model (2), there is one observation on each apple, so there is only one random component, $\mathbf{e_{ti}}$, corresponding to each apple. Here,

$$\begin{aligned} E(\mathbf{y_{ti}}) &= \mu + \tau_t, \text{ since } E(\mathbf{e_{ti}}) = 0 \\ var(\mathbf{y_{ti}}) &= E[(\mathbf{y_{ti}} - (\mu + \tau_t))^2] \\ &= E[\mathbf{e_{ti}}^2] \\ &= \sigma_e^2. \end{aligned}$$

Therefore, $\mathbf{y_{ti}} \sim \mathcal{N}(\mu + \tau_t, \sigma_e^2)$.

To find the model for $\underline{\mathbf{y}}_{\mathbf{ti}}(r \times 1)$, Model (2) can be expanded and written in matrix notation

as follows

$$\begin{pmatrix} \mathbf{y_{11}} = \mu + \tau_1 + \mathbf{e_{11}} \\ \vdots \\ \mathbf{y_{1r_1}} = \mu + \tau_1 + \mathbf{e_{1r_1}} \\ \mathbf{y_{21}} = \mu + \tau_2 + \mathbf{e_{21}} \\ \vdots \\ \mathbf{y_{2r_2}} = \mu + \tau_2 + \mathbf{e_{2r_2}} \end{pmatrix},$$

which is equivalent to

$$\begin{pmatrix} \mathbf{y_{11}} \\ \vdots \\ \mathbf{y_{1r_1}} \\ \mathbf{y_{21}} \\ \vdots \\ \mathbf{y_{2r_2}} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 & 0 \\ \vdots & \\ 1 & 0 \\ 0 & 1 \\ \vdots & \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \mathbf{e_{11}} \\ \vdots \\ \mathbf{e_{1r_1}} \\ \mathbf{e_{21}} \\ \vdots \\ \mathbf{e_{2r_2}} \end{pmatrix}.$$

The above can be written as

$$\underline{\mathbf{y}} = \begin{pmatrix} \underline{1}_{r_1} \\ \underline{1}_{r_2} \end{pmatrix} \mu + \begin{pmatrix} \underline{1}_{r_1} \\ \underline{0}_{r_2} \end{pmatrix} \tau_1 + \begin{pmatrix} \underline{0}_{r_1} \\ \underline{1}_{r_2} \end{pmatrix} \tau_2 + \underline{\mathbf{e}}. \tag{3}$$

Written more concisely, in standard matrix notation, this is

$$\underline{\mathbf{y}} = \mathcal{X}\underline{\beta} + \underline{\mathbf{e}}, \tag{4}$$

where

$$\mathcal{X}(r \times (1+2) = r \times 3) = \begin{pmatrix} \underline{1}_{r_1} & \underline{1}_{r_1} & \underline{0}_{r_1} \\ \underline{1}_{r_2} & \underline{0}_{r_2} & \underline{1}_{r_2} \end{pmatrix}$$

is the design matrix of fixed mean and treatment effects, and

$$\underline{\beta}(3 \times 1) = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}$$

is the vector of regression parameters.

Now, from Model (3)

$$\begin{aligned} E(\underline{\mathbf{y}}) &= \mathcal{X}\underline{\beta}(r \times 1), \text{ since } E(\mathbf{e_{ti}}) = 0 \\ cov(\underline{\mathbf{y}}) &= E[\underline{\mathbf{e}}\underline{\mathbf{e}}^T] \\ &= \sigma_e^2 \mathcal{I}_{\mathbf{r}}, \text{ since } \mathbf{e_{ti}} \sim i.i.d. \\ &= \mathbf{\Omega}(r \times r), \text{ say.} \end{aligned}$$

Therefore, in standard notation, $\underline{\mathbf{y}} \sim \mathcal{N}_r(\mathcal{X}\underline{\beta}, \mathbf{\Omega})$.

The problem with Model (4) is that the columns of $\mathcal{X}$ are not linearly independent, implying that $\mathcal{X}$ is not of full rank. As a result, there are more parameters than equations (3 vs. 2, respectively). Thus, not all the parameters can be uniquely estimated, implying that there are an infinite number of solutions. This makes $\underline{\mathbf{y}}$ unidentifiable. To make it identifiable so that we can estimate the parameters uniquely, we assume $\tau_1 = -\tau_2$. Hence we now have Model (5) with only two fixed effects, μ and τ_1 :

$$\underline{\mathbf{y}} = \mathcal{X}\underline{\beta} + \underline{\mathbf{e}}, \tag{5}$$

where

$$\mathcal{X}(r \times 2) = \begin{pmatrix} \underline{1}_{r_1} & \underline{1}_{r_1} \\ \underline{1}_{r_2} & -\underline{1}_{r_2} \end{pmatrix},$$

which has full rank equal to 2, and

$$\underline{\beta}(2 \times 1) = \begin{pmatrix} \mu \\ \tau_1 \end{pmatrix}:$$

is the corresponding vector of regression parameters.

The parameter estimates of Model (5) can be jointly estimated via maximum likelihood estimation by maximising the log of the likelihood function. However, the likelihood function now changes to include τ_1,

$$\mathcal{L}(\mu, \tau_1, \sigma_e^2) = \prod_{i=1}^{r_1} \frac{1}{\sqrt{2\pi}\sigma_e} e^{-\frac{1}{2}\left(\frac{\mathbf{y_i}-(\mu+\tau_1)}{\sigma_e}\right)^2} \prod_{i=1}^{r_2} \frac{1}{\sqrt{2\pi}\sigma_e} e^{-\frac{1}{2}\left(\frac{\mathbf{y_i}-(\mu-\tau_1)}{\sigma_e}\right)^2}.$$

Since $\mathcal{X}$ is now of full rank, the determinant of $\mathcal{X}^T\mathcal{X}$ exists, and hence its inverse matrix is calculated as follows:

The determinant of $\mathcal{X}^T\mathcal{X}$ is

$$\begin{aligned} det(\mathcal{X}^T\mathcal{X}) &= \begin{vmatrix} r & 2r_1 - r \\ 2r_1 - r & r \end{vmatrix} \\ &= r^2 - (2r_1 - r)^2 \\ &= 4r_1(r - r_1). \end{aligned}$$

From this, the inverse of $\mathcal{X}^T\mathcal{X}$ is

$$(\mathcal{X}^T\mathcal{X})^{-1} = \frac{1}{4r_1(r - r_1)} \begin{pmatrix} r & r - 2r_1 \\ r - 2r_1 & r \end{pmatrix}.$$

Let

$$\bar{\mathbf{y}}_{\mathbf{t}} = \frac{1}{r_t}\sum_{i=1}^{r_t}\mathbf{y_{ti}} = \frac{1}{r_t}\underline{1}_{r_t}^T\underline{\mathbf{y}}_{\mathbf{t}},\ t = 1, 2.$$

Then the unbiased MLEs of the parameters of Model (5) can be shown to be:

$$\underline{\hat{\beta}} = \begin{pmatrix} \hat{\mu} \\ \hat{\tau_1} \end{pmatrix} = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\underline{\mathbf{y}}.$$

Using this we can find $\hat{\mu}$ and $\hat{\tau_1}$:

$$\begin{aligned}
\underline{\hat{\beta}} &= (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\underline{\mathbf{y}} \\
&= \frac{1}{4r_1(r-r_1)}\begin{pmatrix} r & r-2r_1 \\ r-2r_1 & r \end{pmatrix}\begin{pmatrix} \underline{1}_{r_1}^T & \underline{1}_{r_2}^T \\ \underline{1}_{r_1}^T & -\underline{1}_{r_2}^T \end{pmatrix}\begin{pmatrix} \underline{\mathbf{y}}_{\mathbf{1}} \\ \underline{\mathbf{y}}_{\mathbf{2}} \end{pmatrix} \\
&= \frac{1}{4r_1(r-r_1)}\begin{pmatrix} r & r-2r_1 \\ r-2r_1 & r \end{pmatrix}\begin{pmatrix} r_1\bar{\mathbf{y}}_{\mathbf{1}} + r_2\bar{\mathbf{y}}_{\mathbf{2}} \\ r_1\bar{\mathbf{y}}_{\mathbf{1}} - r_2\bar{\mathbf{y}}_{\mathbf{2}} \end{pmatrix} \\
&= \vdots \\
&= \frac{1}{4r_1r_2}\begin{pmatrix} 2r_1r_2(\bar{\mathbf{y}}_{\mathbf{1}} + \bar{\mathbf{y}}_{\mathbf{2}}) \\ 2r_1r_2(\bar{\mathbf{y}}_{\mathbf{1}} - \bar{\mathbf{y}}_{\mathbf{2}}) \end{pmatrix} \\
&= \begin{pmatrix} \hat{\mu} \\ \hat{\tau_1} \end{pmatrix}
\end{aligned}$$

where,

$$\begin{aligned}
\hat{\mu} &= \frac{1}{2}(\bar{\mathbf{y}}_{\mathbf{1}} + \bar{\mathbf{y}}_{\mathbf{2}}); \\
\hat{\tau_1} &= \frac{1}{2}(\bar{\mathbf{y}}_{\mathbf{1}} - \bar{\mathbf{y}}_{\mathbf{2}}).
\end{aligned}$$

So,

$$\begin{aligned}
\hat{\mu} + \hat{\tau_1} &= \frac{1}{2}(\bar{\mathbf{y}}_{\mathbf{1}} + \bar{\mathbf{y}}_{\mathbf{2}}) + \frac{1}{2}(\bar{\mathbf{y}}_{\mathbf{1}} - \bar{\mathbf{y}}_{\mathbf{2}}) \\
&= \bar{\mathbf{y}}_{\mathbf{1}}
\end{aligned}$$

and

$$\begin{aligned}
\hat{\mu} - \hat{\tau_1} &= \frac{1}{2}(\bar{\mathbf{y}}_{\mathbf{1}} + \bar{\mathbf{y}}_{\mathbf{2}}) - \frac{1}{2}(\bar{\mathbf{y}}_{\mathbf{1}} - \bar{\mathbf{y}}_{\mathbf{2}}) \\
&= \bar{\mathbf{y}}_{\mathbf{2}}.
\end{aligned}$$

Now,

$$\begin{aligned}
E(\underline{\hat{\beta}}) &= E\begin{pmatrix} \hat{\mu} \\ \hat{\tau_1} \end{pmatrix} \\
&= \underline{\beta}
\end{aligned}$$

and

$$\begin{aligned} cov(\underline{\hat{\beta}}) &= \sigma_e^2(\mathcal{X}^T\mathcal{X})^{-1} \\ &= \frac{\sigma_e^2}{4r_1r_2}\begin{pmatrix} r & r-2r_1 \\ r-2r_1 & r \end{pmatrix}. \end{aligned}$$

Therefore,

$$var(\hat{\mu}) = \frac{\sigma_e^2 r}{4r_1r_2} = var(\hat{\tau_1})$$

and

$$cov(\hat{\mu}, \hat{\tau_1}) = \frac{\sigma_e^2(r-2r_1)}{4r_1r_2}.$$

Finally, we know that the MLE of σ_e^2 is biased, so in Section 4.1 we gave $(\frac{r}{r-1})\hat{\sigma_e}^2 = s_e^2$, which is unbiased. In this section, where we consider treatment effects,

$$\left(\frac{r}{r-1}\right)\left(\frac{\hat{\sigma_e}^2 r}{4r_1r_2}\right) = s_e^2.$$

Thus, $\hat{\mu} \sim \mathcal{N}(\mu, \frac{\sigma_e^2 r}{4r_1r_2}), \hat{\tau_1} \sim \mathcal{N}(\tau_1, \frac{\sigma_e^2 r}{4r_1r_2})$ and $\frac{4r_1r_2(r-1)s_e^2}{r^2\hat{\sigma_e}^2} \sim \chi^2_{\frac{4r_1r_2(r-1)}{r^2}}$, independently of $\hat{\mu}$ and $\hat{\tau_1}$.

As before, these estimates can be used to derive confidence intervals and test hypotheses about the population. To test population means, the null hypothesis could be, for example, $H_0 : \tau_1 = 0$. If we wanted to test the population variance, we could test $H_0 : \sigma_e^2 = \sigma_0^2$, which is a chi-squared test. However, we are not interested in testing the variance in this instance, therefore this test is not shown here.

For testing either of these hypotheses, we use a likelihood ratio (LR) test, where the LR is in terms of a ratio of two variance estimates. It can be shown to be equivalent to an F-test, with the numerator and denominator having been calculated from classical variance decomposition. This can be shown for the shelf-life data here, where we can test for example, the null hypothesis $H_0 : \tau_1 = 0$. We can show the relationship between the variance-components here using analysis of variance (ANOVA):

The total sum of squared (SS) deviations is a measure of the sum of squared deviation of each observation from the overall mean. Since we have treatment effects in our model,

and we are testing these via $H_0 : \tau_1 = 0$, we must decompose the total sum of squares as follows:

$$\begin{aligned}
\mathrm{SS}_{total} &= \sum_{t=1}^{2}\sum_{i=1}^{r_t}(\mathbf{y_{ti}} - \bar{\mathbf{y}}_.)^2 \\
&= \sum_{t=1}^{2}\sum_{i=1}^{r_t}(\mathbf{y_{ti}} - \bar{\mathbf{y}}_\mathbf{t} + \bar{\mathbf{y}}_\mathbf{t} - \bar{\mathbf{y}}_.)^2 \\
&= \sum_{t=1}^{2}\sum_{i=1}^{r_t}(\mathbf{y_{ti}} - \bar{\mathbf{y}}_\mathbf{t})^2 + \sum_{t=1}^{2}\sum_{i=1}^{r_t}(\bar{\mathbf{y}}_\mathbf{t} - \bar{\mathbf{y}}_.)^2 + 2\sum_{t=1}^{2}\sum_{i=1}^{r_t}(\mathbf{y_{ti}} - \bar{\mathbf{y}}_\mathbf{t})(\bar{\mathbf{y}}_\mathbf{t} - \bar{\mathbf{y}}_.) \\
&= \sum_{t=1}^{2}\sum_{i=1}^{r_t}(\mathbf{y_{ti}} - \bar{\mathbf{y}}_\mathbf{t})^2 + r\sum_{t=1}^{2}(\bar{\mathbf{y}}_\mathbf{t} - \bar{\mathbf{y}}_.)^2 + 0 \\
&= \mathrm{SS}_{error} + \mathrm{SS}_{treatment}.
\end{aligned}$$

The SS_{error} is a measure of the squared deviations between observations within a treatment group and the mean of that group, while the $\mathrm{SS}_{treatment}$ is a measure of the squared deviations between each treatment mean and the overall mean. Now, under the null hypothesis of equal treatment means for our two treatment groups,

$$\frac{\mathrm{SS}_{error}}{\sigma_e^2} \sim \chi^2_{r-2}$$
$$\frac{\mathrm{SS}_{treatment}}{\sigma_e^2} \sim \chi^2_{2-1},$$

where the latter is non-central chi-squared when the null hypothesis is rejected. We thus have two independent estimates of σ_e^2 from the two mean squares (MS)

$$\mathrm{MS}_{error} = \frac{\mathrm{SS}_{error}}{r-2}$$
$$\mathrm{MS}_{treatment} = \frac{\mathrm{SS}_{treatment}}{2-1}.$$

The mean squared error (MS_{error}) is an unbiased estimator of σ_e^2, while the mean square treatment ($\mathrm{MS}_{treatment}$) is only an unbiased estimator if the null hypothesis $H_0 : \tau_1 = 0$ is true. The likelihood ratio, in this case, is just $\frac{\mathrm{MS}_t}{\mathrm{MS}_e} = F(2-1, r-2)$, and it is this F-statistic which is used to assess whether or not the null hypothesis should be rejected. It has a non-central F-distribution with $(2-1)$ and $(r-2)$ degrees of freedom, for the numerator and denominator respectively, and non-centrality parameter equal to

$$r_t\sum_{t=1}^{2}\left(\frac{\mu_t - \mu}{\sigma_e}\right)^2 = r_1\left(\frac{\tau_1}{\sigma_e}\right)^2 - r_2\left(\frac{\tau_1}{\sigma_e}\right)^2.$$

Under the null hypothesis of no treatment differences, the non-centrality parameter is zero. If MS_t is significantly larger than MS_e, there is a large difference between treatments, implying that there is sufficient evidence to reject the null hypothesis. Therefore, by having a treatment effect in the model, σ_e^2 is better estimated than if treatment effect is not included, in which case the variance estimate will be over-inflated because the treatment contribution to σ_e^2 is not teased out. If there is no significant difference between the treatments, the estimate of σ_e^2 will not change, even when the treatment effect is included in the model. This is related to the topic of confounding, which is discussed later. Although the focus of this study is on variances, adding fixed effects to a model affects the estimates of the variance and variance-components, making the understanding of the theory important.

In this section, because we are estimating and testing a specific treatment, inferences made from the analysis are only made about this treatment. Therefore, the treatment effect τ_t is known as a *fixed effect*, and models such as Model (5) are called *fixed-effects models*. The parameters of such models form the basis for studying contrasts, for example, between treatments, groups or people.

4.3 Random effects or variance-components model

Let us put aside the treatment effect for a moment. Suppose we now cut each of the r randomly chosen apples into $j = 1, \ldots, 4$ pieces of approximately equal size. If $\mathbf{y_{ij}}$ is the shelf-life of the j^{th} piece of the i^{th} apple, then $\underline{\mathbf{y}}_\mathbf{i}(4 \times 1) = (\mathbf{y_{i1}}, \ldots, \mathbf{y_{i4}})^T$ is the vector of measurements of the shelf-life of pieces of the i^{th}apple, $\underline{\mathbf{y}}(N \times 1) = (\underline{\mathbf{y}}_\mathbf{1}, \ldots, \underline{\mathbf{y}}_\mathbf{r})^T$ is the vector of measurements of the shelf-life of all apple pieces, and the total sample size is $N = \sum_i^r 4 = 4r$.

In contrast to the previous examples, we now have measurements on pieces of the same apple. As a result, the shelf-lives of the apple pieces are correlated and this correlation must somehow be accounted for in the model. This is done by including two random effects in a statistical model, one for apple pieces, which we already have, and another for

the apples from which those pieces come. Both of these random effects have means of zero but different variances. The residual variance, σ_e^2, is for apple pieces while σ_b^2 is for apples. So here, the sum of these two variances gives the total variance of the shelf-life of the apple pieces. Adding the random apple effect to the model affects the variance-covariance matrix of the observations. As a result, the variance and covariances of $\underline{\mathbf{y}}_\mathbf{i}$, and thus $\underline{\mathbf{y}}$, will not be as straight-forward as previously.

Let $\mathbf{b_i}$ be the random apple effect, such that $\mathbf{b_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_b^2)$. In terms of the impact of the haphazard variation which Cox & Solomon (2003) refer to, $\mathbf{b_i}$ is the random effect measuring the impact of apples on shelf-life.

Let $\mathbf{e_{ij}}$ be the residual corresponding to piece j of the i^{th} apple, such that $\mathbf{e_{ij}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$. Then $\underline{\mathbf{e}}_\mathbf{i}(4 \times 1) = (\mathbf{e_{i1}}, \ldots, \mathbf{e_{i4}})^T \sim \mathcal{N}_4(\underline{0}, \sigma_e^2 \mathcal{I}_\mathbf{4})$ is the within-apple residual vector.

The $\mathbf{e_{ij}}$ and $\mathbf{b_i}$ are independent of each other within each apple and also for different apples.

Let μ be the overall mean shelf-life of the apple pieces, then a model for this data, for the j^{th} piece of the i^{th} apple, is:

$$\mathbf{y_{ij}} = \mu + \mathbf{b_i} + \mathbf{e_{ij}} \text{ for } i = 1, \ldots, r;\ j = 1, \ldots, 4. \tag{6}$$

Here,

$$\begin{aligned} E(\mathbf{y_{ij}}) &= \mu, \text{ since } E(\mathbf{e_{ij}}) = 0 \\ var(\mathbf{y_{ij}}) &= \sigma_b^2 + \sigma_e^2, \text{ because } cov(\mathbf{b_i}, \mathbf{e_{ij}}) = 0. \end{aligned}$$

Therefore, $\mathbf{y_{ij}} \sim \mathcal{N}(\mu, \sigma_b^2 + \sigma_e^2)$, where σ_b^2 is the between-apple component of variance and σ_e^2 is the within-apple component of variance.

Expanding out Model (6), for an apple, gives

$$\begin{pmatrix} \mathbf{y_{i1}} = \mu + \mathbf{b_i} + \mathbf{e_{i1}} \\ \vdots \\ \mathbf{y_{i4}} = \mu + \mathbf{b_i} + \mathbf{e_{i4}} \end{pmatrix},$$

which is equivalent to

$$\begin{pmatrix} \mathbf{y_{i1}} \\ \mathbf{y_{i2}} \\ \mathbf{y_{i3}} \\ \mathbf{y_{i4}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{b_i} + \begin{pmatrix} \mathbf{e_{i1}} \\ \mathbf{e_{i2}} \\ \mathbf{e_{i3}} \\ \mathbf{e_{i4}} \end{pmatrix}.$$

Then, in matrix notation, Model (6) is

$$\underline{\mathbf{y}}_{\mathbf{i}} = \underline{1}_4\mu + \underline{1}_4\mathbf{b_i} + \underline{\mathbf{e}}_{\mathbf{i}}, \text{ for } i = 1, \ldots, r. \tag{7}$$

The mean vector and covariance matrix for $\underline{\mathbf{y}}_{\mathbf{i}}(4 \text{ x } 1)$ are calculated as follows:

$$\begin{aligned}
E(\underline{\mathbf{y}}_{\mathbf{i}}) &= \underline{1}_4\mu(4 \times 1), \text{ since } E(\mathbf{e_{ij}}) = 0 \\
cov(\underline{\mathbf{y}}_{\mathbf{i}}) &= E[(\underline{1}_4\mathbf{b_i} + \underline{\mathbf{e}}_{\mathbf{i}})(\underline{1}_4\mathbf{b_i} + \underline{\mathbf{e}}_{\mathbf{i}})^T] \\
&= \underline{1}_4 E(\mathbf{b_i}\mathbf{b_i}^T)\underline{1}_4^T + E(\underline{\mathbf{e}}_{\mathbf{i}}\underline{\mathbf{e}}_{\mathbf{i}}^T), \text{ since } cov(\mathbf{b_i}, \mathbf{e_{ij}}) = 0. \\
&= \sigma_b^2\underline{1}_4\underline{1}_4^T + \sigma_e^2\mathcal{I}_{\mathbf{4}}.
\end{aligned}$$

Now,

$$\sigma_b^2\underline{1}_4\underline{1}_4^T = \sigma_b^2\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} = \sigma_b^2\mathcal{J}_{\mathbf{4}}.$$

This implies that the covariance matrix of $\underline{\mathbf{y}}_{\mathbf{i}}$

$$\begin{aligned}
cov(\underline{\mathbf{y}}_{\mathbf{i}}) &= \sigma_b^2\mathcal{J}_{\mathbf{4}} + \sigma_e^2\mathcal{I}_{\mathbf{4}} \\
&= \begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 \end{pmatrix} \\
&= \mathbf{\Omega_i}(4 \times 4), \text{ say.}
\end{aligned}$$

So, $cov(\mathbf{y_{ij}}, \mathbf{y_{ik}}) = \sigma_b^2$, for all $j \neq k$, and $\underline{\mathbf{y}}_{\mathbf{i}} \sim \mathcal{N}_4(\underline{1}_4\mu, \mathbf{\Omega_i})$.

Note that here, $\mathbf{\Omega_i}$ has a compound symmetry structure, which is when the correlation, ρ, between all pairs of observations, is the same. In our case, the correlation coefficient between any two pieces of the i^{th} apple is given by

$$\rho(\mathbf{y_{ij}}, \mathbf{y_{ik}}) = \frac{cov(\mathbf{y_{ij}}, \mathbf{y_{ik}})}{\sqrt{var(\mathbf{y_{ij}})var(\mathbf{y_{ik}})}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}, \text{ for all } j \neq k.$$

This known as the intraclass correlation coefficient (ICC) and it is estimated from Model (7) where, within an apple, the covariance between each pair of apple pieces is the same.

In the extreme situation when the between-apple variance $\sigma_b^2 = 0, \rho = 0$, implying that there is no correlation between any two pieces of the same apple. Hence there is maximum variation between two such apple pieces. However, when σ_b^2 is much larger than σ_e^2, (that is, $\sigma_e^2 \to 0$) then the correlation between pieces of the same apple is also large. This in turn means that there is minimum variation between these pieces.

Later, we will see how this changes when considering families rather than apples, and we will also see where the complications arise with family data.

We can now write out the model for $\underline{\mathbf{y}}(N \times 1)$ by expanding out Model (7) for all r apples:

$$\begin{pmatrix} \underline{\mathbf{y_1}} \\ \underline{\mathbf{y_2}} \\ \vdots \\ \underline{\mathbf{y_r}} \end{pmatrix} = \begin{pmatrix} \underline{1}_4 \\ \underline{1}_4 \\ \vdots \\ \underline{1}_4 \end{pmatrix} \mu + \begin{pmatrix} \underline{1}_4 & \underline{0}_4 & \cdots & \underline{0}_4 \\ \underline{0}_4 & \underline{1}_4 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \underline{0}_4 \\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{1}_4 \end{pmatrix} \begin{pmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \vdots \\ \mathbf{b_r} \end{pmatrix} + \begin{pmatrix} \underline{\mathbf{e}}_\mathbf{1} \\ \underline{\mathbf{e}}_\mathbf{2} \\ \vdots \\ \underline{\mathbf{e}}_\mathbf{r} \end{pmatrix}.$$

Let

$$\mathcal{Z}_b(N \times r) = \begin{pmatrix} \underline{1}_4 & \underline{0}_4 & \cdots & \underline{0}_4 \\ \underline{0}_4 & \underline{1}_4 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \underline{0}_4 \\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{1}_4 \end{pmatrix},$$

then

$$\underline{\mathbf{y}} = \underline{1}_N \mu + \mathcal{Z}_b \underline{\mathbf{b}} + \underline{\mathbf{e}}, \tag{8}$$

where

$\underline{\mathbf{b}}(r \times 1) = (\mathbf{b_1}, \ldots, \mathbf{b_r})^T$ is the vector of random effects,

$\mathcal{Z}_b(N \times r)$ is the design matrix corresponding to the random apple effects, and

$\underline{\mathbf{e}}(N \times 1) = (\mathbf{e_{11}}, \ldots, \mathbf{e_{r4}})^T \sim \mathcal{N}_N(\underline{0}, \sigma_e^2 \mathcal{I}_\mathbf{N})$ is the vector of residuals.

Model (8) can be written more succinctly in standard notation, as

$$\underline{\mathbf{y}} = \mathcal{X}\underline{\beta} + \mathcal{Z}_b \underline{\mathbf{b}} + \underline{\mathbf{e}}, \tag{9}$$

where $\underline{\beta}(1 \times 1) = \mu$ is the vector of fixed mean effects,

$\mathcal{X}(N \times 1) = \underline{1}_N$ is the design matrix for the mean vector,

$\mathcal{Z}_b(N \times r)$ is the design matrix corresponding to the random effects, and

$\underline{\mathbf{b}}(r \times 1)$ is the vector of random effects.

Now,

$$\begin{aligned} E(\underline{\mathbf{y}}) &= \underline{1}_N\mu, \text{ since } E(\mathbf{e_{ij}}) = 0 \\ &= \mathcal{X}\underline{\beta}(N \times 1) \\ cov(\underline{\mathbf{y}}) &= \mathcal{Z}_b E(\underline{\mathbf{b}}\underline{\mathbf{b}}^T)\mathcal{Z}_b^T + E(\underline{\mathbf{e}}\underline{\mathbf{e}}^T), \text{ since } cov(\mathbf{b_i}, \mathbf{e_{ij}}) = 0. \\ &= \sigma_b^2 \mathcal{Z}_b \mathcal{Z}_b^T + \sigma_e^2 \mathcal{I}_{\mathbf{N}} \end{aligned}$$

where $\mathcal{Z}_b\mathcal{Z}_b^T(N \times N)$ is a block diagonal matrix consisting of (4×4) blocks of 1 on the diagonal and zeros on the off-diagonal. Thus,

$$\begin{aligned} \mathcal{Z}_b\mathcal{Z}_b^T(N \times N) &= \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & & & & \vdots & & & & \ddots & & & \vdots & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{J}_4 & \mathcal{O}_4 & \cdots & \mathcal{O}_4 \\ \mathcal{O}_4 & \mathcal{J}_4 & \cdots & \mathcal{O}_4 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{O}_4 & \mathcal{O}_4 & \cdots & \mathcal{J}_4 \end{pmatrix}. \end{aligned}$$

In standard notation, $\underline{\mathbf{y}} \sim \mathcal{N}_N(\mathcal{X}\underline{\beta}, \mathbf{\Omega})$, where the block diagonal covariance matrix of $\underline{\mathbf{y}}$ is

$$\begin{aligned} \mathbf{\Omega}(N \times N) &= \sigma_b^2 \mathcal{Z}_b \mathcal{Z}_b^T + \sigma_e^2 \mathcal{I}_{\mathbf{N}} \\ &= \sigma_b^2 \begin{pmatrix} \mathcal{J}_4 & \mathcal{O}_4 & \cdots & \mathcal{O}_4 \\ \mathcal{O}_4 & \mathcal{J}_4 & \cdots & \mathcal{O}_4 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{O}_4 & \mathcal{O}_4 & \cdots & \mathcal{J}_4 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} \mathcal{I}_4 & \mathcal{O}_4 & \cdots & \mathcal{O}_4 \\ \mathcal{O}_4 & \mathcal{I}_4 & \cdots & \mathcal{O}_4 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{O}_4 & \mathcal{O}_4 & \cdots & \mathcal{I}_4 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{\Omega_1} & 0 & \cdots & 0 \\ 0 & \mathbf{\Omega_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{\Omega_r} \end{pmatrix}, \end{aligned}$$

where the elements $\mathbf{\Omega_i}$ are the previously shown per-apple covariance matrices. The diagonalisation can be explained as follows:

We see that the covariances between two pieces of different apples is $cov(\mathbf{y_{ij}}, \mathbf{y_{sk}}) = 0$, for all j, k and $i \neq s$, so the correlation coefficient between any two pieces from different apples is

$$\rho(\mathbf{y_{ij}}, \mathbf{y_{sk}}) = \frac{cov(\mathbf{y_{ij}}, \mathbf{y_{sk}})}{\sqrt{var(\mathbf{y_{ij}})var(\mathbf{y_{sk}})}} = \frac{0}{\sigma_b^2 + \sigma_e^2} = 0, \text{ for all } j, k \text{ and } i \neq s.$$

Therefore, all of the off-diagonal elements of $\mathbf{\Omega}$ are zero. The diagonal elements of $\mathbf{\Omega}$ are the within-apple covariances and thus consist of blocks of elements which are the compound symmetric, per-apple, $\mathbf{\Omega_i}$ matrices. There are r of these– one for each of the r apples under investigation.

Here, we can again decompose the total sum of squares, as we did in Section 4.2. However, instead of a treatment sum of squares, we will have a between-apple sum of squares (SS_b). This in turn implies that the correlation between two pieces of the same apple can be written in terms of the mean squares, namely

$$\rho(\mathbf{y_{ij}}, \mathbf{y_{ik}}) = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_e^2} = \frac{\text{MS}_b - \text{MS}_e}{(r_0 - 1)\text{MS}_b - \text{MS}_e}, \text{ for all } j \neq k,$$

where

$$r_0 = \frac{2r_1(r - r_1)}{r}.$$

This leads to a F-test where the F-statistic under the null hypothesis, $H_0 : \sigma_b^2 = 0$, is $F = \left(\frac{\text{MS}_b}{\text{MS}_e}\right) \sim F(2 - 1, r - 2)$.

As in previous sections, the method of moments estimates again differ from the maximum likelihood estimates, although it is not shown here as estimation is not our focus.

In this example, if we had failed to account for the fact that pieces of the same apple are correlated, then the mean square error of shelf-life of apple pieces would have been over-inflated. This is because all the variation in the lifetimes of apple pieces would not have been properly accounted for. By correctly accounting for the correlations and all the sources of random variation in the lifetimes, the variance of $\mathbf{y_{ij}}$, and thus the variances

and covariances of $\underline{\mathbf{y}}_{\mathbf{i}}$ and $\underline{\mathbf{y}}$, are more accurately estimated. In other words, the estimates will be measured with little error and will thus be close to the true value (Rothman et al., 2008). Therefore, any tests carried out on the data will be more precise, that is, the random error of estimation will be small (Rothman et al., 2008).

In this section, as in the previous ones, we have models which are used to estimate the within-apple/residual variance σ_e^2. The difference in each section is that the value of σ_e^2 varies according to the specified model. So whenever new factors are added to a model, the estimate of this within-apple variance changes according to what kind of factor is added.

In Model (9), the additional factor (apples) represents a random sample from a larger set of factors. Since such models contain only random effect terms, they are called *random-effects models*. The levels of a random effect are not of particular interest in themselves (so we are not interested in particular apples), but may be useful in understanding some of the variation in the underlying population from which they are drawn, and then removing it. Random variables are usually summarised by their variances, as shown in this section (4.3). As a result, such models are also known as *variance-components models*.

Fixed effects are generally estimated in a different way to random effects. They are assumed to have no variance of their own, while each random effect has a corresponding variance-component which must be estimated. As illustrated in Section 4.2, the addition of fixed effects to a model (the treatment effect in our case) affects the estimate of the residual variance, σ_e^2. In particular, by including them in a model and then estimating them, the residual variance is reduced because the variance due to these effects is adjusted for (by removing it from the residual variance).

As demonstrated above, effects are fixed or random depending on whether they are the only effects of interest (fixed effects) or if they are randomly sampled from a larger population of effects and are not, in themselves, of particular interest (random effects). These effects can be modeled separately, as in the models shown in the previous sections, where

fixed effects (Section 4.2) and random effects (Section 4.3) are modeled separately. However, both fixed and random effects can be accounted for in the same model. Such models are known as *mixed-effects models* and are shown in the following sections.

4.4 Mixed-effects model: Two fixed effects and one random effect

Suppose that we again cut each of the r apples into $j = 1, \ldots, 4$ pieces. We again want to compare the shelf-life of pieces of r_1 treated apples to the shelf-life of pieces of r_2 untreated apples ($t = 1, 2$ respectively). Assume that μ is the fixed mean shelf-life of the untreated apple pieces. Here, $r = r_1 + r_2$ apples, so the total sample size $N = \sum_i^r 4 = 4r$. The model for this data is an example of a simple mixed-effects model, containing one fixed effect (treatment) and one random effect (apple). The structure of this model is just a combination of Models (5) , the fixed-effects model, and Model (9), the random effects model. In practice, mixed-effects models are usually more complex, and since we are interested in the effects of adding different fixed and random effects to a model, we will not discuss the simplest model here. Instead we consider a model with two fixed effects and one random effect.

So, suppose we also have information on the weight (in grams) of all the apple pieces and we would like to include it in the model as a covariate. In statistics, we consider covariates carefully for several reasons:

1. some of the covariates may be important predictors of the outcome (shelf-life of the apples, in our case), but their effects maybe masked by the other predictors;

2. the effect of some predictors may not be balanced across the levels of exposure, so we also need to somehow account for this;

3. it is possible to have a covariate that is independently associated with both a predictor (usually the exposure variable) and the outcome variable, in a way that causes it to confound or confuse the effect of the exposure variable on the outcome, even though it is not in the causal pathway between the two. Such a covariate contributes to the covariance in a model and is known as a *confounder*. The distortion produced by a confounder may be

large and can lead to effects being either overestimated or underestimated. Epidemiologists call this distortion *confounding*.

In a statistical model, we 'adjust for', 'account for','correct for' or 'covary for' the effects of such covariates and confounders by including them in the model as fixed effects. This removes the effects of these variables from the sum of squares of the model, leaving behind only the effect that is independent of these variables (Rothman et al., 2008). Adjustment also removes the effect of these confounders from the estimates of the fixed effects, but that is not the focus here so it will not be discussed further.

Let x_{tij} represent the weight of the j^{th} piece of the i^{th} apple, treated with t, which has unknown coefficient α, in the model.

Let $\mathbf{y_{tij}}$ be the shelf-life of the j^{th} piece of the i^{th} apple, where apple i is treated with preservative t. Then, $\underline{\mathbf{y}}_{\mathbf{ti}}(4 \times 1) = (\mathbf{y_{ti1}}, \ldots, \mathbf{y_{ti4}})^T$ is the vector of measurements of the shelf-life of pieces of the i^{th}apple, treated with t, and $\underline{\mathbf{y}}(N \times 1) = (\underline{\mathbf{y}}_{\mathbf{11}}, \ldots, \underline{\mathbf{y}}_{\mathbf{2r_2}})^T$ is the vector of measurements of the shelf-life of all apple pieces. Finally, let τ_t represent the effect of preservative t on the i^{th} apple, such that $\tau_1 = -\tau_2$, which as we showed previously, ensures identifiability in estimation.

The data can then be modeled as follows:

$$\mathbf{y_{tij}} = \mu + \tau_t + \alpha x_{tij} + \mathbf{b_{ti}} + \mathbf{e_{tij}}, \text{ for } t = 1, 2;\ i = 1, \ldots, r_t;\ j = 1, \ldots, 4, \qquad (10)$$

where, $\mathbf{b_{ti}}$ and $\mathbf{e_{tij}}$ are, respectively, the random apple effect and residual for the j^{th} piece of the i^{th} apple, for treatment t.

Assume that $\mathbf{e_{tij}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$ and $\mathbf{b_{ti}} \sim$ i.i.d $\mathcal{N}(0, \sigma_b^2)$, independently of the $\mathbf{e_{tij}}$.

Then, $var(\mathbf{y_{tij}}) = \sigma_b^2 + \sigma_e^2$, as we had for Model (6). However, the expected value of $\mathbf{y_{tij}}$, $E(\mathbf{y_{tij}}) = E(\mu + \tau_{tij} + \alpha x_{tij} + \mathbf{b_{ti}} + \mathbf{e_{tij}}) = \mu + \tau_{tij} + \alpha x_{tij}$, which is different to that for Model (6) as we now have two additional fixed effects.

Therefore, $\mathbf{y_{tij}} \sim \mathcal{N}(\mu + \tau_{tij} + \alpha x_{tij}, \sigma_b^2 + \sigma_e^2)$.

Expanding Model (10) for the i^{th} apple gives:

$$\begin{pmatrix} \mathbf{y_{ti1}} = \mu + \tau_t + \alpha x_{ti1} + \mathbf{b_{ti}} + \mathbf{e_{ti1}} \\ \vdots \\ \mathbf{y_{ti4}} = \mu + \tau_t + \alpha x_{ti4} + \mathbf{b_{ti}} + \mathbf{e_{ti4}} \end{pmatrix},$$

which implies

$$\begin{pmatrix} \mathbf{y_{ti1}} \\ \mathbf{y_{ti2}} \\ \mathbf{y_{ti3}} \\ \mathbf{y_{ti4}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tau_t + \alpha \begin{pmatrix} x_{ti1} \\ x_{ti2} \\ x_{ti3} \\ x_{ti4} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{b_{ti}} + \begin{pmatrix} \mathbf{e_{ti1}} \\ \mathbf{e_{ti2}} \\ \mathbf{e_{ti3}} \\ \mathbf{e_{ti4}} \end{pmatrix}.$$

Thus, for the vector $\underline{\mathbf{y}}_{\mathbf{ti}}(4 \times 1)$, we have:

$$\underline{\mathbf{y}}_{\mathbf{ti}} = \underline{1}_4\mu + \underline{1}_4\tau_t + \alpha\underline{x}_{ti} + \underline{1}_4\mathbf{b_{ti}} + \underline{\mathbf{e}}_{\mathbf{ti}}, \text{ for } t = 1, 2;\ i = 1, \ldots, r_t \tag{11}$$

where $\underline{x}_{ti}(4 \times 1) = (x_{ti1}, \ldots, x_{ti4})^T$ is the vector of weights of each apple piece, and $\underline{\mathbf{e}}_{\mathbf{ti}}(4 \times 1) = (\mathbf{e_{ti1}}, \ldots, \mathbf{e_{ti4}})^T \sim \mathcal{N}_4(\underline{0}, \sigma_e^2\mathcal{I}_4)$ is the within-apple residual vector. Thus, $\underline{\mathbf{e}}_{\mathbf{ti}}$ and $\mathbf{b_{ti}}$ are independent in each apple and independent of each other for different apples.

The expected value and covariance matrix for $\underline{\mathbf{y}}_{\mathbf{ti}}(4 \times 1)$ are:

$$\begin{aligned} E(\underline{\mathbf{y}}_{\mathbf{ti}}) &= \underline{1}_4\mu + \underline{1}_4\tau_t + \alpha\underline{x}_{ti}(4 \times 1), \text{ since } E(\mathbf{e_{tij}}) = 0 \\ cov(\underline{\mathbf{y}}_{\mathbf{ti}}) &= \sigma_b^2\mathcal{J}_4 + \sigma_e^2\mathcal{I}_4, \text{ since } cov(\mathbf{b_{ti}}, \mathbf{e_{tij}}) = 0 \\ &= \mathbf{\Omega_i}(4 \times 4), \text{ say.} \end{aligned}$$

Writing out $\mathbf{\Omega_i}(4 \times 4)$ fully for the i^{th} apple gives the same covariance matrix as for Model (7), the random effects model. Thus,

$$\begin{aligned} cov(\mathbf{y_{tij}}, \mathbf{y_{tik}}) &= \sigma_b^2, \text{ for all } j \neq k, t = 1, 2 \\ \rho(\mathbf{y_{tij}}, \mathbf{y_{tik}}) &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}, \text{ for all } j \neq k, t = 1, 2. \end{aligned}$$

However, here, $\underline{\mathbf{y}}_{\mathbf{ti}} \sim \mathcal{N}_4(\underline{1}_4\mu + \underline{1}_4\tau_t + \underline{x}_{ti}\alpha, \mathbf{\Omega_i})$, as there are additional fixed effects terms.

To write out the vector of all observations, $\underline{\mathbf{y}}(N \times 1)$, in matrix notation, expand Model

(11):

$$\begin{pmatrix} \underline{\mathbf{y}}_{\mathbf{11}} \\ \vdots \\ \underline{\mathbf{y}}_{\mathbf{1r_1}} \\ \underline{\mathbf{y}}_{\mathbf{21}} \\ \vdots \\ \underline{\mathbf{y}}_{\mathbf{2r_2}} \end{pmatrix} = \begin{pmatrix} \underline{1}_4 \\ \vdots \\ \underline{1}_4 \\ \underline{1}_4 \\ \vdots \\ \underline{1}_4 \end{pmatrix} \mu + \begin{pmatrix} \underline{1}_4 \\ \vdots \\ \underline{1}_4 \\ -\underline{1}_4 \\ \vdots \\ -\underline{1}_4 \end{pmatrix} \tau_1 + \alpha \begin{pmatrix} \underline{x}_{11} \\ \vdots \\ \underline{x}_{1r_1} \\ \underline{x}_{21} \\ \vdots \\ \underline{x}_{2r_2} \end{pmatrix}$$

$$+ \begin{pmatrix} \underline{1}_4 & \cdots & \underline{0}_4 & \underline{0}_4 & \cdots & \underline{0}_4 \\ \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ \underline{0}_4 & \cdots & \underline{1}_4 & \underline{0}_4 & \cdots & \underline{0}_4 \\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{1}_4 & \cdots & \underline{0}_4 \\ \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{0}_4 & \cdots & \underline{1}_4 \end{pmatrix} \begin{pmatrix} \mathbf{b_{11}} \\ \vdots \\ \mathbf{b_{1r_1}} \\ \mathbf{b_{21}} \\ \vdots \\ \mathbf{b_{2r_2}} \end{pmatrix} + \begin{pmatrix} \underline{\mathbf{e}}_{\mathbf{11}} \\ \vdots \\ \underline{\mathbf{e}}_{\mathbf{1r_1}} \\ \underline{\mathbf{e}}_{\mathbf{21}} \\ \vdots \\ \underline{\mathbf{e}}_{\mathbf{2r_2}} \end{pmatrix}.$$

Therefore,

$$\underline{\mathbf{y}} = \underline{1}_N \mu + \begin{pmatrix} \underline{1}_{(4r_1)} \\ -\underline{1}_{(4r_2)} \end{pmatrix} \tau_1 + \alpha \underline{x} + \mathcal{Z}_b \underline{\mathbf{b}} + \underline{\mathbf{e}}, \tag{12}$$

where

$\underline{x}(N \times 1) = (x_{111} \ldots x_{2r_24})^T$ is the vector of fixed weights for all the apple pieces,

$\underline{\mathbf{b}}(r \times 1) = (\mathbf{b_{11}}, \ldots, \mathbf{b_{2r_2}})^T \sim \mathcal{N}_r(\underline{0}, \sigma_b^2 \mathcal{J}_\mathbf{r})$ is the vector of random effects for all the apples,

$\underline{\mathbf{e}}(N \times 1) = (e_{111} \ldots e_{2r_24})^T \sim \mathcal{N}_N(\underline{0}, \sigma_e^2 \mathcal{I}_\mathbf{N})$ is the vector of residuals for all the apple pieces, and

$\mathcal{Z}_b(N \times r)$ is the same design matrix of regression coefficients for the random apple effects, seen for Model (8). It again corresponds to $\underline{\mathbf{b}}$.

Written more succinctly in standard notation, Model (12) is

$$\underline{\mathbf{y}} = \mathcal{X} \underline{\beta} + \mathcal{Z}_b \underline{\mathbf{b}} + \underline{\mathbf{e}}, \tag{13}$$

where

$$\mathcal{X}(N \times 3) = \begin{pmatrix} \underline{1}_{(4r_1)} & \underline{1}_{(4r_1)} & \underline{x}_{(4r_1)} \\ \underline{1}_{(4r_2)} & -\underline{1}_{(4r_2)} & \underline{x}_{(4r_2)} \end{pmatrix},$$

is the design matrix for the fixed effects, with

$\underline{x}_{(4r_t)}(4r_t \times 1)$, the vector of weights of all apple pieces that are treated with preservative t, where $t = 1, 2$.

The vector of fixed effect regression parameters is, $\underline{\beta}(3 \times 1) = \begin{pmatrix} \mu \\ \tau_1 \\ \alpha \end{pmatrix}$,

$\underline{\mathbf{b}}(r \times 1) \sim \mathcal{N}_r(\underline{0}, \sigma_b^2 \mathcal{J}_\mathbf{r})$, is the vector of random effects and

$\mathcal{Z}_b(N \times r)$ is the same design matrix for the random apple effects, which is shown for Model (8).

Finally, $\underline{\mathbf{e}}(N \times 1) \sim \mathcal{N}_N(\underline{0}, \sigma_e^2 \mathcal{I}_\mathbf{N})$ is the vector of residuals.

Therefore, $\underline{\mathbf{y}} \sim \mathcal{N}_N(\mathcal{X}\underline{\beta}, \mathbf{\Omega})$, with

$$\begin{aligned} E(\underline{\mathbf{y}}) &= \underline{1}_N \mu + \begin{pmatrix} \underline{1}_{r_1} \\ -\underline{1}_{r_2} \end{pmatrix} \tau_1 + \alpha \underline{x} \\ &= \mathcal{X}\underline{\beta}(N \times 1), \text{ since } E(\mathbf{e_{tij}}) = 0 \\ cov(\underline{\mathbf{y}}) &= \sigma_g^2 \mathcal{Z}_b \mathcal{Z}_b^T + \sigma_e^2 \mathcal{I}_\mathbf{N}, \text{ because } cov(\mathbf{b_{ti}}, \mathbf{e_{tij}}) = 0 \\ &= \mathbf{\Omega}(N \times N) \\ &= \begin{pmatrix} \mathbf{\Omega_1} & 0 & \cdots & 0 \\ 0 & \mathbf{\Omega_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{\Omega_r} \end{pmatrix}, \end{aligned}$$

where $\mathcal{Z}_b \mathcal{Z}_b^T (N \times N)$ is the same block diagonal matrix from Model (9), with (4×4) blocks of 1 on the diagonal, and zeros on the off-diagonal. It is used to calculate the elements, $\mathbf{\Omega_i}$, which are the per-apple covariance matrices defined above.

In the above models, even though we have other terms in the models, the covariance matrices have the same form as those in Section 4.3, where we illustrated the random effects models. This is because the additional terms in the model, τ_t and x_{tij}, are fixed effects and not random effects. This implies that the variance due to these effects is incorporated into the residual variance, so that σ_e^2 in Model (11) will differ from σ_e^2 in Model (7). On the other hand, if the additional terms had been random effects, the variance due to these effects would have been included in the covariance matrices of $\underline{\mathbf{y}}_\mathbf{ti}$ and $\underline{\mathbf{y}}$ as additional variance-components, in the same way as σ_b^2 was included in the covariance matrices in Section 4.3.

In Model (13), as in Model (9), the correlation between any pair of pieces from different

apples is given by $\rho(\mathbf{y_{tij}}, \mathbf{y_{tsk}}) = 0,$ for all $j, k, i \neq s$ and $t = 1, 2$. Therefore, pieces from different apples do not correlate with each other at all.

As pairs of pieces of the same apple have the same covariance (σ_b^2), we can think of each apple as being a 'cluster'. Thus, σ_b^2 represents the between-cluster source of variation, while σ_e^2 represents the within-cluster source of variation. These are the two sources of variation found in any mixed-effects model which has clusters. From the ICC we see that, if the variation between apples is large relative to the variation within an apple, then the correlation between pieces of the same apple will also be large because the within-apple variation is small. Hence, as the between-cluster variation increases, so will the within-cluster correlation.

In Section 4.4, we modeled fixed and random effects simultaneously. We saw that if we take a random-effects model and add fixed effects to it, we get a mixed-effects model where the structure and components of the covariance matrix are the same as those in the random effects model, but the matrix of fixed effects changes due to the addition of fixed effects. In the following section, we keep the same fixed effects structure as in Section 4.4, but add another random effect to the models, to observe the impact on the covariance matrices.

4.5 Mixed-effects model: Two fixed effects and two random effects

Let us now consider the same situation as in Section 4.4, where we cut our r apples into quarters and have the effect of two treatments ($\tau_t,$ for $t = 1, 2$) which are randomly assigned to r_1 treated and r_2 untreated apples. As before, let $\tau_1 = -\tau_2$ for identifiability, and let μ be the mean shelf-life of the untreated apples. There are thus $r = r_1 + r_2$ apples and $N = 4r$ apple pieces in total.

In addition, we again have the weight (grams) of each of the $4r$ apple pieces, denoted by (x_{tij}), with unknown coefficient α.

Suppose we now also consider that each quarter from the same apple is not exactly the same as the other three quarters because we cannot realistically cut an apple into exactly

equal halves, and thus exactly equal quarters. Let us assume that two quarters that come from the same half of an apple are more similar than two quarters that come from different halves of an apple. This implies that there is a random effect of position, and the impact of this random effect must be measured if we are to have accurate estimates, as it is another source of haphazard variation in our data.

Let $\mathbf{f_{ti}}$ be the random effect of the position of the j^{th} piece of the i^{th} apple, treated with t, such that $f = 1, 2, 3, 4$. Here, we can assume that pieces 1 and 2 come from the same half of an apple and pieces 3 and 4 come from the other half of the same apple. Under these assumptions, pieces 1 and 2 will be more correlated with each other, as will pieces 3 and 4. On the other hand, pieces 1 and 3, 2 and 4, and 2 and 3 will not be as highly correlated with each other since they come from different halves of an apple. We will assume here that the correlation between pieces 1 and 2, 2 and 3, and so on, are the same across all the apples. To account for this unknown correlation in the model, as well as in the covariance matrix of the model, we must assign coefficients to $\mathbf{f_{ti}}$.

Suppose that $j, k = 1, 2, 3, 4$ are the j^{th} and k^{th} pieces of an apple, and $z_{f_{ti}jk}$ is the regression coefficient corresponding to $\mathbf{f_{ti}}$, for pieces j and k in apple i.

Assume again that $\mathbf{y_{tij}}$ is the shelf-life of the j^{th} piece of the i^{th} apple, treated with t, so that $\underline{\mathbf{y}}_{\mathbf{ti}}(4 \times 1) = (\mathbf{y_{ti1}}, \ldots, \mathbf{y_{ti4}})^T$ is the vector of measurements of the shelf-life of pieces of the i^{th} apple, treated with t, and $\underline{\mathbf{y}}(N \times 1) = (\underline{\mathbf{y}}_{\mathbf{11}}, \ldots, \underline{\mathbf{y}}_{\mathbf{2r_2}})^T$ is the vector of measurements of the shelf-life of all apple pieces.

Let $\mathbf{e_{tij}}$ be the residual and $\mathbf{b_{ti}}$ be the random apple effect. Here, we again have $\mathbf{e_{tij}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$ and $\mathbf{b_{ti}} \sim$ i.i.d $\mathcal{N}(0, \sigma_b^2)$.

In addition, let $\mathbf{f_{ti}} \sim$ i.i.d $\mathcal{N}(0, \sigma_f^2)$, such that $\mathbf{e_{tij}}, \mathbf{b_{ti}}$ and $\mathbf{f_{ti}}$ are all independent of each other and each one is independent within an apple. We thus have

$$\mathbf{y_{tij}} = \mu + \tau_t + \alpha x_{tij} + \underline{z}_{f_{ti}j}^T \underline{\mathbf{f}}_{\mathbf{ti}} + \mathbf{b_{ti}} + \mathbf{e_{tij}}, \text{ for } t = 1, 2;\ i = 1, \ldots, r_t;\ j = 1, \ldots, 4, \quad (14)$$

where $\underline{z}_{f_{ti}j}^T(1 \times 4) = \{z_{f_{ti}jk}\}$ is the vector of regression coefficients for the j^{th} apple piece, corresponding to the random position effect, $\mathbf{f_{ti}}$, and

$\underline{\mathbf{f}}_{\mathbf{ti}}(4 \times 1) = (\mathbf{f_{ti}}, \ldots, \mathbf{f_{ti}})^T$ is the vector of the random effects of position of each apple piece

relative to each other piece in an apple.

For $\mathbf{y_{tij}}(1 \times 1)$,

$$\begin{aligned}
E(\mathbf{y_{tij}}) &= E(\mu + \tau_t + \alpha x_{tij} + \underline{z}_{f_{tij}}^T \underline{\mathbf{f}}_{\mathbf{ti}} + \mathbf{b_{ti}} + \mathbf{e_{tij}}) \\
&= \mu + \tau_t + \alpha x_{tij}, \text{ since } E(\mathbf{e_{tij}}) = 0 \\
var(\mathbf{y_{tij}}) &= E[(\underline{z}_{f_{tij}}^T \underline{\mathbf{f}}_{\mathbf{ti}} + \mathbf{b_{ti}} + \mathbf{e_{tij}})^2] \\
&= E[(\underline{z}_{f_{tij}}^T \underline{\mathbf{f}}_{\mathbf{ti}})^2] + E[\mathbf{b_{ti}}^2] + E[\mathbf{e_{tij}}^2] + 0, cov(\mathbf{b_{ti}}, \mathbf{f_{ti}}) = cov(\mathbf{b_{ti}}, \mathbf{e_{tij}}) = cov(\mathbf{f_{ti}}, \mathbf{e_{tij}}) = 0 \\
&= \sigma_f^2 + \sigma_b^2 + \sigma_e^2, \text{ since } z_{f_{tijk}} = 1 \text{ for all } j = k, t = 1, 2.
\end{aligned}$$

Thus, $\mathbf{y_{tij}} \sim \mathcal{N}(\mu + \tau_t + \alpha x_{tij}, \sigma_f^2 + \sigma_b^2 + \sigma_e^2)$. So we now have an additional variance component in the model. The effect on the covariance matrices of $\underline{\mathbf{y}}_{\mathbf{ti}}(4 \times 1)$ and $\underline{\mathbf{y}}(N \times 1)$ is shown below.

Expanding Model (14) for the i^{th} apple gives:

$$\begin{pmatrix} \mathbf{y_{ti1}} = \mu + \tau_{ti1} + \alpha x_{ti1} + z_{f_{ti}11}\mathbf{f_{ti}} + \cdots + z_{f_{ti}14}\mathbf{f_{ti}} + \mathbf{b_{ti}} + \mathbf{e_{ti1}} \\ \vdots \\ \mathbf{y_{ti4}} = \mu + \tau_{ti4} + \alpha x_{ti4} + z_{f_{ti}41}\mathbf{f_{ti}} + \cdots + z_{f_{ti}44}\mathbf{f_{ti}} + \mathbf{b_{ti}} + \mathbf{e_{ti4}} \end{pmatrix},$$

which implies

$$\begin{aligned}
\begin{pmatrix} \mathbf{y_{ti1}} \\ \mathbf{y_{ti2}} \\ \mathbf{y_{ti3}} \\ \mathbf{y_{ti4}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu &+ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tau_t + \alpha \begin{pmatrix} x_{ti1} \\ x_{ti2} \\ x_{ti3} \\ x_{ti4} \end{pmatrix} \\
&+ \begin{pmatrix} z_{f_{ti}11} & z_{f_{ti}12} & z_{f_{ti}13} & z_{f_{ti}14} \\ z_{f_{ti}21} & z_{f_{ti}22} & z_{f_{ti}23} & z_{f_{ti}24} \\ z_{f_{ti}31} & z_{f_{ti}32} & z_{f_{ti}33} & z_{f_{ti}34} \\ z_{f_{ti}41} & z_{f_{ti}42} & z_{f_{ti}43} & z_{f_{ti}44} \end{pmatrix} \begin{pmatrix} \mathbf{f_{ti}} \\ \mathbf{f_{ti}} \\ \mathbf{f_{ti}} \\ \mathbf{f_{ti}} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{b_{ti}} + \begin{pmatrix} \mathbf{e_{ti1}} \\ \mathbf{e_{ti2}} \\ \mathbf{e_{ti3}} \\ \mathbf{e_{ti4}} \end{pmatrix}.
\end{aligned}$$

Thus,

$$\underline{\mathbf{y}}_{\mathbf{ti}} = \underline{1}_4 \mu + \underline{1}_4 \tau_t + \alpha \underline{x}_{ti} + \mathcal{Z}_{f_{ti}} \underline{\mathbf{f}}_{\mathbf{ti}} + \underline{1}_4 \mathbf{b_{ti}} + \underline{\mathbf{e}}_{\mathbf{ti}}, \text{ for } t = 1, 2;\ i = 1, \ldots, r_t, \qquad (15)$$

where

$\underline{\mathbf{f}}_{\mathbf{ti}}(4 \times 1) = (\mathbf{f_{ti}}, \ldots, \mathbf{f_{ti}})^T$ is the vector of random position effects of each apple piece relative to the other pieces,

$\underline{\mathbf{e}}_{\mathbf{ti}}(4 \times 1) = (\mathbf{e_{ti1}}, \ldots, \mathbf{e_{ti4}})^T \sim \mathcal{N}_4(\underline{0}, \sigma_e^2 \mathcal{I}_4)$ is the within-apple residual vector,

$\mathbf{b_{ti}} \sim$ i.i.d $\mathcal{N}(0, \sigma_b^2)$ is the random apple effect, and

$\mathcal{Z}_{f_{ti}}(4 \times 4) = \{z_{f_{ti}jk}\}$ is the symmetric matrix of regression coefficients corresponding to the random position effect, $\mathbf{f_{ti}}$.

Since $E(\mathbf{e_{tij}}) = 0$ and $cov(\mathbf{b_{ti}}, \mathbf{f_{ti}}) = cov(\mathbf{b_{ti}}, \mathbf{e_{tij}}) = cov(\mathbf{f_{ti}}, \mathbf{e_{tij}}) = 0$,

$$\begin{aligned}
E(\underline{\mathbf{y}}_{\mathbf{ti}}) &= \underline{1}_4\mu + \underline{1}_4\tau_t + \alpha\underline{x}_{ti}(4 \times 1) \\
cov(\underline{\mathbf{y}}_{\mathbf{ti}}) &= E[(\mathcal{Z}_{f_{ti}}\underline{\mathbf{f}}_{\mathbf{ti}} + \underline{1}_4\mathbf{b_{ti}} + \underline{\mathbf{e}}_{\mathbf{ti}})(\mathcal{Z}_{f_{ti}}\underline{\mathbf{f}}_{\mathbf{ti}} + \underline{1}_4\mathbf{b_{ti}} + \underline{\mathbf{e}}_{\mathbf{ti}})^T] \\
&= \mathcal{Z}_{f_{ti}}E(\underline{\mathbf{f}}_{\mathbf{ti}}\underline{\mathbf{f}}_{\mathbf{ti}}{}^T)\mathcal{Z}_{f_{ti}}^T + \underline{1}_4E(\mathbf{b_{ti}}\mathbf{b_{ti}}{}^T)\underline{1}_4^T + E(\underline{\mathbf{e}}_{\mathbf{ti}}\underline{\mathbf{e}}_{\mathbf{ti}}{}^T), \\
&= \sigma_f^2\mathcal{Z}_{f_{ti}}\mathcal{Z}_{f_{ti}}^T + \sigma_b^2\mathcal{J}_{\mathbf{4}} + \sigma_e^2\mathcal{I}_{\mathbf{4}} \\
&= \mathbf{\Omega_i}(4 \times 4), \text{ say,}
\end{aligned}$$

where

$$\mathcal{Z}_{f_{ti}}\mathcal{Z}_{f_{ti}}^T(4 \times 4) = \begin{pmatrix} 1 & z^2_{f_{ti}12} & z^2_{f_{ti}13} & z^2_{f_{ti}14} \\ z^2_{f_{ti}21} & 1 & z^2_{f_{ti}23} & z^2_{f_{ti}24} \\ z^2_{f_{ti}31} & z^2_{f_{ti}32} & 1 & z^2_{f_{ti}34} \\ z^2_{f_{ti}41} & z^2_{f_{ti}42} & z^2_{f_{ti}43} & 1 \end{pmatrix}.$$

Now,

$$\begin{aligned}
cov(\mathbf{y_{tij}}, \mathbf{y_{tik}}) &= z^2_{f_{ti}jk}\sigma_f^2 + \sigma_b^2, \text{ for all } j \neq k, t = 1, 2, \text{ and} \\
\rho(\mathbf{y_{tij}}, \mathbf{y_{tik}}) &= \frac{z^2_{f_{ti}jk}\sigma_f^2 + \sigma_b^2}{\sigma_f^2 + \sigma_b^2 + \sigma_e^2}, \text{ for all } j \neq k, t = 1, 2,
\end{aligned}$$

so the covariance matrix of each apple, $\mathbf{\Omega_i}(4 \times 4)$, looks as follows:

$$\mathbf{\Omega_i} = \begin{pmatrix} \sigma_f^2 + \sigma_b^2 + \sigma_e^2 & z^2_{f_{ti}12}\sigma_f^2 + \sigma_b^2 & z^2_{f_{ti}13}\sigma_f^2 + \sigma_b^2 & z^2_{f_{ti}14}\sigma_f^2 + \sigma_b^2 \\ z^2_{f_{ti}21}\sigma_f^2 + \sigma_b^2 & \sigma_f^2 + \sigma_b^2 + \sigma_e^2 & z^2_{f_{ti}23}\sigma_f^2 + \sigma_b^2 & z^2_{f_{ti}24}\sigma_f^2 + \sigma_b^2 \\ z^2_{f_{ti}31}\sigma_f^2 + \sigma_b^2 & z^2_{f_{ti}32}\sigma_f^2 + \sigma_b^2 & \sigma_f^2 + \sigma_b^2 + \sigma_e^2 & z^2_{f_{ti}34}\sigma_f^2 + \sigma_b^2 \\ z^2_{f_{ti}41}\sigma_f^2 + \sigma_b^2 & z^2_{f_{ti}42}\sigma_f^2 + \sigma_b^2 & z^2_{f_{ti}43}\sigma_f^2 + \sigma_b^2 & \sigma_f^2 + \sigma_b^2 + \sigma_e^2 \end{pmatrix}.$$

While it also appears to have a compound symmetry structure, this is not the case here because the correlation between different pieces of an apple will differ according to the position of the pieces.

Thus, although $\mathbf{\Omega_i}$ here is symmetric, it differs from $\mathbf{\Omega_i}$ in Sections 4.3 and 4.4, in that:

1. we now have three components of variance, whereas there were only two variance-components previously; and

2. the new component here, σ_f^2, has a coefficient which describes the relationship between two apple pieces, depending on where on the apple they are positioned relative to each other. This was not the case in those previous sections.

So here, $\underline{\mathbf{y}}_{\mathbf{i}} \sim \mathcal{N}_4(\underline{1}_4\mu + \underline{1}_4\tau_t + \alpha\underline{x}_{ti}, \mathbf{\Omega_i})$, where $\mathbf{\Omega_i}$ has the structure shown above.

Expanding Model (15) for $\underline{\mathbf{y}}(N \times 1)$, the vector of shelf lives for all the apples in the sample, gives:

$$\begin{pmatrix} \underline{\mathbf{y}}_{\mathbf{11}} \\ \vdots \\ \underline{\mathbf{y}}_{\mathbf{1r_1}} \\ \underline{\mathbf{y}}_{\mathbf{21}} \\ \vdots \\ \underline{\mathbf{y}}_{\mathbf{2r_2}} \end{pmatrix} = \begin{pmatrix} \underline{1}_4 \\ \vdots \\ \underline{1}_4 \\ \underline{1}_4 \\ \vdots \\ \underline{1}_4 \end{pmatrix} \mu + \begin{pmatrix} \underline{1}_4 \\ \vdots \\ \underline{1}_4 \\ -\underline{1}_4 \\ \vdots \\ -\underline{1}_4 \end{pmatrix} \tau_1 + \alpha \begin{pmatrix} \underline{x}_{11} \\ \vdots \\ \underline{x}_{1r_1} \\ \underline{x}_{21} \\ \vdots \\ \underline{x}_{2r_2} \end{pmatrix}$$

$$+ \begin{pmatrix} \mathcal{Z}_{f_{11}} & \cdots & \mathcal{O}_{\mathbf{4}} & \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{O}_{\mathbf{4}} \\ \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{Z}_{f_{1r_1}} & \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{O}_{\mathbf{4}} \\ \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{O}_{\mathbf{4}} & \mathcal{Z}_{f_{21}} & \cdots & \mathcal{O}_{\mathbf{4}} \\ \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{O}_{\mathbf{4}} & \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{Z}_{f_{2r_2}} \end{pmatrix} \begin{pmatrix} \underline{\mathbf{f}}_{\mathbf{11}} \\ \vdots \\ \underline{\mathbf{f}}_{\mathbf{1r_1}} \\ \underline{\mathbf{f}}_{\mathbf{21}} \\ \vdots \\ \underline{\mathbf{f}}_{\mathbf{2r_2}} \end{pmatrix} +$$

$$+ \begin{pmatrix} \underline{1}_4 & \cdots & \underline{0}_4 & \underline{0}_4 & \cdots & \underline{0}_4 \\ \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ \underline{0}_4 & \cdots & \underline{1}_4 & \underline{0}_4 & \cdots & \underline{0}_4 \\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{1}_4 & \cdots & \underline{0}_4 \\ \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{0}_4 & \cdots & \underline{1}_4 \end{pmatrix} \begin{pmatrix} \mathbf{b_{11}} \\ \vdots \\ \mathbf{b_{1r_1}} \\ \mathbf{b_{21}} \\ \vdots \\ \mathbf{b_{2r_2}} \end{pmatrix} + \begin{pmatrix} \underline{\mathbf{e}}_{\mathbf{11}} \\ \vdots \\ \underline{\mathbf{e}}_{\mathbf{1r_1}} \\ \underline{\mathbf{e}}_{\mathbf{21}} \\ \vdots \\ \underline{\mathbf{e}}_{\mathbf{2r_2}} \end{pmatrix}.$$

Therefore,

$$\underline{\mathbf{y}} = \underline{1}_N\mu + \begin{pmatrix} \underline{1}_{(4r_1)} \\ -\underline{1}_{(4r_2)} \end{pmatrix} \tau_1 + \alpha\underline{x} + \mathcal{Z}_f\underline{\mathbf{f}} + \mathcal{Z}_b\underline{\mathbf{b}} + \underline{\mathbf{e}}, \tag{16}$$

where

$\underline{x}(N \times 1) = (x_{111} \ldots x_{2r_24})^T$ is the vector of fixed weights for all the apple pieces,

$\underline{\mathbf{f}}(N \times 1) = (\mathbf{f_{11}}, \ldots, \mathbf{f_{2r_2}})^T \sim \mathcal{N}_N(\underline{0}, \sigma_f^2\mathcal{Z}_f\mathcal{Z}_f^T)$ is the vector of random effect of position for all the apple pieces, with $\mathcal{Z}_f\mathcal{Z}_f^T(N \times N) = \{z_{f_{ti}jk}^2\}$

$\underline{\mathbf{b}}(r \times 1) = (\mathbf{b_{11}}, \ldots, \mathbf{b_{2r_2}})^T \sim \mathcal{N}_r(\underline{0}, \sigma_b^2\mathcal{J}_\mathbf{r})$ is the vector of random effects for all r apples, and

$\underline{\mathbf{e}}(N \times 1) = (e_{111} \dots e_{2r_24})^T \sim \mathcal{N}_N(\underline{0}, \sigma_e^2 \mathcal{I}_{\mathbf{N}})$ is the overall vector of residuals.

In addition,

$$\mathcal{Z}_f(N \times N) = \begin{pmatrix} \mathcal{Z}_{f_{11}} & \cdots & \mathcal{O}_{\mathbf{4}} & \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{O}_{\mathbf{4}} \\ \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{Z}_{f_{1r_1}} & \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{O}_{\mathbf{4}} \\ \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{O}_{\mathbf{4}} & \mathcal{Z}_{f_{21}} & \cdots & \mathcal{O}_{\mathbf{4}} \\ \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{O}_{\mathbf{4}} & \mathcal{O}_{\mathbf{4}} & \cdots & \mathcal{Z}_{f_{2r_2}} \end{pmatrix}$$

and

$$\mathcal{Z}_b(N \times r) = \begin{pmatrix} \underline{1}_4 & \cdots & \underline{0}_4 & \underline{0}_4 & \cdots & \underline{0}_4 \\ \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ \underline{0}_4 & \cdots & \underline{1}_4 & \underline{0}_4 & \cdots & \underline{0}_4 \\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{1}_4 & \cdots & \underline{0}_4 \\ \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{0}_4 & \cdots & \underline{1}_4 \end{pmatrix}.$$

Writing Model (16), in the more concise standard matrix notation gives:

$$\underline{\mathbf{y}} = \mathcal{X}\underline{\beta} + \mathcal{Z}\underline{\mathbf{b}} + \underline{\mathbf{e}}, \tag{17}$$

where

$\mathcal{X}(N \times 3)$ is the design matrix for the fixed effects and is the same as the corresponding design matrix for Model (13); namely

$$\mathcal{X}(N \times 3) = \begin{pmatrix} \underline{1}_{(4r_1)} & \underline{1}_{(4r_1)} & \underline{x}_{(4r_1)} \\ \underline{1}_{(4r_2)} & -\underline{1}_{(4r_2)} & \underline{x}_{(4r_2)} \end{pmatrix},$$

where

$\underline{x}_{(4r_t)}(4r_t \times 1)$ is the vector of weights of all apple pieces that are treated with preservative t, $t = 1, 2$,

$\underline{\beta}(3 \times 1) = (\mu, \tau_1, \alpha)^T$ is the vector of fixed effects,

$\mathcal{Z}(N \times (N + r))$ is the design matrix for the random position and apple effects, and it has the following structure:

$$\mathcal{Z} = \begin{pmatrix} \mathcal{Z}_{f_{11}} & \underline{1}_4 & \cdots & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \cdots & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 \\ \vdots & \cdots & \ddots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \cdots & \mathcal{Z}_{f_{1r_1}} & \underline{1}_4 & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \cdots & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 \\ \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \cdots & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \mathcal{Z}_{f_{21}} & \underline{1}_4 & \cdots & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \cdots & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \mathcal{O}_{\mathbf{4}} & \underline{0}_4 & \cdots & \mathcal{Z}_{f_{2r_2}} & \underline{1}_4 \end{pmatrix},$$

where

$\mathcal{Z}_{f_{ti}}(4 \times 4) = \{z_{f_{tijk}}\}$ is the symmetric matrix of regression coefficients corresponding to the random position effect, $\mathbf{f_{ti}}$.

Next,

$$\underline{\mathbf{b}}((N+r) \times 1) = \begin{pmatrix} \underline{\mathbf{f}}_{\mathbf{11}} \\ \mathbf{b_{11}} \\ \vdots \\ \underline{\mathbf{f}}_{\mathbf{1r_1}} \\ \mathbf{b_{1r_1}} \\ \underline{\mathbf{f}}_{\mathbf{21}} \\ \mathbf{b_{21}} \\ \vdots \\ \underline{\mathbf{f}}_{\mathbf{2r_2}} \\ \mathbf{b_{2r_2}} \end{pmatrix} \sim \mathcal{N}_{(N+r)}(\underline{0}, \sigma_f^2 \mathcal{Z}_f \mathcal{Z}_f^T + \sigma_b^2 \mathcal{Z}_b \mathcal{Z}_b^T)$$

is the vector of random effects, with $\mathcal{Z}_f \mathcal{Z}_f^T(N \times N) = \{z^2_{f_{tijk}}\}$, and $\mathcal{Z}_b \mathcal{Z}_b^T(N \times N)$, the block diagonal matrix shown for Model (9), with (4×4) blocks of 1 on the diagonal and zeros on the off-diagonal.

Finally, $\underline{\mathbf{e}}(N \times 1) \sim \mathcal{N}_N(\underline{0}, \sigma_e^2 \mathcal{I}_{\mathbf{N}})$ is the vector of residuals.

Therefore,

$$\begin{aligned}
E(\underline{\mathbf{y}}) &= \mathcal{X}\underline{\beta}(N \times 1), \text{ since } E(\mathbf{e_{tij}}) = 0 \\
cov(\underline{\mathbf{y}}) &= E[(\mathcal{Z}_f \underline{\mathbf{f}} + \mathcal{Z}_b \underline{\mathbf{b}} + \underline{\mathbf{e}})(\mathcal{Z}_f \underline{\mathbf{f}} + \mathcal{Z}_b \underline{\mathbf{b}} + \underline{\mathbf{e}})^T] \\
&= \mathcal{Z}_f E(\underline{\mathbf{f}}\underline{\mathbf{f}}^T)\mathcal{Z}_f^T + \mathcal{Z}_b E(\underline{\mathbf{b}}\underline{\mathbf{b}}^T)\mathcal{Z}_b^T + E(\underline{\mathbf{e}}\underline{\mathbf{e}}^T), \text{ as } cov(\mathbf{b_{ti}}, \mathbf{f_{ti}}) = cov(\mathbf{b_{ti}}, \mathbf{e_{tij}}) = cov(\mathbf{f_{ti}}, \mathbf{e_{tij}}) = 0 \\
&= \sigma_f^2 \mathcal{Z}_f \mathcal{Z}_f^T + \sigma_b^2 \mathcal{Z}_b \mathcal{Z}_b^T + \sigma_e^2 \mathcal{I}_{\mathbf{N}} \\
&= \mathbf{\Omega}(N \times N), \text{ say,}
\end{aligned}$$

where $\mathcal{Z}_f \mathcal{Z}_f^T(N \times N) = \{z^2_{f_{tijk}}\}$, and $\mathcal{Z}_b \mathcal{Z}_b^T(N \times N)$ is described above.

Then, $\underline{\mathbf{y}} \sim \mathcal{N}_N(\mathcal{X}\underline{\beta}, \mathbf{\Omega})$, where the block diagonal covariance matrix of $\underline{\mathbf{y}}$ is

$$\mathbf{\Omega}(N \times N) = \begin{pmatrix} \mathbf{\Omega_1} & 0 & \cdots & 0 \\ 0 & \mathbf{\Omega_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{\Omega_r} \end{pmatrix},$$

and the elements $\mathbf{\Omega_i}$ are defined above.

Between Section 4.5 and Section 4.4, we have seen that adding fixed effects to a model does not alter the form of the covariance matrix. In contrast, adding random effects does change the covariance matrix, by including in it additional variance components which represent the variation due to these new effects. As a result, these additional variance-components affect the covariances of observations between pieces of the same apple, and thus the correlation between these pieces. As we saw for Model (15), the correlation coefficient between any two pieces of the same apple, for the models above, is $\rho(\mathbf{y_{tij}}, \mathbf{y_{tik}}) = \frac{z^2_{f_{tij}jk}\sigma^2_f + \sigma^2_b}{\sigma^2_f + \sigma^2_b + \sigma^2_e}$, for all $j \neq k, t = 1, 2$. Here, there is an additional component in both the numerator and denominator, whereas in Section 4.4, the correlation contained only the between-apple component of variance and the residual variance. In addition, the new component, σ^2_f, has a coefficient which contributes to the correlation. Through this coefficient, the correlation between two pieces of the same apple accounts for the position of those pieces relative to the each other. Finally, $cov(\mathbf{y_{tij}}, \mathbf{y_{tsk}}) = 0$ and $\rho(\mathbf{y_{tij}}, \mathbf{y_{tsk}}) = 0$, for all $j, k, i \neq s$ and $t = 1, 2$, so two pieces from different apples do not correlate with each other at all.

Here, the estimated value of the residual variance, σ^2_e also differs from the estimate in Section 4.4, as there is an additional random effect in the models here. The variation due to this new effect adds an additional variance-component, σ^2_f, to the total variance. It thus accounts for some of the variation in the data. Therefore, the total variance now is split between the three components σ^2_e, σ^2_b and σ^2_f, where σ^2_e accounts for the residual variance, as well as the variance due to all the fixed effects in the models.

Dividing up random variation and accounting for all sources of variation in dataset, as demonstrated in the previous sections, is the crux of variance-components methodology. It is necessary since valid statistical inferences about means and differences between means can only be made when the corresponding variances are correctly specified and estimated.

The models shown in Section 4.5 are among the most basic types of mixed-effects models we find. In the last section (4.6), we expand these models to include d covariates (fixed

effects) since, in practice, it is seldom the case that there are only one or two covariates that need to be considered.

Adding covariates to models changes the matrix of fixed effects but does not affect the structure of the covariance matrix, as we have seen. This is fortunate as our previous covariance matrix will not become more complex. However, each additional fixed effect will affect the estimate of σ_e^2, as the variation due to these fixed effects is accounted for in σ_e^2.

4.6 Mixed-effects model: Several fixed effects and two random effects

The number of fixed and random effects presented in this section have been chosen specifically to correspond to those presented in Chapter 7, where the results of the Heartdata analysis are presented. The motive behind this is to reinforce and implement the theory presented here practically, using real-world data. As such, this example is tailored to fit the Heartdata, although we are still considering the statistical theory in terms of apples, as it is better understood in this context first.

The primary objective of the Heartdata is to develop appropriate models and test the variance-components in order to establish heritability and linkage. As such, we are not interested in testing (and do not have) any treatment effects for this data. Due to this, the models in this section exclude the treatment effects that we considered in Sections 4.4 and 4.5, making the models here a little simpler.

Suppose we have $d = 10$ covariates (predictors), whose effects on the apples we must take into consideration in our models.

Assume again that we have r apples, each cut into four pieces, so that $N = \sum_i^r 4 = 4r$ apple pieces in total.

Let $x_{1,ij}, \ldots, x_{10,ij}$ denote ten covariates of interest and let $\alpha_1, \ldots, \alpha_{10}$ be the unknown coefficients corresponding to each of these $x_{d,ij}$.

Assume also that we have two random effects:

$\mathbf{b_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_b^2)$, the random apple effect, and

$\mathbf{f_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_f^2)$, the random effect which measures the impact of the position of apple piece j in apple i, relative to the other three pieces. It has regression coefficient z_{f_ijk}.

Also, $\mathbf{e_{ij}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$ is the random residual for the j^{th} piece of the i^{th} apple, such that $\mathbf{e_{ij}}, \mathbf{b_i}$ and $\mathbf{f_i}$ are all independent of each other between apples, and each one is also independent within an apple.

If $\mathbf{y_{ij}}(1 \times 1)$ is the shelf-life of the j^{th} piece of the i^{th} apple and μ is the overall mean shelf-life of the apple pieces, then $\underline{\mathbf{y}}_\mathbf{i}(4 \times 1) = (\mathbf{y_{i1}}, \ldots, \mathbf{y_{i4}})^T$ is the vector of measurements of the shelf-life of pieces of the i^{th} apple, and $\underline{\mathbf{y}}(N \times 1) = (\underline{\mathbf{y}}_\mathbf{1}, \ldots, \underline{\mathbf{y}}_\mathbf{r})^T$ is the vector of measurements of the shelf-life of all apple pieces. A model for this data, for each apple piece, is

$$\mathbf{y_{ij}} = \mu + \underline{x}_{d,ij}^T \underline{\alpha} + \underline{z}_{f_ij}^T \mathbf{f_i} + \mathbf{b_i} + \mathbf{e_{ij}}, \text{ for } i = 1, \ldots, r;\ j = 1, \ldots, 4;\ d = 1, \ldots, 10, \quad (18)$$

where

$$\begin{aligned} \underline{z}_{f_ij}^T(1 \times 4) &= \{z_{f_ijk}\}, \text{ and} \\ \underline{x}_{d,ij}^T \underline{\alpha}(1 \times 1) &= \begin{pmatrix} x_{1,ij} & x_{2,ij} & \cdots & x_{10,ij} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{10} \end{pmatrix} \\ &= \alpha_1 x_{1,ij} + \alpha_2 x_{2,ij} + \cdots + \alpha_{10} x_{10,ij}. \end{aligned}$$

Therefore,

$$\begin{aligned} E(\mathbf{y_{ij}}) &= \mu + \underline{x}_{d,ij}^T \underline{\alpha}, \text{ since } E(\mathbf{e_{ij}}) = 0 \\ &= \mu + \alpha_1 x_{1,ij} + \alpha_2 x_{2,ij} + \cdots + \alpha_{10} x_{10,ij} \\ var(\mathbf{y_{ij}}) &= E[(\underline{z}_{f_ij}^T \underline{\mathbf{f}}_\mathbf{i})^2] + E[\mathbf{b_i}^2] + E[\mathbf{e_{ij}}^2] + 0, \text{ because } cov(\mathbf{b_i}, \mathbf{e_{ij}}) = cov(\mathbf{b_i}, \mathbf{f_i}) = cov(\mathbf{e_{ij}}, \mathbf{f_i}) = 0 \\ &= \sigma_f^2 + \sigma_b^2 + \sigma_e^2, \text{ since } z_{f_ijk} = 1 \text{ for all } j = k. \end{aligned}$$

Thus, $\mathbf{y_{ij}} \sim \mathcal{N}(\mu + \alpha_1 x_{1,ij} + \alpha_2 x_{2,ij} + \cdots + \alpha_{10} x_{10,ij}, \sigma_f^2 + \sigma_b^2 + \sigma_e^2)$. So unlike in Section 4.5, we now have some additional terms in the mean expression and no treatment effect terms, while the variance contains the same three components as it did there.

If we expand Model (18) for the i^{th} apple, we get:

$$\begin{pmatrix} \mathbf{y_{i1}} = \mu + \alpha_1 x_{1,i1} + \cdots + \alpha_{10} x_{10,i1} + z_{f_i 11}\mathbf{f_i} + \cdots + z_{f_i 14}\mathbf{f_i} + \mathbf{b_i} + \mathbf{e_{i1}} \\ \vdots \\ \mathbf{y_{i4}} = \mu + \alpha_1 x_{1,i4} + \cdots + \alpha_{10} x_{10,i4} + z_{f_i 41}\mathbf{f_i} + \cdots + z_{f_i 44}\mathbf{f_i} + \mathbf{b_i} + \mathbf{e_{i4}} \end{pmatrix},$$

which gives

$$\begin{pmatrix} \mathbf{y_{i1}} \\ \mathbf{y_{i2}} \\ \mathbf{y_{i3}} \\ \mathbf{y_{i4}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} x_{1,i1} & x_{2,i1} & \cdots & x_{10,i1} \\ x_{1,i2} & x_{2,i2} & \cdots & x_{10,i2} \\ x_{1,i3} & x_{2,i3} & \cdots & x_{10,i3} \\ x_{1,i4} & x_{2,i4} & \cdots & x_{10,i4} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{10} \end{pmatrix} + \begin{pmatrix} z_{f_i 11} & z_{f_i 12} & z_{f_i 13} & z_{f_i 14} \\ z_{f_i 21} & z_{f_i 22} & z_{f_i 23} & z_{f_i 24} \\ z_{f_i 31} & z_{f_i 32} & z_{f_i 33} & z_{f_i 34} \\ z_{f_i 41} & z_{f_i 42} & z_{f_i 43} & z_{f_i 44} \end{pmatrix} \begin{pmatrix} \mathbf{f_i} \\ \mathbf{f_i} \\ \mathbf{f_i} \\ \mathbf{f_i} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{b_i} + \begin{pmatrix} \mathbf{e_{i1}} \\ \mathbf{e_{i2}} \\ \mathbf{e_{i3}} \\ \mathbf{e_{i4}} \end{pmatrix}.$$

Therefore,

$$\underline{\mathbf{y}}_{\mathbf{i}} = \underline{1}_4 \mu + \mathcal{X}_{d_i} \underline{\alpha} + \mathcal{Z}_{f_i} \underline{\mathbf{f}}_{\mathbf{i}} + \underline{1}_4 \mathbf{b_i} + \underline{\mathbf{e}}_{\mathbf{i}}, \tag{19}$$

for $i = 1, \ldots, r, d = 1, \ldots, 10$, where

$\mathcal{Z}_{f_i}(4 \times 4) = \{z_{f_i jk}\}$ is the symmetric matrix of regression coefficients for the random position effect, $\mathbf{f_i}$,

$\mathcal{X}_{d_i}(4 \times 10) = \{x_{d,ij}\}$ is the matrix of covariates for the i^{th} apple,

$\underline{\alpha}(10 \times 4) = (\alpha_1, \ldots, \alpha_{10})$ is the vector of regression coefficients corresponding to $\mathcal{X}_{d_i}$,

$\mathbf{b_i}(1 \times 1) \sim$ i.i.d $\mathcal{N}(0, \sigma_b^2)$ is the random apple effect, and

$\underline{\mathbf{e}}_{\mathbf{i}}(4 \times 1) = (\mathbf{e_{i1}}, \ldots, \mathbf{e_{i4}})^T \sim \mathcal{N}_4(\underline{0}, \sigma_e^2 \mathcal{I}_4)$ is the within-apple residual vector.

With this, we can calculate the mean, variance and covariances of $\underline{\mathbf{y}}_{\mathbf{i}}(4 \text{ x } 1)$:

$$\begin{aligned} E(\underline{\mathbf{y}}_{\mathbf{i}}) &= \underline{1}_4 \mu + \mathcal{X}_{d_i} \underline{\alpha}(4 \times 1), \text{ since } E(\mathbf{e_{ij}}) = 0 \\ &= \begin{pmatrix} \mu + \alpha_1 x_{1,i1} + \cdots + \alpha_{10} x_{10,i1} \\ \vdots \\ \mu + \alpha_1 x_{1,i4} + \cdots + \alpha_{10} x_{10,i4} \end{pmatrix} \end{aligned}$$

$$\begin{aligned} cov(\underline{\mathbf{y}}_{\mathbf{i}}) &= \sigma_f^2 \mathcal{Z}_{f_i} \mathcal{Z}_{f_i}^T + \sigma_b^2 \mathcal{J}_4 + \sigma_e^2 \mathcal{I}_4, \text{ since } cov(\mathbf{b_i}, \mathbf{e_{ij}}) = cov(\mathbf{b_i}, \mathbf{f_i}) = cov(\mathbf{e_{ij}}, \mathbf{f_i}) = 0 \\ &= \mathbf{\Omega_i}(4 \times 4), \text{ say,} \end{aligned}$$

with

$$\mathcal{Z}_{f_i}\mathcal{Z}_{f_i}^T(4\times 4)=\begin{pmatrix} 1 & z^2_{f_i12} & z^2_{f_i13} & z^2_{f_i14}\\ z^2_{f_i21} & 1 & z^2_{f_i23} & z^2_{f_i24}\\ z^2_{f_i31} & z^2_{f_i32} & 1 & z^2_{f_i34}\\ z^2_{f_i41} & z^2_{f_i42} & z^2_{f_i43} & 1 \end{pmatrix}.$$

Therefore,

$$\begin{aligned} cov(\mathbf{y_{ij}},\mathbf{y_{ik}}) &= z^2_{f_ijk}\sigma^2_f+\sigma^2_b, \text{ for all } j\neq k, \text{ and}\\ \rho(\mathbf{y_{ij}},\mathbf{y_{ik}}) &= \frac{z^2_{f_ijk}\sigma^2_f+\sigma^2_b}{\sigma^2_f+\sigma^2_b+\sigma^2_e}, \text{ for all } j\neq k. \end{aligned}$$

So, even though the coefficient of σ^2_f is defined differently here, this covariance matrix has the same structure as the corresponding one in Section 4.5, namely

$$\mathbf{\Omega_i}=\begin{cases} \sigma^2_f+\sigma^2_b+\sigma^2_e & \text{for } j=k\\ z^2_{f_ijk}\sigma^2_f+\sigma^2_b & \text{for } j\neq k, \end{cases}$$

Therefore, $\underline{\mathbf{y}}_\mathbf{i}\sim\mathcal{N}_4(\underline{1}_4\mu+\mathcal{X}_{d_i}\underline{\alpha},\mathbf{\Omega_i})$.

We see here that $\mathbf{\Omega_i}$ has not changed from Models (15) to (19). However, $E(\underline{\mathbf{y}}_\mathbf{i})$ has changed in that its elements differ between the two models. They change according to the changes in the fixed effects of the models. Therefore, adding covariates to, or removing them from, a model affects only the mean because the covariates are fixed effects. As a result, expanding Model (19) for the full set of r apples will not alter the structure of the covariance matrix, $\mathbf{\Omega}$, from the one in Section 4.5, but the additional fixed effects will affect the mean of the overall model $\underline{\mathbf{y}}(N\times 1)$.

So, expanding Model (19) gives:

$$\begin{aligned}\begin{pmatrix}\underline{\mathbf{y}}_\mathbf{1}\\ \underline{\mathbf{y}}_\mathbf{2}\\ \vdots\\ \underline{\mathbf{y}}_\mathbf{r}\end{pmatrix}&=\begin{pmatrix}\underline{1}_4\\ \underline{1}_4\\ \vdots\\ \underline{1}_4\end{pmatrix}\mu+\ +\ \begin{pmatrix}\mathcal{X}_{d_1}\\ \mathcal{X}_{d_2}\\ \vdots\\ \mathcal{X}_{d_r}\end{pmatrix}\begin{pmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_{10}\end{pmatrix}+\begin{pmatrix}\mathcal{Z}_{f_1} & \mathcal{O}_\mathbf{N} & \cdots & \mathcal{O}_\mathbf{N}\\ \mathcal{O}_\mathbf{N} & \mathcal{Z}_{f_2} & \cdots & \vdots\\ \vdots & \vdots & \ddots & \mathcal{O}_\mathbf{N}\\ \mathcal{O}_\mathbf{N} & \cdots & \mathcal{O}_\mathbf{N} & \mathcal{Z}_{f_r}\end{pmatrix}\begin{pmatrix}\underline{\mathbf{f}}_\mathbf{1}\\ \underline{\mathbf{f}}_\mathbf{2}\\ \vdots\\ \underline{\mathbf{f}}_\mathbf{r}\end{pmatrix}\\ &\quad+\ \begin{pmatrix}\underline{1}_4\\ \underline{1}_4\\ \vdots\\ \underline{1}_4\end{pmatrix}\mu+\begin{pmatrix}\underline{1}_4 & \underline{0}_4 & \cdots & \underline{0}_4\\ \underline{0}_4 & \underline{1}_4 & \cdots & \vdots\\ \vdots & \vdots & \ddots & \underline{0}_4\\ \underline{0}_4 & \cdots & \underline{0}_4 & \underline{1}_4\end{pmatrix}\begin{pmatrix}\mathbf{b_1}\\ \mathbf{b_2}\\ \vdots\\ \mathbf{b_r}\end{pmatrix}+\begin{pmatrix}\underline{\mathbf{e}}_\mathbf{1}\\ \underline{\mathbf{e}}_\mathbf{2}\\ \vdots\\ \underline{\mathbf{e}}_\mathbf{r}\end{pmatrix}.\end{aligned}$$

This is equivalent to

$$\underline{\mathbf{y}} = \underline{1}_N\mu + \mathcal{X}_d\underline{\alpha} + \mathcal{Z}_f\underline{\mathbf{f}} + \mathcal{Z}_b\underline{\mathbf{b}} + \underline{\mathbf{e}}, \tag{20}$$

where

$\mathcal{X}_d(N \times 10) = \{x_{d,ij}\}$ is the matrix of covariates,

$\underline{\alpha}(10 \times 1) = (\alpha_1, \ldots, \alpha_{10})$ is the vector of regression coefficients corresponding to $\mathcal{X}_d$,

$\underline{\mathbf{f}}(N \times 1) = (\mathbf{f_1}, \ldots, \mathbf{f_r})^T \sim \mathcal{N}_N(\underline{0}, \sigma_f^2\mathcal{Z}_f\mathcal{Z}_f^T)$ is the vector of random effect of position for all the apple pieces, with $\mathcal{Z}_f\mathcal{Z}_f^T(N \times N) = \{z_{f_ijk}^2\}$,

$\underline{\mathbf{b}}(r \times 1) = (\mathbf{b_1}, \ldots, \mathbf{b_r})^T \sim \mathcal{N}_r(\underline{0}, \sigma_b^2\mathcal{J}_\mathbf{r})$ is the vector of random effects for all r apples, and

$\underline{\mathbf{e}}(N \times 1) = (e_{11} \ldots e_{r4})^T \sim \mathcal{N}_N(\underline{0}, \sigma_e^2\mathcal{I}_\mathbf{N})$ is the overall vector of residuals. In addition, $\mathcal{Z}_f(N \text{ x } N)$ and $\mathcal{Z}_b(N \text{ x } r)$ are the matrices of regression coefficients corresponding to $\underline{\mathbf{f}}$ and $\underline{\mathbf{b}}$ respectively.

Model (20) can be written more compactly in standard matrix notation as

$$\underline{\mathbf{y}} = \mathcal{X}\underline{\beta} + \mathcal{Z}\underline{\mathbf{b}} + \underline{\mathbf{e}}, \tag{21}$$

where

$$\mathcal{X}(N \times (d+1) = N \times 11) = \begin{pmatrix} \underline{1}_4 & \mathcal{X}_{d_1} \\ \underline{1}_4 & \mathcal{X}_{d_2} \\ \vdots & \vdots \\ \underline{1}_4 & \mathcal{X}_{d_r} \end{pmatrix}$$

is the design matrix for the fixed effects and differs from that in Model (17) because the fixed effects differ between the two models.

We also have

$$\underline{\beta}((d+1) \times 1 = 11 \times 1) = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{10} \end{pmatrix},$$

the vector of fixed effects corresponding to $\mathcal{X}$, and

$\underline{\mathbf{e}}(N \times 1) \sim \mathcal{N}_N(\underline{0}, \sigma_e^2 \mathcal{I}_{\mathbf{N}})$, the vector of residuals.

$$\mathcal{Z}(N \times (N+r)) = \begin{pmatrix} \mathcal{Z}_{f_1} & \underline{1}_4 & \mathcal{O}_{\mathbf{N}} & \underline{0}_4 \cdots & \mathcal{O}_{\mathbf{N}} & \underline{0}_4 \\ \mathcal{O}_{\mathbf{N}} & \underline{0}_4 & \mathcal{Z}_{f_2} & \underline{1}_4 \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \ddots & \mathcal{O}_{\mathbf{N}} & \underline{0}_4 \\ \mathcal{O}_{\mathbf{N}} & \underline{0}_4 & \mathcal{O}_{\mathbf{N}} & \underline{0}_4 \cdots & \mathcal{Z}_{f_r} & \underline{1}_4 \end{pmatrix}.$$

Corresponding to $\mathcal{Z}$ is $\underline{\mathbf{b}}((N+r) \times 1)$, the vector of random effects. It differs from $\underline{\mathbf{b}}$ in Section 4.5, and looks as follows:

$$\underline{\mathbf{b}} = \begin{pmatrix} \underline{\mathbf{f_1}} \\ \mathbf{b_1} \\ \underline{\mathbf{f_2}} \\ \mathbf{b_2} \\ \vdots \\ \underline{\mathbf{f_r}} \\ \mathbf{b_r} \end{pmatrix}.$$

Given Models (20) and (21),

$$\begin{aligned} E(\underline{\mathbf{y}}) &= \underline{1}_N \mu + \mathcal{X}_d \underline{\alpha}, \text{ since } E(\mathbf{e_{ij}}) = 0 \\ &= \mathcal{X} \underline{\beta} (N \times 1) \\ cov(\underline{\mathbf{y}}) &= \sigma_f^2 \mathcal{Z}_f \mathcal{Z}_f^T + \sigma_b^2 \mathcal{Z}_b \mathcal{Z}_b^T + \sigma_e^2 \mathcal{I}_{\mathbf{N}}, \text{ because } \mathbf{b_i}, \mathbf{f_i}, \mathbf{e_{ij}} \sim i.i.d. \\ &= \mathbf{\Omega}(N \times N), \text{ say,} \end{aligned}$$

where

$\mathcal{Z}_f \mathcal{Z}_f^T (N \times N) = \{z_{f_{ijk}}^2\}$, and

$$\mathcal{Z}_b \mathcal{Z}_b^T (N \times N) = \begin{pmatrix} \mathcal{J_4} & \mathcal{O_4} & \cdots & \mathcal{O_4} \\ \mathcal{O_4} & \mathcal{J_4} & \cdots & \mathcal{O_4} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{O_4} & \mathcal{O_4} & \cdots & \mathcal{J_4} \end{pmatrix}$$

is the block diagonal matrix of 1's, seen in Sections 4.4 and 4.5.

Then, $\underline{\mathbf{y}} \sim \mathcal{N}_N(\mathcal{X}\underline{\beta}, \mathbf{\Omega})$, where the block diagonal covariance matrix of $\underline{\mathbf{y}}$ is

$$\mathbf{\Omega}(N \times N) = \begin{pmatrix} \mathbf{\Omega_1} & 0 & \cdots & 0 \\ 0 & \mathbf{\Omega_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{\Omega_r} \end{pmatrix},$$

and the elements $\mathbf{\Omega_i}$ are defined above. We thus have that $cov(\mathbf{y_{ij}}, \mathbf{y_{sk}}) = 0$ and $\rho(\mathbf{y_{ij}}, \mathbf{y_{sk}}) = 0$, for all j, k and $i \neq s$.

In this section, we have seen how taking out some fixed effects and adding many to a model still does not alter the form of the covariance matrix and thus the correlations between pieces of the same apple. However, the variance estimates of Model (21) will differ from the variance estimates of Model (17) in Section 4.5, even though the forms of the covariance matrices, $\mathbf{\Omega}$, are the same. This is because adding fixed effects reduces the estimate of the residual variance since the variance due to these fixed effects is taken out of the residual variance.

Next in this study we will leave behind the apples and show how the mixed-effects model methodology which has been developed here, applies to the analysis of family-based genetic studies, where the models are even more complicated due to family members being genetically related. In particular, we show how the covariance matrix, $\mathbf{\Omega}$, developed in Sections 4.5 and 4.6 gets more complicated when we have to consider the degree to which family members are related. In the covariance matrix above, we had a coefficient for the random effect of apple-piece position but assumed that in all the apples, the correlation between the same two pieces, for example pieces 3 and 4, was the same. In family genetic studies, such assumptions cannot be made when considering the correlation between family pairs with regard to some trait. This is because the amount of genetic material at a particular genetic locus, which is shared by, for example sibling pairs, will differ from one family to another.

5 Statistical Genetics

> We are far too apt to regard common events as matters of course, and to accept many things as obvious truths which are not obvious truths at all, but present problems of much interest... Why is it when we compare two groups of persons selected at random from the same race, but belonging to different generations of it, we find them to be closely alike? Such statistical differences as there may be, are always to be ascribed to differences in the general conditions of their lives ... The processes of heredity are found to be so wonderfully balanced and their equilibrium to be so stable, that they concur in maintaining a perfect statistical resemblance so long as the external conditions remain unaltered.
>
> If there be any who are inclined to say there is no wonder in the matter, because each individual tends to leave his like behind him, and therefore each generation must resemble the one preceding, I can assure them that they are utterly mistaken. Individuals do *not* equally tend to leave their like behind them ... The question, then, is this:– How is it that although each individual does *not* as a rule leave his like behind him, yet successive generations resemble each other with great exactitude in all their general features? (Galton, 1877:492)

Studies seeking to statistically explain heredity and the relationship between family members can be traced back as far as the 1800s with Sir Francis Galton's explanation of the laws of heredity using the law of deviation (Galton, 1877). In that lecture, he discussed the resemblance of and differences between each generation of sons and their fathers. He explained how traits such as height are inherited, yet the variation in successive generations causes a reversion toward 'mediocrity'. With crude equipment and the relatively little knowledge about genetics in existence in those times, Galton could not trace and identify the specific cause of the observed inheritance. In the 133 years since he gave that first lecture, the study of statistical genetics has leapt forward thanks to technological and computational advancements which have improved the understanding of genetics and made complex statistical analysis more viable.

In October 1990, the Human Genome Project was founded with the aim to ultimately map the entire human genome and thus enable the identification of alleles responsible for the multitude of human diseases found in the world. The Project's primary aims were to gain a better understanding of the role of genetic factors in complex diseases and to gain new insights into human evolution (Palmer (2005), Watson (1990)). Thanks to their efforts and the work of others, much more is now understood about the inheritance of alleles in families. Scientists use this knowledge, combined with advances in statistics, to trace inheritance patterns of both common and rare diseases. Using these patterns, they then try to identify the exact allele(s) responsible for specific diseases; the ultimate aim of any genetic study.

Family studies are carried out because if a trait runs in families, then somewhere on some chromosome, lies a locus where the allele(s) responsible for the trait can be found. The fact that family members share a common living environment and are related (and thus more similar genetically than random unrelated individuals) aids researchers in identifying these alleles.

5.1 Familial aggregation

One of the first steps in a potentially genetic study is called *familial aggregation* and it is based on the tendency for genetic traits, usually diseases, to cluster in families. The aim is to determine whether or not a trait runs in families, without possessing any genetic information for the participants in the study. This is done by observing if aggregation occurs more often than is expected by chance. If the disease does appear to aggregate in families, the next step is to determine if it could be inherited. If it does not appear to aggregate in families, it is not likely to have a strong genetic component. Therefore this first step is very important as it determines the direction of future research.

In analyses, each family unit is treated as a specialised cluster and possible environmental influences (such as age, physical characteristics or even exposure to some toxins) are used

to study the effects on disease risk. Without information on environmental influences, aggregation caused by shared alleles is indistinguishable from aggregation due to shared environment. As a result, statistical models assessing familial aggregation usually combine these two sources of variation and model them as the within-family variance.

For a dichotomous trait such as disease/no disease, aggregation can be assessed via odds ratios and risk ratios, but for quantitative traits, familial aggregation is usually assessed using a correlation or covariance-based measure. Since family pairs are not independent, if the trait is assumed to follow a multivariate normal distribution, then the extent of familial aggregation can be measured by estimating intraclass correlation coefficients. These correlations can be obtained from standard linear regression models using ANOVA methods, as shown in Chapter 4, but we now define the variables as follows:

Consider a set of r extended families, each with n_i members. The total number of family members $N = \sum_i^r n_i$. Let $\mathbf{y_{ij}}$ represent the random quantitative trait value for individual j from family i. Then $\underline{\mathbf{y}}_\mathbf{i}(n_i \times 1) = (\mathbf{y_{i1}}, \ldots, \mathbf{y_{in_i}})^T$ is the random vector of trait values for family i. The overall vector of trait values is given by $\underline{\mathbf{y}}(N \times 1) = (\underline{\mathbf{y}}_\mathbf{1}, \ldots, \underline{\mathbf{y}}_\mathbf{r})^T$. Let $\mathbf{b_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_b^2)$ denote the random effect corresponding to the trait deviation of family i from the overall mean μ, and let $\mathbf{e_{ij}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$ be the trait deviation of individual j from the family mean μ_i. Assume $\mathbf{b_i}$ and $\mathbf{e_{ij}}$ are mutually independent, so $cov(\mathbf{b_i}, \mathbf{e_{ij}}) = 0$ for individual j, and $cov(\mathbf{b_i}, \mathbf{b_s}) = 0, cov(\mathbf{e_{ij}}, \mathbf{e_{ik}}) = 0$ for all $i \neq s, j \neq k$, respectively.

Based on this notation, a model of this data for the j^{th} person in the i^{th} family is:

$$\mathbf{y_{ij}} = \mu + \mathbf{b_i} + \mathbf{e_{ij}}, \text{ for } i = 1, \ldots, r;\ j = 1, \ldots, n_i, \tag{22}$$

where

$$\begin{aligned} E(\mathbf{y_{ij}}) &= \mu \\ var(\mathbf{y_{ij}}) &= \sigma_b^2 + \sigma_e^2. \end{aligned}$$

Therefore, $\mathbf{y_{ij}} \sim \mathcal{N}(\mu, \sigma_b^2 + \sigma_e^2)$.

Of the variance-components, the former, σ_b^2, is the random between-family component of variance and it represents the variation of the 'true' family means about the population mean. The latter, σ_e^2, is the within-family component and it is the variance of the individuals in a family about the mean of that family μ_i.

Model (22) is analogous to Model (6), the random effects model, where the between- and within-apple components of variance were discussed. Now however, we can think of our families as the apples and the individuals in the families as being the pieces of the apples. Therefore, just as we defined a random effect for each apple and we had to account for the variation between the apples, we now have a random family effect and we have to account for the differences between families. We must also account for the difference between individuals in each family, or rather, account for the similarity within each family. As explained before, the more similar family members are, the greater the difference between families will be, relative to the difference within families. We can see this if we write Model (22) for family i in matrix notation:

$$\underline{\mathbf{y}}_{\mathbf{i}} = \underline{1}_{n_i}\mu + \underline{1}_{n_i}\mathbf{b_i} + \underline{\mathbf{e}}_{\mathbf{i}}, \text{ for } i = 1, \ldots, r, \tag{23}$$

where $\underline{\mathbf{e}}_{\mathbf{i}}(n_i \times 1) = (\mathbf{e_{i1}}, \ldots, \mathbf{e_{in_i}})^T \sim \mathcal{N}_{n_i}(\underline{0}, \sigma_e^2 \mathcal{I}_{\mathbf{n_i}})$ is the within-family residual vector.

From this we see that

$$\begin{aligned} E(\underline{\mathbf{y}}_{\mathbf{i}}) &= \underline{1}_{n_i}\mu(n_i \times 1) \\ cov(\underline{\mathbf{y}}_{\mathbf{i}}) &= \sigma_b^2 \mathcal{J}_{\mathbf{n_i}} + \sigma_e^2 \mathcal{I}_{\mathbf{n_i}}, \text{ since } cov(\mathbf{b_i}, \mathbf{e_{ij}}) = 0. \\ &= \mathbf{\Omega_i}(n_i \times n_i). \end{aligned}$$

Specifically, we have

$$\begin{aligned} cov(\mathbf{y_{ij}}, \mathbf{y_{ik}}) &= \sigma_b^2, \text{ for all } j \neq k \\ cov(\mathbf{y_{ij}}, \mathbf{y_{sk}}) &= 0, \text{ for all } j, k \text{ and } i \neq s. \end{aligned}$$

Therefore, $\underline{\mathbf{y}}_{\mathbf{i}} \sim \mathcal{N}_{n_i}(\underline{1}_{n_i}\mu, \mathbf{\Omega_i})$ and

$$\rho(\mathbf{y_{ij}}, \mathbf{y_{ik}}) = \frac{cov(\mathbf{y_{ij}}, \mathbf{y_{ik}})}{\sqrt{var(\mathbf{y_{ij}})var(\mathbf{y_{ik}})}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}, \text{ for all } j \neq k.$$

This is just the intraclass correlation coefficient we saw previously. In the case of families, it is known as the intra-family correlation (Burton et al., 2005). It tells us that, in the extreme case when $\sigma_b^2 = 0, \rho = 0$. This implies that there is no variation in trait values between families and thus no correlation between pairs of family members because all the families are the same. On the other hand, when σ_b^2 is much larger than σ_e^2, the correlation coefficient will also be large as trait values of family members are more similar to each other than to members of other families.

This intra-family correlation coefficient is estimated from a model which assumes that the covariance between each family pair is the same, and it does not account for how closely related a family pair are. Unlike in the case where the apples were cut into equal numbers of pieces, family sizes differ and family pairs have varying degrees of relation with each other.

Incorporating the degree of relationship between pairs of family members (and thus their correlations) will be discussed further in the following sections. However, we must first introduce an important measure of family relatedness.

5.2 Kinship

Through σ_b^2, we have demonstrated how the similarity between family members can be accounted for in a statistical analysis. However, we have still not accounted for the degree to which family members are related. As emphasised previously, this is important because the more closely related two people are, the more similar they are genetically. Family relatedness is accounted for in an analysis through a measure known as the *kinship coefficient*, denoted by φ. It is a function of the degree of the relationship between two people. The more closely related these individuals are, the larger φ will be. Therefore, in

family i, φ_{ijk} accounts for the genetic variation that exists between two family members, j and k, based solely on how closely related they are.

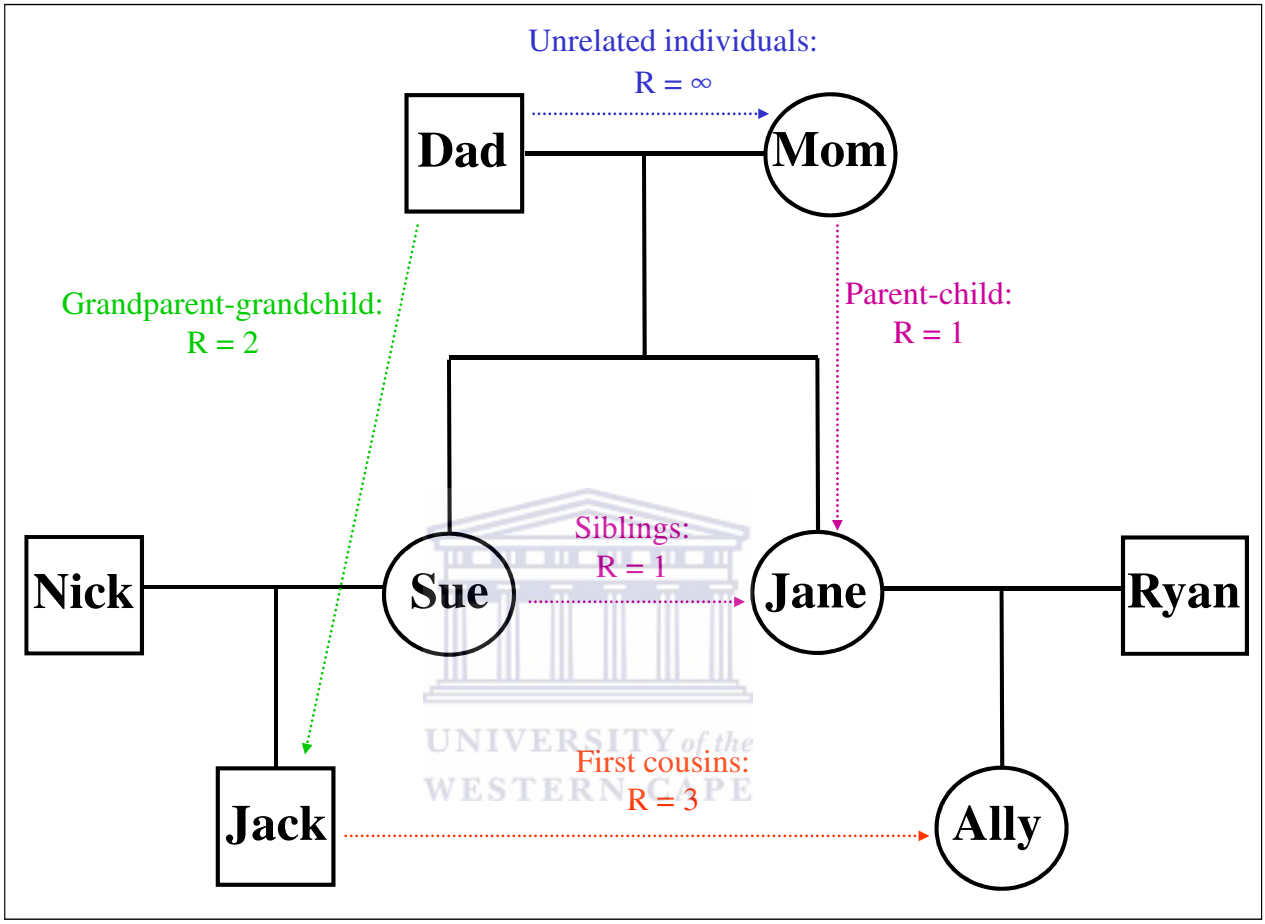


Figure 12: Pedigree depicting degrees of relationship

The degree of relationship between individuals j and k can be described as the number of connections between them in a pedigree. We can calculate this by counting the number of nodes crossed on the path between the two individuals, including person j's node. Figure 12 is of an extended pedigree depicting the different degrees of relationship for different family pairs. We have taken the nuclear pedigree from Figure 3 and extended it to include more family members. The extended pedigree consists of eight members: Dad and Mom are the parents of Sue and Jane, as before; Nick is married to Sue and they have one child, Jack; Ryan is married to Jane and their daughter is Ally. We assume that Nick and Ryan are only related to the rest of the family through their marriages. At the top of the pedigree are the parents, who are unrelated and thus have degree of relation $R = \infty$.

They have two daughters who are first degree relatives with each other and with their parents ($R = 1$). We see this since, the only node between say, Mom and Jane, is Jane's. Similarly, the only node counted between Sue and Jane, is Jane's. When the daughters Sue and Jane have their own children, we see that there are now grandparent-grandchild relationships. These are second degree relationships ($R = 2$), as seen with Dad and Jack where two nodes are counted (Sue's and Jack's). First cousins are third degree relatives, as seen with Jack and Ally where we count the nodes of Sue, Jane and Ally. Continuing in this way, we can identify the degree of relationship between any two individuals in a pedigree. This allows us to compute the corresponding kinship coefficient, φ, which is calculated mathematically as follows:

Let R = degree of relationship between two people. Then, $(\frac{1}{2})^R$ is called the *coefficient of relationship* and can be described as the expected proportion of shared alleles. The kinship coefficient is defined as $\varphi = (\frac{1}{2})^{R+1}$. In analysis, we use $2\varphi = (\frac{1}{2})^R$. A formal definition of kinship will be given in Section 6.2.

For various pairs of individuals, the degree of relationship and the corresponding kinship coefficient are shown in Table 2. These values are used to calculate the covariances

Table 2: Degree of relation (R) and kinship coefficients (φ) for various family pairs

Relative types	**R**	2φ
Monozygotic (identical) twins	0	1
Parent–offspring	1	1/2
Dizygotic (non-identical) twins/Sib–pairs	1	1/2
Half-sibs	2	1/4
Grandparent–grandchild	2	1/4
Uncle/Aunt–niece/nephew	2	1/4
First cousins	3	1/8
Unrelated individuals	∞	0

between pairs of individuals in a pedigree study. This is done through a matrix of kinships coefficients for each family and will be shown in more detail in the sections that follow.

5.3 Segregation analysis

> The partitioning of the variance into its components allows us to estimate the relative importance of the various determinants of the phenotype, in particular the role of heredity versus environment ... The question of 'relative importance' can be answered only if it is expressed in terms of the variance attributable to the different sources of variation. The relative importance of a source of variation is the variance due to that source, as a proportion of the total phenotypic variance. The relative importance of heredity in determining phenotypic values is called the *heritability* of the character (Falconer, 1989:125–126).

Quantitative genetics aims to identify the genetic factors that account for the trait variance observed in quantitative traits, as well as to identify the extent to which these genetic factors contribute to trait variance. This is done through a measure known as *heritability*, which measures the genetic contribution (of all inherited alleles) to trait variability. In general, heritability is about the *cause of variation* in a particular trait (Burton et al., 2005). Therefore, determining heritability is one of the first objectives when studying such traits. Heritability is estimated through segregation analysis, which is the process of fitting genetic models to trait data from family members. The aim of segregation analysis is to find a model that best explains the pattern of familial aggregation observed, by testing to see if alleles are involved in disease aggregation.

Falconer (1989) questions how intrinsically discontinuous variation, which is caused by genetic segregation, translates into the continuous variation of quantitative traits. They give two reasons: the first is that the numerous alleles affecting the trait segregate simultaneously; the second is the superimposition of truly continuous variation which is caused by non-genetic factors. This results in the distinction between alleles which are concerned with Mendelian inheritance (single-genes) and those that result in quantitative traits (many minor genes). The difference between the two, according to Falconer (1989:105), " ... lies in the magnitude of their effects relative to other sources of variation". Single-genes have a large effect. Therefore, they cause a recognisable discontinuity,

even when there is segregation at other loci and non-genetic variation is present. These may thus be studied by Mendelian methods. However, there are also the minor genes whose effects are not large enough to cause a distinct discontinuity. As a result, these genes cannot be studied individually. They result in variation that is caused by the simultaneous segregation of many genes. Traits such as obesity and predisposition to certain diseases, are believed to be caused by the inheritance of minor genes.

Segregation analysis involves developing models that contain parameters quantifying the degree of influence on the trait of both a single-gene locus and loci containing minor genes. In particular, it aims to determine if alleles segregate randomly in families. To assess whether this occurs in a quantitative trait, a hypothesis test can be carried out where the null hypothesis tested is that there is no hereditary variation. In other words, all the variation in trait values between family members can be attributed to environmental factors. Such tests are carried out as a first step into the study of a potentially heritable trait.

Consider again our set of r extended families, each with n_i members, where the total number of family members is $N = \sum_i^r n_i$. Let $\mathbf{y_{ij}}$ represent the random quantitative trait value for individual j from family i, and let $\underline{\mathbf{y}}_{\mathbf{i}}(n_i \times 1) = (\mathbf{y_{i1}}, \ldots, \mathbf{y_{in_i}})^T$, be the random vector of trait values for family i. The overall vector of trait values is then $\underline{\mathbf{y}}(N \times 1) = (\underline{\mathbf{y}}_{\mathbf{1}}, \ldots, \underline{\mathbf{y}}_{\mathbf{r}})^T$.

We want to decompose the between-family variance, σ_b^2, into two components: the heritable genetic variance (σ_g^2), which we will call hereditary variance; and a non-heritable shared-environment variance (σ_f^2). As before, let σ_e^2 denote the residual variance seen previously. Therefore, σ_e^2 is the residual variance after accounting for the heritable and shared environment portions of variation.

Let $\mathbf{g_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_g^2)$ denote the hereditary random effect for family i. It measures the impact of inherited genetic material on the trait.

Let $\mathbf{f_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_f^2)$ be the environmental effect of family i. It measures the impact,

on trait values, of sharing a common environment.

Let $\mathbf{e_{ij}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$ be the environmental or residual effect for individual j from family i.

Assume $\mathbf{g_i}, \mathbf{f_i}$ and $\mathbf{e_{ij}}$ are independent of each other for different families and independent within families as well.

In order to extract the heritable component of variance from the between-family variance, we need to include family relatedness in the segregation model. This is done through specifying a vector of regression coefficients for $\mathbf{g_i}$; in particular the kinship coefficients between each family pair. Specifying kinships as the coefficients of σ_g^2 defines σ_g^2 as hereditary genetic variance and separates it from other between-family variation.

Suppose μ is the overall fixed mean trait value. Thus, a model for this data, for the j^{th} person in the i^{th} family, is

$$\mathbf{y_{ij}} = \mu + \underline{z}_{g_ij}^T \underline{\mathbf{g}}_\mathbf{i} + \mathbf{f_i} + \mathbf{e_{ij}}, \text{ for } i = 1, \ldots, r;\ j = 1, \ldots, n_i, \tag{24}$$

where $\underline{z}_{g_ij}^T (1 \times n_i) = \{z_{g_ijk} = \sqrt{2\varphi_{ijk}}\}$ is the vector of regression coefficients for individual j, corresponding to the hereditary random effect for family i, and φ_{ijk} is the kinship coefficient between individuals j and k in family i. Including the kinship coefficient in the model allows us to account for the degree of relation between each family pair, rather than assuming they all have the same degree of relation. This allows us to model a different coefficient for each family pair, which is similar to what was done in Section 4.5 where we specified a coefficient representing the relationship between two apple pieces, depending on their position in the apple.

From this we get

$$\begin{aligned} E(\mathbf{y_{ij}}) &= \mu \\ var(\mathbf{y_{ij}}) &= \sigma_g^2 + \sigma_f^2 + \sigma_e^2, \text{ as } cov(\mathbf{g_i}, \mathbf{f_i}) = cov(\mathbf{f_i}, \mathbf{e_{ij}}) = cov(\mathbf{g_i}, \mathbf{e_{ij}}) = 0. \end{aligned}$$

Since we are interested mostly in the heritable genetic effects, rather than the shared-environment effects, $\mathbf{f_i}$ is usually omitted from models and its effects are therefore com-

bined with the within-family effects $\mathbf{e_{ij}}$. Thus, Model (24) becomes, for the j^{th} individual in the i^{th} family,

$$\mathbf{y_{ij}} = \mu + \underline{z}_{g_{ij}}^T \underline{\mathbf{g}}_\mathbf{i} + \mathbf{e_{ij}}, \text{ for } i = 1, \ldots, r;\ j = 1, \ldots, n_i. \tag{25}$$

For individual j, $\mathbf{y_{ij}}$ is normally distributed with the following mean and variance:

$$\begin{aligned} E(\mathbf{y_{ij}}) &= \mu, \text{ since } E(\mathbf{e_{ij}}) = 0 \\ var(\mathbf{y_{ij}}) &= \sigma_g^2 + \sigma_e^2, \text{ since } cov(\mathbf{g_i}, \mathbf{e_{ij}}) = 0. \text{ and } 2\varphi_{ijk} = 1 \text{ for all } j = k. \end{aligned}$$

Expanding Model (25) for the i^{th} family gives

$$\underline{\mathbf{y}}_\mathbf{i} = \underline{1}_{n_i}\mu + \mathcal{Z}_{g_i}\underline{\mathbf{g}}_\mathbf{i} + \underline{\mathbf{e}}_\mathbf{i}, \text{ for } i = 1, \ldots, r, \tag{26}$$

where $\mathcal{Z}_{g_i}(n_i \times n_i) = \{\sqrt{2\varphi_{ijk}}\}$ is the symmetric matrix of regression coefficients for the hereditary random effects, $\mathbf{g_i}$.

Therefore,

$$\begin{aligned} E(\underline{\mathbf{y}}_\mathbf{i}) &= \underline{1}_{n_i}\mu(n_i \times 1) \\ cov(\underline{\mathbf{y}}_\mathbf{i}) &= \sigma_g^2 \mathcal{Z}_{g_i}\mathcal{Z}_{g_i}^T + \sigma_e^2 \mathcal{I}_{\mathbf{n_i}}, \text{ because } cov(\mathbf{g_i}, \mathbf{e_{ij}}) = 0 \\ &= \mathbf{\Omega_i}(n_i \times n_i), \text{ say,} \end{aligned}$$

Now,

$$\mathcal{Z}_{g_i}\mathcal{Z}_{g_i}^T(n_i \times n_i) = \begin{pmatrix} 1 & 2\varphi_{i12} & \cdots & 2\varphi_{i1n_i} \\ 2\varphi_{i21} & 1 & \cdots & 2\varphi_{i2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ 2\varphi_{in_i1} & 2\varphi_{in_i2} & \cdots & 1 \end{pmatrix},$$

since $2\varphi_{ijk} = 1$ for all $j = k$. As a result, for segregation analysis, the covariance between individuals j and k in family i is given by

$$\Omega_{ijk} = cov(\mathbf{y_{ij}}, \mathbf{y_{ik}}) = \begin{cases} \sigma_g^2 + \sigma_e^2 & \text{if } j = k \\ 2\varphi_{ijk}\sigma_g^2 & \text{if } j \neq k, \end{cases}$$

where φ_{ijk} is the kinship coefficient between individuals j and k.

Thus the covariance matrix is

$$\mathbf{\Omega_i}(n_i \times n_i) = \begin{pmatrix} \sigma_g^2 + \sigma_e^2 & 2\varphi_{i12}\sigma_g^2 & \cdots & 2\varphi_{i1n_i}\sigma_g^2 \\ 2\varphi_{i21}\sigma_g^2 & \sigma_g^2 + \sigma_e^2 & \cdots & 2\varphi_{i2n_i}\sigma_g^2 \\ \vdots & \vdots & \ddots & \vdots \\ 2\varphi_{in_i1}\sigma_g^2 & 2\varphi_{in_i2}\sigma_g^2 & \cdots & \sigma_g^2 + \sigma_e^2 \end{pmatrix},$$

Given this, for family i the vector of trait values $\underline{\mathbf{y}}_\mathbf{i} \sim \mathcal{N}_{n_i}(\underline{1}_{n_i}\mu, \mathbf{\Omega_i})$.

When analysing familial aggregation, the trait correlation between any relative pair is just the intrafamily correlation coefficient. For the segregation Model (26), the covariance between members of two different families is zero, so their correlation is zero. The correlation between pairs of relatives differs for different family pairs, depending on their relationship, as shown by:

$$\rho(\mathbf{y_{ij}}, \mathbf{y_{ik}}) = \frac{cov(\mathbf{y_{ij}}, \mathbf{y_{ik}})}{\sqrt{var(\mathbf{y_{ij}})var(\mathbf{y_{ik}})}} = \frac{2\varphi_{ijk}\sigma_g^2}{\sigma_g^2 + \sigma_e^2}, \text{ for all } j \neq k.$$

This allows us to separate the heritable genetic variance, σ_g^2, from the residual variance, σ_e^2. We can thus estimate the heritable genetic variance by weighting it according to the degree of relation between the family pair. This is illustrated in the following example.

Suppose that family i in our set of r families is the example family from Figure 3, consisting of Dad, Mom, Sue and Jane. Their covariance matrix will be

$$\mathbf{\Omega_i} = \begin{pmatrix} \sigma_g^2 + \sigma_e^2 & 2\varphi_{Dad,Mom}\sigma_g^2 & 2\varphi_{Dad,Sue}\sigma_g^2 & 2\varphi_{Dad,Jane}\sigma_g^2 \\ 2\varphi_{Dad,Mom}\sigma_g^2 & \sigma_g^2 + \sigma_e^2 & 2\varphi_{Mom,Sue}\sigma_g^2 & 2\varphi_{Mom,Jane} \\ 2\varphi_{Dad,Sue}\sigma_g^2 & 2\varphi_{Mom,Sue}\sigma_g^2 & \sigma_g^2 + \sigma_e^2 & 2\varphi_{Sue,Jane}\sigma_g^2 \\ 2\varphi_{Dad,Jane}\sigma_g^2 & 2\varphi_{Mom,Jane}\sigma_g^2 & 2\varphi_{Sue,Jane}\sigma_g^2 & \sigma_g^2 + \sigma_e^2 \end{pmatrix}.$$

Since $\varphi = \frac{1}{4}$ for all parent-offspring and sib-pair relationships, and it is zero for the Dad-Mom relationship, this matrix becomes

$$\mathbf{\Omega_i} = \begin{pmatrix} \sigma_g^2 + \sigma_e^2 & 0 & \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_g^2 \\ 0 & \sigma_g^2 + \sigma_e^2 & \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_g^2 \\ \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_g^2 & \sigma_g^2 + \sigma_e^2 & \frac{1}{2}\sigma_g^2 \\ \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_g^2 & \sigma_g^2 + \sigma_e^2 \end{pmatrix}.$$

Since we will use this family for further demonstrations, we will shorten the names of the family members to their first initial and use these in the matrices. Therefore, Dad, Mom,

Sue and Jane will be represented by D, M, S and J, respectively. In addition, since the covariance matrices are all square and symmetric, only the upper triangle will be written out in detail.

Even for the simple nuclear family we have used, the covariance matrix does not have a structure which can be modeled using standard statistical software. This is also true when we extend the matrix to include other family members. To demonstrate this, suppose we extend our example family of Dad, Mom, Sue and Jane, to include Jane's husband, Ryan (R), and their daughter, Ally (A). The covariance matrix for the family, including Ryan and Ally, will now be:

$$\mathbf{\Omega_i} = \begin{pmatrix} \sigma_g^2+\sigma_e^2 & 2\varphi_{D,M}\sigma_g^2 & 2\varphi_{D,S}\sigma_g^2 & 2\varphi_{D,J}\sigma_g^2 & 2\varphi_{D,R}\sigma_g^2 & 2\varphi_{D,A}\sigma_g^2 \\ & \sigma_g^2+\sigma_e^2 & 2\varphi_{M,S}\sigma_g^2 & 2\varphi_{M,J}\sigma_g^2 & 2\varphi_{M,R}\sigma_g^2 & 2\varphi_{M,A}\sigma_g^2 \\ & & \sigma_g^2+\sigma_e^2 & 2\varphi_{S,J}\sigma_g^2 & 2\varphi_{S,R}\sigma_g^2 & 2\varphi_{S,A}\sigma_g^2 \\ & & & \sigma_g^2+\sigma_e^2 & 2\varphi_{J,R}\sigma_g^2 & 2\varphi_{J,A}\sigma_g^2 \\ & & & & \sigma_g^2+\sigma_e^2 & 2\varphi_{R,A}\sigma_g^2 \\ & & & & & \sigma_g^2+\sigma_e^2 \end{pmatrix}.$$

Substituting in the corresponding kinship coefficients from Table 2, gives:

$$\mathbf{\Omega_i} = \begin{pmatrix} \sigma_g^2+\sigma_e^2 & 0 & \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_g^2 & 0 & \frac{1}{4}\sigma_g^2 \\ & \sigma_g^2+\sigma_e^2 & \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_g^2 & 0 & \frac{1}{4}\sigma_g^2 \\ & & \sigma_g^2+\sigma_e^2 & \frac{1}{2}\sigma_g^2 & 0 & \frac{1}{4}\sigma_g^2 \\ & & & \sigma_g^2+\sigma_e^2 & 0 & \frac{1}{2}\sigma_g^2 \\ & & & & \sigma_g^2+\sigma_e^2 & \frac{1}{2}\sigma_g^2 \\ & & & & & \sigma_g^2+\sigma_e^2 \end{pmatrix}.$$

Once an extended family is considered, the covariance matrix for the family no longer has a recognisable structure. Here, the more closely related a pair is, the bigger the covariance between them. Similarly, the less closely related they are, the smaller the covariance between them is, until it reaches zero for unrelated individuals.

We can now expand Model (26) to include all the families in the study group, in the same way as we did with the apples in Chapter 4. This gives

$$\underline{\mathbf{y}} = \underline{1}_N\mu + \mathcal{Z}_g\underline{\mathbf{g}} + \underline{\mathbf{e}}, \tag{27}$$

where $\underline{\mathbf{e}}(N\times 1) = (\mathbf{e_{11}},\ldots,\mathbf{e_{rn_r}})^T \sim \mathcal{N}_N(\underline{0}, \sigma_e^2\mathcal{I}_\mathbf{N})$ is the vector of environmental effects, $\underline{\mathbf{g}}(N\times 1) = (\mathbf{g_1},\ldots,\mathbf{g_r})^T \sim \mathcal{N}_N(\underline{0}, \sigma_g^2\mathcal{Z}_g\mathcal{Z}_g^T)$ is the vector of all random hereditary effects,

and

$\mathcal{Z}_g(N \times N) = \{\sqrt{2\varphi_{ijk}}\}$ is the block diagonal symmetric matrix of regression coefficients for the hereditary random effects, $\mathbf{g_i}$, for all r families.

From this we get,

$$\begin{aligned} E(\underline{\mathbf{y}}) &= \underline{1}_N \mu(N \times 1) \\ cov(\underline{\mathbf{y}}) &= \sigma_g^2 \mathcal{Z}_g \mathcal{Z}_g^T + \sigma_e^2 \mathcal{I}_{\mathbf{N}}, \text{ since } cov(\mathbf{g_i}, \mathbf{e_{ij}}) = 0. \\ &= \mathbf{\Omega}(N \times N), \text{ say,} \end{aligned}$$

where $\mathcal{Z}_g \mathcal{Z}_g^T(N \times N)$ is a block diagonal matrix such that

$$\begin{aligned} \mathcal{Z}_g \mathcal{Z}_g^T &= \begin{pmatrix} 2\varphi_{111} & \cdots & 2\varphi_{11n_1} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 2\varphi_{1n_11} & \cdots & 2\varphi_{1n_1n_1} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 2\varphi_{r11} & \cdots & 2\varphi_{r1n_r} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 2\varphi_{rn_r1} & \cdots & 2\varphi_{rn_rn_r} \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{Z}_{g_1}^2 & \mathcal{O}_{\mathbf{N}} & \cdots & \mathcal{O}_{\mathbf{N}} \\ \mathcal{O}_{\mathbf{N}} & \mathcal{Z}_{g_2}^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathcal{O}_{\mathbf{N}} \\ \mathcal{O}_{\mathbf{N}} & \cdots & \mathcal{O}_{\mathbf{N}} & \mathcal{Z}_{g_r}^2 \end{pmatrix}. \end{aligned}$$

From this, $cov(\mathbf{y_{ij}}, \mathbf{y_{sk}}) = 0,$ for all j, k and $i \neq s$.

Thus,

$$\begin{aligned} \mathbf{\Omega}(N \times N) &= \sigma_g^2 \mathcal{Z}_g \mathcal{Z}_g^T + \sigma_e^2 \mathcal{I}_{\mathbf{N}} \\ &= \begin{pmatrix} \sigma_g^2 + \sigma_e^2 & \cdots & 2\varphi_{11n_1}\sigma_g^2 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 2\varphi_{1n_11}\sigma_g^2 & \cdots & \sigma_g^2 + \sigma_e^2 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \sigma_g^2 + \sigma_e^2 & \cdots & 2\varphi_{r1n_r}\sigma_g^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 2\varphi_{rn_r1}\sigma_g^2 & \cdots & \sigma_g^2 + \sigma_e^2 \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} \mathbf{\Omega_1} & 0 & \cdots & 0 \\ 0 & \mathbf{\Omega_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{\Omega_r} \end{pmatrix},$$

since $2\varphi_{ijk} = 1$ for all $j = k$.

Given this, we get $\underline{\mathbf{y}} \sim \mathcal{N}_N(\underline{\mu}, \mathbf{\Omega})$, where $\underline{\mu}$ and $\mathbf{\Omega}$ are defined above.

Since we are interested in inference on the parameters μ, σ_e^2 and specifically σ_g^2, we can carry out a hypothesis test where the null hypothesis is $H_0 : \sigma_g^2 = 0$. If the null hypothesis is rejected, we have sufficient evidence to assume the trait of interest is inherited in our set of families.

The problem with such a hypothesis test is that standard statistical software does not allow us to specify a potentially different coefficient for each covariance in the covariance matrix. As a result, specialised software such as QTDT, SOLAR (Sequential Oligogenic Linkage Analysis Routines) and MENDEL were developed. Full details of these software programs are available on the websites listed in Table 8. For segregation analysis, this software calculates the necessary kinship coefficients, based on the pedigree data, and uses them in the model.

Now that we have introduced segregation analysis, we have the elements we need to discuss heritability, which is an important measure relating to inheritance. Different authors define heritability in different ways. We present the definition used by the majority of them, including Burton (2005).

The variance estimates obtained from the segregation Model (27) can be used to estimate heritability. The proportion of trait variance that is attributed to all hereditary factors (all alleles inherited from parents), compared to environmental factors, is known as the *broad-sense heritability*. It is denoted by H^2, and is defined as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}.$$

It expresses the extent to which an individual's observed trait value is determined by his

genetic material. The larger H^2 is, the more heritable the trait is, because more of the variation in the trait is explained by genetic factors, that is, $\sigma_g^2 > \sigma_e^2$. Consequently, the larger H^2 is, the greater the difference between families will be because the difference within families will be relatively smaller. Similarly, if $\sigma_g^2 = 0, H^2 = 0$, which implies that the trait is not heritable. Finally, if $0 < \sigma_g^2 < \sigma_e^2$, then $H^2 \longrightarrow 0$, which suggests that the trait is weakly heritable. Familial correlations depend on the degree to which family members are related. Therefore, the more closely related they are, the higher the correlations are expected to be.

We have said before that the aim of a genetic study is to determine if a trait is inherited, and if so, to isolate and identify the causal alleles. To do this we need to measure the genetic variation which exists at a particular locus (Burton et al., 2005). Using H^2, we can determine if a trait is heritable, but it considers all forms of hereditary variation, and not specifically the variation due to a particular genetic locus. For this we look to another measure of heritability, explained in Chapter 6 which follows.

Having now described familial aggregation, segregation analysis and heritability, we can move on to linkage analysis, which is the main focus of our study.

6 Linkage Analysis

6.1 Introduction and background

> As early as 1903, Sir Walter Sutton ... pointed out the likelihood that organisms contain many more "unit factors" than chromosomes. Soon thereafter, genetic studies with several organisms revealed that certain genes were not transmitted according to the law of independent assortment, rather these genes seemed to segregate as if they were somehow joined or linked together. Further investigations showed that such genes were part of the same chromosome, and they were indeed transmitted as a single unit.
>
> We now know that most chromosomes consist of very large numbers of genes ... Because the chromosome, not the gene, is the unit of transmission during meiosis, linked genes are not free to undergo independent assortment. In theory, the alleles at all loci of one chromosome should be transmitted as a unit during gamete formation. However, in many instances this does not occur (Klug & Cummings, 2000:137).

Contrary to Mendel's third postulate, alleles found near each other on the same chromosome do not segregate independently, but rather behave as though they are joined or linked. They are thus transmitted from parent to offspring as a single unit during meiosis. This is known as *cosegregation*, and when it occurs, linkage is said to exist. The fact that alleles near each other on a chromosome segregate together more often than expected, while alleles on different chromosomes segregate together purely by chance, is a fundamental concept of linkage analysis. It aims to localise alleles in an extended family by considering the co-inheritance of traits and genetic markers.

Linkage analysis is best at detecting alleles with strong effects on the trait in question. When the alleles at a locus control a quantitative trait, the locus is called a quantitative trait locus (QTL). Linkage analysis attempts to establish whether alleles at a marker locus are linked to alleles at a QTL.

Some examples of linkage analysis in practice include Stone et al. (2002), Abecasis et al.

(2004) and Xu et al. (2009). Stone et al. (2002) investigates a locus potentially associated with a predisposition to severe obesity. According to them, severe obesity is known to be heritable, but identifying the susceptibility alleles has been difficult. Linkage studies in this area have proven to be challenging due to the genetic complexity underlying the predisposition to severe obesity. This is because linkage analysis is sensitive to the degree of genetic complexity of a trait and predisposition to severe obesity has been found to be very complex. Linkage is also generally better at detecting alleles with a high penetrance, strong effects on the trait under observation and low within-family genetic variation. Stone et al. (2002) used several strategies in an attempt to overcome the problem caused by the genetic complexity of the disease. One of these strategies involved selecting their families in a specific way, so as to increase their ability to detect linkage.

Abecasis et al. (2004) and Xu et al. (2009) investigate possible loci associated with schizophrenia in the Afrikaner population of South Africa. According to them, multiple loci and environmental factors affect susceptibility to schizophrenia, but the role of individual loci is not fully understood. In addition, non-familial forms of the disease also exist, affecting findings. Researchers in both studies used linkage analysis to isolate genomic regions in which possible susceptibility alleles lie. To overcome the possible ethnic heterogeneity found in other studies, their study population was restricted to the genetically isolated Afrikaner population.

Going back in history, Mendel believed that during meiosis, each new gamete cell randomly receives one member of each homologous chromosomal pair, originating from either the maternal or paternal parent. It is now known that an event called a *crossover* may occur during meiosis. In a crossover event, parts of the maternal chromosome of a homologous pair 'swaps' genetic material with the corresponding parts of the paternal chromosome, resulting in chromosomes that contain a mixture of DNA from both parents (Burton et al., 2005).

Crossovers between two loci result in *recombination*, the joining of alleles from two different

homologous chromosomes. Recombination ensures that every gamete is unique and is usually observed when an odd number of crossovers occur (Burton et al., 2005).

Figure 13 shows in a simplified manner, how a crossover event occurs between two loci on homologous paternal and maternal chromosomes. During meiosis, the maternal and paternal homologous chromosomes line up next to each other and replicate. Suppose we are looking at two specific diallelic loci on such chromosomes. The first locus has alleles 1 and 2, and the second has alleles 3 and 4, on the respective maternal and paternal chromosomes. Now, part of one of the maternal chromosomes breaks off and joins with part of one of the paternal chromosomes. It does so in such a way that this piece of maternal chromosome 'swaps' with the corresponding piece of the paternal chromosome. Suppose that this breaking and recombining occurs between the two loci of interest. Then, after the crossover event occurs, we see that alleles 3 and 4 'exchange' chromosomes. We thus say that a recombination of these two alleles has occurred at the second locus.

Recombination can be observed in families, as shown in Figure 14. This is based on knowing the *phase* (which allele comes from which parent) of the individuals in the pedigree, for two diallelic markers. When phase is known, the alleles at the markers can be written with a box between them. Alleles in the first block are assumed to come from the paternal parent while those in the second box are assumed to come from the maternal parent. Alleles from different linked loci on the same chromosome are called *haplotypes*.

For the pedigree in Figure 14, Sue gets her 2–3 haplotype from her father and the 2–4 haplotype from her mother. Jane gets the 1–3 haplotype from her father and the 2–4 haplotype from her mother. However, here we see that there is a recombination in the alleles inherited from the father, as the 1 allele is transmitted to Jane instead of the 2 allele.

As the physical distance between two loci on the same chromosome increases, the probability of recombination, θ, between them during meiosis also increases until the limiting value of $\theta = \frac{1}{2}$ is reached. This is the same as the probability that two loci on separate

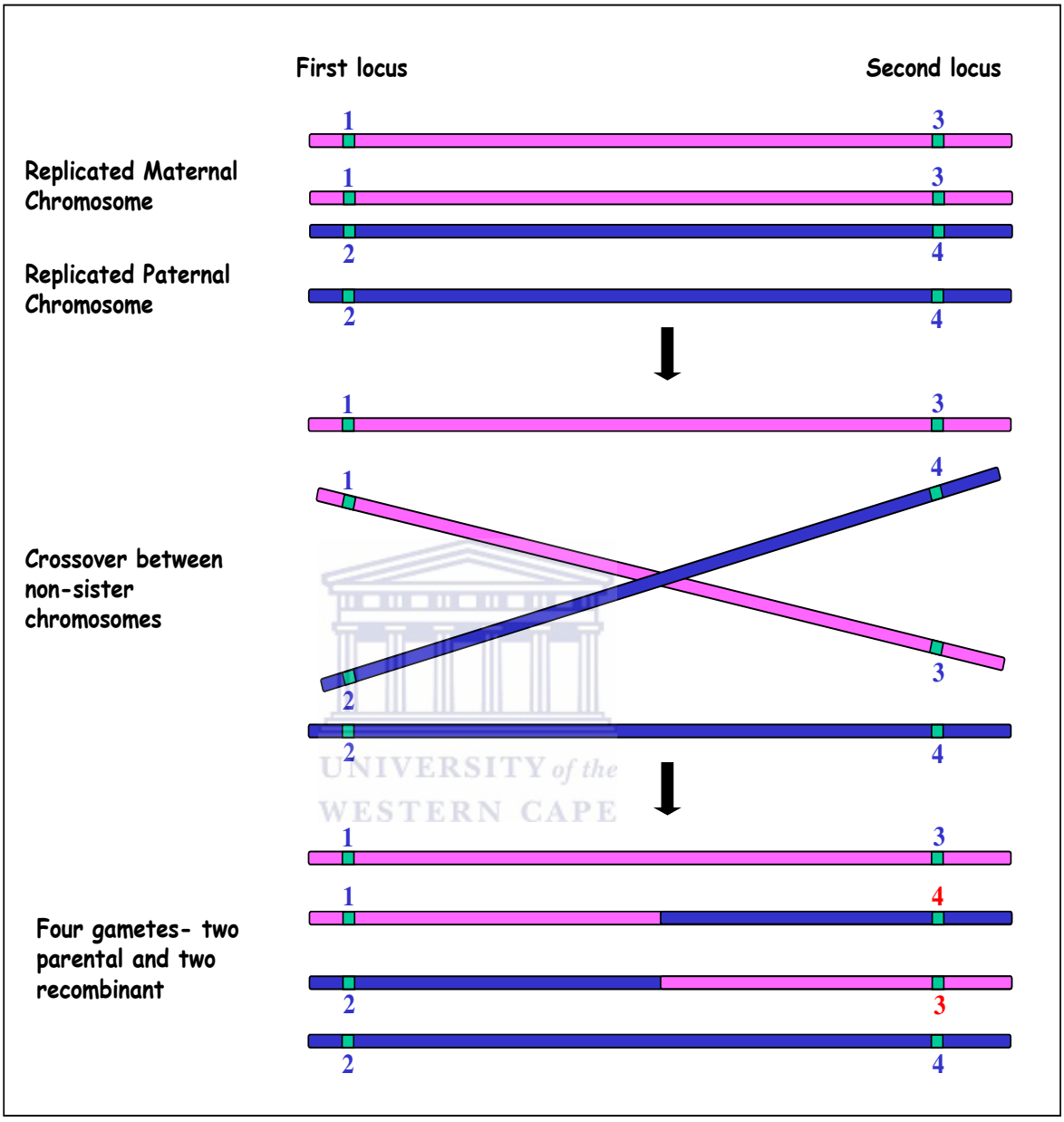


Figure 13: A single crossover event during meiosis

chromosomes will undergo recombination. Two loci that are immediately adjacent to each other on a chromosome are unlikely to undergo recombination. This can also be measured in terms of the recombination fraction, where two loci are linked if recombination between them occurs with a probability less than $\frac{1}{2}$. When $\theta = \frac{1}{2}$, there is no linkage.

The methods of linkage analysis can be applied to both single-gene disorders (parametric or model-based linkage analysis) and complex disease analysis (non-parametric or model-

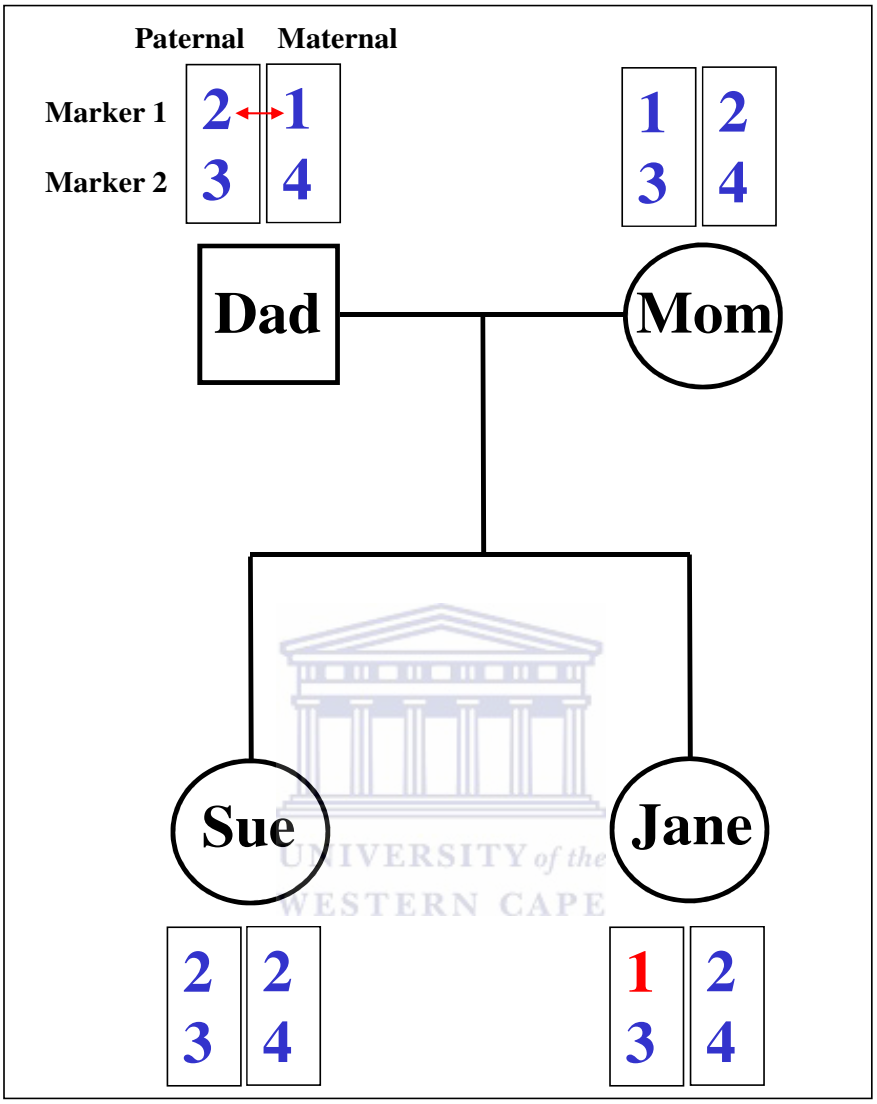


Figure 14: Hypothetical pedigree for demonstrating recombination

free linkage analysis), which are generally caused by both genetic and environmental factors. Here, the terms 'model-free' and 'non-parametric' are genetic terms which mean that no mode of inheritance is assumed (Elston, 1998). They differ from the statistical terms which indicate that no distributional assumptions are made about the data. Since both types of linkage analysis require specifying a statistical model and estimating parameters, the terms 'model-based' and 'model-free' will be used to distinguish between them.

In model-based linkage analysis, a genetic mode of inheritance is assumed, that is, assump-

tions are made about the penetrances and allele frequencies. To describe the evidence for linkage in a model-based linkage analysis, geneticists use a transformation of the likelihood ratio (LR) test, known as the *lod score* or *lods*, which is a function of the recombination fraction. The null hypothesis is that there is no linkage, so $H_0 : \theta = \frac{1}{2}$, while $H_1 : \theta < \frac{1}{2}$.

In the model-free case, no assumptions are made about the penetrances and allele frequencies, so linkage is not assessed through the recombination fraction. This is because, under the null hypothesis, the marker and trait are assumed to segregate independently and randomly of each other, making the true mode of inheritance irrelevant to the validity of any tests.

Model-based methods of linkage, such as the lod score, are used predominantly for single-gene disorders. For complex diseases, it is almost impossible to correctly specify the entire model. In such cases, model-free methods of linkage analysis are preferred. Model-based methods are not discussed further here. In the following section, the focus will be on model-free linkage methods, which are used to increase the robustness, speed and accuracy of calculations, due mostly to modern improvements in computational technology. Of all the model-free methods that can be used, the variance-components methods are of particular interest to us. As with any statistical method, they have their advantages and disadvantages. Before we look at the model-free variance-components methods however, another important measure needs to be introduced.

6.2 Identical by descent

Previously, it was mentioned that one of the reasons why genetic pedigree studies are unique is that they contain information for family members, which needs to be accounted for when carrying out statistical analyses on such data. We have already presented the kinship coefficient, which quantifies the degree of relationship between family pairs. We now present a measure which allows us to quantify the degree of genetic similarity between family pairs, at a specific genetic locus. Thereafter we will illustrate how this information

is used in linkage analysis.

The probability that two family members share 0, 1 or 2 alleles from the same source at any autosomal locus defines their genetic relationship at that locus. Two alleles at a specific locus, each one from a different person, are said to be *identical by descent* (IBD) if they come from a common ancestor. If they are identical in their DNA composition, but it is unknown whether or not they have a common ancestor, then they are said to be *identical by state* (IBS). Two alleles that are IBD must also be IBS.

Let $\pi_{ijk}(a)$ be the probability that individuals j and k in family i share a alleles IBD at a single marker locus. Parent-child pairs must share 1 allele IBD, with probability 1 (assuming the parents are unrelated individuals). Sib-pairs share anything from 0 to 2 alleles IBD at a locus, while monozygotic (identical) twins share both alleles IBD. Unrelated individuals share 0 alleles IBD, with probability 1.

IBD probabilities for sib-pairs and other relative-pairs need to be calculated from their observed genotypes at a particular locus. For example, for a particular locus, consider the three scenarios in Figure 15: In pedigree 1 of Figure 15, Dad has genotype 1/2, Mom has genotype 3/4 and their children have genotypes: Sally: 1/3; Ty: 2/3, and Kay: 2/3. Since both parents are heterozygous and thus informative at this locus, the IBD sharing of each offspring pair can be determined exactly. Sally gets her 1 allele from Dad and the 3 allele from Mom, while Ty receives his 2 allele from Dad and 3 allele from Mom. Sally and Ty receive different alleles from Dad but the same 3 allele from Mom, thus they share 1 allele IBS. They also share this allele IBD, with probability 1. Similarly, Sally shares the 3 allele IBD (and IBS) with Kay, so they also share 1 allele IBD with probability 1. Now, Kay and Ty both receive their 2 allele from Dad and their 3 allele from Mom, implying that they share 2 alleles IBS and they share 2 IBD with probability 1.

In pedigree 2 of Figure 15, Dad has genotype 1/2, Mom has genotype 1/3 and their children have genotypes: Sue: 1/2; and Jane: 1/3. Thus, Sue and Jane share 1 allele IBS. If we did not have the parental genotypes, Sue and Jane's IBD sharing probabilities

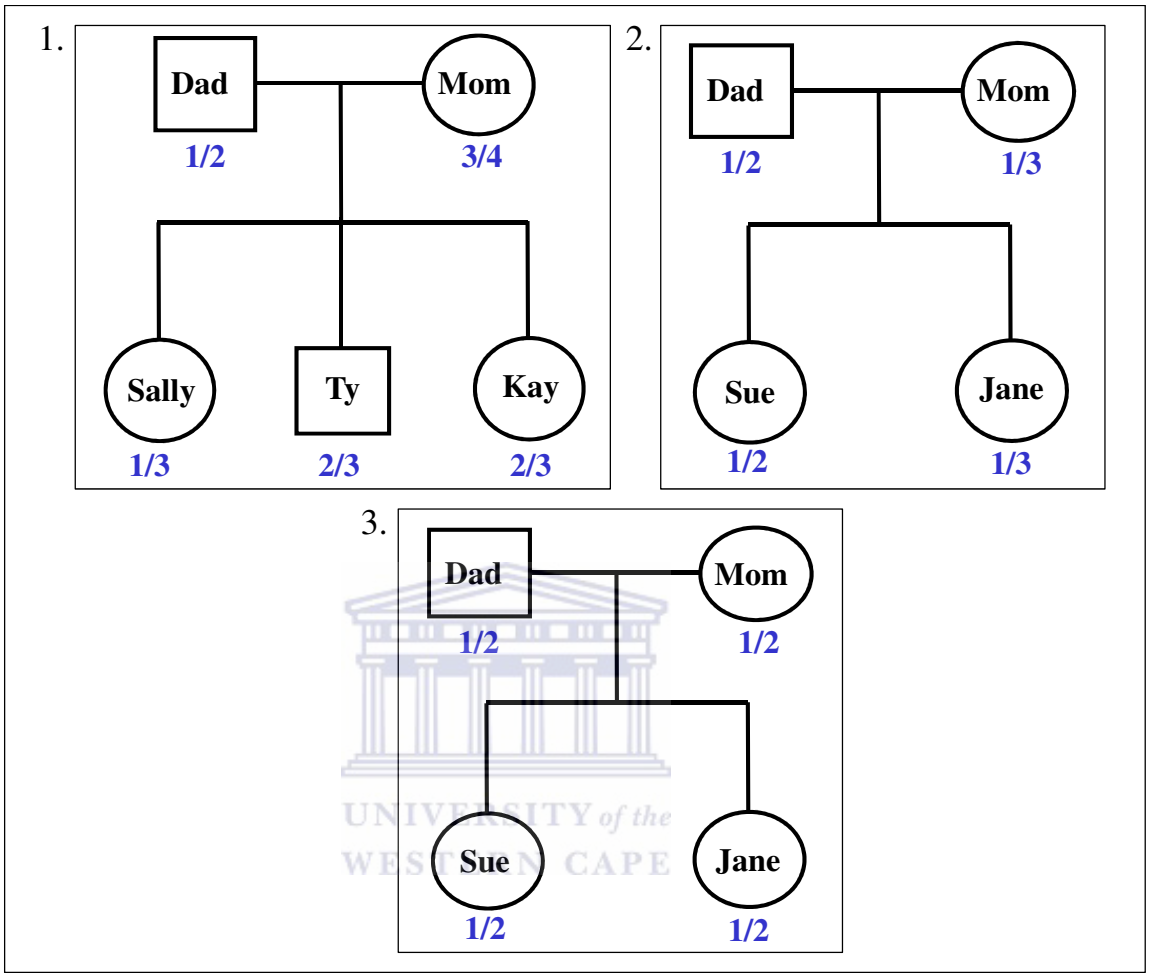


Figure 15: Hypothetical pedigrees for IBD probability calculations

would have to be estimated. However, we know the phase: Sue gets her 2 allele from Dad and her 1 allele from Mom, while Jane gets her 1 allele from Dad and her 3 allele from Mom. Thus, Sue and Jane share 0 alleles IBD at this locus, with probability 1.

For a slightly more complicated scenario, we look at pedigree 3 in Figure 15. Here, Dad, Mom and both children have genotype 1/2. So we know that Sue and Jane share 2 alleles IBS but we do not know if they are IBD. We can use a table to calculate Sue and Jane's IBD sharing at this locus. Now, Sue can get either her 1 or 2 allele from Dad and the other from Mom. Similarly for Jane. Suppose we distinguish between these and count the number of alleles they share, as shown in Table 3.

Table 3: Number of alleles shared IBD by sib-pairs, at a given locus

	Jane	
Sue	Dad 1/Mom 2	Dad 2/Mom 1
Dad 1/Mom 2	2	0
Dad 2/Mom 1	0	2

From Table 3, Sue and Jane have the same genotype, but their phase might be different. Thus they can share either 0 or both alleles IBD, with the following IBD probabilities:

$$\pi_{S,J}(0) = \frac{2}{4} = \frac{1}{2}$$

$$\pi_{S,J}(1) = \frac{0}{4} = 0$$

$$\pi_{S,J}(2) = \frac{2}{4} = \frac{1}{2}.$$

Now suppose we have also obtained genotype information for the extended pedigree in Figure 16, for a specific diallelic marker. This pedigree consists of six members: Dad and Mom are again the parents of Sue and Jane; Ryan is married to Jane and their daughter is Ally. We assume that Ryan is only related to the rest of the family through his marriage. Suppose that the family's genotypes at this marker are: Dad: 1/2; Mom: 1/2; Sue: 1/1; Jane: 1/1; Ryan: 0/0 (unknown); and Ally: 1/2. The IBD sharing of pairs of members from this family can be inferred in the same way as shown above, using the known genotypes.

Table 4 gives the IBD probabilities and kinship coefficients for pairs of members of the example family in Figure 16. In this table, Dad and Ally share 0 alleles IBD for the marker, but Mom and Ally share 1 allele IBD, even though both pairs are grandparent-grandchild relationships. So while the two pairs have different IBD sharing at this locus, we still capture their relationship through the kinship coefficient, which is the same for both pairs. Also, although Ryan's genotypes are unknown, we can calculate his IBD sharing with the rest of the family because he is a married-in; he is only related to the rest of the family through his daughter.

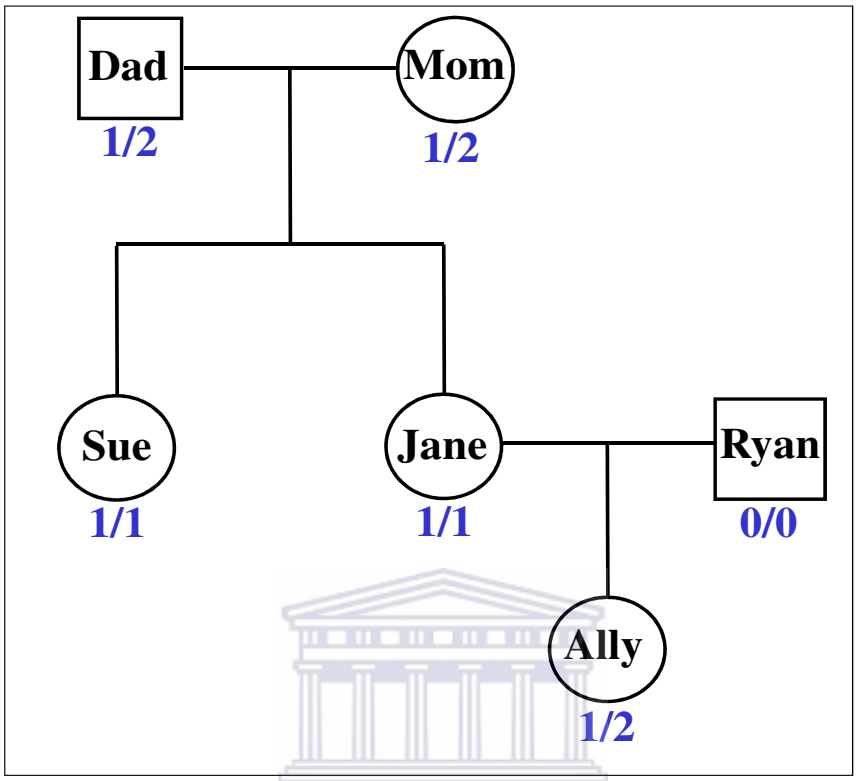


Figure 16: IBD probability calculations for a hypothetical extended pedigree

Table 4: IBD probabilities and kinship coefficients when genotypes are known

		IBD probabilities				
Relative type	**Relative pair**	$\pi(0)$	$\pi(1)$	$\pi(2)$	$\bar{\pi}_{ijk}$	2φ
Parent–offspring	Dad/Mom–Sue/Jane	0	1	0	1/2	1/2
	Jane/Ryan–Ally	0	1	0	1/2	1/2
Sib–pairs	Sue–Jane	0	0	1	1	1/2
Grandparent–grandchild	Dad–Ally	1	0	0	0	1/4
	Mom–Ally	0	1	0	1/2	1/4
Aunt–niece	Sue–Ally	0	1	0	1/2	1/4
Unrelated individuals	Dad–Mom	1	0	0	0	0
	Dad/Mom/Jane/Sue–Ryan	1	0	0	0	0

While the scenarios in Figures 15 and 16 are relatively straight-forward, when some genotypes are not available or informative, IBD sharing cannot always be inferred and, in such cases, theoretical or prior IBD probabilities have to be used. These are based on the same information as kinship coefficients. They are calculated under the null hypothesis of no linkage and assuming that all individuals in a pedigree are heterozygous. For sib-pairs,

the probability that they share 0, 1 or 2 alleles IBD under these assumptions, occurs with probability $\pi_{ijk}(0) = \frac{1}{4}$, $\pi_{ijk}(1) = \frac{1}{2}$ and $\pi_{ijk}(2) = \frac{1}{4}$ respectively. These probabilities are based on the binomial distribution, under the assumption that alleles segregate with equal probability.

Consider the pedigree in Figure 17. This pedigree is from Figure 12 and now contains the genotype information (in blue) for several family members, at a specific diallelic marker locus. The genotypes of Sue, Jane, Jack and Ally are unknown.

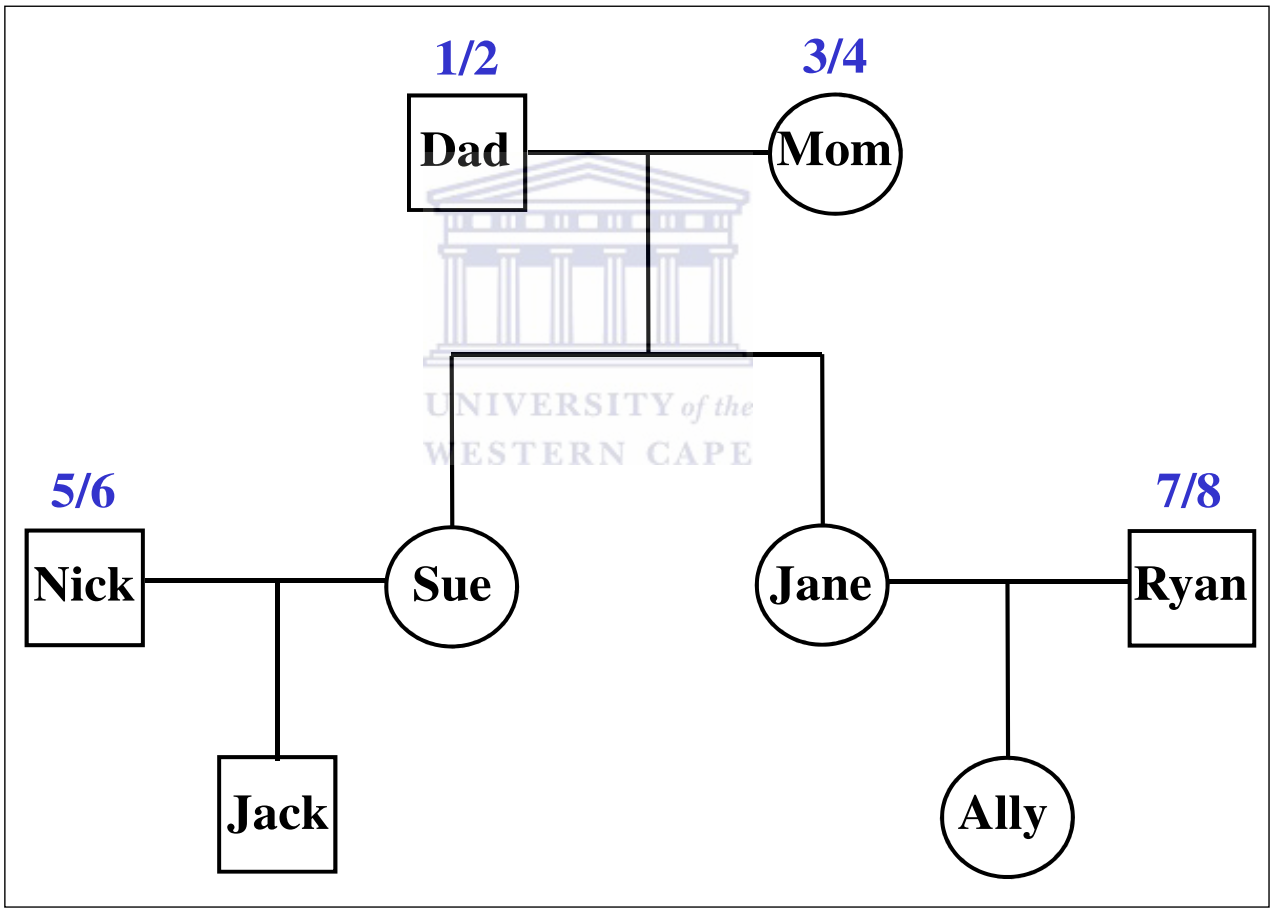


Figure 17: Hypothetical pedigree for theoretical IBD probability calculations

Since Sue and Jane are the offspring of Dad and Mom, each of their genotypes can be either $1/3, 2/4, 1/4$ or $2/3$, with equal probability. Table 5 gives the possible number of alleles they can share IBD. We use this table to calculate the prior probabilities of siblings sharing 0, 1 or 2 alleles IBD.

Table 5: Possible number of alleles shared IBD by sib-pairs

	Jane			
Sue	1/3	2/4	1/4	2/3
1/3	2	0	1	1
2/4	0	2	1	1
1/4	1	1	2	0
2/3	1	1	0	2

From Table 5 we can see that the theoretical or prior IBD probabilities are

$$\pi_{S,J}(0) = \frac{4}{16} = \frac{1}{4}$$
$$\pi_{S,J}(1) = \frac{8}{16} = \frac{1}{2}$$
$$\pi_{S,J}(2) = \frac{4}{16} = \frac{1}{4}.$$

This is true for any such sib-pair.

Given the above four possible genotypes for Jane and given that Ryan has genotype 7/8, their daughter, Ally, can have any one of the following possible genotypes with equal probability:
1/7, 1/8, 3/7, 3/8, 2/7, 2/8, 4/7 or 4/8. So if we want to work out the the theoretical or prior IBD sharing probabilities for a grandparent–grandchild pair, we can tabulate the alleles for, say, Dad and Ally, as in Table 6.

Table 6: Possible number of alleles shared IBD by grandparent-grandchild pairs

	Ally							
Dad	1/7	1/8	3/7	3/8	2/7	2/8	4/7	4/8
1/2	1	1	0	0	1	1	0	0

Therefore,

$$\pi_{D,A}(0) = \frac{4}{8} = \frac{1}{2}$$
$$\pi_{D,A}(1) = \frac{4}{8} = \frac{1}{2}$$
$$\pi_{D,A}(2) = \frac{0}{8} = 0,$$

which is true for any grandparent–grandchild pair.

The kinship coefficient (φ_{ijk}), defined in Section 5.2, can be defined in terms of prior or theoretical IBD probabilities: it is the probability that a pair of alleles at a given locus in family i, one selected randomly from individual j and the other from individual k, are IBD. Mathematically, this is represented as:

$$\varphi_{ijk} = \frac{1}{2}\Big[\frac{\pi_{ijk}(1)}{2} + \pi_{ijk}(2)\Big] = \frac{1}{4}[1.\pi_{ijk}(1) + 2.\pi_{ijk}(2)]$$

Let $\bar{\pi}_{ijk}$ be the expected proportion of alleles shared IBD by individuals j and k in family i. Then,

$$\begin{aligned}\bar{\pi}_{ijk} &= \frac{1}{2}\sum_{a=0}^{2} a.\pi_{ijk}(a) \\ &= \frac{1}{2}[1.\pi_{ijk}(1) + 2.\pi_{ijk}(2)] \\ &= 2\varphi_{ijk}\end{aligned}$$

Therefore, $2\varphi_{ijk}$ gives us the probability that a randomly selected allele from individual i and a randomly selected allele from individual j, originate from the same ancestor. The theoretical (prior) IBD probabilities and kinships are summarised in Table 7.

Table 7: Prior IBD probabilities and kinship coefficients for various family pairs

	IBD probabilities				
Relative types	$\pi(0)$	$\pi(1)$	$\pi(2)$	$\bar{\pi}_{ijk}$	2φ
Monozygotic (identical) twins	0	0	1	1	1
Parent–offspring	0	1	0	1/2	1/2
Dizygotic (non-identical) twins/Sib–pairs	1/4	1/2	1/4	1/2	1/2
Half–Sibs	1/2	1/2	0	1/4	1/4
Grandparent–grandchild	1/2	1/2	0	1/4	1/4
Uncle/Aunt–niece/nephew	1/2	1/2	0	1/4	1/4
First cousins	3/4	1/4	0	1/8	1/8
Second cousins	15/16	1/16	0	1/32	1/32
Unrelated individuals	1	0	0	0	0

It is important to note that Table 7 is a table of theoretical IBD values, which explains why the IBD sharing and kinship coefficients are the same. The IBDs given in Table 7 are only

used in practice when a specific individual's genotype is unknown or uninformative. Once genotypes are available for family members, IBD sharing is fully dependent on them. Thus in Table 4, $\bar{\pi}_{ijk}$, is actually the observed proportion of alleles shared IBD and it differs from 2φ, whereas in Table 7, $\bar{\pi}_{ijk}$ and 2φ are the same, because $\bar{\pi}_{ijk}$ is based on theoretical (prior) IBD sharing. Since IBD sharing depends on genotypes, there are potentially different IBD probabilities at each genetic locus, for any two related individuals.

In Table 4 we showed actual IBD probabilities, calculated from available (observed) genotypes, and in Table 7 we calculated theoretical IBD probabilities by assuming all family members are heterozygous and informative. In practice, there is an intermediate step between actual and theoretical IBD calculations. The theoretical IBD sharing in Table 7 can be estimated more accurately if the population allele frequencies are known. For example, for the pedigree in Figure 17, suppose Mom's genotype is unknown but we do know that allele 1 occurs with a frequency of say, 20%. Then we can use this as a prior probability for Mom's unknown genotype by assuming that her hypothetical 3/4 genotype is actually 1/1; 1/2 or 2/2 with binomial probabilities: $(0.2)^2; (2\text{x}0.2\text{x}0.8);$ and $(0.8)^2$ respectively. However, this method is computationally demanding, so specialised programs are available to carry out these complex calculations.

As with kinship coefficients, IBD values are important when building up statistical models because they make up the coefficients of the locus-specific variance in the covariance matrices. The expected proportion of alleles shared IBD (the expected value of the IBD distribution) is used to extract the variance-component of a specific marker. If the genotypes of a study group are not known for a particular marker, then it is not possible to separate the effect of that marker from the overall effect of all genes (heritability). In other words, we cannot split the trait variance into heritable genetic variance and a locus-specific variance. This is the crux of model-free linkage analysis and is presented in the next section.

6.3 Model-free linkage analysis

In model-free linkage analysis, the relationships between sib-pairs and other relative-pairs in extended pedigrees are used in an attempt to locate the chromosomal position of trait-causing alleles. It is based on the hypothesis that, at the QTL, there is a greater similarity between the traits of relative pairs that share alleles IBD than between those that do not share alleles IBD. The basis of model-free linkage is that genotypic similarities at the QTL are positively correlated with trait similarities. Here, marker and trait data in families are assessed for evidence of linkage.

The model-free methods we will focus on here are the variance-components methods. These methods are sensitive to sampling and distributional assumptions; namely that genetic parameters assume random recruitment and underlying normality. Their advantages include application to large, complex pedigrees instead of just sib-pairs, and they can easily include multiple loci, gene∗gene interactions, as well as gene∗environment interactions (none of which will be considered here). Variance-components methods are also more powerful than some other methods when the model assumptions are met.

Historically, quantitative trait linkage analysis is only carried out on sib-pairs or family trios. One sib-pair method is the Haseman-Elston (HE) method, from which the variance-components methods for family analysis were developed. It is a regression-based method in which the squared trait differences between sib-pairs, $D_i = (\underline{\mathbf{y}}_{ij} - \underline{\mathbf{y}}_{ik})^2$, are regressed on the (estimated) expected proportion of alleles shared IBD by the sib-pair ($\bar{\pi}_{ijk}$), at the locus of interest. If the marker is linked to a QTL, then the sibling pair are expected to share more alleles IBD, as well as have more similar trait values. In this case, there is a negative relationship between D_i and $\bar{\pi}_{ijk}$ (Teare & Barrett, 2005).

On the other hand, the classic variance-components technique involves splitting the total variance into variance due to genetic factors and variance due to environmental factors. This was later improved by including in the model, a variance-component (σ_a^2) for the

hypothesised QTL near the marker. In this way, linkage analysis could be carried out.

While variance-components methods are generally more powerful than HE, the biggest advantage that the HE method has over variance-components methodology is that it is more robust to non-normality and selective recruitment than variance-components. However, it can only be used for sib-pair analysis while variance-components methods can be used to analyse data for pedigrees of any size. Also, HE cannot easily accommodate covariates while an important advantage of variance-components is that it enables us to model means and variances simultaneously, as shown in Chapter 4. This is what makes the variance-components methodology more powerful than traditional sib-pair analysis. This is demonstrated in Pratt et al. (2000), which found that the variance-components methods are consistently more powerful than the HE method. They attributed the large difference in results to the loss of information which occurs when only sib-pairs are used, compared to extended families.

For complex traits, in addition to environmental effects, there may be the aggregate effects of alleles at several loci which contribute a fixed amount to the trait (additive inheritance). Additive inheritance centres around the idea that additive alleles at various loci control quantitative traits, resulting in continuous variation which can be explained by Mendelian inheritance. Here, the hereditary variance, σ_g^2, is divided into locus-specific or additive genetic variance, σ_a^2, and variance due to alleles at many loci. σ_a^2 is important because it is the main cause of resemblance between relatives. In addition, it can be estimated from observed genotype data, making it useful in practice.

As stated in Burton et al. (2005:946), "One of the principal reasons for fitting a variance-components model is to estimate the variance attributable to additive genetic effects." This can be done from a model such as the one which follows.

Suppose there is a particular genetic marker locus that we are interested in. The model for linkage analysis includes both the kinship coefficient and IBD sharing for each family pair, at the locus. This is done by including coefficients for the random effects of family

relatedness and the specific locus. The kinship coefficients extract the hereditary variance, σ_g^2, from the between family variance and are the coefficients of the hereditary random effect. Similarly, using the observed (estimated) proportion of alleles shared IBD as the coefficient for the locus-specific random effect, we extract the locus-specific variance, σ_a^2, from σ_g^2. This allows us to account for the degree to which a pair is related, as well as to account for the alleles they inherit at a specific locus.

Consider our set of r extended families, each with n_i members. The total number of family members $N = \sum_i^r n_i$. Let $\mathbf{y_{ij}}$ represent the random quantitative trait value for individual j, from family i. Let $\underline{\mathbf{y}}_\mathbf{i}(n_i \times 1) = (\mathbf{y_{i1}}, \ldots, \mathbf{y_{in_i}})^T$ be the random vector of trait values for family i and let the overall vector of trait values be $\underline{\mathbf{y}}(N \times 1) = (\underline{\mathbf{y}}_\mathbf{1}, \ldots, \underline{\mathbf{y}}_\mathbf{r})^T$. Let $\mathbf{g_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_g^2)$ denote the hereditary random effect for family i, and let $\mathbf{e_{ij}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$ be the environmental effect for individual j from family i. Now let $\mathbf{a_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_a^2)$ be the random locus-specific additive allelic effect for family i. It measures the impact of the additive allelic effect at the specific locus. Assume $\mathbf{g_i}, \mathbf{a_i}$ and $\mathbf{e_{ij}}$ are independent of each other for different families and independent within families as well.

Then, for the j^{th} individual in the i^{th} family, a model for this data is

$$\mathbf{y_{ij}} = \mu + \underline{z}_{a_ij}^T \underline{\mathbf{a_i}} + \underline{z}_{g_ij}^T \underline{\mathbf{g_i}} + \mathbf{e_{ij}}, \text{ for } i = 1, \ldots, r;\ j = 1, \ldots, n_i, \tag{28}$$

where

μ is the overall mean trait value,

$\underline{z}_{g_ij}(1 \times n_i) = \{z_{g_ijk} = \sqrt{2\varphi_{ijk}}\}$ is the vector of regression coefficients for individual j, corresponding to the hereditary random effect for family i, where φ_{ijk} is the kinship coefficient between individuals j and k in family i.

Finally, $\underline{z}_{a_ij}^T(1 \times n_i) = \{z_{a_ijk} = \sqrt{\bar{\pi}_{ijk}}\}$ is the vector of regression coefficients for individual j, corresponding to the additive random effect of the marker locus, for family i, where $\bar{\pi}_{ijk}$ is the observed (estimated) proportion of alleles shared IBD by individuals j and k in family i.

Therefore, for individual j, $\mathbf{y_{ij}}$ is normally distributed with the following mean and vari-

ance:

$$\begin{aligned} E(\mathbf{y_{ij}}) &= \mu, \text{ since } E(\mathbf{e_{ij}}) = 0 \\ var(\mathbf{y_{ij}}) &= \sigma_a^2 + \sigma_g^2 + \sigma_e^2, \text{ since } cov(\mathbf{a_i}, \mathbf{g_i}) = cov(\mathbf{a_i}, \mathbf{e_{ij}}) = cov(\mathbf{g_i}, \mathbf{e_{ij}}) = 0 \end{aligned}$$

and $2\varphi_{ijk} = \bar{\pi}_{ijk} = 1$ for all $j = k$.

As a result, the variance of the trait for the j^{th} person in the i^{th} family includes the variance-component for the additive genetic effect.

Expanding (28) for the i^{th} family gives

$$\underline{\mathbf{y}}_{\mathbf{i}} = \underline{1}_{n_i}\mu + \mathcal{Z}_{a_i}\underline{\mathbf{a}}_{\mathbf{i}} + \mathcal{Z}_{g_i}\underline{\mathbf{g}}_{\mathbf{i}} + \underline{\mathbf{e}}_{\mathbf{i}}, \text{ for } i = 1, \ldots, r, \tag{29}$$

where

$\mathcal{Z}_{g_i}(n_i \times n_i) = \{\sqrt{2\varphi_{ijk}}\}$ is the symmetric matrix of regression coefficients for the random hereditary effects, and

$\mathcal{Z}_{a_i}(n_i \times n_i) = \{\sqrt{\bar{\pi}_{ijk}}\}$ is the symmetric matrix of regression coefficients for the random additive locus effects.

Therefore,

$$\begin{aligned} E(\underline{\mathbf{y}}_{\mathbf{i}}) &= \underline{1}_{n_i}\mu(n_i \times 1) \\ cov(\underline{\mathbf{y}}_{\mathbf{i}}) &= E(\mathcal{Z}_{a_i}\underline{\mathbf{a}}_{\mathbf{i}} + \mathcal{Z}_{g_i}\underline{\mathbf{g}}_{\mathbf{i}} + \underline{\mathbf{e}}_{\mathbf{i}})(\mathcal{Z}_{a_i}\underline{\mathbf{a}}_{\mathbf{i}} + \mathcal{Z}_{g_i}\underline{\mathbf{g}}_{\mathbf{i}} + \underline{\mathbf{e}}_{\mathbf{i}})^T \end{aligned}$$

By the assumption of mutually independent random effects

$$\begin{aligned} cov(\underline{\mathbf{y}}_{\mathbf{i}}) &= \mathcal{Z}_{a_i}E(\underline{\mathbf{a}}_{\mathbf{i}}\underline{\mathbf{a}}_{\mathbf{i}}^T)\mathcal{Z}_{a_i}^T + \mathcal{Z}_{g_i}E(\underline{\mathbf{g}}_{\mathbf{i}}\underline{\mathbf{g}}_{\mathbf{i}}^T)\mathcal{Z}_{g_i}^T + E(\underline{\mathbf{e}}_{\mathbf{i}}\underline{\mathbf{e}}_{\mathbf{i}}^T) \\ &= \sigma_a^2\mathcal{Z}_{a_i}\mathcal{Z}_{a_i}^T + \sigma_g^2\mathcal{Z}_{g_i}\mathcal{Z}_{g_i}^T + \sigma_e^2\mathcal{I}_{\mathbf{n_i}} \\ &= \mathbf{\Omega_i}(n_i \times n_i), \text{ say.} \end{aligned}$$

As is the case for segregation analysis,

$$\mathcal{Z}_{g_i}\mathcal{Z}_{g_i}^T(n_i \times n_i) = \begin{pmatrix} 1 & 2\varphi_{i12} & \cdots & 2\varphi_{i1n_i} \\ 2\varphi_{i21} & 1 & \cdots & 2\varphi_{i2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ 2\varphi_{in_i1} & 2\varphi_{in_i2} & \cdots & 1 \end{pmatrix}.$$

Now,

$$\mathcal{Z}_{a_i}\mathcal{Z}_{a_i}^T(n_i \times n_i) \;=\; \begin{pmatrix} 1 & \bar{\pi}_{i12} & \cdots & \bar{\pi}_{i1n_i} \\ \bar{\pi}_{i21} & 1 & \cdots & \bar{\pi}_{i2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\pi}_{in_i1} & \bar{\pi}_{in_i2} & \cdots & 1 \end{pmatrix}.$$

So, for model-free linkage analysis, the covariance between individuals j and k in family i is given by

$$\Omega_{ijk} = cov(\mathbf{y_{ij}}, \mathbf{y_{ik}}) = \begin{cases} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \text{if } j = k \\ \bar{\pi}_{ijk}\sigma_a^2 + 2\varphi_{ijk}\sigma_g^2 & \text{if } j \neq k, \end{cases}$$

where φ_{ijk} is the kinship coefficient and $\bar{\pi}_{ijk}$ is the observed (estimated) proportion of alleles shared IBD by individuals j and k in family i.

From this we get,

$$\rho(\mathbf{y_{ij}}, \mathbf{y_{ik}}) \;=\; \frac{\bar{\pi}_{ijk}\sigma_a^2 + 2\varphi_{ijk}\sigma_g^2}{\sigma_a^2 + \sigma_g^2 + \sigma_e^2}, \text{ for all } j \neq k.$$

So now the correlation between two related individuals accounts for the degree to which they are related (hereditary effects) as well as their locus-specific genetic relationship (additive allelic effects).

For our family consisting of Dad, Mom, Sue, Jane, Ryan and Ally, the covariance matrix $\mathbf{\Omega_i}$ has the following structure:

$$\begin{pmatrix} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \bar{\pi}_{D,M}\sigma_a^2 + 2\varphi_{D,M}\sigma_g^2 & \bar{\pi}_{D,S}\sigma_a^2 + 2\varphi_{D,S}\sigma_g^2 & \bar{\pi}_{D,J}\sigma_a^2 + 2\varphi_{D,J}\sigma_g^2 & \bar{\pi}_{D,R}\sigma_a^2 + 2\varphi_{D,R}\sigma_g^2 & \bar{\pi}_{D,A}\sigma_a^2 + 2\varphi_{D,A}\sigma_g^2 \\ & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \bar{\pi}_{M,S}\sigma_a^2 + 2\varphi_{M,S}\sigma_g^2 & \bar{\pi}_{M,J}\sigma_a^2 + 2\varphi_{M,J}\sigma_g^2 & \bar{\pi}_{M,R}\sigma_a^2 + 2\varphi_{M,R}\sigma_g^2 & \bar{\pi}_{M,A}\sigma_a^2 + 2\varphi_{M,A}\sigma_g^2 \\ & & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \bar{\pi}_{S,J}\sigma_a^2 + 2\varphi_{S,J}\sigma_g^2 & \bar{\pi}_{S,R}\sigma_a^2 + 2\varphi_{S,R}\sigma_g^2 & \bar{\pi}_{S,A}\sigma_a^2 + 2\varphi_{S,A}\sigma_g^2 \\ & & & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \bar{\pi}_{J,R}\sigma_a^2 + 2\varphi_{J,R}\sigma_g^2 & \bar{\pi}_{J,A}\sigma_a^2 + 2\varphi_{J,A}\sigma_g^2 \\ & & & & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \bar{\pi}_{R,A}\sigma_a^2 + 2\varphi_{R,A}\sigma_g^2 \\ & & & & & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 \end{pmatrix}.$$

Using the IBD and kinship values from Table 4, the covariance matrix above becomes:

$$\begin{pmatrix} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & 0 & \frac{1}{2}\sigma_a^2 + \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_a^2 + \frac{1}{2}\sigma_g^2 & 0 & 0 + \frac{1}{4}\sigma_g^2 \\ & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \frac{1}{2}\sigma_a^2 + \frac{1}{2}\sigma_g^2 & \frac{1}{2}\sigma_a^2 + \frac{1}{2}\sigma_g^2 & 0 & \frac{1}{2}\sigma_a^2 + \frac{1}{4}\sigma_g^2 \\ & & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & 1\sigma_a^2 + \frac{1}{2}\sigma_g^2 & 0 & \frac{1}{2}\sigma_a^2 + \frac{1}{4}\sigma_g^2 \\ & & & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & 0 & \frac{1}{2}\sigma_a^2 + \frac{1}{2}\sigma_g^2 \\ & & & & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \frac{1}{2}\sigma_a^2 + \frac{1}{2}\sigma_g^2 \\ & & & & & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 \end{pmatrix}.$$

This matrix is now even more complex than the corresponding covariance matrix for segregation analysis (Model (26)) as we now have one more variance component. The level of complexity will also increase with family size, as we now have to include both kinship coefficients and IBD probabilities for all additional family members.

We can expand Model (29), for all the families in our study group, to obtain the model for all the observations. Here, we have

$$\underline{\mathbf{y}} = \underline{1}_N\mu + \mathcal{Z}_a\underline{\mathbf{a}} + \mathcal{Z}_g\underline{\mathbf{g}} + \underline{\mathbf{e}}, \tag{30}$$

where

$\underline{\mathbf{e}}(N \times 1) = (\mathbf{e_{11}}, \ldots, \mathbf{e_{rn_r}})^T \sim \mathcal{N}_N(\underline{0}, \sigma_e^2\mathcal{I}_\mathbf{N})$ is the vector of environmental effects,

$\underline{\mathbf{a}}(N \times 1) = (\mathbf{a_1}, \ldots, \mathbf{a_r})^T \sim \mathcal{N}_N(\underline{0}, \sigma_a^2\mathcal{Z}_a\mathcal{Z}_a^T)$ is the vector of all random locus-specific additive allelic effects,

$\underline{\mathbf{g}}(N \times 1) = (\mathbf{g_1}, \ldots, \mathbf{g_r})^T \sim \mathcal{N}_N(\underline{0}, \sigma_g^2\mathcal{Z}_g\mathcal{Z}_g^T)$ is the vector of all random hereditary effects,

$\mathcal{Z}_a(N \times N) = \{\sqrt{\bar{\pi}_{ijk}}\}$ is the block diagonal symmetric matrix of regression coefficients for the random locus-specific effects, $\mathbf{a_i}$, while

$\mathcal{Z}_g(N \times N) = \{\sqrt{2\varphi_{ijk}}\}$ is the block diagonal symmetric matrix of regression coefficients for the random hereditary effects, $\mathbf{g_i}$, for all r families.

Therefore,

$$\begin{aligned} E(\underline{\mathbf{y}}) &= \underline{1}_N\mu(N \times 1) \\ cov(\underline{\mathbf{y}}) &= \sigma_a^2\mathcal{Z}_a\mathcal{Z}_a^T + \sigma_g^2\mathcal{Z}_g\mathcal{Z}_g^T + \sigma_e^2\mathcal{I}_\mathbf{N}, \text{ since } cov(\mathbf{a_i}, \mathbf{g_i}) = cov(\mathbf{a_i}, \mathbf{e_{ij}}) = cov(\mathbf{g_i}, \mathbf{e_{ij}}) = 0 \\ &= \mathbf{\Omega}(N \times N), \text{ say,} \end{aligned}$$

where $\mathcal{Z}_g\mathcal{Z}_g^T(N \times N)$ is the block diagonal matrix of kinship coefficients, and

$$\mathcal{Z}_a\mathcal{Z}_a^T(N \times N) = \begin{pmatrix} \bar{\pi}_{111} & \cdots & \bar{\pi}_{11n_1} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{\pi}_{1n_11} & \cdots & \bar{\pi}_{1n_1n_1} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \bar{\pi}_{r11} & \cdots & \bar{\pi}_{r1n_r} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \bar{\pi}_{rn_r1} & \cdots & \bar{\pi}_{rn_rn_r} \end{pmatrix}.$$

Since $cov(\mathbf{y_{ij}}, \mathbf{y_{sk}}) = 0$ for all j, k and for family $i \neq s$, and an individual's kinship coefficient and IBD probability with himself are both equal to one, $\mathbf{\Omega}(N \times N)$ is

$$\begin{pmatrix} \sigma_a^2+\sigma_g^2+\sigma_e^2 & \cdots & \bar{\pi}_{11n_1}\sigma_a^2+2\varphi_{11n_1}\sigma_g^2 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{\pi}_{1n_11}\sigma_a^2+2\varphi_{1n_11}\sigma_g^2 & \cdots & \sigma_a^2+\sigma_g^2+\sigma_e^2 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \sigma_a^2+\sigma_g^2+\sigma_e^2 & \cdots & \bar{\pi}_{r1n_r}\sigma_a^2+2\varphi_{r1n_r}\sigma_g^2 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \bar{\pi}_{rn_r1}\sigma_a^2+2\varphi_{rn_r1}\sigma_g^2 & \cdots & \sigma_a^2+\sigma_g^2+\sigma_e^2 \end{pmatrix}.$$

This matrix has the same block diagonal matrix structure as the covariance matrices we have seen before, where each block is made up of the family-specific covariance matrix:

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega_1} & 0 & \cdots & 0 \\ 0 & \mathbf{\Omega_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{\Omega_r} \end{pmatrix}.$$

Here we need to estimate the three variance-components σ_a^2, σ_g^2 and σ_e^2. However, as before, this cannot be done using standard statistical software since such packages do not calculate and allow the IBD and kinship coefficients to be specified in the covariance matrix. So we again have the computational challenge we faced with the segregation model.

Model (30) is a general form of the linear mixed-effects models we have described for segregation analysis and familial aggregation. It is also a general form of most of the models used in practice. Although implemented by Abecasis et al. (2000a) in QTDT, simpler forms of this mixed-model methodology were applied to family data much earlier. An example is the paper by Amos (1994) which uses mixed-models that include kinship coefficients and IBD probabilities in the covariance matrices, in a similar way to what we have shown here. They however split the locus-specific random effect into three categories $(a, d, -a)$ depending on the individual's genotype. The variance components for a and $-a$ are extracted via IBD sharing, while the variance component for d is extracted by the probability of the pair sharing both alleles at that locus IBD. Therefore their model

contained four variance components. However, they could only apply their model to sib-pair data due to the computational limitations involved with calculating IBD sharing for other family pairs.

Almasy and Blangero (1998) extend the results of Amos (1994) to families of arbitrary size. However their research focuses on multipoint quantitative trait linkage, where several markers are simultaneously tested for linkage. This entails calculating multipoint relative-pair IBD sharing, which is complex for large families.

We can now use the variance estimates from Model (30) to estimate another type of heritability. The proportion of trait variance due specifically to the effects of alleles transmitted from parents to offspring, at a particular locus, is known as the *narrow-sense heritability*. It measures the degree to which a trait is passed from parent to child.
Narrow-sense heritability (h^2), which is generally recognised as the heritability of the trait at a specific locus, is defined as

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_g^2 + \sigma_e^2}.$$

$h^2 = 0$, implies that there is no heritability, i.e. variation is not due to alleles at that specific locus.
If $h^2 = 1$, then the trait is very heritable. Therefore all the variation is due alleles at that specific locus.

Although heritability is an important concept, it is also open to misinterpretation. As stated in Burton et al. (2005:946), "It is not about cause itself, but about the cause of variation in a particular trait in a particular population at a particular time." This is because the denominator, which is the total variance of the trait, is a mixture of the variance attributable to genes; to shared environment; residual variance due to unshared and unmeasured factors; as well as measurement error. So heritability for a given trait can vary greatly from one study population to another. In addition, adjustment for covariates also affects the interpretation of heritability. For example, adjusting for an important environmental covariate may decrease the total variance but leave the allelic

variance unchanged. So narrow-sense heritability decreases, giving the appearance that alleles account for more variation in the data than is actually true.

However, heritability is still calculated in practice as it can increase the analytical effect of a study by guiding the selection of the study population. In addition, high heritability can support further study into a trait's genetic determinants. If the trait is found to be heritable, all further research assumes heritability and researchers move on to analyses for locating causal alleles.

To conclude our investigations into linkage analysis, we recall that for single-gene disorders, model-based linkage methods are satisfactory, but these do not apply to complex diseases. Here, much more sophisticated methods such as variance-components analysis are needed. In addition, more sophisticated software is required as standard statistical software is not sufficient for such complex and specialised analyses.

In Chapter 7, we illustrate the analysis of the Heartdata and discuss the results. Through this analysis, the statistical theory we have presented up to this point will be reinforced and its practical application explained.

7 Practical Example

7.1 Data exploration

In this section we show how data from a published genetic study (Revera et al., 2007, Revera et al., 2008, Van der Merwe et al., 2008 and Heradien et al., 2009), which we have called the Heartdata, is analysed. This will be done in the systematic way discussed in the opening chapters of this dissertation. The aim of the analysis is to determine whether a specific marker affects a specific trait after adjusting for the presence/absence of known HCM-causing mutations, in other words, whether it modifies the effect of these mutations. The software packages we used for the analysis of the Heartdata are not the only ones available for this type of analysis. However, since this is not a software review, we will not discuss any packages other than those which we have used. These are among those listed in Table 8. The two quantitative traits analysed here are measures of

Table 8: Software list

Software	**Uniform Resource Locator (URL)**	**Page**
PEDSTATS	http://www.sph.umich.edu/csg/abecasis/PedStats/	111
QTDT	http://www.sph.umich.edu/csg/abecasis/QTDT/	10, 120
Simwalk	http://www.genetics.ucla.edu/software/simwalk/	121
Haploview	http://www.broadinstitute.org/haploview/haploview	116
Merlin	http://www.sph.umich.edu/csg/abecasis/merlin/	121
R-Kinship	http://CRAN.R-project.org/package=kinship	120
Cyrillic	http://www.cyrillicsoftware.com/	120
Prelude, Finale	Downloaded with QTDT and Merlin	121, 121
SOLAR	http://www.sfbr.org/solar/	85
MENDEL	http://www.genetics.ucla.edu/software/	85

left ventricular thickness: *LVMecho* is left ventricular mass in grams (g), as measured by echocardiography; and *cwtscore* is a composite measure of ventricular thickness, measured in millimeters (mm). Since these traits have skew distributions, they were transformed using quantile normalisation and are called *QLVMecho* and *Qcwtscore*, respectively. The quantitative covariates adjusted for are: *Age* (in years); body surface area (*BSA*) in square

meters (m^2); systolic blood pressure (*SystBP*) and diastolic blood pressure (*DiastBP*) as measured in millimeters of mercury (mmHg); and heart rate (*HR*) in beats per minute (bpm).

There are also several dichotomous covariates in the dataset: *Ethnicity* (1=Afrikaner, 2=Mixed ancestry); *Mutation* (which describes whether an individual carries a mutation or not and is coded 1=No, 2=Yes); and *W92*, *T797* (which identify families in which these two possible types of mutations segregate). The third type of mutation, *W403*, is assumed to be present in the absence of the other two. It is therefore redundant in the analysis so that the model becomes identifiable. Lastly, we have data for four genetic markers, called *Marker 11*; *Marker 12*; *Marker 13*; *Marker 14* respectively.

The opening lines of Chapter 3 state that the first step of any analysis is to explore the data to understand it and determine the appropriate method of analysis. The results presented in Chapter 3 are part of the exploratory analysis carried out on the Heartdata. Due to the unique components of this and other genetic datasets, namely genetic marker information as well as pedigree information, analysis of genetic datasets requires specialised software. As a result, the exploratory analysis on the Heartdata was carried out in a program called PEDSTATS (Wigginton and Abecasis, 2005). It is an open-source package created to carry out preliminary analyses on genetic datasets of various sizes.

As shown in Chapter 3, PEDSTATS produces many graphs that allow the exploration and summarisation of the data being analysed. It also reports summary information for all traits and covariate data. These reports include information for trait correlations between sibling pairs and other relative pairs. For dichotomous traits, PEDSTATS reports the proportion of individuals with the particular trait and provides details about the affected individuals. Table 9 gives a summary of the outputs obtained from PEDSTATS and presented here.

PEDSTATS is a command-line utility that is run with the command: **pedstats -d heart.dat -p heart.ped --age** *Age* **--pairs --hardyWeinberg --pdf**

Table 9: Summary of outputs from analysis

Output	Contents	Page
1	Pedigree structure	113
2	Age check	114
2	Hardy-Weinberg check using 34 unrelated individuals	115
4	Quantitative trait statistics	116
5	Covariate statistics	117
6	Pair statistics	118
7	Marker genotype statistics	120
8	IBD sharing for *Marker 13* using example pedigree	122
9	Heritability: *Qcwtscore*	125
10	Testing trait: *Qcwtscore*	126
11	Heritability: *QLVMecho*	128
12	Testing trait: *QLVMecho*	128
13	Linkage: *Qcwtscore*	130
14	Testing trait: *Qcwtscore (Marker 13)*	131
15	Linkage: *QLVMecho*	132
16	Testing trait: *QLVMecho (Marker 13)*	133

The *-d* selects the input data file, which is called *heart.dat*. It contains two columns. The first contains descriptions for the traits (denote by T), covariates (C), affection status (A) and marker genotypes (M). The second column contains the name of the variable that corresponds to the description in the first column. The first five columns of a genetic dataset are omitted from description in the *.dat* file as they always contain the same variables, in the same order. These are: *Family ID*, *Person ID*, *Dad ID*, *Mom ID* and *Sex*, as shown in Table 1.

The *-p* in the command selects the appropriate pedigree file, which contains family structure information, covariate and affection status information, as well as trait and genotype information. It is the file which contains all the data for analysis.

The *--age* selects the age-related variable from the dataset. The word following this command is the name of the variable which contains the age data. In this instance, that variable is just called *Age*.

--pairs specifies that counts must be done on all family pairs, namely, siblings, half-sibs,

parent-child pairs, grandparent-grandchild pairs, cousin pairs and avuncular pairs. In addition, the traits and covariates must be summarised in terms of counts and correlations, for each of the family pairs.

--hardyWeinberg instructs PEDSTATS to run Hardy-Weinberg tests on the data, both for all individuals in the pedigree file, as well as on a group of unrelated individuals which the program chooses from the study group. Since we have related individuals, the test on all individuals is not valid as HWE is based on unrelated people. Thus only the test on the unrelateds will be discussed here.

Finally, *--pdf* instructs the program to generate summary graphs, such as those presented in Chapter 3, and write them to an Adobe PDF file.

These input files and the command form the basis for analysis in other programs such as QTDT and Simwalk, which we discuss later.

Output 1 gives a summary of the pedigree information for the Heartdata example: the number of pedigrees in the study group; the number of individuals; and the distribution of the individuals within the families.

Output 1. Pedigree structure

Individuals:	507
Founders:	156 founders, 351 nonfounders
Gender:	257 females, 250 males
Families:	22
Family Sizes	
Average:	23.05 (5 to 95)
Distribution:	8 (13.6%), 5 (9.1%) and 13 (9.1%)
Generations	
Average:	3.41 (2 to 5)
Distribution:	3 (40.9%), 4 (36.4%) and 2 (13.6%)
Checking family connectedness ...	
... All individuals in each family are connected	

For the Heartdata, there are 507 individuals who come from 22 families. In total, there are

156 founders and 351 non-founders. There are also 257 females and 250 males. Following this, we have information for the average family size and the distribution of family sizes, as well as the average number of generations and the corresponding distribution. As noted previously, there are between 2 and 5 generations in the families here. Finally, PEDSTATS tells us that all of the individuals in each of the families are appropriately related to one another. Thus, a family-relatedness check is successful.

Next we do an age check on our individuals to make sure that there are no data entry errors. This is shown in Output 2, which corresponds to Figure 6. The outliers shown in Figure 6 are identified in Output 2.

Output 2. Age check

Checking for gaps less than 13.0 among relative pairs for covariate *Age* ...

In family F100, individual 3 has age 61.0.
However, 3 should have age at least 70.0, since 3 (61.0) is the mother of 5 (57.0).

Additionally,

3 (61.0) is the mother of 8 (51.0)
3 (61.0) is the mother of 7 (51.0)
3 (61.0) is the mother of 6 (55.0)

In family F101, individual 27 has age 45.0.
However, 27 should have age at least 50.0, since 27 (45.0) is the mother of 29 (37.0).

Additionally,

27 (45.0) is the mother of 30 (34.0)

In family F172, individual 6 has age 45.0.
However, 6 should have age at least 69.0, since 6 (45.0) is the father of 8 (56.0).

Additionally,

6 (45.0) is the father of 7 (42.0)
6 (45.0) is the father of 2 (52.0)

Checking for gaps greater than 70.0 among relative pairs for covariate *Age* ...
Checking for values that differ among twin pairs for covariate *Age* ...
Checking for differences greater than 30.0 in sibling values for covariate *Age* ...

The age checks in Output 2 identify some anomalies, which were discussed in Chapter

3. As mentioned there, it was found that the ages tested above were actually the age at which the individuals' hearts were tested and not their current ages, which explains the seemingly anomalous observations. As such, an age check is not appropriate for this data, but doing it highlighted some potential interpretation problems.

Among others, PEDSTATS ensures that the pedigree information is correct by checking the relatedness between family members (an individual cannot be his own father), checking Mendelian inheritance and by making sure that the sex-codes are consistent for each pedigree.

The following outputs give examples of such summary statistics produced for the Heart-data.

Output 3 gives the text output for the Hardy-Weinberg equilibrium test on *Marker 11*. It corresponds to Figure 5. In Output 3 we have the number of homozygous and het-

Output 3. Hardy-Weinberg check using 34 unrelated individuals

	N_Hom	N_Het	E_Het	N_Alleles	Alleles	P-value
Marker 11	19, 7 rare	15	16	2	3/1	0.7274 E

erozygous allele pairs for *Marker 11*, for 34 unrelated individuals, as well as the number of alleles present and what they are called. There are 19 homozygous pairs (7 of which are minor-allele pairs) and 15 heterozygous pairs, as shown in Figure 5. The third column gives the expected number of heterozygotes under HWE, which is 16. The fourth indicates that there are two alleles for the marker, while the fifth column tells us what these alleles are called (3/1). The last column gives the p-value for the HWE test and indicates that the HWE exact test (E) was run. Since the p-value (0.7274) is not significant, there is not sufficient evidence supporting a departure from Hardy-Weinberg Equilibrium for *Marker 11* for the selected unrelated individuals. So we fail to reject the null hypothesis of HWE.

For the Heartdata, all the markers tested were in HWE except for *Marker 12* which has 31

homozygous pairs, (one of which is the minor-allele pair) and 1 heterozygous allele pair. The expected number of heterozygotes is two, and the two alleles found here are alleles 1 and 2. The p-value for the HWE test was 0.0476, which is significant at the 5% level. Therefore, there is sufficient evidence to suggest that *Marker 12* deviates from HWE.

For HWE testing, some programs, such as Haploview, choose their unrelated individuals in a different way to PEDSTATS. In Haploview, the married-ins are chosen and are all included, but the unrelateds inside families are chosen differently for each run of the test. Hence the result differs with each run. PEDSTATS always chooses the most optimal group of unrelateds, so repeated runs on the same families should produce similar HWE test results.

In Output 4 we have the summary statistics for *LVMecho* and *cwtscore*, and their quantile normalised counterparts, *QLVMecho* and *Qcwtscore*, respectively. Columns 1 and 2 of

Output 4. Quantitative trait statistics

	All	Minimum	Maximum	Mean	Variance	Sibling correlation
LVMecho	187 (36.9%)	48.200	476.600	158.310	4608.852	0.110
cwtscore	176 (34.7%)	97.000	353.500	163.798	2377.505	0.088
QLVMecho	187 (36.9%)	-2.550	2.550	-0.001	0.958	0.127
Qcwtscore	176 (34.7%)	-2.530	2.530	-0.002	0.957	0.123

Output 4 give the names of the trait, the number of observations for this trait and the corresponding percentage (out of 507) in brackets. Columns 3–6 give the minimum value, maximum value, mean and sample variance for each trait. The last column gives the overall sibling correlation for each trait.

The two untransformed traits, *LVMecho* and cwtcore, have large ranges and variances. Once transformed though, they become symmetric and have means close to zero and variances close to one, as seen for the corresponding transformed traits *QLVMecho* and *Qcwtscore*. The sibling correlations for all of these traits are small and positive. Thus, when one siblings' measurement increases (or decreases), so does the corresponding measurement for the other sibling.

The covariates in the dataset can also be summarised as shown for the traits. The Heartdata contains several covariates and these are summarised in Output 5: As for

Output 5. Covariate statistics

	All traits	Minimum	Maximum	Mean	Variance	Sibling correlation
Ethnicity	507 (100.0%)	1.000	2.000	1.722	0.201	1.000
Age	331 (65.3%)	14.000	94.000	40.955	285.904	0.858
Mutation	355 (70.0%)	1.000	2.000	1.428	0.246	0.234
W92	507 (100.0%)	0.000	1.000	0.321	0.219	1.000
T797	507 (100.0%)	0.000	1.000	0.448	0.248	1.000
BSA	197 (38.9%)	1.300	2.500	1.826	0.056	0.138
SystBP	209 (41.2%)	90.000	230.000	124.091	389.073	0.221
DiastBP	209 (41.2%)	60.000	120.000	79.234	106.142	0.084
HR	201 (39.6%)	44.000	120.000	69.134	141.417	0.024

Output 4, the columns of Output 5 give respectively, the covariate name; number and percent of observations; the minimum; maximum; mean; variance; and sibling correlation for each covariate. *Ethnicity*, *Mutation*, *W92* and *T797* are all dichotomous variables, so we can interpret their means as percentages of individuals in the groups. For example, for *Mutation*, 42.8% of the study group harbour a mutation. For the specific mutation types, for *W92* the two categories are 0 and 1, where 1 indicates a presence of mutation *W92* in the family. So a mean of 0.321 implies that 32.1% of the individuals in the study group are from families harbouring mutation *W92* and 78% are from the other families; 44.8% from *T797* and the rest from *W403*.

For *Age*, the youngest family member is tested at age 14 and the eldest at age 94. The group mean age at which hearts are tested is 41 years, and the variance is 286 (SD=17 years). The sibling correlation is high (0.86), so sibings hearts were tested at around the same ages.

For body surface area (*BSA*), systolic BP, diastolic BP and heart rate (*HR*), information is only available for about 40% of the study group. In addition, all four have sibling correlations below 0.25. The mean systolic BP is 124 mmHg and the variance is 389 (SD=20 mmHg), while the mean diastolic BP is 79 mmHg and the variance is 106 (SD=10 mmHg). Thus, while the means for both are in the normal range, the systolic BP variance

is higher.

For both the traits and covariates, pairwise statistics can be produced. These are reported for all family pairs, as shown in Output 6.

Output 6. Pair statistics

Relative Pair Counts:

Sib-pairs:	519 pairs
Half-Sibs:	65 pairs
Cousins:	1047 pairs
Parent-Child:	702 pairs
Grandparent-Grandchild:	546 pairs
Avuncular:	985 pairs

Pair Correlations for Each Trait:

	Sib	HalfSib	Cousin	ParentChild	Grandparent	Avuncular
LVMecho	0.1103	-0.1004	0.0610	0.0417	0.3251	0.0613
cwtscore	0.0876	-0.1915	-0.0177	-0.0367	-0.0016	0.0828
QLVMecho	0.1271	-0.0595	0.0810	0.0422	0.1990	0.0770
Qcwtscore	0.1232	-0.1707	-0.0060	-0.0698	0.0966	0.1328

Pair Counts for Each Trait:

	Sib	HalfSib	Cousin	ParentChild	Grandparent	Avuncular
LVMecho	161	15	271	92	18	241
cwtscore	139	12	238	80	20	234
QLVMecho	161	15	271	92	18	241
Qcwtscore	139	12	238	80	20	234

Pair Correlations for Each Covariate:

	Sib	HalfSib	Cousin	ParentChild	Grandparent	Avuncular
Age	0.8579	0.1489	0.4390	0.7708	0.3068	0.7050
BSA	0.1381	-0.0888	0.0601	0.2642	0.3132	-0.0118
SystBP	0.2215	0.5049	0.3000	0.2250	-0.0009	0.1176
DiastBP	0.0844	0.1002	0.2446	0.1238	0.1345	-0.0309
HR	0.0236	0.3044	0.0235	0.0294	0.1378	0.0462

Pair Counts for Each Covariate:

	Sib	HalfSib	Cousin	ParentChild	Grandparent	Avuncular
Age	407	40	829	229	67	671
BSA	179	26	289	100	20	271
SystBP	202	29	307	110	26	307
DiastBP	202	29	307	110	26	307
HR	191	22	304	107	23	302

The first part of Output 6 reports the number of family pairs in the study group and then goes on to give the pairwise correlation and count results for the traits and quantitative

covariates, for each type of family pair. The Heartdata pedigree file consists of 519 sibling pairs, 65 half-sib pairs, 1047 cousin pairs, 702 parent-child pairs, 546 grandparent-grandchild pairs and 985 avuncular pairs. There is a great deal of overlap between pairs, as shown by, for example, there being more parent-child or sibling pairs than individuals in the study group.

If the trait correlations get smaller as the degree of relation between the family pair gets larger, the trait may be heritable. For the pair correlations for the traits, the sibling correlation for *LVMecho* is 0.11, for half-sibs it is -0.10, for cousins it is 0.06, 0.04 for parent-child pairs, 0.33 for grandparent-grandchild pairs and 0.06 for avuncular pairs. The negative correlation for half-sibs is likely to be the coincidence of an inaccurate estimate, based on a small number of pairs (65).

The next part of the table gives the count data for each type of relative pair, for each of the selected traits. The numbers are similar and, as expected, there are the same number of observations for *LVMecho* and *QLVMecho*, and for *cwtscore* and *Qcwtscore*.

For the paired covariate data, the correlations presented are for the quantitative covariates only as they are not defined for the dichotomous variables. For *Age*, the correlations, in descending order, are 0.86 for siblings, 0.77 for parent-child pairs, 0.71 for avuncular pairs, 0.44 for cousin pairs, 0.31 for grandparent-grandchild pairs and finally 0.15 for half-sibs. These correlations tell us about the ages at which the hearts of the pair were tested. So, for the sibling for example, the high correlation implies that siblings were tested at similar ages.

As with the traits, if a covariate is heritable, we expect the covariate correlation to decrease as the degree of relation between the family pair increases.

Lastly, we have Output 7 which gives the summary information for *Markers 12, 13, 14* and *15*. It shows that, for *Marker 12* for example, 251 out of the 507 individuals in the study group were genotyped, making up about 50% of the sample. Of these, 23 individuals

Output 7. Marker genotype statistics

	Genotypes	Founders	Heterozygosity
Marker 11	326 (64.3%)	26 (16.7%)	52.1%
Marker 12	251 (49.5%)	23 (14.7%)	9.2%
Marker 13	245 (48.3%)	20 (12.8%)	22.4%
Marker 14	295 (58.2%)	27 (17.3%)	44.1%

(14.7%) are founders and only 9.2% of the total number of genotyped individuals are heterozygous. The output for the other three markers is interpreted in the same way. The output for *Marker 14* corresponds to what was seen for Figure 11 in Chapter 3, where 44% of the genotyped individuals were heterozygous. As explained there, the proportion of heterozygous individuals determines marker informativity, and thus how useful a marker is for linkage analysis. Here, *Marker 12* and *Marker 13* have low heterozygosity and thus low informativity. *Marker 11* and *Marker 14* have reasonably high heterozygosity (52.1% and 44.1% respectively).

The summary statistics produced by PEDSTATS and given in Output 5, do not give useful text output for the dichotomous covariates. These can however be visually assessed via the graphical output produced by PEDSTATS, which is presented in Chapter 3, or by using other programs, such as R, to visually assess the data and create summary tables.

Finally, using a package such as Kinship (in R) or Cyrillic, we can draw a pedigree, such as those in Figures 3 and 17, for each of the 22 families.

Once we have explored our data, understood it and carried out any necessary cleaning and transforming, we can start the data analysis. To test heritability and do linkage analysis on the Heartdata, we used a command-line program called QTDT which, as mentioned previously, is one of several programs that may be used. QTDT is appropriate to our study as it applies variance-components methodology in the analysis of family data. It can be applied to quantitative traits, to all family types– from nuclear families to extended pedigrees– and it can run analysis both with and without parental genotypes. However QTDT cannot understand categorical covariates with more than two categories,

so categorical variables with more than two categories, such as our three types of mutation, can be accommodated by splitting them into dummy variables, as we did.

In order to do a linkage analysis on the Heartdata, QTDT requires a matrix of all the IBD probabilities for all family pairs in the sample, which it does not calculate. Simwalk and Merlin are two programs which can calculate IBD matrices. We used Simwalk as our families were too complex for Merlin.

Before Simwalk can be run, we have to run another program, called Prelude, which provides an interface between Simwalk and QTDT. It estimates allele frequencies and generates input files for Simwalk, which are used to calculate IBD probabilities. Prelude is run using the following command: **prelude -d heart.dat -p heart.ped -aa -t0.001**

The *-d* and *-p* specify the input files, as before. The *-aa* specifies that all the individuals should be considered for estimating the allele frequencies and the *-t0.001* specifies the estimated recombination fraction between the markers.

Once we have run this, the data is run through Simwalk, using: **simwalk2**
This creates separate text files with IBD probabilities for each of the 22 families. This is done as explained in Section 6.2 on IBD sharing.

Finally, all the information from the separate IBD files is gathered into one document using a program called Finale and the command: **finale IBD-01.***

This generated IBD file is automatically called *qtdt.ibd* and is used in further analysis. It contains the IBD probabilities for every family pair in our 22 families, for all of the markers under consideration. Output 8 is an example of the information this file contains. It gives us the IBD sharing for some members of one particular family, F100. Although the names are invented, the genotypes, which are for *Marker 13*, are not. They correspond to those in Figure 16 and the IBD sharing corresponds to the values in Table 4.

The first part of Output 8 gives identifies the family and gives the pedigree information

Output 8. IBD sharing for *Marker 13* using example pedigree

Example pedigree and corresponding genotypes

Family ID	ID	DID	MID	Sex	*Marker13*
F100	46 (Mom)	1	2	2	1 2
F100	47 (Dad)	0	0	1	1 2
F100	49 (Sue)	47	46	2	1 1
F100	50 (Jane)	47	46	2	1 1
F100	51 (Ryan)	0	0	1	0 0
F100	52 (Ally)	51	50	1	1 2

IBD sharing between family pairs:

Family ID	ID_j	ID_k	$\pi_{100,jk}(0)$	$\pi_{100,jk}(1)$	$\pi_{100,jk}(2)$
F100	47	46	1	0	0
F100	47	49	0	1	0
F100	47	50	0	1	0
F100	47	52	1	0	0
F100	46	49	0	1	0
F100	46	50	0	1	0
F100	46	52	0	1	0
F100	49	50	0	0	1
F100	49	52	0	1	0
F100	50	52	0	1	0
F100	51	46	1	0	0
F100	51	49	1	0	0
F100	51	50	1	0	0
F100	51	52	0	1	0

for the various family members, including their genotypes for *Marker 13*.

For the second part of Output 8, the first column identifies the family, while the second and third columns give the IDs of two individuals in that family. These are the two family members whose IBD sharing is shown in the remaining columns of the table. These last three columns give the probability of the two individuals sharing 0, 1 or 2 alleles IBD, respectively, for *Marker 13*. For example, if we consider individuals 47 (Dad) and 46 (Mom), they share 0 alleles IBD with probability 1, 1 allele IBD with probability 0 and 2 alleles IBD with probability 0. This is expected as they are unrelated. If we look at individuals 47 and 49 (the daughter of individual 47), then we see that they share 1 allele IBD with probability 1. This is also expected as they are a parent-child pair. The remaining rows are interpreted similarly. In this example, IBD sharing can be determined

exactly as we have all the genotype information we need.

Once we have all the IBD information for all the families in our study group, we can proceed with the data analysis.

7.2 Familial aggregation

As we have mentioned before, the first step in identifying a potentially heritable trait is establishing that the trait runs in families. Then, if it does, proceed to establishing whether the aggregation is partly due to shared genes or entirely due to shared environment. Establishing familiality does not require any familial data and testing it is not possible in QTDT. Since it is essentially a test for clustering, which can be carried out using a mixed-effects model and any statistical package, we can use the 'lme' function (for linear mixed-effects models) in R. To test family clustering we specify a between-family random effect in our model. This model was shown in the section on familial aggregation (Model (22)). In practice, covariates will also be adjusted for in this model. Once we have obtained the estimate for σ_b^2 through such a mixed-effects model, we can formally test for and use the intraclass correlation coefficient to investigate possible familial clustering. If this is established, we can continue on to the next step, which is testing heritability using the segregation analysis model. Since we know that HCM clusters in families, we will not go into detail on testing familial aggregation, but rather move on to testing segregation, heritability and linkage, which are the focus of this study.

7.3 Segregation analysis and heritability

To estimate broad-sense heritability we need to specify the segregation model, adjusting for covariates, then estimate the hereditary random effect σ_g^2 and use this estimate to calculate the broad-sense heritability. Heritability measures the inherited/genetic contribution to the variability of a trait. To assess it we run the Heartdata through QTDT using the command: **qtdt -d heart.dat -p heart.ped -a- -cus- -we -veg --p-values**

> **heritability.txt**

The *-a-* turns off the association model; we are not interested in association analysis here, just linkage. The *-cus-* option specifies the inclusion of all the covariates, as well as sex in the model. The *-we* specifies a model with only environmental variance (e) and compares it to *-veg*, which specifies a model with both hereditary genetic variance (g) and the environmental variance (e). The former specifies the variances for a null model while the latter specifies them for the full model. Including the hereditary genetic variance in the model results in QTDT calculating kinship coefficients for the families. These kinships are included as coefficients in the covariance matrices for each family, as explained in previous sections.

Finally, *--p-values* specifies that all the p-values are printed and not just the significant ones, which is the default option. The output from this analysis is written to a text file which is named by the command: > *heritability.txt*.

Model (25), the segregation analysis model, is used for assessing heritability for the Heart-data. In addition, $d = 10$ covariates are adjusted for in the model, as shown in Section 4.6. Table 10 lists these ten covariates and their corresponding coefficients. Recall that

Table 10: List of covariates and the interpretation of the effect sizes

Covariates		**Effects**	
μ	Overall mean		
$x_{1,ij}$	*Age*	α_1	Effect of being mixed ancestry
$x_{2,ij}$	*Ethnicity*	α_2	Effect of waiting one year to test individual j's heart
$x_{3,ij}$	*Mutation*	α_3	Effect of carrying a mutation
$x_{4,ij}$	*W92*	α_4	Effect of being in a family that carries mutation *W92*
$x_{5,ij}$	*T797*	α_5	Effect of being in a family that carries mutation *T797*
$x_{6,ij}$	*BSA*	α_6	Effect of one unit of *BSA*
$x_{7,ij}$	*SystBP*	α_7	Effect of one unit of systolic blood pressure (*SystBP*)
$x_{8,ij}$	*DiastBP*	α_8	Effect of one unit of diastolic blood pressure (*DiastBP*)
$x_{9,ij}$	*HR*	α_9	Effect of a unit of heart rate (*HR*)
$x_{10,ij}$	*Sex*	α_{10}	Effect of individual j being a female

$x_{2,ij}, x_{6,ij}, \ldots, x_{9,ij}$ are quantitative covariates while the remaining five are dichotomous.

The null hypothesis we test is $H_0 : \sigma_g^2 = 0$ and the alternative hypothesis, which specifies the full model, is $H_1 : \sigma_g^2 > 0$.

We can now run our segregation model on our traits, starting with *cwtscore*. However, recall that the distribution of *cwtscore* is very skew so it had to be transformed since inference based on this data would not be valid, as assumptions of normality are violated. Quantile normalisation was used to transform the data and the transformed observations are found in *Qcwtscore*, which we proceeded to analyse. Output 9 gives the null and full models consecutively.

Output 9. *Qcwtscore*

Null model:

df	Log(Likelihood)	Variances		Means	
162	183.8	σ_e^2	0.484	(Intercept)	-6.349
				Ethnicity:mixed ancestry	0.405
				Age	0.012
				Mutation	0.806
				W92	-0.079
				T797	0.351
				BSA	1.025
				SystBP	0.013
				DiastBP	-0.003
				HR	0.015
				*S*ex: female	-0.241

Full model:

df	Log(Likelihood)	Variances		Means	
161	180.9	σ_e^2	0.324	(Intercept)	-6.218
		σ_g^2	0.161	*Ethnicity*:mixed ancestry	0.303
				Age	0.016
				Mutation	0.830
				W92	-0.079
				T797	0.238
				BSA	1.064
				SystBP	0.011
				DiastBP	-0.002
				HR	0.015
				*S*ex: female	-0.247

From Output 9, we see that null model is based on 162 degrees of freedom and gives a

log-likelihood of 183.8, while the full model is based on 161 degrees of freedom and gives a log-likelihood of 180.9. The variance from the null model, 0.484, is split into two so that, for the full model: the environmental variance $\hat{\sigma}_e^2 = 0.324$ and the hereditary genetic variance $\hat{\sigma}_g^2 = 0.161$. These are, respectively, the portion of variance due to environment effects and the portion which is due to heritable genetic factors. Since the former is larger, most of the variation in the data is due to environmental factors.

Although QTDT estimates effect sizes of covariates, it does not give the standard errors of the estimates because this is not the aim of the package. Since those using it are primarily interested in testing variance components, QTDT is designed to carry out likelihood ratio tests on the variance components. Output 10 gives the result of the heritability test on *Qcwtscore*.

Output 10. Testing trait: *Qcwtscore*

Allele	df(0)	-Lnlk(0)	df(V)	-Lnlk(V)	Chisq	P-value	
N/A	162	183.8	161	180.9	5.88	0.015	(174 probands)

Output 10 tells us that there are no alleles used in the analysis, which is correct as this is a heritability analysis and therefore does not require any genotype information. Next it says that the estimates for the null model were calculated on 162 degrees of freedom, with a log-likelihood of 183.8. Following this, we see that the estimates for the full model were calculated on 161 degrees of freedom with a log-likelihood of 180.9. The chi-square value for the difference between the null and full models is based on likelihood ratios and the test value of 5.9 comes from 2(Lnlk(0)-Lnlk(V)). It gives a p-value of 0.015. Lastly, the output shows that the test was based on 174 probands. Since the p-value is significant at the 5% level, we reject the null hypothesis at this significance level. Hence, there is sufficient evidence to support the hypothesis that $H_1 : \sigma_g^2 > 0$. Therefore, in this study group, it appears that the trait *Qcwtscore* is heritable. The broad-sense heritability estimate for this data is obtained using the variance estimates from the full model, given in Output 9.

Thus, the broad-sense heritability is

$$H^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2} = \frac{0.161}{0.161 + 0.324} = 0.332.$$

Suppose now that our i^{th} family consists of the family members we have in our example pedigree from Figure 17, which we used for the covariance matrix of Model (26) in Section 5.3. Substituting the variance estimates from the full model of *Qcwtscore* into that covariance matrix gives us the following covariance matrix for our extended family, which consists of Dad, Mom, Sue, Jane, Ryan and Ally:

$$\mathbf{\Omega_i} = \begin{pmatrix} 0.485 & 0.000 & 0.081 & 0.081 & 0.000 & 0.040 \\ & 0.485 & 0.081 & 0.081 & 0.000 & 0.040 \\ & & 0.485 & 0.081 & 0.000 & 0.040 \\ & & & 0.485 & 0.000 & 0.040 \\ & & & & 0.485 & 0.081 \\ & & & & & 0.485 \end{pmatrix}.$$

The same analysis as above can be carried out for *QLVMecho* as *LVMecho* also had to be transformed using quantile normalisation. For the null and full models of *QLVMecho*, we turn to Output 11.

From Output 11, we see that the variance estimates, from the full model, are: $\hat{\sigma}_e^2 = 0.323$ for the environmental variance and $\hat{\sigma}_g^2 = 0.119$ for the hereditary genetic variance, adding up to the 0.44 in the null model.

Output 12 gives the result of the heritability test on *QLVMecho*. It tells us that the estimates for the null model were calculated on 172 degrees of freedom, with a log-likelihood of 185.7. Following this, we see that the estimates for the full model were calculated on 171 degrees of freedom with a log-likelihood of 183.7. The 2(Lnlk(0)-Lnlk(V)) chi-square value is 3.93. It gives a p-value of 0.048. Lastly, the output shows that the test was based on 184 probands. Since the p-value is significant at the 5% level, we reject the null hypothesis. So there is sufficient evidence to support the hypothesis that $H_1 : \sigma_g^2 > 0$. Therefore, in this study group, it appears that *QLVMecho* is heritable. The broad-sense

Output 11. *QLVMecho*

Null model:

df	Log(Likelihood)	Variances		Means	
172	185.7	σ_e^2	0.441	(Intercept)	-5.128
				Ethnicity:mixed ancestry	0.228
				Age	0.018
				Mutation	0.505
				W92	-0.146
				T797	0.132
				BSA	1.688
				SystBP	0.005
				DiastBP	-0.005
				HR	0.010
				*S*ex: female	-0.470

Full model:

df	Log(Likelihood)	Variances		Means	
171	183.7	σ_e^2	0.323	(Intercept)	-5.288
		σ_g^2	0.119	*Ethnicity*:mixed ancestry	0.193
				Age	0.020
				Mutation	0.533
				W92	-0.102
				T797	0.103
				BSA	1.803
				SystBP	0.004
				DiastBP	-0.006
				HR	0.009
				*S*ex: female	-0.448

Output 12. Testing trait: *QLVMecho*

Allele	df(0)	-Lnlk(0)	df(V)	-Lnlk(V)	Chisq	P-value	
N/A	172	185.7	171	183.7	3.93	0.048	(184 probands)

heritability estimate for this data is

$$H^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2} = \frac{0.119}{0.119 + 0.323} = 0.269.$$

7.4 Linkage analysis

Linkage analysis aims to assess whether there is evidence that genotypic similarities for relatives, at a specific marker, are independent of trait similarities. This is done through testing the locus-specific trait variance, σ_a^2, at a specific marker; in this example we present the results for *Marker 13* as it was the most significant of all the markers. To test the Heartdata for linkage, we run the following command in QTDT: **qtdt -d heart.dat -p heart.ped -i heart.ibd -a- -cus- -weg -vega --p-values > linkage.txt**

Here, the input data and pedigree files are specified as for segregation analysis. As before, the *-cus-* option specifies the inclusion of sex and all the covariates in the model. We use the segregation model as our null model; *-weg* specifies a null model with both environmental variance (e) and hereditary genetic variance (g). Through *-vega* we add the the locus-specific variance component for *Marker 13*, to the full model. It thus contains the hereditary genetic variance (g), environmental variance (e) and the locus-specific additive allelic variance (a). The output from this analysis is written to a text file called *linkage.txt*. Finally, the IBD file, which is required for this analysis, is specified using *-i*.

Here, since we are interested in testing for an locus-specific effect of *Marker 13*, the null hypothesis is $H_0 : \sigma_a^2 = 0$ and the alternative hypothesis $H_1 : \sigma_a^2 > 0$. As a result, for linkage analysis we use the genotype information for each individual. The null and full models for *Marker 13* for *Qcwtscore* are summarised in Output 13.

From Output 13 we see that the estimates for the variances, from the full model, are 0.342 for the environmental variance, 0.000 for the hereditary genetic variance and 0.134 for the allelic variance of *Marker 13*. Since the first is the largest, most of the variation in the data is due to environmental factors. The hereditary genetic variance is zero, implying that all the genetic variation for *Qcwtscore* is caused by *Marker 13*.

As with segregation analysis, we again see here how the null model variance splits in the full model, where there is a hereditary variance, an environmental variance, as well as

Output 13. *Qcwtscore (Marker 13)*

Null model:

df	Log(Likelihood)	Variances		Means	
161	180.9	σ_e^2	0.324	(Intercept)	-6.218
		σ_g^2	0.161	*Ethnicity*:mixed ancestry	0.303
				Age	0.016
				Mutation	0.830
				W92	-0.079
				T797	0.238
				BSA	1.064
				SystBP	0.011
				DiastBP	-0.002
				HR	0.015
				*S*ex: female	-0.247

Full model:

df	Log(Likelihood)	Variances		Means	
160	178.8	σ_e^2	0.342	(Intercept)	-6.048
		σ_g^2	0.000	*Ethnicity*:mixed ancestry	0.280
		σ_a^2	0.134	*Age*	0.017
				Mutation	0.848
				W92	-0.041
				T797	0.223
				BSA	1.062
				SystBP	0.010
				DiastBP	-0.003
				HR	0.015
				*S*ex: female	-0.244

an allelic variance. In this instance, the hereditary variance of the null model is split between the environmental and allelic variances in the full model, making the remaining hereditary genetic variance negligible in the full model. This is possible because of the different structures of our families; the families have different sizes and different IBD sharing between pairs, making the covariances differ between different families. As a result, the variance does not split into components as cleanly as we expect it to theoretically. Thus, for *Qcwtscore*, all of the genetic variation in the data is due to *Marker 13*, rather than some of it being due to the marker and some it being due to the remaining hereditary genetic variance.

Output 14 is generated by QTDT for the likelihood ratio test on the variances:

Output 14. Testing trait: *Qcwtscore (Marker 13)*

Allele	df(0)	-Lnlk(0)	df(V)	-Lnlk(V)	Chisq	P-value	
All	161	180.9	160	178.8	4.08	0.044	(174 probands)

Output 14 tells us that all the alleles were used in the analysis, which for our data means both alleles for *Marker 13*, as it is diallelic. Next it says that the estimates for the null model were calculated on 161 degrees of freedom, with a log-likelihood of 180.9. The estimates for the full model were calculated on 160 degrees of freedom with a log-likelihood of the 178.8. The chi-square value for the difference between the null and full models here is 4.08 and gives a p-value of 0.044. Lastly, the output shows that the test was based on 174 probands. Since the p-value is significant at the 5% level, we have sufficient evidence to reject the null hypothesis at this significance level. Hence, there is sufficient evidence favouring the alternative hypothesis, $H_1 : \sigma_a^2 > 0$. Therefore, for this study group, it appears that the trait *Qcwtscore* is linked to *Marker 13*.

Taking again our example pedigree from Figure 17, suppose that their IBD sharing and kinship coefficients in Table 4 are actually for *Marker 13*. Thus, using those values, we can illustrate what the covariance matrix, $\mathbf{\Omega_i}$, from Section 6.3 looks like for the six members Dad, Mom, Sue, Jane, Ryan and Ally, for *Marker 13*. Here we assume that Mom, Jane and Ally are 'affected' with the mutation W92. Substituting in the rounded-off values of the variance estimates, gives:

$$\begin{pmatrix} 0.476 & 0.000 & 0.067 & 0.067 & 0.000 & 0.000 \\ & 0.476 & 0.067 & 0.067 & 0.000 & 0.067 \\ & & 0.476 & 0.134 & 0.000 & 0.067 \\ & & & 0.476 & 0.000 & 0.067 \\ & & & & 0.476 & 0.067 \\ & & & & & 0.476 \end{pmatrix}.$$

By the definition given in previous chapters, and using the estimated variance components from the analysis on *Qcwtscore*, the narrow-sense heritability estimate is

$$h^2 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_g^2 + \hat{\sigma}_e^2} = \frac{0.134}{0.134 + 0.0.000 + 0.342} = 0.282.$$

We can carry out the same linkage analysis as above on *QLVMecho*, for *Marker 13*. Again the null model here is the same as the corresponding segregation model and the full model contains the locus-specific variance for *Marker 13*. The null and full models for *QLVMecho* are given in Output 15.

Output 15. *QLVMecho (Marker 13)*

Null model:

df	Log(Likelihood)	Variances		Means	
171	183.7	σ^2_e	0.323	(Intercept)	-5.288
		σ^2_g	0.119	*Ethnicity*:mixed ancestry	0.193
				Age	0.020
				Mutation	0.533
				W92	-0.102
				T797	0.103
				BSA	1.803
				SystBP	0.004
				DiastBP	-0.006
				HR	0.009
				Sex: female	-0.448

Full model:

df	Log(Likelihood)	Variances		Means	
170	182.8	σ^2_e	0.335	(Intercept)	-5.116
		σ^2_g	0.000	*Ethnicity*:mixed ancestry	0.177
		σ^2_a	0.106	*Age*	0.020
				Mutation	0.547
				W92	-0.100
				T797	0.066
				BSA	1.806
				SystBP	0.004
				DiastBP	-0.006
				HR	0.009
				Sex: female	-0.449

Output 15 shows that the estimates for the variances, from the full model, are 0.335 for the environmental variance, 0.000 for the hereditary genetic variance and 0.106 for the additive allelic variance. As with *Qcwtscore*, for *QLVMecho* the null model hereditary variance, $\sigma^2_g = 0.119$, is split up in the full model, between σ^2_e and σ^2_a. The full model hereditary variance is zero, implying that all the genetic variance in the model is due *Marker 13*. As with *Qcwtscore*, a third of the total variance for *QLVMecho* is due to the

marker, while the remainder is due to environmental factors.

Output 16 is generated by QTDT for the likelihood ratio test on the variances: It tells us

Output 16. Testing trait: *QLVMecho (Marker 13)*

Allele	df(0)	-Lnlk(0)	df(V)	-Lnlk(V)	Chisq	P-value	
All	171	183.7	170	182.8	1.91	0.167	(184 probands)

that both alleles for *Marker 13* were used in the analysis. Next it says that the estimates for the null model were calculated on 171 degrees of freedom, with a log-likelihood of 183.7. The estimates for the full model were calculated on 170 degrees of freedom with a log-likelihood of the 182.8. The chi-square value for the difference between the null and full models here is based on likelihood ratios and the test value is 1.91, which gives a p-value of 0.167. The test was based on 184 probands. Since the p-value is not significant at the 5% level, we do not have sufficient evidence to reject the null hypothesis at this significance level. Therefore, for *QLVMecho*, for *Marker 13*, there is not sufficient evidence in favour of $H_1 : \sigma_a^2 \neq 0$. So *QLVMecho* does not appear to have significant additive allelic effects from *Marker 13*, implying that there is no linkage to this marker, despite all the genetic variance being due to it.

In this chapter, we have used data from investigations into hypertrophic cardiomyopathy to illustrate the analysis of quantitative, extended pedigree, data. The most important and interesting concept that we have shown through the systematic anaysis here, is the splitting of variances into components.

In our segregation model for the two traits, we showed how the null model environmental (total) variance is split into two components in the full model. We achieved this split by using the kinship coefficients between family pairs to extract the hereditary genetic variance from the total variance.

For the linkage model of each trait, we used the trait's segregation model, with components σ_e^2 and σ_g^2, as the null model. In the full model, we split the hereditary variance into a

part which accounts for variation due to the alleles of a specific marker, *Marker 13*, and a shared environment part. The IBD sharing between family pairs was used to achieve this split. As a result, the linkage model contains three variance-components. Thus, in going from segregation to linkage, we went from a (null) model with one variance, to a segregation model with two variance-components, to a linkage model with three variance components.

8 Discussion

Genetic studies are usually carried out in order to identify the allele(s) responsible for the trait or disease we are investigating. Linkage analysis is a very important first step in the search for the causal allele(s) as it allows researchers to localise broad regions of a chromosome in which these alleles may lie. Linkage analysis is best for localising relatively rare alleles with high penetrances. These are usually single-gene traits that follow Mendelian inheritance patterns. Complex traits however, usually involve environmental effects and more than one locus, and are thus more difficult to analyze. Hence the causal allele(s) are more difficult to localise.

In this study, we systematically built up the variance-components linkage models in a way that the statistical theory could be understood. In addition, the models we build up are a general form of many of those used in practice. Historically, different methods of linkage analysis were developed as the need for them arose, so linkage models were usually developed for particular sets of data. As a result, the statistical methodology, particularly for model-free linkage using variance-components models, was not written up. This is what prompted this dissertation. The models we have described here are special types of mixed-models and are implemented in QTDT, but no literature exists which systematically explains the statistical theory underlying the software.

Suppose we are investigating hypertrophic cardiomyopathy– a cardiac muscle disease that is characterised by the thickening of the left ventricular wall of the heart. Suppose the trait we are interested in is left ventricular thickness. Consider the variance of left ventricular thickness: some people in our study group have hearts with normal sized left ventricles while others have hearts with unusually thick left ventricles. We want to know why this is so.

Some of the differences may be because ventricular thickness differs between individuals from different ethnic groups. We can remove the effect of ethnicity (adjust for it) by

putting it into our statistical model as a fixed effect (covariate). This enables us to model separate means for each ethnic group. We can then estimate the residual variance, which will be smaller than the unadjusted variance, since the variance due to the ethnicity is removed from it. So the parameters we can estimate from this model are the mean, ethnic effects and variance.

Maybe the family you belong to has an effect on ventricular thickness. We can then either model family in the same way as ethnicity, that is, as a fixed effect, or we can model it as a random effect. If we model it as a fixed effect, we will estimate an effect for each family, which we do not need. We also want to be able to generalise about other families and specify a parsimonious model. So we model family membership as a random effect. This saves us degrees of freedom and allows us to split the variance into a component that represents variation between families (family effect) and a component that represents variation between individuals inside a family (residual effect). We do this by splitting each observation into a mean, an ethnic effect, a family effect and a residual effect. Then, using an appropriate method, we obtain the "best" estimates of our parameters. The residual variance will again be smaller than before as we now also remove the family variance from the total variance. By testing the significance of the between-family variance component, we can determine if we have sufficient evidence that ethnicity aggregates in families.

Our next question pertains to the between-family variance- is it the result of environment factors (for example: mom's cooking is high in fat) or is it hereditary? To investigate this, we split the total variance into three components: hereditary variance; environmental between-family variance and residual variance. Specifically, we split the earlier between-family variance into heritable and non-heritable components. This is not straight-forward since we can't separate the two variances; they are confounded. However, we have information about the covariances between individuals and we can use this. For ventricle thickness, if the covariance between close relatives, siblings say, is higher than the covariance between cousins, and that covariance is larger than the covariance between more

distant relatives, then we have evidence of heritability. We can use this to split the heritable and non-heritable variances by assuming that the trait covariance between any pair of individuals is their kinship coefficient times the heritable/hereditary genetic variance. In this way, we can use the covariances of all available pairs to estimate the components of the between-family variance. To determine if ventricular thickness is hereditary, we can test the hereditary variance. If there is evidence of variation due to heritability, then we have detected the segregation of ventricular thickness in families.

After this, we want to know where on the genome the alleles that are causing ventricular thickness lie. We genotype the individuals in our study group at a series of linked markers in a candidate region, or even more commonly, genotype individuals at a series of markers all across the genome. Based on their genotypes at a particular marker, we calculate IBD probabilities for each pair of individuals. This is the probability of them sharing 0,1 or 2 alleles from common ancestors. We now want to split our hereditary variance into a general hereditary component and a component which explains the variation due to the specific locus. Using a similar argument to the one for segregation analysis, we again need the covariance between pairs of individuals to estimate this locus-specific additive allelic variance component. We specify the models so that, for the covariance of each pair of individuals, the expected proportion of alleles shared IBD (calculated from the IBD distribution for that pair) is the coefficient of the locus-specific variance. The idea now is that the more alleles a pair shares IBD, the stronger the correlation should be. Thus, the expected IBD coefficients will extract the portion of variance which is explained by the alleles shared at that locus. This variance can be estimated and tested to determine if the alleles at that locus contribute significantly to the variance of ventricular thickness. If it does contribute significantly, then we have detected linkage of ventricular thickness to that locus.

For single-gene disorders where the effects are larger, and due to crossovers, linkage analysis can identify smaller regions of interest. However, for complex diseases, linkage analysis can only identify large regions on a chromosome. These regions contain hundreds of loci,

many of which will be potential candidates for the trait under investigation. Thus it can be shown that linkage to a trait can be detected for markers over a relatively large area of the genome. This means that recombination must have occurred here, so linkage can be detected even if the true locus is not close to the marker. Since we are looking for a causal locus, we need to look at progressively narrower regions of the chromosome. As recombination is unlikely to occur here, linkage will no longer be informative. This is why other methods, functional as well as statistical, have to be used to localise the specific allele(s) affecting ventricular thickness. Specifically, tests of allelic association are used. Often, a family-based association analysis for a specific trait is carried out on regions of the genome which are first identified through linkage analysis.

9 References

Abecasis, G.R., Cookson W.O.C. & Cardon, L.R. (2000a). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*, 66: 279–292.

Abecasis, G.R., Cookson W.O.C. & Cardon, L.R. (2000b). Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics*, 8: 545–551.

Abecasis, G.R., Burt, R.A., Hall, D., Bochum, S., Doheny, K.F., Lundy, S.L., Torrington, M., Roos, J.L., Gogos, J.A. & Karayiorgou, M. (2004). Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *American Journal of Human Genetics*, 74: 403–417.

Almasy, L. & Blangero, J. (1998). Multipoint quantitative trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, 62: 1198–1211.

Amos, C.I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics*, 55: 535–543.

Burton, P.R., Tobin, M.D. & Hopper, J.L. (2005). Key concepts in genetic epidemiology. *Lancet*, 366: 941–951.

Cox, D.R. & Solomon, P.J. (2003). *Components of variance*. Boca Raton, FL: Chapman & Hall/CRC.

Davidian, M. (2005). Applied longitudinal data analysis.
Lecture notes from: http://www.stat.ncsu.edu/people/davidian/st732/notes. North Carolina State University, Department of Statistics.

Ellsworth, D.L. & Manolio, T.A. (1999). The emerging importance of genetics in epidemiologic research I, basic concepts in human genetics and laboratory technology. *Annals of*

Epidemiology, 9 (1): 1–16.

Elston, R. (1998). Methods of linkage analysis– and the assumptions underlying them. *American Journal of Human Genetics*, 63: 931–934.

Elston, R. (2004). Introductory Genetics for Statisticians. Weale, M. (ed.), *Genetic Epidemiology I: Fundamentals, Theory, Practice and Latest Developments*. London: The Biomedical & Life Sciences Collection, Henry Stewart Talks Ltd. (Online at http://hstalks.com/bio).

Falconer, D.S. (1989). *Introduction to quantitative genetics*, Third edition. Burnt Mill, Harlow, Essex: Longman Scientific & Technical.

Foulkes, A.S. (2009). *Applied statistical genetics with R*. New York: Springer Science+Business Media, LLC, 78–89.

Galton, F. (1877). Typical laws of heredity, *Nature*, 15: 492–495, 512–514, 532–533.

Heradien, M., Revera, M., van der Merwe, L., Goosen, A., Corfield, V.A., Brink, P.A., Mayosi, B.M. & Moolman-Smook, J.C. (2009). Abnormal blood pressure response to exercise occurs more frequently in hypertrophic cardiomyopathy patients with the R92W troponin T mutation than in those with myosin mutations. *Heart Rhythm*, 6 (11): S18–S24.

Hogg, R.V. & Tanis, E.A. (2006).*Probability and statistical inference*, Seventh edition. USA: Pearson Prentice Hall, 601–609.

Klug, W.S. & Cummings M.R. (2000). *Concepts of genetics*, Sixth edition. New Jersey: Prentice Hall Inc.

Palmer, L.J. (2005). Human genome project. Armitage P. & Colton T. (eds.), *Encyclopedia of Biostatistics*, Second Edition. West Sussex, England: John Wiley & Sons Ltd 4: 2456–2459.

Pratt, S.C., Daly, M.J. & Kruglyak, L. (2000). Exact multipoint quantitative-trait linkage

analysis in pedigrees by variance components, *American Journal of Human Genetics*, 66: 1153–1157.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Revera, M., van der Merwe, L., Heradien, M., Goosen, A., Corfield, V.A., Brink, P.A. & Moolman-Smook, J.C. (2007). Long-term follow-up of R403W(MYH7) and R92W(TNNT2) HCM families: mutations determine left ventricular dimensions but not wall thickness during disease progression. *Cardiovascular Journal of Africa*, 18 (3): 146–153.

Revera, M., van der Merwe, L., Heradien, M., Goosen, A., Corfield, V.A., Brink, P.A. & Moolman-Smook, J.C. (2008). Troponin T and β-myosin mutations have distinct cardiac functional effects in hypertrophic cardiomyopathy patients without hypertrophy. *Cardiovascular Research*, 77: 687–694.

Rothman, K.J., Greenland, S. & Lash, T.L. (2008). *Modern Epidemiology*, Third Edition. Philadelphia, USA: Lippincott Williams & Wilkins, 128–134.

Stone , S., Abkevich, V., Hunt, S.C., Gutin, A., Russell, D.L., Neff, C.D., Riley, R. et al. (2002). A major predisposition locus for severe obesity, at 4p15-p14. *American Journal of Human Genetics*, 70: 1459–1468.

Teare, M.D. & Barrett, J.H. (2005). Genetic linkage studies. *Lancet*, 366: 1036–1044.

Terwilliger, J.D. (2005). Linkage analysis, model-based. Armitage P. & Colton T. (eds.), *Encyclopedia of Biostatistics*, Second Edition. West Sussex, England: John Wiley & Sons Ltd 4: 2819–2831.

Thomas, D.C. (2004). *Statistical methods in genetic epidemiology.* New York: Oxford University Press Inc.

Van der Merwe, L., Cloete, R., Revera, M., Heradien, M., Goosen, A., Corfield, V.A., Brink, P.A. et al. (2008). Genetic variation in angiotensin-converting enzyme 2 gene is associated with extent of left ventricular hypertrophy in Hypertrophic Cardiomyopathy. *Human Genetics*, 124 (1): 57–61.

Watson, J.D. (1990). The human genome project: past, present, and future. *Science*, 248: 44–49.

Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. (2005). A note on exact tests of Hardy-Weinberg Equilibrium. *American Journal of Human Genetics*, 76: 887–893.

Wigginton, J.E. & Abecasis, G.R. (2005). PEDSTATS: descriptive statistics, graphics and quality assessments for gene mapping data. *Bioinformatics*, 21 (6): 3445–3447.

Xu, B., Woodroffe, A., Rodriguez-Murillo, L., Roos, J.L., van Rensburg, E.J., Abecasis, G.R., Gogos, J.A. & Karayiorgou, M. (2009). Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (39): 16746–16751.